Topological analysis reveals multiple pathways in molecular dynamics

Luca Donati*,[1,2] Surahit Chewle,[2] Dominik St. Pierre,[3,2] Vijay Natarajan,[4,2] and Marcus Weber[2]

[1)]*Freie Universität Berlin, Department of Mathematics and Computer Science, Arnimallee 22, D-14195 Berlin, Germany*

[2)]*Zuse Institute Berlin, Takustr. 7, D-14195 Berlin, Germany*

[3)]*Freie Universität Berlin, Department of Biology, Chemistry and Pharmacy, Arnimallee 22, D-14195 Berlin, Germany*

[4)]*Indian Institute of Science, Department of Computer Science and Automation, 560012 Bangalore, India*

(*Electronic mail: donati@zib.de)

Abstract

Molecular Dynamics simulations are indispensable tools for comprehending the dynamic behavior of biomolecules, yet extracting meaningful molecular pathways from these simulations remains challenging due to the vast amount of high dimensional data. In this work, we present Molecular Kinetics via Topology (MoKiTo), a novel approach that combines the ISOKANN algorithm to determine the membership function of a molecular system with a topological analysis tool inspired by the Mapper algorithm. Our strategy efficiently identifies and characterizes distinct molecular pathways, enabling the detection and visualization of critical conformational transitions and rare events. This method offers deeper insights into molecular mechanisms, facilitating the design of targeted interventions in drug discovery and protein engineering.

Keywords: Molecular Dynamics, ISOKANN, Mapper, Topological analysis

I. Introduction

The identification of reaction pathways in chemical processes such as protein folding, protein–ligand binding/unbinding and enzymatic reactions is fundamental to the development of novel drugs and clinical treatments[1–4]. However, the long time scales of these processes, stemming from the ruggedness of the underlying energy landscapes, make it challenging to resolve these pathways via long Molecular Dynamics (MD) simulations. A widely adopted approach, which relies on the sampling of short MD simulations, is the construction of a Markov State Model (MSM)[5–8], from which one can extract pathways linking the states of an initial macro-state to those of a target macro-state, by applying tools such as Transition Path Theory (TPT)[9–12] or MSMPathfinder[13].

MSM-based methods often reveal a large number of transition pathways, offering detailed mechanistic insight[14]. However, these pathways can be difficult to visualize and interpret without clustering tools such as Path Lumping[15] or Latent-space Path Clustering[16], which aim to simplify the network. Recently, new approaches that take advantage of machine learning techniques have been developed[17–19]. Of particular interest is the method proposed in Ref.[20], where MD trajectories connecting the macro-states are generated via enhanced MD simulations[21] and clustered using the Dynamic Time Warping (DTW) algorithm[22]. This novel approach does not rely on MSMs and does not require dimensionality reduction, thereby resulting in a significant advancement over the state-of-the-art.

However, it requires a priori knowledge of the system to configure the enhanced sampling algorithm that generates the trajectories. Moreover, it relies on trajectory clustering, which is substantially more complex than clustering static points because it operates in a high-dimensional space where each object is a time series.

In this article, we propose a new but complementary framework called Molecular Kinetics via Topology (MoKiTo) for identifying pathway networks using tools derived from Topological Data Analysis (TDA). In MoKiTo, MD data can be generated either by conventional MD simulations or by enhanced sampling techniques, and are subsequently ordered and partitioned according to an ordering parameter $\chi : \Gamma \to [0,1]$, where Γ denotes the state space of the molecular system. The data are then connected through shared data points to form a graph which reveals macro-states, transition states, and pathways. Additionally, MoKiTo generates free energy profiles of individual pathways, facilitating their classification and the calculation of free energy differences and transition rates by Square Root Approximation[23–25].

This method is inspired by the Mapper algorithm[26], a TDA tool designed to capture topological structure of large datasets by first ordering the data with respect to a parameter χ and then clustering points within intervals of χ. As an illustrative example, originally presented in Ref.[27], Fig.1 shows a point cloud sampled from a human hand clustered and connected in a graph using an ordering parameter that maps values from the wrist to the fingertips. Here, the wrist and fingertips play the role of metastable basins, and the single scalar χ is sufficient to reveal multiple pathways from a source to distinct targets. This example motivates our use of Mapper for MD data where multiple metastable basins and intricate networks of pathways are often present.

In the MD setting, χ acts as a reaction coordinate, then it should vary monotonically along the dominant slow mode of the system described by the second eigenfunction of the Koopman operator. In two-state scenarios, the committor is the optimal reaction coordinate[28–31] and a natural choice for MoKiTo. Under reversibility and a clear spectral gap, the committor can indeed be approximated by the leading non-trivial Koopman eigenfunction. Alternatively, in this work we adopt as ordering parameter the so-called "membership function", originally introduced in PCCA+ [32]. The membership function, hereafter the χ-function, is defined as an affine normalization of the leading non-trivial Koopman eigenfunction to $[0, 1]$, thus describing the dominant slow mode of the system. Unlike the committor, it operates without having to specify source and target sets but is nevertheless consistent with the definition of natural reaction coordinate[33]. Aside from technical details, which are clarified in the theory section, the significance of the χ-function is to describe the probability that a system's conformational state belongs or does not belong to a macro-state.

To estimate the χ-function, we use ISOKANN[34–36], a data-driven method that trains an artificial neural network until convergence. The advantage of ISOKANN with respect to other methods such as PCCA+[32,37] is that it does not need to discretize the space neither a low-dimensional featurization of the system. It operates directly on molecular coordinates and yields a smooth function that generalizes out of sample and can be evaluated on arbitrarily many configurations.

In MoKiTo, the choice of the ordering parameter remains flexible and arbitrary. Other valid options are the minimum energy path[38–40] or the minimum action path[41–44]. Alternatively, physical coordinates such as interatomic distances, bond angles, and dihedral angles can also serve as ordering parameters in this context. Likewise, the algorithm used to determine the ordering parameter is not fixed. Depending on the system and objectives, one may employ techniques such as Principal Component Analysis (PCA)[45,46], Time-lagged Independent Component Analysis (TICA)[47,48], Time-lagged Autoencoders (TAEs)[49], diffusion maps[50], isometric feature mapping (ISOMAP)[51],

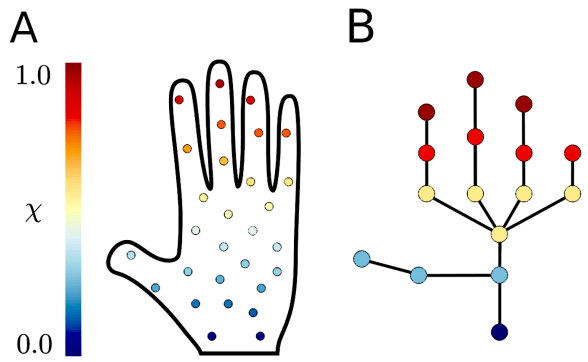


FIG. 1. (A) Data points sampling a "hand" ordered according to a χ-function. (B) Graph realized with the mapper algorithm.

sketch-map[52], or others.

We applied MoKiTo to several molecular systems. Of particular interest is the fourth example, where we studied the villin headpiece subdomain[53–56], revealing both dominant and minor folding pathways. This detailed mapping of the folding landscape not only highlights the complexity of protein folding, but also emphasizes the importance of alternative routes that, despite being less frequent, contribute significantly to the overall dynamics of the protein.

II. Theoretical background

Molecular systems often exhibit a broad separation of time scales, with fast thermal fluctuations that decay much faster than slow conformational rearrangements. A standard approach to characterizing such systems is the spectral analysis of the Koopman operator $\mathscr{K}_\tau$. This operator propagates bounded observables $f : \Gamma \to \mathbb{R}$, where Γ denotes the configuration space of the system, over a lag time τ, or equivalently through its infinitesimal generator $\mathscr{L}$:

$$f_{t+\tau}(x) = \exp(\tau\mathscr{L})\, f_t(x) \tag{1}$$

$$= \mathscr{K}_\tau f_t(x) \tag{2}$$

$$= \mathbb{E}\left[f_t(x_{t+\tau}) | x_t = x\right], \tag{3}$$

where the last line expresses the action of the Koopman operator as a conditional expectation over trajectories. If the dynamics of the molecular system are governed by a confining potential energy

function $V(x) : \Gamma \to \mathbb{R}$, such that the equilibrium density

$$\pi(x) = \frac{1}{Z} e^{-\beta V(x)}, \tag{4}$$

where Z is the normalizing constant, is well defined, then the operators $\mathscr{K}_\tau$ and $\mathscr{L}$ are self-adjoint in the weighted space $L^2(\pi)$, and the dynamics satisfies detailed-balance condition. Under these conditions, the eigenvalues and eigenfunctions of the two operators solve the eigenvalue problems:

$$\mathscr{L}\psi_i = \kappa_i \psi_i \tag{5}$$

$$\mathscr{K}_\tau \psi_i = \lambda_{\tau,i} \psi_i, \tag{6}$$

with $\lambda_{\tau,i} = \exp(\tau \kappa_i)$. The eigenvalues are sorted by decreasing magnitude, so that

$$\kappa_0 = 0 > \kappa_1 > \kappa_2 \ldots \tag{7}$$

$$\lambda_{\tau,0} = 1 > \lambda_{\tau,1} > \lambda_{\tau,2} \ldots > 0, \tag{8}$$

while the eigenfunctions form an orthonormal basis of $L^2(\pi)$ with respect to the weighted scalar product $\langle f, g \rangle_\pi = \int f(x) g(x) \pi(x) dx$. The trivial eigenfunction ψ_0 is constant, representing the equilibrium mode, whereas the leading non-trivial eigenfunction ψ_1 is monotonic with a single node. It encodes the dominant slow relaxation process, orthogonal to equilibrium and provides a natural reaction coordinate for the system, as argued in Ref.[57], since it satisfies the following key properties: (i) it performs dimensionality reduction by mapping each state to a single real value; (ii) it is uniquely determined by the system's dynamics, without requiring an a priori definition of macro-states; and (iii) it provides an optimal description of the dynamics by preserving both the Markovianity and the dominant implied time scales encoded in the Koopman operator. These properties make it convenient to introduce the normalized counterpart of ψ_1, the so-called membership function (or χ-function), originally proposed in Ref.[32] as solution to the PCCA+ problem. The χ-function is defined as an affine rescaling of ψ_1 to the interval $[0,1]$:

$$\chi(x) = a_0 + a_1 \psi_1(x), \tag{9}$$

where

$$\begin{cases} a_0 = \dfrac{\max_x \psi_1(x)}{\max_x \psi_1(x) - \min_x \psi_1(x)}, \\ a_1 = -\dfrac{1}{\max_x \psi_1(x) - \min_x \psi_1(x)}, \end{cases} \tag{10}$$

By construction, the χ-function satisfies the same properties as ψ_1, while offering a natural probabilistic interpretation: $\chi(x)$ can be seen as the degree of membership of state x to one of the two metastable macro-states. Although its definition assumes the existence of two metastable states, its applicability is not restricted to two-state systems. When multiple metastable states are present, the leading non-trivial eigenfunction ψ_1 still encodes the dominant slow relaxation process, distinguishing between two macroscopic sets of states separated by the slowest dynamical barrier. As a result, the χ-function represents an ideal candidate for the ordering parameter in the MoKiTo framework.

