Exploring Transition States of Protein Conformational Changes via Out-of-Distribution Detection in the Hyperspherical Latent Space

Bojun Liu[1,2], Jordan G. Boysen[1], Ilona Christy Unarta[1,2], Xuefeng Du[3], Yixuan Li[2,3], Xuhui Huang[1,2*]

[1]Department of Chemistry, Theoretical Chemistry Institute, University of Wisconsin-Madison, Madison, WI, 53706, USA
[2]Data Science Institute, University of Wisconsin-Madison, Madison, WI, 53706, USA
[3]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, United States, 53706, USA
[*]To whom correspondence should be addressed. E-mail: xhuang@chem.wisc.edu

Abstract

Identifying transitional states is crucial for understanding protein conformational changes that underlie numerous fundamental biological processes. Markov state models (MSMs) constructed from Molecular Dynamics (MD) simulations have demonstrated considerable success in studying protein conformational changes, which are often associated with rare events transiting over free energy barriers. However, it remains challenging for MSMs to identify the transition states, as they group MD conformations into discrete metastable states and do not provide information on transition states lying at the top of free energy barriers between metastable states. Inspired by recent advances in trustworthy artificial intelligence (AI) for detecting out-of-distribution (OOD) data, we present Transition State identification via Dispersion and vAriational principle Regularized neural neTworks (TS-DART). This deep learning approach effectively detects the transition states from MD simulations using hyperspherical embeddings in the latent space. The key insight of TS-DART is to treat the transition state structures as OOD data, recognizing that the transition states are less populated and exhibit a distributional shift from metastable states. Our TS-DART method offers an end-to-end pipeline for identifying transition states from MD simulations. By introducing a dispersion loss function to regularize the hyperspherical latent space, TS-DART can discern transition state conformations that separate multiple metastable states in an MSM. Furthermore, TS-DART provides hyperspherical latent representations that preserve all relevant kinetic geometries of the original dynamics. We demonstrate the power of TS-DART by applying it to a 2D-potential, alanine dipeptide and the translocation of a DNA motor protein on DNA. In all these systems, TS-DART outperforms previous methods in identifying transition states. As TS-DART integrates the dimensionality reduction, state decomposition, and transition state identification in a unified framework, we anticipate that it will be applicable for studying transition states of protein conformational changes.

1. Introduction

Understanding the transition states of protein conformational changes is crucial for gaining insights into various biological processes, including protein folding, misfolding, gene expression, etc. This understanding also facilitates drug design and enzyme engineering. In protein conformational changes, transition states typically encompass a collection of structures located at the peaks or saddle points of free energy barriers that separate different free energy basins. However, due to the low populations and transient features of these transition state structures, it remains challenging to directly investigate them at atomic resolutions using experimental techniques.

Molecular dynamics (MD) simulations can serve as a powerful approach to complement experimental methods in studying protein conformational changes, as they enable the elucidation of conformational dynamics in a high spatial and time resolution. However, all-atom MD simulations typically operate at femtosecond time-step, posing a challenge in capturing protein conformational changes that usually occur in milliseconds or longer. Markov state models (MSMs)[1-13] have effectively addressed this challenge by integrating multiple short MD trajectories to predict long-time scale dynamics. Despite the significant success of MSM in studying conformational changes over the past decade[14-33], identifying transition state structures remains challenging. In MSMs, dynamic processes are modeled as a series of Markovian transitions among metastable conformational states (or free energy minima) at discrete time intervals (called lag times). Each MD conformation, including those in the transition state region, is therefore assigned to a specific metastable state, complicating the unraveling of transition states.

For an MSM containing a large number of small states (or microstates), one approach to identifying transition state structures within the MSM pipeline is to compute the committor probabilities of these small microstates. Microstates with an equal probability (committor equals 0.5) of reaching the initial and final metastable states can be identified as transition states between the two. However, the efficacy of this method relies significantly on the quality and accuracy of constructing the corresponding microstate MSMs. Additionally, it can only identify transition states between pairs of metastable states one at a time. Recently, a deep learning-based approach, MaxEnt-VAMPNets[34], has been developed to identify transition states structures to facilitate the adaptive sampling. This approach utilizes the state assignment probabilities output from VAMPnets[35] to calculate the Shannon entropy for each MD conformation. It assumes that the conformations with higher Shannon entropy values are more likely to be located at low-probability regions (i.e., the summit of free energy barriers). However, the state assignment probabilities from VAMPnets[35] represent the basis functions that can best linearly reconstruct the system's slowest dynamic modes[36]. It is not guaranteed that they can precisely reflect the probabilities of MD conformations transitioning in or out of metastable states.

In the past decades, a number of other MD simulation-based methods have been developed for identifying the transition states of conformational dynamics. For instance, transition path

sampling (TPS)[37,38] directly uncovers transition states using the committor function derived from the transition path ensemble generated through Monte Carlo sampling. Additionally, a deep reinforcement learning approach, integrating an efficient path sampling method called enhanced sampling of reactive trajectories (ESoRT)[39,40], has been employed to identify transition states by framing the problem as a shooting game.[41] Rather than directly capturing transition states, more approaches have been developed to extract the optimal reaction pathways[42-49] or calculate committor functions[50-53], both of which greatly assist in transition states identification. However, despite the robustness and promise of these methods, they may pose challenges in terms of their high computational cost, or the requirement for prior knowledge and accurate characterization of the initial and final states.

In this work, we aim to tackle the problem of identifying the transition states in the MSM framework from a new perspective by incorporating recent advancements from the out-of-distribution (OOD) detection in trustworthy artificial intelligence (AI). OOD detection[54], an important task for trustworthy AI, has emerged and attracted increasing attention in recent years. The major issue targeted by OOD detection is that the model trained on a specific closed-world dataset, i.e., the in-distribution (ID) data, may make overconfident and wrong predictions on unknown examples, i.e., the OOD data, from the open-world. Therefore, a reliable deep learning model should be able to perform a binary ID v.s. OOD detection classification task, and reject OOD data points. This is especially important when applying deep learning models to safety-critical applications[55-60] such as self-driving cars and rare disease detection tasks. OOD detection, which has not been previously employed in the study of biomolecular conformational changes, bears significant potential in identifying transition states. This is because the conformations at the transition states, located at the free energy barriers, are typically less populated and exhibit a distributional shift compared to the highly populated free energy basins. Consequently, these transition states can be considered as OOD data.

Recently, Ming et al. developed a Compactness and DispErsion Regularized learning framework (CIDER)[61] for detecting OOD images. This was achieved by harnessing latent hyperspherical embeddings that are effectively regularized through the joint optimization of compactness loss and dispersion loss. Specifically, the hyperspherical embeddings ($\boldsymbol{z} \in \mathbb{R}^d$) refers to a set of points in d-dimensional Euclidean space that are located at a constant distance from the center, collectively defining a $(d-1)$-dimensional hypersphere, as illustrated in Fig. 1C. Following this, the compactness loss encourages the tightening of image samples in each class on the hypersphere, while the dispersion loss promotes large angular distances between class prototypes. Consequently, OOD samples are expected to lie between class prototypes and can be detected based on cosine similarity-based measures. The concept introduced by CIDER serves as inspiration for our development of a new deep learning approach aimed at detecting transition state structures in protein conformational changes. In the context of biomolecular dynamics, the metastable free energy basins, akin to class prototypes, should be separated from one another, while the transition state structures, representing OOD samples, are expected to reside in between them. However, a major challenge in directly applying CIDER to protein dynamics is

that CIDER was developed under supervised learning settings, requiring labeled data for computing class prototypes. In contrast, protein conformations in MD trajectories typically lack these ground-truth labels. VAMPnets[35] offers a potential solution by bridging the gap between unsupervised learning and the need for labeled data. Furthermore, the VAMP-2 loss function, which is designed to capture the slowest dynamic modes of the system based on the separation of timescales, can facilitate the compression of metastable conformations within each energy basin, playing a similar role to the compactness loss in CIDER.

