Contributing to the Development of the Metadynamics Methodology for Studying Chemical Reactions

Christopher D. Fu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:
Jim Pfaendtner, Chair
Stuart Adler
Charles Campbell

Program Authorized to Offer Degree:
Chemical Engineering

University of Washington

Abstract

Contributing to the Development of the Metadynamics Methodology for Studying Chemical Reactions

Christopher D. Fu

Chair of Supervisory Committee: Jim Pfaendtner, Department of Chemical Engineering

The exploration and characterization of chemical reactions has been and will continue to be an active area of research across many scientific disciplines. However, due to the inherent complexities of reacting systems (i.e., the number of species and mechanisms that occur), unraveling a clear understanding of the various sub-processes that take place can be an arduous and near-impossible task experimentally. With perpetual advancements in hardware, algorithms, and theory, computational tools and methods, e.g. molecular dynamics simulations, are increasing appealing to complement experiments and help drive research forward. Simulations, unfortunately, come with their own set of drawbacks. To paraphrase Hippocrates, time steps are short, simulations long, transitions fleeting, assumptions perilous, and judgment difficult. Herein, we improve upon existing methods, and develop novel ones, to help address these challenges, paving the way for these types of computational studies to be practical tools for studying complex reacting systems.

Table of Contents

List of Figures

List of Tables

Acknowledgements

To begin, I would like to thank my dissertation committee of Jim Pfaendtner (chair), Charles Campbell, Stuart Adler, and Marina Meila for reading this lengthy document over the holidays and providing helpful feedback and advice from my prelim exam up through to my defense.

I would like to thank Magda Balazinska, David Beck, and the other members of the eScience Institute who I got to interact with throughout my time at UW. Participating in the UW Big Data IGERT and eScience Institute events had a significant impact on shaping my career and I cannot imagine that I would have the opportunities ahead of me without it.

While my time spent at UW was only little over three years, the road that led me here started long before I even filled out my application. I specifically want to thank Braden and Abigail Giordano for giving me my first research experience, inspiring me to pursue a career in STEM, and for taking the bullet that is Chris Fu doing wet-lab chemistry. From Ms. G's classroom to the lab at NRL, I am extremely grateful for the guidance and mentorship I received from both of you. I also want to thank Philippe Bardet, the first person to make the mistake of hiring me twice! There is no doubt in my mind that the lessons I learned during the six months I spent in your lab helped shave off at least one year of my PhD and will continue to serve me in the future.

I want to acknowledge the members of the Pfaendtner Research Group, former and current, for their helpful discussions and support throughout my PhD (and prior). I specifically want to thank Jim Pfaendtner, the second person to make the mistake of hiring me twice, for his continued mentorship, advice, and support over the past five years.

Finally, I want to thank my family for their continued support and encouragement. No matter how tough things got or how many dead ends I hit, you helped me to never give up.

Introduction

Chemical reactions are ubiquitous, diverse, and play crucial roles in a variety of systems from the combustion of fuels to fundamental cellular processes. In these different contexts, reactions can range from fairly simple mechanisms to very large, complex networks. For example, combustion reactions can often involve thousands of species and have hundreds of different pathways.[1] Due to the short time scales of these processes, and the underlying complexities of these systems, obtaining a clear understanding of the individual steps at an experimental level is a significant challenge.

The use of computational tools, such as MD simulations, offers the opportunity to model such systems at an atomistic level of detail, providing a much-needed lens with which we can resolve and characterize these mechanisms and networks. Specifically, these physics based simulations can effectively model the kinetics and thermodynamics of such systems. Much as an experimentalist would setup a trial in a reactor, we have the ability to similarly construct a model of our system (reactants) under specific conditions (temperature, volume, etc.), and can then model the evolution of the system over time. Having a complete description of the intricacies of these mechanisms and networks provides the opportunity to control and optimize such systems, opening the door for improving process efficiencies and reducing the production of pollutants, to name a few benefits. Additionally, these computational tools could be applied in a predictive manner to drive experiments by preliminarily screening new systems for certain applications.

The main impediment to relying on MD simulations as a panacea, specifically in the context of modeling chemical reactions, is the significant computational cost. Contrary to classical MD simulations, which rely on using force fields to represent the underlying chemistry of a system, modeling reactions typically requires quantum mechanical methods, which are

usually far more expensive and require smaller simulation time steps. Because it is fairly common for reactive events to be infrequent, and reacting systems can encompass multi-step processes (both in parallel and series), effectively modeling such systems requires very long timescales (order of seconds), relative to the time step of the simulation (order of fs). In addition, characterizing the kinetics of reactive events requires sampling an ensemble of events, as these are stochastic processes, further compounding the expense of an MD approach.[2] While the developments of semi-empirical methods[3,4] and parameterized reactive force fields[5,6] have extended the threshold of accessible timescales, these methods are still often restricted to high temperature situations and may not always suitable.

A common approach to alleviate the computational expense of MD is to employ an enhanced sampling method.[7–11] One method that has shown particular promise in studying reactions is metadynamics (MetaD).[7,12–17] In MetaD, an external bias potential, that is a function of user-defined collective variables (CVs), is constructed on the fly by depositing bias in the form of small Gaussians (along the CVs) over the course of the simulation. This bias acts to enhance the fluctuations of the CVs and to discourage the system from re-visiting areas of space, and instead sample new, rarely accessible areas. Once converged, the deposited bias can be used to provide an estimate of the underlying free energy surface (FES) as a function of the CVs biased. By leveraging the bias information, other desired attributes can be recovered as well. Examples include recovering the potential energy surface (PES) through reweighting[18] and kinetics through infrequent MetaD.[19,20]

While both a valuable and versatile tool, MetaD does introduce some obstacles. The main challenge to applying MetaD (and many enhanced sampling methods for that matter) is correctly identifying which CVs to bias. Proper CVs (at least for biasing) should capture the slowest,

relevant degrees of freedom in the system and be able to distinguish between the different metastable states that exist. Ignoring important CVs can lead to hysteresis in the recovered FES or corrupted dynamics. For simple systems identifying proper CVs can be relatively trivial, but in the context of poorly understood chemistries or complex mechanisms this decision is far from intuitive. Additionally, the efficiency of MetaD decreases exponentially as the dimensionality of the bias potential (i.e., number of CVs biased) increases.[21] Therefore, applying MetaD to characterize a system requires a high-level of intuition about the chemistries involved, as well as the ability to represent these mechanisms in low dimensionality. Such demands not only diminish the benefit of using an MD approach, but also make these approaches ill suited for resolving the mechanisms in poorly understood systems. Due to being restricted to modeling very simple systems, where typically other methods would also suffice, a MetaD based approach to modeling chemical reactions bears a stronger resemblance to a computational parlor trick than a robust and scalable methodology, let alone a panacea. The research presented here focuses on the development of new frameworks and improving upon existing ones in order to overcome these obstacles and make these types of approaches a practical option for studying more complex systems of interest. Specifically, these works have focused in two areas: 1) generalizing and improving the efficiency of the infrequent MetaD method and 2) further developing the methodology and application of the parallel bias metadynamics (PBMetaD) framework.

With regards to infrequent MetaD, previous studies demonstrated how this method could be applied to efficiently recover the reaction rates and energy barriers of an S_N2 reaction.[20] While this study served as a significant proof of concept, the system studied was very simple in that there was only one mechanism to study and identifying the proper CVs for this system is fairly intuitive. In Chapter 1, infrequent MetaD framework is extended to sample the kinetics and

selectivities of systems with multiple, competing pathways, as these systems are quite common to reacting systems and networks.[22] In Chapter 2, different types of CVs, specifically ones that require less user intuition to construct, are shown to be compatible with the infrequent MetaD method to capture reaction kinetics and potentially be more efficient than typical user-intuited CVs.[23] The utility of infrequent MetaD is further highlighted in Chapter 7, where it is used to not only recover the kinetics of toluene oxidation and pyrolysis, but also used to help understand the impact an external electric field has on these reaction kinetics.[24]

With regards to PBMetaD, Pfaendtner and Bonomi previously put forth this method as a way to tackle high dimensional sampling by using a series of mono-dimensional bias potentials to enhance the exploration of phase space.[25] In Chapter 3, this method is applied to model and discover reaction pathways in a complex, multi-step reaction system by using it to bias SPRINT coordinates,[15] a generic CV.[26] Previous works have biased SPRINT coordinates using well-tempered MetaD, but suffered from inefficiencies due to the high dimensional bias potentials and had to employ overly aggressive bias parameters.[14,15] Here it is demonstrated how PBMetaD overcomes these challenges, while still capitalizing on the benefits of SPRINT coordinates. Chapter 4 presents a modification to the PBMetaD framework called PBMetaD with partitioned families (PBMetaDPF), which expedites sampling of systems that require biasing many degenerate CVs.[27] In this context, degenerate CVs are CVs that are known or identified to represent identical free energy profiles (e.g., all of the interatomic distances of an LJ cluster). By allowing for degenerate CVs to contribute to and share the same bias potential, PBMetaDPF provides a more efficient scheme for a specific class of systems. In extending its application, and further developing improvements to it, the PBMetaD framework has become a very appealing tool for exploring reactive systems that require a high dimensional representation.

1 Determining Energy Barriers and Selectivities of a Multi-Pathway System with Infrequent Metadynamics[1]

1.1 Abstract

Estimating the transition rates and selectivity of multi-pathway systems with molecular dynamics simulations is expensive and often requires arduous sampling of many individual pathways. Developing a way to efficiently sample and characterize multi-pathway systems creates an opportunity to apply these tools to study systems that, previously, would have had a prohibitive computational cost. We present an approach that places quartic boundaries at the saddle points to isolate individual pathways without changing their observed rates, reducing the required number of events sampled and estimated rate uncertainty. In addition to recovery rates, the selectivity between pathways is also accurately predicted as well. To further reduce the computational cost of the analysis, we have paired this approach with the infrequent metadynamics method. The method is demonstrated on model systems and stiffened alanine dipeptide. Furthermore, we present an appropriate method for recovering the energy barriers of specific transitions paths by taking the slope of an Arrhenius plot generated from the infrequent metadynamics results at various temperatures. We also compare this method against previously another method in literature to demonstrate its superior performance. In the future these methods can be used in a variety of contexts where competing escape pathways with different barriers are relevant.

[1] Reproduced in part with permission from C.D. Fu, L.F.L. Oliveira, and J. Pfaendtner. Determining energy barriers and selectivities of a multi-pathway system with infrequent metadynamics. Journal of Chemical Physics, 146, 014108 (2017). Copyright 2017 American Institute of Physics.

1.2 Introduction

Molecular dynamics (MD) simulations are very effective at characterizing the thermodynamics and kinetics of a system. In particular, one property of interest is the mean first passage time, which is the average time it takes for a barrier to be crossed. While MD simulations can be very powerful in quantifying the dynamics of systems through the mean first passage time, the computational cost of these simulations is usually prohibitive because the transition of interest is typically an infrequent or rare event. The problem arises from a time scale issue because the time scale of integration (fs) is often orders of magnitude smaller than the time scale of the process. Additionally, estimating the mean first passage time of a particular path requires many transition events to be sampled. The expense is further compounded when dealing with transitions that cross high-energy barriers or compete with other pathways. Such systems are widely found in chemical and biochemical reactions as well as conformational transitions in biomolecules.

One means to address this issue to has been through the use of enhanced sampling methods.[7,19,28–32] Specifically, methods such as forward flux sampling,[28] transition path sampling,[29] and various forms of the string method,[30,33,34] have had success in characterizing transitions paths and calculating estimates of rate coefficients. However, such methods typically struggle to identify all of the parallel, competing transition paths in a system, let alone characterize the selectivity of these paths.[28] This is a potential benefit of the methods such as hyperdynamics[32] and infrequent metadynamics method,[7,19,20] as the bias deposited in these simulations does not preclude a system from exploring any relevant pathway. Our approach uses the metadynamics (MetaD) family of methods[7] to address the time scale issue.

In MetaD a history-dependent bias is applied along a few chosen collective variables over the course of the simulation. The deposited bias prevents a system from being trapped in a low-

energy basin and, instead, allows for the system to explore regions of the phase space that would normally be rarely visited in an unbiased simulation. The infrequent metadynamics method, developed by Tiwary and Parrinello,[19] recovers the rates of rare events from biased simulations. The infrequent metadynamics method has been shown to reduce the cost of simulations, characterize the mean first passage time over a particular barrier, and estimate the mean escape time from a basin. Fleming et al. further applied infrequent metadynamics to characterize the transition rates of chemical reactions and illustrated that there rate results could also estimate the energy barrier of transitions.[20]

We extended the application of infrequent metadynamics to study systems with multiple, competing pathways and demonstrated its ability to determine rates associated with individual pathways and predict the overall selectivity of systems from biased simulations. However, such an analysis encounters some complications, which we will enumerate below, and requires numerous events to be sampled in order to get a converged rate estimate, beyond what is typically practical. To address this, we developed an approach that isolates a single pathway of interest, but still recovers a transition rate consistent with that of a system without intervention. We applied this approach to two systems, namely, an analytical system of varying energy barriers described by a cosine function, and the conformational change of alanine dipeptide in vacuum. Lastly, we prove that the so-called first passage method, an analysis for characterizing the activation energy associated with a transition path, is invalid and further validate the approach outlined by Fleming et al.[20]

The paper is organized as follows: in Section II we outlines the details and key steps involved in carrying infrequent metadynamics, pathway isolation, and evaluating selectivity. In Section III we apply our approach to model systems and demonstrate its ability to reduce the number of

simulations required and the uncertainty of the estimated rates. Section IV demonstrates that the first passage method is not suitable for evaluating energy barriers across a variety of energy barriers and Section V is devoted to the conclusions and potential applications.

1.3 Theory and Methodology

1.3.1 Infrequent Metadynamics and Multi-Pathway Sampling

We begin by introducing the infrequent metadynamics method. Because many multi-pathway systems of potential interest have relatively high energy barriers, it is practical that our approach be compatible with an enhanced sampling method such as metadynamics. The infrequent metadynamics method uses knowledge of the bias deposited throughout the simulation to evaluate the effective escape time, t^{eff} (inverse of the rate of escape), for an individual event, and can be recovered by:[19,20]

$$t^{eff} = \alpha * t^{MD} \quad \text{(Eq. 1.1)}$$

$$\alpha = \frac{1}{t^{MD}} * \int_0^{t^{MD}} dt' \, e^{\beta V_{bias}(s,t')} \quad \text{(Eq. 1.2)}$$

where t^{MD} is the molecular dynamics time when the transition occurred, α is the acceleration factor, V_{bias} is the bias deposited at that location and time in the simulation, and β is $1/k_bT$. The infrequent metadynamics method has been shown to reduce the cost of simulations, determine the mean first passage time over a particular barrier, and estimate the mean escape time from a basin.[2,19,20,35]

While infrequent metadynamics has been used to estimate the mean escape time from a stable state with multiple transition pathways,[2,19,32] characterizing the individual rates of each pathway is not as straightforward. One difference is that the pathways in such systems are competing, and therefore the collection of events for a particular pathway is not independent of the other transition events. However, this problem can trivially be dealt with. In simulating stochastic

processes, each individual transition sampled is an independent event and the population of transition rates collected should follow a Poisson distribution.[2] For systems with multiple transition pathways, the cumulative rate of escape from a stable state, is the sum of the average rates associated with each pathway.[2] Therefore, the rates of individual pathways can be determined from the cumulative escape rate from the stable state and the number of events sampled for each pathway. The average escape rate of each pathway, υ_i, can be found from:

$$\upsilon_i = \frac{n_i}{t_{tot}^{esc}} \qquad \text{(Eq. 1.3)}$$

where n_i is the number of events that passed through a given pathway and t_{tot}^{esc} is the sum of the effective escape times for all of the events recorded for a given system. It is worth noting that for a single pathway system, Eq. 3 reduces to a regular expression for the average rate of escape of a system. While all of the systems studies were limited to only two pathways, this analysis can easily be extended to systems with more accessible pathways.

While Eq. 3 provides a way to calculate the rates on individual pathways, there is a drawback to this method. If sampling a fixed number of events for a given system, the uncertainty of the rates calculated is strongly determined by the number of events sampled for each pathway. Because these are stochastic processes being simulated, the number, or ratio, of events sampled can fluctuate in a given sample population leading to fairly large uncertainties. In order to overcome the error introduced by these sampling fluctuations, a large number of events must be sampled through each pathway. The sampling of events for each pathway becomes a significant and expensive obstacle as the difference in the energy barrier heights and number of alternative pathways increases. As the difference in energy barriers increases, transition events are less likely to occur through the higher energy pathway, making it difficult to characterize. In order to sample just a few events through a transition pathway that is more than a few k_bT higher than

another competing pathway many more simulations would need to be run because the bulk of transition events would proceed through the lower energy pathway. The problem is exacerbated if the number of pathways increases – even if there are similar barrier heights. The obvious brute force solution of increasing sampling to convergence then becomes impractical for most systems.

1.3.2 Pathway Isolation

To address these issues, we developed an approach within the framework of infrequent metadynamics that isolates a single pathway of interest, but still recovers a transition rate consistent with that of a system without intervention. In order to sample specific transition pathways, we placed quartic boundaries at the saddle points of all paths except the one of interest. Thus, allowing transitions to occur only through one specified pathway. Our approach is similar in concept to the boxed MD method (BXD),[10] which employs reflecting boundary conditions in discretized windows of the relevant phase space. In our approach, however, the wall is placed at the saddle point, so that the available phase space of the initial stable state, and by extension the transition rate, is unaffected. This observation is consistent with transition state theory (TST) as shown by the following equation:

$$k_{A\to B} = \kappa\omega\frac{Z_{AB}^{TS}}{Z_A} \qquad \text{(Eq. 1.4)}$$

where $k_{A\to B}$ is the rate from state A to state B, ω is a constant related to the system's temperature and mass of the reaction coordinates, κ is the transmission coefficient (≤ 1), and Z_{AB}^{TS} and Z_A are the partition functions of the transition state and state A, respectively. While placing the walls at the saddle points of transition pathways is a clean way to achieve this effect, characterizing the transition state accurately can often be challenging. An alternative approach to determine how to construct the boundaries would be to run a few trial infrequent metadynamics simulations and monitor the collective variable values before and after a transition event. When a

transition occurs, a shift or change should be observed in the typical range of the collective variable values. The boundary can then be placed so as not to interfere with the collective variable values corresponding to the initial state, before the shift, but still prevent the collective variable values from reaching that of the new state. Because the wall is placed so it does not interfere with the volume of phase space of state A, the partition function of state A, and therefore $k_{A\to B}$, is unaffected by the wall.

Additionally, we treat the wall as an extension of the underlying PES and do not include its contribution in the acceleration factor (i.e., only the metadynamics transient is used to calculate a in Eq 2). To verify that the quartic walls did not affect the recovered rates, simulations were performed with two available pathways and then with single pathways. The rates were calculated in the same way with Eq. 3, except now both the number of events sampled and sum of escape times both correspond to only the one path. Because the paths are sampled independently, the number of events needed to characterize a single pathway is fixed and there are no wasted simulations over-characterizing a single path. Further details are provided below.

1.3.3 Evaluating Selectivity

In addition to the individual path rates, the selectivity of one pathway over another is a value of interest for multi-pathway systems. We demonstrated our method is capable of recovering the selectivity of two reaction channels in a two-pathway system. In order to assess how successful the single pathway systems were at describing the selectivity (S), parameters from Arrhenius plots were used to recover values for the pre-exponential factor (A) and energy barrier ($\Delta E^{\neq}$), and these were used to estimate the selectivity using the following equation for a first order process:

$$S_{1,2} = \frac{r_1}{r_2} \propto \frac{\#\ events\ through\ path\ 1}{\#\ events\ through\ path\ 2} \propto \frac{A_1}{A_2} e^{\frac{(\Delta E_2^{\neq} - \Delta E_1^{\neq})}{RT}} \qquad \text{(Eq. 1.5)}$$

A prediction of selectivity based solely on the known difference in energy barriers was also made, neglecting the contribution of the pre-exponential factors in Eq. 5, i.e. in the limit of assuming identical pre-factors ($A_1=A_2$).

1.3.4 Sampling Procedure & Error Analysis

We applied this approach to an analytical system of varying energy barriers and the conformational change of alanine dipeptide in vacuum. By coupling it with the infrequent metadynamics method, we further demonstrate that our approach is a viable method for reducing the computational cost of investigating multi-pathway systems. Hundreds to thousands of events were sampled for each system studied and for each event the effective escape time and identity of transition path were recorded.

In accordance with the procedure explained by Salvalaglio et al.,[2] the Kolmogorov-Smirnov (KS) test[36] was applied to confirm that the distribution of collected escape times of a given data set followed a Poisson distribution. An ensemble was deemed to follow a Poisson distribution if the *p*-value recovered exceeded 0.05. This is a vital step to analyzing the distribution of times recovered as it ensures that the individual events of a distribution are uncorrelated.

Error analysis was carried out for each data set following the bootstrapping procedure outlined in ref 10. Because all of our population sizes were on the order of magnitude of hundreds to thousands, the bootstrapping analysis was done with one thousand subsets of 100 random events drawn (with replacement) from the collected data sets. If one of the subsets did not pass the KS test ($p > 0.05$), this subset was discarded and another was drawn to replace it; the subset reject rate was always small, never exceeding 8%.

1.4 Results and Discussion

1.4.1 An Analytical System

The first system investigated was a one-dimensional potential described by a cosine function with two accessible pathways. The system was constructed in a piece-wise manner so it was differentiable at all points, but allowed for us to study different combinations of energy barriers. Two different potential functions were analyzed with different barrier heights (6-4 energy units and 5-4 energy units barriers) in order to verify the consistency of this method in predicting selectivity, transition rates, and energy barriers (from Arrhenius plots).

The simulations were carried out using the LimPy package (Langevin Integrator Metadynamics Python)[37]. LimPy is a python package that carries out molecular dynamics simulations using a Langevin integrator following the algorithm outlined by Bussi and Parrinello,[38] and is capable of carrying out variants of the MetaD family of enhanced sampling methods, including the infrequent metadynamics algorithm to post-process the results.

Each system was simulated over the temperature range of 0.63 to 1.68 $1/k_bT$ and over 2400 events were collected. Simulations were carried out with friction factor 5, a time step of 0.001, and a mass of 1. The simulations were performed using the well-tempered variant of MetaD. We used an initial Gaussian height of 0.25 energy units, a Gaussian width of 0.25, a bias factor of 9, and a deposition stride of 10^4 time steps. Figure 1.1 illustrates the combination of systems simulated for the 5-4 energy barrier pathways and the Arrhenius-like behavior of the system. The higher, 5 energy unit, pathway is colored in blue in Figures 1.1 A-C and is referred to as the Left barrier. The lower, 4 energy unit, pathway is colored in red in Figures 1.1 A-C and is referred to as the Right barrier.

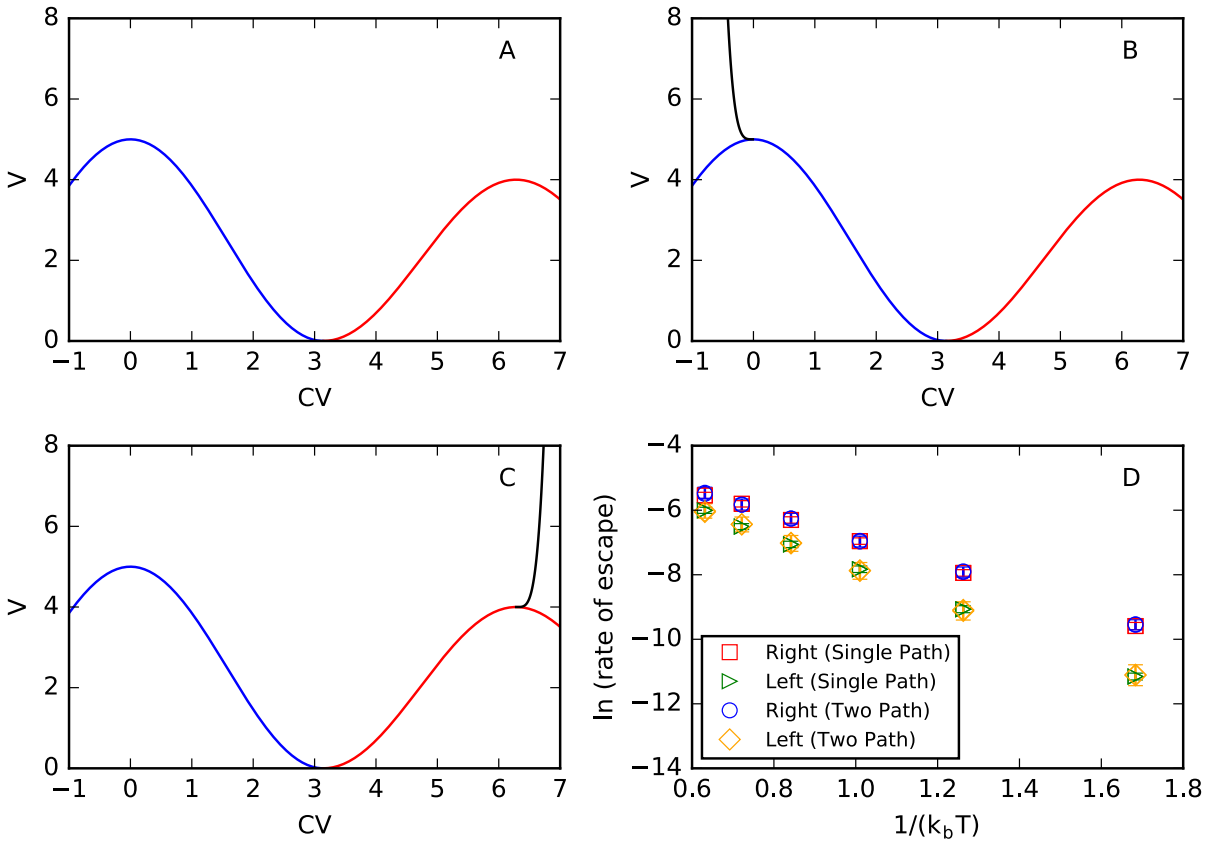


Figure 1.1: (A) Depiction of the cosine potential system with two available pathways with barrier heights of 5 energy units (Left) and 4 energy units (Right). (B) Single pathway system with a quartic boundary potential placed at the saddle point of the Left barrier. (C) Single pathway system with a quartic boundary potential placed at the saddle point of the Right barrier. (D) Arrhenius plot with the rate data for the systems illustrated in A-C with the red squares corresponding to transitions rates across the Right barrier of the single pathway system (1B), green triangles corresponding to the transition rates across the Left barrier of the single pathway system (1C), blue circles corresponding to the transitions across the Right barrier of the two pathway system (1A), and the orange diamonds corresponding to the transitions across the Left barrier of the two pathway system (1A).

Following classical TST, the entropic contributions are captured in the pre-exponential factor, or the exponential of the intercept of the Arrhenius plot. Therefore, the activation enthalpy of the system can be recovered from the slope of the Arrhenius plot, and is strongly related to the potential energy of the system.[39] Table 1.1 shows that energy barriers calculated from the Arrhenius plot parameters are in close agreement with energy barriers set in the program, matching what is expected from TST.

Table 1.1: Calculated energy barriers and pre-exponential factors from Arrhenius plots for cosine potentials.

System	Actual Energy Barrier (energy units)	Arrhenius Plot Energy Barrier (energy units)	Pre-Exponential Factor $(\text{time step})^{-1}$
Single Path	5.00	4.85 (0.12)	0.052 (0.006)
	4.00	3.88 (0.12)	0.048 (0.006)
Two Path	5.00	4.87 (0.32)	0.050 (0.017)
	4.00	3.85 (0.18)	0.048 (0.010)

The values in parenthesis are the estimates of the standard deviations and were generated through linear regression.

As shown in Figure 1.1D and Table 1.1, there is remarkably good agreement between the rates collected from the blocked and unblocked system, as well as the energy barriers for each pathway. We estimated the uncertainty of the energy barriers and the pre-exponential factors through linear regression of the Arrhenius plots. Because the difference in energy barriers is relatively small, we were able to exhaustively sample both pathways of the system without the barriers. However, as the difference in energy barriers increases ($> k_B T$), sampling the higher energy pathway becomes challenging, which is where the isolating pathways can help. We extended our analysis to the 6-4 energy unit systems to demonstrate this point. Figure 1.2D and

Table 1.2 both show how our method can be extended to systems with larger differences in energy barriers.

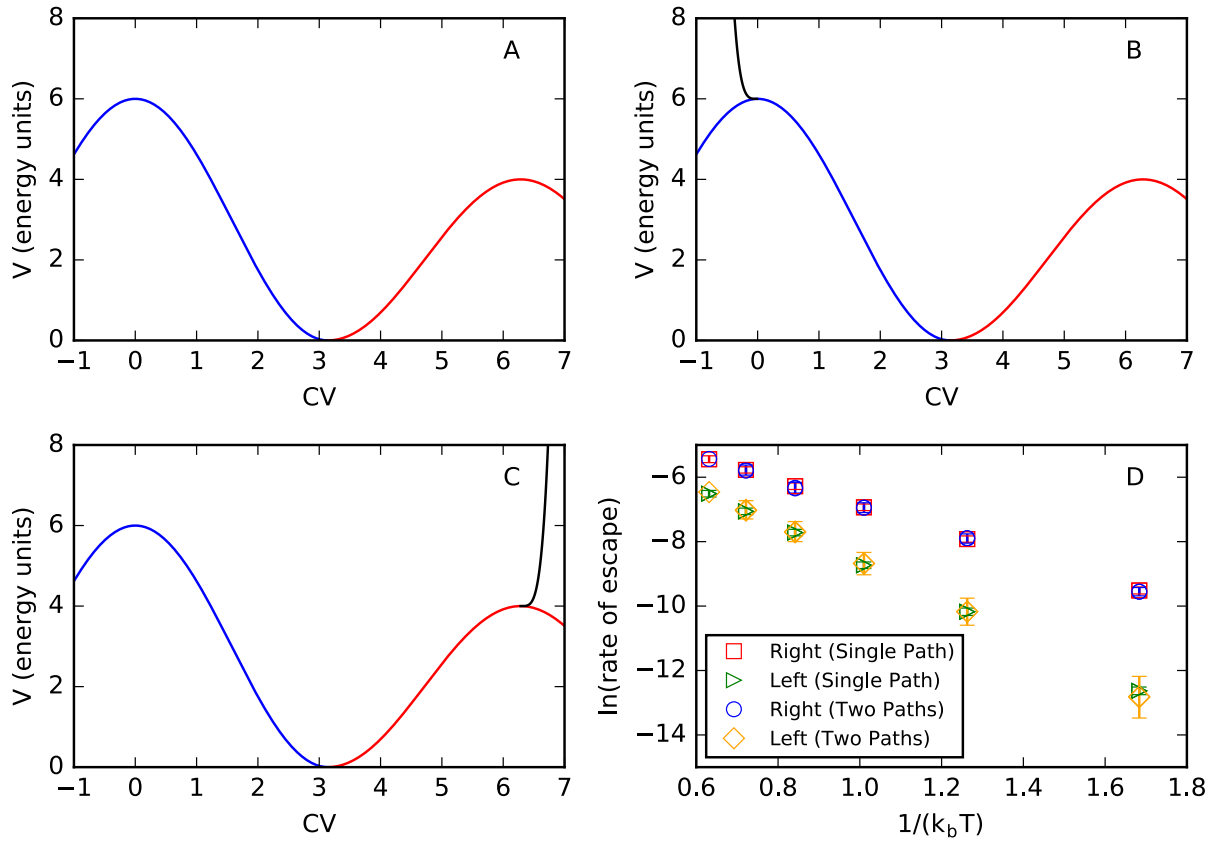


Figure 1.2: (A) Depiction of the cosine potential system with two available pathways with barrier heights of 6 energy units (Left) and 4 energy units (Right). (B) Single pathway system with a quartic boundary potential placed at the saddle point of the Left barrier. (C) Single pathway system with a quartic boundary potential placed at the saddle point of the Right barrier. (D) Arrhenius plot with the rate data for the systems illustrated in A-C with the red squares corresponding to transitions rates across the Right barrier of the single pathway system (1B), green triangles corresponding to the transition rates across the Left barrier of the single pathway system (1C), blue circles corresponding to the transitions across the Right barrier of the two pathway system (1A), and the orange diamonds corresponding to the transitions across the Left barrier of the two pathway system (1A).

Table 1.2. Calculated energy barriers and pre-exponential factors from Arrhenius plots for cosine potentials.

System	Actual Energy Barrier (energy units)	Arrhenius Plot Energy Barrier (energy units)	Pre-Exponential Factor $(\text{time step})^{-1}$
Single Path	6.00	5.80 (0.13)	0.058 (0.008)
	4.00	3.89 (0.12)	0.050 (0.006)
Two Path	6.00	5.99 (0.62)	0.070 (0.042)
	4.00	3.88 (0.18)	0.049 (0.011)

The values in parenthesis are the estimates of the standard deviations and were generated through linear regression.

It is for the 6-4 energy barrier pathway where we see the benefit of isolating pathways. Although thousands of events were collected for each trial, this large number was excessive for sampling the blocked paths, but necessary for sampling the multiple pathway systems as shown in Figure 1.3.

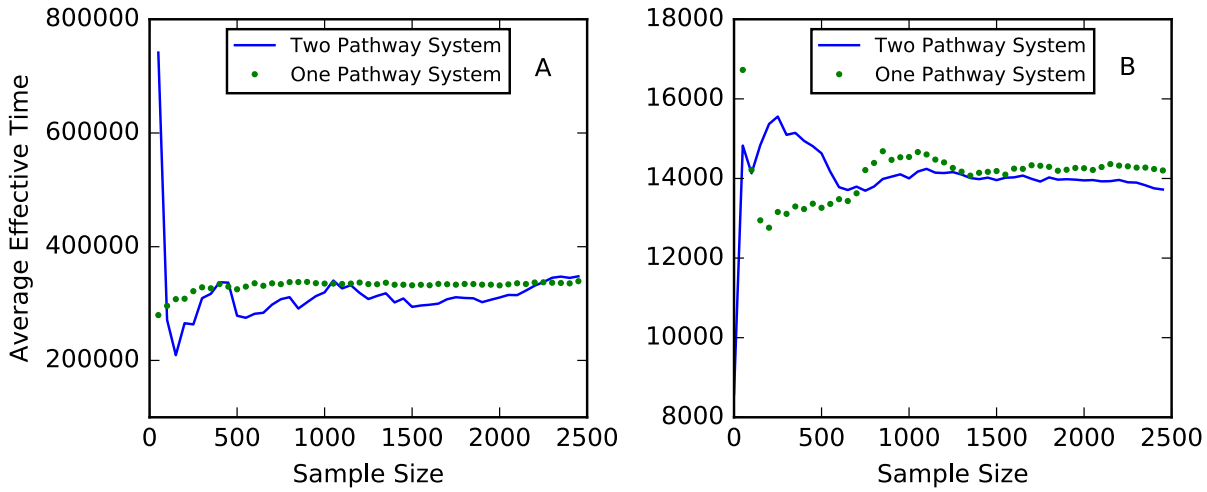


Figure 1.3: Convergence of the average escape times for the Left (6 energy unit barrier) pathway (A) and the Right (4 energy unit barrier) pathway (B) as a function of the number of transition events sampled in each system. The green dots and blue lines refer to the systems with one and two accessible paths, respectively.

In addition to expediting the convergence of the estimated rate, the uncertainty values of the rates recovered from the blocked systems are significantly smaller than those with two available pathways. This can be seen in the error bars in Figure 1.2D (and shown in supplementary Appendix 1). It is clear that for analyzing systems with multiple accessible pathways our approach improves the confidence of the results, but also reduces the number of trials needed.

Figure 1.4 reports the expected theoretical selectivity as a function of the measured selectivity. The selectivity predicted from the Arrhenius plot parameters of the blocked system shows a strong agreement and does not have a preferred tendency to over or under predict selectivity values. While the uncertainties of the pre-exponential factors, generated from linear regression, are relatively large, and therefore open up the possibility that they could be same, this is unlikely. If the pre-exponential factors for the systems were the same, then the difference between energy barriers alone would dictate the selectivity. However, this is shown to be a poor approximation

as the selectivity is consistently over-estimated (see Figure 1.4). Because the energy barriers were set through specifying the cosine potential, the barriers have no temperature dependence.

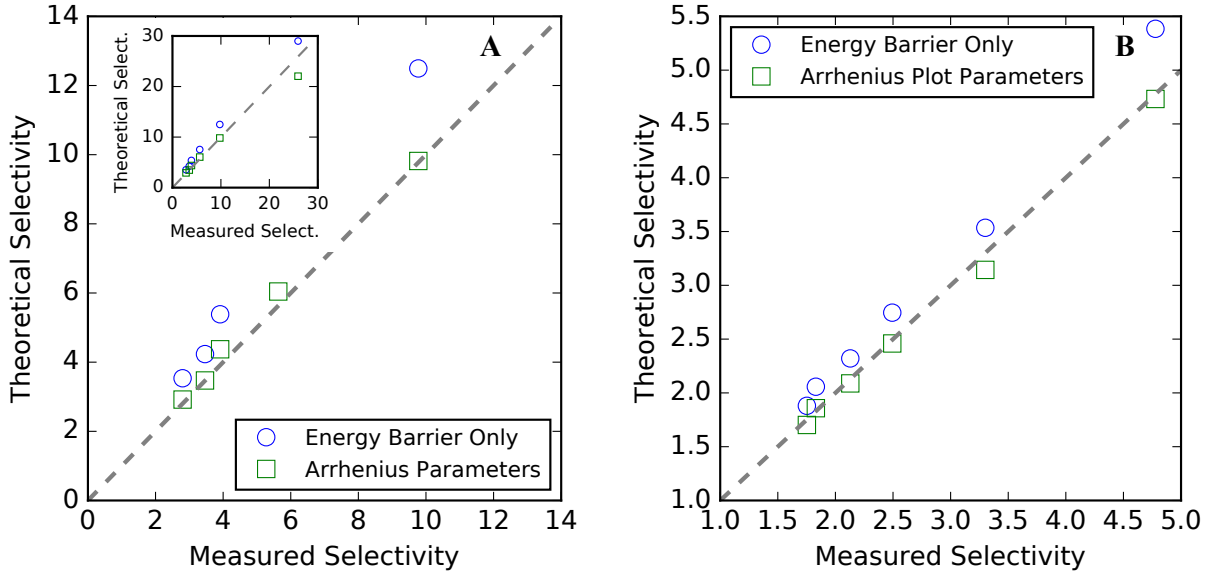


Figure 1.4: Expected selectivity (dimensionless) plotted as a function of measured selectivity for (A) the 6-4 energy unit cosine potential for temperatures ranging from 0.63 to 1.68 $1/k_bT$ and (B) the 5-4 energy unit cosine potential for temperatures ranging from 0.63 to 1.68 $1/k_bT$. Measured selectivity is the ratio of events recorded for the two-pathway system. The green squares are the selectivity values predicted from the Arrhenius plot parameters. The blue circles are the selectivity values predicted from the known difference in energy barrier heights (2 energy units). The grey line illustrates a perfect 1:1 fit. The inlet shows the same, except extended to a larger range of selectivities explored.

1.4.2 Alanine Dipeptide in Vacuum

To further verify the validity of this approach, we applied it to the conformation changes of alanine dipeptide in vacuum. These simulations were carried out using GROMACS 5.1[40] and PLUMED 2.2.[41] We stiffened the dihedral potentials in the Amber14sb force field in order to create a clear two-pathway system with energy barriers within a few k_bT. This combination of

energy barriers allowed us to collect a large number of events through each pathway without needing to perform an excessive number of simulations or to perform simulations at high temperatures. We reconstructed the potential energy surface using the reweighting function of PLUMED, which follows the method outlined by Tiwary and Parrinello.[42] In the simulations, we biased the Φ and Ψ torsional angles with an initial Gaussian height of 1.2552 kJ, Gaussian width of 0.2 radians in each dimension, bias factor of 9, and deposition stride of 1 ps. Simulations were run for 2 μs with a time step of 1 fs. The short range electrostatic and Van der Waals interactions were cut off at 2.0 nm. In order to ensure convergence, we made a comparison between the FES at 50%, 75% and 100% completion as shown in Figure 1.5.

