Summer 2021

Exploring Biochemical Mechanisms with Hybrid Quantum Mechanics/Molecular Mechanics and Enhanced Sampling Methods

Edwin Enciso
Western Washington University, edwin.enciso11@gmail.com

Exploring Biochemical Mechanisms with Hybrid Quantum Mechanics/Molecular Mechanics and Enhanced Sampling Methods

By

Edwin Enciso

Accepted in Partial Completion
of the Requirements for the Degree
Master of Science

ADVISORY COMMITTEE

Dr. Jay McCarty, Chair

Dr. Jeanine Amacher

Dr. Tim Kowalczyk

GRADUATE SCHOOL

David L. Patrick, Dean

Master's Thesis

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Western Washington University, I grant to Western Washington University the non-exclusive royalty-free right to archive, reproduce, distribute, and display the thesis in any and all forms, including electronic format, via any digital library mechanisms maintained by WWU.

I represent and warrant this is my original work, and does not infringe or violate any rights of others. I warrant that I have obtained written permissions from the owner of any third party copyrighted material included in these files.

I acknowledge that I retain ownership rights to the copyright of this work, including but not limited to the right to use all or part of this work in future works, such as articles or books.

Library users are granted permission for individual, research and non-commercial reproduction of this work for educational purposes only. Any further digital posting of this document requires specific permission from the author.

Any copying or publication of this thesis for commercial purposes, or for financial gain, is not allowed without my written permission.

Edwin Enciso

7/16/2021

Exploring Biochemical Mechanisms with Hybrid Quantum Mechanics/Molecular Mechanics and Enhanced Sampling Methods

A Thesis
Presented to
The Faculty of
Western Washington University

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science

by
Edwin Enciso
July 2021

Abstract

In the field of molecular dynamics (MD), a long-standing issue is the time frame required in order to fully observe a chemical reaction. Enhanced sampling methods have been the primary way of overcoming this issue for the past 40 years. In this experiment our goal was to combine new and existing sampling methods in order to create an efficient and accurate way of retrieving kinetics data from simulations. In order to do this, we examined two test cases: the enzymes chorismate mutase and cytosine deaminase. We did this using hybrid quantum mechanics/molecular mechanics simulations coupled with enhanced sampling methods. The electronic structure of the quantum region is computed at the semiempirical level. We used metadynamics enhanced sampling method to study the reaction within the enzyme active site by biasing along the reaction coordinate. This allowed us to estimate the barrier heights and transition state coordinates. We also used the variationally enhanced sampling method to probe reaction kinetics. This latter approach suffered from convergence issues leading us to consider more advanced collective variables or improvements to how the bias is optimized.

Acknowledgements

I'd like to give thanks to :

Zach McGrew, who helped in properly compiling CP2K and PLUMED onto both the WWU cluster and the McCarty group's workstation.

Jeanine Amacher and Tim Kowalczyk, for their involvement in the thesis committee and their notes and advice on the proposal of this thesis.

Jay McCarty, whose work this thesis is based on, and for the time and energy they have invested into giving me guidance.

Table of Contents

List of Tables and Figures

1.0 Introduction

A long-standing issue in the field of molecular dynamics (MD) when attempting to simulate biochemical systems is the time scales required and the associated computational cost that they require.[1] In order to simulate these systems accurately with numerical methods we must consider the quickest degrees of freedom that a given biomolecule will exhibit.[2] These movements are often on the scale of femto- or picoseconds, while the time required for a reaction to complete can be on the order of minutes or longer. “Brute-force” simulations for one of these reactions would require an unreasonably large number of steps to complete while producing an overwhelming volume of data.

In the face of these issues, computational biochemists have utilized enhanced sampling methods. These methods typically employ an external bias, accelerating the transition between metastable states of the system. This improves the exploration of the system’s configuration space by amplifying natural fluctuations in the system, which are called, collective variables (CVs).[2] The identification of these CVs is an important task as well, as amplifying the fluctuations which lead to the reaction of interest is the crux of these methods. Fluctuations in unimportant degrees of freedom need to be avoided as they will waste time and lower computational efficiency and could lead to hysteresis. This is one of the major pitfalls of collective variable-based enhanced sampling methods.

1.1 Overview of Enhanced Sampling Methods

Enhanced sampling methods have been utilized and improved upon over the past few decades, with one of the most notable methods, umbrella sampling, being created in 1977 by Torrie & Valleu.[3] These methods can be generally categorized as either CV based methods or non-CV based methods. This thesis focuses on the former of the two. Umbrella sampling was the

first of these CV based methods and laid groundwork for the creation of metadynamics, a popular enhanced sampling method currently utilized in molecular dynamics.[4] In these methods the CVs are described as functions of atomic coordinates (**s(R)**). These CVs must be robust enough as to describe whatever process that's being examined. Data from MD simulations can be used to create a free energy surface or a configuration space which can describe the system's state in relation to the process of interest. The free energy surface is defined as the logarithm of the equilibrium distribution which would define the CVs, up to a constant.

$$F(s) = -\frac{1}{\beta}\log \int d\boldsymbol{R}\, \delta(s - s(\boldsymbol{R}))e^{-\beta U(\boldsymbol{R})} \quad (1)$$

Here, β is equal to $(k_B T)^{-1}$, and $U(\boldsymbol{R})$ is a potential energy function. In order to witness something like an unbinding or binding event, the CV would need to describe the system in the bound and unbound state. These two states would need to be well defined by the CV and separated enough that they can be differentiated. A simple CV might be the distance between the ligand's center of mass and the protein's center of mass. A more complex CV might utilize differences in charges or coordination numbers between the two states.

These states of interest will be separated by a potential energy barrier. Overcoming this barrier in the course of a molecular dynamics simulation is a rare event and can take a prohibitive amount of time. To overcome this timescale issue, a bias potential is introduced which aims to lower this barrier. The way this bias is utilized is dependent on the method, but in all cases the bias aims to lower the potential energy barrier by enhancing natural fluctuations in the system.[2]

1.2 Metadynamics

Metadynamics is one of the more popular enhanced sampling methods which has been modified and extended in numerous ways in order to create methods such as funnel

metadynamics or well-tempered metadynamics.[5–7] It has also been used in conjugation with other methods, like replica exchange models, or machine learning algorithms.[7,8,22] Metadynamics builds a history-dependent bias over iterations, and this building is dynamic, controlled by the set of variables in the following equation:

$$V_n(s) = V_{n-1}(s) + G(s, s_n)\exp\left(\frac{-1}{\gamma-1}\beta V_{n-1}(s_n)\right) \quad (2)$$

where $V_0 = 0$, $G(s,s_n)$ is a Gaussian kernel, and the exponent acts as a scaling factor, with the parameter γ being referred to as the bias factor.[7] It is also known that the scaling factor decreases with each iteration by 1/n, meaning the portion of the Gaussian added each iteration decreases in size.[6,9]

In order to pull any useful information out of these biased simulations, the sampled distributions must be reweighted to account for the added bias in order to get physical real-world results. First, examine the equilibrium distribution of any set of CVs:

$$P(s) = \int d\boldsymbol{R}\delta[s - s(\boldsymbol{R})]P(\boldsymbol{R}) = \langle\delta[s - s(\boldsymbol{R})]\rangle \quad (3)$$

Here, $P(\boldsymbol{R}) = \exp[-\beta U(\boldsymbol{R})]/Z$, the Boltzmann distribution which corresponds to the potential energy function $U(\boldsymbol{R})$, with $Z = \int dR\exp[-\beta U(R)]$ being the partition function.[7] When we add the bias function shown in Equ. 1, we get a new equilibrium distribution which follows:

$$P_V(s) = \int d\boldsymbol{R}\delta\big(s - s(\boldsymbol{R})\big)P_V(\boldsymbol{R}) = \frac{\exp\big(-\beta\big(F(s)+V(s)\big)\big)}{\int \boldsymbol{ds}\ \exp\big(-\beta\big(F(s)+V(s)\big)\big)} \quad (4)$$

where, $P_V(\boldsymbol{R}) = \frac{\exp(-\beta(U(\boldsymbol{R})+V(s(\boldsymbol{R}))))}{Z_V}$ is the modified Boltzmann distribution of the biased ensemble with a biased partition function $Z_V = \int d\boldsymbol{R}\ \exp[-\beta(U(\boldsymbol{R}) + V(s(\boldsymbol{R})))]$, and $F(\boldsymbol{s})$ is the unbiased free energy surface (FES). A remarkable feature of metadynamics is that it can be described asymptotically by an ordinary differential equation that has the solution in the limit of long time:

$$V(s,t) = -\left(1 - \frac{1}{\gamma}\right) F(s) + c(t) \quad (5)$$

where

$$c(t) = \frac{1}{\beta} \log \frac{\int ds\, e^{-\beta F(s)}}{\int ds\, e^{-\beta F[(s)+V(s,t)]}} \quad (6)$$

is a time dependent constant that is independent of **s**.[9,28] From here we can extract the FES up to a constant *as* $F(\mathbf{s}) = -(1/\beta) \log N_V(\mathbf{s}) - V(\mathbf{s})$, where $N_V(\mathbf{s})$ is a histogram which is built over the course of the biased simulation.

Due to metadynamics' ability to easily extract unbiased FES's from biased simulation. it has gained much popularity since its advent in 2002 by Laio and Parrinello.[4] The method is extremely scalable only being limited by the molecular dynamics methods and forcefields utilized in simulation. However, it has limitations like all methodologies, with one of the primary limitations being that metadynamics deposits bias indiscriminately.[10] Metadynamics does not differentiate a transition state from a stable state, and so it can deposit bias on transition state, effectively increasing the potential barrier which separates states. Due to this shortcoming, conventional metadynamics does not do a good job at estimating system kinetics or dynamics observables such as time correlation functions. The way one overcomes this in practice is to deposit a new Gaussian kernel infrequently, so-called infrequent metadynamics.[14] This method takes advantage of the separation of time scales between the time spent visiting local minima and the short time spent at the transition state. However, this reduction in the bias deposition rate limits the usefulness of this approach for *ab initio* or QM/MM methods which are more computationally intensive compared to conventional MD.

1.3 Variationally Enhanced Sampling (VES)

Recently, Variationally Enhanced Sampling (VES) was introduced and is similar in spirit to metadynamics. Both build an external bias which is dependent on the CVs chosen. The advantage that VES brings is that a bias is constructed on the fly to sample any target probability

distribution of interest. For QM/MM simulations, we can exploit the feature to design a bias which does not fill the free energy surface indiscriminately as metadynamics does. Instead, it only fills it to a predefined level which gives us more control over the magnitude by which we accelerate the thermodynamics of the system, along with ensuring that we have a transition state without bias. This allows us to obtain accurate kinetics from the biased simulation at a reduced computational cost relative to infrequent metadynamics.[10] It still has some of the limitations of past methods however, such as the importance in the selection of CVs. If good CVs are not chosen, then accurate barrier heights cannot be determined leading to incorrect predictions for the reaction rates.