In this work, we have developed a novel end-to-end approach called Transition State identification via Dispersion and vAriational principle Regularized neural neTworks (TS-DART) for detecting transition state structures of protein conformational changes from MD simulations. This approach utilizes the penultimate layer of a deep neural network with L2-norm constraint to serve as hyperspherical latent representations of the biomolecular conformations. The loss function comprises two terms: VAMP-2 loss and dispersion loss. By minimizing the VAMP-2 loss, the penultimate hyperspherical embeddings of MD conformations can preserve all pertinent kinetic geometries and be compacted in terms of their kinetic metastability, capturing the slowest dynamics. Subsequently, the optimization of dispersion loss further regularizes the hyperspherical latent space, ensuring the metastable state centers (or free energy minima) are uniformly distributed across the hypersphere. Consequently, all transition state conformations, located between free energy basins, could be simultaneously and automatically identified based on their cosine similarities to the state centers. We have demonstrated the efficacy of our method by applying it to three systems: the 2D Müller potential, alanine dipeptide and the translocation of a DNA motor protein along double-strand DNA (dsDNA). The code of TS-DART is available at https://github.com/xuhuihuang/ts-dart.

2. Results and Discussions

Hyperspherical latent representations at the penultimate layer in TS-DART. The schematic representation of TS-DART's model architecture is shown in Fig. 1A. Different from VAMPnets[35], which directly employs two parallel encoders to utilize transition pairs of MD conformations to produce the SoftMax probabilities of state assignments, TS-DART introduces a novel enhancement in its model architecture. Specifically, it incorporates an additional L2-norm/scale layer at the penultimate layer to extract the hyperspherical latent representations of MD conformations (Fig. 1A). These hyperspherical latent representations can be effectively regularized by the joint optimization of VAMP-2 loss and dispersion loss, enabling the robust transition states identification on the latent hypersphere (Fig. 1C). In particular, the L2-norm/scale layer consists of two parts. The feature vectors at the penultimate layer are first divided by their L2-norms, and then rescaled by a scaling factor γ (Fig. 1B). As a result, the feature embeddings at the penultimate layer are successfully confined on a hypersphere of radius γ, referring to as the hyperspherical latent representations. To illustrate the hyperspherical representations of MD trajectories, we utilize a 2D Müller potential[62] with three minima as an example (Fig. 2A). In the latent space of the trained TS-DART model for the Müller potential

(Fig. 2B), all MD conformations perfectly lie on a hypersphere. Three distinct free energy basins can be clearly identified in this hypersphere, with dashed lines denoting the mean vectors of each basin in Fig. 2B. Furthermore, these three free energy basins in the hypersphere (Fig. 2D) correspond clearly to the three energy basins of the Müller potential (see basin 1-3 in Fig. 2A). Pseudo-code for the TS-DART algorithm is provided in SI Scheme S1, and additional details on training the TS-DART model can be found in the Methods section.

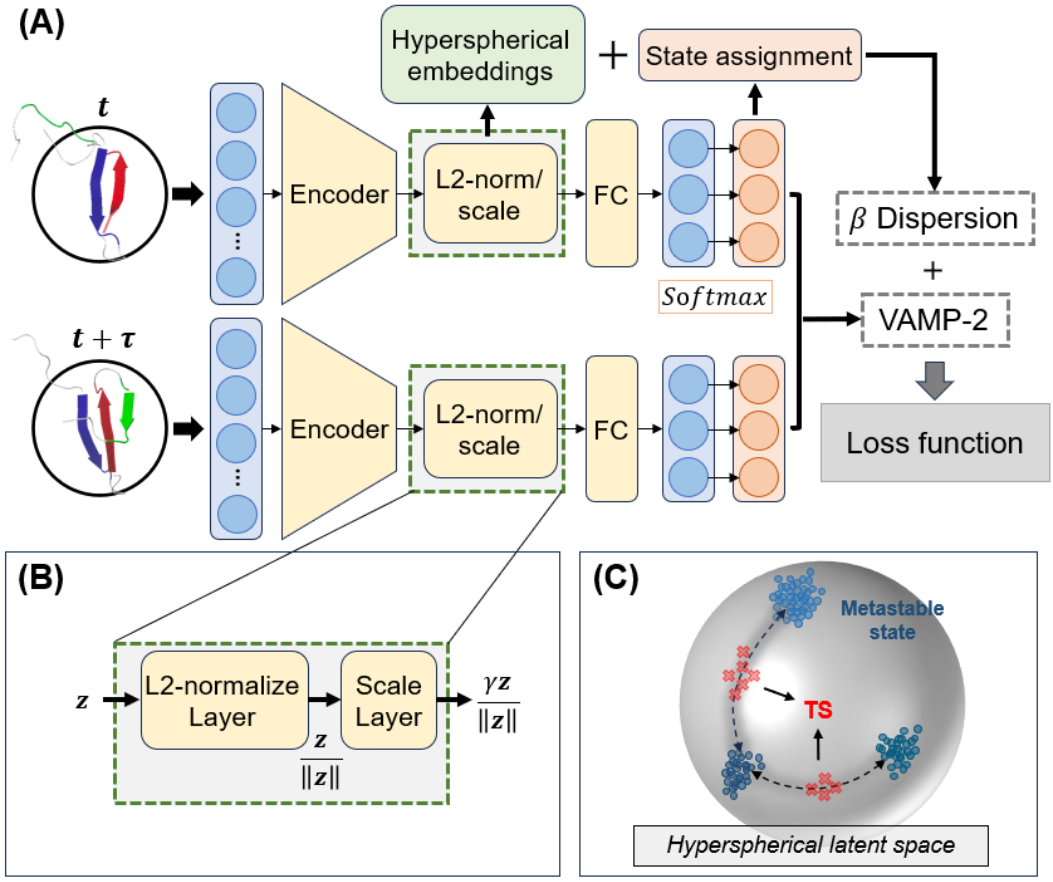

Figure 1. Schematic representation of TS-DART for transition state identification. (a) Overview of the TS-DART framework. (b) Utilization of the L2-norm/scale layer to confine feature embeddings to be within a hypersphere of radius γ. (c) Identification of the transition states in the hyperspherical latent space.

Utilizing VAMP-2 loss to capture the slowest dynamics within the hypersphere. Like in VAMPnets[35], the minimization of the VAMP-2 loss ($\mathcal{L}_{vamp}$ in Eq. 1) ensures our TS-DART model can capture the slowest dynamic modes underlying the conformational changes of interest.

$$\mathcal{L}_{vamp} = -\left\| \bar{\mathbf{C}}_{00}^{-\frac{1}{2}} \bar{\mathbf{C}}_{01} \bar{\mathbf{C}}_{11}^{-\frac{1}{2}} \right\|_F^2 - 1, \tag{1}$$

where $\bar{\mathbf{C}}_{01}$ represents the time-lagged correlation matrix, $\bar{\mathbf{C}}_{00}$ and $\bar{\mathbf{C}}_{11}$ denote the time-instantaneous covariance matrices at time t and $t+\tau$ (τ is the lag time). The details for

computing these correlation matrices are presented in the Methods section. According to this property, the SoftMax outputs from TS-DART can provide optimal state assignments that are aligned with the free energy basins, allowing the on-the-fly labeling of MD conformations during the training. In addition, with the help of the VAMP-2 loss function term, the hyperspherical latent representations in the trained TS-DART model can retain all relevant kinetic geometries and compact the conformations on the hypersphere in terms of their kinetic metastability. This is demonstrated by the latent space representation of the Müller potential (Fig. 2D), where the three identified free energy minima and their order of connection correspond precisely with the arrangements of the three energy basins of the Müller potential (Fig. 2C). Additionally, we observe that the hyperspherical latent space effectively condenses MD conformations into three distinct clusters, each corresponding to a specific metastable free energy basin. For instance, the largest free energy basin (basin 1 in Fig. 2A) has been more significantly compressed compared to the other two basins due to its large metastability, resulting in a small but deep free energy minimum in the latent space (Fig. 2B). We attribute this observation to the presence of a sole fully connected layer extending from the latent bottleneck to the outputs, compelling the latent representations to optimally capture the slowest dynamics through the VAMP-2 loss. This unique design of the penultimate layer for representation learning has already exhibited considerable promise in diverse deep learning fields ranging from computer vision to natural language processing[63,64]. Finally, to elucidate the distinct roles of the VAMP-2 and dispersion terms in the loss function, we conducted a control experiment by omitting the dispersion loss. The aforementioned observations remain the same despite the absence of the dispersion loss (Fig. 2E & 2F). However, it is noteworthy that the three free energy basins exhibit uneven distribution in the latent space when the dispersion loss is excluded. Further discussion on this point will be presented in the next section.

