MLP$_{fold}$: Identification of transition state ensembles in molecular dynamics simulations using machine learning

Preetham Venkatesh
Department of Biochemistry
University of Washington
preetham@uw.edu

Govardhan Reddy
Solid State Chemistry Unit
Indian Institute of Science
greddy@iisc.ac.in

Abstract

Molecular dynamics simulations generate a large amount of raw data, often requiring computationally expensive analysis to extract thermodynamic information. Here, we propose a method, MLP$_{fold}$, to identify the transition state ensemble of a system through an automated labeling process and supervised learning using a simple multilayer perceptron (MLP). This method replicates the conventional P_{fold} calculation but without running any additional simulations. MLP$_{fold}$ was tested on numerous model potentials and Brownian dynamics simulation of the Ubiquitin hairpin and shows promise in predicting committor probabilities and identifying transition states.

1 Introduction

Molecular dynamics (MD) simulations are a powerful tool for studying biomolecular processes such as protein folding. An important objective when performing these simulations is identifying the transition state ensemble (TSE), *i.e.* the set of conformations that populate the top of the folding free energy barrier, which are least stable and equally likely to fold and unfold. In protein folding, the TSE is essential in describing the dynamics and mechanisms that drive the formation of a well-defined globular structure. However, due to their short lifetime, probing the transition states experimentally is very challenging, which led to the development of computational tools to identify the transition state such as accelerated molecular dynamics [1], transition path sampling [2], string method [3], and the nudged elastic band method [4].

In chemical kinetics, a 'transition coordinate' is defined as the coordinate along which the system progresses most slowly. The "probability of folding" method[5] was developed to translate this abstract concept into a measurable quantity. It proposes performing many short simulations starting from a set of putative transition state conformations. The simulations are stopped if they enter either the folded or unfolded state. The statistical likelihood of folding before unfolding, i.e., the fraction of simulations that entered the folded state, is considered the P_{fold} value for that conformation. It represents the kinetic distance of the conformation to the folded and unfolded states. By definition, the set of conformations with $P_{fold} = 0.5$ constitute the transition state ensemble.

The P_{fold} method has been widely used in analyzing MD simulations since it was first proposed - however, it has some drawbacks that limit its applicability. The computational cost of running many additional simulations can be prohibitive, especially for larger systems, and identifying a set of putative conformations can be challenging and prone to bias. Therefore, there is a need for a method that identifies transition states without requiring additional simulations and can evaluate all conformations sampled in a trajectory. Here, we propose a new method, MLP$_{fold}$, where we use a single trajectory with multiple transitions to generate a training data set, then train a simple

Machine Learning for Structural Biology Workshop, NeurIPS 2022.

neural network to predict outcomes of a simulation given the positions and velocities of atoms in a conformation. The method shows promise in identifying transition states in simulations of a toy model and the Ubiquitin hairpin.

2 Methods

2.1 2D Toy Model

2.1.1 2D Model Potential

The model potentials tested have two degrees of freedom, x and y, and has two minimas governed by the potential U (x,y)[6] given by

$$U(x,y) = W(x^6 + y^6) - f1 * G(x, x_1, \sigma_1) * G(y, y_1, \sigma_1) - f2 * G(x, x_2, \sigma_2) * G(y, y_2, \sigma_2) \quad (1)$$

where f is the depth factor and,

$$G(x, x_0, \sigma) = e^{-\frac{x-x_0}{2\sigma^2}} \quad (2)$$

The parameter W controls the scale of the potential, and the parameters x_1, y_1, x_2, y_2 determine the location of minimas, and $\sigma_1, \sigma_2, f1, f2$ control the shape of the two Gaussian wells around the minimas. The values of these parameters used for the different model potentials are in Table A1. The simulation is initialized randomly in the region surrounding the minimas.

2.1.2 Training Dataset and Neural Network Training

A training dataset is built where each data point corresponds to a single frame of the trajectory. The input features are the x and y components of the position and velocity, *i.e.* the input vector is (x,y,v_x, v_y). A basin is defined as the region surrounding a minima with an energy difference of less than k_BT, which is the kinetic energy available to the particle in the 2D potential. The output label is the basin that the particle first enters from this frame in the trajectory. A flowchart showing this labeling schematic is shown in Figure A1. Every tenth data point is held out in a validation dataset.

A simple MLP with two hidden layers of size 8 and 4, respectively is used for the two-minima model potentials. The outer layer predicts the probability of the model entering state 1 or state 2, and the state with the higher probability is picked as the outcome for the statistical-likelihood calculation. ReLU activation function is applied before a forward pass from the hidden layers. The model was trained for 1000 epochs using an Adam optimizer [7] with a learning rate of 1e-3 and Adam hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $eps = 1e-8$. Cross-entropy loss was used to train the model. The training curve is shown in Figure A2.

2.1.3 Committor Probability Approximation

We fit the x and y components of all the velocities in the trajectory to a normal distribution. For each position in the dataset, 500 random velocities were sampled from this distribution. The neural network then predicted the outcome for each pair of positions and sampled velocities. The predicted committor probability for state 1 for a given conformation is the statistical likelihood of entering state 1 ($n_{state1}/(n_{state1} + n_{state2})$), where n_{statei} is the number of predictions the model makes where the particle enters state i starting from this conformation). Transition states were then identified as positions where the committor probability is 0.5 ± 0.02. The deviation tolerated from 0.5 is different for other model potentials, and is shown on the respective figures.