VES utilizes a convex functional of Valsson and Parrinello, which when minimized, will provide a bias potential that samples according to a target probability distribution. In this work, we chose a target probability which will locally fill the FES up to a predetermined level.[11] The functional is as follows:

$$\Omega(V) = \frac{1}{\beta}\log\frac{\int d\boldsymbol{s}\exp[-\beta(F(s)+V(s)]}{\int d\boldsymbol{s}\exp[-\beta F(s)]} + \int d\boldsymbol{s}\, p(\boldsymbol{s})V(\boldsymbol{s}) \tag{4}$$

where *p(s)* is an arbitrarily chosen probability distribution which is normalized. Minimizing the functional with respect to *V(s)* tields the following relation which is valid up to an irrelevant constant:

$$V(s) = -F(\boldsymbol{s}) - \frac{1}{\beta}\log p(\boldsymbol{s}) \tag{5}$$

when $p(\boldsymbol{s}) \neq 0$ and $V(\boldsymbol{s}) = \infty$ otherwise. The arbitrary distribution, *p(s)* is important because at the minimum, the sampling of the CVs will follow the distribution ***p(s)***, allowing us to customize the sampling of the FES more effciently.[11] With this all, the variational form for ***V(s)*** is as follows:

$$V(s) = v(s)S(-v(s) - F_c) \tag{6}$$

where F_c is the predetermined level which we want to stop depositing bias at, ***v(s)*** is a function built up of basis set functions, like Legendre or Chebyshev polynomials, whose parameters are optimized in order to minimize $\Omega(V)$, and *S(x)* is a Fermi-type switching function of the form:

$$S(X) = \frac{1}{1+e^{\lambda X}} \tag{7}$$

where λ is a parameter that determines the speed at which the switching function goes to zero, with units of inverse energy. This switching function ensures that deposition occur only at locations on the FES where the FES itself is lower than the cutoff energy, F_c.[10]

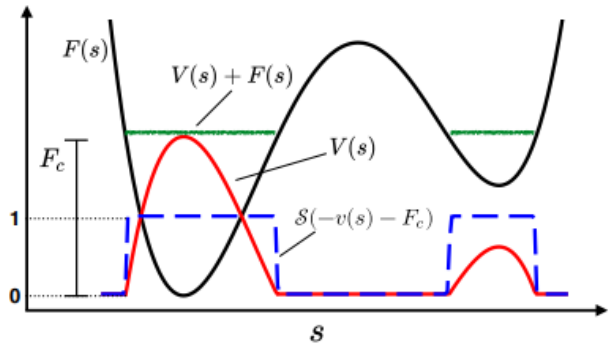

Fig 1. A free energy surface with one collective variable. The variables of eq. 6 are shown with, the switching function S(x) (dashed blue line), for the chosen cut off, F_c (green line). F(s) is the free energy surface, when added to the bias, V(s) (red line) we fill the FES up to the cut off. Taken with permission from "Variationally Optimized Free Energy Flooding for Rate Calculation" by McCarty et al., 2015. [10]

1.4 Project Goals

The aim of this thesis is to produce a novel methodology using a combination of existing methods which can extract information regarding the kinetics, mechanisms, and free energy surface of a biochemical system of interest in condensed phases at reduced computational cost. In my case I will be examining two biochemical systems using our methods: 1) chorismate mutase, and 2) cytosine deaminase.

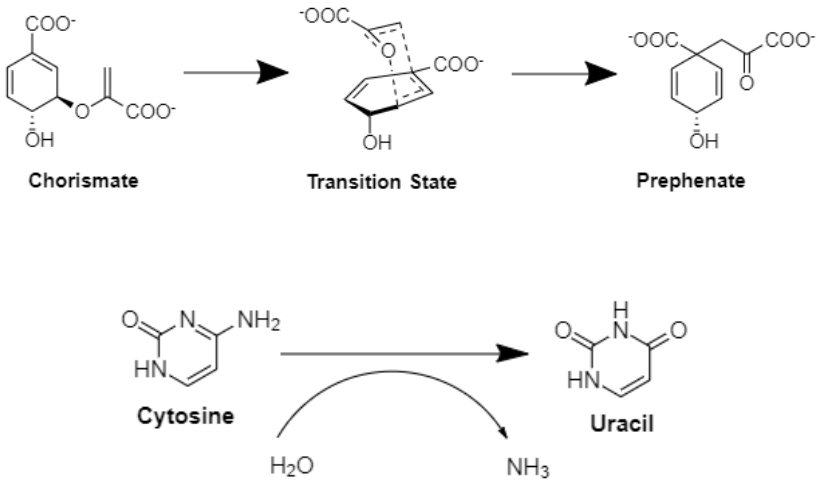

Fig 2. The reactions which chorismate mutase and cytosine deaminase catalyze, respectively.

These systems represent test-cases to develop and test new enhanced sampling methods. The methods which will be utilized are: 1) variationally enhanced sampling (VES), a enhanced sampling method which fills the free energy surface of system up to a predefined cutoff; 2) multi class harmonic linear discriminant analysis (MC-HLDA), a method of selecting important fluctuations in the system which to bias;[13,17] 3) parallel replica dynamics, a method of running multiple simulations in tandem such that they share information regarding their respective locations in the configuration space;[23] 4) QM/MM hybrid simulations. This form of molecular dynamics allows us to focus our computational resources into specific regions of the system, primarily the active site, by simulating these regions using quantum mechanics, while keeping costs low by simulating everything else using molecular mechanics. The methodology will be tested on two biochemical systems as the scalability of this method depends on the complexity of the system at hand and the CVs that can be utilized.

2.0 Chorismate Mutase

Chorismate mutase was selected as a model system as it has been very well studied in terms of MD simulations of the system.[12,16,19,20,25] It provided me with a starting point where the techniques I was interested in could be implemented more easily, as the system is simple. The enzyme does not covalently bond to the substrate, only a few residues need to be included in the QM region and it doesn't feature any metal ions as we will see in our second test case. Chorismate mutase has been used in numerous studies regarding QM/MM methods. Not only because of its simplicity, but of its importance in numerous biological contexts. I am specifically working with chorismate mutase from *bacillus subtilis*, a well-studied model organism for gram-positive bacteria. The reaction which chorismate mutase catalyzes is shown in fig. 2 above. It is a Claisen rearrangement in which chorismate interconverts to prephenate. This reaction is a part of the Shikimate pathway, which creates tyrosine from phosphoenol pyruvate and erythrose-4-phosphate. Fungi, bacteria, and plants use the pathway to build aromatic amino acids.[24]

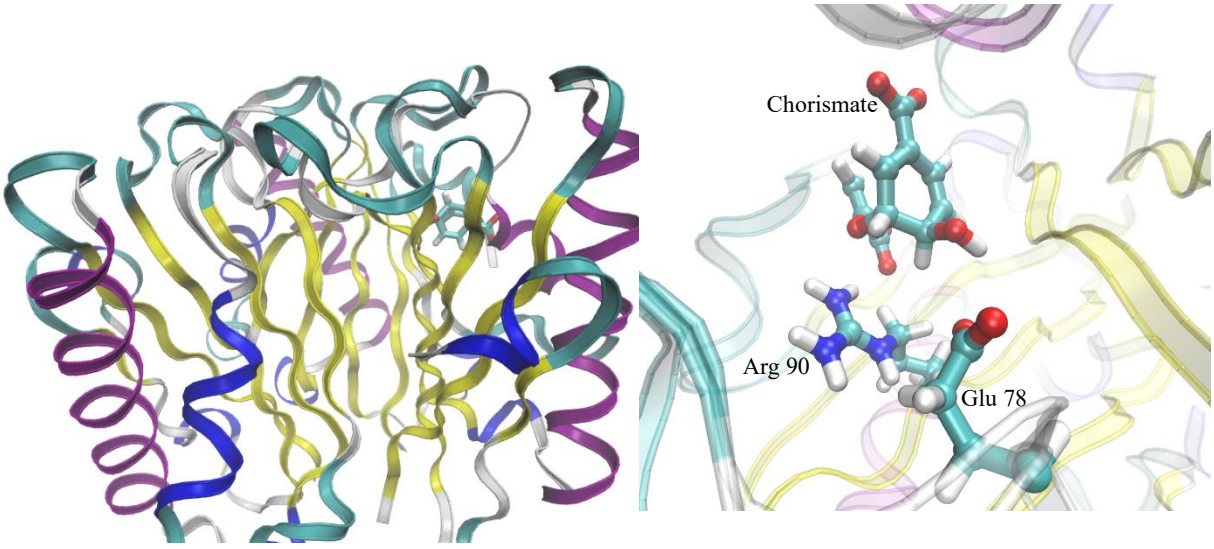

Fig. 3 On the left: Chorismate mutase's structure. On the right: The active site of chorismate mutase showing the substrate, chorismate, and the two residues added into the QM region, Arg 90 and Glu 78.

2.1 Method

The files for the system were obtained from the Protein Data Bank under the id of "2CHT". The coordinates of the ligand and the proteins themselves were merged using the ambertools software package.[29] Excess subunits and inhibitors were removed during this merging step. The system was placed inside a cube with side lengths of approximately 80 angstroms along with periodic boundaries. The system was then solvated by water molecules and neutralized via the addition of sodium ions, finishing up the setup needed for the molecular dynamics region. The MM region used the amber force fields for proteins, specifically 'ff19SB'[37], while the QM region utilized semi-empirical methods, specifically AM1, RM1 and PM6.

Once the topology files were prepared for the system, equilibration began with a short energy minimization using steepest descent in the CP2K software.[27] We then performed a 5 ps molecular dynamics run at a constant volume and temperature, the NVT ensemble, using the velocity rescaling thermostat.[30] This was followed by a 5 ps MD run at a constant pressure, the NPT ensemble, using the same thermostat along with the barostat of Martyna et al.[36]. Next, the QM region was delineated, first starting with just the substrate, via the CP2K input file. A simple CV was utilized, the difference between the length of the bond that needs to be broken, and the bond that needs to be formed.