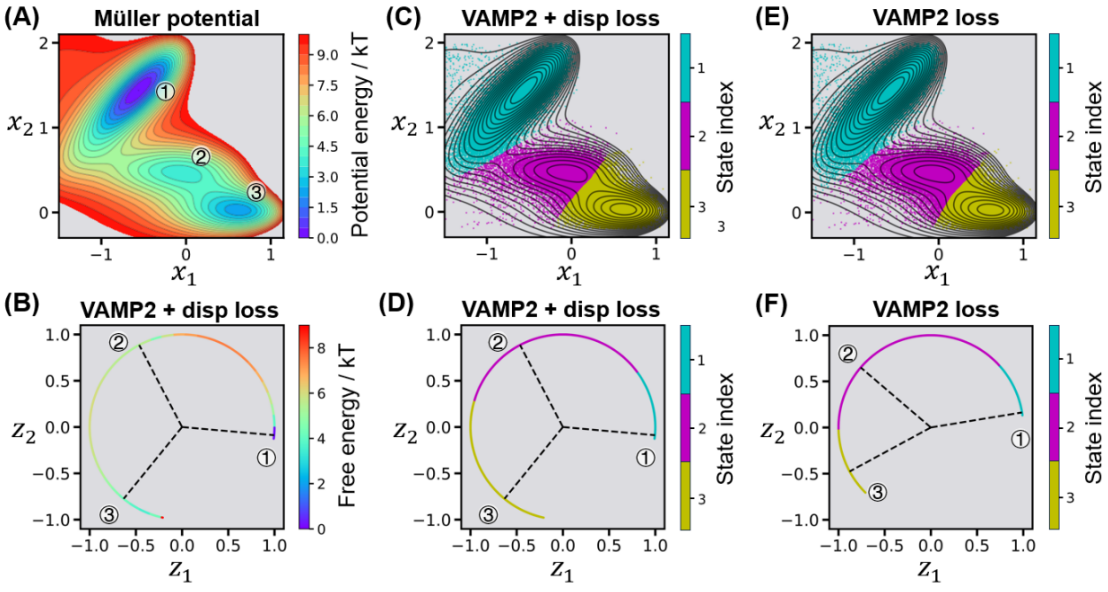

Figure 2. Demonstration of TS-DART on the Müller potential for learning latent hyperspherical representations. (a) The 2D-Müller potential. (b) Projections of MD conformations onto the latent

hyperspherical space, and the free energy ($-kT \ln P(\theta)$) is displayed, where $P(\theta)$ corresponds to the probabilities of MD conformations at the polar angle θ on the hypersphere. The dashed lines indicate the mean vectors of each of the three metastable states. (c) Visualization of the output state assignments (state 1 to 3) of the TS-DART model overlaid on the Müller potential. (d) The same as (b) except that the state assignments rather than potential of mean force in the latent space are displayed. (e) and (f) are the same as (c) and (d), respectively, except that the results of a control experiment that only includes the VAMP-2 loss when training the TS-DART model are shown.

Implementing dispersion loss for uniform latent distribution of metastable state centers and defining an OOD score for transition state identification. The dispersion loss was initially introduced by us and has successfully been employed in OOD detection for image classification tasks[61]. In TS-DART, we introduce the dispersion loss aiming to encourage the state centers (i.e., free energy minima) to be uniformly distributed across the hypersphere by maximizing the angular distances between these centers. For example, in the presence of the dispersion loss (Fig. 2D), the centers of 3 free energy basins of the Müller potential (labeled as state 1-3) are well separated and uniformly distributed in the latent space. In sharp contrast, without the dispersion loss (Fig. 2F), the three state centers exhibit an uneven distribution in the latent space. The dispersion loss is defined as follows:

$$\mathcal{L}_{dis} = \frac{1}{C}\sum_{i=1}^{C} \log \frac{1}{C-1}\sum_{j=1}^{C} \mathbf{1}\{j \neq i\} e^{\boldsymbol{\mu}_i^{\mathrm{T}} \boldsymbol{u}_j/\sigma} \tag{2}$$

where C corresponds to the number of states, $\boldsymbol{\mu}_c$ is a unit vector, representing the mean direction of all conformations (state center) in state c, and σ is a scaling hyperparameter and specifically defined as 0.1. To compute the dispersion loss, it is necessary to first estimate the state center vectors $\{\boldsymbol{\mu}_c\}_{c=1}^{C}$. For the robustness and the efficiency of training, we employ an exponential-moving-average (EMA)[65] method to estimate $\{\boldsymbol{\mu}_c\}_{c=1}^{C}$ on-the-fly and update them frequently during the training (see the Methods section and SI Scheme S1 for details).

After we obtained the regularized hyperspherical latent representations through the joint optimization of VAMP-2 and dispersion loss, the metastable states are compacted and uniformly separated. As a result, transition states in between these metastable states will share equal angular distances to their nearest state centers (Fig. 2D). This prompts us to define an OOD score based on the cosine similarity to quantify the angular distances in the hyperspherical latent space (Eq. (3)), and thus help identify all the transition states automatically and simultaneously.

$$OOD\ score = -\max\{\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\mathcal{U}}\} + 1 \tag{3}$$

where $\boldsymbol{\mathcal{U}}$ denotes as $[\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_C]$. In light of the provided definition, the OOD score ranges from a minimum of 0, with higher values indicating increased out-of-distribution characteristics, thereby highlighting the structures in transition states. As shown in Fig. 3C & 3D, MD conformations situated at two different transition state regions (between basins 1 & 2, and between basins 2 & 3 of the Müller potential in Fig. 2A) are identified simultaneously (Fig. 3D), as they display equal and the largest OOD scores in the latent space (Fig. 3B). We then chose

MD conformations with the largest OOD scores as transition state structures (see Fig. 3C-D for the Müller potential example, and the Methods section for additional details).

Previous studies[54] in the trustworthy AI field have utilized other similarity-based metrics in the latent representations for OOD detection. However, these methods may not be ideally suited for direct application in identifying transition states for protein conformational changes. For example, one can directly detect OOD samples at the state boundaries on the hypersphere and treat them as transition states structures. However, this method may not be well-suited for identifying transition states in our TS-DART model. This is because the VAMP-2 loss in TS-DART is designed to optimize the slowest dynamic modes, rendering state boundaries less sensitive and thereby limiting its ability to accurately pinpoint transition state conformations. As shown in SI Fig. S2B, MD conformations located at the state boundaries in the Müller potential do not accurately correspond to true transition states. Recent approaches[61,66] including CIDER aim to detect low-density regions in the latent space as OOD samples, which may serve as potential transition state structures in our context. However, these methods may overlook the transition states separated by relatively low free energy barriers or incorrectly classify the low-density regions as transition states. For example, choosing an appropriate density threshold to concurrently identify two transition states of the Müller potential is challenging. With a density threshold set at $\exp(-7.2kT)$, only the transition state between states 1 and 2 can be identified (SI Fig. S3C). Reducing the threshold to $\exp(-6kT)$ enables the identification of the transition state between states 2 and 3. However, in this case, the transition state between states 1 and 2 becomes poorly defined (SI Fig. S3D).

