

UMAP as a Dimensionality Reduction Tool for Molecular Dynamics Simulations of Biomacromolecules: A Comparison Study

Published as part of *The Journal of Physical Chemistry virtual special issue "Ruth Nussinov Festschrift"*.

Francesco Trozzi, Xinlei Wang, and Peng Tao*

Cite This: *J. Phys. Chem. B* 2021, 125, 5022–5034

Read Online

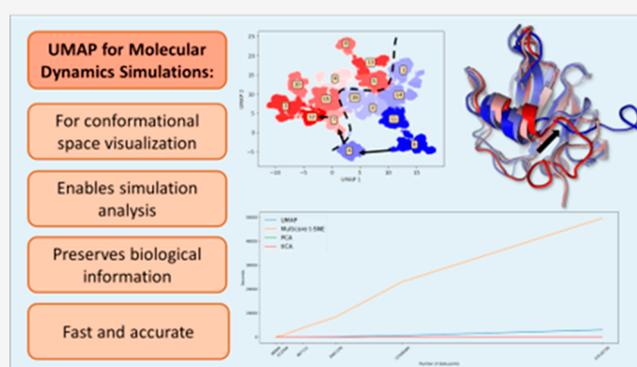
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: Proteins are the molecular machines of life. The multitude of possible conformations that proteins can adopt determines their free-energy landscapes. However, the inherently high dimensionality of a protein free-energy landscape poses a challenge to deciphering how proteins perform their functions. For this reason, dimensionality reduction is an active field of research for molecular biologists. The uniform manifold approximation and projection (UMAP) is a dimensionality reduction method based on a fuzzy topological analysis of data. In the present study, the performance of UMAP is compared with that of other popular dimensionality reduction methods such as t-distributed stochastic neighbor embedding (t-SNE), principal component analysis (PCA), and time-structure independent components analysis (tICA) in the context of analyzing molecular dynamics simulations of the circadian clock protein VIVID. A good dimensionality reduction method should accurately represent the data structure on the projected components. The comparison of the raw high-dimensional data with the projections obtained using different dimensionality reduction methods based on various metrics showed that UMAP has superior performance when compared with linear reduction methods (PCA and tICA) and has competitive performance and scalable computational cost.



1. INTRODUCTION

Proteins are molecular engines present in all life forms on Earth. Protein structures have been described hierarchically as protein primary, secondary, and tertiary structures. Main protein functional information is expected to be derived from this structural information.^{1–4} However, protein molecules are in constant dynamics, which also is a key factor in the regulation of the protein functions.⁵

Molecular dynamics (MD) simulations provide dynamical information on protein conformations to map the protein conformational space and thus rationalize protein function.^{6–8} The sampling of the protein conformations collected during the simulations compose the protein conformational space.

The high degrees of freedom of protein molecules present challenges also referred to as the curse of dimensionality. To face this challenge, various dimensionality reduction methods have been applied to MD simulations under the assumption that a few degrees of freedom through coordinate projections could account for the majority of the protein functions.^{9–20} The projections obtained can then be used as collective variables (CVs) to build a Markov state model (MSM). MSMs have been applied to identify protein functional states on the free-energy surface and to describe the transitions among them.^{21–27}

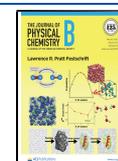
Dimensionality reduction methods can be broadly categorized into two groups: linear and nonlinear.^{28–30} Linear methods, such as principal component analysis (PCA) and time-structure independent component analysis (tICA), construct new CVs by performing linear combinations of the input variables. On the contrary, nonlinear methods, such as the t-distributed stochastic neighbor embedding (t-SNE) method and autoencoders, construct new CVs by mapping the input variables to a nonlinear function. Ultimately, because of the highly curved shape of protein free-energy landscapes, nonlinear dimensionality reduction methods should be more beneficial to process MD trajectories, as compared with linear methods.^{11,16,19}

All dimensionality reduction methods have their own advantages and limitations. Zhou et al.¹⁹ compared several algorithms widely used for the analysis of MD simulations and

Received: March 8, 2021

Revised: April 26, 2021

Published: May 11, 2021



demonstrated the overall superior performance of the t-SNE method. In their study, the t-SNE method was found to be able to correctly reproduce the kinetic barrier and structural similarity of different clusters. However, the relatively high computational cost of the t-SNE method forces the user to significantly reduce the sampling of protein trajectories to obtain results in a reasonable time frame. Furthermore, because of the intrinsic property of the Kullback–Leibler (KL) divergence as its loss function, t-SNE is not guaranteed to always preserve distances correctly among data points in the low-dimensional space when the distances among these data points are large in the high-dimensional space.

In 2018, McInnes et al.³¹ developed a new fuzzy topology-based dimensionality reduction method named uniform manifold approximation and projection (UMAP), which could serve as an alternative method to t-SNE. UMAP has been used to process data from single-cell experiments as a dimensionality reduction method with either equal or better quality than t-SNE.^{32–35}

In the present study, we aim to demonstrate the applicability and efficiency of UMAP in the computational studies of biomacromolecules. Using a well-studied protein as the model system, we performed a comparative study along with other popular dimensionality reduction tools including PCA, tICA, and t-SNE to investigate the applicability of UMAP in the context of analyzing and processing data obtained from MD simulations of biomacromolecules to gain insight into their structure–function relations.

In this study, VIVID (VVD), which is a well-characterized circadian clock protein as a member of the light oxygen voltage (LOV) domain family,³⁶ is used as the model protein. VVD is an allosteric protein and could be activated upon photo-excitation. It has two distinct functional states: dark and light states. The VVD dark state could be excited by blue light to form a covalent bond with its flavin cofactor and undergoes a global conformational change, mainly in its N-terminus region, leading to a cascade of circadian-clock-related signaling events.^{37–39}

2. MATERIALS AND METHODS

2.1. Dimensionality Reduction Methods. **2.1.1. Principal Component Analysis.** PCA reduces the dimensionality of the data by projecting each data point onto a few principal components as a lower-dimensional representation of the original data while preserving the data's variation.³⁹ The components in PCA are linear combinations of input variables and are orthogonal to each other. Given two variables, x and y , their sample covariance measures how these two variables deviate from their averages \bar{x} and \bar{y} in relation to each other based on n observations

$$\sigma(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (1)$$

In PCA, a $p \times p$ covariance matrix, C , is constructed for a given data set with p variables, in which each element C_{ij} is represented by the covariance between two variables, as expressed in eq 1. In this symmetric matrix, each element is a sample covariance between two variables, x_i and x_j , expressed as $C_{ij} = \sigma(x_i, x_j)$. The eigenvectors of C are the components of PCA. The eigenvalues of C measure the contribution of each component in the data set. The larger the magnitude of eigenvalue, the higher the contribution of its corresponding

component, that is, the eigenvector. In general, the eigenvectors with the largest eigenvalues are designated as principal components to form 2D or 3D space for data projection. The PCA was performed using scikit-learn implemented in Python.⁴⁰

2.1.2. Time-Structure Independent Component Analysis. The tICA method aims to identify the slowest degrees of freedom and therefore in preserving the kinetic information present in the MD trajectories by maximizing the autocorrelation function.^{41–43} Given a time series of molecular coordinates provided by the MD trajectories, $\mathbf{x}(t) = (x_1(t), \dots, x_n(t))$, tICA aims to reduce the dimensionality of the trajectories and to identify hidden key structural changes by decomposing the generalized eigenvalue problem $\bar{C}F = CFK$, where $K = \text{diag}(k_1, \dots, k_n)$ and $F = (f_1, \dots, f_n)$ are the eigenvalue and eigenvector matrices, respectively. C and \bar{C} are the covariance matrix and the time-lagged covariance matrix of the coordinate vector, respectively

$$C = \langle (x(t) - \langle x(t) \rangle)^t (x(t) - \langle x(t) \rangle) \rangle \quad (2)$$

$$\bar{C} = \langle (x(t) - \langle x(t) \rangle)^t (x(t + t_0) - \langle x(t) \rangle) \rangle \quad (3)$$

where $\langle \dots \rangle$ denotes the average. To obtain a symmetric time-lagged covariance matrix, $\frac{1}{2}(\bar{C} + \bar{C}^t)$ is calculated. The latter step assumes the time reversibility of the process, which is satisfied in MD simulations. The projected vectors of the MD are

$$a(t) = (a_1(t), \dots, a_n(t))^t = Fx(t)^t \quad (4)$$

The featurization and dimensionality reduction were performed using the MSMBuilder package.⁴⁴

