

DeepWEST: Deep learning of kinetic models with the Weighted Ensemble Simulation Toolkit for enhanced sampling

Anupam Anand Ojha,[†] Saumya Thakur,[‡] Surl-Hee Ahn,[¶] and Rommie E.
Amaro^{*,†}

[†]*Department of Chemistry, University of California San Diego, La Jolla, CA, USA*

[‡]*Department of Chemistry, Indian Institute of Technology Bombay, Mumbai, Maharashtra,
India*

[¶]*Department of Chemical Engineering, University of California Davis, Davis, CA, USA*

E-mail: ramaro@ucsd.edu

Abstract

Recent advances in computational power and algorithms have made molecular dynamics (MD) simulations reach greater timescales. However, for observing conformational transitions associated with biomolecular processes, MD simulations still have limitations. Several enhanced sampling techniques seek to address this challenge, including the weighted ensemble (WE) method, which samples transition between metastable states using many weighted trajectories to estimate kinetic rate constants. However, initial sampling of the potential energy surface has a significant impact on the performance of WE, i.e., convergence and efficiency. We thereby introduce deep-learned kinetic modeling approaches that extract statistically relevant information from short MD trajectories to provide a well-sampled initial state distribution for WE simulation. This hybrid approach overcomes any statistical bias to the system as it runs short

unbiased MD trajectories and identifies meaningful metastable states of the system. It is shown to provide a more refined free energy landscape closer to steady-state that could efficiently sample kinetic properties such as rate constants.

1 Introduction

Molecular dynamics (MD) simulations have found their applications in science and engineering, such as chemistry and biochemistry, statistical mechanics, condensed matter physics, and material science.¹⁻¹⁰ In recent years, MD simulations have significantly impacted in studying complex biological processes such as protein folding, drug discovery, receptor-ligand binding and unbinding, protein-membrane interactions, and protein-protein interactions. MD simulations can effectively analyze key mechanistic insights into highly complex dynamics of biological systems of interest in atomistic detail.¹¹⁻¹⁸ However, the task of estimating the kinetics and thermodynamics of such systems comes with its own set of challenges.

Existing classical force fields are sometimes insufficient to estimate specific properties of interest, such as polarization and charge delocalization effects for complex biological systems.¹⁹⁻²¹ Timesteps for MD simulations are restricted within the femtosecond range to correctly integrate the equations of motion as they cannot exceed the time period of highest-frequency thermal oscillation. Biologically relevant systems undergo complex conformational transitions, which are essential to their functions. It is challenging to capture these transitions through conventional MD, especially if they are relatively slow processes (milliseconds or longer), involve large-scale conformational rearrangements, or include protein association and/or (re)folding. We resort to computationally expensive long-scale MD simulations to observe such transitions or “rare events”.²²⁻²⁴

Several enhanced sampling methods have been developed to overcome the difference in timescales of conventional MD simulations compared to the timescales of biological processes.²⁵⁻²⁹ One category of enhanced sampling methods adds a bias potential to the potential energy surface (PES) that decreases the energy barrier of transition between metastable

states and accelerates the conformational search. We could roughly categorize these methods as collective variable (CV)-based and CV-free enhanced sampling methods. CV-based enhanced sampling approaches include and are not limited to metadynamics (metaD),^{30,31} variationally enhanced sampling,³² and Markov state models (MSMs)^{33,34} while CV-free enhanced sampling methods include parallel tempering or replica exchange molecular dynamics (REMD),^{35,36} selective integrated tempering, and Gaussian accelerated molecular dynamics (GaMD).³⁷ All the previously mentioned enhanced sampling approaches are capable of accelerated PES conformational search. Thereby, they prove to be effective in extensive thermodynamic sampling. On the contrary, such enhanced sampling approaches add a bias potential to the system’s potential energy and thereby lead to altered dynamics. Extracting kinetic and mechanistic insights is often subjected to assumptions such as low residence times in the transition states regions, quasistationarity characteristics of metaD, construction of a master equation for non-Arrhenius and multistate kinetics, and usage of Kramers’ rate theory in the overdamped regime.^{38–42}

On the other hand, several path sampling methods exist for an extensive sampling of kinetic properties, which are broadly divided into complete path sampling and segment-based sampling. Transition path sampling (TPS) and dynamic importance sampling (DIMS) are based on complete reactant to product path sampling. Segment-based sampling approaches are based on splitting the reactant to product path into several segments or “bins” where bin-to-bin transitions are sampled by running independent MD simulations of fixed duration within these bins. These methods include and are not limited to weighted ensemble (WE) methodology,⁴³ milestoning approaches,^{44–46} adaptive multilevel splitting (AMS),^{47–49} and transition interface sampling (TIS).⁵⁰ These path sampling methods typically focus on sampling the transition regions, and thereby, an entire PES of the system is often neglected. The WE method enhances the sampling of rare events by running independent and unbiased parallel simulations in short configurational spaces or “bins”. These simulations communicate through each other through replication and resampling, leading to the precise estimation of

kinetic observables. The WE method is further explained in the section 2.2.

Our recent work demonstrated the effectiveness of a hybrid enhanced sampling method, the GaMD-WE method, which combines GaMD and WE for an extensive sampling of both the thermodynamics and kinetics of biological systems of interest.⁵¹ GaMD is performed initially to sample the PES of the system by applying several boost potentials and is followed by reweighing to recover the original PES. Configurations selected from the recovered PES are the starting structures for the WE simulations. Several independent boost potentials are applied in parallel GaMD simulations and are reweighted accordingly to recover the PES of the system. The GaMD run with the most PES coverage provides the starting structures for WE simulations. We aim to further decrease the computational cost of thermodynamic and kinetic sampling of systems of interest by running unbiased MD simulations and implementing deep learning with the Markovian variational approach for faster and enhanced hybrid sampling.^{52,53} This way, we can bypass the entire reweighing process to recover the original PES.

Providing a well-sampled initial state distribution to WE simulations is often a concern. Generally, multiple starting structures are provided at the beginning of a simulation, and they often fail to establish extensive communication with each other, thereby leading to an initial bias in the simulation. Markov State Models (MSMs) have shown to be successful in generating equilibrium distribution and thereby removing the initial bias in the simulations.⁵⁴ The history-augmented Markov state model (haMSM) is another efficient method for the removal of bias in stationary distributions, and it has been implemented recently in WE simulations, where a steady state analysis is performed, and the weights are distributed accordingly to restart the simulation.⁵⁵ In short, DeepWEST is an attempt to construct MSMs from short MD simulations, providing a well-sampled initial distribution for the WE simulations. We propose a hybrid method that uses the variational approach for Markov principle (VAMP) and neural networks to identify metastable states from unbiased and short MD simulations as a precursor for running WE simulations. Selected conformations

from these states then serve as starting structures for WE simulations to further sample the kinetics and thermodynamics of the system. This hybrid methodology further reduces the computational cost as compared to our previously developed hybrid GaMD-WE method, accelerates the WE approach further by providing well-sampled initial configurations, and provides a better comprehensive picture of the thermodynamic and kinetic properties of the systems of interest by introducing no statistical bias to the free energy landscape to the system. In the forthcoming sections, we will describe the WE method, the VAMP approach, and the hybrid method in detail. We will also demonstrate the capability of the DeepWEST approach to sample kinetic and thermodynamic properties faster than the WE method alone and compare the findings with the already established GaMD-WE approach.

2 Methods

2.1 Variational Approaches for Markovian Processes

Markovian processes are stochastic processes where the future state of the system, $\mathbf{x}_{t+\tau}$ depends only on the current state, \mathbf{x}_t . Here, t is the time step, and τ is the lag time. Various Markov modeling approaches have been developed recently to extract key information for complex dynamical processes. These methods include and are not limited to Markov state models (MSMs),⁵⁶⁻⁵⁸ Markov transition models, variation approach to conformational dynamics (VAC),⁵⁹ time-lagged independent component analysis (TICA),⁶⁰ variational diffusion maps,^{61,62} and variational approach for Markov processes (VAMP).^{52,63} Dynamical systems often display high non-linearity in their system coordinates. To analyze non-linear and high-dimensional dynamical systems, we employ the Koopman operator, \mathbb{K} , that linearly transforms the vector space spanned by observables in the form of time-series data generated by MD simulations. This facilitates the prediction and estimation of nonlinear dynamical properties through the traditional methods employed for linear dynamical systems.⁶⁴⁻⁶⁶ A majority of the Markovian modeling approaches exploit the fact that there exists a non-

linear transformation of these features such that the dynamics can be approximated as a linear Markov model.^{67,68} Let χ_0 and χ_1 be the feature transformations for the trajectory coordinates \mathbf{x}_t and $\mathbf{x}_{t+\tau}$ respectively, and \mathbb{E} be the expectation value such that the matrix \mathbb{K} determines the dynamics of the system according to equation 1. According to the VAMP theory, when the subspaces spanned by the features, χ_0 and χ_1 , are identical to the top left and top right singular functions of \mathbb{K} , we obtain the best finite-dimensional linear model.

$$\mathbb{E}[\chi_1(x_{t+\tau})] \approx \mathbb{K}^T \mathbb{E}[\chi_0(x_t)] \quad (1)$$

To generate well-sampled initial conformations for WE simulations, we resort to a neural network architecture that employs the variational approach for Markov processes (VAMP), known as VAMPnets, which is a non-biased trajectory learning approach towards faster estimation of kinetics and thermodynamics of systems of interest as compared to other hybrid approaches.⁶⁹ MD data analysis is primarily performed in subsequent steps of featurization, dimension reduction, discretization, and coarse-graining.⁷⁰⁻⁷³ Featurization is the process where the simulation data is often subjected to the removal of translational and rotational motion and/or transformed into internal coordinates. Featurization follows a dimensionality reduction where a high dimensional trajectory data is reduced to slow collective variables.⁷⁴⁻⁷⁶ However, initial steps of featurization, such as choosing appropriate metrics for training the trajectory dataset such as using the cartesian coordinates or internal coordinates, selecting suitable CVs that describe the conformation space such as dihedral angles RMSD, the radius of gyration, removal of solvent coordinates, selection of heavy atoms, and realignment are to be determined before training the trajectory using VAMPnets. The resultant metric space is then discretized to fewer states, and the process is called discretization.^{71,77,78} Finally, a coarse-graining of the Markov state model (MSM) is performed since the internal dynamics of a particular set of microstates can be faster than the modeling timescales.^{79,80} All of the above steps involved in the process of generating MSMs are performed by VAMPNets that learn the non-linear collective variables or reaction coordinates

that separate the metastable states of biological systems.

A simple VAMPnets model comprises two parallel neural network architectures that are employed to learn feature transformations using the VAMP approach. Each network receives the initial MD trajectory coordinates and time-lagged correspondent of the same coordinates, \mathbf{x}_t and $\mathbf{x}_{t+\tau}$ respectively. Non-linear dimensionality reduction is then performed on these two sets of trajectory inputs for each time step, t . As per the VAMP principle, a VAMP-2 score is defined that attains its maximum value when the top left and right components of the Koopman operator \mathbb{K} are equivalent to the subspaces spanned by these features.⁵² Deployed neural networks are trained to maximize the VAMP-2 score in order to achieve the best finite-dimensional linear model. In the above process, it achieves the segregation of the trajectory frames and assigns them to particular clusters or Markov states, which then accelerates the process of rare-event sampling and transition state analysis. Key features for VAMPnets include the choice of the lag time, the number of output nodes, and the network depth of the architecture. Few pre-optimization runs with varying lag times were performed on the simulation trajectory to find the lag time that could resolve the slowest processes. The choice of a lag time relies upon the eigenvalue decomposition of the Markov propagators. When selecting a large lag time that exceeds the timescale of the slowest processes, it becomes difficult to fit the noisy data. However, a short lag time leads to them getting stuck in one of the suboptimal maxima of the training score. The selection of output nodes represents the separable metastable regions from the trajectory data. A higher number of output nodes may not be suitable for small trajectory data since the clustering will lead to discretizing the transition regions. The clustering of metastable states also depends heavily on the network depth of the architecture. A deeper network describes complex functions and is more difficult to train.

2.2 Weighted ensemble method

Standard unbiased MD simulations are limited by the timescale and infrequency of significant biological events. Such events are rarely captured in the trajectory data, and it becomes challenging to analyze further and estimate the kinetics of such transitions. The weighted ensemble simulation approach is an enhanced sampling method designed to sample such rare occurrences or transitions such as protein-protein association reactions and conformational changes by replicating and resampling an ensemble of weighted trajectories.^{43,81} The whole configurational space is subdivided into macrostates or “bins” that sequentially lead the system from an initial state to the target state. Many short simulations or “walkers” carrying probabilities or “weights” are run and the system evolves throughout the simulation through “resampling” that ensures an equal number of trajectories in each bin, i.e., by splitting or merging walkers. These walkers explore the configurational space extensively, eventually sampling rare transitions. Appropriate progress coordinates are used to define the bin boundaries. Transitions between these bins are recorded while the system evolves in time to estimate the kinetics and thermodynamics. The WE method has been demonstrated to accurately estimate the kinetics of rare events for multiple biological systems.^{82,83}

Resampling is a crucial component of the WE approach that leads to the statistically unbiased future evolution of the system.^{82,84,85} Walkers are resampled after every iteration via appropriate replication and reassignment of weights. Weighted trajectories are initiated from assigned bins and are propagated for a short interval, τ . These trajectories are either replicated where there are too few trajectories or deleted where there are more than the required number of trajectories per bin. Since there is no statistical bias in the simulation, thermodynamic and kinetic properties can be directly obtained by the evolution of the weights of the walkers in each bin. Simulation results can be periodically checked for convergence by estimating rate constants for various possible transitions between the macrostates. This approach also eliminates the need for choosing the resampling time, τ based on the Markovian property and thereby provides the flexibility to select τ based on the system size.