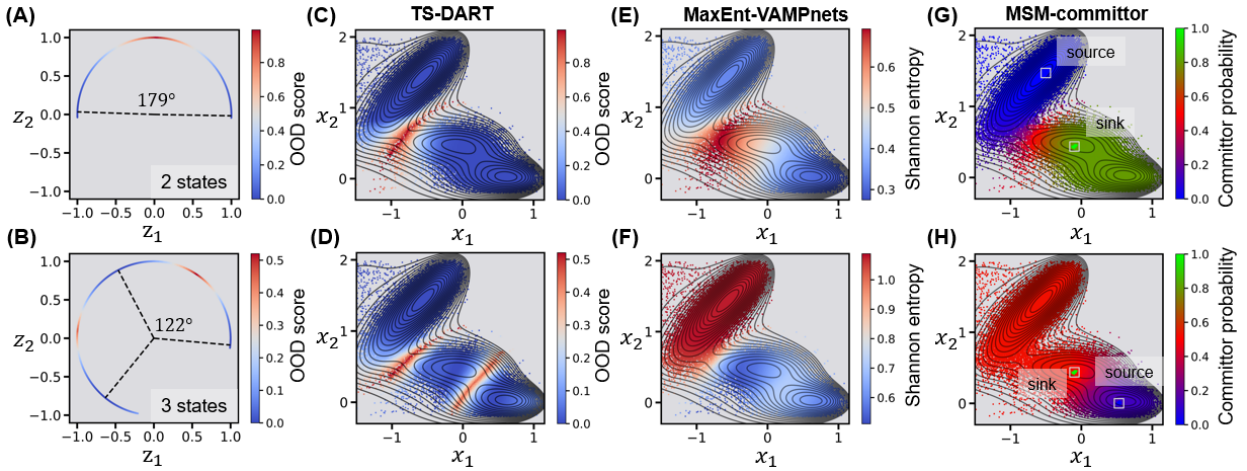

Figure 3. TS-DART outperforms MaxEnt-VAMPNets and committor probabilities in identifying transition states for the Müller potential. (a-b) Hyperspherical latent representations with OOD scores obtained from TS-DART are shown. Dashed lines point to the centers of metastable states. (c-d) MD conformations with their OOD scores obtained from TS-DART are overlaid with the Müller potential. (e-f) MD conformations with Shannon entropy obtained from MaxEnt-VAMPNets are overlaid with the Müller potential. (g-h) MD conformations with committor probabilities obtained from the 1,000-state MSM overlaid with the Müller potential. Two grey squares represent the centers of the source and sink states,

utilized as input into the transition path theory for computing committor probabilities. For further details, please refer to the SI Sec. 3.

TS-DART outperforms committor probabilities and MaxEnt-VAMPNets in identifying transition states for the Müller potential. The Müller potential exhibits three potential energy basins, featuring two transition states, positioned between them (Fig. 2A). We trained two TS-DART models by specifying 2 and 3 metastable states, respectively. Fig. S1A displays the validation curves of VAMP-2 loss (top) and dispersion loss (bottom) during the training of TS-DART. The first 50 epochs represent a pre-training process with pure VAMP-2 loss optimization (please refer to the training details in the Methods section). The results show that the VAMP-2 loss converges quickly within one epoch, indicating that the model has already captured the slowest dynamics of the systems within the pre-training process. Then, via the joint optimization of VAMP-2 loss and dispersion loss, the dispersion loss is minimized, while the VAMP-2 loss maintains the same value, demonstrating the efficacy of the model. With the trained models, we first plotted the OOD scores of all MD conformations in the latent hyperspherical space (Fig. 3A-3B). As shown in Fig. 3C-3D, the MD conformations positioned between the state centers (shown in red in Fig. 3A-3B) consistently align with those located at the summit of the energy barriers. This facilitates the straightforward identification of transition state conformations. Notably, in the two-state classification scenario, the transition states conformations located at the highest energy barrier are identified (Fig. 3C). In the three-state case, all transition states situated between the three energy basins are simultaneously captured (Fig. 3D).

We next demonstrate that our TS-DART outperforms two previously developed methods, MaxEnt-VAMPNets[34] and MSM's committor probabilities (MSM-committor)[14], in identifying the transition states for the Müller potential. For MaxEnt-VAMPNets, the VAMPnets model was trained by specifying the number of metastable states as two and three, respectively (see SI Sec. 2 for details). Regarding committor probabilities, the Müller potential was first discretized into 1,000 microstates using k-centers clustering. Committor probabilities were then computed based on the 1,000-state MSM by specifying the source and sink states (see SI Sec. 3 for details). Fig. 3E & 3G display the results from MaxEnt-VAMPNets (2-state model) and MSM-committor for identifying the transition state associated with the highest energy barrier. Although both methods successfully uncover the correct transition state region (see the red region in Fig. 3 E & 3G), their state boundaries are not as clear as revealed by the OOD scores from our TS-DART (Fig. 3C). Fig. 3F & 3H represent the results from MaxEnt-VAMPNets (3-state model) and MSM-committor aimed at detecting the transition state corresponding to the second highest energy barrier. Strikingly, both MaxEnt-VAMPNets and MSM's committor probabilities mistakenly identified the energy basin instead of the energy barrier as the transition state region. We anticipate that this misidentification occurs because the Shannon entropy computed from VAMPnets' outputs lacks direct physical connections with the true uncertainty measures of MD data, either in terms of transitioning in or out of energy basins. For MSM-committor, metastable energy basins that are far away from the sink and source states, separated by high energy barriers, are wrongly categorized as transition states. These remote states might have weak connections

with sink and source states, and therefore share an equal probability of transitioning to either, leading to the misidentification as shown in Fig. 3H.

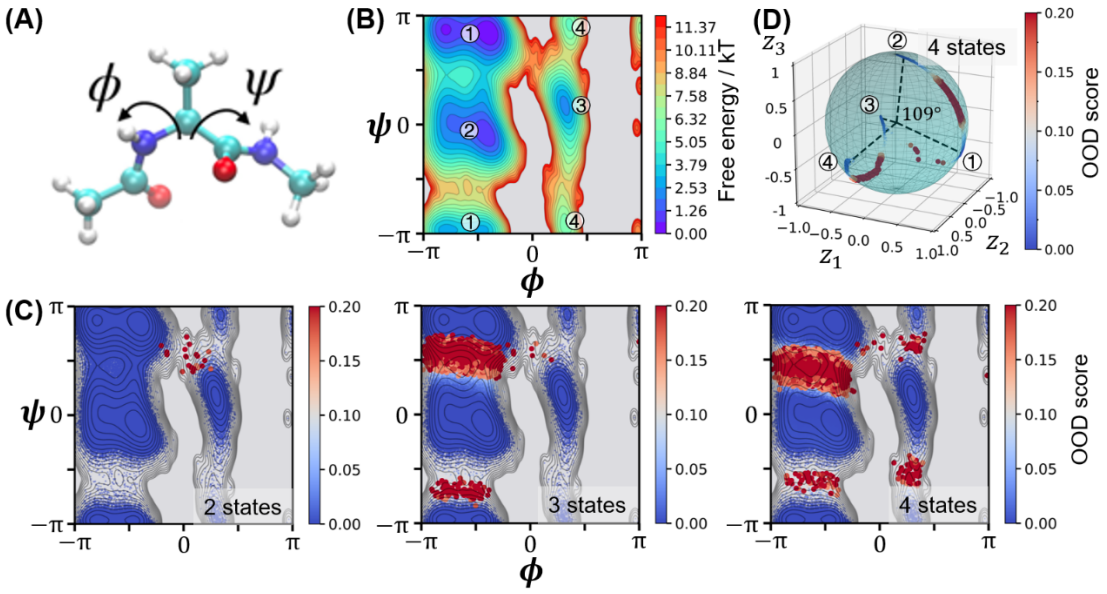

Figure 4. TS-DART identifies transition states for alanine dipeptide. (a) A representative conformation of alanine dipeptide. (b) Projection of the free energy landscape for alanine dipeptide onto two backbone torsion angles, ϕ and ψ. (c) TS-DART models with 2, 3, and 4 states successfully identified transition states (points with large OOD scores, shown in red) located at different free energy barriers shown in left, middle and right panels, respectively. (d) Hyperspherical representations in a 3D latent space (2D hypersphere) for the 4-state TS-DART model, with dashed lines indicating the centers of metastable states. Notably, for visualization, the maximum value of all color bars is constrained to the same value (0.2).

