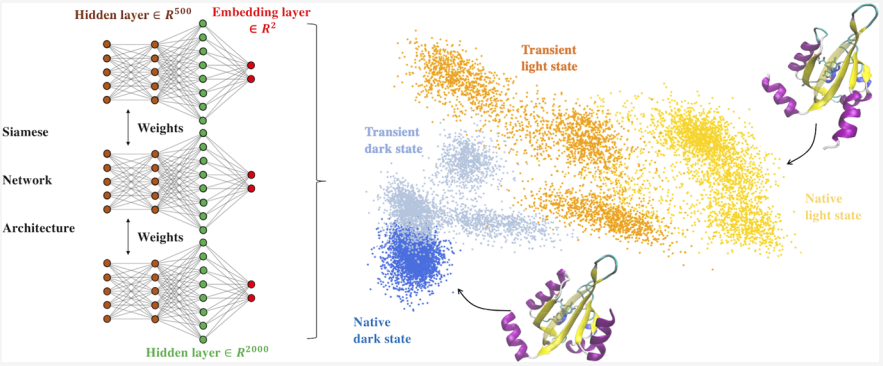
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

ivis Dimensionality Reduction Framework for Biomacromolecular Simulations

Hao Tian and Peng Tao*

Cite This: https://dx.doi.org/10.1021/acs.jcim.0c00485 Read Online



ABSTRACT: Molecular dynamics (MD) simulations have been widely applied to study macromolecules including proteins. However, the high dimensionality of the data sets produced by simulations makes thorough analysis difficult and further hinders a deeper understanding of biomacromolecules. To gain more insights into the protein structure−function relations, appropriate dimensionality reduction methods are needed to project simulations onto low-dimensional spaces. Linear dimensionality reduction methods, such as principal component analysis (PCA) and time−structure-based independent component analysis (t-ICA), could not preserve sufficient structural information. Though better than linear methods, nonlinear methods, such as t-distributed stochastic neighbor embedding (t-SNE), still suffer from the limitations in avoiding system noise and keeping inter-cluster relations. ivis is a novel deep learning-based dimensionality reduction method originally developed for single-cell data sets. Here, we applied this framework for the study of light, oxygen, and voltage (LOV) domains of diatom *Phaeodactylum tricornutum* aureochrome 1a (PtAu1a). Compared with other methods, ivis is shown to be superior in constructing a Markov state model (MSM), preserving information of both local and global distances, and maintaining similarity between high and low dimensions with the least information loss. Moreover, the ivis framework is capable of providing new perspectives for deciphering residue-level protein allostery through the feature weights in the neural network. Overall, ivis is a promising member of the analysis toolbox for proteins.

INTRODUCTION

Molecular dynamics (MD) simulations have been widely used in biomolecules to provide insights into their functions in atomic-scale mechanisms.[1] For this purpose, an extensive time scale is generally preferred for the simulations to study protein dynamics and functions. Due to the emergence of graphics processing units (GPU) and their application in biomolecular simulations, the MD simulation time scale has reached from nanoseconds to experimentally meaningful microseconds.[2,3] However, simulation data for biomacromolecules such as proteins are high-dimensional and suffer from the curse of dimensionality,[4] which hinders an in-depth analysis, including extracting slow time-scale protein motions,[5] identifying representative protein structures,[6] and clustering kinetically similar macrostates.[7] To make these analyses feasible, it will be informative to construct a low-dimensional space to characterize protein dynamics in the best way possible.

In recent years, new dimensionality reduction algorithms have been developed and can be applied to analyze protein simulations, construct representative distributions in a low-dimensional space, and extract intrinsic relations between the protein structure and functional dynamics. These methods can be generally categorized into linear and nonlinear methods.[8,9] Linear dimensionality reduction methods produce new variables as the linear combination of the input variables, such as principal component analysis (PCA)[10] and time−

Received: May 6, 2020
Published: August 21, 2020

https://dx.doi.org/10.1021/acs.jcim.0c00485
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

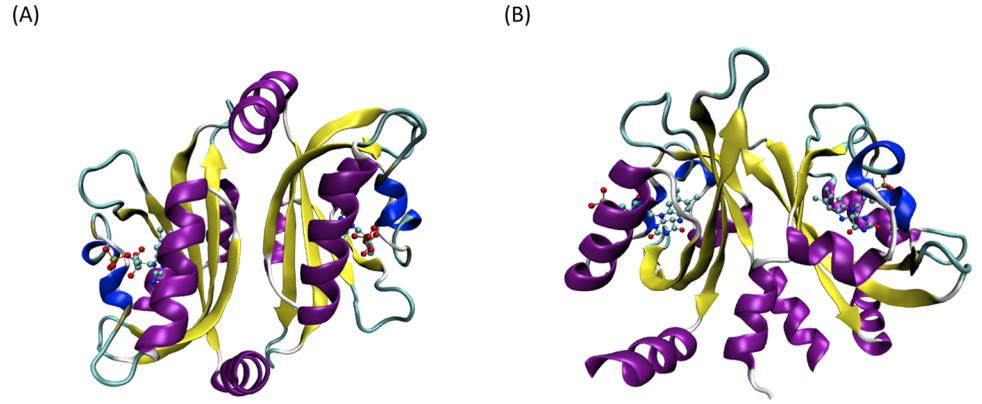


Figure 1. Native structures of AuLOV. (A) Two monomers in the native dark state. (B) Dimer in the native light state. The sequence of secondary structure starts from Ser240 to Glu367.

structure-based independent component analysis (t-ICA).[11] Nonlinear methods construct variables through a nonlinear function, including t-distributed stochastic neighbor embedding (t-SNE)[12] and auto encoders.[13] It was reported that nonlinear methods were more powerful in reducing dimensionality while preserving representative structures.[14]

Information is inevitably lost to a certain degree through the dimensionality reduction process.[15] It is expected that the distances among data points in the low-dimensional space resemble the original data in the high-dimensional space. The Markov state model (MSM) is often applied to study the dynamics of biomolecular systems. MSM is constructed by clustering states in the reduced dimensional space to obtain long-time kinetic information.[16] However, many dimensionality reduction methods, such as PCA and t-ICA, fail to keep the similarity characteristics in the low dimension, which would cause a misleading clustering analysis based on the projections of low-dimensional space.[17] Therefore, more appropriate dimensionality reduction methods are needed to build a proper MSM.

A novel framework, ivis[18] is a recently developed dimensionality reduction method for single-cell data sets. ivis is a nonlinear method based on Siamese neural networks (SNNs).[19] The SNN architecture consists of three identical neural networks and ranks the similarity to the input data. The loss function used for the training process is a triplet loss function[20] that calculates the Euclidean distance among data points and simultaneously minimizes the distances between data of the same labels while maximizing the distances between data of different labels. Due to this intrinsic property, the ivis framework is capable of preserving both local and global structures in a low-dimensional space.

With the success in single-cell expression data, the ivis framework is promising as a dimensionality reduction method for simulations of biomacromolecules to investigate their functional dynamics such as allostery. Diatom *Phaeodactylum tricornutum* aureochrome 1a (PtAu1a) is a recently discovered light, oxygen, or voltage (LOV) protein from the photosynthetic stramenopile alga Vaucheria frigida.[21] This protein consists of an N-terminal domain, a C-terminal LOV core, and a basic region leucine zipper (bZIP) DNA-binding domain. PtAu1a is a monomer in the native dark state. The interaction between its LOV core and bZIP prohibits DNA binding.[22] Upon light perturbation, a covalent bond forms between the C4a position of the cofactor flavin mononucleotide (FMN) and sulfur in cysteine 287, triggering a conformational change that leads to the LOV domain dimerization. In the current study, the PtAu1a LOV domain (AuLOV) with two flanking helices (A′α and Jα helices) is simulated through MD simulations. The structures of both the dark and light states are shown in Figure 1. The main difference between AuLOV and most other LOV proteins is that the LOV domain lies in the C-terminal in AuLOV while it lies in the N-terminal in other LOV proteins.[23,24] Therefore, the conformational changes in AuLOV are expected to differ from other LOV proteins, raising the question of how the allosteric signal transmits in AuLOV. In the current study, the ivis framework, together with other dimensionality reduction methods, is applied to project the AuLOV simulations onto reduced dimensional spaces. The performances of the selected methods are assessed and compared, validating the ivis as a superior framework for dimensionality reduction of biomacromolecule simulations.

METHODS

Molecular Dynamics (MD) Simulations. The crystal structures of AuLOV dark and light states were obtained from the Protein Data Bank (PDB)[25] with PDB IDs 5dkk and 5dkl, respectively. The light structure sequence starts from Gly234, while the dark structure sequence starts from Phe239 in chain A and Ser240 in chain B. For consistency, residues before Ser240 were removed to keep the same number of residues in all chains. Therefore, simulations of dark and light states can be treated similarly. Both structures contain FMN as a cofactor. The FMN force field from a previous study[26] was used in this study. Two new states, named the transient dark state (forcing the cysteinyl−flavin C4a adduct in the dark state structure) and the transient light state (breaking the cysteinyl−flavin C4a adduct in the light state structure), were constructed to fully explore the protein conformational space. Two monomers (Figure 1A) and a dimer (Figure 1B) were simulated in the dark and light states, respectively.