2.1.3. t-Distributed Stochastic Neighbor Embedding Method. t-SNE is an unsupervised nonlinear dimensionality reduction method.⁴⁵ t-SNE builds its reduced representation first by constructing a probability distribution of distances between any two observations i and j in the high-dimensional manifold as

$$p_{ij} = \frac{e^{-\|x_i - x_j\|^2 / 2\sigma_i^2}}{\sum_{k \neq i} e^{-\|x_i - x_k\|^2 / 2\sigma_i^2}} \quad (5)$$

and a Student's t probability distribution in the lower dimensional space. Let y_i and y_j be the unknown lower dimension representations of observations i and j , respectively. The Student's t distribution in t-SNE is used to avoid overcrowding of data points in the lower dimensional space.

$$q_{ij} = \frac{e^{-\|y_i - y_j\|^{-1}}}{\sum_{k \neq i} e^{-\|y_i - y_k\|^{-1}}} \quad (6)$$

The aim of t-SNE is to maximize the similarity between these two density distributions over y_i values. The metric used to assess the dissimilarity between the high- and low-dimensional distributions is the KL divergence

$$\text{KL}(P_i \| Q_i) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (7)$$

The optimization of the low-dimensional representation is achieved by the minimization of the KL divergence. One disadvantage of using KL divergence is that its loss function mainly preserves only local distances, and there is no guarantee

regarding the preservation of large high-dimensional distances in a low-dimensional space. The t-SNE projections were performed using the scikit-learn package implementation.⁴⁰

2.1.4. Uniform Manifold Approximation and Projection. UMAP is a fuzzy topology-based dimensionality reduction method.³¹ Similar to t-SNE, UMAP constructs probability distributions in the high-dimensional manifold as

$$p_{ij} = e^{-d(x_i, x_j)/\sigma_i} \quad (8)$$

An important difference in the UMAP probability distributions is the local distance metric, which is unique for every pair of points. The distance probability in the low-dimensional space in UMAP is given by

$$q_{ij} = (1 + a(y_i - y_j)^{2b})^{-1} \quad (9)$$

Another main difference between UMAP and t-SNE is the loss function to be minimized. KL divergence is used as loss function in t-SNE. In UMAP, cross-entropy (CE) is the loss function and is defined as

$$\text{CE}(X, Y) = \sum_i \sum_j \left[p_{ij}(X) \log \left(\frac{p_{ij}(X)}{q_{ij}(Y)} \right) + (1 - p_{ij}(X)) \log \left(\frac{1 - p_{ij}(X)}{1 - q_{ij}(Y)} \right) \right] \quad (10)$$

The CE function provides the advantage of being able to preserve the correlation between distances in the high and low dimensions for both small and large distances. UMAP projections were performed using the Python implementation available at <https://github.com/lmcinnes/umap>.

2.1.5. UMAP Hyperparameter Selection. Two crucial hyperparameters for UMAP usage are the number of neighbors and the minimum distance. The first parameter balances the accuracy of the local structure versus global structure of the data by varying the number of points in the local neighborhood. Small values of this hyperparameter reflect high accuracy in the local data structure, whereas large values reflect high accuracy in representing the global data structure, at the cost of the local one. The minimum distance parameter dictates the minimum distance between data points. Small values allow data clustering, whereas high parameters favor scattered data and should better preserve the global data structure.³¹ The UMAP hyperparameters were selected based on a benchmarking performed using the Pearson correlation and the cluster similarity score as criteria. We found that with a number of neighbors of 1000 and a minimum distance of 1, UMAP delivered the best performance for our data set, as shown in Figure S1.

2.2. Molecular Dynamics Simulations. The crystal structures of VVD in its dark (PDB ID: 2PD7³⁶) and light (PDB ID: 3RH8³⁶) states were taken from the Protein Data Bank (PDB).⁴⁶ All of the structures were cut to start at residue 37 for consistency. All structures were modeled with the flavin mononucleotide (FMN). In the light state, FMN was modeled with the photoinduced covalent bond between the FMN and a proximal CYS and the protonated NS. The force-field parameters for the FMN in the dark and light states were obtained from a previous study.⁴⁷ In this study, a total of four systems were simulated: a VVD dark crystal structure with the

FMN modeled in the dark state (native dark state), a VVD light crystal structure with the FMN modeled in the light state (native light state), a VVD dark crystal structure with the FMN modeled in the light state (transient light state), and a VVD in the light state with the FMN modeled in the dark state (transient dark state).

The protonation state of the histidine has been confirmed using the ProteinPrepare tool at playmolecule.com.⁴⁸ The preparation of the structures and the heating step were performed using CHARMM c41b1.⁴⁹ In particular, hydrogen atoms were added to the structures. The structures were then solvated using TIP3P water molecules and neutralized by adding chloride atoms and sodium cations. After the addition of the solvent, the size of the simulation box was 64.70 Å³. The structures were minimized first using the steep descent method for 200 steps followed by the adopted basis Newton–Raphson minimization for 1000 steps. The NVT dynamics of 24 ps was carried out to increase the temperature of the system from 0 to 300 K. For each structure, three 10 ns NPT equilibration dynamics simulations starting with random initial velocities were carried out. The final coordinates and velocities were used to start a production simulation of 1.1 μs trajectory, in which the first 100 ns was considered as equilibration and excluded from the final analysis. A total of 12 μs of MD trajectories was generated. The simulations were carried out using OpenMM 7.3 on a graphics processing unit (GPU).⁵⁰ A Monte Carlo barostat was used in the NPT simulations to maintain constant pressure.⁵¹ The NVT simulations were performed using the Langevin integrator.^{50,51} For the integrator, a friction coefficient of 1 ps⁻¹ was implemented. For all simulations, the covalent bonds containing hydrogen atoms were constrained using the SHAKE method.⁵² A step size of 2 fs was used. Frames were saved every 100 ps for the simulations. Periodic boundary conditions were applied, and particle mesh Ewald method was used to calculate the long-range interactions.⁵³ The cutoff used for the long-range interactions was 12 Å.

2.3. Analyses. **2.3.1. Root-Mean-Square Deviation.** For a system represented in Cartesian coordinates, the root-mean-square deviation (RMSD) is calculated to measure the deviation from a reference structure by taking the square root of the averaged difference between the atomic coordinates vectors of a reference structure, r_i^0 , and the structure in the i th frame among total of N frames, r_i

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^N (r^0 - U_i r_i)^2}{N}} \quad (11)$$

U_i is the rotation matrix to superimpose the structure in the i th frame against the reference structure.

2.3.2. Pearson Correlation Analysis. The Pearson correlation is a measure of linear correlation between two variables.⁵⁴ In this study, Pearson correlation is applied to evaluate how well the distances in the high-dimensional manifold are preserved in the embedded manifold. The distance metric to evaluate distances between points in both the high- and low-dimensional manifolds is the Manhattan distance. This type of distance has been shown to be a better metric of distance in high-dimensional spaces.⁵⁵ The first step in building the Pearson correlation is building the covariance. The covariance is then divided by the square root of the product of the variance of each variable.

$$\rho(x, y) = \frac{\sigma(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}} \quad (12)$$

The Pearson correlation is a dimensionless variable with values in the range of $[-1, +1]$. The negative values represent anticorrelation, and positive values represent correlation between two variables.

2.3.3. Cluster Similarity Score. The similarity between clusters in the high and low dimensions is measured using a cluster similarity score. The cluster similarity score is computed by comparing the population of the clusters obtained in the reduced dimensional space with the ones of the clusters in the high-dimensional space. Two populations will be compared to check whether the same data point is present in both high- and low-dimensional clusters. Once the cluster with the highest similarity in the low dimension is identified, it will be excluded in the similarity search for the subsequent clusters. This guarantees the unique pairing between clusters in the high- and low-dimensional spaces. For each cluster, the number of points that are allocated in the same cluster in both the high- and low-dimensional spaces are summed and given as percentage values to the total number of data points in the trajectory.

$$\text{CS} = \frac{\sum_{c=1}^{\text{tot}} \xi_c}{\text{frames}} * 100 \quad (13)$$

2.3.4. Silhouette Coefficient. The silhouette coefficient (SC) is a metric to evaluate the clustering performance.⁵⁶ This coefficient is calculated by comparing the mean distance between a cluster and the points in the nearest cluster (x) and the mean distance of the points within a cluster (y).

$$\text{SC} = \frac{(y - x)}{\max(x, y)} \quad (14)$$

The SC is maximized when clusters are well separated from each other.