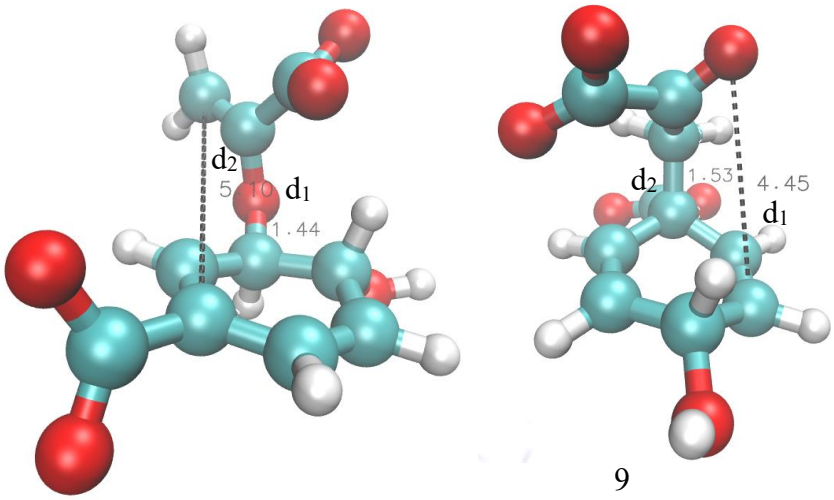

Fig 4. Chorismate (left) and prephenate (right) with the two bond lengths involved in the collective variable used for the simulations. The collective variable is the difference between these bonds: $d_1 - d_2$. The sign of the CV can tell us if we have chorismate or prephenate, as the CV should be positive if we have chorismate and negative for prephenate.

The associated input files for running the simulation in CP2K were modified from CP2K tutorials on their site.[27] The open-source, community-developed PLUMED library[31], version 2.6.0-dev[32], was used to setup the input files for the metadynamics and VES simulations. These input files are included in the supplementary material. The metadynamics simulations that will be discussed had the following parameters: they were ran for 60 ps, with a timestep of half a femtosecond. The electronic structure method utilized was the semiempirical PM6[33], after some brief benchmarking showed it gave us a reasonable tradeoff between accuracy and performance, compared to other methods we looked at, AM1 and RM1. We also investigated the effect of introducing important amino acid side chains into the QM region. The system had two residues, Arg 90 and Glu 78, in the quantum mechanical region along with the substrate. These two residues have been included in the QM/MM simulations of this system in the literature.[16,25] I was not aiming to get a converged free energy surface here, instead I am demonstrating that the CV can be used to sample the reactant state, product state and the transition state and provide an estimated height of the free energy barrier that makes the reaction a rare event. To that end I picked these residues as they've been shown to provide a good amount of stabilization to the transition state.[20]

2.2 Results & Discussion

Using CP2K and PLUMED, multiple molecular dynamics simulations were run using both metadynamics, and VES. These simulations were able to successfully replicate the reaction using three semi-empirical methods. With these simulations I was also able to create a path variable out of the coordinates from the successful simulations. A path variable is a CV which describes the configurations of the system from the start of a reaction to the end.[26] The system is pushed towards these configurations using the enhanced sampling method of choice. I was also

able to get some barrier heights from the metadynamics simulations, while not fully converged the barrier heights seem consistent with one another.

Metadynamics was initially used to gauge the performance of simulating the system on a high-performance GPU workstation (8 CPU threads and 1 Nvidia RTX 2080 GPU). We also made use of the Western CSE cluster. I tested three semi-empirical methods that have been used in other QM/MM studies into the system.[12,15,16,20] The three methods utilized and the free energy surface that was recreated from those simulations are shown below in fig. 4. Table 1 contains the pertinent parameters I used to deliberate which method would work best for us. An accurate, fully-converged, free energy surface from metadynamics requires many forward and reverse reactions to be sampled as the method is very slowly converging. The free energy is estimated from the metadynamics bias according to equation 2, where the bias potential is the sum of all the deposited Gaussian kernals. In this case I did not have enough sampling to present a fully converged free energy surface; however, this was not my primary goal. These simulations did not run for long enough as they were primarily used to gauge the stability and performance of these different methods and so these free energy surfaces are not representative of the true surface. However, the degree of agreement seen between the three surfaces certainly is an affirming result, as significant differences in these results due to a choice in a semi empirical method would mean that likely something went wrong in the setup of these files, most likely with allocation of electrostatic charges. The simplicity of the system should allow it to work with just about any semi-empirical method without issue, but this does not mean they all work equally as well. In the end, I decided upon PM6 as all the methods seemed to retrieve similar free energy surfaces, but PM6 had the fastest steps on average, reflected by the lowest CPU time per step, and crossed the potential energy barrier quicker than the others in terms of number of steps.

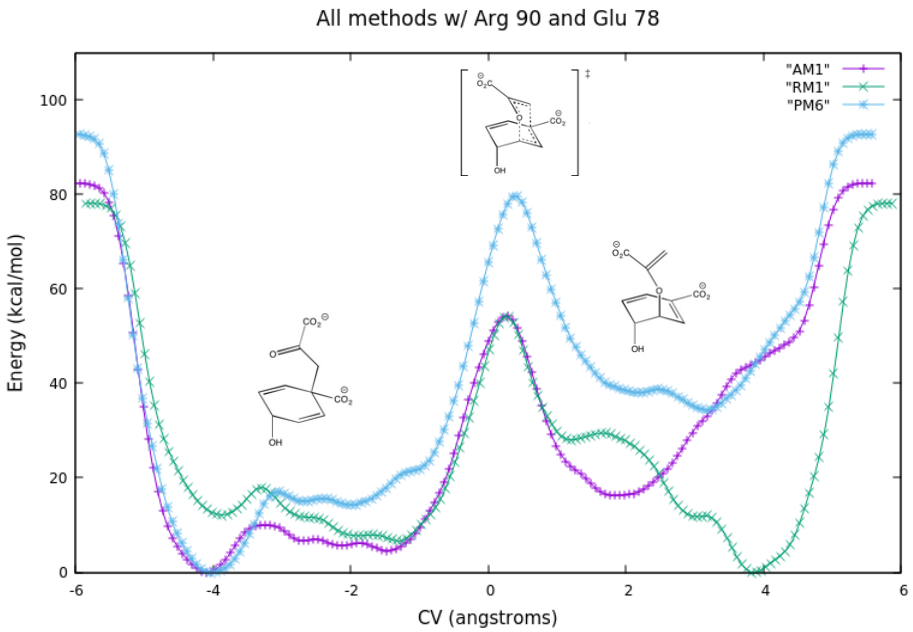

Fig 5. The free energy surfaces obtained from three different metadynamics simulations, each utilizing a different semi-empirical method. AM1 is shown in purple, RM1 in green and PM6 in blue. These simulations ran for 30 ps, with depositions of bias occurring every 50 steps. The residues Arg 90 and Glu 78 were added into the QM region in these simulations. The reactant, transition state and product are shown over their respective basin/peak.

Method	AM1	RM1	PM6
Barrier Height	~49 kcal/mol	~47 kcal/mol	~57 kcal/mol
Time to First Crossing Event	18.75 ps	13.31 ps	9.96 ps
Avg. CPU Time Per Step	9.14 s	6.58 s	6.33 s

Table 1. The results of the benchmarking of the various semiepiracal methods. The first crossing event is the simulation time taken in order for the system to cross the potential energy barrier and enter the next basin.

Along with this, more simulations were ran adding residues into the quantum mechanical region of the simulation along with the substrate. Arg 90 and Glu 78 were added to the QM region, shown in fig. 3 ,starting first with Arg 90. These residues were connected to the rest of the system via a link atom, typically the second carbon in the backbone of the residue. The link atom was created between the alpha carbon MM atom and the beta carbon QM atom using the Integrated Molecular Orbital and Molecular Mechanics (IMMOM) method as implemented by CP2K. These residues have been shown to create hydrogen bonds with the substrate in a manner which activate it, and so adding it into the QM region should allow us to see the decrease in

energy required to perform the reaction.[25] We see the largest drop when Arg 90 was first added, and much smaller drop when Glu 78 was added.

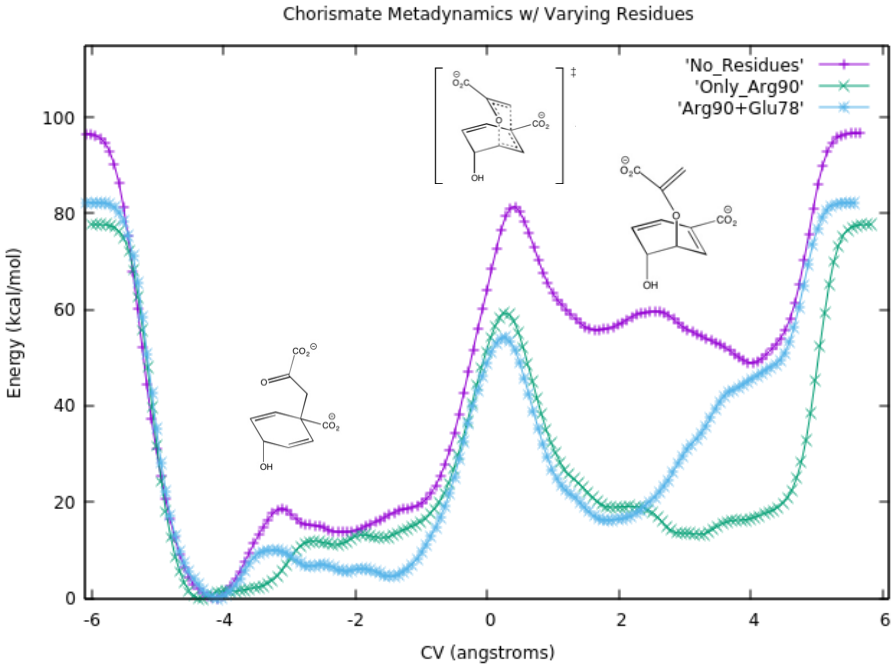

Fig 6. Free energy surfaces taken from metadynamics simulations with varying amounts of residues in the QM region. In purple is the fes for just the substrate. In green is the fes for substrate and the Arg90 residue. In blue is the fes for the substrate along with the Arg 90 and Glu 78 residue.

The stabilization of the transition state induced by the addition of residues into the QM region has been well-documented in other studies and most agree that Glu 78 and Arg 90 are the largest contributors to this effect.[25]

Along with this all, I used the metadynamics results to build a path variable out of the frames of the simulation which best represented the transition between the reactant and product. The path variable is simply a set of coordinates which will be used to bias the position of the atoms towards the coordinates in the set. These coordinates correspond to the process of interest, and will go from the start of the process to the end of it. The frames are chosen by finding sets of coordinates in which the RMSD between each set is kept equal between all sets, so that the transition between sets is smooth and realistic. The method provides two metrics that describe the progress of the simulation and allows one to check in without rendering the system in a

visualizing software.[26] For a more complex system, a path variable is a good tool as simple variables will usually be insufficient to get the reaction done, and using many variables increases the computational resources needed for the simulation. Path variables will not be needed in the case of chorismate mutase, however we plan on using them on cytosine deaminase.

