ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

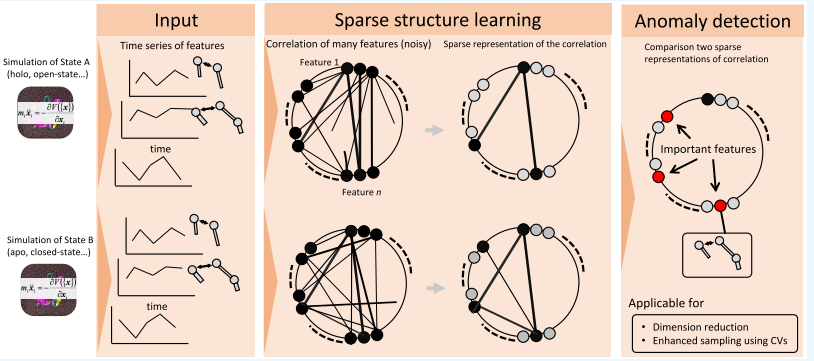
Application of Anomaly Detection to Identify Important Features of Protein Dynamics

Yu Yamamori* and Kentaro Tomii

Cite This: *ACS Omega* 2025, 10, 22789−22801

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Molecular dynamics (MD) simulations are a popular tool for the study of protein dynamics. Recent machine-learning-based structure prediction methods, such as AlphaFold, can provide a broad variety of initial protein structures for MD simulation. Hence, the development of methods to enhance the practicality of MD simulation (such as efficient sampling or detection of collective variables) is increasingly important. Identifying a small number of elements or features that can describe biological phenomena from MD trajectories serves as a basis for these methods. In this study, we applied the anomaly detection method based on sparse structure learning of the element correlation within MD trajectories to identify important features associated with state transitions. This approach was tested on the correlation of residue−residue distances from the open- and closed-state simulations of T4 lysozyme and the holo- and apo-state simulations of the PDZ3 domain. This has clear implications for understanding cooperative motion through its combination with a dimension reduction technique.

1. INTRODUCTION

Molecular dynamics (MD) simulations have become increasingly popular for understanding the mechanism of biomolecules. Moreover, it is the only tool that can easily observe the behavior of biomolecules in atomic spatial resolution and molecular vibration time resolution.[1,2] The primary outcome of an MD simulation is a time series of atomic coordinates and velocities known as a "*trajectory*". The objective of MD simulations is to deduce structural, kinetic, or thermodynamic properties from this trajectory. For example, thermodynamic properties are usually computed as a time average of physical quantities obtained from the trajectory, as the execution of MD simulations can be considered as sampling in phase space. Advances in computational resources and techniques[3−8] have extended the spatiotemporal limit of MD simulations, enabling simulations of the entire cellular environment at the atomic level[9,10] or multiple unfolded-to-folded state transitions of a globular protein, if state-of-the-art supercomputers are available.[11] The growing number of experimentally determined structures[12] and recent advances in machine learning structure prediction methods[13−17] have contributed to the increasing attention to and applications of MD simulations. However, there is a gap between the scope of many MD simulations and those of biologically significant phenomena. Additionally, extracting meaningful interpretations from the trajectory remains a crucial task.

Enhanced sampling methods offer powerful solutions for bridging the spatiotemporal gap between simulations and real phenomena.[18−20] Several approaches have been developed to improve the efficiency of MD simulations, including umbrella sampling,[21] multicanonical MD,[22] replica exchange MD (REMD),[23−26] metadynamics,[27−29] and temperature-accelerated MD (TAMD).[30] Currently, REMD and metadynamics are the two most widely adopted methods. The original REMD method[23] employs noninteracting replicas of the system which are simultaneously simulated at different temperatures, ranging

Received: December 24, 2024
Revised: April 20, 2025
Accepted: April 30, 2025
Published: May 29, 2025

from the target temperature to the temperature that enables the system to cross the free energy barrier. The temperatures of the replica pairs are periodically exchanged by following the Metropolis method to achieve random walks along the temperature space. The REMD family comprises numerous variations, owing to its high extensibility. Hamilton REMD[24] is a prominent example characterized by replicas with varying parameters of the energy function and exchanges parameters to randomly traverse the parameter space. Metadynamics improves sampling by introducing an explicit evolution of collective variables (CVs) and a CV-based potential function with a history-dependent component. TAMD is like metadynamics in that it uses the time evolution of predefined CVs, but it distinguishes itself by considering two temperatures, one associated with the Cartesian coordinates and the other with the CVs, with the latter being set to a high temperature to promote efficient CV space sampling. If the parameters satisfy the necessary conditions, the simulation can yield a probability distribution as a function of the CV. Additionally, variations of REMD use predefined CVs.[31] For many of these promising methods, successful simulation depends on identifying appropriate CVs capable of accurately describing relevant phenomena.

The second issue is the interpretation of the trajectory, which is closely linked to the identification of appropriate CVs. In many practical scenarios, the selection of CVs is often guided by a chemical or physical intuition for the target system. However, this work focused on automatic methods for extracting CVs from a trajectory. Dimension reduction techniques are commonly employed for this purpose.[32,33] Principal component analysis (PCA) is the most widely used method.[34−37] In PCA, diagonalizing a covariance matrix of the selected coordinates maximized the fluctuations of the first component. In general, Cartesian coordinates are used for PCA after the removal of overall translation and rotation motion. Other choices of coordinates are internal coordinates such as dihedral angles,[38,39] interatomic distances,[40,41] and potential energy terms.[42] Another noteworthy method is time-lagged independent component analysis (tICA), which was originally introduced in signal processing and aims to maximize the time scales of the first component to capture the slower motions of biomolecules.[43−45] Kernel tICA,[46] relaxation mode analysis,[47,48] and dynamics mode decomposition[49] are other methods used to discover slow CVs. Methods employing nonlinear dimensionality reduction methods such as Isomap[50] and diffusion maps[51,52] were used to analyze MD simulation to transcend the limits of linear projection methods. Another noteworthy method used a graph-based approach.[53] More recent CV-finding methods based on deep learning[19,32] include time-lagged autoencoder[54] and variational approach to Markov process network (VAMPnets).[55]

In most cases, the problems of sampling and interpretation are closely related to that of CV-finding. The first step of CV-finding is to determine the set of input coordinates. For the best use of CV-finding methods, it is desirable to know the smallest number of coordinates that can describe a specific biological process. These key coordinates are often called "order parameters"[56] or "features".[57] A small number of key coordinates is expected to exist, and detecting them is helpful in many biological processes such as allosteric transitions including cooperative motions.[58] Attempts to identify features automatically, independent of specific chemical or biological knowledge of the target system, can be framed as a feature-finding problem. Previous attempts to achieve this include functional mode analysis (FMA), which detects collective motions related to a particular protein function,[59] automatic mutual information noise omission method (AMINO), which uses a mutual information-based distance metric,[56] the sparse group lasso (SGL) method,[60] and molecular systems automated identification of cooperativity (MoSAIC) method, which uses the correlation relationship between input coordinates and clustering algorithm.[57] These methods are classified as preprocesses of CV-finding methods such as PCA and tICA or of enhanced sampling such as metadynamics and contribute to improving the accuracy and efficiency of CV-finding and automation of enhanced sampling methods. Besides, these "features" themselves are expected to be helpful in interpreting trajectories.

In this study, we propose a method that applies anomaly detection[61] to MD trajectories to identify important features in protein dynamics. We assume the following: the structures of a biomolecule (a protein) at different states are available, and MD simulations can be performed from the initial structures of each state, although the transition between them may be outside the scope of a brute-force MD simulation. This is a typical situation in today's MD simulations. Among anomaly detection methods, our method estimates the sparse structure or sparse precision matrices from the correlation relationship of input coordinates with the graphical lasso for two states and identifies a small number of coordinates with a high "anomaly" that comes from the difference between two states. We intend that the few highly anomalous coordinates serve as the necessary coordinates to describe the differences between the two states. The feature-finding methods discussed here[56−60] can be summarized as approaches that select a small number of degrees of freedom from a large set, such as distances between atoms/residues of the protein or sets of dihedral angles of the protein. Our proposed method shares its use of sparse relationships with the SGL method[60] and its use of correlation with the MoSAIC method.[57] The proposed method is set apart by its potential to discover important features even from simulations at a time scale that does not include state transitions by comparing relatively short trajectories from different states and to capture accurate correlations by eliminating pseudocorrelations. To validate these assertions, we tested our proposed method on two examples involving a state transition: the open-closed transition of the T4 lysozyme and the dynamic allostery of the PDZ3 domain.

