The Pennsylvania State University

The Graduate School

Eberly College of Sciences

KINETIC MODELING OF THE ENSEMBLE OF PATHWAYS PRESENT IN CONFORMATIONAL DYNAMICS

A Thesis in

Chemistry

by

David K. Wolfe

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2019

The thesis of David K. Wolfe was reviewed and approved* by the following:

Edward P. O'Brien
Assistant Professor of Chemistry
Thesis Advisor

William G. Noid
Associate Professor of Chemistry

Lasse Jensen
Professor of Chemistry

Philip C. Bevilacqua
Distinguished Professor of Chemistry and Biochemistry and Molecular Biology
Head of Department of Chemistry

*Signatures are on file in the Graduate School

ABSTRACT

Chemical reactions in biology occur in high dimensional space and at physiological temperatures, thereby ensuring that there are multiple reaction pathways connecting an ensemble of reactant and an ensemble of product states. Current methods in the computational chemistry field attempt to describe such a reaction using a single reaction pathway at zero Kelvin, or as an ensemble of pathways restricted locally to be near the minimum energy path. Here, we develop a new computational method that allows us to describe multiple pathways by using a novel combination of Markov state modeling and chain-of-states methods. The method is verified first on an analytical surface, followed by application to several dipeptide molecules. We find that our method is able to accurately predict the kinetics of the system and we show that the consideration of multiple pathways more accurately describes the trend in overall rate at a series of temperatures and the overall observed flux.

TABLE OF CONTENTS

List of Figures

List of Abbreviations

A	Arrhenius Prefactor
β	inverse temperature
BACE	Bayesian Agglomerative Clustering Engine
COS	Chain-of-States
D	diffusion coefficient
D_o	diffusion coefficient within the reactant well
FTSM	Finite Temperature String Method
k_B	Boltzmann's constant
k_{obs}	calculated rate of reaction
k_{TST}	classical transition state theory rate
k_k	Kramer's rate
$\vec{\boldsymbol{K}}$	master equation rate matrix
$\vec{\boldsymbol{R}}$	The transpose of the master equation rate matrix
NEB	Nudged Elastic Band
PCCA	Perron Coupled Cluster Analysis
PCCA+	Perron Coupled Cluster Analysis Plus
QM	Quantum Mechanical
QM/MM	Quantum Mechanical/Molecular Mechanical
SOS	String-of-States
T	temperature
$\vec{\boldsymbol{T}}$	Markovian transition matrix
WHAM	Weighted Histogram Analysis Method
$\Delta E^{\ddagger}$	free energy change to reach the transition state
η	viscosity
γ	collision frequency
ζ	friction coefficent
κ	transmission factor
$\vec{\pi}$	stationary distribution
ω_o	angular frequency within well
$\omega_{\ddagger}$	angular frequency as the transition state

Chapter 1. INTRODUCTION

1.1 General Introduction and Motivation

In the field of computational chemistry, one often desires to simulate systems on timescales that are comparable to experimental timescales. However, even with rapidly increasing computational power reaching the millisecond or second timescale on par with experiment using standard molecular dynamics (MD) simulations is not feasible. These simulations numerically integrate Newton's equations of motion for every atom in the system, leading to large increases in required computer time as the system size increases. To solve this problem, enhanced sampling techniques have been developed to access longer timescales Umbrella Sampling (US), transition path sampling (TPS)[1–3], and chain of states (COS) methods.[4,5] TPS explores state space by stitching together short unbiased trajectories along a reaction coordinate. COS methods on the other hand connect distinct regions in phase space, or portions of the underlying free energy surface in the case of chemical reactions, using a series of replicas constrained along the selected reaction coordinate. The current single pathway COS methods find the minimum energy path (MEP) or minimum free energy path (MFEP) only, often leaving significant regions of phase space unexplored.

In the case that the reaction occurs at zero Kelvin, the MEP is assumed to provide a reasonable description of the reaction, but most chemical reactions occur at high temperatures where other pathways may be accessible and entropic contributions can play a significant role. Therefore, one would expect to gain a more accurate model of a chemical system by accounting for higher energy pathways in addition to the minimum energy pathway. To achieve a comprehensive model accounting for multiple pathways, we develop a method to construct a network model underlying a process using a combination of Markov state models[6–8] (MSM) and COS methods. This method is then applied to several simple systems and verified to successfully predict kinetic behavior.

The remainder of this thesis is laid out as follows. In the next section, the overall method, including basic descriptions of the methods used and workflow, is described. In Chapter 2, each method is discussed briefly from a theoretical standpoint. In Chapters 3 and 4, the method is applied first to a simple toy model, followed by the simple biomolecular test systems alanine, glycine, and proline dipeptide. The method is verified here from a purely classical perspective, but in the future the method will be extended to a quantum mechanical – molecular mechanical (QM/MM) approach as is discussed briefly at the end of Chapter 5.

1.2 Overview of the Method

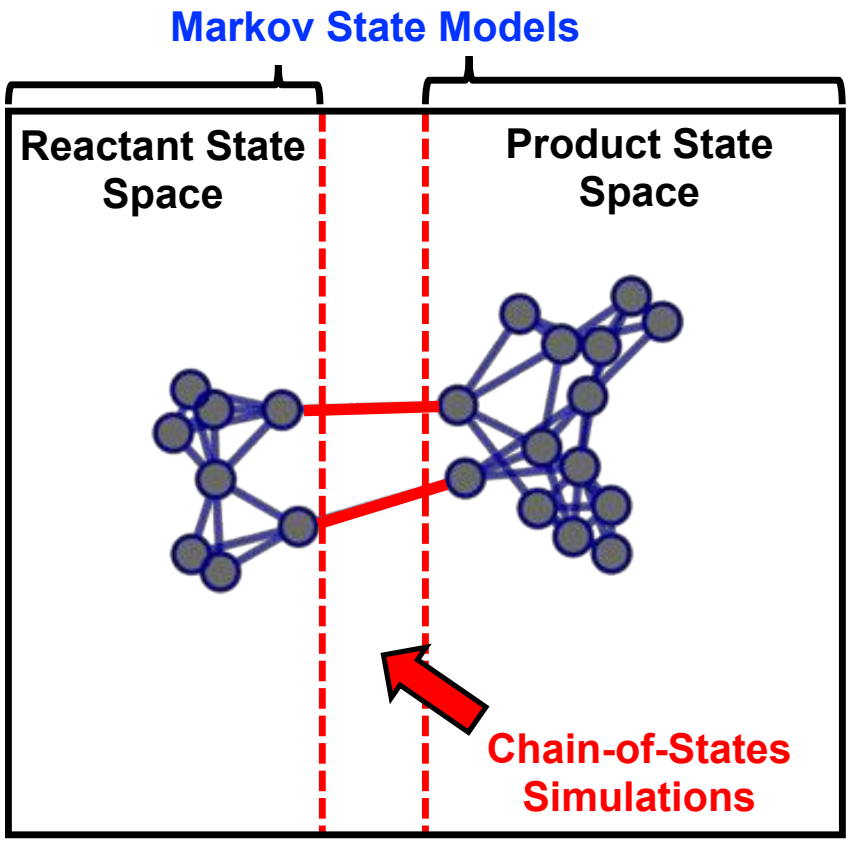

Figure 1. Decomposition of Markov state space. The entire discrete Markov state-space is decomposed into long-lived regions where classical MD will be run and MSMs constructed, and the transition path region, where COS methods will be applied. k_{obs} represents the overall transition rate through all pathways connecting reactant and product Markov state space.

While the application of the method within this paper focuses on conformational sampling with no actual chemical step, future applications will focus on the simulation of chemical reactions. Therefore, throughout this section on the development of the proposed workflow for the method we assume the perspective that there is a chemical reaction step. Simulation of chemical reactions requires expensive quantum mechanics (QM) simulations in order to account for bond breakage, bond formation, and electron motion. However, the computational expense of these simulations prevents complete exploration of the underlying free energy landscape.[9,10] Instead, we will perform conformational sampling using classical all-atom MD simulations within the product and reactant ensembles separately. From these simulations we will identify metastable regions of reactant state space and then we will use COS methods to drive a reaction between that state and another metastable region in product state space. This compartmentalization of the discrete state space is depicted for a general reaction network in Figure 1. The kinetic information obtained from both types of simulations will be gathered in a comprehensive master equation (ME) model in which the underlying rate constants come from both MM and QM/MM simulations. The master equation framework will then be used to calculate rates, pathway probabilities, fluxes, and to investigate the underlying conformational changes driving the reaction from reactant state space to product state space.

1.2.1 Markov model construction

The output of a MD simulation is a timeseries of atomic positions that can be projected onto sets of lower dimensional order parameters including distances, bond angles, dihedral angles and the RMSD between internal structures. MSMs have become popular in computational chemistry, particularly in biomolecular applications,[11–15] as a way of projecting simulation data onto a set of discrete states from which various kinetic and thermodynamic properties can be predicted. One of the advantages of MSMs is that one can identify metastable states without having a picture of the underlying free energy landscape to base assumptions on. This is a significant advantage in

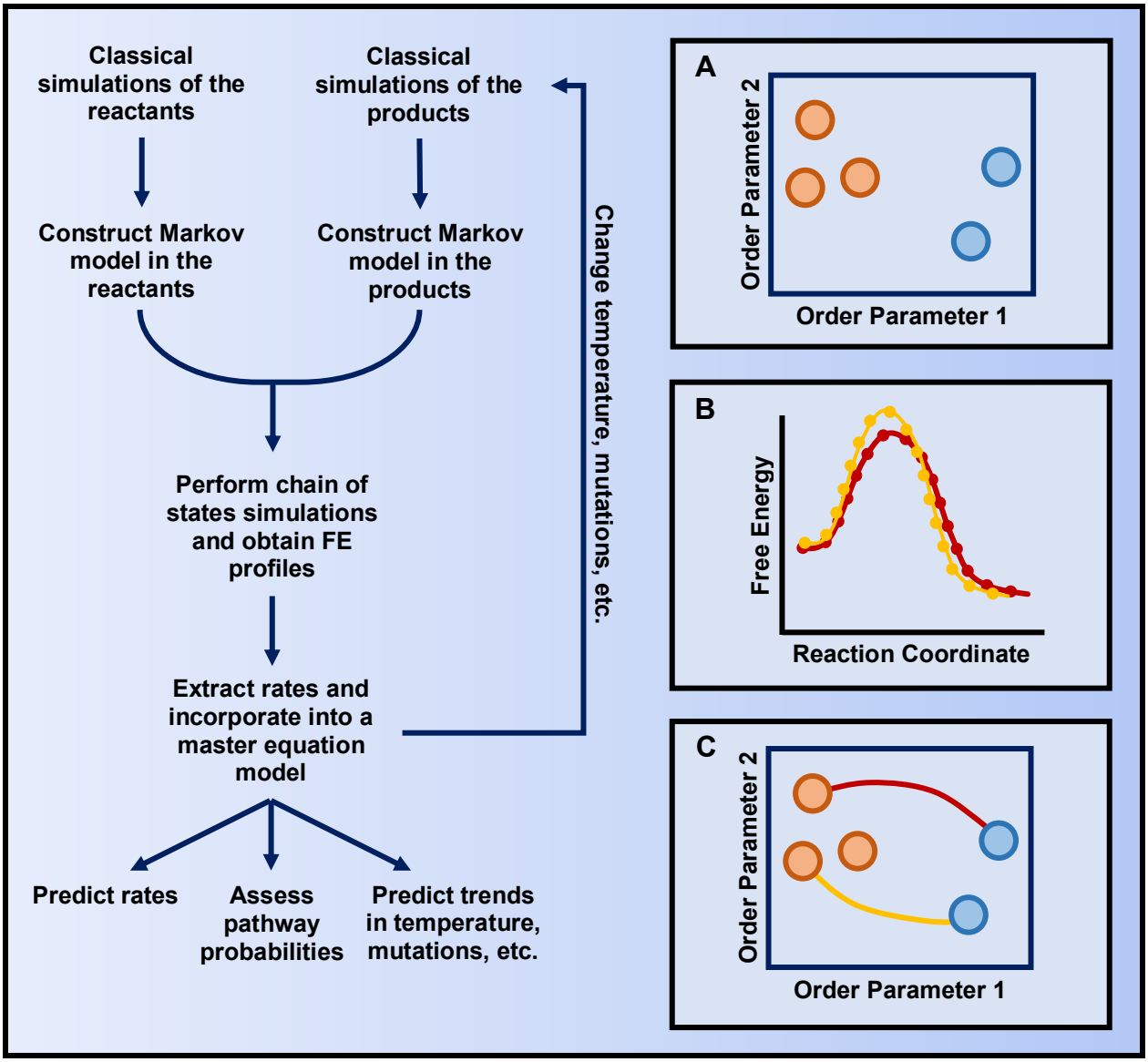

Figure 2. Outline of the workflow. The method consists of three different stages, (a) construction of MSMs in the reactants and products, (b) obtaining free energies along the reaction coordinate from COS, and (c) incorporation into a combined master equation model.

applications to large systems where it is more difficult to obtain complete sampling, at least using current methods. In order to build an MSM within the reactant and product ensembles, we first project the timeseries onto a set of coarse order parameters. In practice these order parameters are often chosen based on some underlying knowledge of the chemical system, but optimal order parameters may also be identified using principle component analysis (PCA) or time independent component analysis (tICA) developed specifically for MSMs.[16] More recently, machine learning and other optimization techniques have been used to discover optimal order parameters and reaction coordinates.[17,18] Metastable states are identified from the timeseries of order parameters using a series of clustering steps, and rates of intra-ensemble conversion are extracted. The Markov model construction is depicted in Figure 2A.

1.2.2 Quantum Mechanical-Molecular Mechanical Chain-of-States simulations

Following the identification of metastable conformational states, we will incorporate multiple pathways into our model by driving a reaction between each pair of reactant and product states using COS methods. Expanding on the information provided in section 1.1, COS methods perform a constrained minimization along a chain of replicas of the system stretching along a reaction coordinate. This chain of states provides a representation of the system as it progresses along a reaction coordinate, although it is not always clear how to select such a coordinate without extensive unbiased simulations. The free energy along the reaction coordinate, Figure 2B, will be obtained inherently within the COS method, in some cases, or by using umbrella sampling along the coordinate post-minimization.[19–21] As previously stated, simulation of chemical reactions requires expensive QM calculations and in large biomolecular systems this can become infeasible computationally. So, we will apply a multiscale quantum mechanical molecular mechanical approach when considering reactions in complex, biomolecular systems. QM/MM treats the region near the chemical reaction with QM simulations, and the surrounding region with classical MM simulations, as this portion has no direct contribution to the chemical step.[22,23] The end result of this step is a one-dimensional free energy profile describing the reaction between each pair of Markov states in the network.

1.2.3 Construction of master equation models

The pathway specific rate constants will then be extracted from the free energy profile for each pathway in the network. Classical transition state theory (TST),[24,25] position dependent diffusion

models[26–28] along the reaction coordinate, Kramers[29–31] theory, and Grote-Hynes theory[32,33] among others have been used in past studies to estimate transition rates. In order to avoid increased computational expense related to diffusion maps and other sampling based techniques, we will use a reaction rate theory approach, TST or Kramers, to extract a rate from the free energy profile. These theories have been shown to be accurate given accordance to a set of assumptions discussed later in Chapter 2, but in general one can expect to get within an order of magnitude of the true rate given a good estimate of the transition state free energy. The rate constants extracted from the free energy profiles describing reactant to product transitions are then incorporated into a master equation (ME) model that already includes the intra-ensemble transition rates extracted from the classical simulations. This marriage of classical and QM/MM simulations, Figure 2C, provides a more complete of description of the complex landscape of chemical reactions than one would achieve from a traditional, single pathway method.

1.2.4 Prediction of rates and the study of protein motion

Our ultimate goal is to develop a method that more accurately describes the full ensemble of reaction paths for a chemical reaction and then apply that method in a useful way. The prediction of the absolute rate of a reaction allows a direct comparison to experiment. However, in addition to being difficult to predict with a high degree of accuracy, good agreement can sometimes be obtained through a cancellation of errors inherent in the many methods applied.[34] So we instead will construct a

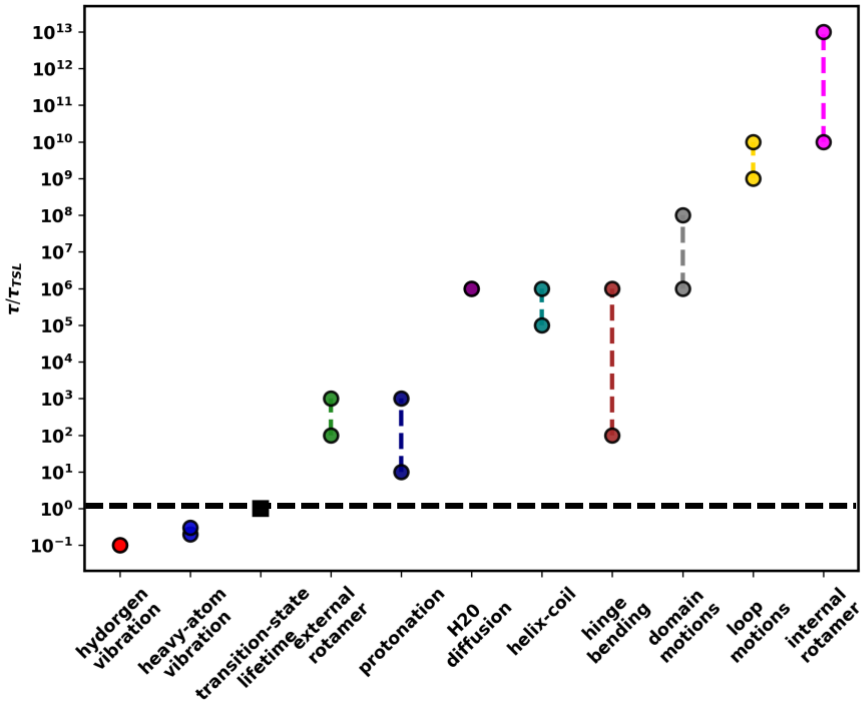

Figure 3. Timescales of protein motion. The timescales of protein motions relative to the transition path time (~100 fs), the dashed horizontal line.

complete master equation model at a series of temperatures in order to predict the Arrhenius behavior, or by introducing a series of mutations comparable with experimental studies in enzymes.[35,36] In the case that we are unable to calculate an absolute rate then our method can still be a valuable contribution to the field if we can reproduce a ratio of rates across a temperature range, or the rank ordering of kinetic changes induced by mutations.

There is a debate in recent literature over the influence of dynamics on chemical reactions in enzymes with reasonable arguments both in support and opposition.[37–42] Here we refer to enzyme dynamics as longer timescale conformational motions on the order of ns-μs, Figure 3, representative of structural rearrangements rather than the fast vibrational motions on the timescale of the reaction usually estimated to be ~100 fs. Given our proposed workflow, we expect that observing conformational changes on the order of several hundred nanoseconds will be observable within the reactant and product ensembles. Mapping a mutation or temperature induced change in kinetics to a specific conformational transition is a far-reaching application of our proposed method.

