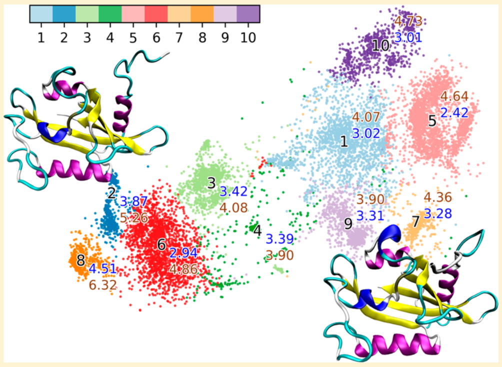
JCTC Journal of Chemical Theory and Computation
Article
Cite This: J. Chem. Theory Comput. 2018, 14, 5499−5510
pubs.acs.org/JCTC

t-Distributed Stochastic Neighbor Embedding Method with the Least Information Loss for Macromolecular Simulations

Hongyu Zhou, Feng Wang, and Peng Tao*

Department of Chemistry, Center for Scientific Computation, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas 75275, United States

ABSTRACT: Dimensionality reduction methods are usually applied on molecular dynamics simulations of macromolecules for analysis and visualization purposes. It is normally desired that suitable dimensionality reduction methods could clearly distinguish functionally important states with different conformations for the systems of interest. However, common dimensionality reduction methods for macromolecules simulations, including predefined order parameters and collective variables (CVs), principal component analysis (PCA), and time-structure based independent component analysis (t-ICA), only have limited success due to significant key structural information loss. Here, we introduced the t-distributed stochastic neighbor embedding (t-SNE) method as a dimensionality reduction method with minimum structural information loss widely used in bioinformatics for analyses of macromolecules, especially biomacromolecules simulations. It is demonstrated that both one-dimensional (1D) and two-dimensional (2D) models of the t-SNE method are superior to distinguish important functional states of a model allosteric protein system for free energy and mechanistic analysis. Projections of the model protein simulations onto 1D and 2D t-SNE surfaces provide both clear visual cues and quantitative information, which is not readily available using other methods, regarding the transition mechanism between two important functional states of this protein.



■ INTRODUCTION

Molecular dynamics (MD) simulations have been widely applied on macromolecules, especially biomacromolecules to provide atomistic insights into their structure−function relations.[1] Those insights are unattainable by most experimental approaches. Recently, with the significant improvement of computational powers due to graphical processing units (GPUs), the simulated time scale for all-atom MD simulations has been extended from nanoseconds to milliseconds scales.[2,3] Long-time MD simulations can provide meaningful predictions and insights into the mechanism of protein functions, because the slow time scale motions in dynamics are critical for protein functions.[4] However, biomacromolecules including proteins normally have hundreds to thousands of degrees of freedom. The curse of dimensionality[5] induces the difficulties for many analyses of long-time MD simulations, including extracting the important essential motions,[6] clustering different states based on kinetics or structures,[7,8] visualization of the free energy landscape, etc.[9,10] These analyses could retrieve the important information from the simulation data and provide insights into the protein function-related dynamics. Therefore, an effective low-dimensional description of MD simulations could be beneficial in many cases.

Geometrically, appropriate low-dimensional descriptors could be developed based on the assumption that the dynamics of protein in a long time scale simulation can be modeled by several slow modes.[11] Some theoretical studies support this assumption with regard to protein dynamics, which can be modeled by Markov state models (MSMs) based on their *Markovian* property.[8,12] In many cases, describing important dynamics using several predefined collective variables (CVs) is an efficient approach.[13] Those CVs are also referred to as the reaction coordinates for rare events including chemical reactions. However, defining the CVs to quantify protein dynamics is more complicated than chemical reactions.[13] Compared with small molecules, proteins have higher dimensionality, and inappropriate CVs could disguise protein kinetic barriers.[14] Valid CVs should be suitable to capture key dynamical events in simulations in order to obtain meaningful insight. Natural contacts,[15] root-mean-square deviations (RMSDs), radius of gyration (Rg),[16] and structural reaction coordinates including *P* and *Q* values[17] are all possible CVs and suitable to describe the protein dynamics from different perspectives.

Some dimensionality reduction methods can be applied on an ensemble of configurations to obtain appropriate low-dimensional descriptors for key protein dynamics. Without predefined CVs, these methods reconstruct the coordinates based on the geometrical high-dimensional properties of the system[18] and are normally categorized as either linear or nonlinear.[19,20] The coordinates employed in the linear dimensionality reduction methods are linear combinations of

Received: June 27, 2018
Published: September 25, 2018

DOI: 10.1021/acs.jctc.8b00652

input variables, including principal component analysis (PCA), also referred to as quasi-harmonic analysis in MD simulations,[21] and time-structure based independent component analysis (t-ICA).[4] Nonlinear dimensionality reduction methods construct coordinates as a nonlinear function of the input variables, including diffusion map,[22] isomap,[20] autoencode neural networks,[23] etc. A thorough comparison for diffusion map,[22] isomap,[20] the locally linear embedding (LLE) method,[19] and PCA was reported in a previous study.[24] In general, nonlinear dimensionality reduction methods are more suitable than the linear ones for systems with dynamics lying on highly curved and convoluted manifolds.[25]

These methods can be applied on biomacromolecules to obtain suitable descriptors for further analyses including free energy surface plotting. However, structural information loss is inevitable in dimensionality reduction processes. Different methods preserve different structural information through projections. For example, PCA can maximize the variance for each component, and t-ICA maximizes the time-lagged auto correlation for a given lag time.[26] Previous studies suggest that t-ICA has better performance than PCA for extracting the slowest dynamical modes.[27,28] However, because these methods are not designed to maintain the similarity between high-dimensional data and low-dimensional descriptors, the clusters of high-dimensional structures are usually not well characterized by the projected representations. For example, the *k*-means clustering method[29] using Cartesian coordinates could overlap significantly in PCA projection surface. In addition, projection onto low-dimensional surfaces could lead to inappropriate clustering analysis of simulation data, because inadequate projections could hide the important kinetic barriers, and result in incorrect thermodynamics calculations.[30]

One state-of-the-art method to reduce the dimensionality while maintaining the similarity between low-dimensional descriptors and high-dimensional data is the t-distributed Stochastic Neighbor Embedding (t-SNE) method.[31] In the t-SNE method, Gaussian probability distributions over high-dimensional space are constructed and used to optimize a Student t-distribution in low-dimensional space. The low-dimensional embedding descriptors can be obtained by minimizing the Kullback−Leibler divergence[32] between the distributions on high- and low-dimensional spaces using a gradient descent algorithm. In the t-SNE method, the low-dimensional space maintains the pair-wised similarity to the high-dimensional space, leading to a clustering on the embedding space close to the clustering in the high-dimensional space without losing significant structural information. This method has been widely applied in bioinformatics,[33] such as gene expression analysis,[34] single-cell visualization,[35] and cell types detections.[36]

With some promising development,[37,38] the t-SNE method could be applied to investigate protein dynamics and clustering protein structures and visualize free energy surfaces. In this study, the t-SNE method is demonstrated as an excellent dimensionality reduction algorithm for protein simulations and should be applicable to other biomacromolecules in general. Vivid (VVD) is a photosensitive circadian clock protein belonging to the Light-Oxygen-Voltage (LOV) domain.[39] Upon blue light activation, a covalent bond is formed between residue Cys108 and the cofactor flavin adenine dinucleotide (FAD) in VVD and leads to two distinct states (referred to as dark and light states) with significant conformational changes mainly involving its N-terminus.[39,40] Up to now, the mechanism, in which the formation of the above covalent bond leads to global conformational change in VVD, is still elusive. Following population shift hypothesis,[41,42] the t-SNE method is applied to construct low-dimensional descriptors to faithfully represent the free energy landscape of VVD related to the switching between the dark and light states. Combining with the clustering analysis and the time-resolved fitting analysis, the dynamics of trajectories can be tracked on the t-SNE projection surface. In this study, we demonstrate t-SNE as a superior dimensionality reduction method for MD simulation analysis through comparison with other methods. The exceptional performance of the t-SNE method validates it as a faithful method to probe the free energy landscape correlated to protein functions.