The efficiency of WE simulations is attributed to the weights of the starting conformations, and long-scale simulations of multiple short trajectories are required to obtain accurate kinetic and thermodynamic properties of systems of interest. Choice of initial starting conformations significantly affects the accuracy and simulation time, and it can be remarkably enhanced by providing an intelligently sampled set of initial starting conformations. The motivation for DeepWEST originates from this notion and aims at obtaining these conformations through deep learning approaches using VAMP. The Weighted Ensemble Simulation Toolkit with Parallelization and Analysis (WESTPA) is an open-source package to perform the WE simulations.⁸⁶ This toolkit is highly interoperable, compatible with a wide range of MD engines, and provides an integrated protocol for efficient storage and analysis of the estimates generated. Our current work incorporates the WESTPA package for running WE simulations. Since we are focused on obtaining the rate constants between multiple states in these systems, we run equilibrium WE simulations to sample their kinetic and thermodynamic properties. However, steady-state WE simulations can be employed for systems where rate constants are calculated between two states. Such simulations employ “target state recycling”, i.e., the walkers are fed into the initial state once they reach the target state. Steady-state WE simulations with starting states from equilibrium WE are more commonly used for sampling kinetic and thermodynamic properties.

2.3 DeepWEST

The deep learning of kinetic models with the Weighted Ensemble Simulation Toolkit (DeepWEST) method aims to provide well-sampled initial distribution to WE simulations. It is achieved by running a relatively short MD simulation and processing the trajectory data through the VAMP approach by employing neural networks. Initial conformations for WE simulations and their probabilistic values are extracted from the resultant MSMs. WE simulations are then run with this distribution, generating a more refined free energy landscape closer to the steady-state distribution. We have developed a DeepWEST package that au-

tomates the entire process of running MD simulations for trajectory analysis, deep learning of kinetic models for generating MSMs through VAMPNets, and well-sampled initial conformations for running WE simulations. The entire workflow can be described below:

1. Systems of interest are downloaded from the Protein Data Bank (PDB) server and are prepared for MD simulations using appropriate force field parameters, periodic box vectors, and solvation using the Amber 14 package.⁸⁷
2. Once the system is prepared, it is followed by energy minimization, simulated heating, and equilibration using the OpenMM simulation engine.⁸⁸
3. MD simulation is then performed using the Amber simulation engine for the desired amount of simulation time. Trajectories are saved with the desired frequency.
4. The trajectory data is subjected to featurization, dimensionality reduction, discretization, and kinetic modeling using VAMPnets to generate metastable states.
5. VAMPnets deploy a “fuzzy clustering” of MD trajectories in “n” output states. Each conformation in the MD trajectory has a probability for each of these output states that sum to one. The cluster with the maximum likelihood for each conformation is the one that is assigned to the conformation. Post clustering, conformations within each cluster are further binned based on one of the CVs (dihedral angle, ϕ , in the case of alanine dipeptide and RMSD in the case of chignolin). From each subcluster, a certain number of conformations are selected from each bin to generate a well-sampled initial state sampling for WE simulations. Once the conformations are chosen from these clusters, they are assigned the weights based on their fraction of the population of the cluster to which they belong.
6. WE simulation folder is created with all the required starting structures, weights, and topology files to run a WE simulation. To avoid any steric clashes during the initial

WE simulations, starting structures are processed in subsequent steps of minimization of heavy atoms (N, C, O, and C_α) followed by minimization of the entire system.

7. WE simulations are then run with the starting structures with desired probabilities assigned for each conformation.

The Results section shows that our method estimates the kinetics of the systems of interest, i.e., alanine dipeptide and chignolin, more quickly and more accurately than the WE method alone. Long-scale unbiased MD simulations or brute force methods are often used as a reference for evaluating the performances of new methods on model systems.⁸⁹ Five independent brute force simulations each of 2 μ s for alanine dipeptide and 4 μ s for chignolin were performed to obtain the reference values of rate constants over aggregate simulation time. DeepWEST performs equally well and often surpasses the brute force and the hybrid GaMD-WE approach with the advantage of being computationally inexpensive and easy to implement while adding no statistical bias to the system.

3 Results

To demonstrate the effectiveness of the current approach, we have tested the DeepWEST method on three systems, namely alanine dipeptide with explicit solvation (Figure 1a), chignolin with implicit solvation (Figure 1b), and the NTL9 protein with implicit solvation (Figure 1c). Kinetics and thermodynamics obtained from these systems using our DeepWEST approach are compared against the WE and our previously developed hybrid GaMD-WE approach. For the systems mentioned above, we demonstrate that the DeepWEST approach surpasses the WE approach in estimating the rate constants between the metastable states of interest and even outperforms the already established hybrid GaMD-WE approach in many occurrences. We also demonstrate that DeepWEST samples the free energy landscape equivalently compared to the already established approaches. Error bars representing 95% confidence intervals are calculated for all WE, GaMD-WE, and DeepWEST runs. Simula-

tions are run until there is no significant change in the average rate constant or “convergence” is reached. Rate constants are plotted with an approximate interval of 50 iterations for each run.

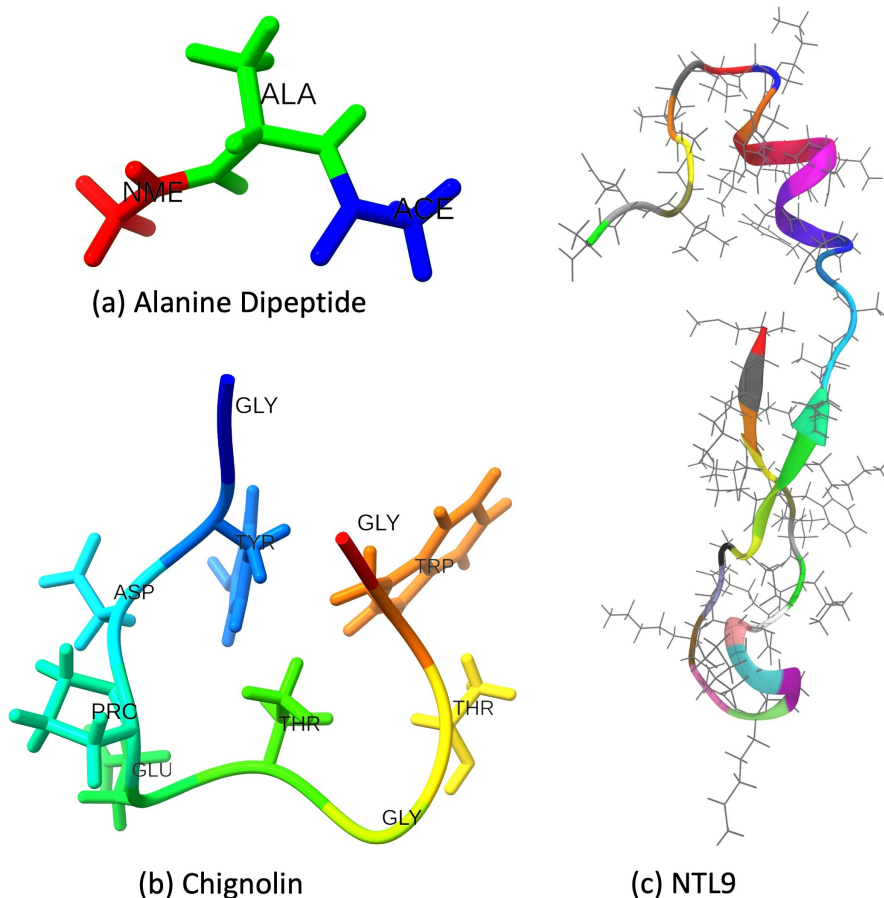


Figure 1: Model systems tested by WE, GaMD-WE, and DeepWEST approaches

3.1 Alanine dipeptide

Alanine dipeptide has been commonly used as a model system for testing new methods.^{13,51,90–92} It is a 22-atom system with an acetyl group at the N-terminus and N-methyl amide at the C-terminus (Figure 1a). Initial coordinates for alanine dipeptide were obtained from <http://ftp.imp.fu-berlin.de/pub/cmb-data/alanine-dipeptide-nowater.pdb>. AMBER ff14SB forcefield was used to prepare the system for MD simulations.⁹³ It was then subjected to explicit solvation using the TIP3P water model.⁹⁴ Alanine dipeptide was subjected to 50 ns

of unbiased MD simulations. Trajectories were trained using VAMP neural networks with a time lag of 80 *ps*, training ratio of 0.9, and a batch size of 1000 that learned discretization in three metastable states. A list of hyperparameters used for network training is provided in Table 1. Starting structures were selected from the metastable states as per the DeepWEST protocol, followed by three independent WE simulation of 12 μ s each with a total simulation time of 36 μ s. Resampling time, τ was kept to be identical as used in WE and GaMD-WE runs, i.e., 10 *ps*.^{51,82} CVs were set to be the dihedral angles, ϕ and ψ , which were evenly spaced in bins of 0.17 *rad* with ϕ and ψ ranging between $[-3.14 \text{ rad}, 3.14 \text{ rad}]$. The target number of walkers per bin, n_w was set to be 4 for the WE simulation.

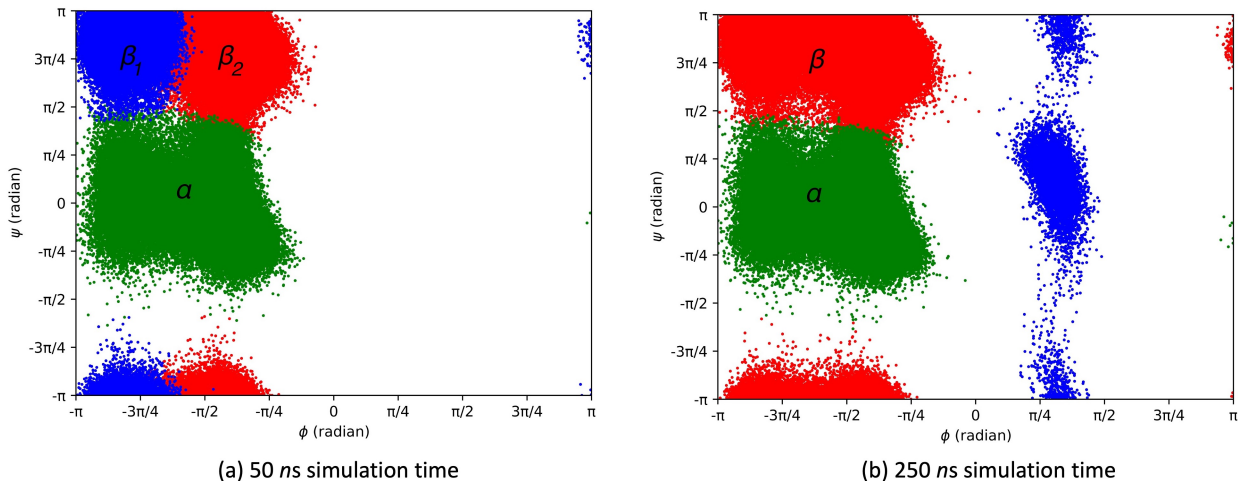


Figure 2: Three metastable states of alanine dipeptide as an output of VAMPNets (a) separated by α (green) and β (red and blue) regions of the Ramachandran plot and (b) separated by negative (green and red) and positive (blue) values of ϕ and further separated by α (green) and β (red) regions of the Ramachandran plot.