Using VES, we were able to simulate the system and see multiple crossing events. More work still needs to be done in order to compute mean first passage times and to retrieve proper kinetics from these simulations. However, the preliminary work did provide a good starting system to work on before trying the method on cytosine deaminase. For these simulations a basis set of 20 Legendre polynomials were used to build the bias. The bias was written as:

$$V(s) = \sum_{i=0}^{20} \alpha_i P_i(s) \quad (8)$$

where α_i are the set of coefficients to be optimized by the variational procedure and $P_i(s)$ are the Legendre polynomials. The target distribution was a uniform distribution of 80 kcal/mol which spanned the phase space between -8 angstroms to 8 angstroms with respect to the CV, which was chosen after testing with an energy cutoff of 80 kcal/mol. The default value of the λ in the Fermi switching function given by equation 7 was kept at 10 kcal/mol. This value ensures a reasonably sharp cutoff in the probability distribution at the energy cutoff. First, we started by testing different cutoff energies, testing between 30 kcal/mol and 80 kcal/mol. We suspected the optimal energy cutoff would be somewhere between 60 kcal/mol and 80 kcal/mol, as this is roughly what we see in the metadynamics results. I ended up going with 80 kcal/mol so that we could get the initial crossing quickly and hopefully start seeing multiple crossover events. Since QM/MM simulations are more computationally intensive than MD, the basis set coefficients need to be updated frequently during the simulation. Here the coefficients are updated according to the recursion relation:

$$\alpha^{n+1} = \alpha^n - \mu[\Omega'(\bar{\alpha}) + \Omega''(\bar{\alpha})[\alpha^n - \bar{\alpha}]] \quad (9)$$

where μ is a fixed step size set to 0.3 and

$$\Omega' = \langle \frac{\partial V(s)}{\partial \alpha_i} \rangle_V - \langle \frac{\partial V(s)}{\partial \alpha_i} \rangle_p \quad (10)$$

where the first set of brackets indicates an average over the biased simulation and the second brackets an average over the target probability distribution. In this work the average over the biased simulation is accumulated over 100 steps (50 fs). This averaging is short because of the computational demands of QM/MM simulations. The short averaging window led to some issue regarding the manner in which the bias is averaged over the course of the simulation which will now be discussed in more detail.

I found that during the start of the simulation, there was very steady growth in the bias built up. In the top left of fig. 6, the bias grows steadily, only increasing in increments of about 10 kcal/mol. However, in the top right of fig 6., the bias between 60 ps and 90 ps is shown, and the bias changes extremely between 5 ps time frames. The changes do correspond with the position of CV according to the CV vs. time graph at the bottom of fig. 6. We see at 85 ps, that the system has moved into a new basin, one that lies between the CV values of 0 angstroms and -2 angstroms, and by 90 ps the system moved back towards positive CV values, which is reflected in the movement of the bias towards the positive side of the x axis.

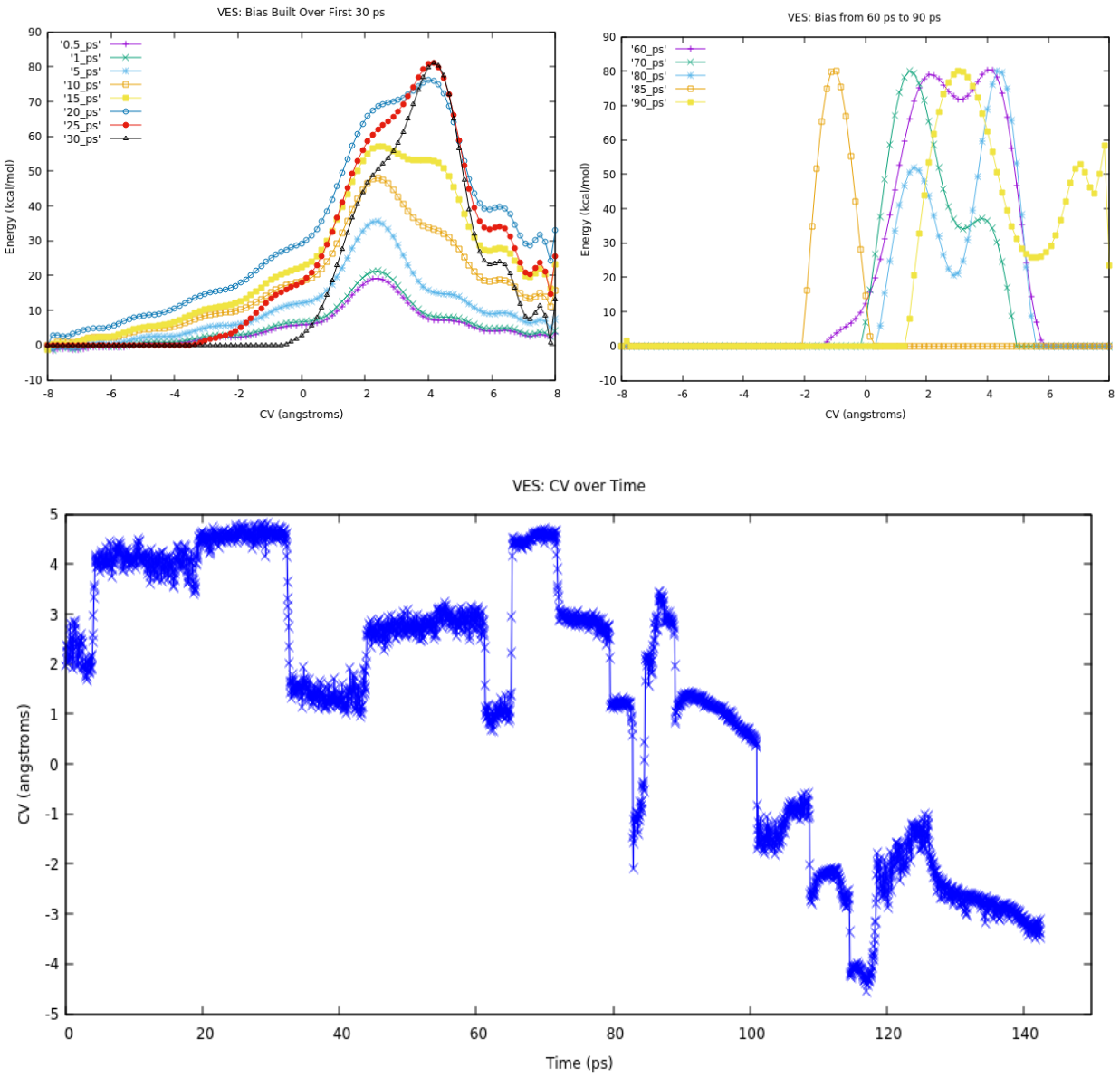

Fig 7. Preliminary VES results. On top left: The bias from the start of the simulation to 30 ps of simulation time. On top right: The bias from the 60 ps to 90 ps. On bottom: The value of the collective variable over the course of the simulation.

The erratic behavior of the bias from 60 – 90 ps is due to a lack of sampling during the short window over which the bias is updated variationally. Overall, I obtained a fairly good amount of data out of these simulations, but there is still some issue that needs to be fixed in the implementation of this method. At first all seems well, but as the time grows it seems that the bias acts more erratically.

2.3 Conclusions

In this chapter, I have presented a study of chorismate mutase using QM/MM simulations combined with two enhanced sampling methods. The first method being the widely used metadynamics. The results show that metadynamics is able to simulate the system and lead to efficient barrier crossings along the reaction coordinate with the choice of the collective variable being the difference of the bond lengths between the bond we would like the break and the bond would like to create. This led me to create a path variable that can be used to further bias the simulation along the reaction coordinate. Unfortunately, the slow convergence of metadynamics means that in the 60 ps of simulation time, only a small number of crossing events occurred, meaning the uncertainty in the free energy as computed by equation 5 is quite large. Nonetheless, the information obtained from these preliminary metadynamics simulations can be useful for comparing different semiempirical methods. Comparing three common semiempirical methods, AM1, RM1 and PM6, the estimated barrier heights are similar but PM6 transition occurred in less time, both in terms of simulation time and real-world time. It is well known that semi-empirical methods are less accurate then other methods such as DFT, but have a reduced computational cost. The estimated barrier heights from these semi-empirical methods can be compared to DFT calculations (possibly in vacuum) to estimate the accuracy of these methods for studying biochemical reactions.

2.4 Future Work

After all these results were retrieved and discussed, the focus of the project was shifted from chorismate mutase to cytosine deaminase to try the same methodology. The degree of complexity of cytosine deaminase in terms of active site features and reaction mechanism is much greater then chorismate mutase. Once that is done, focus will return to chorismate in order

to utilize parallel replica dynamics and MC-HLDA, which will again act as preliminary trial before trying the same method on cytosine deaminase. Once a more complex and optimal CV has been successfully implemented, then parallel replica dynamics can be used to further optimize the use of computational resources. After this benchmarking will be done for a final time to retrieve statistics regarding the speed of the simulation using varying computational resources.

Results indicate that we need to improve our implementation of VES to properly simulate the system and retrieve kinetics data. The primary issue I have been running into is the way that the bias is averaged between iterations. Early on it seems to build up bias well, but once the system moves into other areas of the phase space, the simulation seems to "forget" about previous bias that has been deposited. This ends up creating a situation in which the CV can fluctuate wildly as new bias is rapidly deposited in unexplored phase space and deconstructed in previously explored phase space. This is where multiple walkers could help us leverage the computational resources that we already have. We could start many walkers in different locations, have them gently build up bias in their respective starting location, and once one of them moves to another basin, we can stop the simulation, and move them back to their starting location. This could alleviate the issues being seen when moving into a new basin with VES, while still exploring a majority of the phase space.

3.0 Cytosine Deaminase

The work done on chorismate mutase was to test enhanced sampling methods (metadynamics and VES) on a simple reaction coordinate to speed up QM/MM simulations. Cytosine deaminase is a more complex system and has not been well-studied. Cytosine deaminase presented us with a major challenge in comparison to chorismate mutase. Cytosine deaminase is seen naturally in fungi and prokaryotes, making up a part of the pyrimidine salvage pathway.[18] In this thesis I am working with yeast cytosine deaminase specifically. Cytosine deaminase has also been shown more interest, as it is a candidate in anticancer gene therapy. Cytosine deaminase can produce 5-fluorouracil from the prodrug 5-fluorocytosine. 5-fluorocytosine is reasonably non-toxic to humans, while the product, 5-fluorouracil, is highly toxic to the gastrointestinal and hematopoietic systems. Producing the product in the tumor will minimize any unintended toxic effects of using 5-fluorouracil as a chemotherapeutic.[21]

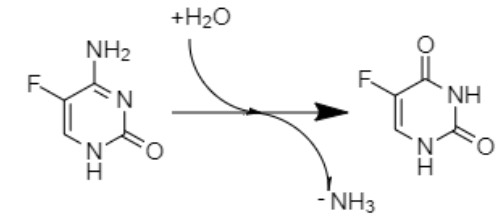

Fig 8. The deamination of 5-fluorocytosine to create 5-fluorouracil.