■ METHODOLOGY

Molecular Dynamics Simulation. The initial structures of the dark and light states of VVD were obtained from the Protein Data Bank (PDB)[43] with the IDs as 2PD7 and 3RH8, respectively. The dark and light state crystal structure sequences start from Met36 and His37, respectively. For consistency, residue 36 in the dark state was removed to maintain the same number of residues in both states. Both structures include a flavin adenine dinucleotide (FAD) as ligand. FAD and flavin mononucleotide (FMN) are two types of cofactors commonly existing in the LOV domain. Because FMN and FAD carry similar biological roles, the adenosine monophosphate (AMP) moiety was removed from the FAD in VVD crystal structures to form an FMN. An FMN force field from a previous study was used for the simulations carried out in this study.[44] A total of four simulation configurations were constructed, including dark state conformation with or without the photoinduced covalent bond and light state conformation with or without the photoinduced covalent bond. The VVD-FMN complex in each configuration was solvated using explicit water model (TIP3P)[45] and neutralized with sodium cations and chloride anions. Initially, 10 ns of isothermal−isobaric ensemble (NPT) MD simulations were carried out for each configuration. Subsequently, three independent 1.1 μs of canonical ensemble (NVT) Langevin MD simulations using different random seeds at 300 K were conducted for each configuration. The first 100 ns simulations were discarded as equilibrium, and the following 1 μs simulation was used for the dimensionality reduction analysis. These led to a total of 12 μs simulations of VVD for the analysis. For all simulations, the SHAKE method was used to constrain all bonds associated with hydrogen atoms. A step size of 2 fs was used, and simulation trajectories were saved every 1 ns. Cubic simulation box and periodic boundary conditions were applied for all MD simulations. Electrostatic interactions were calculated using the particle mesh Ewald (PME) method.[46] All simulations were carried out using the CHARMM[47] simulation package version 41b1 with the support of GPU calculations based on OpenMM.[48]

Relaxation Time Scale. MSMBuilder[12] was used to build the Markov state model (MSM) and estimate relaxation time scale. To apply MSM, the microstates are clustered for different description surfaces using the *k*-means clustering method, and the transition probability matrix was estimated among different states. Consequently, the eigenvalues and eigenvectors are calculated for the transition probability matrix. According to the Frobenius theorem,[49] for the stochastic transition probability matrix, the first eigenvalue is 1.0, and all other eigenvalues are less than 1.0. The relaxation time scale is estimated based on the second eigenvalue as the following equation

$$t(\tau) = -\frac{\tau}{\ln \lambda_1} \tag{1}$$

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

Root-Mean-Square Deviation (RMSD). The conformational change during the MD simulations can be measured by RMSD with regards to a reference structure. For a molecular structure represented by Cartesian coordinates, the RMSD is defined as the following:

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

The Cartesian coordinate vector r_i^0 is the i^{th} atom in the reference structure. For each simulation, the RMSD values with reference to the dark and light state crystal structures were calculated to quantity the sampling following a previous study.[50]

Principal Component Analysis (PCA). The normal modes for principal component analysis are extracted from a trajectory by diagonalizing the correlation matrix of the atomic positions. The correlation matrix is a measure of the Pearson correlated value of a set of atoms. Each matrix element is defined as

$$\mathrm{C}_{\mathrm{ij}} = \frac{\mathrm{c}_{\mathrm{ij}}}{\mathrm{c}_{\mathrm{ii}}^{1/2}\mathrm{c}_{\mathrm{jj}}^{1/2}} = \frac{\langle \mathrm{r}_\mathrm{i}\mathrm{r}_\mathrm{j}\rangle - \langle \mathrm{r}_\mathrm{i}\rangle\langle \mathrm{r}_\mathrm{j}\rangle}{[(\langle \mathrm{r}_\mathrm{i}^2\rangle - \langle \mathrm{r}_\mathrm{i}\rangle^2)(\langle \mathrm{r}_\mathrm{j}^2\rangle - \langle \mathrm{r}_\mathrm{j}\rangle^{1/2})]} \tag{3}$$

where C_{ij} is the measure of correlated movement between atoms i and j, c_{ij}, c_{ii}, and c_{jj} are the correlation matrix elements, and r_i and r_j are Cartesian coordinate vectors from the least-squares fitted structures with translational and rotational motions being projected out. Matrix elements are between −1 and 1 with negative values indicating negative correlation and positive values indicating positive correlation between the motions of atoms i and j.

Time-Structure Based Independent Component Analysis (t-ICA). The t-ICA method was developed to identify the slowest dynamics in the simulation with the maximum autocorrelation value. For an n-dimensional time series $\boldsymbol{x}(t) = {}^t(x_1(t),\ldots,x_n(t))$, t-ICA is performed by solving the following generalized eigenvalue problem

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

where $\mathbf{K}$ and $\mathbf{F}$ are the eigenvalue and eigenvector matrices, respectively. $\mathbf{C}$ is the covariance matrix, and $\bar{\mathbf{C}}$ is the time-lagged covariance matrix at time τ, which are defined as

$$\mathbf{C} = \langle(\boldsymbol{x}(t) - \langle\boldsymbol{x}(t)\rangle)^t(\boldsymbol{x}(t) - \langle\boldsymbol{x}(t)\rangle)\rangle \tag{5}$$

$$\bar{\mathbf{C}} = \langle(\boldsymbol{x}(t) - \langle\boldsymbol{x}(t)\rangle)^t(\boldsymbol{x}(t+\tau) - \langle\boldsymbol{x}(t)\rangle)\rangle \tag{6}$$

The independent component vectors obtained from t-ICA are uncorrelated and have the maximum autocorrelation value at a given time.

t-Distributed Stochastic Neighbor Embedding (t-SNE) Method. The t-SNE method is a nonlinear dimensionality reduction method, particularly well-suited for projecting high-dimensional data onto low-dimensional space for analysis and visualization purposes. Distinguished from other dimensionality reduction methods, the t-SNE method was designed to project high-dimensional data onto low-dimensional space with minimum structural information loss, so that the points close to each other on the low-dimensional surface represent states that are similar in the high-dimensional space.

Following the original article,[31] the t-SNE method is briefly described here. This method starts with converting the high-dimensional Euclidean distance between data points (the Cartesian coordinates of each frame in the simulation) into the conditional probability $p_{j|i}$. Given x_i and x_j as two data points representing two structures in Cartesian coordinates, the probability density distribution of its neighboring data points for x_i is assumed as a Gaussian function centered at x_i with variance σ_i. The probability of x_j to be selected as the neighbor of x_i is a conditional probability calculated as

$$p_{j|i} = \frac{\exp(-\| x_i - x_j \|^2/2{\sigma_i}^2)}{\sum_{k\neq i} \exp(-\| x_i - x_k \|^2/2{\sigma_i}^2)} \tag{7}$$

The above conditional probability is a nonsymmetric measurement as $p_{i|j}$ and $p_{j|i}$ are usually different. Therefore, the similarity of data points x_i and x_j is calculated as the joint probability defined as p_{ij}

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{8}$$

In low-dimensional space, the joint probability describing similarity is computed for y_i and y_j as the counterparts of the high-dimensional structures x_i and x_j. In the t-SNE method, Student's t-distribution with one degree of freedom is employed to calculate the joint probability between y_i and y_j, with q_{ij} as

$$q_{ij} = \frac{(1 + \| y_i - y_j \|^2)^{-1}}{\sum_{k\neq i} (1 + \| y_i - y_k \|^2)^{-1}} \tag{9}$$

If the map points y_i and y_j correctly model the similarity between the high-dimensional data points x_i and x_j, the joint probability q_{ij} should be close to p_{ij}. Therefore, the aim for the t-SNE method is to find a low-dimensional representation that minimizes the difference between q_{ij} and p_{ij} for all data points i and j.

One way to compare the differences between high-dimensional data and low-dimensional representations is using the Kullback−Leibler (KL) divergence over all data points to construct the cost functions C to evaluate the projection from high-dimensional structure (P) to low-dimensional representation (Q) as

$$C = KL(P\|Q) = \sum_{i\neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{10}$$

The cost function C could be minimized using the gradient descent method.