The crystal structures with added hydrogen atoms were solvated within a rectangular water box using the TIP3P water model.[27] Sodium and chlorine ions were added for charge neutralization. Energy minimization was done for each water box. The system was further subjected to 20 picoseconds (ps) of MD simulations to increase the temperature from 0 to 300 K and another 20 ps simulation for equilibrium. In all, 10 nanoseconds (ns) of isothermal−isobaric ensemble (NPT) MD simulations under 1 bar pressure were conducted. The

https://dx.doi.org/10.1021/acs.jcim.0c00485
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

canonical ensemble (NVT) is usually applied in the production runs to investigate the allosteric process.[28,29] In all, 1.1 microseconds (μs) of a canonical ensemble (NVT) Langevin MD simulation at 300 K was carried out for each production run. The Langevin dynamics friction coefficient that couples the system to a heat bath was set to 1 ps^{-1},[30,31] with minimum perturbation to the dynamical properties of the protein system.[32] For all production simulations, the first 100 ns simulation is treated as the equilibration stage and not included in the analysis. For each structure, three independent MD simulations were carried out and a total of 12 μs simulations were used in the analysis. All chemical bonds associated with hydrogen atoms were constrained with the SHAKE method. A step size of 2 femtoseconds (fs) was used and simulation trajectories were saved for every 100 ps. The periodic boundary condition (PBC) was applied in simulations. Electrostatic interactions were calculated with the particle mesh Ewald (PME) algorithm[33] and a cutoff of 1.2 nanometers (nm). Simulations were conducted using graphics processing unit accelerated calculations of OpenMM[34] with the CHARMM[35] simulation package, version c41b1, and the CHARMM27 force field.[36]

Feature Processing. In MD simulations, protein structures are represented as atom positions in Cartesian coordinates. However, this representation is neither rotation invariant nor feasible for analysis purposes due to the significant number of atoms with a total of 3 N degrees of freedom. To represent the protein structures with rotational invariance and essential structural information, pair-wised backbone Cα distances were selected to represent the overall protein configuration. Following our previously proposed feature-processing method,[37] distances were encoded as a rectified linear unit (ReLU)[38]-like activation function and further expanded as a vector.

$$\mathrm{ReLU}(x) = \max(0, x) \tag{1}$$

Dimensionality Reduction Methods. *ivis.* ivis is a deep learning-based method for structure-preserving dimensionality reduction. This framework is designed using Siamese neural networks, which implement a novel architecture to rank similarity among input data. Three identical networks are included in the SNN. Each network consists of three dense layers and an embedding layer. The size of the embedding layer was set to 2, aiming to project high-dimensional data into a two-dimensional (2D) space. The scaled exponential linear unit (SELU)[39] activation function is used in the dense layers

$$\mathrm{selu}(x) = \lambda\begin{cases} x & \text{if } x > 0 \\ \alpha\exp(x) - \alpha, & \text{if } x \le 0 \end{cases} \tag{2}$$

The LeCun normal distribution is applied to initialize the weights of these layers. For the embedding layer, the linear activation function is used, and weights are initialized with Glorot's uniform distribution. To avoid overfitting, dropout layers with a default dropout rate of 0.1 are used for each dense layer.

A triplet loss function is used as the loss function for training

$$L_{\mathrm{tri}}(\theta) = \left[\sum_{a,p,n} D_{a,p} - \min(D_{a,n}, D_{p,n}) + m\right]_+ \tag{3}$$

where a, p, and n are the anchor points, positive points, and negative points, respectively. D and m are the Euclidean distance and margin, respectively. Anchor points are points of interest. The triplet loss function aims to minimize the distance between anchor points and positive points while maximizing the distances between anchor points and negative points. The distances between positive points and negative points are also taken into account, as shown in $\min(D_{a,n}, D_{p,n})$ in the above equation.

The k-nearest neighbors (KNNs) are used to obtain data for the triplet loss function. k is a tuning parameter and is set to 100. For each round of calculation, one point in the data set is selected as an anchor. A positive point is randomly selected among the nearest k neighbors around the anchor, and a negative point is randomly selected outside the neighbors. For each training epoch, the triplet selection is updated to maximize the differences in both local and global distances.

If the data set can be classified into different groups based on their intrinsic properties, ivis can also be used as a supervised learning method by combining the distance-based triplet loss function with a classification loss. Supervision weight is a tuning parameter to control the relative importance of loss function in labeling classification.

The neural network is trained using the Adam optimizer function with a learning rate of 0.001. Early stopping is a method to prevent overfitting in a training neural network and is applied in this study to terminate the training process if the loss function does not decrease after 10 consecutive epochs.

Time−Structure-Independent Component Analysis (t-ICA). The t-ICA method finds the slowest motion or dynamics in molecular simulations and is commonly used as a dimensionality reduction method for macromolecular simulations.[11] For a given n-dimensional data, t-ICA is employed by solving the following equation

$$\bar{C}F = CKF \tag{4}$$

where K is the eigenvalue matrix, C is the correlation matrix, and F is the eigenvector matrix. $\bar{C}$ is the time lag correlation matrix defined as

$$\bar{C} = \langle\langle x(t) - \langle x(t)\rangle)^t (x(t+\tau) - \langle x(t)\rangle)\rangle \tag{5}$$

The results calculated by t-ICA are linear combinations of input features that are highly autocorrelated.

Principal Component Analysis (PCA). PCA is a method that finds the projection vectors that maximize the variance by conducting an orthogonal linear transformation.[10] In the new coordinate system, the greatest variance of the data lies on the first coordinate and is called the first principal component. Principal components can be solved through the singular value decomposition (SVD).[40] Given data matrix X, the covariance matrix can be calculated as

$$C = X^{\mathrm{T}}X/(n-1) \tag{6}$$

where n is the number of samples. C is a symmetric matrix and can be diagonalized as

$$C = VLV^{\mathrm{T}} \tag{7}$$

where V is a matrix of eigenvectors and L is a diagonal matrix with eigenvalues λ_i in descending order.

t-Distributed Stochastic Neighbor Embedding (t-SNE). t-SNE is a nonlinear dimensionality reduction method that tries to embed similar objects in high dimensions to points close to each other in a low-dimensional space.[12] t-SNE has been demonstrated to be a suitable dimensionality reduction method for protein simulations.[41] The calculation process

https://dx.doi.org/10.1021/acs.jcim.0c00485
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

consists of two stages. First, conditional probability is calculated to represent the similarity between two objects as

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)} \quad (8)$$

where σ_i is the bandwidth of the Gaussian kernels.

While the conditional probability is not symmetric since $p_{j|i}$ is not equal to $p_{i|j}$, the joint probability is defined as

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \quad (9)$$

To better represent the similarity among objects in the reduced map, the similarity q_{ij} is defined as

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k\neq i}(1 + \|y_i - y_k\|^2)^{-1}} \quad (10)$$

Combined with the joint probability p_{ij} and similarity q_{ij}, Kullback–Leibler (KL) divergence is used to determine the coordinates of y_i as

$$\mathrm{KL}(P\|Q) = \sum_{i\neq j} p_{ij}\log\frac{p_{ij}}{q_{ij}} \quad (11)$$

The KL divergence measures the differences between high-dimensional data and low-dimensional points, which is minimized through the gradient descent method.

A drawback of the traditional t-SNE method is the slow training time. To speed up the computational time of the dimensionality reduction process, multicore t-SNE[42] is used and abbreviated as t-SNE in this study.

Performance Assessment Criteria. Several assessment criteria are applied to quantify and compare the performance of each dimensionality reduction method.

Root-Mean-Square Deviation (RMSD). The RMSD is used to measure the conformational change in each frame with regard to a reference structure. Given a molecular structure, the RMSD is calculated as

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}(r_i^0 - Ur_i)^2}{N}} \quad (12)$$

where r is a vector represented in Cartesian coordinates and r_i^0 is the ith atom in the reference structure.

Pearson Correlation Coefficient (PCC). The Pearson correlation coefficient[43] reflects the linear correlation between two variables. PCC has been rigorously applied to estimate the linear relation between distances in the original space and the reduced space.[44] The L2 distance, which is also called the Euclidean distance, is used for the distance calculation and is shown as follows

$$d_2(p, q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (13)$$

Based on the L2 distance expression, PCC is calculated as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})}} \quad (14)$$

where n is the sample size; x_i, y_i, $\bar{x}$, and $\bar{y}$ are the distances and the mean value of distances, respectively. x and y represent distances in the original data and the projected data, respectively.

Spearman's Rank-Order Correlation Coefficient. Spearman's rank-order correlation coefficient is used to quantitatively analyze how well distances between all pairs of points in the original spaces have been preserved in the reduced dimensions. Specifically, the Spearman correlation coefficient measures the difference in distance ranking, which is calculated as follows

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \quad (15)$$

where d_i is the difference in paired ranks and n equals the total number of samples.