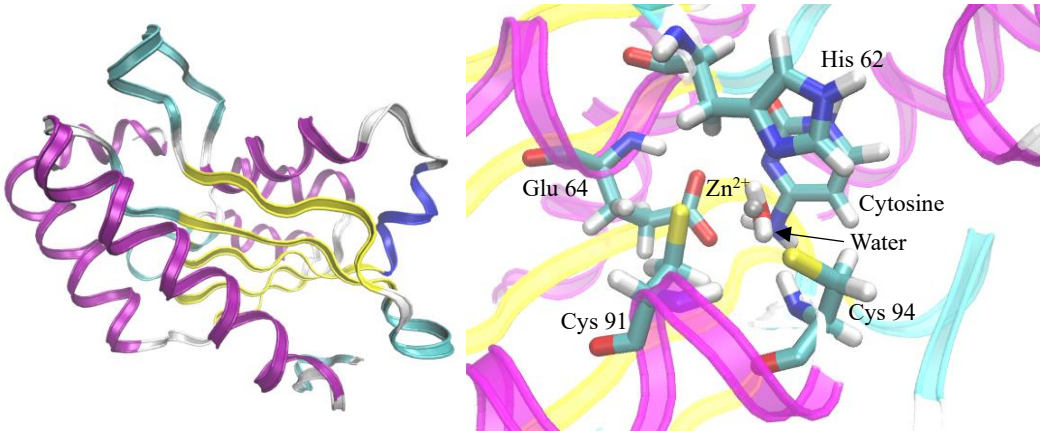

Fig 9. Structure (left) and active site (right) of cytosine deaminase, along with the residues and molecules being modeled by quantum mechanics.

The system is more complex in two aspects: first, the active site of cytosine deaminase contains more residues, which covalently bond to the substrate. There is a zinc ion present, which complicates the simulations as the semi-empirical methods we are using are not optimized for metal ions. Along with this, the mechanism of cytosine deaminase is more complex, there are more intermediate steps, and the individual steps of the mechanism will require different collective variables. The mechanism is shown in full in fig. 9, shown below. An understanding of the mechanism could lead to insights regarding protein engineering or mutational studies.

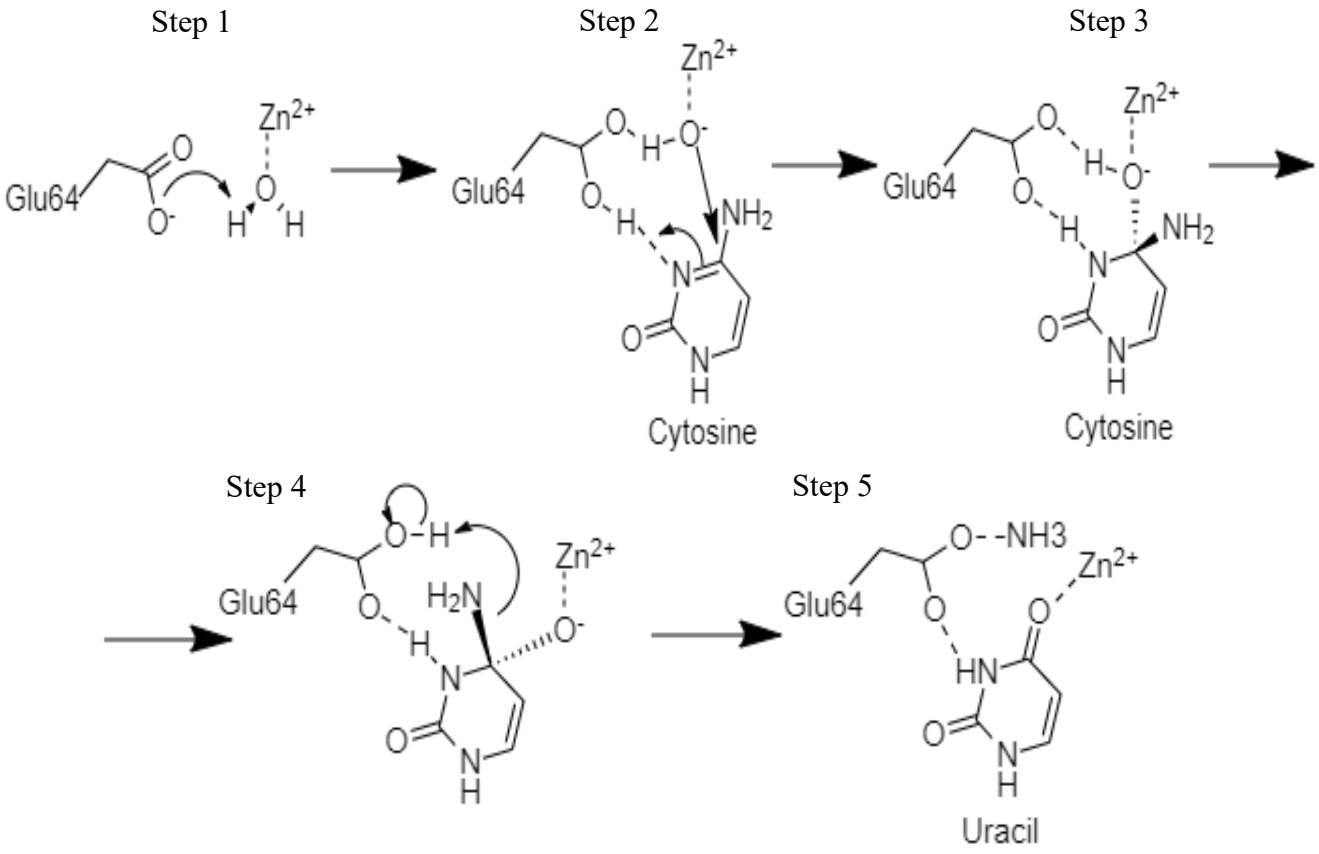

Fig 10. The full mechanism of cytosine deaminase. Adapted from "Crystal Structure of Yeast Cytosine Deaminase: Insights Into Enzyme Mechanism and Evolution." by Ko et al., 2003.[35]

3.1 Method

The files for this system were obtained from the protein data bank under the id '1P6O'. We removed one of the subunits leaving us with only a single active site and removed the extra molecules present. We combined it with the ligand, which was simply cytosine for the starting trials. It should be noted that yeast cytosine deaminase is a functional dimer, with a single active site per protein chain. Each of the active site contains a catalytic zinc coordinated by several residues and a water molecule. For computational investigation into the catalytic mechanism, I will only be considering a single protomer unit as it constitutes the smallest catalytic unit. Once the system was built, it was subjected to the same operations as the chorismate mutase system was, all the selections in forcefields and software was also the same. First, I performed solvation by water molecules and neutralization via sodium ions, and then a series of equilibration steps were done, and finally the delineation of the quantum mechanical region, which in the case of cytosine deaminase is much more complex in comparison to chorismate mutase. There are multiple residues which are included in the QM region: His 62, Cys 91, Cys 94, and Glu 64. As before, all residues included in the QM region were capped between the alpha carbon (in the MM region) and the beta carbon (in the QM region). These are all connected to the MM region via link atoms. Along with these residues the QM region also contains a water molecule, which is deprotonated in the starting steps of cytosine deaminase's mechanism, and a zinc atom, around which many of the nearby residues coordinate. A flowchart describing the process is shown below in fig. 11. We found that the system seemed to be most stable when using the PM6 method, and so we used it for all the simulations regarding cytosine deaminase.

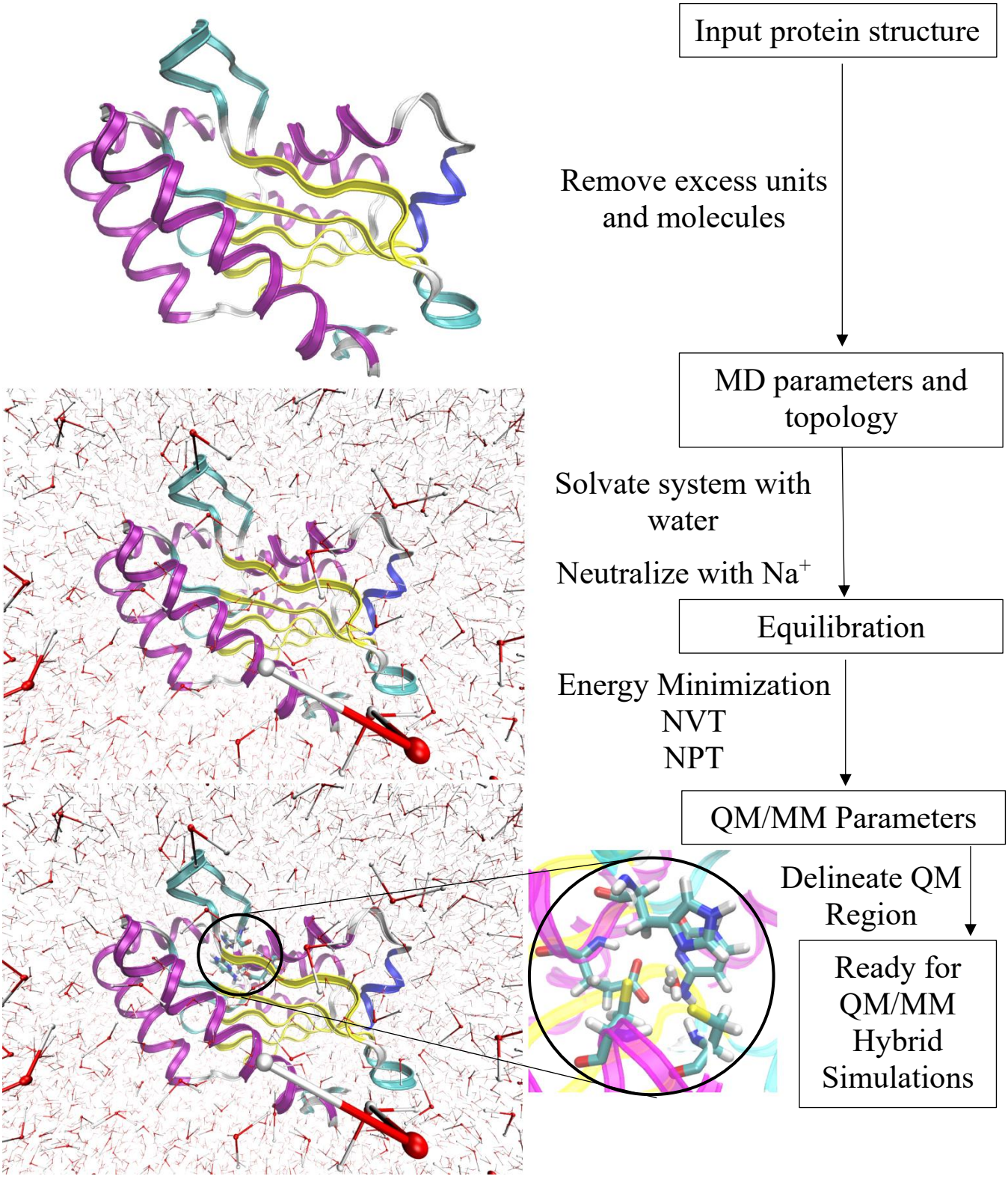

Fig 11. An overview of the basics of the full system's creation. First protein structures are gathered and edited to only contain the relevant systems. Next, solvation and neutralization are performed, and the basic MD equilibrations can be done once the MD parameters and topology are gathered. Finally, the QM/MM parameters are specified and the system is ready for QM/MM MD simulations.