2.2 Ubiquitin hairpin

2.2.1 SOP-SC Model

The first 18 residues of the N-terminal hairpin of Ubiquitin (PDB ID:1UBQ [8]) were modeled using a coarse-grained Self Organized Polymer-Side Chain (SOP-SC) model [9], as described previously

[10]. The model represents each amino acid residue using two interaction centers (beads) located at the C_α position and the center of mass of the side chain, which interact using a residue-dependent statistical potential. The Brownian dynamics simulation (described in Section A.1) was run at 85 K.

2.2.2 Training Dataset

Each conformation in the simulation was described using two parameters: end-to-end distance (R_{ee}) and fraction of native contacts (f_{tot}). These two collective variables were binned into 100 and 50 equally spaced bins, respectively, and a free energy surface was generated by calculating the probability of finding the system in a given bin (Figure A9b).

The folded, unfolded, and intermediate states were defined by identifying local minimas in the free energy surface. A basin was defined as the region surrounding the minima with an energy difference less than $0.3k_BT$. The thermal energy for the system is $1.5k_BT$, but since the intermediate minima is shallower than that, it was set at $0.3k_BT$. The outcome of each conformation was recorded as described for the toy model. Each conformation was described using the velocities of the 36 beads and pairwise euclidean distances between the beads, which was linearized into a 738-length feature vector, comprising of $36 * 3 = 108$ velocities and 630 pairwise distances. We used pairwise distances instead of positions to keep the input features rotation and translation invariant.

2.2.3 Neural Network Training

The neural network architecture comprised two hidden layers with 512 and 128 heads and one output layer comprising three heads corresponding to the three states. ReLU activation was applied before each forward pass. The model was trained using an Adam optimizer [7] with a learning rate of 1e-3 and Adam hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $eps = 1e - 8$. The model was trained for 500 epochs. A weighted cross-entropy loss function was used, where the weights were a normalized inverse of the number of instances of a class. This was done as the intermediate state had low sampling.

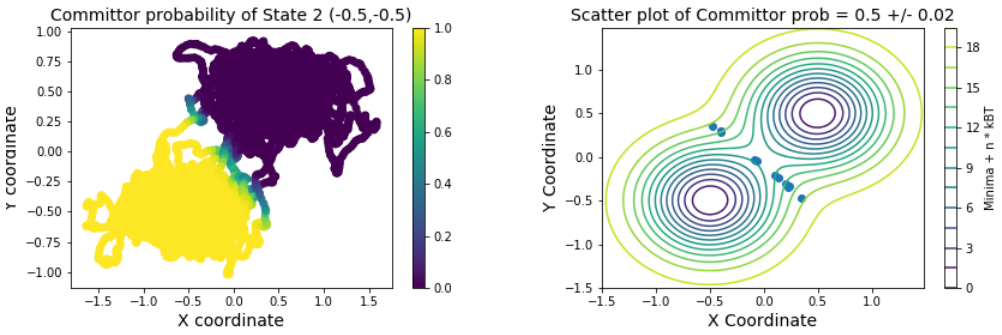
2.2.4 Committor Probability Approximation

The x, y and z velocities were fit to a Maxwell-Boltzmann distribution. For each conformation in the dataset, 500 initial velocities were generated from this distribution. The neural network then predicted the outcome for each set of pairwise distances and velocities. The statistical likelihood of the predictions is the predicted committor probability for each state, and the transition states between any two states are defined as conformations where the committor probability is 0.5 $\pm$ 0.05 for both states.

3 Results

3.1 MLP$_{fold}$ identifies states that lie on the saddle region of potential

Figure 1 shows the committor probability and transition states identified using the above method for the 2D toy model. The regions surrounding the minimas at $(0.5, 0.5)$ and $(-0.5, -0.5)$ have their committor probability of entering state 2 as 0 and 1.0, respectively. This is expected, as the Gaussian deposits in the potential will guide the particle into the basin if the simulation is started in this region as the kinetic energy is not sufficient to exit the basin. The system exhibits a sharp change in its committor probability near the saddle region, $y = -x$, for the given potential. Here, the outcome of the simulation is highly sensitive to the velocities. For a particle at $(0, 0)$, if the velocity is in the direction of state 2, it will enter state 2 and vice versa for state 1. Since the velocity distribution has a mean of 0, a particle in this saddle region is equally likely to enter either of the states. Therefore, the transition state ensemble for this system is the $y = -x$ line. Figure 1 shows that our predicted TSE lies along this line, demonstrating our method's capability in identifying transition states. The prediction is also robust to the choice of deviation from P = 0.5, as demonstrated in Figure A3. We also show that MLP$_{fold}$ can identify the transition state at different sampling temperatures, different basin heights and widths, and position of minimas (see the Appendix).

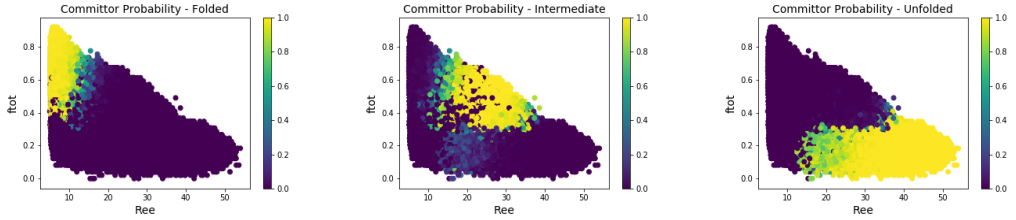


(a) (b)

Figure 1: a) Committor probability of visiting the basin characterized by a minima at (-0.5,-0.5) for each coordinate visited during trajectory at $\beta = 9.5$. b) Scatter plot of coordinates identified as transition states defined as having a committor probability of 0.5 ± 0.02 projected onto a contour plot of the potential.