Chapter 2: THEORETICAL BACKGROUND

2.1 Simulation Techniques

2.1.1 Molecular Dynamics

MD simulations offer an atomistic description of molecular systems on timescales that are generally not discernable by traditional experimental techniques. Newton's equations of motion are integrated numerically on a per atom basis using the force acting on each atom as determined by the underlying MD force field. In this work, the CHARMM[43] force field will be used exclusively. Significant interactions terms are included in the potential energy function, Eq. (2.1.1), and the forces are calculated from the first derivative of this function.

$$V = \begin{matrix} V_{bonds} + V_{angles} + V_{dihedrals} + V_{Urey-Bradley} + V_{impropers} \\ +V_{non-bonded} + V_{electrostatics} \end{matrix} \quad (2.1.1)$$

The result of a MD simulation is a timeseries of atomic coordinates from which structural properties can be calculated. The time step in an all-atom MD simulation is generally at a maximum of 1-2 fs, requiring billions to trillions of integration steps to reach the timescales of certain motions relevant for biological functions. As a result, MD simulations are generally only able to observe events on the ns-μs timescale, though recent developments in computing have enabled the folding of small proteins of the millisecond timescale to be observed.[44,45]

2.1.2 Langevin Dynamics

Langevin dynamics are a variant of MD simulations that includes a stochastic term during integration. The stochastic term in the Langevin equation allows one to model the effect of solvent buffeting on the system without explicitly including the molecules, useful for simulation of systems in vacuum. This also allows Langevin dynamics to be used as a thermal bath that restricts the system to a temperature. The Langevin equation is given in Eq. (2.1.2), where $F_i(x_i)$ is the first derivative of the MD potential, γ is the collision frequency related to the friction as $\frac{\zeta}{m}$, and $\mathcal{N}(0,1)$ is a Gaussian random variable with mean zero and variance one.

$$m_i\left(\frac{d^2x_i}{dt^2}\right) = -F_i(x_i) + \gamma\Delta t\left(\frac{dx_i}{dt}\right) + \sqrt{2\gamma\Delta t k_B T}\mathcal{N}(0,1) \quad (2.1.2)$$

An additional benefit of including the stochastic term explicitly in the integration scheme is that it allows thermal activation in some friction regimes, though the main results of this thesis come

from simulations near the spatial diffusion regime ($\gamma \gg 1$). We also note that at equilibrium there will be no net effect of the stochastic force as a result of thermal averaging.

2.1.3 Brownian Dynamics

Brownian dynamics is the term given to Langevin dynamics in the overdamped, or high friction, limit. In the high friction limit, the inertial term of the Langevin equation becomes insignificant and does not contribute significantly to dynamics. The resulting equation describing overdamped Langevin, or purely diffusive, motion in one dimension is given in Eq. (2.1.3).[46,47]

$$x_{i+1} = x_i - \frac{1}{\gamma}\Delta t F_i(x_i) + \sqrt{2\gamma\Delta t k_B T}\mathcal{N}(0,1) \tag{2.1.3}$$

The lack of an inertial term allows one to describe the system in the Markovian limit, implying that the future time evolution of the system is decorrelated from its previous progress. A Brownian integrator is not explicitly included in molecular dynamics software, but an approximation of overdamped dynamics can be achieved simply by setting the collision frequency to a large value (water viscosity is ~50-60 ps^{-1}). This will essentially remove the inertial term from the Langevin equation, which is reasonable for approximating the diffusive motion of a biomolecule on an underlying free energy surface.

2.2 Markov Processes and Master Equations

2.2.1 Markov Chains

The most basic definition of a Markov process is that the future progress depends only on the present state of the system and not on the previous progress through state space. This is commonly termed the memoryless property of Markov processes and can be better described as the conditional probability of progressing forward losing any 'memory' of the progress of the past evolution of the system. Given that s_i is the current state of the system in state space $S = \{s_0, s_1, \dots, s_N\}$ at time t_i and $P(s_i, t_i)$ is the probability of occupying the discrete state s_i at time t_i,

$$P(s_i, t_i | s_{i-1}, t_{i-1}; s_{i-2}, t_{i-2}; \dots; s_0, t_0) = P(s_i, t_i | s_{i-1}, t_{i-1}). \tag{2.2.1}$$

Assuming that this property is followed throughout the entire process, the current state of the system at time t_n can be obtained from the product of the previous steps,

$$P(s_n, t_n | s_{n-1}, t_{n-1}; \dots; s_0, t_0) = \prod_{i=1}^{n} P(s_i, t_i) P(s_{i-1}, t_{i-1}) \tag{2.2.2}$$

The full representation of the Markov chain in state space, S, is often given in matrix form, $\boldsymbol{T}_{NxN}$, called the transition matrix of the Markov chain. Each matrix element t_{ij} is the elementary probability of transitioning to state s_j from the state s_i given that a transition occurred in time Δt.

$$\boldsymbol{T} = \begin{bmatrix} t_{00} & t_{01} & \cdots & t_{0N} \\ t_{10} & t_{11} & \cdots & t_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ t_{N0} & t_{N1} & \cdots & t_{NN} \end{bmatrix}. \tag{2.2.3}$$

Therefore, the time evolution of the discrete-time Markov model is governed by Eq. (2.2.4), where $\vec{p}(t)$ is the vector of state populations at time t,

$$\vec{p}^{\,T}(t + \Delta t) = \vec{p}^{\,T}(t)\vec{\boldsymbol{T}}. \tag{2.2.4}$$

2.2.2 Master Equations

A master equation approach can be thought of a continuous time version of a Markov chain, but here the evolution of the system is governed by transition rates as opposed to elementary transition probabilities given a set jump time, Δt.[48,49] Assuming the Markovian condition is satisfied, meaning that the rates are independent of previous conditions in this case, the population of state s_i evolves as,

$$\frac{dp_i}{dt} = -p_i \sum_{j \neq i} \omega_{ij} + \sum_{j \neq i} \omega_{ji} p_j \,, \tag{2.2.5}$$

where ω_{ij} is the rate of transitioning from s_i to s_j per unit time, and p_i is the probability of being in s_i at time t. Eq. (2.2.5) can also be extended to a matrix form describing the entire state space represented by the master equation. The matrix of interest is the master equation rate matrix,

$$\boldsymbol{K} = \begin{bmatrix} -\sum_{j \neq 0} \omega_{0j} & \omega_{10} & \cdots & \omega_{N0} \\ \omega_{01} & -\sum_{j \neq 1} \omega_{1j} & \cdots & \omega_{N1} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{0N} & \omega_{1N} & \cdots & -\sum_{j \neq N} \omega_{Nj} \end{bmatrix}. \tag{2.2.6}$$

The evolution of the system through the full Markov-state space is governed by,

$$\frac{d\vec{p}}{dt} = \boldsymbol{K}\vec{p}. \tag{2.2.7}$$

The solution of Eq. (2.2.7) given the initial condition $\vec{p}(0)$ is,

$$\vec{p}(t) = e^{\boldsymbol{K}t}\vec{p}(0). \tag{2.2.8}$$

For convenience later on in our workflow we also define the network rate matrix,

$$\boldsymbol{R} = \boldsymbol{K}^T = \begin{bmatrix} -\sum_{j\neq 0} \omega_{0j} & \omega_{01} & \dots & \omega_{0N} \\ \omega_{10} & -\sum_{j\neq 1} \omega_{1j} & \dots & \omega_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{N0} & \omega_{N1} & \dots & -\sum_{j\neq N} \omega_{Nj} \end{bmatrix}. \quad (2.2.9)$$

As can be seen here and in section 2.2.1 both the master equation and Markov chain approach describe the evolution of a system through a discrete, underlying state-space, $S = \{s_0, s_1, \dots, s_N\}$. The difference is in the formulation as a discrete-time vs. continuous time process. Both representations are given because, while this work will use primarily the master Equation formulation, much of the theory underlying the construction of MSMs relies on the discrete-time version. Throughout the rest of this chapter we will use the master Equation representation, but there is also an analogous Markov Chain representation.

2.2.3 Stationary Solution of the Master Equation

At steady-state, the forward and backward fluxes associated with s_i are equal and the net flux of the system is universally zero. Convergence of a system to equilibrium is required in order to calculate thermodynamic properties from statistical mechanics. In almost all systems, true steady-state conditions will never be obtained due to finite sampling issues. However, this issue is relieved somewhat by coarse-graining the system so that one only needs to know information about transitions between the major metastable states. In many applications it is common to enforce equilibrium statistics using a detail balance constraint, which assumes the observation of an $j \rightarrow i$ transition everytime an $i \rightarrow j$ transition is observed. Using the master equation rate matrix one can calculate occupation probabilities at steady-state, and from these the relative free energies of the system.

The vector of steady-state populations is denoted $\vec{\pi}$ to distinguish it from $\vec{p}$, the time-dependent vector of state probabilities. $\vec{\pi}$ is obtained by setting Eq. (2.2.7) equal to zero and solving the resulting eigenvalue problem, assuming that there is no net flux at the steady-state. In practice Eq. (2.2.10) is solved by diagonalizing $\boldsymbol{K}$ to obtain the underlying eigenspectrum. $\vec{\pi}$ is the normalized eigenvector corresponding to the zero eigenvalue of the solution of,

$$\frac{d\vec{\pi}}{dt} = \boldsymbol{K}\vec{\pi} = 0. \quad (2.2.10)$$

Due to the discretization of state-space, and to a small degree numerical errors, there will almost never be an eigenvalue exactly equal to zero. However, given that the eigenvalues are all less than zero, one simply chooses the largest eigenvalue. In all cases in this work the eigenvalue of interest was found to be a very small, negative number. Also implicit in the accuracy of the stationary probability calculated from a discrete model is that the transitions between s_i's are Markovian. Deviations from the true probabilities will occur for non-Markovian systems.

The free energy of a state can be estimated as,

$$A_i = -\beta^{-1} \log\left(\frac{\pi_i}{\max(\vec{\pi})}\right), \tag{2.2.11}$$

where $\beta = \frac{1}{k_B T}$ is the temperature. Note that this free energy is only a free energy difference from the lowest energy state of the model and could perhaps be better represented as ΔA_i.

2.2.4 First Passage Times

The concept of first passage times exists in many fields that use a discrete or network representation of state-space. In computational chemistry, the first passage time is often used to compare computational results to experiment by way of rates. In these cases, the extraction of a first passage time requires the definition of states using a set of selected order parameters and knowledge of the system being studied. Continuing from the previous definitions, a more general definition of the first passage time is the first time at which the system enters a state $s_i \in \Omega$ given any prior sequence of states $s_i \notin \Omega$, where Ω is a finite region of state space. More specifically, one defines a boundary, $\partial\Omega$, encompassing the entire region and then considers entry into Ω to be the time at which $\partial\Omega$ is first crossed. Assuming that the transition obeys single-exponential kinetics, the rate is the inverse of the mean first passage time, which is always true for two-state kinetics.[30,50]

To calculate the mean first passage time, one needs to sum over the mean time of all possible sequences connecting two states, or sets of states in our case,

$$t_{i_0} = \sum_{i_0 \ldots i_N} (\tau_{i_0 i_1} .. \tau_{i_{N-1} i_N}) p_{i_0 i_1} .. p_{i_{N-1} i_N} \tag{2.2.12}$$

where t_{i_0} is the first passage time given the process was in state i at $t = 0$, $\tau_{i_j i_k}$ is the mean time for a $j \rightarrow k$ transition, and $p_{i_j i_k}$ is the interstate transition probability. For a general distribution, the mean transition time is simply,

$$\tau_{ij} = \int_0^\infty t f_{ij}(t) dt\,. \tag{2.2.13}$$

For a Markovian process the distribution $f_{ij}(t)$ is modeled as,

$$f_{ij} = \frac{1}{\tau_i} e^{-\frac{t}{\tau_i}} = \kappa_i e^{-\kappa_i t}, \tag{2.2.14}$$

where the notation τ_i references the dwell time in state i. As pointed out in other studies, technically one can still calculate a first passage time regardless of the functional form of the distribution as long as their means are the same.[51,52] We modify the first passage time equation to represent an absorbing Markov chain in the product states and convert to matrix form to arrive at,

$$< \vec{t} > = \sum_{k=1}^{N} \boldsymbol{T}^k \vec{\alpha}^{-1}, \tag{2.2.15}$$

where $\vec{\alpha} = \left(\sum_{i \neq j} \omega_{ij}\right)^{-1}$. Noting that Eq. (2.2.15) is a geometric power series that converges as $k \to \infty$, since there is net probability lost to the absorbing states the equation for the first passage time is,[50,51,53,54]

$$< \vec{t} > = \ (\boldsymbol{I} - \ \boldsymbol{T})^{-1} \vec{\alpha} = \ -\boldsymbol{R}^{-1} \vec{1}. \tag{2.2.16}$$

The overall rate between the reactants and products is calculated from the vector of first passage times by weighting each state dependent first passage time by the probability to occupy that state at equilibrium,

$$k_{obs} = \ (\vec{\pi}^T < \vec{t} >)^{-1}. \tag{2.2.17}$$

2.3 Construction of Master Equation Models

2.3.1 Selection of Order Parameters

The first task one comes to when constructing Markov models from molecular simulation data is how to project an inherently high dimensional process onto a simple time series of order parameters. If one has extensive knowledge of the system in question it may be possible to project the data onto some bond distances, RMSD between secondary structures, etc. and obtain an accurate model. However, in many cases one will not know much, if anything, about the system studied. Here projection techniques such as principle component analysis (PCA), or a more sophisticated method such as time-structure independent component analysis (tICA) may be used. PCA selects mutually orthogonal order parameters that maximally describe the variance in the MD data using the eigenvalue decomposition of the correlation matrix. This tends to identify large amplitude motions in terms of molecular coordinate changes. However, these large

amplitude motions may turn out to be poor descriptions of the kinetically significant motions present in the system.

tICA extends the concept of PCA by identifying sets of order parameters that optimally describe the significant timescales of the system and optimally reproduce its eigenfunctions. One must supply time series of possible order parameters chosen manually prior to the decomposition step. Usually the first two linear combinations identified by tICA become the order parameters underlying the model.

2.3.2 Discretization of State Space

The identification of kinetically stable states requires some knowledge of the kinetics of the system as projected onto the chosen order parameters. Intuitively, one way to capture kinetics is to count transitions between adjacent grid points on the underlying free energy surface, as is sometimes done in toy models, followed by the pruning of unvisited grid points. In some of the more well-known techniques used to construct MSMs, this step is done by discretizing state space into microstates representing locally similar structures using a variety of geometric clustering algorithms. For the reasonably simple models studied in this thesis, we will use two basic spatial clustering algorithms, k-means and k-mediods,[55,56] that are derivatives of standard k-centers clustering algorithms. These methods group geometrically similar points into representative centers. For MD simulation data this is advantageous because it tends to result in a well discretized state space in areas of extensive sampling with less weight given to poorly sampled, transient regions which either do not contribute to our model, or will be accounted for with the COS simulations.

The end result of the discretization step is a set of several hundred or thousand microstates representative of the portions of state space most frequently sampled by the system in the MD simulations. The kinetics underlying the system can then be recovered by counting transitions between individual microstates. Transition counting is implemented by observing the progress of the trajectories projected onto the set of order parameters. The system is considered to be in a microstate i if the coordinates of the system are within a specified cutoff of the coordinates of the structure represented by the microstate. A transition between two microstates, $i \rightarrow j$, is observed when the trajectory leaves state i and reaches state j sometime in the future without entering any other states in the intervening time. The observed transition counts are stored in a matrix, $\boldsymbol{C}$, with elements C_{ij} equal to the number of observed transitions between states i and j. A visual depiction of the microstate discretization and counting process in shown in Figure

3A. The microscopic transition probabilities can be obtained by normalization of the rows of the matrix, $\boldsymbol{C}$, to obtain the matrix of underlying transition probabilities, $\boldsymbol{T}$. An element of the transition matrix,

$$T_{ij} = \frac{C_{ij}}{\sum_j C_{ij}}, \tag{2.3.1}$$

is the probability that the system will transition to state j at time $t + \Delta t$ given that the system is in state i at time t.

Several considerations must be made when counting transitions, including the choice of state definitions, the total number of microstates needed to represent that system, and the metric used to calculate distances. For biomolecular applications, the total root mean squared distance (RMSD) between all atoms in the representative structures is often used, while for data projected onto specific order parameters often times simple Euclidean distances can be appropriate. The number of microstates and cutoff distances are more difficult to determine as one needs enough microstates to fully describe the underlying surface, but not so many that there are few transitions observed between any pair of states. Similar problems occur when choosing a cutoff distance; the distance must be small enough that one can confidently believe that the system is actually in that microstate, but at the same time large enough that the system spends an appreciable amount of time in that state.

Many of these problems can be alleviated by increasing the amount of data one has available, either by increasing the length of the MD trajectories or by obtaining more trajectories. Long trajectories that are naturally in equilibrium are preferable to a large number of short trajectories for which the equilibrium detailed balance constraint is enforced by recording a $j \rightarrow i$ everytime that an $i \rightarrow j$ transition is observed. Long trajectories may also explore regions of the underlying free energy surface that are simply inaccessible on the timescale of shorter trajectories. Failure to meet the detail balance condition results in transition matrices with significant imaginary eigenvector components, inducing errors in state decomposition and rate

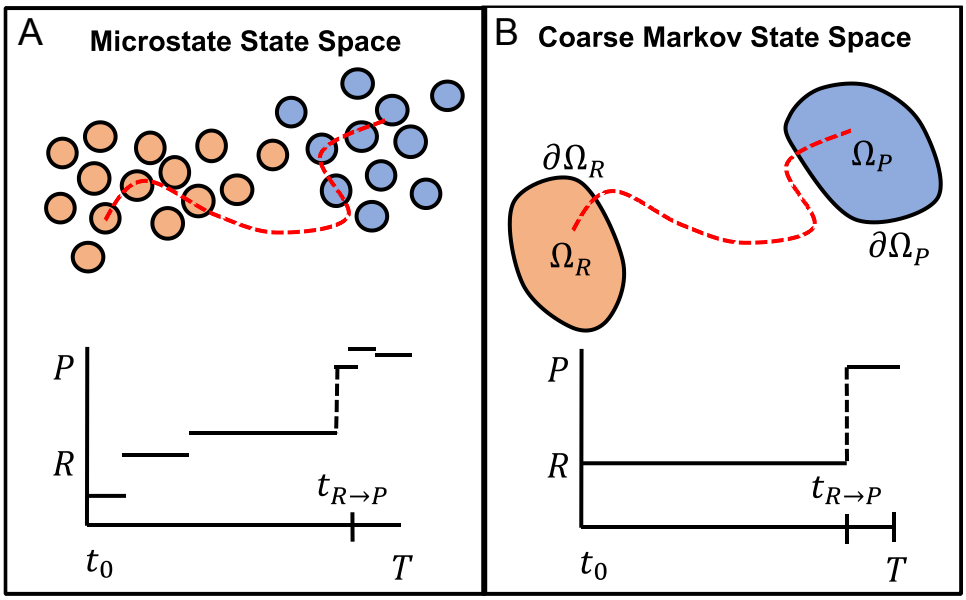

Figure 4. Construction of coarse master equation models. (A) A discrete microstate space is constructed with frequent, short-lived transitions. (B) The microstates are grouped by kinetic similarities and a boundary $\partial\Omega$ is selected to define a core region.

predictions. These problems are especially prevalent in the microstate count matrix, but are less significant if a coarse-graining step is applied.