The different steps of the mechanism of cytosine deaminase necessitated different CVs for each of the steps. I decided it would be best to take the mechanism one step at a time, and so I started first with the deprotonation of the water in the active site. For this step, a simple difference of bond lengths was used.

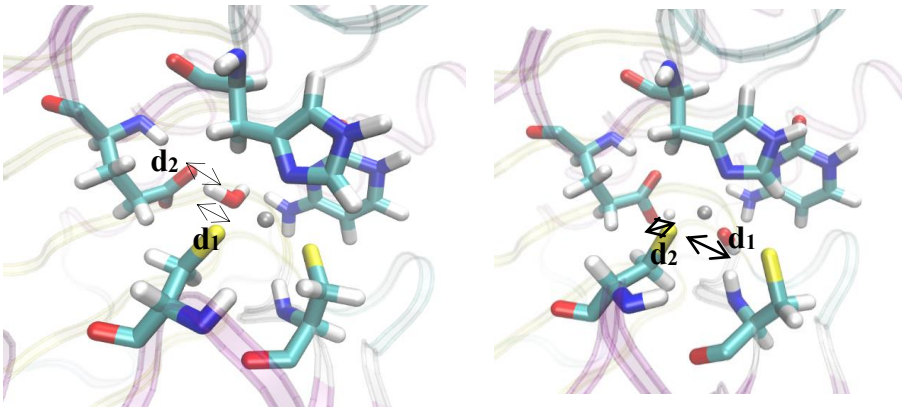

Fig 12. The bond lengths used for the first collective variables, the water molecule is deprotonated, and the proton is taken over to one of the oxygens on Glu 64. The collective variable is the difference of the two bond lengths: $d_1 - d_2$.

The next step involved getting the proton to transfer to the cytosine's unprotonated nitrogen. For this I utilized a combination of three CVs: first, the CV from the last step, second, a difference of bonds made up of the bond to be formed and the bond to be broken, and finally, the torsion angle of Glu 78's side chain.

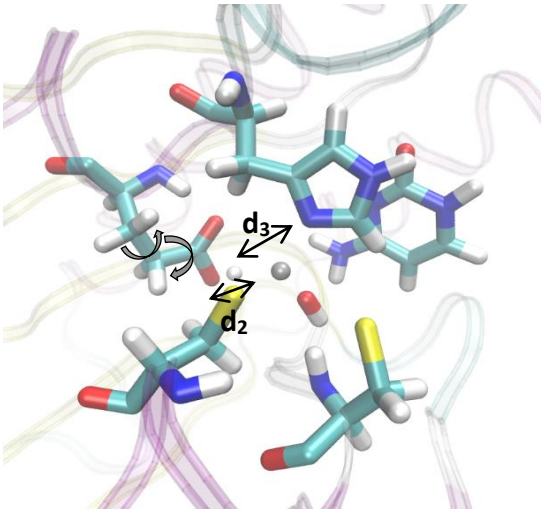

Fig 13. The bond lengths and angles used for the collective variables. One of the bond distances, d_2, is used again in the second collective variable along with the bond distance d_3. The difference of bond lengths used is: $d_2 - d_3$. Along with this we have two angles which are varied in order to adjust the torsion angle of Glu78's side chain.

I am still working with this CV at the time, as I have had trouble getting the side chain to rotate over towards the cytosine's nitrogen.

3.2 Results & Discussion

I am still working on simulating the full mechanism, but I have been able to gather data on the first two steps. The first step was simple enough to get done, as I only need to move a single proton, and bond it to an oxygen which was located fairly close to it. The difference in bond length collective variable we had been using on chorismate mutase works well here too.

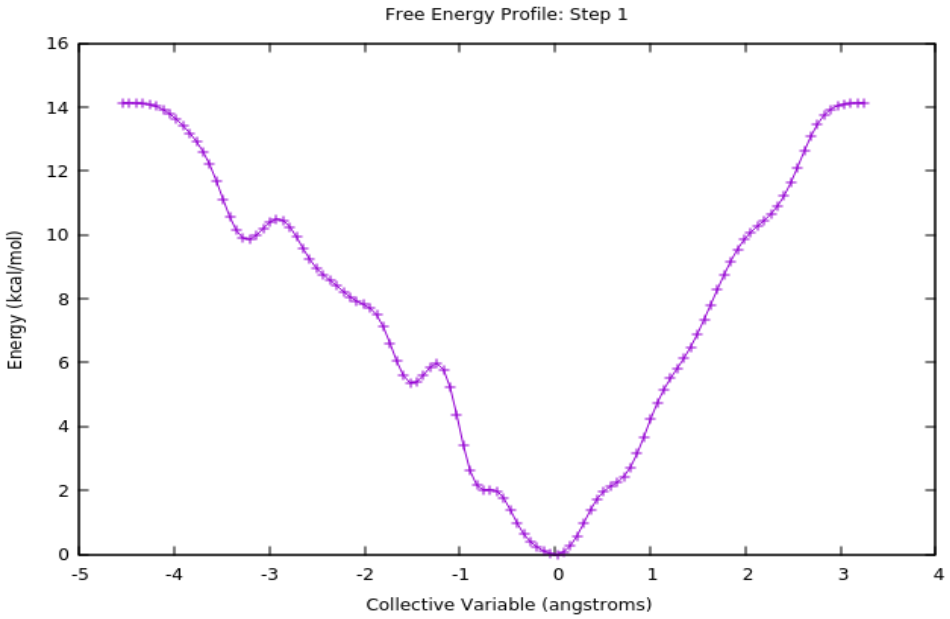

Fig 14. The free energy profile for the first step in the mechanism of cytosine deaminase. The collective variable here is the difference in the bond lengths between the bond we broke, between one of the water's protons and the oxygen, and the bond we would like to form, between the newly free proton and the oxygen present on the Glu64.

The free energy surface here is not converged of course but it provided me with an idea of the barrier I am trying to cross. In other literature, the size of this barrier has been reported as about 12 kcal/mol.[18] We see about a 10 kcal/mol barrier here, so I am not too far off from literature values. While this step was easily facilitated by a single variable, the next is not able to be done in a similar manner.

For the next step I needed to transfer the proton from the Glu 64 over to the unprotonated nitrogen on cytosine. This proved to be a bit more difficult than anticipated, as I ran into a few issues. I first tried using another difference of bond lengths, however I found that the proton

would return to the oxygen from which I split it from and pull the whole water molecule in tow towards the cytosine. I found that this happened often and so I introduced restraints to try and keep the distance between the two large enough to stop rebinding. These restrains would often cause issues and spikes in energy inside the QM region, and this might lead to unphysical results, so I moved onto other ideas. I kept the collective variable from the last step since I saw that without the CV, that the proton would return to the rest of the water molecule and take it in tow to the cytosine's nitrogen. In addition to this I added another collective variable, the torsion angle of Glu 78's side chain. I am still working with this collective variable at the moment.

3.3 Conclusions

In this chapter, I have performed metadynamics simulations of the first step in the conversion of cytosine to uracil in the active site of the cytosine deaminase enzyme. The primary challenge facing us is the creation of collective variable which can facilitate the individual steps of the mechanism. My strategy has been to break the reaction into the most elementary steps possible, in hopes of finding obvious parameters that would work as collective variables. In contrast to chorismate mutase, there is no clear reaction coordinate. The first step was successful and after a metadynamics simulation where bias was deposited along the reaction coordinate, I was able to transfer a proton from water to Glu 64, and I seemed to retrieve a reasonable barrier height which agrees with the literature value of ~10 kcal/mol. However, the product of this step is an unstable intermediate that quickly returns to the starting structure when bias is removed. Starting from this intermediate, I have been attempting to use three collective variables, but these simulations have typically crashed due to the addition of too much bias, or the reaction will simply not occur in a reasonable time. The latter issue being a result of the 3-D metadynamics which I have been attempting. One option would be to use the nudged elastic band[34] (NEB)

method in order to explore the reaction pathway and gain some insight into what a efficient collective variable might be. Regardless, trial and error will always be an option as we experiment with other collective variables.

3.4 Future Work

The most pressing goal will be to get the whole system simulated using coarse collective variables like those that have been shown. Once this is done, one can move onto making a path variable for it like the one made for chorismate mutase. This will allow data to be retrieved from the many metastable states of the reaction mechanism, which is needed in order to develop a more sophisticated CV that can facilitate the whole mechanism and allow the system to be simulated in one go. Somewhere in this process the hope is that parallel replica dynamics can be implemented in order to use computational resources more efficiently. Beyond this the second goal is to perform VES on the system. This will allow kinetics data to be retrieved for the individual steps of the reaction and will allow one to identify the rate-limiting step. This will still need to be worked out on the chorismate side of this study, but first simulating cytosine deaminase using metadynamics will help greatly. Metadynamics can provide us with a rough estimate of what our energy cutoffs should be. This will save us both time and computational resources that might have otherwise been spent guessing and checking different cutoff values.

4.0 Conclusion

In this thesis I have investigated the use of two enhanced sampling methods, the popular metadynamics, and the newer variationally enhanced sampling in their ability to do QM/MM simulations of biochemical systems. Both methods are dependent on a collective variable. As a starting point, I worked with chorismate mutase and investigated the interconversion of chorismate to prephenate. In this case, I utilized a simple CV, the difference in bond lengths between the bonds that need to be broken and formed (C-C bond and C-O bond shown in fig. 4). While, QM/MM simulations have great potential to study enzyme reactions, they are mainly limited by the computational cost required to run them. I tested a number of semiempirical methods in order to find a trade-off between accuracy and efficiency. The methods tested were the AM1, RM1 and PM6. All three methods give similar free-energy barrier heights, validating the methods used. The reaction occurred in the shortest amount of time using the PM6 method, both in terms of simulated time and real world time. It would be interesting to see how these methods compare to newer semi-empirical methods and to DFTB. From the successful reaction, I constructed a path CV of equally spaced frames along the reaction coordinate that can be used to sample the reaction along a specific path.