2. METHODS

2.1. Anomaly Detection Based on Sparse Structure Learning. The general purpose of anomaly detection is to identify abnormal patterns or elements in the given data set. In our study, we identified a small number of elements with anomalies by comparing two independent multidimensional time series that contain the same set of elements.[61] The method has two stages: initially, the learnings of the sparse correlation relationship of elements of each time series are determined. The correlation of elements in each time series is approximated as a multidimensional Gaussian distribution constrained by the sparsity of its precision matrix (inverse covariance matrix). The precision matrix is estimated through the maximum a posteriori (MAP) estimation method. If the i,jth element of the estimated (sparse) precision matrix is nonzero, the correlation relationship is estimated between the ith and jth elements of the time series. Subsequently, a small

https://doi.org/10.1021/acsomega.4c11546
ACS Omega 2025, 10, 22789−22801

number of elements with anomalies are identified by comparing two sparse correlation relationships. We summarized the method as follows based on the original paper.[61]

Initially, we introduced the first stage of the method, sparse structure learning of one time series. Suppose that we have a time series $D = \{\mathbf{x}^M(n)|n = 1, ..., N\}$, where $\mathbf{x} = (...x_i...)$ denotes an M-dimension random variable, x_i denotes the ith element of $\mathbf{x}$, n denotes the index of the discrete time, and N is the number of sampling snapshots. In the case of application to an MD trajectory, D corresponds to a trajectory, and x can be considered a function of Cartesian coordinates of atoms such as interatomic distances. Without losing the generality, we assume that each element of $\mathbf{x}$ is standardized as

$$x_i \leftarrow \frac{x_i - \mu_i}{\sqrt{\Sigma_{i,i}}} \tag{1}$$

where x_i is the ith element of x, μ is the mean of $\{x_i\}$ defined as $\mu_i = \frac{1}{N}\sum_{n=1}^{N} x_i(n)$, and $\Sigma_{i,i}$ is the variance defined as $\Sigma_{i,i} = \frac{1}{N}\sum_{n=1}^{N} x_i^2(n)$; the operation "$\leftarrow$" means substitution. We assume that the probability distribution function of $\mathbf{x}$ is an M-dimensional Gaussian distribution with mean $\mathbf{0}$ and variance Λ^{-1}, $p(\mathbf{x}) = N(\mathbf{x}|\mathbf{0},\Lambda^{-1})$, where $\Lambda^{M\times M}$ is the precision matrix, $|\Lambda|$ means the determinant of Λ, and $\Sigma^{M\times M}$ is the variance-covariance matrix, where the relationship $\Sigma = \Lambda^{-1}$ is satisfied. We aim to estimate Λ using the MAP estimation based on D. (Estimation of Λ directly means estimations of $p(\mathbf{x})$ and Σ.) To make the precision matrix Λ sparse, the prior probability distribution for Λ must be set. In this case, we define the prior probability distribution $P(\Lambda)$ as

$$P(\Lambda) = \frac{\rho}{2}\exp\left(-\sum_{i,j}^{M}\rho(|\Lambda_{i,j}|)\right) = \frac{\rho}{2}\exp(-\rho\ ||\Lambda||_1) \tag{2}$$

where $\rho(\subset[0, 1])$ is the parameter that controls the degree of sparsity of Λ. If ρ is sufficiently near 1, most of the elements of Λ must be 0. Based on the MAP estimation, the following maximizing problem is to be solved:

$$\Lambda^* \leftarrow \arg\max_{\Lambda}\{\ln P(\Lambda)p(\mathbf{x})\} \tag{3}$$

The object function for maximization can be transformed as

$$\begin{aligned}
&\ln P(\Lambda)\prod_{n=1}^{N} N(\mathbf{x}|\mathbf{0},\ \Lambda)\\
&\quad = \ln\frac{\rho}{2}\exp(-\rho\ ||\Lambda||_1)\\
&\qquad + \sum_{n=1}^{N}\ln\frac{|\Lambda|^{1/2}}{(2\pi)^{M/2}}\exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)\\
&\quad = \ln\frac{\rho}{2} - \rho\ ||\Lambda||_1 + \frac{N}{2}\ln|\Lambda| - \frac{1}{2}\mathrm{Tr}\left(\Lambda\sum_{n=1}^{N}x^Tx\right) + C\\
&\quad = -\frac{N}{2}\mathrm{Tr}(S\Lambda) + \frac{N}{2}\ln|\Lambda| - \rho\left(\sum_{i,j}|\Lambda_{i,j}|\right) + C
\end{aligned} \tag{4}$$

Subsequently, ignoring constant factors and constant terms, we redefine the object function $f(\Lambda)$ for maximization as

$$f(\Lambda) \equiv \ln|\Lambda| - \mathrm{Tr}(S\Lambda) - \rho\left(\sum_{i,j}|\Lambda_{i,j}|\right) \tag{5}$$

where S is the sample variance-covariance matrix from the data D, $S = [S_{i,j}]$ ($S_{i,j} = \frac{1}{N}\sum_{n=1}^{N} x_i(n)x_j(n)$), and C is a constant. $[]_{i,j}$ means the i,j element of the matrix. Then, estimating sparse correlation of x_i means finding Λ that maximizes $f(\Lambda)$. This maximization problem can be solved using the block gradient method,[61] and the precision matrix Λ is obtained under the given value of ρ for the time series D. If the i,j-th element of Λ is not zero, it means that the correlation relationship between x_i and x_j is estimated.

Subsequently, the second stage, the identification of elements with anomalies, is as follows. Assume that we have two precision matrices Λ and Λ' estimated based on two independent multidimensional time series D and D' that contain the same set of elements. The definition of the anomaly $a(\mathbf{x}) = (a_1, ..., a_M)$ of each element of $\mathbf{x}$ is given as

$$a_i \equiv \int d\mathbf{x}_{-i}p(\mathbf{x}_{-i}|D)\int dx_i p(x_i|\mathbf{x}_{-i},\ D)\ln\frac{p(x_i|\mathbf{x}_{-i},\ D)}{p(x_i|\mathbf{x}_{-i},\ D')} \tag{6}$$

where $\mathbf{x}_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_M)$. Equation 6 represents the Kullback−Leibler divergence between $p(x_i|\mathbf{x}_{-i}D)$ and $p(x_i|\mathbf{x}_{-i}, D')$. Using the assumption that $p(\mathbf{x})$ is a multiple Gaussian distribution $N(\mathbf{x}|\mathbf{0},\Lambda^{-1})$, the anomaly of the ith element of $\mathbf{x}$ is obtained as

$$a_i = \frac{1}{2}\ln\frac{\Lambda_{i,i}}{\Lambda'_{i,i}} - \frac{1}{2}\left\{\frac{[\Lambda S\Lambda]_{i,i}}{\Lambda_{i.i}} - \frac{[\Lambda' S\Lambda']_{i,i}}{\Lambda'_{i.i}}\right\} \tag{7}$$

where Λ and Λ' are the precision matrices estimated based on D and D', respectively. The multiple Gaussian distributions do not match the real distribution. However, we justify our assumption as we are only considering the presence or absence of correlation between pairs of elements.

2.2. Application of Anomaly Detection to MD Trajectories. We used the minimum distances of atoms (except hydrogen atoms) between the two residues as inputs for the method. The primary outputs of MD simulations are the time series of Cartesian coordinates of atoms in a system. However, the direct use of Cartesian coordinates is inconvenient, because the overall translation and rotation motion must be removed. Thus, we used minimum atomic distances (except hydrogen atoms) between residues as inputs. We refer to these distances as residue−residue distances. Although all possible pairs should be considered, we subjected the residue pairs to the subjection of the calculation to restrict the number of elements being considered. As input pairs, the union of pairs with residue−residue distances which satisfy the following conditions in two states before and after transition were used: $D_{i,j} \le d$ and $|i - j| > n$, where the indices (i, j) indicate the residue pair (i,j), and $D_{i,j}$ is the residue−residue distance. For the values of d and n, we adopted 8.0 Å and 10, respectively, after a few trials.

Our choice of threshold inter-residue distances is worth mentioning. Several trial-and-error tests were performed to determine the threshold, including the commonly used threshold of the residue−residue contact, where the distance between two atoms except hydrogen atoms of different residues is less than 4.5 Å. We selected 8.0 Å, because we aimed to target a relatively large number of residue−residue

https://doi.org/10.1021/acsomega.4c11546
ACS Omega 2025, 10, 22789−22801

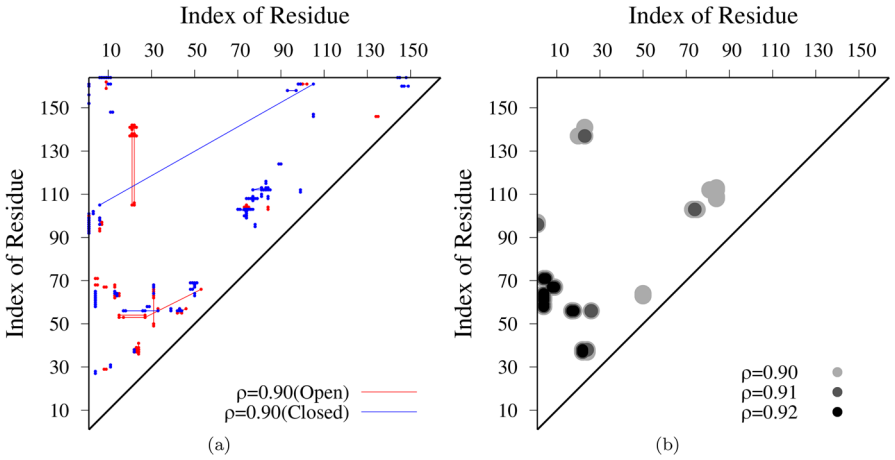


Figure 1. Sparse structure and residue−residue pairs with the high anomalies of T4 lysozyme. (a) Sparse representation of the correlation of residue−residue distances of open states and closed states of T4 lysozyme. Each axis indicates the indices of the residues. Each point represents one residue−residue pair. A line that connects points indicates that a correlation of the corresponding residue−residue distances was found by the method when the sparsity was set at 0.90. Red indicates the open state and blue indicates the closed state. (b) Residue−residue pairs with high anomalies between the open and closed states of T4 lysozyme. The density of the points corresponds to the sparsity which was used for anomaly detection.

pairs around the normal residue−residue contacts. We believed that including a larger number of pairs would ensure robust results. Thus, we adopted a larger distance instead of the usual threshold. When we adopted the threshold 4.5 Å, these results were basically consistent with those when the threshold was 8.0 Å, that is, in both cases, as the sparsity increased, nearly the same residue pairs were identified as those with high anomaly (Figure S3 shows figures corresponding to Figure 1 when the threshold was 4.5 Å).