Mantel Test. The Mantel test is a nonparametric method that was originally used in genetics;[45] it tests the correlation between two distance matrices. A common problem in evaluating the correlation coefficient is that distances are dependent on each other and therefore cannot be determined directly. The Mantel test overcomes this obstacle through permutations of the rows and columns of one of the matrices. The correlation between two matrices is calculated at each permutation. MantelTest GitHub repository[46] was used to implement the algorithm.

Shannon Information Content (IC). While chemical information in the original space could be lost to a certain degree in the reduced space, dimensionality reduction methods are expected to keep the maximum information. Shannon information content is applied to test the information preservation in the reduced space, which is defined as

$$I(x) = -\log_2(P) \quad (16)$$

where P is the probability of a specific event x.

To avoid the possible dependency among different features in the reduced dimensions, the original space was reduced to one dimension (1D) to calculate the IC. The values in the 1D were sorted and put into 100 bins of the same length. The bins were treated as events and the corresponding probabilities were calculated as the ratio of the number of samples in each bin to the total number of samples.

Markov State Model (MSM). The Markov state model has been widely used to partition the protein conformational space into kinetically separated macrostates[47] and estimate the relaxation time to construct long-time-scale dynamics behavior.[6] MSMBuilder[48] (version 3.8.0) was employed to implement the Markov state model. The k-means clustering method was used to cluster 1000 microstates. A series of lag times at equal intervals was set to calculate the transition matrix. The corresponding second eigenvalue was used to estimate the relaxation time scale, which was calculated as

$$t(\tau) = -\frac{\tau}{\ln\lambda_1} \quad (17)$$

where λ_1 is the second eigenvalue and τ is the lag time.

The generalized matrix Rayleigh quotient (GMRQ),[49] generated using the combination of cross-validation and variational approaches, was used to assess the effectiveness of MSM in dimensions and dimensionality reduction methods. State decompositions are different through various dimensionality reduction methods. A good method is expected to lead to a Markov state model with larger GMRQ values.

https://dx.doi.org/10.1021/acs.jcim.0c00485
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

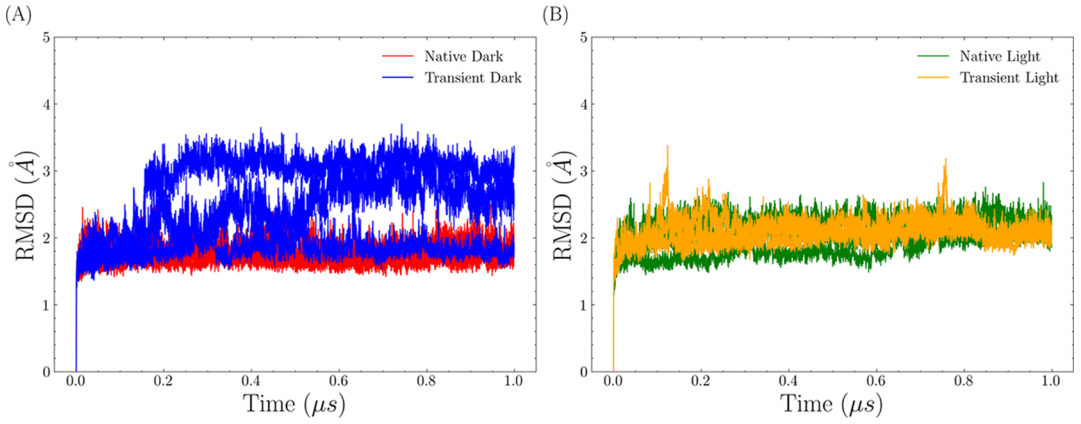


Figure 2. RMSDs of AuLOV MD trajectories. (A) Native dark and transient dark states. (B) Native light and transient light states. For each state, three independent simulations were carried out.

Machine Learning Methods. Random Forest (RF). Random forest[50] is a supervised machine learning method that was used in this study for trajectory-state classification. A random forest model consists of multiple decision trees, which are a class of partition algorithm that recursively groups data samples of the same label. Features at each split are selected based on the information gained. A final prediction result of the random forest is made from the results in each decision tree through a voting algorithm. For random forest models at each depth, the number of decision trees was set to 50. Scikit-learn (version 0.20.1)[51] was used for RF implementation.

Artificial Neural Network (ANN). An artificial neural network was used to learn the nonlinear relationships of coordinates on the reduced 2D dimension. An ANN is generally formed with an input layer, a hidden layer, and an output layer. In each layer, different neurons (nodes) are assigned and connected with adjacent layer(s). During the training process, input data are fed through the input and hidden layers and prediction results are made in the output layer. For each training step, the error between the predicted result and the actual result is propagated from the output layer back to the input layer, which is also called back-propagation,[52] and the weight in every neuron is updated. When there is more than one hidden layer, ANN is also referred to as a deep neural network (DNN), which requires more computation power. To minimize the training cost, only two hidden layers, each with 64 nodes, were used. The Adam optimizer[53] was used for weight optimization. ANN was implemented with Keras (version 2.2.4-tf).[54]

RESULTS

Data Set of Cα Distances Represents the Protein Structures. There are two native states of AuLOV: native dark state and native light state. To explore the protein response with regard to the formation of the covalent bond between cysteine 287 and FMN, two new states were constructed as a transient dark state and a transient light state by forcing the cysteinyl−flavin C4a adduct in the native dark state and breaking this adduct in the native light state, respectively. The RMSDs of MD simulations are plotted in Figure 2. For each trajectory, the RMSD values were calculated with regard to the first frame. Averaged RMSDs were 1.75, 2.04, 2.39, and 2.08 Å in native dark, transient dark, transient light, and native light states, respectively. Compared with the result in the native dark state, the higher RMSD value in the transient dark state indicates that the light-induced covalent bond Cys287-FMN increases the protein flexibility and dynamics. The transient light state displays the highest averaged RMSD value, indicating the highest flexibility or the largest conformational change.

The pair-wised distances of backbone Cα in simulations were extracted as features representing the character of protein configurations. There are 254 residues in the AuLOV structure, and a total of 254 × 253/2 = 32 131 Cα distances were calculated. Before further analysis, features were transformed into vectors with our proposed technique outlined in the Methods section. Considering the nonbonded chemical interaction, 10.0 Å was selected as the threshold for feature transformation. There are 10 000 frames in each trajectory, leading to a sample size of 120 000 in the overall data set. Full data sets were applied in all analyses. To gain more statistical significance, each MD trajectory was split into five subtrajectories at equal intervals. The performance assessments were conducted for each subtrajectory independently. The mean and standard deviation values of the five subsets were calculated.

Information is Well Preserved in the ivis Dimensionality Reduction Method. Several hyperparameters of the ivis model were selected based on the recommended values for different observation sizes. Given the large number of sample size, k was set to 100 and the number of early stopping epochs was 10. The Maaten neural network architecture was selected, which consists of three dense layers with 500, 500, and 2000 neurons, respectively. To select the best parameter of supervision weight, the full trajectory data set was randomly split into a training set (70%) and a testing set (30%). ivis models were trained on the training set and validated on the testing set. The prediction result with different supervision weights is plotted in Figure 3. The ivis model performed poorly at 0.0 supervision weight, which corresponds to unsupervised ivis, with an average accuracy below 25%. The average accuracy values for other supervision weights were stable and over 95%. Specifically, there was no significant increase in the accuracy value after the 0.1 supervision weight, which was chosen as the hyperparameter for the supervised ivis model. An unsupervised ivis framework with the same value of other hyperparameters was applied for comparison.

Four dimensionality reduction models (supervised ivis, PCA, t-SNE, and t-ICA) were applied to the MD simulations to project a high-dimensional (32 131) space to a 2D surface

https://dx.doi.org/10.1021/acs.jcim.0c00485
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

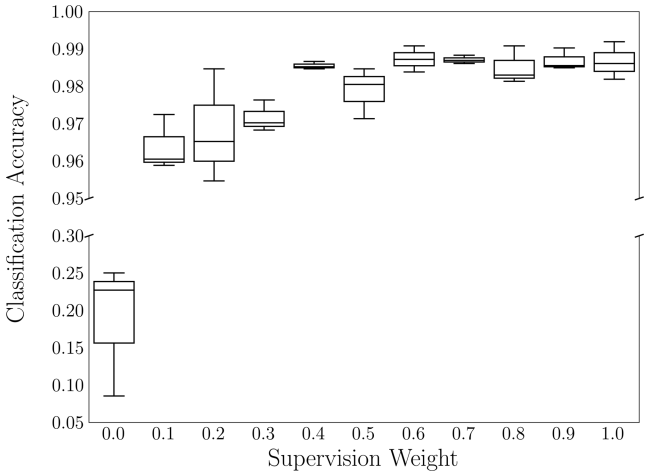


Figure 3. Classification accuracy using the ivis framework with different supervision weights. With the 0.0 supervision weight, it is referred to as an unsupervised ivis model. Classification accuracy is high for any nonzero supervision weight. Therefore, 0.1 was chosen as the hyperparameter for supervised ivis.