To compare the performance of the DeepWEST method with the WE and the hybrid GaMD-WE method, we have identified three metastable states in alanine dipeptide as shown in Figure 2(a). Note that Figure 2(b) shows the three metastable states when the simulation time was 250 *ns*. For the DeepWEST method, we have used the 50 *ns* simulation time as an input trajectory to be trained by the network architecture. The metastable states were chosen in such a way that each state represents an important region of the PES. These regions are α_R , α_L , and P_{II} . α_R lies within the α region while P_{II} lies in the β region of the

Table 1: Hyperparameters used in VAMPNets for alanine dipeptide

Hyperparameters	Description	Value
τ	Time lag between the two MD trajectory datasets	80 ps
Batch size	Number of samples in a batch for gradient descent	1000
Train ratio	Percentage of trajectory points used for training	0.9
Network depth	Number of hidden layers in the network	8
Layer width	Width of the hidden layer	100
Learning rate	Learning rate used for the ADAM optimiser	1e-4
Output size	Number of metastable states as an output	3
n_{epochs}	Number of iterations over the training set	100
ϵ	Threshold for eigenvalues	1e-5

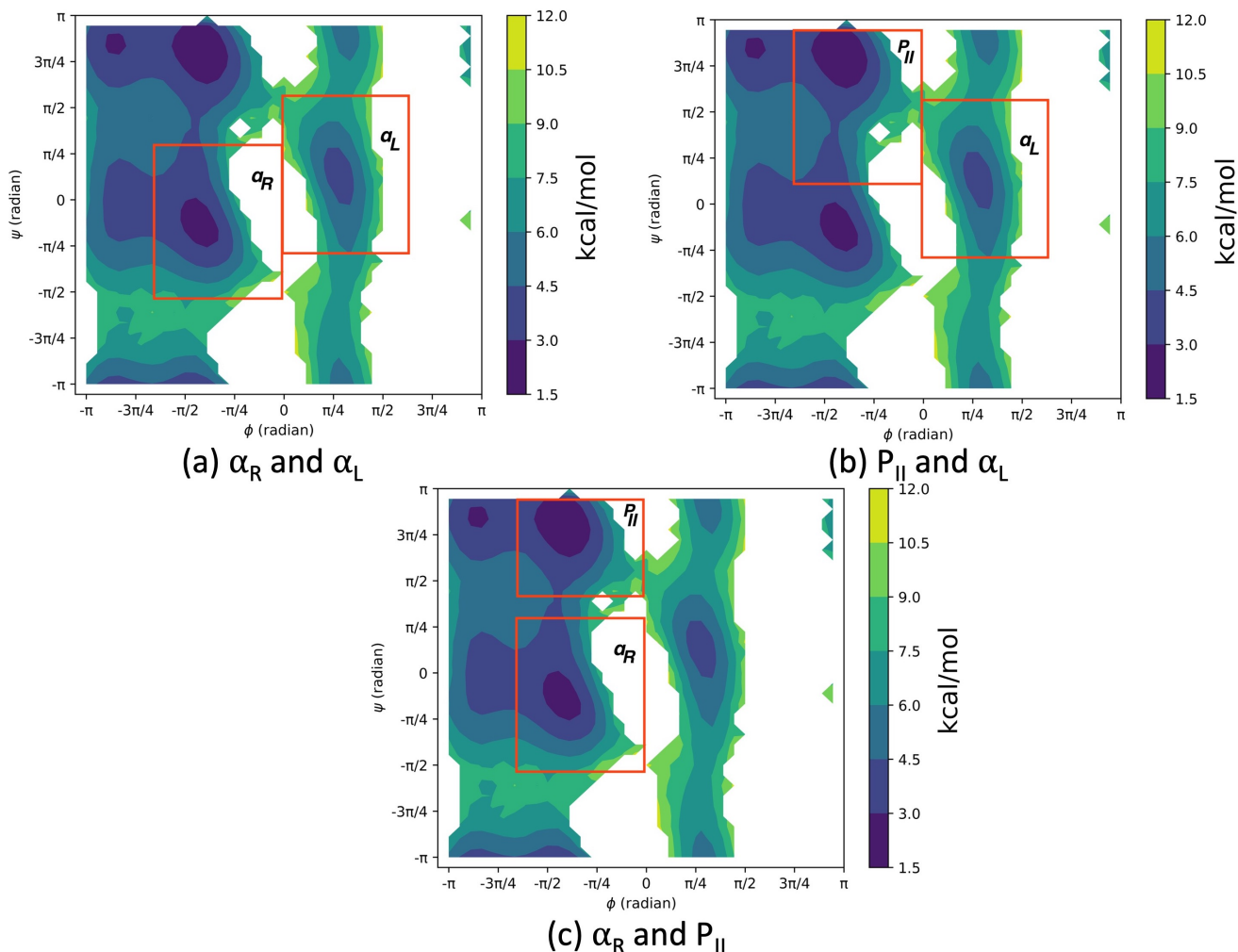


Figure 3: Average free energy profile of alanine dipeptide from three separate 12 μ s runs DeepWEST simulations. Rate constants to be calculated between the following regions of interest in alanine dipeptide: (a) $\alpha_R \Leftrightarrow \alpha_L$ (b) $P_{II} \Leftrightarrow \alpha_L$ and (c) $\alpha_R \Leftrightarrow P_{II}$

Ramachandran plot (Figure 2 and 3). α_L lies in the region where ϕ attains positive radian values. α_R is defined as $-2.09 \text{ rad} \leq \phi \leq 0 \text{ rad}$ and $-1.75 \text{ rad} \leq \psi \leq 0.87 \text{ rad}$, α_L is defined as $0 \text{ rad} \leq \phi \leq 2.09 \text{ rad}$ and $-0.87 \text{ rad} \leq \psi \leq 1.75 \text{ rad}$, and P_{II} is defined as $-2.09 \text{ rad} \leq \phi \leq 0 \text{ rad}$ and $1.31 \text{ rad} \leq \psi \leq 3.14 \text{ rad}$. VAMPNets demonstrated success in the discretization of MD trajectory data in three metastable states separating the α and β regions of the Ramachandran plot (Figure 2a) and further separating the conformations with the negative and positive values of the ϕ angles (Figure 2b).

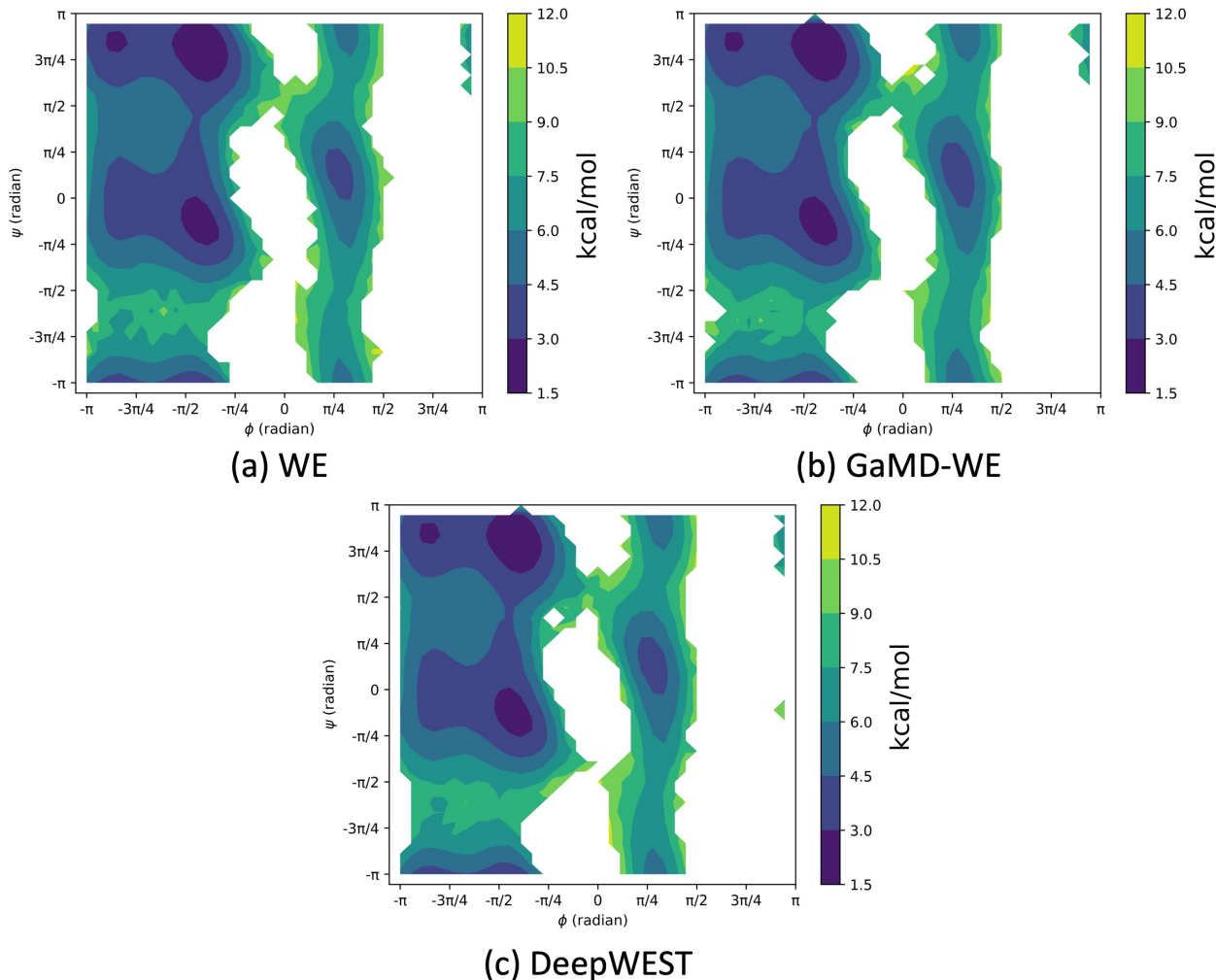


Figure 4: Average free energy profiles of alanine dipeptide from three separate $12 \mu\text{s}$ runs of WE, GaMD-WE, and DeepWEST simulations, respectively.

Two different WE approaches, namely the conventional WE and the hybrid GaMD-WE approach, were compared to assess the performance of our newly developed DeepWEST

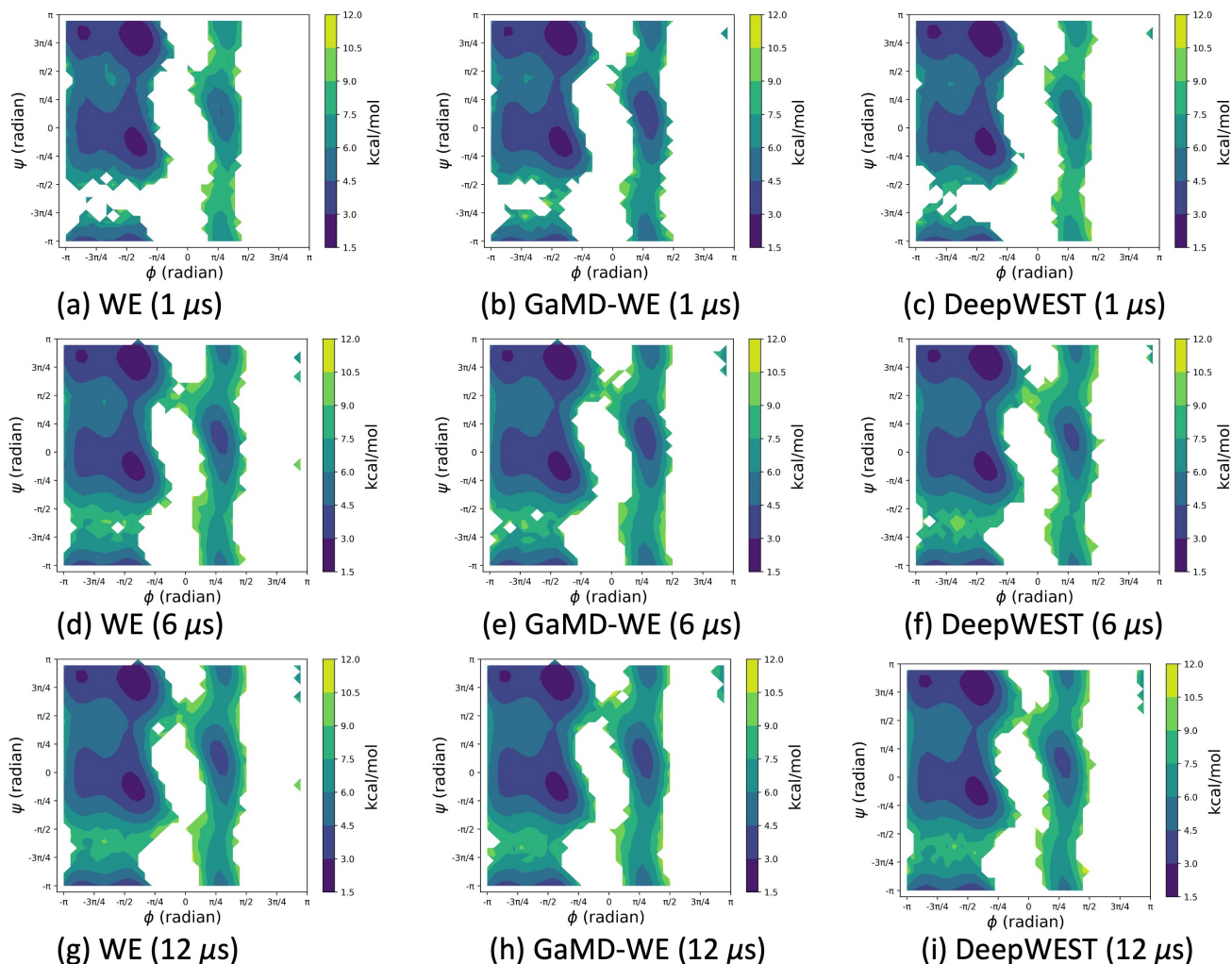


Figure 5: Average free energy profiles of alanine dipeptide from three separate runs of WE, GaMD-WE, and DeepWEST simulations at different stages the simulation.

Table 2: Rate constants (ns^{-1}) between different metastable states of alanine dipeptide obtained after $12 \mu\text{s}$ of aggregate simulation time. In the case of brute force simulation, the first value indicates the average rate constant while the value within the brackets indicates the 95% confidence interval computed using Bayesian bootstrapping. For WE, GaMD-WE, and the DeepWEST methods, the error bars represent 95% confidence intervals obtained from the standard deviation of three independent runs.

Transition	Brute force	WE	GaMD-WE	DeepWEST
$\alpha_L \rightarrow P_{II}$	0.335, [0.275, 0.408]	0.325 ± 0.022	0.328 ± 0.012	0.336 ± 0.015
$P_{II} \rightarrow \alpha_L$	0.01, [0.008, 0.012]	0.009 ± 0.002	0.008 ± 0.001	0.009 ± 0.001
$\alpha_R \rightarrow P_{II}$	6.812, [6.723, 6.897]	6.971 ± 0.046	6.737 ± 0.037	6.788 ± 0.155
$P_{II} \rightarrow \alpha_R$	2.823, [2.777, 2.868]	3.012 ± 0.052	2.941 ± 0.042	2.975 ± 0.024
$\alpha_R \rightarrow \alpha_L$	0.01, [0.008, 0.012]	0.009 ± 0.002	0.008 ± 0.002	0.009 ± 0.001
$\alpha_L \rightarrow \alpha_R$	0.305, [0.254, 0.367]	0.375 ± 0.27	0.325 ± 0.073	0.445 ± 0.427

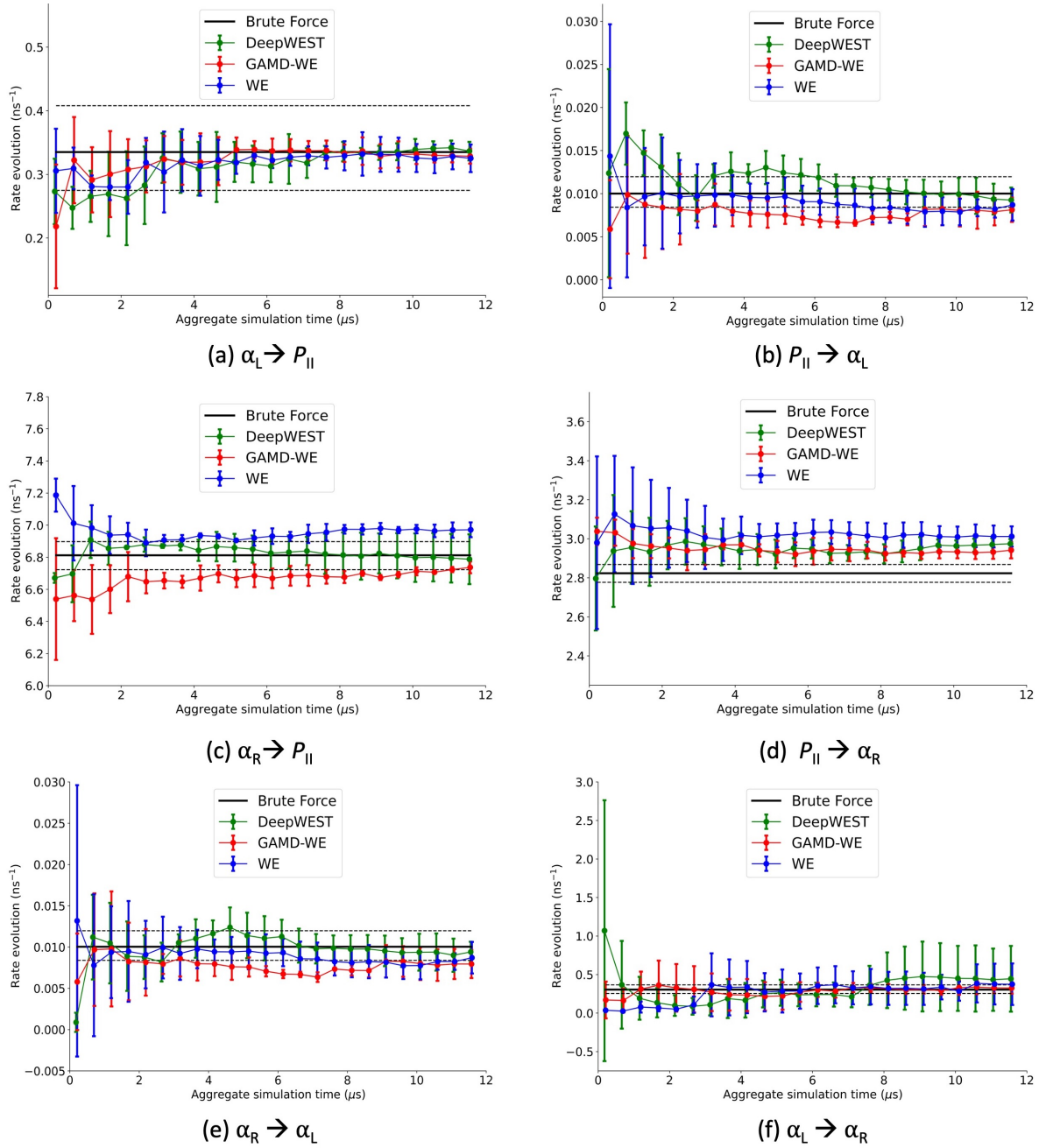


Figure 6: Evolution of rate constants over aggregate simulation time for brute force simulation (black), WE simulations (red), GAMD-WE approach (blue), and DeepWEST approach (green). For WE, GaMD-WE, and the DeepWEST methods, the error bars represent 95% confidence intervals obtained from the standard deviation of three independent runs.

approach. WE was run for a total simulation time of 36 μs averaged over three independent runs of 12 μs each. Similarly, the GaMD-WE approach was run for 50 ns of GaMD followed by 11.95 μs of WE, totaling a simulation time of 36 μs averaged over three independent runs of 12 μs each. Likewise, the DeepWEST approach was run for an unbiased MD simulation of 50 ns followed by 11.95 μs of WE simulation, totaling a simulation time of 36 μs averaged over three independent runs of 12 μs each. Figure 2 demonstrates the ability of the deep-learned kinetic modeling approach to cluster the MSMs and also ensures that adequate sampling has been achieved to select starting conformations. The average free energy is given by $-k_B T \ln P$ where k_B is the Boltzmann constant, T is the temperature, and P is the probability. Figure 4a and Figure 4b represent the average free energy profiles of alanine dipeptide for the WE and the hybrid GaMD-WE approach, respectively, while Figure 4c represents the free energy profile for the DeepWEST approach. Figure 4c shows almost identical PES coverages obtained through the DeepWEST approach as compared to the WE (Figure 4a) and the hybrid GaMD-WE approach (Figure 4b). This, in turn, demonstrates the ability of the DeepWEST approach in accessing the entire free energy landscape of the system with enough sampling in the P_{II} , α_{L} , and the right-handed α -helix or α_{R} region. Figure 5 demonstrates the average free energy profiles of alanine dipeptide from three separate runs of WE, GaMD-WE, and DeepWEST simulations at different stages the simulation.

To explore the performance of the DeepWEST approach in estimating kinetic rates, brute force simulations are carried out to obtain reference values. Three independent 12 μs simulations were run, and Bayesian bootstrapping was performed for 95 % confidence intervals. Rate constants were obtained from different methods, i.e., the brute force calculations, WE, GaMD-WE, and the DeepWEST method for transitions between metastable regions of interest (Table 2). For the transitions between the central metastable region, α_{R} to P_{II} , the DeepWEST method outperformed both the WE and the GaMD-WE methods in estimating the rate constant. The rate constant for the DeepWEST method converged faster compared

to the brute force value (Figure 6c). However, as observed in Figure 6d, all the methods slightly overestimated the rate constant for the transition from P_{II} to α_{R} . In the transitions from α_{L} to P_{II} (Figure 6a), P_{II} to α_{L} (Figure 6b), α_{R} to α_{L} (Figure 6e), and α_{L} to α_{R} (Figure 6f), comparable convergence and accuracy of the kinetic rates were observed for the DeepWEST approach as compared to the WE and the GaMD-WE methods.

3.2 Chignolin

Chignolin (PDB ID: 1UAO), a designed protein, is a model system consisting of 10 amino acid residues (GYDPETGTWG) (Figure 1b).⁹⁵ It forms a stable β -hairpin structure in implicit solvent, and MD simulations reveal the slow unfolding of chignolin.⁹⁶ Initial coordinates for chignolin were obtained from <https://files.rcsb.org/download/1UAO.pdb1.gz>. The system was then subjected to Generalized Born (GB) implicit solvation using the model II radii.^{97–99} AMBER ff14SB force field was used to prepare the system for MD simulations.⁹³ Chignolin was subjected to 100 *ns* of constant volume unbiased MD simulation with a collision frequency of 1 *ps*⁻¹, employing a Langevin thermostat at a constant temperature of 300 K. To interpret the kinetics of a system that demonstrates folding and unfolding behavior, we encounter the problem of trajectory alignment with respect to a unique reference structure. Moreover, a large amount of noise would be introduced while the networks transform the data via rotations and translations. Hence for the chignolin system, internal coordinates were chosen as a network input. Nearest-neighbor heavy-atom distances between all non-redundant residues separated by two or more residues served as the network input resulting in 28 nodes. MD trajectories were trained using VAMP neural networks with a time lag of 40 *ps*, training ratio of 0.9, and a batch size of 1000 that learned discretization in three metastable states by performing a hierarchical decomposition of the state space. A list of hyperparameters used for network training is provided in Table 3. Starting structures were selected from the metastable states as per the DeepWEST protocol, followed by three independent WE simulations of 40 μs each with a total simulation time of 120 μs . Resampling

time, τ was kept to be identical as used in WE and GaMD-WE runs, i.e., 20 *ps*.⁵¹ CVs for WE simulations were set to be mass-weighted root-mean-square deviation (RMSD) and mass-weighted radius of gyration (R_g) respectively. Both CVs were evenly spaced in bins of 0.02 *nm* ranging between [0 *nm*, 1 *nm*]. The target number of walkers per bin, n_w was set to 4 for the WE simulation.

Table 3: Hyperparameters used in VAMPNets for chignolin

Hyperparameters	Description	Value
τ	Time lag between the two MD trajectory datasets	40 <i>ps</i>
Batch size	Number of samples in a batch for gradient descent	1000
Train ratio	Percentage of trajectory points used for training	0.9
Network depth	Number of hidden layers in the network	6
Layer width	Width of the hidden layer	100
Learning rate	Learning rate used for the ADAM optimiser	1e-4
Output size	Number of metastable states as an output	3
n_{epochs}	Number of iterations over the training set	100
ϵ	Threshold for eigenvalues	1e-5

To compare the performance of the DeepWEST method with the hybrid GaMD-WE and WE methods, we have identified three metastable states in chignolin as shown in Figure 7. These regions are folded, unfolded, Intermediate I (I_1), and Intermediate II (I_2). The folded region is defined as $\text{RMSD} \leq 0.20$ *nm*, while the unfolded region is defined as $\text{RMSD} \geq 0.55$ *nm*. I_1 is defined to be 0.20 *nm* \leq $\text{RMSD} \leq 0.30$ *nm* and 0.425 *nm* \leq $R_g \leq 0.525$ *nm* while I_2 is defined to be 0.60 *nm* \leq $\text{RMSD} \leq 0.70$ *nm* and 0.70 *nm* \leq $R_g \leq 0.80$ *nm*.

Chignolin represents a common protein structural motif that undergoes folding and unfolding during brute force MD simulations. We tested the DeepWEST method by estimating rate constants between regions of interest, especially in the folded and unfolded regions. WE was run for a total simulation time of 120 μs averaged over three independent runs of 40 μs each. Similarly, the GaMD-WE approach ran six independent 500 *ns* of GaMD with varying boost potentials, and starting conformations were selected from the GaMD run that had the largest PES coverage. It was then followed by 39.50 μs of WE simulations, totaling a simulation time of 120 μs averaged over three independent runs of 40 μs each. Likewise, the

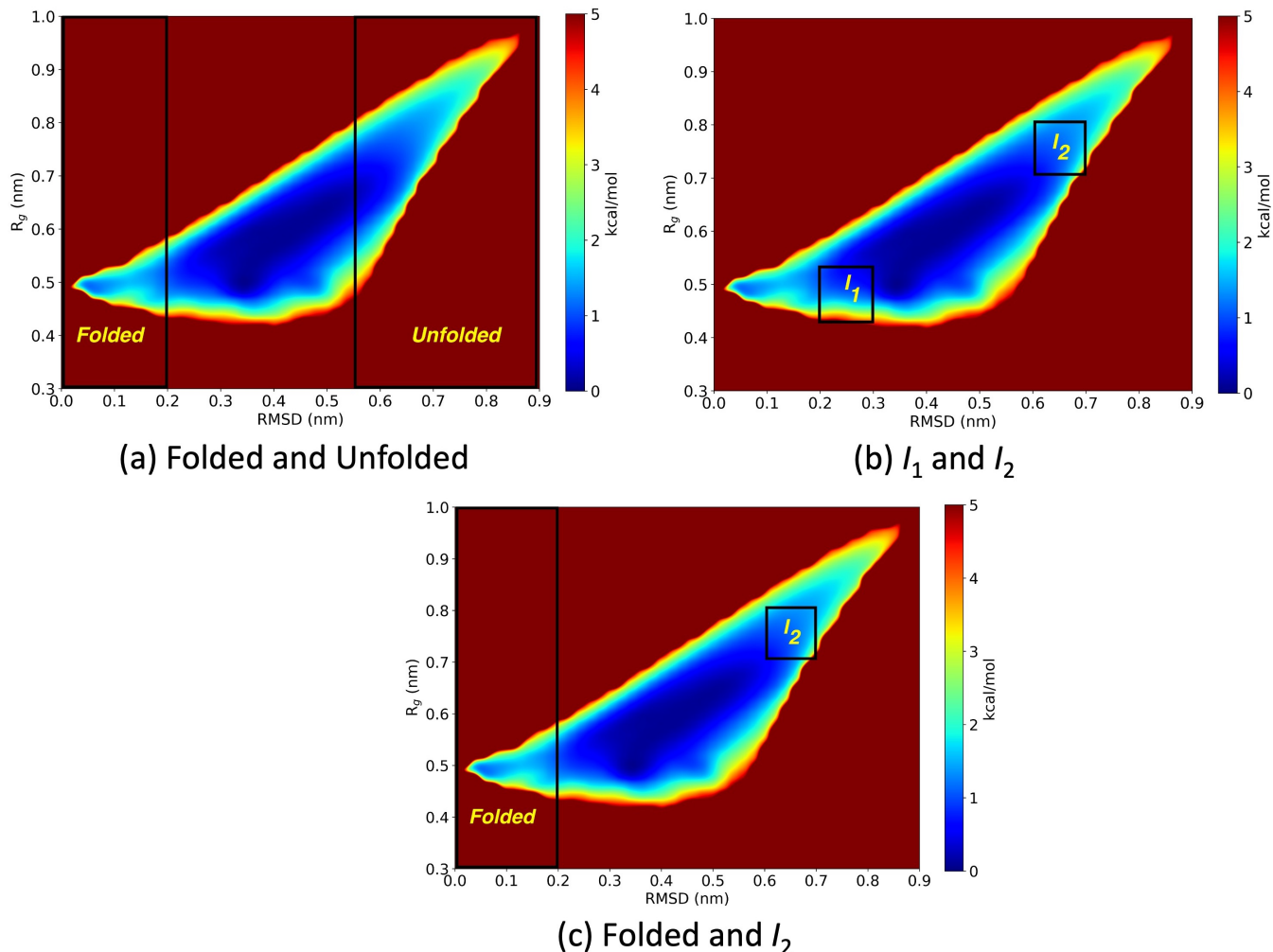


Figure 7: Average free energy profile of chignolin from three separate 40 μ s runs of DeepWEST simulations. Rate constants to be calculated between the following regions of interest in chignolin: (a) Folded \Leftrightarrow Unfolded (b) $I_1 \Leftrightarrow I_2$ and (c) Folded $\Leftrightarrow I_2$

Table 4: Rate constants (ns^{-1}) between different metastable states for chignolin obtained after 40 μ s of aggregate simulation time. In the case of brute force simulation, the first value indicates the average rate constant while the value within the brackets indicates the 95% confidence interval computed using Bayesian bootstrapping. For WE, GaMD-WE, and the DeepWEST methods, the error bars represent 95% confidence intervals obtained from the standard deviation of three independent runs.

Transition	Brute force	WE	GaMD-WE	DeepWEST
Folded \rightarrow Unfolded	1.016, [0.983, 1.05]	0.873 ± 0.036	1.044 ± 0.017	0.998 ± 0.049
Unfolded \rightarrow Folded	0.144, [0.137, 0.15]	0.159 ± 0.002	0.148 ± 0.003	0.149 ± 0.013
$I_1 \rightarrow I_2$	0.513, [0.501, 0.526]	0.515 ± 0.009	0.509 ± 0.017	0.513 ± 0.018
$I_2 \rightarrow I_1$	0.316, [0.307, 0.325]	0.327 ± 0.012	0.325 ± 0.015	0.324 ± 0.007
$I_2 \rightarrow$ Folded	0.134, [0.127, 0.14]	0.145 ± 0.003	0.135 ± 0.005	0.136 ± 0.011
Folded $\rightarrow I_2$	0.519, [0.502, 0.536]	0.49 ± 0.009	0.506 ± 0.016	0.511 ± 0.015

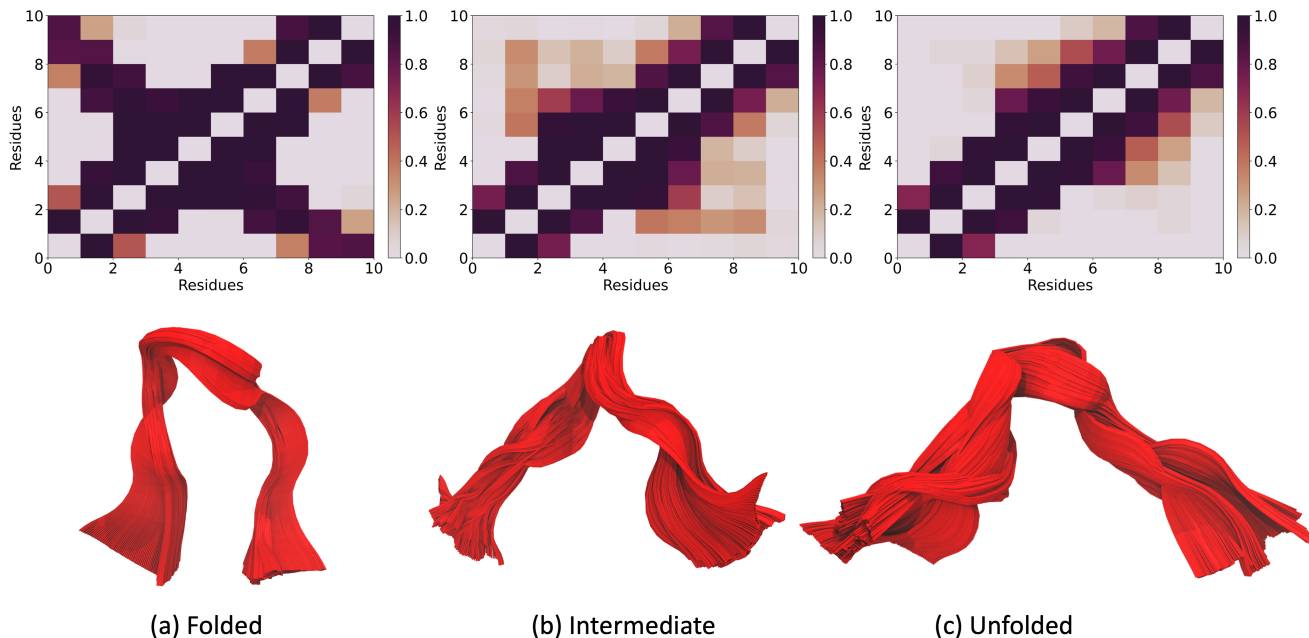


Figure 8: Three metastable states of chignolin as an output of VAMPNets. The upper panel shows the mean contact map for each metastable state while the lower panel shows the 3D representation for each of the states.

DeepWEST approach ran an unbiased MD simulation of 100 *ns* followed by 39.90 μ s of WE simulation, totaling a simulation time of 120 μ s averaged over three independent runs of 40 μ s each. Figure 8 demonstrates the ability of the deep-learned kinetic modeling approach to cluster the MSMs for chignolin. Initial conformations were selected from these metastable states for WE simulations. Figure 9a and Figure 9b represent the average free energy profiles of chignolin from the WE and the hybrid GaMD-WE approach, respectively, while Figure 9c represents the free energy profile from the DeepWEST approach. Figure 10 demonstrates the average free energy profiles of chignolin from three separate runs of WE, GaMD-WE, and DeepWEST simulations at different stages the simulation. The PES coverages for chignolin from the three approaches are comparable and demonstrates that the DeepWEST method can sample the thermodynamics of the system accurately.

To assess the performance of the DeepWEST approach with respect to WE and the hybrid GaMD-WE approach, we estimated the convergence of rate constants between metastable states, especially defined within folded and unfolded regions of chignolin (Table 4). The

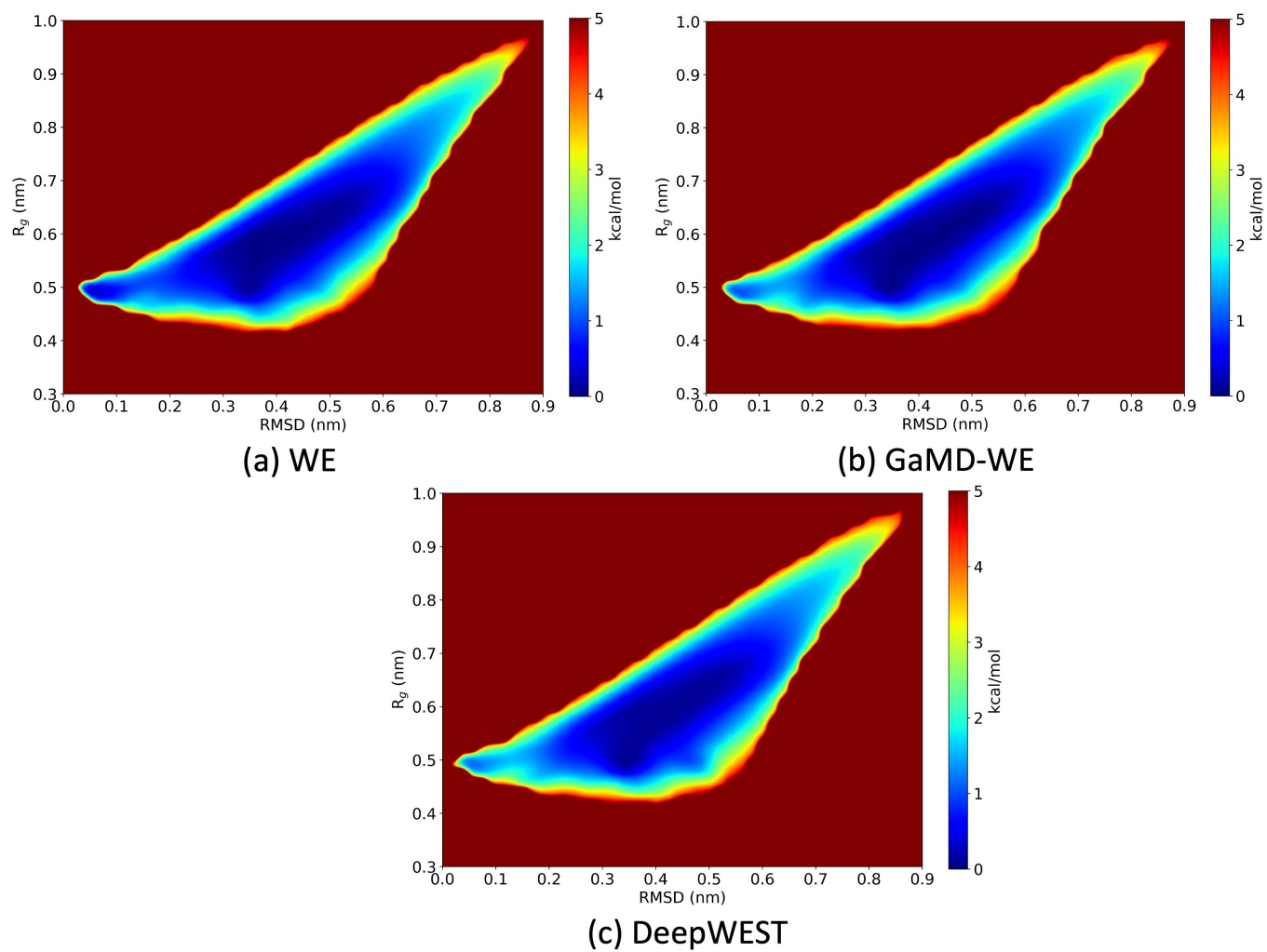


Figure 9: Average free energy profiles of chignolin from three separate 40 μ s runs of WE, GaMD-WE, and DeepWEST simulations, respectively.

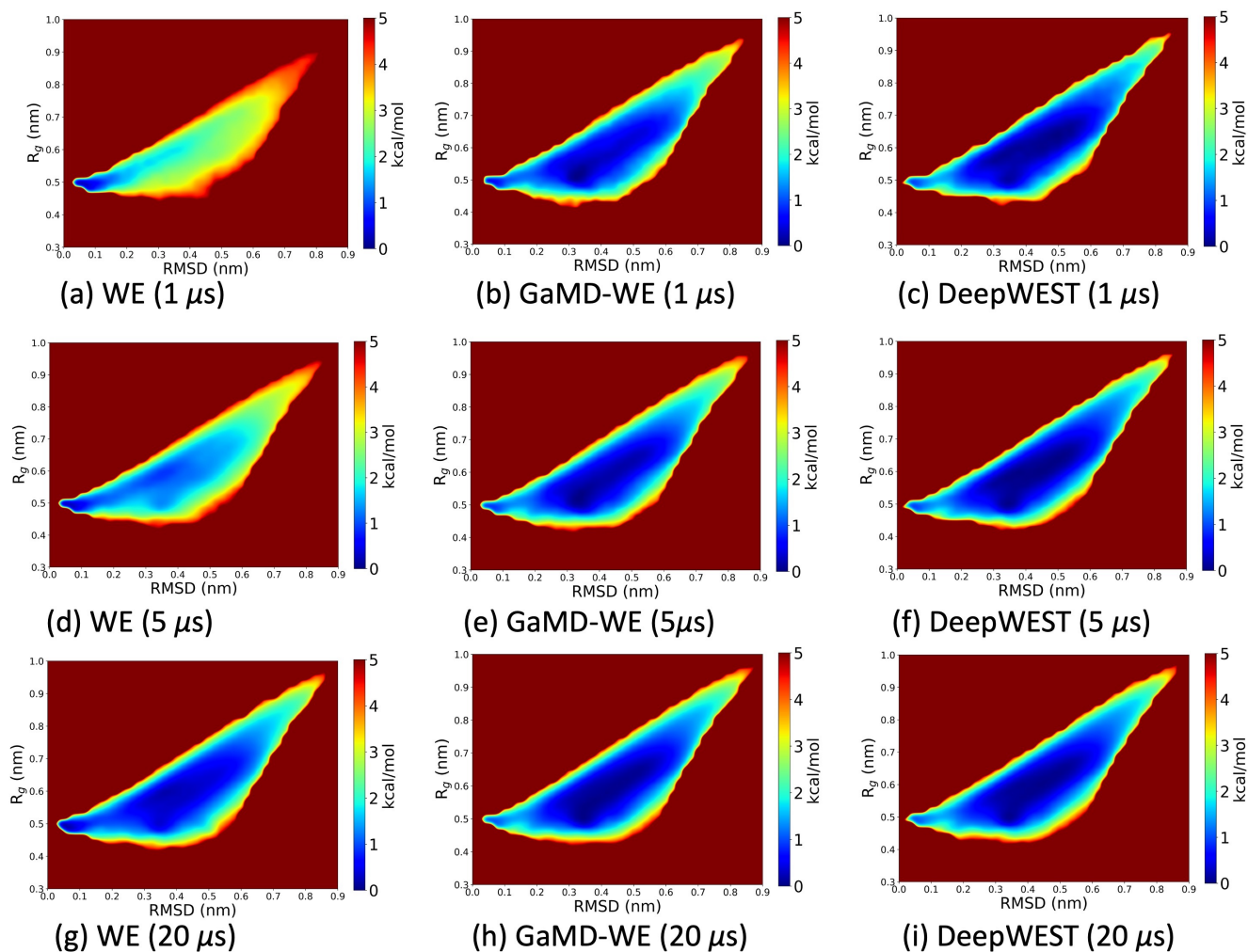


Figure 10: Average free energy profiles of chignolin from three separate runs of WE, GaMD-WE, and DeepWEST simulations at different stages the simulation.

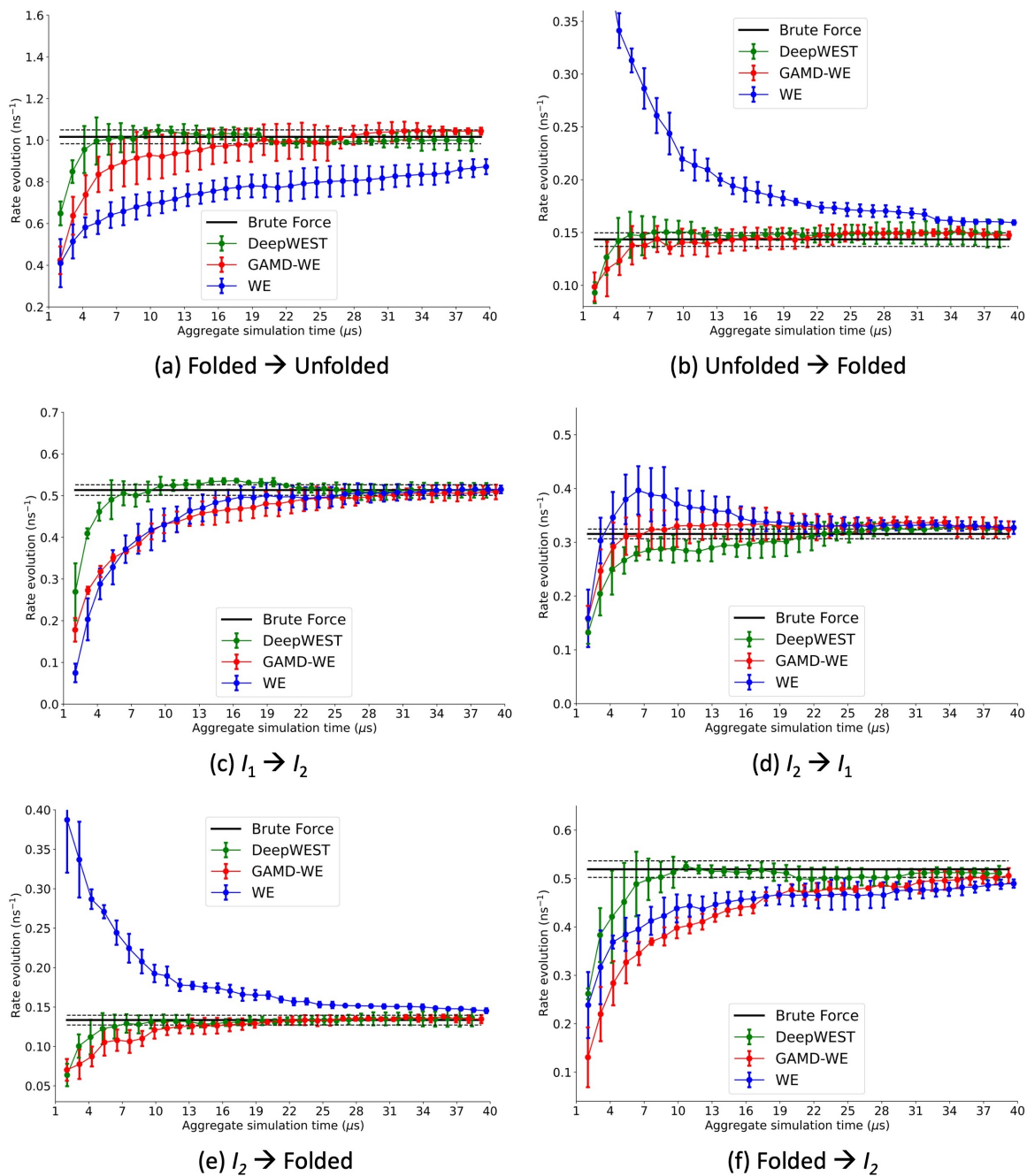


Figure 11: Evolution of rate constants over aggregate simulation time for brute force simulation (black), WE simulations (red), GAMD-WE approach (blue), and DeepWEST approach (green). For WE, GaMD-WE, and the DeepWEST methods, the error bars represent 95% confidence intervals obtained from the standard deviation of three independent runs.

two intermediate regions, I_1 and I_2 are separated apart in the PES, and convergence of rate constants between these regions is difficult to achieve. Figure 11c demonstrates faster and more accurate convergence of the DeepWEST approach as compared to the WE and GaMD-WE approach for the transition from I_1 to I_2 . However, the GaMD-WE method outperformed the DeepWEST method in estimating the rate constant for the transition from I_2 to I_1 (Figure 11d). The DeepWEST method showed faster convergence and greater accuracy in estimating the rate constants between the I_2 and the folded states (Figure 11e and 11f). Faster convergence for the DeepWEST method was also observed from the folded to the unfolded state (Figure 11a). However, the GaMD-WE and the DeepWEST method equally outperformed the WE method in estimating the rate constant in the transition from the unfolded to the folded state (Figure 11a). In conclusion, for all the cases, both the hybrid methods, especially the DeepWEST method, outperformed the WE method in achieving faster convergence and accuracy of rate constants between the metastable regions of interest.

3.3 NTL9

To further test the capabilities of the DeepWEST method for a higher dimensional problem, we chose the N-terminal domain of ribosomal protein L9 (NTL9) system. The NTL9 protein is a 627-atom system consisting of 39 amino acid residues (Figure 1c) which has a folding time of $\approx 1.5 \text{ ms}$ ¹⁰⁰. Initial coordinates for the unfolded NTL9 protein was obtained from https://github.com/westpa/westpa2_tutorials/blob/main/tutorial-3.3/istates/ntl9.pdb. The system was then subjected to Generalized Born (GB) implicit solvation using the model II radii⁹⁷⁻⁹⁹. AMBER ff19SB force field was used to prepare the system for MD simulations¹⁰¹. The NTL9 protein was then subjected to 10 μs of constant volume unbiased MD simulation with a collision frequency of 1 ps^{-1} , employing a Langevin thermostat at a constant temperature of 300 K. Following the folding and unfolding behavior of the NTL9 system, internal coordinates were chosen as a network input. Nearest-neighbor heavy-atom distances between all non-

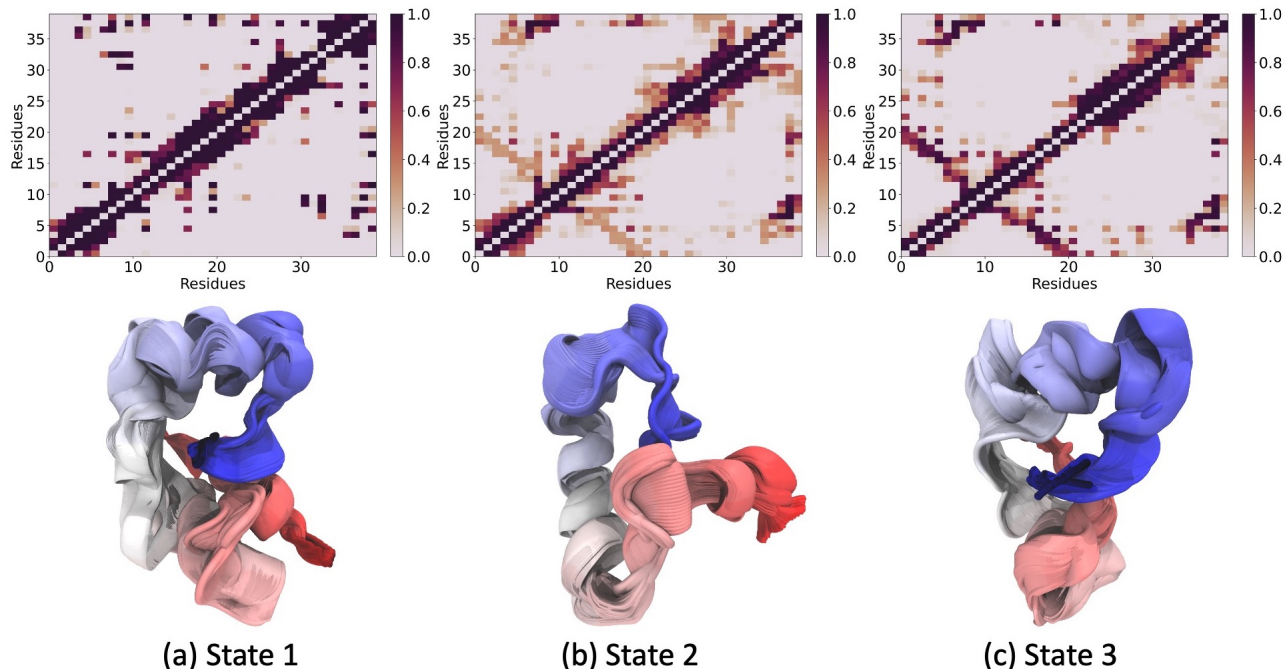


Figure 12: Three metastable states of the NTL9 system as an output of VAMPNets. The upper panel shows the mean contact map for each metastable state while the lower panel shows the 3D representation for each of the states.

redundant residues separated by two or more residues served as the network input resulting in 666 nodes. MD trajectories were trained using VAMP neural networks with a time lag of 60 *ns*, training ratio of 0.9, and a batch size of 1000 that learned discretization in three metastable states by performing a hierarchical decomposition of the state space (Figure 12). Starting structures were selected from the metastable states as per the DeepWEST protocol followed by the WE simulation of 90 μ s with a resampling time, τ of 40 *ps*. Therefore, the total simulation time for the DeepWEST method was 120 μ s. CVs for WE simulations were set to be mass-weighted root-mean-square deviation (RMSD) and mass-weighted radius of gyration (R_g) respectively. Both CVs were evenly spaced in bins of 0.05 *nm* ranging between [0 *nm*, 2 *nm*]. The target number of walkers per bin, n_w was set to 8 for the WE simulation. Similarly, the conventional WE simulation was run for 130 μ s with the same parameters (i.e. resampling time, CVs and binning parameters) as described previously for the DeepWEST simulation. The WE simulations initiated from the unfolded state to the folded state with

the CVs set to be the mass-weighted RMSD and mass-weighted R_g respectively. Since the reference structure for computing RMSD is the initial unfolded state, RMSD is expected to increase during the simulation as the system folds. The unfolded region is defined as $\text{RMSD} \leq 0.60 \text{ nm}$, while the unfolded region is defined as $\text{RMSD} \geq 1.0 \text{ nm}$.