One remaining parameter to be selected is the bandwidth of Gaussian distribution σ_i that is centered over each high-dimensional data x_i. Because the density of high-dimensional data varies for different points in most cases, it is unlikely that a single value of σ_i could be used for all data points. A binary search of σ_i is carried out for each data point to match a fixed hyperparameter "perplexity" that is specified by users. This perplexity is defined as

$$\mathrm{Perp}(P_i) = 2^{-\sum_j p_{ij} \log_2 p_{ij}} \tag{11}$$

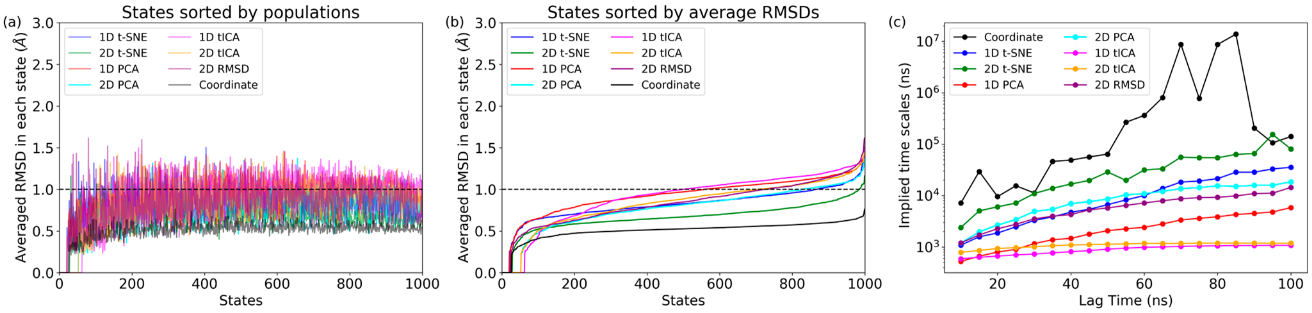


Figure 1. Comparison of several dimensionality reduction methods: (a) averaged RMSDs of the microstates clustered using different methods sorted by population of each microstate; (b) averaged RMSDs of the microstates clustered using different methods sorted by the averaged RMSDs; (c) estimated time scale for different MSMs constructed based on different methods using different lag times (ranging between 5 ns and 100 ns).

For each data point x_i, σ_i is optimized until the perplexity matches the value specified by users. Usually, a larger data set requires a larger perplexity value. The performance of t-SNE is fairly robust with sufficiently large hyperparameters.[31]

Besides the searching for the bandwidth of Gaussian distributions, the perplexity is also used to determine the number of nearest neighbors for a particular data point x_i using a tree-based Barnes-Hut implementation of the t-SNE method.[51] The most time-consuming step in the t-SNE method is the calculation of joint probability for each pair structure. For large data sets, the computational cost for this step may become prohibitively expensive. In the Barnes-Hut implementation of the t-SNE method, given a perplexity value μ as an integer number, only the 3 μ nearest neighbors for each data point x_i are considered. For the structures not belonging to the 3 μ nearest neighbors of x_i, the joint probability was treated as zero. Through this approximation, the computational cost of t-SNE is significantly reduced with moderate decreasing of the performance. In this study, to evaluate the best performance of t-SNE, the perplexity is specified as $N/3$ to ensure that the joint probability of all data points with regard to each x_i is calculated. The Scikit-learn package[52] with t-SNE implementation is employed in this study to carry out all the calculations.

■ RESULTS

Initial Comparison of t-SNE with Other Methods. A total of eight representations using different dimensionality reduction methods are applied on the model system for comparison purposes: one-dimensional (1D) and two-dimensional (2D) models for t-SNE, PCA, and t-ICA methods, respectively, as well as 2D RMSD and full Cartesian coordinates. The *k*-means clustering method was used to divide a total of 12 μs VVD simulations into 1,000 microstates in each representation only using the collective variables or order parameters associated with that representation. For microstates in each representation, an averaged RMSD value is calculated by averaging all pairwise RMSD values among all structures within each specific microstate using Cartesian coordinates. This averaged RMSD value measures the structure similarity for each microstate. In general, smaller averaged RMSDs represent better clustering results. The averaged RMSDs values are plotted for each representation in the order of decreasing cluster size in Figure 1a. For comparison purposes, all averaged RMSDs are sorted and plotted in Figure 1b.

An appropriate discretization should have smaller averaged RMSDs overall, warranting better structural similarity and kinetic accessibility inside each microstate. It was suggested that an adequate microstate should have an averaged RMSD lower than 1.0 Å.[53,54] Large averaged RMSD values of microstates may lead to inadequate MSMs. As shown in Figure 1b, the clustering in the Cartesian space has the best performance due to the least structural information loss. The result shows that every microstate is clustered with the averaged RMSD significantly lower than 1.0 Å using Cartesian coordinates. However, Cartesian coordinates are not suitable for any further analysis, especially as reaction coordinates to construct free energy surface.

After dimensionality reduction processes, some structural information will be inevitably lost, and some kinetic barriers are obscured during the projection. One criterion to assess information loss is measuring the similarity with the Cartesian coordinates results. With this regard, the 2D t-SNE model is closer to the Cartesian coordinates clustering result (Figure 1b) than all other models. Therefore, it is the best dimensionality reduction model presented in this study. Surprisingly, the 1D t-SNE model is significantly better than the remaining models and comparable with the 2D PCA model. This demonstrates that the t-SNE method is intrinsically better than many other dimensionality reduction methods with minimum information loss. 1D t-ICA and 1D PCA are the least effective methods presented in this study (Figure 1a). Overall, the performance of each dimensionality reduction method by comparing the structure similarity in each microstate is ranked as Cartesian Space (benchmark) > 2D t-SNE > 1D t-SNE ≈ 2D PCA > 2D t-ICA ≈ 2D RMSD > 1D PCA ≈ 1D t-ICA.

In addition to the averaged RMSDs, another metric to compare different dimensionality reduction methods is the approximate relaxation time scale estimated using MSM.[55] Based on the 1,000 microstates of each representation, the relaxation time scale can be estimated from the eigenvalue of the transition probability matrix among these microstates. The relaxation time scale is an approximate time length needed for any system to reach its steady state. Experimental studies suggest that the time for conformational changes could take up to hundreds of milliseconds.[56] Applying different lag times, the relaxation time scale can be estimated based on the transition probabilities among microstates. The relaxation time scale estimated based on lag times ranging from 5 to 100 ns is shown in Figure 1c. With the smallest structural information loss among all representations, the clustering analysis using Cartesian coordinates is expected to be the closest to the experimental relaxation time scale. Using dimensionality reduction, structural information loss may lead to inadequate

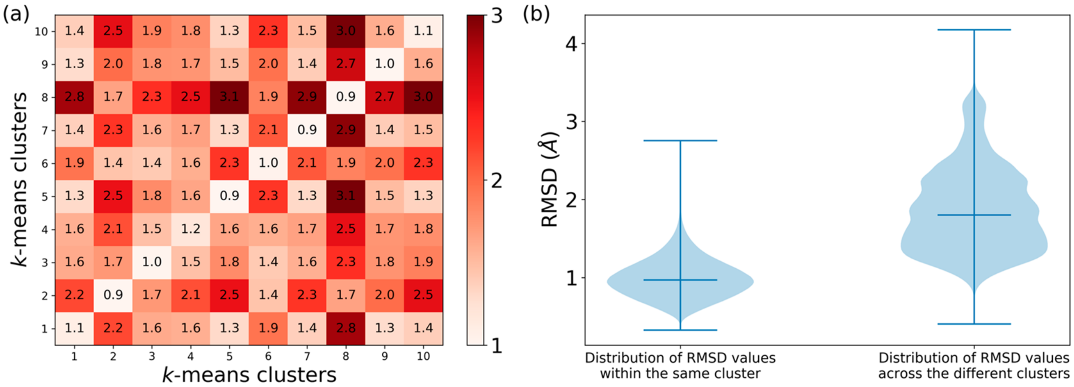


Figure 2. Ten clusters obtained from *k*-means clustering based on Cartesian coordinates: (a) averaged RMSD value for all structure pairs from cluster pair; (b) distribution of RMSD values based on structure pairs either within the same cluster (plot on left-hand side) or across the different clusters (plot on right-hand side). The middle horizontal line in (b) is the averaged RMSD value of each distribution.