(Figure 4). The supervised ivis dimensionality reduction method, as well as PCA, successfully separated dark and light states while keeping the corresponding transition states close (Figure 4A,B). These states are important for dynamical analysis as they could be used to reveal the free-energy and kinetic transition landscape for the target system. For t-SNE (Figure 4C) and t-ICA (Figure 4D) projections, the transient dark state and native dark state overlap significantly, thus hindering the extraction of thermodynamics and kinetics information. Therefore, the supervised ivis dimensionality reduction method and PCA are demonstrated to be proper in representing the chemical information in the low dimension among the investigated methods.

The *k*-means clustering was used in the reduced dimensions to partition a total of 120 000 frames from AuLOV MD trajectories into 1000 microstates. Within each cluster, the RMSDs were calculated for each structure pair. An RMSD value of each cluster is defined as the average RMSD value among all structure pairs within that cluster. The results of five dimensionality reduction models are shown in Figure 5A. The average RMSD value of an appropriate microstate should be lower than 1.0 Å.[55,56] From this perspective, unsupervised ivis and supervised ivis show similar values in each microstate and are the best two methods among the selected methods. As reported previously,[41] t-SNE also exhibited good performance in measuring the similarity with the Cartesian coordinates.

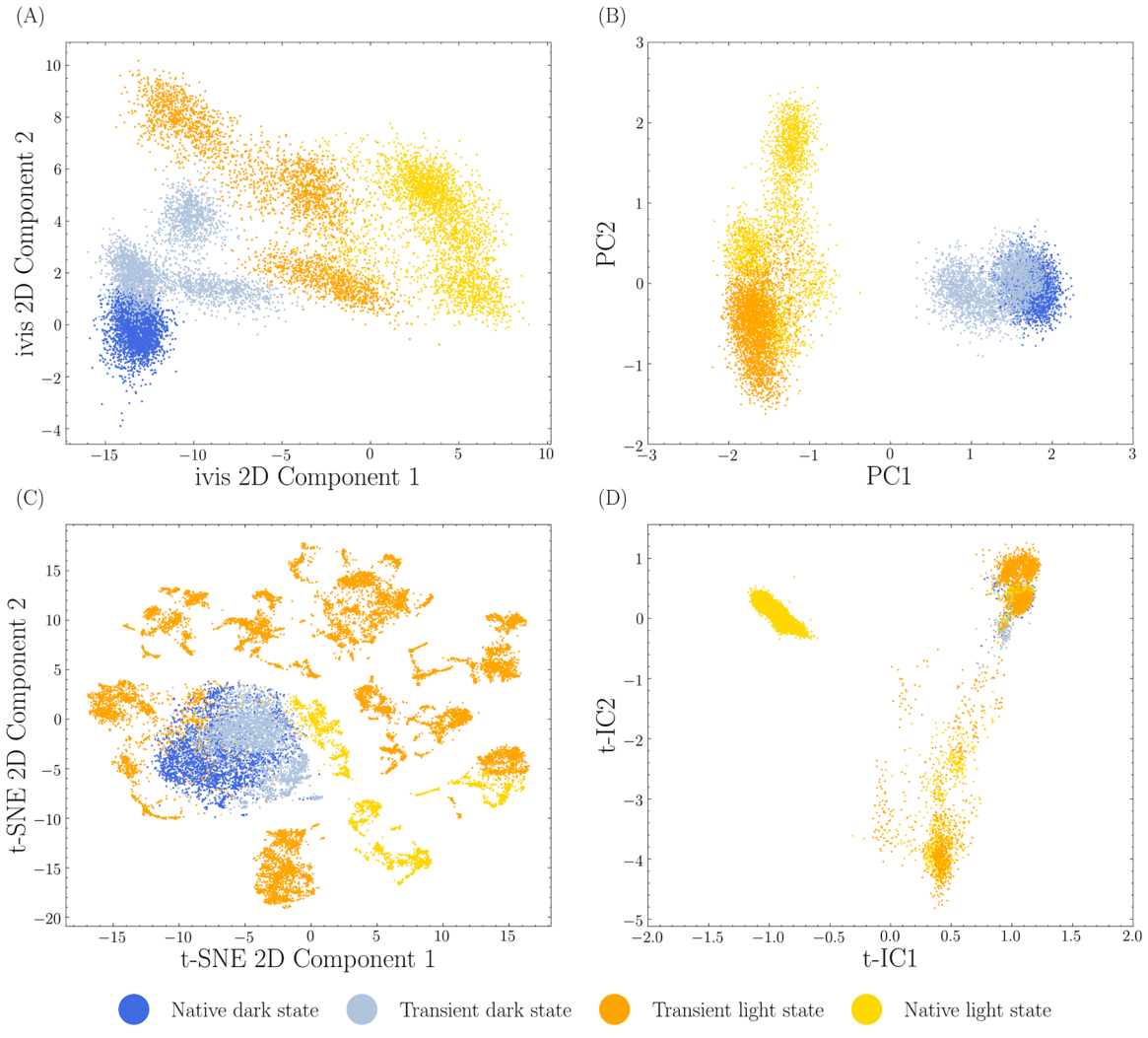


Figure 4. Two-dimensional projections of four dimensionality reduction methods: (A) supervised ivis, (B) PCA, (C) t-SNE, and (D) t-ICA.

https://dx.doi.org/10.1021/acs.jcim.0c00485
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

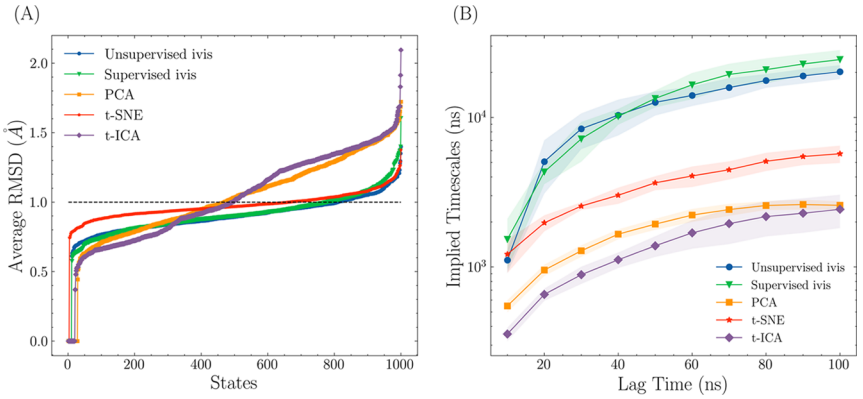


Figure 5. Analysis results of 2D projections for different dimensionality reduction methods. (A) Average values of RMSDs in microstates clustered within a projected 2D dimensional space. (B) Estimated implied time scales from Markov state models with regard to different lag times. For each model, the mean value of the implied time scale is calculated among five subsets and is plotted in solid color. The standard deviation is calculated to show the stability for each lag time and is illustrated using a light color.

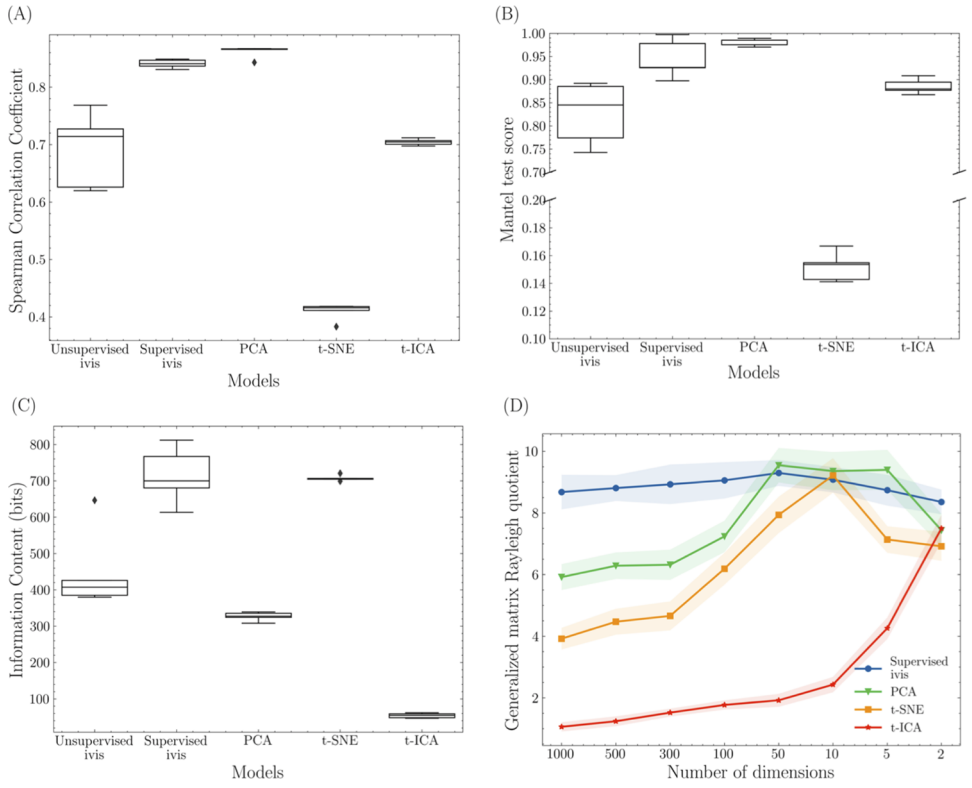


Figure 6. Results of quantitative analysis. (A) Spearman correlation coefficient results of different dimensionality reduction methods. The mean values for unsupervised ivis, supervised ivis, PCA, t-SNE, and t-ICA are 0.69, 0.84, 0.86, 0.41, and 0.70, respectively. The height of each box represents the interquartile range. (B) Mantel test scores of different dimensionality reduction methods. The mean values for unsupervised ivis, supervised ivis, PCA, t-SNE, and t-ICA are 0.83, 0.95, 0.98, 0.15, and 0.89, respectively. (C) Shannon information content of different dimensionality reduction methods. The mean values for unsupervised ivis, supervised ivis, PCA, t-SNE, and t-ICA are 449.0, 714.6, 327.0, 707.3, and 54.0, respectively. To avoid dependent variables in the information content calculation, high-dimensional Cα distances were projected to 1D. (D) Generalized matrix Rayleigh quotient with different dimensions and dimensionality reduction methods.