Transition states for alanine dipeptide. Alanine dipeptide consists of 10 heavy atoms, with its conformational changes often visualized through two backbone torsion angles, ϕ and ψ (Fig. 4A & 4B). To identify its transition states, we trained 3 TS-DART models for alanine dipeptide using the Cartesian coordinates of its 10 heavy atoms as input features, and the number of metastable states was set to be 2, 3, and 4, respectively (see Fig. 4C and refer to the Methods section for additional details). In the 2-state TS-DART model, we successfully identified the transition state located on the highest free energy barrier (left panel of Fig. 4C). This free energy barrier corresponds to the slowest dynamic transition between β (free energy basin 1 in Fig. 4B) and α_L (free energy basin 3 in Fig. 4B) metastable states of alanine dipeptide. In the 3-state model, we identified additional transition state regions situated between free energy basins 1 and 2. It is noteworthy that since the torsion angles exhibit periodicity every 2π, there are two transition state regions that separate basins 1 and 2 (α_R) (middle panel of Fig. 4C). In the 4-state model, all the transition states were simultaneously identified, with additional transition states detected corresponding to the third slowest dynamic mode, separating the two free energy basins (3 and 4) located on the right-hand side (right panel of Fig. 4C). In this 4-state model, we chose a

three-dimensional latent space (2D hypersphere), and the centers of the 4 metastable states (indicated by the dashed lines, see Fig 4D) form a tetrahedral geometry with the inter-state angular distance approximately equals to 109 degrees. In the implementation of the TS-DART method, we recommend choosing a three-dimensional latent space (2D hypersphere) for models containing 4 or more states.

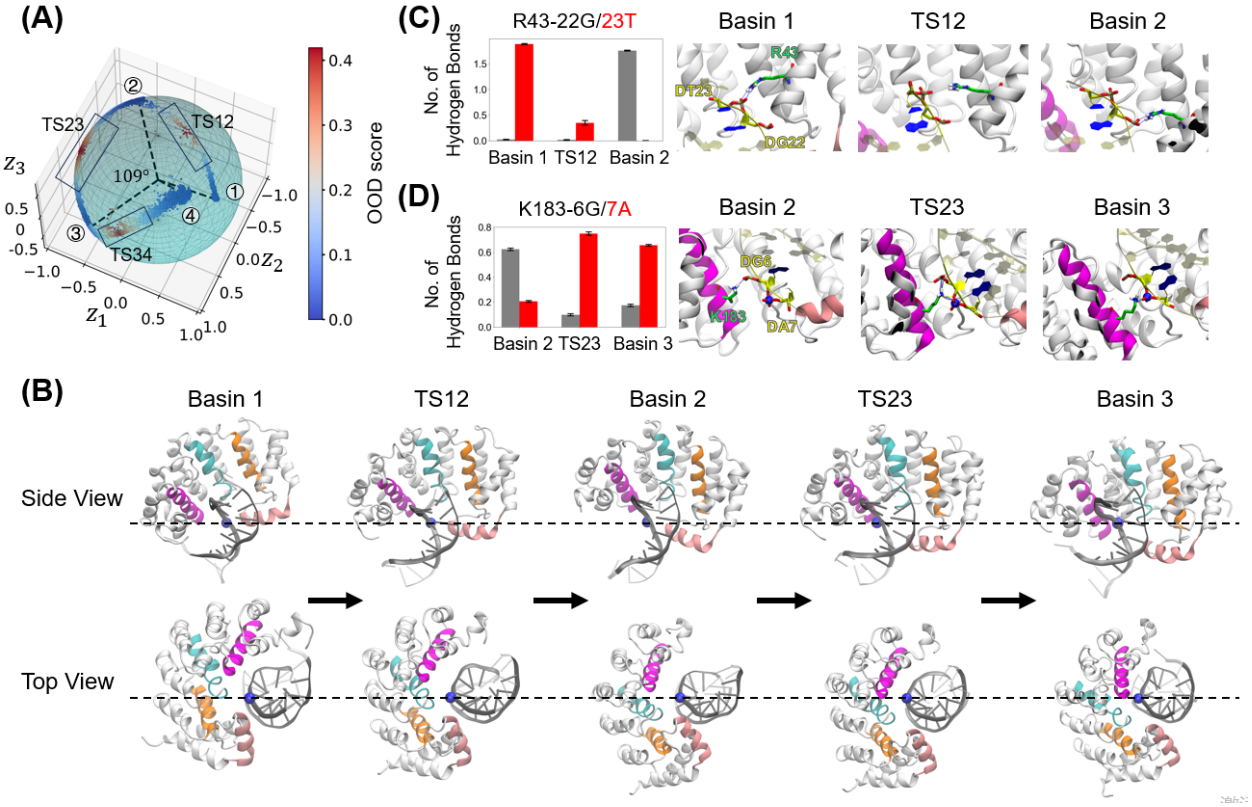

Figure 5. TS-DART identifies transition states of a DNA motor (AlkD) translocating along a double-stranded DNA (dsDNA) over one base pair. (a) Hyperspherical representations in a 3D latent space of AlkD-DNA complex system from the TS-DART model, with dashed lines indicating the center vectors of metastable states and the rectangle box outlining the selected transition states structures. (b) Representative conformations of the three basins and the transition state are shown from two different point of views. AlkD and dsDNA are shown as white and gray, respectively. The α-helices that are in contact with the dsDNA are shown in individual colors and the phosphor atom of the A7 phosphate group is shown as a blue sphere. The average number of hydrogen bonds between a residue of AlkD and two adjacent nucleotides of the dsDNA were calculated for each basin and transition state. (c) The average number of hydrogen bonds between residue 43 of AlkD and i) DG22 (gray bar graph), ii) DT23 (red bar graph) were calculated and visualized for basin 1, TS12, and basin 2. (d) The average number of hydrogen bonds between residue 183 of AlkD and i) DG6 (gray bar graph), ii) DA7 (red bar graph) were calculated and visualized for basin 2, TS23, and basin 3.

Transition states for the translocation of a motor protein on DNA. *Bacillus cereus* alkylpurine glycosylase D (AlkD) is a DNA motor protein that can translocate along the dsDNA

and plays a crucial role in repairing DNA damage[67]. We employed TS-DART to investigate the transition states of the diffusion dynamics of AlkD along a double-stranded DNA for the distance of one base pair. We have followed our previous study[68] to select 684 pairwise distances as the input features (see SI Fig. S4) and 4 states to train TS-DART. Fig. 5A displays the hyperspherical latent representations of AlkD-DNA complexes obtained from the TS-DART model, where four successively connected free energy basins (pointed by dashed lines) are uncovered and uniformly separated across the hypersphere. These four free energy basins are consistent with those revealed by MSMs in our previous study[68] (see SI Fig. S5). Specifically, free energy basin 1, 2, 3 correspond to the pre-translocation, an intermediate state exhibiting a rotation of AlkD on dsDNA, and post-translocation state, respectively (see Fig. 5B). Notably, we didn't show basin 4 as it represents a hyper-translocation state, where the AlkD has translocated along the dsDNA beyond one base pair. We identify three transition states with high OOD scores (>0.112, see Fig. 5A and the Methods section for additional details) that separate adjacent free energy basins in the hyperspherical latent space. Specifically, the transition state (TS12) separating the pre-translocation (basin 1) and the intermediate state (basin 2) exhibits a partial rotation (see the top view of Fig. 5B). As shown in Fig. 5C, this rotational movement results in the disruption of the hydrogen bond initially formed between residue R43 and base 23T (basin 1). Subsequently, a new hydrogen bond forms with an adjacent base, 22G (basin 2). Within TS12, both of these hydrogen bonds are significantly disrupted (Fig. 5C), leading to a reduction of over two hydrogen bonds between AlkD and dsDNA (SI Fig. S6). Consequently, TS12 represents a substantial free energy barrier that AlkD must overcome. Indeed, the transition from the pre-translocation to the intermediate state corresponds to the rate-limiting step (occurring at ~17.8 μs)[68]. In contrast, in the transition state TS23, AlkD forms a similar number of hydrogen bonds with the dsDNA compared to basin 2 and 3 (Fig. 5D and SI Fig. S6). This results in a relatively fast transition from the intermediate state to the post-translocation state, involving the translation of AlkD on dsDNA (occurring at ~1.3 μs)[68]. Our transition state analysis reveals the important role of hydrogen bonds in governing the dynamics of the asymmetric translocation of AlkD on dsDNA.