3.2 Predicted committor probabilities distinguish the three states in the ubiquitin hairpin folding energy surface

Figure 2 shows the predicted committor probabilities for the ubiquitin hairpin simulation. For each of the three states, the regions where their committor probabilities are 1.0 are clearly seen in yellow. In addition, it also reveals potential pathways for the folding of the β-hairpin system. The conformations that correspond to a committor probability of approximately 0.5 for the intermediate state can be seen in blue-green in Figure 2b, which gives insight into the system folding pathway. Additionally, a wide region of midway committor probabilities are seen for both the unfolded and intermediate state between $12.0 \leq R_{ee} \leq 22.0$ and $0 \leq f_{tot} \leq 0.3$. This region is not a basin in the free energy surface and would be an interesting region to explore in future MD simulations. It also reveals conformations that are approximately equidistant kinetically from all the three states ($R_{ee} \approx 11.0$ and $f_{tot} \approx 0.3$).



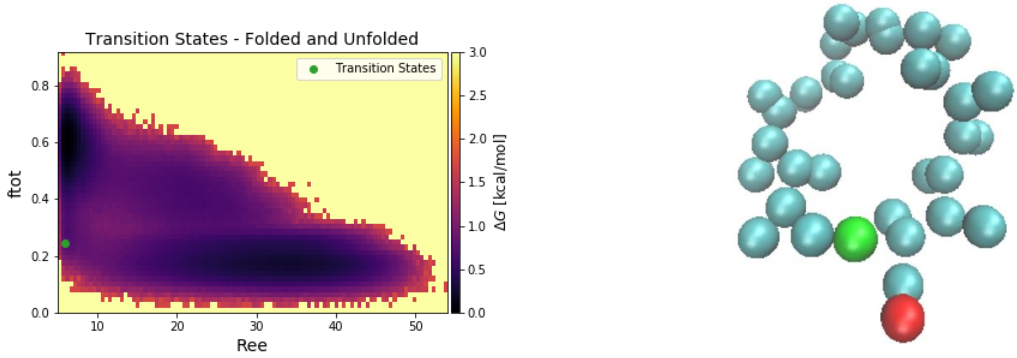
(a) (b) (c)

Figure 2: Calculated committor probabilities for a) folded, b) intermediate, and c) unfolded state of ubiquitin β-hairpin projected onto R_{ee} and f_{tot}.

3.3 Structural analysis of identified transition states provides insight into folding pathways

We sought to characterize the conformations that are identified as the transition state by the model and use it to draw inferences about the folding pathway of the hairpin. Only a single conformation was found to have a committor probability of 0.5 ± 0.05 for both the folded and unfolded state. Interestingly, it does not lie on the barrier separating the two states when viewed on a free energy surface (R_{ee} 10, f_{tot} 0.3) but rather on a narrow path that connects the two states. This path has the same R_{ee} as the folded state but low f_{tot}, suggesting that the N and C terminus coming together was the first step when the system transitions directly from the unfolded to folded state without going through an intermediate. Identification of this transition state highlights the utility of using committor probabilities over simply picking the peak of the barrier separating two states in a projected free energy surface. The predictions for the transition states between the folded and intermediate states,

and especially the unfolded and intermediate states, are less reliable likely due to the low sampling of the intermediate state (Figure A10).



(a) (b)

Figure 3: Identified transition states (green) and a representative structure for the TS between folded and unfolded states (a and b). The green and red beads in the structure represent the C_α atoms of the first and last residue respectively to show the end-to-end distance.

4 Discussion and Future Work

We developed and tested a novel algorithm that analyses a single trajectory with multiple transitions to create a training dataset for a neural network model that learns relationships between inputs comprising of positions and velocities to predict simulation outcomes. We then utilize this trained model to calculate committor probabilities in a manner identical to the well-established P_{fold} method but with the major advantage of not requiring any additional simulations.

An exhaustive committor probability calculation as done here would not be possible using the traditional P_{fold} method, which would have required running 500,000,000 simulations (500 trials * 100,000 conformations) to obtain a comparable committor probability landscape. To the best of our knowledge, this is the first such work that seeks to remove the need for additional simulations in P_{fold} analysis. The CPU training time for a model is in the order of minutes to hours, and the inference time for a single initialization state is less than a millisecond. The quick inference time enables committor probability predictions with higher statistical significance as we can sample a greater number of velocities per conformation.

One of the limitations of our work is that for some systems, it might not be possible to have multiple transitions in a single trajectory. We are working to understand how the number of transitions affects the method's performance. We are also thinking about better data visualization methods for the committor probability landscape, as a scatter plot for every conformation can hide information due to overlaps. To demonstrate the advantage of this method for complex systems, we are currently working on applying this method to a simulation of the full ubiquitin protein. The low sampling of a state also affects the model performance: this can be seen in the predictions of the transition states involving the intermediate state (Figure A10). Increased sampling or an improved loss function are some of the ways we plan to mitigate this. Finally, we plan to compare the accuracy of this method in identifying transition states to the conventional P_{fold} method for both the ubiquitin hairpin and the full protein, as well as with any available experimental data. Nonetheless, MLP$_{fold}$ shows promise as a method to identify transition states in a rapid fashion without needing additional simulations which meets an important need in the community.