A choice must also be made in how one represents the time-evolution of the system, either as a continuous-time master equation or as a discrete-time Markov chain. In this work, we choose to use a master Equation approach and build the models at the base timestep of the simulations. We also note that no matter how frequently data is printed during the MD simulations the time in the master equation model will never be truly continuous, but we make the point to distinguish this construction method from the lag-time based models.

In our approach, approximately Markovian dynamics are enforced by only considering metastable coarse-grained states. However, in the case of a discrete-time model, Markovian behavior is generally enforced by choosing a lagtime, Δt, such that the longest timescales of the system are converged. The lagtime is chosen based on the convergence of implied timescales, calculated from the eigenspectrum on the Markov transition matrix as,

$$t_{ITS}^{i} = -\frac{\Delta t}{\log(\lambda_i)}, \tag{2.3.2}$$

where λ_i is an eigenvalue of the transition count matrix and t_{ITS}^{i} is the timescale associated with the i^{th} eigenvalue.[8]

Following the choice of Δt, the transition matrix is constructed using the aforementioned methods using the portions of the trajectories recorded at times $t = 0, t + \Delta t, t + 2\Delta t, \ldots, T$. Such a model will naturally be inaccurate in describing any behavior faster than the selected lagtime. From a theoretical standpoint, the transitions between most states in the model will most likely not be Markovian, but the overall dynamics between major metastable states approximates a

Markov chain. In fact, the lagtime is generally chosen based only on the few largest timescales, assuming that the faster processes do not matter to the overall kinetics. Although lag-time based models are popular there is increasing evidence that the selection of lag time in the model can greatly affect the results one obtains in addition to lose of resolution in faster timescales. For this reason, we choose to use the coarse master equation description, though we then need to choose state definitions more carefully.

2.3.4 Decomposition into Kinetically Similar States

Identifying a set of states that are similar geometrically may not directly map to kinetically differentiated states. The identification of representative coarse states, also called macrostates or Markov states in this thesis, have been an area of active development in the field. The most popular methods of coarse-graining are based on the Perron Coupled Cluster Algorithm[57] (PCCA) due to their simplicity and theoretical agreement with the idea of dominant timescales. These methods seek to identify optimal lumpings of microstates using the eigenspectrum of the Markov transition matrix. Starting with the second eigenvector (the first corresponds to the stationary probability), the total state space is split based on the sign of the eigenvector elements. The sign change provides a natural separatrix for probability flow between those two grouping. Further decomposition requires use of the sign changes in the third, fourth, and so on eigenvectors. An optimal number of states is determined based on the cluster of eigenvalues near one, referred to as the Perron cluster.

PCCA and its derivative PCCA+[58], have been the most popular methods used in general MSM construction and are featured in MSM software[59,60]. A more recently developed method, the Bayesian Agglomerative Clustering Engine[61] (BACE), has been shown to provide better grouping in more complex systems compared to PCCA and PCCA+. Simple tests with random matrices also prove this fact. Implementation of BACE also requires the microstate transition matrix, but instead of calculating the eigenspectrum one instead compares the Bayesian probabilities of a pair of states coming from the same distribution of elementary transition probabilities. The metric in BACE, Eq. (2.3.3), compares the log-likelihood that two rows of the transition matrix come from the same distribution. Here $C_i = \sum_j C_{ij}$ is the sum over the observe transitions out of state i. $D(p||q)$ given in Eq. (2.3.4) is the divergence of the transition probability distribution out of a state, $\vec{p}_i$, and the expected probability distribution if two states were merged,

$$\log\left[\frac{\left(P_{diff} \mid \boldsymbol{C}\right)}{\left(P_{same} \mid \boldsymbol{C}\right)}\right] = C_i D(\vec{p}_i|q) + C_j D\left(\vec{p}_j|q\right). \tag{2.3.3}$$

Where $D(\vec{p}_i|q)$ is

$$D(\vec{p}_i|q) = \sum_k p_{ik} \log\left[\frac{p_{ik}}{q_k}\right], \tag{2.3.4}$$

and

$$q = \frac{C_i\vec{p}_i + C_j\vec{p}_j}{C_i + C_j}. \tag{2.3.5}$$

Rows are iteratively combined until one reaches a desired number of final states representing the metastable states on the free energy landscape. The number of states can be chosen by first performing the clustering until only one state remains and tracking the metric in Eq. (2.3.3). Theoretically, there will a large jump in the metric once one begins merging or over coarse-graining states that are significantly different. We choose the final number of states for our models by tracking this metric, though other methods based on the slowest timescale and metastability metric have been used. We generally observe a distinguishable discontinuity in the Bayes metric corresponding to coarse-graining different energy basins into a single state allowing identification of multiple metastable minima.

2.3.4 Estimation of Transition Rates

The rate of transition between the states identified through kinetic clustering are estimated by counting jumps between the cores of the states. The edges of the metastable states are defined to be within a cutoff in energy from the local minima within each such state. This corresponds to identifying an energy minimum on the free energy landscape defined by the selected order parameters. Intuitively, the ideal cutoff would be $E < 1\ k_B T$ from the minima, but for systems with broad, shallow wells multiple metastable states within the same general region can result just from large spatial separation. A simple radial cutoff can also be used, but this will fail to consider the natural shape of the free energy surface near the minima, leading to transitions that are recorded when in fact that system has not yet entered the desired region.

The rules for recording transitions are the same as outline in Section 2.3.2, with the exception that self-transitions are also recorded. That is, for every time step that an $i \rightarrow j$ event does not occur an $i \rightarrow i$ transition is recorded. The result is a coarse matrix of transition counts, $\hat{\boldsymbol{C}}$, with diagonal elements $\hat{C}_{ii}$ representative of the dwell-time within state i. The transition rates ω_{ij} are then be obtained using Eq. (2.3.6) where Δt is the printing frequency of the simulations, or the lag time for a discrete-time model.[53]

$$\omega_{ij} = -\frac{\frac{1}{\Delta t}\log\left(\frac{\hat{C}_{ii}}{\sum_j \hat{C}_{ij}}\right)}{\sum_{i\neq j}\hat{C}_{ij}}\hat{C}_{ij} \tag{2.3.6}$$

We note that there are several procedures for obtaining the transition rates, but they produce the same result given the Markovian assumption that global decay out of state i follows $e^{-\sum_{i\neq j}\omega_{ij}*t}$ is met.[62]

At this point, we have constructed a coarse-grained master Equation[49] governing the dynamics in the chosen order parameter space with an underlying rate matrix defined in Eq. (2.3.6). The methods outlined in Section 2.2 can now be applied to calculate free energy differences, first passage times, and other properties given that one has sampling adequate to define most pairwise interconversion rates. Now we proceed to developing our method to account for the case where one is not able to actually obtain a significant number of transitions from naive simulation.

2.4 Chain-of-States Methods

2.4.1 Introduction to Chain-of-States

The standard simulation techniques discussed in previous sections will allow one to study chemical processes on short time scales where it is feasible to directly integrate the equations of motion. However, in many fields one is interested in processes that occur on a much longer time scales than are routinely accessible by standard simulations. Within the field of computational biophysics, processes such as protein folding, conformational transitions, and enzyme catalysis dynamics occur on the μs-ms timescale which requires the use of an advanced sampling or rare events technique. Many of these techniques, including umbrella sampling[63], metadynamics[18], and transition path sampling (TPS), drive the system of interest along a low-dimensional reaction coordinate. Chain-of-states methods were developed to find the MEP between two predetermined endpoint states, making these techniques readily applicable to driving reactions between metastable regions of state space.

Chain-of-States simulations seek to connect disparate regions of state space using a 'chain' of copies of the system of interest driven along a reaction coordinate. Each copy, or image, of the system is set up as a separate MD simulation, and a constraint is applied to keep a discretized description along the entire reaction coordinate. In some methods, including nudged elastic band[64–67] (NEB), the constraint takes the form of a harmonic spring force connecting the images applied after a short minimization interval. Another class of methods based on the original

string method[68] instead enforce equidistant images along the reaction coordinate by reparametrizing the entire string during the minimization process. A large number of variations of the above methods including the growing string method,[69,70] string method in collective variables[71], and replica path[72] have been implemented to study conformational motion, enzyme catalysis, and protein folding among others. However, the minimization procedures used often neglect the entropic contributions of off path sampling, leading to less accurate free energies.

2.4.2 Finite-Temperature String Method

The finite-temperature string method[20,73–76] was developed, in part, to take into account such entropic contributions and obtain free energy estimates at a given temperature instead of relying on a zero Kelvin minimization scheme. Theoretically, the FTSM finds the optimal isocommittor surface at each discretized point along the reaction coordinate and is usually derived in accordance with TPT.[77,78] In terms of the TPT definition of the committor, a trajectory launched from any point on the isocommittor surface has an equal probability of committing to the products before of the reactants. Along the minimum energy path the forces acting in the transverse direction are uniformly zero at all images along the string. Given that the system is constrained to sample regions that are at a small perpendicular distance to the local string image one can represent the isocommittor surface using a local hyperplane approximation,

$$\xi_i = \hat{n}_i \cdot \left[\vec{r}_i - \vec{\phi}_i\right], \tag{2.4.1}$$

where ξ_i is the local value of the reaction coordinate at image i, $\vec{r}_i$ is the instantaneous position of the image, $\vec{\phi}_i$ is the current position of the image center, and $\hat{n}_i$ is the local path tangent to the string,

$$\hat{n}_i = \begin{cases} \dfrac{\vec{\phi}_{i+1} - \vec{\phi}_i}{\left\|\vec{\phi}_{i+1} - \vec{\phi}_i\right\|}, & i = 1 \\ \dfrac{\vec{\phi}_{i+1} - \vec{\phi}_{i-1}}{\left\|\vec{\phi}_{i+1} - \vec{\phi}_{i-1}\right\|}, & 1 < i < N \\ \dfrac{\vec{\phi}_i - \vec{\phi}_{i-1}}{\left\|\vec{\phi}_i - \vec{\phi}_{i-1}\right\|}, & 1 = N. \end{cases} \tag{2.4.2}$$

During the minimization process, the system at each image is restrained to sample parallel to the path using the potential,

$$U_{\parallel}^{i} = \frac{k_{par}}{2}\left(\xi_i^{\parallel}(\vec{r}_i) - \xi_i^{\parallel}\left(\vec{\phi}_i\right)\right)^2, \tag{2.4.3}$$

where k_{par} is a harmonic force constant and $\xi_i^{\parallel}(\vec{r}_i)$ is the position of the system projected into the hyperplane defined in Eq. (2.4.1). A constraint is also applied in the transverse direction to limit off path sampling using,

$$V_{\perp}^{i} = \frac{k_{prp}}{2}\left(\xi_i^{\perp}(\vec{r}_i) - \xi_i^{\perp}(\vec{\phi}_i)\right)^2. \tag{2.4.4}$$

Here k_{prp} is a harmonic force constant for perpendicular constraint and $\xi_i^{\perp}$ is the projection perpendicular to the path defined as,

$$\xi_i^{\perp} = \left[\vec{r}_i - \vec{\phi}_i\right]\hat{n}_i \cdot \hat{n}_i. \tag{2.4.5}$$

Sometimes, a flat bottom potential is used in the place of a harmonic potential to allow free sampling perpendicular to the path, in this case the potential in Eq. (2.4.4) is modified,

$$V_{\perp}^{i} = \frac{k_{prp}}{2}\min\left(\left[\left(\xi_i^{\perp}(\vec{r}_i) - \xi_i^{\perp}(\vec{\phi}_i)\right) - d_{prp}\right], 0\right), \tag{2.4.6}$$

where d_{prp} is the perpendicular distance from the image center at which the constraint is first applied.

After sampling using these constraints for a period of time, the string is updated using the running average of the position of each image over the sampling period. A reparameterization step follows to enforce equal spacing of the images along the reaction coordinate, taking the place of the spring force constraint in NEB-like methods. The potentials in Eq. (2.4.3) and Eq. (2.4.5) are updated to be centered at the new images and the minimization continues until convergence is reached. The string is generally considered to be converged when it is no longer changing significantly between updates as measured by the RMSD from the positions of the initial string images.[20] A stricter convergence criteria is to follow the decay of the forces to zero in order to meet constraint $\Delta U_{\perp} = 0$ at all positions along the string. However, in practice it is often hard to reach full convergence using the force constraint due to finite simulation times.

While conceptually simple and well developed theoretically, the results of FTSM are dependent on many parameters including k_{par}, k_{prp},d_{prp}, the frequency of reparameterization, the number of time steps average over in sampling, and, most importantly, the initial guess for the reaction path. To account for this, one often performs extensive parameterization runs prior to the production runs, which can become expensive for large systems. Several variations of FTSM have been implemented to alleviate some of these concerns including FTSM with umbrella sampling[19,79] and the Adaptive String Method.[80]

2.4.3 Free Energy Profiles from FTSM

The ultimate goal of the FTSM implementation in our method is to accurately measure free energies along the reaction coordinate in order to facilitate rate constant calculations. As previously mentioned, one of the attractive features of FTSM is that the method offers on-the-fly estimations of the free energy during the minimization process. The free energy is calculated using thermodynamic integration of the parallel forces along the string,

$$A(\xi_i) = \int_0^{\xi_i} \left(\frac{dA_i^{\parallel}}{d\xi}\right) d\xi\,, \tag{2.4.6}$$

where $A_i^{\parallel}$ is the free energy parallel to the path at image i. The average force acting on the string in the parallel direction can be obtained by differentiating Eq. (2.4.3) and averaging over the sampling window of duration, T.[20]

$$\frac{dA_i^{\parallel}}{d\xi} = \frac{k_{par}}{T}\int_t^{t+T} (1 - \xi_i) * \left(\hat{n}_i \cdot \left[\vec{r}_i(t) - \vec{\phi}_i\right]\right) dt \tag{2.4.7}$$

The free energies obtained from this integration are usually approximate and depend on the chosen force constants and parameters. In fact, it is generally accepted that one should obtain free energies post minimization using another method, usually umbrella sampling.[20,21,81]

2.4.4 Free Energy Profiles from Umbrella Sampling

Umbrella sampling is an enhanced sampling method that induces sampling in high energy regions using restrained simulations. First, consider a simple one-dimensional reaction coordinate, x. Umbrella simulations are restrained to be near single value of the reaction coordinate using the potential,

$$U_b(x) = \frac{k_u}{2}(x - x_0)^2, \tag{2.4.8}$$

where U_b denotes an umbrella window and the harmonic force constant for the umbrella potential is denoted k_u to distinguish it from the FTSM force constants. The windows are distributed along the reaction coordinate such that there is overlap in the distributions of x between adjacent windows. The same concept can be generalized for a one-dimensional reaction coordinate that is actually a function of several collective variables (CVs),[19,63,82]

$$U_b(\xi) = \frac{k_u}{2}(\xi - \xi_0)^2, \tag{2.4.9}$$

where ξ can be either a single CV or a mass-weighted linear combination of multiple CVs. k_u is chosen to ensure sufficient overlap between the umbrella windows.

The umbrella sampling simulations are then unbiased to obtain free energies using the Weighted Histogram Analysis Method[82,83] (WHAM). Given a one-dimensional reaction coordinate the single-histogram WHAM equations with N_u windows,

$$P_{\{k_u\}}(\{U_b\},\xi) = \frac{\sum_{l=1}^{N_u} N_l e^{-\beta \sum_{m=1}^{Nu} k_u^m U_b^m}}{\sum_{m=1}^{N_u} n_m \, e^{(f_m - \beta \sum_{m=1}^{Nu} k_u^m U_b^m)}} \tag{2.4.10}$$

and

$$e^{-f_m} = \sum_{\{U_b\},\xi} P_{\{k_u\}}(\{U_b\},\xi) \tag{2.4.11}$$

can be iterated to self-consistency.

2.5 Estimation of Rates from Free Energy Profiles

2.5.1 The Assumptions of Reaction Rate Theories

Traditional reaction rate theories have been applied to estimate rate constants for reactions projected onto a one-dimensional reaction coordinate starting in the reactant ensemble (state A in Figure 4) and ending in the product ensemble (state B).[30,84,85] The most well known of these theories is possibly transition state theory (TST). The central assumption of TST is that there exists a dividing surface between the reactants and products through which the flux is minimal, shown here at x = 0 in Figure 4. A further assumption is that any trajectory initiated in the reactants then passing through the dividing surface never recrosses before reaching the products. In practice, one is often unable to meet the second criteria, especially in complex systems where the dynamics must be projected onto a low-dimensional reaction coordinate. The direct result of this fact is that in real systems the rate calculated from a flux through the dividing surface method will be an overestimate.

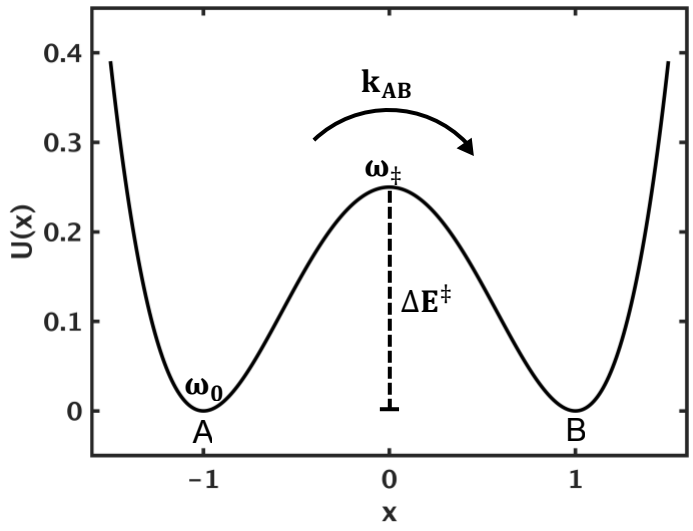

Figure 5. Assumptions of Reaction Rate Theory.[30] A reaction rate for a particle diffusing between A and B according to the potential, $U(X) = \frac{1}{4}x^4 - \frac{1}{2}x^2$, is calculated from the flux through the saddle point at x = 0.

Also important in TST and other theories based on flux through a

dividing surface is a separation of time scales between the relaxation time within the reactant ensemble and escape time, implying that the relaxation time with the reactant state is much faster than the decay time scale, $\tau_{AA} \ll \tau_{AB}$. This ensures that the reactant ensemble is in equilibrium at all times during the reaction supplying the justification for using equilibrium and memoryless assumptions. In terms of the thermodynamic energy, the requirement is that the magnitude of the transition state energy barrier, $\Delta E^{\ddagger}$, is several times larger than the thermal noise. Various forms of TST have been derived to address many situations that will not be discussed in entirety here. We instead use the most common form of TST, usually called simple transition state theory.

2.5.2 Simple Transition State Theory

Simple transition state theory takes the aforementioned assumptions and applies the additional constraint that the potential energy of the reactant well can be approximated by a harmonic potential,

$$U_0(\xi) = U(\xi_0) + \frac{\omega_0^2}{2}(\xi - \xi_0)^2, \tag{2.5.1}$$

where ξ_0 is the value of the reaction coordinate in the reactants and ω_0 is the oscillation frequency. The frequency ω_0 can be obtained from the second-derivative of the full Hamiltonian projected onto the reaction coordinate at ξ_0, $\omega_0 = \left(\frac{1}{M}U''(\xi_0)\right)^{\frac{1}{2}}$. Calculating the full Hessian matrix is often expensive in large systems, so we instead fit the local energy profile obtained from the COS simulations to the second term in Eq. (2.5.1) to obtain ω_0.