A metric to compare different dimensionality reduction methods is the implied relaxation time scale calculated from the Markov state model. To build MSM, MD simulations were projected onto a 2D space and 1000 microstates were sampled through *k*-means with the corresponding estimated relaxation time scales. For each method, the slowest time scale in each lag time was extracted based on different lag times ranging from 10 to 100 ns and is shown in Figure 5B. The convergence of time scales is important for eigenvalue and eigenvector calculations.[57] For all five models, relaxation time scales converged, indicating the Markovianity of the MSMs. Both supervised ivis and unsupervised ivis models show long time scales, indicating the effectiveness of MSM built on the reduced spaces.

Euclidean distances between data points in the low-dimensional space are expected to reflect the similarity in the high dimension. To quantify the degree of this relationship kept in reduced dimensional space, Spearman correlation coefficients were calculated between Euclidean distance pairs

https://dx.doi.org/10.1021/acs.jcim.0c00485
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

in the original space and those in the reduced space. The results are shown in Figure 6A. While PCA preserved the Euclidean distances well with an average coefficient of 0.86, the supervised ivis model showed a comparable high coefficient of 0.84. The unsupervised ivis model also exhibited the ability to preserve the linear relationship. The poor performance of the t-SNE model may be due to the fact that t-SNE is a nonlinear method and therefore suffers from the problem that distance in the high-dimensional space is not linearly projected to the low-dimensional space, as reported in other studies.[58,59]

While ivis models showed good ability in keeping the linear projection relation, the Spearman correlation coefficient fails to overcome the problem that features are not independent. The pair-wised distances are subjected to the molecular motion of Cα, wherein changing the coordinate of one Cα atom would affect the distances related to this atom. Therefore, to address this issue, the Mantel test was used to randomize the Euclidean distances. Permutations of rows and columns in the Euclidean distance matrix were done 10 000 times, with the Pearson correlation coefficient calculated each time. The results of the Mantel test are plotted in Figure 6B. Both unsupervised ivis and supervised ivis showed remarkable results in preserving the correspondence relationship in a randomized order, with the mean coefficients of 0.83 and 0.95, respectively.

During the process of dimensionality reduction, information is inevitably lost to some degree. To measure the retaining information through the dimensionality reduction process, the Shannon information is applied to the coordinates in the low-dimensional space. However, when dealing with multiple variables, especially for the dependent Cα distances, the total Shannon information is not equal to the sum of the Shannon information of each variable. To reduce the computation complexity, high-dimensional features were reduced to 1D for calculation and the results are plotted in Figure 6C. It shows that the supervised ivis model is superior in preserving information content with the least information loss. It is also worth noting that t-SNE showed better performance than the unsupervised ivis model.

To study the performance of the Markov state model on dimensions and dimensionality reduction methods, the generalized matrix Rayleigh quotient was calculated for each dimension and method (Figure 6). The results of four methods showed different trends. Supervised ivis and t-ICA methods were the least and most affected by the number of dimensions, respectively. For PCA and t-SNE, the optimal parameter of the number of dimensions is in the tens. Two dimensions are typically used for MSM construction and visualization purpose.[60−62] In this regard, supervised ivis exhibited the highest GMRQ value.

ivis Helps in Understanding Biological Systems and the Allosteric Mechanism. The transition probabilities among macrostates in ivis projections are shown in Figure 7. Based on the similarity to crystal structures, macrostates 2 and 8 are referred to as the native dark state and native light state, respectively. Other macrostates are considered as transient states. The probabilities of the native dark state and native light state to remain to themselves are among the highest ones and indicate the high stability of these two states. It is interesting to observe that macrostate 9 may have the highest stability among all Markov states.

The effectiveness of MSM depends on the projected 2D space, where appropriate discrete states are produced by clustering the original data points in the projection space. The

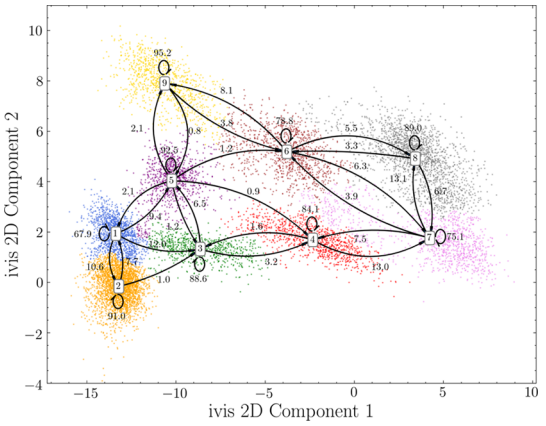


Figure 7. Transition probabilities between macrostates in the ivis dimensionality reduction 2D projections.

number of macrostates is determined based on the implicated time scales using different lag times in different reduced spaces. In this study, nine, nine, seven, nine, and seven macrostates were selected for unsupervised ivis, supervised ivis, PCA, t-SNE, and t-ICA, respectively. The samples were clustered through Perron-cluster cluster analysis (PCCA). The data set was further split into a training set (70%) and a testing set (30%). Two machine learning methods (random forest and artificial neural network) were applied to predict the macrostates of each data point based on the pair-wised Cα distances. Prediction accuracy results are plotted in Figure 8A,B. It shows that the supervised ivis framework is the best among the five dimensionality reduction methods. Surprisingly, while the unsupervised ivis model was trained without class labels in the loss function, the high prediction accuracy of this model demonstrates its good performance on the 2D projections. Random forest is often applied to distinguish the macrostates since it provides feature importance, which is important for the interpretation of biological systems. The accumulated feature importance of a random forest model on the supervised ivis model is plotted in Figure 8C. The top 490 features account for 90.2% of the overall feature importance.

The high prediction accuracy of the supervised ivis framework suggests that supervised ivis is more promising in elucidating the conformational differences among macrostates. The neural network architecture on the first dense layer of the supervised ivis model was 32 131 × 500, where 32 131 and 500 represent the number of Cα distances and dense layers, respectively. To identify key residues and structures that are important in the dimensionality reduction process, 32 131 feature weights on the last layer were treated as the feature importance and shown as the protein contact map in Figure 9. The contact map is symmetrical along the diagonal. The upper triangular part is divided into four regions as follows: region 1: pair wised Cα distances within chain A, region 4: Cα distances within chain B, and regions 2 and 3: Cα distances between chains A and B. Our results show characteristics similar to those in a previous study.[17] Local protein structures are encoded to features close to the diagonal. Global structures are encoded to features further from the diagonal. In Figure 9, the local information is shown as a red rectangle as the Cα and Dα helices in the AuLOV system, and global information is shown as a black rectangle as the Gβ and Hβ strands. While regions 2 (protein interactions from chain A to chain B) and 3 (protein interactions from chain B to chain A) are mostly symmetrical,

https://dx.doi.org/10.1021/acs.jcim.0c00485
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

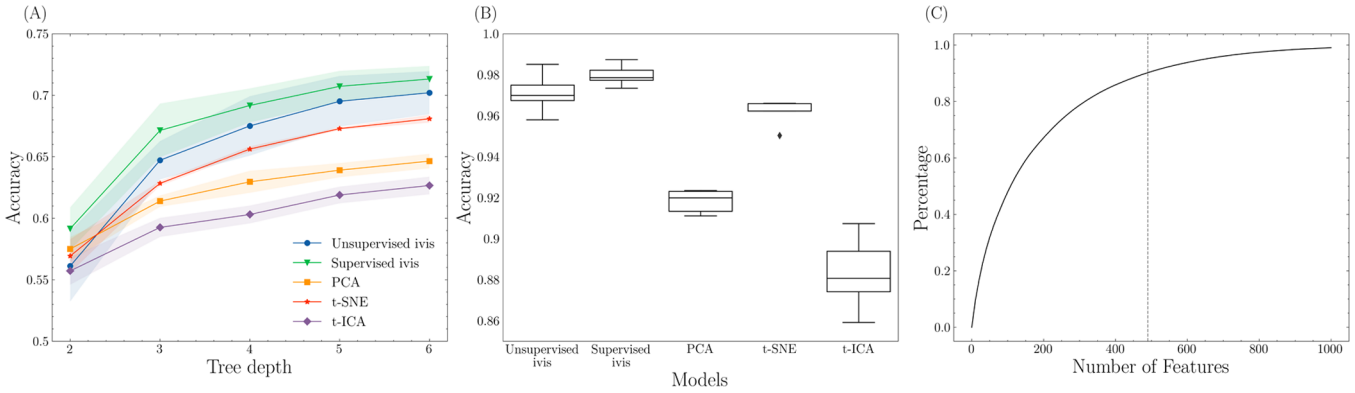


Figure 8. Prediction accuracy of different machine learning models. (A) A random forest and (B) an artificial neural network were used on the reduced 2D spaces to predict the labels of macrostates from MSM. (C) Accumulated feature importance of a random forest model applied in the projections of the supervised ivis framework at depth 5.