4.1 Code availability

The python-based code and jupyter notebooks to reproduce our results for the toy model are available at `https://github.com/preetham-v/MLPfold`.

References

[1] Arthur F Voter. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Physical Review Letters*, 78(20):3908, 1997.

[2] Peter G Bolhuis, David Chandler, Christoph Dellago, and Phillip L Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual review of physical chemistry*, 53(1):291–318, 2002.

[3] E Weinan, Weiqing Ren, and Eric Vanden-Eijnden. String method for the study of rare events. *Physical Review B*, 66(5):052301, 2002.

[4] Hannes Jónsson, Greg Mills, and Karsten W Jacobsen. Nudged elastic band method for finding minimum energy paths of transitions. 1998.

[5] Rose Du, Vijay S Pande, Alexander Yu Grosberg, Toyoichi Tanaka, and Eugene S Shakhnovich. On the transition coordinate for protein folding. *The Journal of chemical physics*, 108(1):334–350, 1998.

[6] Sun-Ting Tsai, En-Jui Kuo, and Pratyush Tiwary. Learning molecular dynamics with simple language model built upon long short-term memory neural network. *Nature communications*, 11(1):1–11, 2020.

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] Senadhi Vijay-Kumar, Charles E Bugg, and William J Cook. Structure of ubiquitin refined at 1.8 åresolution. *Journal of molecular biology*, 194(3):531–544, 1987.

[9] Zhenxing Liu, Govardhan Reddy, Edward P O'Brien, and D Thirumalai. Collapse kinetics and chevron plots from simulations of denaturant-dependent folding of globular proteins. *Proceedings of the National Academy of Sciences*, 108(19):7787–7792, 2011.

[10] Govardhan Reddy and D Thirumalai. Dissecting ubiquitin folding using the self-organized polymer model. *The Journal of Physical Chemistry B*, 119(34):11358–11370, 2015.

A Appendix

A.1 Brownian dynamics simulation of Ubq hairpin

Brownian Dynamics simulations were used to describe the folding kinetics using the equations of motion were integrated using Ermak-McCammon algorithm,

$$\vec{r_i}(t+h) = \vec{r_i}(t) + \frac{h}{\zeta}\vec{F_c} + \vec{\Gamma} \tag{3}$$

where $\vec{\Gamma}$ is a random force with a Gaussian distribution of mean zero and variance $\langle\vec{\Gamma}(h)^2\rangle = \frac{2k_BTh}{\zeta}$. The friction coefficient $\zeta = 50\frac{m}{\tau_H}$ approximately corresponds to the value in water and $h = 0.005\tau_H$. The characteristic unit of length a = 1.0 Å, energy ϵ = 1.0 kcal/mol and mass m = 1.8 x 10^{-22} g which is the typical mass of the bead. The unit of time in the simulations was $\tau_L(=\sqrt{ma^2/\epsilon})$ = 0.51 ps.

A.2 Four-well model potential

The 4-well model has the parameters $W = 1e-3$, the depth factor $f = 1$ and $\sigma = 0.6$ are common for all wells, and the minimas are located at $(1,1)$, $(-0.5,0)$, $(1,-1.5)$, and $(-2,-2)$. A particle of unit mass is simulated to undergo Langevin dynamics in these different potentials.

A.3 Tables and Figures

	W	$f1$	σ	(x_1, y_1)	$f2$	$\sigma 2$	(x_2, y_2)
$\beta = 12.5$	1e-4	2	0.4	(0.5,0.5)	2	0.4	(-0.5,-0.5)
$\beta = 9.5$	1e-4	2	0.4	(0.5,0.5)	2	0.4	(-0.5,-0.5)
$\beta = 7.5$	1e-4	2	0.4	(0.5,0.5)	2	0.4	(-0.5,-0.5)
Different Depths	1e-4	1	0.4	(0.5,0.5)	2	0.4	(-0.5,-0.5)
Different Widths	1e-4	2	0.8	(1,1)	2	0.4	(-0.5,-0.5)
Asymmetric positions	1e-4	2	0.4	(0.5,0.5)	2	0.4	(-0.8,0)