The expression for the rate passing through the dividing surface (transition state) derived in the canonical ensemble is,

$$\frac{\omega_0}{2\pi}\exp\left(-\beta\Delta E^{\ddagger}\right). \tag{2.5.2}$$

Eq. (2.5.2) is a basic expression for the rate that does not include the effects of friction or of dynamical effects. Adjustments for friction and dynamical effects are added in as corrections within the prefactor; the exponential term is unchanged. While expressions have been derived for the effects of friction on conformation motion as outlined in section 2.5.3, dynamical effects and tunneling in chemical reactions are often lumped together in a transmission coefficient that acts as an adjustment factor for minor corrections.[24,25]

2.5.3 Kramers' Theory

The theory published by Kramers[31] took a different approach to obtaining a rate over a barrier by representing the flow of probability along the reaction coordinate in term of the Langevin equation, Eq. (2.1.2). Assuming that the flux across the dividing surface was zero at equilibrium, Kramers described the time evolution of the probability density along the reaction coordinate using a Fokker-Planck equation,

$$\frac{\partial \rho(x,v,t)}{\partial t} = \left[-\frac{\partial}{\partial x}v + \frac{\partial}{\partial v}\left(\frac{U'(x) + m\gamma v}{m}\right) + \frac{\gamma k_B T}{m}\left(\frac{\partial^2}{\partial v^2}\right)\right]\rho(x,v,t), \tag{2.5.3}$$

where ρ is the phase space density, m is the mass of a particle, and γ is the damping rate. By making the assumptions that the potential energy at the transition state could be approximated by a potential of the form of Eq. (2.5.1) and that $\frac{\partial \rho(x,v,t)}{\partial t} = 0$ at equilibrium Kramers solved Eq. (2.5.3) for the probability density at steady-state, $\rho_{SS}(x,v)$. The rate of escape is then obtained by integrating,

$$k_k = \int_{-\infty}^{\infty} v\rho_{SS}(0,v)dv. \tag{2.5.4}$$

The resulting rate takes the form,

$$k_k = \frac{\left[\frac{\gamma^2}{4} - \omega_{\ddagger}^2\right]^{\frac{1}{2}} - \frac{\gamma}{2}}{\omega_{\ddagger}} \frac{\omega_0}{2\pi} \exp\left(-\beta \Delta E^{\ddagger}\right), \tag{2.5.6}$$

where $\omega_{\ddagger}$ is the frequency at the transition state barrier. At the overdamped or spatial-diffusion limit, $\gamma > \omega_{\ddagger}$, Eq. (2.5.6) reduces to,

$$k_k^{overdamped} = \frac{\omega_0 \omega_{\ddagger}}{2\pi\gamma} \exp\left(-\beta \Delta E^{\ddagger}\right). \tag{2.5.7}$$

We find that at high enough friction values Eq. (2.5.7) generally provides reasonable rates for the simple systems tested. The predictions begin to break down if the energy profile at the transition state is flat or nonsymmetric, rather than parabolic, or if projecting a high-dimensional process onto the reaction coordinate results in a lower effective friction (reflective in the damping rate, γ).

The value of γ in Eq. (2.5.6) and Eq. (2.5.7) is calculated from the autocorrelation function of the reaction coordinate in the reactant well. We run umbrella simulations in the reactant and products, project the data onto the local path tangent, and estimate the local diffusion coefficient using,[26,86,87]

$$D = \frac{\langle \xi^2 \rangle - \langle \xi \rangle^2}{\tau_{\xi}}, \tag{2.5.8}$$

where τ_{ξ} is the decorrelation time of the local reaction coordinate,

$$\tau_\xi = \frac{\int_0^\infty \langle \partial\xi(t)\partial\xi(0)\rangle dt}{\langle \xi^2\rangle - \langle \xi\rangle^2}. \tag{2.5.9}$$

The effective mass required for the conversion of the diffusion coefficient in Eq. (2.5.8) to the damping rate γ is calculated from the equipartition of energy along the reaction coordinate. The local diffusion parameter will vary along the reaction coordinate in reality, leading to most systems being better described by position dependent diffusion models. However, it has been shown that for overdamped dynamics, as used in this work, that an ensemble averaged diffusion constant can provide reliable results[28] and given that the process spends a disproportionate amount of time in the reactants, we should obtain a reasonable degree of accuracy with this approximation. Our calculation of the effective damping rate also assumes that the Einstein relation, $D = \frac{k_B T}{\zeta}$, is a valid approximation.

Kramers' rate theory provides a description of the kinetics of a system at varying viscosity, making the application of the theory usable for estimating rates of conformational transitions and protein folding from simulations in explicit solvent where Eq. (2.5.7) is applicable. However, for the case of chemical reactions most of the assumptions used in the derivation of Eq. (2.5.6) will not hold, mainly that the velocities are thermally averaged along the reaction coordinate resulting in a diffusive process. Chemical reactions are thought to occur on the fs-ps timescale, which is much faster than the relaxation time of the system. In the future, application of our methodology to chemical reactions will require using a more applicable rate theory, such as Grote-Hynes theory, which takes into account memory friction along the reaction coordinate.[32,33,88]

2.5.4 Markovian Milestoning

The calculation of rates from a Markovian milestoning[50,51] procedure differs from the assumptions of reaction rate theories. Here, we no longer use the energy as a function of the reaction coordinate to estimate flux

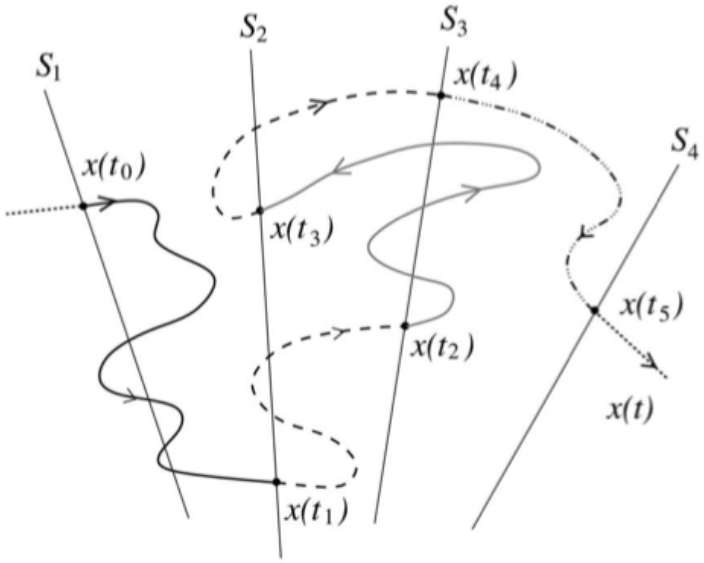

Figure 6. Demonstration of the milestoning procedure. In milestoning the trajectory is assigned to a milestone, S, until it hits a different milestone. This permits the trajectory to cross the same milestone multiple times without recording a transition, leading to longer dwell times. Figure taken from Vanden-Eijden and Venturoli, (2009).

through a narrow dividing surface. Instead, one uses simulation data along the reaction coordinate to construct a Markov process, using essentially the same transition counting procedure as described in section 2.3. The metastable states in the master equation model are replaced by hypersurfaces along the reaction coordinate such that transitions between the hypersurfaces are independent of all previous transitions. The isocommittor surfaces identified using FTSM serve as ideal milestones because, theoretically, any trajectory launched from that surface has the same transition probability, meeting the requirements of statistical independence. An illustration of the milestoning process is given in Figure 6. In order to construct the rate matrix, $\boldsymbol{R}$, one must record both the number of transitions between milestones and the total time the trajectories are associated with the milestone. The rate is calculated as,

$$\omega_{ij} = \frac{N_{ij}}{T_i}, \tag{2.5.10}$$

where N_{ij} is the number of transitions observed between milestones i and j and T_i is the total dwell time associated with milestone i.

The rate along a pathway is calculated using Eq. (2.2.16) with the assumption that in this case that the 'reactants' is the first milestone and the 'products' is the last milestone. We find that in some cases this approximation works quite well, mainly when the path found by FTSM lacks excessive curvature and the dynamics of the system are overdamped, consistent with other studies.[50] However, while in theory FTSM finds the optimal milestone, this is with the caveat that FTSM finds the optimal milestones given the set of initial conditions and parameters supplied. This means that the set of milestones depends on the number of images, the force constants, and especially the initial guess.

Despite these problems, calculating rates with Markovian milestoning is attractive for several reasons. First, the methods based on reaction rate theory introduce assumptions that the energy along the reaction coordinate is parabolic, which is unlikely to be true in complex systems. Second, the friction coefficient required from calculations with Kramers' theory, as does Grote-Hynes, requires projections onto the reaction coordinate followed by estimation from the autocorrelation function and the rate is dependent on the accuracy of these approximations. Milestoning allows one to calculate rates directly from simulation data without assumptions that are not often met in real systems. Perhaps most importantly a milestoning approach would allow us to calculate a rate even if there are minor intermediates along the reaction pathway or, even more problematic, undiscovered metastable states.

Chapter 3: METHODS

3.1 Application on a Theoretical Surface

3.1.1 Brownian Dynamics Simulations

Brownian dynamics simulations were carried out by integrating Eq. (2.1.3) in two dimensions with isometric diffusion constant $D = \frac{k_B T}{\zeta} = 1$ ps and $\Delta t = 0.001$ ps, according to the potential,

$$F(x,y) = \sum_i d_i * \exp\left(a_i(x - x_i^c)^2 + b_i(y - y_i^c)\right), \tag{3.1.1}$$

where the constants take on the values,

$$d = [-11, -12, -15, -14, 7, 5, 5, 5, 5], \tag{3.1.2}$$

$$x^c = [4, -3, -3, 4, 0.25, -10, 10, 0, 0], \tag{3.1.3}$$

$$y^c = [4, 4, -6, -6, 0, -3, 0, 1\ 0, -10], \tag{3.1.4}$$

$$a = [0.1, 0.05, 0.1, 0.05, 0.1, 0.05, 0.05, 0.01, 0.01], \tag{3.1.5}$$

$$b = [0.1, 0.05, 0.05, 0.05, 0.1, 0.01, 0.01, 0.05, 0.05]. \tag{3.1.6}$$

Sets of 200 statistically independent trajectories were run at the series of temperatures, $T = \{250, 275, 300, 325, 350\}$, for $2*10^9$ timesteps and recording the (x, y) coordinates every 1000 timesteps.

3.1.2 Master Equation Construction

The state space was discretized by clustering the time series of x and y coordinates into 1000 centers using the k-means clustering algorithm. The 200 sets of centers were combined and clustered again using k-means to get a set of centers representative of all of the trajectories. Transitions were counted between the centers with a cutoff radius of 1 in Euclidean distance from the centers' coordinates. The reactants region, defined as $\Lambda_R = \{-10 < x < 10: -10 < y \leq 0\}$, and the products region, $\Lambda_P = \{-10 < x < 10: 0 < y < 10\}$, were further coarse-grained into macrostates using the BACE clustering algorithm. The boundary of each macrostate, $\partial\Omega$, was defined as the contour in Euclidean space encompassing all areas of the surface within $1\ k_B T$ of the most probable point in the macrostate. The coarse-grained count matrix was constructed by counting transitions between the macrostates as defined by $\partial\Omega$. The master equation rate matrix, Eq. (2.2.9), was obtained from the coarse transition matrix using Eq. (2.3.6) and used to predict the first passage time, $< t >_{\Lambda_R \to \Lambda_P}$. The true first passage time was obtained by fitting the survival probability function,

$$s_R(t) = e^{-k_{obs}t}, \tag{3.1.7}$$

obtained directly from the exhaustive simulations and extracting the rate.

3.1.3 Construction of One-Dimensional Pathways

One dimensional free energy profiles were constructed along the two minimum energy paths connecting the reactants and product sets. The trajectories were projected along the y-axis and binned with $x_{path1} = \{-10 < x < 0\}$ and $x_{path2} = \{0 < x < 10\}$. The free energy in each bin was calculated from normalized occupation probabilities along that pathway. We note this is equivalent to integrating Eq. (3.1.1) with respect to x over the defined regions.

3.1.4 Error Estimation Procedure

The error analysis procedure used here follows, in general, the work of Sharma *et. al.*[62] In summary, in order to sample the ensemble of transition matrices underlying our models we first convert the master equation rate matrix to a matrix of transition probabilities, including self-transition probabilities. Then we simulate virtual trajectories through Markov state space, initiating each chain according to the stationary probability of occupying a state. The next state is chosen by generating a random number in the range $0 < r < 1$ and checking which region of the number line each element falls in. The next state is the state with the transition probability corresponding to that region of the number line, similar to the procedure in the well-known Stochastic Simulation, or Gillespie, Algorithm.[89,90] Each virtual trajectory was simulated for a length of time equal to the duration of the Brownian dynamics trajectories and 200 virtual trajectories were considered to be a single 'sample' of the transition matrix distribution. This procedure was repeated until 1000 resamples were obtained and first passage times, stationary probabilities, fluxes, and all other properties were calculated for each sampled transition matrix and the 95% confidence intervals were estimated for each quantity.

3.2 Application to Dipeptides in Vacuum

3.2.1 Molecular Dynamics Simulations

All-atom Langevin dynamics simulations were carried out in vacuum for alanine, proline, glycine dipeptides using the Charmm MD software. All simulations used the CHARMM36 force field with CMAP corrections, a dielectric constant of 1, and collision frequency $\gamma = 50$ ps^{-1}. 100 statistically

independent trajectories of ~1 μs in length were obtained at five different temperature for each dipeptide structure. The set of temperatures (in units of K) were as follows: for alanine, T = $\{400,450,500,550,600\}$; for proline, T = $\{600, 650, 700, 750, 800\}$; for glycine, T = $\{250, 275, 300, 325, 350\}$. The resulting total aggregated simulation time across all dipeptides was ~1.5 ms. Integration was performed using the standard leap-frog integrator in Charmm with a time step of 1 fs and saving coordinates every 1 ps. The trajectories were projected onto the (θ, φ) dihedral angles in the cases of alanine and glycine, and the (ξ, φ) dihedrals in the case of proline where the ξ angle is a virtual dihedral reporting on the cis-trans rotation of the amine group.[91] All analysis was carried out using the MDAnalysis[92] module in python.

3.2.2 Construction of Master Equation Models

The state space for each dipeptide was discretized by first clustering a randomly selected portion of each of the 100 independent trajectories into 1500 representative centers using the k-means clustering algorithm. The sets of 1500 centers were combined and again clustered using k-means to produce a set of 3000 centers representative of all trajectories. The 3000 centers were trimmed to remove any states within 2° of another to prepare for transition counting. The resulting number of centers ranged between ~1200 and ~ 2500. The microstate transition count matrix was constructed by counting transitions with a cutoff radius of 1°.

For the kinetic clustering step, the state space of proline and alanine was further split into 'reactants' and 'products' regions to simulate application of our method to a system where one cannot obtain sampling of the entire state space. The regions for alanine are defined as $\Lambda_P = \{0 < \phi < 130;\ -180 < \psi < 180\}$ and $\Lambda_R = \{-180 < \phi < 0 \cup 130 < \phi < 180;\ -180 < \psi < 180\}$. The regions for proline were defined as $\Lambda_P = \{-180 < \xi < -50 \cup 50 < \xi < 180;\ -180 < \psi < 180\}$ and $\Lambda_R = \{-50 < \xi < 50;\ -180 < \psi < 180\}$. Due to the symmetry of the glycine free energy surface no distinct regions were readily discernable so the entire state space was clustered together. Clustering was carried out using the BACE algorithm and the number of macrostates was chosen based on the largest jump in the Bayes information metric between iterations. If the selected grouping included single microstates acting as an entire macrostates, we proceeded to use one fewer state in our model until boundaries could be defined (requiring greater than 2 microstates).

The coarse-grained transition count matrix was constructed using both an energy based cutoff. For the energy-based cutoff, the macrostate boundaries were defined as the convex region in dihedral space encompassing the region within the specified energy cutoff of the most probable

point in the macrostate. Following the construction of the coarse transition matrix the master equation rate matrix was constructed using Eq. (2.3.6). The rates calculated from the master Equation model and the true rates from the MD simulations were compared to discern how well each constructed model can self-consistently calculate the overall rate.

3.2.3 Finite-Temperature String Simulations

Following the construction of the master equation models, FTSM simulations were performed connecting the reactant macrostates to the product macrostates in order assess the overall accuracy of our method in systems where one knows, approximately, the exact rate. Due to time and resource constraints, FTSM was performed for proline only for the purpose of serving as an example in this work, though simulations for all dipeptides were performed elsewhere. For each pair of reactant and product states four linear initial guesses were generated in four directions, to promote exploration of pathways over each edge of the periodic state space. This resulted in 16 pathways per temperature for a total of 80 pathways total for proline dipeptide. FTSM optimization was performed using the force constants reported in Appendix A. The positions of the string images were updated every 1 ps using an evolution frequency of 100 fs, and skipping the first 100 fs of sampling after each update to allow for decorrelation. The minimization was carried out for 100 ns per pathway and convergence was ascertained by the convergence of the RMSD from the initial guess.

Following minimization, pathways were removed from the model if the associated transition tube overlapped with the boundary of any macrostate other than the endpoints. The reasoning being that intra-reactant and intra-product rates are already accounted for in the master equation model prior to the FTSM step. This resulted in 3 final pathways per temperature that visually follow the minimum energy paths on the underlying free energy surface.

3.2.4 Free Energy Profiles from Umbrella Sampling

US simulations were performed post-FTSM to estimate the free energies of the images along the profiles. Harmonic constraints were applied to the relative dihedral angles for each dipeptide. Simulation were performed in Charmm with the CHARMM36 force field with CMAP corrections for a length of 10 ns per window. Otherwise the parameters were the same as those used for the simulations in Sec. 3.2.1. The resulting restrained trajectories were analyzed with the MDAnalysis[92] package in python and the results were unbiased using WHAM in two dimensions.

The free energy of each image along of the path was taken from the free energy of umbrella windows, which approximates the true PMF when using strong force constants.[82]

3.2.5 Estimation of Rates

The transition rate constant for each pathway was calculated in several ways. First, the classical TST rate was estimated using Eq. (2.5.2) where the curvature was obtained by fitting the profile within a cutoff of the well to a parabolic function. Second, the rate was calculated using Kramer's theory in the overdamped regime and transition state theory. The curvature term near the reactant well and the barrier top were found by fitting the profile within 1 $k_B T$ of the well and barrier top respectively to a parabolic function. The friction coefficient was estimated from the diffusion constant calculated using Eq. (2.5.8), utilizing a set of 10, 1 ns trajectories printing every 10 fs, centered at each endpoint. Third, the umbrella simulation data was used perform the milestoning procedure along the reaction coordinate. The milestones were defined to be the edges of the Voronoi cells produced by the dihedral coordinates of the final string. Only transitions between adjacent milestones were recorded to prevent any biasing of the results due to the restraining force. The first passage time was predicted using Eq. (2.2.17).