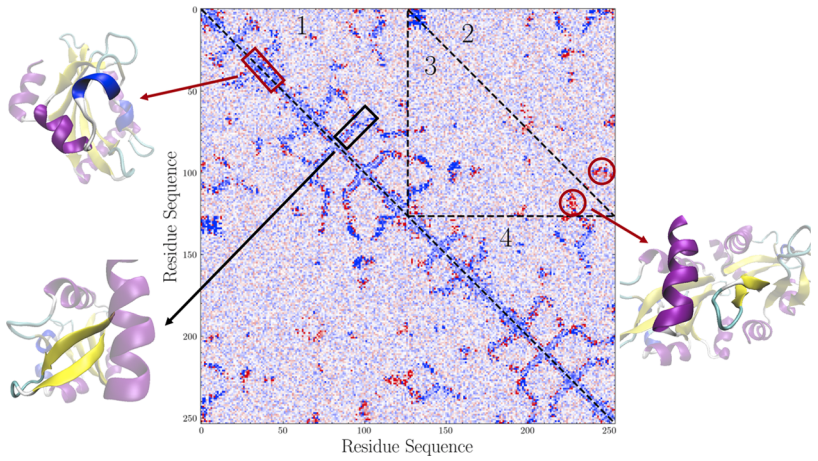


Figure 9. Protein contact map with the corresponding protein structures. Feature weights of the first dense layer in the supervised ivis dimensionality reduction method were extracted and were colored red (positive), white (close to zero), and blue (negative). The residue sequence starts from Ser240 in chain A and ends in Glu367 in chain B.

we found an asymmetrical behavior (red circle in Figure 9) in which the interaction between Jα in chain A and linkers in chain B is stronger than the interaction between Jα in chain B and linkers in chain A.

To examine the important residues identified in the protein contact map, for each Cα distance, the corresponding feature weight was accumulated to the two related residues. Therefore, the significance of residues and structures is quantified. The top 20 residues are listed in Table 1, with important residues that are experimentally identified[22,63−66] shown in bold font. The accumulated importance of the secondary structure is shown in Table 2, which shows that the A′α helix, the Jα helix, and protein linkers are important in AuLOV allostery.

ivis is More Computationally Efficient Than t-ICA and t-SNE. A key factor in comparing different dimensionality reduction methods is their computational cost, for it could be prohibitively expensive when dealing with a large-size and

Table 1. Top 20 Residues Identified in the Supervised Ivis Framework

residue	importance (%)	residue	importance (%)
ILE 242	1.12	PHE 241	1.10
LEU 245	1.07	**ALA 248**[a]	1.04
GLN 250	1.01	SER 314	1.00
GLN 246	0.99	**ASN 251**	0.98
THR 247	0.97	PRO 268	0.97
ASN 329	0.96	**GLN 350**	0.96
MET 313	0.95	ALA 244	0.94
PHE 331	0.93	ASN 311	0.93
SER 240	0.92	**GLN 365**	0.92
CYS 351	0.91	ALA 335	0.90

[a]Experimentally confirmed important residues are shown in bold font.

Table 2. Accumulated Feature Importance of Secondary Structures

secondary structure	importance (%)
A′α	13.17
Aβ	6.34
Bβ	2.36
Cα	8.50
Dα	4.14
Eα	1.40
Fα	5.50
Gβ	6.52
Hβ	8.67
Iβ	7.98
Jα	10.44
linkers	24.98

https://dx.doi.org/10.1021/acs.jcim.0c00485
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

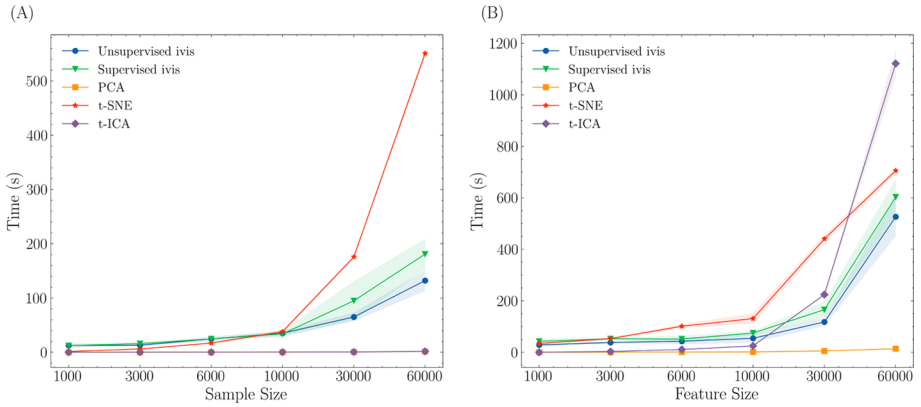


Figure 10. Computation time of each dimensionality reduction method spent on fitting high-dimensional data. (A) Runtime result of 1000 feature sizes with different sample sizes. Results of PCA and t-ICA were overlapped because of the time scale. (B) Runtime result of a sample size of 10 000 with different feature sizes.

high-dimensional data set. To compare the computational efficiency of different dimensionality reduction methods with regard to sample size and feature size, three randomly generated data sets with uniform distributions between 0 and 1 were applied for each data set size. The relation between runtime and sample size, with a feature size of 1000, is shown in Figure 10A. While t-SNE is stable and fast in small data sets (≤10 000 sample size), its runtime grows the fastest among the five models and is not feasible for a large data set. t-ICA and PCA overlapped with each other since these two models are less affected by the sample size. Unsupervised ivis and supervised ivis exhibited similar runtime results. The relation between runtime and feature size with a sample size of 10 000 is shown in Figure 10B. t-ICA and t-SNE show similar trends in the runtime growth trend, as they perform fast in small feature sizes (≤10 000), but they are not practical in higher dimensions. While both ivis models are slower than PCA, the runtimes of these two models are acceptable for a large sample size and high dimension. The training process of supervised ivis is further displayed in Figure 11. Triplet loss was stable after 4 epochs and stopped at 32 epochs, with early stopping of 10.

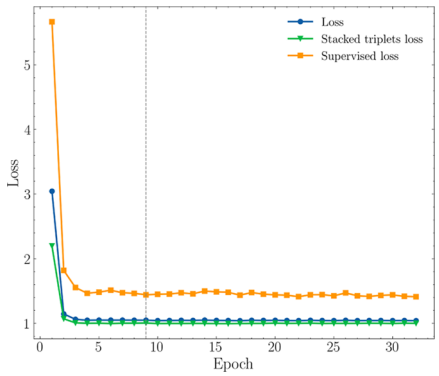


Figure 11. Triplet loss of each epoch for the supervised ivis framework with a supervision weight of 0.1 and early stopping of 10. The model is trained on a data set of 10 000 samples with 60 000 dimensions. The dashed gray line indicates the expected termination in training with early stopping of 5.

DISCUSSION

As a deep learning-based algorithm, the ivis framework was originally designed for single-cell experiments to provide a new approach for visualization and explanation purposes. In this study, ivis is applied on the MD simulations of allosteric protein AuLOV for dimensionality reduction. Combined with several performance criteria, ivis is demonstrated to be effective in keeping both local and global features while offering key insights for the mechanism of protein allostery.

Various dimensionality reduction methods have been used in protein systems, such as PCA, t-ICA, and t-SNE. As linear methods, PCA and t-ICA aim to capture the maximum variance and autocorrelation of protein motion, respectively. However, nonlinear dimensionality reduction methods, such as t-SNE, have been shown to be superior to linear methods in keeping the similarity between high dimensions and low dimensions.[41] Nevertheless, the limitations of t-SNE, such as susceptibility to system noise[67] and poor performance in extracting the global structure, hinder further interpretations for biological systems. Compared with these dimensionality reduction methods, ivis is outstanding in preserving distances in the low-dimensional spaces and could be utilized for biological system explanations.

Dimensionality reduction methods have different strengths in preserving structural information and can be applied to various data sets. While there is no universal standard measuring the performance of different methods, an appropriate method should reflect the distance and similarity between projections in a low-dimensional space. Similar structures in the high-dimensional space should be close in the low-dimensional space. This criterion is important in the construction of the Markov state model, which requires clustering discrete microstates on the projections. Improper projections would lead to poor MSMs, thus obscuring the protein motions and hindering further structure−function study.[68] Moreover, an adequate MSM requires the similarity between structures in each microstate. To evaluate the effectiveness of MSM, the average RMSD value is often used as a good indicator for dimensionality reduction methods. In this regard, both unsupervised ivis and supervised ivis are suitable to build MSM in a low-dimensional space. Estimated relaxation time scale reflects the number of steady states and is used to construct kinetically stable macrostates. The time scale of protein motion ranges from milliseconds to seconds in

https://dx.doi.org/10.1021/acs.jcim.0c00485
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

experiments. Among all of the tested dimensionality reduction methods, the ivis framework showed the longest time scale, with over 10^{-5} s. This experimentally meaningful time scale, combined with the average RMSDs, suggests the success of ivis in the construction of MSM.