Other choices for internal coordinates should be considered. We conducted residue−residue interaction energies as one option. This was inspired by studies such as residue interaction network analysis.[62,63] In this case, we used the time series of atomic interaction energy decomposed per residue for the analysis of the simulation of T4 lysozyme: the residue−residue interaction energies are too coarse to compare to previous studies (data not shown).[41,57,58] Nonetheless, we believe that there is room to explore better alternatives for selecting internal coordinates.

To the best of our knowledge, this is the first study to apply the sparse-structure-learning-based anomaly detection algorithm[61] for the analysis of MD trajectories to find important features. To achieve this, the choice of internal coordinates needs to be examined.

2.3. MD Simulations and Analyses. All MD simulations were performed using GROMACS-2022.[64,65] The unit cell was cubic, and the periodic boundary condition was used with a minimum image convention. Electrostatic interaction was handled using the smooth particle mesh Ewald method.[66] The LINCS algorithm was used for the protein to fix the lengths of all bonds involving hydrogen atoms, and SETTLE was used to keep the water molecules rigid.[67,68] For proteins, the AMBER03 force field[69] was used, and for water, the TIP3P model was used.[70] The solution was neutralized, and the NaCl concentration was set at 0.150 M. The number of ions was determined by the SLTCAP server.[71] The equation of motion was integrated with the leapfrog stochastic dynamics method and the inverse friction constant of 0.1 ps to keep the system at 300 K. The pressure was maintained at 1 bar using a Parrinello−Rahman barostat[72] at a coupling time of 5.0 ps. To enlarge a time step at 4 fs, we used the hydrogen mass repartitioning method,[73] and LINCS was set at the sixth order. For the equilibration of the initial system, a recently proposed 10-step protocol was applied and the final 10th equilibration step was performed for 1 ns.[74] For the production run, the configuration was stored every 10 ps. For the analyses of trajectories, the residue−residue distance was calculated using the MDtraj package.[75] Our program conducted sparse structure learning and anomaly detection. For further analysis, PCA was conducted on the residue−residue distances that were identified as high-anomaly residue pairs.

Our intention to apply PCA to the identified features should be noted. The features obtained by anomaly detection are only the list of important elements, such as the list of residue−residue pairs, and we were concerned that these may not provide a clear or simple explanation of the phenomena. We intended to combine PCA with the feature-finding method to capture the important modes and provide a clearer explanation. Although PCA was the first choice for our method of dimension reduction, we would like to use other dimension reduction methods (such as tICA or VAMPnets) in the future that may provide more clarification, as these methods are particularly adept at handling dynamic features.

2.4. T4 Lysozyme and PDZ3 Domain. The first target of our method was the T4 lysozyme, a well-studied system that is supposed to contain a cooperative motion.[41,57,58] T4 lysozyme contains 164 residues, which perform the open-closed transition between two domains. Each state has a lifetime of a few microseconds, and the transition occurs on a nanosecond time scale. We performed several (see section 3.1) 1-μs-long MD simulations of T4 lysozyme in water solution. As the initial structures, an open state structure (PDB ID 150L)[76] and a closed state structure (PDB ID 2LZM)[77] were adopted.

The second target was the PSD-95 PDZ3 domain, which plays an important role in signal transduction. PDZ domains are highly conserved structural modules that consist of two or three α-helices and five β-strands. Moreover, in PDZ domains, the C-terminus of the target protein commonly binds to a groove between the second β-strand and the second α-helix. The PDZ3 domain is a well-studied PDZ domain since it is

https://doi.org/10.1021/acsomega.4c11546
ACS Omega 2025, 10, 22789−22801

regarded as an example of a dynamic allostery.[78−83] Previous MD simulations suggest that the cooperative mechanism was observed in the changes induced by ligand peptide binding similar to the open-closed motion of T4 lysozyme.[80] We conducted three MD simulations, each lasting for 1 μs, of the PDZ3 domain (residues 306−415) with a ligand peptide (KQTSV) in solution and unbounded PDZ3 domain in solution, respectively. As the initial structures for bound and unbound states, crystal structures with PDB IDs 1BE9 and 1BEF were adopted.[84]

3. RESULTS

3.1. Open and Closed States of T4 Lysozyme.

Simulations were conducted for each state to compare the open and closed states, each lasting one μs. We conducted six simulations in total: three initiated from the open state and three from the closed state. During the simulations, one trajectory from the open state rapidly transitioned to the closed state, leading to two stable open-state trajectories, three stable closed-state trajectories, and one trajectory containing both states. For our analysis, we focused on the five stable trajectories (two open-state and three closed-state) while excluding the trajectory that underwent the transition. We conducted the same analyses for all trajectories to check the robustness of the results. Sparse structure learning of the correlations of residue−residue distances was performed for five independent trajectories. Anomaly detection was conducted for the combination of sparse representation of the correlation of the open- and closed-state trajectories. Therefore, we have six sets of results. The input pairs of distances were determined by the following condition: two residues at which the minimum distance between atoms (except hydrogen atoms) belonging to each residue is less than 8 Å, and one residue is separated by 10 or more residues from the other residue. The input pairs were selected based on the initial structures, and 972 pairs was input. The residue−residue pairs with anomalies were detected at various sparsity values including 0.80, 0.90, 0.91, and 0.92. Initially, we checked the number of nonzero elements of the sparse representation of the correlation of residue−residue distances (the number of possible nonzero elements is approximately 5.0×10^5). When the sparsity is 0.80, the number is greater than 500 for all six cases. These are too large for our aim, which is to identify a small number of features. The following sections focus on the cases where the sparsity is equal to or greater than 0.90.

Figure 1a shows the results of sparse structure learning of time series of the residue−residue distances of the correlation between the open- and closed-state T4 lysozyme. This figure corresponds to the results of one of six combinations of the open-state and closed-state simulations (see Figure S4 for other combinations). Each colored point indicates the residue−residue pair. The line connecting the points indicates that the method detected a correlation between the corresponding residue−residue distances. The blue and red colors correspond to the closed and open states, respectively. The number of lines (the number of nonzero elements of the sparse representation of the correlation) was 122 for the open state and 121 for the closed state at a sparsity of 0.90. Figure 1b shows the residue−residue pairs with anomalies identified from the result of Figure 1a using the method described in section 2.1. The density of the colored points corresponds to the sparsity used for anomaly detection (0.90, 0.91, and 0.92). The number of residue−residue pairs was 33, 19, and 14 with sparsities of 0.90, 0.91, and 0.92, respectively. These values are small enough for our aim.

The 33 pairs of residues identified at a sparsity of 0.90 are visualized on the structure of T4 lysozyme (Figure 2). A group

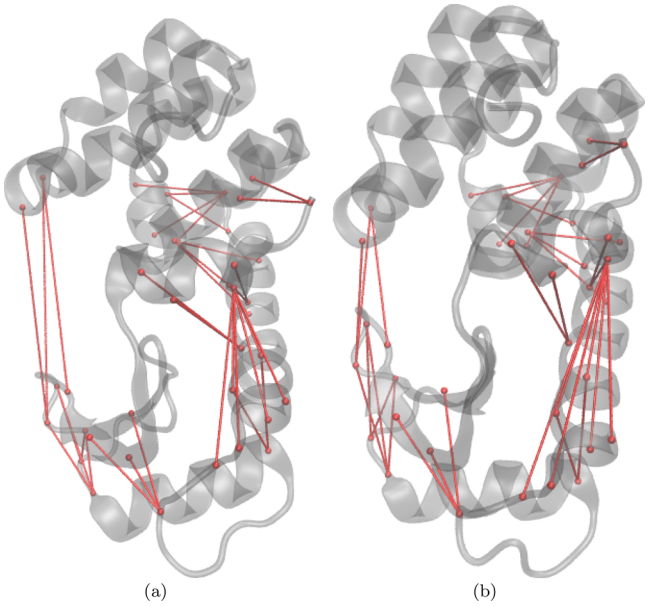


Figure 2. Structures of T4 lysozyme with anomaly residue−residue pairs. (a) The 33 residue−residue pairs with anomalies detected at ρ = 0.90 are projected on the open-state structure. (b) The same pairs are projected on the closed-state structure.

of pairs was observed that connected the first helix (residues 2−12) and the third helix (residues 59−81). Moreover, the pairs that connect the first helix and the fifth helix (residues 92−107) correspond to the "hinge" region of the open-close motion of T4 lysozyme. A group of pairs was observed connecting the loop bridging the first β strand (residues 14−20) to the second β strand (residues 24−28) and the ninth helix (residues 136−142). This corresponds to the "mouth" of the open-close motion. These three groups are the main groups of the identified pairs. Principal component analysis (PCA) of the minimum distances of atoms (except hydrogen atoms) between the identified residue pairs was performed to recognize the implication of the identified pairs. As input data, we combined the time series of the residue−residue distances of the identified pairs from the open- and closed-state simulations. Figure 3a−c gives the logarithm of the probability distribution of the open and closed-state trajectories projected on PC1 vs PC2 space, PC1 vs PC3 space, and PC2 vs PC3 space, respectively. The eigenvalue of PC1 contributed over 80% (Figure 3d).