The calculations of the SCs were performed using the implementation available in the scikit-learn package.⁴⁰

2.3.5. Machine-Learning Classification: Random Forest. The random forest method is an ensemble learning method comprising multiple decision trees for classification.⁵⁷ In each step of developing decision tree model, the model uses the parameters $\Phi = (j, t)$ composed of the data features j and a threshold t to divide the data in two parts based on the threshold.

$$Q_{\text{left}}(\theta) = (x, y) | x_j \leq t, \quad Q_{\text{right}}(\theta) = (x, y) | x_j \geq t \quad (15)$$

with x being the training data and y being the training label. The Gini impurity criterion was used to assess the quality of the model. The Gini impurity score represents the likelihood of an incorrect classification of a new random variable of feature t according to the existing label distribution.

$$G = \sum_k p_k (1 - p_k) \quad (16)$$

By constructing multiple random decision trees, the random forest method minimizes the potential bias toward a certain set of features in each specific decision tree model.^{58,59} The random forest method implemented in the scikit-learn Python package was used in this study.⁴⁰

2.3.6. Markov State Model. The MSM is used to estimate the conditional transition probabilities among nonoverlapping

states.⁶⁰ The collection of the transition probabilities among n states is represented as the transition matrix, T , with its element calculated as $T_{ij} = \frac{c_{ij}}{\sum_k c_{ik}}$, where c_{ij} is the count of the number of times the trajectories transition from a state i to a state j within a certain time interval, Δt , called the lag time, τ .

In this study, the first two components of each dimensionality reduction method were used as CVs to construct the MSM. The MSMBuilder Python package was used to build the MSM.⁴⁴ The default hyperparameters provided by MSMBuilder were used for the analysis. The ergodic cutoff was turned on, and the maximum likelihood method was used to achieve the reversibility of the transition matrix. A lag time of 30 ns was chosen.

2.3.7. Transition Path Theory. To study the path of the conformational changes along the allosteric process, the transition path theory (TPT) was used.^{61–63} The central element of TPT is the committor probability, q_i^+ . The committor probability represents the probability of the state i belonging to the macrostate A to transition to the macrostate B instead of staying in the macrostate A.⁶³ Per definition, q_i^+ for state i belonging to A or B is 0 or 1, respectively.⁶⁴ The committor probability for the intermediate states can be calculated as

$$-q_i^+ + \sum_{i \in I} T_{ik} q_k^+ = -\sum_{i \in B} T_{ik} \quad (17)$$

with the committor probability increasing along the path. The transition probability matrix built from the MSM contains all transitions among different macrostates including the ones that return to the initial state. The effective flux, f_{ij} , contains the probability flux that contributes only to the transition A to B

$$f_{ij} = \pi_i q_i^- T_{ij} q_j^+ \quad (18)$$

The net flux, which does not account for detours, is computed by

$$f_{ij}^+ = \max\{0, f_{ij} - f_{ji}\} \quad (19)$$

The MSMBuilder implementation of TPT was used in this study.⁴⁴

The implementation of the cluster similarity score, Pearson correlation analysis, clustering RMSD, and grid-search for random forest is available at https://github.com/smutaogroup/UMAP_analyses.

3. RESULTS

3.1. Comparison Between UMAP and Other Dimensionality Reductions. **3.1.1. Preservation of the Data Structure.** Protein conformations and dynamics are the key factors underlying protein functions. Our investigation started by comparing the preservation of the structure of the data in the high-dimensional manifold, expressed in Cartesian coordinates, into a lower dimensional space. The correct representation of the high-dimensional data structure into a lower dimension is crucial for the interpretation of biological information such as reproduction of free-energy barriers, evaluating transitions between conformations, and so on.

To evaluate the preservation of the data structure in low-dimensional space, the Pearson correlation analysis was carried out. Pearson correlation is a measure of linear correlation between two independent variables, X and Y . In this analysis, X and Y are the distances in the high- and low-dimensional

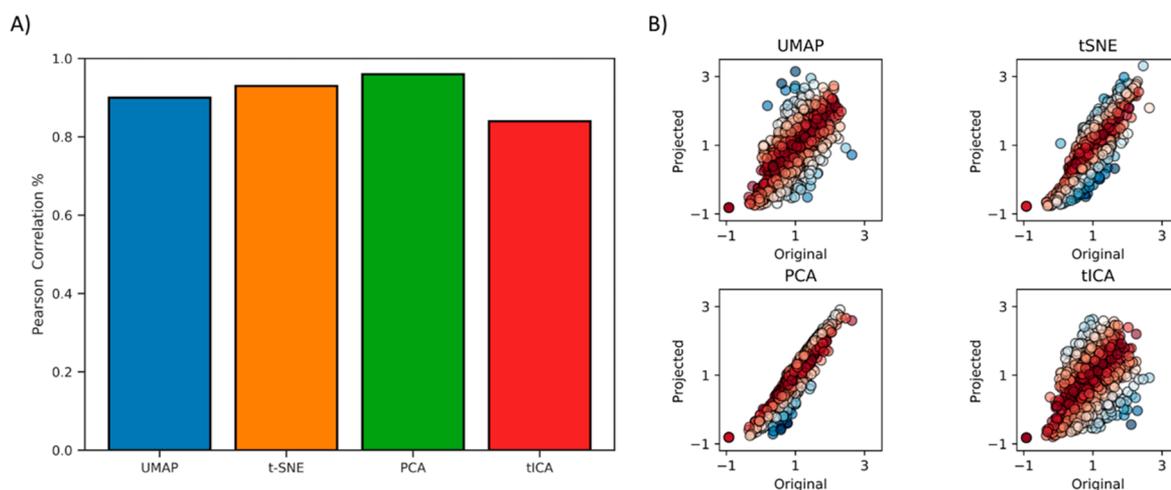


Figure 1. Pearson correlation analysis of UMAP, t-SNE, PCA, and tICA calculated based on the 2D reduced representations. (A) Pearson correlation values between projected and high-dimensional trajectories. (B) Scatterplots where the X axis represents the distance in the high-dimensional space and the Y axis represents the distance in the low-dimensional space. The coloring represents the agreement between the original and projected distances. Red and blue represent agreement and disagreement, respectively.

spaces, respectively. The distances in the projected low-dimensional space should reflect the original distances in the high-dimensional space. Frames saved every 1 ns were extracted from the MD trajectories and reduced to a 2D representation using UMAP, t-SNE, PCA, and tICA.

The Pearson correlation scores for UMAP, t-SNE, PCA, and tICA are 0.87, 0.90, 0.89, and 0.84, respectively (Figure 1). Thus the distances between points in the projections obtained from all methods are highly correlated with the distances in the high-dimensional Cartesian space (Figure 1A). Figure 1B shows the distance values for the same pair of data points in the low- and high-dimensional spaces. We observe that whereas PCA has a very narrow distribution, indicating high correlation overall, the deep-blue points for small distances suggest that this method does not provide an accurate representation of the local data structure.

3.1.2. Local Structure Assessment via Microclustering Analysis. An accurate representation of the local data structure through clustering analysis is crucial for further analyses regarding protein kinetics. The large deviation among structures within the same cluster could result in inaccurate free-energy barriers and interfere with consequent analysis such as Markov state modeling. To evaluate the quality of clustering analyses using various dimensionality reduction methods, we investigate the preservation of the local data structure in the following clustering analyses using these methods.

The 12 μ s of VVD trajectories were clustered into 1000 microstates using the *k*-means clustering method and the CVs of the different embedding as input variables. The averaged RMSD of all structures within each microstate was calculated to measure the structural similarity within each microstate. To retain structural and dynamical information, each microstate should contain a similar degree of similarity corresponding to the cluster in the high-dimensional manifold. Figure 2 shows the comparison of the different 2D representation in terms of the averaged RMSD with the nonreduced Cartesian representation. UMAP outperforms other dimensionality reduction methods in terms of the similarity within each microstate, achieving a high degree of similarity to the

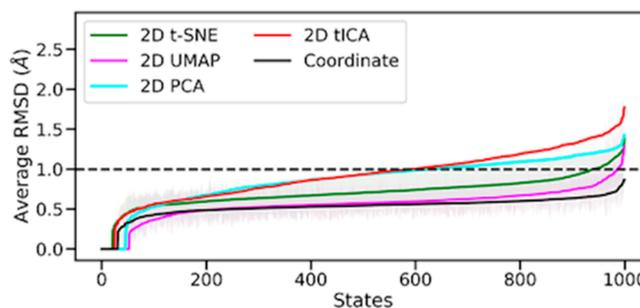


Figure 2. Averaged RMSD of 1000 microstates for various 2D representations and Cartesian coordinates. Microstates were sorted based on the average RMSD values.

nonreduced Cartesian representation. The second best method in this aspect is t-SNE, followed by PCA and tICA. Both UMAP and t-SNE methods consistently have RMSD values < 1 Å, which is desired for an ideal dimensionality reduction method and has been proposed as a threshold of structural dissimilarity within a macrostate needed to avoid the presence of energy barriers within the structural cluster.^{64,65}

3.1.3. Division of the Conformational Space into Macrostates. The protein conformational landscape is characterized by a series of low free-energy basins comprising low-energy conformations. To correctly cluster different structures into metastable states, the *k*-means clustering method was used to build clusters with the mean RMSD within the cluster smaller than 1 Å. As previously mentioned, structures with an RMSD smaller than 1 Å are expected to belong to the same free-energy basin and therefore the same metastable state. This procedure ensures that no artificial free-energy barriers are hidden within each cluster. In this study, a total of 16 clusters were found to be the lowest number of clusters that had a mean intracluster RMSD within 1 Å and a maximal intercluster difference (Figure 3). These 16 clusters are hereby referred to as macrostates.