We show that the hyperspherical latent representations from TS-DART model's penultimate layer serve as a good reduced kinetic space for understanding the slow dynamics of protein conformational changes. Previous methods, such as tICA[69,70], and SRVs[71], which are rooted in the variational approach, aim to identify decorrelated orthogonal collective variables (CVs) for dimensionality reduction. However, these methods may hinder the comprehensive understanding of intricate collaborative dynamic motions. In contrast, alternative approaches like RC flow[72] are specifically designed to reveal latent kinetic manifolds that preserve full-state kinetic information. Nevertheless, these methods pose greater challenges in terms of training and may struggle to discover clearer state boundaries. Here, we highlight the robustness of our hyperspherical latent representations for capturing the relevant reduced kinetics and underscores that these representations benefit from two perspectives. Firstly, the utilization of penultimate layer of a deep neural network for representation learning is simple and robust. This design choice facilitates the learned latent representations to be directly regularized through the optimization of

the loss function defined on the output layer, such as VAMP-2 score in our framework. Secondly, and of more significance, representation learning on a unit sphere helps the model better capture and describe kinetics. Recent advancements[73-78] in deep learning field have demonstrated that the hyperspherical latent space performs better than traditional Euclidean space in applications ranging from variational autoencoders to convolutional neural networks for image classification tasks. In our specific context of biomolecular dynamics, the hyperspherical latent space is particularly important, offering greater capacity for describing more complex kinetic geometries including specific kinetic symmetries and periodicity of the kinetic data. In the future, we anticipate that the hyperspherical latent representations in TS-DART will have broad applications in the study of biomolecular dynamics. For example, it can provide a good platform for analyzing the parallel transition pathways of complex dynamic systems on the hypersphere. Furthermore, recent developments[72,79] have utilized autoencoder or normalizing flow architectures to learn a reduced kinetic model in the latent space, such as Brownian dynamics, by introducing specific dynamical assumptions or constraints in the loss function. We expect that the penultimate hyperspherical latent space in TS-DART can be deployed for learning a continuous dynamic model by integration more physical-driven novel designs of loss functions.

3. Conclusion

We introduce TS-DART, a novel deep learning approach designed for detecting transition states from MD simulations by utilizing hyperspherical embeddings in the latent space. Inspired by recent advancements in trustworthy AI for identifying OOD data, TS-DART treats transition state structures as OOD data. This approach allows TS-DART to discern transition state conformations that separate multiple metastable states, facilitated by the introduction of a dispersion loss function term. Moreover, the hyperspherical embeddings of MD conformations in TS-DART retain all relevant kinetic geometries and are compacted in terms of their kinetic metastability through the incorporation of a VAMP-2 loss function term. Collectively, our TS-DART method establishes an end-to-end pipeline capable of simultaneously and automatically identifying all transition states across multiple free energy barriers underlying protein conformational changes. To demonstrate the efficacy of TS-DART, we apply it to the 2D Müller potential, alanine dipeptide, and the translocation of a DNA motor protein on dsDNA. We anticipate that TS-DART can find widespread application in identifying transition states for protein conformational changes.

4. Methods

Remove-mean time-instantaneous and time-lagged correlation matrices. Given a set of basis functions $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_m)^T$, and a MD trajectory of length T ($\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_T\}$), the remove-mean time-instantaneous and time-lagged correlation matrices are defined as follows:

$$\begin{cases} \bar{\mathbf{C}}_{00} = \dfrac{1}{T-\tau}\mathbf{X}^{\mathrm{T}}\mathbf{X} - \boldsymbol{\pi}_0\boldsymbol{\pi}_0^{\mathrm{T}} \\ \bar{\mathbf{C}}_{11} = \dfrac{1}{T-\tau}\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \boldsymbol{\pi}_1\boldsymbol{\pi}_1^{\mathrm{T}} \\ \bar{\mathbf{C}}_{01} = \dfrac{1}{T-\tau}\mathbf{X}^{\mathrm{T}}\mathbf{Y} - \boldsymbol{\pi}_0\boldsymbol{\pi}_1^{\mathrm{T}} \end{cases} \quad (4)$$

where $\mathbf{X}$ and $\mathbf{Y}$ are two $T-\tau$ by m matrices, defined as $[\boldsymbol{\chi}(\boldsymbol{x}_1), \dots, \boldsymbol{\chi}(\boldsymbol{x}_{T-\tau})]^{\mathrm{T}}$ and $[\boldsymbol{\chi}(\boldsymbol{x}_{\tau+1}), \dots, \boldsymbol{\chi}(\boldsymbol{x}_T)]^{\mathrm{T}}$ (τ is the lag time). $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ are mean vectors of $\mathbf{X}$ and $\mathbf{Y}$, which equal to $\frac{1}{T-\tau}\mathbf{X}^{\mathrm{T}}\mathbf{1}$ and $\frac{1}{T-\tau}\mathbf{Y}^{\mathrm{T}}\mathbf{1}$, respectively. In our method, we parameterized the basis functions, $\boldsymbol{\chi}$, using the SoftMax outputs of two parallel networks with shared parameters from TS-DART.

Exponential-moving-average (EMA) to estimate the metastable state centers. We denote the hyperspherical embeddings of conformations as $\{\boldsymbol{z_i}\}_{i=1}^{N}$, where $\boldsymbol{z}_i \in \mathbb{R}^d$, and $\{1,2,\dots,C\}$ as the metastable state indices, the metastable state centers correspond to vectors $\{\boldsymbol{\mu}_c\}_{c=1}^{C}$ that can be computed via a EMA manner:

$$\boldsymbol{\mu}_c \coloneqq Normalize(\theta\boldsymbol{\mu}_c + (1-\theta)\boldsymbol{z}_i),\ c = \tilde{y}_i \quad (5)$$

where $\tilde{y}_i \in \{1,2,\dots,C\}$ represents the state index of the conformation i, θ is the state center update factor and specifically selected as 0.5.

MD simulation datasets. The analytical form of the Müller potential[62] (Fig. 2A) is as follows:

$$V_{Müller}(x_1, x_2) = \sum_{i=1}^{4} A_i \exp(a_i(x_1-\bar{x}_i)^2 + b_i(x_1-\bar{x}_i)(x_2-\bar{y}_i) + c_i(x_2-\bar{y}_i)^2) \quad (6)$$

where $(A_1,\dots,A_4) = (-10,-5,-8.5,0.75)$, $(a_1,\dots,a_4) = (-1,-1,-6.5,0.7)$, $(b_1,\dots,b_4) = (0,0,11,0.6)$, $(c_1,\dots,c_4) = (-10,-10,-6.5,0.7)$, $(\bar{x}_1,\dots,\bar{x}_4) = (1,0,0.5,-1)$, $(\bar{y}_1,\dots,\bar{y}_4) = (0,0.5,1.5,1)$. We performed a Brownian dynamics simulation (time step equals 2×10^{-4}, damping factor equals 1) to sample this Müller potential at the temperature of 0.9. A reflective boundary condition is adopted: $x_1 \in [-1.5, 1.2]$, $x_2 \in [-0.2, 2]$. The simulation trajectory contains 3×10^5 frames with the saving interval of 0.01. For alanine dipeptide, we obtained the MD simulation dataset from a previous study[80]. It contains three 250-ns MD trajectories, with a saving interval of 1 ps. As a result, the entire dataset contains 750,000 MD conformations. All conformations were aligned to the first frame according to the minimal root mean square deviation. The input features consist of the x, y and z coordinates of the 10 heavy atoms (totally 30 input features). For the translocation of a DNA motor protein (AlkD) on DNA system, we obtained the MD simulation dataset from a previous study[68]. It contains 200 50-ns and 100 45-ns MD trajectories, with a saving interval of 20 ps. As a result, the entire dataset contains 725,300 MD conformations. The input features consist of the pairwise distances of 684 atom pairs, which are constituted by phosphate atoms of five base pairs in the center region of dsDNA and heavy atoms of five protein helixes within 12 Å of nucleotides. See SI Fig. S4 for illustration.

Training details of the TS-DART models. We specified the scaling factor (γ) of the hyperspherical embeddings as 1, the scaling hyperparameter in dispersion loss as 0.1, the weight (β) of dispersion loss as 0.01 for Müller potential and alanine dipeptide datasets, 0.05 for AlkD-DNA dataset, the state center update factor (θ) as 0.5, and the lag time (τ) as 1 time step for Müller potential, 1 ps for alanine dipeptide, 8 ns for AlkD-DNA dataset. In addition, we established a criterion for determining the dimensionality (denote as d) of the latent hyperspherical embeddings (i.e., $(d-1)$-sphere): $d = 3$, if number of states is more than 3, otherwise, $d = 2$. For the specific training hyperparameters of TS-DART on Müller potential, alanine dipeptide and AlkD-DNA datasets, please refer to SI Sec. 1 for details.

The validation of $\boldsymbol{\beta}$. The selection of β is trivial in TS-DART model. However, choosing a proper magnitude of β is important for the fully optimization of both VAMP-2 and dispersion losses. In this work, we performed the ablation tests on β within Müller potential, alanine dipeptide and AlkD-DNA systems by training the TS-DART model with different magnitudes of β. The determination of the magnitude of β is guided by the following two criteria: (1) The dispersion loss can converge to the minimum boundary. (2) There is no significant deviation in VAMP-2 loss before and after integrating the dispersion loss optimization. Please see SI Fig. S7 for details.

The selection of transition states structures based on the OOD scores. With obtaining OOD scores from TS-DART model that effectively distinguish between transition and metastable states, there are multiple ways for selecting transition state structures based on these scores for practical applications. For example, one can directly select a certain number of conformations with highest OOD scores, which is particularly useful for adaptive sampling. Also, one can define a threshold to identify these transition states structures. Here, a specific threshold is recommended:

$$thres = -\cos(\theta/4) + 1 \tag{7}$$

where θ represents the angle between neighbor state center vectors on the hypersphere after the optimization of TS-DART model. Based on the above definition, the identified transition states boundaries will position at the middle between the summit of the free energy barriers (i.e., mean vectors of state center vectors) and the free energy minima (i.e., state center vectors) on the hypersphere. In this work, we successfully selected the transition states structures of AlkD translocation along a DNA system (see Fig. 5A) by applying a threshold (0.112), computed by adopting θ of 109.45°.

5. Acknowledgement

We acknowledge the support from the NIH/NIGMS under award number 1 R01GM147652-01A1 and the support from the Hirschfelder Professorship Fund from University of Wisconsin-Madison to X.H. Yixuan Li is supported by the AFOSR Young Investigator Program under award number FA9550-23-1-0184, National Science Foundation (NSF) Award No. IIS-2237037 & IIS-

2331669, Office of Naval Research under grant number N00014-23-1-2643, and faculty research awards/gifts from Google and Meta. Xuefeng Du is supported by the Jane Street Graduate Research Fellowship.

6. References

1 Prinz, J.-H. *et al.* Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics* **134** (2011).

2 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* **126** (2007).

3 Husic, B. E. & Pande, V. S. Markov state models: From an art to a science. *Journal of the American Chemical Society* **140**, 2386-2396 (2018).

4 Bowman, G. R., Pande, V. S. & Noé, F. in *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* (eds Gregory R. Bowman, Vijay S. Pande, & Frank Noé) 1-6 (Springer Netherlands, 2014).

5 Pan, A. C. & Roux, B. Building Markov state models along pathways to determine free energies and rates of transitions. *The Journal of chemical physics* **129**, 064107 (2008).

6 Buchete, N.-V. & Hummer, G. Coarse master equations for peptide folding dynamics. *The Journal of Physical Chemistry B* **112**, 6057-6069 (2008).

7 Wang, W., Cao, S., Zhu, L. & Huang, X. Constructing Markov State Models to elucidate the functional conformational changes of complex biomolecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **8**, e1343 (2018).

8 Huang, X., Bowman, G. R., Bacallado, S. & Pande, V. S. Rapid equilibrium sampling initiated from nonequilibrium data. *Proceedings of the National Academy of Sciences* **106**, 19765-19769 (2009).

9 Malmstrom, R. D., Lee, C. T., Van Wart, A. T. & Amaro, R. E. Application of molecular-dynamics based markov state models to functional proteins. *Journal of chemical theory and computation* **10**, 2648-2657 (2014).

10 Morcos, F. *et al.* Modeling conformational ensembles of slow functional motions in Pin1-WW. *PLoS computational biology* **6**, e1001015 (2010).

11 Zhang, B. W. *et al.* Simulating replica exchange: Markov state models, proposal schemes, and the infinite swapping limit. *The Journal of Physical Chemistry B* **120**, 8289-8301 (2016).

12 Konovalov, K. A., Unarta, I. C., Cao, S., Goonetilleke, E. C. & Huang, X. Markov state models to study the functional dynamics of proteins in the wake of machine learning. *JACS Au* **1**, 1330-1341 (2021).

13 Liu, B., Qiu, Y., Goonetilleke, E. C. & Huang, X. Kinetic network models to study molecular self-assembly in the wake of machine learning. *MRS Bulletin*, 1-9 (2022).

14 Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L. & Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations.

Proceedings of the National Academy of Sciences **106**, 19011-19016 (2009).
15 Bowman, G. R., Voelz, V. A. & Pande, V. S. Taming the complexity of protein folding. *Current opinion in structural biology* **21**, 4-11 (2011).
16 Da, L.-T. *et al.* A jump-from-cavity pyrophosphate ion release assisted by a key lysine residue in T7 RNA polymerase transcription elongation. *PLoS computational biology* **11**, e1004624 (2015).
17 Da, L.-T., Wang, D. & Huang, X. Dynamics of pyrophosphate ion release and its coupled trigger loop motion from closed to open state in RNA polymerase II. *Journal of the American Chemical Society* **134**, 2399-2406 (2012).
18 Da, L.-T. *et al.* Bridge helix bending promotes RNA polymerase II backtracking through a critical and conserved threonine residue. *Nature communications* **7**, 1-10 (2016).
19 Silva, D.-A. *et al.* Millisecond dynamics of RNA polymerase II translocation at atomic resolution. *Proceedings of the National Academy of Sciences* **111**, 7665-7670 (2014).
20 Malmstrom, R. D., Kornev, A. P., Taylor, S. S. & Amaro, R. E. Allostery through the computational microscope: cAMP activation of a canonical signalling domain. *Nature communications* **6**, 7588 (2015).
21 Kohlhoff, K. J. *et al.* Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nature chemistry* **6**, 15-21 (2014).
22 Deng, N.-j., Dai, W. & Levy, R. M. How kinetics within the unfolded state affects protein folding: An analysis based on Markov state models and an ultra-long MD trajectory. *The Journal of Physical Chemistry B* **117**, 12787-12799 (2013).
23 Wan, H., Ge, Y., Razavi, A. & Voelz, V. A. Reconciling simulated ensembles of apomyoglobin with experimental hydrogen/deuterium exchange data using bayesian inference and multiensemble markov state models. *Journal of chemical theory and computation* **16**, 1333-1348 (2020).
24 Buch, I., Giorgino, T. & De Fabritiis, G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences* **108**, 10184-10189 (2011).
25 Lawrenz, M., Shukla, D. & Pande, V. S. Cloud computing approaches for prediction of ligand binding poses and pathways. *Scientific reports* **5**, 1-5 (2015).
26 Silva, D.-A., Bowman, G. R., Sosa-Peinado, A. & Huang, X. A role for both conformational selection and induced fit in ligand binding by the LAO protein. *PLoS computational biology* **7**, e1002054 (2011).
27 Plattner, N. & Noé, F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nature communications* **6**, 1-10 (2015).
28 Wang, B., Sexton, R. E. & Feig, M. Kinetics of nucleotide entry into RNA polymerase active site provides mechanism for efficiency and fidelity. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1860**, 482-490 (2017).
29 Khaled, M., Gorfe, A. & Sayyed-Ahmad, A. Conformational and dynamical effects of Tyr32 phosphorylation in K-Ras: molecular dynamics simulation and Markov state