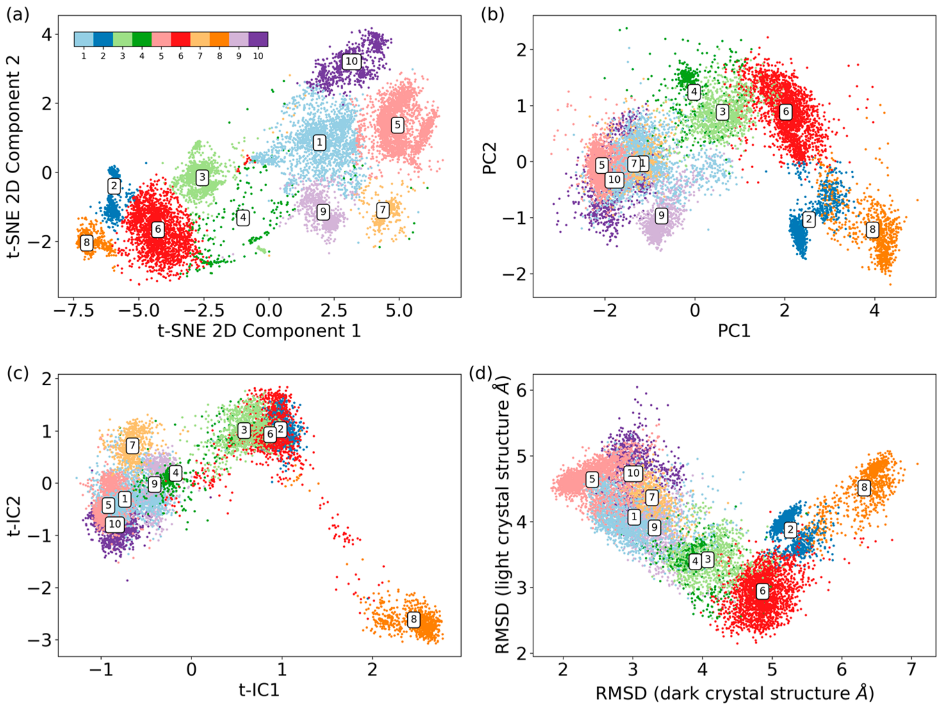


Figure 3. Ten *k*-means clusters of VVD systems using Cartesian coordinates represented by different dimensionality reduction methods: (a) 2D t-SNE method; (b) 2D PCA method; (c) 2D t-ICA method; (d) 2D RMSD values with reference to the dark and light state crystal structures, respectively.

clustering of microstates and neglecting of some kinetic barriers within the microstates due to the assumption that the kinetic barriers among different conformers are negligible within each microstate. These potentially neglected kinetic barriers could lead to an inaccurate time frame required to reach the steady state as an estimated relaxation time scale. This may be the reason that the t-ICA method results in a significantly lower relaxation time scale than other methods. Overall, the estimated time scale using 2D t-SNE is the closest to the one based on Cartesian coordinates, suggesting that the estimation of overall kinetic barrier among microstates generated by 2D t-SNE is the closest to the real relaxation time scale.

Representation of High-Dimensional *k*-Means Clusters. The above analyses demonstrate the effectiveness of the t-SNE method for dividing the structures from simulations into microstates. Compared with PCA and t-ICA, the t-SNE method is better at preserving the kinetic barriers and is the closest to the results using full Cartesian coordinates, showing minimum structural information loss. In this section, the t-SNE method is further tested to distinguish the high-dimensional clusters and construct the free energy surface.

Here, we redo the clustering analysis aiming for a smaller number of clusters using the *k*-means method and Cartesian coordinates for better representation. The averaged RMSD for all pairwise structures belonging to the same cluster is applied as the validation metric for the clustering analysis quality as

$$\mathrm{RMSD}_{\mathrm{same}} = \mathrm{Mean}(\mathrm{RMSD}_{i=1...N, j=1...N}\ \forall\ i, j \in C_{m, m=1...M}) \tag{12}$$

where i and j are indices for all N data points, and C_m represents any of the M clusters. Similarly, the structure dissimilarity among different clusters is represented as

DOI: 10.1021/acs.jctc.8b00652

$$\mathrm{RMSD}_{\mathrm{diff}} = \mathrm{Mean}(\mathrm{RMSD}_{i=1\ldots N, j=1\ldots N}\ \forall\ i \in C_{m,m=1\ldots M} \text{ and } j \in C_{n,n=1\ldots M} \text{ and } m \neq n) \tag{13}$$

The number of clusters is chosen as ten, because it is the smallest number of clusters to achieve $\mathrm{RMSD}_{\mathrm{same}}$ less than 1.0 Å (0.990 Å as the middle horizontal line in the left-hand side plot of Figure 2b). The populations for those clusters are 16.6%, 7.4%, 8.5%, 3.6%, 19.6%, 17.8%, 4.8%, 6.4%, 7.5%, and 7.8%, respectively. The RMSDs for all cluster pairs for these ten states are shown in Figure 2a. It is clear that $\mathrm{RMSD}_{\mathrm{diff}}$ (1.869 Å) is significantly larger than $\mathrm{RMSD}_{\mathrm{same}}$ (0.990 Å) as illustrated in Figure 2b. This indicates that the structural similarities inside the clusters are much higher than that between different clusters. Adequate low-dimensional descriptors should be able to project these ten Cartesian coordinates based clusters onto the free energy surface with clear distinguishability. If a low-dimensional free energy surface does not distinguish these clusters clearly, some kinetic barriers among these clusters could be significantly obscured.

The above ten clusters are plotted using different collective variables shown in Figure 3. The 2D t-SNE model has remarkable results in distinguishing different states as free energy basins. All ten clusters are well separated from each other and depicted in different colors (Figure 3a). Compared to our previous study of VVD,[50] distribution of these ten states on a 2D RMSD plot shows the similarities of each state to the dark and light state crystal structures (Figure 3d). Cluster 8 could be considered as a hidden state, which is different from both the dark and light state crystal structures. Clusters 2, 6, 3, and 4 could be grouped as the light region as these states are close to the light state crystal structure. Clusters 1, 5, 7, 9, and 10 could be grouped as the dark region as these states are close to the dark state crystal structure. Overall, the hidden state and the states in the light region are well-separated on the 2D RMSD surface and PCA surface. However, the states in the dark region significantly overlap with each other when projected onto these surfaces (Figures 3b and 3d). The t-ICA model captures the slowest dynamics in the simulations and results in a clear separation of the hidden state and the dark and the light regions (Figure 3c). However, in the t-ICA model, the dark clusters 1, 5, 10, 7, and 9 significantly overlap with each other, as well as for the light clusters 2, 3, and 6 (Figure 3c). These results demonstrate that the t-SNE method offers superior performance in representing the free energy surface and interrogating the differences among high-dimensional clusters compared to the PCA, t-ICA, and RMSD models.

Because the different free energy basins are clearly separated in the 2D t-SNE projections, the free energy surface using the two t-SNE collective variables generated in the 2D t-SNE model is constructed (Figure 4). Each state is clearly represented by separate minimum energy basins, suggesting that the t-SNE collective variables could represent the high-dimensional distribution with minimum information loss. It is worth pointing out that this free energy surface does not distinguish between nonbonded and bonded configurations, which will be elaborated below.

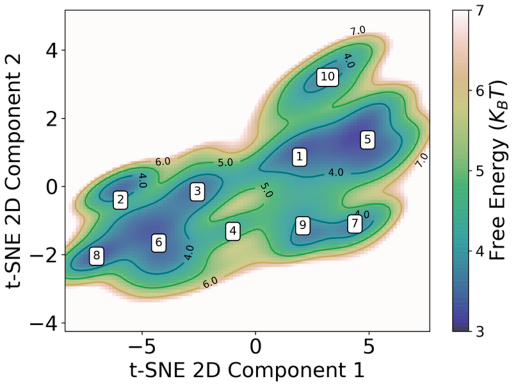


Figure 4. Free energy landscape representation of 2D t-SNE projections.

Conformational Changes Revealed by t-SNE. In the representations generated using the t-SNE method, data points that are distinct from each other are separated by large pairwise distances, and data points that are similar to each other are separated by small pairwise distances.[30] It was noted that smaller pairwise distances are more faithful to represent similar data points than large pairwise distances to represent distinct data points in t-SNE representations.[30] In other words, if two points on a low-dimensional surface generated using the t-SNE method are very close to each other, they are likely very similar to each other in the original high-dimensional space. However, if two points are far away from each other on a low-dimensional t-SNE surface, the distance between them may not accurately represent how different they are in the high-dimensional space. This is due to a general issue of dimensionality reduction that the "global structures" of data are difficult to be preserved. To address this issue, as stated in the methodology, the perplexity μ was set as $N/3$. So that for any structure x_i, the joint probability or similarity was calculated with regard to all other data points to preserve the "global structures" of the original data set. As a comparison, the t-SNE distributions with reference to the crystal structures of dark and light states, respectively, are plotted in Figure 5 with regard to different perplexity μ.