Table A1: Parameters for two-well model potential

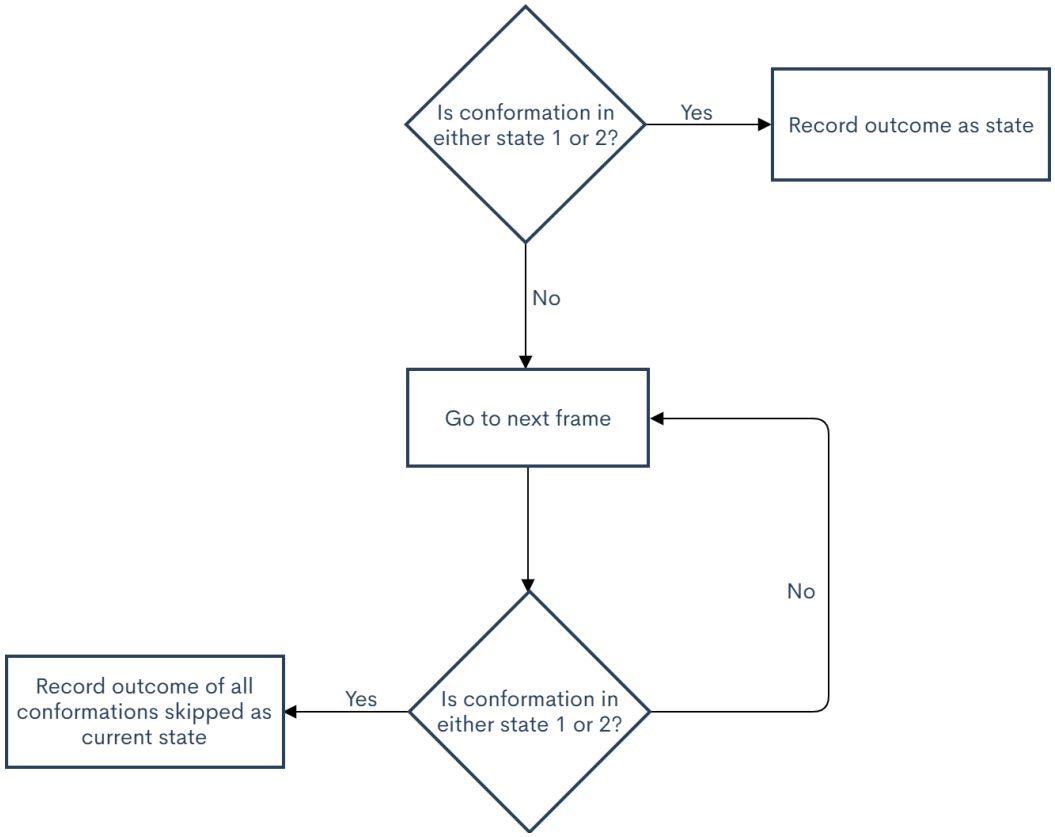


Figure A1: Flowchart to determine outcome labels for conformations in a trajectory. If a given simulation at time step t is already in one of the states, it is recorded as so. If $t + 1$ is not in one of the states, then we go to the next frame, till we find a conformation enters one of the states at time step $t + n$. All conformations from $t + 1$ to $t + n$ are labelled as having their outcome be the same as the one at $t + n$.

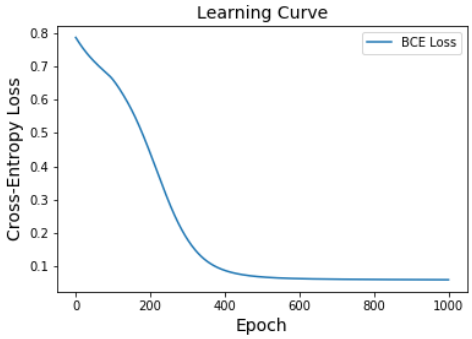
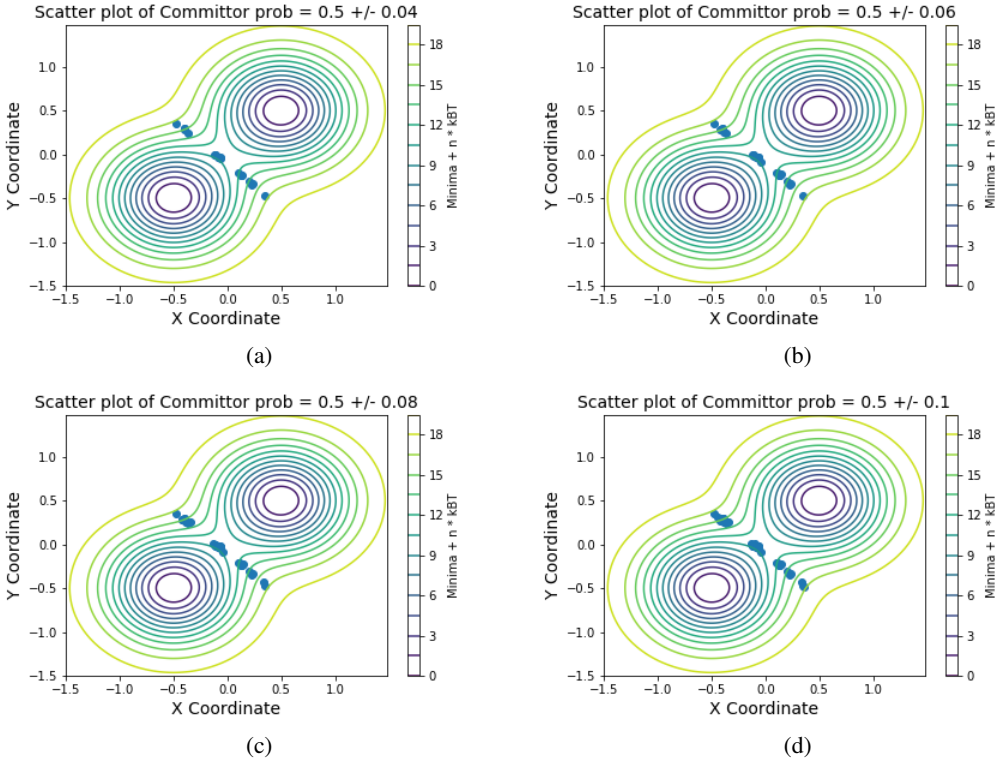


Figure A2: The learning curve of the neural network architecture described previously. The model saturates within 500 epochs, achieving a final loss value of 0.0583 with a final accuracy on training dataset of 98.69%. The corresponding final loss and accuracy of the held out validation dataset is 0.0587 and 98.65% respectively.