models analysis. *The Journal of Physical Chemistry B* **123**, 7667-7675 (2019).
30 Barros, E. P., Demir, Ö., Soto, J., Cocco, M. J. & Amaro, R. E. Markov state models and NMR uncover an overlooked allosteric loop in p53. *Chemical science* **12**, 1891-1900 (2021).
31 Feng, J., Selvam, B. & Shukla, D. How do antiporters exchange substrates across the cell membrane? An atomic-level description of the complete exchange cycle in NarK. *Structure* **29**, 922-933 (2021).
32 Son, C. Y., Yethiraj, A. & Cui, Q. Cavity hydration dynamics in cytochrome c oxidase and functional implications. *Proceedings of the National Academy of Sciences* **114**, E8830-E8836 (2017).
33 Jiang, H. *et al.* Markov state models reveal a two-step mechanism of miRNA loading into the human argonaute protein: selective binding followed by structural re-arrangement. *PLoS computational biology* **11**, e1004404 (2015).
34 Kleiman, D. E. & Shukla, D. Active Learning of the Conformational Ensemble of Proteins Using Maximum Entropy VAMPNets. *Journal of Chemical Theory and Computation* (2023).
35 Mardt, A., Pasquali, L., Wu, H. & Noé, F. VAMPnets for deep learning of molecular kinetics. *Nature communications* **9**, 5 (2018).
36 Wu, H. & Noé, F. Variational approach for learning Markov processes from time series data. *Journal of Nonlinear Science* **30**, 23-66 (2020).
37 Bolhuis, P. G., Chandler, D., Dellago, C. & Geissler, P. L. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual review of physical chemistry* **53**, 291-318 (2002).
38 Dellago, C., Bolhuis, P. G. & Geissler, P. L. Transition path sampling. *Advances in chemical physics* **123**, 1-78 (2002).
39 Zhang, J., Yang, Y. I., Yang, L. & Gao, Y. Q. Dynamics and kinetics study of "In-Water" chemical reactions by enhanced sampling of reactive trajectories. *The Journal of Physical Chemistry B* **119**, 14505-14514 (2015).
40 Zhang, J. *et al.* Rich dynamics underlying solution reactions revealed by sampling and data mining of reactive trajectories. *ACS central science* **3**, 407-414 (2017).
41 Zhang, J. *et al.* Deep reinforcement learning of transition states. *Physical Chemistry Chemical Physics* **23**, 6888-6895 (2021).
42 Jónsson, H., Mills, G. & Jacobsen, K. W. in *Classical and quantum dynamics in condensed phase simulations* 385-404 (World Scientific, 1998).
43 Weinan, E., Ren, W. & Vanden-Eijnden, E. String method for the study of rare events. *Physical Review B* **66**, 052301 (2002).
44 Weinan, E., Ren, W. & Vanden-Eijnden, E. Finite temperature string method for the study of rare events. *J. Phys. Chem. B* **109**, 6688-6693 (2005).
45 Vanden-Eijnden, E. Towards a theory of transition paths. *Journal of statistical physics* **123**, 503-523 (2006).
46 Maragliano, L., Fischer, A., Vanden-Eijnden, E. & Ciccotti, G. String method in

collective variables: Minimum free energy paths and isocommittor surfaces. *The Journal of chemical physics* **125** (2006).

47 Pan, A. C., Sezer, D. & Roux, B. Finding transition pathways using the string method with swarms of trajectories. *The journal of physical chemistry B* **112**, 3432-3440 (2008).

48 Roux, B. String method with swarms-of-trajectories, mean drifts, lag time, and committor. *The Journal of Physical Chemistry A* **125**, 7558-7571 (2021).

49 He, Z., Chipot, C. & Roux, B. Committor-consistent variational string method. *The Journal of Physical Chemistry Letters* **13**, 9263-9271 (2022).

50 Lai, R. & Lu, J. Point Cloud Discretization of Fokker--Planck Operators for Committor Functions. *Multiscale Modeling & Simulation* **16**, 710-726 (2018).

51 Khoo, Y., Lu, J. & Ying, L. Solving for high-dimensional committor functions using artificial neural networks. *Research in the Mathematical Sciences* **6**, 1-13 (2019).

52 Li, Q., Lin, B. & Ren, W. Computing committor functions for the study of rare events using deep learning. *The Journal of Chemical Physics* **151** (2019).

53 Prinz, J.-H., Held, M., Smith, J. C. & Noé, F. Efficient computation, sensitivity, and error analysis of committor probabilities for complex dynamical processes. *Multiscale Modeling & Simulation* **9**, 545-567 (2011).

54 Yang, J., Zhou, K., Li, Y. & Liu, Z. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334* (2021).

55 Amodei, D. *et al.* Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

56 Dietterich, T. G. Steps toward robust artificial intelligence. *Ai Magazine* **38**, 3-24 (2017).

57 Leike, J. *et al.* AI safety gridworlds. *arXiv preprint arXiv:1711.09883* (2017).

58 Smuha, N. A. The EU approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International* **20**, 97-106 (2019).

59 Shneiderman, B. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* **10**, 1-31 (2020).

60 Mohseni, S. *et al.* Practical machine learning safety: A survey and primer. *arXiv preprint arXiv:2106.04823* **4** (2021).

61 Ming, Y., Sun, Y., Dia, O. & Li, Y. How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection? In *The Eleventh International Conference on Learning Representations,* (2023).

62 Müller, K. & Brown, L. D. Location of saddle points and minimum energy paths by a constrained simplex optimization procedure. *Theoretica chimica acta* **53**, 75-93 (1979).

63 Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597-1607 (2020).

64 Wang, T. & Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 9929-9939 (2020).

65 Li, J., Xiong, C. & Hoi, S. MoPro: Webly Supervised Learning with Momentum Prototypes. In *International Conference on Learning Representations* (2021).
66 Sun, Y., Ming, Y., Zhu, X. & Li, Y. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, 20827-20840 (2022).
67 Jones Jr, L. E. *et al.* Differential effects of reactive nitrogen species on DNA base excision repair initiated by the alkyladenine DNA glycosylase. *Carcinogenesis* **30**, 2123-2129 (2009).
68 Peng, S. *et al.* Target search and recognition mechanisms of glycosylase AlkD revealed by scanning FRET-FCS and Markov state models. *Proceedings of the National Academy of Sciences* **117**, 21889-21895 (2020).
69 Schwantes, C. R. & Pande, V. S. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *Journal of chemical theory and computation* **9**, 2000-2009 (2013).
70 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* **139** (2013).
71 Chen, W., Sidky, H. & Ferguson, A. L. Nonlinear discovery of slow molecular modes using state-free reversible VAMPnets. *The Journal of chemical physics* **150**, 214114 (2019).
72 Wu, H. & Noé, F. Reaction coordinate flows for model reduction of molecular kinetics. *arXiv preprint arXiv:2309.05878* (2023).
73 Xu, J. & Durrett, G. Spherical latent spaces for stable variational autoencoders. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4503-4513 (2018).
74 Davidson, T. R., Falorsi, L., De Cao, N., Kipf, T. & Tomczak, J. M. Hyperspherical variational auto-encoders. *34th Conference on Uncertainty in Artificial Intelligence (UAI-18)*, 856-865 (2018).
75 Bojanowski, P. & Joulin, A. Unsupervised learning by predicting noise. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 517-526 (2017).
76 Mettes, P., Van der Pol, E. & Snoek, C. Hyperspherical prototype networks. *Advances in neural information processing systems* **32** (2019).
77 Liu, W. *et al.* Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212-220 (2017).
78 Wang, F., Xiang, X., Cheng, J. & Yuille, A. L. NormFace: L2 Hypersphere Embedding for Face Verification. *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1041-1049 (2017).
79 Wang, D., Wang, Y., Evans, L. & Tiwary, P. Introducing dynamical constraints into representation learning. *arXiv preprint arXiv:2209.00905* (2022).
80 Nüske, F. *et al.* Markov state models from short non-equilibrium simulations—Analysis and correction of estimation bias. *The Journal of Chemical Physics* **146** (2017).