To evaluate how well the clustering in the low-dimensional space represents the original data in the high-dimensional space, we calculated a similarity score by comparing the

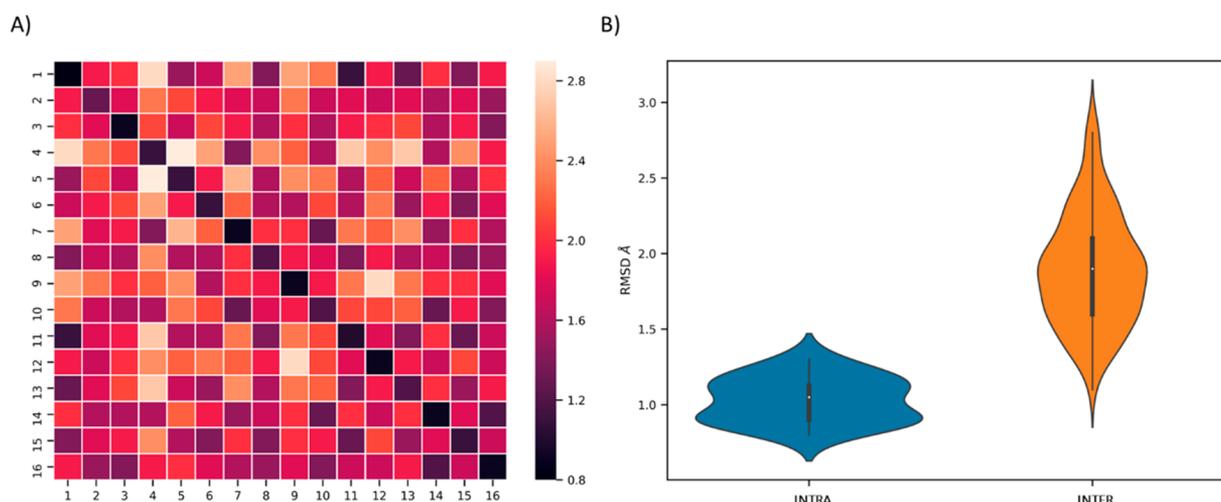


Figure 3. Selection of the number of macrostates based on the cluster RMSD. (A) Heatmap of the RMSD within each state. (B) Violin plot of RMSD values within states (blue) and interstates (orange).

population of each cluster in the reduced embedding with the corresponding cluster in the high-dimensional manifold. The first step was the assignment of the clusters in the low-dimensional embedding to the corresponding high-dimensional clusters. After the assignment was made, for each pair of clusters, one in the high-dimensional space and one in the low-dimensional space, the number of points shared by both clusters was counted. The total number of these shared points was summed and converted to the percentage to the total number of data points. Figure 4 shows that UMAP improves

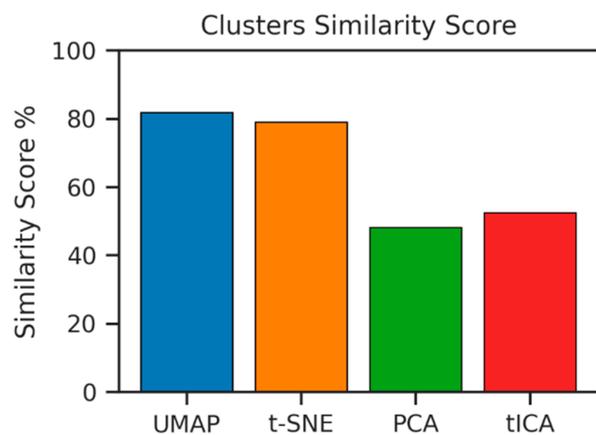


Figure 4. Similarity, expressed in percentage, between cluster populations in low-dimensional representations and high-dimensional Cartesian space.

the t-SNE performance and outperforms PCA and tICA in the similarity score. This demonstrates that UMAP could appropriately assign protein structures into their corresponding functional metastable states (macrostates) for further analyses.

Another important criterion for the dimensionality reduction method is cluster separation. Specifically, an adequate low-dimensional projection should retain the separation that these macrostates have in the high-dimensional space. The projection of the macrocluster is plotted in Figure S2.

To quantitatively assess the quality of clustering projections, we employed the SC. As described in the Materials and Methods section, the SC is a clustering quality assessment

criterion that evaluates the separation distances between clusters. SCs range from -1 to $+1$, where positive values indicate better separation between clusters and negative values indicating their overlap. The SC for each method and in the Cartesian space are listed in Table S1. In Figure 5, we plot the

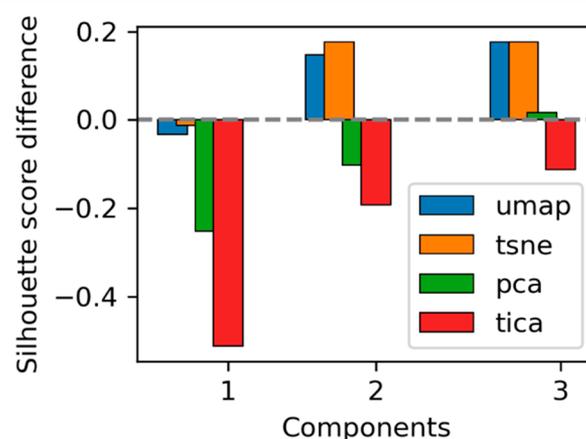


Figure 5. Comparison of the silhouette coefficient for UMAP, t-SNE, PCA, and tICA projections versus Cartesian space results. Bar heights represent the deviation in the coefficient from the Cartesian case. Positive values represent the higher separation of the clusters in the projected space. Negative values represent the overcrowding of the clusters in projected spaces.

difference between the SC of each dimensionality reduction method and the SC of the original Cartesian space. In this comparison, a negative value indicates that a projection using a certain dimensionality reduction method increases the overlap among the macrostates compared with the results in Cartesian space. A positive value indicates a higher separation, possibly an overseparation, among macrostates in the projection compared with the results in Cartesian space. When only one dimension is used to project the conformational space, UMAP and t-SNE offer a more faithful projection of the macroclusters, demonstrated by the small values of the SC difference. When two dimensions are considered, a similar amount of divergence in different directions is observed for the projection for all methods. Interestingly, the nonlinear

methods UMAP and t-SNE tend to overseparate the clusters, whereas the linear PCA and tICA tend to increase their overlap. When three dimensions are used, both UMAP and t-SNE still lead to higher separation than the Cartesian results. PCA method results in a slightly higher separation. tICA still results in a higher overlapping than the Cartesian results.

3.1.4. Machine-Learning Classification. In our previous studies, it was demonstrated that machine-learning-based classification for the macrostates is an effective approach to delineate protein allosteric mechanism related to individual residues.^{27,66–68} To build effective machine-learning classification models, it is desired to have dimensionality reduction methods that could enhance the quality of classification models. We used the random forest method as a machine-learning classification model. The best combination of hyperparameters for the input data from each dimensionality reduction method was identified using grid searches (Table S2). Using UMAP and t-SNE, the machine-learning classification models for the macrostates generated using the two most dominant dimensions reached 95% accuracy (Figure 6).

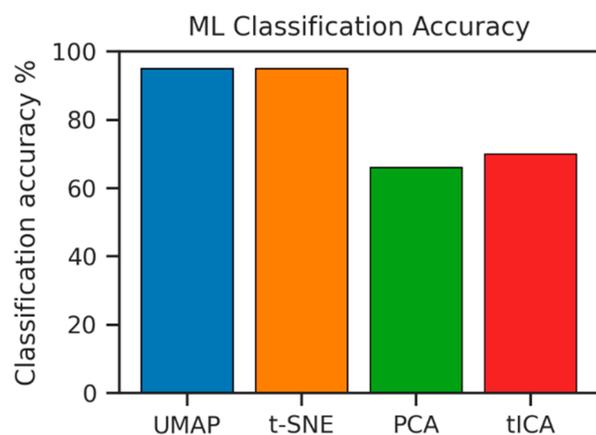


Figure 6. Machine-learning prediction accuracy of the different macrostates based on the 2D input of the low-dimensional representation using random forest.

As a comparison, the similar prediction accuracy was much smaller for PCA and tICA, 66 and 70%, respectively. This is in the agreement with their performance of the projections of the 16 macrostates (Figure 5), in which macrostates generated using PCA and tICA methods had more overlaps than those from UMAP and t-SNE methods.