The results in Figure 5 show that increasing perplexity for more nearest neighbors to be calculated significantly increases the preservation of global structure through projection. With a small perplexity value as 10, the cluster 6 is projected adjacent to the clusters 1, 7, and 9 (Figure 5a). However, the conformation is significantly different between the cluster 6 (light state) and the clusters 1, 7, and 9 (dark state). With the perplexity value as 100, clusters 2 and 4 are close to each other (Figure 5b). With the perplexity value as 1000, the t-SNE method gives a well-behaved representation (Figure 5c) that converges to the most comprehensive analysis with the perplexity value as $N/3$ with N as 12,000 for VVD (Figure 5d).

With the larger perplexity value, the KL divergence between low-dimensional description with the high-dimensional data decreases. The KL divergences are 1.556, 0.887, 0.364, and 0.143 for the perplexity values as 10, 100, 1000, and $N/3$, respectively. Smaller KL divergence values mean that the low-dimensional description can better represent the high-dimensional data structure. Clusters 5 and 6 have the lowest averaged RMSDs with reference to the dark and light state crystal structures, respectively. With the largest perplexity, clusters 5 and 6 lay at the two opposite locations on the 2D t-SNE surface (Figure 5d). Therefore, the clusters that lay between clusters 5 and 6, including clusters 1, 3, 4, 7, and 9, may represent a gradual conformational change from the dark state region (plotted in blue) starting from the cluster 5 toward the light state region (plotted in red) ending at the cluster 6. With

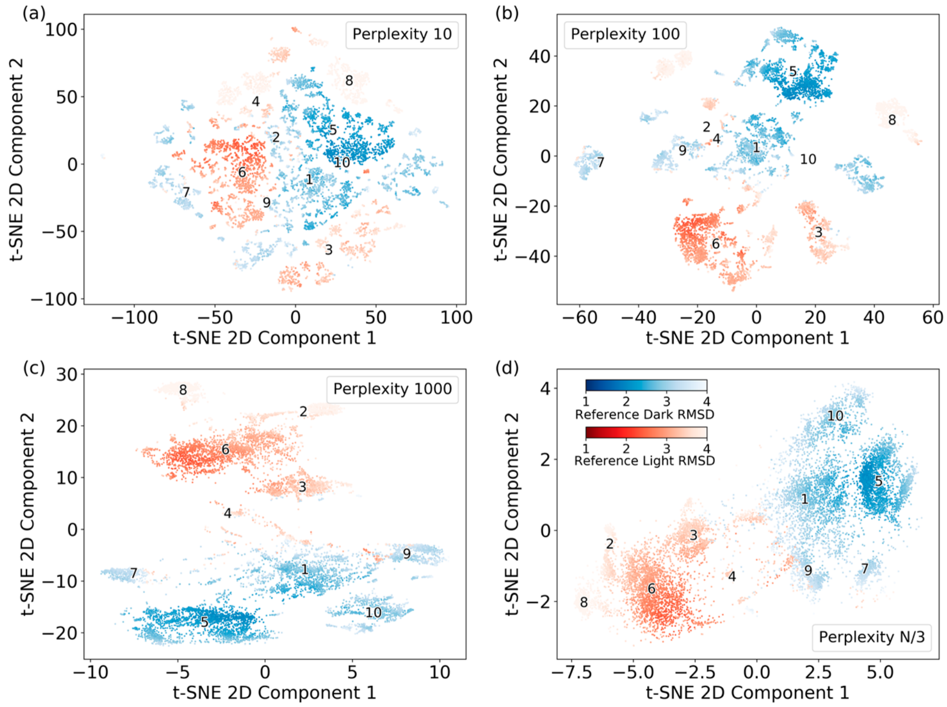


Figure 5. t-SNE 2D projection free energy surface using different perplexity μ values. The joint probability calculated for 3 μ nearest neighbors is as follows: (a) perplexity value as 10; (b) perplexity value as 100; (c) perplexity value as 1000; (d) perplexity value as $N/3$. N is the total number of data points. Different colors indicate that a structure is either close to the dark (blue) or light (red) state crystal structure in terms of RMSD values.

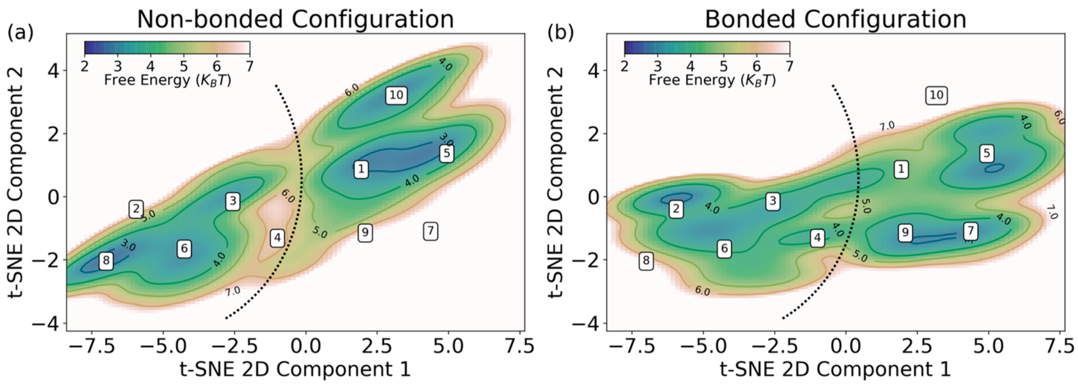


Figure 6. Free energy surfaces estimated from (a) t-SNE 2D projection from nonbonded configuration samplings and (b) t-SNE 2D projection from bonded configuration samplings.

a large perplexity value, the t-SNE method could preserve the global structure of the original data set, and the 2D t-SNE projection reveals potentially feasible transitions between the dark and light states of VVD.

Revealing the Covalent Bond Effects Based on 2D t-SNE Projection. The above analyses demonstrate the advantage of the t-SNE method compared with other dimensionality reduction methods in constructing free energy surfaces and capturing the structural changes. Next, the t-SNE method is further applied on VVD protein simulations to reveal the influence of the key photoinduced covalent bond between VVD and its cofactor FAD on the overall free energy landscape.

The free energy surfaces are plotted on the 2D t-SNE projection for the simulations of nonbonded and bonded configurations, respectively, (Figure 6). Direct comparison between two plots shows that the photoinduced covalent bond significantly changes the free energy landscape of VVD protein. It should be noted that ten high-dimensional k-means clusters labeled by numbers on both plots are not expected to be the free energy minima on either surface, because the clustering was carried out using the simulations from all configurations. In the nonbonded configurations (Figure 6a), dark state clusters 1, 5, and 10 and light state clusters 3, 4, and 6, as well as the hidden state cluster 8, are all extensively sampled. One light state cluster 2 and two dark state clusters 7 and 9 are not sampled well. In the bonded configuration (Figure 6b), dark state clusters 1, 5, 7, and 9 and light state clusters 2, 3, 4, and 6 are all sampled sufficiently. Light state cluster 10 and the hidden state cluster 8 are not sampled well. The difference between the sampling results of nonbonded and bonded configurations reveals the impact of the photoinduced covalent bond on the free energy landscape of the system.