Figure 4 shows the PC components. The thickness of the red line is proportional to the absolute value of the element of the eigenvector, and the direction of the arrow indicates the sign of the element of the eigenvector. Only PC1 can divide the open and closed states: the open state located in the area where PC1 > 0, and the closed state in the area where PC1 < 0 (see Figure 3a,b). PC1 corresponds to the motion of the opening (closing) of the mouth and shrinking (spreading) of the hinge region of the open-close motion. The mouth opens and hinge shrinks in the positive direction (Figure 4a). PC2 represents a combination of two motions: in the positive direction, the hinge shrinks while the distance between the fifth

https://doi.org/10.1021/acsomega.4c11546
ACS Omega 2025, 10, 22789−22801

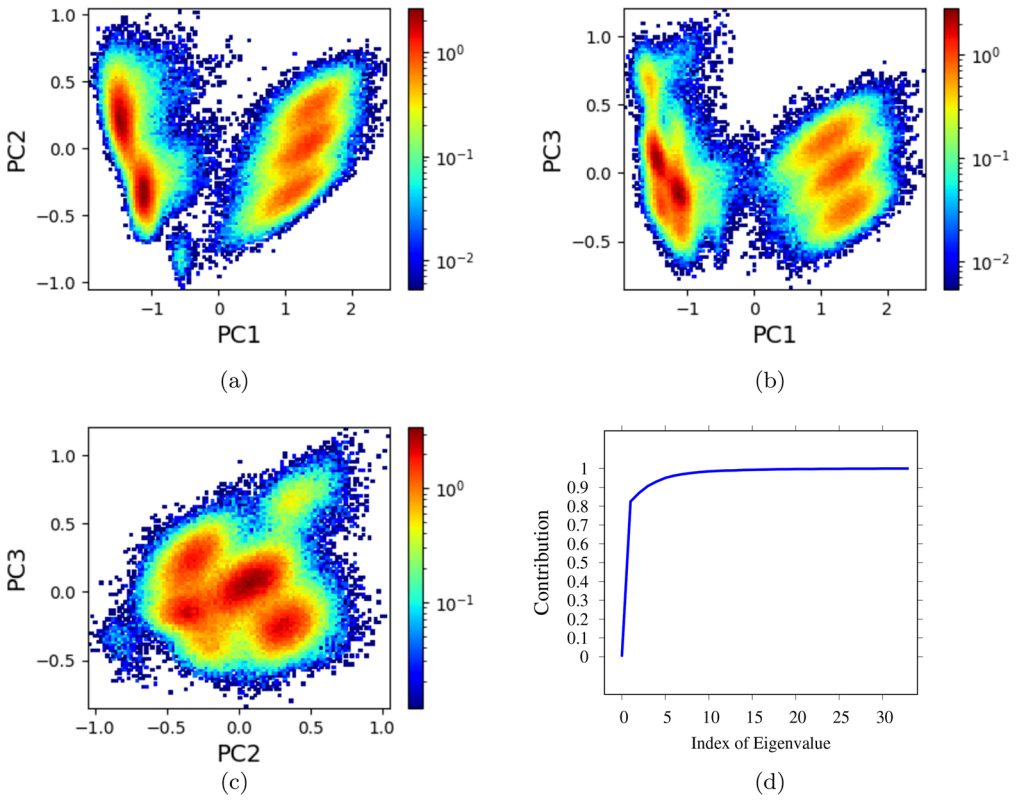


Figure 3. Logarithm of probability distribution projected into PC spaces based on T4 lysozyme simulations. The logarithm of probability of distribution spanned (a) PC1 and PC2, (b) PC1 and PC3, and (c) PC2 and PC3. (d) Contribution of eigenvalues to cumulative PCs. The contribution of PC1 is 0.83.

helix and the N-terminal region increases; in the negative direction, the hinge spreads while the distance between the fifth helix and the N-terminal region decreases. The hinge spreads and the fifth helix and N-terminal distance are in the positive direction (Figure 4b). Where two open and closed state regions meet in Figure 3a, both values of PC1 and PC2 are small, which means that the mouth begins to close, the hinge spreads, and the fifth helix and N-terminus begin to close. This implies that the decreasing of the distance between the fifth helix and the N-terminal can trigger the mouth closing/hinge spreading event. Furthermore, the same implication can be observed in Figure 3b. The PC3 corresponds to the combination of the distancing (closing) between the fifth helix and the N-terminal and the spreading (shrinking) motion of the hinge region. The hinge spreads and the fifth helix and N-terminal close in the positive direction (Figure 4c). Figure 3b shows that where two state regions meet the mouth closed, the hinge opened and the distance between the fifth helix and the N-terminal closed. These findings were observed for the other five data sets. (Figures corresponding to Figure 1 can be found in Figure S2.)

3.2. Holo and Apo States of PDZ3 Domain. The PDZ3 domain is an example of allostery without a significant conformational change in backbone structure. The hidden dynamical allostery was revealed by Lee et al., where the removal of the third helix (residues 394−399) at its C-terminal drastically reduces the binding affinity.[78] It has three α-helices and five β-strands; we refer to them as α1, α2, α3, β1, β2, ..., respectively. We performed three 1-μs-long MD simulations for both holo and apo states. We aimed to identify the residue−residue pairs that characterize the phenomena based on the trajectories from MD simulations. To achieve this, anomaly detection analyses based on the sparse structure learning technique were performed. This analysis used the same parameters as those used in the T4 lysozyme case. Anomaly detection was performed for each combination of the apo- and holo-state trajectories. Specifically, we analyzed three apo-state trajectories and three holo-state trajectories, resulting in nine sets of results corresponding to all possible combinations of these trajectories ($9 = 3 \times 3$). The number of initial input coordinates (pairs of residue−residue distances) was 731. We checked the number of nonzero elements of the sparse representation of the correlation of residue−residue distances. The number of nonzero elements was sufficiently small when the sparsity was set to $\geq$0.80.

Figure 5a shows the sparse representation of the correlation of residue−residue distances of the holo and apo states of the PDZ3 domain. This figure corresponds to the results of one of nine combinations of the apo-state and the holo-state simulations. (See Figure S4 for other combinations.) The colored points indicate one residue−residue pair (red and blue for the holo and apo states, respectively). The line connecting the red and blue points indicates that the distances of two pairs are correlated at the holo and apo states, respectively, as in the case of T4 lysozyme. The number of correlated pairs was 131 at the holo state and 113 at the apo state, when the sparsity was set at 0.90. Figure 5b shows the residue−residue pairs with anomalies. The density of the colored points corresponds to the sparsities of 0.80 and 0.90. The numbers of identified residue−residue pairs were 183 and 29, respectively.

Figure 6 visualizes 29 residue−residue pairs identified at a sparsity of 0.90 on two states of PDZ3 domain structures.

https://doi.org/10.1021/acsomega.4c11546
ACS Omega 2025, 10, 22789−22801

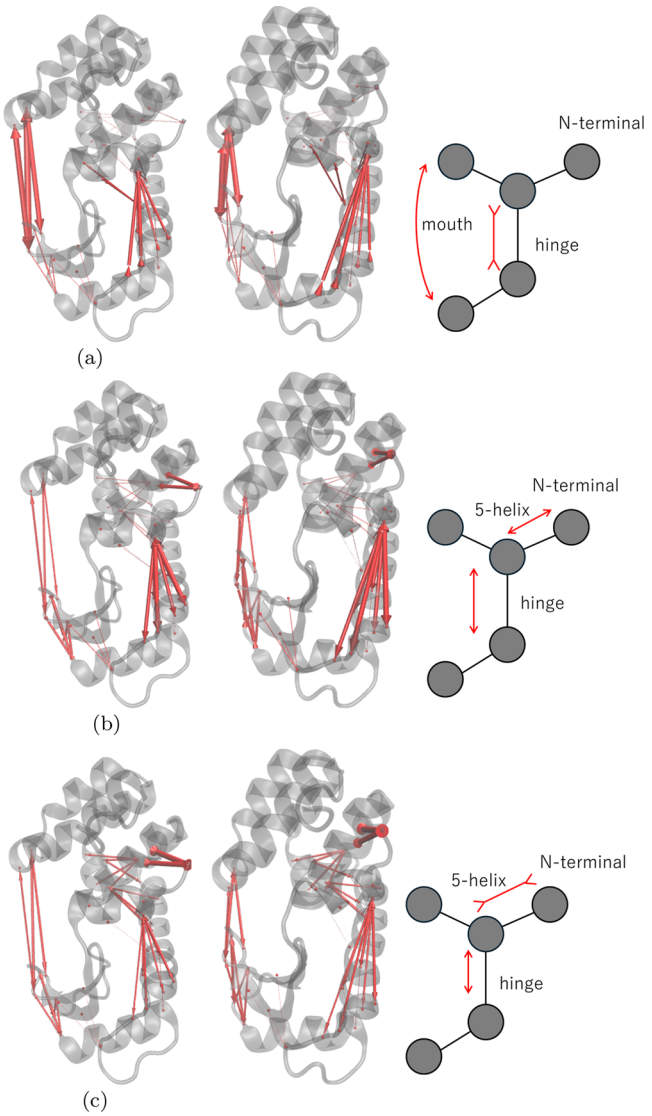


Figure 4. Components of PCs of the identified residue−residue distances of T4 lysozyme. Components of PC1 (a), PC2 (b), and PC3 (c) are visualized on the open-state (left) and closed-state (middle) structures of T4 lysozyme. The thickness of the line is proportional to the value of the element of the eigenvector, and the direction of the arrow indicates the sign of the element of the eigenvector. The diagrams represent simplified structures of T4 lysozyme and the arrows beside the diagrams are the main components of each PC (right). The direction of the arrows indicates that the component of PC is positive.