3.1.5. Kinetics. Proteins are in constant motion, whether when carrying out their biochemical functions or not. One common approach to probe protein dynamics is building MSMs to estimate probabilities for protein transition among different macrostates. To build an effective MSM, it is important for a dimensionality reduction method to retain information about how proteins transition among these macrostates. To evaluate the retaining of such information, we analyzed the relaxation time scales in MSM, also referred to as implied time scales, using different dimensionality reduction methods with comparison to the results using Cartesian coordinates.¹⁹ The relaxation time scale can be interpreted as the time needed for a system to change its state.⁶⁹ Because protein functions are presumed to be strongly correlated with protein slow motions, a well-behaving dimensionality reduction method is expected to preserve slow degrees of freedom of

protein simulations for the accurate description of protein kinetics related to their functions.

Overall, all methods behaved well to produce the implied time-scale ranges close to the Cartesian coordinate results, except for PCA (Figure 7). UMAP produces implied time

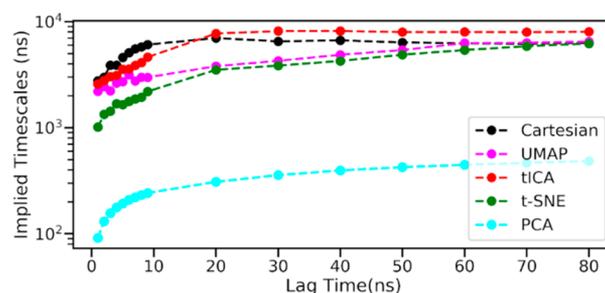


Figure 7. Comparison of implied time scales of different methods.

scales that were the closest to the Cartesian coordinates results, especially with the lag time longer than 60 ns. The t-SNE method was the second best, and its implied time scales also converged to the Cartesian coordinates results with the lag time longer than 70 ns. Although tICA also produced a result close to the one of Cartesian coordinates, it overestimated the implied time scales of the system, and its results did not converge to the Cartesian coordinate results. This comparison demonstrates that UMAP could retain protein dynamics information to describe the kinetics of the target system.

The transition matrix produced in each MSM provides transition probability between each pair of macrostates as detailed kinetics information on the system. To further evaluate the performance of each dimensionality reduction method to retain kinetics information on the system, we implemented a transition matrix error analysis by comparing the transition matrices from different dimensionality reduction methods with the Cartesian coordinate results as the reference. For this comparison, we used a total of 16 macrostates identified in Section 3.1.3. The absolute value for the difference of each transition matrix element is listed and illustrated as heatmaps (Figure 8). To quantify the deviation from the high-dimensional transition matrix, the deviation sum was calculated for each method as 8.26, 9.33, 14.06, and 9.39 for UMAP, t-SNE, PCA, and tICA, respectively. UMAP outperforms the other dimensionality reduction methods. This result indicates that UMAP well captures the system kinetics, related to the free-energy surface. Moreover, tICA performs similarly to t-SNE and UMAP.

3.2. How Many Dimensions Are Needed? Ideally, a good dimensionality reduction method should retain both structural and kinetic information on a trajectory using a minimal number of dimensions possible. Therefore, the amount of information that could be retained by the most dominant components generated in each method was evaluated. One of the goals of this analysis is to determine how many dimensions are optimal for retaining the structural and kinetic properties of the model system using each method. Both the Pearson correlation and the transition matrix error for each method were used to evaluate the structural and kinetic information retention. The Pearson correlation is used to identify the structural similarities between the original data in the high-dimensional space and the low-dimensional embedding using various numbers of dimensions.

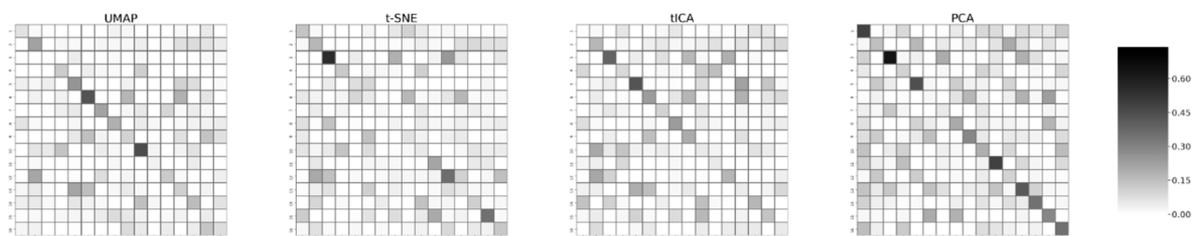


Figure 8. Heatmap representation of the divergence of the different transition matrices obtained using different dimensionality reduction methods from the high-dimensional transition matrix.

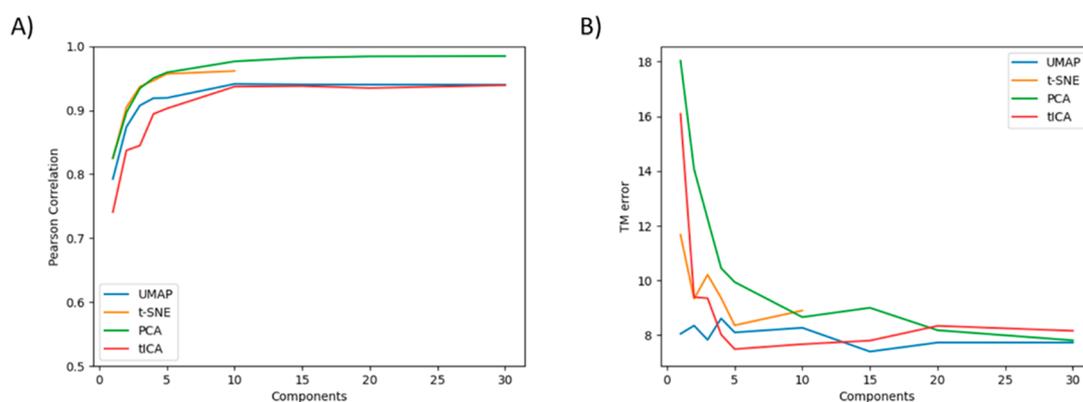


Figure 9. Performance of different methods regarding the number of components used in the projection. (A) Pearson correlation between the high-dimensional representation and the reduced representation of the data at varying numbers of projected dimensions. (B) Transition matrix error between the high-dimensional representation and the reduced representation of the data at varying numbers of projected dimensions.

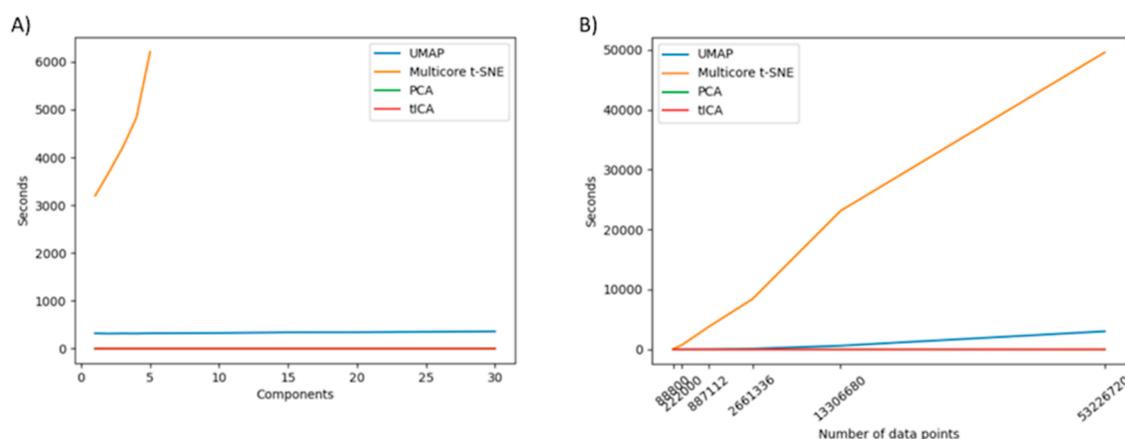


Figure 10. Benchmark using different dimensionality reduction methods. (A) Time in seconds required for dimensionality reduction at various numbers of projected dimensions. (B) Time in seconds required for 2D projections using different number of frames as data points.

Using the Pearson correlation analysis, the correlation between the original data and the new data projected onto different numbers of components starting with one was calculated and plotted for the selected methods (Figure 9A). For UMAP, t-SNE, and PCA, the Pearson correlation is >0.80 when the original data are projected onto the most dominant dimension and >0.90 with the top two dominant dimensions being used. Both PCA and t-SNE display the best Pearson correlation when using one or two components. For tICA, the Pearson correlation is much lower compared with the other three methods, making it less ideal for the analysis of the protein conformational space.

To evaluate the kinetic information retention using various numbers of dimensions, the transition matrix error analysis was

carried out for each method when using different numbers of components for data projection. The total of absolute transition matrix error for each case is plotted for comparison (Figure 9B). When using only one or two components for data projection, PCA displays the highest errors, probably due to its linearity nature. tICA also displays significant errors when using only one or two components for data projection. Surprisingly, t-SNE also displays significant errors when using only one or two components for data projection. UMAP consistently displays the lowest transition matrix error when using one, two, or three components for data projection. Overall, the above analyses ensure the applicability of UMAP with a minimal loss of structural and kinetic information, two critical aspects in the study of protein biology.