It should be noted that the transition region between the dark state crystal structure (cluster 5) and the light state crystal structure (cluster 6) has a lower free energy barrier in the

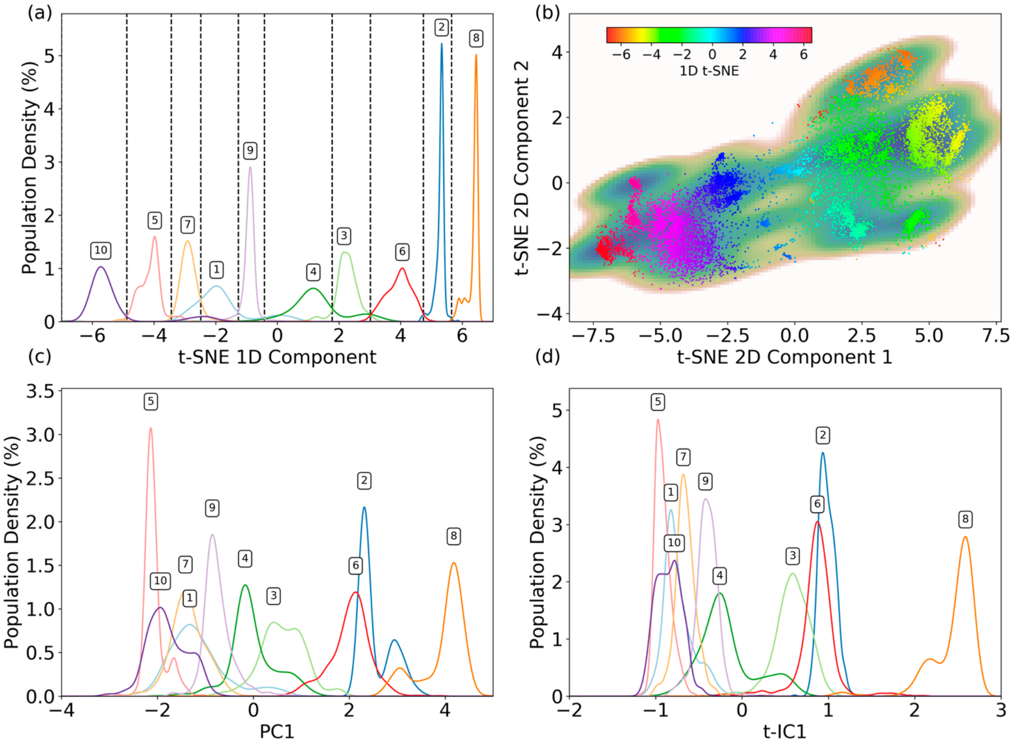


Figure 7. Ten clusters distribution on various 1D spaces: (a) t-SNE 1D projection; (b) t-SNE 2D projection colored by the 1D t-SNE projection value; (c) PCA 1D projection; (d) t-ICA 1D projection.

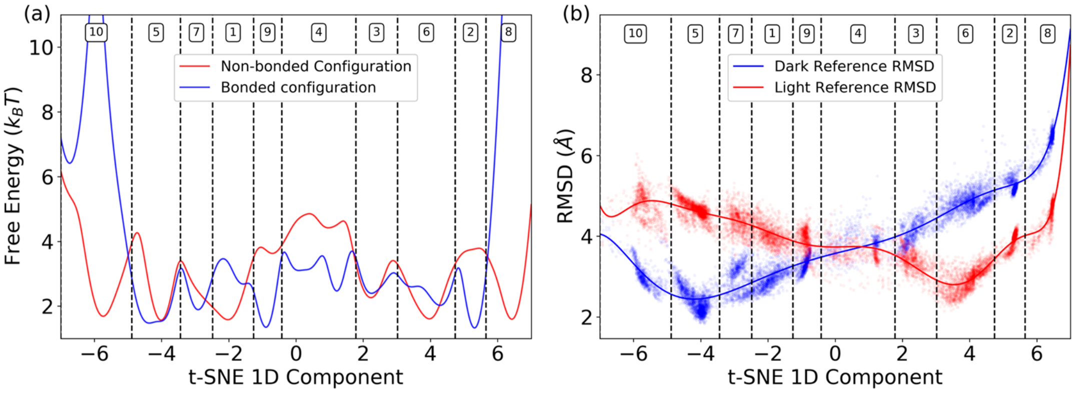


Figure 8. (a) t-SNE 1D projection separated by ten clusters to represent the free energy distribution for bonded and nonbonded configurations and (b) t-SNE 1D projection with regard to the RMSD values referencing to the dark and light state crystal structures, respectively.

bonded configuration (Figure 6b) than the one in the nonbonded configuration indicated by dashed lines (Figure 6a). Without this covalent bond, the cluster 4 and the region between clusters 3 and 1 are less sampled as shown in the nonbonded configuration, resulting in a free energy barrier around 5 to 6 k_BT. In the bonded configuration, the sampling of this region is increased, leading to a lower free energy barrier around 4 k_BT. This significant change of the transition free energy barrier provides a theoretical framework to explain the mechanism in which the photoinduced covalent bond facilitates the transition from the dark state to the light state. To evaluate the transition barrier more accurately, the 1D t-SNE projection was applied to construct free energy profiles as follows.

Revealing the Covalent Bond Effects Based on 1D t-SNE Projection. The 1D t-SNE projection is applied on the VVD simulations with the distribution of each cluster projected and plotted in Figure 7a. All ten clusters are well separated from each other with minimum overlap among them, demonstrating the superior performance of the t-SNE method as an effective dimensionality reduction method. To compare 1D and 2D t-SNE projections, the 1D t-SNE vector is represented as a color spectrum to illustrate distribution of clusters on the 2D t-SNE surface (Figure 7b). The projections of all ten clusters onto the 1D t-SNE space as color spectrum are clearly distinguishable. It should be noted that unlike PCA or t-ICA methods, the 1D t-SNE vector is not either of the two vectors generated in the 2D t-SNE model. As a comparison, these states are also projected onto PC1 or t-IC1 vectors from the PCA and t-ICA models, respectively (Figure 7c and 7d). In these 1D projections, the distributions of ten clusters significantly overlap with each other, indicating that the PC1 or t-IC1 vectors could not capture the difference among these states.

DOI: 10.1021/acs.jctc.8b00652

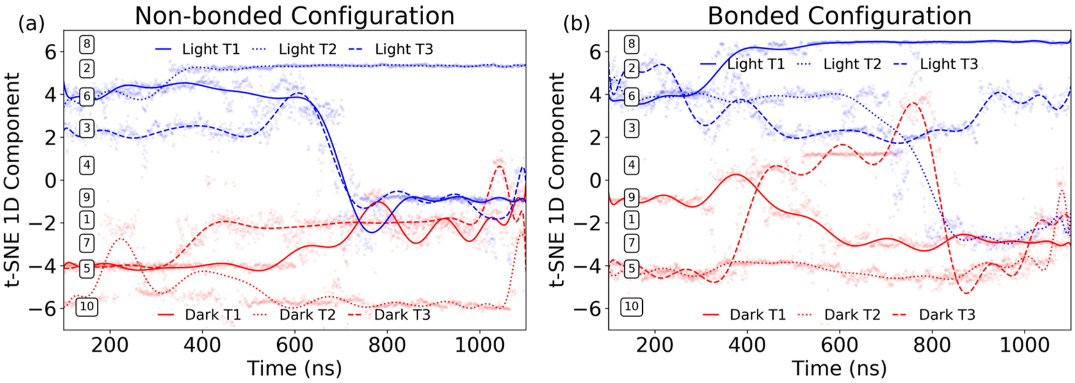


Figure 9. Transition of all 12 trajectories among different clusters: (a) six trajectories with the nonbonded configuration as trajectories starting from the light state (light T1, T2, T3 in blue) and trajectories starting from the dark state (dark T1, T2, T3 in red) and (b) six trajectories with the bonded configuration as trajectories starting from the light state (light T1, T2, T3 in blue) and trajectories starting from the dark state (dark T1, T2, T3 in red).

Using 1D t-SNE projection, the free energy profiles are plotted for the nonbonded and bonded configurations, respectively (Figure 8a). It is obvious that the photoinduced covalent bond significantly alters the free energy surface of VVD protein. Because the distances among clusters along the t-SNE vector reflect the actual similarities of these clusters, it is likely that clusters between cluster 5 (representing the dark state crystal structure of VVD) and cluster 6 (representing the light state crystal structure of VVD), including clusters 7, 1, 9, 4, and 3, could serve as transition regions between the functional dark and light states of VVD. This also agrees with the distribution of these clusters on the 2D t-SNE projection as illustrated in Figures 5 and 7b. The free energy profiles plotted in Figure 8a suggest lower barriers for the transition between dark and light states. This is consistent with the decreasing free energy barrier in the transition region in the bonded configuration compared to the nonbonded configuration shown in 2D free energy surfaces (Figure 6).

To further evaluate feasibility of the transition pathway between VVD dark and light states, the RMSDs of all VVD simulations are plotted in 1D t-SNE with reference to the crystal structures of VVD dark (blue) and light states (red), respectively (Figure 8b). The distribution profile (represented as solid line) is fitted for each RMSD plot as well. Both RMSD distributions reveal a smooth change of ten clusters on this 1D t-SNE distribution. Cluster 4 is the most likely to serve as the transition state region, because this cluster has relatively equal RMSDs with reference to the dark and light state crystal structures. Clusters 10, 2, and 8 are unlikely to be involved in the dark/light states transition, because these clusters deviate from both dark and light state crystal structures.