(a) (b) (c) (d)

Figure A3: The transition states identified remain in the same saddle point region when deviation up to a) 0.04, b) 0.06, c) 0.08 and d) 0.1 are tolerated. This is a consequence of the sharp behavior of the committor function in this region.

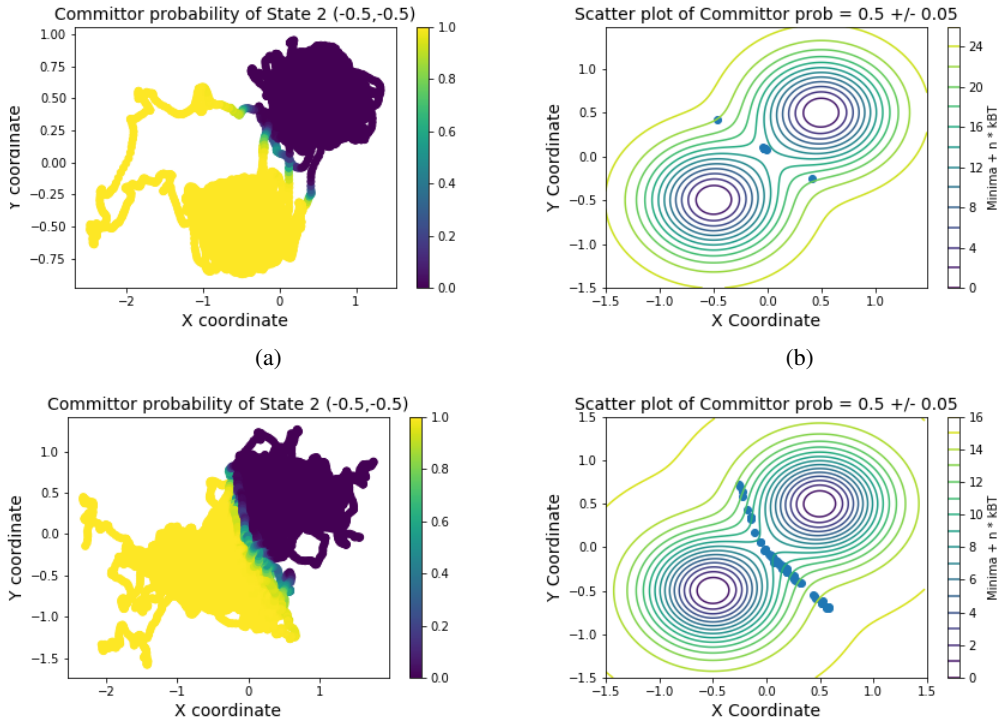


Figure A4: Predicted committor probabilities and transition states at low (a and b) and high (c and d) temperatures. The predicted transition state ensemble lies largely on the saddle region of the potential.

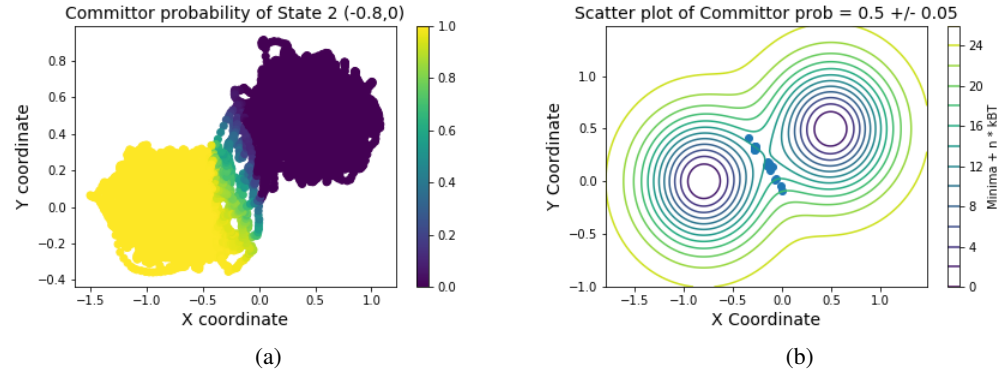


Figure A5: Simulation conducted with potential having identically shaped Gaussian depositions with minimas at $(-0.8, 0)$ and $(0.5, 0.5)$. a) and b) show the predicted committor probability and transition states identified respectively. The TSE is on the saddle region of the potential as expected.