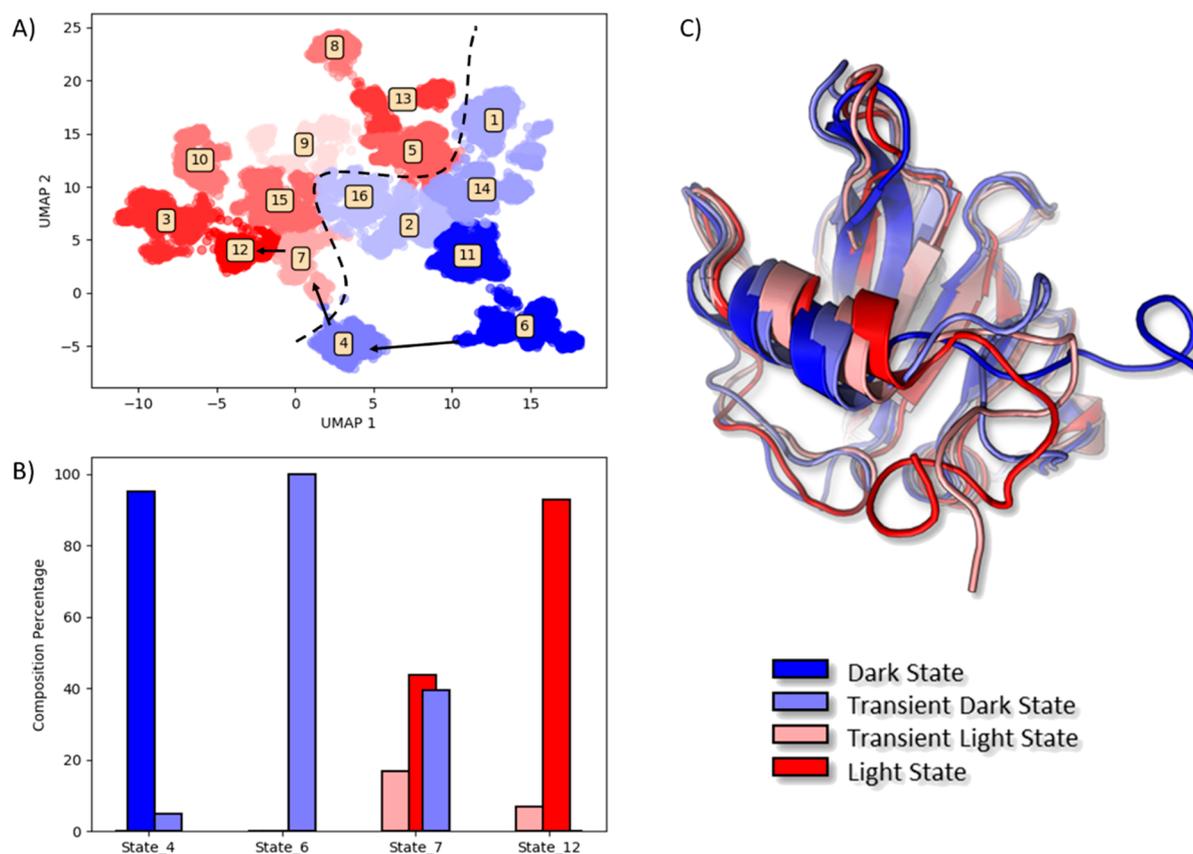


Figure 11. Demonstration of the protein function analysis using the UMAP method. (A) UMAP 2D projection. Reduced space was clustered in 16 macrostates according to the previously presented criteria. The clusters were color-coded based on their population. Dark states are blue, and light states are red. The dashed line represents the division between dark and light areas. Arrows represent the pathway for allosteric conversion from fully dark to fully light states. (B) Population state analysis of the macrostates involved in the VVD allosteric process. (C) Visualization of the four representative states involved in the allosteric process.

3.3. Benchmark. From a practical point of view, another crucial factor in choosing a dimensionality reduction method is its computational cost. We carried out some benchmark calculations to compare the computational cost of UMAP, t-SNE, PCA, and tICA for various numbers of components for projection (Figure 10A). For PCA, tICA, and UMAP, the computational cost remains close to constant regardless of the number of components used for projection. However, the computational cost increases exponentially with the number of components used for projecting when using t-SNE. In our benchmark calculation, we could perform the calculation only up to five dimensions with t-SNE method. The computational cost with only one or two components for data projection is also significantly higher than that of all of the other methods by factors of 5 to 8. This greatly limits the applicability of t-SNE for protein dynamics analyses. Although the computational cost for PCA and tICA is close to negligible as linear methods, the computational cost for UMAP as a nonlinear dimensionality reduction method is not much higher, making UMAP a very feasible option for dimensionality reduction analysis. All benchmark calculations were performed with NVIDIA GPUs, which were configured with dual Intel Xeon E5-2695v4 2.1 GHz 18-core “Broadwell” processors and 256 GB of DDR4-2400 memory.

The benchmark calculations were also carried for two components projection with different numbers of data points used, as 2D is the most widely used for protein dynamics analyses. Regarding the speed for varying the number of points,

the number of data points was varied by progressively increasing the stride between consecutive frames in the trajectories. Figure 10B shows that UMAP achieves similar speed performance as PCA and tICA. Whereas its computational time increases with the increase in the sample number, its speed remains close to the linear methods. The time required by t-SNE significantly exceeds the time required by tICA, PCA, and UMAP.

3.4. Leading to Insight into Protein Function Mechanism. The purpose of using dimensionality reduction methods for protein dynamics analyses is providing mechanistic insights into protein structures and functions. Because there is no universal standard to evaluate such performance, some demonstrative analyses were carried out for the model system used in this study. VVD is a well characterized circadian clock protein that has been shown to undergo conformational changes depending on the light condition. Following the rationale of the sections previously discussed, the conformational space of this protein sampled in the MD simulations has been clustered in 16 macrostates projected onto the 2D surface using the top two components from UMAP analysis (Figure 11A). In this plot, blue is used to indicate the dark state, lighter blue is used to indicate the transition dark state, red is used to indicate the light state, and lighter red is used to indicate the transition light state. The light (red) and dark (blue) states in the UMAP projections are well separated (Figure 11A), which is ideal for studying proteins with distinct functional states.

With a well-separated representation of functional states in the reduced dimensions, the TPT could be used to provide detailed kinetics of transitions among different macrostates. Using TPT, it was identified that the major transition pathway from the dark state (State 6) to the light state (State 12) gradually transitions via the transition pathway through State 4 belonging to the transient dark state and State 7 belonging to the transient light state (Figure 11B).

The α -helix movement and undocking of the N-terminus as key changes are illustrated in the representative structures (Figure 11C). These movements were recognized to be crucial steps in VVD allostery in the comprehensive mechanistic study done on this system by Zhou et al.²⁷ This analysis not only demonstrates the ability of UMAP to capture the fundamental biological properties of the system but also showcases that UMAP can be used as a visualization tool for the protein conformational space.

4. DISCUSSION

MD simulations have been used as indispensable approaches for the studies of protein functions within the dynamics framework. Although the time scales affordable for the MD simulations have been increasing significantly in recent years, they are still far from being comparable to the actual time scales for the protein biological functions. Even with this limitation, the curse of the dimensionality still prevents direct analyses of many properties of protein dynamics. Therefore, dimensionality reduction methods have been serving as essential tools to process protein MD simulations to gain insight into protein dynamical properties including both conformational space and kinetics information.

RMSD is a simple and effective measurement of conformational deviation between two structures. Using this quantity, the ability of each dimensionality reduction method to represent the conformational space sampled from the simulation could be accurately evaluated. The reason that the nonlinear dimensionality reduction methods, including UMAP and t-SNE, could produce much better clustering results than the linear methods, including PCA and tICA, is probably because most protein conformational changes have intrinsic nonlinearity, such as bond bending, dihedral angle rotations, and global motion of protein structures. UMAP produces better results than t-SNE when compared with the Cartesian coordinate results as the benchmark (Figure 2), suggesting that this method is approaching methodological limit of dimensionality reduction methods in general. This is also supported by the highest similarity score of UMAP among all four methods (Figure 4). It should be noted that the comparison presented here is by no means complete or exhaustive. Therefore, it should not be concluded that UMAP is the best option to represent the conformational space sampling in all cases.