Among them, most connect the loop between β2 and β3 (residue 330−335) with α2 (residue 372−380), while other pairs connect the loop between β2 and β3 with α3 (residue 394−399), β3 (residue 336−341), and β5 (residue 385−392). Additionally, pairs connecting α3 with α2, β2, and β3 and the loop between β2 and β3 and pairs connecting α2 with the loop between β2 and β3 were observed. These observations show that our analysis detected the connections between the loop between β2 and β3, α2, and α3 and the connections in which they work as a hub. As mentioned in section 2.4, β2 and α2 consist of the groove bound to the ligand peptide, and experiments revealed that α3 is a relevant region to the allostery of the PDZ3 domain.[78]

To further explore the implication of the identified pairs, PCA was conducted on the distances of the identified pairs. As input data, we combined the time series of the residue−residue distances of 29 identified pairs from both holo- and apo-state simulations. Figure 7a−c gives the logarithm of the probability distribution of the holo- and apo-state trajectories projected on PC1 vs PC2 space, PC1 vs PC3 space, and PC2 and PC3 space, respectively. The contribution of PC1 is 0.426, the cumulative contribution of PC1 and PC2 is 0.626, and the cumulative contribution of PC1, PC2, and PC3 is 0.723 (Figure 7d).

The PC1, PC2, and PC3 components are visualized in Figure 8a−c, respectively. The thickness of the red line is proportional to the absolute value of the element of the eigenvector, and the direction of the arrow indicates the sign of the element of the eigenvector. PC1 includes the closing (distancing) motions between the loop between β2 and β3 to α2 and the distancing (closing) motion between β3 and α3. PC2 includes the closing (distancing) motion between the loop between β2 and β3, β4, and α2. PC3 is a rather minor motion, such as the N-terminal region. These findings were observed for the other eight data sets (figures corresponding to Figure 5 can be found in Figure S5).

PC1 consists of the network within the binding site (α2, the loop between β2 and β3) and α3. Only PC2 can divide the holo and apo states clearly: the apo state locates where PC2 is larger than around −0.2 and the holo state where PC2 is smaller than around −0.2. This is because PC2 consists of the network of the distances within the binding site (α2, the loop between β2 and β3 and β4), and the decrease in the value of PC2 indicates the core formation of the binding site. PC3 corresponds to the distance between the N-terminal region and the loop between α1 and β4 and indicates only that in the apo state the distance sometimes can be longer than that in the holo state.

4. DISCUSSION

We proposed a new method applying sparse structure learning-driven anomaly detection[61] on the trajectories of MD simulations to identify an important small number of elements associated with the transition between two states. Our first target was the open-close transition of T4 lysozyme. Analysis of six cases using independent trajectories of the open and closed states detected 23 to 44 residue−residue pairs with high anomaly when the sparsity was set at 0.90 (Figure S2). This small number of pairs is concentrated around the "mouth" and "hinge" regions of the open-closed motion and implies that our method can identify the features associated with the transition motion of two states. PCA was performed for the residue−residue distances of the identified pairs to further investigate this implication. PC1 (whose contribution was over 80%) divides the open and closed states and corresponds to the cooperative motion of the opening (closing) of the mouth and shrinking (spreading) of the hinge. This suggests that PCA provides a single meaningful axis that can describe the transition motion. Results similar to those of previous studies were observed.[41,57,58] The comprehensive work on opening and closing motions of T4 lysozyme[58] using a large amount of simulations and the detailed analysis show the open-to-closed state as follows: (1) distancing between Glu5 and Lys60 leading to the exposure of Phe4, (2) closing between Phe4 in the first helix and Phe104 in the fifth and third helix, (3) straightening of the third helix at Phe67, causing the closing

https://doi.org/10.1021/acsomega.4c11546
ACS Omega 2025, 10, 22789−22801

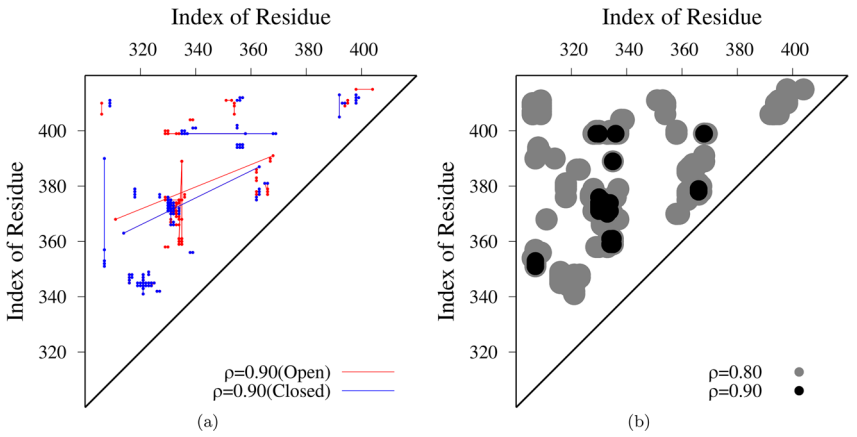


Figure 5. Sparse structure and residue−residue pairs with high anomalies of the PDZ3 domain. (a) Sparse representation of correlation of residue−residue distances of the holo state and apo state of the PDZ3 domain. Each axis indicates the indices of the residues. Each point indicates one residue−residue pair. A line that connects points indicates that the method revealed a correlation of the corresponding residue−residue distances when the sparsity was set at 0.90. Red indicates the holo state, and blue indicates the apo state. (b) Residue−residue pairs with high anomalies between the holo and apo states of the PDZ3 domain. The density of the points corresponds to the sparsity which was used for anomaly detection (ρ = 0.80 and ρ = 0.90).

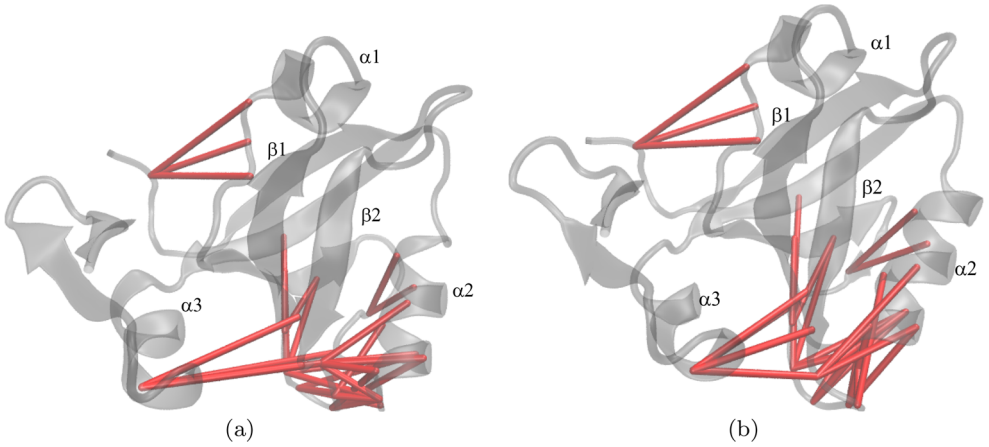


Figure 6. Structures of the PDZ3 domain with anomaly residue−residue pairs. (a) The 29 residue−residue pairs with anomalies detected at ρ = 0.90 are projected on the stable structure in the holo state. (b) The same pairs are projected on the stable structure in the apo state.

between Phe4 and Phe67, (4) opening of the mouth (forming the salt bridges: Glu22-Arg137, Asp20-Arg145, and Glu11-Arg145) and closing of the hinge (forming the salt bridges: Arg14-Asp20 and Glu11-Arg14), and (5) the rearrangement of β1. In our analysis using a sparsity of 0.90, we identified the distances between the residues in helix 1 and N-terminal residues (residue pairs: Met1-Arg96 and Met1-Cys97). The motions were included in the higher PC (PC2 or PC3) in all six cases. This can be seen corresponding to the second stage of the scenario. The distances between helix 1 and Phe67 in helix 3 (residue pairs: Met1-Phe67, Arg8-Phe67, and Ile9-Phe67) were observed in all cases, which seems to correspond to the third stage. As expected, the corresponding opening/closing motion of the mouth and hinge were identified in all six cases (residue pairs: Glu22-Gln141 and Phe4-Lys60).

The second case is dynamic allostery induced by peptide binding of the PDZ3 domain. This example is well-studied, and previous studies using correlation analysis and mutation MD simulations reported that the interactions between the loop between β2-β3 and α3 are important.[81] As mentioned in section. 3.2, our analysis identified the correlation of the distance between the loop between β2-β3 and α2 and distance α2 and α3. In both cases, our proposed method shows reasonable results that are consistent with previous findings.