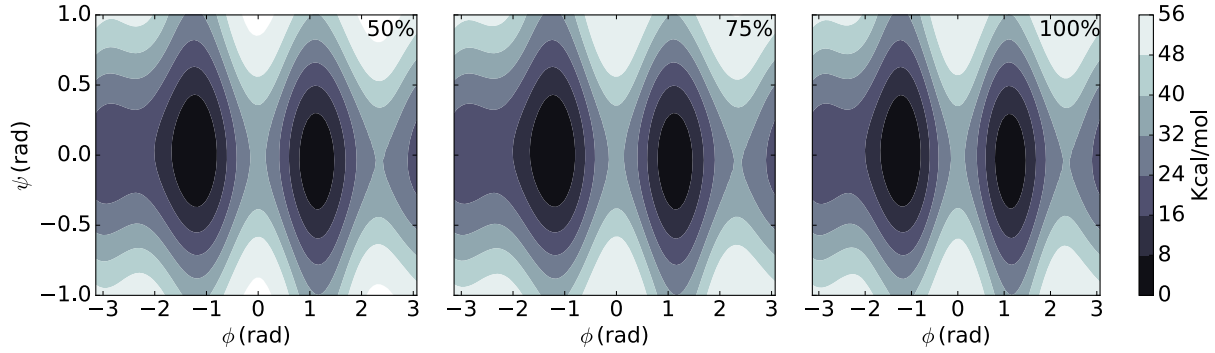


Figure 1.5: Free energy surfaces (FES) of alanine dipeptide in vacuum with stiffened dihedral potentials (Amber14sb force field) at 50%, 75%, and 100% of the total simulation time (2μs).

While there are some fluctuations in the higher energy regions of the system, the wells and transition pathways of interest are fully resolved and do not change. We verified these by running the climbing image nudged elastic band method (CI-NEB)[43] on each FES to calculate and compare the energy barriers, and the deviations were well below k_bT. Once the converged FES and PES were resolved, we applied infrequent metadynamics to systems with and without saddle point boundaries. The simulations were initiated from a state where Φ and Ψ were 1.13 and -0.057 radians, respectively. The infrequent metadynamics simulations were halted when the

Φ angle dropped below -0.5 radians (Left barrier) or exceeded 2.7 radians (Right barrier) and were carried out with deposition stride of 10 ps. We used PLUMED to recover the acceleration factors for each simulation, as well as place the quartic boundaries. We placed the quartic boundaries at what were perceived to be the saddle points in Figure 1.5, and additionally monitored a few biased simulations to ensure the wall location was not placed in phase space visited by the system before committing to the new basin. In general, if an exact transition state is difficult to discern for a given system, this is an alternative, viable method for determining where to place the boundaries so it will not limit the phase space of the original basin (per Eq. 4).

The Arrhenius plot shown in Figure 1.6 displays the rates recovered for the blocked and unblocked systems, which are remarkably similar. There is also very good agreement between the energy barriers calculated from the Arrhenius plot and the energy barrier of the potential energy surface, well within k_bT, as reported in Table 1.3. The uncertainty of the energy barriers calculated from the Arrhenius plots were generated from linear regression.

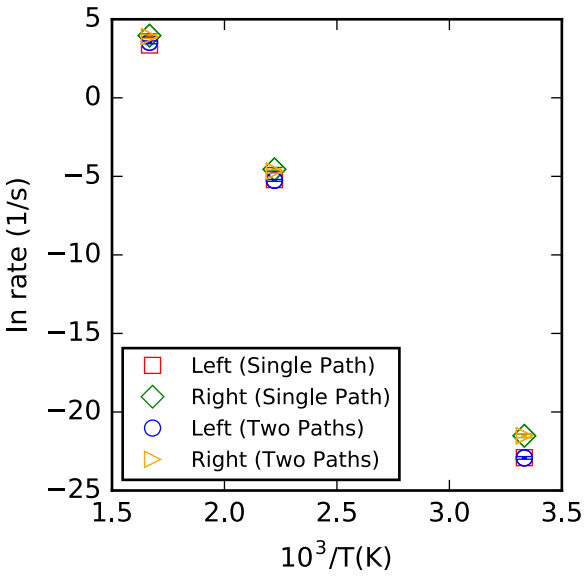


Figure 1.6: Arrhenius plot for the stiffened alanine dipeptide with a deposition stride of 10 ps. The squares represent the rates calculated from the single path system systems and the triangles represent the rates calculated from the two path systems. The red squares and green diamonds represent the Left and Right pathways, respectively, in the single path systems. The blue circles and orange triangles represent the Left and Right pathways, respectively, in the two-pathway system.

Table 1.3: Calculated energy barriers and pre-exponential factors from Arrhenius plots for stiffened alanine dipeptide in vacuum.

Systems	Actual Energy Barrier (kcal/mol)	Arrhenius Plot Energy Barrier (kcal/mol)	Pre-Exponential Factor (1/s)
Single Path	31.1	31.4 (0.04)	1.1 (0.05) $*10^{13}$
	29.9	30.2 (0.12)	5.0 (0.7) $*10^{12}$
Two Path	31.1	31.3 (0.25)	8.3 (2.6) $*10^{12}$
	29.9	30.3 (0.02)	6.0 (0.13) $*10^{12}$

The values in parenthesis are the estimates of the standard deviations and were generated through linear regression.

In determining the selectivity of the system, we observed that considering the potential energy barriers alone is insufficient. When comparing the measured ratio of events to the selectivity calculated based entirely on the difference in barrier heights alone, Figure 1.7 shows that the selectivity is over-predicted. As was observed with the analytical system, the intercept and slope from the Arrhenius plots of the blocked system are much better at predicting the selectivity of the system (see Eq. 5). Running simulations at more temperatures can reinforce the confidence of the Arrhenius plot parameters, but it is clear from Figure 1.6 that even running as little as three temperatures provides a better estimate of selectivity than energy barriers alone.

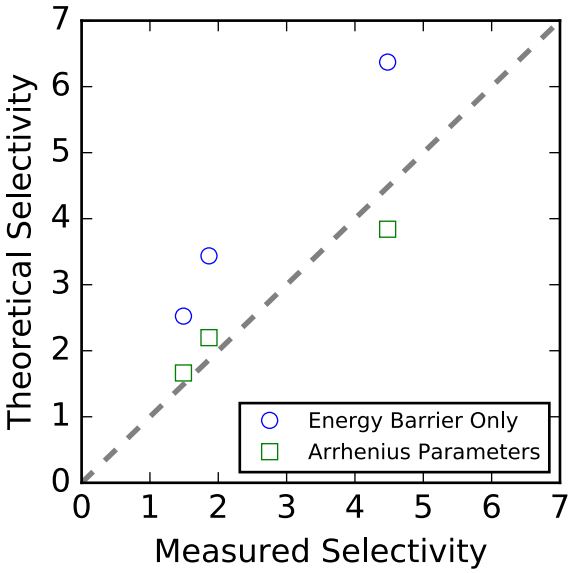


Figure 1.7: Expected selectivity plotted as a function of measured selectivity for the stiffened alanine dipeptide for temperatures 300 K, 450 K, and 600 K. Measured selectivity is the ratio of events recorded for the two-pathway system. The green squares are the selectivity values predicted from the Arrhenius plot parameters for the 10 ps deposition stride. The blue circles are the selectivity values predicted from the known difference in energy barrier heights (1.2 kcal/mol). The grey line shows what would be an ideal 1:1 match.

In accordance with TST, it is likely that entropic differences between the two pathways are the cause for deviation between the observed selectivity and those predicted from just the difference in energy barriers. This conclusion is further reinforced by differences in the pre-exponential factors. As shown in Table 1.4 the pathway leading to the left does have slightly higher entropic barrier, as there is a larger difference between the free and potential energy. However, accounting for this 0.1 kcal/K in entropy still underestimates the ratio of pre-exponential factors. It is possible that larger entropic effects are not currently recovered because the uncertainty in the FES and PES measured are ~ k_bT, which would have a substantial impact

on the selectivity. Despite the uncertainty regarding the impact of entropy on the selectivity for this particular system, the consistency of the rates recovered with and without boundaries demonstrates that the addition of the boundaries is not the source of this error.

Table 1.4: Free and potential energy barriers of the two accessible pathways for the conformation changes of alanine dipeptide.

Pathway	Free Energy Barrier (kcal/mol)*	Potential Energy Barrier (kcal/mol)*
Left	31.6	31.1
Right	30.3	29.9

*Energy barriers calculated by CI-NEB using the converged FES and PES.

1.5 Energy Barrier Characterization: First Passage Method

In addition to characterizing the mean first passage time, the energy barrier of a transition is also of interest. As shown above in Table 1.1 & Table 1.4, and noted by Fleming et al.,[20] the energy barrier of a transition pathway can be recovered from constructing an Arrhenius plot of rate data collected and calculating the slope. This approach, while accurate (agreement < k_bT), has the drawback that the system must be simulated at multiple temperatures, which adds to the computational expense. However, it does not require a converged FES, which is a common method.[13,21] An alternative, less expensive method for estimating the energy barrier of a transition is the so-called mean first passage method, which calculates the maximum bias deposited up until a transition occurred in a simulation, or the apparent free energy barrier.[20,44,45] Fleming et al. observed a connection between the hills deposition stride and the apparent free energy barrier recovered from a metadynamics simulation.[20] However, the energy barrier studied was relatively small ~10 k_bT. We extended this study by simulating barrier heights up to 84 k_bT on an analytical system in order to see if this trend is a general phenomenon of metadynamics simulations or just an artifact of small energy barriers.

We used the Muller Brown PES[46] in the LimPy package to determine whether there is any connection between deposition stride and apparent free energy barrier recovered for a range of barrier heights. We tested the first passage method on systems with energy barriers of 8.4, 17, 42, and 84 k_bT, using deposition strides of 5, 50, 500, and 5000 time steps. The simulations were carried out with a time step size of 0.01, friction factor of 5, a temperature of 1.68 $1/k_bT$, and a mass of 1. The Muller Brown PES was constructed following the equation outlined by Zuckerman et al. which allowed for scalable energy barriers.[47] While the parameters were preserved as much as possible across all four strides, the hill height and bias factor were scaled to accommodate the higher energy barriers to ensure that the transitions would occur in a reasonable time (Table 1.5).

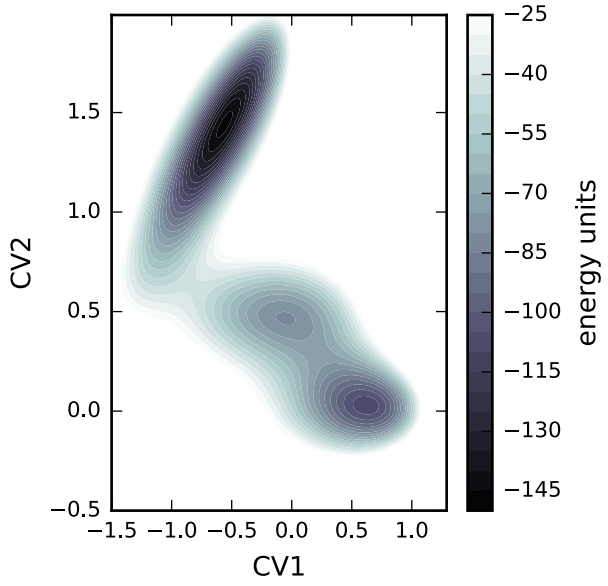


Figure 1.8: Example of a Muller Brown PES tested. The simulations originate in the center at (-0.05, 0.467) and proceeded to the lower right well marked when the CV2 value dropped below 0.

Table 1.5: MetaD parameters for barrier heights on the Muller Brown PES

Barrier Height (k_bT)	Gaussian Height (k_bT)	Gaussian Width (s)	Bias Factor (g)
8.4	0.84	0.02	9
17	0.84	0.02	9
42	2.10	0.02	25
84	4.21	0.02	34

A clear connection between the calculated apparent free energy barrier and the deposition stride of the simulation is shown in Figure 1.9 and Table 1.6. Even up to energy barriers of 84 k_bT the distribution of apparent free energy barriers varies with deposition stride, with the faster deposition rates leading to higher distributions, matching what Fleming et al. observed. We hypothesize that the underlying cause for the trend is that faster deposition rates leads to over-filling of the basin with the bias potential, highlighted by the deposition stride of 5 and 50 time steps.

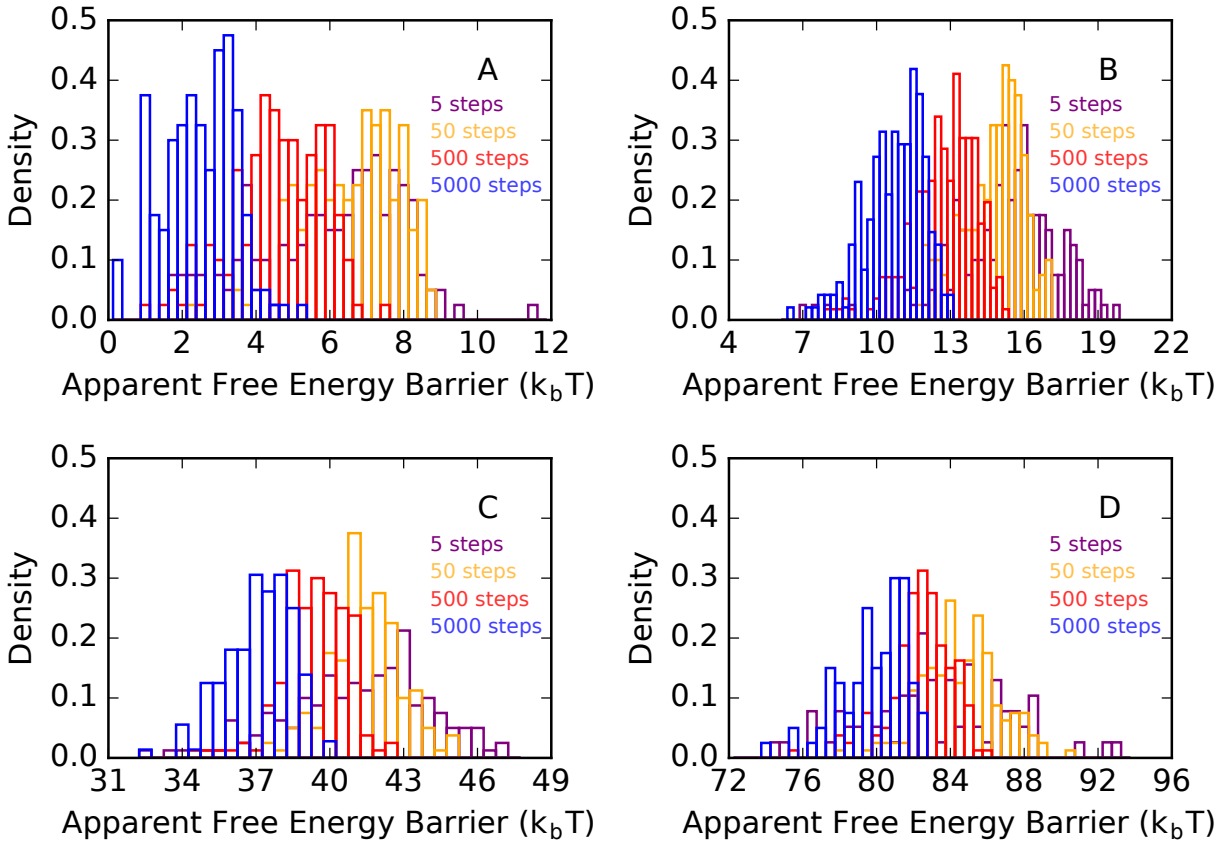


Figure 1.9: Density plots depicting the dependence of apparent free energy barrier measured on the deposition stride for energy barriers of (A) 8.4 k_bT (B) 17 k_bT (C) 42 k_bT (D) 84 k_bT all at a temperature of 1.68 $1/k_bT$. Deposition strides of 5, 50, 500, and 5000 time steps were tested and are visualized with purple, orange, red, and blue bars respectively.

Table 1.6: Mean apparent free energy barriers for different combinations of barrier heights and deposition paces. Standard deviation of the distributions (Figure 1.9) in parentheses.

Set Barrier Height (k_bT)	5 Step Pace (k_bT)	50 Step Pace (k_bT)	500 Step Pace (k_bT)	5000 step Pace (k_bT)
8.4	5.9 (1.9)	6.5 (1.3)	4.5 (1.2)	2.4 (1.0)
17	15.2 (2.1)	14.7 (1.2)	12.6 (1.3)	10.6 (1.3)
42	40.9 (2.9)	41.1 (1.6)	39.2 (1.3)	36.9 (1.3)
84	81.9 (7.7)	84.3 (2.1)	82.2 (1.6)	79.6 (1.7)

It is known that the error of a MetaD simulation scales with the inverse of the deposition stride, meaning slower deposition strides will reduce the error of simulation.[44] In contrast, Figure 1.9 and Table 1.6 show that the slowest deposition rate returns the largest deviation from the true free energy barrier when estimated in this manner. The connection between deposition stride and apparent free energy barrier deviation likely arises because these simulations are not converged, but truncated when an escape event occurs. Also, infrequent metadynamics relates the transition times from biased simulations to their unbiased counterparts. As a result, a slower deposition stride would lead to a longer biased transition time and, correspondingly, a smaller amount of bias to be deposited. Therefore, prior to the transition event, a slower deposition stride would lead to a smaller apparent free energy barrier (height of bias deposited), and thus result in a larger deviation from the true free energy barrier. Due to the strong dependence of the apparent free energy barrier on the deposition stride, we propose future studies avoid the use of this method for approximating free energy barriers of a transition states, or at minimum assess the error of the approach by performing multiple simulations at different values of the MetaD stride.

1.6 Conclusions

We have demonstrated a method that uses boundaries at the saddle points of transition paths to efficiently isolate, sample, and characterize individual paths in multi-pathway systems without affecting the recovered transition rate. The described approach reduces the number of events needed to effectively characterize the transition rates of pathways with energy barriers significantly higher in energy ($> k_BT$) than other competing pathways in the system and reduces the overall error in the recovered rates. Furthermore, we demonstrated that our approach is capable of capturing the selectivity and defining the energy barriers of such systems. This approach is very well suited for investigating reaction mechanisms and networks which have many states connected by different transition paths. Independent of the difference in barrier heights, the number of events necessary for estimating the transition rates in a system scale directly with the number of paths to characterize. This feature drastically reduces the number of simulations to run. We paired our approach with the infrequent metadynamics method to further reduce the computational cost. By reducing the cost of simulations and the number needed to be run, this approach is a promising way to characterize systems with high energy barriers and many pathways.

2 Assessing Generic Collective Variables for Determining Reaction Rates in Metadynamics Simulations[2]

2.1 Abstract

A persistent challenge in using the metadynamics method is deciding which degrees of freedom, or collective variables, should be biased because these selections are not obvious and require intuition about the system being studied. There are, however, collective variables, which can be constructed with only basic knowledge about the system studied, that provide an opportunity to alleviate this issue. We simulated two different reacting systems where two types of such collective variables (SPRINT coordinates and the collective variable-driven hyperdynamics method) were biased following the infrequent metadynamics method in order to recover the rates of reactions. We demonstrate that both generic collective variables are capable of reproducing the reaction rates of both systems and can enhance the efficiency of the simulation when compared to typical collective variables.

[2] Reproduced in part with permission from C.D. Fu, L.F.L. Oliveira, and J. Pfaendtner. Assessing generic collective variables for determining reaction rates in metadynamics simulations. Journal of Chemical Theory and Computation, 13: 968-973, 2017. Copyright 2017 American Chemical Society.

2.2 Introduction

While molecular dynamics (MD) simulations are highly effective at exploring and describing the thermodynamics and kinetics of complex systems, its broad application can be limited due to the significant computational cost of the simulations. This expense is particularly cumbersome when recovering kinetic information because these are stochastic processes, requiring many events to be sampled. There have been many innovations to compensate for this problem, both in improvements in hardware and in methodology. To address this challenge, it is becoming commonplace to use sampling methods to recover detailed kinetics and thermodynamics of a wide range of systems.[7,28–30,32,48,49] Recently, the metadynamics (MetaD) family of enhanced sampling methods has shown to be effective in recovering the mean first passage escape times of systems that are kinetically trapped in deep free-energy minima.[19,20,35] This method, referred to as infrequent metadynamics, has been applied to describe the kinetics of a variety of systems including chemical reactions,[20] biological systems,[50–53] and has been applied to describe systems with multiple, competing pathways.[22,35]

A common challenge in executing MetaD simulations is selecting the proper degrees of freedom, or collective variables (CVs), to bias. A proper set of CVs should be able to distinguish between the different stable states of a system and capture the slowest degrees of freedom for the transitions of interest. While in some systems this can be an intuitive decision, often times the proper selection of all of the relevant CVs is not obvious and requires a significant level of knowledge regarding the different states and transitions that exist in a given mechanism. This challenge can limit the application of MetaD method as an exploratory tool because unknown transitions and states could be missed if a CV is omitted. One approach to circumvent this challenge is to sample many slow degrees of freedom simultaneously as in the original bias

exchange MetaD[54] framework, variational enhanced sampling method,[8,55] and recently introduced parallel bias MetaD.[25]

An alternative approach to mitigate this issue has come up through the use of, what we will refer to as, "generic CVs". These are CVs that are constructed following a set of rules that require minimal information about the chemistry involved in a system. Constructing generic CVs does not need to exclude chemical intuition, as this can often help optimize the simulation parameters, however there is a need for a set of CVs that require minimal chemical intuition to setup, but constrain the phase space explored in a system. In this letter we focus on two different generic CVs: social permutation invariant (SPRINT) coordinates[15] and collective variable-driven hyperdynamics (CVHD).[56,57] It has been shown that both CVs can be biased in accordance with the MetaD algorithm to accelerate a simulation and discover transitions and states of a system with a low level of chemical knowledge.[14,15,56,58] We evaluate both CVs to determine if either are a valid CV set for calculating accurate reaction rates, producing a Poisson distribution of reaction rates. This characteristic of stochastic processes, to our knowledge, has not previously been shown for simulations that biased either generic CV. We also compare how efficient these CVs are to calculate and bias.

2.3 Generic Collective Variable Overview

As the construction and theory of these CVs is explained thoroughly elsewhere,[15,48,56,57,59,60] we will only present a brief overview of the details involved. SPRINT coordinates are generated by constructing a contact matrix of the atoms in a system and use the largest eigenvalue and corresponding eigenvector through the following equations:

$$a_{ij} = \frac{1-(r_{ij}/r_0)^n}{1-(r_{ij}/r_0)^m} \qquad \text{Eq. 2.1}$$

$$v_i^{max} = \frac{1}{(\lambda^{max})^M} \sum_j a_{ij}^M v_j^{max} \quad \text{Eq. 2.2}$$

$$S_i = \sqrt{N} \lambda_{max} v_i^{max,sorted} \quad \text{Eq. 2.3}$$

where, a_{ij} is the contact matrix (constructed from the distance between atoms *i* and *j* (r_{ij}), the corresponding typical bonded distance (r_0), and common switching function integers n (6) and m (12)), a_{ij}^M is the number of walks of length M between atoms *i* and *j*, λ_{max} is the largest eigenvalue with v_j^{max} being the corresponding eigenvector, *N* is the number of atoms in the matrix, and S_i is the SPRINT coordinate of atom *i*. The "sorted" superscript indicates that the eigenvector is sorted in increasing order, with atoms of the same type together. The only information required regarding the chemistry involved is the typical bond lengths (r_0), which can easily be determined, if not already suggested by previous studies.[14,15] One decision point where chemical intuition can play a role, is selecting which atoms to include in constructing the contact matrix. While typically all atoms are used, if a system is well understood then atoms not participating in the expected transitions can be excluded and therefore reduce the computational cost of implementation.

CVHD, unlike SPRINT coordinates, only biases one CV, η, through the following relationships:

$$\chi_i = \begin{cases} 0 & if\ (r_i < r_i^{min}) \\ \frac{r_i - r_i^{min}}{r_i^{max} - r_i^{min}} & if\ (r_i^{min} < r_i < r_i^{\max}) \\ 1 & if\ (r_i > r_i^{max}) \end{cases} \quad \text{Eq. 2.4}$$

$$\chi_t = \left(\sum_{i=1}^N \chi_i^p\right)^{1/p} \quad \text{Eq. 2.5}$$

$$\eta = \begin{cases} \frac{1}{2}(1 - \cos(\pi \chi_t^2)) & if\ (\chi_t < 1) \\ 1 & if\ (\chi_t > 1) \end{cases} \quad \text{Eq. 2.6}$$

where χ_i is indicative of a local distortion (example here being the degree of distortion of a bond stretching, although other CVs could also be used. χ_t is the total or global distortion of the system where in this instance *p* was set to 8,[56] and η is the CV biased and was transformed in Eq. 6 to be differentiable at the end points of the spectrum. Like SPRINT coordinates, the construction of these terms requires knowledge of the initial state and some idea of a change that marks a transition. In the example given in Eq. 4, the transition of a bond breaking is trivially determined because the maximum and minimum bond distances of typical bonds as well as which atoms are bonded are both well known.

It is important to note that constructing both sets of generic CVs is largely dependent on knowing the initial state of the simulation. Because both CVs have been well documented in exploring and discovering transitions, we did not apply them in the most general approach, but rather in a way to optimize their efficiency. We are concerned with their ability to (1) accurately calculate reaction rates (2) yield a Poisson distribution of rates even when biased and (3) accelerate a simulation. The transition times for individual simulations were calculated following the infrequent metadynamics algorithm presented here:

$$t^{eff} = \alpha * t^{MD} \qquad \text{Eq. 2.7}$$

$$\alpha = \frac{1}{t^{MD}} * \int_0^{t^{MD}} dt' \, e^{\beta V_{bias}(s,t')} \qquad \text{Eq. 2.8}$$

where t^{MD} is the molecular dynamics time when the transition occurred, α is the acceleration factor, V_{bias} is the bias deposited at that location and time in the simulation, and β is $1/k_bT$. The infrequent metadynamics method has been shown to reduce the cost of simulations, determine the mean first passage time over a particular barrier, and estimate the mean escape time from a basin.[2,19,20,35] In order to ensure that the distributions recovered followed a Poisson distribution, we employed the Kolmogorov-Smirnov (KS) test[36] following the procedure outlined by

Salvalaglio et al.[2] This is an important step as it ensures that the distributions of times collected are uncorrelated, and, to our knowledge, neither of these generic CVs have been evaluated in the context of the KS test when rates are recovered. Error analysis was conducted following the bootstrapping procedure outlined by Salvalaglio et al.[2] and Fleming et al.[20] The bootstrapping analysis was carried out with one thousand subsets, where each subset size is 50% of the total population (drawn with replacement). Each subset was subjected to the KS test at a significance level of p=0.05, where samples that failed the test were discarded and replaced by another subset. For all of the populations collected in this study the reject rate was small and never exceeded 11%.

In addition to ensuring the consistency between recovered rates, we also compared the values of the acceleration factor (α) recovered from biasing different CVs. The α value is indicative of to what extent the simulation was accelerated relative to unbiased MD. Additionally, we also compared the computational efficiency of the runs by taking the ratio of the t^{MD} to the actual runtime, t^{run}, as this is indicative of how expensive a CV is to evaluate. These metrics help evaluate and determine if there is any loss or gain, in terms of computational expense, to using generic CVs over the reaction specific ones. We simulated an S_N2 reaction of $CH_3Cl + Cl^- \rightarrow CH_3Cl + Cl^-$ in vacuum and a Diels-Alder reaction of $C_2H_4 + C_4H_6 \rightarrow C_6H_{10}$, in which typical CVs and generic CVs were biased. Reaction rates, as well as the above mentioned simulation metrics, were recorded and compared.

2.4 S_N2 Reaction

The first system we analyzed was the S_N2 reaction of $CH_3Cl + Cl^- \rightarrow CH_3Cl + Cl^-$, as this system has obvious collective variables to bias, and has been thoroughly explored with the infrequent metadynamics method.[20] These simulations were carried out using the semi-empirical

PM6 method[3] in the AMBER[61] program with the aid of the PLUMED[41] plugin . All simulations were carried out using a Gaussian height of 0.3 kJ/mol and a bias factor of 9. In order to prevent the atoms straying far apart, we placed a wall at 0.5 nm for both C-Cl bonds. Parameters for constructing the generic CVs and their corresponding Gaussian widths are listed below, however details for parameter selection are described in SI.

For this reaction, the specific CVs biased were the two carbon-chlorine bonds, which were biased with a Gaussian width of 0.0025 nm. A transition was noted to occur when the distance between the originally bonded C-Cl atoms exceeded 0.25 nm, following what was outlined by Fleming et al.[20] For the SPRINT coordinates, we constructed a contact matrix using just the carbon and chlorine atoms, as it is clear that the hydrogen atoms will not participate in the transition. The contact matrix was calculated using r_0 values of 0.265 nm for C-C and 0.222 nm for both C-Cl and Cl-Cl. The SPRINT coordinates were biased with a Gaussian width of 0.025. A transition was noted to occur when the value of the SPRINT coordinate of the initially non-bonded chlorine atom exceeded 0.70, indicating a bond had formed. Local distortion terms for CVHD were used for the bonded and the non-bonded chlorine atoms and their distances to the carbon atom. The bonded C-Cl distortion was constructed with minimum and maximum bond lengths of 0.15 nm and 0.22 nm, respectively. The local distortion term for the non-bonded C-Cl distortion was constructed using:

$$\chi_i = \begin{cases} 0 & if\ (r_i > r_i^{mmax}) \\ \frac{r_i^{max} - r_i}{r_i^{max} - r_i^{min}} & if\ (r_i^{min} < r_i < r_i^{\max}) \\ 1 & if\ (r_i < r_i^{min}) \end{cases} \qquad \text{Eq. 2.9}$$

in order to account for the C-Cl bond forming rather than breaking. For this term, r_i^{max} was set to be 0.5 nm (where the wall was placed) and r_i^{min} was set to be 0.20 nm. The CV η was biased with a Gaussian width of 0.0250 nm. In addition to implementing CVHD using the carbon-

chlorine bond distances, it is worth noting that the formula for the local distortions is easily generalized. In this spirit, we followed the protocol of Bal and Neyts[56] and implemented CVHD by using the SPRINT coordinates as the local distortions. This was done following Eq. 9 in ref. 11 reproduced here below, except with SPRINT coordinates.

$$\chi_i = \begin{cases} \left|\frac{S_i - S_i^{ref}}{\Delta S_i}\right| & if\ (S_i \in [S_i^{ref} \pm \Delta S_i]) \\ 1 & if\ otherwise \end{cases} \quad \text{Eq. 2.10}$$

Because this system is symmetric with one chlorine atom replacing another, the initial SPRINT coordinate of one chlorine atom is the final state of the other, and vice-versa. The difference between these two values was used for the denominator, ΔS_i (0.60), and the reference values S_i^{ref} were set to be the initial SPRINT coordinates of 0.77 and 0.11 for the non-bonded and bonded chlorine atoms, respectively. The SPRINT coordinate for the carbon atom was omitted for this because the connectivity stayed relatively the same throughout the simulation. For all CVHD runs a transition was noted to occur when η equaled a value of one. For all generic CV simulations, we also tracked the distance of the originally bound C-Cl atoms (the same criteria for noting a transition when the reaction specific CVs were biased) in order to verify that our generic CVs were accurately capturing transitions.

For each set of CVs tested, we carried out simulations at temperatures of 300 K, 450 K, and 600 K. At each temperature we ran simulations employing deposition strides of 1 ps, 20 ps, and 100 ps (results for all three strides are included in SI). The *p*-values of each population exceeded 0.05, indicating that each biasing each set of CVs was able to produce a Poisson distribution. Figure 2.1 illustrates the strong agreement between the rates calculated from biasing the different sets of CVs, even for the systems where the CVHD variable η was constructed from distortions

of SPRINT coordinates. Additionally, the energy barrier and pre-exponential factors calculated from the Arrhenius plots are in strong agreement as well (shown in SI).

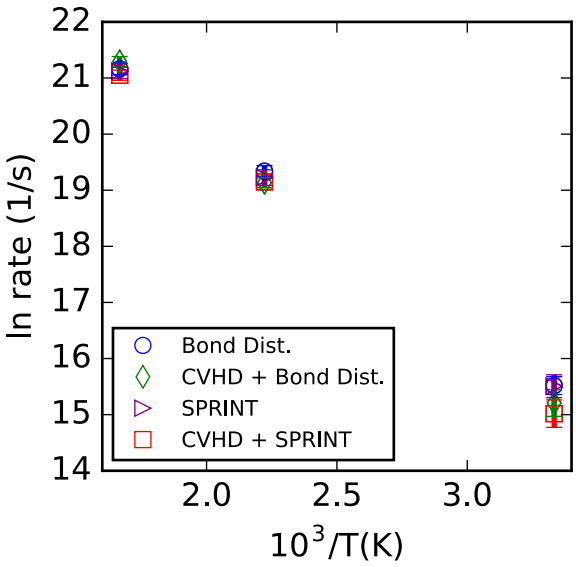


Figure 2.1: Arrhenius plot constructed from the rates recovered from biasing bond distances (blue circles), CVHD using bond distances (green diamonds), SPRINT coordinates (purple triangles), and CVHD using SPRINT coordinates (red squares). These simulations were biased with a deposition stride of 100 ps. Rates are the inverse of the mean escape time from bootstrapping.

While Figure 2.1 shows that the different CV sets are equivalent in estimating the reaction rate for this system, Table 2.1 shows that the efficiencies and speeds of the CVs varies drastically. Only the data sets corresponding to simulation conditions of 300 K with 1 ps stride and 600 K and 100 ps are shown, as these were the two ends of the spectrum in terms of bias deposited.

Table 2.1. Acceleration factors and MD efficiencies of simulations with different biased CVs for the S_N2 reaction.

System	Collective Variable Biased	Acceleration Factor α (s/s)*	MD Efficiency (fs/s)*
T = 300 K, 1 ps stride	Bond Distances	60 (3)	275 (0.3)
	CVHD + Bond Distances	274 (19)	272 (0.3)
	SPRINT	174 (11)	252 (0.3)
	CVHD + SPRINT	1107 (60)	267 (0.3)
T = 600 K, 100 ps stride	Bond Distances	1.0 (0.0)	260 (2.0)
	CVHD + Bond Distances	2.4 (0.1)	253 (1.9)
	SPRINT	1.0 (0.0)	247 (2.5)
	CVHD + SPRINT	7.5 (0.4)	264 (0.2)

*Values in parenthesis are the estimated standard deviations from bootstrapping.

Table 2.1 clearly shows that in systems where bias is essentially required in order to see a transition (300 K, 1 ps stride) and in systems that approach regular MD (600 K, 100 ps stride) that biasing the CVHD collective variable leads to larger values of α, and therefore is more effective at accelerating the simulation. This enhancement in acceleration factor for the CVHD simulations is attributed to the fact that the bias is deposited along a single dimension, η, instead of two dimensions like the two bond distances or multiple SPRINT coordinates. Furthermore, there is little difference in the computational expense of evaluating the different CVs as the spread of MD efficiencies is within tens of femtoseconds per second, which is much smaller than the spread in the α values.

Understanding the differences in acceleration factors between the simulations that biased SPRINT coordinates and bond distances is less intuitive, because the SPRINT coordinate simulations biased an extra CV, which is known to decrease the efficiency of a simulation.[21] However, we believe that, in this particular case, the free energy surface topology is responsible for this phenomenon. Figures 2.2A and 2.2B show the different free energy surface of the S_N2 reaction projected along the SPRINT coordinates and bond distances, respectively. Details of the FES calculations are provided in the SI.

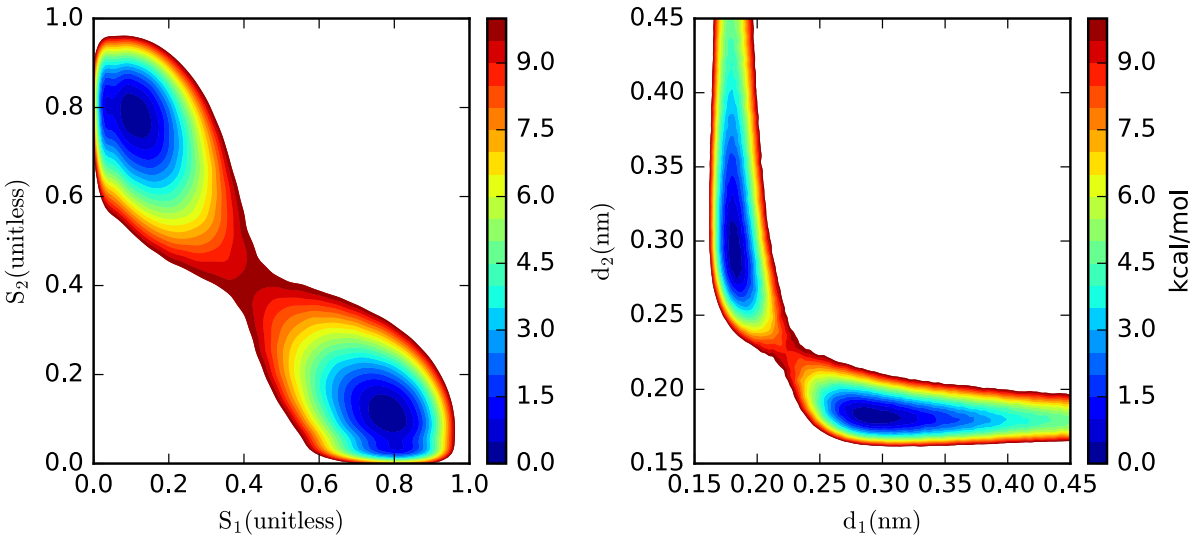


Figure 2.2: Free energy surface of the S_N2 reaction projected on the A) the SPRINT coordinates of the chlorine atoms and B) the two C-Cl bond distances.

In addition to having a larger free energy basin volume, the SPRINT coordinate system exhibited larger CV fluctuations, correspondingly a higher system diffusivity, and as a result a faster bias deposition rate (shown in SI). We believe the combination of these factors all contribute to the larger acceleration factor for the SPRINT coordinate simulations over the typical bond distances.

2.5 Diels-Alder Reaction

While the S_N2 system proved that the generic CVs have the capacity to capture the slowest, relevant degrees of freedom, it is a very straightforward process being simulated with a fairly low energy barrier, making it difficult to speculate if the generic CVs would work universally and if the efficiencies would scale. To further explore these questions, we simulated a Diels-Alder reaction of $C_2H_4 + C_4H_6 \rightarrow C_6H_{10}$ by using SPRINT coordinates, CVHD, and the specific bond distances between the reacting carbon atoms as the biased CVs, shown in Figure 2.3. This system presented the challenge of having two bonds that must form simultaneously to create the desired product without forming any undesired side products, which was not a possibility in the S_N2 system, in addition to having C-C double bonds. Because this system has more active atoms and degrees of freedom, this is a typical situation where biasing a generic CV was preferable because finding the correct reaction specific CVs took arduous sampling and experimentation to elicit.

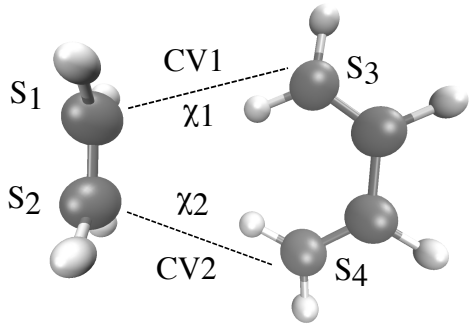


Figure 2.3: Outline of the Diels-Alder reaction with the different CV sets labeled as followed: CV1 and CV2 refer to the reaction specific CVs (bond distances), χ_1 and χ_2 refer to the local distortions used to construct η for CVHD, and the S_i's refer to the four SPRINT coordinates biased.