In addition to representing conformational changes well, it is more important and challenging to retain the kinetics information on protein dynamics when projecting the simulation trajectories onto reduced dimensional space. Because this is an active research area, a universal standard to evaluate kinetics information is yet to be determined. Therefore, the convergence test of implied time scales in MSM is used as the benchmark for kinetics information retention. Although it is not surprising that tICA demonstrates a closer trend to the results of Cartesian coordinates, as tICA was developed to capture the slow and global motion of proteins, it

is somewhat suspicious that the tICA results do not converge to the Cartesian coordinate results with longer lag times. Overall, UMAP is demonstrated as a more balanced option than both the tICA and t-SNE methods. With smaller lag times, UMAP results are also close to the Cartesian coordinate results, similar to tICA (Figure 7). With longer lag times, UMAP results display better convergence to the Cartesian coordinates results than t-SNE (Figure 7). The superiority of UMAP for kinetics information retention is also supported by the high accuracy of the machine-learning-based prediction model using UMAP projections (Figure 6). The fact that UMAP results are better than other methods in both conformation and kinetics representations is promising. Although there is no direct evidence, the satisfactory performance of UMAP to preserve protein kinetics information could be partially due to the good representation of protein conformational space.

Because of the limitation of the human perception of dimensionality, the dimensions of graphical representation of protein dynamics analysis have been limited to two. This is validated by the evaluation of the Pearson correlation between the projected data in low-dimensional spaces and the original data and transition matrix error analyses (Figure 9). Both t-SNE and PCA methods display high Pearson correlation with one or two dimensions for data projection. As a linear method, PCA is not always suitable for protein dynamics simulations, leaving t-SNE as a better choice in this regard. UMAP produces comparable Pearson correlations with one or two dimensions for data projection. Interestingly, the UMAP method produces much lower transition matrix error than all of the other three methods, making it a well-balanced method for low-dimensional space projection. Considering the almost constant computational cost of the UMAP method, which is similar to those of PCA and tICA and much lower than the exponentially increasing computational cost of t-SNE (Figure 10), UMAP clearly serves as a viable option for dimensionality reduction analyses of complexed biomolecular systems, including proteins.

5. CONCLUSIONS

In this study, the suitability and performance of the UMAP as a new fuzzy topology-based dimensionality reduction method for the simulation of macromolecules was systematically evaluated and compared with other widely used dimensionality reduction methods, including PCA, tICA, and t-SNE. Using the Cartesian coordinates representation as the benchmark, it was demonstrated that UMAP could well retain the protein conformational information after the projection of original data. More importantly, the UMAP could also retain the protein kinetics information, which is critical to gain insight into protein functions within the dynamics framework. The balanced performance of UMAP to preserve protein kinetics is achieved through building an MSM based on the UMAP projection with well-preserved conformational space information. As a nonlinear dimensionality reduction method, UMAP displays similar overall performance and is more balanced between conformational and kinetics information retention than t-SNE. In addition, the computational cost of UMAP remains close to constant regardless of the number of dimensions being used for data projection. As a comparison, t-SNE requires exponentially increasing computational cost regarding the number of dimensions for the target data projection. Overall, the UMAP method is a well-behaving and

balanced dimensionality reduction method for in-depth biomacromolecule simulation analyses to gain insight into both structure–function and dynamics–function relations.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jpbc.1c02081>.

Figure S1. Hyperparameter evaluation for UMAP performance. Figure S2. Projections of the macroclusters found in the Cartesian space into low-dimensional space obtained using UMAP, t-SNE, PCA, and tICA. Table S1. Best hyperparameters for the ML classification model with different dimensionality reduction projections as input data. Table S2. Silhouette coefficients calculated on the projections performed using UMAP, t-SNE, PCA, and tICA (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Peng Tao – Department of Chemistry, Center for Research Computing, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas 75275, United States; orcid.org/0000-0002-2488-0239; Email: ptao@smu.edu

Authors

Francesco Trozzi – Department of Chemistry, Center for Research Computing, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas 75275, United States; orcid.org/0000-0001-5538-9189

Xinlei Wang – Department of Statistical Science, Southern Methodist University, Dallas, Texas 75275, United States

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.jpbc.1c02081>

Funding

The research reported in this paper was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award no. R15GM122013.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Computational time was generously provided by Southern Methodist University's Center for Research Computing.

■ ABBREVIATIONS

UMAP, Uniform manifold approximation and projection; t-SNE, t-distributed stochastic neighbor embedding; PCA, principal component analysis; tICA, time-structure independent component analysis; MD, molecular dynamics; CV, collective variable; MSM, Markov state model; KL, Kullback–Leibler; VVD, VIVID; CE, cross-entropy; RMSD, root mean square deviation; PC, Pearson correlation; SC, silhouette coefficient; ML, machine learning; TPT, transition path theory

■ REFERENCES

(1) Joshi, T.; Xu, D. Quantitative Assessment of Relationship between Sequence Similarity and Function Similarity. *BMC Genomics* **2007**, *8* (1), 222.

(2) Fowler, D. M.; Araya, C. L.; Fleishman, S. J.; Kellogg, E. H.; Stephany, J. J.; Baker, D.; Fields, S. High-Resolution Mapping of Protein Sequence–Function Relationships. *Nat. Methods* **2010**, *7* (9), 741–746.

(3) Hegyi, H.; Gerstein, M. The Relationship between Protein Structure and Function: A Comprehensive Survey with Application to the Yeast Genome. *J. Mol. Biol.* **1999**, *288* (1), 147–164.

(4) Orengo, C. A.; Todd, A. E.; Thornton, J. M. From Protein Structure to Function. *Curr. Opin. Struct. Biol.* **1999**, *9* (3), 374–382.

(5) Hensen, U.; Meyer, T.; Haas, J.; Rex, R.; Vriend, G.; Grubmüller, H. Exploring Protein Dynamics Space: The Dynasome as the Missing Link between Protein Structure and Function. *PLoS One* **2012**, *7* (5), No. e33931.

(6) Karplus, M.; Kuriyan, J. Molecular Dynamics and Protein Function. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (19), 6679–6685.

(7) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-Timescale Molecular Dynamics Simulations of Protein Structure and Function. *Curr. Opin. Struct. Biol.* **2009**, *19* (2), 120–127.

(8) Hansson, T.; Oostenbrink, C.; Van Gunsteren, W. F. Molecular Dynamics Simulations. *Curr. Opin. Struct. Biol.* **2002**, *12* (2), 190–196.

(9) García, A. E. Large-Amplitude Nonlinear Motions in Proteins. *Phys. Rev. Lett.* **1992**, *68* (17), 2696–2699.

(10) Stein, S. A. M.; Loccisano, A. E.; Firestone, S. M.; Evanseck, J. D. Chapter 13 Principal Components Analysis: A Review of Its Application on Molecular Dynamics Data. *Annu. Rep. Comput. Chem.* **2006**, *2*, 233–261.

(11) Tian, H.; Tao, P. Ivis Dimensionality Reduction Framework for Biomacromolecular Simulations. *J. Chem. Inf. Model.* **2020**, *60* (10), 4569–4581.

(12) Song, Z.; Zhou, H.; Tian, H.; Wang, X.; Tao, P. Unraveling the Energetic Significance of Chemical Events in Enzyme Catalysis via Machine-Learning Based Regression Approach. *Commun. Chem.* **2020**, *3* (1), 1–10.

(13) Das, P.; Moll, M.; Stamati, H.; Kavradi, L. E.; Clementi, C. Low-Dimensional, Free-Energy Landscapes of Protein-Folding Reactions by Nonlinear Dimensionality Reduction. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (26), 9885–9890.

(14) Brown, W. M.; Martin, S.; Pollock, S. N.; Coutsias, E. A.; Watson, J.-P. Algorithmic Dimensionality Reduction for Molecular Structure Analysis. *J. Chem. Phys.* **2008**, *129* (6), 064118.

(15) Stamati, H.; Clementi, C.; Kavradi, L. E. Application of Nonlinear Dimensionality Reduction to Characterize the Conformational Landscape of Small Peptides. *Proteins: Struct., Funct., Genet.* **2010**, *78* (2), 223–235.

(16) Ferguson, A. L.; Panagiotopoulos, A. Z.; Kevrekidis, I. G.; Debenedetti, P. G. Nonlinear Dimensionality Reduction in Molecular Simulation: The Diffusion Map Approach. *Chem. Phys. Lett.* **2011**, *509* (1–3), 1–11.

(17) Duan, M.; Fan, J.; Li, M.; Han, L.; Huo, S. Evaluation of Dimensionality-Reduction Methods from Peptide Folding–Unfolding Simulations. *J. Chem. Theory Comput.* **2013**, *9* (5), 2490–2497.

(18) Doerr, S.; Ariz-Extreme, I.; Harvey, M. J.; De Fabritiis, G. Dimensionality Reduction Methods for Molecular Simulations. *arXiv*, November 2, **2017**, arXiv:1710.10629, ver. 2. <https://arxiv.org/abs/1710.10629> (accessed on March 8, 2021).

(19) Zhou, H.; Wang, F.; Tao, P. T-Distributed Stochastic Neighbor Embedding Method with the Least Information Loss for Macromolecular Simulations. *J. Chem. Theory Comput.* **2018**, *14* (11), 5499–5510.