We consider this proposed method a type of "feature-finding" method.[56−60] Such approaches can provide the basis for methods enhancing the applicability and interpretability of MD simulations including efficient sampling and CV-finding methods. A key advantage of our method is its computational efficiency compared with other feature-finding methods because it requires fewer resources for simulations. It allows for the comparison of time-series data from two different states, enabling the discovery of important features even when state transitions exceed simulation capabilities or when only relatively short trajectories are available. For example, our method, which relies on relatively short simulations (1 μs) from different states, allows us to obtain the results more easily than methods that analyze one longer simulation (50 μs)[57] which must involve the state transition. Based on the results from two test cases, our method can not only identify a small number of important features but also provide clearer insights for the state transition when combined with dimension reduction techniques. Additionally, the most time-consuming part of the analysis, the sparse structure learning of time series,

https://doi.org/10.1021/acsomega.4c11546
ACS Omega 2025, 10, 22789−22801

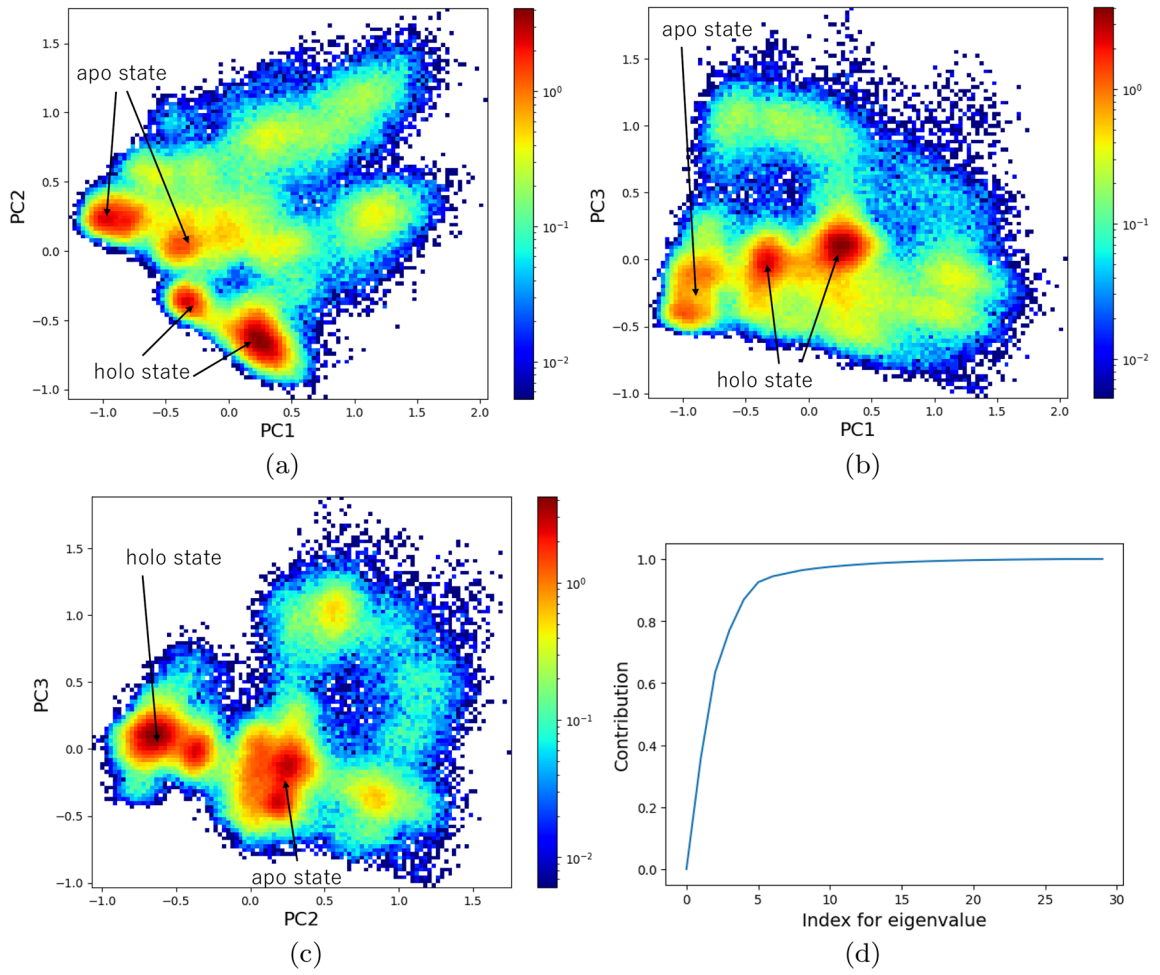


Figure 7. Logarithm of probability distribution projected on PC spaces based on PDZ3 domain simulations. The logarithm of probability distribution spanned along (a) PC1 and PC2, (b) PC1 and PC3, and (c) PC2 and PC3. (d) The contribution of eigenvalues to cumulative PCs. The contribution of PC1 is 0.426.

was completed within a few hours for each case using one CPU. As a further application, we are considering integrating our detection method with enhanced sampling techniques, such as metadynamics or TAMD, which require predefined collective variables.

Since MoSAIC is also a feature-finding method based on the correlation relationship between input coordinates, we will now compare it with our method.[57] Overall, the results of MoSAIC corresponded well with those of our anomaly detection method and its combination with PCA. In that study, the same enzyme T4 lysozyme was targeted as in our research. A 50 μs MD trajectory including the open-closed transition was adopted for the MoSAIC analyses with parameters at γ = 0.5 and n = 5. Three main motions were found as the clusters. Cluster 1 corresponds to the antiparallel motion of opening/closing motion of mouth and hinge, which corresponds to PC1 in our analysis. Cluster 2 is related to the rearrangement around the N-terminal region, which corresponds to the distancing between the N-terminal region and the fifth helix included in PC2 and PC3 in our analysis. Cluster 3 identified a twisting motion between the β-sheets and the second helix. Similarly, in our analysis, anomaly detection identified the set of residue pairs with high anomalies around the first and second strand and the second helix. Therefore, we conclude that the results of anomaly detection or the combination with PCA correspond well to those obtained by MoSAIC applied to a much longer trajectory.

It may be helpful to mention whether meaningful results could have been obtained without combining with PCA. In both cases, the combination of the feature-finding method and PCA allowed us to obtain a small number (2−3) of modes, which leads to a convincing explanation of the transition. If anomaly detection alone had produced a similarly small number of features by increasing the sparsity ρ, the resulting information would have less information. For instance, in the T4 lysozyme case, when the sparsity was increased to 0.92, three features (the residue−residue pairs) were obtained (residues 4−63, 4−62, and 4−50). However, they were all associated with the open/close motion of the mouth of T4 lysozyme, and information about the spreading/shrinking hinge or the rearrangement of the N-terminal region were lacking. In the PDZ3 domain case, when ρ was increased up to 0.93, only four residue−residue pairs (residues 335−389, 335−361, 335−359, and 334−359) remained. These pairs correspond to the pairs between β3 and β4, between β3 and β5, and between the loop between β2 and β3 and β4, and we could not ascertain the implications of α3, which is known to be important for dynamic allostery. Therefore, it can be concluded that the combination of the feature-finding method with dimension reduction techniques, such as PCA, is effective.

https://doi.org/10.1021/acsomega.4c11546
ACS Omega 2025, 10, 22789−22801

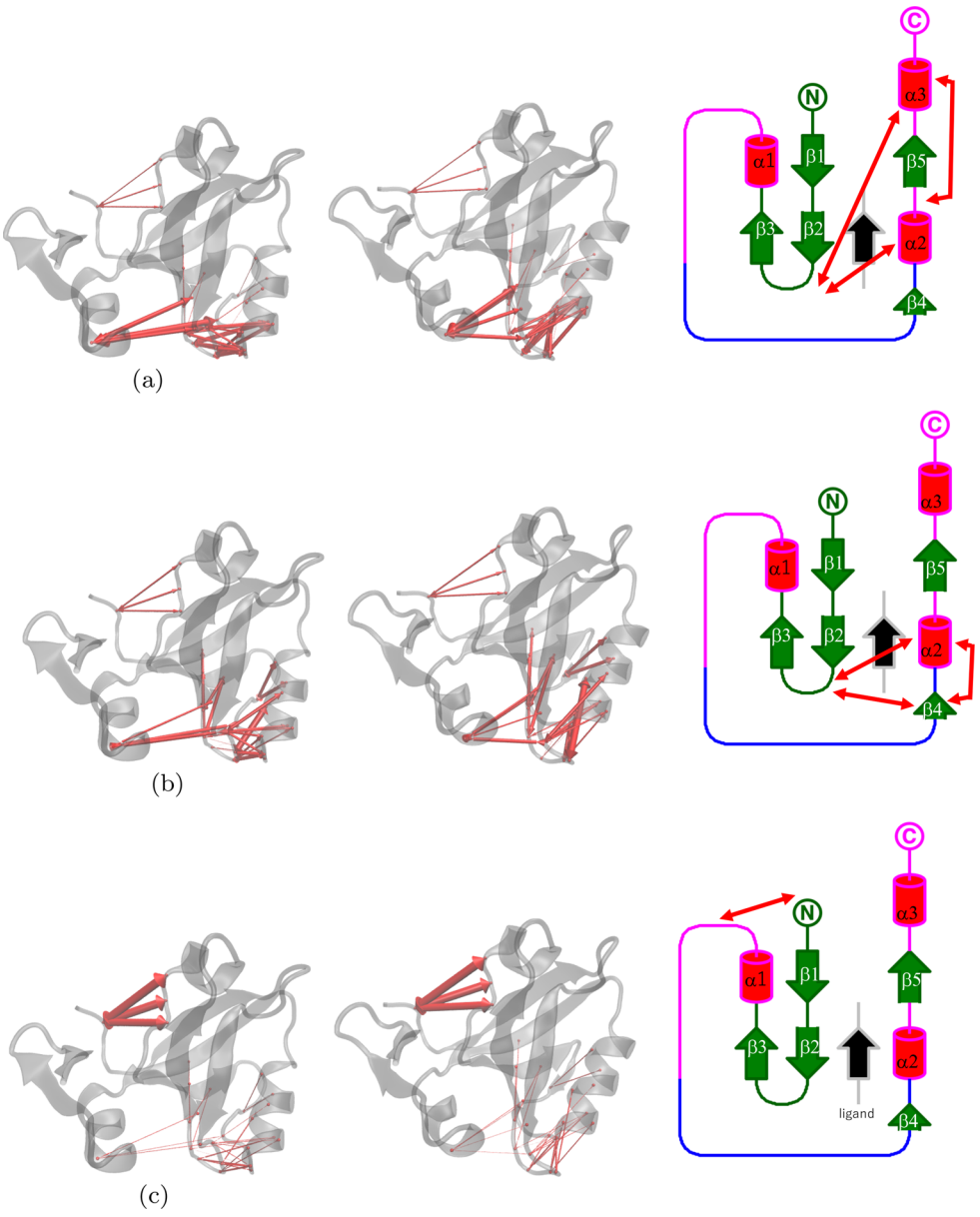


Figure 8. Components of PCs of the identified residue−residue distances of the PDZ3 domain. Components of PC1 (a), PC2 (b), and PC3 (c) are visualized on the holo-state (left) and the apo-state (middle) structures of the PDZ3 domain, respectively (only protein structures without ligand peptide are shown). The thickness of the line is proportional to the value of the element of the eigenvector, and the direction of the arrow indicates the sign of the element of the eigenvector. The diagrams represent simplified structures of the PDZ3 domain and the arrows are the main component(s) of each PC (right).