Each system simulated was conducted with a Gaussian height of 5.0 kJ/mol, a bias factor of 25, with walls placed at 0.5 nm for the CV1 and CV2 bond distances. For the simulations with the reaction specific carbon-carbon bonds biased (CV1 and CV2 in Figure 2.3), we used a Gaussian width of 0.01 nm for each CV and a deposition pace of 10 ps. A transition was noted to occur when both CV1 and CV2 dropped a value of 0.16 nm, which was observed to occur simultaneously. For the simulations that employed CVHD, we used a Gaussian width of 0.01 (dimensionless), and a deposition pace of 10 ps. The local distortions are the formations of C-C bonds, therefore we followed Eq. 9 with r_{max} set to 0.6 nm (to ensure neither distortion went to zero), and r_{min} of 0.16 nm, approximately the maximum length of a C-C bond. A transition was noted to occur when the CV η equaled a value of one. For the simulations that biased SPRINT coordinates, S_1 and S_2 were biased with Gaussian widths of 0.1 (dimensionless) and S_3 and S_4 were biased with Gaussian widths of 0.04 (dimensionless). SPRINT coordinates were calculated by using a contact matrix restricted to just the carbon atoms in the system with r_0 set to be 0.265 nm for all interactions, although only the four carbon atoms participating in the reaction were biased. Because the Gaussians deposited were in four dimensions, we used a deposition pace of one ps in order to expedite the simulation. A transition was noted to occur when the values of the SPRINT coordinates corresponding to S1 and S2 in Figure 2.3 exceeded values of 3.4, indicating that both carbon atoms had bonded to the accompanying butadiene molecule. Simulations were carried out at 300 K, 450 K, 600 K, and 900 K for each set of CVs biased. The Arrhenius plot generated from these simulations is shown in Figure 2.4.

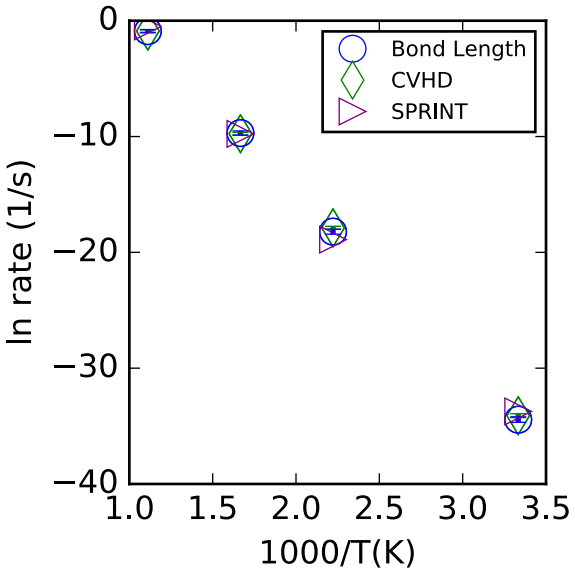


Figure 2.4: Arrhenius plot constructed from the rates recovered from biasing bond distances (blue circles), CVHD using bond distances (green diamonds), and SPRINT coordinates (purple triangles) for the Diels-Alder reaction. Error bars for the SPRINT coordinates are omitted because the sample sizes were too small to perform error analysis via bootstrapping.

As was observed for the S_N2 system, Figure 2.4 shows that the choice of which CV to bias has no effect on the recovered reaction rate as there is significant overlap in the data points plotted and strong agreement in the Arrhenius plot parameters (included in SI). It is clear that generic CVs such as SPRINT or CVHD are fully capable of capturing the slowest degrees of freedom such that the kinetics of a transition can be recovered. Table 2.2 highlights the acceleration factors and MD efficiencies of the different CVs. In the interest of brevity, we present the limiting cases of 300 K and 900 K, which were the systems that required the most and least bias, respectively, in order for the reaction to occur (other runs included in SI).

Table 2.2. Acceleration factors and MD efficiencies of simulations with different biased CVs for the Diels-Alder reaction.

Temperature	CV Biased	MetaD Boost α (s/s)*	MD Efficiency (fs/s)*
T = 300 K	Bond Distances	5.24 (1.11) *10^21	438 (0.4)
	CVHD + Bond Distances	2.13 (0.31) *10 ^22	447 (0.2)
	SPRINT**	2.05 (0.35) *10^21	426 (2.5)
T = 900 K	Bond Distances	2.74 (0.34) *10^7	430 (0.2)
	CVHD + Bond Distances	1.57 (0.13) *10^8	432 (0.2)
	SPRINT**	1.19 (0.95) *10^7	412 (5.4)

*Values in parenthesis are the estimated standard deviations from bootstrapping. **Simulations were carried out with a bias deposition stride of one ps instead of 10 ps.

From the two limiting cases in terms of bias deposited, Table 2.2 shows that CVHD is able to generate acceleration factors that almost an order of magnitude larger than those generated from biasing the reaction specific CVs or SPRINT coordinates. As noted for the S_N2 system, we attribute the enhanced acceleration from CVHD to the fact that bias is deposited along one CV, η, rather than the two bond lengths that make up the distortions. Contrary to the S_N2 reaction, we observe that the SPRINT coordinates are the least computationally efficient and yield the smallest acceleration factors, despite increasing the bias deposition pace tenfold. However, this result is attributed to the fact that four SPRINT coordinates were biased, as it is known that the efficiency of MetaD decreases exponentially with the number of CVs biased.[25] The MD efficiencies of the CVs tested are approximately the same across both temperatures, with some expected loss occurring at the 300 K simulation because these simulations were considerably longer. We will note though that because SPRINT coordinates simulations required

a four dimensional CV space, these simulations were significantly more memory intensive and proved to be an obstacle in collecting a large number of trials.

2.6 Conclusions

In this letter, we have demonstrated that SPRINT coordinates and CVHD are capable of capturing the kinetics of chemical reactions by applying these CVs to two different systems of varying complexity. These CVs, unlike those typically biased, require little intuition of the chemistry involved in a particular system, and in some situations no knowledge of the final product. When comparing the performance of these generic CVs to typical CVs, the generic CVs proved to produce the same reactions rates with comparable efficiencies. CVHD consistently produced larger acceleration factors than the reaction specific CVs, with the increase reaching up to an order of magnitude. The acceleration factors recovered from biasing SPRINT coordinates strongly depends on the number of CVs biased, with the values decreasing as the number of CVs biased increases, but its overall performance is comparable to those of the reaction specific CVs.

Therefore, biasing generic CVs with the infrequent metadynamics algorithm is not only a viable option for characterizing reaction rates of poorly understood and computationally expensive systems, but is in fact very well suited for such applications.

3 Lifting the Curse of Dimensionality on Enhanced Sampling of Reaction Networks with Parallel Bias Metadynamics[3]

3.1 Abstract

A common challenge to applying metadynamics to the study of complex systems is selecting the proper collective variables to bias. The advent of generic collective variables, specifically social permutation invariant (SPRINT) coordinates, has helped addressed this challenge by reducing the level of a priori knowledge required to just basic chemical fundamentals. However, the efficiency of biasing SPRINT coordinates can be severely handicapped by the high dimensionality of the bias potential. Here we circumvent this deficiency by biasing SPRINT coordinates using the parallel bias metadynamics framework. We demonstrate the efficacy of this method to efficiently explore a complex system, without any prior knowledge about transition pathways, by applying it to study the decomposition of γ-ketohydroperoxide and generating a comprehensive reaction network of relevant pathways. The reduction in both computational cost and chemical intuition makes this method a promising option for studying complex reacting systems.

[3] Reproduced in part with permission from C. D. Fu and J. Pfaendtner. Assessing generic collective variables for determining reaction rates in metadynamics simulations. Journal of Chemical Theory and Computation, 14: 2516-2525, 2018. Copyright 2018 American Chemical Society.

3.2 Introduction

Elucidating the relevant pathways and elementary steps that participate in reacting systems such as combustion, pyrolysis, and atmospheric chemistries, to name a few, is an active area of research from both an experimental and computational standpoint. Completely resolving these intricate reaction networks through experimental methods is typically challenging as the number of species can easily extend well into the thousands and involve many parallel, and competing, pathways. Exploring such systems through computational means offers the promise of providing insights into the various elementary steps that are common to these processes, but difficult to observe. Additionally, continuing improvements in both hardware and methodology make computational exploration a much more appealing choice for exploring potential energy landscapes.

Molecular dynamics (MD) simulations are capable of capturing the individual mechanisms of these systems at high resolution, and have been shown to describe a variety of complex reacting systems.[62–64] However, because it is not unusual for such systems to have high energy barriers, to contain multiple pathways, or to require a high-level of theory to accurately define the potential energy surface (PES), MD simulations are usually handcuffed to a significant computational cost that is impractical for most applications. Improvements in software[65] and the use of semi-empirical methods[4] and reactive forcefields[5,6] have made longer timescales more accessible, but such methods are not always suitable. Another means to reduce the computation cost, without sacrificing chemical accuracy, is the use of enhanced sampling methods such as metadynamics

(MetaD),[7] boxed molecular dynamics,[10] umbrella sampling,[9] hyperdynamics,[32,66] bond-boost,[48] and variationally enhanced sampling.[8]

Specifically, MetaD has been shown to effectively reduce the simulation time required to characterize a system, but still allow for desired properties like free energy surfaces (FES), PESs, and mean first passage times over barriers to be recovered.[7] In MetaD a history dependent bias potential, constructed on the fly, acts on a few user-defined collective variables (CVs) to discourage a system from re-visiting states and make other, normally rarely visited, states more accessible. While beneficial in characterizing pathways, a challenge to using MetaD is that the chosen CVs biased must be able to distinguish between the different states and capture the slowest degrees of freedom of transitions, which requires a significant level of chemical intuition about the system (i.e. products, reaction coordinates). Additionally, in the context of studying multiple reactions in a mechanism, what may be a viable set of CVs for the first reaction, may not be suitable for the subsequent steps, restraining this approach to only sample individual steps at a time. While other studies have focused on methods for identifying promising combinations of CVs,[67] recent advancements in the development of so-called "generic CVs" offer promise in overcoming this challenge.

Unlike typical CVs (e.g. bond distances, reaction coordinates), generic CVs are calculated following a general set of equations that only require information about the current state of the system, without direct knowledge of potential product states, reducing the level of chemical intuition required. One example of a generic CV is the one used in collective variable-driven hyperdynamics[56,58] (CVHD), which is based on the SISYPHUS method.[57] CVHD involves collapsing many local distortions into a single CV (η), which is biased, and has been shown to discover reaction pathways. Even though this global CV can encompass many local CVs and can

extend to sample multiple reactive events in a given simulation, this still requires the user to identify all of the relevant fluctuations at the onset, which can be challenging for systems with many possible CVs and unforeseen transitions. Another type of generic CV, the social permutation invariant (SPRINT) coordinates[15] are calculated by using the equilibrium distances between atom types and the distances between each of the atoms in a system to construct a contact matrix, along with additional operations (see Methods section). Unlike CVHD, which biases one CV that is tailored to a specific system, SPRINT coordinates are calculated through the same process, regardless of system, and there is one coordinate per atom accounted for in the contact matrix. Both of these generic CVs have been biased following the MetaD framework to discover and describe the kinetics of reactive pathways, as well as explore reactive processes.[14,15,23,56,58]

While biasing SPRINT coordinates offers the benefit of being straightforward to apply with little required knowledge about (meta)stable states in the system, a potential downside is dimensionality explosion in the bias potential as each SPRINT CV (1 per atom) adds an additional dimension. Because the efficiency of MetaD is known to decrease exponentially with the dimensionality of the bias potential,[21] biasing SPRINT coordinates either yields a diminishing return, in terms of efficiency, or forces the user to apply overly aggressive bias parameters (e.g., large hill height, large hill widths, fast deposition pace).[14,15,23] A previous study by Zheng and Pfaendtner,[14] which biased SPRINT coordinates to explore high temperature methanol oxidation, specifically acknowledged the need to improve the efficiency of the method, despite only having a seven-dimensional bias potential with an aggressive hill height of 2.8 kcal/mol and hill width of 1.5 (unitless). Such issues would only further be compounded when applied to larger systems.

In this paper, we address this issue of dimensionality by biasing the SPRINT coordinates following the parallel bias MetaD (PBMetaD) framework introduced by Pfaendtner and Bonomi.[25] Contrary to typical MetaD, PBMetaD utilizes multiple low dimensional biases to sample and reconstruct the FES of a high dimensional CV space, rather than a single high dimensional bias potential.[25] This approach has previously been demonstrated to reconstruct the FES of a variety of systems,[25,68–70] but, to our knowledge, has not been applied to act on generic CVs, specifically SPRINT coordinates. Herein, we apply the PBMetaD method to bias the SPRINT coordinates of all of the atoms in the system to study the decomposition of γ-ketohydroperoxide (KHP) and the elementary reactions that follow. Modeling this system we demonstrate the effectiveness of this method to explore energy surfaces and discover transition pathways efficiently without requiring chemical insight into the reaction pathways. Furthermore, using an ensemble of these simulations, we recover a reaction network that encompasses the relevant reaction pathways. The methodology presented in this workflow is independent with the level of theory used to describe the chemistry and should be suitable to describe a variety of complex reacting systems.

The paper is organized as follows. We provide a brief overview of the PBMetaD theory, how the SPRINT coordinates were constructed, and simulation details. Following this, the reaction pathways we sampled are presented and the impact of certain simulation parameters are explored and discussed. The paper concludes with our recommendation for how this framework should broadly be applied to recover a comprehensive reaction network and a perspective of future applications of this method.

3.3 Methods

The system simulated in this study was the decomposition of KHP in vacuum, as well as reactions stemming from 1,2-dioxolan-3-ol (CYCP), which is commonly formed from KHP. While previous studies of the Korcek reaction mechanism typically used higher levels of theory,[71–73] we modeled these simulations with the semi-empirical PM6 Hamiltonian,[3] using the AMBER 14[61] program, because the focus of this study is regarding the development of a methodology, rather than the gaining insight into PES of KHP. Presence of transition states and reaction pathways were therefore verified with additional calculations using the Gaussian 09[74] program to ensure our results self-consistent with the PM6 Hamiltonian, instead of artifacts of the bias. The PLUMED[41] plugin was used to evaluate and bias the SPRINT coordinates of all the atoms in the system following the PBMetaD framework. Because the procedures for calculating SPRINT coordinates and carrying out PBMetaD are provided elsewhere,[15,25] we will only provide a brief overview of the key details and equations involved.

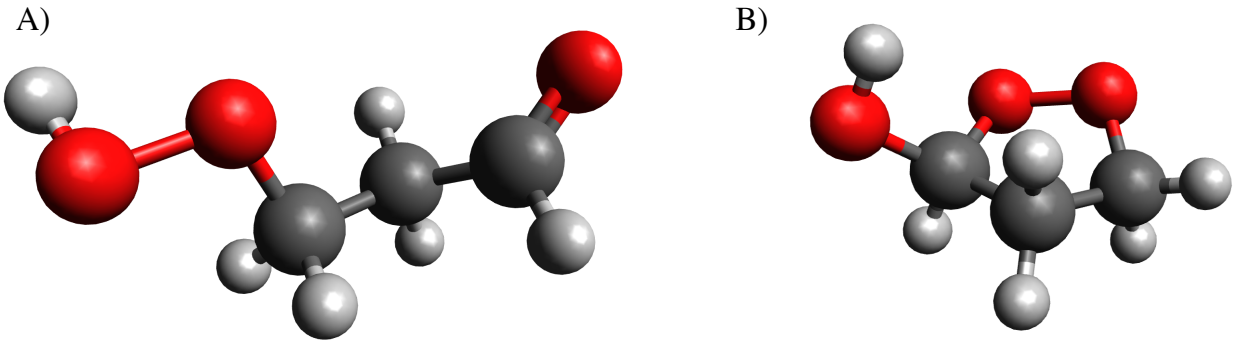


Figure 3.1: Images of the different starting species for simulations A) γ-ketohydroperoxide (KHP) and B) 1,2-dioxolan-3-ol (CYCP).

SPRINT coordinates are evaluated by first constructing a square contact matrix, *a,* using the following switching function to define entry a_{ij} for a pair of atoms

$$a_{ij} = \frac{1-(r_{ij}/r_0)^n}{1-(r_{ij}/r_0)^m} \qquad \text{Eq. 3.1}$$

where r_0 is usually selected based on an average of equilibrium bond distances between specific pairs of atom types, r_{ij} corresponds to the actual distance between two atoms (i and j), and the values of n and m are switching function integers, in this case set to be 6 and 12, respectively. The values for r_0 were set to be 2.65 Å (for C-C and C-O) or 2.22 Å (for C-H, O-H, H-H) depending on which atomic interaction was being described.[14] Note that this matrix is both symmetric and non-negative.[15] Selecting proper values for r_0, which can trivially be determined, is the only chemical insight needed to calculate SPRINT coordinates. No information regarding potential reactions is required. Next, this matrix is transformed into a vector through the following operation

$$v_i^{max} = \frac{1}{(\lambda^{max})^M} \sum_j a_{ij}^M v_j^{max} \qquad \text{Eq. 3.2}$$

where λ^{max} is the largest eigenvalue, v_j^{max} its corresponding eigenvector, and a_{ij}^M is the number of walks of length M between atoms i and j. Lastly, the SPRINT coordinates, S_i, are then calculated from the following equation

$$S_i = \sqrt{N}\lambda_{max} v_i^{max,sorted} \qquad \text{Eq. 3.3}$$

where N is the number of atoms in the contact matrix, and the "sorted" superscript indicates that the eigenvector is sorted in increasing order where identical, indistinguishable atom types are grouped together. However, because PBMetaD creates mono-dimensional bias potentials for each SPRINT coordinate (i.e. one for each index of the vector S), re-ordering within groups can potentially lead to different bias potentials to be applied to SPRINT coordinates that are associated with different atoms over the course of the simulation. We considered the potential consequences of sorting and not sorting within atom types and attempted both methods. Because it is the re-ordering step that provides the invariant nature of the SPRINT coordinates, not re-ordering the vector just treats all atoms as distinguishable from one another and should have not

have a detrimental impact on the result, save maybe a loss in efficiency. Whereas exchanges of bias potentials solely based on the index in the vector is different from other bias exchange schemes[54,75] and thus its impact was investigated.

In order to circumvent the obstacle of dimensionality, all 12 of the SPRINT coordinates were biased following the PBMetaD method, rather than classical MetaD. A classical MetaD simulation involves constructing a single, *n*-dimensional bias potential, where *n* refers to the number of CVs biased as shown here

$$V_G(s_i, t) = \int_0^t dt' \, W(t') \, \exp\left(\sum_{i=1}^{n} \frac{-(s_i(R) - s_i(R(t')))^2}{2\sigma_i^2}\right) \qquad \text{Eq. 3.4}$$

where V_G is the constructed bias potential, $d\square'$ is the deposition stride, s_i refers to one CV of the *n* that are biased, σ_i refers to the width of the Gaussian along dimension *i*, and $W(t')$ refers to the hill height deposited at time t'. In the well-tempered variant, $W(t')$ is scaled down as the bias accumulates shown here

$$W(t') = W * \exp\left(-\frac{V_G(s,t)}{k_B \Delta T}\right) \qquad \text{Eq. 3.5}$$

In contrast with these forms, PBMetaD instead employs a series of low-dimensional bias potentials that operate on individual CVs. Therefore, instead of having one *n*-dimensional bias potential, an equivalent PBMetaD simulation would employ *n* 1-dimensional bias potentials. The individual bias potentials follow the same form as the WTMetaD, except an additional, so-called "conditional weight", term is used to modify the hill height of an individual bias potential based on the value of that CVs bias, relative to the bias deposited along the other CVs shown here

$$W_i(t') = W * \exp\left(-\frac{V_G(s_i,t)}{k_B \Delta T}\right) * \frac{\exp\left(-\frac{V_i(s_i,t)}{k_B T}\right)}{\sum_{j=1}^{n} \exp\left(-\frac{V_j(s_j,t)}{k_B T}\right)} \qquad \text{Eq. 3.6}$$

here W refers to the initial hill height, the first exponential term refers to the scaling from the well-tempered variant (shown in Eq.5), and the conditional weight is given by the fraction of the

negative of exponent of the individual CV bias (normalized by k_BT), divided by the sum of these for all of the CVs biased. The conditional weight acts to deposit larger hills along CVs with lower values of bias at the time of deposition.

3.3.1 Simulation Details

The PBMetaD simulations utilized a bias factor of 75 and a deposition stride of 1 ps. We applied a harmonic restraint on the radius of gyration of all of the atoms in the system to prevent species from drifting too far apart, in violation of the Perron-Frobenius theorem, which would result in the SPRINT coordinates of multiple atoms to drop to zero. In the context of SPRINT coordinates, the Perron-Frobenius theorem specifies that the maximum eigenvalue, λ^{max}, has the properties of being real, positive, and nondegenerate while its corresponding eigenvector, v_j^{max}, is comprised on non-zero elements, provided that the system (i.e. matrix a_{ij}) can be represented as a connected graph.[15] The upper limit restraint was implemented using PLUMED and was set to be at 0.3 nm with a strength of $1*10^7$ (kJ/mol)/nm^2 in order to ensure this. In order to explore the impact of MetaD parameters on the recovered networks, we ran simulations that were biased with an initial hill height of 1 kJ/mol and a sigma value of 0.1 (unitless) and another set with an initial hill height of 0.5 kJ/mol and a sigma value of 0.05 (unitless). While the sigma value is typically set to be a fraction of the mean fluctuation of a CV (typically 1/3-1/2), it is likely that CV fluctuations will significantly change following reactive events, thus we compare the results from using these different parameter sets. Simulations that were carried out with the aggressive bias parameters (1 kJ/mol hill and 0.1 sigma) were run for 200, 150, and 100 ns at temperatures of 300, 600, and 800 K, respectively. We added an additional 50 ns of run time for simulations carried out with the more conservative bias parameters (0.5 kJ/mol hill and 0.05 sigma) to account for the anticipated slower biased transition rates.

Following the scheme where SPRINT coordinates are not re-ordered (i.e. treating each atom as being distinguishable regardless of type), we initiated simulations from KHP for each set of conditions noted above. These simulations were carried out with a time step of 0.001 ps and employed a Langevin thermostat with a collision frequency of 10 ps^{-1}. Because distinguishing the atoms from one another would only have the drawback of being less efficient, due to the removal of the invariant aspect, we exhaustively sampled each of these systems with a minimum of 45 reactive trajectories, each of which was run between 100-200 ns depending on the temperature.

Additionally, we studied the impact of the initial reactant on the structure of the recovered reaction network by starting simulations from CYCP (see Figure 3.1), a common product formed from KHP. These simulations were biased with the conservative set of PBMetaD parameters (0.5 kJ/mol hill and 0.05 sigma). Due to the high number of competing pathways observed that extend from CYCP, many of which form bi-molecular or tri-molecular products, we relaxed the restraint on the radius of gyration to 0.5 nm in order to ensure this was not impacting the pathway selectivity. Similar to the KHP decomposition, we carried out these simulations without re-ordering the SPRINT coordinates and recovered a minimum of 45 trajectories per system.

We also investigated the impact of re-ordering the SPRINT coordinates. Under this scheme, we launched simulations starting with KHP and CYCP, using the aggressive and conservative bias parameters respectively, at all three temperatures. Because we had a substantial amount of data for simulations without re-ordering, only a subset of the total simulations was required to identify any differences. Therefore, we sampled 16 trajectories per system investigated under the re-ordering scheme.

3.3.2 Merging Trajectories into Networks

Reactive events in a trajectory were detected by monitoring changes in the neighbor lists of each atom using the MDTraj[76] Python package. Changes that persisted for more than one ps (1000 steps, or 10 consecutive trajectory frames) were noted to be events and the structures were analyzed using the Open Babel code.[77] Trajectories were further cleaned to weed out false transitions that were typical of systems with more than one species, which is explained further in the SI.

For an individual trajectory, the different species observed were set as nodes and edges connecting the nodes were drawn between species that followed each other in a trajectory. Note forward and reverse reactions were labeled as different pathways. Trajectories were then accumulated together by adding new nodes and edges to the network. Networks were visualized using the Graphviz[78] Python package. We will note that in the case of forming ethylene and performic acid from KHP, the ethylene and performic acid step was sometimes omitted, instead just showing the oxirane and formic acid product, due to the lifetime of the ethylene and performic acid was less than 1 ps. However, we were able to recover these species by performing NEB calculations[43,79] between KHP and the oxirane and formic acid product.[43]

To gain insight into the relevancy of various pathways and a basis for comparing the networks generated for the different systems tested, we performed an error analysis on these networks by applying a bootstrapping procedure. For a given system with M number of trajectories, we randomly selected M trajectories with replacement and assembled this sample into a reaction network. For each pathway, we counted the number of different trajectories, within the sample, the pathway was present in and assigned a value of zero to any pathway present in the complete network, but absent in the network constructed by the sample of trajectories. These counts were

then normalized by M and multiplied by 100 to yield what we term the "justified presence" (JP) for each path.

$$JP_i = \frac{Number\ of\ Trajectories\ Pathway\ i\ is\ Observed}{M} * 100 \qquad \text{Eq. 3.7}$$

We repeated this process 1000 times for each system and recorded the mean and standard deviation of the JP values for each pathway. In this study, we used this JP metric as a facsimile for relevancy in the constructed networks. Pathways with a mean JP value of three or higher were determined to be relevant to the network, as this corresponds with being present in more than one trajectory on average. Pathways with a mean JP value below this cutoff were omitted, as these were either rarely sampled or appeared on the fringes of the network and are, therefore, not critical to the construction of the core network.

3.4 Results and Discussion

3.4.1 Analyzing Networks

Reaction networks comprised of the pathways with a JP value exceeding three for simulations starting with KHP, at 800 K, and biased with the aggressive MetaD parameters are presented in Figure 3.2 and Figure 3.3. After we identified the relevant species and pathways, we verified their "true presence" by performing transition state searches using the Gaussian 09 program,[74] followed by intrinsic reaction coordinate (IRC) calculations, and additionally calculated the energy barriers of the discovered pathways (shown in Figures 3.2 and 3.3) all of which were done using the PM6 Hamiltonian. As it is clear from Figures 3.2 and 3.3, biasing SPRINT coordinates following the PBMetaD algorithm allows for us to recover an extensive reaction network, with a variety of chemistries, which are consistent with the PM6 Hamiltonian used to describe the chemistry. While we only show subset of the pathways sampled, Figures 3.2 and 3.3 also illustrates how we are able to sample multiple reactive events per trajectory, as the network

contains multiple levels and pathways, but also capture important details like reversible steps. Furthermore, the range of the energy barriers sampled extends up to 50 kcal/mol, which are often difficult to sample in standard MetaD simulations, let alone with unbiased MD. We attribute the accessibility of the multiple high-energy pathways in individual simulations to the continued accumulation of the parallel bias potentials, which act to dramatically reduce the simulation time required to sample pathways as the simulation progresses. Unlike the CVHD method, which monitors the reaction via the CV (event occurs when η=1) and restarts the bias deposition on a newly defined global distortion term, our method can allow for bias to continue to accumulate throughout a simulation and relegate reactive event identification to a post-processing task, rather than on the fly. Additionally, similar core pathways are observed across the three temperatures simulated. The only significant deviation between the temperatures is that only at 800 K are we able to sample other pathways that compete with the formation of CYCP from KHP. However this is not surprising because the formation of CYCP is approximately 18 kcal/mol, which is the energetically lowest pathway by ~10 kcal/mol. Additionally, even when the competing pathways are sampled, they typically occur after CYCP has formed and then undergone the reverse reaction back to KHP. As anticipated, we observe that higher temperature systems allow us to proceed deeper into the network, as the rates of reaction of reduced, and higher energy barrier pathways become sampled as the selectivity toward the lower barrier pathways decreases.

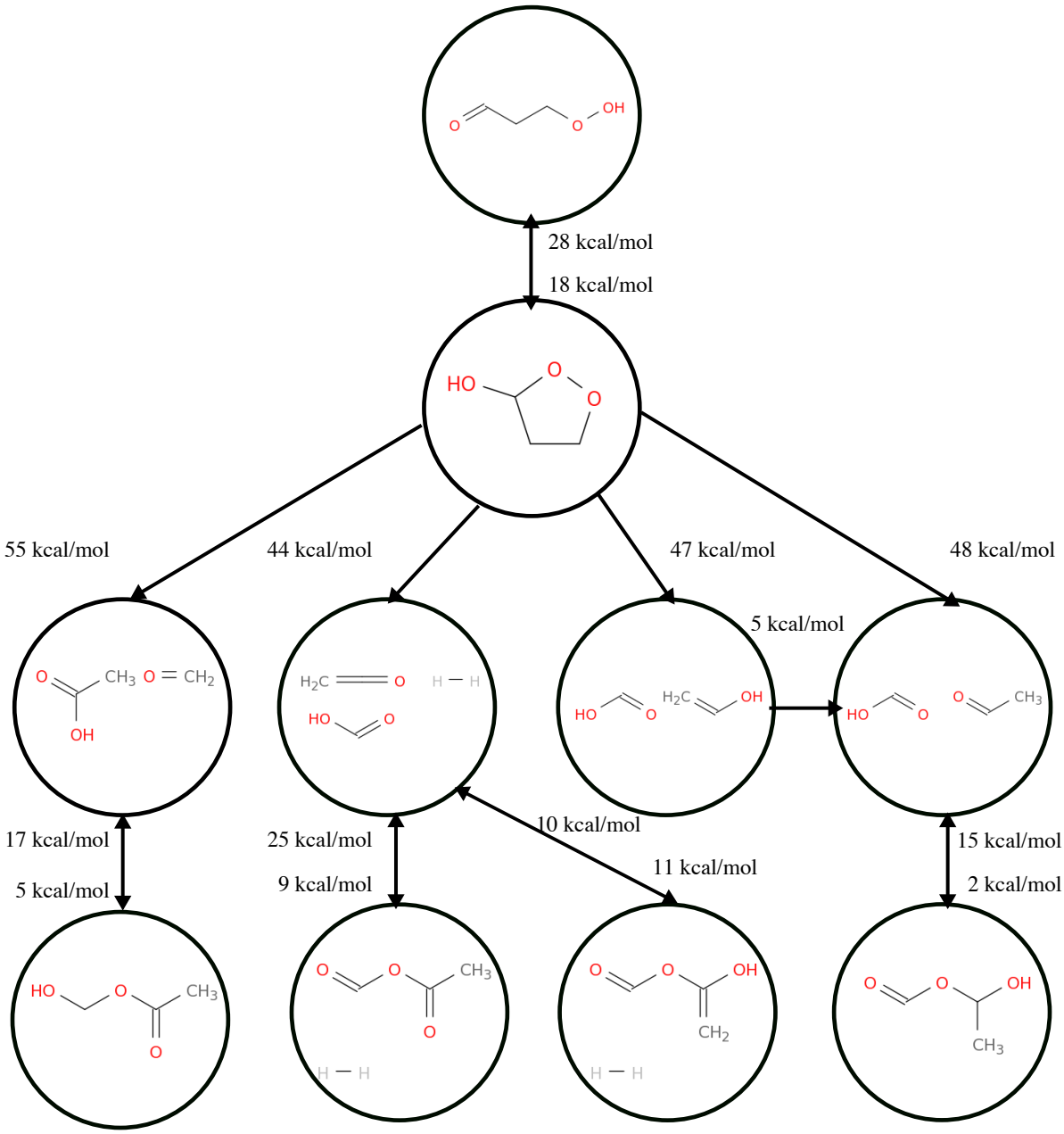


Figure 3.2: Reaction network for simulations initiated from KHP at 800 K proceeding through CYCP, biased with the aggressive bias parameters. Only pathways the proceed after CYCP with a JP value exceeding 3.0 in within two transition events of KHP are shown. Forward reactions proceed top to bottom, and vice-versa for reverse reactions. Energy barriers, calculated with Gaussian 09, are provided for transitions sampled.

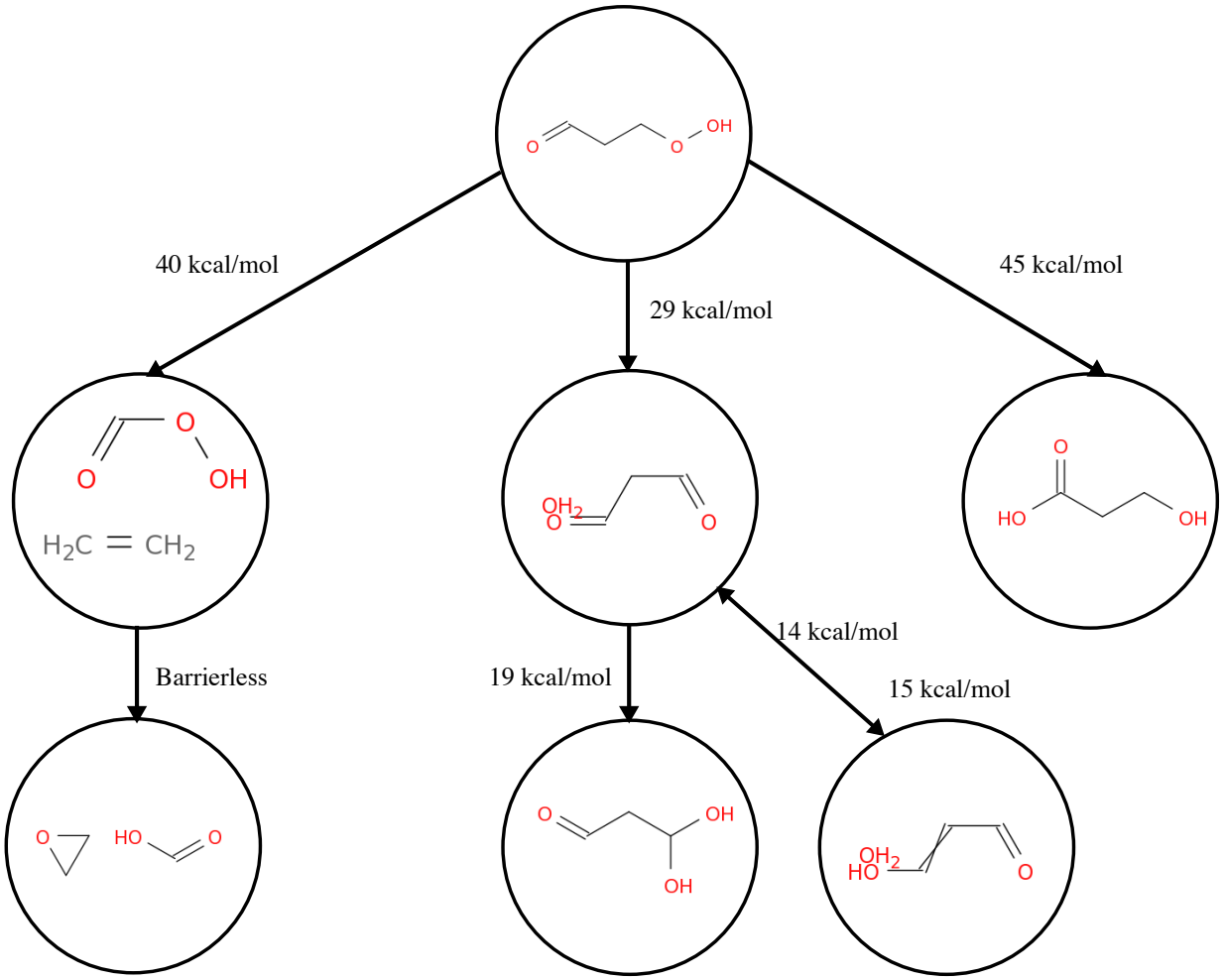


Figure 3.3: Reaction network for simulations initiated from KHP at 800 K proceeding through other pathways, biased with the aggressive bias parameters. Only pathways with a JP value exceeding 3.0 in within two transition events of KHP are shown. Forward reactions proceed top to bottom, and vice-versa for reverse reactions. Energy barriers, calculated with Gaussian 09, are provided for transitions sampled.

While able to sample a variety of chemistries, we note that, when biasing SPRINT coordinates, users should use caution in instances where the system breaks down into multiple, smaller molecules (> 2). We observed a small fraction of the simulations, in which three or more

species formed, yielded some false transition events, and these transitions were omitted from our results (see SI). We attribute this finding to the fact that, even with a restraint on the radius of gyration, the reactants drifted too far apart and violated the Perron-Frobenius theorem. This was further verified as multiple SPRINT coordinate values dropped to zero, which additionally manifested itself into an unconverged energy calculation for those steps in the simulation. Therefore, it is important to reiterate the necessity of a restraint or adjustment of the switching function parameters in order to ensure the system is always "connected" when biasing SPRINT coordinates.

In addition to its importance to hydrocarbon oxidation, we opted to the model the decomposition of KHP because the main reaction pathways, via the Korcek reaction mechanism, have previously been characterized, as well as other additional pathways, which can act as a benchmark for the performance of our method.[73] Specifically, a previous study by Suleimanov and Green characterized reaction pathways stemming from KHP and CYCP at an M062X/6-311++G* level of theory using a combination of graph theory, the freezing string method, and the Berny algorithm.[72] Even though we modeled our system using a semi-empirical method, we initially verified that the PM6 Hamiltonian could describe the pathways listed by Suleimanov and Green, using the Gaussian 09 program. We are able to find many of the pathways that were proposed by Suleimanov and Green, including the key pathways of the Korcek reaction mechanism,[73] as well as other documented chemistries.[80] As Table 3.1 shows, the pathways that are missing in our reaction network are consistently the higher barrier pathways (when modeled with PM6), which is not surprising, but actually anticipated, as these pathways would have very low selectivities compared to the other pathways sample. Also, as Figures 3.2 and 3.3 show, our method produced additional competing pathways stemming from KHP and CYCP, likely due to

the use of a different Hamiltonian, which would further lower the relative selectivities for those higher energy barrier pathways. Therefore, even though our reaction network is missing some branches, we are able to capture the kinetically relevant, lower energy barrier pathways, which are significant to the network for the temperature ranges we are simulating. The ability to capture finite temperature effects and provide a commentary of the relevancy of pathways is an advantage of this approach.

Table 3.1: Pathways discovered by Suleimanov and Green,[72] along with the barrier heights defined by the PM6 Hamiltonian and whether or not the pathway was sampled in our reaction network.

Reactant	Product	$\Delta E^{\neq}_{PM6}$ (kcal/mol)*	Status
γ-ketohydroperoxide	1,2-dioxolan-3-ol**	18.3	Found
	Succinaldehyde + Water	28.6	Found
	Ethylene + Performic Acid	40.3	Found
	Formic Acid + Acetaldehyde	44.4	Missed
	O=C=CH-CH2-O-OH + H2	68.5	Missed
	HO-O-CH2-CH=CH-OH	52.5	Found***
1,2-dioxolan-3-ol**	γ-ketohydroperoxide**	27.7	Found
	Formic Acid + Acetaldehyde**	47.6	Found
	Acetic Acid + Acetaldehyde**	54.5	Found
	3,3-dihydroxypropanal	55.2	Missed

*Energies include zero point energies (unscaled).

**Energies reported are the Boltzmann averages of the pathways for both enantiomers.

***JP value below 3.0.

An additional, advantageous feature that arises from using MD to discover pathways is the ability to natively capture differences in chirality and distinguish between enantiomers. As Figure 3.4 shows, CYCP can actually form one of two enantiomers, based on the relative position of the alcohol group. As expected, KHP can react to form either isomer, and both isomers are capable of continuing on to any of the sampled pathways as shown by Table 3.2. The

presence of an additional low energy pathway connecting KHP to CYCP would explain why this progression was so heavily sampled in our trajectories. While Table 3.2 shows that the overall impact of the chiral species the on the overall kinetics or progression of the network is not significant (energies differ by a few kcal/mol on average), being able to elicit competing pathways which produce chiral species was not addressed in the study by Suleimanov and Green, and it is unclear how other path finding methods would fare. Therefore, this framework is potentially very well suited for studying reaction mechanisms that do not produce racemic product mixtures.

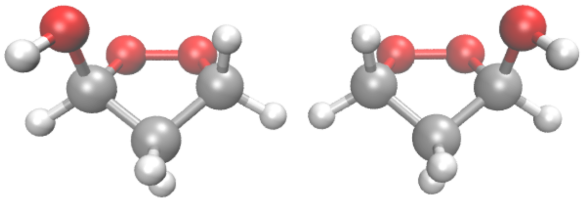


Figure 3.4: The S and R enantiomers of the 1,2-dioxolan-3-ol.

Table 3.2: Barrier Heights for Pathways Away from the Enantiomers of 1,2-dioxolan-3-ol.