Finally, each trajectory was individually analyzed to track real time transition among different states represented by clusters. The propagation of each trajectory is projected onto the 1D t-SNE surface with labels corresponding to clustering states (Figure 9). Six 1 μs trajectories for nonbonded and six 1 μs trajectories for bonded configuration are plotted in Figures 9a and 9b, respectively. For each configuration, three trajectories starting from the dark state are referred to as dark T1, T2, and T3 and plotted in red, and three trajectories starting from the light state are referred to as light T1, T2, and T3 and plotted in blue.

In the nonbonded configurations, no dark trajectory samples any light state clusters (clusters 2, 3, and 6) or hidden state close to the light state (cluster 8). Only dark T3 trajectory briefly reaches cluster 4 as the proposed transition state region before falling back to the dark state region (Figure 9a). Interestingly, dark T2 trajectory dwells in the hidden state cluster 10 for a significant portion of the simulation. On the contrary, light T1 and T3 trajectories show clear transitions from the light to the dark state region through the transition state region and do not return back to the light state region.

In the bonded configuration, light T1 trajectory mainly dwells in cluster 8 as the hidden state, and light T3 trajectory samples clusters 6 and 3 (Figure 9b). Interestingly, light T2 trajectory also shows the transition from the light to the dark state region through the transition state region and does not return back to the light state region. For the simulations starting from the dark state, dark T1 trajectory briefly approaches cluster 4 as the transition state region before dwelling in the dark state region. Dark T2 trajectory mainly dwells in the cluster 5. Dark T3 trajectory, however, slowly crosses cluster 4 and briefly reaches cluster 6 as a light state and quickly transforms back to the dark state afterward. Compared to the nonbonded configuration, the presence of the photoinduced covalent bond does increase the probability of transformation from the dark state to the light state.

■ DISCUSSION

Developed by Geoffrey Hinton and Laurens van der Maaten, the t-SNE method is a nonlinear dimensionality reduction method and has been widely applied in many fields including artificial intelligence, cancer research, biomedical signal processing, and bioinformatics.[33−36] In the current study, the t-SNE method is applied on molecular dynamics simulations of circadian protein VVD to demonstrate the effectiveness of this method in probing free energy surfaces and reveal potential allosteric effects associated with the photoinduced covalent bond in VVD. For many dimensionality reduction methods being applied on molecular simulations, structural information loss is inevitable when describing 3*N*-dimensional structures by only one or two dimensions. The widely applied PCA method identifies the eigenvector to capture the maximum variance of the protein fluctuation during simulation. The t-ICA method identifies the eigenvector with the maximum autocorrelation time to represent the slowest dynamical motions. Both PCA and t-ICA are linear dimensionality reduction methods. However, for protein systems, nonlinear dimensionality reduction methods could be more suitable by preserving

DOI: 10.1021/acs.jctc.8b00652
J. Chem. Theory Comput. 2018, 14, 5499−5510

maximum structural and dynamical information.[25] Recently, some nonlinear dimensionality reduction methods including diffusion map,[25] isomap,[57] autoencode, and time-lagged autoencode[26] have been developed. These methods have different strengths in extracting critical structural and dynamical information. For example, time-lagged autoencoders could outperform t-ICA methods by embedding the nonlinear transformation to search the conformational changes with maximum autocorrelation time.[26] Compared to these methods, the t-SNE method is superior in extracting the pairwise distance information from high-dimensional structures and constructing low-dimensional descriptors. Practically, pairwise distances are the most commonly used order parameters to construct free energy surfaces. In a recent study, a *k*-nearest neighbor estimator was applied to estimate the free energy of a high-dimensional system through a low-dimensional embedding manifold by design without explicit projection.[58] As a comparison, the t-SNE method, also a type of stochastic neighboring embedding method, explicitly projects the densities in the high-dimensional space onto a low-dimensional space with minimum structural information loss.

There is no universal standard to compare different dimensionality reduction methods. Many studies applied different metrics for comparison.[4,55,59] In principle, an adequate low-dimensional descriptor should have the following properties. If two points are very close on the projected surface, they should correspond to the similar high-dimensional structures. The *k*-means clustering method partitions multidimensional data into different clusters. For simulations of biomacromolecules such as proteins, these clusters are referred to as microstates and applied in constructing Markov state models (MSMs). To be an adequate 1D or 2D descriptor for protein simulations, it should lead to low averaged RMSDs in each microstate generated using this descriptor, to maintain the structural similarity within each microstate.

As demonstrated in this study, the t-SNE method has the best performance, while t-ICA has the worst performance based on the comparison of the structure similarity inside each cluster. This result is somewhat surprising, because t-ICA should outperform PCA in capturing the slowest dynamical motions in theory. There are two possible explanations for this observation. First, the lag time of t-ICA may not be adjusted thoroughly to achieve the best performance. Second, some small conformational changes may be associated with slow transition time but are treated as the “fast dynamics”, which worsens the performance of the t-ICA method. In the current study, the large conformational changes among dark, light, and hidden conformations are captured as the slowest dynamics. However, for the smaller conformational changes among the states within the dark or light regions, the t-ICA method cannot distinguish them very well. The nonlinear design of the t-SNE method enables this method to maximally preserve the data distribution, resulting in both 2D and 1D t-SNE analyses with the best performance. This validates the t-SNE as a superior alternative method for analysis of molecular dynamics simulations for biomacromolecules.

In MSM, the relaxation time is an estimated time to approach steady state. Experimentally, the relaxation time scale to accomplish the transition among different conformations can be up to milliseconds to seconds for proteins.[56] In general, the one based on the Cartesian coordinates implies the transition time scale that is the closest to the experimental observation (Figure 1b). All other MSMs based on various dimensionality reduction methods imply significantly shorter transition time scales. The 2D t-SNE model implies a transition time scale closest to the one based on Cartesian coordinates. The microstates are constructed based on the assumption that no significant kinetic barrier exists within each microstate. Therefore, inadequate construction of the microstates could cause that some original kinetic barriers in the high-dimensional Cartesian space are disguised or distorted upon projection, leading to inferior performance of other models.

Due to the preservation of pairwise distance distribution, the t-SNE method is excellent to represent and distinguish high-dimensional clusters. Clustering analysis, also considered as an unsupervised learning, has been widely applied on MD simulations including structure similarity based clustering (e.g., *k*-means) and kinetic based clustering (e.g., MSMs).[10,53] Because of the structural similarity, each cluster often corresponds to a minimum on the free energy surface. As demonstrated in Figure 3, PCA or t-ICA methods as well as 2D RMSD could not represent ten clusters generated using Cartesian coordinates well, showing significant overlap among some clusters when using these projections. As a comparison, both 2D and 1D t-SNE models (Figure 3a and 7a, respectively) have much better performance in distinguishing different clusters. This strongly suggests that the t-SNE method could serve as a general dimensionality reduction tool to capture the difference among high-dimensional clusters and represent the free energy surface for biomacromolecules.

The t-SNE method is further applied to quantitatively evaluate the impact of the photoinduced covalent bond on protein VVD allostery and identify the potential conformational switching pathways. Both the 1D t-SNE free energy profile and the 2D t-SNE free energy surface suggest that the covalent bond could lower the free energy in the transition region (cluster states 4 and 9) by ∼1 k_BT. The decreasing free energy in the transition region is likely to facilitate a functionally important conformational transition from the dark to the light state. Overall, the t-SNE method should be one important addition in the simulation analysis toolbox to distinguish clusters and represent the free energy surface for biomacromolecule simulations and can be combined with other methods for more informative analyses.

■ CONCLUSION

In this study, the t-SNE method was applied as a superior dimensionality reduction method for the analysis of molecular dynamics simulations of proteins. The advantage of the t-SNE method in retaining the pairwise distance distribution information, capturing the conformational changes, distinguishing the high-dimensional clusters, and representing free energy surface was demonstrated through comparison with other commonly used dimensionality reduction methods. It is also demonstrated that even with only one dimension, the t-SNE method has a better performance than many other methods, rendering this method as one of the best options for the analysis of biomacromolecules simulations. Using the 1D t-SNE model, a time dependent fitting analysis was carried out to track the real time state changes of each trajectory. Overall, the t-SNE method could retain the structural and dynamical information with minimum information loss compared to other commonly used dimensionality reduction methods and could be applied for the analyses of simulations for many other biomacromolecules.