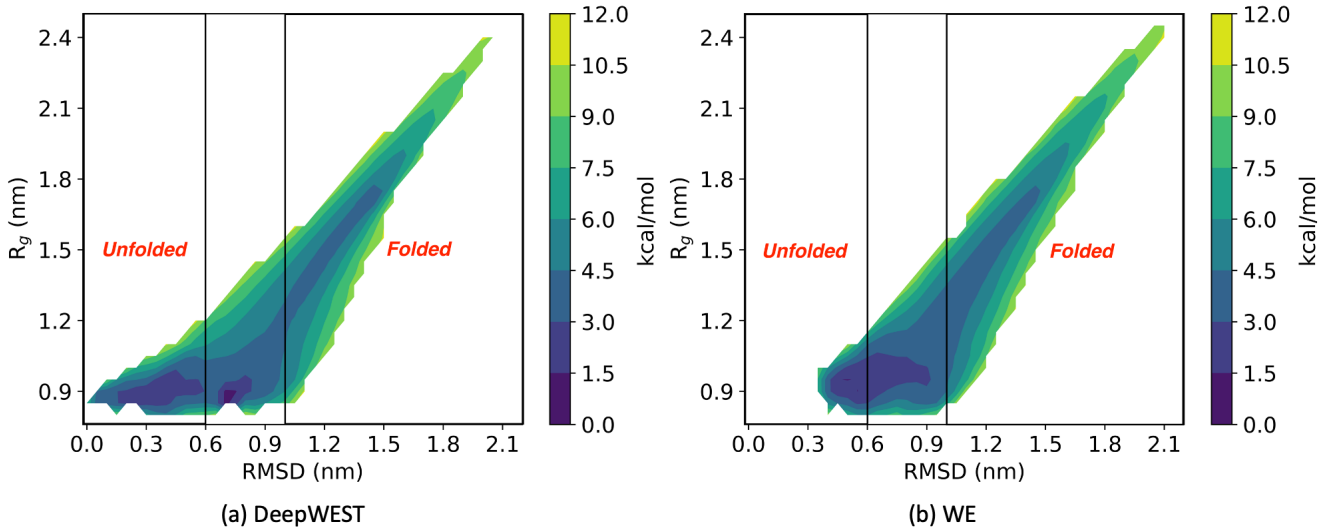


Figure 13: Average free energy profiles of NTL9 from 100 μs runs of DeepWEST and WE simulations, respectively.

For the same amount of simulation time, i.e., 130 μs , the DeepWEST method proved more effective than the conventional WE in covering the free energy landscape of a more complex protein, i.e., the NTL9 system. Figure 13a and Figure 13b represent the free energy landscapes for the NTL9 system for the DeepWEST and the conventional WE method, respectively, where the lowest energy state was set to zero. It is evident from Figure 13a that the DeepWEST approach is more effective than the conventional WE in exploring other metastable states in the NTL9 system, particularly in the unfolded region. The rate constants between the unfolded and the folded states were not measured, but we expect the DeepWEST method to outperform the conventional WE method in kinetic rate measurements.

4 Discussion

The WE method predominantly depends on appropriate CVs to sample the system. In most cases, the best choice of CVs is unknown. In cases where CVs cannot sufficiently describe the system’s dynamics, hybrid approaches and most enhanced sampling CV-free approaches may generate a well-sampled initial configurational space.^{26,51,102–105} Initial sampling of the PES has a significant impact on the accuracy and convergence of WE. Our previously developed hybrid GaMD-WE method addresses the concern of substantial initial sampling but reweighing is an additional concern associated with it. To obtain the correct PES of the system, GaMD resorts to reweighing, and a substantial amount of energetic noise is introduced with an increase in the system size. In that case, GaMD needs to be run for a longer timescale for varying degrees of boost potentials, increasing the overall computational cost drastically. As demonstrated through previous examples, the newly developed DeepWEST approach is a powerful method that obtains both the kinetics and thermodynamics of the system. Compared to our earlier developed hybrid approach, it is a hassle-free method for enhanced sampling prior to running WE simulations. First, DeepWEST eliminates the process of running various accelerated MD simulations with varying degrees of boost potentials. Second, it also eliminates the process of PES reweighing to uncover the original energy landscape by running unbiased MD simulations. Most importantly, deep neural architectures learn the clustering of metastable states from short MD trajectories which could be further processed to extract starting conformations for WE simulations. Both methods involved in the DeepWEST approach add no statistical bias to the system, and thereby, kinetics and thermodynamics obtained from this approach are exact. However, a longer MD simulation time may be required to discretize metastable states as the system size increases to provide statistically relevant starting conformations for WE simulations. Recent improvements in the WE method have been made to achieve faster kinetics.^{55,104} An even more accurate and faster kinetics and thermodynamic sampling could be achieved if DeepWEST is combined with these improvements in WE.

Currently, we have shown the DeepWEST method to work on simpler systems, i.e., alanine dipeptide in explicit solvation and chignolin in implicit solvation. We have also shown a greater free energy landscape coverage for the NTL9 protein for the DeepWEST method against the conventional WE method. However, extending this method to more complex biological systems comes with challenges. Deep neural networks learn feature transformations and cluster trajectories based on input CVs. The network architecture for such systems primarily depends on the choice of time lag and the number of metastable states we want to cluster. Identifying the slowest processes in complex systems comes with the choice of accurate CVs, which in most cases is unknown. However, recent developments in enhanced MD sampling and MSM approaches would be worth considering. Implementing multi-ensemble Markov models (MEMMs) that conduct large ensembles of even shorter simulations and further accelerate trajectories by adding boosted potentials through various accelerated enhanced sampling methods in DeepWEST could aid the clustering of complex proteins.^{58,106} Implementation of hydrogen-mass repartitioning (HMR) for solvated systems to accelerate the dynamics would prove indispensable in faster learning of metastable states.²⁷

5 Conclusion

A hybrid approach for a faster learning of kinetics combining the deep learning of MD trajectory with the WE simulations is presented here. It employs a deep learning architecture to sample statistically relevant conformations representative of the rare events from short MD simulation trajectories. The method is tested on three different model systems, namely, alanine dipeptide with explicit solvation, chignolin with implicit solvation, and the NTL9 protein with implicit solvation. Our method significantly outperforms the WE and the GaMD-WE approach for the chignolin model system, and is comparable in performance for the alanine dipeptide system. Because of its reduced simulation time and slighter sophistication as compared to the GaMD-WE approach, DeepWEST allows for fast estimation of kinetics

with enhanced coverage. We have also developed an end-to-end package, DeepWEST, that automates the entire process of running MD trajectories, extracting statistically relevant conformations using VAMP theory and neural networks, and preparing the WE simulation. The DeepWEST package is available at <https://github.com/anandojha/DeepWEST>. In conclusion, we have demonstrated a proof-of-concept of a hybrid data-driven approach that can be incorporated in the WE approach to obtain improved results with significantly lesser time and complexities.

Acknowledgement

A.A.O. acknowledges support the Molecular Sciences Software Institute (MolSSI) fellowship under NSF grant OAC-1547580. R.E.A. acknowledges support from NSF Extreme Science and Engineering Discovery Environment (XSEDE) CHE060063 and NIH GM132826. All simulations were performed using the Triton Shared Computing Cluster (TSCC) and Expanse at the San Diego Supercomputing Center (SDSC). The latter allocation is supported by NSF XSEDE CHE060063 to R.E.A.

References

- (1) Durrant, J. D.; McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC biology* **2011**, *9*, 1–9.
- (2) Karplus, M.; Petsko, G. A. Molecular dynamics simulations in biology. *Nature* **1990**, *347*, 631–639.
- (3) Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nature structural biology* **2002**, *9*, 646–652.
- (4) Wood, W. W.; Erpenbeck, J. J. Molecular dynamics and Monte Carlo calculations in statistical mechanics. *Annual Review of Physical Chemistry* **1976**, *27*, 319–348.

- (5) Wereszczynski, J.; McCammon, J. A. Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition. *Quarterly reviews of biophysics* **2012**, *45*, 1–25.
- (6) Zheng, Q.; Chu, W.; Zhao, C.; Zhang, L.; Guo, H.; Wang, Y.; Jiang, X.; Zhao, J. Ab initio nonadiabatic molecular dynamics investigations on the excited carriers in condensed matter systems. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2019**, *9*, e1411.
- (7) Smith, B.; Akimov, A. V. Modeling nonadiabatic dynamics in condensed matter materials: some recent advances and applications. *Journal of Physics: Condensed Matter* **2019**, *32*, 073001.
- (8) Steinhauser, M. O.; Hiermaier, S. A review of computational methods in materials science: examples from shock-wave and polymer physics. *International journal of molecular sciences* **2009**, *10*, 5135–5216.
- (9) Massobrio, C.; Du, J.; Bernasconi, M.; Salmon, P. S. Molecular dynamics simulations of disordered materials. *Cham: Springer International Publishing* **2015**,
- (10) Akimov, A. V.; Prezhdo, O. V. The PYXAID program for non-adiabatic molecular dynamics in condensed matter systems. *Journal of chemical theory and computation* **2013**, *9*, 4959–4972.
- (11) Piggot, T. J.; Holdbrook, D. A.; Khalid, S. Conformational dynamics and membrane interactions of the E. coli outer membrane protein FecA: a molecular dynamics simulation study. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2013**, *1828*, 284–293.
- (12) Lindahl, E.; Sansom, M. S. Membrane proteins: molecular dynamics simulations. *Current opinion in structural biology* **2008**, *18*, 425–431.

- (13) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. C.; Zhestkov, Y., et al. Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and a β -hairpin peptide. *The Journal of Physical Chemistry B* **2004**, *108*, 6582–6594.
- (14) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Dill, K. A. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Modeling & Simulation* **2006**, *5*, 1214–1226.
- (15) Cui, Q.; Sulea, T.; Schrag, J. D.; Munger, C.; Hung, M.-N.; Naïm, M.; Cygler, M.; Purisima, E. O. Molecular dynamics—Solvated interaction energy studies of protein–protein interactions: The MP1–p14 scaffolding complex. *Journal of molecular biology* **2008**, *379*, 787–802.
- (16) Rakers, C.; Bermudez, M.; Keller, B. G.; Mortier, J.; Wolber, G. Computational close up on protein–protein interactions: how to unravel the invisible using molecular dynamics simulations? *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2015**, *5*, 345–359.
- (17) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry* **2016**, *59*, 4035–4061.
- (18) Liu, K.; Watanabe, E.; Kokubo, H. Exploring the stability of ligand binding modes to proteins by molecular dynamics simulations. *Journal of computer-aided molecular design* **2017**, *31*, 201–211.
- (19) Jorgensen, W. L.; Jensen, K. P.; Alexandrova, A. N. Polarization effects for hydrogen-bonded complexes of substituted phenols with water and chloride ion. *Journal of chemical theory and computation* **2007**, *3*, 1987–1992.

- (20) Tkatchenko, A.; DiStasio Jr, R. A.; Car, R.; Scheffler, M. Accurate and efficient method for many-body van der Waals interactions. *Physical review letters* **2012**, *108*, 236402.
- (21) Tkatchenko, A.; Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Physical review letters* **2009**, *102*, 073005.
- (22) Hartmann, C.; Banisch, R.; Sarich, M.; Badowski, T.; Schütte, C. Characterization of rare events in molecular dynamics. *Entropy* **2014**, *16*, 350–376.
- (23) Anderson, J. B. Predicting rare events in molecular dynamics. *Advances in Chemical Physics* **1995**, *91*, 381–432.
- (24) Passerone, D.; Parrinello, M. Action-derived molecular dynamics in the study of rare events. *Physical Review Letters* **2001**, *87*, 108302.
- (25) Zwier, M. C.; Chong, L. T. Reaching biological timescales with all-atom molecular dynamics simulations. *Current opinion in pharmacology* **2010**, *10*, 745–752.
- (26) Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q. Enhanced sampling in molecular dynamics. *The Journal of chemical physics* **2019**, *151*, 070902.
- (27) Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E. Long-time-step molecular dynamics through hydrogen mass repartitioning. *Journal of chemical theory and computation* **2015**, *11*, 1864–1874.
- (28) Tuckerman, M. E.; Martyna, G. J.; Berne, B. J. Molecular dynamics algorithm for condensed systems with multiple time scales. *The Journal of chemical physics* **1990**, *93*, 1287–1291.
- (29) Feenstra, K. A.; Hess, B.; Berendsen, H. J. Improving efficiency of large time-scale

- molecular dynamics simulations of hydrogen-rich systems. *Journal of computational chemistry* **1999**, *20*, 786–798.
- (30) Laio, A.; Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics* **2008**, *71*, 126601.
- (31) Bussi, G.; Laio, A.; Parrinello, M. Equilibrium free energies from nonequilibrium metadynamics. *Physical review letters* **2006**, *96*, 090601.
- (32) Valsson, O.; Parrinello, M. Variational approach to enhanced sampling and free energy calculations. *Physical review letters* **2014**, *113*, 090601.
- (33) Chodera, J. D.; Noé, F. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology* **2014**, *25*, 135–144.
- (34) Husic, B. E.; Pande, V. S. Markov state models: From an art to a science. *Journal of the American Chemical Society* **2018**, *140*, 2386–2396.
- (35) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters* **1999**, *314*, 141–151.
- (36) Yang, L.; Qin Gao, Y. A selective integrated tempering method. *The Journal of chemical physics* **2009**, *131*, 12B606.
- (37) Miao, Y.; McCammon, J. A. *Annual reports in computational chemistry*; Elsevier, 2017; Vol. 13; pp 231–278.
- (38) Miao, Y. Acceleration of biomolecular kinetics in Gaussian accelerated molecular dynamics. *The Journal of chemical physics* **2018**, *149*, 072308.
- (39) Stelzl, L. S.; Hummer, G. Kinetics from replica exchange molecular dynamics simulations. *Journal of chemical theory and computation* **2017**, *13*, 3927–3935.

- (40) Buchete, N.-V.; Hummer, G. Peptide folding kinetics from replica exchange molecular dynamics. *Physical Review E* **2008**, *77*, 030902.
- (41) Tiwary, P.; Parrinello, M. From metadynamics to dynamics. *Physical review letters* **2013**, *111*, 230602.
- (42) Yang, L.; Liu, C.-W.; Shao, Q.; Zhang, J.; Gao, Y. Q. From thermodynamics to kinetics: enhanced sampling of rare events. *Accounts of chemical research* **2015**, *48*, 947–955.
- (43) Zuckerman, D. M.; Chong, L. T. Weighted ensemble simulation: review of methodology, applications, and software. *Annual review of biophysics* **2017**, *46*, 43–57.
- (44) Votapka, L. W.; Stokely, A. M.; Ojha, A. A.; Amaro, R. E. SEEKR2: Versatile Multiscale Milestoning Utilizing the OpenMM Molecular Dynamics Engine. *Journal of Chemical Information and Modeling* **2022**,
- (45) Jagger, B. R.; Ojha, A. A.; Amaro, R. E. Predicting ligand binding kinetics using a Markovian milestoning with voronoi tessellations multiscale approach. *Journal of Chemical Theory and Computation* **2020**, *16*, 5348–5357.
- (46) Vanden-Eijnden, E.; Venturoli, M. Markovian milestoning with Voronoi tessellations. *The Journal of chemical physics* **2009**, *130*, 194101.
- (47) Cérou, F.; Guyader, A. Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications* **2007**, *25*, 417–443.
- (48) L’Ecuyer, P.; Demers, V.; Tuffin, B. Splitting for rare-event simulation. Proceedings of the 2006 winter simulation conference. 2006; pp 137–148.
- (49) Cérou, F.; Del Moral, P.; Furon, T.; Guyader, A. Sequential Monte Carlo for rare event estimation. *Statistics and computing* **2012**, *22*, 795–808.

- (50) Van Erp, T. S.; Bolhuis, P. G. Elaborating transition interface sampling methods. *Journal of computational Physics* **2005**, *205*, 157–181.
- (51) Ahn, S.-H.; Ojha, A. A.; Amaro, R. E.; McCammon, J. A. Gaussian-Accelerated Molecular Dynamics with the Weighted Ensemble Method: A Hybrid Method Improves Thermodynamic and Kinetic Sampling. *Journal of Chemical Theory and Computation* **2021**, *17*, 7938–7951.
- (52) Wu, H.; Noé, F. Variational approach for learning Markov processes from time series data. *Journal of Nonlinear Science* **2020**, *30*, 23–66.
- (53) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436–444.
- (54) Bacci, M.; Caffisch, A.; Vitalis, A. On the removal of initial state bias from simulation data. *The Journal of Chemical Physics* **2019**, *150*, 104105.
- (55) Russo, J. D.; Zhang, S.; Leung, J. M.; Bogetti, A. T.; Thompson, J. P.; DeGrave, A. J.; Torrillo, P. A.; Pratt, A.; Wong, K. F.; Xia, J., et al. WESTPA 2.0: High-performance upgrades for weighted ensemble simulations and analysis of longer-timescale applications. *Journal of Chemical Theory and Computation* **2022**, *18*, 638–649.
- (56) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics* **2011**, *134*, 174105.
- (57) Bowman, G. R.; Bolin, E. R.; Hart, K. M.; Maguire, B. C.; Marqusee, S. Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. *Proceedings of the National Academy of Sciences* **2015**, *112*, 2734–2739.
- (58) Wu, H.; Paul, F.; Wehmeyer, C.; Noé, F. Multiensemble Markov models of molecular thermodynamics and kinetics. *Proceedings of the National Academy of Sciences* **2016**, *113*, E3221–E3230.

- (59) Schütte, C.; Fischer, A.; Huisinga, W.; Deuffhard, P. A direct approach to conformational dynamics based on hybrid Monte Carlo. *Journal of Computational Physics* **1999**, *151*, 146–168.
- (60) Schwantes, C. R.; Pande, V. S. Modeling molecular kinetics with tICA and the kernel trick. *Journal of chemical theory and computation* **2015**, *11*, 600–608.
- (61) Boninsegna, L.; Gobbo, G.; Noé, F.; Clementi, C. Investigating molecular kinetics by variationally optimized diffusion maps. *Journal of chemical theory and computation* **2015**, *11*, 5947–5960.
- (62) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *The Journal of chemical physics* **2013**, *139*, 07B604_1.
- (63) Noé, F.; Nuske, F. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Modeling & Simulation* **2013**, *11*, 635–655.
- (64) Williams, M. O.; Kevrekidis, I. G.; Rowley, C. W. A data-driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science* **2015**, *25*, 1307–1346.
- (65) Korda, M.; Putinar, M.; Mezić, I. Data-driven spectral analysis of the Koopman operator. *Applied and Computational Harmonic Analysis* **2020**, *48*, 599–629.
- (66) Korda, M.; Mezić, I. Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control. *Automatica* **2018**, *93*, 149–160.
- (67) Mezić, I. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics* **2005**, *41*, 309–325.

- (68) Koopman, B. O. Hamiltonian systems and transformation in Hilbert space. *Proceedings of the national academy of sciences of the united states of america* **1931**, *17*, 315.
- (69) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nature communications* **2018**, *9*, 1–11.
- (70) Doerr, S.; Harvey, M.; Noé, F.; De Fabritiis, G. HTMD: high-throughput molecular dynamics for molecular discovery. *Journal of chemical theory and computation* **2016**, *12*, 1845–1852.
- (71) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A software package for estimation, validation, and analysis of Markov models. *Journal of chemical theory and computation* **2015**, *11*, 5525–5542.
- (72) Harrigan, M. P.; Sultan, M. M.; Hernández, C. X.; Husic, B. E.; Eastman, P.; Schwantes, C. R.; Beauchamp, K. A.; McGibbon, R. T.; Pande, V. S. MSMBuilder: statistical models for biomolecular dynamics. *Biophysical journal* **2017**, *112*, 10–15.
- (73) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical journal* **2015**, *109*, 1528–1532.
- (74) Schmid, P. J. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics* **2010**, *656*, 5–28.
- (75) Wu, H.; Nüske, F.; Paul, F.; Klus, S.; Koltai, P.; Noé, F. Variational Koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations. *The Journal of chemical physics* **2017**, *146*, 154104.

- (76) Molgedey, L.; Schuster, H. G. Separation of a mixture of independent signals using time delayed correlations. *Physical review letters* **1994**, *72*, 3634.
- (77) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *The Journal of chemical physics* **2007**, *126*, 04B616.
- (78) Husic, B. E.; Pande, V. S. Ward clustering improves cross-validated Markov state models of protein folding. *Journal of chemical theory and computation* **2017**, *13*, 963–967.
- (79) Kube, S.; Weber, M. A coarse graining method for the identification of transition rates between molecular conformations. *The Journal of chemical physics* **2007**, *126*, 024103.
- (80) Yao, Y.; Cui, R. Z.; Bowman, G. R.; Silva, D.-A.; Sun, J.; Huang, X. Hierarchical Nyström methods for constructing Markov state models for conformational dynamics. *The Journal of chemical physics* **2013**, *138*, 05B602.1.
- (81) Huber, G. A.; Kim, S. Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophysical journal* **1996**, *70*, 97–110.
- (82) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. The “weighted ensemble” path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *The Journal of chemical physics* **2010**, *132*, 054107.
- (83) Copperman, J.; Zuckerman, D. M. Accelerated estimation of long-timescale kinetics from weighted ensemble simulation via non-markovian “microbin” analysis. *Journal of chemical theory and computation* **2020**, *16*, 6763–6775.
- (84) Donyapour, N.; Roussey, N. M.; Dickson, A. REVO: Resampling of ensembles by variation optimization. *The Journal of Chemical Physics* **2019**, *150*, 244112.

- (85) Aristoff, D. Analysis and optimization of weighted ensemble sampling. *ESAIM: Mathematical Modelling and Numerical Analysis* **2018**, *52*, 1219–1238.
- (86) Zwier, M. C.; Adelman, J. L.; Kaus, J. W.; Pratt, A. J.; Wong, K. F.; Rego, N. B.; Suárez, E.; Lettieri, S.; Wang, D. W.; Grabe, M., et al. WESTPA: An interoperable, highly scalable software package for weighted ensemble simulation and analysis. *Journal of chemical theory and computation* **2015**, *11*, 800–809.
- (87) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham III, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications* **1995**, *91*, 1–41.
- (88) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D., et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology* **2017**, *13*, e1005659.
- (89) van der Spoel, D.; Lindahl, E. Brute-force molecular dynamics simulations of villin headpiece: comparison with NMR parameters. *The Journal of Physical Chemistry B* **2003**, *107*, 11178–11187.
- (90) Mironov, V.; Alexeev, Y.; Mulligan, V. K.; Fedorov, D. G. A systematic study of minima in alanine dipeptide. *Journal of Computational Chemistry* **2019**, *40*, 297–309.
- (91) Rosky, P. J.; Karplus, M. Solvation. A molecular dynamics study of a dipeptide in water. *Journal of the American Chemical Society* **1979**, *101*, 1913–1937.
- (92) Sinko, W.; Miao, Y.; de Oliveira, C. A. F.; McCammon, J. A. Population based

- reweighting of scaled molecular dynamics. *The Journal of Physical Chemistry B* **2013**, *117*, 12759–12768.
- (93) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation* **2015**, *11*, 3696–3713.
- (94) MacKerell Jr, A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S., et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- (95) Honda, S.; Yamasaki, K.; Sawada, Y.; Morii, H. 10 residue folded peptide designed by segment statistics. *Structure* **2004**, *12*, 1507–1518.
- (96) Satoh, D.; Shimizu, K.; Nakamura, S.; Terada, T. Folding free-energy landscape of a 10-residue mini-protein, chignolin. *FEBS letters* **2006**, *580*, 3422–3426.
- (97) Feig, M.; Im, W.; Brooks III, C. L. Implicit solvation based on generalized Born theory in different dielectric environments. *The Journal of chemical physics* **2004**, *120*, 903–911.
- (98) Onufriev, A. V.; Case, D. A. Generalized Born implicit solvent models for biomolecules. *Annual review of biophysics* **2019**, *48*, 275–296.
- (99) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Structure, Function, and Bioinformatics* **2004**, *55*, 383–394.
- (100) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. Molecular simulation of ab initio protein folding for a millisecond folder NTL9 (1- 39). *Journal of the American Chemical Society* **2010**, *132*, 1526–1528.

- (101) Tian, C.; Kasavajhala, K.; Belfon, K. A.; Raguetta, L.; Huang, H.; Miguez, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q., et al. ff19SB: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *Journal of chemical theory and computation* **2019**, *16*, 528–552.
- (102) Bussi, G.; Laio, A. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics* **2020**, *2*, 200–212.
- (103) Fu, H.; Chen, H.; Wang, X.; Chai, H.; Shao, X.; Cai, W.; Chipot, C. Finding an optimal pathway on a multidimensional free-energy landscape. *Journal of Chemical Information and Modeling* **2020**, *60*, 5366–5374.
- (104) Ray, D.; Stone, S. E.; Andricioaei, I. Markovian Weighted Ensemble Milestoning (M-WEM): Long-time Kinetics from Short Trajectories. *Journal of Chemical Theory and Computation* **2021**,
- (105) Miao, Y.; Bhattarai, A.; Wang, J. Ligand Gaussian accelerated molecular dynamics (LiGaMD): Characterization of ligand binding thermodynamics and kinetics. *Journal of chemical theory and computation* **2020**, *16*, 5526–5547.
- (106) Wu, H.; Mey, A. S.; Rosta, E.; Noé, F. Statistically optimal analysis of state-discretized trajectory data from multiple thermodynamic states. *The Journal of Chemical Physics* **2014**, *141*, 12B629_1.

TOC Graphic