Metadynamics is a slowly converging process, and the bias potential did not converge during any of the 60 ps simulations. In order to get kinetics, I attempted to use the VES procedure to drive the system from reactant to product without any biasing of the transition state. The VES procedure uses a stochastic optimization procedure to construct the bias on the fly. Because of the computational cost of QM/MM simulations, the bias needs to be updated frequently, and this caused problems with the convergence of the bias from the stochastic

optimization procedure. This is a limitation in the VES method that needs to be solved before this method can be applied to QM/MM simulations and will be the focus of future work.

The final part of this thesis is the investigation of a more complex reaction: the conversion of cytosine to uracil in the yeast cytosine deaminase protein. The reaction involves multiple transition states and it is not intuitively obvious what a good collective variable to use would be. My strategy was to break the reaction into a series of steps, each with a clear CV. Using bond distances as a CV again, I was able to drive the protonation of Glu 64 by deprotonation of a water in the active site, using the metadynamics. The estimated barrier height of ~10 kcal/mol, agrees with the literature value of ~12 kcal/mol. Unfortunately, the intermediate state is unstable without the bias present, and it quickly returns to starting state. Several metadynamics simulations using different combinations of CVs were ran in attempts to reach the second transition state. Unfortunately, these simulations were unsuccessful, either due to slow convergence of metadynamics or due to the instability of the intermediate. A way forward would be to try an alternative approach, such as nudged elastic band (NEB) to sample a hypothetical raction coordinate.

Once the issues are dealt with I believe that these methods can be applicable to a number of studies. Mutational studies could utilize it to compare the kinetics of many iterations of enzymes. Systems in which CV choices are obvious will be perfect models for these methods, perhaps in more controlled environments like crystal lattices. There are many factors that come into play when trying to simulate these systems: the computational resources afforded to a team, the force fields and models that have been built over the history of this discipline, the software that is available and trusted, and the protein data that is provided to the community by structural biologists, along with many more. These all need to be taken account of and understood so that

one is not just working with an incomprehensible “black box” so to speak. These methods: VES, metadynamics, MC-HLDA and parallel replica dynamics can work in tandem. If these methods can be simplified and packaged into a set of programs then we believe it can be invaluable tool to researchers who may find MD methods inaccessible and to those who are experienced in the field.

Bibliography

(1) Voter, A. F.; Montalenti, F.; Germann, T. C. Extending the Time Scale in Atomistic Simulation of Materials. Annu. Rev. Mater. Res. 2002, 32 (1), 321–346. https://doi.org/10.1146/annurev.matsci.32.112601.141541.
(2) Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q. Enhanced Sampling in Molecular Dynamics. J. Chem. Phys. 2019, 151 (7), 070902. https://doi.org/10.1063/1.5109531.
(3) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. J. Comput. Phys. 1977, 23 (2), 187–199. https://doi.org/10.1016/0021-9991(77)90121-8.
(4) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. Proc. Natl. Acad. Sci. 2002, 99 (20), 12562–12566. https://doi.org/10.1073/pnas.202427399.
(5) Limongelli, V.; Bonomi, M.; Parrinello, M. Funnel Metadynamics as Accurate Binding Free-Energy Method. Proc. Natl. Acad. Sci. 2013, 110 (16), 6358–6363. https://doi.org/10.1073/pnas.1303186110.
(6) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. 2008. https://doi.org/10.1103/PhysRevLett.100.020603.
(7) Valsson, O.; Tiwary, P.; Parrinello, M. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. Annu. Rev. Phys. Chem. 2016, 67 (1), 159–184. https://doi.org/10.1146/annurev-physchem-040215-112229.
(8) Lamim Ribeiro, J. M.; Provasi, D.; Filizola, M. A Combination of Machine Learning and Infrequent Metadynamics to Efficiently Predict Kinetic Rates, Transition States, and Molecular Determinants of Drug Dissociation from G Protein-Coupled Receptors. J. Chem. Phys. 2020, 153 (12), 124105. https://doi.org/10.1063/5.0019100.
(9) Dama, J. F.; Parrinello, M.; Voth, G. A. Well-Tempered Metadynamics Converges Asymptotically. Phys. Rev. Lett. 2014, 112 (24), 240602. https://doi.org/10.1103/PhysRevLett.112.240602.
(10) McCarty, J.; Valsson, O.; Tiwary, P.; Parrinello, M. Variationally Optimized Free Energy Flooding for Rate Calculation. Phys. Rev. Lett. 2015, 115 (7), 070601. https://doi.org/10.1103/PhysRevLett.115.070601.
(11) Valsson, O.; Parrinello, M. A Variational Approach to Enhanced Sampling and Free Energy Calculations. Phys. Rev. Lett. 2014, 113 (9), 090601. https://doi.org/10.1103/PhysRevLett.113.090601.
(12) A Definitive Mechanism for Chorismate Mutase | Biochemistry https://pubs.acs.org/doi/10.1021/bi050886p (accessed Oct 10, 2020).
(13) Ponte, F.; Piccini, G.; Sicilia, E.; Parrinello, M. A Metadynamics Perspective on the Reduction Mechanism of the Pt(IV) Asplatin Prodrug. J. Comput. Chem. 2020, 41 (4), 290–294. https://doi.org/10.1002/jcc.26100.
(14) Accuracy of Molecular Simulation-Based Predictions of koff Values: A Metadynamics Study | The Journal of Physical Chemistry Letters https://pubs.acs.org/doi/10.1021/acs.jpclett.0c00999 (accessed Oct 10, 2020).
(15) Kříž, K.; Řezáč, J. Benchmarking of Semiempirical Quantum-Mechanical Methods on Systems Relevant to Computer-Aided Drug Design. J. Chem. Inf. Model. 2020, 60 (3), 1453–1460. https://doi.org/10.1021/acs.jcim.9b01171.

(16) Lee, Y. S.; Worthington, S. E.; Krauss, M.; Brooks, B. R. Reaction Mechanism of Chorismate Mutase Studied by the Combined Potentials of Quantum Mechanics and Molecular Mechanics. J. Phys. Chem. B 2002, 106 (46), 12059–12065. https://doi.org/10.1021/jp0268718.
(17) Mendels, D.; Piccini, G.; Parrinello, M. Collective Variables from Local Fluctuations. J. Phys. Chem. Lett. 2018, 9 (11), 2776–2781. https://doi.org/10.1021/acs.jpclett.8b00733.
(18) Zhang, X.; Zhao, Y.; Yan, H.; Cao, Z.; Mo, Y. Combined QM(DFT)/MM Molecular Dynamics Simulations of the Deamination of Cytosine by Yeast Cytosine Deaminase (YCD). J. Comput. Chem. 2016, 37 (13), 1163–1174. https://doi.org/10.1002/jcc.24306.
(19) Kangas, E.; Tidor, B. Electrostatic Complementarity at Ligand Binding Sites: Application to Chorismate Mutase. J. Phys. Chem. B 2001, 105 (4), 880–888. https://doi.org/10.1021/jp003449n.
(20) Ranaghan, K. E.; Ridder, L.; Szafczyk, B.; Sokalski, W. A.; Hermann, J. C.; Mulholland, A. J. Insights into Enzyme Catalysis from QM/MM Modelling: Transition State Stabilization in Chorismate Mutase. Mol. Phys. 2003, 101 (17), 2695–2714. https://doi.org/10.1080/00268970310001593286.
(21) Yao, L.; Li, Y.; Wu, Y.; Liu, A.; Yan, H. Product Release Is Rate-Limiting in the Activation of the Prodrug 5-Fluorocytosine by Yeast Cytosine Deaminase. Biochemistry 2005, 44 (15), 5940–5947. https://doi.org/10.1021/bi050095n.
(22) . Stanton, C. L.; Kuo, I.-F. W.; Mundy, C. J.; Laino, T.; Houk, K. N. QM/MM Metadynamics Study of the Direct Decarboxylation Mechanism for Orotidine-5‘-Monophosphate Decarboxylase Using Two Different QM Regions: Acceleration Too Small To Explain Rate of Enzyme Catalysis. J. Phys. Chem. B 2007, 111 (43), 12573–12581. https://doi.org/10.1021/jp074858n.
(23) Shea, J.-E.; Levine, Z. A. Studying the Early Stages of Protein Aggregation Using Replica Exchange Molecular Dynamics Simulations. Methods Mol. Biol. Clifton NJ 2016, 1345, 225–250. https://doi.org/10.1007/978-1-4939-2978-8_15.
(24) Rohr, J. Shikimic Acid. Metabolism and Metabolites. Von E. Haslam. Wiley, Chichester, 1993. 387 S., Geb. 75.00 £. – ISBN 0-471-93999-4. Angewandte Chemie 1995, 107 (5), 653–653. https://doi.org/10.1002/ange.19951070532.
(25) Lyne, P. D.; Mulholland, A. J.; Richards, W. G. Insights into Chorismate Mutase Catalysis from a Combined QM/MM Simulation of the Enzyme Reaction. J. Am. Chem. Soc. 1995, 117 (45), 11345–11350. https://doi.org/10.1021/ja00150a037.
(26) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in Free Energy Space. The Journal of Chemical Physics 2007, 126 (5), 054103. https://doi.org/10.1063/1.2432340.
(27) Kühne, T. D.; Iannuzzi, M.; Del Ben, M.; Rybkin, V. V.; Seewald, P.; Stein, F.; Laino, T.; Khaliullin, R. Z.; Schütt, O.; Schiffmann, F.; Golze, D.; Wilhelm, J.; Chulkov, S.; Bani-Hashemian, M. H.; Weber, V.; Borštnik, U.; Taillefumier, M.; Jakobovits, A. S.; Lazzaro, A.; Pabst, H.; Müller, T.; Schade, R.; Guidon, M.; Andermatt, S.; Holmberg, N.; Schenter, G. K.; Hehn, A.; Bussy, A.; Belleflamme, F.; Tabacchi, G.; Glöß, A.; Lass, M.; Bethune, I.; Mundy, C. J.; Plessl, C.; Watkins, M.; VandeVondele, J.; Krack, M.; Hutter, J. CP2K: An Electronic Structure and Molecular Dynamics Software Package - Quickstep: Efficient and Accurate Electronic Structure Calculations. J. Chem. Phys. 2020, 152 (19), 194103. https://doi.org/10.1063/5.0007045.
(28) Tiwary, P.; Dama, J. F.; Parrinello, M. A Perturbative Solution to Metadynamics Ordinary Differential Equation. J. Chem. Phys. 2015, 143 (23), 234112.https://doi.org/10.1063/1.4937945.
(29) D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H.

Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, C. Jin, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xue, D.M. York, S. Zhao, and P.A. Kollman (2021), Amber 2021, University of California, San Francisco.
(30) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. J. Chem. Phys. 2007, 126 (1), 014101. https://doi.org/10.1063/1.2408420.
(31) The PLUMED consortium. Promoting transparency and reproducibility in enhanced molecular simulations, Nat. Methods 16, 670 (2019)
(32) G.A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, G. Bussi. PLUMED2: New feathers for an old bird, Comp. Phys. Comm. 185, 604 (2014), preprint available as arXiv:1310.0980
(33) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J Mol Model* **2007**, *13* (12), 1173–1213. https://doi.org/10.1007/s00894-007-0233-4.
(34) Mills, G.; Jónsson, H. Quantum and Thermal Effects in H2 Dissociative Adsorption: Evaluation of Free Energy Barriers in Multidimensional Quantum Systems. *Phys. Rev. Lett.* **1994**, *72* (7), 1124–1127. https://doi.org/10.1103/PhysRevLett.72.1124.
(35) Ko, T.-P.; Lin, J.-J.; Hu, C.-Y.; Hsu, Y.-H.; Wang, A. H.-J.; Liaw, S.-H. Crystal Structure of Yeast Cytosine Deaminase: Insights Into Enzyme Mechanism and Evolution. *Journal of Biological Chemistry* **2003**, *278* (21), 19111–19117. https://doi.org/10.1074/jbc.M300874200.
(36) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant Pressure Molecular Dynamics Algorithms. *J. Chem. Phys.* **1994**, *101* (5), 4177–4189. https://doi.org/10.1063/1.467468.
(37) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J Chem Theory Comput* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

Supplementary Information

PLUMED Metadynamics input file

```
UNITS LENGTH=A TIME=fs ENERGY=kcal/mol #define units
d1: DISTANCE ATOMS=5663,5675 # define params used in colvar
d2: DISTANCE ATOMS=5670,5671
S1: COMBINE ARG=d1,d2 COEFFICIENTS=1,-1 PERIODIC=NO # define colvar
METAD ...
  LABEL=metd
  ARG=S1 # colvar used in metad
  SIGMA=.5 # width of gaussian
  HEIGHT=1.5 # height of gaussian
  PACE=100 # rate of deposition
  GRID_MIN=-12 # bounds of phase space defined by colvar
  GRID_MAX=12
  GRID_BIN=150 # number of bins in the phase space
  BIASFACTOR=18 # value of bias factor
  TEMP=298 # temperature of sim
... METAD

PRINT ARG=S1,d1,d2,metd.* FILE=COLVARS STRIDE=10 # params to be printed out and the rate
at which they are printed.
FLUSH STRIDE=1
```

PLUMED VES input file

```
# set units
UNITS LENGTH=A TIME=ps ENERGY=kcal/mol

# define colvars of sys
d1: DISTANCE ATOMS=5663,5675
d2: DISTANCE ATOMS=5670,5671
S1: COMBINE ARG=d1,d2 COEFFICIENTS=1,-1 PERIODIC=NO

# Set upper barrier to ensure the system gets pushed into the right basis
U1: UPPER_WALLS ARG=S1 AT=4.60 KAPPA=120.0

# Select basis set to use for building bias
bf1: BF_LEGENDRE ORDER=20 MINIMUM=-8 MAXIMUM=8

### HERE WE RUN VES
td: TD_UNIFORM
```

```
VES_LINEAR_EXPANSION ...
  ARG=S1
  BASIS_FUNCTIONS=bf1
  LABEL=variational
  TEMP=297
  BIAS_CUTOFF=80.0
  BIAS_CUTOFF_FERMI_LAMBDA=10.0
  TARGET_DISTRIBUTION=td
... VES_LINEAR_EXPANSION

OPT_AVERAGED_SGD ...
  RESTART=NO
  BIAS=variational
  STRIDE=100
  LABEL=var-S
  STEPSIZE=0.3
  COEFFS_FILE=coeffs.dat
  INITIAL_COEFFS=i_coeffs.dat
  BIAS_OUTPUT=10
  TARGETDIST_STRIDE=40
  TARGETDIST_OUTPUT=40
  COEFFS_OUTPUT=1
... OPT_AVERAGED_SGD

# Stop simulation after the transition
#COMMITTOR ARG=S1 BASIN_LL1=4.0 BASIN_UL1=5.0 STRIDE=600

### FINALLY PRINT EVERYTHING
PRINT FILE=COLVAR ARG=S1,d1,d2,variational.*,U1.* STRIDE=100
```

Metadynamics input file: Cytosine Deaminase, Global and Force Evaluation Sections

```
&GLOBAL
 PROJECT MONITOR
 PRINT_LEVEL LOW
 RUN_TYPE MD
&END GLOBAL

&FORCE_EVAL
 METHOD QMMM
 STRESS_TENSOR ANALYTICAL
 &DFT
  CHARGE -1
  &QS
   METHOD PM6
   &SE
    &COULOMB
     CUTOFF [angstrom] 10.0
    &END
    &EXCHANGE
     CUTOFF [angstrom] 10.0
    &END
   &END
  &END QS
  &SCF
   MAX_SCF 30
   EPS_SCF 1.0E-6
   SCF_GUESS ATOMIC
   &OT
    MINIMIZER DIIS
    PRECONDITIONER FULL_SINGLE_INVERSE
   &END
   &OUTER_SCF
    EPS_SCF 1.0E-6
    MAX_SCF 10
   &END
   &PRINT
    &RESTART OFF
    &END
    &RESTART_HISTORY OFF
    &END
   &END
  &END SCF
 &END DFT
 &MM
  &FORCEFIELD
   PARMTYPE AMBER
   PARM_FILE_NAME system_qm-charge.parm7
   &SPLINE
    EMAX_SPLINE 1.0E8
    RCUT_NB [angstrom] 10
   &END SPLINE
  &END FORCEFIELD
  &POISSON
   &EWALD
    EWALD_TYPE PME
    ALPHA .40
    GMAX 80
   &END EWALD
  &END POISSON
 &END MM
 &SUBSYS
  &CELL
   ABC [angstrom] 73.6556960  73.5178610  73.8039010
   ALPHA_BETA_GAMMA 90 90 90
  &END CELL
  &TOPOLOGY
   CONN_FILE_FORMAT AMBER
   CONN_FILE_NAME system_qm-charge.parm7
  &END TOPOLOGY
   !NA+ is not recognized by CP2K, so it is necessary to define it here using KIND
 &KIND NA+
   ELEMENT Na
  &END KIND
  &KIND NS
   ELEMENT N
  &END KIND
  &KIND NS3
   ELEMENT N
  &END KIND
  &KIND NS4
   ELEMENT N
  &END KIND
  &KIND NS1
   ELEMENT N
  &END KIND
 &KIND NS2
   ELEMENT N
  &END KIND
  &KIND ZN2+
   ELEMENT ZN
  &END KIND
  &KIND NV
   ELEMENT N
  &END KIND
  &COLVAR
   &DISTANCE_FUNCTION
    COEFFICIENT -1
    ATOMS 2424 2426 2428 2430
   &END DISTANCE_FUNCTION
  &END COLVAR
 &END SUBSYS
 &QMMM
  ECOUPL COULOMB
  &CELL
   ABC 40 40 40
   ALPHA_BETA_GAMMA 90 90 90
  &END CELL
 &QM_KIND  H
  MM_INDEX 2425
  MM_INDEX 2426
  MM_INDEX 2427
  MM_INDEX 2429
  MM_INDEX 2430
  MM_INDEX 896
  MM_INDEX 897
  MM_INDEX 901
  MM_INDEX 903
  MM_INDEX 905
  MM_INDEX 920
  MM_INDEX 921
  MM_INDEX 923
  MM_INDEX 924
  MM_INDEX 1343
  MM_INDEX 1344
  MM_INDEX 1382
  MM_INDEX 1383
  MM_INDEX 2441
  MM_INDEX 2440
 &END QM_KIND
 &QM_KIND  N
  MM_INDEX 2428
  MM_INDEX 2431
  MM_INDEX 2434
  MM_INDEX 899
  MM_INDEX 902
 &END QM_KIND
 &QM_KIND  C
  MM_INDEX 2432
  MM_INDEX 2435
  MM_INDEX 2436
  MM_INDEX 2437
  MM_INDEX 895
  MM_INDEX 898
  MM_INDEX 900
  MM_INDEX 904
  MM_INDEX 919
  MM_INDEX 922
  MM_INDEX 925
  MM_INDEX 1342
  MM_INDEX 1381
 &END QM_KIND
 &QM_KIND  O
  MM_INDEX 2433
  MM_INDEX 926
  MM_INDEX 927
  MM_INDEX 2439
 &END QM_KIND
 &QM_KIND  ZN
  MM_INDEX 2424
 &END QM_KIND
 &QM_KIND  S
  MM_INDEX 1345
  MM_INDEX 1384
 &END QM_KIND
 &LINK
  MM_INDEX  893
  QM_INDEX  895
  LINK_TYPE IMOMM
 &END LINK
 &LINK
  MM_INDEX  917
  QM_INDEX  919
  LINK_TYPE IMOMM
 &END LINK
 &LINK
  MM_INDEX  1340
  QM_INDEX  1342
  LINK_TYPE IMOMM
 &END LINK
 &LINK
  MM_INDEX  1379
  QM_INDEX  1381
  LINK_TYPE IMOMM
 &END LINK
 &END QMMM
&END FORCE_EVAL
```

Metadynamics input file: Cytosine Deaminase, Motion and Restart Section

```
&MOTION
 &MD
 ENSEMBLE NVT
 TIMESTEP [fs] 0.5
 STEPS    10000  !5000 fs = 5ps
 TEMPERATURE 298
 &THERMOSTAT
  TYPE NOSE
  REGION GLOBAL
  &NOSE
   TIMECON [fs] 100.
  &END NOSE
 &END THERMOSTAT
 &END MD
 &FREE_ENERGY
  &METADYN
  USE_PLUMED .TRUE.
  PLUMED_INPUT_FILE ./plumed_diff.inp
  &END METADYN
 &END FREE_ENERGY
 &PRINT
  &RESTART                    ! This section controls the printing of restart files
   &EACH                      ! A restart file will be printed every 10000 md steps
    MD 1000
   &END
  &END
  &RESTART_HISTORY                  ! This section controls dumping of unique restart files during the run keeping
                                      all of them.Most useful if recovery is needed at a later point.
   &EACH                      ! A new restart file will be printed every 10000 md steps
    MD 1000
   &END
  &END
  &TRAJECTORY                   ! Thes section Controls the output of the trajectory
   FORMAT XYZ                   ! Format of the output trajectory is XYZ
   &EACH                      ! New trajectory frame will be printed each 100 md steps
    MD 250
   &END
  &END
 &END PRINT
&END MOTION

&EXT_RESTART
 RESTART_FILE_NAME NPT-1.restart
 RESTART_COUNTERS .FALSE.
 RESTART_THERMOSTAT .FALSE.
&END
```