	PM6 Energy Barrier Heights* (kcal/mol) for Enantiomers	
Product	R-CYCP	S-CYCP
Ethenone + Formic Acid + H2	46.0	43.8
Acetic Acid + Formaldehyde	55.7	54.1
Acetaldehyde + Formic Acid	48.3	47.2
KHP	29.4	27.2
Vinyl Alcohol + Formic Acid	47.6	46.2

*Zero point energies (unscaled) included

3.4.2 Investigating Bias Parameters

Due to the variety of species and chemistries that can occur in given system, a potential obstacle in applying this method is properly selecting the MetaD parameters to not only characterize one transition, but several transitions. In order to determine how sensitive our network results are with regard to the bias parameters, we ran the same set of simulations but reduced the hill height to 0.5 kJ/mol and the sigma value for each SPRINT coordinate to 0.05 (unitless). We only reduced the bias, rather than increase it, to verify that using such an aggressive parameter set did not result in lower energy barrier pathways to be missed or blocked. Because we are mainly concerned with how the bias parameters impact sampling of the key, competing pathways, we will restrict discussion to pathways stemming from KHP and CYCP. We present the results of the JP values of pathways leading away from KHP and CYCP for systems starting with the KHP in Tables 3.3 and 3.4, respectively. When analyzing the impact of the bias parameters in Table 3.3, we observe that our method is very robust with regard to which pathways are sampled as well as their JP values. For both the aggressive and conservative bias parameter sets, the most sampled pathway is to one of the CYCP isomers. Only at 800 K is another pathway, with a JP value exceeding three, leading away from KHP observed. Across all three temperatures, deviations between the JP values of pathways are within the bootstrapped uncertainty, thus demonstrating that, at least with regard to sampling the first reactive event, the bias parameters have a negligible impact on pathway selectivity. Even when analyzing the distribution of reactive events stemming from CYCP, the results from both sets of bias parameters are very similar. Note the lack of sampling at 300 K using the conservative bias parameters is because most of the simulations only captured the KHP to CYCP transition before terminating and thus truly highlights the incredible impact the bias has on expediting the sampling of reactive events. However, at the higher temperatures, when simulations were able to

evolve past CYCP, we again observe very consistent results between both systems, typically within the bootstrapped uncertainty. The main deviation between the two bias parameters can be seen for the reverse reaction back to the original KHP species, which shows a difference of two standard deviations. However, because MetaD uses a history-dependent bias potential to discourage systems from re-visiting states, we suspect that the aggressive parameter simulations deposited more bias in phase space occupied by the KHP species, making this reverse pathway less sampled. The presence of the bias would also explain why the reverse reaction is sampled less frequently, despite it having the lowest energy barrier. This phenomenon leads us to conclude that there is an important tradeoff when selecting bias parameters. More aggressive parameters will allow for more events to be sampled in a given trajectory, as shown by the 300 K results, but this comes at the expense of potentially missing reverse reactions.

Table 3.3: Justified presence for sampled pathways leading away from KHP for different temperatures and bias parameters. KHP is the starting structure for these simulations.

		Justified Presence for Ketohydroperoxide Reactant Pathways				
Bias Parameters	Temp (K)	S-CYCP $\Delta E^{\neq} = 18.0$	R-CYCP $\Delta E^{\neq} = 18.6$	Succinaldehyde + Water $\Delta E^{\neq} = 28.6$	Ethylene + Performic Acid $\Delta E^{\neq} = 40.3$	3-hydroxypropionic Acid $\Delta E^{=} = 44.6$
1 kJ/mol Hill $\sigma = 0.1$	300	54 (7)	46 (7)	0	0	0
	600	39 (7)	63 (7)	0	0	2 (2)
	800	62 (7)	44 (7)	12 (5)	4 (3)	4 (3)
0.5 kJ/mol Hill $\sigma = 0.05$	300	56 (7)	44 (7)	0	0	0
	600	49 (7)	53 (7)	0	0	2 (2)
	800	58 (7)	49 (7)	19 (6)	2 (2)	0

* Values in parenthesis are from bootstrapping.

** Energies shown are calculated using the PM6 Hamiltonian with zero-point energies (unscaled)

Table 3.4: Justified presence for sampled pathways leading away from CYCP for different temperatures and bias parameters. KHP is the starting structure for these simulations.

		Justified Presence for 1,2-dioxolan-3-ol Reactions					
Bias Parameters	Temp (K)	KHP $\Delta E^{\neq} = 27.7$	Vinyl Alcohol + Formic Acid $\Delta E^{\neq} = 46.6$	Acetal-dehyde + Formic Acid $\Delta E^{\neq} = 47.6$	Ethenone + Formic Acid + H_2 $\Delta E^{\neq} = 44.3$	Ethylene + Performic Acid $\Delta E^{\neq} = 52.3$	Acetic Acid + Formal-dehyde $\Delta E^{\neq} = 54.5$
1 kJ/mol Hill $\sigma = 0.1$	300	0	34 (7)	15 (5)	6 (4)	4 (3)	39 (7)
	600	8 (4)	16 (5)	21 (6)	4 (3)	2 (2)	52 (7)
	800	29 (7)	15 (5)	8 (4)	15 (5)	0	38 (7)
0.5 kJ/mol Hill $\sigma = 0.05$	300	0	0	0	0	0	4 (3)
	600	12 (5)	24 (6)	6 (4)	6 (4)	0	51 (7)
	800	43 (7)	13 (5)	15 (5)	2 (2)	2 (2)	30 (7)

* Values in parenthesis are from bootstrapping.

** Energies shown are calculated using the PM6 Hamiltonian with zero-point energies (unscaled) and Boltzmann averaged for the two enantiomers.

3.4.3 Impact of Starting Species

In addition to testing the impact of the bias parameters, we analyzed the impact of the initial starting species as well by starting simulations from the S-CYCP species. Because the PBMetaD utilizes many, in this case 12, low dimensional bias potentials, it is possible that changing the starting species could influence how the bias accumulates, and therefore have an effect on which pathways are sampled. We present the JP values for pathways leading away from S-CYCP species for simulations initiated from that state in Table 3.5. Once again, in context of capturing pathways, we see that the same pathways are sampled when simulations are launched from S-CYCP instead of KHP, demonstrating the robustness of this method. However, the distribution of JP values is very different than that of the equivalent bias parameters in Table 3.3. The acetic

acid and formaldehyde pathway, which was heavily sampled for simulations initiated from KHP, now shows up sparingly and, instead, the acetaldehyde and formic acid pathway is sampled more. We attribute this shift to the removal of bias that accumulated during the KHP stage. The pathway leading to acetic acid and formaldehyde has the highest energy barrier, and thus was likely more accessible during the initial KHP simulations because of the accumulated of bias prior to the formation CYCP. For systems initiated from CYCP, this accumulated bias is absent, and thus makes the acetic acid and formaldehyde pathway less accessible.

Table 3.5: Justified presence for sampled pathways leading away from S-CYCP for different temperatures for simulations starting from S-CYCP.

		Justified Presence for 1,2-dioxolan-3-ol Reactions					
Starting Species	Temp (K)	KHP $\Delta E^{\neq} = 27.7$	Vinyl Alcohol + Formic Acid $\Delta E^{\neq} = 46.6$	Acetaldehyde + Formic Acid $\Delta E^{\neq} = 47.6$	Ethenone + Formic Acid + H_2 $\Delta E^{\neq} = 44.3$	Ethylene + Performic Acid $\Delta E^{\neq} = 52.3$	Acetic Acid + Formaldehyde $\Delta E^{\neq} = 54.5$
S-CYCP	300	0	2 (2)	96 (3)	2 (2)	0	0
	600	7 (4)	7 (4)	73 (7)	9 (4)	0	4 (3)
	800	46 (7)	15 (5)	34 (7)	4 (3)	0	6 (4)

* Values in parenthesis are from bootstrapping.

** Energies shown are calculated using the PM6 Hamiltonian with zero-point energies (unscaled) and Boltzmann averaged for the two enantiomers.

3.4.4 Impact of Reordering

The final variation we tested was the impact of re-ordering the SPRINT coordinates within atom types, in order to capitalize on the invariant aspect. While these simulations were initially carried out using the same MD and bias parameters as those with static SPRINT coordinate ordering, we observed that these parameters were not suitable. Using a time step of 0.001 ps and a collision frequency of 10 ps^{-1}, we observed that the temperature and energy experienced

significant drift over the course of the simulation, often resulting in crashed runs. We attribute the drift the exchanging bias potentials between SPRINT coordinates that occurs with re-ordering. Because the bias potential is only associated with an index in the vector of SPRINT coordinates, and not a specific atom, the re-ordering scheme will cause different bias potentials to act on SPRINT coordinates of different atoms any time their sorted order changes. While the concept of exchanging bias potentials is not unusual, such schemes typically impose a criterion for accepting and rejecting exchanges based on the energetics, rather than CV value.[54,75] In order to combat this, we increased the collision frequency to 1000 ps^{-1}.

When analyzing the impact of re-ordering SPRINT on pathway discovery, we observed many of the same pathways, but with different distributions. Simulations initiated from KHP all proceeded to one of the CYCP isomers, with no strong preference for one isomer over another. However, as Table 3.6 shows, the distribution of pathways explored following CYCP is different from that of the static ordering scheme. The favored pathway is consistently the one forming acetaldehyde and formic acid, instead of acetic acid and formaldehyde. While the energy barrier for the former is lower than the latter, none of the trajectories were capable of re-discovering the KHP product, despite it being the lowest energy barrier pathway. Even though the reverse pathway to KHP was not typically the favored product in the static ordering trajectories, this pathway was still discovered and significantly sampled at 800 K. We attribute this difference to change in how the bias is deposited under the re-ordering scheme (see SI).

Table 3.6: Justified presence for sampled pathways leading away from CYCP for different temperatures for simulations starting from KHP with re-ordering SPRINT Coordinates

	Justified Presence for 1,2-dioxolan-3-ol Reactions					
Temp (K)	Vinyl Alcohol + Formic Acid $\Delta E^{\neq} = 46.6$	Acetaldehyde + Formic Acid $\Delta E^{\neq} = 47.6$	Ethenone + Formic Acid + H_2 $\Delta E^{\neq} = 44.3$	Ethylene + Performic Acid $\Delta E^{\neq} = 52.3$	Acetic Acid + Formaldehyde $\Delta E^{\neq} = 54.5$	*Oxirane + Acetic Acid $\Delta E^{\neq} = N/A$
300	25 (12)	56 (12)	17 (10)	0	0	0
600	37 (12)	63 (12)	0	0	0	0
800	18 (10)	37 (12)	18 (11)	4 (7)	4 (6)	11 (10)

* Transition state not found

** Energies shown are calculated using the PM6 Hamiltonian with zero-point energies (unscaled) and Boltzmann averaged for the two enantiomers.

For the simulations initiated from the CYCP we also observe similar pathways, but different JP distributions. As shown in Table 3.6, the formation of acetic acid and formaldehyde is the favored pathway at 300 K and 600 K, and heavily sampled at 800 K, despite this having the highest energy barrier. In addition, none of the trajectories sampled the KHP transition pathway, despite utilizing the conservative bias parameters. The omission of the KHP pathway at all of the temperatures, coupled with a clear proclivity for the of acetic acid and formaldehyde, indicates that re-ordering the SPRINT coordinates influences pathway is not the optimal choice for pathway discovery when compared to the static ordering scheme. Additionally, we sample a new pathway of CYCP forming oxirane and formic acid, however this transition state could not be found in either our Gaussian transition state searches or NEB calculations. This finding opens the possibility that this scheme may yield false transition events.

Table 3.7: Justified presence for sampled pathways leading away from CYCP for different temperatures for simulations starting from CYCP with re-ordering SPRINT Coordinates

	Justified Presence for 1,2-dioxolan-3-ol Reactions						
Temp (K)	Vinyl Alcohol + Formic Acid $\Delta E^{\neq} = 46.6$	Acetaldehyde + Formic Acid $\Delta E^{\neq} = 47.6$	Ethenone + Formic Acid + H_2 $\Delta E^{\neq} = 44.3$	Ethylene + Performic Acid $\Delta E^{\neq} = 52.3$	Acetic Acid + Formaldehyde $\Delta E^{\neq} = 54.5$	Oxirane + Acetic Acid $\Delta E^{\neq} = N/A$	*Water + Succinaldehyde $\Delta E\neq = 48.14$
300	0	17 (11)	0	0	82 (10)	0	0
600	0	43 (12)	0	0	57 (12)	0	0
800	4 (7)	37 (12)	11 (10)	0	24 (12)	11 (10)	4 (6)

* Note the Water + Succinaldehyde pathway was not noted in the previous tables, this pathway was discovered in the non-reordering scheme, but had a JP value of less than three.

** Energies shown are calculated using the PM6 Hamiltonian with zero-point energies (unscaled) and Boltzmann averaged for the two enantiomers.

3.4.5 Recommendations

Using this Korcek reaction as a test case, we have explored and tested the potential impact of the system settings and bias parameters on the resulting networks recovered. For future applications we recommend that atoms be denoted as distinguishable and therefore not reorder the SPRINT coordinates, unless preliminary testing shows that free energy profiles of SPRINT coordinates of the same atom type are truly equivalent, some examples being Lennard Jones clusters or an S_N2 reaction ($CH_3Cl + Cl^-$). However, in the context of studying poorly understood systems, we find that associating individual bias potentials with the SPRINT coordinates of specific atoms to be fairly robust. One drawback we observed was the poor sampling of the pathway from CYCP to KHP, especially at low temperature, as this was by far the lowest energy pathway. We attribute the poor sampling to the fact that biasing the SPRINT coordinates can cause the alcohol group to rotate more (~8 kcal/mol barrier), and explore phase space where it is

out of the plane of the ring. If the alcohol group is out of proximity of the O atoms in the ring, the reverse reaction is not possible. In addition, the action of the multiple bias potentials acting on the members of the ring structure may enhance the ring fracture. However, we are still able to sample this pathway at higher temperatures, indicating that biasing SPRINT coordinates does not preclude this reaction from happening altogether. Therefore, for future investigations we recommend that simulations be carried out at a range of temperatures. Additionally, we note that while the starting species does not have a drastic impact of sampling key pathways, it can have an influence on the distribution of pathways sampled. Especially for systems that show many competing pathways extending from a single node, it can be prudent to use such nodes as starting points in order to ensure all of the relevant pathways are sampled.

3.5 Conclusions

In this paper, we have outlined a procedure for efficiently using SPRINT coordinates as CVs to bias. Where as under the classical MetaD approach such an effort would have been inefficient and memory intensive due to the high dimensionality of the bias potential, the PBMetaD framework not only overcomes these obstacles, but also offers the capability of sampling a series of transition events using parameters set at the initial state of the simulation. PBMetaD now enables SPRINT coordinates to be suitable CVs to bias for systems with a large number of atoms, rather than be limited to small cluster systems. By studying the decomposition of KHP, we demonstrate how an ensemble of these simulations can sample a variety of pathways and chemistries, which are consistent with literature. The insights gained from these simulations can serve as a foundation for further analysis like infrequent metadynamics, to obtain kinetics, or provide structures for a path CV analysis, to obtain converged free energy barriers for individual steps. Furthermore, the robustness of this method, with regards to the bias parameters, and the

simplicity of defining the SPRINT coordinates allow for this method to be easily implemented for a variety of systems. By coupling the efficiency of high-dimensional sampling of PBMetaD with the flexibility of SPRINT coordinates, the method presented provides a suitable for applying MD to a variety of systems which were previously impractical to sample.

4 Biasing Smarter, Not Harder, By Partitioning Collective Variables Into Families in Parallel Bias Metadynamics[4]

4.1 Abstract

Molecular simulations of systems with multiple copies of identical atoms or molecules may require the biasing of numerous, degenerate collective variables (CVs) to accelerate sampling. Recently, a variation of metadynamics (MetaD) named parallel bias metadynamics (PBMetaD) has been shown to make biasing of many CVs more tractable. We extended the PBMetaD scheme so that it partitions degenerate CVs into families that share the same bias potential, consequently expediting convergence of the free-energy landscape. We tested our method, named Parallel Bias MetaD with Partitioned Families, on 3, 21, and 78 CV systems and obtained an approximately proportional increase in convergence speed compared to standard PBMetaD.

[4] Reproduced in part with permission from A. Prakash, C. D. Fu, M. Bonomi, and J. Pfaendtner. Biasing Smarter, Not Harder, By Partitioning Collective Variables Into Families in Parallel Bias Metadynamics. Journal of Chemical Theory and Computation, 14: 4985-4990, 2018. Copyright 2018 American Chemical Society.

4.2 Introduction

Molecular simulations have immense potential to provide crucial molecular- and nano-scale details of physical, chemical, and biological processes. However, simulations of slow molecular transitions, like protein folding/unfolding and chemical reactions, continue to be stymied due to high free-energy barriers and rugged free-energy landscapes that limit sampling. Several enhanced sampling methods, like metadynamics (MetaD),[81,82] umbrella sampling,[18] hyperdynamics,[66] variationally enhanced sampling,[8] and adaptive force biasing,[83] have tried to alleviate this problem by applying a bias along predefined coarse-grained descriptors or collective variables (CVs) of the system to accelerate sampling. Since applying bias in a high-dimensional CV space is often inefficient, researchers have traditionally resorted to choosing a minimal set of CVs.[84]

Identifying a small set of CVs that can effectively differentiate between relevant states of a system is a primary challenge for many enhanced sampling approaches. While a variety of CV-selection methods, like time-lagged independent component analysis (TICA),[85] reconnaissance metadynamics,[86] and spectral gap optimization of order parameters (SGOOP),[67] have recently been developed to address this challenge, researchers primarily still rely on their physico-chemical intuition of the system to select optimal CVs. Frequently, more than one candidate CV is biased which presents the challenge of efficiently biasing them with limited computational resources. To this end, replica-based methods like multiple walker metadynamics,[87] altruistic metadynamics,[88] and the flying Gaussian method[89] have been developed to exploit parallel simulations that share the bias potential to accelerate sampling even further. However, these methods do not address the problem of the high dimensionality of the bias potential, which requires extensive sampling, due to the larger phase space that needs to be explored, for convergence.

Some MetaD-based methods have been formulated to address the challenge of biasing a large number of CVs. In bias exchange[90] MetaD, a replica exchange approach is used, in which multiple replicas of the system, each biasing only one or few CVs, are simulated in parallel. Conformations of the system are periodically exchanged using a metropolis criterion. Other methods use several low dimensional bias potentials to bias individual CVs in lieu of a single high-dimensional bias potential.[25,84] Parallel bias Metadynamics (PBMetaD) is one such method that has been used to bias 4-40 CVs in the same simulation.[25,26,68,91,92]

This letter introduces PBMetaD with partitioned families (PBMetaDPF) for systems that require biasing multiple CVs that share identical properties. As an example, a simulation can contain multiple copies of a protein in a box. Since these proteins are identical subunits, they are expected to share identical properties. Thus, a CV describing property X of the n^{th} subunit would also describe property X of any other subunit of the system. Consequently, a system with N identical subunits would contain N CVs describing property X for each subunit. These CVs can be considered as indistinguishable or *degenerate,* as their equilibrium probability distributions are identical. In PBMetaDPF, the gain in efficiency over PBMetaD is achieved by grouping degenerate CVs into one family so that CVs in the same family share and contribute to the same bias potential. Finally, the free-energy profile for each CV family can be determined using the standard MetaD integration and standard reweighting techniques[42,93] can be used to study other degrees of freedom. The concept of a shared bias potential is inspired by multiple walker MetaD[87] where the potential of a CV evolves by contributions from *walkers* of the CV in parallel replicas of the system. In contrast, in our method all contributions to this shared potential come from multiple CVs within one replica only.

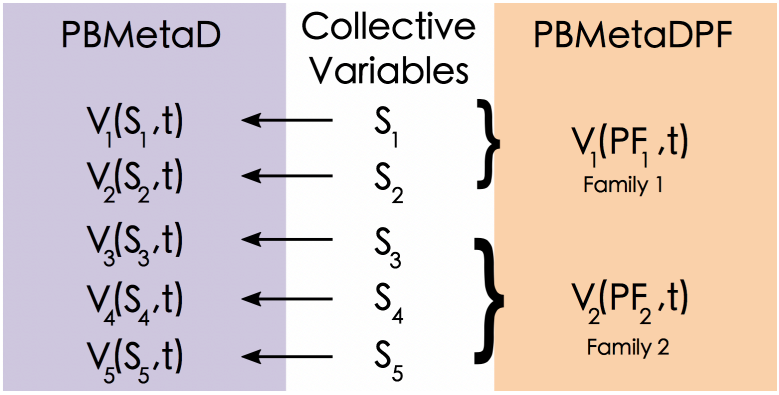
4.3 Theory

To highlight the differences between the PBMetaD and PBMetaDPF approaches, we will briefly review the theory of PBMetaD. In classical MetaD, or its well-tempered variant (WTMetaD),[82] a single multidimensional bias potential is constructed as a function of user-specified CVs, where the dimensionality of the potential is equal to the number of CVs. In contrast, PBMetaD constructs multiple monodimensional bias potentials applied along each individual CV. The method uses a new scheme to permit the instantaneous application of an arbitrary number of bias potentials, which was shown to converge (empirically) to the exact underlying free-energy surfaces.[25] To achieve this, the bias potential for the i^{th} CV(s_i), under the PBMetaD framework, is constructed through the following equation

$$V_G(s_i,t) = \int_0^t dt'\, W * \exp\left(-\frac{V_G(s_i R(t'),t')}{k_B \Delta T}\right) * \exp\left(\frac{-(s_i(R(t)) - s_i(R(t')))^2}{2\sigma^2}\right) * W_{\square B}(s_i,t') \qquad \text{Eq. 4.1}$$

$$W_{PB}(s_i,t) = \frac{\exp\left(-\frac{V_G(s_i R(t),t)}{k_B T}\right)}{\sum_{j=1}^{n} \exp\left(-\frac{V_G(s_j R(t),t)}{k_B T}\right)} \qquad \text{Eq. 4.2}$$

where W is the initial Gaussian height, σ_i the width of the Gaussian, dt' the pace of Gaussian deposition, k_B the Boltzmann constant, T the system temperature, ΔT is an input parameter with units of temperature which controls the rate at which Gaussians are scaled down, and n is the total number of CVs in the system. The first three terms in Eq. 1 follow the algorithm of a typical WTMetaD simulation where the Gaussian height is reduced as bias accumulates, while the last term, shown in Eq. 2, is a conditional weight to account for the effect of the bias deposited along the other CVs due the correlations among CVs.[84] This conditional weight distributes a new Gaussian across all CVs, with the CVs with lower bias at the time of deposition receiving a larger contribution. In this algorithm, each bias potential evolves independently, and the only interaction occurs via the conditional weight term.



Scheme 4.1: Diagrammatic view of the differences between PBMetaD and PBMetaDPF sampling schemes. Under the PBMetaD biasing scheme, an individual bias potential is evolved for each CV and the CV only acts under its own potential. In contrast, the PBMetaDPF schemes allows for all of the members of a given family to contribute to the formation of a single bias potential that, in turn, acts on all of the members of a particular family.

While PBMetaD offers a more scalable way to perform high-dimensional sampling, converging numerous bias potentials independently can still require lengthy simulation times, as evidenced by Prakash et al.[68] In PBMetaDPF, we expedite the convergence of these energy landscapes, which comprise indistinguishable particles, by partitioning degenerate CVs into families (PFs). For book-keeping purposes, we will refer to CVs as s_{PFf-k}, where *f* refers to the PF it belongs to and *k* refers to the CV index in that PF (e.g. s_{PF1-2} is the second CV belonging to PF one). CVs partitioned into the same PF deposit bias similar to the multiple walkers framework,[87] where the Gaussians deposited along the different individual CVs of a particular PF all contribute to the formation of a single bias potential that acts on all the CVs belonging to that PF. The bias potential for any CV belonging to partitioned family PF1 (s_{PF1-x}; where x is any member of family 1), which has m members, is recovered through:

$$V_G(s_{PF1-x},t) = \sum_{k=1}^{m}\int_0^t dt'\, W * \exp\left(-\frac{V_G(s_{PF1-k}R(t'),t')}{k_B\Delta T}\right) * \exp\left(\frac{-(s_{PF1-x}(R(t))-s_{PF1-k}(R(t')))^2}{2\sigma^2}\right) * W_{PB}(s_{PF1-k},t') \quad \text{Eq. 4.3}$$

Here, the bias potential of each family is constructed by the contributions of every member. In other words, the bias potential of the family acts on each member of the family. Consequently, only one free-energy profile is recovered per PF instead of recovering one free-energy profile for each CV. Note that the denominator of the conditional weight term still sums over all the CVs biased in a system, as is done in regular PBMetaD.

4.4 Results

4.4.1 3-Particle Lennard-Jones System

To assess the accuracy and efficiency of our approach, we used a simple three-particle Lennard-Jones (LJ) system (σ = 0.39 nm, ε = 30 kJ/mol). For both PBMetaD and PBMetaDPF, we biased all three interatomic distances with initial Gaussian heights of 2.0 kJ/mol, Gaussian widths of 0.01 nm, a bias factor of 10, and a deposition pace of 1 ps. Sixteen independent biased simulations were run in the NVT ensemble for 2 μs. Further, to confirm that the PBMetaD framework is suitable for describing such systems, we performed parallel tempering (PT) simulations to provide an independent reference free-energy profile (see Appendix 4 for details).

We monitored the root mean squared deviation (RMSD) of the free-energy profiles recovered from PBMetaD and PBMetaDPF relative to that obtained with PT to assess both convergence speed and accuracy. The RMSD between two profiles was defined as:[44]

$$RMSD\left(F_{ref}, F\right) = \sqrt{\frac{1}{\Omega}\int dS[\left(F_{ref}(S) - \bar{F}_{ref}\right) - (F(S) - \bar{F})]^2} \qquad \text{Eq. 4.4}$$

where S is the CV value, $F_{ref}(S)$ and $F(S)$ are the two free-energy profiles being compared, $\bar{F}$ and $\bar{F}_{ref}$ are the free-energy averaged over the region Ω. For the three-particle and 13-particle systems, the region of interest was defined as the CV space within 30 kJ/mol of the global minimum of the reference PT profile, while for the seven-particle system the region of interest was defined as the CV space within 100 kJ/mol of the global minimum of the reweighted

interatomic distance profile from WTMetaD. For each PBMetaD simulation, we recovered three free-energy profiles, one for each interatomic distance. In the case of PBMetaDPF, a single free-energy profile for a given simulation is naturally recovered because the method groups the three CVs into the same family.

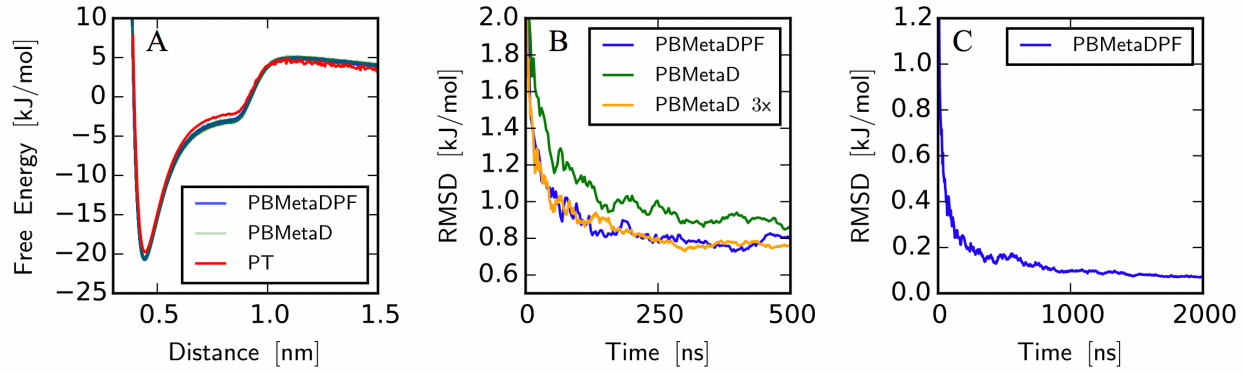


Figure 4.1: (A) Mean-aligned free-energy profiles of the interatomic distance between LJ particles. In total, the 16 PBMetaDPF profiles, the 48 PBMetaD profiles, and one parallel tempering (PT) profile are plotted. (B) The average RMSD of PBMetaDPF profiles (blue), PBMetaD profiles (green), and of PBMetaD with a projected convergence rate of three times faster (orange), all RMSD calculations are relative to the reference PT profile. (C) The average RMSD of PBMetaDPF relative to the converged PBMetaD profile over the course of the simulation.

As shown in Figure 4.1A, both PBMetaD and PBMetaDPF accurately reproduced the free-energy profile along the interatomic distance obtained with PT. The RMSD of the free-energy profile recovered from PBMetaDPF simulations is well-within k_bT (~ 2.5 kJ/mol at 300 K) when compared to PT (Figure 4.1B) and PBMetaD (Figure 4.1C) profiles. The error primarily stems from differences in the higher free-energy regions (Figure A4.5) since it is difficult for the

parallel tempering run to converge in that area. However, the small value of free energy difference shows that partitioning CVs into families does not introduce additional errors, for this system. In fact, on average, PBMetaDPF converges to the reference PT profile approximately three times faster than PBMetaD, as shown by the overlap between the red and green lines in Figure 4.1B. A three-fold acceleration in convergence is attributed to the fact that the bias potential in PBMetaDPF is constructed by three CVs as opposed to a single CV in PBMetaD.

4.4.2 13-Particle Lennard-Jones System

To further demonstrate the accelerated convergence offered by the PBMetaDPF, we simulated a 13-particle LJ system (σ = 0.39 nm, ε = 11 kJ/mol). A lower ε value was chosen so that PT simulations could converge in a reasonable amount computational time. In PBMetaDPF, all interatomic distances (78 CVs) were biased and grouped into the same family. Again, we performed 16 independent simulations using the same bias parameters as the three-particle system. We also performed a PT simulation of the system (see details in Appendix 4) to provide an independent reference for the free-energy profiles.

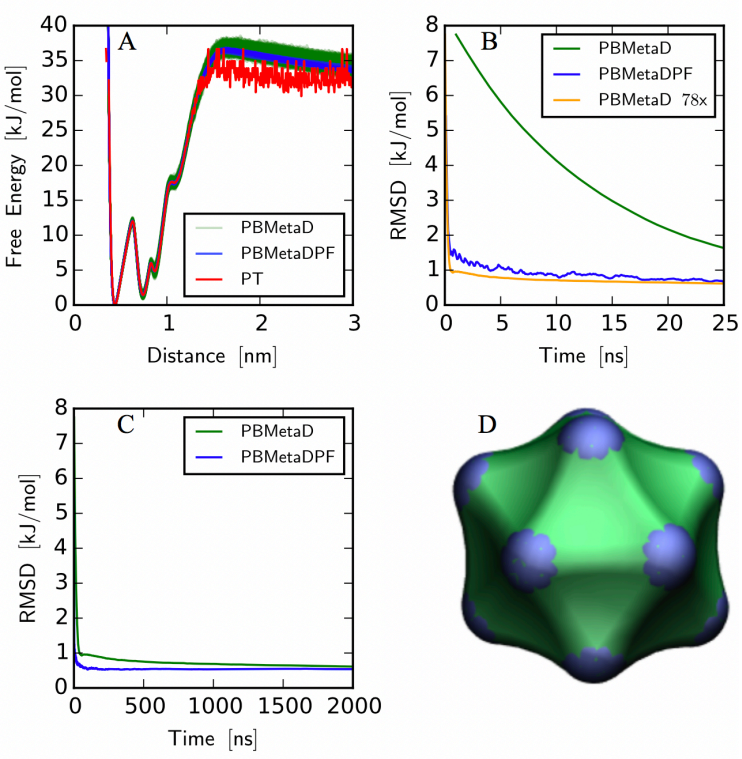


Figure 4.2: (A) All of the mean-aligned free-energy profiles for PBMetaD after 4 μs (78 profiles x 16 trials) and PBMetaDPF after 4 μs (16 trials) and one profile from parallel tempering (PT). (B) The average RMSD, with respect to the converged PT profile, of PBMetaDPF profiles (blue), PBMetaD profiles (green), and of PBMetaD with a projected convergence rate of 78 times faster (orange) over the course of the simulation. (C) The average RMSD of PBMetaD profiles (green) and PBMetaDPF profiles (blue), all RMSD calculations are relative to the converged PT profile over the course of the simulation. (D) Structure corresponding to the global free-energy minimum.

PBMetaD (all 78 CVs) and PBMetadPF converged to the same free-energy profiles (Figure 4.2A). However, the PBMetaD profiles in Figure 4.2A exhibit more fluctuations between free-energy profiles than the single PBMetaDPF profile. This highlights the challenge of converging 78 independent free-energy profiles, which necessitates that each atom pair explore

the entire CV space. However, we do observe that in the long-time limit (> 2 μs), both PBMetaD and PBMetaDPF give approximately equivalent average RMSD values (Figure 4.2B). Similar to the LJ3 system, it is shown that partitioning the CVs into a single family allows accelerated convergence of the free-energy profile to (Figure 4.2C). Most notably, the speed in convergence is again proportional to the number of CVs in the family, 78 in this case.

In addition to recovering the free-energy with respect to the biased CVs (interatomic distances), the coordinates from the trajectory and metadynamics biases can be utilized to obtain stable structures of the LJ particle system. To find the most stable structure, the frames of the PBMetaD and PBMetaDPF trajectories were clustered (see Appendix 4). The clusters were reweighted (described in detail in the Appendix 4) using the PBMetaD and PBMetaDPF bias and the method of Torrie and Valleau[18] to find the most stable structure – an icosahedron (Figure 4.2D), which had highest probability of all the structures (~ 100%). This result is consistent with previous analyses of the system where the icosahedron structure of LJ_{13} was shown to be at least 2.85ε (= 31.35 kJ/mol for this system) more stable than other structures.[15,94]

Before this, a MetaD approach that biases 78 CVs for exploring the structural minima of aggregating systems such as the one above would have been intractable. But by partitioning the CVs into a single family, scaling this approach to larger systems is now possible.

4.4.3 7-Particle Lennard-Jones System

Lastly, to demonstrate the effectiveness of PBMetaDPF to explore and describe systems with multiple, metastable states, we applied it to a 7-particle LJ system constrained to two dimensions. This system is well-studied using a variety of methods and is known to have four stable structural minima and 19 transition states.[95–97] Here, we apply PBMetaD and PBMetaDPF to explore the potential energy landscape by biasing the 21 interatomic distances, running each

simulation for 2 μs. As in previous treatments of interatomic distances, we partition them into a single PF for PBMetaDPF. For comparison, we also carried out unbiased MD simulations as well as a WTMetaD simulation where the second and third moments of the coordination numbers were biased, following the work of Nava et al.[95] Both PBMetaD and PBMetaDPF yielded identical free-energy profiles for the interatomic distances. These profiles were identical to the free-energy profiles recovered from reweighting the WTMetaD onto the interatomic distances (Figure 4.3A). In this case too, PBMetaDPF could converge the free-energy surface for the interatomic distances faster (Figure 4.3B), and the degree of acceleration is commensurate with the reduction of profiles (21 times faster). Unsurprisingly, the unbiased MD simulation (~ 2 μs) did not generate the correct free-energy surface (Figure A4.2) for this system emphasizing the need for enhanced sampling techniques.

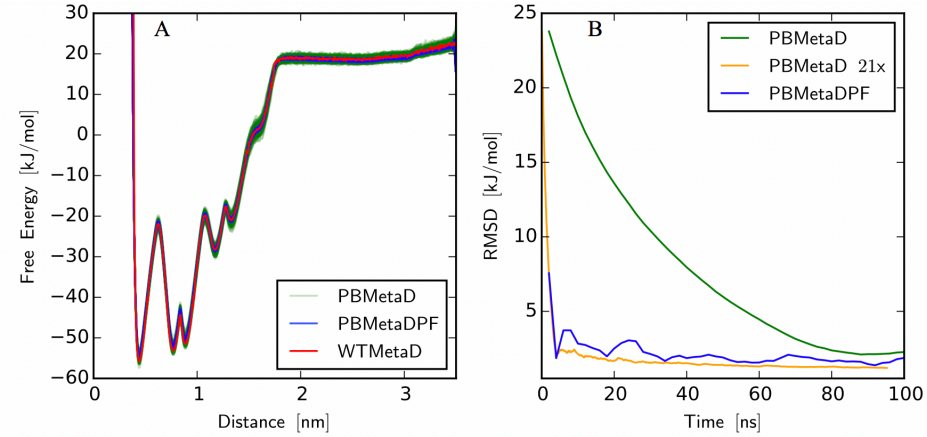


Figure 4.3: (A) Mean-aligned free-energy profiles of the interatomic distance between LJ particles. In total, the 16 PBMetaDPF profiles, the 336 PBMetaD profiles, and one WTMetaD profile (reweighted) are plotted. (B) The average RMSD of PBMetaDPF profiles (blue), PBMetaD profiles (green), and average RMSD of PBMetaD projected to converge 21 times faster (orange) relative to the converged WTMetaD profile over the course of the simulation. The area of interest was restricted to be 100 kJ/mol of the minimum of the reweighted WTMetaD profile.

In previous MetaD studies of this system, the second and third moments of the coordination number were selected and biased, as these CVs were able to differentiate between the four stable structural minima.[95,96] The limit to two CVs was also more amenable to methods that were unable to bias more than a few CVs at the same time. However, we chose to bias the interatomic distances using the PBMetaD and PBMetaDPF frameworks since it is a far more general approach and the CVs are more interpretable. After the simulations were converged, we reweighted the second and third moments of the coordination number to demonstrate the consistency of our results with prior work.[95,96] As shown in Figure 4.4A-C, the correct FESs were recovered with reweighting using both PBMetaD and PBMetaDPF. The FESs recovered

using the two methods are identical in shape (Figure A4.2) and exhibit identical approach to convergence over the course of the simulation (Figure 4.4 B and C).

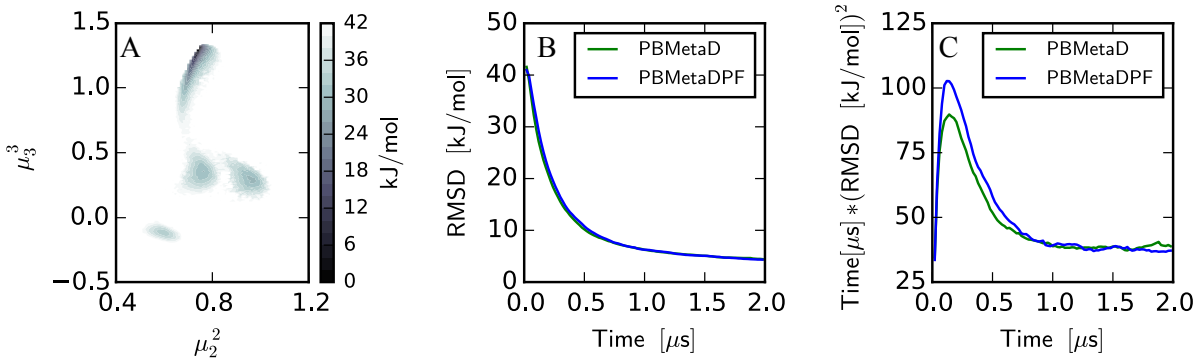


Figure 4.4: (A) Free-energy surface for the 7-particle LJ system reweighted for the second and third moments of coordination numbers using PBMetaDPF. (B) The average RMSD from 16 PBMetaD and PBMetaDPF simulations (each) reweighted for the second and third moments of coordination numbers with respect to a WTMetaD simulation biasing those same CVs. (C) A demonstration of the absence of systematic error in reweighting both PBMetaD and PBMetaDPF. The area of interest was restricted to be 40 kJ/mol of the minimum of the reweighted WTMetaD surface.

Further, we analyzed the trajectories using the clustering and reweighting method described in Appendix 4 to find the most stable structural minima. We were able to obtain the minima obtained in previous investigations of the system (Figure 4.5). Further, the probabilities of the structural minima predicted using PBMetaD and PBMetaDPF were within 0.5 % of each other (Table A4.1).

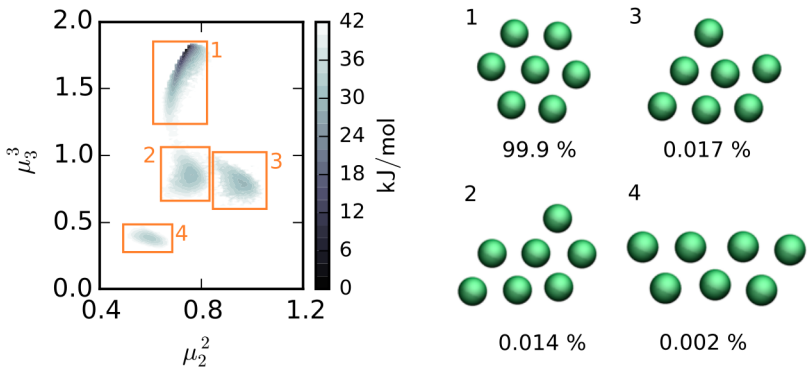


Figure 4.5: (left) Free-energy surface recovered from PBMetaDPF simulation of the 2D 7-particle LJ system after reweighting for second and third moments of the coordination number. (right) Representative structures for the regions highlighted in orange on the free-energy surface along with the probability of occurrence of each structure in the 2D phase space plotted on the right.

This study highlights that PBMetaDPF can be used to effectively sample the simple yet challenging structural phase space of the 7-particle LJ system. Furthermore, biasing the interatomic distances is a far more general, and intuitive approach for studying these complex structures. This eliminates the use of unintuitive CVs like moments of coordination numbers. Moreover, unlike PBMetaD, PBMetaDPF offers a more scalable approach by leveraging the excessive number of degenerate CVs to expedite converging an individual FES, rather than converging 21 profiles independently (as also proved in the LJ_{13} system). Lastly, the free-energy surface with respect to other CVs of interest can always be reweighted and constructed after converging the FES for the distance CVs.

4.5 Other Potential Applications

In this study, we have applied PBMetaDPF to simple LJ systems. However, we can predict several other applications for this method where researchers were restricted to using a

single CV because previous methods were not amenable to multiple, degenerate CVs. In a recent study of ion-channel dynamics,[90] the authors treated degenerate ion distances as separate CVs in bias-exchange metadynamics and noted that the result was the same. Conceivably, the PBMetaDPF scheme could be used to bias these degenerate variables. A similar use case could be obtaining the potential of mean force (PMF) profiles of ion-pairing systems, using distances between ions as their CV.[98] Most ion PMFs are calculated with only a single ion-pair in water which requires an additional correction term to account for differences between simulated (one pair) and experimental ion concentrations.[99] PBMetaDPF can be used in simulations to bias the distances with multiple ions pairings, treating unique ion-pair distances as PFs, which would allow the recovery of free-energy estimates at realistic concentrations, potentially providing efficient routes to describe ion solvation/desolvation.[100]

Another potential area of application is in aggregation studies, where the free-energy of association of a particle is computed to predict the behavior of aggregates at higher concentrations for experiments. For example, the dimerization free-energy of peptoids in the dilute limit was used to predict how peptoids at higher concentrations aggregate in solution.[101] Similarly, the dimerization free of two MgO nanoparticles in vacuum was calculated to understand the crystal growth that occurs in saturated solutions.[102] We suggest that more crowded environments should be simulated and all the inter-particle distances should be biased to recover better estimates of the free energies of association at realistic, experiment level concentrations. While the concentration levels would typically require biasing too many CVs, partitioning the families of indistinguishable particle interactions should allow for this type of approach to be feasible. Many aggregation studies simulate multiple particles in a box, for example when looking at amyloid peptide aggregation with MD simulations.[103,104] We propose

that using PBMetaDPF to bias inter-particle distances, and radius of gyration of all residues exhaustively sample the system and reduce the time required to sample all structural minima.

4.6 Conclusions

In this letter, we presented PBMetaDPF, an extension of the PBMetaD approach to expedite convergence for a special sub-class of enhanced sampling problems. The method partition degenerate CVs into families, such that they contribute to the same bias potential, in the spirit of the multiple walker scheme. For 3-, 7-, and 13-particle LJ systems, we demonstrated that partitioning CVs into one family led to accelerating the convergence of free-energy profiles. Further, in the case of the 7-particle LJ system, we show that standard reweighting techniques can be used to calculate the free-energy as a function of other CVs not directly biased and detect stable structural minima. Thus, even after partitioning CVs into families, the method maintains its compatibility with commonly-used reweighting protocols. Finally, and most importantly, grouping the CVs into one family led to an increase in the convergence speed of the free-energy profiles that was proportional to the number of CVs grouped into the family. Lastly, PBMetaDPF has been implemented in the open-source PLUMED[105] library (developers' version soon to be made available to the public), allowing for this approach to not only be readily accessible, but also be easily applied to variety of systems and problems.

5 Diagnosing the Impact of External Electric Fields on Toluene Oxidation and Pyrolysis[5]

5.1 Abstract

In this study we utilize biased, reactive molecular dynamics (ReaxFF MD) simulations to quantitatively assess the impact of external electric fields on the kinetics of toluene oxidation and pyrolysis. We observe that the application of a strong external electric field significantly accelerates the kinetics of toluene oxidation, while having no effect on the pyrolysis reactions. When viewed through the lens of harmonic transition state theory, this phenomenon can be ascribed to the increase in the change of entropy between the transition state and reactants for the oxidation reactions. This conclusion is further verified with a model that relates the change in entropy as a function of field strength to predict the kinetics of toluene oxidation, which accounts for the total variance in the data recovered from simulation.

[5] Reproduced in part with permission from C. D. Fu, Y. He, and J. Pfaendtner. Diagnosing the Impact of External Electric Fields on Toluene Oxidation and Pyrolysis. Journal of Physical Chemistry A, under review. Copyright 2018 American Chemical Society.

5.2 Introduction

Exploring and characterizing reacting systems through computational means, specifically reactive molecular dynamics (MD) simulations, is of growing interest to the reactions community, as these simulations can provide unique insights and resolution into the mechanistic details of these complex processes. Specifically, reactive simulations provide an opportunity to sample and explore the effects a multitude of factors, such as solvent, temperature, density, etc., have on reaction pathway kinetics and thermodynamics. While MD is typically encumbered with a high computational cost, resulting in only running short simulations or requiring massive amounts of computing power, this burden is gradually becoming alleviated due to continued improvements in hardware, theory, and methodology. Specifically, the development of the ReaxFF reactive force fields[5,6] have shown tremendous promise in their ability to accurately describe reactive processes, but at a computational expense more comparable to classical MD than *ab initi*o simulations. Over the past decade, the ReaxFF force field parameterized for hydrocarbon combustion and pyrolysis has been applied to explore and accurately describe a variety of systems.[5,6,56,58,106–109]

One facet of ReaxFF simulations that has evolved in recent years is its application to explore how external electric fields can impact reaction pathways. Experimental and computational investigations into how electric fields can effect reactions has been a long standing area of interest,[17,110,111] specifically, and not surprisingly, with applications in the field of catalysis.[112] Recently, Tan et al.[106] put forth a very interesting finding that applying an external electric field can accelerate the oxidation of toluene, matching what the phenomenon that was observed experimentally. Additionally, Jiang et al.[109] utilized ReaxFF simulations to demonstrate how electric fields can impact pathway selectivity and even make available pathways unique to specific field strengths.

While the studies of Tan et al.[106] and Jiang et al.[109] put forth novel findings, with interesting implications for oxidation chemistries, a commonality that is shared between them are the obstacles they encountered and the type of coarse analysis that is used to explore these systems. Even though ReaxFF simulations extend the accessibility of the simulation time scales to nanoseconds, this is often insufficient to sample reactive events, unless the systems are simulated at exceedingly high temperatures (>2000 K).[5,6,64,107] Moreover, for a given system, typically a few to up to 10 simulations are carried, where the evolution of the counts of the different species are used to evaluate reaction kinetics and track reaction pathways, providing a fairly broad overview of the mechanisms taking place in the system.

Herein, we perform a deeper study regarding the impact of an applied external electric field on the kinetics of the initiation reactions of toluene oxidation and pyrolysis, providing a detailed mechanistic explanation for the results observed by Tan et al. Additionally, we sample the reaction rates of these different mechanisms in a temperature regime that falls more in line with experiment (1000 K – 2000 K).[111,113,114] Efficient sampling of the reaction kinetics of the different events is facilitated by using the so-called infrequent Metadynamics (IMetaD) method, described below. By sampling the oxidation of toluene, and the two dominant, competing pyrolysis reactions, shown in Figure 5.1 under across this range of temperatures and electric fields, we are able uncover the impact, or lack thereof, an electric field has on these different pathways.

The paper is organized as follows. We provide an overview of the both the ReaxFF force field and simulation protocol, in addition to how IMetaD was used to recover accurate reaction kinetics from biased simulations. We then present the kinetic results in the form of Arrhenius plots, and then provide a mechanistic and quantitative explanation as to how an electric field

impacts reaction kinetics. The paper concludes with our review of the ramifications of these results and possible interesting areas for future investigation.

5.3 Methods

5.3.1 Simulation Details

Because the ReaxFF methodology has been established and applied for well over a decade, we refer readers to those sources summarizing its implementation.[5,115,116] As we are studying a similar, and in some cases identical, system setup as Tan et al.[106] we followed their simulation protocol. Oxidation systems were constructed with packmol,[117] with 100 O_2 molecules and one toluene molecule placed in a cubic box of length of 2.5 nm. For pyrolysis systems the 100 H_2 molecules were used in the same setup in place of O_2. Using the LAMMPS software package, NVT simulations were carried out using the ReaxFF force field developed by Chenoweth et al.,[5] as this has previously been used to accurately describe this system and others like it.[5,106,107] As was done by Tan et al., the systems underwent an energy minimization, and then prior to the production runs were equilibrated at 5 K for 50 ps using a 0.1 fs time step. All NVT simulations were carried out using the Berendson thermostat[118] with a damping coefficient of 0.1 ps. Charges were equilibrated using the corresponding Qeq method implemented in LAMMPS,[119–121] which is also standard for ReaxFF simulations. It was ensured that no reactions occurred during the equilibration phase. All production runs were carried out with unique seeds for establishing the velocities at the desired temperature and all runs used a bond order cutoff of 0.3 nm. Because our primary concern was evaluating the kinetics of the initiation reactions, and not beyond, the bonds of the atoms in toluene were monitored, and the simulation was terminated if any of the bonds broke (C-H bond exceeding 0.2 nm, C-C exceeding 0.4 nm). In the coordination of the atoms in toluene to the oxygen atoms were also tracked to ensure no new bonds formed either. When applied, an electric field was set to the requisite field strength along

the z-axis, and the field strengths tested were $1.0x10^{-5}$, $1.0x10^{-3}$, $1.0x10^{-2}$, $5.0x10^{-2}$, $1.0x10^{-1}$, $1.25x10^{-1}$, $1.5x10^{-1}$, $1.75x10^{-1}$, and $2.0x10^{-1}$ V/Å, in addition to conditions without any electric field. For the oxidation reactions, at each field strength tested, reaction kinetics were sampled reaction from 1000 – 2000 K in increments of 250 K. For the pyrolysis reactions, conditions of no electric field and a strong electric field of 0.2 V/Å were modeled across the temperature range of 1250 – 2000 K in increments of 250 K.

5.3.2 Infrequent Metadynamics

In order to sample reactive events on a more reasonable time scale, but still allow for the recovery of accurate reaction kinetics, we used the IMetaD framework, as this has shown great promise in sampling the kinetics of chemical reactions.[20,23,122] As both metadynamics (MetaD) and IMetaD are explained in detail elsewhere,[2,7,19,20] we will only provide a brief overview of the theory. Over the course of a simulation, a bias potential in the form of Gaussians, or "hills", are slowly deposited to act on a few select degrees of freedom, or collective variables (CVs). CVs are chosen to be the slowest degrees of freedom that we wish to enhance in order to expedite the exploration in phase space. The IMetaD variant, as developed by Tiwary and Parrinello,[19] works in a similar fashion, and facilitates the recovery of an estimate of unbiased kinetics from biased simulations. The major deviation from a typical MetaD workflow is that IMetaD uses an ensemble of simulations to repeatedly sample a transition pathway via separate simulations in order to get a converged estimate of mean transition time. For a given simulation, knowledge of the deposited bias potential can be used to recover an estimate of the unbiased kinetics (escape time) via

$$t^{eff} = \alpha * t^{MD} \qquad \text{Eq. 5.1}$$

$$\alpha = \frac{1}{t^{MD}} * \int_0^{t^{MD}} dt' e^{\beta V_{bias}(s,t')} \qquad \text{Eq. 5.2}$$

where t^{eff} is the effective escape time, t^{MD} is the run time of the MD simulation when the transition event occurs, α is the so-called acceleration factor, β is $1/k_BT$, and V_{bias} is the value of the bias potential at the location in CV-space (s) and time (t') in the simulation. This framework has successfully been used to describe the kinetics of a variety of chemical reactions[20,23,122] and capture other attributes of interest such as pathway selectivity.[22] It is important to note that IMetaD shares many similarities with existing methods such as bond-boost,[48] hyperdynamics,[66] and variational flooding,[49] but deviates in that IMetaD uses a more flexible transient bias potential rather than a static one. Another similar method is the collective variable-driven hyperdynamics (CVHD) method developed by Bal and Neyts.[56,58] While this method has shown promise in both compatibility with ReaxFF[58] and accurately assessing unbiased kinetics,[23] it requires a collapsed, one dimensional CV and is typically used in exploring long time scale transitions rather than focus on recovering rates of individual transitions. Additionally, IMetaD calls for an additional post-processing step as a check that the distributions of escape times recovered are uncorrelated and Poisson, which ensures that the bias is not corrupting the dynamics of the system.[2] The population of sampled events is verified to be Poisson by applying the Kolmogorov-Smirnov (KS) analysis ($p>0.05$), as prescribed by Salvalaglio et al.[2] For all of the data sets corresponding to the different conditions and systems tested, the recovered p-values all satisfied this requirement and even a more strict criterion of $p>0.10$ (p-values for each data set provided in SI).

The PLUMED[41,123] plug-in was used to apply the bias potential for these simulations. For all systems, an initial hill height of 1 kJ/mol, a bias factor of 10, and a deposition pace of 10,000 steps were used. The parameter choices are fairly conservative given the anticipated energy barriers of these reactions (>> 10 k_bT), but proved to be robust in sampling across the

different reactions and temperatures as the p-values recovered from the KS-test for each sample exceeded 0.1, let alone the typical 0.05 cutoff. In addition, we provide the results of a representative pyrolysis reaction in Appendix 5 (Table A5.6) where the pace was varied and consistent kinetics were recovered, verifying the suitability of these parameters.[50] A minimum of 32 events was sampled for each particular system explored (mechanism, temperature, field strength). For the pyrolysis systems, we either biased the C-C bond connecting the methyl group to the benzene ring to sample the production of a methyl and phenyl radical or we biased one of the C-H bond distances of the methyl group to sample the production of a hydrogen and a benzyl radical. The widths of Gaussians were set to one half the mean fluctuation of the CVs, which was observed to slightly change with temperature. In order to effectively sample a specific reaction, we followed the methodology of Fu et al.[22] and quartic restraints were placed on all other toluene bonds in order to effectively block any competing reaction pathways. For the oxidation systems, we biased only one of the C-H bond distances of the methyl group to sample the production of a hydroperoxy radical and a benzyl radical. Because this biasing scheme does not preclude the competing pyrolysis reaction from occurring, we tracked the coordination of that H atom to the oxygen atoms in the system (see Appendix 5 for switching function details). If the coordination state of the H-atom was below 0.4, we attributed the reaction to a pyrolysis reaction. Across the different oxidation systems sampled, no more than three pyrolysis events were recovered in a given data set. As such, the average reaction rate of the oxidation reaction for these systems is no longer the inverse of the mean escape time recovered, but instead

$$\upsilon_{oxidation} = \frac{n_{oxidation}}{t_{tot}^{esc}} \qquad \text{Eq. 5.3}$$

where $\upsilon_{oxidation}$ is the average oxidation reaction rate, $n_{oxidation}$ is the total number of oxidation events sampled in the data set, and t_{tot}^{esc} is the sum of the total effective escape time for

all events sampled in the data set, oxidation and pyrolysis.[2,22] Note, the same process could be carried out to evaluate the reaction rate of the competing pyrolysis reaction by substituting the number of pyrolysis reactions sampled in the data set into the numerator. However, because we at most only encountered this competing pathway three times in a given set of conditions, it is insufficient to rely on this as a converged estimate compared to the ~30 events of the oxidation pathway. Error analysis was carried out following the bootstrapping method outlined by Fleming et al.,[20] by generating 1000 randomly selected (with replacement) subsamples for each data set, where the size of each subsample was one half of the total data set size. If a subsample did not pass the KS-test (p-value <0.05), it was removed and replaced with a new subsample. The reject rate was typically small, with the largest being 17%. The error for the mean escape time is then described as the standard deviation of the 1000 accepted, bootstrapped subsamples (see Appendix 5 for full results).

5.4 Results and Discussion

5.4.1 Toluene Oxidation Kinetics

Following the protocol above, we sampled the kinetics of the toluene oxidation reaction where one of the hydrogen atoms of the methyl group is abstracted by an O_2 molecule in our system. The kinetics of this reaction was sampled across the temperature range of 1000 – 2000 K, and fit to an Arrhenius equation. In the absence of an external electric field, our Arrhenius fit yielded an activation energy of 42.0 kcal/mol and pre-exponential factor of 2.4×10^{12} s^{-1}. Experiment values in literature values ranging between ~40-43 kcal/mol for activation energies, depending on the source and setup,[113,114] which are in agreement with the values we recover. We do note that or recovered pre-exponential values are elevated compared to literature, but this is common with ReaxFF simulations.[6,107] In addition to establishing the accuracy of the ReaxFF force field in

describing the oxidation kinetics, it is worth noting that evaluating the kinetics of these reactions in this temperature range was made possible by using IMetaD. At 1000 K the calculated mean effective time for the reaction is on the order of milliseconds, but our average simulation time was only ~3-4 ns, similar to the length of the unbiased simulations carried out by Tan et al.[106] Therefore, in this instance, the IMetaD method not only allowed for us to recover accurate kinetics and kinetic parameters at these lower temperatures, but the typical length of these simulations is on the same order of a typical, high temperature ReaxFF simulation.

Once it was established that our workflow yielded accurate kinetics, we then sampled systems with an external electric field applied along the z-axis. Arrhenius plots for a subset of the electric fields tested are depicted in Figure 5.1, along with the Arrhenius plot for the no field condition. It is clear from Figure 5.1 that increasing the strength of the applied electric field leads to faster reaction kinetics, matching what has been observed in previous studies,[106,111] with the strongest field conditions accelerating the reaction by an order of magnitude.

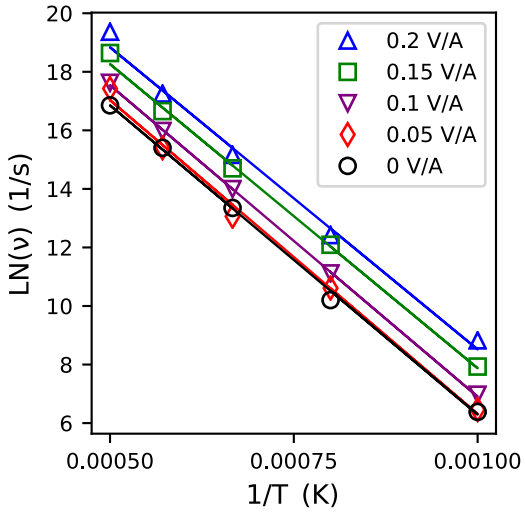


Figure 5.1: Arrhenius plot for oxidation of toluene under high field strength conditions and no electric field (black circles). Error bars were smaller than the symbols, and were omitted for clarity.

However, it is worth noting that for electric field strengths of 0.05 V/Å and below, no appreciable acceleration is observed (see Appendix 5 for lower field strengths) as there is very good agreement between those data points and the no field condition. This result indicates that there appears to be some threshold value for electric field strength to have a non-negligible impact. We further verified this result by evaluating the difference in the times measured for the various electric field conditions relative to the no field condition via

$$Mean\ \%\ Error = \frac{1}{n}\sum_{i=1}^{n}\frac{|\tau_{no\ field}(T_i)-\tau_{field}(T_i)|}{\tau_{no\ field}(T_i)} * 100 \qquad \text{Eq. 5.4}$$

where n represents the number of temperatures tested, $\tau_{no\ field}(T_i)$ is the mean reaction time for the no field condition at temperature T_i, $\tau_{field}(T_i)$ is the mean reaction time calculated for a given electric field strength at temperature T_i. The error calculated between the various field and no field conditions is then compared to the calculated uncertainty for each data set (i.e., the mean

ratio of bootstrapped standard deviation normalized by measured reaction time), which is shown in Figure 5.2. From analyzing the trends of the two error measurements in Figure 5.2, it is clear that for a given data set the relative uncertainty measured from bootstrapping stays relatively constants between 20-30%, regardless of the strength or presence of the electric field. It is worth noting that the relative uncertainty recovered from bootstrapping is typical for the IMetaD framework with equivalent sample sizes.[20,23,122,124] Additionally, in agreement with Figure 5.1, we see that for fields strengths of 0.05 or lower, on average, the deviation from the no field condition is on the same order as the error that is to be expected from carrying out this analysis. However, for field strengths of 0.1 V/Å and above, we see substantial deviation that is well outside this error range.

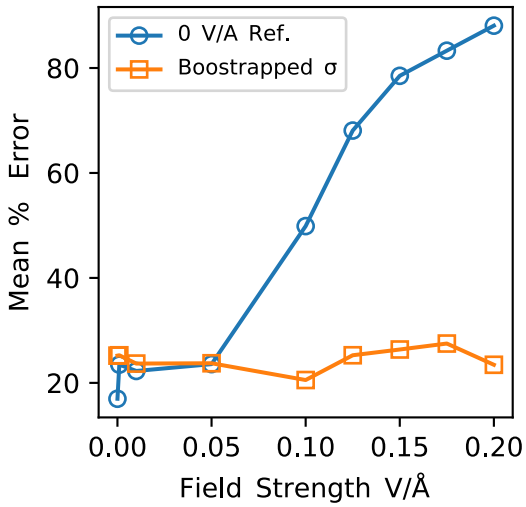


Figure 5.2: Mean percent error (absolute value) of reaction times as a function of field strength. The orange squares represent the mean relative uncertainty calculated from bootstrapping. The blue circles represent the mean percent error between the escape times calculated under the different field strengths, relative to the no field condition (absolute values taken for each difference).

5.4.2 Impact of Electric Fields

While it is clear from Figures 5.1 and 5.2 that with a sufficiently strong electric field toluene oxidation can be accelerated, it is difficult to discern from these figures alone how the electric fields give rise to these faster kinetics. In order to get a better mechanistic understanding of this phenomenon, for each electric field strength tested, we calculated the activation energy and pre-exponential factor from fitting the data to an Arrhenius equation, and plotted the values as a function of field strength in Figures 5.3A and 5.3B. In addition, we evaluated the R^2 values to assess the quality of these fits to the data, and these are plotted as blue circles in Figure 5.3C. From analyzing Figure 5.3A, we see that the electric field has little no effect on the activation energy of the oxidation reaction. While one may note that there appears to be a slight trend in the activation energy decreasing as the field strength increases, the overall change is not only negligible compared to size of the error bars, but also to thermal fluctuation in this temperature range (>2 kcal/mol). This observation is further verified in Figure 5.3C (green squares) where we applied an Arrhenius fit to the different electric field data sets, but supplied the activation energy of the no field condition instead of leaving this as a fitted parameter. The modified Arrhenius equation shows strong agreement with the sampled data as the R^2 values are all above 0.94 and show no discernable trend with field strength.

However, we do observe a trend in the recovered pre-exponential factors as a function of field strength. Keeping in mind that Figure 5.3B is the natural log of the pre-exponential factors, we see that the values increase with field strength, most notably in the range of 0.1 – 0.2 V/Å. Because the error bars for these values are fairly large, we again verified this trend by applying an Arrhenius fit to the different electric field data sets, but this time we fixed the pre-exponential factor value to be that of the no field condition and evaluated the R^2 values for these fits (shown in Figure 5.3C as orange triangles). For reference, we also evaluated the R^2 values for using the

complete Arrhenius equation fitted to the no field data to explain the data recovered from electric fields (shown as red diamonds in Figure 5.3C). From Figure 5.3C, we observe as distinct falloff in the R^2 values of the Arrhenius equation created with the no field pre-exponential at a field strength of 0.1 V/Å and higher. This falloff not only matches our observation from Figure 5.1, but also closely matches the decrease in fit quality that is observed for using both the activation energy and pre-exponential factor from the no field condition to capture the data. Therefore, we can be confident that the pre-exponential factor changes with electric field strength and that the activation energy is unaffected.

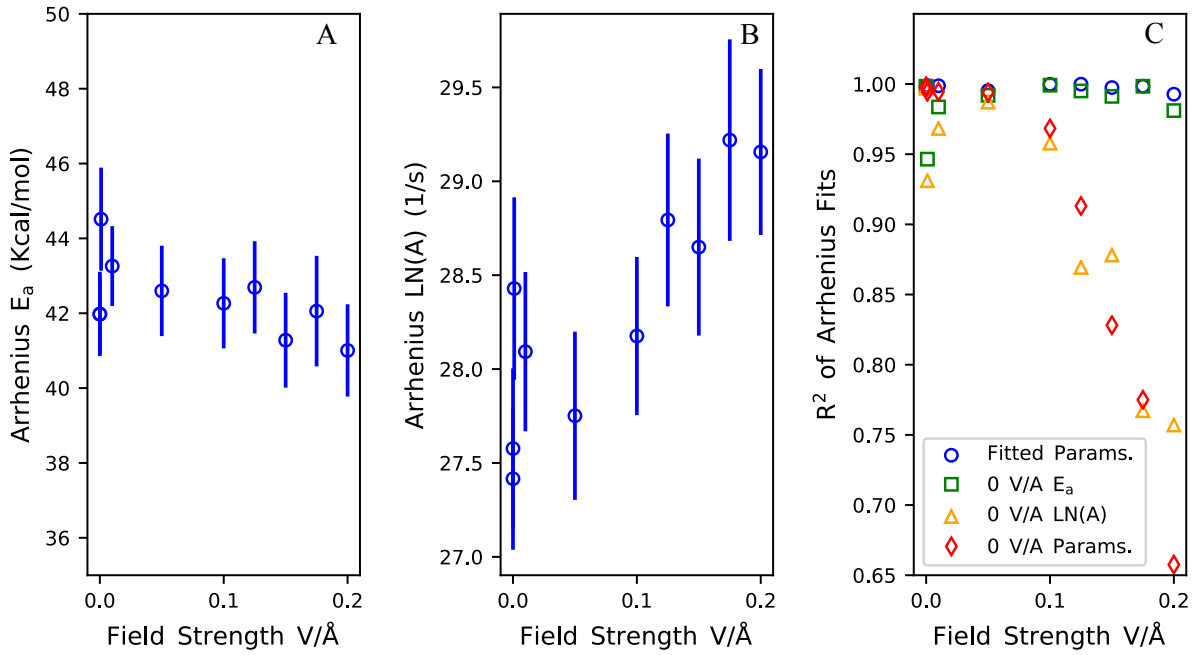


Figure 5.3: A) Calculated activation energies from Arrhenius fits plotted as a function of electric field strength of the system. Error bars shown represent are one standard deviation. B) Calculated natural log of the pre-exponential factor from Arrhenius fits plotted as a function of the electric field strength of the system. Error bars shown represent are one standard deviation. C) R^2 values for different Arrhenius equation fits to the different field strength data sets. The blue circles represent the R^2 values of Arrhenius equations generated from the data under a specific electric field condition. The green squares represent the R^2 values of Arrhenius equations generated from using the activation energy from the no field condition for all fits, and the pre-exponential factor calculated from the different field conditions. The orange triangles represent the R^2 values of Arrhenius equations generated from using the pre-exponential factor from the no field condition for all fits, and the activation energy calculated from the different field conditions. The red diamonds represent the R^2 values of fitting the data different field strength conditions to the Arrhenius parameters (i.e. activation energy and pre-exponential factor) calculated from the no field data set.

5.4.3 A Collision Theory Interpretation

With the insight that the electric fields are influencing the pre-exponential factors of the oxidation mechanism, it is beneficial to relate this to a mechanistic understanding of what aspects of the system are changing to produce this result. Because the impact of the electric field is accounted for in the simulation as a force through the following equation

$$F = q \cdot \vec{E} \qquad \text{Eq. 5.5}$$

where F is the force of the electric field exerted on an atom, q is the charge on an atom, and E is the electric field strength (V/Å), which in this system is applied along the z-axis only. Visual analysis of the MD trajectories showed a general net diffusion of toluene in the positive Z direction while O_2 did not exhibit such behavior. Analyzing the partial charges of the toluene atoms showed that, on average, the entire toluene molecule has a net positive charge and, consequently, the O_2 in close proximity have a negative charge. Because the atomic partial the Qeq method used adjusts the partial charges of atoms based on their neighbors, this result is not unexpected. Moreover, it is then intuitive that due to the net positive charge of toluene, external electric fields would not only cause a net diffusion of toluene along the field's axis, but that stronger fields would lead to greater velocities along this direction (see SI). Taking into account the net diffusion and increasing component velocity, and knowing that the electric field influences pre-exponential factor, it is natural to apply a collision theory lens to this phenomenon as

$$k = A * exp\left(\frac{-E_a}{RT}\right) = Z * \rho * exp\left(\frac{-E_a}{RT}\right) \qquad \text{Eq. 5.6}$$

where k is the rate constant, A is the pre-exponential factor, E_a is the activation energy, R is the molar gas constant, Z is the collision frequency, and ρ is the steric factor. From Eq. 6 it is clear that $Z * \rho$ is equivalent to A.

We evaluated the collision frequency by running a simulation at 1000 K, at field strengths of 0, $1.0x10^{-5}$, $1.0x10^{-3}$, $1.0x10^{-2}$, $2.0x10^{-2}$, $5.0x10^{-2}$, $7.5x10^{-2}$, $1.0x10^{-1}$, $1.25x10^{-1}$, $1.5x10^{-1}$, $1.75x10^{-1}$, and $2.0x10^{-1}$ V/Å, with the velocities initiated from the same seed so any changes can solely be attributed to the electric field. These simulations were run for three nanoseconds, with only the last two nanoseconds used for analysis. For these simulations, we evaluated the collision frequency by counting the number of instances (frames) where the minimum distance between any of the hydrogen atoms in the methyl and any of the oxygen atoms dropped below a certain threshold, then normalized by time. Collision frequencies for the different cutoffs are shown for the different field strengths in Figure 5.4A. In order to capture the net change in collision frequency when an electric field is applied, Figure 5.4B shows the ratio of the calculated collision frequency in an electric field to the collision frequency in the absence of a field. Regardless of the cutoff used for the calculation, we see that as field strength increases, the calculated Z increases in an exponential manner.

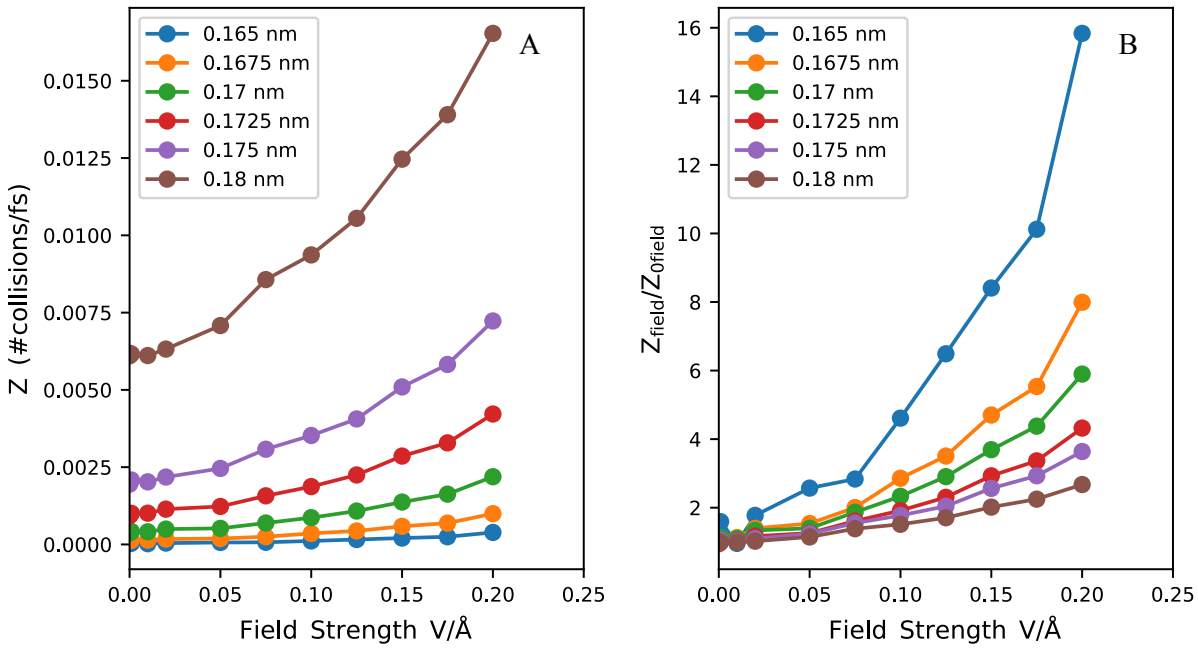


Figure 5.4: A) Collision frequency calculated as a function of electric field strength. The different lines plotted represent different distance cutoffs for determining what defines a collision. B) Collision frequencies of different field strengths and cutoffs normalized by the collision frequency calculated under no electric field.

Based on our conclusion that the activation energy is unaffected by the electric field, and following Eq. 6, then the trends observed in Figure 5.4B should be analogous the ratio of the reaction rate in an electric field to the reaction rate in the absence of a field, or, more precisely, the ratio of the pre-exponential factors. In Figure 5.5, we present the average ratio of the reaction rates in a given electric field to the corresponding rates in the absence of an electric field (averaged across the five temperatures sampled in each field) and compare this ratio of the collision frequencies as a function of field strength with a cutoff of 0.1675 nm. For reference, we also include the ratio of pre-exponential factors that were calculated from the Arrhenius fits where the activation energy of the no field system was used in the fit. As anticipated, we see very

good agreement between the increase in collision frequency and the acceleration in reaction rate. Moreover, we see that the ratio of the pre-exponential factors serves as a reasonable facsimile for denoting the effective boost afforded by the electric fields, further verifying that the change in reaction rates is not due to any change in the activation energy of reaction. We can therefore conclude that by inducing a net diffusion and positive net velocity in toluene, relative to the O_2 molecules, the electric field increases the frequency with which the toluene interacts and collides with O_2, resulting in a faster rate of reaction.

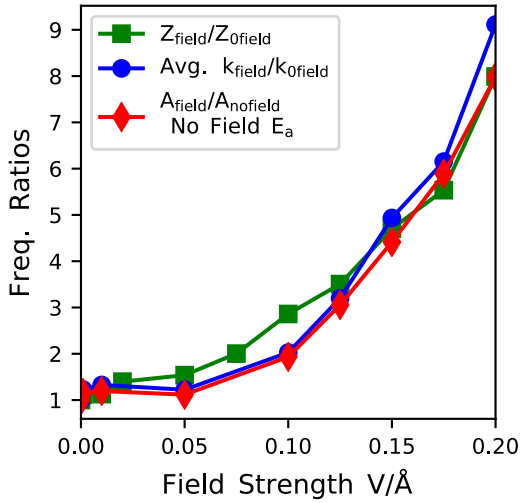


Figure 5.5: Increase in relevant frequencies as a function of electric field strength, presented as ratios relative to the no field condition. The green squares show the increase in collision frequency, Z (0.1675 nm cutoff), as a function of electric field strength. The blue circles represent the ratio of the calculated rates under electric fields relative to the zero electric field conditions (values are averaged across the five temperatures in each data set). The diamonds represent the ratios of the pre-exponential factors calculated under the different electric field conditions relative to the pre-exponential factor calculated for the zero field condition. These pre-exponential factors were fitted to the data with the Arrhenius equation using the activation energy calculated for the zero field condition.

5.4.4 A Transition State Theory Interpretation

In addition to a collision theory interpretation, it is also instructive to analyze this phenomenon through the lens of transition state theory. As noted in Figure 5.5, ratio of the rates of reactions, and most notable the pre-exponential factors exhibit an exponential increase with electric field strength. Relying on a thermodynamics interpretation of the reaction rate k, we can define this change in rate as

$$k = \frac{k_B T}{h} \mathcal{V}_o^{\upsilon-1} \exp\left(-\frac{\Delta G^{\neq}}{k_B T}\right) = \frac{k_B T}{h} \mathcal{V}_o^{\upsilon-1} \exp\left(\frac{\Delta S^{\neq}}{k_B}\right) \exp\left(-\frac{\Delta H^{\neq}}{k_B T}\right) \quad \text{Eq. 5.7}$$

$$\frac{k_1}{k_2} = \exp\left(\frac{\Delta {S_1}^{\neq} - \Delta {S_2}^{\neq}}{k_B}\right) \exp\left(-\frac{\Delta {H_1}^{\neq} - \Delta {H_2}^{\neq}}{k_B T}\right) \quad \text{Eq.5.8}$$

where $\Delta G^{\neq}$ is the change in free energy between the transition state and reactant state, $\Delta S^{\neq}$ is the change in entropy between the transition state and reactant state, $\Delta H^{\neq}$ is the change in enthalpy between the transition and reactant state, h is Planck's constant, $\mathcal{V}_o$ is the volume per molecule, and υ refers to the reaction order.[125] Imposing our assumption that the activation energy E_a, which is analogous the $\Delta H^{\neq}$, the last exponential term becomes unity. We can then simplify Eq. 8 to be either of the following two equations

$$\frac{k_{field}}{k_{no\ field}} = \exp\left(\frac{\Delta\Delta {S_{field}}^{\neq}}{k_B}\right) \quad \text{Eq. 5.9}$$

$$\ln\left(\frac{k_{field}}{k_{no\ field}}\right) = \frac{\Delta\Delta {S_{field}}^{\neq}}{k_B} \quad \text{Eq. 5.10}$$

where here we define $\Delta\Delta {S_{field}}^{\neq}$ to be the difference in the changes of entropy that occurs from the presence of the electric field. Consequentially, plotting the log of $k_{field}/k_{no\ field}$ as a function of electric field strength yields a strong linear trend as shown in Figure 5.6 ($R^2 = 0.98$). Note, this fit was only applied to field conditions of 0.1 V/Å and above, as these were the field conditions that experience a change in reaction rates. As indicated by Eq. 10, the slope of this

line indicates the change in entropy due to the presence of the external electric field, normalized by the Boltzmann factor k_B.

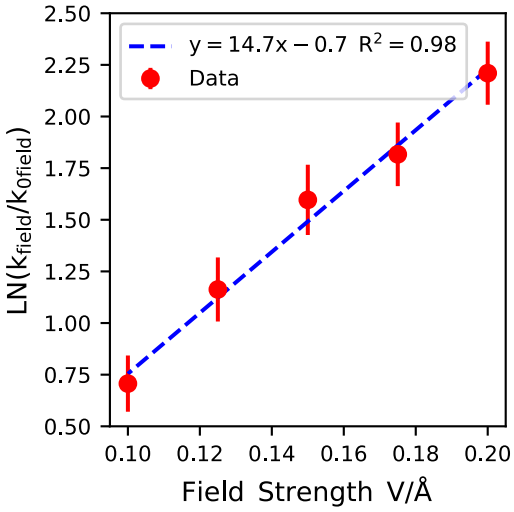


Figure 5.6: Natural log of the ratio of rates, normalized by the zero field condition, as a function of electric field strength. The red dots represent the "boost" for each field condition (averaged across the five temperatures), and the error bars represent one standard deviation. The blue dashed line is the a linear fit applied to the data. The slope of the line represents the change in entropy (divided by the gas constant) that occurs per V/Å. Note, the fit is only applied to field strengths of 0.1 V/Å or higher as these were the only data sets that were accelerated from the electric field (see Figure 5.3).

The combination of Eq. 10 and Figure 5.6 provide interesting implications as to how an electric field impacts the transition state of the oxidation reaction. The positive, linear nature of the slope of the line in Figure 5.6 indicates that presence of a strong electric field increases the change in entropy of the transition state relative to the reactants proportionally to the strength of the electric field. Viewing the change in entropy at a statistical mechanics level produces the following

$$A \propto \mathcal{V}_o^{v-1} \frac{\bar{q}^{\circ\neq}}{(q^{\circ}_{Toluene})(q^{\circ}_{O_2})} \qquad \text{Eq. 5.11}$$

where $\bar{q}^{\circ\neq}$ is the partition function of the transition state modulo the vibrational degree of freedom along the reaction coordinate, $q^{\circ}_{Toluene}$ is the partition function of toluene, and $q^{\circ}_{O_2}$ refers to the partition function of O_2, note that q° denotes the exclusion of the electronic degree of freedom from the partition function. Due to the methodology of how we carry out these simulations, it is not possible to discriminate whether or not the changes induced by the electric field are isolated to the reactants, transition state, or a combination of the two. However, it is possible to identify the partition functions that are affected by the electric field. Based on the positive trend of toluene z-velocity as a function of electric field (see SI), we can assert that there is a change in the translational partition function induced by the electric field. This argument is further strengthened as we also observe a linear relationship between the change in entropy and electric field. In addition to the translational partition function, we also find it likely that the rotational partition function changes as well. We monitored the orientation of various aspects of toluene with respect to the z-axis both with no electric field and a field strength of 0.2 V/Å (see SI). As expected, under the no field condition the orientation of toluene relative to the z-axis was completely random. However, under the 0.2 V/Å field we observe a slight change in orientation by 2-5 degrees depending on which aspect of toluene is monitored, although it still explored the full range of orientations. This slight change indicates that the orientation of toluene is not completely random when in the presence of a strong electric field, and thus a small loss in the rotational entropy of the molecule. Therefore, while we cannot ascribe the root cause of the change in entropy to be solely a decrease in a specific partition function of the reactants, nor an increase in a specific partition function of the transition states, we do find it very likely that the presence of a strong electric field induces a net increase in the change of the combination of the

translational and rotational partition functions of the transition state relative to the reactants occurs.

Despite not being able to specifically decompose the $\Delta\Delta S_{field}^{\neq}$ into the specific partition functions, the line of fit in Figure 5.6 does present a valuable opportunity to develop a thermodynamically driven model to predict the reaction rates (or times) for toluene oxidation for fields strengths exceeding 0.1 V/Å, the conditions where we observed a significant change in reaction rates. From this fit we can derive the following equation

$$\frac{\Delta\Delta S_{field}^{\neq}}{k_B} = \begin{cases} 14.7 * E_{field} - 0.7 & if\ (E_{field} \geq 0.1 \frac{V}{\text{Å}}) \\ 0 & if\ (E_{field} < 0.1 \frac{V}{\text{Å}}) \end{cases} \qquad \text{Eq. 5.12}$$

which provides us with the change in the entropic change of the reaction due to the presence of the electric field. With the knowledge that the activation energy is unchanged, along with the Arrhenius parameters for the no field condition, and Eq. 12 we can then complete out predictive model to estimate the mean reaction time, or rate, for a given temperature and electric field to be

$$k(T, E_{field}) = \exp\left(LN(A_{no\ field}) + \frac{\Delta\Delta S_{field}^{\neq}}{k_B}\right) * \exp\left(\frac{-E_{a\ no\ field}}{k_B T}\right) \qquad \text{Eq. 5.13}$$

where k is the mean reaction rate (1/s) recovered as a function of temperature T, and electric field strength, E_{field}, $LN(A_{no\ field})$ is the natural log of the pre-exponential factor recovered for the no field condition (y-intercept of the Arrhenius plot), $\frac{\Delta\Delta S_{field}^{\neq}}{k_B}$ is provided from Eq. 12, and $\frac{-E_{a\ no\ field}}{k_B}$ is the activation energy of the no field condition normalized by the gas constant (slope of the Arrhenius plot). Figure 5.7 shows a parity plot of the predicted reaction rates using Eq. 13 plotted against the actual reaction rates (inverse mean reaction time) that we measured from our simulations for all of the different temperatures and electric field strengths we sampled. Note that the we observe excellent agreement between the predicted and measured values, but we are able

to account for almost all of the variance in data (R^2=0.94). In addition to providing a useful model for predicting oxidation kinetics for our system at intermediate temperatures and electric field strengths, this result further substantiates our conclusion that the electric field has a purely entropic contribution. In light of this insight, it was all the more important that we sample this reaction in the lower temperature range of 1000 – 2000 K, rather than the higher temperature range that is typical of ReaxFF simulations,[5,106,109] as the reaction kinetics would be far more sensitive to enthalpic changes in this range, if any were to occur.

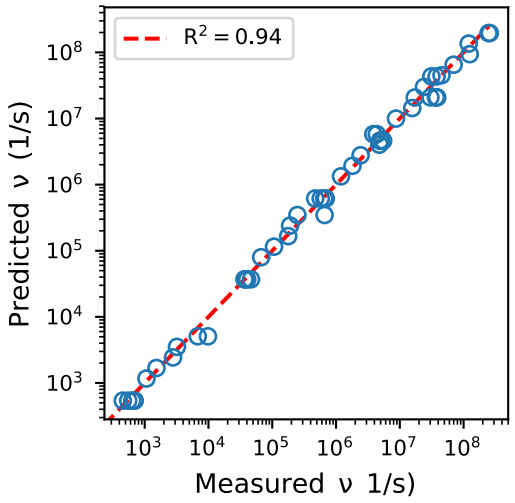


Figure 5.7: Parity plot of the calculated reaction rate from Eq. 13 versus the measured reaction rate from simulations (inverse mean reaction time) shown as blue circles. The red dashed line is a y=x line provided for reference.

5.4.5 Toluene Pyrolysis

In addition to the oxidation of toluene, we also studied the competing pyrolysis reactions of toluene. It is important to reiterate that quartic restraints were placed on the bonds that we did not wish to break, in order to ensure that we are only sampling the pathway we desired for a given

simulation, and that there was no oxygen present in the system (H_2 instead). Following the protocol described above, we initially sampled the competing pyrolysis reactions across a temperature range of 1250 – 2000 K, and then fit the data to an Arrhenius equation. Much like the oxidation mechanism, we find that the fitted Arrhenius parameters fall in line with literature values as activation energies range from 88-89 kcal/mol and 83-99 kcal/mol, and pre-exponential factors range from $3.1x10^{15}$-$1.4x10^{16}$ s^{-1} and $3.0x10^{15}$-$3.1x10^{15}$ for the methyl and H radical formation reactions, respectively[113,114] Additionally, and just as importantly, we also find that the selectivities between these two mechanisms also falls in line with literature, ranging between 0.07-0.3 across the temperature range of 1000-2000 K.[113]

Table 5.1: Arrhenius Parameters for Competing Pyrolysis Reactions

Reaction	Ea (kcal/mol)[a]	Pre-exponential Factor (1/s)[a]
Toluene --> C6H5 + CH3	87.9 (2.2)	2.00 (1.4) *10^{15}
Toluene --> C6H5CH2 + H	92.1 (2.0)	9.64 (6.4) *10^{14}

[a] Values in parenthesis are the estimated standard deviations.

Once we established that we could recover accurate kinetics, we then sampled these reaction pathways in an electric field of 0.2 V/Å, then fit the results to an Arrhenius equation. The results for the field and no field conditions are shown as an Arrhenius plot in Figure 5.8. Unlike the oxidation mechanism, we see that even in the presence of a very strong electric field, that the kinetics for both reactions is unaffected. In fact, across the entire temperature range of interest we observe remarkable agreement between the field and no field conditions.

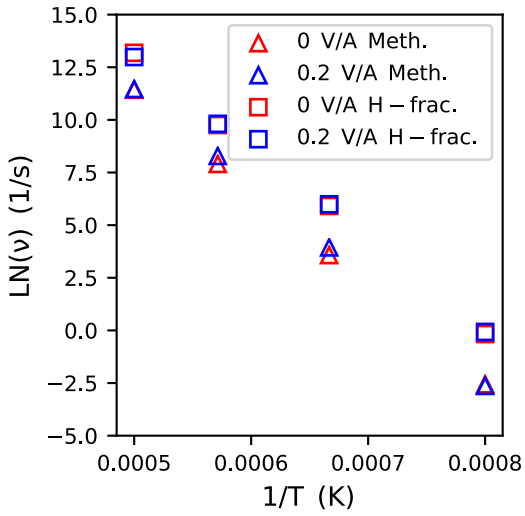


Figure 5.8: Arrhenius plot of competing pyrolysis reactions with and without an electric field. Triangles represent the reaction where the C-C bond of the methyl group to the aromatic ring breaks. Squares represent the reaction where one of the methyl hydrogen atoms breaks off. Data corresponding to no electric field are shown in red and data corresponding to 0.2 V/Å are shown in blue.

Given what we observed in the oxidation reactions, this result is not entirely unexpected. While strong electric fields had a significant impact of the accelerating oxidation kinetics, we noted that its impact was an entropic one on a bimolecular reaction. Both of these pyrolysis reactions, however, are unimolecular decompositions, that are enthalpically dominated. Therefore, the electric fields would not have a substantial impact of either of these reaction pathways. We do note that our results vary somewhat with Tan et al.,[106] who noted an effect from an electric field both for oxidation and pyrolysis. The benefits of enhanced sampling, as used in the current work, mitigate the effect of statistical uncertainty due to small sample sizes, which may be partially responsible for the differences. In addition, the system they studied had O_2 present, whereas we modeled this system in the absence of O_2, it could be that O_2 might play

some role in enhancing the pyrolysis mechanism by inducing charge in toluene, which may be of interest to investigate further in a future work.

5.5 Conclusions

Herein, we exhaustively sampled the oxidation and pyrolysis mechanisms of toluene in the presence of an external electric field. By utilizing the powerful IMetaD enhanced sampling method, we were able to accurately sample the kinetics of these reactions in a temperature range that falls in line with typical experiments. For all three mechanisms studied we found strong agreement with experimental values for kinetic parameters, while accelerating the simulations by orders of magnitude through the application of a bias potential. These results indicate promise in using the ReaxFF force field with the metadynamics family of enhanced sampling methods to efficiently and effectively characterize reacting systems.

Our investigation yields a clear mechanistic understanding into how an electric field influences some reaction kinetics while not others. We observe that an external electric field, provided it is sufficiently strong, can have a profound impact on oxidation kinetics. The electric field is shown to have a pre-dominantly entropic contribution to the system, increasing the change in entropy that occurs between the transition state and reactants. While we cannot discern whether the field influences strictly the reactant state, transition state, or both, we identify that this effect evolves in changing the translational and rotational partition functions of toluene. In addition, we demonstrated how collision theory is also able to explain the accelerated kinetics for the oxidation reactions. In line with these conclusions, we observed that electric fields do not have a significant impact on the unimolecular pyrolysis reactions of toluene, as these are enthalpically driven.

This invites the possibility that an electric field may be an effective way to enhance the selectivity of oxidation reactions over pyrolysis, which can be useful in the low temperature

regimes where pyrolysis can dominate for some hydrocarbons.[6] Among other areas for future investigation are more complex hydrocarbons such as Jet-Propellant 10 (JP-10), which ReaxFF has already been shown to model accurately.[107,108] Additionally, our focus for this study was on the initiation reactions, but further exploring the potential ramifications that an electric field could have on the reactions that follow is also of interest. Specifically, pairing these studies with other enhanced sampling techniques[26] could facilitate exploration of these systems at lower temperatures at a manageable computational expense.

6 Significance and Perspective for Future Applications

6.1 Summary of Work

This document contains the work of five projects, each which helped to further develop the metadynamics methodology. In Chapter 1, we outlined a modified version of infrequent metadynamics for systems that contain multiple, competing pathways. By placing quartic boundaries at or near the transition state of undesired pathways, these events can be blocked, forcing the system to only evolve through a specific desired pathway. This method was shown to preserve kinetics, pathway selectivities, as well as reduce the uncertainty of rate estimates and the number of samples required to characterize the system.

In Chapter 2, we assessed the performance of different types of collective variables to characterize reaction kinetics. In addition to being the first study to test if generic CVs such as SPRINT coordinates and CVHD were capable of recovering uncorrupted kinetics, we also evaluated their relative efficiencies. We found that collapsing distortions into a single CV via CVHD is typically the most efficient CV (i.e. largest acceleration factors) due its low dimensionality. Additionally, we found that the combination of CVHD and SPRINT coordinates can provide a significant speedup in sampling events without corrupting dynamics. This study demonstrated that infrequent MetaD is extremely versatile due to its compatibility with many types of CVs requiring varied levels of chemical intuition to construct.

In Chapter 3, we presented a novel workflow of using PBMetaD to bias SPRINT coordinates. We applied this method to explore the γ-ketohydroperoxide decomposition reaction network, and were able to capture the Korcek reaction mechanism, as well as other pathways. By adjusting the various simulation parameters, we showed that this framework is robust and compares favorably against non-MD based approaches when it comes to discovering pathways. This framework is

one of the few approaches for reaction pathway discovery that does not rely on chemical intuition or pre-determined reaction rules.

We then put forth a modified version of PBMetaD in Chapter 4. By partitioning collective variables into families (PBMetaDPF), we produced a new method that is able to accurately reproduce the free energy profiles of the collective variables biased, but at a lower computational expense. We showed that partitioning CVs leads to a linear speedup in convergence time (tested up to 78 CVs), with no additional error or hysteresis in the profiles recovered. Furthermore, in Appendix 6 we demonstrate that this method works on systems with multiple, different families. This method offers much promise in facilitating the exploration of very high-dimensional systems in a variety of contexts.

The last project, described in Chapter 5, applied infrequent MetaD to characterize the reaction kinetics of toluene oxidation and pyrolysis with ReaxFF force fields. Beyond recovering energy barriers that are strikingly close to experimental values, we were able to recover reaction events on the time scale of seconds, despite only running simulations for nanoseconds in length. Furthermore, we were able to discern that the application of a strong electric fields leads to a significant effect on the change in the entropy of reaction for oxidation reactions, significantly accelerating the rate of reaction. We also determined that for the range of electric sampled fields (<= 0.2 V/Å), there is no effect on toluene pyrolysis kinetics.

6.2 Moving Forward: Chemical Manifest Destiny

A longstanding obstacle in applying metadynamics is identifying the proper collective variables to bias. Further compounded by the poor scaling with dimensionality, metadynamics is typically ill suited for modeling systems that required high-dimensional representations (>3 CVs). In the context of chemical reactions, these challenges are particularly difficult to overcome

as systems can have multiple reaction pathways with very complex mechanisms. The projects and methods presented in this document address the underlying challenges of using molecular dynamics simulations and metadynamics to explore and characterize complex reacting systems. Furthermore, when considering the work contained in this document, the whole is greater than the sum of its parts. Combining the different methods and workflows of these projects presents a viable workflow for exploring a complex, multi-step reaction network, and then characterizing the kinetics of the pathways discovered.

Regardless of our level of intuition regarding a specific system, we have demonstrated that the PBMetaD method acting on SPRINT coordinates is a robust and efficient way to explore multi-step reaction networks. Through characterizing reversible pathways, as well was competing pathways (including those leading to enantiomers), we showed this method can recover a clear outline of the relevant reaction pathways, intermediates, and products for a given system without relying on *a priori* knowledge regarding potential mechanisms.[26] There is particular promise in enhancing the scalability of this framework to larger systems by utilizing partitioned families for SPRINT coordinates of particular atom types, and would be an interesting future investigation.

The roadmap of pathways and products from PBMetaD and SPRINT simulations can then be used to supply the necessary chemical intuition for infrequent MetaD simulations. Equipped with knowledge of the stable states and relevant degrees of freedom, sampling kinetics is fairly trivial. At the very least, one could merely bias the SPRINT coordinates and accurately recover the desired reaction rates and energy barriers.[23] Where the presence of competing pathways may have previously increased the uncertainty of the rate calculations, as well as the required number of simulations run, these alternative paths can merely be blocked to isolate the desired pathways for analysis.[22]

The combination of using PBMetaD and SPRINT to discover pathways and using of infrequent MetaD to quantify pathway kinetics is an effective and general workflow for completely resolving the desired characteristics of a complex system. Without relying on any assumptions regarding mechanisms, this framework is easily extendable to a variety of systems that previously would have been intractable to model both in size and complexity such as battery chemistry, atmospheric chemistries (pyrolysis, combustion, etc.), and reactions in condensed phases, to name a few. Additionally, we envision that the output from these investigations can be very powerful when paired with a kinetic Monte Carlo framework to model the behavior of the system at larger scales.

Finally, as we demonstrated in Chapter 5, using metadynamics to explore systems that are described with ReaxFF force fields is an effective way to efficiently and accurately model a system. As more ReaxFF force fields are developed for other types of chemistries, so too will the diversity of systems our methods can reach as well. There is also significant promise in further investigating the impact electric fields can have on various reaction mechanisms by using ReaxFF and MetaD. As the cost of these calculations continue to become cheaper, these frameworks will only become more appealing.

Appendix 1

Table A1.1: Mean escape times collected for 6-4 energy units cosine system with and without barriers.

Temperature ($1/k_BT$)	Energy Barrier (energy units)	System	Mean Escape Time
1.68	6	Single Path	306112 (36315)
		Two Path	365486 (240952)
	4	Single Path	13568 (1470)
		Two Path	14099 (1823)
1.26	6	Single Path	26274 (3015)
		Two Path	26080 (11011)
	4	Single Path	2752 (269)
		Two Path	2668 (387)
1.01	6	Single Path	6203 (634)
		Two Path	5842 (2045)
	4	Single Path	1026 (105)
		Two Path	1038 (158)
0.84	6	Single Path	2251 (230)
		Two Path	2233 (678)
	4	Single Path	533 (54)
		Two Path	571 (92)
0.72	6	Single Path	1164 (118)
		Two Path	1139 (315)
	4	Single Path	322 (31)
		Two Path	329 (51)

Values in parenthesis represent estimated standard deviation from bootstrapping method.

Table A1.2: Mean escape times collected for 5-4 energy units cosine system with and without barriers.

Temperature (K)	Energy Barrier (energy units)	System	Mean Escape Time
1.68	5	Single Path	69903 (7379)
		Two Path	67128 (21734)
	4	Single Path	14608 (1619)
		Two Path	14064 (2193)
1.26	5	Single Path	8672 (942)
		Two Path	9147 (2586)
	4	Single Path	2818 (285)
		Two Path	2746 (434)
1.01	5	Single Path	2512 (257)
		Two Path	2645 (685)
	4	Single Path	1051 (105)
		Two Path	1066 (179)
0.84	5	Single Path	1164 (115)
		Two Path	1128 (276)
	4	Single Path	546 (54)
		Two Path	524 (87)
0.72	5	Single Path	670 (62)
		Two Path	627 (144)
	4	Single Path	331 (33)
		Two Path	338 (58)

Values in parenthesis represent estimated standard deviation from bootstrapping method.

Modification for Amber14sb forcefield:

The type 9 dihedral terms corresponding to the Φ (C-N-CX-C) and Ψ (N-CX-C-N). Following the equation 4.61 in the Gromacs manual[126] the dihedral potential equations are defined by the following equation:

$$V_d(\phi_{ijkl}) = k_\phi(1 + \cos(n\phi - s)) \qquad \text{Eq. A1.1}$$

The Φ dihedral was manipulated by changing the k_Φ terms were changed to -30 and 30 for n=2 and 3, respectively. The Ψ dihedral was manipulated by multiplying the k_Φ terms by a factor of 10 for n=1-3.

Appendix 2

Determining Collective Variable Parameters

For simulations that biased the C-Cl bond distances directly, we utilized the same parameters as the ones used by Fleming et al.,[20] where the Gaussian width corresponded to one half the mean fluctuation of the bonded C-Cl bond distance (in order to be conservative in the bias deposition this was used for both bonds, despite larger fluctuations existing for the non-bonded C-Cl distance). For all generic CVs biased, the Gaussian width was set to be one half of the mean fluctuation of the CVs from unbiased simulations.

For simulations that biased the SPRINT coordinates the contact matrix was constructed using values of 6 and 12 for n and m, respectively, as these were described as typical values of Pietrucci et al.[15] Values for r_0 were selected to be 0.265 nm and 0.222 nm for C-C and C-Cl, respectively. These values are consistent with previous studies that biased SPRINT coordinates with similar interactions.[14,15]

For simulations that biased the CVHD variable, η, we defined the local distortions to based on the maximum and minimum bond lengths observed from a series of steered MD simulations. The bonded C-Cl distance never dropped below 0.15 nm and a transition was consistently observed when the bond distance exceeded 0.22 nm. Therefore these were selected as to be the r_i^{min} and r_i^{max}, respectively, for the bonded C-Cl distortion term. The r_i^{min} nonbonded C-Cl distortion was set to 0.20 nm because a transition was noted to occur (in the steered MD simulations) when the bond distance consistently dropped below 0.2 nm. The value for r_i^{max} was chosen to be 0.5 nm, corresponding to the restraint placed on the bond distance.

Table A2.1: Rates, *p*-values, and rejects from bootstrapping for the S_N2 reaction

Temp (K)	CV Biased	Pace (ps)	Teff (ns)[a]	p-value[b]	Rejects[c]
300	Bond Distances	1	276 (26)	0.59	4
		20	271 (35)	0.61	7
		100	183 (29)	0.36	106
	CVHD+Bond Distances	1	297 (31)	0.64	2
		20	296 (30)	0.58	2
		100	249 (59)	0.58	24
	SPRINT	1	276 (30)	0.53	7
		20	302 (43)	0.41	48
		100	184 (38)	0.59	21
	CVHD + SPRINT	1	281 (23)	0.58	8
		20	312 (38)	0.54	11
		100	300 (73)	0.51	30
450	Bond Distances	1	3.49 (0.42)	0.52	17
		20	4.22 (0.45)	0.65	4
		100	3.99 (0.39)	0.6	2
	CVHD+ Bond Distances	1	5.03 (0.47)	0.45	19
		20	4.33 (0.42)	0.65	3
		100	4.90 (0.43)	0.64	0
	SPRINT	1	4.56 (0.57)	0.43	37
		20	4.14 (0.58)	0.58	7
		100	4.47 (0.62)	0.5	17
	CVHD + SPRINT	1	4.92 (0.42)	0.48	13
		20	5.06 (0.45)	0.59	4
		100	4.82 (0.50)	0.53	16

600	Bond Distances	1	0.45 (0.06)	0.58	3
		20	0.66 (0.07)	0.68	1
		100	0.64 (0.07)	0.7	1
	CVHD + Bond Distances	1	0.71 (0.07)	0.69	1
		20	0.63 (0.06)	0.62	1
		100	0.56 (0.05)	0.63	1
	SPRINT	1	0.74 (0.07)	0.62	4
		20	0.64 (0.07)	0.47	7
		100	0.67 (0.09)	0.51	7
	CVHD + SPRINT	1	0.70 (0.06)	0.63	2
		20	0.74 (0.07)	0.59	3
		100	0.72 (0.07)	0.56	5

[a]Means and uncertainties generated from bootstrapping

[b]p-values generated from bootstrapping

[c]Number of rejects out of 1000 samples

Table A2.2: Arrhenius Plot parameters for the S_N2 reaction for different CVs biased and bias deposition rates

CV Biased	Pace (ps)	Arrhenius Plot Energy Barrier (kcal/mol)[a]	Pre-Exponential Factor (1/s)[a]
Bond Distances	1	7.66 (0.18)	$1.45\ (0.35) *10^{12}$
	20	7.16 (0.20)	$6.68\ (1.58) *10^{11}$
	100	6.70 (0.22)	$4.55\ (1.33) *10^{11}$
CVHD + Bond Distances	1	7.19 (0.17)	$6.15\ (1.29) *10^{11}$
	20	7.34 (0.16)	$8.04\ (1.61) *10^{11}$
	100	7.35 (0.28)	$8.30\ (2.40) *10^{11}$
SPRINT	1	7.05 (0.17)	$5.35\ (1.11) *10^{11}$
	20	7.33 (0.22)	$7.95\ (2.07) *10^{11}$
	100	6.67 (0.29)	$4.05\ (1.33) *10^{11}$
CVHD + SPRINT	1	7.14 (0.13)	$5.92\ (1.02) *10^{11}$
	20	7.18 (0.18)	$5.88\ (1.22) *10^{11}$
	100	7.07 (0.29)	$5.49\ (1.63) *10^{11}$

[a]Uncertainties generated from linear interpolation

Table A2.3: Rates, p-values, and rejects from bootstrapping for the Diels-Alder reaction

Temp (K)	CV Biased	Teff (ns)[a]	p-value[b]	Rejects[c]	Sample Size
300	Bond Distances	9.05 (2.16) $*10^{14}$	0.66	0	35
	CVHD + Bond Distances	6.47 (1.09) $* 10^{14}$	0.53	9	101
	SPRINT[d]	4.48 $*10^{14}$	0.009[e]	N/A	4
450	Bond Distances	8.15 (1.76) $*10^{7}$	0.49	53	32
	CVHD + Bond Distances	5.84 (0.76) $*10^{7}$	0.6	1	146
	SPRINT[d]	1.61 $*10^{8}$	0.34	N/A	8
600	Bond Distances	1.63 (0.30) $*10^{4}$	0.58	11	64
	CVHD + Bond Distances	1.73 (0.20) $*10^{4}$	0.61	2	173
	SPRINT[d]	1.75 $*10^{4}$	0.94	N/A	8
900	Bond Distances	2.47 (0.35)	0.63	1	81
	CVHD + Bond Distances	2.44 (0.26)	0.62	1	187
	SPRINT[d]	1.63	0.87	N/A	15

[a]Means and uncertainties generated from bootstrapping

[b]p-values generated from bootstrapping

[c]Number of rejects out of 1000 samples

[d]No bootstrapping or error analysis was carried out due to small sample size

[e]Failed KS-Test ($p < 0.05$) but due to small sample size

Table A2.4: Arrhenius Plot parameters for the Diels-Alder reaction for different CVs biased and bias deposition rates

Collective Variable Biased	Arrhenius Plot Energy Barrier (kcal/mol)[a]	Pre-Exponential Factor (1/s)[a]
Bond Distances	29.9 (0.26)	$5.97\ (1.54) * 10^{6}$
CVHD + Bond Distances	29.5 (0.14)	$5.37\ (0.84) * 10^{6}$
SPRINT	29.5 (1.26)	$4.52\ (6.48) * 10^{6}$

[a]Uncertainties generated from linear interpolation

Table A2.5: Comparison of the free energy basin volume, deposited hill volume, diffusivity, and average acceleration factors for systems with different biased CVs and temperatures.

System	Basin Volume (normalized by kT)	Deposited Hill Volume (normalized by kT)	Diffusivity (CV^2/ps)	Average Alpha
Bond Dist. 300K	$6.90 * 10^{-2}$	$4.72 * 10^{-6}$	$2.04 * 10^{-6}$	60
Bond Dist. 600K	$3.45 * 10^{-2}$	$2.36 * 10^{-6}$	$2.19 * 10^{-6}$	1
Sprint. 300K	$2.53 * 10^{-1}$	$4.72 * 10^{-4}$	$7.75 * 10^{-6}$	174
Sprint. 600K	$1.26 * 10^{-1}$	$2.36 * 10^{-4}$	$7.31 * 10^{-6}$	1

Convergence of Free Energy Surface

Converged free energy surfaces of the S_N2 reaction are shown in Figure A.2. To verify that these surfaces are converged, we took the data at 50%, 75%, and 100% of the simulation shown here. Outside of the high energy regions, which are rarely sampled and are not relevant for the process sampled, there is very strong agreement between the three different FESs. Furthermore, we implemented the climbing image nudged elastic band method (CI-NEB) to calculate the energy barriers of the different FESs. It is clear from the Table S6 that there is calculated energy barriers for the forward and reverse reactions are very similar, well within the margin (< kT).

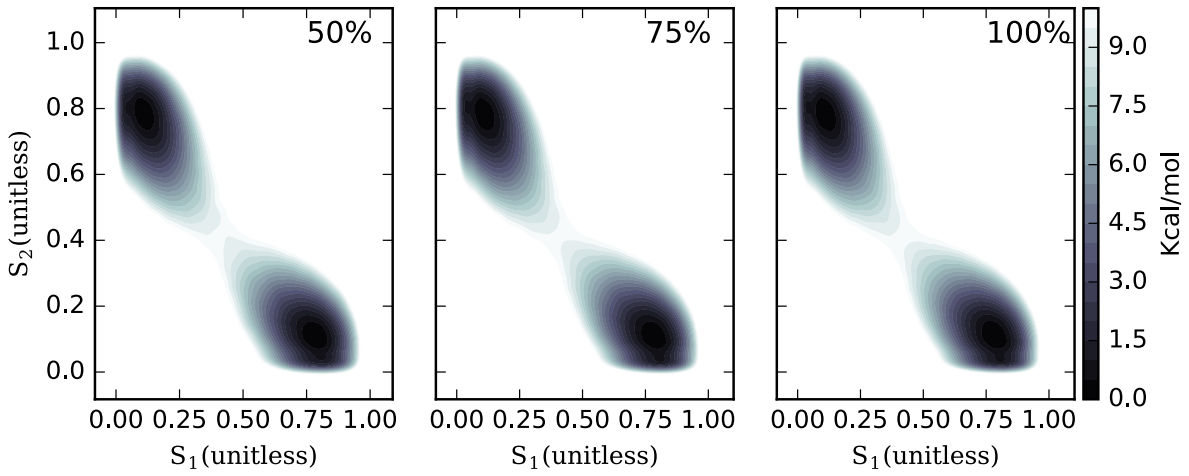


Figure A2.1: Free energy surfaces (FESs) of the S_N2 reaction projected upon the SPRINT coordinates of the chlorine atoms (biased) at 50%, 75%, 100% of the simulation.

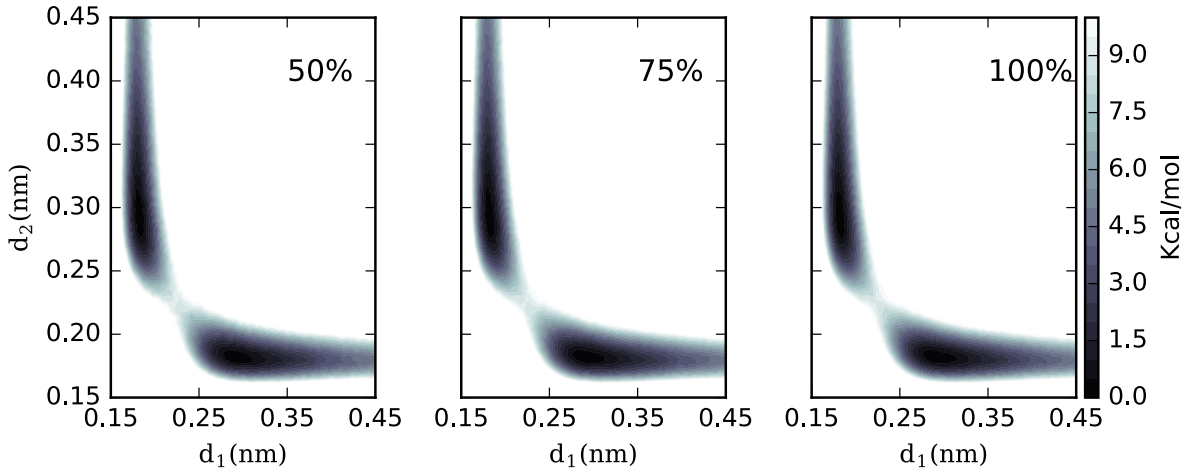


Figure A2.2: Free energy surfaces (FESs) of the S_N2 reaction projected upon the C-Cl bond distances (biased) at 50%, 75%, 100% of the simulation.

Table A2.6: Free energy barriers for different FES at 50%, 75%, and 100% of the simulation length.

Biased CV	Percent of Total Simulation Length	Forward Reaction Free Energy Barrier (kcal/mol)	Reverse Reaction Free Energy Barrier (kcal/mol)
SPRINT Coordinates	100	9.57	9.61
	75	9.53	9.59
	50	9.52	9.58
C-Cl Bond Distances	100	8.91	8.95
	75	8.91	9.01
	50	8.88	9.04

Appendix 3

Procedure for Identifying Reactive Events

Identifying transition events and intermediate species in trajectories is a common challenge in analyzing reactive systems. Even with outputting every 100 fs to the trajectory file, the length (> 1,000,000 frames) and number (> 400) of trajectories collected makes manually screening these videos intractable. Therefore, we implemented a series of steps to automatically clean the trajectories to either reduce the millions of frames in a trajectory to a few key frames of likely reactive events.

The first step is to post-process the trajectory using the MDTraj[76] python package. At each frame in the trajectory, a neighbor list is created for every atom, with the cutoff for a neighbor set to be 0.16 nm, a typical bond C-C bond length. Frames with identical neighbor-lists for all of the atoms are taken to represent an identical set of species. If a change in any of the neighbor-lists is noted, and the change persists for 10 consecutive frames without returning to the original reference set, then a reactive event is noted to likely have occurred at that frame. These individual frames are then pulled from the trajectory as individual pdb files using the CPPTRAJ[127] module in AMBER. While effective, this method proved to be overly sensitive for segments that included two or three molecules, as it sometimes identified non-reactive interactions between molecules to be likely reactive events (i.e. molecules drifting close together and then far apart). However, even with these extra frames, the number of frames of interest was typically reduced to the order of hundreds or thousands of frames rather than the 1-2.5 million of a full trajectory.

To compensate for the sensitivity of the neighbor-list, we utilized the Open Babel[77] and Pybel[128] modules in python to read and convert the pdb files to SMILES strings, which are two-

dimensional representations of the species. If consecutive frames of interest shared equivalent SMILE string representations, reactive events were determined to not have occurred and these repeated structure frames were discarded. While this step further refined the collection of likely events, SMILE string generation is based on neighbor-lists and cutoffs, rather than a bond order analysis. Therefore, sometimes bonds stretched far from equilibrium, but not broken (which is anticipated since these are biased simulations), can result in SMILE strings showing changes in false species.

To address this, we utilized the Gaussian 09[74] program to carry out geometry optimizations on the remaining structure in order to relax the bonds back to their equilibrium distances. The resulting log files were then converted to SMILE strings using the Open Babel python package. Any optimizations that failed to converge for any reason were manually expected to identify the proper structure/SMILE string (common for frames with multiple molecules). Consecutive frames that the show equivalent SMILE strings were again discarded, leaving only the frames at which new species are formed in the trajectory along with corresponding species present in that frame.

Table A3.1: Reactions Observed with Calculated Energy Barriers From Gaussian

Reactant (SMILE string)	Product (SMILE string)	PM6 Energy Barrier (kcal/mol)*
C(=O)([CH3])O[CH]=O.[H][H]	C(=O)([CH3])[OH].[C]=O.[H][H]	35.6
C(=O)([CH3])O[CH]=O.[H][H]	C(=[CH2])=O.O=[CH][OH].[H][H]	24.9
C(=[CH2])([OH])O[CH]=O.[H][H]	C(=[CH2])=O.O=[CH][OH].[H][H]	10.0
C(=[CH2])=O.O=[CH][OH].[H][H]	C(=O)([CH3])O[CH]=O.[H][H]	8.8
C(=[CH2])=O.O=[CH][OH].[H][H]	C(=[CH2])([OH])O[CH]=O.[H][H]	10.6
[C@@H]([CH3])([OH])O[CH]=O	[CH](=O)[CH3].O=[CH][OH]	15.9
[C@H]([CH3])([OH])O[CH]=O	[CH](=O)[CH3].O=[CH][OH]	14.6
[CH2]([CH2][CH]=O)O[OH]	[CH2]([CH3])O[OH].[C]=O	59.5
[CH2]([CH2][CH]=O)O[OH]	[CH2]=[CH2].[OH]O[CH]=O	40.2
[CH2]=[CH2].[OH]O[CH]=O	[CH2]1[CH2]O1.[CH](=O)[OH]	0.0
[CH2]([CH2][CH]=O)O[OH]	[CH2]([CH2]C(=O)[OH])[OH]	44.6
[CH2]([CH2][CH]=O)O[OH]	[CH](=O)[CH2][CH]=O.[OH2]	28.6
[CH2]([CH2][CH]=O)O[OH]	[CH2]1[CH2][C@@H](OO1)[OH]	18.6
[CH2]([CH2][CH]=O)O[OH]	[CH2]1[CH2][C@H](OO1)[OH]	18.0
[CH2]([CH3])O[OH].[C]=O	[CH2]([CH3])[OH].C(=O)=O	0.0
[CH2]([OH])OC(=O)[CH3]	[CH2]([OH])[OH].[CH2]=C=O	46.7
[CH2]([OH])OC(=O)[CH3]	[CH2]=O.[CH3]C(=O)[OH]	17.4
[CH2]([OH])OC(=O)[CH3]	[CH2]=O.[CH2]=C=O.[OH2]	46.4
[CH2]([OH])OC(=O)[CH3]	[CH][OH].[CH3]C(=O)[OH]	61.5
[CH2]([OH])[OH].[CH2]=C=O	[CH2]([OH])OC(=O)[CH3]	23.7
[CH2]([OH])[OH].[CH2]=C=O	[CH2]=O.[CH2]=C=O.[OH2]	33.5
[CH2]1[CH2]O1.[CH](=O)[OH]	[CH2]O[CH2].[CH](=O)[OH]	51.9
[CH2]1[CH2][C@@H](OO1)[OH]	C(=[CH2])=O.O=[CH][OH].[H][H]	46.0
[CH2]1[CH2][C@@H](OO1)[OH]	[CH2]=O.[CH3]C(=O)[OH]	55.7
[CH2]1[CH2][C@@H](OO1)[OH]	[CH](=O)[CH3].O=[CH][OH]	48.3
[CH2]1[CH2][C@@H](OO1)[OH]	[CH2]([CH2][CH]=O)O[OH]	29.4
[CH2]1[CH2][C@@H](OO1)[OH]	[CH](=[CH2])[OH].O=[CH][OH]	47.6
[CH2]1[CH2][C@H](OO1)[OH]	[CH2]=O.[CH3]C(=O)[OH]	54.1
[CH2]1[CH2][C@H](OO1)[OH]	[CH](=O)[CH3].O=[CH][OH]	47.2
[CH2]1[CH2][C@H](OO1)[OH]	[CH](=[CH2])[OH].O=[CH][OH]	46.2
[CH2]1[CH2][C@H](OO1)[OH]	[CH2]([CH2][CH]=O)O[OH]	27.2
[CH2]1[CH2][C@H](OO1)[OH]	C(=[CH2])=O.O=[CH][OH].[H][H]	43.8
[CH2]=O.[CH2]=C=O.[OH2]	[CH2]([OH])[OH].[CH2]=C=O	15.1
[CH2]=O.[CH3]C(=O)[OH]	[CH2]([OH])OC(=O)[CH3]	5.1
[CH](=O)[CH2][CH]([OH])[OH]	[C@@H]1([CH2][C@H](O1)[OH])[OH]	36.2

[CH](=O)[CH2][CH]=O.[OH2]	[CH](=O)[CH]=[CH][OH].[OH2]	14.5
[CH](=O)[CH2][CH]=O.[OH2]	[CH](=O)[CH2][CH]([OH])[OH]	19.4
[CH](=O)[CH3].O=[CH][OH]	[C@H]([CH3])([OH])O[CH]=O	1.7
[CH](=O)[CH3].O=[CH][OH]	[C@@H]([CH3])([OH])O[CH]=O	1.7
[CH](=O)[CH]=C=O.[OH2].[H][H]	[CH](=[CH]C(=O)[OH])[OH].[H][H]	0.0
[CH](=O)[CH]=[CH][OH].[OH2]	[CH](=O)[CH]=C=O.[OH2].[H][H]	17.0
[CH](=O)[CH]=[CH][OH].[OH2]	[CH](=O)[CH2][CH]=O.[OH2]	13.6
[CH](=[CH2])[OH].O=[CH][OH]	[CH](=O)[CH3].O=[CH][OH]	4.5
[CH](=[CH]C(=O)[OH])[OH].[H][H]	[CH](=O)[CH]=C=O.[OH2].[H][H]	19.3
[CH][OH].[CH3]C(=O)[OH]	[CH2]=O.[CH3]C(=O)[OH]	35.6
[CH][OH].[CH3]C(=O)[OH]	[CH2]([OH])OC(=O)[CH3]	3.1

Appendix 4

A4.1 PBMetaD and PBMetaDPF Simulations

All simulations were performed using GROMACS[129] 2016 simulation engine and PLUMED 2 (Developer's version).[41,123] All systems with Lennard-Jones (LJ) particles were set up in vacuum in a periodic cubic box of side equal to 20 nm, with the requisite number of atoms. There were no charges in the system. Van der Waals cut-off was set to 9.5 nm. The energy of the system was minimized using the steepest descent algorithm. For production run, the system was simulated in the NVT ensemble (T = 300 K) using the Bussi-Donadio-Parrinello thermostat[130] (temperature coupling time constant, τ = 0.1 ps). A 2 fs timestep was used.

The OPLS-AA force field was used for all simulations. The Lennard-Jones interactions were calculated using the following form:

$$V(r) = 4\varepsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right) \qquad \text{Eq. A4.1}$$

For the 3- and 7-particle systems, the σ and ε were set to 0.393 nm and 30 kJ/mol, respectively. For the 13-particle system, the σ and ε were set to 0.393 nm and 11 kJ/mol, respectively. A smaller ε was used for the 13-particle system so that it was possible to converge higher energy values (~100 kJ/mol) within accessible computational power.

A4.2 Parallel Tempering Simulations (LJ_3 and LJ_{13})

Parallel tempering simulations of the 3-particle LJ system were set up with 16 replicas, and temperatures for the NVT simulation ranging from 300-1000 K. The temperature spacing

(300.0, 318.85, 339.54, 362.33, 387.51, 415.42, 446.48, 481.18, 520.12, 564.03, 613.78, 670.48, 735.48, 810.50, 897.73, 1000.00) was calculated using the formula proposed by Prakash et al.[131]

$$\frac{1}{T_i} = \frac{1}{T_{i-1}} - \sqrt{\frac{k}{T_{i-1}}} \qquad \text{Eq. A4.2}$$

Where T_i = temperature of the current replica, T_{i-1} = temperature of the previous replica, k = constant optimized for a specific starting temperature, end temperature and number of replicas. The authors derived it for biomolecular systems where the heat capacity changes with temperature. Since the LJ systems also exhibit varying heat capacities with temperatures, this scheme was chosen in favor of a geometric temperature distribution (which is useful for systems with constant heat capacities).[131]

Exchanges were attempted every 250 simulation steps. Other details of the NVT simulation remained the same as the metadynamics runs. A converged profile was obtained after ~ 2 μs/replica simulation time.

Parallel tempering simulations of the 13-particle LJ system were set up with 32 replicas, and temperatures for the NVT simulation ranging from 300-5000 K. The temperature spacing (300, 302.98, 306.02, 309.10, 312.22, 315.40, 318.62, 321.89, 325.22, 328.59, 332.02, 335.50, 339.04, 342.63, 346.29, 350.0, 383.88, 422.98, 468.43, 521.70, 584.71, 660.0, 751.0, 862.57, 1001.31, 1177.04, 1404.35, 1705.97, 2118.86, 2706.67, 3587.15, 5000.00) was calculated using the method of Prakash et al.[131] Exchanges were attempted every 125 simulation steps. Other details of the NVT simulation remained the same as the metadynamics runs. A converged profile was obtained after ~ 2 μs/replica simulation time.

A4.3 Clustering

The trajectories for LJ systems (13 particles and 7 particles) were clustered using a 2-step method. First, using the GROMACS tool *gmx cluster* (method gromos[132], cut-off 0.05 nm), cluster numbers were assigned to frames. Since this tool considers all atoms as distinguishable particle when calculating the RMSD of structures, another clustering method was required to cluster the resulting structures. Second, using the Bag-of-Bonds method,[133] where the inter-atomic distances are used as features (78 features for the 13-particle system, and 21 features for the 7-particle system), and the average linkage method of hierarchical clustering, and limiting the clustering to structures of all atoms only, the most similar clusters from the first step were combined into one. Thus, new cluster numbers were assigned to every frame. The entire simulation trajectory was used.

A4.5 Reweighting

After assigning cluster numbers, the weight for each cluster was calculated using the Torrie-Valleau[18] method. The weights were calculated using the formula:

$$W = \sum \exp(\beta V) \qquad \text{Eq. A4.3}$$

where, V = bias from the PBMetaDPF potential, $\beta = 1/k_BT$ (T = temperature, k_B = Boltzmann's constant), W= weight of each cluster which is the sum of all the weights for the frames where the cluster was identified. For the calculation of V, the simulation was rerun using the GROMACS command *mdrun rerun* and the information from the Gaussian hills deposited during the production run. This provided an estimate of V at each frame of the trajectory. The Gaussian deposition pace was set to 1000000 during rerun to prevent the deposition of new Gaussian hills and changing free-energy estimates.

The weights (Ws) were then normalized to obtain a percentage representation of weights. At this step, only frames corresponding to the regions of phase space that we were interested in (red boxes in Figure 5) were reweighted for. As mentioned earlier, clustering was done using the entire trajectory, however reweighting was done on a subset of frames only.

A4.6 Well-tempered Metadynamics Simulations (LJ_7 SYSTEM)

In order to create an independent reference for the LJ_7 system we carried out a well-tempered metadynamics (WTMetaD) simulation biasing the second and third moments of coordination, as was done by Nava et al.[95] These simulations had an initial Gaussian height of 2 kJ/mol, Gaussian width of 0.01 along each CV, bias deposition pace of 500 steps, and a bias factor of 10. For consistency, we applied the same restraints on interatomic distances placed at 3.5 nm. The coordination number was calculated using the formula

$$s = \frac{1-\left(\frac{r}{r_0}\right)^n}{1-\left(\frac{r}{r_0}\right)^m} \qquad \text{Eq. A4.4}$$

where, r_0 = 0.6 nm, n = 8, and m = 16, for this system.[96]

Additionally, these simulation results were also reweighted to capture a free-energy profile for the interatomic distance. We followed the reweighting procedure described above, except we accounted for contributions from all 21 interatomic distances to create the free-energy profile.

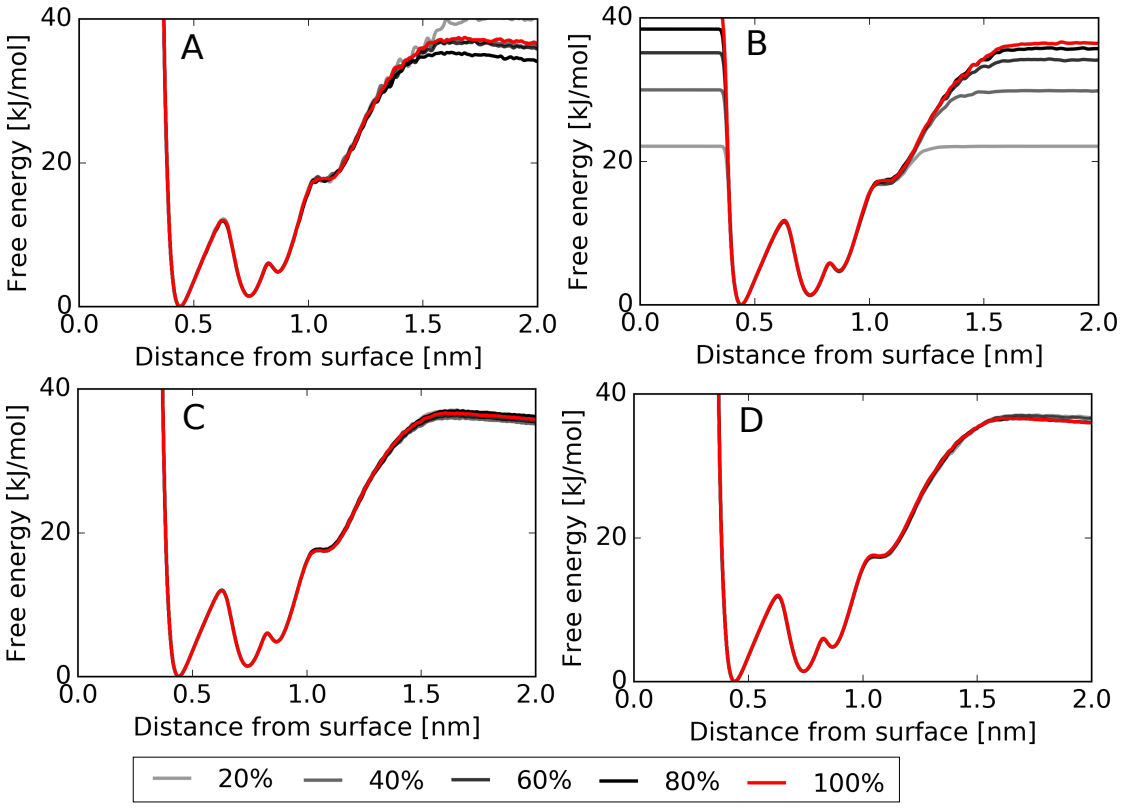


Figure A4.1: For the 13-particle LJ system, evolution of the free-energy profiles for (A & C) PBMetaDPF and (B & D) PBMetaD (averaged over 78 profiles) for the first (A & B) 100 ns and (C & D) total simulation time of 2 μs.

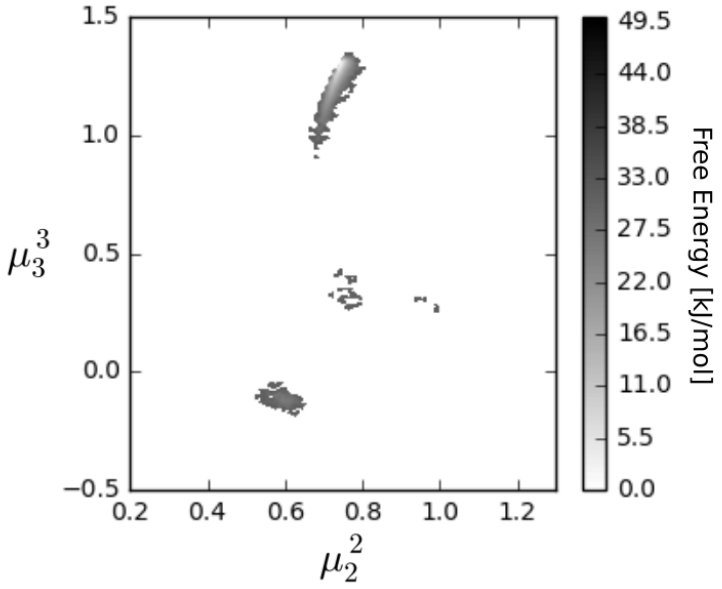


Figure A4.2: (left) Free-energy surface recovered from MD simulations without enhanced sampling of the 2D 7-particle LJ system for ~ 2 μs at 300 K.

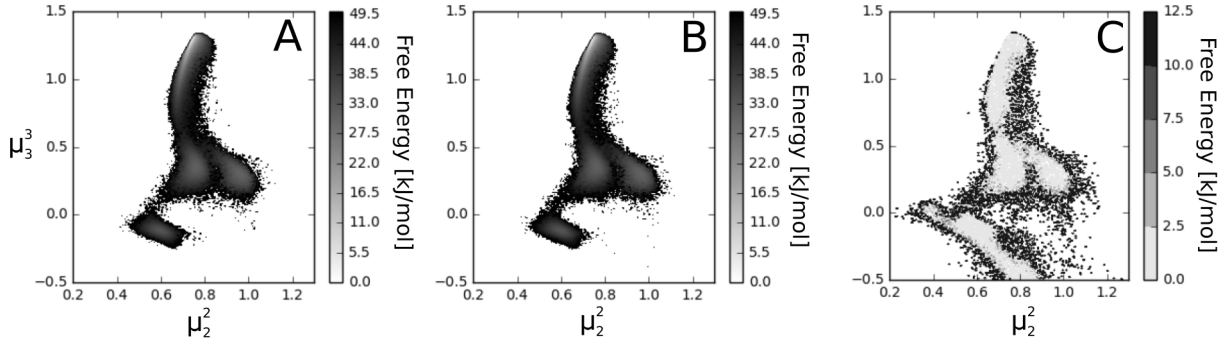


Figure A4.3: Free-energy surface for the 7-particle LJ system reweighted for the second and third moments of coordination numbers using (A) PBMetaDPF and (B) PBMetaD. (C) Difference in free-energy between PBMetaD and PBMetaDPF free-energy surfaces.

Table A4.1: Weights calculated from biased trajectories of PBMetaD and PBMetaDPF simulations.

Weight of top cluster/ Region	Region 1	Region 2	Region 3	Region 4
PBMetaDPF trajectory, biased weight	99.93%	92.5%	95.9%	97.9%
PBMetaDPF trajectory, unbiased weight	94.50%	80.89%	86.67%	91.62%
PBMetaD trajectory, biased weight	99.72 %	98.07%	98.50%	98.19%
PBMetaD trajectory, biased weight	94.50 %	97.28%	96.98%	87.83%

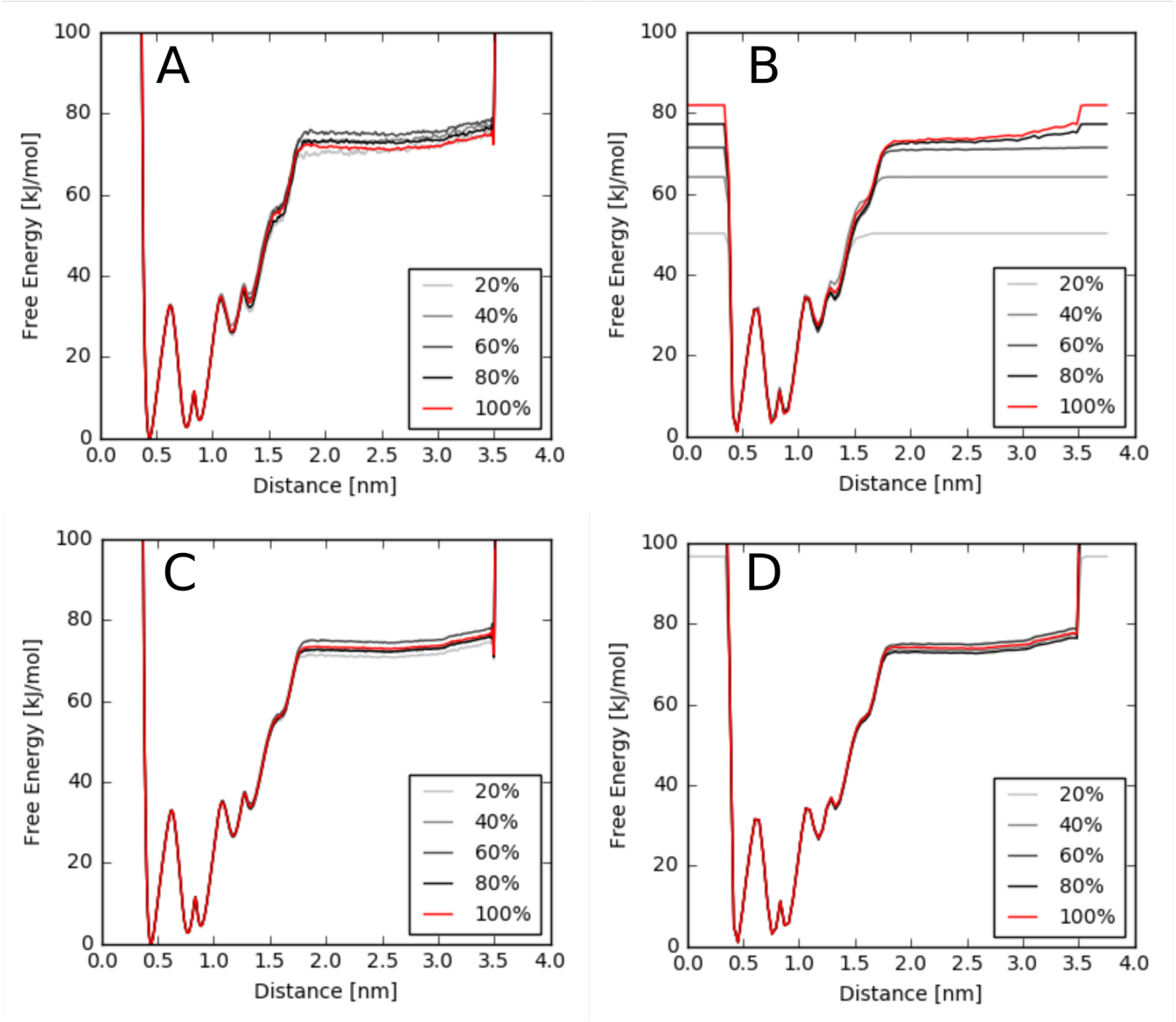


Figure A4.4: For the 7-particle LJ system, evolution of the free-energy profiles for (A & C) PBMetaDPF and (B & D) PBMetaD (averaged over 21 profiles) for the first (A & B) 125 ns and (C & D) total simulation time.

Appendix 5

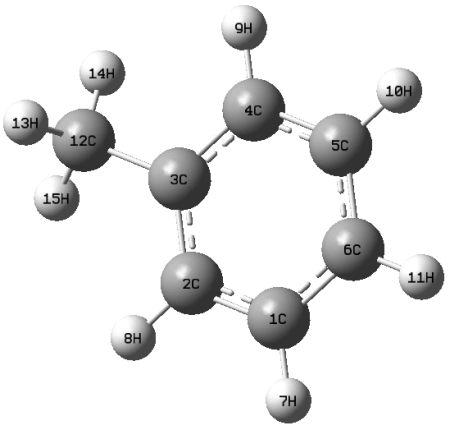


Figure A5.1: Toluene molecule with atomic labels for reference.

Table A5.1: Average atomic partial charge of toluene with and without an electric field.

	System	
	Oxidation, No Field*	Oxidation 0.2 V/A**
Atom	Mean Partial Charge (multiple of electron)*	
1C	-0.049 (0.012)	-0.050 (0.014)
2C	-0.082 (0.013)	-0.081 (0.015)
3C	0.077 (0.012)	0.072 (0.014)
4C	-0.082 (0.013)	-0.081 (0.014)
5C	-0.049 (0.012)	-0.050 (0.014)
6C	-0.056 (0.012)	-0.057 (0.014)
7H	0.185 (0.019)	0.0181 (0.02)
8H	0.168 (0.022)	0.163 (0.024)
9H	0.167 (0.022)	0.164 (0.024)
10H	0.185 (0.019)	0.181 (0.020)
11H	0.185 (0.019)	0.181 (0.021)
12C	-0.386 (0.046)	-0.375 (0.047)
13H	0.227 (0.035)	0.219 (0.036)
14H	0.225 (0.028)	0.218 (0.031)
15H	0.225 (0.028)	0.218 (0.031)

*Averaged over 2.4 ns biased simulation (C12-H13 biased)
** Averaged over 1.9 ns biased simulation (C12-H13 biased)
*** standard deviation of partial charges shown in parenthesis.

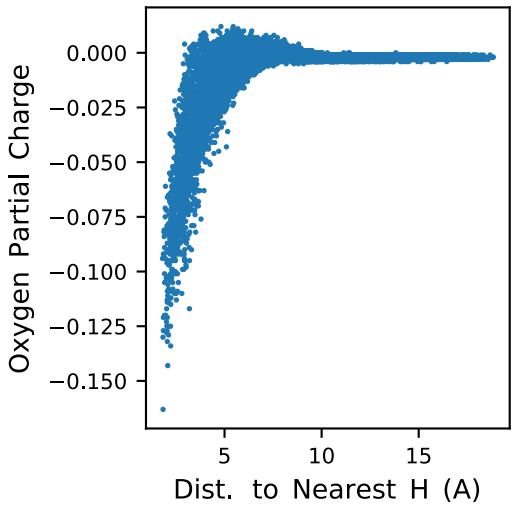


Figure A5.2: Partial charges of oxygen plotted against the distance to the nearest hydrogen in toluene. The data plotted corresponds to the 200 oxygen atoms present in a given frame of a trajectory. The trajectory is 2.4 ns long and the frames were selected to be every 20 ps over the course of the simulation. Note this is also from a biased simulation. Charge is shown in terms of electron charge (i.e. a proton has a value of 1.0).

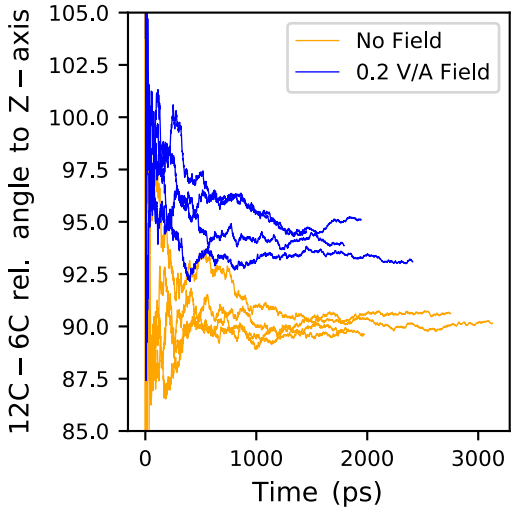


Figure A5.3: The average orientation of vector between atoms 12C and 6C (see Figure A5.1) and the z-axis is plotted for four simulations without an electric field (orange) and with an electric field of 0.2 V/A (blue lines). Note the orientation is periodic between 0 and 180 degrees. Orientation defined as the arccosine of dot product of the vector defined for the two atoms and the vector (0,0,1), normalized by their magnitudes.

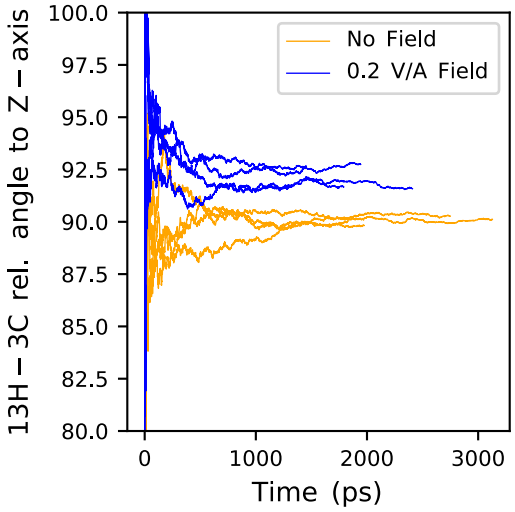


Figure A5.4: The average orientation of vector between atoms 13H and 3C (see Figure A5.1) and the z-axis is plotted for four simulations without an electric field (orange) and with an electric field of 0.2 V/A (blue lines). Note the orientation is periodic between 0 and 180 degrees. Orientation defined as the arccosine of dot product of the vector defined for the two atoms and the vector (0,0,1), normalized by their magnitudes.

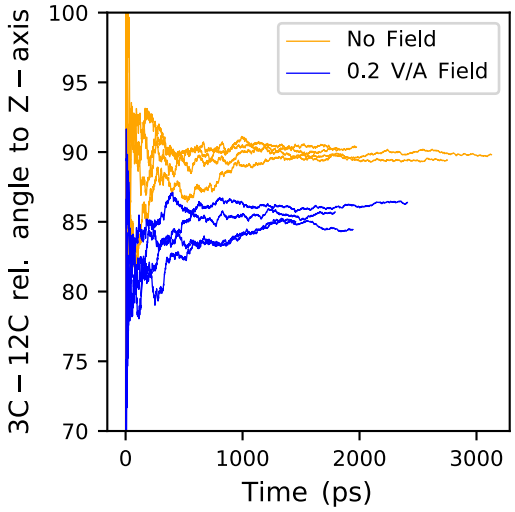


Figure A5.5: The average orientation of vector between atoms 3C and 12C (see Figure A5.1) and the z-axis is plotted for four simulations without an electric field (orange) and with an electric field of 0.2 V/A (blue lines). Note the orientation is periodic between 0 and 180 degrees. Orientation defined as the arccosine of dot product of the vector defined for the two atoms and the vector (0,0,1), normalized by their magnitudes.

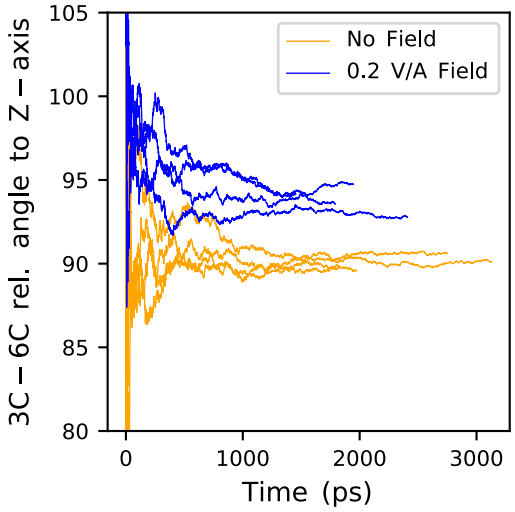


Figure A5.6: The average orientation of vector between atoms 3C and 6C (see Figure A5.1) and the z-axis is plotted for four simulations without an electric field (orange) and with an electric field of 0.2 V/A (blue lines). Note the orientation is periodic between 0 and 180 degrees. Orientation defined as the arccosine of dot product of the vector defined for the two atoms and the vector (0,0,1), normalized by their magnitudes.

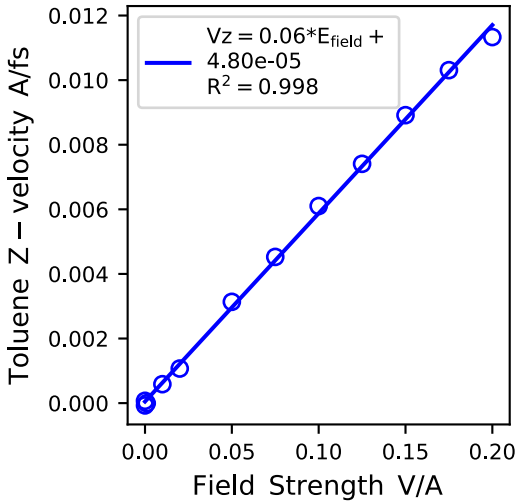


Figure A5.7: Average Z-velocity of toluene as a function of field strength. Note the linear trend.

Table A5.2: Infrequent Metadynamics results for set of oxidation conditions.

Field (V/A)	Temp. (K)	Mean Reaction Time (s)*	# Events	P-Value	# Rejects**
0.2	2000	3.88 (0.95) $*10^{-9}$	32	0.99	20
	1750	3.20 (0.77) $*10^{-8}$	32	0.48	49
	1500	2.61 (0.42)$*10^{-7}$	128	0.51	28
	1250	4.03 (1.1) $*10^{-6}$	32	0.2	100
	1000	1.47 (0.36) $*10^{-4}$	32	0.6	24
0.2***	2000	4.06 (0.95) $*10^{-9}$	32	0.99	3
	1750	2.65 (0.97) $*10^{-8}$	32	0.46	2
	1500	2.30 (0.59) $*10^{-7}$	32	0.99	22
	1250	1.51 (0.48)$*10^{-6}$	32	0.93	40
	1000	1.02 (0.28)$*10^{-4}$	32	0.65	40
0.175	2000	8.31 (2.40) $*10^{-9}$	32	0.74	25
	1750	4.17 (1.15) $*10^{-8}$	32	0.99	29
	1500	2.09 (0.53) $*10^{-7}$	32	0.9	48
	1250	5.25 (1.26) $*10^{-6}$	32	0.94	24
	1000	3.12 (0.97) $*10^{-4}$	32	0.91	40
0.15	2000	8.01 (2.60) $*10^{-9}$	32	0.51	72
	1750	5.83 (1.55)$*10^{-8}$	32	0.95	51
	1500	4.12 (0.68) $*10^{-7}$	128	0.46	32
	1250	5.61 (1.89)$*10^{-6}$	32	0.25	58
	1000	3.59 (0.82) $*10^{-4}$	32	0.49	45
0.125	2000	1.43 (0.46) $*10^{-8}$	32	0.87	59
	1750	6.38 (1.40) * 10^{-8}	32	0.73	51
	1500	5.52 (1.16) $*10^{-7}$	32	0.57	44
	1250	9.36 (2.64) $*10^{-6}$	32	0.56	24
	1000	6.53 (1.48) $*10^{-4}$	32	0.56	41
0.1	2000	2.22 (0.51) $*10^{-8}$	32	0.99	22
	1750	1.15 (0.21) $*10^{-7}$	32	0.61	63
	1500	8.36 (1.21) $*10^{-7}$	128	0.52	8
	1250	1.49 (0.30) $*10^{-5}$	64	0.75	14
	1000	9.39 (2.56) $*10^{-4}$	32	0.67	35
0.05	2000	2.70 (0.68)$*10^{-8}$	32	0.64	60
	1750	2.04 (0.50) $*10^{-7}$	32	0.98	22
	1500	2.12 (0.49) $*10^{-6}$	32	0.66	41
	1250	2.47 (0.55) $*10^{-5}$	32	0.53	54
	1000	1.56 (0.37) $*10^{-3}$	32	0.94	49
0.01	2000	2.71 (0.77) * 10^{-8}	32	0.59	35
	1750	1.85 (0.42) * 10^{-7}	32	0.97	15

	1500	1.45 (0.43) * 10^{-6}	32	0.93	0
	1250	2.15 (0.42) * 10^{-5}	32	0.68	66
	1000	1.81 (0.32) * 10^{-3}	32	0.63	76
0.001	2000	2.60 (0.68) * 10^{-8}	32	0.95	28
	1750	1.82 (0.40) *10^{-7}	32	0.88	12
	1500	1.53 (0.34) *10^{-6}	32	0.96	27
	1250	2.72 (0.72) *10^{-5}	32	0.17	143
	1000	2.20 (0.65) *10^{-3}	32	0.67	31
0.00001	2000	3.31 (0.77) *10^{-8}	32	0.49	66
	1750	2.01 (0.54)*10^{-7}	32	0.99	15
	1500	1.74 (0.46) *10^{-6}	32	0.46	49
	1250	2.67 (0.75) *10^{-5}	32	0.5	40
	1000	1.43 (0.30) * 10^{-3}	27	0.69	46
0	2000	4.79 (1.1) *10^{-8}	32	0.96	40
	1750	2.04 (0.44) *10^{-7}	32	0.75	25
	1500	1.59 (0.19)*10^{-6}	128	0.95	5
	1250	3.71 (0.98) *10^{-5}	32	0.14	171
	1000	1.68 (0.35) *10^{-3}	32	0.94	33

* Standard deviation from bootstrapping shown in parenthesis.
**Number of rejected samples that occurred to reach a 1000 subsamples in bootstrapping.
*** Thermostat turned off along z-axis to confirm electric field was not corrupting dynamics of system.

Table A5.3: Infrequent Metadynamics results for set of sampling pyrolysis reaction to form methyl radical.

Field (V/A)	Temp. (K)	Mean Reaction Time (s)*	# Events	P-Value	# Rejects **
0.2	2000	1.05 (0.23) * 10^{-5}	32	0.32	74
	1750	2.52 (0.44) * 10^{-4}	32	0.48	61
	1500	1.94 (0.38) * 10^{-2}	32	0.86	5
	1250	14.0 (2.6)	32	0.84	7
0	2000	1.09 (0.27) *10^{-5}	32	0.85	21
	1750	3.69 (0.77) *10^{-4}	32	0.9	33
	1500	2.80 (0.62) *10^{-2}	32	0.99	12
	1250	12.92 (2.88)	32	0.96	15

* Standard deviation from bootstrapping shown in parenthesis.
**Number of rejected samples that occurred to reach 1000 subsamples in bootstrapping.

Table A5.4: Infrequent Metadynamics results for set of sampling pyrolysis reaction to form hydrogen radical.

Field (V/A)	Temp. (K)	Mean Reaction Time (s)*	# Events	P-Value	# Rejects**
0.2	2000	2.29 (0.60) $*10^{-6}$	32	0.98	30
	1750	5.48 (1.56) $*10^{-5}$	32	0.99	20
	1500	2.50 (0.78) $*10^{-3}$	32	0.81	28
	1250	1.07 (0.27)	32	0.74	3
0	2000	1.88 (0.39) $*10^{-6}$	32	0.96	17
	1750	5.80 (1.55) $*10^{-5}$	32	0.75	25
	1500	2.71 (0.78) $*10^{-3}$	32	0.56	15
	1250	1.19 (0.34)	32	0.45	84

* Standard deviation from bootstrapping shown in parenthesis.
**Number of rejected samples that occurred to reach 1000 subsamples in bootstrapping.

Table A5.5: Widths of Gaussians for the different collective variables and temperatures sampled.

Reaction	Collective Variable (Bond Dist.)	Temp (K)	σ (nm)*
Toluene --> $C_6H_5 + CH_3$	3C-12C	2000	0.002
		1750	0.002
		1500	0.002
		1250	0.001
Toluene --> $C_6H_5CH_2 + H$	12C-13H	2000	0.0027
		1750	0.0027
		1500	0.0027
		1250	0.0012
Toluene + O_2 --> $C_6H_5CH_2 + HOO$	12C-13H	2000	0.0027
		1750	0.0027
		1500	0.0027
		1250	0.0012
		1000	0.0012

* Taken from one half the fluctuation of the CV.
**Same value applied regardless of electric field strength.

Discriminating Between Oxidation and C-H Bond Fracture

It was observed that in some of the higher temperature oxidation systems that the C-H bond biased would break forming a hydrogen radical and $C_6H_5CH_2$, rather than having the hydrogen abstracted by O_2. Because this merely represents a competing pathway that exists in the system, the correct rate of oxidation reaction can be recovered through Eq. ## in the main text. However, this depends on us having a clean description of whether or not a reaction occurs via oxidation or pyrolysis. As such, we tracked the coordination of the hydrogen that was removed to all of the oxygen atoms via the following equations

$$s_{ij} = \frac{1-(\frac{r_{ij}}{r_0})^6}{1-(\frac{r_{ij}}{r_0})^{12}} \qquad \text{Eq. A5.1a}$$

$$\sum_{i\in A}\sum_{j\in B} s_{ij} \qquad \text{Eq. A5.1b}$$

where here *i* is the hydrogen atom (13H) that is removed from toluene, *j* refers to all of the neighboring oxygen atoms, r_{ij} refers to the interatomic distance between atom pair *ij*, r_0 was set to be 0.13 nm and interactions beyond 0.30 nm were ignored. Note this is a slightly stretched beyond a typical H-O distance to account for the fact our simulations terminate at when a reaction occurs and so the peroxy radical may not have time to fully relaxed (i.e. bond is stretched at time C-H bond breaks). Because a value of r_{ij} 0.13 nm would yield a value of s_{ij} equal to 0.50, we treated anything system with a s_{ij} below 0.4 at the point of reaction to be a pyrolysis reaction. Such a treatment would ensure that oxygen was nowhere near the point of reaction, and matched our visualization of the trajectories.

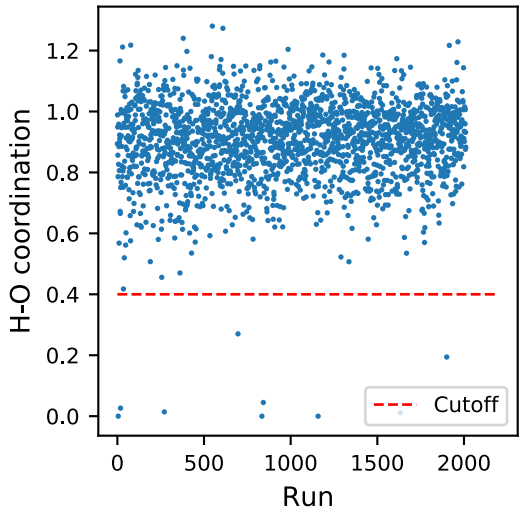


Figure A5.8: H-O Coordination for all oxidation simulations at the time the reaction occurs. Coordination values of 0.4 and above are treated as oxidation reactions and those below 0.4 are treated as occurring as C-H bond fractions without being attacked by O_2.

Verification of Parameters

Table A5.6: Kinetic results with from biased simulations with varied deposition pace.

Pace (steps)	τ (s)*	P-value	μ/σ**	Ln(2)*μ/τ_{median}***	# Samples	# Reject****
50000	7.39 (1.52) x10^4	0.713	1.24	0.83	28	72
10000	6.53 (1.40) x10^4	0.968	0.83	1.30	56	11

* Standard deviation from bootstrapping shown in parenthesis
** Ratio of the mean time (same as τ) and standard deviation of sample
*** Ratio of natural log of 2 multiplied by the ratio of the mean and median of the sample
**** Number of rejected subsamples out of 1000

In order to ensure that the bias parameters for the infrequent metadynamics simulations were robust and did not corrupt the dynamics by depositing bias too fast (i.e., in the transition region) we present kinetic results of a representative pyrolysis reaction using 1 kJ/mol, bias factor of 10, and Gaussian width of 0.002 nm (C-C bond of methyl group). Here we carried out the simulations at a pace of 10000 and 50000 steps to verify that the kinetics recovered are

independent of bias parameters, specifically the deposition pace.[50]The reaction times recovered were in strong agreement, well within the uncertainty from bootstrapping. Also, it is worth noting that the p-values are well above the 0.05 cutoff, and that the other ancillary statistics, namely μ/σ and $\mathrm{Ln}(2)*\mu/\tau_{median}$ approach unity.[2] The reject rates are also low as well.

Appendix 6 Methodological Notes

A6.1 Infrequent Metadynamics

Infrequent MetaD is a powerful method that allows for the kinetics of a transition to be assessed from biased simulations. While this method is suitable for a variety of systems from protein dynamics[2,35,50,124,134–136] to chemical reactions,[20,22,23,122] there are many steps to implementing it correctly and, consequentially, many potential pitfalls to avoid. Herein, I will provide a brief overview of key process details that I believe lead to the most *robust* results.

A6.1.1 CV Selection and Stop Criterion

As with all MetaD studies, the most important decision is determining which CVs should be biased. Because infrequent MetaD is typically applied in the context of assessing the kinetics between two states, we are typically blessed with some chemical intuition of the system or can rely on published studies of similar systems as a starting point.[45] Once a set of candidate CVs have been selected, it is prudent to run a steered MD simulation driving the system from the starting state to the final state. This process can be trivially done using the moving restraint in PLUMED.[105] In addition to verifying that the desired transition occurred and that the final state is indeed stable, the trajectories should also be analyzed to determine the specific CV values that should be used to indicate that a transition has occurred. The criteria for a transition should be selected so the system fully commits to the new state (i.e. no re-crossing event occurs), see Figure A6.1.

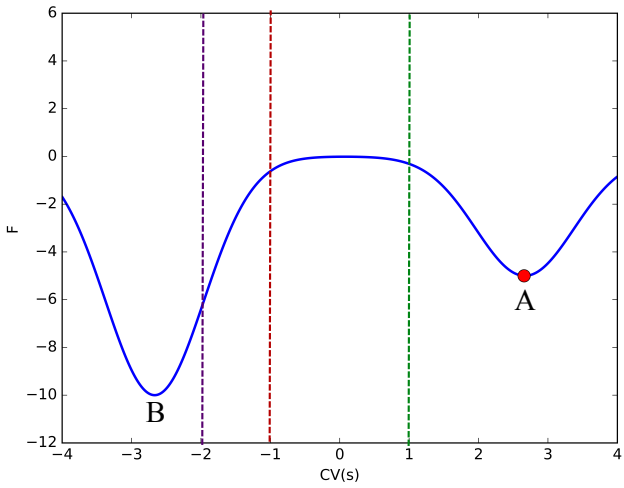


Figure A6.1: Analytical potential with two states A and B with the system starting in basin A, as shown by the red dot. The green, red, and purple lines indicate three potential CV values to indicate when a transition occurs.

For the model system shown in Figure A6.1, we see two clear states A and B, with a broad transition state region between CV values of -1 and 1. For evaluating the kinetics of a transition from state A to state B, the best choice for a "stop criterion" is a CV value of s<=-2 (purple line). The green dashed line at s<=1 is unsuitable as while this indicates that the system has escaped basin A, the system has not committed to basin B, and thus could easily return to basin A. A stop criterion of s<=-1 (red line) might be suitable, but as it is still near the transition state it is not clear. For this system, the region of s<=-2 can only be reached once the system has fully committed to basin. Because in many instances our understanding of the system of interest is not nearly as resolved as the one shown in Figure 6.1, it is best to choose a very conservative stop criterion to ensure a transition has occurred, and then evaluate the rates with less strict values and monitor the convergence of the transition times with stop criterion. It is important to note that there have been many recent developments in sophisticated, generic CVs such as SGOOP[67] which may help simplify this process in the future.

A6.1.2 Parameters, Replicas and Analysis

Because the transition events sampled are stochastic processes, it is necessary to sample an ensemble of events and evaluate the average transition time. In order to get a robust estimate of the average transition time, it is typically best to sample at least 10 transition events for a given set of conditions because anything less makes it difficult to perform error analysis with meaningful error bars.

One of the core assumptions of the infrequent MetaD method is that bias is not deposited in the transition region of the system.[2,19,66] If this assumption is violated, the recovered kinetics will be corrupt. Therefore, the overall bias deposition rate must be modulated to ensure this assumption holds. It is standard to run a series of trials with varying deposition paces and monitoring at what pace the recovered kinetics begins to converge.[20,50] An alternative approach would be to carry out unbiased simulations at high temperatures so the kinetics can be assessed at a reasonable cost, and compare these directly with the biased runs at the same temperature. If an agreement is found, then these bias parameters are typically suitable at lower temperatures, provided the system does not substantially change.

Because the ensemble of these events are expected to follow a Poisson distribution,[2] we apply the two-sample Kolmogorov-Smirnov (KS-test) test[36] to evaluate whether or not the bias corrupts the dynamics of the system. It is standard to determine that the dynamics are not corrupted if the recovered p-value exceeds a value of 0.05.[2] However, there are a few additional notes that should be included with this procedure. At least one of the systems being modeled should have 30 events, or more, collected to verify the bias parameters are suitable. The KS-test is known to lose power for sample sizes less than 30,[137] and there are documented cases where as more events are collected the p-value of the sample decreases.[136] When the distribution of events is analyzed through the KS-test, it is prudent to only apply it when ALL of the simulations that

have been launched have concluded. Prematurely applying this test while runs are still in progress will not only lead to over sampling faster events, but may also yield incorrect p-values (false positives or false negatives). Moreover, while the KS-test is the standard practice, there are other metrics for assessing the quality of results that should at the very least be monitored: the ratio of the average and standard deviation of the sample (referring to transition times) and the ratio of the sample average transition time multiplied by the ln(2) to the median of transition time of the sample. Each of these ratios should be close to unity. These alternative metrics are typically not relied on because they are very sensitive to noise in small samples, but they should at least be checked to ensure that the values are somewhat close to unity (i.e. between 0.5-2.0).

A6.2 Parallel Bias Metadynamics with Multiple Partitioned Families

Following up on our previous work, the PBMetaDPF method was applied to simple LJ systems with to explore how the method scales and if there are any potential pitfalls that might arise when multiple families are used. Note in the original study, all systems explored partitioned all CVs into a single family.[27] Herein, we explore how this method scales with the number of families, and if the distribution of members within the families impacts the overall convergence of the system. The same LJ7 system used in the study by Prakash et al. is used as a test case.[27] Interatomic distances were biased, this time the 21 CVs were partitioned into either three families or seven families. For systems with three families, we tested the following distributions of CVs: 7-7-7, 12-7-2, 19-1-1, and 10-10-1, and the overall convergence of this system compared to the convergence of PBMetaD is shown in Figure A6.2A. For systems with seven families, we tested the following distributions of CVs: 3-3-3-3-3-3-3, 8-8-1-1-1-1-1, 4-4-4-4-3-3-1, and 15-1-1-1-1-1-1 and the overall convergence of this system compared to the convergence of PBMetaD is shown in Figure A6.2B.

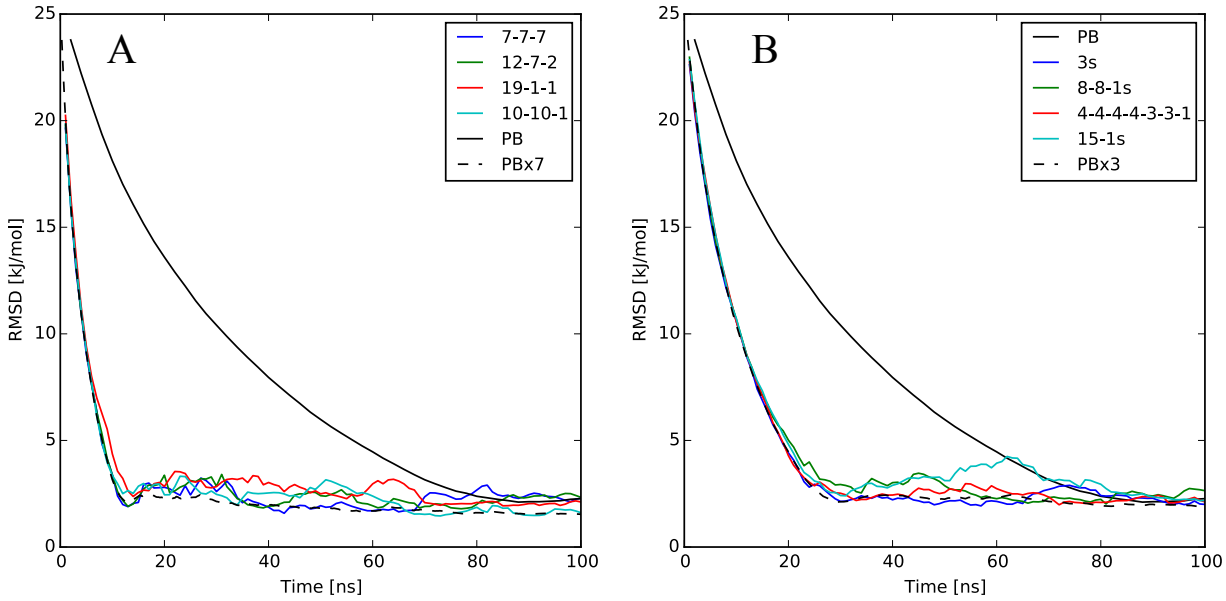


Figure A6.2: A) Convergence of LJ7 system for PBMetaD and PBMetaDPF using three families with various CV distributions. B) Convergence of LJ7 system for PBMetaD and PBMetaDPF using seven families with various CV distributions. The convergence in both plots is shown as RMSD relative to a WTMetaD simulation as was done in Prakash et al.[27] PBx3 and PBx7 indicates the convergence of the PBMetaD run, but accelerated by a factor of three and seven respectively to compare to the PBMetaDPF runs

From Figure A6.2, it is clear PBMetaDPF performs its best when the total number of families is minimized as using three families results in a convergence speed up by a factor of seven, while using seven families only expedites convergence by a factor of three. Note that we found in the original study that partitioning all of the CVs into one family led to a speed up by a factor of 21. It is also worth noting that the distribution of CVs across the families appears to have no impact on the convergence of the system, as the RMSD profiles are nearly indistinguishable in both plots.

There are however instances when the number of families cannot be reduced, such as when there are multiple types of particles in a system. Here we analyze how the PBMetaDPF method performs when we have different families that actually represent different interactions. To start, we studied the interactions of a system with three argon atoms (σ=0.34 nm, ε=20 kJ/mol) and one krypton atom (σ=0.39 nm, ε=10 kJ/mol). This represents a system with 6 interatomic distances, which were partitioned into two families, Kr-Ar interactions and Ar-Ar interactions. This system was biased with PBMetaD and PBMetaDPF with a pace of 500 steps, initial hill height of 2 kJ/mol, and a bias factor of 10 at a temperature of 300 K. Parallel tempering simulations of the 4-particle LJ system were set up with 32 replicas, and temperatures for the NVT simulation ranging from 300-8000 K. The temperature spacing (300, 302.98, 306.02, 309.10, 312.22, 315.40, 318.62, 321.89, 325.22, 328.59, 332.02, 335.50, 339.04, 342.63, 346.29, 350.00, 383.5, 422.08, 466.87, 519.25, 581.07, 654.77, 743.61, 852.09, 986.55, 1156.09, 1374.25, 1661.89, 2052.57, 2603.22, 3417.28, 8000.0 K) was evaluated using the procedure described in Appendix 4. From analyzing Figure A6.3, we see that both PBMetaD and PBMetaDPF yield the same free energy profiles for both types of interatomic distances, also matching the profiles recovered from PT. Specifically, Figure A6.3.B shows that reducing this six CV system to two families accelerates the convergence of each profile by a factor of three, also matching what we observed for the LJ7 system with different numbers of families.

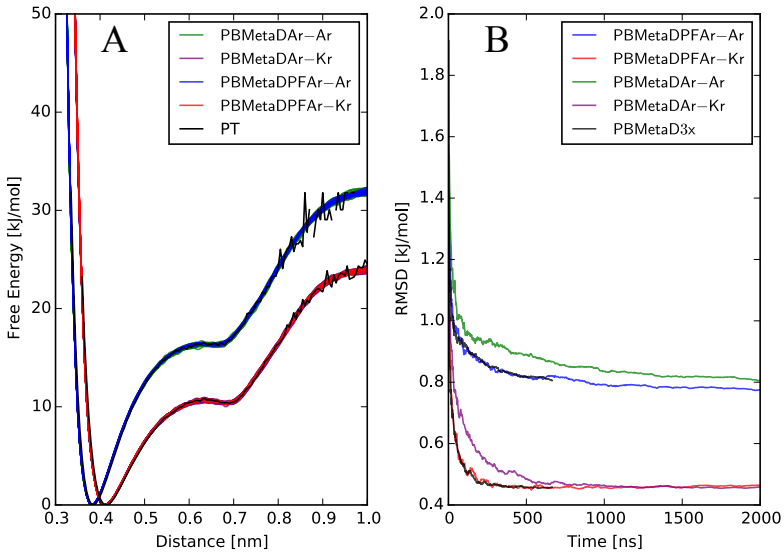


Figure A6.3: A) Free energy profiles recovered for Kr-Ar and Ar-Ar interactions from PBMetaD, PBMetaDPF, and PT. B) Convergence of the Kr-Ar and Ar-Ar profiles from PBMetaD and PBMetaDPF shown as RMSD vs time, where the RMSD is relative to the profile recovered from parallel tempering.

We further explored this by studying a system with two argon (Ar) atoms (σ=0.34 nm, ε=20 kJ/mol), one krypton (Kr) atom (σ=0.39 nm, ε=10 kJ/mol), and one neon (Ne) atom (σ=0.28 nm, ε=25 kJ/mol), using the same bias and PT parameters as before. As expected, we observe good agreement between the free energy profiles of the different atomic interactions across the different enhanced sampling methods as shown in Figure A6.4. Figure A6.4B shows that there is only a slight speed up in terms of convergence between PBMetaDPF and PBMetaD. This phenomenon is expected though because the six CV system is only slight reduced to a four families, leading to a speed up of ~1.5x. Therefore, this is a system where PBMetaDPF is applicable, but its impact is not exactly profound or appreciable.

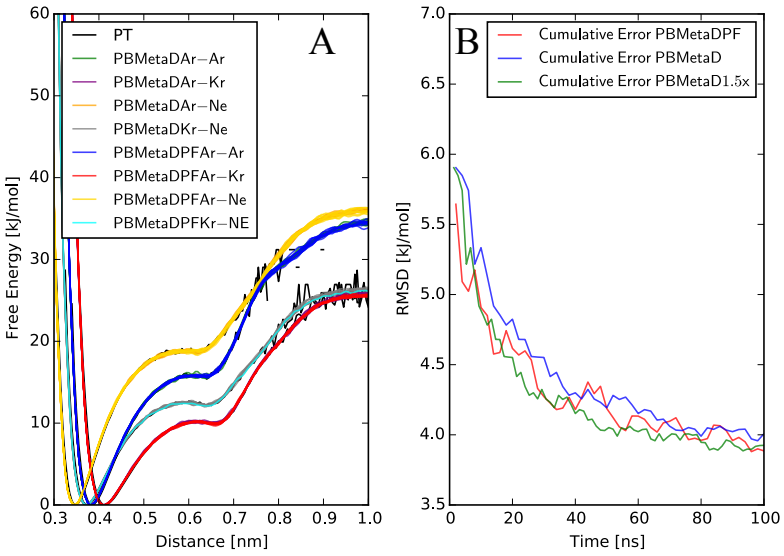


Figure A6.4: A) Free energy profiles recovered for Kr-Ar and Ar-Ar interactions from PBMetaD, PBMetaDPF, and PT. B) Convergence of the Kr-Ar and Ar-Ar profiles from PBMetaD and PBMetaDPF shown as RMSD vs time, where the RMSD is relative to the profile recovered from parallel tempering.

While here we demonstrate conclusively that PBMetaDPF is applicable to systems with multiple partitioned families representing different interactions, it is clear that it is not a panacea to high dimensional sampling. PBMetaDPF scales exclusively with the number of families, rather than the distribution of CVs within each family. Therefore, as with all MetaD-based methods, the CVs biased should be chosen carefully. Having great disparity in the number of CVs in each family should ideally be avoided as additional families reduce the potential gain in efficiency from PBMetaDPF.

References

(1) Gao, C. W.; Vandeputte, A. G.; Yee, N. W.; Green, W. H.; Bonomi, R. E.; Magoon, G. R.; Wong, H. W.; Oluwole, O. O.; Lewis, D. K.; Vandewiele, N. M.; et al. JP-10 Combustion Studied with Shock Tube Experiments and Modeled with Automatic Reaction Mechanism Generation. *Combust. Flame* **2015**, *162* (8), 3115–3129.

(2) Salvalaglio, M.; Tiwary, P.; Parrinello, M. Assessing the Reliability of the Dynamics Reconstructed from Metadynamics. *J. Chem. Theory Comput.* **2014**, *10* (4), 1420–1425.

(3) 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*, 1173–1213.

(4) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-Consistent-Charge Density-Functional Tight-Binding Method for Simulations of Complex Materials Properties. *Phys. Rev. B* **1998**, *58* (11), 7260–7268.

(5) Chenoweth, K.; van Duin, A. C. T.; Goddard, W. A. ReaxFF Reactive Force Field for Molecular Dynamics Simulations of Hydrocarbon Oxidation. *J. Phys. Chem. A* **2008**, *112* (5), 1040–1053.

(6) Ashraf, C.; van Duin, A. C. T. Extension of the ReaxFF Combustion Force Field toward Syngas Combustion and Initial Oxidation Kinetics. *J. Phys. Chem. A* **2017**, *121* (5), 1051–1068.

(7) Valsson, O.; Tiwary, P.; Parrinello, M. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu. Rev. Phys. Chem.* **2016**, *67*, 159–184.

(8) Valsson, O.; Parrinello, M. Variational Approach to Enhanced Sampling and Free Energy Calculations. *Phys. Rev. Lett.* **2014**, *113* (9), 090601–090605.

(9) Kästner, J. Umbrella Sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (6), 932–942.

(10) Glowacki, D. R.; Paci, E.; Shalashilin, D. V. Boxed Molecular Dynamics: A Simple and General Technique for Accelerating Rare Event Kinetics and Mapping Free Energy in Large Molecular Systems. *J. Phys. Chem. B* **2009**, *113* (52), 16603–16611.

(11) Dama, J. F.; Rotskoff, G.; Parrinello, M.; Voth, G. A. Transition-Tempered Metadynamics: Robust, Convergent Metadynamics via on-the-Fly Transition Barrier

Estimation. *J. Chem. Theory Comput.* **2014**, *10* (9), 3626–3633.

(12) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (5), 826–843.

(13) Ensing, B.; Vivio, M. De; Liu, Z.; Moore, P.; Klein, M. Metadynamics as a Tool for Exploring Free Energy Landscapes of Chemical Reactions. *Acc. Chem. Res.* **2006**, *39* (2), 73–81.

(14) Zheng, S.; Pfaendtner, J. A CPMD + Metadynamics Study of High Temperature Methanol Oxidation Reactions Using Generic Collective Variables. *J. Phys. Chem. C* **2014**, *118*, 10764–10770.

(15) Pietrucci, F.; Andreoni, W. Graph Theory Meets Ab Initio Molecular Dynamics: Atomic Structures and Transformations at the Nanoscale. *Phys. Rev. Lett.* **2011**, *107* (8), 085504.

(16) Cassone, G.; Sponer, J.; Sponer, J. E.; Pietrucci, F.; Saitta, A. M.; Saija, F. Synthesis of (d)-Erythrose from Glycolaldehyde Aqueous Solutions under Electric Field. *Chem. Commun.* **2018**, *54* (26), 3211–3214.

(17) Cassone, G.; Pietrucci, F.; Saija, F.; Guyot, F.; Saitta, A. M. One-Step Electric-Field Driven Methane and Formaldehyde Synthesis from Liquid Methanol. *Chem. Sci.* **2017**, *8* (3), 2329–2336.

(18) 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.

(19) Tiwary, P.; Parrinello, M. From Metadynamics to Dynamics. *Phys. Rev. Lett.* **2013**, *111* (23), 230602.

(20) Fleming, K. L.; Tiwary, P.; Pfaendtner, J. A New Approach for Investigating Reaction Dynamics and Rates with Ab Initio Calculations. *J. Phys. Chem. A* **2016**, *120* (2), 299–305.

(21) Laio, A.; Gervasio, F. L. Metadynamics: A Method to Simulate Rare Events and Reconstruct the Free Energy in Biophysics, Chemistry and Material Science. *Rep. Prog. Phys.* **2008**, *71* (12), 126601.

(22) Fu, C. D.; L. Oliveira, L. F.; Pfaendtner, J. Determining Energy Barriers and Selectivities of a Multi-Pathway System with Infrequent Metadynamics. *J. Chem. Phys.* **2017**, *146* (1), 014108.

(23) Fu, C. D.; Oliveira, L. F. L.; Pfaendtner, J. Assessing Generic Collective Variables for

Determining Reaction Rates in Metadynamics Simulations. *J. Chem. Theory Comput.* **2017**, *13* (3), 968–973.

(24) Fu, C.; He, Y.; Pfaendtner, J. Diagnosing the Impact of External Electric Fields on Toluene Oxidation and Pyrolysis. *J. Phys. Chem. A* **2019**, *under revi*.

(25) Pfaendtner, J.; Bonomi, M. Efficient Sampling of High-Dimensional Free-Energy Landscapes with Parallel Bias Metadynamics. *J. Chem. Theory Comput.* **2015**, *11* (11), 5062–5067.

(26) Fu, C. D.; Pfaendtner, J. Lifting the Curse of Dimensionality on Enhanced Sampling of Reaction Networks with Parallel Bias Metadynamics. *J. Chem. Theory Comput.* **2018**, *14*, 2516–2525.

(27) Prakash, A.; Fu, C. D.; Bonomi, M.; Pfaendtner, J. Biasing Smarter, Not Harder, by Partitioning Collective Variables into Families in Parallel Bias Metadynamics. *J. Chem. Theory Comput.* **2018**, *14* (10), 4985–4990.

(28) Allen, R. J.; Valeriani, C.; Rein Ten Wolde, P. Forward Flux Sampling for Rare Event Simulations. *J. Phys. Condens. Matter* **2009**, *21* (46), 463102.

(29) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53* (1), 291–318.

(30) E, W.; Ren, W.; Vanden-Eijnden, E. String Method for the Study of Rare Events. *Phys. Rev. B* **2002**, *66* (5), 052301.

(31) Yang, L.; Liu, C. E.; Shao, Q.; Zhang, J.; Gao, Y. Q. From Thermodynamics to Kinetics: Enhanced Sampling of Rare Events. *Acc. Chem. Res.* **2015**, *48* (4), 947–955.

(32) Voter, A. F. A Method for Accelerating the Molecular Dynamics Simulation of Infrequent Events. *J. Chem. Phys.* **1997**, *106* (11), 4665–4677.

(33) Zimmerman, P. M. Automated Discovery of Chemically Reasonable Elementary Reaction Steps. *J. Comput. Chem.* **2013**, *34* (16), 1385–1392.

(34) E, W.; Ren, W.; Vanden-Eijnden, E. Finite Temperature String Method for the Study of Rare Events. *J. Phys. Chem. B* **2005**, *109* (14), 6688–6693.

(35) Tiwary, P.; Limongelli, V.; Salvalaglio, M.; Parrinello, M. Kinetics of Protein–Ligand Unbinding: Predicting Pathways, Rates, and Rate-Limiting Steps. *Proc. Natl. Acad. Sci.* **2015**, *112* (5), E386–E391.

(36) Massey, F. J. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Am. Stat. Assoc.* **1951**, *46* (253), 68–78.

(37) Fu, C.; Pfaendtner, J. LimPy: Langevin Integrator Metadynamics in Python http://zenodo.org/record/62068 (accessed Sep 11, 2016).

(38) Bussi, G.; Parrinello, M. Accurate Sampling Using Langevin Dynamics. *Phys. Rev. E* **2007**, *75* (5), 056707.

(39) Davis, M.; Davis, R. *Fundamentals of Chemical Reaction Engineering*; Dover Publications Inc: Mineola, 2003.

(40) Abraham, M. J.; Murtola, T.; Schulz, R.; Pall, S.; Smith, J. C.; Hess, B.; Lindah, E. Gromacs: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.

(41) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.* **2014**, *185* (2), 604–613.

(42) Tiwary, P.; Parrinello, M. A Time-Independent Free Energy Estimator for Metadynamics. *J. Phys. Chem. B* **2015**, *119* (3), 736–742.

(43) Henkelman, G.; Uberuaga, B. P.; Jonsson, H. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *J. Chem. Phys.* **2000**, *113* (22), 9901–9904.

(44) Laio, A.; Rodriguez-fortea, A.; Gervasio, L.; Ceccarelli, M.; Parrinello, M. Assessing the Accuracy of Metadynamics. *J. Phys. Chem. B* **2005**, *109* (14), 6714–6721.

(45) Zheng, S.; Pfaendtner, J. Enhanced Sampling of Chemical and Biochemical Reactions with Metadynamics. *Mol. Simul.* **2014**, *41* (1–3), 55–72.

(46) Muller, K.; Brown, L. D. Location of Saddle Points and Minimum Energy Paths by a Constrained Simplex Optimization Procedure. *Theor. Chim. Acta* **1979**, *53* (1), 75–93.

(47) Zuckerman, D. M.; Woolf, T. B. Dynamic Reaction Paths and Rates through Importance-Sampled Stochastic Dynamics. *J. Chem. Phys.* **1999**, *111* (21), 9475–9484.

(48) Miron, R. A.; Fichthorn, K. A. Accelerated Molecular Dynamics with the Bond-Boost Method. *J. Chem. Phys.* **2003**, *119* (12), 6210–6216.

(49) Piccini, G.; McCarty, J.; Valsson, O.; Parrinello, M. Variational Flooding Study of a SN2 Reaction. *J. Phys. Chem. Lett.* **2017**, *8* (3), 580–583.

(50) Tung, H.-J.; Pfaendtner, J. Kinetics and Mechanism of Ionic-Liquid Induced Protein

Unfolding: Application to the Model Protein HP35. *Mol. Syst. Des. Eng.* **2016**, *1* (4), 382–390.

(51) Tiwary, P.; Berne, B. J. Kramers Turnover: From Energy Diffusion to Spatial Diffusion Using Metadynamics. *J. Chem. Phys.* **2016**, *144* (13), 20–22.

(52) Tiwary, P.; Mondal, J.; Morrone, J. A.; Berne, B. J. Role of Water and Steric Constraints in the Kinetics of Cavity-Ligand Unbinding. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (39), 12015–12019.

(53) Sprenger, K. G.; Pfaendtner, J. Using Molecular Simulation to Study Biocatalysis in Ionic Liquids. In *Methods in Enzymology*; Voth, G. A., Ed.; Academic Press: Cambridge, MA, 2016; Vol. 577, pp 419–441.

(54) Domene, C.; Barbini, P.; Furini, S. Bias-Exchange Metadynamics Simulations: An Efficient Strategy for the Analysis of Conduction and Selectivity in Ion Channels. *J. Chem. Theory Comput.* **2015**, *11* (4), 1896–1906.

(55) Shaffer, P.; Valsson, O.; Parrinello, M. Enhanced, Targeted Sampling of High-Dimensional Free-Energy Landscapes Using Variationally Enhanced Sampling, with an Application to Chignolin. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (5), 1519712113-.

(56) Bal, K. M.; Neyts, E. C. Merging Metadynamics into Hyperdynamics: Accelerated Molecular Simulations Reaching Time Scales from Microseconds to Seconds. *J. Chem. Theory Comput.* **2015**, *11* (10), 4545–4554.

(57) Tiwary, P.; Van De Walle, A. Accelerated Molecular Dynamics through Stochastic Iterations and Collective Variable Based Basin Identification. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2013**, *87* (9), 094304.

(58) Bal, K. M.; Neyts, E. C. Direct Observation of Realistic-Temperature Fuel Combustion Mechanisms in Atomistic Simulations. *Chem. Sci.* **2016**, *7*, 5280–5286.

(59) Fichthorn, K. a; Miron, R. a; Wang, Y.; Tiwary, Y. Accelerated Molecular Dynamics Simulation of Thin-Film Growth with the Bond-Boost Method. *J. Phys. Condens. Matter* **2009**, *21* (8), 084212.

(60) Perez, D.; Uberuaga, B. P.; Shim, Y.; Amar, J. G.; Voter, A. F. Chapter 4 Accelerated Molecular Dynamics Methods: Introduction and Recent Developments. *Annu. Rep. Comput. Chem.* **2009**, *5*, 79–98.

(61) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.;

Ferguson, D.; Seibel, G.; Kollman, P. AMBER , a Package of Computer Programs for Applying Molecular Mechanics , Normal Mode Analysis , Molecular Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Comput. Phys. Commun.* **1995**, *91* (1–3), 1–41.

(62) Wang, L. P.; McGibbon, R. T.; Pande, V. S.; Martinez, T. J. Automated Discovery and Refinement of Reactive Molecular Dynamics Pathways. *J. Chem. Theory Comput.* **2016**, *12* (2), 638–649.

(63) Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. Discovering Chemistry with an Ab Initio Nanoreactor. *Nat. Chem.* **2014**, *6* (December), 1044–1048.

(64) Do, M.; Kro, L. C.; Kopp, W. A.; Ismail, A. E.; Leonhard, K. Automated Discovery of Reaction Pathways, Rate Constants, and Transition States Using Reactive Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2015**, *11* (6), 2517–2524.

(65) Ufimtsev, I. S.; Martinez, T. J. Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients and First Principles Molecular Dynamics. *J. Chem. Theory Comput.* **2009**, *5* (1), 2619–2628.

(66) Voter, A. F. Hyperdynamics: Accelerated Molecular Dynamics of Infrequent Events. *Phys. Rev. Lett.* **1997**, *78* (20), 3908–3911.

(67) Tiwary, P.; Berne, B. J. Spectral Gap Optimization of Order Parameters for Sampling Complex Molecular Systems. **2015**, *113* (11), 2839–2844.

(68) Prakash, A.; Sprenger, K. G.; Pfaendtner, J. Essential Slow Degrees of Freedom in Protein-Surface Simulations: A Metadynamics Investigation. *Biochem. Biophys. Res. Commun.* **2017**.

(69) Löhr, T.; Jussupow, A.; Camilloni, C. Metadynamic Metainference: Convergence towards Force Field Independent Structural Ensembles of a Disordered Peptide. *J. Chem. Phys.* **2017**, *146* (16), 165102.

(70) Bonomi, M.; Camilloni, C.; Vendruscolo, M. Metadynamic Metainference: Enhanced Sampling of the Metainference Ensemble Using Metadynamics. *Sci. Rep.* **2016**, *6* (1), 31232.

(71) Ranzi, E.; Cavallotti, C.; Cuoci, A.; Frassoldati, A.; Pelucchi, M.; Faravelli, T. New Reaction Classes in the Kinetic Modeling of Low Temperature Oxidation of N-Alkanes. *Combust. Flame* **2015**, *162* (5), 1679–1691.

(72) Suleimanov, Y. V.; Green, W. H. Automated Discovery of Elementary Chemical Reaction Steps Using Freezing String and Berny Optimization Methods. *J. Chem. Theory Comput.* **2015**, *11* (9), 4248–4259.

(73) Jalan, A.; Alecu, I. M.; Aguilera-Iparraguirre, J.; Merchant, S. S.; Yang, K. R.; Merhcant, S. S.; Truhlar, D. G.; Green, W. H. New Pathways for Formation of Acids and Carbonyl Products in Low Temperature Oxidation. *J. Am. Chem. Soc.* **2013**, *135*, 11100–11114.

(74) M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ort, and D. J. F. Gaussian 09, Revision E.01. Gaussian, Inc: Wallingford CT 2016.

(75) Piana, S.; Laio, A. A Bias-Exchange Approach to Protein Folding. *J. Phys. Chem. B* **2007**, *111* (17), 4553–4559.

(76) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L. P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109* (8), 1528–1532.

(77) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3* (10), 33.

(78) Ellson, J.; Gansner, E. R.; Koutsofios, E.; North, S. C.; Woodhull, G. Graphviz and Dynagraph – Static and Dynamic Graph Drawing Tools.

(79) Iannuzzi, M.; Schiffmann, F.; Vandevondele, J. CP 2 K : Atomistic Simulations of Condensed Matter Systems. **2014**, *4* (February), 15–25.

(80) Jalan, A.; Allen, J. W.; Green, W. H. Chemically Activated Formation of Organic Acids in Reactions of the Criegee Intermediate with Aldehydes and Ketones. *Phys. Chem. Chem. Phys.* **2013**, *15* (39), 16841–16852.

(81) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (20), 12562–12566.

(82) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **2008**, *100* (2), 1–4.

(83) Comer, J.; Gumbart, J. C.; Hénin, J.; Lelièvre, T.; Pohorille, A.; Chipot, C. The Adaptive Biasing Force Method: Everything You Always Wanted To Know but Were Afraid To

Ask. *J. Phys. Chem. B* **2015**, *119* (3), 1129–1151.

(84) Gil-Ley, A.; Bussi, G. Enhanced Conformational Sampling Using Replica Exchange with Collective-Variable Tempering. *J. Chem. Theory Comput.* **2015**, *11* (3), 1077–1085.

(85) Noé, F.; Clementi, C. Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation. *J. Chem. Theory Comput.* **2015**, *11* (10), 5002–5011.

(86) Tribello, G. A.; Cuny, J.; Eshet, H.; Parrinello, M. Exploring the Free Energy Surfaces of Clusters Using Reconnaissance Metadynamics. *J. Chem. Phys.* **2011**, *135* (11), 114109.

(87) Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. Efficient Reconstruction of Complex Free Energy Landscapes by Multiple Walkers Metadynamics. *J. Phys. Chem. B* **2006**, *110* (8), 3533–3539.

(88) Hošek, P.; Toulcová, D.; Bortolato, A.; Spiwok, V. Altruistic Metadynamics: Multisystem Biased Simulation. *J. Phys. Chem. B* **2016**, *120* (9), 2209–2215.

(89) Šućur, Z.; Spiwok, V. Sampling Enhancement and Free Energy Prediction by the Flying Gaussian Method. *J. Chem. Theory Comput.* **2016**, *12* (9), 4644–4650.

(90) Domene, C.; Barbini, P.; Furini, S. Bias-Exchange Metadynamics Simulations: An Efficient Strategy for the Analysis of Conduction and Selectivity in Ion Channels. *J. Chem. Theory Comput.* **2015**, *11* (4), 1896–1906.

(91) Heller, G. T.; Aprile, F. A.; Bonomi, M.; Camilloni, C.; De Simone, A.; Vendruscolo, M. Sequence Specificity in the Entropy-Driven Binding of a Small Molecule and a Disordered Peptide. *J. Mol. Biol.* **2017**, *429* (18), 2772–2779.

(92) Bonomi, M.; Camilloni, C.; Vendruscolo, M. Metadynamic Metainference: Enhanced Sampling of the Metainference Ensemble Using Metadynamics. *Sci. Rep.* **2016**, *6* (1), 31232.

(93) Branduardi, D.; Bussi, G.; Parrinello, M. Metadynamics with Adaptive Gaussians. *J. Chem. Theory Comput.* **2012**, *8* (7), 2247–2254.

(94) Doye, J. P. K.; Miller, M. A.; Wales, D. J. Evolution of the Potential Energy Surface with Size for Lennard-Jones Clusters. *J. Chem. Phys.* **1999**, *111* (18), 8417–8428.

(95) Nava, M.; Palazzesi, F.; Perego, C.; Parrinello, M. Dimer Metadynamics. *J. Chem. Theory Comput.* **2017**, *13* (2), 425–430.

(96) Tribello, G. A.; Ceriotti, M.; Parrinello, M. A Self-Learning Algorithm for Biased Molecular Dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (41), 17509–17514.

(97) Wales, D. J. Discrete Path Sampling. *Mol. Phys.* **2002**, *100* (20), 3285–3305.

(98) Fennell, C. J.; Bizjak, A.; Vlachy, V.; Dill, K. A. Ion Pairing in Molecular Simulations of Aqueous Alkali Halide Solutions. *J. Phys. Chem. B* **2009**, *113* (19), 6782–6791.

(99) General, I. J. A Note on the Standard State's Binding Free Energy. *J. Chem. Theory Comput.* **2010**, *6* (8), 2520–2524.

(100) Laage, D.; Hynes, J. T. On the Residence Time for Water in a Solute Hydration Shell: Application to Aqueous Halide Solutions. *J. Phys. Chem. B* **2008**, *112* (26), 7697–7701.

(101) Ma, X.; Zhang, S.; Jiao, F.; Newcomb, C. J.; Zhang, Y.; Prakash, A.; Liao, Z.; Baer, M. D.; Mundy, C. J.; Pfaendtner, J. Tuning Crystallization Pathways through Sequence Engineering of Biomimetic Polymers. *Nat. Mater.* **2017**, *16*, 767–774.

(102) Spagnoli, D.; Banfield, J. F.; Parker, S. C. Free Energy Change of Aggregation of Nanoparticles. *J. Phys. Chem. C* **2008**, *112* (38), 14731–14736.

(103) Mo, Y.; Lu, Y.; Wei, G.; Derreumaux, P. Structural Diversity of the Soluble Trimers of the Human Amylin(20-29) Peptide Revealed by Molecular Dynamics Simulations. *J. Chem. Phys.* **2009**, *130* (12), 125101–212413.

(104) Buchanan, L. E.; Dunkelberger, E. B.; Tran, H. Q.; Cheng, P.-N.; Chiu, C.-C.; Cao, P.; Raleigh, D. P.; De Pablo, J. J.; Nowick, J. S.; Zanni, M. T.; et al. Mechanism of IAPP Amyloid Fibril Formation Involves an Intermediate with a Transient β-Sheet.

(105) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.* **2014**, *185* (2), 604–613.

(106) Tan, S.; Xia, T.; Shi, Y.; Pfaendtner, J.; Zhao, S.; He, Y. Enhancing the Oxidation of Toluene with External Electric Fields: A Reactive Molecular Dynamics Study. *Sci. Rep.* **2017**, *7* (1), 1–11.

(107) Chenoweth, K.; Duin, A. C. T. Van; Dasgupta, S.; Iii, W. A. G. Initiation Mechanisms and Kinetics of Pyrolysis and Combustion of JP-10 Hydrocarbon Jet Fuel Initiation Mechanisms and Kinetics of Pyrolysis and Combustion of JP-10 Hydrocarbon Jet Fuel. **2009**, 1740–1746.

(108) Ashraf, C.; Shabnam, S.; Jain, A.; Xuan, Y.; van Duin, A. C. T. Pyrolysis of Binary Fuel Mixtures at Supercritical Conditions: A ReaxFF Molecular Dynamics Study. *Fuel* **2019**, *235* (July 2018), 194–207.

(109) Jiang, X. Z.; Feng, M.; Zeng, W.; Luo, K. H. Study of Mechanisms for Electric Field

Effects on Ethanol Oxidation via Reactive Force Field Molecular Dynamics. *Proc. Combust. Inst.* **2018**, *000*, 1–11.

(110) Murgida, G. E.; Wisniacki, D. A.; Tamborenea, P. I.; Borondo, F. Control of Chemical Reactions Using External Electric Fields: The Case of the LiNC ⇄ LiCN Isomerization. *Chem. Phys. Lett.* **2010**, *496* (4–6), 356–361.

(111) Eriksson, K.; Brooks, B.; Glover, J. Reactions of Oxygen and Toluene in an Electrical Discharge Reactor. *J. Chem. Technol. Biotechnol.* **1991**, *50* (4), 483–491.

(112) Che, F.; Gray, J. T.; Ha, S.; Kruse, N.; Scott, S. L.; McEwen, J. S. Elucidating the Roles of Electric Fields in Catalysis: A Perspective. *ACS Catal.* **2018**, *8* (6), 5153–5174.

(113) Dagaut, P.; Pengloan, G.; Ristori, A. Oxidation, Ignition and Combustion of Toluene: Experimental and Detailed Chemical Kinetic Modeling. *Phys. Chem. Chem. Phys.* **2002**, *4* (10), 1846–1854.

(114) Baulch, D. L.; Bowman, C. T.; Cobos, C. J.; Cox, R. A.; Just, T.; Kerr, J. A.; Pilling, M. J.; Stocker, D.; Troe, J.; Tsang, W.; et al. Evaluated Kinetic Data for Combustion Modeling: Supplement II. *J. Phys. Chem. Ref. Data* **2005**, *34* (3), 757.

(115) Agrawalla, S.; van Duin, A. C. T. Development and Application of a ReaxFF Reactive Force Field for Hydrogen Combustion. *J. Phys. Chem. B* **2011**, *115* (6), 960–972.

(116) Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M.; et al. The ReaxFF Reactive Force-Field: Development, Applications and Future Directions. *npj Comput. Mater.* **2016**, *2* (September 2015).

(117) Martinez, L.; Andrade, R.; Birgin, E. G.; Martinez, J. M. Packmol: A Package for Building Initial Configurations for Molecular Dynamics Simulations. *J. Comput. Chem.* **2008**, *30* (13), 2157–2164.

(118) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690.

(119) Rappé, A. K.; Goddard, W. A. Charge Equilibration for Molecular Dynamics Simulations. *J. Phys. Chem.* **1991**, *95* (8), 3358–3363.

(120) Aktulga, H. M.; Fogarty, J. C.; Pandit, S. A.; Grama, A. Y. Parallel Reactive Molecular Dynamics: Numerical Methods and Algorithmic Techniques. *Parallel Comput.* **2012**, *38*

(4–5), 245–259.

(121) Nakano, A. Parallel Multilevel Preconditioned Conjugate-Gradient Approach to Variable-Charge Molecular Dynamics. *Comput. Phys. Commun.* **1997**, *104*, 59–69.

(122) Oliveira, L. F. L.; Fu, C. D.; Pfaendtner, J. Density Functional Tight-Binding and Infrequent Metadynamics Can Capture Entropic Effects in Intramolecular Hydrogen Transfer Reactions. *J. Chem. Phys.* **2018**, *148* (15), 154101.

(123) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; et al. PLUMED: A Portable Plugin for Free-Energy Calculations with Molecular Dynamics. *Comput. Phys. Commun.* **2009**, *180* (10), 1961–1972.

(124) Casasnovas, R.; Limongelli, V.; Tiwary, P.; Carloni, P.; Parrinello, M. Unbinding Kinetics of a P38 MAP Kinase Type II Inhibitor from Metadynamics Simulations. **2017**, 4780–4788.

(125) Peters, B. *Reaction Rate Theory and Rare Events*, 1st ed.; Elsevier, 2017.

(126) Abraham, M. J.; van der Spoel, D.; Lindahl, E.; Hess, B.; Team, and the G. development. GROMACS User Manual Version 5.1.2. **2016**.

(127) Roe, D. R.; Cheatham III, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Synamics Trajectory Data. *J Chem Theory Com* **2013**, *9* (7), 3084–3095.

(128) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: A Python Wrapper for the OpenBabel Cheminformatics Toolkit. *Chem. Cent. J.* **2008**, *2* (1), 5.

(129) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *Journal of Computational Chemistry*. 2005, pp 1701–1718.

(130) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126* (1), 014101.

(131) Prakash, M. K.; Barducci, A.; Parrinello, M. Replica Temperatures for Uniform Exchange and Efficient Roundtrip Times in Explicit Solvent Parallel Tempering Simulations. *J. Chem. Theory Comput.* **2011**, *7* (7), 2025–2027.

(132) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chemie Int. Ed.* **1999**, *38*

(1–2), 236–240.

(133) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Müller, K. R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6* (12), 2326–2331.

(134) Wang, Y.; Valsson, O.; Tiwary, P.; Parrinello, M.; Lindorff-Larsen, K. Frequency Adaptive Metadynamics for the Calculation of Rare-Event Kinetics. **2018**, No. February 2018, 1–15.

(135) Wang, Y.; Martins, J.; Lindorff-Larsen, K. Biomolecular Conformational Changes and Ligand Binding: From Kinetics to Thermodynamics. *Chem. Sci.* **2017**, *8*, 6466–6473.

(136) Wang, Y.; Papaleo, E.; Lindorff-Larsen, K. Mapping Transiently Formed and Sparsely Populated Conformations on a Complex Energy Landscape. *Elife* **2016**, *5* (AUGUST), 1–35.

(137) Yap, B. W. Power Comparisons of Shapiro-Wilk , Kolmogorov-Smirnov , Lilliefors and Anderson- Darling Tests. **2011**, No. November 2014.