AUTHOR INFORMATION

Corresponding Author

*E-mail: ptao@smu.edu.

ORCID

Feng Wang: 0000-0001-7808-7538
Peng Tao: 0000-0002-2488-0239

Funding

The work was supported by the National Institutes of Health [GM122013]. Computational time was generously provided by Southern Methodist University's Center for Scientific Computation.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4* (2), 187−217.

(2) 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. OpenMM 4: a reusable, extensible, hardware independent library for high performance molecular simulation. *J. Chem. Theory Comput.* **2013**, *9* (1), 461−469.

(3) 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. Millisecond-scale molecular dynamics simulations on Anton. In *Proceedings of the conference on high performance computing networking, storage and analysis*; ACM: 2009; p 39, DOI: 10.1145/1654059.1654126.

(4) 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.

(5) 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; ACM: 1998; pp 604−613.

(6) Ichiye, T.; Karplus, M. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Struct., Funct., Genet.* **1991**, *11* (3), 205−217.

(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* (10), L75−L77.

(8) Chodera, J. D.; Noé, F. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135−144.

(9) Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002**, *9* (9), 646.

(10) Papaleo, E.; Mereghetti, P.; Fantucci, P.; Grandori, R.; De Gioia, L. Free-energy landscape, principal component analysis, and structural clustering to identify representative conformations from molecular dynamics simulations: the myoglobin case. *J. Mol. Graphics Modell.* **2009**, *27* (8), 889−899.

(11) Lin, T.; Zha, H. Riemannian manifold learning. *IEEE Trans. Pattern Anal. Mach. Intell* **2008**, *30* (5), 796−809.

(12) 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.

(13) Hayward, S.; Go, N. Collective variable description of native protein dynamics. *Annu. Rev. Phys. Chem.* **1995**, *46* (1), 223−250.

(14) Noguti, T.; Gō, N. Collective variable description of small-amplitude conformational fluctuations in a globular protein. *Nature* **1982**, *296* (5859), 776.

(15) Dove, S. L.; Joung, J. K.; Hochschild, A. Activation of prokaryotic transcription through arbitrary protein−protein contacts. *Nature* **1997**, *386* (6625), 627.

(16) Lobanov, M. Y.; Bogatyreva, N.; Galzitskaya, O. Radius of gyration as an indicator of protein structure compactness. *Mol. Biol.* **2008**, *42* (4), 623−628.

(17) Cho, S. S.; Levy, Y.; Wolynes, P. G. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (3), 586−591.

(18) Zheng, W.-S.; Lai, J.-H.; Yuen, P. C. Linear dimension reduction techniques. In *Encyclopedia of Biometrics*; Springer: Germany, 2009; pp 899−904, DOI: 10.1007/978-3-642-27733-7.

(19) Roweis, S. T.; Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **2000**, *290* (5500), 2323−2326.

(20) Tenenbaum, J. B.; De Silva, V.; Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290* (5500), 2319−2323.

(21) Levy, R.; Srinivasan, A.; Olson, W.; McCammon, J. Quasi-harmonic method for studying very low frequency modes in proteins. *Biopolymers* **1984**, *23* (6), 1099−1112.

(22) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (21), 7426−7431.

(23) Hinton, G. E.; Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313* (5786), 504−507.

(24) 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.

(25) 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.

(26) Wehmeyer, C.; Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.* **2018**, *148* (24), 241703.

(27) Schwantes, C. R.; Pande, V. S. Modeling molecular kinetics with tICA and the kernel trick. *J. Chem. Theory Comput.* **2015**, *11* (2), 600−608.

(28) Mitsutake, A.; Iijima, H.; Takano, H. Relaxation mode analysis of a peptide system: Comparison with principal component analysis. *J. Chem. Phys.* **2011**, *135* (16), 164102.

(29) Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; Wu, A. Y. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24* (7), 881−892.

(30) Noé, F.; Wu, H.; Prinz, J.-H.; Plattner, N. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *J. Chem. Phys.* **2013**, *139* (18), 184114.

(31) Maaten, L. v. d.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9* (Nov), 2579−2605.

(32) Joyce, J. M. Kullback-leibler divergence. In *International Encyclopedia of Statistical Science*; Springer: Germany, 2011; pp 720−722, DOI: 10.1007/978-3-642-04898-2_327.

(33) Haghverdi, L.; Buettner, F.; Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **2015**, *31* (18), 2989−2998.

(34) Wilson, N. K.; Kent, D. G.; Buettner, F.; Shehata, M.; Macaulay, I. C.; Calero-Nieto, F. J.; Castillo, M. S.; Oedekoven, C. A.; Diamanti, E.; Schulte, R. Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. *Cell Stem Cell* **2015**, *16* (6), 712−724.

(35) Amir, E.-a. D.; Davis, K. L.; Tadmor, M. D.; Simonds, E. F.; Levine, J. H.; Bendall, S. C.; Shenfeld, D. K.; Krishnaswamy, S.; Nolan, G. P.; Pe'er, D. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **2013**, *31* (6), 545.

(36) Grün, D.; van Oudenaarden, A. Design and analysis of single-cell sequencing experiments. *Cell* **2015**, *163* (4), 799−810.

(37) Rydzewski, J.; Nowak, W. Ligand diffusion in proteins via enhanced sampling in molecular dynamics. *Phys. Life Rev.* **2017**, *22−23*, 58−74.

(38) Zhang, J.; Chen, M. Unfolding hidden barriers by active enhanced sampling. *Phys. Rev. Lett.* **2018**, *121* (1), 010601.

(39) Zoltowski, B. D.; Crane, B. R. Light activation of the LOV protein vivid generates a rapidly exchanging dimer. *Biochemistry* **2008**, *47* (27), 7012−7019.

(40) Zoltowski, B. D.; Vaccaro, B.; Crane, B. R. Mechanism-based tuning of a LOV domain photoreceptor. *Nat. Chem. Biol.* **2009**, *5* (11), 827−834.

(41) Kar, G.; Keskin, O.; Gursoy, A.; Nussinov, R. Allostery and population shift in drug discovery. *Curr. Opin. Pharmacol.* **2010**, *10* (6), 715−722.

(42) Arora, K.; Brooks, C. L. Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (47), 18496−18501.

(43) 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.

(44) 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.

(45) 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* (2), 926−935.

(46) 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.

(47) Brooks, B. R.; Brooks, C. L.; MacKerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **2009**, *30* (10), 1545−1614.

(48) Eastman, P.; Pande, V. OpenMM: a hardware-independent framework for molecular simulations. *Comput. Sci. Eng.* **2010**, *12* (4), 34−39.

(49) Pillai, S. U.; Suel, T.; Cha, S. The Perron-Frobenius theorem: some of its applications. *IEEE Signal Process. Mag.* **2005**, *22* (2), 62−75.

(50) Zhou, H.; Zoltowski, B. D.; Tao, P. Revealing Hidden Conformational Space of LOV Protein VIVID Through Rigid Residue Scan Simulations. *Sci. Rep.* **2017**, *7*, 46626.

(51) Van Der Maaten, L. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **2014**, *15* (1), 3221−3245.

(52) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(53) 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.

(54) 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.

(55) Doerr, S.; Ariz-Extreme, I.; Harvey, M. J.; De Fabritiis, G. Dimensionality reduction methods for molecular simulations. 2017, ArXiv e-prints, arXiv. https://arxiv.org/abs/1710.10629 (accessed Oct 1, 2018).

(56) Lamb, J. S.; Zoltowski, B. D.; Pabit, S. A.; Crane, B. R.; Pollack, L. Time-resolved dimerization of a PAS-LOV protein measured with photocoupled small angle X-ray scattering. *J. Am. Chem. Soc.* **2008**, *130* (37), 12226−12227.

(57) You, Z.-H.; Lei, Y.-K.; Gui, J.; Huang, D.-S.; Zhou, X. Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics* **2010**, *26* (21), 2744−2751.

(58) Rodriguez, A.; d'Errico, M.; Facco, E.; Laio, A. Computing the Free Energy without Collective Variables. *J. Chem. Theory Comput.* **2018**, *14* (3), 1206−1215.

(59) Tribello, G. A.; Ceriotti, M.; Parrinello, M. Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (14), 5196−5201.