Chapter 4 RESULTS:

4.1 Verification on a Theoretical Surface

To test the applicability of our method on an ideal test case a theoretical free energy surface was considered. This surface features 4 clearly distinguishable free energy basins, Figure 7A, and two distinct pathways, Figure 7B, connecting the less stable 'reactant' region ($y < 0$) to the 'product' region. Brownian dynamics simulations were executed using the numerically calculated derivative of the energy and coarse master equation models were constructed from the trajectories. In all cases, four total metastable states were found; two states in the reactant region and two states in the product region corresponding to the wells on the underlying free energy surface. The exact overall rate was calculated from the survival probability of the system in the reactant region depicted in Figure 7A. The ability of the master equation model to self-consistently calculate the overall rate was assessed by applying Eq. (2.4.17) to calculate the rate of transitioning between the set of reactant states and the set of product states. One should also consider whether or not the master equation model is capable of predicting trends in the physical behavior of the system to any degree of accuracy. Consistently predicting a trend in the transition rate lessens the likelihood that an accurately

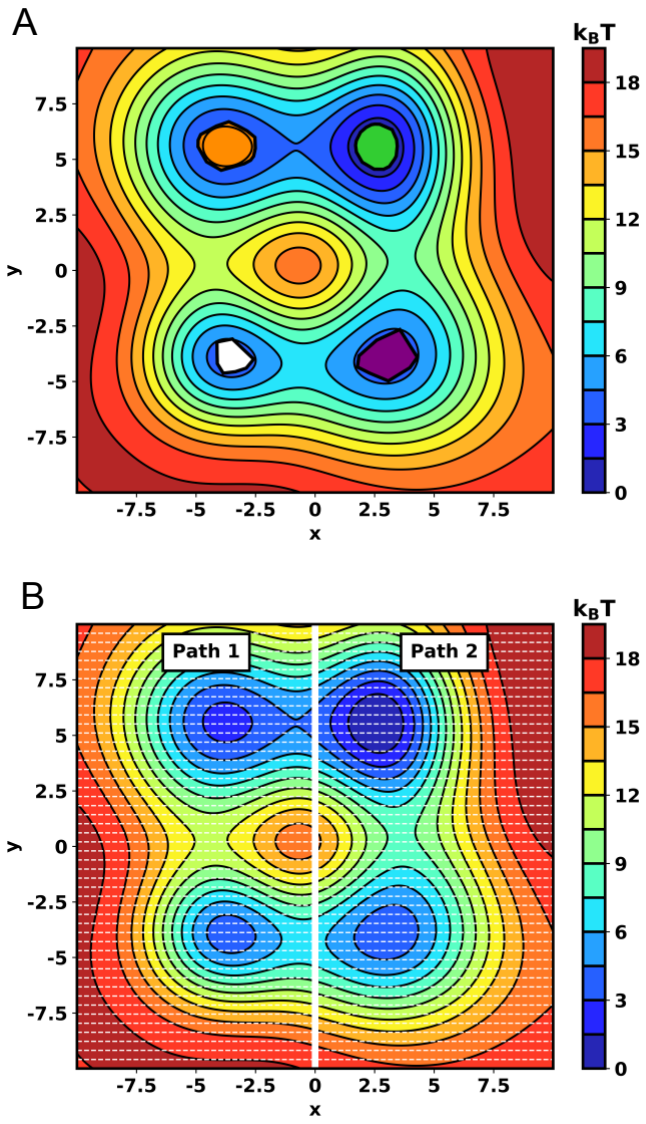

Figure 7. Decomposition of toy model into parallel pathways. The free energy underlying the toy model (A), produced from Eq. (3.1.1) and (B) decomposed into the two major pathways connecting the reactants ($y \leq 0$) and the products ($y > 0$).

calculated rate is a result of cancellation of errors, common in rate predictions.[34] Here, we predict the Arrhenius behavior of the system at a series of temperatures with the expectation that the rate is proportional to the Arrhenius factor, $e^{-\frac{\Delta E^{\ddagger}}{k_B T}}$.

Accurate prediction of rates from free energy profiles

To further test the accuracy of our method in combination with 1D energy profiles, we estimated rates from one-dimensional free energy profiles along the two dominant reaction pathways labeled in Figure 7B. To obtain the free energy profiles, the system was projected on the y-axis between the endpoints of the pathway, binning the Brownian trajectories in 50 windows along the

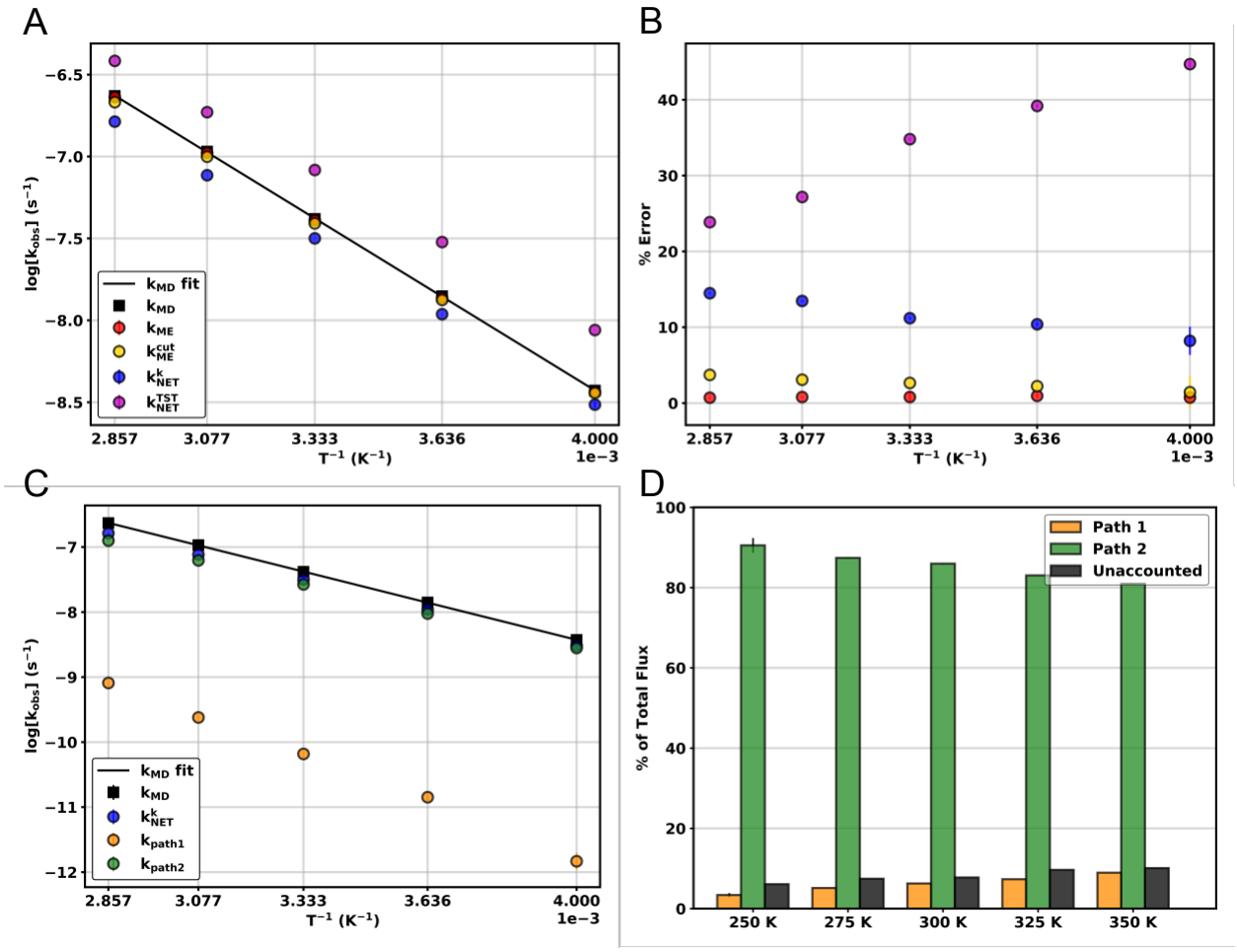

Figure 8. Effects of multiple pathways on rate predictions in a toy model. The Arrhenius behavior of the observed rates predicted using the master equation model (red), trimmed master equation model (gold), and the network model containing rates extracted using TST and Kramers' theories. (B) The respective errors for each transition. (C) The prediction of rates is decomposed into the individual pathways shown in Figure 6B, and the path fluxes (D) are predicted for each pathway along with the flux that the network model does not account for. All error bars represent the 95% confidence intervals predicted from the resampling procedure followed by bootstrapping.

reaction coordinate, and calculating free energies from the probabilities of occupying a bin. The rate constant for each pathway was estimated using Kramers' theory and incorporated into the coarse master equation. The rates predicted from Kramers' theory scale logarithmically with the inverse temperature, labelled k_{NET}^{k} in Figure 7A in reference to the rate predicted from our network model. We note that in the original master equation model there are, while small, technically non-zero transition rates between the reactants and products not directly connected by path 1 or path 2 in Figure 7B. The highest possible accuracy of our combined network model, and any predicted trends, will correspond to the master equation without these rates. We also predict the overall transition rate after 'cutting' out these rates from the model, k_{ME}^{cut} in Figure 7.

As expected, the complete master equation predicts the overall transition rate quite accurately, while we observe underestimations from the rates predicted by the network model as a result of the missing rates. The trend towards higher accuracy at lower temperatures is somewhat surprising initially, as one would expect less complete sampling and hence poorer predictions. However, the toy model used here allowed extensive sampling such that we have good statistics even at lower temperatures. The lower temperatures would tend to reduce the effects of the removed pathways by forcing transitions to proceed directing along path 1 and 2 only, resulting in a more accurate prediction of the rate using the network model.

We observe larger errors in the predicted rate when using classical TST, k_{NET}^{TST}, to extract the rates from the free energy profiles. This is because TST does not account for the effect of damping along the reaction coordinate and the resulting slowdown in the rate as stated in Section 2.5. We proceed to use the microscopic rates extracted with Kramers' theory throughout the rest of this section, though we do note that in some situations the TST estimates could possibly give better estimates of the absolute rate than Kramers' theory. This would occur if we failed to account for a significant portion of the flux in the system, resulting in a cancellation of errors. The behavior of the rates estimated from TST serves as a baseline calculation that depends only on the free energy barrier and local curvature with the reactant well, not on the local diffusion. In the future, TST could be used to check whether any observed discrepancies from Arrhenius behavior is a result of errors in the free energy calculation or in the estimation of local diffusion constants.

Multiple pathways are required to accurately predict the overall kinetics in a toy potential

To test whether our method is more accurate than a single pathway method we compared the results with both pathways to the model when only one of pathways is included, Figure 8C. We find that using both pathways does produce a higher degree of accuracy, though only slightly so

when comparing to using the MFEP (path 2 in Figure 7A). If one considers only path 1, the results are quite poor both in terms of the overall error in the calculated rate and the deviation from the expected Arrhenius behavior. We also compare the predicted flux through each pathway in the full network model, Figure 8D. We find that ~90% of the total flux flows along path 1, similar to a simple estimate one would obtain by comparing the relative energy barriers of the paths. We also see an increase in flux through both path 1 and other unaccounted for pathways as the temperature increases. These changes are small, but as will be shown in later applications even small changes in the flux can have affects on the predicted Arrhenius behavior.

One may point out that in this case the MFEP does predict both the overall rate and the Arrhenius behavior reasonably well. However, while it is obvious in this case where the MFEP path is and how one should obtain it, in more complex systems it is unlikely that this will be the case. Including multiple pathways will increase the likelihood of finding not only the MFEP but other paths contributing significantly to the overall reactant to product flux. In addition, finding reaction pathways with COS methods relies, almost deterministically, on an initial guess that would make finding the global MFEP difficult. Obtaining only path 1 could be thought of as the worst case result one could obtain using a single pathway method.

Here, we have shown that our method can predict both overall transition rates and trends robustly in an idealized case. We also compare the results from the network model to the results obtained from the equilibrium master equation model as an additional robustness check and see no significant deviations. We anticipate that in more complex systems finding the true MFEP will be more difficult than for our analytical model and multiple pathways will be required to predict the true dynamics of the system.

4.2 Conformational Dynamics of Dipeptides in Vacuum

We follow the test on an analytical surface by applying our method to a series of small dipeptide molecules[93–95], including the alanine dipeptide, a ubiquitous test case in computational chemistry and physics. The use of multiple dipeptides as test systems allows us to consider the effects of a variety of timescales, order parameters, number of pathways, and underlying free energy surfaces with the goal of gaining insight into optimal application of our method. The use of these chemical systems invoke an implementation of FTSM in a manner more similar to a large biomolecule than one would obtain applying FTSM to a toy model. Each dipeptide was studied at a series of

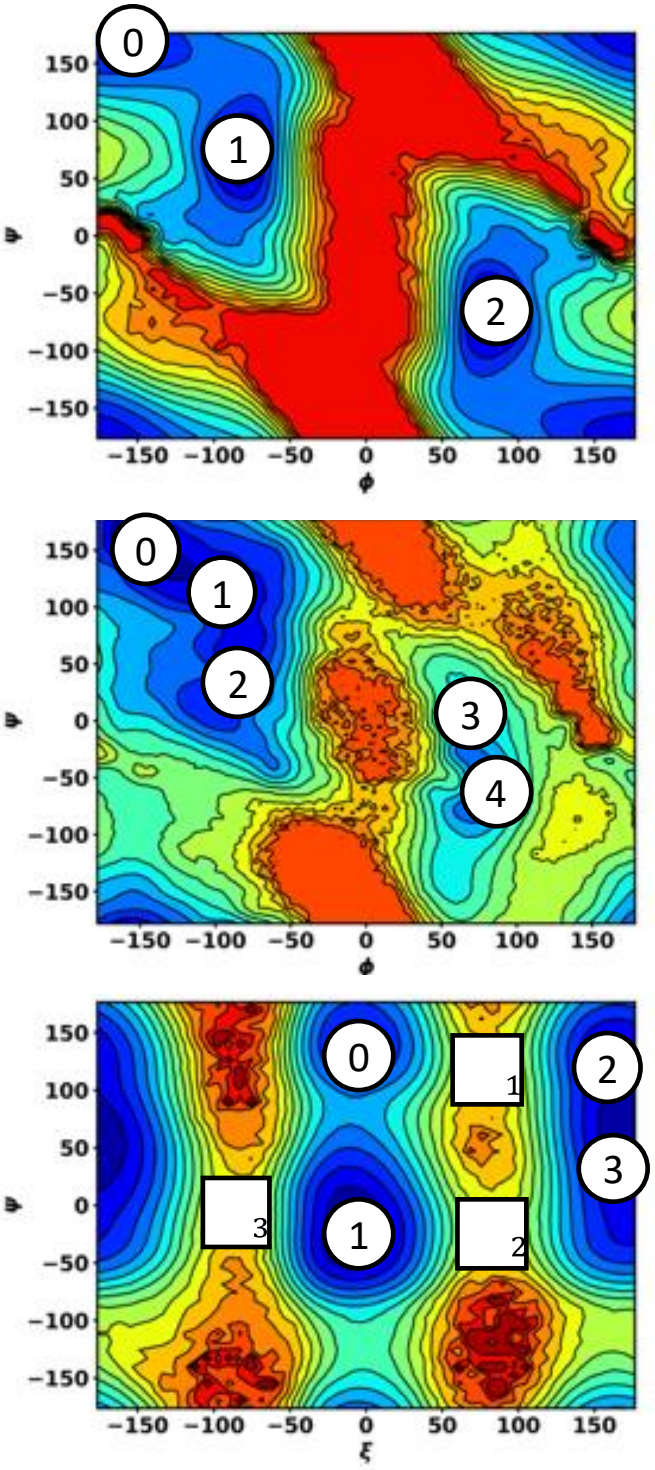

Figure 9. Free energy surfaces of dipeptides in vacuum Free energy surfaces predicted from binning exhaustive trajectories using Eq.(2.2.11)for (A) glycine, (B) alanine, and (C) proline dipeptides plotted at 300K, 500K and 700K respectively. White circles represent the states in the network models across all temperatures. For proline the location of the transition states for each pathway found using FTSM is labeled. Contour lines represent 1 $_{B}$.

temperatures and the trend in the overall transition rate as a function of temperature is again of interest. The effect of different choices made within our workflow was also assessed, including the clustering algorithm used, the size of the identified macrostate, and the method of calculating a rate constant post-FTSM simulations. As in the previous section, we compare the ability of the constructed master equation to self-consistently compute an overall transition rate and the effects of including multiple pathways.

Master equation construction reveals changes in model topology at different temperatures.

Long, unbiased MD simulations were run for each of the dipeptides and the underlying free energy surfaces were obtained by binning the complete set of trajectories at each temperature. The placement of the metastable energy basins remain mostly unchanged across all temperatures. Representative surfaces are shown in Figure 9 for reference. We find that in all cases our Markov model construction places states in the visible energy basins. In the case of alanine dipeptide specifically, Figure

9A, we discover additional states as the temperature is increased. This occurs in alanine, but not in proline or glycine, because there are significant, flat regions that become populated at higher temperatures resulting in additional states that are kinetically distinct mostly as a result spatial separation, rather than stabilization from an energy minima. Though the temperature range studied here for alanine (400K – 600K) is somewhat extreme this shows that in some cases the topology of the model can change based on the physical characteristics of the system. This contrasts with methods that use reweighting schemes to estimate state probabilities and transition rates from replica exchange simulations and other enhanced sampling methods to construct models under the assumption of constant topology.

State definitions affect the accuracy of the predicted kinetics

Following identification of the relevant metastable states within each system, we address whether the definition of the boundary of a state can have significant effects on the kinetics predicted by the master equation model. Intuitively, the state definitions will affect the kinetics simply because the state dwell times change, but there can also be other effects. For example, having states so large that their boundaries almost overlap would result in a significant number of non-Markovian transitions between the states. We address this question by varying the size of the Markov state by redefining the boundary of each metastable state, first using a simple radial cutoff in dihedral space, and then an energy-based cutoff. Here we discuss that energy-based cutoff results, but the radial based cutoffs produce virtually the same results.

The implementation of an energy-based cutoff is motivated by the assumption that the system can freely sample regions within one unit of thermal energy. Theoretically, one would then set the ideal boundary to be a contour containing all state-space within 1 $k_B T$ of the local minimum within the metastable state. In practice, we find that cutoffs of around ~0.5 $k_B T$ and lower work best. This is because in broad, flat wells one can identify multiple, stable states within the same energy well that are separated enough spatially that the state can be considered significantly different, as in alanine. In such cases, larger cutoffs tend to cause overlap between states and significant errors in the predicted kinetics.