(20) Tribello, G. A.; Gasparotto, P. Using Dimensionality Reduction to Analyze Protein Trajectories. *Front. Mol. Biosci.* **2019**, *6* (JUN), 46.

(21) Prinz, J. H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134* (17), 174105.

(22) Bowman, G. R.; Pande, V. S.; Noé, F. An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation. *Adv. Exp. Med. Biol.* **2014**, *797*, 148.

- (23) Shukla, D.; Hernández, C. X.; Weber, J. K.; Pande, V. S. Markov State Models Provide Insights into Dynamic Modulation of Protein Function. *Acc. Chem. Res.* **2015**, *48* (2), 414–422.
- (24) Shukla, S.; Shamsi, Z.; Moffett, A. S.; Selvam, B.; Shukla, D. Application of Hidden Markov Models in Biomolecular Simulations. In *Hidden Markov Models*; Springer, 2017; pp 29–41.
- (25) Shamsi, Z.; Moffett, A. S.; Shukla, D. Enhanced Unbiased Sampling of Protein Dynamics Using Evolutionary Coupling Information. *Sci. Rep.* **2017**, *7* (1), 1–13.
- (26) Zhou, S.; Wang, Q.; Wang, Y.; Yao, X.; Han, W.; Liu, H. The Folding Mechanism and Key Metastable State Identification of the PrP127–147 Monomer Studied by Molecular Dynamics Simulations and Markov State Model Analysis. *Phys. Chem. Chem. Phys.* **2017**, *19* (18), 11249–11259.
- (27) Zhou, H.; Dong, Z.; Verkhivker, G.; Zoltowski, B. D.; Tao, P. Allosteric Mechanism of the Circadian Protein Vivid Resolved through Markov State Model and Machine Learning Analysis. *PLoS Comput. Biol.* **2019**, *15* (2), No. e1006801.
- (28) Roweis, S. T.; Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **2000**, *290* (5500), 2323–2326.
- (29) Cunningham, J. P.; Ghahramani, Z. Linear Dimensionality Reduction: Survey, Insights, and Generalizations. *J. Mach. Learn. Res.* **2015**, *16* (89), 2859–2900.
- (30) Sugiyama, M. Nonlinear Dimensionality Reduction. *Introduction to Statistical Machine Learning* **2016**, 429–446.
- (31) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3* (29), 861.
- (32) Becht, E.; McInnes, L.; Healy, J.; Dutertre, C. A.; Kwok, I. W. H.; Ng, L. G.; Ginhoux, F.; Newell, E. W. Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nat. Biotechnol.* **2019**, *37* (1), 38–47.
- (33) Cao, J.; Spielmann, M.; Qiu, X.; Huang, X.; Ibrahim, D. M.; Hill, A. J.; Zhang, F.; Mundlos, S.; Christiansen, L.; Steemers, F. J.; Trapnell, C.; Shendure, J. The Single-Cell Transcriptional Landscape of Mammalian Organogenesis. *Nature* **2019**, *566* (7745), 496–502.
- (34) Packer, J. S.; Zhu, Q.; Huynh, C.; Sivaramakrishnan, P.; Preston, E.; Dueck, H.; Stefanik, D.; Tan, K.; Trapnell, C.; Kim, J.; Waterston, R. H.; Murray, J. I. A Lineage-Resolved Molecular Atlas of *C. Elegans* Embryogenesis at Single-Cell Resolution. *Science* **2019**, *365* (6459), eaax1971.
- (35) Diaz-Papkovich, A.; Anderson-Trocmé, L.; Ben-Eghan, C.; Gravel, S. UMAP Reveals Cryptic Population Structure and Phenotype Heterogeneity in Large Genomic Cohorts. *PLoS Genet.* **2019**, *15* (11), No. e1008432.
- (36) Zoltowski, B. D.; Schwerdtfeger, C.; Widom, J.; Loros, J. J.; Bilwes, A. M.; Dunlap, J. C.; Crane, B. R. Conformational Switching in the Fungal Light Sensor Vivid. *Science* **2007**, *316* (5827), 1054–1057.
- (37) Zoltowski, B. D.; Crane, B. R. Light Activation of the LOV Protein Vivid Generates a Rapidly Exchanging Dimer. *Biochemistry* **2008**, *47* (27), 7012–7019.
- (38) Zoltowski, B. D.; Vaccaro, B.; Crane, B. R. Mechanism-Based Tuning of a LOV Domain Photoreceptor. *Nat. Chem. Biol.* **2009**, *5* (11), 827–834.
- (39) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2* (1–3), 37–52.
- (40) Pedregosa, F.; Michel, V.; Varoquaux, G.; Thirion, B.; Dubourg, V.; Passos, A.; Perrot, M.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Courneau, D.; Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Thirion, B.; Grisel, O.; Dubourg, V.; Passos, A.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (41) Naritomi, Y.; Fuchigami, S. Slow Dynamics in Protein Fluctuations Revealed by Time-Structure Based Independent Component Analysis: The Case of Domain Motions. *J. Chem. Phys.* **2011**, *134* (6), 065101.
- (42) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9* (4), 2000–2009.
- (43) Sultan, M. M.; Pande, V. S. TICA-Metadynamics: Accelerating Metadynamics by Using Kinetically Selected Collective Variables. *J. Chem. Theory Comput.* **2017**, *13* (6), 2440–2447.
- (44) Harrigan, M. P.; Sultan, M. M.; Hernández, C. X.; Husic, B. E.; Eastman, P.; Schwantes, C. R.; Beauchamp, K. A.; McGibbon, R. T.; Pande, V. S. MSMBuilder: Statistical Models for Biomolecular Dynamics. *Biophys. J.* **2017**, *112* (1), 10–15.
- (45) Van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9* (11), 2579–2605.
- (46) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. The Protein Data Bank. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58* (6), 899–907.
- (47) Freddolino, P. L.; Gardner, K. H.; Schulten, K. Signaling Mechanisms of LOV Domains: New Insights from Molecular Dynamics Studies. *Photochem. Photobiol. Sci.* **2013**, *12* (7), 1158–1170.
- (48) Martínez-Rosell, G.; Giorgino, T.; De Fabritiis, G. Play-Molecule ProteinPrepare: A Web Application for Protein Preparation for Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2017**, *57* (7), 1511–1516.
- (49) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Cafilisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614.
- (50) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L. P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput. Biol.* **2017**, *13* (7), No. e1005659.
- (51) Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L. P.; Shukla, D.; Tye, T.; Houston, M.; Stich, T.; Klein, C.; Shirts, M. R.; Pande, V. S. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **2013**, *9* (1), 461–469.
- (52) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341.
- (53) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593.
- (54) Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson Correlation Coefficient. In *Noise Reduction in Speech Processing*; Springer, 2009; pp 1–4.
- (55) Aggarwal, C. C.; Hinneburg, A.; Keim, D. A. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **2001**, 1973, 420–434.
- (56) Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20* (C), 53–65.
- (57) Liaw, A.; Wiener, M. Classification and Regression by Random Forest. *R News* **2002**, *2* (December), 18–22.
- (58) Turney, P. Bias and the Quantification of Stability. *Mach. Learn.* **1995**, *20* (1), 23–33.
- (59) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32.
- (60) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140* (7), 2386–2396.

- (61) E., W.; Vanden-Eijnden, E. Towards a Theory of Transition Paths. *J. Stat. Phys.* **2006**, *123* (3), 503–523.
- (62) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. Transition Path Theory for Markov Jump Processes. *Multiscale Model. Simul.* **2009**, *7* (3), 1192–1219.
- (63) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (45), 19011–19016.
- (64) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and Challenges in the Automated Construction of Markov State Models for Full Protein Systems. *J. Chem. Phys.* **2009**, *131* (12), 124101.
- (65) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know about Markov State Models but Were Afraid to Ask. *Methods* **2010**, *52* (1), 99–105.
- (66) Wang, F.; Shen, L.; Zhou, H.; Wang, S.; Wang, X.; Tao, P. Machine Learning Classification Model for Functional Binding Modes of TEM-1 β -Lactamase. *Front. Mol. Biosci.* **2019**, *6*, 47.
- (67) Tian, H.; Tao, P. Deciphering the Protein Motion of S1 Subunit in SARS-CoV-2 Spike Glycoprotein through Integrated Computational Methods. *J. Biomol. Struct. Dyn.* **2020**, 1–8.
- (68) Tian, H.; Trozzi, F.; Zoltowski, B. D.; Tao, P. Deciphering the Allosteric Process of the *Phaeodactylum Tricornutum* Aureochrome 1a LOV Domain. *J. Phys. Chem. B* **2020**, *124* (41), 8960–8972.
- (69) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 2. Example Applications to Alanine Dipeptide and a β -Hairpin Peptide. *J. Phys. Chem. B* **2004**, *108* (21), 6582–6594.