We justify our method for setting the suitable value of ρ: lowering the sparsity value considerably leads to an excessive number of features, which makes interpretation difficult, while increasing by too much risks omitting important features. In this study, we increased the sparsity value until the number of selected features reached a reasonable range, typically around a few dozen. From our trial-and-error, we suggested a sparsity value of 0.90 as the initial value.

Considering how the sparsity value is set may lead us to the following insight: adjusting the sparsity enables the selection of only essential features under high sparsity while allowing the capture of local motion under low sparsity condition. We think that the interpretation is correct in some cases. By setting a higher sparsity value (larger ρ), we can indeed extract only the most relevant features describing the core biological process (for example, the open-close motion of mouth in the case of T4 lysozyme). Conversely, the smaller sparsity allows the inclusion of additional features, such as the spreading/shrinking of the hinge or the rearrangement of the N-terminal region in the case of T4 lysozyme. However, we cannot establish that the value of sparsity alone can directly quantify the biological significance of the selected features. Instead, it provides us with a means to balance interpretability and completeness by fine-tuning the number of extracted features. While a higher sparsity highlights only the most dominant features, a lower sparsity may reveal additional mechanistic details. Thus, sparsity adjustment serves as a useful tool for refining our understanding of the system, but biological validation is still necessary to confirm the functional relevance of the identified features.

https://doi.org/10.1021/acsomega.4c11546
ACS Omega 2025, 10, 22789−22801

Finally, we mention the applicability of our method to other systems and its flexibility for the selection of the state. We believe that our method is applicable to systems that undergo structural changes, including cooperative motions when the selected states are relatively stable. In some cases, it is expected that 1 μs of length will be insufficient to achieve a converged sparse structure. For example, we also conducted our analysis for the third target: cavity formation of myeloid differentiation protein 2 (MD2). The combination of anomaly detection and PCA captured the cavity-formation motion; however, when the length of the simulations was set to equal to two other cases, the sparse structures obtained from the MD2 simulation were not as robust as those of the T4 lysozyme and PDZ3 domain. The results suggested, in the case of MD2, 1 μs length was insufficient to achieve a converged sparse structure (Figures S6–S8). We present the general approach used to set the length of the simulation and determine its sufficiency. One possibility extends the simulation length until a convergence is reached. The other possibility involves running independent simulations and assessing the consistency of the obtained results.

5. CONCLUSION

The use of MD simulation in the studies of biomolecules has become increasingly prevalent owing to advancements in experimental structure-determination techniques and machine-learning-based structure prediction methods. However, interpretation of results and sampling problems remain critical. Recently, interest has increased in applying machine learning and associated techniques to sampling and analysis processes. In most cases, both problems are related to the challenges of finding collective variables. CV-finding requires deciding the set of input coordinates, and it leads to interest in capturing a small number of coordinates (sometimes called "features") that can describe a specific biological process. In this Letter, we introduce a novel approach using an anomaly detection method based on sparse structure learning to compare MD trajectories, facilitating the automatic extraction of important features. The application of this method to the open-close states of T4 lysozyme and the holo and apo states of the PDZ3 domain illustrates its capability to identify coordinates that delineate differences in each state and the usefulness of combining it with dimension reduction methods such as PCA. Moreover, this method can function as a preprocessing step for enhanced sampling techniques like metadynamics, thereby contributing to improved efficiency and automation of studies using MD simulation.

■ ASSOCIATED CONTENT

Data Availability Statement

The program for anomaly detection based on sparse structure learning is available via our github repository (https://github.com/fazzz/anom_ssl2). To compile the program, The LAPACK library is required, and the basic usage is as follows: `assl2 -Sparsity r -ndim n time-series-A time-series-B > out`, where you can specify the sparsity using the option "-Sparsity", dimension of time series using the option "-ndim", "time-series-A" and "time-series-B" refers to the TEXT files that contains *n* dimensional time series data. The detailed description of usage is available in the github page. MD trajectories of the T4 lysozyme and the PDZ3 domain are available from the following figshare repositories: https://figshare.com/articles/dataset/T4_lysozyme_1-micro_second_x_8_trajectories/14920767 and https://figshare.com/articles/dataset/PDZ3_domain_1micro_second_simulation_x_6_trajectories/24864603. Each repository contains 1 μs trajectory stored every 10 ns in XTC format which only contains the coordinates of the protein and input files of GROMACS including mdp files for the production run, topology file, and initial structure files.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c11546.

Sparse structure of T4 lysozyme from six combinations of trajectories, residue−residue pairs with the high anomaly of T4 lysozyme from six combinations of trajectories, sparse structure of T4 lysozyme when the threshold for inter-residue distance was 4.5 Å, sparse structure of the PDZ3 domain from nine combinations of trajectories, residue−residue pairs with the high anomaly of the PDZ3 domain from nine combinations of trajectories, sparse structures of MD2 from nine combinations of trajectories, residue−residue pairs with the high anomaly of MD2 from nine combinations of trajectories, and components of PCs of the identified residue−residue distances of MD2 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Yu Yamamori − *Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan;* Present Address: Center for Computational Science (CCS), University of Tsukuba, 1-1-1, Tennodai Tsukuba, Ibaraki 305-8577, Japan. Email: yamamori@ccs.tsukuba.ac.jp; orcid.org/0000-0003-2854-212X; Email: yu.yamamori@aist.go.jp

Author

Kentaro Tomii − *Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan;* orcid.org/0000-0002-4567-4768

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c11546

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was partially supported by the Research Support Project for Life Science and Drug Discovery (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED under Grant Number JP24ama121028. We would like to thank Editage (www.editage.jp) for English language editing.

■ REFERENCES

(1) Leimkuhler, B.; Matthews, C. *Molecular Dynamics; Interdisciplinary Applied Mathematics*; Springer International Publishing: 2015.
(2) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99*, 1129−1143.
(3) Shaw, E. A. *International symposium on High-performance parallel and distributed computing*; 2013, pp 129−130.