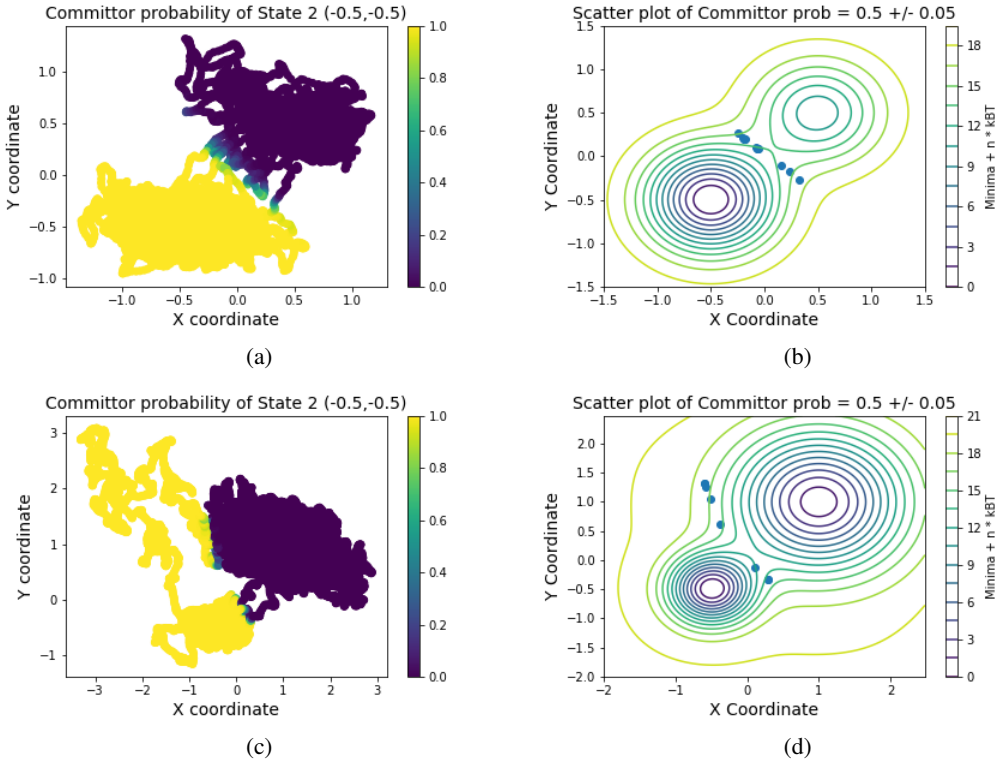


Figure A6: a) Predicted committor probability and b) identified transition states projected on contour of potential for simulations conducted at $\beta = 9.5$ with the basin at $(-0.5, -0.5)$ being twice as deep as the basin at $(0.5, 0.5)$. c) and d) show the same for simulations conducted at $\beta = 9.5$ with the gaussian deposit with minima $(1, 1)$ having the twice the standard deviation as the one at $(-0.5, -0.5)$.

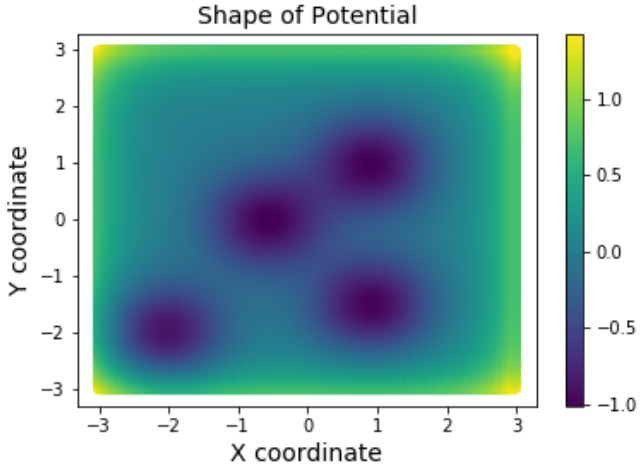
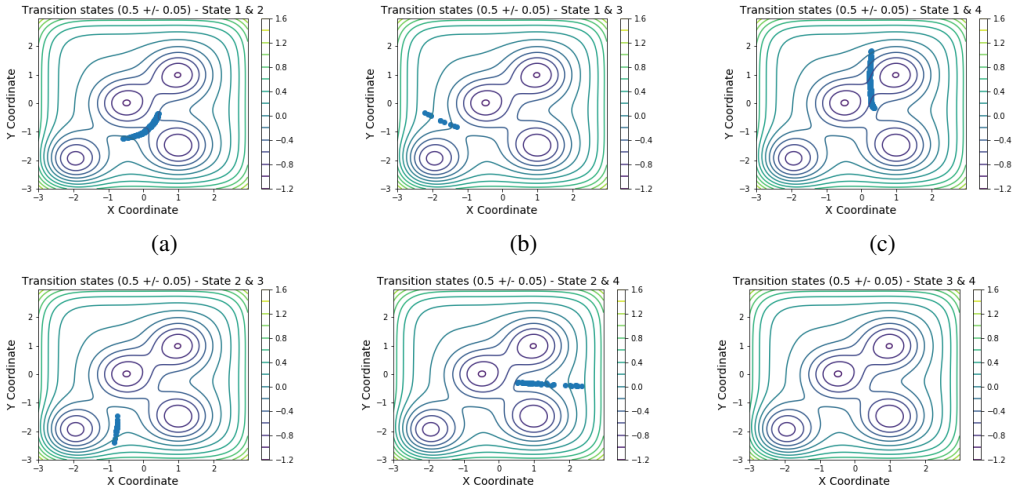


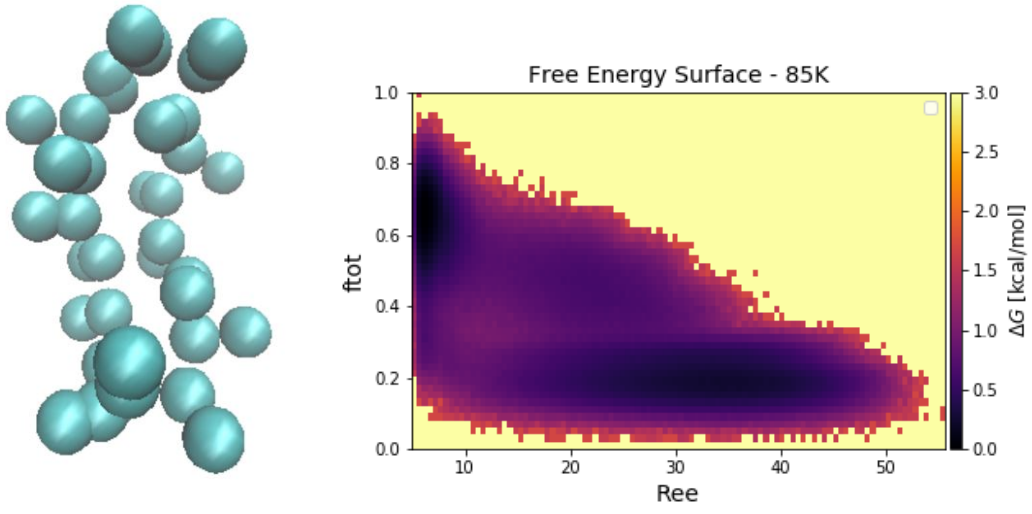
Figure A7: The model potential for the 4-well system



(a) (b) (c)

(d) (e) (f)

Figure A8: Identified transition states for 4-well model between a) State 1 and 2,) State 1 and 3, c) State 1 and 4, d) State 2 and 3, e) State 2 and 4, and f) State 3 and 4. Here, state 1, 2, 3, and 4 refer to the minimas at (-0.5,0), (1,-1.5), (-2,-2) and (1,1) respectively. The transition state between two states is defined as the positions where the committor probability is 0.5 ± 0.05 for the two states. MLP$_{fold}$ identifies transition states that lie in the saddle regions between these minimas, demonstrating potential to identify TSE in multi-state systems. No transition state is identified in f) as there is no direct transition possible between states 3 and 4.



(a) (b)

Figure A9: a) A representative structure of the coarse-grained Ubiquitin β-hairpin. b) 2D Free Energy obtained from the simulation projected on collective variables R_{ee} and f_{tot}.

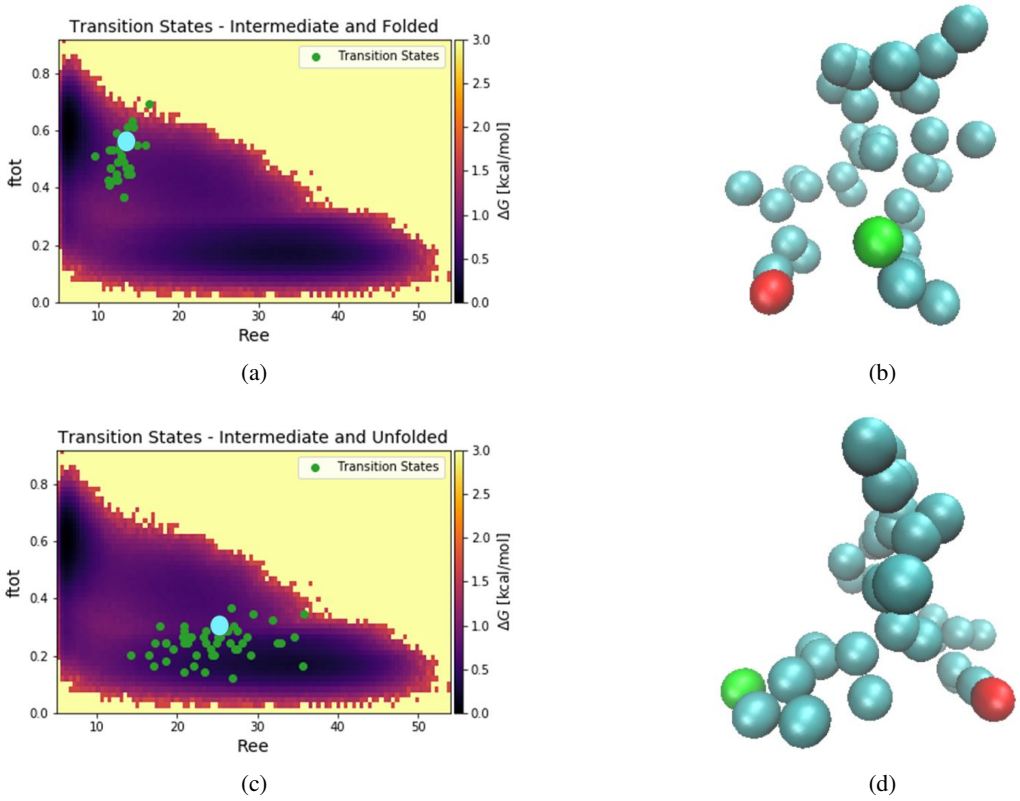


Figure A10: Identified transition states (green) and a representative structure for the TS between Folded and Intermediate (a and b) and Unfolded and Intermediate (c and d). The green and red beads in the structure represent the C-alpha atoms of the first and last residue respectively to show the end-to-end distance. Due to low sampling of the intermediate state, the predictions for it are relatively poor resulting in a scattered Transition State Ensemble that includes conformations from the Unfolded state.