We explore the effect of the size of the state in terms of the energy cutoff across all of the dipeptides at all temperatures by first comparing the overall transition rates calculated with Eq. (2.2.17). The rates predicted from the full master equation model from the exhaustive simulations are displayed in Figure 10. The full set of data, including the error in the overall predicted rate is shown in Appendix B1-3. We find that the overall transition rates predicted from the master

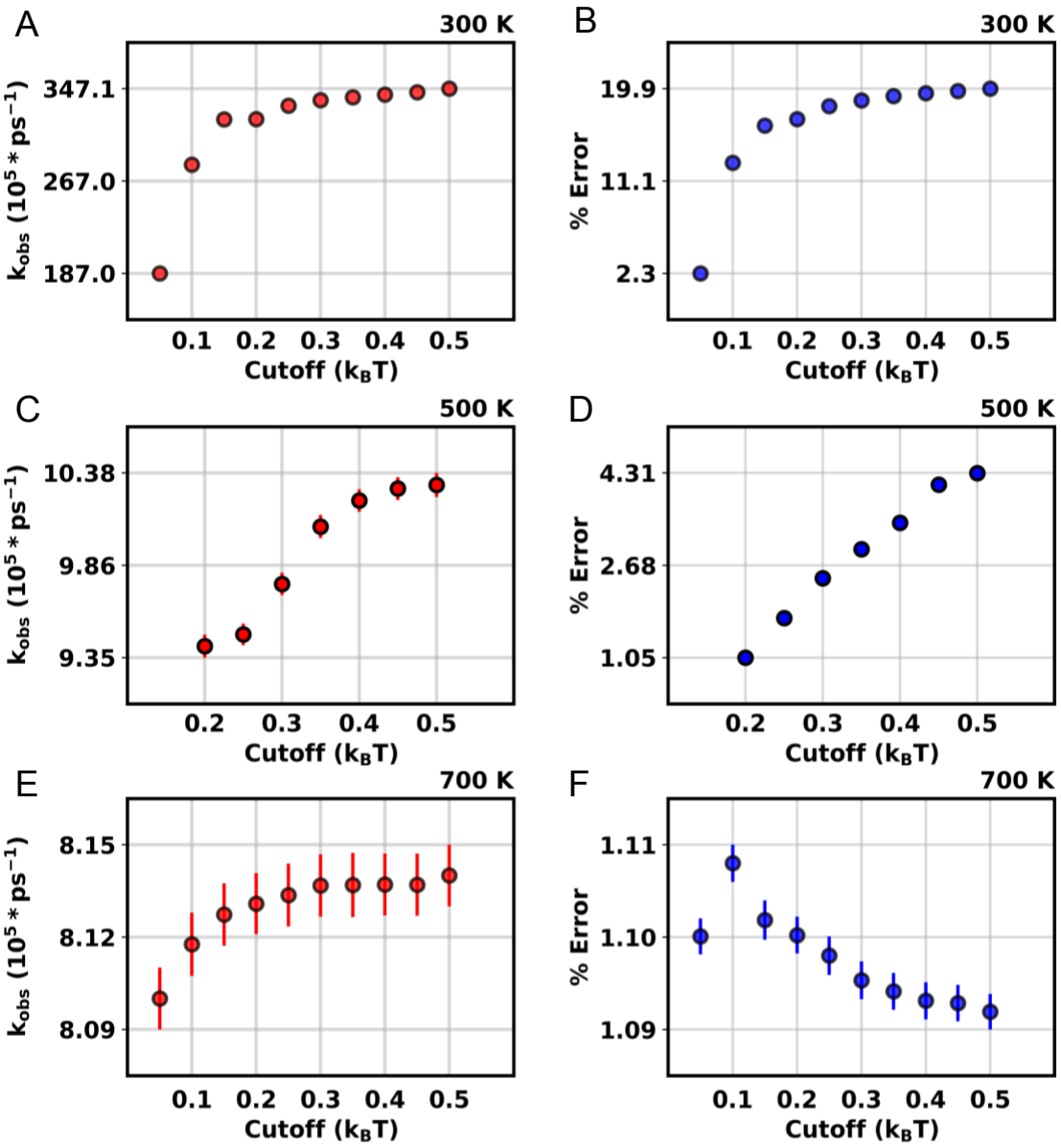

Figure 10. Convergence of predicted rates as a function of state definitions size. The calculated transition rate from the full master equation model for (A) glycine at 300K, (C) alanine at 500K, and (E) proline at 700K. The calculated rate is compared to the true rate from the exhaustive simulations and reported in (B,D,F) for glycine, alanine, and proline respectively. The error bars represent the 95% confidence intervals predicted from the resampling method followed by bootstrapping.

equation model begins to converge as the size of the state increases. This convergence can be attributed to the size of the state progressing past the portions of state space that are locally accessible due to stochastic forces. At smaller cutoffs, the states are so small that the trajectories may simply pass through space in the vicinity of the state but never actually be counted as entering. At state sizes larger than those shown in Figure 10, the boundaries of the states begin to overlap and our models break down. We find that for glycine and alanine, (Figure 10 B,D) there

is a trend towards increasing accuracy of the master equation at predicting the true transition rate as the size of the state is decreased, but we observe an opposite trend for proline (Figure 10F). We also note the larger relative change in accuracy in the predicted rate in the case of alanine and glycine compared to proline, indicating that not all systems will display such drastic changes as observed in the former cases.

Overall, we find no consistent trend in the accuracy of all models across all systems, but instead note that there are ranges of state sizes at which the rates predicted from the models are converged. This is a similar to the results of other studies that found the timescales of a model are converged for some series of state sizes, but diverge at too large or too small of values[96], and also similar to the concept of having converged timescales in other lag-time based construction methods.

Robustness of the trends predicted from the network model can break down at small state sizes

We also use our model to predict the Arrhenius behavior in the overall transition rate using a series of state definitions. In the case of proline, Figure 11E, we find few differences in the predicted behavior as state size is varied. However, for both glycine and alanine we find that there are significant deviations from linearity at small state sizes, Figure 11A and 11C respectively.

In the case of glycine, this can be explained by observing that there is an on-path intermediate state connecting the more stable region of the reactants to the products. At small state sizes, many transitions can simply pass through that region without being counted. In addition, at higher temperatures (>300K), the intermediate state becomes more stable through a combination of its geometric size and increased flux passing through it. A small state size disrupts this phenomenon. There is also a switch in the predicted stability of the selected states in glycine as the state size is increased (Appendix B4). In the case of alanine, the topology of the model changes as temperature is increased possibly resulting in the deviations from linearity. Alanine is also interesting because several of its states are identified based on spatial separation in dihedral space rather than strictly by energy stability, introducing the possibility of observing different dwell times within each state as temperature is increased.

Also of interest is the convergence of the stabilities of the Markov states as the size of the state is increased. For proline, Figure 11F, we again find no significant differences as state size

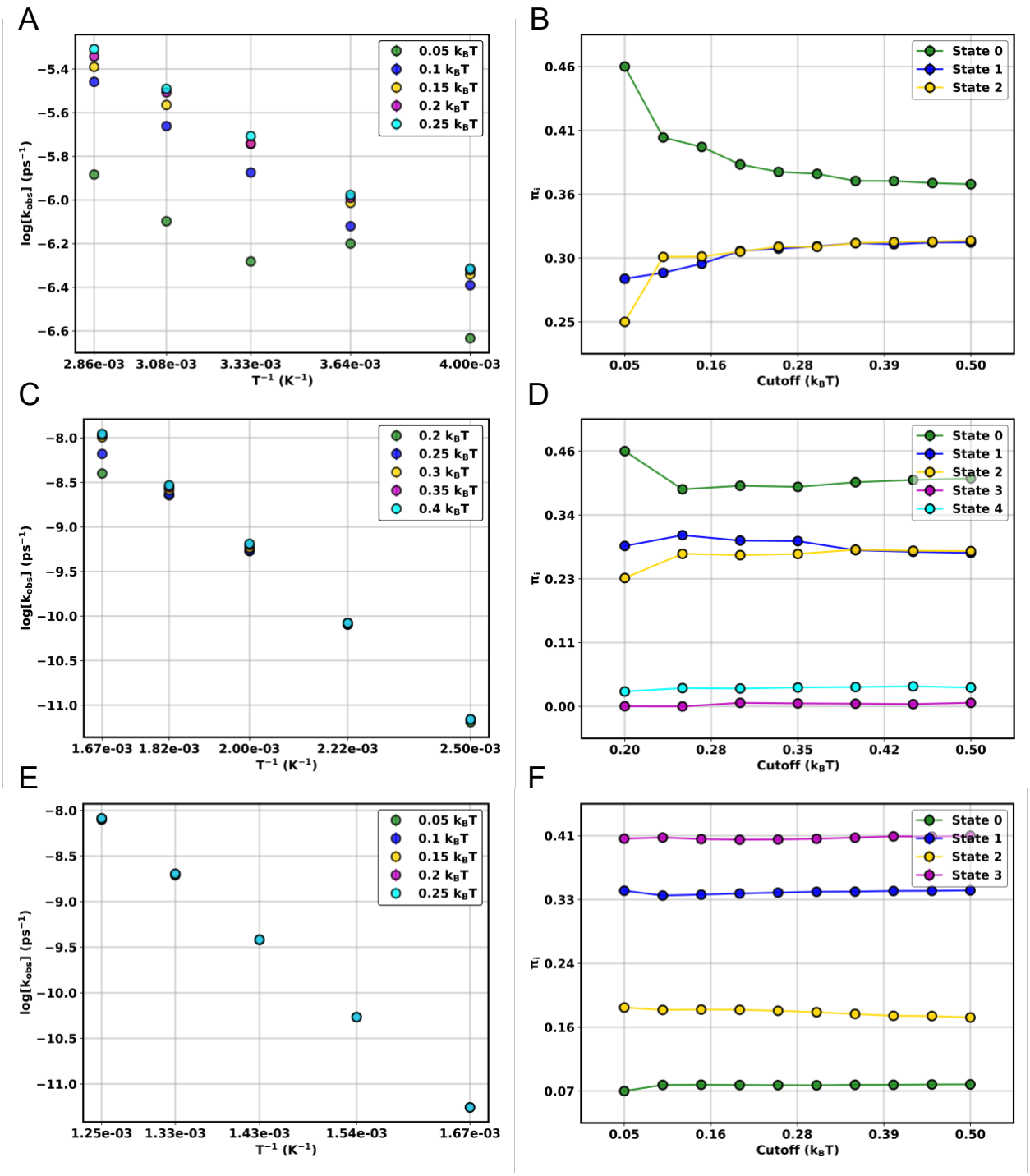

Figure 11. Convergence of Arrhenius behavior and stationary probability. The rate calculated from the full master equation model is plotted logarithmically as a function of inverse temperature for (A) glycine, (C) alanine, and (E) proline. The stationary probability of each Markov state shown in Figure 8 for (B) glycine, (D) alanine, and (F) proline plotted as a function of state size. The solid lines are to guide the eye. Error bars representing the 95% confidence intervals are smaller than the markers.

is varied. However, for the glycine models we find that the stabilities of the states, Figure 11B, are actually in the opposite order than what one would have expected from the free energy surface

in Figure 9A. As the state size increases the probabilities converge to the expected relative ranking. This is especially apparent in the glycine models due to the symmetry of the free energy surface, which implies that states 1 and 2 should be approximately the same in stability if our models are accurate. The state stabilities of alanine, Figure 11D, also converge as the state size is increased. There also appears to be a correlation between the convergence of the state stabilities, the predicted rates, and the behavior as a function of temperature, which is especially apparent in glycine and alanine.

Overall, we find that we are able to construct accurate models over a range of state definitions with indications that, at least in the systems shown here, that very small state sizes can cause inaccuracies. This conclusion would tend to contradict the observed increase in accuracy observed in Figure 10. This can be explained by recognizing that the errors shown in Figure 10 represent the ability of a model to self-consistently predict a transition rate given that a set of states has already been defined, implying that the ability to predict a rate with a high degree of accuracy may not always correspond to having a well-constructed underlying model. Small state definitions may also not correspond well to experimental observables given that any experimental signal with represent a macroscopic value, naturally corresponding to definitions matching an entire metastable region of state space. We also expect that the general convergence of the properties shown here should also be reflected in real systems for well-constructed models, even if the exact manner of convergence varies from system to system.

Lack of separation in timescales can result in poor agreement in overall rate predictions. Following the analysis of properties predicted at various state sizes, a final cutoff radius was chosen based on the convergence point of the predicted rate, stationary probability, and Arrhenius behavior. For proline, the stationary probability and Arrhenius predictions were converged almost immediately, so only the convergence in the predicted rate was used. A single cutoff was chosen and kept consistent across all temperatures in order to enforce consistency in the expected trends since the absolute value in the overall rate depends on the cutoff radius. In all cases the last convergence point was in the overall predicted rate as a function of temperature. The Markovinity of the models were assessed by fitting the survival probability of each state to an exponential function. In all cases the R^2 values from fitting the survival probability to a single exponential function were above 0.99 (not shown).

The overall rates were predicted using the models constructed with the chosen cutoff for each dipeptide at each temperature and compared to the rates predicted from the exhaustive MD

simulations. We find that we are able to predict the overall transition rate to a reasonable degree of accuracy with the exception of glycine. In this case we note that the energy barrier between the reactant and product ensembles are very low, in fact they are almost exactly the same for the on pathway intermediate state (state 1 in Figure 9A). The result is a lack of separation of timescales between the intra-reactant transition and the reactant to product transitions implying a breakdown of the normal assumptions of a reaction rate.

Inclusion of COS simulations connecting disparate regions allows an accurate description of kinetics in proline dipeptide.

As a full example, our proposed method was carried out in full on the proline dipeptide for the purposes of this thesis. Proline was chosen based on the simplicity and consistency of the models constructed across all the temperatures and the low number of viable pathways. Simulations have also been carried out for glycine and alanine as well but are reported elsewhere. FTSM simulations were carried out by connecting all reactant states to all products using four initial guesses per pair of states. Convergence of the simulations was judged based on the position of string images in dihedral space. Following the pruning of any pathways that overlapped with a non-endpoint state, we were left with three total pathways for each model. These pathways visually follow the minimum energy paths on the underlying free energy surface, Figure 12A.

The free energy profiles obtained from the US simulations were compared to the equilibrium free energy surface, Figure 12B, and found to be in generally good agreement.

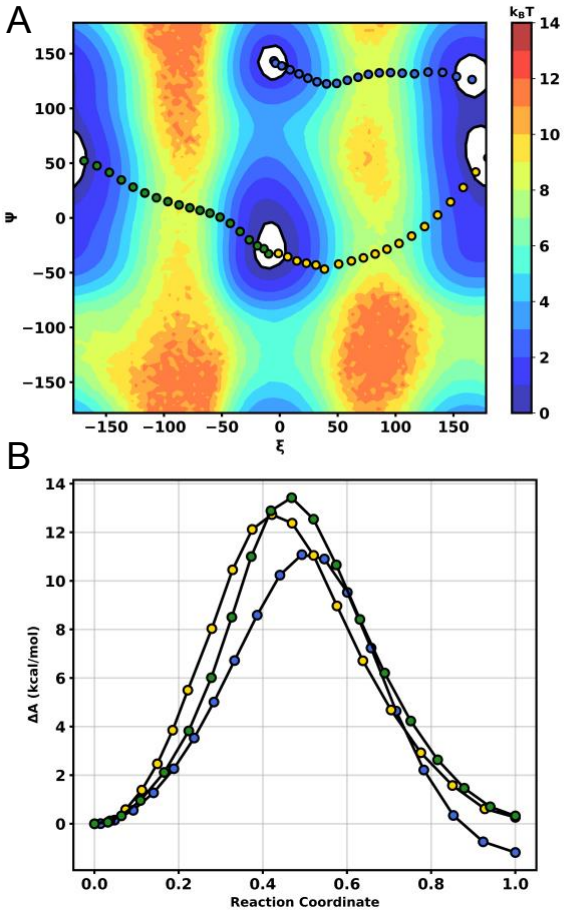

Figure 12. Full network model for proline dipeptide. (A) Free energy surface and associated network at 800K for proline dipeptide. The Markov states are shown at the selected cutoff of 0.35 k_BT as white polygons. The MEP is represented by blue markers, the IEP by gold, and the HEP by green. (B) Free energy profiles obtained from US with a force constant of 50 kcal/mol*deg^2. Coloring corresponds to the pathways in (A).

We do note, however, that this is more of a qualitative test as the free energies are estimated from a binning procedure and the bin size can significantly affect the obtained free energy. Still, the fact that we can obtain transition state free energies that approximate the true value to within $1\ \frac{\text{kcal}}{\text{mol}}$ supports the ability of our combined method to capture conformational dynamics accurately.

Rate constants were extracted from the free energy profiles using several methods including classical TST, Kramers' theory, and Markovian milestoning. The rates obtained from milestoning using the string images as milestone centers differ from the true values by several orders of magnitude. The rates also change when the placement of the milestones along the string is altered. The milestoning procedure is not discussed further here, but is instead included in Chapter 5 as a future avenue for rate predictions. We find that we are still able to predict the true rate well with the master equation model constructed from the exhaustive sampling, but see more significant deviations in the predicted rate after incorporating the rates extracted from the free energy profiles into the network model. Based on the results of the previous section we proceed to use Kramers' theory, but we would note that compared to TST the qualitative expectation $k_k < k_{TST}$ is met. For the rest of this section, it can be assumed that the rate estimated from our network model, k_{NET} is Figure 10A, includes microscopic rate constants obtained using Kramers' theory.

We account for the overall underestimate of the rate predicted by the network model, Figure 10A, by first noting again that there are rates in the full master equation model that do not have representative pathways in the network model, resulting in a natural underestimate. The second consideration is that the rates come from a low dimensional projection onto a one-dimensional reaction coordinate, followed by the extraction of a rate constant from a free energy profile using the assumption of a harmonic well, which may not always be a good assumption. In addition, accurately obtaining a local diffusion coefficient by projecting constrained simulations onto the local reaction coordinate can be difficult especially when one needs to compare projections using different normalized path tangents, as we do at different temperatures. We obtain local diffusion coefficients on the order of ~ 0.2 – 0.5 $\text{rad}^2\text{ps}^{-1}$, which are on a similar order of magnitude to those reported elsewhere for the torsional modes of dipeptides.[26] For future applications, any inaccuracies in the free energy barrier obtained from the COS/US procedure will dominate over most considerations involved in calculating a prefactor in reaction rate theories. This is fortunate because running constrained QM/MM simulations in the endpoint region for long enough to get converged correlation functions will be difficult from a computational standpoint.

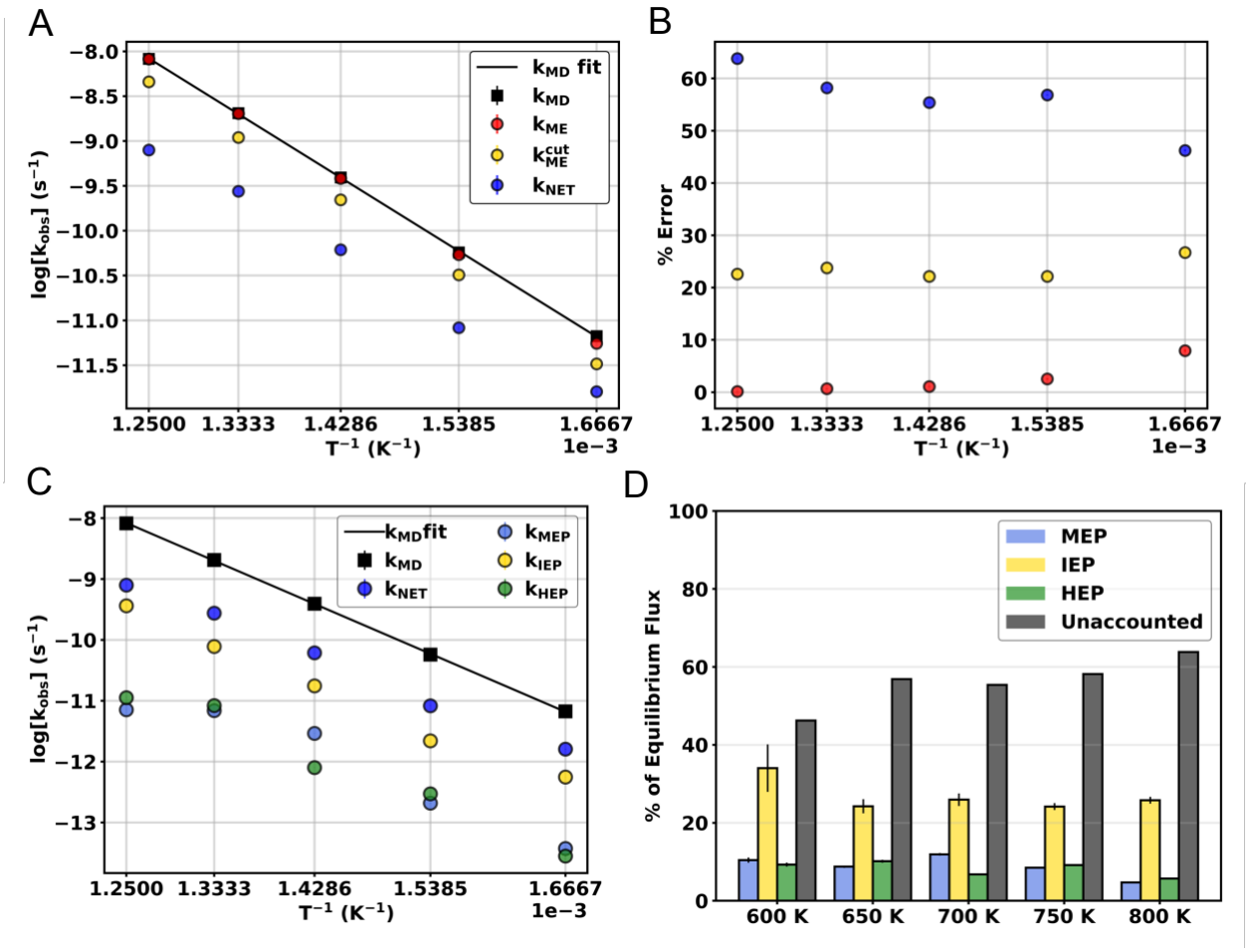

Figure 13. The effects of multiple pathways into the proline dipeptide master equation model. The Arrhenius behavior of the observed rates predicted using the master equation model (red), trimmed master equation model (gold), and the network model containing rates extracted using Kramers' theories. (B) The respective errors for each transition. (C) The prediction of rates is decomposed into the individual pathways shown in Figure 12A, and the path fluxes (D) are predicted for each pathway along with the flux that the network model does not account for. All error bars represent the 95% confidence intervals predicted from the resampling procedure followed by bootstrapping.

Incorporation of multiple pathways into the network model predicts Arrhenius behavior and overall flux more accurately than a single pathway method.

We assess the contributions of each pathway to the prediction of the overall rate and Arrhenius behavior in similar manner to the toy model in section 4.1. As expected, incorporating multiple pathways into the master equation model results in a lower overall error in the predicted rate, Figure 13C, compared to the rate predicted from any single pathway. Also noteworthy, is that we

find the pathway with the lowest free energy barrier does not reproduce either the overall transition rate, or the expected Arrhenius behavior accurately. This observation can be explained by decomposing the equilibrium flux between reactants and products into the flux along individual pathways.

As shown in Figure 13D, there are significant contributions from pathways other than the path with the lowest energy barrier. This fact results from the differences in stabilities of the end point states of the pathways. State 1 in Figure 10C is associated with the MFEP, but accounts for only about 10%-20% of the equilibrium population within the reactant ensemble. Therefore, there is more flux through the pathway associated with state 2 in Figure 10C in spite of the higher associated free energy barrier. We also note that our network model does not recapitulate the total flux even though we have identified the major transition paths on the free energy surface. We postulate that as more pathways are added or the temperature is lowered we will account for the majority of the flux, though there is a maximum number of pathways that could be included in our model as a result of the pruning procedure.

These results show that even in a simple system one could miss important contributions to observed behavior if only a single pathway is used, even if that pathway is the MFEP. In more complex systems with rougher underlying free energy landscapes it will be less clear where the MFEP will lie *a priori* and how that pathway contributes to the overall flux. In such cases, our method will be as good, or superior to, single pathway methods at capturing the overall reactant to product flux.

Chapter 5 FUTURE DIRECTIONS:

Given the number of analyses performed here and the end goal of the method development discussed in the introduction, we break down the future directions into two sections. First, we discuss further theoretical advances that could be introduced to make our method more robust and transferable across systems, and second, we discuss the possible applications of this method to study biological systems.

5.1 Future Method Development.

The examples herein show that the results obtained from a Markov model are dependent on the parameters used to construct the model, consistent with the results of other studies. We also find that there appears to be optimal ranges for the parameters used to construct the models, or, in the most pessimistic outlook, that there is an optimal range where the results will be minimally affected by changing a parameter during construction. While in this study the same cutoff was applied uniformly to all states in the model, we speculate that each individual state has an optimal definition. The next logical step is to construct a model by optimizing the entire set of parameters including number of states and state definitions. In practice, one could apply a simulated annealing approach by constructing a model with one randomly varied parameter and iterating to convergence of all the microscopic transition rates. Alternatively, if one has additional information about the system, either from experiment or from other computational studies, this information could be added as an additional constraint.

Adopting such an optimization procedure would greatly improve transferability on model construction across systems, and prevent one from drawing different conclusions from models constructed with different parameters. Practical implementation of the optimization procedure is feasible using the tools developed during this work, though further optimization of the underlying codes, especially in the parallelization routines, would be warranted. Such an advance would be a valuable contribution to the field of Markov modeling, though in practice this would be difficult, if not impossible, unless some metric can be found to optimize to. Markov state selection is an ongoing field of study in itself, and it is still unclear how one should choose the boundary of the Markov state after identification, though several procedures have been suggested.

It would also be ideal to implement a robust milestoning approach to predict transition rates from the FTSM and US procedure. The milestoning rate matrix constructed along the reaction coordinate using umbrella sampling data could be unbiased with the Dynamic Histogram Analysis Method or one of its variations.[97,98] This would alleviate inaccuracies in the rate extraction procedure resulting from the shape of the profile.

Another significant issue that one will face in more complex systems is the slow convergence of FTSM during the minimization procedure. An approach similar to Adaptive Finite Temperature String,[80] which optimizes the parallel and perpendicular force constants to match umbrella windows during minimization, could be adopted. Such a procedure would both eliminate the need for post-minimization US simulations and provide faster convergence to the minimized pathway by adopting force constants optimal for the local shape of the surface, which may change greatly over the course of the entire minimization process.

5.2 Future Applications to Biomolecular Systems.

Given the issues above, any applications discussed here will necessarily occur concurrently to or after the issues with the methods used in our workflow are addressed. We anticipate that our method will be most readily applicable to the study of conformational dynamics in proteins and other biomolecules, where the timescales of the transition are beyond what is feasible from direct sampling. One such system that could serve as a test case is adenylate kinase (AK), which displays a prominent loop opening and closing on the μs-ms timescale thought to affect ligand binding and has been studied using MSM techniques in the past.[15] As verification of our method in a larger, solvated system one could connect the ensemble of open and closed conformations using COS and predict the rate of loop and opening and closing, though experimental estimations of the rate vary.

Our ultimate goal in development of this method is to apply it to assess the ensemble of possible reaction pathways in enzymes and evaluate whether conformational motion can play a significant role in modulating the rate. A specific example of a system where our method could be used is the enzyme lactate dehydrogenase, which also displays a loop motion that has been thought to modulate the activity of the enzyme.[99–101] In an experimental study, the carbohydrate Trehalose was introduced to the system to increase the viscosity of the solvent and it was shown that the catalytic rate was slowed in the presence of high Trehalose concentrations.[102] This would work as ideal test case for our model because the differences in reaction rates can be attributed to changes in the MM environment rather than the QM environment needed to simulate chemical reactions with the COS simulations. However most importantly, the study of Hernandez-Mera and Sampedro[102] provides a set of experiment data at a series of temperature and solvent viscosities to compare the predictions of our method to, the ultimate measure of the overall usefulness of the method developed herein.

Acknowledgments

Funding for the work done in this thesis was provided by the National Institutes of Health, project number 5R35GM124818-02 and the National Science Foundation, award number 1553291. The findings and conclusions here do not necessarily reflect the view of the above funding agencies.

APPENDIX A: Parameters Used in Simulations and Model Construction

Table A1: Finite-Temperature String parameters for proline dipeptide

State 1	State 2	Initial Guess	k_{par} (kcal/mol*deg^2)	k_{prp} (kcal/mol*deg^2)	d_{prp} (A)	Evolution frequency	Update frequency
0	2	I	10000	10000	0.1	100	1000
0	2	II	1000	1000	0.2	100	1000
0	2	III	10000	10000	0.1	100	1000
0	2	IV	1000	1000	0.2	100	1000
0	3	I	1000	1000	0.2	100	1000
0	3	II	1000	1000	0.2	100	1000
0	3	III	1000	1000	0.2	100	1000
0	3	IV	1000	1000	0.2	100	1000
1	2	I	5000	5000	0.2	100	1000
1	2	II	5000	5000	0.2	100	1000
1	2	III	5000	5000	0.2	100	1000
1	2	IV	5000	5000	0.2	100	1000
1	3	I	5000	5000	0.2	100	1000
1	3	II	5000	5000	0.2	100	1000
1	3	III	5000	5000	0.2	100	1000
1	3	IV	5000	5000	0.2	100	1000

Table A2: Master equation model parameters for glycine dipeptide:

Temperature (K)	Reactant States	Product States	Energy Cutoff (k_BT)
250	2	1	0.3
275	2	1	0.3
300	2	1	0.3
325	2	1	0.3
350	2	1	0.3

Table A3: Master equation model parameters for proline dipeptide:

Temperature (K)	Reactant States	Product States	Energy Cutoff (k_BT)
600	2	2	0.35
650	2	2	0.35
700	2	2	0.35
750	2	2	0.35
800	2	2	0.35

Table A4: Master equation model parameters for alanine dipeptide:

Temperature (K)	Reactant States	Product States	Energy Cutoff (k_BT)
400	2	1	0.45
450	2	1	0.45
500	3	1	0.45
550	3	1	0.45
600	3	2	0.45

APPENDIX B: Supplemental figures

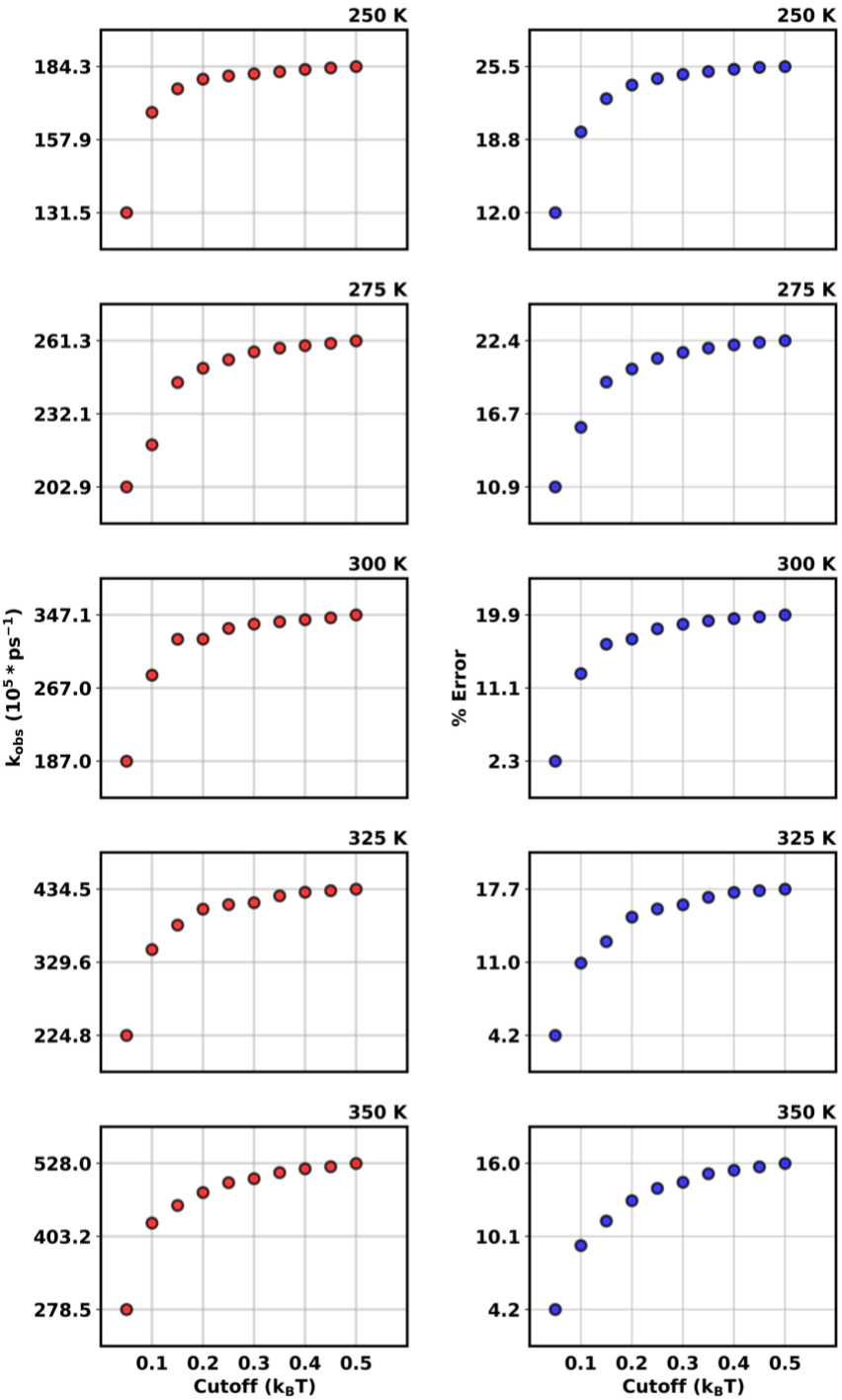

Figure B1: Glycine dipeptide prediction transtion rates, all cutoffs. Predicted transition rate plotted as a function of energy cutoff at all temperatures.

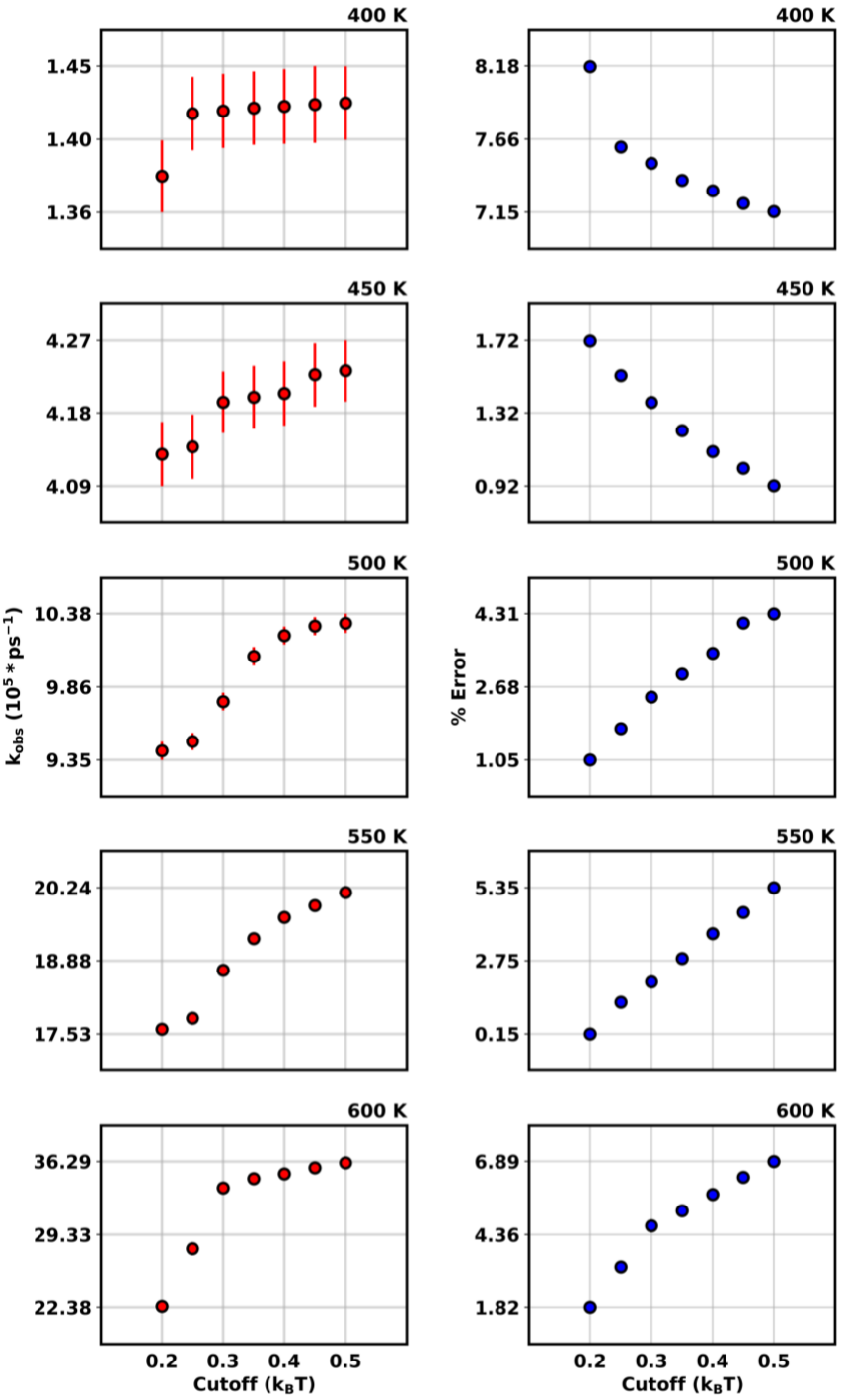

Figure B2: Alanine dipeptide prediction transtion rates, all cutoffs. Predicted transition rate plotted as a function of energy cutoff at all temperatures.

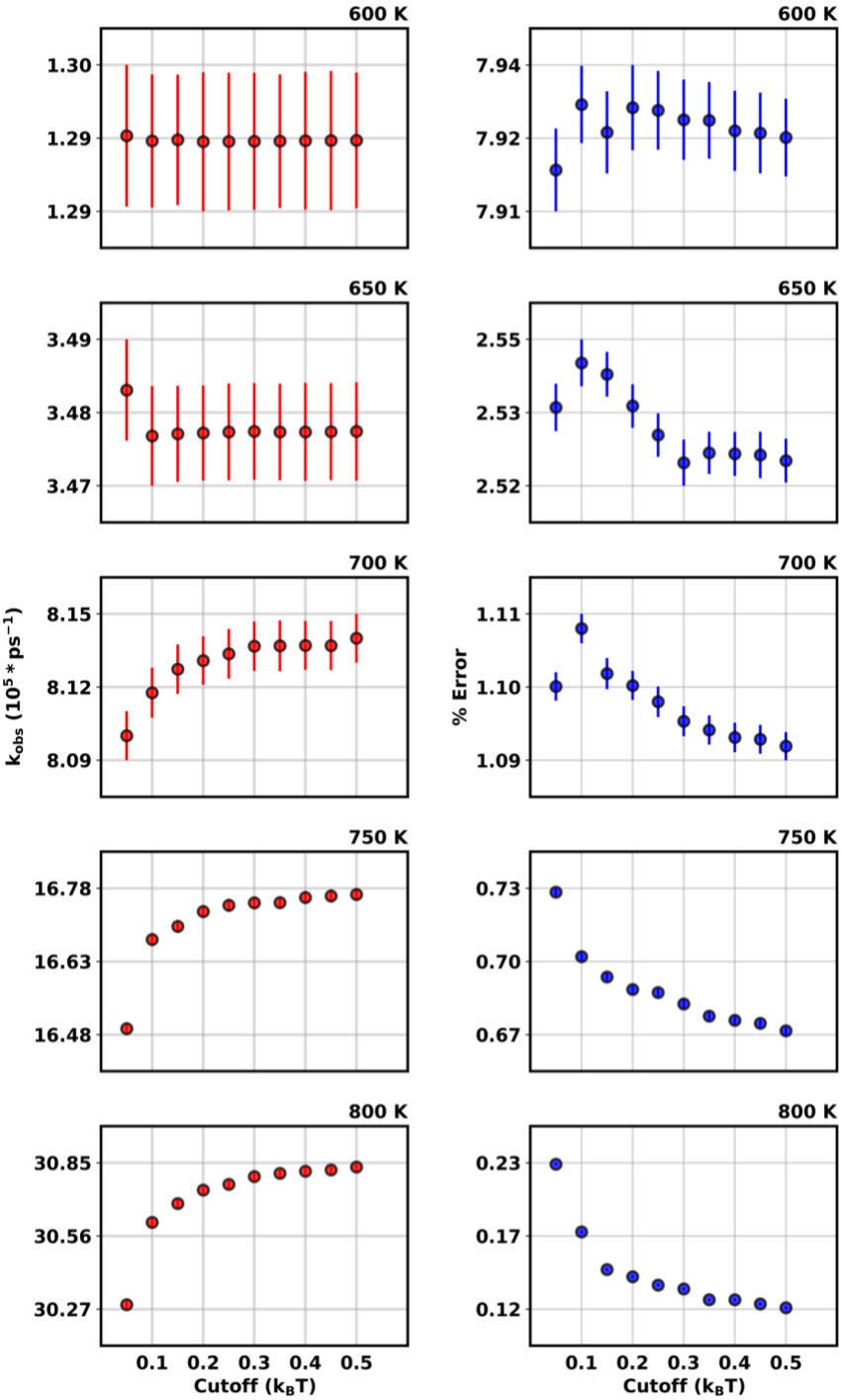

Figure B3: Proline dipeptide prediction transtion rates, all cutoffs. Predicted transition rate plotted as a function of energy cutoff at all temperatures.

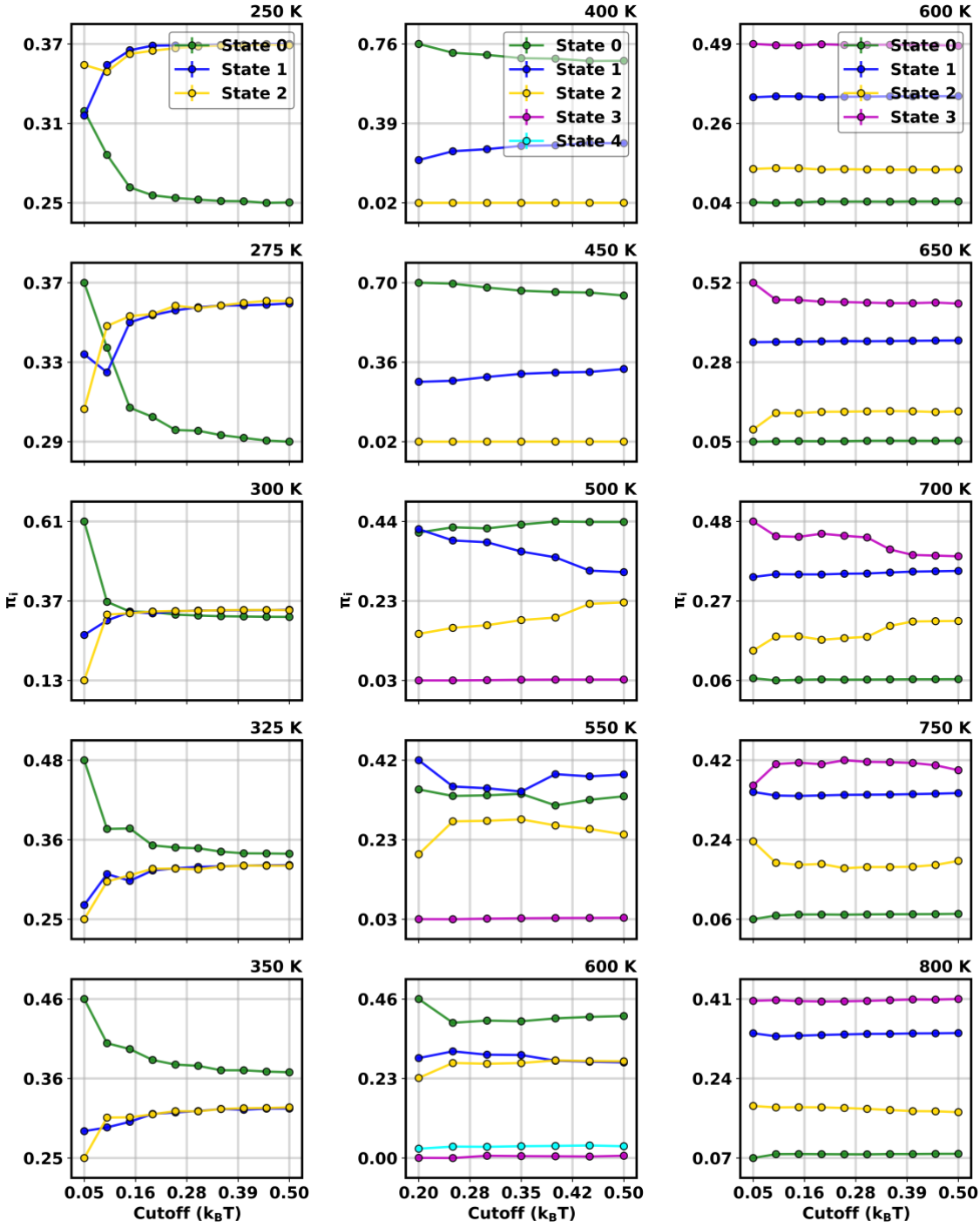

Figure B4: Stationary probabilities as a function of energy cutoff. Stationary probability for each state as a function of cutoff at all temperatures. First column) glycine, second column) alanine, and third column) proline dipeotides.

BIBLIOGRAPHY

(1) Chandler, P. G. B. and C. D. and P. L. G. and D. *J. Phys. Condens. Matter* **2000**, *12* (8A), A147.

(2) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. *J. Chem. Phys.* **1998**, *108* (5), 1964–1977.

(3) Dellago, C.; Bolhuis, P. G.; Chandler, D. *J. Chem. Phys.* **1999**, *110* (14), 6617–6625.

(4) Bergonzo, C.; Simmerling, C. *Annu. Rep. Comput. Chem.* **2011**, *7*, 89–97.

(5) Tao, P.; Hodošček, M.; Larkin, J. D.; Shao, Y.; Brooks, B. R. *J. Chem. Theory Comput.* **2012**, *8* (12), 5035–5051.

(6) Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods* **2010**, *52* (1), 99–105.

(7) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121* (1), 415–425.

(8) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134* (17), 174105.

(9) Ryde, U. In *Methods in Enzymology*; 2016; Vol. 577, pp 119–158.

(10) Janoš, P.; Trnka, T.; Kozmon, S.; Tvaroška, I.; Koča, J. *J. Chem. Theory Comput.* **2016**, *12* (12), 6062–6076.

(11) Plattner, N.; Doerr, S.; De Fabritiis, G.; Noé, F. *Nat Chem* **2017**, *advance on*.

(12) Meng, Y.; Shukla, D.; Pande, V. S.; Roux, B. *Proc. Natl. Acad. Sci.* **2016**, *113* (33), 9193–9198.

(13) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci.* **2009**, *106* (45), 19011–19016.

(14) Gao, K.; Zhao, Y. *J. Phys. Chem. B* **2017**, *121* (14), 2952–2960.

(15) Zheng, Y.; Cui, Q. *J. Chem. Theory Comput.* **2018**, *14* (3), 1716–1726.

(16) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. *J. Chem. Phys.* **2013**, *139* (1), 15102.

(17) Sultan, M. M.; Pande, V. S. *J. Chem. Phys.* **2018**, *149* (9), 94106.

(18) Hovan, L.; Comitani, F.; Gervasio, F. L. *J. Chem. Theory Comput.* **2019**, *15* (1), 25–32.

(19) Song, H. D.; Zhu, F. *J. Phys. Chem. B* **2017**, *121* (15), 3376–3386.

(20) Ovchinnikov, V.; Karplus, M. *J. Chem. Phys.* **2014**, *140* (17).

(21) Fang, D.; Ito, S.; Okamoto, Y.; Ovchinnikov, V.; Cui, Q. *Mol. Simul.* **2016**, *42* (13), 1056–1078.

(22) Senn, H. M.; Thiel, W. *Angew. Chemie Int. Ed.* **2009**, *48* (7), 1198–1229.

(23) van der Kamp, M. W.; Mulholland, A. J. *Biochemistry* **2013**, *52* (16), 2708–2728.

(24) Truhlar, D. G. *Arch. Biochem. Biophys.* **2015**, *582*, 10–17.

(25) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. *Science (80).* **2004**, *303* (5655), 186 195.

(26) Hummer, G. *New J. Phys.* **2005**, *7* (1), 34.

(27) Best, R. B.; Hummer, G. *Phys. Rev. Lett.* **2006**, *96* (22), 228104.

(28) Zheng, W.; Best, R. B. *J. Phys. Chem. B* **2015**, *119* (49), 15247–15255.

(29) Peters, B. Peters, B. B. T.-R. R. T. and R. E. S., Ed.; Elsevier: Amsterdam, 2017; pp 435–450.

(30) Hänggi, P.; Talkner, P.; Borkovec, M. *Rev. Mod. Phys.* **1990**, *62* (2), 251–341.

(31) Kramers, H. A. *Physica* **1940**, *7* (4), 284–304.

(32) Grote, R. F.; Hynes, J. T. *J. Chem. Phys.* **1980**, *73* (6), 2715–2732.

(33) Peters, B. Peters, B. B. T.-R. R. T. and R. E. S., Ed.; Elsevier: Amsterdam, 2017; pp 451–471.

(34) Peters, B. Peters, B. B. T.-R. R. T. and R. E. S., Ed.; Elsevier: Amsterdam, 2017; pp 1–17.

(35) Kohen, A.; Cannio, R.; Bartolucci, S.; Klinman, J. P.; Klinman, J. P. *Nature* **1999**, *399* (6735), 496–499.

(36) Ceccarelli, C.; Liang, Z. X.; Strickler, M.; Prehna, G.; Goldstein, B. M.; Klinman, J. P.; Bahnson, B. J. *Biochemistry* **2004**, *43* (18), 5266–5277.

(37) Agarwal, P. K.; Billeter, S. R.; Rajagopalan, P. T. R.; Benkovic, S. J.; Hammes-Schiffer, S. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (5), 2794–2799.

(38) Hammes-Schiffer, S.; Benkovic, S. J. *Annu. Rev. Biochem.* **2006**, *75* (1), 519–541.

(39) Warshel, A.; Bora, R. P. *J. Chem. Phys.* **2016**, *144* (18), 180901.

(40) Nagel, Z. D.; Klinman, J. P. *Nat Chem Biol* **2009**, *5* (8), 543–550.

(41) Schwartz, S. D.; Schramm, V. L. *Nat Chem Biol* **2009**, *5* (8), 551–558.

(42) Schramm, V. L.; Schwartz, S. D. *Biochemistry* **2018**, *57* (24), 3299–3308.

(43) Brooks, B. R.; Brooks 3rd, C. L.; Mackerell Jr, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614.

(44) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science (80).* **2011**, *334* (6055), 517 520.

(45) Piana, S.; Shaw, D. E. *J. Phys. Chem. B* **2018**, *122* (49), 11440–11449.

(46) Chekmarev, D. S.; Ishida, T.; Levy, R. M. *J. Phys. Chem. B* **2004**, *108* (50), 19487–19495.

(47) Ermak, D. L.; McCammon, J. A. *J. Chem. Phys.* **1978**, *69* (4), 1352–1360.

(48) Peters, B. Peters, B. B. T.-R. R. T. and R. E. S., Ed.; Elsevier: Amsterdam, 2017; pp 363–401.

(49) Buchete, N.-V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112* (19), 6057–6069.

(50) Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *130* (19), 194101.

(51) Vanden-Eijnden, E.; Venturoli, M.; Ciccotti, G.; Elber, R. *J. Chem. Phys.* **2008**, *129* (17), 174102.

(52) Berezhkovskii, A. M.; Szabo, A. *J. Chem. Phys.* **2019**, *150* (5), 54106.

(53) Caniparoli, L.; O'Brien, E. P. *J. Chem. Phys.* **2015**, *142* (14), 145102.

(54) Ninio, J. *Proc. Natl. Acad. Sci. U. S. A.* **1987**, *84* (3), 663–667.

(55) Bowman, G. R. Bowman, G. R., Pande, V. S., Noé, F., Eds.; Springer Netherlands: Dordrecht, 2014; pp 7–22.

(56) Park, H.-S.; Jun, C.-H. *Expert Syst. Appl.* **2009**, *36* (2, Part 2), 3336–3341.

(57) Deuflhard, P.; Weber, M. *Linear Algebra Appl.* **2005**, *398*, 161–184.

(58) Röblitz, S.; Weber, M. *Adv. Data Anal. Classif.* **2013**, *7* (2), 147–179.

(59) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. *J. Chem. Theory Comput.* **2011**, *7* (10), 3412–3419.

(60) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. *J. Chem. Theory Comput.* **2015**, *11* (11),

5525–5542.

(61) Bowman, G. R. *J. Chem. Phys.* **2012**, *137* (13), 134111.

(62) Sharma, A. K.; Bukau, B.; O'Brien, E. P. *J. Am. Chem. Soc.* **2016**, *138* (4), 1180–1195.

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

(64) Jónsson, H.; Mills, G.; Jacobsen, K. W. In *Classical and Quantum Dynamics in Condensed Phase Simulations*; 1998; pp 385–404.

(65) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. *J. Chem. Phys.* **2000**, *113* (22), 9901–9904.

(66) Maragakis, P.; Andreev, S. A.; Brumer, Y.; Reichman, D. R.; Kaxiras, E. *J. Chem. Phys.* **2002**, *117* (10), 4651–4658.

(67) Herbol, H. C.; Stevenson, J.; Clancy, P. *J. Chem. Theory Comput.* **2017**, *13* (7), 3250–3259.

(68) E, W.; Ren, W.; Vanden-Eijnden, E. *Phys. Rev. B* **2002**, *66* (5), 052301.

(69) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. *J. Chem. Phys.* **2004**, *120* (17), 7877–7886.

(70) Zimmerman, P. M. *J. Chem. Phys.* **2013**, *138* (18), 184102.

(71) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. *J. Chem. Phys.* **2006**, *125* (2), 24106.

(72) Lee Woodcock, H.; Hodošček, M.; Sherwood, P.; Lee, Y. S.; Schaefer III, H. F.; Brooks, B. R. *Theor. Chem. Acc.* **2003**, *109* (3), 140–148.

(73) E, W.; Ren, W.; Vanden-Eijnden, E. *J. Phys. Chem. B* **2005**, *109* (14), 6688–6693.

(74) Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *130* (19).

(75) Ren, W.; Vanden-Eijnden, E.; Maragakis, P.; E, W. *J. Chem. Phys.* **2005**, *123* (13), 134109.

(76) Ovchinnikov, V.; Karplus, M.; Vanden-Eijnden, E. *J. Chem. Phys.* **2011**, *134* (8), 85103.

(77) E., W.; Vanden-Eijnden, E. *J. Stat. Phys.* **2006**, *123* (3), 503.

(78) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. *Multiscale Model. Simul.* **2009**, *7* (3), 1192–1219.

(79) Zhu, F.; Hummer, G. *Proc. Natl. Acad. Sci.* **2010**, *107* (46), 19814 LP-19819.

(80) Zinovjev, K.; Tuñón, I. *J. Phys. Chem. A* **2017**, *121* (51), 9764–9772.

(81) Ovchinnikov, V.; Nam, K.; Karplus, M. *J. Phys. Chem. B* **2016**, *120* (33), 8457–8472.

(82) Zhu, F.; Hummer, G. *J. Comput. Chem.* **2012**, *33* (4), 453–465.

(83) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13* (8), 1011–1021.

(84) Peters, B. Peters, B. B. T.-R. R. T. and R. E. S., Ed.; Elsevier: Amsterdam, 2017; pp 227–271.

(85) Peters, B. *J. Phys. Chem. B* **2015**, *119* (21), 6349–6356.

(86) Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. *J. Chem. Phys.* **1996**, *104* (15), 5860–5868.

(87) Zhu, F.; Hummer, G. *J. Chem. Theory Comput.* **2012**, *8* (10), 3759–3768.

(88) Satija, R.; Makarov, D. E. *J. Phys. Chem. A* **2019**.

(89) Gillespie, D. T. *J. Comput. Phys.* **1976**, *22* (4), 403–434.

(90) Gillespie, D. T. *J. Phys. Chem.* **1977**, *81* (25), 2340–2361.

(91) Fischer, S.; Dunbrack, R. L.; Karplus, M. *J. Am. Chem. Soc.* **1994**, *116* (26), 11931–11937.

(92) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. *J. Comput. Chem.* **2011**, *32* (10), 2319–2327.

(93) Head-Gordon, T.; Head-Gordon, M.; Frisch, M. J.; Brooks, C. L.; Pople, J. A. *J. Am. Chem. Soc.* **1991**, *113* (16), 5989–5997.

(94) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. *Biopolymers* **1992**, *32* (5), 523–535.

(95) Lee, J.; Lee, I.-H.; Joung, I.; Lee, J.; Brooks, B. R. **2017**, *8*, 15443.

(96) Guarnera, E.; Vanden-Eijnden, E. *J. Chem. Phys.* **2016**, *145* (2), 24102.

(97) Rosta, E.; Hummer, G. *J. Chem. Theory Comput.* **2015**, *11* (1), 276–285.

(98) Stelzl, L. S.; Kells, A.; Rosta, E.; Hummer, G. *J. Chem. Theory Comput.* **2017**, *13* (12), 6328–6342.

(99) Pan, X.; Schwartz, S. D. *J. Phys. Chem. B* **2015**, *119* (17), 5430–5436.

(100) Pineda, J. R. E. T.; Antoniou, D.; Schwartz, S. D. *J. Phys. Chem. B* **2010**, *114* (48), 15985–15990.

(101) Reddish, M. J.; Peng, H.-L.; Deng, H.; Panwar, K. S.; Callender, R.; Dyer, R. B. *J. Phys. Chem. B* **2014**, *118* (37), 10854–10862.

(102) Hernández-Meza, J. M.; Sampedro, J. G. *J. Phys. Chem. B* **2018**, *122* (15), 4309–4317.