(4) Shaw, D. E. Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. *International Conference for High Performance Computing, Networking, Storage and Analysis*; 2014, pp 41−53.
(5) Shaw, D. E. Twenty Microseconds of Molecular Dynamics Simulation before Lunch. *International Conference for High Performance Computing, Networking, Storage and Analysis*; 2021, pp 14−19.
(6) Jung, J.; Kobayashi, C.; Kasahara, K.; Tan, C.; Kuroda, A.; Minami, K.; Ishiduki, S.; Nishiki, T.; Inoue, H.; Ishikawa, Y.; Feig, M.; Sugita, Y. New parallel computing algorithm of molecular dynamics for extremely huge scale biological systems. *J. Comput. Chem.* **2021**, *42*, 231−241.
(7) Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878−3888.
(8) Stone, J. E.; Hallock, M. J.; Phillips, J. C.; Peterson, J. R.; Luthey-Schulten, Z.; Schulten, K. Evaluation of emerging energy-efficient heterogeneous computing platforms for biomolecular and cellular simulation workloads. *International Parallel and Distributed Processing Symposium*; 2016, pp 89−100.
(9) Yu, I.; Mori, T.; Ando, T.; Harada, R.; Jung, J.; Sugita, Y.; Feig, M. Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. *eLife* **2016**, *5*, e19274.
(10) Feig, M.; Sugita, Y. Whole-Cell Models and Simulations in Molecular Detail. *Annu. Rev. Cell Dev. Biol.* **2019**, *35*, 191−211.
(11) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 5915−5920.
(12) Fernandez-Leiro, R.; Scheres, S. H. Unravelling biological macromolecules with cryo-electron microscopy. *Nature* **2016**, *537*, 339−346.
(13) Jumper, J.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.
(14) Baek, M.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871−876.
(15) Ahdritz, G. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nat Methods* **2024**, *21*, 1514.
(16) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123−1130.
(17) Arantes, P. R.; Polêto, M. D.; Pedebos, C.; Ligabue-Braun, R. Making it Rain: Cloud-Based Molecular Simulations for Everyone. *J. Chem. Inf. Model.* **2021**, *61*, 4852−4856.
(18) Bernardi, R. C.; Melo, M. C.; Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta* **2015**, *1850*, 872−877.
(19) Sidky, H.; Chen, W.; Ferguson, A. L. Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation. *Mol. Phys.* **2020**, *118*, 1737742.
(20) Hénin, J.; Lelièvre, T.; Shirts, M. R.; Valsson, O.; Delemotte, L. Enhanced sampling methods for molecular dynamics simulations. *Living Journal of Computational Molecular Science* **2022**, *4*, 1583.
(21) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.
(22) Nakajima, N.; Nakamura, H.; Kidera, A. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *J. Phys. Chem. B* **1997**, *101*, 817−824.
(23) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.
(24) Sugita, Y.; Kitao, A.; Okamoto, Y. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.* **2000**, *113*, 6042.
(25) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13749−13754.
(26) Wang, L.; Friesner, R. A.; Berne, B. J. Replica exchange with solute scaling: A more efficient version of replica exchange with solute tempering (REST2). *J. Phys. Chem. B* **2011**, *115*, 9431−9438.
(27) Laio, A.; Parrinello, M. *Escaping free-energy minima* **2002**, *99*, 12562−12566.
(28) Laio, A.; Rodriguez-Fortea, A.; Gervasio, F. L.; Ceccarelli, M.; Parrinello, M. Assessing the accuracy of metadynamics. *J. Phys. Chem. B* **2005**, *109*, 6714−6721.
(29) Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **2008**, *100*, 020603.
(30) Maragliano, L.; Vanden-Eijnden, E. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett.* **2006**, *426*, 168−175.
(31) Yamamori, Y.; Kitao, A. MuSTAR MD: multi-scale sampling using temperature accelerated and replica exchange molecular dynamics. *J. Chem. Phys.* **2013**, *139*, 145105.
(32) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361−390.
(33) Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.* **2021**, *121*, 9722−9758.
(34) 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.
(35) García, A. E. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* **1992**, *68*, 2696.
(36) Amadei, A.; Linssen, A. B.; Berendsen, H. J. Essential dynamics of proteins. *Proteins: Struct., Funct., Bioinf.* **1993**, *17*, 412−425.
(37) Hayward, S.; Kitao, A.; Go, N. Harmonicity and anharmonicity in protein dynamics: A normal mode analysis and principal component analysis. *Proteins: Struct., Funct., Bioinf.* **1995**, *23*, 177−186.
(38) Mu, Y. G.; Nguyen, P. H.; Stock, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins: Structure Function and Bioinformatics* **2005**, *58*, 45−52.
(39) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chem. Phys.* **2007**, *126*, 244111.
(40) Abseher, R.; Nilges, M. Are there non-trivial dynamic cross-correlations in proteins? *J. Mol. Biol.* **1998**, *279*, 911−920.
(41) Ernst, M.; Sittel, F.; Stock, G. Contact- and distance-based principal component analysis of protein dynamics. *J. Chem. Phys.* **2015**, *143*, 244114.
(42) Koyama, Y. M.; Kobayashi, T. J.; Tomoda, S.; Ueda, H. R. Perturbational formulation of principal component analysis in molecular dynamics simulation. *Phys. Rev. E* **2008**, *78*, 046702.
(43) Naritomi, Y.; Fuchigami, S. Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis. *J. Chem. Phys.* **2013**, *139*, 215102.
(44) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*, 015102.
(45) McGibbon, R. T.; Husic, B. E.; Pande, V. S. Identification of simple reaction coordinates from complex dynamics. *J. Chem. Phys.* **2017**, *146*, 044109.
(46) Schwantes, C. R.; Pande, V. S. Modeling molecular kinetics with tICA and the kernel trick. *J. Chem. Theory Comput.* **2015**, *11*, 600−608.

https://doi.org/10.1021/acsomega.4c11546
ACS Omega 2025, 10, 22789−22801

(47) Mitsutake, A.; Takano, H. Relaxation mode analysis for molecular dynamics simulations of proteins. *Biophys. Rev.* **2018**, *10*, 375−389.
(48) Nagai, T.; Mitsutake, A.; Takano, H. Principal Component Relaxation Mode Analysis of an All-Atom Molecular Dynamics Simulation of Human Lysozyme. *J. Phys. Soc. Jpn.* **2013**, *82*, 023803.
(49) Chen, K. K.; Tu, J. H.; Rowley, C. W. Variants of dynamic mode decomposition: Boundary condition, Koopman, and fourier analyses. *J. Nonlinear Sci.* **2012**, *22*, 887−915.
(50) Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 9885−9890.
(51) Preto, J.; Clementi, C. Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. *Phyical Chemistry and Chemical Physics* **2014**, *16*, 19181−19191.
(52) Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 13597−13602.
(53) Kattnig, D. R.; Nielsen, C.; Solov'Yov, I. A. Molecular dynamics simulations disclose early stages of the photo-activation of cryptochrome 4. *New J. Phys.* **2018**, *20*, 083018.
(54) Wehmeyer, C.; Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.* **2018**, *148*, 241703.
(55) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **2018**, *9*, 5.
(56) Ravindra, P.; Smith, Z.; Tiwary, P. Automatic mutual information noise omission (AMINO): generating order parameters for molecular systems. *Mol. Syst. Des. Eng.* **2020**, *5*, 339−348.
(57) Diez, G.; Nagel, D.; Stock, G. Correlation-Based Feature Selection to Identify Functional Dynamics in Proteins. *J. Chem. Theory Comput.* **2022**, *18*, 5079−5088.
(58) Post, M.; Lickert, B.; Diez, G.; Wolf, S.; Stock, G. Cooperative Protein Allosteric Transition Mediated by a Fluctuating Transmission Network. *J. Mol. Biol.* **2022**, *434*, 167679.
(59) Hub, J. S.; de Groot, B. L. Detection of Functional Modes in Protein Dynamics. *PLoS Comput. Biol.* **2009**, *5*, e1000480.
(60) Bai, F.; Puk, K. M.; Liu, J.; Zhou, H.; Tao, P.; Zhou, W.; Wang, S. Sparse group selection and analysis of function-related residue for protein-state recognition. *J. Comput. Chem.* **2022**, *43*, 1342−1354.
(61) Idé, T.; Lozano, A. C.; Abe, N.; Liu, Y. Proximity-Based Anomaly Detection using Sparse Structure Learning. *Proceedings of the 2009 SIAM International Conference on Data Mining*; 2009, *1*, pp 97−108.
(62) Serçinoğlu, O.; Ozbek, P. gRINN: a tool for calculation of residue interaction energies and protein energy network analysis of molecular dynamics simulations. *Nucleic Acids Res.* **2018**, *46*, W554−W562.
(63) Contreras-Riquelme, S.; Garate, J.-A.; Perez-Acle, T.; Martin, A. J. M. RIP-MD: a tool to study residue interaction networks in protein molecular dynamics. *PeerJ.* **2018**, *6*, e5998.
(64) Berendsen, H. J.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **1995**, *91*, 43−56.
(65) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1−2*, 19−25.
(66) 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.
(67) Miyamoto, S.; Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952−962.
(68) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463.
(69) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999−2012.
(70) 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.
(71) Schmit, J. D.; Kariyawasam, N. L.; Needham, V.; Smith, P. E. SLTCAP: A Simple Method for Calculating the Number of Ions Needed for MD Simulation. *J. Chem. Theory Comput.* **2018**, *14*, 1823.
(72) Parrinello, M.; Rahman, A. Crystal Structure and Pair Potentials: A Molecular-Dynamics Study. *Phys. Rev. Lett.* **1980**, *45*, 1196.
(73) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.* **1999**, *20*, 786−798.
(74) Roe, D. R.; Brooks, B. R. A protocol for preparing explicitly solvated systems for stable molecular dynamics simulations. *J. Chem. Phys.* **2020**, *153*, 054123.
(75) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L. P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528−1532.
(76) Zhang, X. J.; Matthews, B. W. Conservation of solvent-binding sites in 10 crystal forms of T4 lysozyme. *Protein Sci.* **1994**, *3*, 1031−1039.
(77) Weaver, L. H.; Matthews, B. W. Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. *J. Mol. Biol.* **1987**, *193*, 189−199.
(78) Petit, C. M.; Zhang, J.; Sapienza, P. J.; Fuentes, E. J.; Lee, A. L. *Hidden dynamic allostery in a PDZ. domain* **2009**, *106*, 18249−18254.
(79) Stevens, A. O.; He, Y. Allosterism in the PDZ Family. *Int. J. Mol. Sci.* **2022**, *23*, 1454.
(80) Ali, A. A.; Gulzar, A.; Wolf, S.; Stock, G. Nonequilibrium Modeling of the Elementary Step in PDZ3 Allosteric Communication. J. Phys. *Chem. Lett.* **2022**, *13*, 9862−9868.
(81) Vargas-Rosales, P. A.; Caflisch, A. Domino Effect in Allosteric Signaling of Peptide Binding. *J. Mol. Biol.* **2022**, *434*, 167661.
(82) Kumawat, A.; Chakrabarty, S. Hidden electrostatic basis of dynamic allostery in a PDZ domain. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E5825−E5834.
(83) Kumawat, A.; Chakrabarty, S. Protonation-Induced Dynamic Allostery in PDZ Domain: Evidence of Perturbation-Independent Universal Response Network. *J. Phys. Chem. Lett.* **2020**, *11*, 9026−9031.
(84) Doyle, D. A.; Lee, A.; Lewis, J.; Kim, E.; Sheng, M.; MacKinnon, R. Crystal Structures of a Complexed and Peptide-Free Membrane Protein−Binding Domain: Molecular Basis of Peptide Recognition by PDZ. *Cell* **1996**, *85*, 1067−1076.

https://doi.org/10.1021/acsomega.4c11546
ACS Omega 2025, 10, 22789−22801