It is expected that Euclidean distances between data points in the high-dimensional space should be proportional to the distances between the projected points in the low-dimensional space. In the current study, the long distance in the original dimensional space represents a high degree of dissimilarity in protein structure and the related two data points are more likely to be in different protein folding states. A well-behaved dimensionality reduction method should keep this correspondence in the low-dimensional space. Several assessments are applied to quantify this relationship. Spearman's rank-order correlation coefficient is calculated to test the linear relationship of pair-wised distances of data points. A potential problem is that distances are not independent. Rather, the change in position of one residue would lead to the change in the related $n - 1$ pair-wised distances. Therefore, to overcome this problem, the Mantel test is used to randomly permute rows and columns of distance matrix. The result of the Mantel test showed a similar trend as that in the Spearman correlation coefficient, which indicates that all methods are free from the dependency of distances and maintain good stability. The concept of the Shannon information in information theory is utilized to compare the information content in each projection space. The results of the above criteria show that ivis is capable of effectively separating different classes in the low-dimensional space and preserving high-dimensional distances with the least information loss. While the high-dimensional data set is usually projected onto a 2D surface, the effectiveness of MSMs on different dimensions was tested. Through the results of GMRQ, different methods showed various results. It is proposed that suitable dimensions are dependent on the biological system and dimensionality reduction method. However, two-dimensional space is still desired for visualization purposes if it can represent sufficient biological information.

The protein contact map demonstrates the superiority of the ivis dimensionality reduction method that ivis can retain both local and global information. Unexpectedly, the asymmetrical nature of the AuLOV dimer is revealed by comparing the protein−protein interactions. Several important residues are identified by the ivis framework. Met313, Leu331, and Cys351 have been reported as light-induced rotamers near cofactor FMN.[22] These key residues are located on the surface of the β-sheet, which is consistent with and proves the concept of the signaling mechanism that signals originating from the core of Per-ARNT-Sim (PAS) generate conformational change mainly within the β-sheet.[63,64] Gln365 is important for the stability of the Jα helix through hydrogen bonding with Cys316.[65] Leu248, Gln250, and Asn251 were also found to be important in modulating allostery within a single chain, reported as the A′α linker, while Asn329 and Gln350 function as FMN stabilizers.[66] Through the AuLOV dimerization, A′α and Jα helices undergo conformational changes and are expected to account for large importance, as shown in Table 2. However, the protein linkers, as well as the Cα helix and Hβ and Iβ strands, also showed high importance. The significance of protein linkers in the current study is consistent with both experimental and computational findings[69−72] that protein linkers are indispensable components in allostery and biological functions. Together, these unexpected structures are vital in AuLOV allostery and worth further study. Overall, several key residues and secondary structures identified by the ivis framework agree with the experimental finding, which consolidates the good performance of ivis in elucidating the protein allosteric process.

Computational cost should be considered when comparing dimensionality reduction methods, since it is computationally expensive for large data sets, especially for proteins. From this perspective, different models are benchmarked using a dummy data set. Results showed that PCA requires the least computational resource, not subjected to either sample size or feature size. This might be due to the reason that PCA implemented in Scikit-learn uses SVD for acceleration. Further, since the size of the data set was large, randomized truncated SVD was applied to reduce the time complexity to $O(n_{\max}^2 \cdot n_{\text{components}})$ with $n_{\max} = \max(n_{\text{samples}}, n_{\text{features}})$.[73] While t-SNE is comparable with ivis regarding several assessments, the computational cost could be prohibitively expensive for large data sets as t-SNE has a time complexity of,[74] where N and D are the number of samples and features, respectively. Though tree-based algorithms have been developed to reduce the complexity to,[75] it is still challenging for the high-dimensional protein system. ivis exhibited less computational cost in higher sample sizes and dimensions. Further, as shown in Figure 11, the loss of the ivis model converges fast and the overall computational cost could be further reduced with early stopping iterations. Combined with the performance criteria and runtime comparison, the ivis framework is demonstrated as a superior dimensionality reduction method for protein systems and can be an important member in the analysis toolbox for the MD trajectory.

CONCLUSIONS

As originally developed for single-cell technology, the ivis framework is applied in this study as a dimensionality reduction method for molecular dynamics simulations for biological macromolecules. ivis is superior to other dimensionality reduction methods in several aspects, ranging from preserving both local and global distances, maintaining similarity among data points in high-dimensional space and projections, to retaining the most structural information through a series of performance assessments. ivis also shows great potential in interpreting biological systems through the feature weights in the neural network layer. Overall, ivis reached a balance between dimensionality reduction performance and computational cost and is therefore promising as an effective tool for the analysis of macromolecular simulations.

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 75205, United States;* orcid.org/0000-0002-2488-0239; Email: ptao@smu.edu

Author

Hao Tian − *Department of Chemistry, Center for Research Computing, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas 75205, United States;* orcid.org/0000-0002-0186-9811

https://dx.doi.org/10.1021/acs.jcim.0c00485
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.0c00485

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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. Computational time was generously provided by Southern Methodist University's Center for Research Computing. The authors thank Xi Jiang from the Biostatistics Ph.D. program in the Statistics department of SMU for fruitful discussions.

■ REFERENCES

(1) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334*, 517−520.
(2) Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J. et al. In *Millisecond-Scale Molecular Dynamics Simulations on Anton*, Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, 2009; pp 1−11.
(3) 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.; et al. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **2013**, *9*, 461−469.
(4) Indyk, P.; Motwani, R. In *Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality*, Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, 1998; pp 604−613.
(5) Ichiye, T.; Karplus, M. Collective Motions in Proteins: A Covariance Analysis of Atomic Fluctuations in Molecular Dynamics and Normal Mode Simulations. *Proteins: Struct., Funct., Bioinf.* **1991**, *11*, 205−217.
(6) 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*, No. e1006801.
(7) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. Ten-Microsecond Molecular Dynamics Simulation of a Fast-Folding WW Domain. *Biophys. J.* **2008**, *94*, L75−L77.
(8) Roweis, S. T.; Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **2000**, *290*, 2323−2326.
(9) Tenenbaum, J. B.; de Silva, V.; Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **2000**, *290*, 2319−2323.
(10) Levy, R.; Srinivasan, A.; Olson, W.; McCammon, J. Quasi-Harmonic Method for Studying Very Low Frequency Modes in Proteins. *Biopolymers* **1984**, *23*, 1099−1112.
(11) 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*, No. 02B617.
(12) van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(13) Hinton, G. E.; Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504−507.
(14) 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*, 1−11.
(15) Zhao, X.; Kaufman, A. Multi-Dimensional Reduction and Transfer Function Design Using Parallel Coordinates. Volume graphics. *Vol. Graph.* **2010**, 69−76.
(16) Suárez, E.; Adelman, J. L.; Zuckerman, D. M. Accurate Estimation of Protein Folding and Unfolding Times: Beyond Markov State Models. *J. Chem. Theory Comput.* **2016**, *12*, 3473−3481.
(17) Doerr, S.; Ariz-Extreme, I.; Harvey, M. J.; De Fabritiis, G. Dimensionality Reduction Methods for Molecular Simulations, arXiv preprint arXiv:1710.10629. arXiv.org e-Print archive. https://arxiv.org/abs/1710.10629 (submitted Oct 29, 2017).
(18) Szubert, B.; Cole, J. E.; Monaco, C.; Drozdov, I. Structure-Preserving Visualisation of High Dimensional Single-Cell Datasets. *Sci. Rep.* **2019**, *9*, No. 8914.
(19) Koch, G.; Zemel, R.; Salakhutdinov, R. *Siamese Neural Networks for One-Shot Image Recognition* ICML Deep Learning Workshop, 2015.
(20) Hermans, A.; Beyer, L.; Leibe, B. In Defense of the Triplet Loss for Person Re-Identification, arXiv preprint arXiv:1703.07737. arXiv.org e-Print archive. https://arxiv.org/abs/1703.07737 (submitted Mar 22, 2017).
(21) Takahashi, F.; Yamagata, D.; Ishikawa, M.; Fukamatsu, Y.; Ogura, Y.; Kasahara, M.; Kiyosue, T.; Kikuyama, M.; Wada, M.; Kataoka, H. AUREOCHROME, A Photoreceptor Required for Photomorphogenesis in Stramenopiles. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 19625−19630.
(22) Heintz, U.; Schlichting, I. Blue Light-Induced LOV Domain Dimerization Enhances the Affinity of Aureochrome 1a for Its Target DNA Sequence. *eLife* **2016**, *5*, No. e11860.
(23) Losi, A.; Gärtner, W. Bacterial Bilin- and Flavin-Binding Photoreceptors. *Photochem. Photobiol. Sci.* **2008**, *7*, 1168−1178.
(24) Crosson, S.; Rajagopal, S.; Moffat, K. The LOV Domain Family: Photoresponsive Signaling Modules Coupled to Diverse Output Domains. *Biochemistry* **2003**, *42*, 2−10.
(25) Berman, H. M.; Bhat, T. N.; Bourne, P. E.; Feng, Z.; Gilliland, G.; Weissig, H.; Westbrook, J. The Protein Data Bank and the Challenge of Structural Genomics. *Nat. Struct. Biol.* **2000**, *7*, 957−959.
(26) Freddolino, P. L.; Gardner, K. H.; Schulten, K. Signaling Mechanisms of LOV Domains: New Insights From Molecular Dynamics Studies. *Photochem. Photobiol. Sci.* **2013**, *12*, 1158−1170.
(27) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(28) Buchenberg, S.; Knecht, V.; Walser, R.; Hamm, P.; Stock, G. Long-Range Conformational Transition of a Photoswitchable Allosteric Protein: Molecular Dynamics Simulation Study. *J. Phys. Chem. B* **2014**, *118*, 13468−13476.
(29) Tsuchiya, Y.; Taneishi, K.; Yonezawa, Y. Autoencoder-Based Detection of Dynamic Allostery Triggered by Ligand Binding Based on Molecular Dynamics. *J. Chem. Inf. Model.* **2019**, *59*, 4043−4051.
(30) Cerutti, D. S.; Duke, R.; Freddolino, P. L.; Fan, H.; Lybrand, T. P. A Vulnerability in Popular Molecular Dynamics Packages Concerning Langevin and Andersen Dynamics. *J. Chem. Theory Comput.* **2008**, *4*, 1669−1680.
(31) Ben-David, M.; Huang, H.; Sun, M. G.; Corbi-Verge, C.; Petsalaki, E.; Liu, K.; Gfeller, D.; Garg, P.; Tempel, W.; Sochirca, I.; et al. Allosteric Modulation of Binding Specificity by Alternative Packing of Protein Cores. *J. Mol. Biol.* **2019**, *431*, 336−350.
(32) Braun, E.; Gilmer, J.; Mayes, H. B.; Mobley, D. L.; Monroe, J. I.; Prasad, S.; Zuckerman, D. M. Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living J. Comput. Mol. Sci.* **2019**, *1*, No. 5957.
(33) 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*, 8577−8593.
(34) Eastman, P.; Pande, V. OpenMM: A Hardware-Independent Framework for Molecular Simulations. *Comput. Sci. Eng.* **2010**, *12*, 34−39.
(35) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.