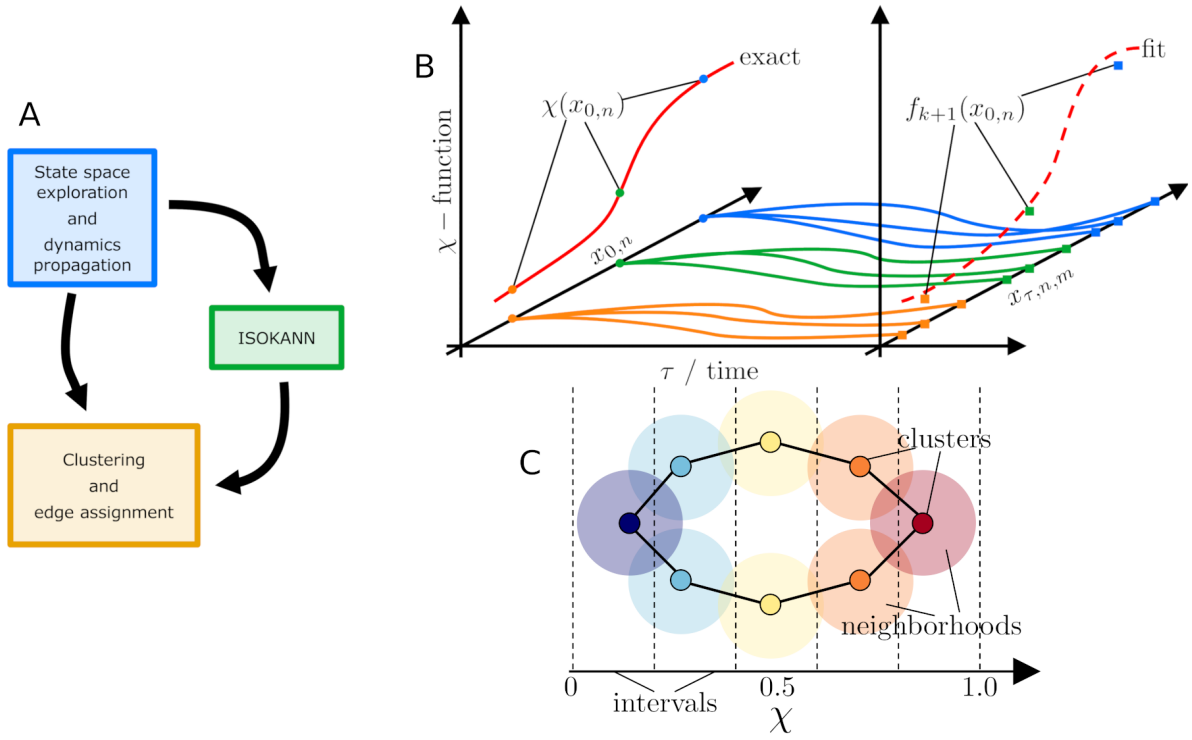
A valid alternative is the committor function $q(x)$, solution of the backward Kolmogorov equation

$$\mathscr{Q}q = 0\,, \tag{11}$$

with absorbing boundary conditions $q(x) = 0$ for $x \in A$ and $q(x) = 1$ for $x \in B$, where A and B denote two disjoint sets representing the macro-states of the system. The committor is acknowledged as the optimal reaction coordinate for a two-state system with clearly defined boundary sets A and B[28–31], as its isosurfaces coincide with the transition surfaces of the dynamics of the system. Furthermore, the committor admits the spectral representation

$$\begin{aligned} q &= \sum_i b_i \psi_i && (12)\\ &\approx b_0 + b_1 \psi_1\,, && (13) \end{aligned}$$

where $b_i = \langle q, \psi_i \rangle_\pi$ are the expansion coefficients of the committor in the eigenfunction basis and where the approximation holds for a two-state reversible system with a large spectral gap. In this case, the committor reduces to an affine transformation of the leading non-trivial eigenfunction, just as the membership function does. Therefore, both $\chi(x)$ and $q(x)$ induce the same ordering of configurations and essentially the same iso-surfaces. However, the χ-function can be obtained without any prior specification of sets A and B, making it more general than the committor and the preferred choice for the ordering parameter in MoKiTo. This advantage becomes particularly relevant in multi-state landscapes, where suitable definitions of A and B may not exist, while the χ-function remains unambiguously defined.



Description of the method.

FIG. 2. (A) MoKiTo workflow diagram. Constructing the MKM using a three-stage procedure. (B) In the case of unknown χ-function, it is necessary first to propagate short trajectories starting from $x_{0,n}$, then to apply an arbitrary function f_k and estimate the average for each state $x_{0,n}$, and then to apply the shift-scale function S as in Eq. 17. (B) ISOKANN scheme. Given the exact χ-function, it is possible to calculate $\chi(x_{0,n})$ as shown in the left panel. The states thus found are then connected via a function fitted with an FNN. (C) Clustering and edge assignment scheme. The states representing the state space are first subdivided into intervals according to the χ-function. Then, the states of the same interval are clustered by CNN clustering algorithm and edges are found by overlapping the neighborhoods.

III. Methods

MoKiTo is a framework that allows one to extract kinetic information from a set of short MD trajectories to construct graphs that highlight the dominant pathways between macro-states. This procedure, as outlined in Fig. 2-(A), can be summarized in three stages:

- The first stage focuses on exploration of the state space and propagation of dynamics by means of MD simulations.

- The second stage involves definition of an ordering parameter, the so-called χ-function, by means of the ISOKANN algorithm.

- The third stage uses a clustering algorithm to cluster the MD data filtered by the χ-function. Then the edges of the graph are assigned according to the overlap of neighborhoods.

A. State space exploration and dynamics propagation

The objective of this stage is to identify N representative states of the state space Γ, which can be achieved through different methods depending on the system under investigation. As we will show in our numerical experiments, for systems characterized by low energy barriers, e.g., short chains of amino acids, conventional Molecular Dynamics or Monte Carlo simulations are the most convenient solution. Instead, for bigger systems such as proteins, we recommend the use of enhanced techniques, such as Simulated Tempering MD (STMD) simulations[58], replica exchange MD[59], umbrella sampling[60], or metadynamics[61–63]. Here, in the fourth example of chicken villin headpiece protein, we opted for STMD simulations, where the problem of getting stuck in local minima is overcome by dynamically adjusting the temperature during the simulation. STMD simulations are advantageous over the other proposed methods because they do not require a set of collective variables to be chosen a priori. However, states generated by STMD simulations require further minimization and equilibration to ensure they represent the equilibrium distribution of the canonical ensemble.

Once a set of N representative states has been built, it is necessary to propagate the dynamics. For this purpose, we perform M conventional MD simulations of length τ for each representative state. It is important to emphasize that, at this stage, the simulations must accurately reflect the true dynamics of the system under investigation. Consequently, enhanced sampling techniques that modify the potential energy surface or alter the system's temperature are not applicable. From now on, we will use the notation X_0 to denote the set of N initial states and X_τ to denote the set of $N \times M$ final states of MD trajectories of length τ.

B. ISOKANN

To compute the χ-function for a multidimensional system, we use ISOKANN[34], an iterative algorithm inspired by the Von Mises iteration method[64], that proceeds according to the update rule

$$f_{k+1}(x) \;=\; S\mathscr{K}_\tau f_k(x)\,, \tag{14}$$

where S is a linear transformation, $\mathscr{K}_\tau$ is the Koopman operator and $f_k(x)$ is a real-valued function. The function S, based on the definition of χ-function in Eq. 9, is known as shift-scale function:

$$S\mathscr{K}_\tau f_k(x) = \frac{\mathscr{K}_\tau f_k(x) - \min\left(\mathscr{K}_\tau f_k(x)\right)}{\max\left(\mathscr{K}_\tau f_k(x)\right) - \min\left(\mathscr{K}_\tau f_k(x)\right)}\,. \tag{15}$$

This transformation prevents the convergence of $\mathscr{K}_\tau f_k(x)$ to the dominant trivial eigenfunction, ensures that the updated function remains within the interval $[0,1]$, and guides the convergence toward the desired χ-function:

$$\lim_{k\to\infty} f_{k+1}(x) \;=\; \chi(x). \tag{16}$$

Since we do not know an analytical expression of $\mathscr{K}_\tau$, nor a matrix representation of it, we approximate the conditional expectation in Eq. 3 as

$$f_{k+1}(x_{0,n}) \;=\; S\frac{1}{M}\sum_{m=1}^{M} f_k(x_{\tau,n,m}|x_0 = x_{0,n})\,, \tag{17}$$

where $x_{\tau,n,m} \in X_\tau$ is the final state of the mth trajectory started in $x_{0,n} \in X_0$.

In this formulation, we assume an analytical expression for $f_k(x)$. However, the application of Eq. 17 yields only the scalar values $f_{k+1}(x_{0,n})$ at the sample points $x_{0,n}$. Since we do not have an explicit analytical form for $f_{k+1}(x)$, the question arises: how can we apply the Koopman iteration at the next step? To address this, we seek an analytical function that best fits the N scalar values $f_{k+1}(x_{0,n})_{n=1}^{N}$ using regression techniques. For low-dimensional systems, methods such as spline interpolation or radial basis functions are often preferable due to their simplicity and low number of trainable parameters. In contrast, for high-dimensional systems, Feedforward Neural Networks (FNNs) are recommended due to their superior computational power and scalability. In the examples presented in this manuscript, we employed an FNN, whose training procedure is described in the SI. Fig. 2-(B) summarizes the ISOKANN procedure.

C. Clustering and edge assignment.

We subdivide the χ-function into N_I disjoint intervals containing states with similar χ-value. The number of intervals into which the χ-function is subdivided is arbitrary, but ideally there should be a sufficient number of states representing similar macroscopic behavior, i.e. macrostates and transition states. This partial clustering based on the χ-function is useful to reduce the complexity of the data while preserving important properties such as Markovianity[65].

Once the intervals of the χ-function have been defined, we cluster states in state space Γ. Several algorithms could be used. Here, we chose the Common Nearest Neighbor (CNN) clustering algorithm[66], an unsupervised clustering algorithm that uses local density information to identify clusters of data points without prior knowledge of the number of clusters. CNN clustering tends to work well with non-linearly separable data and has already been shown to be suitable for the study of molecular systems[67,68]. The key assumption of this algorithm is that two states are more likely to belong to the same cluster if they share a significant number of neighbors. The algorithm is then controlled by two key parameters: the radius of the neighborhood ε and the number of nearest neighbors θ. Two states are considered neighbors if they are less than a distance ε apart from each other:

$$x_i \text{ and } x_j \text{ are neighbors if } |x_i - x_j| < \varepsilon\,. \tag{18}$$

Then, for every pair of states, the intersection of their respective nearest neighbor sets is determined: if x_i and x_j share at least θ neighbors, they belong to the same cluster. The algorithm determines K clusters $\Omega_1, \Omega_2, ..., \Omega_K$, where K is an output parameter, however, like many density-based methods, its performance is sensitive to the choice of θ and ε. Following the indications in Refs.[66,67], the parameters should be chosen on the basis of the histogram of pairwise Root Mean Square Distances (RMSDs) between the states in the dataset: ε should be set to a value slightly smaller than the first maximum of this histogram, while θ is varied until adequated sized clusters are found.

The last step in the construction of the graph is the assignment of edges, i.e. finding the connections between pairs of clusters Ω_i and Ω_j $\forall i, j = 1, 2, ..., K$. For this purpose, we assume that a transition in an infinitesimal time span can only take place between similar conformations, both macroscopically and microscopically. Thus, to determine edges, we look for clusters belonging to consecutive intervals that have common states in their neighborhood. Indeed, a transition between

clusters in consecutive intervals indicates a transition between states with similar macroscopic properties and clusters that share states in their neighborhood, ensuring that the transition occurs between states with similar conformational structure. In practice, the procedure begins by aligning the states within a cluster to minimize the RMSD and computing the average structure of the cluster. Then, the neighborhood of the cluster, whose size is determined by a threshold r_n, is identified by calculating the RMSD between the average structure and all the states in the X_0 dataset. The procedure to determine the clusters and the edges is schematized in Fig. 2-(C).

The resulting graph, referred to as the Molecular Kinetics Map (MKM), consists of clusters ordered according to the values of the χ-function. The initial cluster Ω_0 (with $\chi \approx 0$) and the final cluster Ω_K (with $\chi \approx 1$) represent the dominant macro-states of the system; typically reactants and products in a chemical reaction, or, in the context of biomolecules, the unfolded-folded states of a protein, bound-unbound conformations, and so on. The MKM can then be used to identify the principal pathways connecting the macro-states. We define an arbitrary pathway χ^p connecting the initial cluster Ω_1 to the final cluster Ω_K as a sequence of clusters along the reaction coordinate χ as

$$\chi^p = \{\Omega_{p_1}, \Omega_{p_2}, ..., \Omega_{p_L}\}, \tag{19}$$

where L is the length of the pathway, $\Omega_{p_1} = \Omega_1$ and $\Omega_{p_L} = \Omega_K$ are the initial and final clusters of the MKM respectively, and each $\Omega_{p_j} = \Omega_2, \Omega_3, \ldots, \Omega_{K-1}$.

D. Free energy profiles and calculation of transition rates.

We assume that the N initial states $x_0, n \in X_0$ are distributed according to the canonical equilibrium distribution

$$\pi(x) = \frac{1}{Z} \exp\left(-\frac{1}{\beta} V(x)\right), \tag{20}$$

where Z is the canonical partition function that normalizes the distribution and $\beta = 1/k_B T$, with temperature T and Boltzmann constant k_B. Then, given a pathway χ^p as defined in Eq. 19, the probability distribution projected onto the clusters Ω_{p_i} of the pathway is defined as

$$\pi(\Omega_{p_i}) = \frac{\int_\Gamma \delta[\Omega_{p_i}(x) - \Omega_{p_i}] \exp(-\beta V(x))\, dx}{\int_\Gamma \exp(-\beta V(x))\, dx}, \tag{21}$$

where δ is the Dirac δ-function. Eq. 21 describes the Boltzmann weight of each cluster of the pathway and it is associated with the free energy

$$E(\Omega_{p_i}) = -\frac{1}{\beta}\log\pi(\Omega_{p_i})\,. \tag{22}$$

Plotting the free energy along each pathway as a function of the reaction coordinate χ provides a natural way to classify pathways according to their thermodynamic likelihood; i.e., pathways traversing lower free energy barriers are more probable. In addition, applying the Square Root Approximation method (SqRA)[23–25] we approximate the transition rate between adjacent clusters of the same pathway as

$$k_{p_i,p_{i+1}} \propto \sqrt{\frac{\pi_{p_{i+1}}}{\pi_{p_i}}}\,, \tag{23}$$

Furthermore, we calculate the effective transition rate from the initial to the final state of a pathway as

$$k^p = \left(\sum_{i=1}^{L-1}\frac{1}{k_{p_i,p_{i+1}}}\right)^{-1}\,, \tag{24}$$

which corresponds to the harmonic mean of the inverse transition rates along the pathway, and provides a coarse-grained estimate of the overall kinetics.

IV. Results

A. Two-dimensional system

As an illustrative example, we considered a two-dimensional system governed by overdamped Langevin dynamics and defined by the potential energy function

$$V(x,y) = 10(x^2-1)^2 + 5xy + 10(y^2-1)^2 + 2.2x\,, \tag{25}$$

illustrated in Fig. 3-(A). The potential is characterized by 4 local minima of different height: the deepest is corner $C_1 = (-1,1)$, followed by $C_4 = (1,-1)$, $C_3 = (-1,-1)$ and $C_2 = (1,1)$, in order from lowest to highest . To give physical meaning to the problem, we assumed that the potential has energy units $\mathrm{kJ\,mol^{-1}}$, and that generates forces $-\nabla_x V$ and $-\nabla_y V$ on a fictitious particle of mass $m = 1\,\mathrm{amu}$ that moves on a flat surface. We also assumed standard thermodynamic parameters: the temperature of the system was $T = 300\,\mathrm{K}$ with molar Boltzmann constant

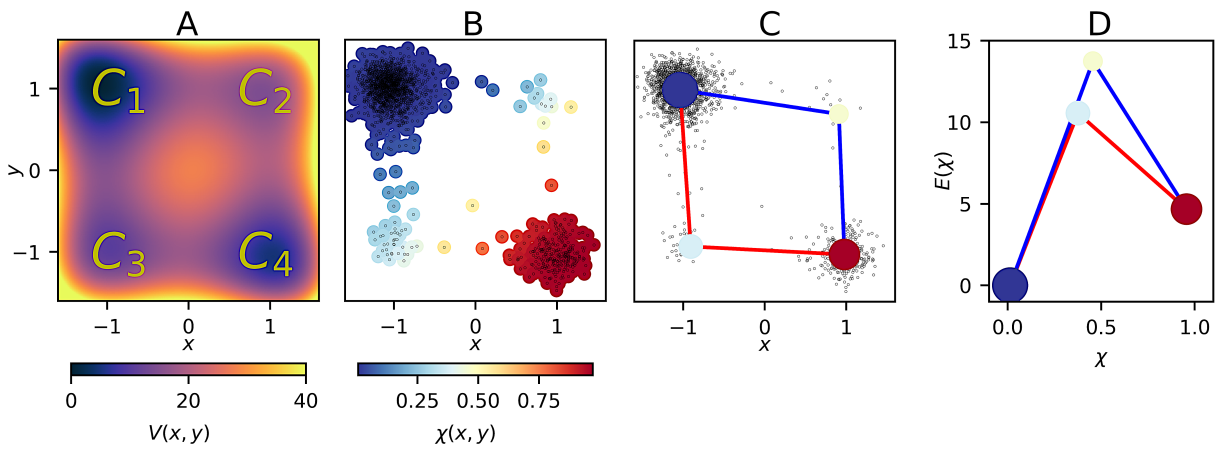


FIG. 3. Results of the two-dimensional system. (A) Potential energy function of the two-dimensional system; (B) States of the two-dimensional system extracted from a trajectory and colored according to the membership function $\chi(x,y)$. (C) MKM of the two-dimensional system projected onto the Cartesian space. The black dots represent the initial states X_0; (D) Free energy profiles of the two-dimensional system.

$k_B = 8.314 \times 10^{-3}\,\mathrm{kJ\,K^{-1}\,mol^{-1}}$. This choice of the parameters makes sure metastability, indeed the system's thermal energy $\beta^{-1} = k_B T = 2.49\,\mathrm{kJ\,mol^{-1}}$ is significantly smaller than the height of the barriers along x and y. The interaction of the particle with the environment is modeled via a friction coefficient $\gamma = 1\,\mathrm{ps^{-1}}$ and a diffusion constant $D = {}^{k_B T}/_{m\gamma} = 2.49\,\mathrm{nm^2\,ps^{-1}}$ in each direction.

State space exploration and dynamics propagation. We solved the overdamped Langevin dynamics equations of motion

$$\begin{cases} dx_t = -\beta D \nabla_x V(x_t, y_t)\, dt + \sqrt{2D}\, dW_x \\ dy_t = -\beta D \nabla_y V(x_t, y_t)\, dt + \sqrt{2D}\, dW_y \end{cases}, \tag{26}$$

where W_x and W_y are two independent and uncorrelated Wiener processes, applying the Euler-Maruyama scheme[69] with a timestep of $\Delta t = 0.001\,\mathrm{ps}$.

First, we generated a sufficiently long trajectory of 1×10^7 timesteps which covers the relevant regions of the potential. Then, we extracted 4000 initial states equally spaced from the trajectory, i.e. one each 1000 timesteps, and carried out 10 short trajectories of 10 timesteps from each initial state.

As suggested in the Methods section, we organized the data into two arrays: X_0 of shape

(4000,2) containing the coordinates of the initial states, and X_τ of shape (4000,10,2) containing the coordinates of the final states of the short trajectories.

ISOKANN. To construct the χ-function, we applied ISOKANN. For regression, we used an FNN with three layers and the sigmoid function as activation function. The FNN was implemented using PyTorch[70], the input layer had 2 nodes, one for each coordinate of the system, the hidden layer had 128 nodes and the output layer had 1 node corresponding to the χ-value. The optimization of the FNN parameters was performed using the Stochastic Gradient Descent (SGD) algorithm[71] and minimizing the mean squared error (squared ℓ^2-norm). At each ISOKANN iteration, we trained for 15 epochs, iterating over randomly generated batches of size 100, with an initial learning rate of 0.001. This choice of hyperparameters is the result of a random search and leads to a convergence of the χ-function in 4 iterations (SI, Fig. S1). To validate the model, we monitored training and validation losses and observed that they decrease in parallel without an increase in the gap or a rebound in validation loss, indicating no evidence of overfitting.

The χ-function, evaluated at each initial point X_0, is illustrated in Fig. 3-(B). The corners C_1 and C_4, colored by blue ($\chi \approx 0$) and red ($\chi \approx 1$) respectively, are the two main macro-states, i.e. the regions of state space most visited by the particle. Corners C_3 and C_2 are colored with a gradient of colors blue-yellow ($\chi \approx 0.5$), and can be interpreted as transition state states.

MKM construction. To construct the MKM representing the macro-states and the main pathways of the dynamics, we proceeded in two steps: first we grouped the states by dividing the χ-function into 3 regular intervals and finding for each interval 1703, 32 and 263 states (x,y) respectively, then we applied the CNN clustering algorithm to group the states into smaller clusters having similar χ-value. For the CNN clustering algorithm, we chose as radius of the neighborhood $\varepsilon = 1.0$ and as number of nearest neighbors $\theta = 5$, finding $N_c = 4$ clusters. To connect the clusters, we searched for areas of overlap between neighborhoods of clusters belonging to consecutive intervals, using as a threshold the Euclidean distance $r_n = 0.6$. The MKM is shown in Fig. 3-(C).

Observations. The potential energy function has four macro-states and by means of the χ-function we identify two macroscopic regions at corners C_1 and C_4, and one transition region that includes both corners C_2 and C_3. There are two main pathways from corner C_1 to corner C_4 and vice versa. In Fig. 3-(D), we show the free energy profiles as defined in Eq. 22. Since the blue and red clusters

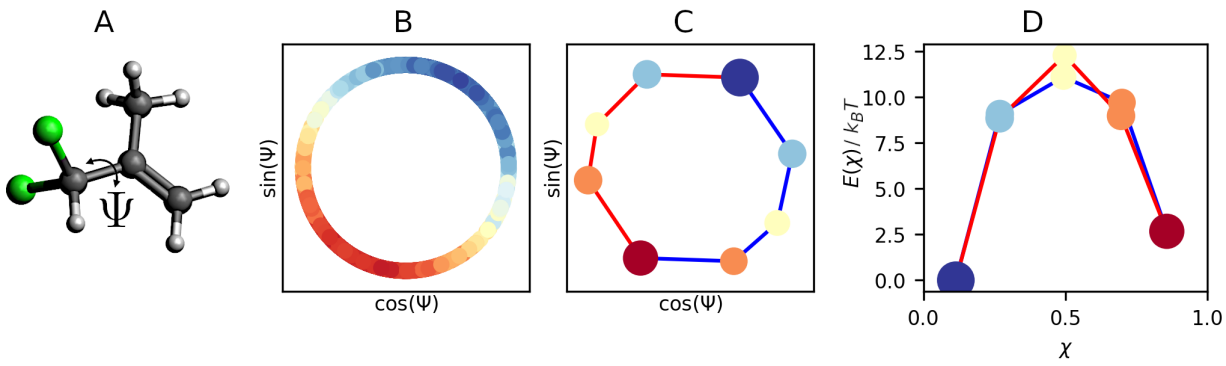


FIG. 4. Results of 33-Dichloroisobutene molecule. (A) 33-Dichloroisobutene molecule; (B) χ-function projected onto the main torsion angle of the 33-Dichloroisobutene; (C) MKM projected onto the main torsion angle of the 33-Dichloroisobutene; (D) Free energy profiles of the 33-Dichloroisobutene molecule.

are the largest, their energy levels are low, the yellow clusters, instead, are higher and correspond to transition states. From this graph, we also deduce that the red path ($C_1 - C_3 - C_4$) is the most likely path, as the system needs less energy to visit corner C_3 than corner C_2. By SqRA (eqs. 23, 24), we estimated the effective transition rates of the two paths: $k_+^r = 0.07$ and $k_+^b = 0.15$ for the red and blue path respectively. Thus the reaction along the blue path is about 2.14 times faster than the red one. We estimated also the transition rates for the reverse reactions: $k_-^r = 0.18$ and $k_-^b = 0.40$ respectively.

B. 33-Dichloroisobutene

As the first molecular system example, we studied 3,3-Dichloroisobutene ($C_4H_6Cl_2$), a dichloro derivative of isobutene, represented in Fig. 4-(A). The compound has 12 atoms, for a total of 36 dimensions, and the rotation around the torsion angle Ψ (C_3-C_2-C_4-Cl_1) is known to be the slowest process of the system. Thus, we used the torsion angle Ψ as relevant coordinate to visualize the results.

State space exploration and dynamics propagation. We performed MD simulations using the package OpenMM[72] with the Generalized Amber Force Field[73]. To simulate an explicit solvent, we used the TIP3P-FB water model[74] with a padding distance of 1.4 nm which generates a box of 682 water molecules for 33-Dichloroisobutene. We assumed Langevin dynamics and applied the Langevin leapfrog integrator[75] with $\gamma = 1.0\,\mathrm{ps}^{-1}$ as friction coefficient, and $\Delta t = 2\,\mathrm{fs}$ as integrator

timestep. Non-bonded interactions between atoms, such as Coulomb forces and Lennard-Jones forces, were calculated by Particle-Mesh Ewald (PME) method[76] and interactions between atoms more than 1 nm apart were truncated. The lengths of all bonds involving a hydrogen atom have been constrained. Before doing the first simulation, we brought the system to a local energy minimum, then we equilibrated the system with a 20 ps simulation to obtain a state belonging to the NVT ensemble, with temperature equal to $300 \pm 6\,\mathrm{K}$.

From a trajectory of 40×10^6 timesteps, corresponding to 80 ns, we extracted 4000 states $x_0 = \{r_0, v_0\}$ (positions and velocities of each atom including the solvent) every 8 ps. Then the states x_0 have been used as initial states for 10 new short trajectories of length 0.02 ps.

ISOKANN. The procedure for constructing the χ-function via ISOKANN was the same as in the previous example. However, instead of providing the Cartesian coordinates of the atoms, we used the pairwise distances between all the atoms of the system (without the water). This increases the number of dimensions of the χ-function to $12\cdot(12-1)/2 = 66$, but ensures that χ is invariant with respect to translations and rotations. As a model to approximate the χ-function, we used an FNN with 4 layers (2, 66, 33, 1 nodes), however, as activation function we used Leaky ReLU which performs better than sigmoid in high-dimensionality regression tasks. We set the initial learning rate for the SGD algorithm to 0.01 and applied a weight decay of 0.01 for regularization. The model converged after 279 iterations, achieving training and validation losses smaller than 10^{-3} ; no train–validation divergence or validation-loss rebound was observed, and χ was stable over ISOKANN iterations (SI, Fig. S2).

The χ-function, evaluated in each initial point x_0, is reported in Fig. 4-(B). For ease of reading, we have projected the χ-function onto the unit circle, i.e. the values $\cos(\Psi)$ and $\sin(\Psi)$, where Ψ is the torsion angle. We clearly distinguish macro-states colored with red and blue, and the transition states colored with yellow. The correlation between χ-function and torsion angle Ψ is equal to 0.9, confirming that the latter is a good choice as a relevant coordinate to describe the slowest process of the system.

Clustering and edge assignment. To construct the MKM, we discretized the χ-function into 5 equal intervals, then applied the CNN clustering algorithm. Clustering was done by pre-computing the Root Means Square Distance (RMSD) matrix. Then, after analyzing the distribution of RMSDs, we identified the CNN clustering parameters: $\varepsilon = 0.09, 0.08, 0.09, 0.07, 0.09$, and $\theta = 5$

(for each interval). With this setting, we obtained 8 clusters, 2 for macro-states and 6 for transition states. To find edges, we used as threshold for the neighborhoods $r_n = 0.05$. The MKM projected onto the unit circle of the angle Ψ is shown in Fig. 4-(C).

Observations. We observe two main clusters: the red one ($\Psi \approx \pi/4$) corresponds to a conformational state where both chlorines are staggered with the methylene group; the blue one ($\Psi \approx 3\pi/4$) corresponds to a conformational state where both chlorines staggered with the methyl group. The two macro-states are connected by two pathways that correspond to the rotations of the torsion angle Ψ. The free energy profiles, illustrated in Fig. 4-(D), also reveal the energy barriers of the paths. Since the system is perfectly symmetrical with respect to the rotation of the Ψ angle, the two pathways overlap. In other words, no direction is preferred, and the system can rotate clockwise or counterclockwise with the same probability. However, we observe that there is a higher barrier between the red cluster and subsequent orange clusters than between the blue cluster and subsequent light blue clusters. Then the configurational states belonging to the red cluster are the most stable of the system. The effective transition rates of the two pathways are: $k^r_+ = 0.11$ and $k^b_+ = 0.11$ for the red and blue path respectively; while for the reverse reactions are: $k^r_- = 0.16$ and $k^b_- = 0.16$ respectively. This confirms the symmetry of the system and the equivalence of the two pathways.

C. Hexapeptide VGVAPG

The VGVAPG is an elastin-derived hexapeptide[77], already used to test methods for MD simulations[78,79]. The peptide has 73 atoms, corresponding to 219 dimensions in the Euclidean space. As relevant coordinate, denoted by r_{ee} in the figures, we used the Euclidean distance between the nitrogen atom of the N-terminus and the carboxyl-carbon of the C-terminus (Fig. 5-(A).

State space exploration and dynamics propagation. The MD simulations were carried out with the same settings as for 33-Dichloroisobutene, but the water box was increased to 782 water molecules and the force field was the AMBER ff-14sb[80]. The length of the first simulation was 500×10^6 timesteps, corresponding to $1\,\mu s$, from which, we extracted 1000 initial states for the short trajectories. For each initial state, we produced 10 short trajectories of 1000 timesteps, corresponding to 2 ps.

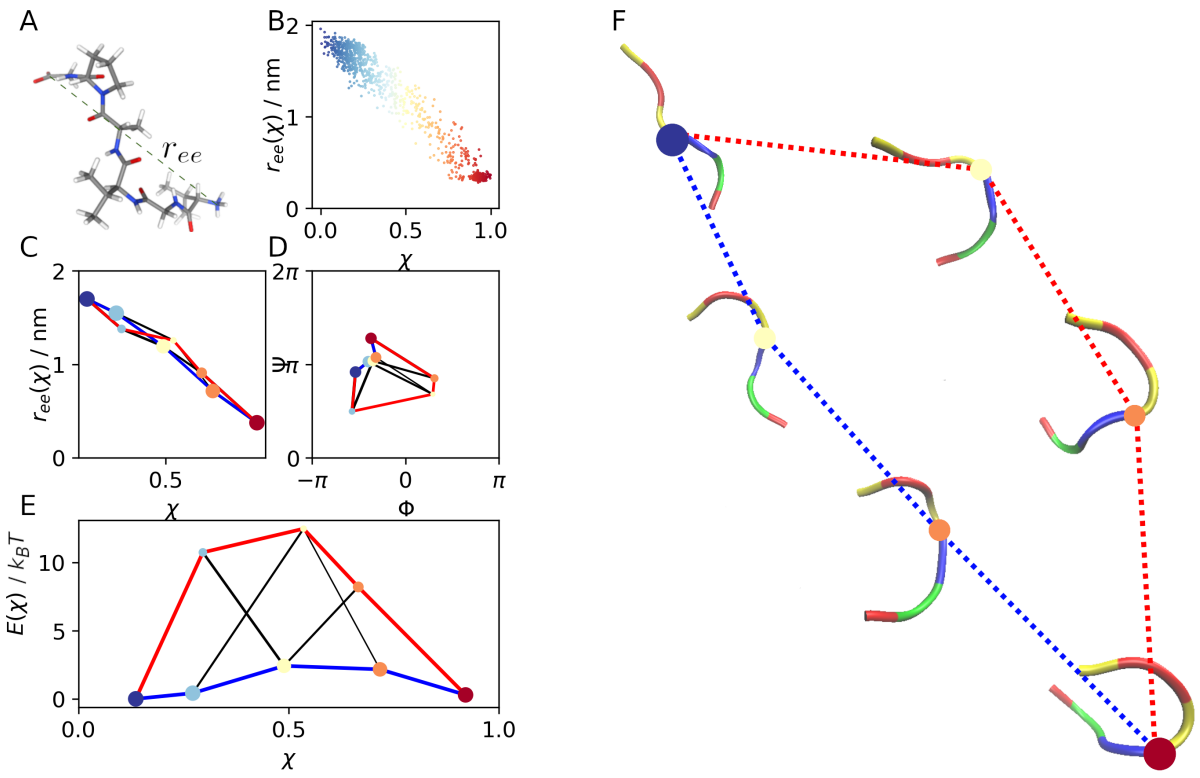


FIG. 5. Results of VGVAPG hexapeptide. (A) VGVAPG molecule; (B) χ-function projected onto the end-to-end distance of the molecule. (C) MKM with representative structures of the VGVAPG; (D) MKM projected onto the end-to-end distance of the VGVAPG. (E) MKM projected onto the Ramachandran plot of the second residue (Glycine 1); (F) Free energy profiles of VGVAPG. The colors of the molecular structure represent the residues: Val (yellow), Gly (red), Pro (blue) and Ala (green).

ISOKANN. We performed a random search to find the best hyper-parameters of the neural network and determined as optimal parameters, 1752 nodes in the hidden layer, 0.001 as initial learning rate and 0.005 as weight decay. The ISOKANN algorithm was performed for 57 iterations, training and validation losses stabilized around 4×10^{-4} with no train–validation divergence, and χ remained stable (SI, Fig. S3). The χ-function is plotted in Fig. 5-(B). We observe a large macro-state, corresponding to $\chi \approx 0.0$ (blue), which includes configurations whose relevant coordinate r_{ee} ranges from 0.7 to 2 nm; the transition states, with $\chi \approx 0.5$ (yellow), range from 0.35 to 1.7 nm and the macro-state corresponding to $\chi \approx 1.0$ (red), includes configurations whose relevant coordinate ranges from 0.25 to 0.35 nm. The correlation between χ and r_{ee} is 0.98.

MKM construction. We divided the χ-function into 5 equal intervals between 0 and 1. Then, by analyzing the distribution of RMSDs, we determined the parameters for the CNN clustering: $\varepsilon = 0.3, 0.23, 0.15, 0.17, 0.3$ and $\theta = 5$ for each interval. Thus, we found 1, 2, 2, 2 and 1 cluster for each interval respectively. The MKM, obtained with $r_n = 0.3$, is reported in Fig. 5, where we propose different representations: in Fig. 5-(C), we show the complete MKM projected onto the end-to-end distance of the peptide; in Fig. 5-(D) we show the MKM projected onto the Ramachandran plot of the first Glycine (G1) of the peptide; in Fig. 5-(E), we show the free energy profiles; in Fig. 5-(F), we show the main representative structures of the backbone of the peptide (omitting the less relevant clusters of the MKM);

Observations. The blue cluster ($\chi \approx 0$) comprises completely open structures with $r_{ee} > 1.5\,\text{nm}$, while the red cluster ($\chi \approx 1$) represents closed structures with $r_{ee} \approx 0.3\,\text{nm}$. Since distance r_{ee} is highly correlated with χ, we cannot distinguish multiple paths from Fig. 5-(C). Thus, to better characterize the dynamics and describe the opening-closing mechanism of the peptide, we analyzed the Ramachadran plot of each residue. Here, in Fig. 5-(D), we report the Ramachadran plot of the second residue (the first Glycine in the chain VGVAPG), as it shows the most interesting dynamics. First, we observe that most of the clusters, in particular the clusters belonging to the blue path, are located in quadrant II and III where $\Phi < 0$. We therefore deduce that the closure of the hexapeptide along the blue pathway does not lead to a significant rotation of the Glycine torsion angles. However, one orange cluster ($\chi \approx 0.7$) is located in quadrant IV ($\Phi \approx 0.95$), indicating that the red pathway involves a wide rotation of the Φ torsion angle of the Glycine: first about 120 degrees anti-clockwise, then again about 110 degrees clockwise. The effective transition rates of the two pathways are: $k^r_+ = 0.09$ and $k^b_+ = 0.23$ for the red and blue path respectively; while for the reverse reactions are $k^r_- = 0.12$ and $k^b_- = 0.24$ respectively. We conclude that the red pathway is more energy-intensive and, consequently, less probable.

D. Villin headpiece subdomain

As a last example, to demonstrate the applicability of our approach to large systems, we studied the villin headpiece subdomain[53–56] which is one of the most studied protein for understanding protein folding. Villin consists of 35 residues (582 atoms), and in its folded structure, it forms 3 α-helices as shown in Fig. 6-(A): residues 4-8 form helix H_1 (green), residues 15-18 form helix

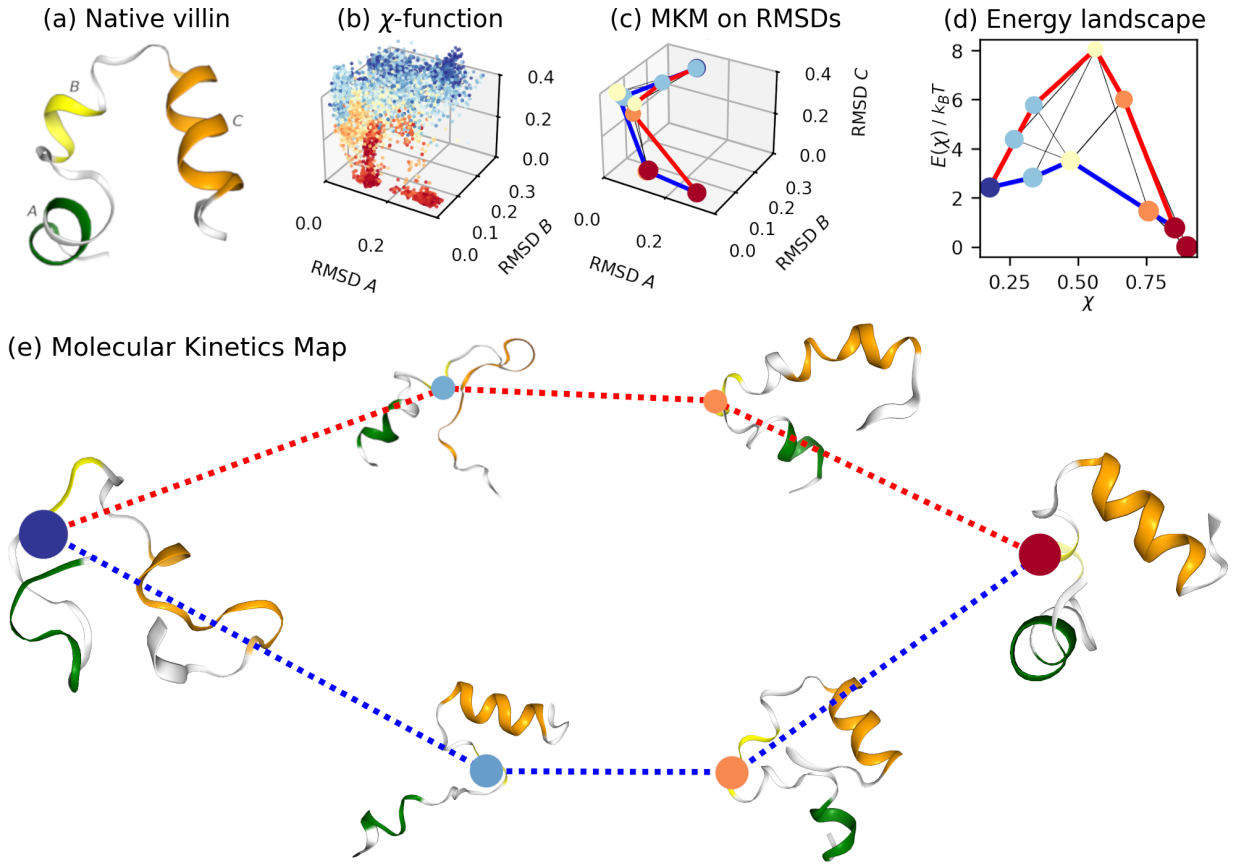


FIG. 6. Results of villin headpiece subdomain. (A) The X-ray crystal structure of villin headpiece solved at pH 6.7, green, yellow and orange colors identify the helix H_1 (green), H_2 (yellow) and H_3 (orange) respectively; (B) χ-function projected onto the RMSD of segments A and B. (C) MKM with representative structures of the villin protein; (D) MKM projected onto the RMSD of segments A and B of the villin protein; (E) Free energy profiles of the villin protein.

H_2 (yellow), residues 23-32 form helix H_3 (orange). To have a 2-dimensional representation of the molecular system, we used the RMSD of the segment A (residues 3–21), which includes H_1 and H_2, and segment B (residues 15–33), which includes H_2 and H_3, with respect to the corresponding segments of the X-ray crystal structure solved at pH 6.7 deposited in the RCSB protein data bank repository (PDB ID: 1YRF[81]) as done in Ref.[55]. Thus, a structure with low RMSD values corresponds to a folded structure, a structure with high RMSD values is an unfolded structure, and partially folded structures correspond to a situation where only the RMSD of one segment has low values.

State space exploration and dynamics propagation. The folding timescale of the villin protein is 2.8 μs, but a complete exploration of the space of states requires a conventional MD simulation of more than 2.8 μs[56]. Alternatively, we carried out STMD simulations utilizing the dedicated OpenMM module for exploring the state space and selecting representative structures. First, we prepared a complete extended structure with PyMol[82]. Then, we minimized the structure and equilibrated the system for 20 ps reaching a partially folded structure. At this point, we carried out 6 independent replicas of 1 μs with temperatures ranging from 273 K to 500 K. All the other parameters and options were as in the previous examples, the box contained 2713 water molecules. From each replica, we extracted 1000 structures, for a total of 6000 structures, which constitute the set X_0 of initial states. However, since temperature was a dynamic variable, we further equilibrated these structures for 100 ps to have a sample representing the Canonical Ensemble at $T = 300\,\mathrm{K}$. Afterward, we ran 10 short MD simulations of 1000 timesteps (0.2 ps) for each initial state.

ISOKANN. The ISOKANN algorithm was applied as before, but we changed the input coordinates. Indeed, since the system has 582 atoms, the number of pairwise distances is $582\cdot(582-1)/2 = 169071$. Modern neural networks are able to handle this dimensionality, however, as a matter of efficiency and to show the versatility of the method, we preferred to reduce the dimensionality by using the internal coordinates (bonds, angles and torsion angles) of the backbone (140 atoms). In this way, we reduced the number of dimensions to 1716. We used a neural network with four layers (1716, 858, 429, 1), the initial learning rate of the SGD algorithm was 0.01 and the weight decay 0.005. Convergence of the training occurred in 95 iterations with a training and validation loss in the order of 10^{-3} (SI, Fig. S4). Convergence was reached within 95 iterations, training and validation losses settled on the order of 10^{-3}, decreasing in parallel with no divergence, and χ remained stable across iterations (SI, Fig. S4).

In Fig. 6-(B), we report the χ-function projected onto the two collective variables. The bottom left corner (small values of RMSDs), contains folded structures with $\chi \approx 1$ (red). As we move away from the corner, we observe transition structures with $\chi \approx 0.5$ (yellow) up to a large area containing unfolded structures with $\chi \approx 0$ (blue).

MKM construction. We divided the χ-function into 5 equal intervals and determined the optimal parameters $\varepsilon = 0.9, 0.5, 0.3, 0.5, 0.5$ and $\theta = 10, 60, 50, 150, 20$ for the CNN clustering algorithm and $r_n = 0.6$ to determine the edges. Thus we found 10 clusters, two clusters correspond to fully

folded and unfolded structures, while the others contain partially folded structures. The MKM is shown in Fig. 6 in several representations: Fig. 6-(C) shows the projection onto the RMSDs of the segments A and B; Fig. 6-(D) shows the free energy profiles as function of the χ-values. (E) shows the edges between the most relevant representative structures of the protein (omitting the less relevant clusters of the MKM);

Observations. We identify multiple folding pathways in the transition from unfolded to folded state, each with a distinct energetic profile. We have highlighted in blue the pathway that requires less energy, and in red the one that requires more, then the blue pathway is the most likely folding route. This pathway corresponds to a scenario in which helix H_3 reaches its folded state more rapidly than the other two helices. Conversely, the red pathway exhibits a process in which helix H_1 stabilizes prior to helices H_2 and H_3. In both cases, the formation of helix H_2 is the slowest process and it starts at $\chi > 0.5$. The effective transition rates of the two pathways are: $k_+^r = 0.21$ and $k_+^b = 0.26$ for the red and blue pathway respectively; while for the reverse reactions are $k_-^r = 0.18$ and $k_-^b = 0.22$ respectively.

E. Sensitivity analysis to χ-discretization and CNN parameters

To assess the robustness of MoKiTo, we quantified how the MKMs depend on the number of χ-intervals N_I and on the CNN parameters (ε, θ). Instead of directly comparing the paths, which is difficult due to changes in the number of clusters and edges as the parameters vary, we examined the spectrum of the infinitesimal generator $\mathscr{Q}$. To this end, we used the Square Root Approximation (SqRA) of the infinitesimal generator[23,24], which allows the operator $\mathscr{Q}$ to be discretised into a rate matrix $\mathbf{Q}$, whose entries Q_{ij} are proportional to the transition rates between connected clusters and are defined as

$$Q_{ij} \propto \sqrt{\frac{\pi_j}{\pi_i}}, \tag{27}$$

where π_i and π_j are the Boltzmann weights of the clusters. Then, we have solved the eigenvalue problem for the rate matrix $\mathbf{Q}$, for each system under investigation. The first three eigenvalues $\kappa_1, \kappa_2, \kappa_3$ show weak sensitivity to the parameters (SI, Figs. Fig. S5, S6, S7). As N_I increases, with fixed ε and θ, the eigenvalues quickly reach a plateau: for the two-dimensional example, the plateau is already reached with $N_I = 3$, while for molecular systems, $N_I = 5$ is sufficient. Varying

ε and θ, with fixed N_I, reveals that the number of clusters decreases monotonically with ε, and the corresponding eigenvalues shift toward zero until convergence. The parameter θ has only a minor effect on both cluster count and eigenvalue spectrum. Therefore, within wide ranges, the main dynamic characteristics detected by MoKiTo are stable, with ε representing the most influential parameter as it controls the granularity of the MKM.

V. Discussion

The dynamics of a molecular system are highly complex and can be represented as a network of pathways connecting clusters of similar configurational states. The idea behind MoKiTo is that molecular dynamics can be decomposed into dynamic processes, known also as Koopman modes, associated with different relaxation time scales. The process associated with the leading non-trivial Koopman eigenfunction ψ_1 represents long-term transitions, such as protein folding/unfolding or biomolecular binding/unbinding events, and indicates that the molecular system evolves over time along a preferred direction. Then, we use the χ-function, a transformation of ψ_1 via Eq. 9, to order the data according to their macroscopic features, facilitating subsequent clustering based on structural similarities.

The first example is particularly useful in showing the power of MoKiTo. The potential energy function has four macro-states and the χ-function reveals correctly the two macroscopic regions and the transition region. Although it would be possible to cluster the initial states X_0 without using the χ-function, the reaction coordinate facilitates the clustering of states with similar macroscopic properties by quantifying how far the intermediate states are away from the starting state (χ-value 0) or close to the end state (χ-value 1).

The other examples show that MoKiTo also applies well to molecular systems of several orders of size. Dichloroisobutene is a small molecule with two macro-states and two possible pathways, the clockwise and counterclockwise rotation of the torsion angle Ψ. MoKiTo captures these properties and reveals that the red configurational states in Fig. 4-(D), i.e. with both chlorine atoms staggered with the methylene group, are the most stable states of the system. This is the expected result, since strain is minimized by distancing chlorines and the methyl group, which is more sterically demanding than the methylene group due to an additional hydrogen.

The third example shows that MoKiTo can be used to assess the quality of a reaction coordinate. The dynamics of VGVAPG are characterized by the opening and closing of the salt-bridge between

the positively charged N-terminus and the negatively charged C-terminus. Intuitively, one might think that the distance between the extreme atoms of the peptide is a sufficient reaction coordinate. Instead, via MoKiTo we revealed more complex dynamics involving two possible rotations of the second residue of the hexapeptide.

Finally, the fourth example shows how to use MoKiTo to identify pathways in the folding/unfolding process of a protein. In contrast to previous examples, where we used conventional MD simulations to sample the representative states of the state space, we used STMD simulations. Indeed, the key requirement for MoKiTo is to have a broad representative set of states regardless of how these were sampled. By contrast, to estimate the χ-function, the short trajectories should respect the true dynamics of the system, but it is not necessary that these trajectories reach a state of thermodynamic equilibrium. We identified two dominant pathways from the unfolded to the folded state. In the pathway highlighted in blue in Fig. 6, helix H_3 bends earlier than H_1 and H_2; by contrast, the red pathway exhibits an earlier bend of H_1. These observations are consistent with previous results in Ref.[55]. From Fig. 6-D, we estimated the difference between the activation free energy barriers of the two pathways to be $4.5\,k_BT$, in close agreement with the value $4.8\,k_BT$ reported in Ref.[83]. Also, mixed pathways appear from the MKM in Fig. 6, where helices form cooperatively at similar time scales, confirming the more recent findings in Ref.[84].

VI. Conclusion

Traditional tools to pathway analysis such as MSMs and TPT are rigorous and widely used, but they typically require explicit discretization and dimensionality reduction. MoKiTo offers a topology perspective: given an ordering parameter χ representing the leading slow mode, it constructs an MKM and identifies reaction pathways as graph objects, together with energy diagrams to weigh the pathways. We learn χ with ISOKANN, which can be trained on short trajectories that do not reach thermal equilibrium and does not require prior definition of macro-states as discrete state space subsets, eliminating the need for detailed prior knowledge of the system. Across our tests, MoKiTo recovered known mechanisms of villin headpiece subdomain, confirming previous results obtained by other methods.

MoKiTo alone does not provide absolute kinetics (rates, MFPT, flows) and still requires the support of external tools, e.g. SqRA, to calculate relevant kinetic quantities. For this reason, MoKiTo should be considered a complementary tool that elucidates the mechanism and topology

of pathways by interacting with established kinetic models.

There is room for improvement for MoKiTo. First, the backward Kolmogorov equation for the χ-function can be turned into a variational principle. This would provide a formulation of the problem that depends only on equilibrium expectations that can be evaluated from independently sampled Boltzmann configurations, including enhanced-sampling or Monte Carlo data, hence trajectory-free. Second, the user parameters in MoKiTo can be chosen more objectively by analyzing the eigenvalue spectrum of the generator on the MKM. For this purpose, we envision automated selection of $(N_I, \varepsilon, \theta)$ based on spectral stability and robustness of the leading eigenvectors. This development would turn MoKiTo into a self-tuning tool providing greater confidence in the recovered pathways.

Data Availability Statement

The software used for this study is openly available on GitHub at `https://github.com/donatiluca/MoKiTo`. The complete dataset, including original MD trajectories and torch arrays representing pairwise distance matrices, is archived on Zenodo (DOI: `10.5281/zenodo.14229803`) and securely stored on the Zuse Institute Berlin server, and can be made available upon reasonable request.

Supporting Information

- Additional information regarding the implementation of the algorithm used to train the FNN within the ISOKANN method.
- Additional Figs. S1, S2, S3, S4 showing the convergence of the χ-function for each example presented in the manuscript, are available in the SI appendix.

Acknowledgments

This research has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Cluster of Excellence MATH+, project AA1-15: “Math-powered drug-design”, additionally it has been partially funded by the Bundesministerium für Bildung und Forschung (BMBF, Federal Ministry of Education and Research) within the project “CCMAI –

Computermodellierung und künstliche Intelligenz zur Aufklärung von physiologischer und pathologischer Rezeptorfunktion". We thank the DFG and the BMBF for their support. VN was supported by Berlin MATH+ under the Visiting Scholar program and by a research stay from the Alexander von Humboldt Foundation.

References

[1]K. A. Dill and J. L. MacCallum, "The protein-folding problem, 50 years on," Science **338**, 1042–1046 (2012).

[2]C. Levinthal, "Are there pathways for protein folding?" Journal de Chimie Physique et de Physico-Chimie Biologique **65**, 44 (1968).

[3]P. S. Kim and R. L. Baldwin, "Intermediates in the folding reactions of small proteins," Annual Review of Biochemistry **59**, 631–660 (1990).

[4]C. R. Matthews, "Pathways of protein folding," Annual Review of Biochemistry **62**, 653–683 (1993).

[5]C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, "A direct approach to conformational dynamics based on hybrid monte carlo," Journal of Computational Physics **151**, 146–168 (1999).

[6]N.-V. Buchete and G. Hummer, "Coarse master equations for peptide folding dynamics," The Journal of Physical Chemistry B **112**, 6057–6069 (2008), pMID: 18232681.

[7]G. R. Bowman, V. S. Pande, and F. Noé, eds., *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, Vol. 797 of Advances in Experimental Medicine and Biology (Springer, Heidelberg, 2014).

[8]B. G. Keller, S. Aleksic, and L. Donati, "Markov state models in drug design," in *Biomolecular Simulations in Structure-based Drug Discovery*, edited by F. L. Gervasio (Wiley-Interscience, Weinheim, 2018) p. 67.

[9]W. E and E. Vanden-Eijnden, "Towards a theory of transition paths," J. Stat. Phys. **123**, 503–523 (2006).

[10]P. Metzner, C. Schütte, and E. Vanden-Eijnden, "Transition path theory for markov jump processes," Multiscale Model. Simul. **7**, 1192–1219 (2009).

[11]F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations," Proc. Natl. Acad. Sci. U. S. A. **106**, 19011–19016 (2009), epub 2009 Nov 3, PMID: 19887634, PMCID: PMC2772816.

[12]Y. Meng, D. Shukla, V. S. Pande, and B. Roux, "Transition path theory analysis of c-src kinase activation," Proceedings of the National Academy of Sciences **113**, 9193–9198 (2016).

[13]D. Nagel, A. Weber, and G. Stock, "Msmpathfinder: Identification of pathways in markov state models," J. Chem. Theory Comput. **16**, 7874–7882 (2020).

[14]F. Paul, F. Noé, and T. R. Weikl, "Identifying conformational-selection and induced-fit aspects in the binding-induced folding of pmi from markov state modeling of atomistic simulations," The Journal of Physical Chemistry B **122**, 5649–5656 (2018).

[15]L. Meng, F. K. Sheong, X. Zeng, L. Zhu, and X. Huang, "Path lumping: An efficient algorithm to identify metastable path channels for conformational dynamics of multi-body systems," J. Chem. Phys. **147**, 044112 (2017).

[16]Y. Qiu, M. S. O'Connor, M. Xue, B. Liu, and X. Huang, "An efficient path classification algorithm based on variational autoencoder to identify metastable path channels for complex conformational changes," J. Chem. Theory Comput. **19**, 4728–4742 (2023).

[17]S. Bray, V. Tänzel, and S. Wolf, "Ligand unbinding pathway and mechanism analysis assisted by machine learning and graph methods," J. Chem. Inf. Model. **62**, 4591–4604 (2022).

[18]S. Motta, L. Callea, L. Bonati, and A. Pandini, "Pathdetect-som: A neural network approach for the identification of pathways in ligand binding simulations," J. Chem. Theory Comput. **18**, 1957–1968 (2022).

[19]S. Bandyopadhyay and J. Mondal, "A deep encoder–decoder framework for identifying distinct ligand binding pathways," J. Chem. Phys. **158**, 194103 (2023).

[20]D. Ray and M. Parrinello, "Data-driven classification of ligand unbinding pathways," Proc. Natl. Acad. Sci. U. S. A. **121**, e2313542121 (2024).

[21]M. Invernizzi and M. Parrinello, "Exploration vs convergence speed in adaptive-bias enhanced sampling," Journal of Chemical Theory and Computation **18**, 3988–3996 (2022).

[22]H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE transactions on acoustics, speech, and signal processing **26**, 43–49 (1978).

[23]L. Donati, M. Heida, B. G. Keller, and M. Weber, "Estimation of the infinitesimal generator by square-root approximation," J. Phys. Condens. Matter **30**, 425201 (2018).

[24]L. Donati, M. Weber, and B. G. Keller, "Markov models from the square root approximation of the fokker–planck equation: calculating the grid-dependent flux," J. Phys. Condens. Matter **33**, 115902 (2021).

[25]L. Donati, M. Weber, and B. G. Keller, "A review of Girsanov reweighting and of square root

approximation for building molecular Markov state models," J. Math. Phys. **63**, 123306 (2022).

[26]G. Singh, F. Memoli, and G. Carlsson, "Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition ," in *Eurographics Symposium on Point-Based Graphics*, edited by M. Botsch, R. Pajarola, B. Chen, and M. Zwicker (The Eurographics Association, 2007).

[27]P. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, and G. Carlsson, "Extracting insights from the shape of complex data using topology," Sci. Rep. **3** (2013).

[28]C. Dellago, P. G. Bolhuis, and D. Chandler, "Transition path sampling and the calculation of rate constants," The Journal of Chemical Physics **108**, 1964–1977 (1998).

[29]P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, "Transition path sampling: Throwing ropes over rough mountain passes, in the dark," Annual review of physical chemistry **53**, 291–318 (2002).

[30]C. Dellago, P. G. Bolhuis, and P. L. Geissler, "Transition path sampling," in *Advances in Chemical Physics* (John Wiley & Sons, Ltd, 2002) Chap. 1, pp. 1–78, https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471231509.ch1.

[31]P. V. Banushkina and S. V. Krivov, "Optimal reaction coordinates," WIREs Computational Molecular Science **6**, 748–763 (2016).

[32]P. Deuflhard and M. Weber, "Robust Perron cluster analysis in conformation dynamics," Linear Algebra Appl. **398**, 161–184 (2004).

[33]R. T. McGibbon, B. E. Husic, and V. S. Pande, "Identification of simple reaction coordinates from complex dynamics," The Journal of Chemical Physics **146**, 044109 (2017), https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/1.4974306/15522976/044109_1_online.pdf.

[34]R. J. Rabben, S. Ray, and M. Weber, "ISOKANN: Invariant subspaces of Koopman operators learned by a neural network," J. Chem. Phys. **153**, 114109 (2020).

[35]A. Sikorski, E. Ribera Borrell, and M. Weber, "Learning koopman eigenfunctions of stochastic diffusions with optimal importance sampling and isokann," Journal of Mathematical Physics **65**, 013502 (2024).

[36]L. Donati, C. Schütte, and M. Weber, "The kramers turnover in terms of a macro-state projection on phase space," Mol. Phys. **0**, e2356748 (2024).

[37]M. Weber, *Meshless Methods in Conformation Dynamics*, Ph.D. thesis, FU Berlin (2006).

[38]K. Fukui, "A formulation of the reaction coordinate," Journal of Physical Chemistry **74**, 4161–

4163 (1970).

[39]A. Tachibana and K. Fukui, "Theoretical study of the reaction coordinate in chemical reactions," Theoretica Chimica Acta **57**, 81–90 (1980).

[40]W. Quapp and D. Heidrich, "Analysis of the concept of minimum energy path on the potential energy surface of chemically reacting systems," Theoretica Chimica Acta **66**, 245–260 (1984).

[41]R. Olender and R. Elber, "Yet another look at the steepest descent path," Journal of Molecular Structure: THEOCHEM **398–399**, 63–71 (1997).

[42]M. Heymann and E. Vanden-Eijnden, "The geometric minimum action method: A least action principle on the space of curves," Communications on Pure and Applied Mathematics **61**, 1052–1117 (2008).

[43]P. Eastman, N. Grønbech-Jensen, and S. Doniach, "Simulation of protein folding by reaction path annealing," Journal of Chemical Physics **114**, 3823–3841 (2001).

[44]W. E, W. Ren, and E. Vanden-Eijnden, "Minimum action method for the study of rare events," Communications on Pure and Applied Mathematics **57**, 637–656 (2004).

[45]I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences **374**, 20150202 (2016).

[46]B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural computation **10**, 1299–1319 (1998).

[47]G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for markov model construction," The Journal of chemical physics **139** (2013).

[48]C. R. Schwantes and V. S. Pande, "Modeling molecular kinetics with tica and the kernel trick," Journal of chemical theory and computation **11**, 600–608 (2015).

[49]C. Wehmeyer and F. Noé, "Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics," The Journal of Chemical Physics **148**, 241703 (2018), https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/1.5011399/16653374/241703_1_online.pdf.

[50]R. R. Coifman and S. Lafon, "Diffusion maps," Applied and computational harmonic analysis **21**, 5–30 (2006).

[51]P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, "Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction," Proceedings of the National Academy of Sciences **103**, 9885–9890 (2006).

[52]M. Ceriotti, G. A. Tribello, and M. Parrinello, "Simplifying the representation of complex free-energy landscapes using sketch-map," Proceedings of the National Academy of Sciences **108**, 13023–13028 (2011).

[53]D. Vardar, A. Chishti, B. Frank, E. J. Luna, A. Noegel, S. W. Oh, M. Schleicher, and C. McKnight, "Villin-type headpiece domains show a wide range of f-actin-binding affinities," Cell motility and the cytoskeleton **52**, 9–21 (2002).

[54]Y. Tang, M. J. Grey, J. McKnight, A. G. Palmer III, and D. P. Raleigh, "Multistate folding of the villin headpiece domain," Journal of molecular biology **355**, 1066–1077 (2006).

[55]H. Lei, C. Wu, H. Liu, and Y. Duan, "Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations," Proceedings of the National Academy of Sciences of the United States of America **104**, 4925–30 (2007).

[56]K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How fast-folding proteins fold," Science **334**, 517–520 (2011).

[57]B. E. Husic and V. S. Pande, "Markov state models: From an art to a science," Journal of the American Chemical Society **140**, 2386–2396 (2018).

[58]E. Marinari and G. Parisi, "Simulated tempering: A new monte carlo scheme," Europhysics Letters **19**, 451 (1992).

[59]Y. Sugita and Y. Okamoto, "Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape," Chem. Phys. Lett. **329**, 261–270 (2000).

[60]M. Souaille and B. Roux, "Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations," Comp. Phys- Comm. **135**, 40–57 (2001).

[61]T. Huber, A. Torda, and W. van Gunsteren, "Local elevation: A method for improving the searching properties of molecular dynamics simulation," J. Comput. Aided Mol. Des. **8** (1994).

[62]A. Laio and M. Parrinello, "Escaping free-energy minima." Proc. Natl. Acad. Sci. U.S.A. **99**, 12562–6 (2002).

[63]A. Barducci, G. Bussi, and M. Parrinello, "Well-tempered metadynamics: A smoothly converging and tunable free-energy method." Phys. Rev. Lett. **100**, 020603 (2008).

[64]R. von Mises and H. Pollaczek-Geiringer, "Praktische verfahren der gleichungsauflösung ." Zamm-zeitschrift Fur Angewandte Mathematik Und Mechanik **9**, 152–164 (1929).

[65]M. Weber, "Implications of PCCA+ in Molecular Simulation," Computation **6** (2018).

[66]B. Keller, X. Daura, and W. F. van Gunsteren, "Comparing geometric and kinetic cluster algo-

rithms for molecular simulation data," J. Chem. Phys. **132**, 074110 (2010).

[67]O. Lemke and B. G. Keller, "Density-based cluster algorithms for the identification of core sets," The Journal of Chemical Physics **145**, 164104 (2016).

[68]O. Lemke and B. G. Keller, "Common nearest neighbor clustering—a benchmark," Algorithms **11** (2018), 10.3390/a11020019.

[69]B. Leimkuhler and C. Matthews, *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods.* (Springer, Interdisciplinary Applied Mathematics; Vol. 39, 2015).

[70]A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32* (Curran Associates, Inc., 2019) pp. 8024–8035.

[71]H. E. Robbins, "A stochastic approximation method," Annals of Mathematical Statistics **22**, 400–407 (1951).

[72]P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, "Openmm 7: Rapid development of high performance algorithms for molecular dynamics," PLOS Computational Biology **13**, 1–17 (2017).

[73]J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," J. Comp. Chem. **25**, 1157–1174 (2004).

[74]L.-P. Wang, T. J. Martinez, and V. S. Pande, "Building force fields: An automatic, systematic, and reproducible approach," J. Phys. Chem. Lett. **5**, 1885–1891 (2014).

[75]J. A. Izaguirre, C. R. Sweet, and V. S. Pande, "Multiscale dynamics of macromolecules using normal mode langevin," Pac. Symp. Biocomput. **15**, 240–251 (2010).

[76]T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems," J. Chem. Phys. **98**, 10089–10092 (1993).

[77]N. Floquet, S. Héry-Huynh, M. Dauchez, P. Derreumaux, A. M. Tamburro, and A. J. P. Alix, "Structural characterization of vgvapg, an elastin-derived peptide," Peptide Science **76**, 266–280 (2004).

[78]L. Donati and B. G. Keller, "Girsanov reweighting for metadynamics simulations," J. Chem. Phys. **149**, 072335 (2018).

[79]J. Giraldo-Barreto, S. Ortiz, E. Thiede, K. Palacio-Rodríguez, B. Carpenter, A. Barnett, and

P. Cossio, "A bayesian approach to extracting free-energy profiles from cryo-electron microscopy experiments," Scientific Reports **11** (2021), 10.1038/s41598-021-92621-1.

[80]J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb," J. Chem. Theory Comput. **11 (8)**, 3696–3713 (2015).

[81]T. K. Chiu, J. Kubelka, R. Herbst-Irmer, W. A. Eaton, J. Hofrichter, and D. R. Davies, "High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein," Proceedings of the National Academy of Sciences **102**, 7517–7522 (2005).

[82]Schrödinger, LLC, *The PyMOL Molecular Graphics System, Version 2.0*, Schrödinger, LLC (2015), accessed: 2024-10-08.

[83]R. Harada and A. Kitao, "The fast-folding mechanism of villin headpiece subdomain studied by multiscale distributed computing," Journal of Chemical Theory and Computation **8**, 290–299 (2012).

[84]E. Wang, P. Tao, J. Wang, and Y. Xiao, "A novel folding pathway of the villin headpiece subdomain hp35," Physical Chemistry Chemical Physics **21** (2019), 10.1039/C9CP01703H.

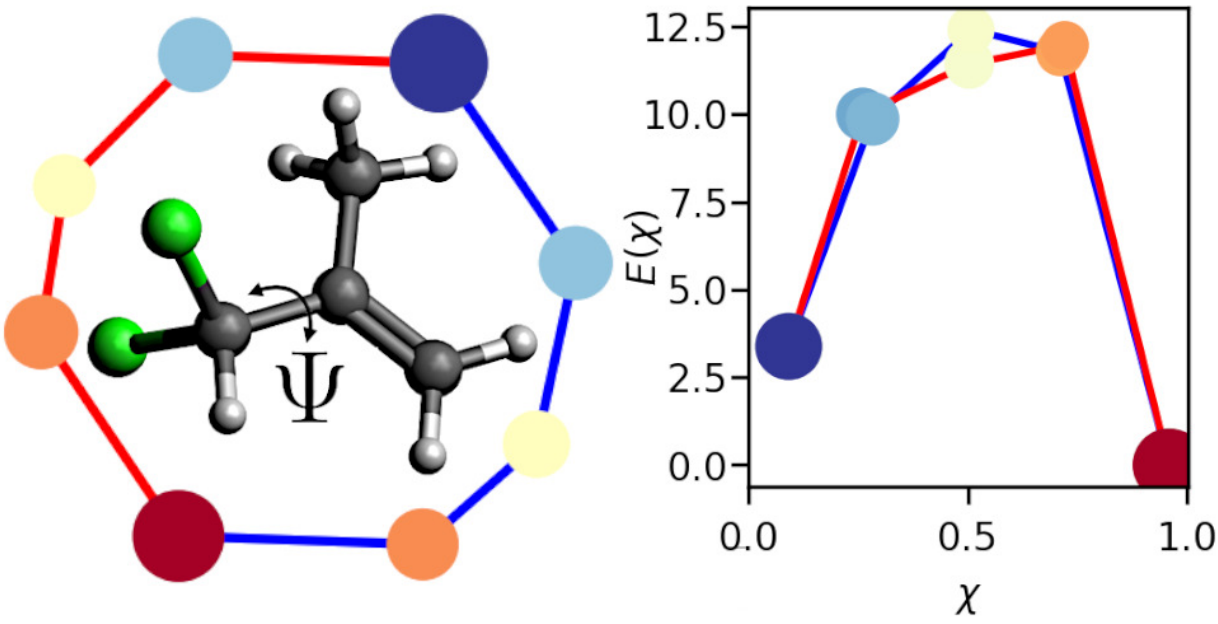


FIG. 7. For Table of Contents Only.