(36) Foloppe, N.; MacKerell, A. D., Jr. All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data. *J. Comput. Chem.* **2000**, *21*, 86−104.
(37) 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**, DOI: 10.1080/07391102.2020.1802338.
(38) Nair, V.; Hinton, G. E. In Rectified Linear Units Improve Restricted Boltzmann Machines, Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010; pp 807−814.
(39) Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. In *Self-Normalizing Neural Networks*, Advances in Neural Information Processing Systems, 2017; pp 971−980.
(40) Golub, G. H.; Reinsch, C. *Linear Algebra*; Springer, 1971; pp 134−151.
(41) 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*, 5499−5510.
(42) Ulyanov, D. Multicore-TSNE, 2016. https://github.com/DmitryUlyanov/Multicore-TSNE (accessed March 17, 2020).
(43) Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. *Noise Reduction in Speech Processing*; Springer, 2009; pp 1−4.
(44) Adler, J.; Parmryd, I. Quantifying Colocalization by Correlation: The Pearson Correlation Coefficient Is Superior to the Mander's Overlap Coefficient. *Cytometry, Part A* **2010**, *77*, 733−742.
(45) Diniz-Filho, J. A. F.; Soares, T. N.; Lima, J. S.; Dobrovolski, R.; Landeiro, V. L.; Telles, M. P. d. C.; Rangel, T. F.; Bini, L. M. Mantel Test in Population Genetics. *Genet. Mol. Biol.* **2013**, *36*, 475−485.
(46) Carr, J. W. MantelTest, 2013. https://github.com/jwcarr/MantelTest (accessed April 5, 2020).
(47) McGibbon, R. T.; Schwantes, C. R.; Pande, V. S. Statistical Model Selection for Markov Models of Biomolecular Dynamics. *J. Phys. Chem. B* **2014**, *118*, 6475−6481.
(48) 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*, 10−15.
(49) McGibbon, R. T.; Pande, V. S. Variational Cross-Validation of Slow Dynamical Modes in Molecular Kinetics. *J. Chem. Phys.* **2015**, *142*, No. 124105.
(50) Liaw, A.; Wiener, M.; et al. Classification and Regression by randomForest. *R News* **2002**, *2*, 18−22.
(51) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(52) Hecht-Nielsen, R. *Neural Networks for Perception*; Elsevier, 1992; pp 65−93.
(53) Kingma, D. P.; Ba, J. A Method for Stochastic Optimization, arXiv preprint arXiv:1412.6980. arXiv.org e-Print archive. https://arxiv.org/abs/1412.6980 (accessed Dec 22, 2014).
(54) Chollet, F. et al. Keras, 2015. https://keras.io (accessed April 1, 2020).
(55) 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*, 99−105.
(56) 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*, No. 124101.
(57) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. Automatic Discovery of Metastable States for the Construction of Markov Models of Macromolecular Conformational Dynamics. *J. Chem. Phys.* **2007**, *126*, No. 155101.
(58) Schubert, E.; Gertz, M. In *Intrinsic t-Stochastic Neighbor Embedding for Visualization and Outlier Detection*, International Conference on Similarity Search and Applications, 2017; pp 188−203.
(59) Zhou, Y.; Sharpee, T. Using Global t-SNE to Preserve Inter-Cluster Data Structure. *bioRxiv* **2018**, No. 331611.
(60) Wang, F.; Zhou, H.; Wang, X.; Tao, P. Dynamical Behavior of β-Lactamases and Penicillin-Binding Proteins in Different Functional States and Its Potential Role in Evolution. *Entropy* **2019**, *21*, No. 1130.
(61) 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*, No. 47.
(62) Romano, P.; Guenza, M. GRadient Adaptive Decomposition (GRAD) Method: Optimized Refinement Along Macrostate Borders in Markov State Models. *J. Chem. Inf. Model.* **2017**, *57*, 2729−2740.
(63) 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*, 1054−1057.
(64) Möglich, A.; Ayers, R. A.; Moffat, K. Structure and Signaling Mechanism of Per-ARNT-Sim Domains. *Structure* **2009**, *17*, 1282−1294.
(65) Herman, E.; Sachse, M.; Kroth, P. G.; Kottke, T. Blue-Light-Induced Unfolding of the Jα Helix Allows for the Dimerization of Aureochrome-LOV From the Diatom *Phaeodactylum tricornutum*. *Biochemistry* **2013**, *52*, 3094−3101.
(66) Arinkin, V.; Granzin, J.; Röllen, K.; Krauss, U.; Jaeger, K.-E.; Willbold, D.; Batra-Safferling, R. Structure of a LOV Protein in Apo-State and Implications for Construction of LOV-Based Optical Tools. *Sci. Rep.* **2017**, *7*, No. 42971.
(67) Amid, E.; Warmuth, M. K. A More Globally Accurate Dimensionality Reduction Method Using Triplets, arXiv preprint arXiv:1803.00854. arXiv.org e-Print archive. https://arxiv.org/abs/1803.00854 (submitted March 1, 2018).
(68) 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*, No. 174105.
(69) Ma, B.; Tsai, C.-J.; Haliloğlu, T.; Nussinov, R. Dynamic Allostery: Linkers Are Not Merely Flexible. *Structure* **2011**, *19*, 907−917.
(70) Reddy Chichili, V. P.; Kumar, V.; Sivaraman, J. Linkers in the Structural Biology of Protein-Protein Interactions. *Protein Sci.* **2013**, *22*, 153−167.
(71) George, R. A.; Heringa, J. An Analysis of Protein Domain Linkers: Their Classification and Role in Protein Folding. *Protein Eng., Des. Sel.* **2002**, *15*, 871−879.
(72) Gokhale, R. S.; Khosla, C. Role of Linkers in Communication Between Protein Modules. *Curr. Opin. Chem. Biol.* **2000**, *4*, 22−27.
(73) Halko, N.; Martinsson, P.-G.; Tropp, J. A. Finding Structure with Randomness: Stochastic Algorithms for Constructing Approximate Matrix Decompositions, arXiv:0909.4061. arXiv.org e-Print archive. https://arxiv.org/abs/0909.4061 (submitted on Sept 22, 2009).
(74) Ding, J.; Condon, A.; Shah, S. P. Interpretable Dimensionality Reduction of Single Cell Transcriptome Data with Deep Generative Models. *Nat. Commun.* **2018**, *9*, No. 2002.
(75) Van Der Maaten, L. Accelerating t-SNE Using Tree-Based Algorithms. *J. Mach. Learn. Res.* **2014**, *15*, 3221−3245.

https://dx.doi.org/10.1021/acs.jcim.0c00485
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX