Unravel Allostery Mechanism Through Molecular Dynamic Simulations with Machine Learning Approaches

Hongyu Zhou
Southern Methodist University, hongyuz@smu.edu

Follow this and additional works at: https://scholar.smu.edu/hum_sci_chemistry_etds

Part of the Other Chemistry Commons

Recommended Citation
https://scholar.smu.edu/hum_sci_chemistry_etds/9

This Dissertation is brought to you for free and open access by the Chemistry at SMU Scholar. It has been accepted for inclusion in Chemistry Theses and Dissertations by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.
UNRAVEL ALLOSTERY MECHANISM
THROUGH MOLECULAR DYNAMIC SIMULATIONS
WITH MACHINE LEARNING APPROACHES

Approved by:

Prof. Peng Tao
Assistant Professor

Prof. Elfi Kraka
Professor

Prof. Brian Zoltowski
Associate Professor

Prof. Shouyi Wang
Associate Professor
UNRAVEL ALLOSTERY MECHANISM
THROUGH MOLECULAR DYNAMIC SIMULATIONS
WITH MACHINE LEARNING APPROACHES

A Dissertation Presented to the Graduate Faculty of
Dedman College
Southern Methodist University
in
Partial Fulfillment of the Requirements
for the degree of
Doctor of Philosophy
with a
Major in Theoretical and Computational Chemistry
by
Hongyu Zhou
B.A., Pharmaceutical Science, Tianjin University, China

May 18, 2019
ACKNOWLEDGMENTS

This work could not have been accomplished without the experience of my wonderful advisor, Prof. Peng Tao, along with my groupmates here at SMU including Zilin Song, Zheng Dong, Li Shen, Feng Wang and Francesco Trozzi. I'm also forever grateful to my family for being so patient with me and my friends for reminding me to take breaks and to be great instead of just reading about greatness all of the time.
Allostery is referred to as protein function changes due to external perturbations, and ubiquitous in the living cells. Investigation of the allosteric mechanism is essential for developing drugs or assisting protein engineering. Recently, the population shift model has been proposed to explain the allosteric mechanism. This model emphasizes the importance of conformational distribution in the allostery. In the past several years, I have been continuously working on developing computational methods to explore and quantify protein allostery mechanism. By fully utilizing protein simulation results, several machine learning related methods have been applied to protein systems to investigate the underlying conformational changes in different allosteric states. These methods include Rigid Residue Scan (RRS), Machine Learning based Classification, Relative Entropy based Dynamic Allosteric Network, Markov State Models and Directed Kinetic Transition Networks. These models investigate allostery from three different perspectives, including the thermodynamic, the kinetics and the structures.

Rigid Residue Scan (RRS) has been applied as the initial attempt to search for important residues. The importance of configurational entropy is emphasized in the RRS method. Key allosteric residues are identified through the comparison between unperturbed and perturbed simulations. After RRS, several machine-learning models, including artificial neural networks, random forest, and support vector machines, have been applied for a classification of various conformational ensembles. Among them, neural network models have the best performance in distinguishing different states. The random forest models align best with the population shift hypothesis, which demonstrate that allosteric states could be altered from one distribution to
another upon an external perturbation. Random forest models can also provide the feature importance for each feature. These important features align well with other experimental studies.

To incorporate both structural and thermodynamic view of allostery, an allosteric network model has also been developed to model the protein as the network where each residue represents different vertex. The significance of an allosteric effect upon a perturbation is treated as the weight of each edge connecting residues (vertices). The shortest pathway algorithms and the community detections algorithms have been applied here to separate the protein into several allosteric communities. These analyses could provide several possible communication pathways and allosteric communities. Based on these results, possible allosteric mechanism can be unraveled.

Furthermore, the kinetics of allostery has been investigated through Markov State Models (MSMs). The MSMs represent protein structures as individual Markov states. Following the kinetic theory of protein folding, each state represents a different local minimum on the free energy surface. By estimating the transition probability matrix, a steady state distribution (equilibrium) of each Markov state can be quantified. DKTN model is a further improvement of MSMs. The underlying assumption of DKTN model is replaced as the rate theory instead of the Markovian property. Analogy from DKTN to Continuous-Time Markov Chain (CTMC) demonstrates that the MSM is a special case of DKTN, and DKTN model could evolve into real equilibrium.

Overall, allosteric mechanisms could be thoroughly studied using the above models through three different perspectives, including the thermodynamics, the kinetics and the structures. These models have been applied on various proteins. A good agreement from our studies with the experimental results suggests an efficiency and validity of our models. The possible allosteric mechanism for the Light-Oxygen-Voltage (LOV) domain protein VIVID has been proposed. In summary, these models enhanced the understanding of the allosteric mechanism in terms of conformational distributions, and could be applied as a standard toolbox for studying the allosteric mechanism.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>............................................................. ix</td>
<td></td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>.................................................................. xii</td>
<td></td>
</tr>
<tr>
<td>LIST OF SYMBOLS AND ACRONYMS</td>
<td>............................................................... xiii</td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>Introduction</td>
<td>................................................................. 1</td>
</tr>
<tr>
<td>1.1</td>
<td>Background to Allostery</td>
<td>............................................................... 1</td>
</tr>
<tr>
<td>1.2</td>
<td>Molecular Dynamic Simulation</td>
<td>.............................................................. 4</td>
</tr>
<tr>
<td>2.</td>
<td>Methodologies</td>
<td>.................................................................. 7</td>
</tr>
<tr>
<td>3.</td>
<td>Identify Important Residues Using Rigid Residue Scan (RRS)</td>
<td>.......................................................... 13</td>
</tr>
<tr>
<td>3.1</td>
<td>RRS on Model Proteins: PDZ Domain</td>
<td>............................................................... 13</td>
</tr>
<tr>
<td>3.2</td>
<td>Application of RRS Method on VVD Protein</td>
<td>.......................................................... 16</td>
</tr>
<tr>
<td>4.</td>
<td>Distribution and Population-Shift: Random Forest Model</td>
<td>....................................................... 20</td>
</tr>
<tr>
<td>4.1</td>
<td>Population Shift Model</td>
<td>.................................................................. 20</td>
</tr>
<tr>
<td>4.2</td>
<td>Machine Learning: Random Forest</td>
<td>................................................................. 22</td>
</tr>
<tr>
<td>4.3</td>
<td>Classification Using PDZ2</td>
<td>.................................................................. 24</td>
</tr>
<tr>
<td>5.</td>
<td>Network Model in Proteins: Allosteric Network</td>
<td>......................................................... 28</td>
</tr>
<tr>
<td>5.1</td>
<td>Network Model in Proteins</td>
<td>.................................................................. 28</td>
</tr>
<tr>
<td>5.2</td>
<td>Relative Entropy based Dynamical Allosteric Network Model (REDAN)</td>
<td>................. 29</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>5.3</td>
<td>Model Protein System: PDZ2</td>
<td>30</td>
</tr>
<tr>
<td>6.</td>
<td>Markov Chain in Protein: Markov State Model</td>
<td>35</td>
</tr>
<tr>
<td>6.1</td>
<td>Markov State Models (MSMs) Based on Molecular Dynamics Simulation</td>
<td>35</td>
</tr>
<tr>
<td>6.2</td>
<td>Application of MSM on VVD</td>
<td>37</td>
</tr>
<tr>
<td>7.</td>
<td>Directed Kinetic Transition Network (DKTN) Model</td>
<td>46</td>
</tr>
<tr>
<td>7.1</td>
<td>Limitation of Discrete Markov State Model</td>
<td>46</td>
</tr>
<tr>
<td>7.2</td>
<td>Transition Kinetic Model and Continuous Time Markov Chain (CTMC)</td>
<td>48</td>
</tr>
<tr>
<td>7.3</td>
<td>Application of DKTN on VVD protein system</td>
<td>58</td>
</tr>
<tr>
<td>8.</td>
<td>Other Projects: Related to Allostery</td>
<td>63</td>
</tr>
<tr>
<td>8.1</td>
<td>Dimensionality Reduction and t-SNE</td>
<td>63</td>
</tr>
<tr>
<td>8.2</td>
<td>Direct Pathway Dynamical Sampling</td>
<td>69</td>
</tr>
<tr>
<td>9.</td>
<td>Conclusion</td>
<td>73</td>
</tr>
</tbody>
</table>

REFERENCE                                                                 | 76   |

PUBLICATIONS                                                                | 88   |

APPENDIX (FIRST AUTHOR ONLY)                                               | 89   |
LIST OF FIGURES

Figure 1: Protein in the water box.............................................................................................................. 5
Figure 2: Quantify the protein dynamics by (a) entropy and (b) distribution based on principal component analysis............................................................................................................................................................ 15
Figure 3: Key residues recognized based on protein entropic response ............................................. 16
Figure 4: 2D RMSD plots for Dark and Light states: (a) Unperturbed simulations, and (b) RRS 112, (c) RRS 119, (d) RRS 124 simulations respectively................................................................................................................................. 18
Figure 5: (a) Cluster analysis of 2D RMSD distribution of all simulations combined, (b) Distribution of all simulations among 15 clusters .................................................................................................................. 19
Figure 6: Schematic representation of population shift model .............................................................. 21
Figure 7: Machine learning models for PDZ2. (a) Decision tree parameter tuning, (b) DT model test results, (c) artificial neural network model tuning, (d) ANN model testing results, (e) benchmark dummy classifier ........................................................................................................................................................................................................................................... 25
Figure 8: Probability distribution for unbound and bound state for (a) DT model and (b) NN model respectively. ............................................................................................................................................................................................................................................................................................................................................................ 26
Figure 9: Distribution differences for different features. (a) 2D RMSD distribution, (b) Cα distance between residue Lys38 and His71, (c) dihedral angle between residue Pro1 and Lys2, (d) Cα distance between Asn16 and Arg31 ........................................................................................................................................................................................................................................................................................................................................................................... 27
Figure 10: The significance of distribution changes and free energy surface changes quantified by PRE. (a) Residue pair (N14:A74) with the highest PRE in the protein; (b) The free energy surface of the N14:A74 distance distribution; (c) The residue pair (D56:V64); (d) The free energy surface (D56:V64); (e) N14 and A74 in protein; (f) shortest pathway decomposition with cutoff 12Å; (g) the shortest pathway illustration with cutoff 12Å; (h) the shortest pathway illustration with cutoff 5Å. ........................................................................................................................................................................................................................................................................................................................................................................... 31
Figure 11: Projection of PDZ2 bound and unbound state using PCA. (a) Projection onto PC1 and PC2 surface; (b-p) Projection of different community pair onto pair-specific PC1/PC2 surface ........................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................... 33
Figure 12: VVD protein with four different configurations. (a) 2D-RMSD plot with reference to the native dark and light structures, respectively; (b) Estimated relaxation timescale based on transition probabilities ................................................................. 38

Figure 13: MSM based on VVD simulations, (a) 300 microstates; (b) lumped macrostates; (c) representative structures ............................................................................................................. 39

Figure 14: Chapman-Kolmogorov test of the MSMs ................................................................. 39

Figure 15: Transition Probability Matrix for (a) all trajectories, (b) non-bonded configuration, (c) bonded configuration ............................................................................................................. 40

Figure 16: Training models for the macrostates using different ML models: (a) Artificial Neural Networks; (b) Decision Tree; (c) Random Forest; (d) one-vs-one random forest ................................. 41

Figure 17: Distribution between selected features: (a) Feature 213; (b) Feature 227; (c) Feature 4976 ................................................................................................................................. 42

Figure 18: The demonstration of the estimation of $\tau_{\mu \nu}$ for the microstate $\mu$ and $\nu$ ......................... 49

Figure 19: A model system for comparison between the DKTN and CTMC models ......................... 54

Figure 20: Propagation of distribution starting from different initial conditions: (a) Starting exclusively from state A, (b) Starting exclusively from state D. ............................................ 55

Figure 21: Continuous-time Markov chain (CTMC) model: The transition probability matrix and result MSM model at (a) 0s; (b) 10s; and (c) equilibrium ......................................................... 55

Figure 22: The half time and evolution for different system including whole edges, removing edge A to B, removing edge B to C, removing edge B to D, and removing edge C to D. (a) the system starting at state A, the distribution of state D; (b) the system starting at state D, the distribution of state A ............................................................................................................. 56

Figure 23: Meta-stable state classification: (a) Cross-Validation; (b) GMM model; (c) k-Means model; (d) Averaged pair-wised RMSD of each state pair ......................................................... 59

Figure 24: Established DKTN model based on microstates: (a) The equilibrium reaction rate (also referred to as flux) between microstates in the equilibrium; (b) The rate constants between microstates ............................................................................................................. 60

Figure 25: System evolution over time starting in microstate 3 with (a) 0%, (b) 50%, (c) 75.0% and (d) 99.9% to the equilibrium. System starting in microstate 8 with (e) 0%, (f) 50%, (g) 75.0%, and (h) 99.9% to the equilibrium ......................................................................................... 60
Figure 26: Ten clusters obtained from $k$-means clustering: (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 or across the different clusters. ........................................ 65

Figure 27: Ten $k$-means clusters of VVD systems 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. ............................... 66

Figure 28: 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. ........................................................................................................ 67

Figure 29: Alanine dipeptide as test case: (A) alanine dipeptide structure and two dihedral angles ($\phi$ and $\psi$) as reaction coordinates; (B) initial minimum energy pathway for alanine dipeptide isomerization; (C) distribution of conventional molecular dynamics simulations of alanine dipeptide on its potential energy surface with reference to $\phi$ and $\psi$. ........................................... 70

Figure 30: DPDS simulations of alanine dipeptide with different $K_{rms}$ and PES coverage of each simulation: (A) $K_{rms} = 0.1$; (B) $K_{rms} = 100$; (C) $K_{rms} = 1000$; (D) plot of PES coverage along the simulation time.......................................................... 71

Figure 31: DPDS simulations of alanine dipeptide with different $K_{ang}$ (while $K_{rms} = 0.1$ kcal/(mol·Å²)) and PES coverage of each simulation: (A) $K_{ang} = 10$ kcal/mol; (B) $K_{ang} = 100$; (C) $K_{ang} = 1000$; (D) plot of PES coverage along the simulation time. .............................................. 72
LIST OF TABLES

Table 1: Key allosteric residues for PDZ2 from NMR study .............................................. 15

Table 2: Selected RRS simulations with major effect on the 2D RMSD distributions for either Dark or Light state or both .................................................................................................................. 18

Table 3: Top 5 features with the highest importance during the classification ....................... 42

Table 4: Accumulated feature importance between each community .................................... 43

Table 5: The changes of Community A and Community B during transition between states ..... 44

Table 6: Individual edge importance for the transition $A \rightarrow D$ ........................................ 57

Table 7: Important conformational changes for certain transitions ....................................... 61
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRS</td>
<td>Rigid Residue Scan</td>
</tr>
<tr>
<td>MSMs</td>
<td>Markov State Models</td>
</tr>
<tr>
<td>CTMC</td>
<td>Continuous-Time Markov Chain</td>
</tr>
<tr>
<td>LOV</td>
<td>Light-Oxygen-Voltage</td>
</tr>
<tr>
<td>REDAN</td>
<td>Relative Entropy based Dynamic Allostery Network</td>
</tr>
<tr>
<td>DKTN</td>
<td>Directed Kinetic Transition Network model</td>
</tr>
<tr>
<td>MWC</td>
<td>Monod, Wyman, and Changeux model</td>
</tr>
<tr>
<td>KNF</td>
<td>Koshland, Nemethy, and Filmer model</td>
</tr>
<tr>
<td>MD</td>
<td>Molecular Dynamic</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root-Mean-Square Deviation</td>
</tr>
<tr>
<td>RMSF</td>
<td>Root-Mean-Square Fluctuation</td>
</tr>
<tr>
<td>PDZ2</td>
<td>PDZ second domain</td>
</tr>
<tr>
<td>VVD</td>
<td>Vivid protein</td>
</tr>
<tr>
<td>TN</td>
<td>Transition Network</td>
</tr>
<tr>
<td>MTT</td>
<td>Mean-Time-to-Transfer</td>
</tr>
<tr>
<td>ODE</td>
<td>Ordinary Differential Equation</td>
</tr>
<tr>
<td>PDZ3</td>
<td>PDZ third domain</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>ID3</td>
<td>Iterative Dichotomiser 3</td>
</tr>
<tr>
<td>NN</td>
<td>Neural Networks model</td>
</tr>
<tr>
<td>DT</td>
<td>Decision Tree model</td>
</tr>
<tr>
<td>RF</td>
<td>Random Forest model</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>t-ICA</td>
<td>Time-structural Independent Component Analysis</td>
</tr>
<tr>
<td>PCN</td>
<td>Protein Contact Network</td>
</tr>
<tr>
<td>RIN</td>
<td>Residue Interaction Network</td>
</tr>
<tr>
<td>ENM</td>
<td>Elastic Network Model</td>
</tr>
<tr>
<td>GN</td>
<td>Girvan-Newman algorithm</td>
</tr>
<tr>
<td>KL</td>
<td>Kernighan-Lin algorithm</td>
</tr>
<tr>
<td>t-SNE</td>
<td>t-Distribution Stochastic Neighboring Embedding</td>
</tr>
</tbody>
</table>
This is dedicated to my father, mother, sisters and brothers-in-law.
1. Introduction

1.1 Background to Allostery

Allostery is a ubiquitous process for all kinds of proteins referring to protein regulation due to perturbation at distal sites. Although the word “allostery” was first coined in 1961\(^1\), the phenomenon has been studied even before this term was introduced. In 1904, an interesting biological relationship was observed between carbon dioxide, oxygen and by Christian Bohr, and was named as Bohr effect\(^2\). Currently, the Bohr effect is known as an instance of allosteric effects with the conformational change. After the term “allostery” appeared in 1961, two well-known models were proposed as symmetric MWC model by Monod, Wyman, and Changeux (MWC)
\(^3\) and sequential KNF model by Koshland, Nemethy, and Filmer (KNF).\(^4\) Both models assume that allosteric proteins have multiple subunits and ligand binding will induce a conformational change of the protein into another state. The KNF model assumes that ligand binding will change the conformation of subunits sequentially. According to the MWC model, the multiple subunits will be altered concurrently, and the ratio of different conformational states are determined by their thermal equilibrium.\(^4\)-\(^5\) Both models successfully explained many allosteric protein regulations, leading to the consensus that conformational changes are the “signature” of allosteric effects.\(^6\) Things have changed in the later 90s, when some proteins undergoing functional changes without conformational changes were identified, and therefore defined as “dynamic allosteric protein”.\(^7\)-\(^10\) Over the years, more and more proteins have been
found to be dynamic allosteric proteins, leading to the hypothesis that all proteins could be more or less allosteric.\textsuperscript{11-13}

Currently, to fully understand allostery, researchers are investigating different aspects including the thermodynamics, the kinetics, and the structures of protein allostery.\textsuperscript{14-17} In the thermodynamics view of allostery, starting from MWC model, the role of entropy was further highlighted and treated as the driving force in the allostery.\textsuperscript{18} For example, in the PDZ domain and Pin1, the reorganization of active site is driven by the lower conformational entropy cost due to the phenomenon that ligand binding reduces the side chain flexibility.\textsuperscript{18} Therefore, the allosteric conformational changes could be driven by entropy changes, and the ratios among different conformational states were found to be proportional to the entropy differences.\textsuperscript{19-20}

On the other hand, allosteric regulation has also been commonly regarded as the kinetic effect that lead to the conformational or dynamical changes.\textsuperscript{18} This view is particularly dominant in the enzyme-involved allostery, in which the turnover rate $k_{cat}$ and Michaelis constant $K_M$ are used to characterize the enzyme kinetics. Upon the kinetic view, enzyme-catalyzed reaction can be viewed as three steps, including the binding of the ligand, the reaction of reactants to products, and the releasing of the products. Allosteric ligands can affect the dynamics and rate constants in those steps and lead to different functions.\textsuperscript{21-23}

Another common view of allostery is the structure based view of allostery.\textsuperscript{14} In the structure based view, allostery based communication networks or pathways are considered critical to understand the allosteric functions. Unlike the thermodynamics and kinetics views, in the structural based view, a deterministic pathway or a network model is responsible for the “communication” between a distal allosteric site and the activate site. Moreover, the ultimate goal for the structure based view is to predict the allosteric site and the allosteric communication
mechanism given by the protein structure, protein function and activate site. The structure based view of allostery has also been evolved from a single pathway communication to the multiple pathways cooperation mechanism, where network could come into place between distal allosteric sites with activate sites.

One of the state-of-the-art models to explain the allosteric effect is the “conformational ensemble and population shift model” proposed by Nussinov group in 1999. The population shift model highlights the importance of the distribution of different conformational states in the allostery, and uses the free energy landscape to model the ratio between different states. This model hypothesizes that allosteric proteins could exist in multiple states with different populations, where the population can be altered upon the ligand binding and perturbations. After the model was proposed, the population shift model has been widely used to explain allostery and advance the allostery theory from two-states to ensemble of multiple states. Overall, this model takes advantage of both thermodynamic view and structural view. The population shift model can be represented as the free energy landscape to incorporate the effect of ligand binding. Meanwhile, the population shift model also pointed that the classical structural view of allostery, or the pathway between allosteric site and activate site does not involve allostery directly, but identifies the high correlation between two sites.

In this thesis, I summarize the past work conducted on allostery from different aspects. Specifically, starting from identifying important residues as a first step, we examined the population shift model hypothesis for a typical dynamical driven allosteric protein using machine learning. Meanwhile, from the kinetic point of view, we applied Markov State Models (MSM) on the allosteric protein to verify the kinetics and further develop Directed Kinetic Transition Network (DKTN) model and made an analogy to the Continuous Time Markov Chain (CTMC)
to model the kinetic transformation between different conformational states.\textsuperscript{36-38} On the structural point of view, we developed the Relative Entropy based Dynamic Allosteric Network (REDAN) model to simulate the allosteric effects and energy flow inside the protein as a network. Further pathway and community analysis could reveal the communication between different communities inside the protein.\textsuperscript{39} Overall, these studies systemically investigated the allostery from different perspectives and advanced the understanding of allostery through computational studies.

1.2 Molecular Dynamic Simulation

Before getting into the details of unravelling allosteric mechanism, the computational tools applied in the above studies are the Molecular Dynamic (MD) simulations. In general, the molecular dynamics is a computational simulation method for studying the physical movement of atoms or molecules based on the classical mechanics.\textsuperscript{40} The atoms are allowed to interact with each other for fixed periods of time, providing the view of the dynamic evolution of the system. Currently, the molecular dynamics method has engaged in many fields, including biology, geology, chemistry etc.\textsuperscript{41-42} Biological applications of molecular dynamic simulations have been widely used to study the evolution of protein systems, yielding of around \textasciitilde20, 000 publications per year. The time scales and computational scales for large macromolecule systems for MD simulation are more suitable compared with \textit{ab initio} mechanics dynamics.\textsuperscript{43} Specifically, molecular dynamic simulation is suitable for investigating of allosteric mechanism. In principle, long molecular dynamics simulations can provide all the details regarding the communication between allosteric and active sites, especially for simulations longer than microseconds.\textsuperscript{44-45} Therefore, in my studies, molecular dynamics simulation is included as the major tool to investigate protein allosteric mechanism.
A typical molecular dynamic simulation is composed of the following steps. Start from solvating the protein into the water box; keep adding ions to maintain an adequate ionic strength; minimize the energy into a stable structure between water box and protein system; then heat and equilibrate the system and finally conduct the simulation with several nanoseconds or with the microseconds scale as shown in Figure 1. After the simulation, the analysis of protein trajectories is an open-ended question. Although standard analyses like Root-Mean-Square Deviation (RMSD) or Root-Mean-Square Fluctuation (RMSF) for residues are commonly conducted in many analyses, currently, the analyses involving Markov State or Machine Learning are more prevalent to deal with the large amount of data from the simulations, commonly in the size of Terabytes.46-47

Figure 1: Protein in the water box

Overall, in the studies presented in this dissertation, molecular dynamic simulations combined with Machine Learning and Markov State Models are the primary tools for understanding protein allostery. I investigated allostery from several aspects including the
thermodynamic view, the structural view and the kinetic view, from the methodology developments to the applications. This research advanced the understanding of allostery in terms of population distribution, and shed light on the allosteric communication mechanisms for several proteins.
2. Methodologies

The focus of the current dissertation is methodology developments for the allostery from different perspectives, including the kinetics, the thermodynamics and the structures. Detail methods will be introduced from Chapter 3 to Chapter 8. In this chapter, an overview and a comparison of these methods are provided.

**Rigid Residue Scan**

Rigid Residue Scan (RRS) is the first method developed in our group to analyze the importance of residues upon an external perturbation.\(^4^8\) Comparing the difference between an unperturbed simulation with the corresponding perturbed simulation, the importance of particular residue is quantified. Rigid body simulation is used as the perturbation because of two reasons. First, protein dynamics are consisted by the internal dynamics of each residue. By removing the internal dynamics of a residue, the protein dynamics could be affected. The importance of a residue is estimated using the differences between these two simulations. Moreover, protein dynamics is also related to protein vibrations. Rigid body also affects the vibration of the protein. Therefore, the allostery of the protein can be affected too.

Two proteins are investigated using RRS method, including PDZ second domain and VIVID protein from LOV (Light-Oxygen-Voltage) domain. The detailed results of the RRS on these two proteins are demonstrated in the next Chapter. Overall, the results of both proteins lead to agreement between the computational studies and the experimental data, which indicates the
efficiency and validity of the RRS method. Overall, RRS method has the following advantages. First, RRS method provides a systematically evaluation of the individual residue’s importance between a perturbed simulation with the corresponding unperturbed simulation. Second, important residues have been identified with a good agreement between the experimental studies and the current study. Meanwhile, certain disadvantages exist. First, the computational cost of the RRS is expensive. RRS requires a large number of simulations for a thoroughly scan for all residues in the protein. Second, the allosteric mechanism remains unclear despite the important residues being identified. To address those limitations, the Machine Learning approaches were developed on those simulations to investigate the underlying mechanism.

**Machine Learning Approaches Combined with Molecular Dynamic Simulation**

Revisiting the population-shift model in the introduction section, the key hypothesis in the model is population distributions, which are altered upon an external perturbation. Distribution, or more generally, a population distribution, has been widely studied in Machine Learning, computer science and statistics. Decision tree is one of the machine learning models which takes advantage of the probability during a classification of different labels. To verify the population-shift hypothesis on dynamics allosteric proteins, PDZ second domain (PDZ2) was investigated as the model system. Although there is certain debate regarding an existence of the conformational ensemble of the PDZ2 protein, Decision Tree and Neural Network models clearly suggest the conformational ensemble indeed exists in the PDZ2 protein. The population distributions of those conformations are shifted upon a peptide binding perturbation. The detail result of the machine learning study is described in Chapter 4. The features with the top importance have a highly agreement with other computational studies. The distribution plots
clearly suggest that the population distributions of those features will be altered upon a peptide binding perturbation.

Compared with the RRS method, certain advantages exist for the Machine Learning based approaches. First, the computational costs are lower than the Rigid Residue Scan. The method also provides the distribution information where the previous method cannot. The important residues identified in these methods are different. The residues identified by ML are related to the distribution changes, while the residues identified by RRS are related to configurational entropy changes. However, limitations also exist. The distribution differences identified from the machine learning approaches are not necessary the distributions related to allostery. Meanwhile, the allosteric mechanism is still unclear. Detailed results, advantages and limitations will be discussed in Chapter 4. Because of those disadvantages, a network model was adopted in proteins to address the allosteric mechanism, leading to the Relative Entropy based Dynamic Allosteric Network model (REDAN).

**Relative Entropy based Dynamic Allosteric Network**

Followed the studies on machine learning models, the allosteric mechanism is further investigated using a network analysis. Protein network models are widely used in bioinformatics and computational chemistry to study connections of residues. These residues are treated as vertices, and the interactions between them represent edges. To incorporate distribution information in the network model, “Relative Entropy” is introduced to calculate the distribution difference between a perturbed simulation with the corresponding unperturbed simulation. The relative entropy value indicates the significance of the allosteric effect upon an external perturbation. Upon constructing the network model, several analyses could be performed on the model including the pathway analysis and the community analysis. The pathway analysis
decomposes a residue pair with a large relative entropy value and a large distance into a series of residues. Accumulation of the allosteric effects among the adjacent residues in the pathway is a possible communication pathway regarding the long-distance residue pair. Community analysis, on the other hand, decomposes the allosteric effects inside a protein into global interactions between communities. Therefore, the allosteric mechanism is localized into the interactions between communities without redundant analyses of local interactions within the community. The details of the REDAN model are described in Chapter 5.

Comparing with the previous models, one significant strength of REDAN model is the incorporation of the distribution information inside a protein. The utilization of various algorithms on a network could provide a possible allosteric mechanism hypothesis for a protein. However, from the contemporary view of allostery, the kinetic information is also important to understand conformational changes. Although REDAN model has incorporated both the structural and the thermodynamic information, the kinetics is absent in the current model. Therefore, to further investigate the kinetics in the allostery, Markov State Models (MSMs) have been applied to study the kinetics.

**Markov State Models Combined with Machine Learning Approaches**

Based on the Markov Chain theory, the Markov State Models (MSMs) have been widely applied on MD simulations to study transition probabilities, or more general, the kinetics between different conformational states. These conformational states are referred as “microstates” in the MSMs. In Chapter 6, the MSMs have been applied on the Vivid protein (VVD) to investigate the conformational change mechanism. Various analyses have been conducted on the VVD protein. First, the Markovian property of the MSMs has been verified. Next, the mapping between microstates into the Markov States has been conducted.
the REDAN model, the community analysis has also been conducted on the VVD protein combined with the MSMs. Based on the community analysis result and the transition probabilities of the MSMs, the allosteric mechanism of VVD protein was proposed that two loops have been identified as functional related region.

Comparing with the above analysis, the MSMs provide the kinetic information and transition probabilities between different microstates. These transition probabilities are important to understand the allosteric mechanism. Combined with the community analysis, certain regions could be identified, and the potential allosteric mechanism could be revealed. Overall, the MSMs could served as the complementary analysis method to the REDAN model to reveal the kinetic information regarding the allosteric protein with conformational changes. However, certain limitations exist for the MSMs, including the selection of lag time, the requirements of large number of samplings etc. Therefore, to further improve MSMs, the Directed Kinetic Transition Network (DKTN) model has been proposed.

**Directed Kinetic Transition Network model**

As opposite to MSMs, DKTN model does not rely on the underlying Markovian Property of MD simulations. Instead, as a further improvement of Transition Network (TN) model, the foundation of DKTN model is the reaction rate theory. The conformational changes between the ensemble of microstates are treated as first-order chemical reaction. Based on the detailed balanced constraint, two metrics are collected from simulations as the equilibrium distribution and the mean-time-to-transfer (MTT) between adjacent microstates. The “reaction rate” for an adjacent microstate is the inverse of the MTT value. Different conformational states are treated as vertices. The edges represent the transformation of different microstates. The backwards and forwards rate constants are calculated from the simulation and detailed balanced constraint.
Comparing with MSMs, the MSMs is a special case of DKTN model where DKTN model is evolved into equilibrium. Overall, the DKTN model can model a system from non-equilibrium to equilibrium by solving the ordinary differential equations (ODE). Meanwhile, the lag time in the MSMs is not needed in the DKTN mode because DKTN model is mathematically equivalent to the Continuous Time Markov Chain (CTMC). Unlike the MSMs model, the equilibrium distribution is utilized in the DKTN model. The limitation of DKTN model is the requirement of long time samplings, which ensures that the simulation approaches equilibrium. The comparisons between MSMs and DKTN are listed in Chapter 7.

Summary

In this dissertation, five different methods have been discussed regarding allostery, including RRS, ML, REDAN, MSMs and DKTN. These methods were developed sequentially to unravel allostERIC mechanism through three different perspectives: the thermodynamics, the structures, and the kinetics. The RRS investigates the effects of configurational entropy. ML models further investigates population distributions, or equivalently, the free energy landscape for different ensemble conformations. Both methods focus on the thermodynamics view of allostery. REDAN model, on the other hand, incorporates the thermodynamics with the structural view, which forms a unified view of allostery. The above three methods are sufficient to identify important residues, unravel allostERIC mechanism, discover possible communication pathways etc. However, the kinetics is absent in the above methods. To address this limitation, the MSMs and DKTN model are introduced and developed specifically for studying the kinetics of the conformational changes. By forming a unified view of the structure, the thermodynamics and the kinetic of allostery using those methods, the underlying allostERIC mechanism can be revealed. Those methods will be discussed in detail in the following Chapters.
3. Identify Important Residues Using Rigid Residue Scan (RRS)

3.1 RRS on Model Proteins: PDZ Domain

To obtain a solid understanding of the allosteric mechanism, the first step is to identify the important residues which can potentially affect allostery. Many experimental studies demonstrated that certain residues could adjust the allosteric effects better than others\textsuperscript{50-52}. Therefore, identifying important residues enhances the understanding of protein allosteric mechanism. In this study, the goal is to predict allostery related residues. To identify important residues, we developed a method named the Rigid Residue Scan (RRS), where we made two key assumptions. First, protein function, or specifically, the allosteric effect is related to protein dynamics. Second, protein dynamics is correlated with the internal dynamics of each residue. Meanwhile, the internal dynamics of allosteric related residues is more important than other residues regarding protein overall dynamics.

According to the population shift model and recent observations of dynamic-driven allosteric proteins\textsuperscript{8, 11}, the importance of protein dynamics is highlighted where Molecular Dynamic simulations come in place. The second assumption mentioned above may not be as intuitive as the first assumption. However, the subsequent analysis suggests that the assumption is valid for several proteins\textsuperscript{33, 53} The internal dynamics of allosteric related residues has higher impact than other residues regarding the protein overall dynamics.

Basically, the RRS method consists two stages. First, several unperturbed simulations are required for quantifying protein overall dynamics. Second, a deep scan simulation of treating
each residue as rigid body is conducted.\textsuperscript{54} Each perturbed simulation is compared with the unperturbed simulation to quantify the importance of each residue related to the overall dynamics. Unlike mutation studies, the rigid body dynamics is a theoretical treatment which removes the internal dynamics of a residue during the simulation to verify the impact on the overall dynamics. Although it cannot be verified directly through experiments, the important residues identified through our methods are highly correlated with the experimental results with statistically significance on our model systems, third PDZ domain (PDZ3) and second PDZ domain (PDZ2) system.

We chose PDZ3 and PDZ2 proteins as model systems because they have been widely investigated through experimental studies.\textsuperscript{50, 55-57} Meanwhile, they are well-known dynamic-driven allosteric proteins which do not involve significant conformational changes when binding with a ligand. In our study, rigid body is treated for each residue in those proteins and conducted 192 simulations for the PDZ2 protein. To verify the impact of individual residues on the overall protein dynamics, we investigated the impact through different aspects, including covariance matrix, distribution difference based on principal component analysis, and configurational entropy as shown in Figure 2.

Among them, the configuration entropy is calculated from the covariance matrix of a simulation as shown in the reference\textsuperscript{58} using quasi-harmonic analysis module in CHARMM. The configurational entropy of protein is calculated as the Eq 13 from the reference\textsuperscript{59}. It should be noted that the configuration entropy is an approximation of the target value within the harmonic limits. Because all the protein simulations are superimposed regarding the reference structure, the translations and rotations are projected out for entropy calculations.
Figure 2: Quantify the protein dynamics by (a) entropy and (b) distribution based on principal component analysis

Figure 2 clearly indicates the dynamics difference between different simulations using (a) entropy and (b) distribution based on principal component analysis, respectively. One interesting observation of the configurational entropy differences is that the residues with similar unbound and bound configurational entropy are identified as the allosteric related residues. For two unperturbed simulations, the difference is 0.016 kcal/(mol*K) with a higher PDZ2 entropy from the bound state. For the 11 residues being held rigid, seven among those 11 residues are reported as important residues from an experimental NMR study as shown in Table 1 and displayed in Figure 3.

Table 1: Key allosteric residues for PDZ2 from NMR study

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a: residues in bold type are recognized in the present study</td>
</tr>
</tbody>
</table>
Figure 3: Key residues recognized based on protein entropic response

Although in the current study, the direct relationship between the experimental study and the RRS is not revealed, the significant overlap between these studies suggests our method is valid to identify key allosteric residues. Overall, RRS is along the way to uncover allosteric mechanism, or more specifically, to uncover the allosteric residues in the thermodynamic view, which focuses on changes of configurational entropy upon a perturbation of individual residues.

3.2 Application of RRS Method on VVD Protein

After developing the RRS method on the model proteins, the method was further applied on the FAD-binding LOV (light, oxygen, voltage) protein VVD. VVD is a protein related to the adaptation of blue-light responses in Neurospora crassa, with two conformations referred to as dark and light states upon the responding of light. The conformational change mechanism of the VVD protein is a typical two-states allosteric phenomenon. In the current study, the RRS method is applied on the VVD protein to identify several allosteric related residues.

Compared with the PDZ domain proteins, VVD protein involves the conformational changes upon a perturbation. To quantify the conformational changes more directly, 2D Root-Mean-
Square Deviation (RMSD) is applied to display the structure using the crystal dark and light structures as the reference structures. Generally, RMSD is used to quantify differences between two structures. Although RMSD is not a good reaction coordinate in general, 2D RMSD particularly suits here because the reference dark and light structure are significantly distinguished with each other, leading to a pseudo-orthogonal space in revealing conformational samplings. Moreover, in the t-Stochastic Neighboring Embedding (t-SNE) dimension reduction application of VVD protein, the clustering of 2D RMSD indicates that the structural similarity of 2D RMSD is well held by comparing with PCA or t-ICA projections. Furthermore, in the Markov State Model (MSM) application of VVD protein, the Chapman-Kolmogorov test shows that the 2D-RMSD projection also maintains a good Markovian property. Overall, it is clear that 2D-RMSD is a suitable reaction coordinate for the VVD protein. In this chapter, Rigid Residue Scan method is applied to identify key residues by projecting onto 2D-RMSD surface.

Obviously, for the unperturbed simulation, the unbound state and ligand-bound state simulate different conformational space as shown in Figure 4a. For the perturbed simulations, certain residues can clearly change the distribution upon perturbation from dark to light state (Figure 4b), or vice visa (Figure 4c), or overlapping the sampling conformational space for both ligand-bound and unbound simulations (Figure 4d). RRS 112, 119 and 124 simulations are not the only cases which display this particular allosteric effect. Other RRS simulations displaying similar effect in the 2D RMSD distributions are listed in Table 2.
Interestingly, the rigid residue perturbations have larger effects on the light states than the dark states. As shown in Table 2, for more than 30 residues, the sampling spaces for the light state have been affected. Only several residues have more perturbations on the dark states. Also, from the sampling space and the entropy analysis, the light states also show the large flexibility than the dark state. Overall, the result suggests light state is more likely to be influenced comparing with dark state upon perturbations.

Table 2: Selected RRS simulations with major effect on the 2D RMSD distributions for either Dark or Light state or both

<table>
<thead>
<tr>
<th>Affecting</th>
<th>Mainly Dark state</th>
<th>Mainly Light state</th>
<th>Both states</th>
</tr>
</thead>
</table>

Besides the above analysis, in the current study, the clustering analysis has also been conducted to identify the hidden conformational states, which could be important for the protein
functions. Based on the structural similarity in each state, the overall sampling space has been clustered into 15 conformational states corresponding to different conformations as shown in Figure 5. The dark states including the unperturbed simulation and RRS simulations dwell in state 1. For the light states, they are diverse in several clusters. The states with the lowest RMSD to the reference crystal light structure are the state 4, 7 and 8. Along with the RRS result, the clustering result also suggests that light state is more flexible, and samples much larger conformational space than the dark state.

![Cluster analysis of 2D RMSD distribution of all simulations combined](image)

**Figure 5:** (a) Cluster analysis of 2D RMSD distribution of all simulations combined, (b) Distribution of all simulations among 15 clusters

In this study, the RRS method has been applied on VVD protein. Several key residues have been identified with a good agreement with some experimental studies. Those studies suggest the potential role of those residues regarding the conformation changes. Furthermore, from both the unperturbed and perturbed simulations, the result suggests that the flexibility of the light states is significantly higher than the dark states. Therefore, we further applied Markov State Models (MSM) to investigate the covalent bond and the kinetics of the conformational changes in Chapter 6.
4. Distribution and Population-Shift: Random Forest Model

4.1 Population Shift Model

RRS method is dedicated to identify important allostery related residues by the positive control between an unperturbed simulation with a perturbed simulation with the rigid body treatment. However, the allosteric mechanism stays uncovered with RRS method. To better understand the underlying mechanism, the population shift model is revisited on the model protein system: PDZ2 domain. As introduced in Chapter 1, the population shift model becomes dominant after developing in 1990s. Compared with the previous “induced-fit” hypothesis, the conformational ensembles and population shifts model proposed the following assumptions.

1. The protein exists in an ensemble of conformational sub-states around their native states in solutions.
2. The population of each state reflects the free energy landscape. The perturbation of a ligand will alter the population of each sub-state.

Specially, in the “induced-fit” model, a static unbound protein accommodates an incoming substrate by adapting its ligand-binding site’s shape from one conformation to another. On the contrary, the population shift model assumes that both states will pre-exist in the solution already. The ligand binding event will change the population between those states, favoring one conformation over the other. Previously, because the limitation of the Nuclear Magnetic Resonance (NMR) resolution, only the dominant conformation can be detected,
leading the result of “induced-fit” hypothesis as identical with population shift model.\textsuperscript{67-68} However, with the improvement of resolution, current NMR methods are able to detect multiple low population states and provide direct evidence of population shift model rather than the “induced-fit” hypothesis.\textsuperscript{4, 28, 69-71} The basic concepts of population shift model is illustrated in Figure 6.\textsuperscript{11}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{population_shift_model}
\caption{Schematic representation of population shift model}
\end{figure}

Clearly, the population shift model fits well for the conformational driven allosteric proteins. More and more evidences suggest the validity of population shift model in terms of distribution changes upon a perturbation.\textsuperscript{69-71} However, the dynamical driven allosteric proteins may conflict with the population shift model at the first glance like PDZ2 protein which does not involve significant conformational changes. Therefore, such debate exists regarding the existence of different distributions or conformational states in dynamical driven allosteric proteins.\textsuperscript{9, 18} In the framework of the population shift model, the explanations for dynamical driven allosteric proteins are the followings reasons.\textsuperscript{68} Two of them are the dominant state distribution and crystallization effects.\textsuperscript{68}
The dominant state distribution suggests that it is possible that allostery can change the population of two states without changing the dominant states. For example, allostery can change the distribution between two states from 99% vs. 1% to 80% vs. 20%, without changing the dominant state. Therefore, even though overall distributions have been significantly affected, as long as the dominant state stays same, the characterized structure will be the same. Moreover, the crystal packing can stabilize the inactive state or destabilized the allosteric states (or vice versa). Meanwhile, pH and small molecules solutes in the crystallization buffer can stabilize the opposite state or destabilize the allosteric state, resulted in the limited conformational changes after crystallization.

Those possible reasons explain the absence of the conformational changes in the allosteric protein with the existence of different states and distributions. However, certain debates exist that the dynamical driven allosteric proteins do not involve different conformations or distributions, the allosteric effects are caused solely by the changes of dynamics of protein before and after ligand binding without changing the conformation at all. To better understand allosteric mechanism and to verify the population shift model hypothesis, machine learning was applied for the classification between two states and revealed the distribution of each conformational state.

4.2 Machine Learning: Random Forest

Machine Learning (ML) methods were applied in the current study because of two reasons. First, distribution, or more specific, the probability distribution, is one of the central concepts in computer science and statistics to investigate datasets. Second, the simulation is able to generate a large dataset (~TB) which contains many instances of the structures. Manually inspection of large datasets is not feasible, where Machine Learning comes into place. Overall,
ML approaches is suitable for two scenarios. The task is too complicated to be programmed or the task is beyond a human capacity. In this case, because molecular dynamic simulations could generate extremely large amount of data, only machine learning is capable to analyze the results.

Classifications based on Machine Learning have been widely applied in various fields including computational biology and chemistry. In the past studies, random forest, or more broadly, decision trees have been applied because the classification of decision tree (DT) models is based on the distribution differences in two datasets. In general, decision tree models can be applied to identify key factors that contribute the most to the target states, based on the information gain or relative entropy. Comparing with other machine learning models, decision tree models are insensitive to the high dimensions because the information gain is calculated for each dimension individually based on the ID3 (Iterative Dichotomiser 3) algorithm. Meanwhile, decision tree models are applied on input features directly without any non-linear transformation, leading to that the results from the decision tree models are easier to be interpreted than other machine learning models.

Random forest model is an improvement of decision tree model based on ensemble learning. One limitation of the decision tree models is bias. Decision tree models could be biased to a particular feature leading to over-fitting estimation. For example, for two highly correlated features, DT tree identifies one of them as “important” features with large information gain. Subsequently, the dataset is split based on this particular feature. As a result, the other feature would be discarded because the information gain of this particular feature on the remaining dataset will be small. However, because both features are highly correlated, it does not make sense for one being classified as “important”, when the other feature is classified as “non-important”. To address this issue, random forest model is constructed by multiple decision
trees. Each one of them is applied to classify the data based on random selected features. Subsequently, multiple decision trees will vote together. The importance of each feature is the average value of those trees. Overall, random forest model is known to less over-fitting than a single decision tree model.  

4.3 Classification Using PDZ2

In the first study of applying ML, PDZ2 is selected as the template protein because it is a widely investigated protein with sufficient experimental/computational data as comparison. In this study, two questions were attempted to be answered.

1. To understand the population-shift model in the dynamical driven allosteric protein.
2. To develop a theoretical prediction model to recognize different allosteric states.

Those two questions can be achieved by the ML classifications.

The data cleaning of the binary simulation data is involved with the following steps. First, an appropriate coordinate needs to be selected to describe a protein structure. This step is named as featurization. Different featurization methods may lead to different conclusions. Sometime the results can be biased. In this study, Cα pairwise distances and backbone dihedral angles are selected to be features, because those two features are independent to the translation and rotation. On the contrary, Cartesian coordinates are not suitable as the direct input to the machine learning models, because Cartesian coordinates are dependent to the translation and rotation. In the first study, total 26 trajectories for unbound and bound states were conducted and 4371 features were extracted including the pairwise distance and dihedral angle.

Besides the decision tree model, the neural networks model has also been applied in this study for comparison purpose. Shown in Figure 7, although the neural networks model has a
slightly better performance than the decision tree model, the feature important cannot be extracted from the neural networks model. Overall, the training and validation accuracy are above 80%, which suggests the simulations of two different states can be well classified.

Figure 7: Machine learning models for PDZ2. (a) Decision tree parameter tuning, (b) DT model test results, (c) artificial neural network model tuning, (d) ANN model testing results, (e) benchmark dummy classifier.

One interesting observation of those different ML models is the probability distribution generated by DT and NN models as shown in Figure 8. In the neural networks model, two peaks are classified respectively for unbound and bound state. The perturbation upon peptide binding significantly changes the distribution from one state to the other. Meanwhile, decision tree model identified 5 peaks, which exists in two state simulations. However, upon peptide binding, the population of each peak is altered. In the unbound state, these peaks decrease upon binding, whereas in the bound state, these peaks increase. Close observation with Figure 8a and Figure 6 suggests that the probability distribution based on the decision tree model validates the population shift hypothesis.
Figure 8: Probability distribution for unbound and bound state for (a) DT model and (b) NN model respectively.

Meanwhile, besides confirming the validity of population shift model on a dynamic driven allosteric protein PDZ2, the important features can also be identified from the decision tree models, which indicate the most distinguishable part in the protein upon peptide binding as shown in Figure 9. It is clear that the principal component analysis (Figure 9a) shows that the overall distribution between unbound state and bound state are similar with each other as the consequence of the dynamical driven allosteric proteins. However, certain features suggest that the differences between unbound state and bound state simulation are significant (Figure 9b-d). These features are also highlighted in the RMSF studies and the computational studies, confirming the efficiency of our machine learning approaches.
Figure 9: Distribution differences for different features. (a) 2D RMSD distribution, (b) Cα distance between residue Lys38 and His71, (c) dihedral angle between residue Pro1 and Lys2, (d) Cα distance between Asn16 and Arg31

Overall, in this study, DT and NN model have been applied to investigate the mechanism of PDZ2 protein based on the population shift hypothesis. Although PCA suggests that the overall distribution of two protein simulations is close to each other, the DT model identified certain features with significant distribution differences upon the peptide binding. Meanwhile, the probability distribution identified by DT model is similar to the hypothesis from the population shift model, because those populations are altered upon a perturbation. Moreover, important features and residues can also be identified through ML with good agreement with other experimental and computational studies.\textsuperscript{50,55-57} These demonstrate the efficiency of applying the ML approaches on the protein system.
5. Network Model in Proteins: Allosteric Network

5.1 Network Model in Proteins

Network is everywhere in the world, both conceptually and physically. Common networks include power grid, social network, google search pages, data centers, internet etc. In general, network model is widely used in many areas, including biology, chemistry, computer science, engineering etc. The network, or more general, the graph theory is one of the central concepts and data structures in computer science to model the real world. Several algorithms have been developed based on the graph theory including shortest pathway algorithms, communities detection algorithms, independent set detection, minimum spanning tree detection, etc. Meanwhile, many algorithm problems in the graph theories are known to be NP-complete problems, which require significant computational resources to solve. Although the network and graph theories are developed in the field of computer science, biology has adopted this concept and developed various network models to model gene expression and proteins.

To understand protein structure-function relationship, many network methods have been developed to model protein structure. Proteins, comparing with other macromolecules, are easier to be modelled as the network, because the basic building blocks of protein, amino acids, interact with each other in various ways. Those protein networks can be categorized into static networks and dynamical networks, whereas the static networks represent single structures, and dynamical networks represent the correlation based on molecular dynamic simulations.
Static networks include protein contact network (PCN) and residue interaction network (RIN) models.\textsuperscript{107-109} Those models reveal the protein stability and identify hubs, domains and function related residues. Besides, the Elastic Network Model (ENM) is also a common protein network model to estimate the vibration and frequency of protein by model the residue – residue interaction as a harmonic string.\textsuperscript{95, 110} On the contrary, the dynamical network model focuses on the correlation between residues based on molecular dynamics simulations, and the shortest pathway and community detection algorithm are applied on the dynamics network analysis.\textsuperscript{106} Because those network models are not developed specifically for protein allostery, the information about the change of distributions is not emphasized in those studies. Therefore, inspired by the previous machine learning results, a new network model is proposed based on the distribution information between two states. This network model specifically describes how the protein is altered upon perturbation. Because the distribution indicates the free energy landscape, the distribution changes upon perturbation can be regarded as the changes of free energy landscape upon perturbation.

\textbf{5.2 Relative Entropy based Dynamical Allosteric Network Model (REDAN)}

The REDAN model treats each residue or the amino acid as a node in the network, whereas the distribution difference between pairwise residue is treated as the weights of the edge. Here, inspired by the decision tree model, the relative entropy is used to quantify the difference between two distributions.\textsuperscript{39, 111} Relative entropy, or Kullback–Leibler divergence, is a common metric to quantify the distribution difference upon perturbation.\textsuperscript{112} In the current framework, relative entropy indicates the significance of the allosteric effect upon perturbation. The higher relative entropy indicates the larger allosteric effect upon perturbation in the pairwise distance. Overall, this network is a complete (fully-connected) network for each residue in the protein
which combined the structure information and the distribution information during the simulation. Therefore, the current network investigated the allosteric protein from both structural point of view and thermodynamic point of view, in which the shortest pathway and communities can suggest the possible communication pathways and the mechanisms.

The shortest pathway is calculated based on the specific cutoff value on the network, similar with the dynamical network model. The large allosteric effects between two distal residues are more likely to be connected through several intermediate residues, which can be identified through the shortest pathway algorithm. The pathway with the smallest weight implies the largest allosteric effect through the propagation channel. Here, the shortest pathway is identified by the standard Dijkstra’s algorithm. Besides the top shortest pathway, the top k pathways have also been identified using Yen’s algorithm as shown in the literature. Meanwhile, other than the shortest pathway algorithm, the community detection algorithms have also been applied on the current network to identify individual communities which can conserve the total relative entropy value inside the community. The optimization of the lowest costs inside communities is obvious a NP-hard problem by reduction to the subset-sum problem. Therefore, the exact solution is infeasible to be obtained where the heuristic algorithm is used. Both Girvan-Newman (GN) algorithm and the Kernighan-Lin (KL) algorithm have been applied to identify different communities. Combined the shortest pathway and community analysis, the network provides a deeper understanding for the allosteric mechanism.

5.3 Model Protein System: PDZ2

The allosteric mechanism of PDZ2 is revisited using the network analysis as shown in Figure 10. First, the residue pair with the highest relative entropy is identified as N14 with A74 in the protein as the relative entropy 2.019 in Figure 10a. The free energy landscape for the
distance of N14:A74 is calculated and plotted in Figure 10b. Clearly, with the perturbation of peptide binding, the free energy landscape of N14:A74 has shifted significantly. However, for certain distance like D56:V64, the distribution and free energy landscape are similar upon perturbation, which indicates that the perturbation is more sensitive to certain part of protein rather than being equally imposed on each residue as the consequence of allostery. Because the distance between N14 and A74 is more than 20Å, it is unlikely that these two residues interact with each other through a direct manner. The shortest pathway decomposition shown in Figure 10e suggests a possible allosteric communication pathway, which connects two residues through several intermediate residues. Obviously, the allosteric effects are likely to be a cooperative effects of multiple pathways rather than a single dominant pathway. Accordingly, the cutoff value can provide more flexibility regarding different pathways shown in Figure 10g and Figure 10h.

Figure 10: The significance of distribution changes and free energy surface changes quantified by PRE. (a) Residue pair (N14:A74) with the highest PRE in the protein; (b) The free energy surface of the N14:A74 distance distribution; (c) The residue pair (D56:V64); (d) The free energy surface (D56:V64); (e) N14 and A74 in protein; (f) shortest pathway decomposition with cutoff 12Å; (g) the shortest pathway illustration with cutoff 12Å; (h) the shortest pathway illustration with cutoff 5Å.
These pathways can be regarded as the different scale of interactions. For example, the pathways with the cutoff value 5Å focus on the allosteric effects leading by the covalent interaction or hydrogen bonding with each other. A cutoff 12Å provides the allosteric effects by the electrostatic interactions. It is more comprehensive to study the pathways with different cutoff value rather than a particular cutoff value as shown in the reference. Regarding the investigation target, certain pathways could be more interesting than other pathways.

Other than the shortest pathway calculation, the communities can provide a comprehensive decomposition of the total allosteric effects into the allosteric effects caused by the interaction between different communities. The minimization of total relative entropy value inside each community is certain to be a NP-hard problem where only heuristic solution can be proposed for the optimization problem. Although the KL and GN algorithm can provide different communities results, the secondary structure is mostly likely to be conserved in those communities. Therefore, those communities can provide insight into the allosteric effects related to the overall protein structures. As shown in Figure 11, the community detection algorithm can separate the overall distribution difference into the distribution differences over different communities. Figure 11 (b,g,k,n,p) suggests that the distribution inside each community is similar between unbound and bound state, while the distribution differences mainly comes from the distribution differences between communities.
Figure 11: Projection of PDZ2 bound and unbound state using PCA. (a) Projection onto PC1 and PC2 surface; (b-p) Projection of different community pair onto pair-specific PC1/PC2 surface

In this study, a new network model named relative entropy based dynamical allosteric network (REDAN) model was developed to quantitatively characterize protein allosteric effects upon perturbations. Comparing with the other protein network models, one key difference is the application of relative entropy on the pairwise distance distribution to quantify the significance of allosteric effects inside protein. As a consequence, the network can represent the significance of allosteric network inside the protein from the structure point of view. Moreover, because the distribution is directly related to the free energy, the changes in the distribution is essentially the changes of free energy surface due to the external perturbations.

Shortest pathway algorithm and community detection algorithm can shed light on the allosteric mechanism. The allosteric effects between two distal residues can be decomposed through a set of intermediate residues, where the smaller allosteric effects accumulate to larger allosteric effects. The cutoff value in the network provide flexibility for different interactions
with different scales. Meanwhile, the allosteric community analysis could further identify the communities. Because the secondary structure tends to be conserved inside the communities, the investigation of allosteric mechanism could focus on the interaction between those communities rather than the interactions inside the community. Overall, the REDAN model can be applied on any proteins upon perturbations, and identify the residue pairs with the most significant allosteric effects, search the potential allosteric communication pathways, and reveal the allosteric communities. These applications could greatly improve the understanding of the allosteric mechanism for many proteins.
6. Markov Chain in Protein: Markov State Model

6.1 Markov State Models (MSMs) Based on Molecular Dynamics Simulation

In the above analyses, both machine learning and network model were used to investigate allostery from both structural point of view and thermodynamic point of view. However, investigation of allostery through the kinetic point of view is absent in the above studies. Nevertheless, one of the frameworks named Markov State Models (MSMs) has been developed to investigate molecular dynamic simulation from the kinetic point of view.\textsuperscript{46, 116-118} After first proposed in early 2000, the method is widely applied in many protein systems after 2010,\textsuperscript{119-121} and becoming a standard toolkit for analyzing molecular dynamic simulations.

Before introducing markov state model in chemistry, markov state, or more generally, markov chain is already a widely applied model in many fields.\textsuperscript{122} One of the applications is the google page ranking algorithm. The heart of the google page ranking algorithm depends on the markov state hypothesis, which states that the probability of traversing from one page to another page is independent to the previous history.\textsuperscript{123} As a result, the “importance” of each page can be quantified through the stationary distribution of these pages. The key property of markov state is the Markovian property.\textsuperscript{124} A stochastic process has the Markovian property only when the conditional probability distribution of the future state depends on the present state, not on the history. With the Markovian property, the markov chain can be categorized into finite states / continuous states and discrete-time / continuous-time markov chain. The most widely used markov chain is the finite states discrete-time markov chain, such as Markov State Models...
One of the interesting properties of finite state discrete markov chain is the stationary distribution. Upon the calculation of the transition probability matrix between each state in the markov chain model, the following property holds for the matrix column sum as 1.0, which is known to be Frobenius theorem.125

1. $\lambda = 1.0$ is one of the eigenvalue of the transition probability matrix $P$.
2. Any eigenvalue of the transition probability matrix $P$ is smaller than 1.0.
3. The eigenvector corresponding to the eigenvalue 1.0 is the stationary distribution of the markov chain.

With the above property, it is clear that the stationary distribution of any markov chain can be obtained through the transition probability matrix, which is the heart of google page ranking algorithm. Overall, markov chain is applied in many fields including computer science, biology, chemistry, etc. The hidden markov state is also widely applied in biology to study the gene expression.121, 126-127

Markov state model is particular suited for the protein simulations because theoretically the states in the phase space hold the markov properties. Although in reality, due to the dimension reduction and clustering, the markov property does not exactly hold. However the markov state model is still a good practice to estimate the protein kinetics during the simulation.88, 118 Traditionally, analysis of MD trajectories is mainly involved with the “look and see” analyses of the rare events through the molecular movies or projection of the simulation onto several chemical related collective variables.118 Although it can directly display the chemical relative information through those collective variables or reaction coordinates, those artificially constructed variables could hide the barrier or distort the free energy surfaces or oversimplify the surface. Therefore, markov states models are introduced to overcome this limitation. The
Markov State Models or MSMs model the protein kinetics through a N*N probability matrix with a discrete “lag time” τ. The probability matrix entry $P_{ij}$ indicates the conditional probability of finding the state j starting from the state i after the τ time. Comparing with transitional analysis methods, MSMs abandons the view of single trajectory where focus on the ensemble of trajectories, providing the calculation schema of some observances or statistical property through the transition probability matrix.\textsuperscript{118} Overall, comparing with other analyses methods, MSMs can model the kinetic information of the state transformation, which is particularly important to investigate allosteric mechanism through the kinetic point of view.

6.2 Application of MSM on VVD

In the study of investigating VVD conformational changes through the kinetic view, Markov State Models have been applied with the combination of the machine learning classification models. In the current framework, unlike the normal classification method which focuses on the structural similarity, the markov state models take the advantage of the kinetic information. Each markov state represents a different kinetic basin in the free energy landscape. The classification model can be further applied to these clusters to investigate the differences between each cluster and unravel the allosteric mechanism.

In the current VVD simulations, total four different configurations of simulations were conducted to investigate the role of the covalent bond (FAD – Cys108) in the conformational changes. More specifically, the dark structure with and without the covalent bond and the light structure with and without the covalent bond were subjected to MD simulations. The dark structure without the covalent bond, and the light structure with the covalent bond are referred to as the native dark and native light states, respectively. And the dark states with covalent bond and the light state with the covalent bond are referred to as the transient dark and transient light
configuration, respectively. For each configuration, three independent 1μs simulations were conducted. Therefore, total of 12 μs have been conducted. Those simulations are plotted in Figure 12. The dark state configuration shows less fluctuation than the light state configuration, which is consistent with RRS results.

Figure 12: VVD protein with four different configurations. (a) 2D-RMSD plot with reference to the native dark and light structures, respectively; (b) Estimated relaxation timescale based on transition probabilities

To apply markov state analysis, those structures in the simulations are clustered into different microstates first before the estimation of the transition probability matrix. Here, k-means clustering analysis was applied to divide the configuration space into 300 clusters. The lag time is estimated in Figure 12b as 30ns, when the relaxation timescale started to be converged. The number of macrostates chosen to be eight after several try-and-errors, which can best represent the free energy landscape without losing too much generalization. Overall, those macrostates with the representative structures are shown in Figure 13. Clearly, the representative structure of those macrostates is mainly different in the N-terminal, where the backbone structures are similar. The observation is consistent with the experimental studies, which showed that the N-terminal is the most distinguishable part between dark and light structure.60, 128-129
After building the markov state model, the markov property of the model has to be validated before extracting any information from this model. To validate the Markovian property, the typical procedure involves a Chapman-Kolmogorov test shown in Figure 14. The detail procedure and equations of the Chapman-Kolmogorov test are demonstrated in literature equation 66. In the current MSMs, although the markov property does not hold precisely, the MSMs can be used as the approximation to estimate the kinetic property of the VVD system.

After constructing the markov state model, the covalent bond effects can be investigated through the differences between the transition probability matrix for non-bonded and bonded configuration respectively as shown in Figure 15. As shown in Figure 15b and 15c, the formation of the covalent bond can significantly change the transition probabilities among macrostates.
Therefore, the covalent bond could alter the free energy surface and barrier between states. Specifically, with the covalent bond, the transition probability between 3:2, 7:5, 5:4 have increased from (3%, 4%, 11%) to (14%, 10%, 14%), respectively. The significant increase of transition probability demonstrates that the effects of covalent bond on the overall protein dynamics, and facilitate the conformational change from dark conformation to the light conformations.

![Figure 15: Transition Probability Matrix for (a) all trajectories, (b) non-bonded configuration, (c) bonded configuration](image)

Meanwhile, the steady-state distribution of non-bonded configuration also has a significant difference from the bonded configuration. The hidden state 6 is only observed in the non-bonded configuration simulations. With the covalent bond, state 8 is observed as the stable state to the light state without reaching the hidden state. According to those observations, the covalent bond is crucial to the light state conformation changes.

After constructing the markov state through MSMs, to further investigate the structural differences between each state and identify the most important region related to conformational change, the above machine learning and community methods were applied in the current study as well. Unlike the previous study, in which the classification is only conducted as binary classification, in the current study, the classification was conducted over all eight states.
Different models were applied for the classification purpose. Meanwhile, because this is a multi-label classification, the one-vs-one (ovo) classification strategy was applied to avoid being biased by any label. The result of the classification is shown in Figure 16.

Figure 16: Training models for the macrostates using different ML models: (a) Artificial Neural Networks; (b) Decision Tree; (c) Random Forest; (d) one-vs-one random forest

Comparing with the dummy classifier with accuracy $1/8 = 12.5\%$, all four ML models have successfully achieved higher training accuracy as 95.0\%, 98.3\%, 98.1\% and 99.1\%. The highest validation accuracy is the neural network model as 90.1\%. The random forest and one-vs-one random forest model show lower accuracy, but still significantly higher than the dummy classifier, suggesting the efficiency of those classification models. Upon constructing the ML models, similar to the previous study using ML models and network models, the most distinguishable feature can be extracted from those models as Figure 17. For several features including Feature 213 and Feature 227, the pairwise distance distribution is significantly
different in those states, while feature 4976 shows significant similarity among all the states. The top 5 features are listed in Table 3.

![Feature Distribution](image1)

**Figure 17:** Distribution between selected features: (a) Feature 213; (b) Feature 227; (c) Feature 4976

The residues corresponding to the top 5 features have been identified in several studies. Although those residues information is interesting to learn, without a systematic analysis to combine the information from those residues, the underlying mechanism is still unclear. Therefore, similar to the community analysis in the network model, using the feature importance as the weights, the residues are divided by the communities. The following investigation will focus on the structural differences between communities.

**Table 3:** Top 5 features with the highest importance during the classification

<table>
<thead>
<tr>
<th>features</th>
<th>OVERALL</th>
<th>Rank: 1</th>
<th>Rank: 2</th>
<th>Rank: 3</th>
<th>Rank: 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>213 (T38 – G105)</td>
<td>2.83%</td>
<td>S3-S5: 4.28%</td>
<td>S1-S7: 4.03%</td>
<td>S5-S6: 4.00%</td>
<td>S2-S6: 4.00%</td>
</tr>
<tr>
<td>227 (T38 – K119)</td>
<td>2.38%</td>
<td>S2-S4: 4.55%</td>
<td>S3-S4: 4.00%</td>
<td>S1-S2: 4.00%</td>
<td>S2-S6: 4.00%</td>
</tr>
<tr>
<td>189 (T38 – K31)</td>
<td>1.75%</td>
<td>S3-S7: 4.28%</td>
<td>S2-S5: 4.08%</td>
<td>S3-S5: 4.05%</td>
<td>S1-S7: 4.02%</td>
</tr>
<tr>
<td>29 (H37 – V67)</td>
<td>1.52%</td>
<td>S1-S7: 4.05%</td>
<td>S3-S6: 4.00%</td>
<td>S2-S8: 3.99%</td>
<td>S3-S8: 3.99%</td>
</tr>
<tr>
<td>355 (L39 – E102)</td>
<td>1.51%</td>
<td>S5-S6: 5.72%</td>
<td>S4-S7: 5.48%</td>
<td>S1-S6: 5.31%</td>
<td>S4-S8: 4.44%</td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pcbi.1006801.001

The overall structure is divided into four different communities. Interestingly, community A includes N-terminus from H37 to G43. Community B identified the A’α/Aβ and Eα/Fα loops, and the residue C108 which is bonded to the ligand. Communities C and D consist the majority of protein including Aβ strand, Bβ strand, Dα helix, Fα helix and Gβ strand (Community C), and
A’α helix, Cα helix, Eα helix, Hβ stand, and Iβ strand (Community D). The feature importance between communities and inside each community is listed in Table 4. Clearly, the feature importance inside each community is minimized to less than 0.2%, whereas the feature importance between communities is maximized.

Table 4: Accumulated feature importance between each community

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Commu. A</td>
<td>0.041%</td>
<td>15.885%</td>
<td>37.085%</td>
<td>36.445%</td>
</tr>
<tr>
<td>Commu. B</td>
<td>0.107%</td>
<td></td>
<td>5.236%</td>
<td>3.867%</td>
</tr>
<tr>
<td>Commu. C</td>
<td></td>
<td>0.217%</td>
<td></td>
<td>0.924%</td>
</tr>
<tr>
<td>Commu. D</td>
<td></td>
<td></td>
<td></td>
<td>0.191%</td>
</tr>
</tbody>
</table>

The feature importance between Community A with the rest of protein accumulate to 89.415% importance in the one-vs-one random forest classifier. The result is not surprising because N-terminal is the most flexible part between native dark and light states of VVD. However, interestingly, although Community B only contains several residues, the feature importance of Community B with C and D still accounts for 9.103% of total importance. As a comparison, despite that fact that Communities C and D comprise the majority of protein, the feature importance between them and among them only take less than 1% of total feature importance. The above result clearly suggests the importance of Community B, or more specifically, the A’α/Aβ and Eα/Fα loops. To further investigate the role of A’α/Aβ and Eα/Fα loops in the conformation changes, the feature importance of individual states transition is listed in Table 5. From Table 5 and Figure 15, it is clear that for the adjacent states, Community B has the significant feature importance, while in the non-adjacent states, the feature importance of Community A dominant. Therefore, it is reasonable to hypothesize that the Community B changes before N-terminal switches. In other word, the A’α/Aβ and Eα/Fα loops serve as the switch to the N-terminal changes. The transition has a higher probability than for N-terminus
directly shifting. Moreover, as stated above, the bonded configuration has a higher probability to change from dark to light comparing with the non-bonded configuration. Therefore, it’s reasonable to hypothesize that the covalent bond could alter the flexibility of Community B first which could subsequently change the Community A.

Table 5: The changes of Community A and Community B during transition between states

<table>
<thead>
<tr>
<th>Adjacent macrostates</th>
<th>A with C and D</th>
<th>B with C and D</th>
</tr>
</thead>
<tbody>
<tr>
<td>State 3 (crystal dark) → State 2</td>
<td>38.35%</td>
<td>46.38%</td>
</tr>
<tr>
<td>State 2 → State 7</td>
<td>50.12%</td>
<td>30.22%</td>
</tr>
<tr>
<td>State 3 → State 7</td>
<td>92.65%</td>
<td>0.87%</td>
</tr>
<tr>
<td>State 7 → State 5</td>
<td>59.85%</td>
<td>19.37%</td>
</tr>
<tr>
<td>State 5 → State 4 (crystal light)</td>
<td>71.15%</td>
<td>17.60%</td>
</tr>
<tr>
<td>State 4 → State 1</td>
<td>50.38%</td>
<td>38.01%</td>
</tr>
<tr>
<td>State 7 → State 4</td>
<td>85.09%</td>
<td>0.32%</td>
</tr>
<tr>
<td>State 1 → State 8</td>
<td>94.97%</td>
<td>0.52%</td>
</tr>
<tr>
<td>State 1 → State 6</td>
<td>74.41%</td>
<td>3.72%</td>
</tr>
<tr>
<td>Non-Adjacent macro-states</td>
<td></td>
<td></td>
</tr>
<tr>
<td>State 2 → State 8</td>
<td>79.71%</td>
<td>0.01%</td>
</tr>
<tr>
<td>State 3 → State 4</td>
<td>91.27%</td>
<td>0.01%</td>
</tr>
<tr>
<td>State 3 → State 6</td>
<td>81.98%</td>
<td>0.01%</td>
</tr>
<tr>
<td>State 3 → State 1</td>
<td>78.13%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

In this study, we adopted the Markov State Models (MSMs) to investigate the kinetics of protein conformational change. Unlike the normal clustering methods, MSMs take advantage of the kinetic information, and each state is corresponding to a different minimum in the free energy surface. Besides that, the transition probability matrix for unbound and bound state trajectories provides more information regarding the impact of covalent bond on the overall protein dynamics. Meanwhile, by integrating the MSMs with the machine learning and community analysis, certain regions in the protein can be identified to be crucial to the protein conformational changes. Experimental observations also support our findings in A’α/Aβ and Eα/Fα loops. As a conclusion, this study provides a systematic way to examine the allosteric
mechanism related to conformational change by incorporating MSMs, machine learning, and community analysis.
7. Directed Kinetic Transition Network (DKTN) Model

7.1 Limitation of Discrete Markov State Model

The above study demonstrates the advantage of Markov State Models, which the kinetic information can be incorporated into the MSMs. With the combination of ML models, the important residues and communities can be identified. Using VVD protein as the case study, the underlying mechanism for the conformational changes was revealed. However, there are certain limitations of MSMs which are addressed in the current study. One potential issue of applying MSMs is that the key approximation, markovian properties, may not hold for the discrete states. Although some theoretical studies demonstrated that the approximation could be precise when the conformational spaces are appropriately discretized\textsuperscript{118, 132-133}, in practice, depending on certain dimensionality reduction methods and clustering algorithms applied in MSMs, producing appropriate discretization is still challenging in many cases. In most cases, different dimensionality reduction methods could lead to dramatically different MSM based on the same simulations result\textsuperscript{134-136}. Another important factor to be considered for MSMs is the lag time $\tau$. Because the transition probability needs to be estimated based on a given lag time, the selection of a proper lag time is critical to the quantity of MSMs. Unfortunately, the selection of lag time may not be asymptotic, which makes the determination of lag time to maximize markovian property of the system challenging.

Upon those limitations, the basic assumption of the kinetic model is revisited from the Markovian property to the rate theory, namely the first order chemical reactions. The model
named Directed Kinetic Transition Network (DKTN) is developed based on the Transition Network (TN) model, where the TN is a discrete representation of a conformational space, and represents the conformational changes through a network of sub-transitions.\textsuperscript{137-138} The discretized states in the TN model are similar to the states in the MSM as a representation of intermediates between stable conformations. However, different from MSMs, the edges in the TN model represent the free energy differences of two states instead of the transition probabilities. To further improve TN model, the reaction rate is estimated from the average state transformation time and the detailed balance constraints in the DKTN model. The edges in the DKTN model are directed edges with two directions modeling the reversible transitions. Therefore, the DKTN system could be represented as a solvable ordinary differential equation (ODE) system. We will also show that DKTN model can be reformulated into a continuous time Markov chain (CTMC), which can be regarded as a special MSM without the lag time $\tau$ and inconstant transition probability matrix. The building blocks for the new model are the estimation of distributions and "mean time of transition" between different states. Both quantities can be estimated directly from the simulation. Overall, comparing with the TN model and the MSMs model, the DKTN model can model the system evolving from non-equilibrium state to the equilibrium state. The TN model or MSM model can be regarded as the special case of DKTN model where the system reaches the equilibrium. The efficiency of this model is demonstrated using VVD, to reveal some key secondary structure changes related to the conformational switching process between the dark and light conformation.
7.2 Transition Kinetic Model and Continuous Time Markov Chain (CTMC)

Constructing Directed Kinetic Transition Network (DKTN) model

To construct DKTN model, the basic building blocks include the “mean time to transition” and the equilibrium distribution. Based on these two metrics and the detailed balance constraints, the overall model can be established. Similar with the MSM model, the first step of DKTN model is also clustering the structures into the microstates based on the structural similarity. Similar with the MSM model, the averaged RMSD less than 1.0Å within the microstate implies the sufficient kinetic similarity. First of all, the equilibrium Boltzmann distribution of each microstate $\pi_s$ is estimated as the percentage of the number of snapshots in state S within all number of snapshots: $\pi_s = \frac{N_s}{\sum_s N_s}$.

The free energy of each microstate could be estimated as the following equation

$$E_s = -k_B T \ln \pi_s$$  \hspace{1cm} (1)

Regarding the edge connecting two different microstates $\mu \rightarrow \nu$, and the detailed balance constraint as a condition for equilibrium and reversible property of molecular dynamics simulation, the following equation will hold

$$k_{\mu \nu} = k_{\nu \mu} = \pi_\mu k'_{\mu \nu} = \pi_\nu k'_{\nu \mu}$$ \hspace{1cm} (2)

where $\pi_\mu, \pi_\nu$ is the Boltzmann distribution of microstate $\mu$ and $\nu$, and $k'_{\mu \nu}$ and $k'_{\nu \mu}$ is the reaction rate constants for the transitions as $\mu \rightarrow \nu$ and $\nu \rightarrow \mu$. The $k_{\nu \mu}$ and $k_{\mu \nu}$ is equilibrium reaction rate for the transitions $\mu \rightarrow \nu$ and $\nu \rightarrow \mu$, also referred to as the equilibrium flux in the TN models. Following the detailed balance constraint, $k_{\nu \mu}$ and $k_{\mu \nu}$ are equivalent in the equilibrium.
The reaction rate represents how fast a transition between two microstates occurs, and is the inverse of the mean time for the transitions between them. Therefore, in the equilibrium, the mean time for the transition between two states is given by the inverse of the flux. Obviously, the "meantime to transition" for μ→ν and ν→μ are identical in the equilibrium, also defined as τ_{νμ} as the following equation.

\[ \tau_{νμ} = k_{νμ}^{-1} = k_{μν}^{-1} \]  

(3)

In the current study, the "meantime to transition" (MTT) or τ_{νμ} is estimated through the collection of transitions in the equilibrium simulation as the average value of the transition time between any two adjacent microstates in the simulations.

As shown in Figure 18, all the adjacent transitions as μ→ν or ν→μ should be collected in the given simulations. For each transition, it is assumed that the starting timestamp for state μ or ν is t_s, and the ending timestamp for the other state ν or μ is t_e, respectively. For this instance of transition, the transition time between μ and ν is defined as (t_e - t_s)/2. Collecting all instances of μ→ν or ν→μ transitions, the τ_{νμ} or MTT between μ and ν is estimated as
\[
\tau_{\nu\mu} = \frac{1}{n} \sum_{i=1}^{n} \frac{t_{ei} - t_{si}}{2}
\]

After the estimation of \(\tau_{\nu\mu}\), the reaction rate constants for transitions \(\mu \rightarrow \nu\) and \(\nu \rightarrow \mu\) can be rewritten as

\[
k'_{\nu\mu} = (\pi_{\nu} \tau_{\nu\mu})^{-1}, \quad k'_{\mu\nu} = (\pi_{\mu} \tau_{\nu\mu})^{-1}
\]

These reaction rate constants \(k'_{\nu\mu}\) and \(k'_{\mu\nu}\) are used as the directed edge constant connected two microstates in the DKTN model.

Overall, the basic building blocks of DKTN model include microstates (nodes V), transitions between microstates (edges E), and the reaction rate constants for the transitions (edge weights W). Unlike the undirected TN models, which are static networks representing the equilibrium flux only, the DKTN model represents the kinetic property of system by chemical kinetic models. After constructing the DKTN model, the process of system evolution from non-equilibrium to equilibrium can be obtained by solving the ODE equation.

**Solving Equilibrium Distribution**

In the current study, the DKTN model is constructed from the simulation for reaction rate using equilibrium distribution and MTT. However, the only necessary information required in the DKTN model is the reaction rates. Upon successfully constructing DKTN model, the equilibrium distribution and MTT can be solved reversely from the reaction rates. In other word, if some reaction rates can be obtained experimentally, we indeed could improve the accuracy of DKTN model by replacing the estimated reaction rates with the measured reaction rates.
In general, for a given DKTN model that can evolve into an equilibrium distribution defined as $p$, the following equations could hold for any edges connecting two nodes $\mu$ and $\nu$ as $E_{\mu\nu}$ and $E_{\nu\mu}$ (the weights of edges) in the DKTN model, and we have the following equation.

\[ p_\mu E_{\mu\nu} - p_\nu E_{\nu\mu} = 0 \ orall E \]  \hspace{1cm} (6)

\[ \sum_{\mu \in \nu} p_\mu = 1 \]  \hspace{1cm} (7)

The above equations are linear equations, which can be solved analytically using the linear programing.\textsuperscript{139} The solution of $p$ for the above equations is identical with the $\pi$ estimated from the simulations. This guarantees that the established DKTN model could reach the steady equilibrium Boltzmann distribution from given starting conditions.

Meanwhile, the exactly time evolved function can be obtained by solving the ODE equations as the following.

\[ \frac{dp_1}{dt} = \left( - \sum_{i=1}^{n} E_{1i} \right) p_1 + E_{21}p_2 + E_{31}p_3 + \cdots + E_{n1}p_n \]

\[ \frac{dp_2}{dt} = E_{12}p_1 + \left( - \sum_{i=1}^{n} E_{2i} \right) p_2 + E_{32}p_3 + \cdots + E_{n2}p_n \]

\[ \vdots \]

\[ \frac{dp_n}{dt} = E_{1n}p_1 + E_{2n}p_2 + E_{3n}p_3 + \cdots + \left( - \sum_{i=1}^{n} E_{ni} \right) p_n \]  \hspace{1cm} (8)

where $p_i$ represents the distribution of microstate $i$ at any given time. If there is no edge between microstate $\mu$ and $\nu$, then both $E_{\mu\nu}$ and $E_{\nu\mu}$ are zero. Assume that the reaction rate
constant for any microstate returns to it itself is zero, which indicated by \( E_{ii} = 0 \ \forall \ i \in V \), the above equations can be written in matrix form as

\[
p'_t = Ap_t
\]

where

\[
p_t = \begin{pmatrix} p_1(t) \\ \vdots \\ p_n(t) \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix}
E_{12} & (-\sum_{l=1}^{n} E_{1l}) & E_{31} & \cdots & E_{n1} \\
E_{13} & E_{23} & (-\sum_{l=1}^{n} E_{2l}) & \cdots & E_{n2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
E_{1n} & E_{2n} & E_{3n} & \cdots & (-\sum_{l=1}^{n} E_{nl})
\end{pmatrix}
\]

The above first order ODE system was widely studied, and for a specific initial condition \( p(0) = p_0 \), the result for the above equation can be obtained as the matrix exponential of \( A \) as the following.

\[
p_t = e^{At} p_0
\]

where \( e^{At} \) is given by the following power series.

\[
e^{At} = I + At + \frac{1}{2!} t^2 A^2 + \cdots + \frac{1}{n!} t^n A^n + \cdots
\]

The matrix exponential can be solved through eigen decomposition.

\[
A = U D U^{-1}
\]

\[
e^A = U e^D U^{-1} = U \begin{bmatrix}
e^{\lambda_1} & 0 & \cdots & 0 \\
0 & e^{\lambda_2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & e^{\lambda_n}
\end{bmatrix} U^{-1}
\]

After solving the above ODE system, the DKTN model can produce an exact time evolution description of the distribution among microstate until reaching equilibrium. Therefore, the
DKTN model is a dynamical model, which could evolve over time. This property makes DKTN model different from both the traditional TN models, which can only describe equilibrium conditions, and MSMs, in which the system evolution is discretized by the lag time.

**Relationship with Continuous Time Markov Chain (CTMC)**

By reading the matrix form \( \mathbf{p}_t = \mathbf{A} \mathbf{p}_t \) of DKTN model, the connection with CTMC is obvious since the evolution of CTMC also follows \( \mathbf{P}_t' = \mathbf{Q} \mathbf{P}_t \), where \( \mathbf{Q} \) is known as rate matrix, \( \mathbf{P}_t \) is known as the transition probability matrix. By comparing those two equations, it is obvious that the evolution in DKTN model is identical with the CTMC. The ordinary transition matrix \( \mathbf{A} \) in DKTN model functions similar to the transition rate matrix \( \mathbf{Q} \) in the CTMC model. Noted that the transition rate matrix in the CTMC model is row-sum-zero matrix\(^3\), however, the ordinary transition matrix \( \mathbf{A} \) in DKTN is column-sum-zero matrix. Therefore, we can define \( \mathbf{Q} \) by \( \mathbf{A}^T \) to reformulate DKTN into CTMC as \( \mathbf{Q} \equiv \mathbf{A}^T \).

The special characteristic of DKTN model \( \pi_t \mathbf{A} \mathbf{A} = \pi_j \mathbf{A} \mathbf{A} \) for all \( i \) and \( j \) is known as the reversibility of CTMC model. Upon the reversibility, certain properties hold for reversible CTMC model also hold for DKTN model. One of them is the following \( \pi_t \mathbf{e}^A \mathbf{A} = \pi_j \mathbf{e}^A \mathbf{A} \) for all \( i \) and \( j \). The equation can be rewritten as \( \frac{1}{\pi_i} \mathbf{e}^A \mathbf{A} = \frac{1}{\pi_j} \mathbf{e}^A \mathbf{A} \), where \( \mathbf{e}^A \mathbf{A} \) represents the distribution for microstate \( j \) at time \( t \) starting with microstate \( i \), and \( \mathbf{e}^A \mathbf{A} \) represents the distribution for microstate \( i \) at time \( t \) starting with microstate \( j \). Therefore, at any given time \( t \), the percentage to the equilibrium for microstate \( j \) in \( i \rightarrow j \) transition is identical with the percentage to the equilibrium for microstate \( i \) in \( j \rightarrow i \) transition.
Demonstration of DKTN model using model system

Here, we provide a simple model system illustrating the similarity between the DKTN and CTMC models. This model system contains 4 states, named as A, B, C, and D. The reaction rate constants among these states with arbitrary values are illustrated in Figure 19.

![Diagram of the model system](image)

Figure 19: A model system for comparison between the DKTN and CTMC models

The above system contains 4 edges, connecting A and B, B and C, B and D and C and D, respectively. First, the equilibrium distribution \( \pi_A \) to \( \pi_D \) can be determined for this DKTN model through solving the linear equations as the following.

\[
0.1\pi_A - 0.15\pi_B = 0 \\
1.0\pi_B - 2.0\pi_C = 0 \\
0.2\pi_B - 0.1\pi_D = 0 \\
0.4\pi_C - 0.1\pi_D = 0 \\
\pi_A + \pi_B + \pi_C + \pi_D = 1.0
\]

(13)

The solutions to the above equations are 0.3, 0.2, 0.1, 0.4, respectively, independent with the initial condition. Next, the ODE transition matrix \( A \) can be obtained, and the time evolution of the system for a special initial condition can be solved. Propagation of this system starting from initial conditions with only state A or state D respectively, are plotted in Figure 20. As a notice,
unlike the final equilibrium distribution, the time evolved distribution is dependent on the initial condition.

Figure 20: Propagation of distribution starting from different initial conditions: (a) Starting exclusively from state A, (b) Starting exclusively from state D.

By reformulating the DKTN into the CTMC model, the transition probability matrix evolve over time as the following. The transition probability matrix among microstates for initial condition at time 0s, at time 10s and at equilibrium are calculated as $e^{At_{t=0}}$, $e^{At_{t=10}}$, $e^{At_{t=\infty}}$, respectively. The transition probability matrix is shown in Figure 21. The transition probability matrix only become constant when the system becomes equilibrium.

Figure 21: Continuous-time Markov chain (CTMC) model: The transition probability matrix and result MSM model at (a) 0s; (b) 10s; and (c) equilibrium.
To qualify how fast the system reaches the equilibrium, using the similar idea from chemical kinetics as the half time, a half time is defined here as the time required for certain states to reach halfway to the equilibrium concentration. The halftime for certain states can be estimated from the equation. Upon the calculation of halftime, the “effective reaction rate” can be calculated as $k_{eff} = \frac{\ln 2 \pi \rho}{t_{half}}$.

Using effective reaction rate as metrics, we can quantify the importance of each conformational change to the overall transitions. As shown in the following picture, removing different edges will slow down the time required for the states to reach the equilibrium. Different edges have different importance regarding the overall time to reach the equilibrium.

$$k_{eff} = \frac{\ln 2 \pi \rho}{t_{half}}$$  \hspace{1cm} (14)

Figure 22: The half time and evolution for different system including whole edges, removing edge A to B, removing edge B to C, removing edge B to D, and removing edge C to D. (a) the system starting at state A, the distribution of state D; (b) the system starting at state D, the distribution of state A.
Notice, although the above are two different reactions and two different states: the distribution of D for the reaction $A \rightarrow D$ and the distribution A for the reaction $D \rightarrow A$. The trends are identical with each other despite the difference of y axis distributions. This is because of the reversibility and detailed balance constraints. The impact to the effective transition from A to D is represented as the difference of the half time and effective reaction rate constant after removing different edge.

Table 6: Individual edge importance for the transition $A \rightarrow D$

<table>
<thead>
<tr>
<th>Effective transition</th>
<th>Half Time (s)</th>
<th>Effective Reaction Rate Constant (s$^{-1}$)</th>
<th>Edge Importance$^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole system</td>
<td>7.878</td>
<td>0.0352</td>
<td></td>
</tr>
<tr>
<td>Removing $A \leftrightarrow B$</td>
<td>$\infty$</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Removing $B \leftrightarrow C$</td>
<td>9.698</td>
<td>0.0286</td>
<td>18.76%</td>
</tr>
<tr>
<td>Removing $B \leftrightarrow D$</td>
<td>11.205</td>
<td>0.0247</td>
<td>29.68%</td>
</tr>
<tr>
<td>Removing $C \leftrightarrow D$</td>
<td>10.173</td>
<td>0.0275</td>
<td>22.55%</td>
</tr>
</tbody>
</table>

1. Half time is numerically estimated as the time for state D to reach half equilibrium.
2. Effective reaction rate constant is calculated as $ln2\pi_D/t_{half}$.
3. The importance of each edge is calculated as decreasing of effective reaction rate constant as $(k_{eff}^{System} - k_{eff}^{Remove Edge})/k_{eff}^{System}$. This value is also equal to $1 - (t_{half}^{System}/t_{half}^{Remove Edge})$.

As listed in Table 6, the importance of each edge is calculated as the percentage of the decreasing effective reaction rate constant after removing the individual edge, and can also be represented as the increasing percentage of halftime, $1 - (t_{half}^{System}/t_{half}^{Remove Edge})$. For this model system, we can see that for the transition $A \rightarrow D$. The edges $A \leftrightarrow B$ has 100% importance, because after removing this edge, the state D becomes unreachable from the state A. Removing edges between states B and D will decrease the effective reaction rate constant for the transition $A \rightarrow D$ by 29.68%, which is the second highest after 100%. Therefore, the edges between states B and D are more important than the edges between states B and C, as well as the edges between
C and D. It may be concluded that the processes contributing the most to the transitions between A and D is the element conformational changes between states A and B, as well as B and D.

For a complex biological system, identifying the important conformational changes between two biologically significant states could provide some novel insights. The application of the DKTN model on the VVD protein will be presented in details to illustrate the effectiveness and value of this theoretical model for biomacromolecules.

7.3 Application of DKTN on VVD protein system

The DKTN model is applied on the VVD protein system. Several steps are conducted, including the clustering of microstates, estimating the reaction rate, constructing the DKTN model, and solving the DKTN models. For the first step, the microstates are clustered using two steps, first clustering the metastable states using Gaussian Mixture Model (GMM), then the microstates is clustered based on the result of Gaussian Mixture Models.
Comparing with $k$-means or $k$-means based classifications, it is assumed in GMM that the underlying clusters should follow a gaussian distribution. In principle, the MD simulations follow the Boltzmann distribution, which is also a type of gaussian distribution. Therefore, this assumption should hold for the MD simulations. Therefore, the number of metastable states can be determined using GMM model as the reference.\textsuperscript{140-141} After classification of the metastable states, the microstates are further classified from the metastable states using 1.0Å as the cutoff value. Total 34 microstates have been classified, and the reaction rates for each of them are estimated using the DKTN model. Overall the DKTN model is constructed as the following.
Figure 24: Established DKTN model based on microstates: (a) The equilibrium reaction rate (also referred to as flux) between microstates in the equilibrium; (b) The rate constants between microstates

Figure 25: System evolution over time starting in microstate 3 with (a) 0%, (b) 50%, (c) 75.0% and (d) 99.9% to the equilibrium. System starting in microstate 8 with (e) 0%, (f) 50%, (g) 75.0%, and (h) 99.9% to the equilibrium.

Using the DKTN model, we can calculate the time and process to reach the equilibrium from dark and light states (Figure 25). This suggests a process for the non-equilibrium system to reach the equilibrium. After solving the ordinary differential equations for DKTN model with different initial conditions (starting from microstate 3 or microstate 8), the time evolution of each system
can be obtained analytically. The system's evolution to the equilibrium starting from microstate 3 was shown in Figure 25(a)-(d), representing 0%, 50%, 75.0% and 99.9% of diffusion of initial structure to the equilibrium, respectively. Similarly, the system's evolution to the equilibrium starting from microstate 8 is illustrated in Figure 25(e)-(h). Starting from the microstate 3 (dark conformation), system took around 139.54μs to reach equilibrium, while starting from the microstate 8 (light conformation), system took much less time, around 58.24μs to reach equilibrium. This demonstrates that the light conformation took less time than dark conformation to undergo conformational switching. Besides the above application of DKTN model on the evolution of system into equilibrium, the important conformational changes can also be determined using the changes of effective reaction rate upon removing different edges.

<table>
<thead>
<tr>
<th>Top 10 conformational changes decreased effective reaction rate for certain transition</th>
<th>Transition from microstate 3 to microstate 8</th>
<th>Transition from metastable state 1 (dark) to metastable state 2 (light)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1:17 (23.042%)</td>
<td>1:17 (22.226%)</td>
</tr>
<tr>
<td>2</td>
<td>3:17 (17.747%)</td>
<td>7:21 (19.179%)</td>
</tr>
<tr>
<td>3</td>
<td>7:21 (17.656%)</td>
<td>3:17 (17.013%)</td>
</tr>
<tr>
<td>4</td>
<td>8:20 (16.792%)</td>
<td>2:17 (15.437%)</td>
</tr>
<tr>
<td>5</td>
<td>2:17 (16.054%)</td>
<td>8:20 (12.665%)</td>
</tr>
<tr>
<td>6</td>
<td>4:17 (9.309%)</td>
<td>4:17 (8.912%)</td>
</tr>
<tr>
<td>7</td>
<td>17:33 (8.740%)</td>
<td>17:33 (7.757%)</td>
</tr>
<tr>
<td>8</td>
<td>3:4 (5.440%)</td>
<td>3:4 (5.138%)</td>
</tr>
<tr>
<td>9</td>
<td>8:33 (5.108%)</td>
<td>6:7 (4.319%)</td>
</tr>
<tr>
<td>10</td>
<td>17:21 (4.044%)</td>
<td>17:21 (4.178%)</td>
</tr>
</tbody>
</table>

In Table 7, the top 10 important conformational changes are listed as microstate A → microstate B with the decrease of effective reaction rate. The importance is measured in two
scenarios. The conformational change for two microstates or two metastable states. Nevertheless, the top important edges are close with each other. Among top 10 edges, only 1 of them is different in two columns. Overall, the microstates transitions from microstate 17 → 1/2/3 and microstates transition from 8 → 20 and 7 → 21 are the most important transition effects for the overall conformational changes. The further structural difference comparison of those microstates suggests the importance of N-terminal, Hβ/Iβ loop and A’α/Aβ loop, which is in agreement with our previous study using MSMs, validating of the current DKTN model.

In this study, DKTN model was introduced as a theoretical method to study kinetics in the protein conformational changes. The rate constant for any transition observed in the simulation could be accurately estimated using this method, providing critical kinetic information with regard to individual states. In addition, the DKTN model could also be used to identify dominant transition pathway between any state pairs and to provide potential target for potential kinetic regulations of the system. Comparing with MSMs, the DKTN method is logical expansions of TN model, and can considered as a general MSM model without the lag time or constant transition probability matrix. In addition, the DKTN model is a more general model than the transition network model, which can be considered as a special case in the DKTN model and only deal with equilibrium states. Both advantages and limitations of DKTN model are discussed with solutions. Overall, the DKTN model could be an effective computational tool to model complex dynamical processes related to macromolecules such as protein folding and allostery.
8. Other Projects: Related to Allostery

In this chapter, several side projects related to allostery in the past years will be briefly discussed including a dimensional reduction method: t-Stochastic Neighboring Embedding (t-SNE) and an enhanced sampling method: DPDS.

8.1 Dimensionality Reduction and t-SNE

For the high dimensional data like protein simulations, to better visualize and represent in the low dimensional space, dimensionality reduction or feature selection are the common strategies. Dimensionality reduction is commonly used in many data related fields, and an adequate dimensional reduction method is necessary to correctly characterize the features of data. For the MD simulations, using the “collective variable” to represent the simulations is a common strategy. However, because the collective variables or in other words, “reaction coordinates” usually are chosen manually. Inappropriate collective variables would certainly disguise important kinetic information and lead to unreasonable reaction barriers.\textsuperscript{118, 142-143} Therefore, to better represent high dimensional data, dimensionality reduction methods are commonly used to embed the high dimensional data onto the low dimensional space. In general, the dimensionality reduction methods can be categorized into the linear and non-linear categories.\textsuperscript{144-145} Linear dimensionality reduction methods are widely used in the MD simulations, such as principal component analysis (PCA), independent component analysis (ICA), time-dependent independent component analysis (t-ICA).\textsuperscript{146} However, in the previous study, non-linear dimensionality reduction methods generally have better performance than the linear dimensionality reduction
methods because the dynamics of systems lying on the highly curved and convoluted manifolds. Commonly used non-linear dimensionality reduction methods include diffusion map, isomap, and autoencoder neural networks.

These non-linear dimensionality reduction methods are usually based on an assumption that the data lie on a certain high-dimensional manifold. Although the assumption holds for many high-dimensional data, there are not enough evidence to suggest the structures of MD simulations also hold such assumption. Instead, one of the state-of-the-art dimensionality reduction methods, t-SNE, focuses on the similarity between low dimensional data and high dimensional data to avoid significant information loss in the low dimensional space. Because t-SNE does not require any presumptions of the manifold of the high-dimensional space, it could be a good fit for MD simulations. In the t-SNE method, Gaussian probability distributions over high-dimensional space are constructed and used to optimize the student t-distributions in the low dimensional space by minimizing the Kullback-Leibler divergence between these distributions. Overall, because of the similarity between those two distributions, the clusters in the high-dimensional space are usually also well-separated in the low-dimensional space, better representing the similarities among the data rather than PCA or ICA. In this study, the t-SNE is introduced to analyze the MD simulation with the comparison with PCA, t-ICA and 2D-RMSD to illustrate the advantage of t-SNE with other methods.
Figure 26: Ten clusters obtained from k-means clustering: (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 or across the different clusters.

Ten clusters are first classified using k-means clustering method based on the trajectories. Ten clusters are chosen because it is the minimum number of clusters to ensure the average RMSD within the same cluster is smaller than 1.0Å. These ten clusters are directly classified from the high dimensional Cartesian coordinates, and also represent different energy basins in the high dimensional space. The populations for these 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 pairwise RMSD difference between each cluster pair is shown in Figure 26a. The structural difference within the same cluster is much lower than the structural different among different clusters. Therefore, as a good low-dimensional descriptor, these clusters should be well separated to represent the barriers between different clusters. The results for different dimensionality reduction methods are shown in Figure 27.
Figure 27: Ten k-means clusters of VVD systems 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.

In Figure 27, different dimensionality reduction methods have been applied. The 2D t-SNE model has better results comparing with PCA, t-ICA and 2D RMSD. All ten clusters are well separated from each other (represented in different colors). Using the previous study of VVD protein, Cluster 8 is the hidden state which is different from both dark and light states. Clusters 2, 3, 4, and 6 belong to the light region, because the RMSD value are closer to the crystal light structure. Other states including 1, 5, 7, 9 and 10 belong to the dark region instead. Overall, the 2D RMSD surface and PCA surface could distinguish the hidden state and the light region. However, the states in the dark state region are not well separated in these two low-dimensional spaces. The barriers among those states are likely disguised. Meanwhile, t-ICA method successfully captures the slowest dynamics in the simulation as t-IC1 between the hidden state and the light state. The subtle barriers within the light region and dark region are well represented on the surface. Overall, the 2D t-SNE better represents different states than PCA, t-ICA and 2D RMSD in this study.
Besides the 2D t-SNE, even with one dimension, the t-SNE can correctly capture the essential dynamics of protein, as shown in Figure 28. With only one dimension, all ten clusters are still well separated from each other with the minimum overlap among them. As a comparison, the 1D PCA and t-ICA are also plotted in Figure 28. Clearly, the ten clusters significantly overlap with each other in PCA and t-ICA, which suggests that the 1D PCA and 1D t-ICA are not suitable to show the free energy surface. With 1D t-SNE, one additional dimension can be added to better demonstrate the relationship between the time and free energy landscape, which can be used to investigate protein folding pathways or the conformational changes.

Figure 28: 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.

In this study, the t-SNE method was introduced for the analysis of molecular dynamics simulations, and obtained the superior performance comparing with other linear dimensionality reduction methods. The advantage of the t-SNE method is retaining the pairwise distance distribution information among the high dimensional data by maximizing the similarity between the high dimensional Gaussian distributions and low dimensional t-distribution using Kullback-Leibler divergence. Overall, by using VVD protein as the case study, it is clear that t-SNE can
better represent different clusters and free energy landscape than PCA, t-ICA and 2D RMSD. Meanwhile, the 1D t-SNE model can be used with the time dependent fitting analysis 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 comparing to other commonly used dimensionality reduction methods, and could be applied as a common toolkit for the analyses of simulations for other biomacromolecules.
8.2 Direct Pathway Dynamical Sampling

Another concern for investigating allostery is the efficiency for sampling. In principle, following the ergodic hypothesis, the equilibrium or the possible conformational space can be achieved only through sufficiently long simulations.\textsuperscript{150} In practice, because of the different energy basins on the free energy landscape, the simulation can be easily trapped in a local minimum without reaching other minima. In general, the transition state regions, or the high energy regions are difficult to be sampled. Many enhanced sampling methods have been developed to address this issue. Umbrella sampling methods, although not very efficient, are the most widely used enhanced sampling method to gain sufficient sampling in the high energy region with additional potential function using order parameters.\textsuperscript{151} Another commonly used enhanced sampling method, metadynamics, samples the high energy region by revisiting already sampled states, which also requires pre-defined order parameters.\textsuperscript{152} Transition Path Sampling (TPS) is another widely used sampling method, which samples the transitions between two stable regions, which should be close enough to be sampled in one transition.\textsuperscript{153} Adaptive sampling is also a widely used enhanced sampling method combining with Markov State Models (MSMs).\textsuperscript{154} Although it can produce the MSM with the statistical significance, it does not sample the long-time transition instead of the ensemble of short trajectories to model the equilibrium.

On the contrary, our group use the techniques of chain-of-states methods to enhance the samplings on the high energy region, named direct pathway dynamical sampling (DPDS) method.\textsuperscript{155} In this framework, the MD simulations are carried out within the chain-of-states framework\textsuperscript{156} to directly sample the transition pathways connecting the two end stable states. To control the smoothness of pathways and distance between states, two parameters ($K_{\text{rms}}$ and $K_{\text{ang}}$) are used. Comparing with other enhanced sampling methods, the DPDS method has the
advantage that no a priori reaction coordinates are needed. However, the limitation of DPDS method is requiring the extra theoretical studies to weight those sampling in order to achieve an unbiased sampling. Three examples are used to demonstrate the efficiency of DPDS methods in sampling the high-energy areas.

The following Figure 29 is the case study of the isomerization of the alanine dipeptide. The reaction coordinate to describe the reaction is the backbone dihedral angle ($\phi$ and $\psi$) as shown in Figure 29a. Two conformers exist for alanine dipeptide as $C^\text{eq}_7$ and $C^\text{ax}_7$, corresponding to two minimums. With 1μs unbiased MD simulation (shown in Figure 29c), the two energy basins are heavily sampled without reaching any high energy regions. Clearly, the high energy region between those two energy basins do not have sufficient samplings to calculate the free energy landscape.

![Figure 29](image)

Figure 29: Alanine dipeptide as test case: (A) alanine dipeptide structure and two dihedral angles ($\phi$ and $\psi$) as reaction coordinates; (B) initial minimum energy pathway for alanine dipeptide isomerization; (C) distribution of conventional molecular dynamics simulations of alanine dipeptide on its potential energy surface with reference to $\phi$ and $\psi$.

Therefore, using the chain-of-states framework as the basic idea, a pathway between two end stable states was established, and the direct samplings on those pathways were conducted. Based on the constraints between each replica on the pathway, some replica is forced to stay in the high energy region without reaching the low energy basins. Therefore, shown as Figure 30, the whole conformational space has a significant sampling increase comparing with the unbiased simulations in terms of the coverage.
Figure 30: DPDS simulations of alanine dipeptide with different \(K_{\text{rms}}\) and PES coverage of each simulation: (A) \(K_{\text{rms}} = 0.1\); (B) \(K_{\text{rms}} = 100\); (C) \(K_{\text{rms}} = 1000\); (D) plot of PES coverage along the simulation time.

Meanwhile, the \(K_{\text{rms}}\) parameter can have significant impact on the enhanced sampling efficiency. With a small \(K_{\text{rms}}\) parameter, the enhanced simulation results are close to the unbiased simulation as shown in Figure 30a. With the increasing \(K_{\text{rms}}\) values, the sampling conformational space deviates from the unbiased simulations. The transition regions are not well sampled using small \(K_{\text{rms}}\). With larger \(K_{\text{rms}}\) values of 100 and 1000, the samplings between basins including the transition from the larger basin to itself was much enhanced. The coverage for each parameter is computed in Figure 30d. The \(K_{\text{rms}}\) value of 1000 has the largest sampling efficiency comparing with other \(K_{\text{rms}}\) values. On the other hand, \(K_{\text{ang}}\) value can be used to control the smoothness of the pathway as shown in Figure 31. Comparing with \(K_{\text{rms}}\), \(K_{\text{ang}}\) value has the larger impact on the "shape" of sampling pathways. The most efficient \(K_{\text{ang}}\) values are 10, 100 and 1000 kcal/mol.
Figure 31: DPDS simulations of alanine dipeptide with different $K_{\text{ang}}$ (while $K_{\text{rms}} = 0.1 \text{kcal/(mol·Å2)}$) and PES coverage of each simulation: (A) $K_{\text{ang}} = 10 \text{kcal/mol}$; (B) $K_{\text{ang}} = 100$; (C) $K_{\text{ang}} = 1000$; (D) plot of PES coverage along the simulation time.

With $K_{\text{ang}}$ value as 10 shown in Figure 31a, the samplings of the transition path region in the attraction basin b significantly increase (blue circle). With $K_{\text{ang}}$ value as 100, the path connecting two basins is also revealed (red circle). Continuing increasing the $K_{\text{ang}}$ values, the two basins are directly connected with each other. Therefore, using $K_{\text{ang}}$ value as the controller, different valleys on the potential energy surface have been reached and sampled extensively.

Besides the above case study, the $\beta$-hairpin folding and $\beta$-alanine intramolecular condensation reaction have also been applied as the case studies of DPDS method. In all three cases, DPDS method provides great efficiency in enhancing the samplings comparing with the unbiased simulation. We also carried out the TPS sampling of alanine dipeptide, although TPS can also provide extensive samplings on the PES, the pathway information obtained from TPS is not as straightforward as the DPDS. Moreover, with the control of $K_{\text{ang}}$ value and $K_{\text{rms}}$ value, various pathways on the PES can be revealed. Because no a prior reaction coordinates are required, the setup and the analysis of DPDS are easier than other enhanced sampling methods.
9. Conclusion

Allostery, referred to as the “second secret of life”, has been thoroughly studied both experimentally and theoretically. However, the mechanism of allostery is still obscure for many proteins. Moreover, the idea that “allostery is an intrinsic property of protein” has attracted a lot of attentions from the community in the recent years. The understanding of allostery has been evolved from the conformational changes to the thermodynamic or kinetic differences during the last 20 years. Currently, the investigation of allostery can be summarized into three aspects including the thermodynamics view, the kinetics view and the structural point view. The thermodynamics view focuses on the enthalpy and entropy as the driving force for the conformational changes and function changes. The kinetics view focuses on the kinetics and the transition probability among different conformations. The structural point of view focuses on the communication inside of the protein in terms of pathways and communities.

In this thesis, different projects conducted in the last five years related to allostery have been summarized. My past work proposed several new methods in all three aspects to investigate the allostery. Based on the rigid body dynamics, Rigid Residue Scan (RRS) was used to investigate the thermodynamics of the protein upon the rigid body perturbation. The allosteric residues would respond differently from other residues upon the rigid body perturbation. Moreover, the machine learning approaches were introduced to investigate the difference between different allosteric state populations. Neural networks and decision trees models can provide the different distributions for unbound and bound state simulations. Although PDZ2 is a widely investigated
dynamics driven allosteric protein, these machine learning studies support the merit of population shift hypothesis. Further comparisons between NN and DT/RF models lead to the selection of random forest models on the future studies because the RF model can provide the feature importance with no bias on any features. The features importance can provide the specific information regarding the structures of proteins, which connect the structural point of view to the thermodynamics point of view.

Based on this idea, the Relative Entropy based Dynamic Allosteric Network (REDAN) was constructed to illustrate the protein distribution changes within the protein structural point of view. This method is the first allosteric network model that establishes the connection between distribution changes and protein structures. This network model can be regarded as the free energy prorogation network because such distribution essentially represents the free energy landscape of protein. The model protein system PDZ2 was investigated using REDAN, and the pathway and community analyses revealed the potential communication mechanism inside the PDZ2 protein.

Furthermore, as the most applied theoretical kinetic model for MD simulation, Markov State Models (MSMs) have also been applied to investigate the conformational changes of VVD. Comparing with above network analysis or RRS, the advantage of MSMs is the involvement of the kinetic and transition probability between different states. Using the MSMs and the combination of the machine learning models, important residues and potential mechanism were proposed.

The methodology improvement of MSMs led to the Directed Kinetic Transition Network (DKTN) model. The DKTN model, which is established based on the equilibrium distribution and the short-time state transition, is equivalent to a Continuous Time Markov Chain model
(CTMC). In comparison with the MSMs, the DKTN model can utilize both the equilibrium information and short-time transition information, and is capable of modeling the process from non-equilibrium to equilibrium state, which is not feasible using MSMs.

Besides the above projects, I also introduced t-SNE to the MD simulation for dimensionality reduction purpose and developed the Direct Pathway Dynamic Sampling (DPDS) method for enhancing sampling. Those methods have also been applied on various model system, and demonstrate the advantages in comparison with existing models.

Overall, my research provide several newly developed methods to investigate allostery from different aspects. By using RRS, RF, REDAN and DKTN models, the mechanisms for the functional changes, population shift, and the conformational changes of allosteric proteins can be investigated systematically. Those methods could shed more light on the allosteric mechanisms.
REFERENCE


46. 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.


86


PUBLICATIONS


APPENDIX (FIRST AUTHOR ONLY)


RESEARCH ARTICLE

Rigid Residue Scan Simulations Systematically Reveal Residue Entropic Roles in Protein Allostery

Robert Kalescky1,*, Hongyu Zhou1,*, Jin Liu1,2,*, Peng Tao1,*

1 Department of Chemistry, Center for Drug Discovery, Design, and Delivery (CD4), Center for Scientific Computation, Southern Methodist University, Dallas, Texas, United States of America, 2 Department of Pharmaceutical Sciences, University of North Texas System College of Pharmacy, University of North Texas Health Science Center, Fort Worth, Texas, United States of America

☯ These authors contributed equally to this work.
€ Current address: Center for Scientific Computation, Southern Methodist University, Dallas, Texas, United States of America
* jin.liu@unthsc.edu (JL); ptao@smu.edu (PT)

Abstract

Intra-protein information is transmitted over distances via allosteric processes. This ubiquitous protein process allows for protein function changes due to ligand binding events. Understanding protein allostery is essential to understanding protein functions. In this study, allostery in the second PDZ domain (PDZ2) in the human PTP1E protein is examined as model system to advance a recently developed rigid residue scan method combining with configurational entropy calculation and principal component analysis. The contributions from individual residues to whole-protein dynamics and allostery were systematically assessed via rigid body simulations of both unbound and ligand-bound states of the protein. The entropic contributions of individual residues to whole-protein dynamics were evaluated based on covariance-based correlation analysis of all simulations. The changes of overall protein entropy when individual residues being held rigid support that the rigidity/flexibility equilibrium in protein structure is governed by the La Châtelier’s principle of chemical equilibrium. Key residues of PDZ2 allostery were identified with good agreement with NMR studies of the same protein bound to the same peptide. On the other hand, the change of entropic contribution from each residue upon perturbation revealed intrinsic differences among all the residues. The quasi-harmonic and principal component analyses of simulations without rigid residue perturbation showed a coherent allosteric mode from unbound and bound states, respectively. The projection of simulations with rigid residue perturbation onto coherent allosteric modes demonstrated the intrinsic shifting of ensemble distributions supporting the population-shift theory of protein allostery. Overall, the study presented here provides a robust and systematic approach to estimate the contribution of individual residue internal motion to overall protein dynamics and allostery.
Author Summary

Allostery is a fundamental dynamics property of many proteins, and plays a critical role in protein functions. Despite extensive experimental and theoretical studies of protein allostereic mechanisms, the current understanding and predicting power of protein allostery are still limited. One of the main challenges in studying protein allostery is effectively narrowing down residues for further site-directed mutagenesis study. Our goal is to develop effective computational tools to systematically evaluate significance of individual residue in protein dynamics and allostery without any \textit{a priori} knowledge about protein allostereic mechanism. In this study, we significantly enhanced a simulation protocol developed in our lab, rigid residue scan (RRS), through combination of configurational entropy calculation, principal component analysis (PCA), and projection of ensembles onto coherent allostereic modes. Detailed analysis of the impact of removing individual residue internal motions on overall protein dynamics led to identification of key allostereic residues. Our prediction of key allostereic residues has good agreement with experimental studies of an allostereic protein as a model system, which displays allostery through binding events. Interestingly, the entropy calculations suggest that the La Châtelier’s principle in chemical equilibrium may also govern the rigidity/flexibility equilibrium in protein structure, which is related to protein allostery. Our study has demonstrated these methods to be very valuable tools to effectively identify initial key residues for proteins with crystallographic structures and limited information of their allostereic mechanisms.

Introduction

Allostery is the process by which signals are transmitted from distal ligand binding sites to functional sites in proteins. The concept of allostery originated from early attempts to explain the fact that the binding of oxygen molecules to hemoglobin deviates from the typical Michaelis-Menten kinetics model.[1–3] Following the term “allostereic” being coined and reviewed during early 60’s,[4, 5] two protein allostery theories were proposed and referred to as the Monod–Wyman–Changeux (MWC)[6] and Koshland–Nemethy–Filmer (KNF)[7] models. In these models, allostery theories were formed based on significant conformational changes of hemoglobin observed in crystallographic structures. In addition to hemoglobin, allostery with conformational change has been observed in other proteins such as aspartate transcarbamoylase,[8] insulin,[9] trypsin,[10] and caspases[11]. In these proteins, the binding signal is assumed to be transmitted through protein conformational change.

Multiple allostery theories have evolved based on experimental and theoretical studies.[12–27] The classical “induced fit” model[28–30] fits well to protein conformational changes upon ligand binding observed in hemoglobin.[31] However, a more recent “population shift” model of protein allostery[32–35] is strongly supported by sophisticated NMR experiments.[36–38] In this model, no conformational changes can be detected throughout the process in which proteins carry out their functions. Instead, allostery-triggering events alter the distribution of the protein ensemble among distinctive sub-states.

Many computational methods have been developed to delineate protein allostereic mechanisms in atomic detail and to facilitate development of allostery theories. Some methods are mainly based on protein tertiary structure comparison using topology or graph theory for analysis.[39–45] Some methods analyze energy-based residue-residue interactions to explore residue coupling.[46–50] Normal mode analysis (NMA)[51] is employed based on the elastic network model (ENM)[52, 53] or the Gaussian network model (GNM).[54] These models
provide coarse-grained protein structure descriptions, which reduce the computational cost to probe the protein’s vibrational modes. Modes with low frequency and large magnitude presumably correspond to allosteric mechanisms. Molecular dynamics (MD) simulation is the most widely used and direct means to simulate protein dynamics. Thus, it is frequently used with certain modifications to investigate protein allostery as a dynamical process. Ota and Agard proposed the MD-based anisotropic thermal diffusion (ATD) method to probe energy dissipation pathways in proteins. In this method, a single residue is heated in a protein at an extremely low temperature (approximately 10K) to probe energy dissipation pathways. Sharp and Skinner developed a pump-probe MD method that perturbs protein dynamics by exerting oscillating forces on target residues. Long-time MD simulations were carried out and subjected for further analysis to reveal protein allosteric effects in several other studies.

Deep understanding of protein allostery remains elusive despite the experimental and theoretical studies done thus far. More methodological development is needed to quantitatively evaluate the effect of individual amino acid residues on overall protein dynamics. Although mutagenesis studies can provide valuable information about the impact of changing specific residues on protein activity, systematically posing perturbation on individual residues provides an alternative way to probe the effect of the internal motions of specific residues on protein dynamics or to discover the function of individual residues without changing their chemical entity. Applying rigid constraints on selected degrees of freedom in protein structure has been implemented to probe protein allostery. Alternatively, we recently developed a simulation method, referred to as rigid residue scan (RRS), to systematically probe the impact of each individual residue on overall protein dynamics through rigid body MD simulations using an efficient integrator. In this study, the entropy calculation and principal component analysis are combined with the RRS method to evaluate the effects of internal motions from individual residues on overall protein dynamics as well as allostery upon ligand binding.

Results

Root-Mean-Square Deviation (RMSD) and Average Structures from Simulations

The all-atom RMSD for the unperturbed unbound and bound states (without rigid residue perturbation) of PDZ2 are shown in Fig 1, which indicates that both structures are stable throughout the simulations. The RMSD plots of all RRS simulations are listed in S1 Table. In general, the RRS simulations are stable throughout the simulations, with the majority of the simulations having average RMSD under 2 Å (Fig 2).

To assess the overall conformational change upon binding and rigid residue perturbation, averaged structures of PDZ2 were calculated for each simulation. Using the averaged structure of PDZ2 from unperturbed unbound simulation as reference structure, the all-atom RMSD of all other averaged structures were calculated and plotted in Fig 2. For most of the RRS simulations, both unbound and bound, the average structures have RMSD between 1.5 and 2.0 Å, with very few exceptions.

Changes of Protein Entropy upon Perturbation

Entropy contributions from PDZ2 are estimated using the method described in the Computational Methods section for all the unperturbed and rigid body perturbed simulations. The heat maps of cross-correlation matrices (based on which the entropy was calculated) are provided for all the simulations in S2 Table. Using the entropy of PDZ2 from unperturbed
simulation of unbound state as reference, the relative entropies of PDZ2 ($\Delta S$) from all the simulations are plotted in Fig 3 and listed in S3 Table. The $\Delta S$ are also sorted with ascending order and listed in S4 Table.

For both unbound and bound states, the entropy of PDZ2 significantly increases for most of the RRS simulations. The changes of unbound PDZ2 entropy in rigid residue simulations comparing to unperturbed simulations vary from $-0.100$ to $0.254$ kcal/(mol•K) with average as $0.058$ kcal/(mol•K) and unsigned average as $0.066$ kcal/(mol•K). For the bound state, the PDZ2 entropy differences in rigid residue simulations comparing to unperturbed simulation vary from $-0.065$ to $0.405$ kcal/(mol•K) with average as $0.060$ kcal/(mol•K) and unsigned average as $0.071$ kcal/(mol•K). Overall, in 80 unbound and 68 bound RRS simulations, the $\Delta S$ of PDZ2 is positive. This is counterintuitive, because treating a residue as a rigid body removes the internal degrees of freedom of that residue and should reduce the overall entropy. Furthermore, we calculated the PDZ2 entropy difference between two states ($\Delta\Delta S$) by subtracting the unbound state entropy from the bound state entropy with the same residue being held rigid.

Fig 1. RMSD for the unperturbed (no rigid residue perturbation) molecular dynamics simulations of both unbound and bound PDZ2. The RMSD is determined relative to the initial simulation structure of each simulation.

doi:10.1371/journal.pcbi.1004893.g001

Fig 2. All-atom RMSD of average structures of 190 RRS MD simulations.

doi:10.1371/journal.pcbi.1004893.g002
The absolute ΔΔS values range from 0.001 to 0.341 kcal/(mol•K) (S4 Table). For the two unperturbed simulations, this difference is 0.016 kcal/(mol•K) with a higher PDZ2 entropy from the bound state (Residue 0 in S3 and S4 Tables). For the 11 residues being held rigid, the absolute ΔΔS is smaller than 0.016 kcal/(mol•K) (S4 Table). Seven among these 11 residues, D15, T28, V40, T81, R31, L78, L18, were reported as important allosteric residues from an NMR study of PDZ2 bound to the RA-GEF-2 peptide (Table 1).[72]

The error bar of PDZ2 entropy calculations was estimated for both unbound and bound states of PDZ2 in unperturbed simulations and seven RRS simulations (Table 2). For these simulations, the PDZ2 entropy was calculated based on seven sets of 30 ns (total of 210 ns).
trajectories. The standard deviation (σ) and 85% Confidence Interval (CI) of each state is rather small, indicating the convergence of simulation within 30 ns of trajectories. The standard deviation (σ) and 85% CI of ΔΔS for unperturbed states are 0.037 and 0.039 kcal/(mol•K), respectively. It is noticeable that the errors of entropy calculations, although small when compared to the total entropy, are comparable to the differences between simulations. For the unbound state, total of 28 RRS simulations have ΔS smaller than the 85% CI of unperturbed unbound state. For the bound state, total of 53 RRS simulations have ΔS smaller than the corresponding 85% CI (S3 Table). Among seven identified residues, there are three residues (D15, T28, V40) with ΔS values higher than the 85% CI of corresponding unperturbed states. Although the canceling of the error could improve the reliability of the analysis, these comparisons indicate that the uncertainty of calculated configurational entropies requires further improvement, for example by including anharmonicity and higher order correlations, to increase the reliability of the calculations.

Velocity autocorrelation analysis was carried out for the unperturbed simulations and the seven RRS simulations listed in Table 2 to estimate the relaxation time in these simulations. Only one trajectory of each simulation was subjected to the analysis. All the selected simulations display a relaxation time around 20 ps (S1 Fig), showing that RRS simulations have similar relaxation time to the unperturbed simulations.

### Individual Residue Response upon Perturbation

From each simulation, the entropy contribution of each residue to total protein entropy was evaluated. Such individual residue entropy contributions are plotted as heat maps for the RRS simulations of both unbound and bound states of PDZ2 (Fig 5). To make plots clear, the contribution from each individual residue in unperturbed simulations was used as reference in unbound and bound states, respectively. The response of each individual residue varies significantly. The most prominent features in both heat maps are the blue diagonal lines, reflecting the fact that the entropy contribution from each residue diminishes when that residue is held

### Table 2. Estimation of error bar for PDZ2 entropy calculations (kcal/(mol•K))

<table>
<thead>
<tr>
<th>Residue</th>
<th>Average S (σ, 85% Confidence Interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unbound</td>
</tr>
<tr>
<td>none</td>
<td>5.202 (0.046, 0.029)</td>
</tr>
<tr>
<td>15</td>
<td>5.213 (0.086, 0.054)</td>
</tr>
<tr>
<td>18</td>
<td>5.154 (0.045, 0.028)</td>
</tr>
<tr>
<td>22</td>
<td>5.208 (0.058, 0.036)</td>
</tr>
<tr>
<td>28</td>
<td>5.180 (0.038, 0.024)</td>
</tr>
<tr>
<td>40</td>
<td>5.269 (0.064, 0.040)</td>
</tr>
<tr>
<td>61</td>
<td>5.206 (0.084, 0.052)</td>
</tr>
<tr>
<td>81</td>
<td>5.208 (0.039, 0.024)</td>
</tr>
</tbody>
</table>

* For each state, seven sets of 30 ns trajectories were used.

**Key residues for PDZ2 allostery upon RA-GEF2 peptide binding from NMR study [72].**


* Residues in bold type are recognized from the present study.

Table 2. Estimation of error bar for PDZ2 entropy calculations (kcal/(mol•K)).

<table>
<thead>
<tr>
<th>Residue</th>
<th>Average S (σ, 85% Confidence Interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unbound</td>
</tr>
<tr>
<td>none</td>
<td>5.202 (0.046, 0.029)</td>
</tr>
<tr>
<td>15</td>
<td>5.213 (0.086, 0.054)</td>
</tr>
<tr>
<td>18</td>
<td>5.154 (0.045, 0.028)</td>
</tr>
<tr>
<td>22</td>
<td>5.208 (0.058, 0.036)</td>
</tr>
<tr>
<td>28</td>
<td>5.180 (0.038, 0.024)</td>
</tr>
<tr>
<td>40</td>
<td>5.269 (0.064, 0.040)</td>
</tr>
<tr>
<td>61</td>
<td>5.206 (0.084, 0.052)</td>
</tr>
<tr>
<td>81</td>
<td>5.208 (0.039, 0.024)</td>
</tr>
</tbody>
</table>

* For each state, seven sets of 30 ns trajectories were used.
as a rigid body during the simulation. The most recognizable features besides blue diagonal lines are the horizontal lines in both heat maps, either in red or blue. These horizontal red or blue lines indicate that response from some residues to rigid body perturbation is consistent regardless which residue being held rigid in the perturbed simulations. To further illustrate this feature, the average entropic response from each residue in all RRS simulations was calculated and plotted in Fig 6 for the unbound and bound states, respectively. The average entropic responses are also listed in S5 Table and with descending order in S6 Table. The individual residue entropies were also normalized using the number of atoms in each residue following a previous study.[73] The normalized individual residue entropies are illustrated in S2 Fig and S3 Fig. The patterns described above remain the same with the normalized entropies, showing that the differences of residue responses are inherent to each residue and not scaled with residue size.

The average entropic responses range between 0.0352 and $-0.0167 \ \text{kcal/(mol} \cdot \text{K})$ for the unbound state, and between 0.0183 and $-0.0161 \ \text{kcal/(mol} \cdot \text{K})$ for the bound state. Among the top ten residues with largest average entropic responses in the unbound state, seven residues, R31, T28, V61, L18, V26, V22, N27, were among the 21 important allosteric residues reported in an NMR study of PDZ2 (Table 1).[72] However, for the bound state, only four residues, V22, V85, V61, V26, among top ten residues were reported as important allosteric residues from the same NMR study of PDZ2 (Table 1).[72] Noticeably, three residues, V22, V61, and V26, are among the top ten residues of both unbound and bound states.

### Quasi-harmonic Analysis and Principal Component Analysis (PCA)

Quasi-harmonic analysis was carried out for both unperturbed and RRS simulations. The distributions of density of states from quasi-harmonic analysis of unperturbed simulations are plotted for both unbound and bound states in Fig 7. Obviously, the binding with the peptide...
does not significantly affect the distribution of density of states. Similarly, the rigid residue perturbation does not significantly affect the distribution of density of states either (S4 Fig). We further carried out the PCA to evaluate the contribution of each quasi-harmonic mode to overall dynamics, and plotted accumulative contribution of these modes for both unperturbed unbound and bound states in Fig 8. Low frequency modes significantly contribute to overall protein dynamics. For the unbound state, total of 83 modes with frequency under 18.7 cm\(^{-1}\) contribute 90% to the overall dynamics. For the bound state, total of 52 modes with frequency under 13.4 cm\(^{-1}\) contribute 90% to the overall dynamics. In both cases, translational and rotational modes are excluded. Because protein allostery is highly dynamical process that couples the dynamics of distal parts of the protein, it is logical to assume that mainly low frequency modes, which involve overall dynamics of proteins, play important roles in protein allostery.
To identify significant low-frequency modes, PCA was carried out for the seven 30 ns trajectories as well as total 210 ns trajectories for the unperturbed unbound and bound states, respectively. For the unperturbed unbound state, the dot products were calculated between the five lowest frequency quasi-harmonic modes (PC1 to PC5) of each 30 ns trajectory with the PC1 to PC5 modes from the whole 210 ns trajectory (S7 Table). The same calculations were also carried out for the unperturbed bound state (S7 Table). The unsigned averaged dot product of each mode is listed in Table 3. Among five modes, only PC1 modes (with the lowest frequency quasi-harmonic mode) in both unperturbed unbound and bound states have significant overlap between each trajectory and overall trajectory. The overlaps for PC2 through PC5 are significantly less than PC1 (Table 3), indicating that these modes and all other modes with higher frequencies do not have physical significance.

To further evaluate significance of PC1 modes in the unperturbed unbound state, the dot products among PC1 modes from seven 30 ns trajectories were calculated to produce a 7×7 matrix (S8 Table). The absolute values of off-diagonal matrix elements range from 0.694 to 0.960 with unsigned average values (standard deviation) as 0.842 (0.078). The similar analysis of the unperturbed bound state results in a matrix with absolute values of off-diagonal matrix elements range from 0.653 to 0.938 with average values (standard deviation) as 0.781 (0.089) (S8 Table). It should be noted that PC1 modes from 210 ns trajectories of unperturbed unbound and bound states do not overlap significantly with each other (with magnitude of dot product as −0.214). Therefore, two PC1 modes from two states could serve as coherent allosteric modes revealing effect of PDZ2 upon ligand binding. The PC1 modes calculated using 210 ns trajectories are listed in Table 3.

Table 3. Unsigned average dot products of five lowest frequency quasi-harmonic modes (PC1 through 5) between each 30 ns trajectory and whole 210 ns trajectory.

<table>
<thead>
<tr>
<th>Modes</th>
<th>Unbound</th>
<th>Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC1</td>
<td>0.929 (0.044)</td>
<td>0.899 (0.051)</td>
</tr>
<tr>
<td>PC2</td>
<td>0.422 (0.171)</td>
<td>0.459 (0.227)</td>
</tr>
<tr>
<td>PC3</td>
<td>0.193 (0.116)</td>
<td>0.204 (0.150)</td>
</tr>
<tr>
<td>PC4</td>
<td>0.206 (0.128)</td>
<td>0.176 (0.103)</td>
</tr>
<tr>
<td>PC5</td>
<td>0.287 (0.177)</td>
<td>0.254 (0.175)</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pcbi.1004893.t003
ns trajectories of the unperturbed unbound and bound states are used for further analysis in this study.

Ensemble Distributions

The simulations of unperturbed states are projected onto a 2D surface using two PC1 vectors from unperturbed states to illustrate the distribution of ensemble representing each state (Fig 9). Despite the close similarity between PDZ2 structures from unbound and bound states, the clear separation between two states on this 2D surface provides more insight into the allosteric difference between these states. The separation of two distributions on the 2D surface was represented by an average distance (0.734) between two attraction basins (Fig 9). All RRS simulations are also projected onto this 2D surface using the same two PC1 vectors to probe the impact of rigid residue perturbation on the distribution of ensemble on the same surface (S9 Table). For all the RRS simulations, the separation between unbound and bound states distributions resembles the unperturbed states, suggesting that the allosteric effect triggered by ligand binding event is robust upon rigid residue perturbation. However, the average distance between two distributions varies significantly among RRS simulations (Fig 10, S10 Table and sorted values in S11 Table), revealing the different contribution from each residue in the protein allostery. It is notable that the distribution distances of RRS simulations when residues R31, V40, or L78—all identified as key allosteric residues in the NMR study [72]—being held rigid are significantly shorter than the one of unperturbed states.

Discussion

In the present study, we applied entropy analysis, quasi-harmonic analysis, and PCA on the RRS simulations of PDZ2 domain from PSD-95 protein to investigate the relationship between
each individual residue and overall protein dynamics aiming to decipher protein allosteric mechanism within this framework.

Residue Internal Dynamics and Overall Protein Dynamics

Individual amino acid residues, which are basic building blocks for protein structures, and therefore serve as the main target for many residue based protein allostery analysis methods, [47, 74–76] in which residue based interaction energy is the target for analysis. However, because protein allostery is mainly considered a dynamical process, it should be informative to investigate the internal dynamics of each individual residue and their impact on overall protein dynamics. Presumably, the internal degrees of freedom or dynamics of key allosteric residues should play unique roles in allostery with specific impact on overall protein dynamics. The RRS simulations combining with entropy analysis make it feasible to systematically evaluate the contribution from individual residue internal degrees of freedom to overall protein dynamics. Comparison between the unbound and bound states connects such contribution with protein allostery upon binding.

Rigid body constraint, which effectively removes the internal degrees of freedom in residue, should theoretically reduce the disorder of the protein as well as the protein entropy. On the contrary, rigid residue constraints lead to the increase of PDZ2 entropies in most RRS simulations. This counterintuitive observation indicates that the internal dynamics of each individual residue in a well-folded protein cooperatively contribute to the overall protein dynamics. In a recent simulation study of protein structures,[77] it was also reported that rigidifying some of protein degrees of freedom often cause more flexibility in other parts and lead to increasing protein entropy. The basic La Châtelier’s principle in chemical equilibrium was referred to govern the rigidity/flexibility equilibrium in protein structure.[77] Seemingly, our observation of increasing protein entropies in rigid residue simulations also agrees with the La Châtelier’s principle.

Without rigid body constraints, the binding with peptide leads to slight increase of PDZ2 entropy (0.016 kcal/(mol•K)). This is also in agreement with that the RMSD of PDZ2 in bound state is slightly higher than the one of unbound state (Fig 1). For only 11 residues among 94 PDZ2 residues, the PDZ2 entropy difference between unbound and bound states from RRS
simulations is smaller than 0.016 kcal/(mol•K). Seven among these 11 residues D15, T28, V40, T81, R31, L78, L18 (Fig 11), were recognized as important for PDZ2 allostery upon binding by the NMR study.[72] All seven residues displayed significant dynamical parameter change upon binding. Although the direct relationship between the present study and NMR study of PDZ2 is not obvious, the overlap between two studies are unlikely to be random coincident. It should not be overlooked that the uncertainty of calculated configurational entropies undermines the reliability of predictions based on these calculations. Nevertheless, the current development is only a small step towards deeper understanding of protein allostery in terms of configurational entropy change. Improvement of configurational entropy calculations by including anharmonicity and higher order correlations will be applied to increase the reliability of the calculations.

The remaining residues were also identified by various computational studies as key allosteric residues. R79 was identified as one of “Hot Residues” for allostery in a study using ENM to probe PDZ2 allostery[58] as well as one of “nodes” to form an allostery communication network in another study using protein structure network model.[42] Both residues N14 and E90 were identified as part of an interacting cluster localized at the ligand binding pocket.[47] This strongly suggests that the RRS simulations could reveal the significance of internal dynamics of some key allosteric residues with regard to overall protein dynamics.

L18, as one of identified residues, when being held rigid, leads to the entropy increase of 0.045 kcal/(mol•K) for unbound state and 0.031 kcal/(mol•K) for bound state, which may be resulted from La Châtelier’s principle of protein rigidity/flexibility equilibrium.[77] However, this residue makes hydrophobic contact with the C-terminal valine of the binding peptide, which may counteract some of the entropy increasing effect. K13, another key residue for PDZ domain for interaction with binding peptide, forms multiple hydrogen bonds with

---

**Fig 11. Key residues recognized based on protein entropic response to rigid body perturbation.**

doi:10.1371/journal.pcbi.1004893.g011
Gly82, the backbone and side chain of Thr81. When K13 being held rigid, the entropy of PDZ2 increases 0.013 kcal/(mol•K) for unbound state and decreases 0.052 kcal/(mol•K) for bound state. These observations indicate that in RRS, strong interactions between residues being held rigid and ligand may help to counteract the general trend of increasing entropy in rigid body simulations.

Response of Individual Residue to Perturbation

Another informative analysis presented in this study is the response from each individual residue to perturbations on proteins. By its definition, the mechanism of protein allostery is to be elucidated at overall protein structure level. However, this should not prevent any attempt to evaluate general response of individual residues to external or internal perturbations. Using the estimate of entropic contribution from individual residue in each simulation, the intrinsic difference of response to perturbation from each residue is revealed. Residue R31 is a clear case in the unbound state. Regardless which residue being held rigid, the entropic contribution from R31 is higher than the value in the unperturbed unbound state. Coincidentally, R31 has also been identified based on overall protein entropy change and NMR study\[72\] discussed in the previous section. On the contrary, this positive entropic response from R31 is inhibited upon binding with the peptide (Fig 5 and S3 and S4 Tables). Another six residues (T28, V61, L18, V26, V22, N27) also showed overwhelming positive response similar to R31, and were also recognized in the NMR study of PDZ2 allostery (Table 1).

The individual residue response pattern is significantly different in the bound state. Residues with consistent entropic response to rigid residue perturbations are very different from those in the unbound state. The observations that some residues display consistent response suggest that these residues are more sensitive to perturbations than other residues. The difference between unbound and bound states shows that the binding event inherently changes the sensitivity of each residue upon the perturbations, reflecting the very nature of protein allostery influenced by binding and perturbation events.

Ensemble Distributions upon Coherent Allosteric Modes

Given the sheer number of modes generated from the quasi-harmonic analysis, one definitely suffers the risk of studying trivial and random patterns when focusing on only a few modes related to overall protein dynamics. However, the PCA analyses using multiple trajectories of unperturbed states clearly validate the physical significance of two PC1 modes from two states. The fact that two PC1 modes are virtually orthogonal to each other signifies the relevance between these two modes and protein allostery upon binding. The projection of RRS simulations onto these coherent allosteric modes support the entropy driven theory of protein allostery,[78, 79] which explain the protein allostery phenomenon as shifting of ensemble distribution upon perturbation instead of significant conformational change observed in either crystallographic\[31\] or NMR studies.[80] The various distances between two states projecting on the coherent allosteric modes demonstrate the shift of protein ensemble distribution upon rigid residue perturbation, revealing detailed information about individual residue with regard to overall allostery.

Computational Cost of RRS Simulations

It is undeniable that the computational cost of the proposed method is exceptionally high compared to many other methods to elucidate protein allostery mechanisms. However, the current goal of our research is to develop an unbiased protocol to assess potential individual residue’s contribution towards overall protein dynamics and consequently allostery, with little or no a priori information about protein allosteric mechanisms. In the experiment, the most feasible way to
probe the contribution of each individual residue to protein function is mutagenesis study. However, some knowledge of importance of residues is necessary for the mutagenesis study. Otherwise, all positions should be considered. In addition, it is somewhat arbitrary to choose what amino acids to which the wild type residues should be mutated. From the mutagenesis study, the perturbation added to protein contains two parts: removing the wild type residue and adding the mutated residue. These two parts are distinct but inseparable in the mutagenesis study. Without any a priori knowledge about relationship between protein sequence and allostery, one would mutate every residue to all other 19 natural amino acids to obtain the most comprehensive and unbiased evaluation of relationship between each residue and protein allostery. This is actually what has been done in an experimental study of allostery of another PDZ domain, in which total of 1577 mutants were generated for 83 out of 115 residues.[81] However the equivalent strategy could not be applied routinely to other proteins due to its obvious high cost. In addition to the exceedingly high cost of doing all possible mutations, the effects of removing the wild type residue and adding the mutated residue are still inseparable. One of the main advantages from the strategy presented in this study is that only the internal motion of each individual residue is removed, while the chemical content of the wild type residue is intact. This strategy provides a practical mean to investigate the intrinsic dynamical effect of each individual residue to overall protein dynamics. At this early stage, the RRS simulations were carried out for all residues for the sake of completion. Further improvement of the method is ongoing to significantly reduce the computational cost while keeping the confidence level of the results.

Concluding Remarks

In this study, we further developed a recently proposed RRS method through combination of configurational entropy calculation and PCA to systematically evaluate the contribution of internal degrees of freedom of individual residue to overall protein dynamics and potential allostery upon ligand binding. Through the changes of the entropy from whole protein upon rigid residue perturbation, key residues were recognized as those when being held as rigid bodies, the protein entropy difference between unbound and bound states is smaller than the entropy difference from unperturbed simulations. These key residues have good agreement with a previous NMR study of the same protein bound to the same peptide.[72] Entropic response from individual residue upon perturbations was also evaluated. In the unbound state simulations, residues generally displaying increased entropic contribution upon rigid residue perturbation are in good agreement with the same NMR study.[72] The different patterns of individual residue response in the unbound and bound states suggest that the binding event inherently changes the sensitivity of each residue upon the perturbations. PCA of unperturbed states of PDZ2 revealed two quasi-harmonic coherent allosteric modes, which are robust upon analysis of multiple trajectories of each state. The projection of RRS simulations onto coherent allosteric modes reveals the intrinsic shifting of ensemble distributions upon rigid residue perturbations, and supports the population-shift point of view about protein allostery. Overall, the combination of entropy calculation and PCA with the RRS method provides a systematic approach to estimate the individual residue contribution to protein dynamics as well as allostery. Further development is actively under development to reduce the computational cost with deeper understanding of the protein allostery.

Methods

MD Simulations

The assessment of the role of individual residues in overall protein dynamics is carried out using rigid residue scan, a systematic simulation method developed in our group.[68] In the
RRS method, rigid body constraints are applied to each residue in the target protein in separate simulations (referred to as perturbed simulations). Thus, there are as many rigid body MD simulations (perturbed simulations) as there are residues comprising the target protein.

The second PDZ domain (PDZ2) from the human tyrosine phosphatase 1E (hPTP1E) is used as a test protein in this study to further develop the RRS method. The allosteric mechanisms of PDZ2 have been the subject of a number of studies with various residues identified as key allosteric residues.[42, 47, 61] Initial crystal structures for the unbound and bound states with RA-GEF2 peptide (EQVSAV) were obtained from the Protein Data Bank (PDB) with IDs 3LNX and 3LNY, respectively.[82] PDZ2 structures in both 3LNX and 3LNY contain 94 residues. For the residues with multiple copies in the PDB files, the first coordinate set was used to prepare the simulation systems.

The structures from the PDB were processed with hydrogen atoms added and solvated in water (TIP3P)[83] with charge balancing ions of sodium and chlorine added. Additional ions were included to adjust the ionic strength in simulation cells to about 0.02 M. The system was then subjected to energy minimization with 200 steps of steepest descent and 9491 steps of adopted basis Newton-Raphson minimization, which yielded a total gradient of less than 0.001 kcal/(mol•Å). This was followed by an equilibration step that raised the temperature of the system from 100 K to 300 K over 12 picoseconds (ps). Then the systems were equilibrated via 10 nanosecond (ns) isothermal-isobaric ensemble (NPT) MD simulations at 300 K and 1 atm.

The frame from the simulation trajectory with dimensions closest to the average dimension for the entire trajectory was selected. This set of coordinates and its corresponding velocities were used as the initial conditions for 34 ns canonical ensemble (NVT) Langevin MD simulations also at 300 K. The first 4 ns of each NVT simulation was treated as equilibrium, and therefore not included in the reported analysis. The NVT simulations consisted of normal MD simulations without rigid residue constraint for the unbound and bound PDZ2 (referred as unperturbed simulations) and the rigid residue scan over all 94 residues in PDZ2. There are total of 190 simulations, including 188 rigid residue simulations and two unperturbed simulations of unbound and bound states of PDZ2. Considering 30 ns of each NVT simulation for analysis, this work comprises 5,700 ns of simulation time. A 2 femtosecond (fs) simulation time step was used in all simulations. To estimate the error bar of the entropy calculations and validate coherent allosteric modes, additional 180 ns simulations were carried out for the unperturbed states and rigid residue simulations corresponding to seven residues (15, 18, 22, 28, 40, 61, and 81). Each set of 180 ns simulations were carried out with three independent 60 ns trajectories for better sampling and shorter computing times. All simulations used cubic periodic boundary conditions, and electrostatic interactions were modeled using the particle mesh Ewald method.[84] All simulations were carried out using CHARMM version 38b1 and version 27 of the CHARMM force field.[85]

Analysis of MD Trajectories

RMSD. The RMSD was used to measure the variability of a set of Cartesian coordinates over the course of a MD trajectory relative to a reference set. Specifically, the coordinates of each atom comprising the protein or a subset are compared with that of a reference structure. The RMSD for a given simulation is defined as

\[ R = \sqrt{\frac{\sum_{i=1}^{N} (r_i^0 - U r_i)^2}{N}}, \]

where \( N \) is the number of atoms, \( r_i^0 \) is the Cartesian coordinate vector for atom \( i \), and \( U \) is the best-fit alignment transformation matrix between a given structure and its reference structure.
Cross-correlation (normalized covariance) matrix. The cross-correlation matrix is a measure of the correlated movement of a set of atoms. Each matrix element is defined as

\[ C_{ij} = \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\sqrt{\langle (r_i^2 - \langle r_i \rangle^2 \rangle \langle (r_j^2 - \langle r_j \rangle^2 \rangle \rangle}} \]

where \( C_{ij} \) is the measure of correlated movement between atoms \( i \) and \( j \), \( c_{ii}, c_{jj}, \) and \( c_{ij} \) are the covariance matrix elements, and \( r_i \) and \( r_j \) are Cartesian coordinate vectors from the least-square fitted structures, therefore with translation and rotation 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 \( j \). Correlation is defined as related movement along the line between two points. Correlated movement along orthogonal paths yields a cross-correlation matrix element of zero. [86] After discarding the first 4 ns NVT simulation as equilibrium for each simulation, total of 15,000 frames were extracted from the remaining 30ns simulation with 2ps interval, and processed to generate cross-correlation matrix.

Entropy analysis. All-atom quasi-harmonic analysis was employed to analyze MD trajectories to probe protein dynamics using vibrational normal modes on an effective quasi-harmonic potential.[87] The element of force constant matrix \( F \) on the effective quasi-harmonic potential for normal modes calculation is given by [70]

\[ F_{ij} = k_B T \frac{1}{[C^{-1}]_{ij}} \]

where \( k_B \) is the Boltzmann constant, \( T \) is the temperature, and \( [C^{-1}] \) is the inverse of covariance matrix \( C \). Therefore, the normal modes and corresponding frequency \( \omega \) of the molecule on the effective quasi-harmonic potential can be calculated through the solution of secular equation

\[ \det(F - \omega^2 M) = 0, \]

where \( M \) is the mass matrix of protein. The configurational entropy of protein, \( S_{\text{config}} \) could be estimated from the \( 3n-6 \) nonzero quasi-harmonic frequencies through [71]

\[ S_{\text{config}} = k_B \sum_{i=1}^{3n-6} \frac{\hbar \omega_i / kT}{e^{\hbar \omega_i / kT} - 1} - \ln(1 - e^{-\hbar \omega / kT}) \]

It should be noted that the \( S_{\text{config}} \) is an approximation of the target value using the exact equation within harmonic limit. The all-atom covariance matrices were employed for entropy calculations. Because the protein structures from each simulation are superimposed to the first frame of the simulation before the calculation of covariance matrix, the overall translation and rotation motion was projected out for the entropy calculation.

To estimate entropy contribution from each residue, all-atom covariance matrix elements corresponding to correlation between atoms within the same residue (including side chain and backbone) are selected to form a sub covariance matrix, which is processed in the same way described above.[85, 88, 89] By constructing a sub covariance matrix for each residue from all-atom covariance matrix of whole protein, no alignment of the different conformations of the subsystem was performed. It should also be noted that the entropy corresponding to the correlation between the target residue and the rest of the system is not included in the individual residue entropy.

Principal component analysis (PCA). For each simulation, PCA was performed by projecting each of the extracted 15,000 frames from 30 ns trajectory onto the normal modes generated by quasi-harmonic analysis. Both translation and rotation components were projected out for each frame. The residues being held rigid were included in the above analyses just as any
other residues. All the analyses described in this section were carried out using CHARMM version 38b1.[85]

**Supporting Information**

**S1 Fig.** Autocorrelation functions of unperturbed and seven rigid residue scan simulations. Relaxation time around 20 ps was displayed in all simulations. (TIF)

**S2 Fig.** Heat maps of normalized individual residue entropy contribution under rigid residue perturbation for unbound (left) and bound (right) states. The entropy contribution from each residue in unperturbed simulations (with index as 0 in both plots) is set as reference. (TIF)

**S3 Fig.** Average of normalized entropic response from each residue in all rigid residue scan simulations. (TIF)

**S4 Fig.** Distribution of density of states for all rigid residue scan simulations. (TIF)

**S1 Table.** RMSD plots of PDZ2 from rigid residue scan for both unbound and bound states. (PDF)

**S2 Table.** Heat maps, histograms of Cα carbons cross-correlation matrices for all residues in PDZ2 from unperturbed and rigid residue scan simulations. (PDF)

**S3 Table.** Relative entropies (ΔS) and differences (ΔΔS) of PDZ2 between unbound and bound states. (PDF)

**S4 Table.** Relative entropies (ΔS) and differences (ΔΔS) of PDZ2 between unbound and bound states sorted with ascending order. (PDF)

**S5 Table.** Average entropic response of individual residues upon rigid body perturbations. (PDF)

**S6 Table.** Average entropic response of individual residues upon rigid body perturbations sorted with descending order. (PDF)

**S7 Table.** Dot products of five lowest frequency quasi-harmonic modes (PC1-PC5) from seven sets of 30 ns trajectories with the PC1 to PC5 from whole 210 ns trajectory. (PDF)

**S8 Table.** Dot products among PC1 modes from seven 30 ns trajectories. (PDF)

**S9 Table.** Projections of simulations onto 2D-surface using two PC1 modes from unperturbed unbound and bound states. (PDF)
S10 Table. Average distance between distributions of unbound and bound states projected onto 2D-surface using two PC1 modes.

(PDF)

S11 Table. Average distance between distributions of unbound and bound states projected onto 2D-surface using two PC1 modes sorted with ascending order.

(PDF)

Acknowledgments

The authors acknowledge Southern Methodist University’s Center for Scientific Computation and Texas Advanced Computing Center (TACC) at the University of Texas at Austin for providing high performance computing resources.

Author Contributions

Conceived and designed the experiments: JL PT. Performed the experiments: RK HZ. Analyzed the data: RK HZ JL PT. Wrote the paper: RK HZ JL PT.

References


41. Raimondi F, Felline A, Seeber M, Mariani S, Fanelli F. A Mixed Protein Structure Network and Elastic Network Model Approach to Predict the Structural Communication in Biomolecular Systems: The PDZ2...


57. Gu J, Bourne PE. Identifying allosteric fluctuation transitions between different protein conformational states as applied to Cdc15 Dependent Kinase 2. BMC Bioinf. 2007; 8:45.


Revealing Hidden Conformational Space of LOV Protein VIVID Through Rigid Residue Scan Simulations

Hongyu Zhou, Brian D. Zoltowski & Peng Tao

VIVID (VVD) protein is a Light-Oxygen-Voltage (LOV) domain in circadian clock system. Upon blue light activation, a covalent bond is formed between VVD residue Cys108 and its cofactor flavin adenine dinucleotide (FAD), and prompts VVD switching from Dark state to Light state with significant conformational deviation. However, the mechanism of this local environment initiated global protein conformational change remains elusive. We employed a recently developed computational approach, rigid residue scan (RRS), to systematically probe the impact of the internal degrees of freedom in each amino acid residue of VVD on its overall dynamics by applying rigid body constraint on each residue in molecular dynamics simulations. Key residues were identified with distinctive impacts on Dark and Light states, respectively. All the simulations display wide range of distribution on a two-dimensional (2D) plot upon structural root-mean-square deviations (RMSD) from either Dark or Light state. Clustering analysis of the 2D RMSD distribution leads to 15 representative structures with drastically different conformation of N-terminus, which is also a key difference between Dark and Light states of VVD. Further principle component analyses (PCA) of RRS simulations agree with the observation of distinctive impact from individual residues on Dark and Light states.

Light is a ubiquitous environmental signal of metabolism regulation for the majority of lives on earth. Light, Oxygen, Voltage (LOV) photoreceptor domains, first designated in 1997 are small, commutable proteins, and couple blue-light triggered control mechanism in response to the light stimulation\(^1\). The LOV domains are involved with control of phototropism\(^1,3\), chloroplast relocation\(^4\), stomatal opening\(^5\), rapid inhibition of stem growth\(^6\), and gametogenesis\(^7\), for higher plants, and circadian temporal regulation in bacteria and fungi. LOV domains are present in many multi-domain proteins, including DNA binding domains (i.e., leucine zipper, bHLH, and zinc finger)\(^8\), STAS domains\(^9\), and kinases\(^10–12\), and form a subset of Per, ARNT, Sim (PAS) superfamily, a sensor module found in all three kingdoms of life\(^13\). All LOV domains contain common flavin chromophore, which could covalently bond to an adjacent cysteine residue upon blue-light activation. Due to their ubiquitous and modular nature in sensory proteins, LOV domains have been exploited widely for the design of optogenetic tools\(^14\).

VIVID (VVD) protein is a LOV domain that regulates filamentous fungus Neurospora crassa circadian system. In Neurospora, VVD and another LOV domain, White Collar-1 (WC-1) are involved with light responses regulation. WC-1 forms a complex (referred as WCC) with nonphotosensitive WC-2 to activate transcription of the clock oscillator protein Frequency (FRQ). The main function of VVD is forming an autoregulatory negative feedback loop that connects the molecular output of circadian clock back to the input of the clock, which is also referred as “gating”\(^15\). Gating regulation is necessary for the appropriate response to daily light intensity fluctuation\(^16–18\). Photo-activated VVD reduces activation of the WCC to tune the Neurospora blue-light response. Flavin adenine dinucleotide (FAD) is required by both VVD and WC-1 for their light sensing functions. Some other LOV proteins, for example blue-light receptors in plants, phototropins, employ flavin mononucleotide (FMN) as chromophore for their light sensing function\(^17,19–21\).
The mechanisms that LOV domains respond to light activation and send signals to other parts of organisms are of great general interest. The common feature of LOV domains photo activation is that upon blue-light activation, the chromophore, either FAD or FMN, forms covalent bond with a cysteine residue of LOV domains, and induces a large conformational change within the LOV domain, thereby regulating activity of accessory signaling domains or protein:protein interactions. The states before and after photo activations could be referred as Dark and Light states, respectively. Using nuclear magnetic resonance (NMR) spectroscopy, Harper et al. showed that a 20 residues long helix, designated as Jα, undergoes conformational change upon photoactivation of Avena sativa Phototropin1 LOV2 (AsPhot1 LOV2) domain\(^{22,23}\). Similar conformational change upon photoactivation was directly observed through crystallographic characterization of VVD Dark and Light states. Zoltowski et al. showed that photoinduced formation of a covalent bond between Cys108 sulfur and C4a position of FAD drives an N-terminal cap conformational change critical for VVD function (Fig. 1)\(^{24}\).

Interestingly, the VVD N-terminal cap locates at a position similar to the Jα helix in LOV2 domain, indicating possible conservation of an allosteric pathway. Through time-resolved small-angle X-ray scattering analysis, Lamb et al. revealed another VVD Dark state structure with extended N-terminal cap in addition to the previously identified compact Dark state structure\(^{25}\). In the crystallographic structure of VVD Light state, the N-terminal cap is driven away from the protein core into a more extended structure, which leads to VVD dimerization\(^{26}\). A conserved residue Cys71, critical for VVD conformational change upon photo activation, is further away from highly conserved Cys108 (S-S distance as 14 Å), while a much closer Cys76 towards Cys108 (S-S distance as 6 Å) has no effect on VVD function with mutation Cys76Ser displaying same activity as the wild type in vivo\(^{24}\). It is also reported that mutation Cys71Ser does not display detectable conformational change upon photo activation\(^{24}\). Further study revealed that mutation Cys71Val induces conformational change of VVD similar to the photo activated state of wild type VVD. However Cys71Val VVD does not function as the light-excited VVD.

LOV domains have been subjected to molecular dynamics (MD) simulation studies to explore the mechanisms of switching between Dark and Light states upon photoactivation. Peter et al. carried out MD simulations of AsPhot1 LOV2 domain and showed that the Jα helix could be driven away from the protein core through disruption of location intramolecular interaction upon formation of covalent bond between cysteine and FMN\(^{27}\). They also simulated LOV1 domain of Chlamydomonas reinhardtii and found out that the covalent bond between Cys57 and FMN induced by photoactivation could perturb salt bridge interaction between Glu51 and Lys91 through backbone chain\(^{28}\). They pointed out that the local structural perturbation impacts the dynamical behavior of hydrophobic and hydrophilic sides of LOV1 differently. Based on extensive MD simulations of AsPhot1 LOV2, Freddolino et al. showed that the dynamics of a conserved residue Glu513 are coupled with Jα helix through hydrogen bond interaction and alter the overall dynamical correlation patterns of whole protein\(^{29}\). Peter et al. carried out MD simulations of VVD Dark and Light states as monomer and dimer, and suggested that residue Gln182 close to FAD plays an important role in the transition of VVD from Dark state to Light state through hydrogen bonding interactions after the formation of photo-induced covalent bond between Cys108 and FAD\(^{30}\).

The significant conformational changes of these proteins upon photoactivation are features shared by many LOV domains leading to their wide usage in optogenetic tools. Currently the utility of these devices is limited due to residual activity in the typically inhibitory dark-state. Ideally, LOV allostery could be tuned to create an inactive off-state that digitally switches to a highly functional on-state in response to the lighting conditions. To do so it is essential to identify allosteric sites to target for mutations with substantial impact on overall protein dynamical behavior or structure. For example, the Cys71Val mutant of VVD resides in an intermediate structure that may sample space along the allosteric trajectory between VVD’s Dark and Light states\(^{31}\). Although numerous residues and related mutations were under investigations for LOV domains, a systematic investigation of all the residues of LOV domain is yet to be taken place for their potential contribution to overall protein dynamics and conformational distributions. To identify key residues that link and manipulate Dark-Light state trajectories a systematic analysis of the contribution of all residues to conformational landscape is needed.

Recently, we have developed a simulation method, referred to as rigid residue scan (RRS) to systematically estimate the impact from residues on overall protein dynamics by exerting rigid body constraint\(^{32,33}\) on each individual residue in separate MD simulations\(^{34}\). The examples presented in the RRS development articles...
The RMSD contour plot of the unperturbed Dark and Light MD simulations using RMSD values with references to both optimized Dark and Light states.

**Results**

**Root-Mean-Square Deviation (RMSD).** The mass weighted RMSD of all atoms for Dark and Light states of VVD are plotted in Fig. 2A. Initially, the Light state has higher RMSD values than the Dark state agreeing with the expectation that Light state is more flexible than the Dark state. However, the RMSD values of the Dark state exceed the Light state around 120 ns of simulation, suggesting a larger conformational space accessible for the Dark state than the Light state. To illustrate and compare the conformational distributions of both Dark and Light states, we projected two simulations onto a two dimensional (2D) contour plots using RMSD values with reference to the optimized Dark and Light states, respectively (Fig. 2B). The unperturbed Dark and Light state simulations display different distributions on the 2D RMSD plots. For 210 ns of simulations, the Light state displays only one basin of attraction, but the Dark state displays three major basins of attraction, with the far right one adjacent to the Light state on the same plot.

The 1D and 2D RMSD plots of RRS simulations (all 30 ns) are provided in Supplementary Data (Tables S1 and S2). The RRS simulations of the Dark and Light states with the same residue being held rigid are projected on the same plot. As comparison, the projections of the first 30 ns unperturbed MD simulations are listed as well. For trajectories carried out in most RRS simulations, the Light state is more flexible than the Dark state with higher RMSD values in 1D plots. The 2D RMSD plots of these RRS simulations reveal more interesting details about the impact of rigid body on overall protein dynamics of two states. For the unperturbed states, the Dark and Light states are well separated (Fig. 3A), suggesting that the Dark and Light state structures are well reserved during the course of 30 ns simulations, respectively. Although the 2D RMSD plots are similar to the unperturbed states for many RRS simulations, some rigid residue perturbations dramatically change the overall dynamics of proteins in very different ways (Table S2). For example, when residue Glu112 is treated as rigid body during simulations, the distributions of both Dark and Light states on 2D RMSD plot shift towards each other and have significant overlap (Fig. 3B). When residue Asn119 is treated as rigid body during simulations, the distribution of the Light state on 2D RMSD plot shifts towards the Dark state, while the distribution of the Dark state trajectory remains close to the optimized Dark state as reference structure (Fig. 3C). In a third case, when residue Val124 is treated as rigid body during simulations, the distribution of the Dark state on 2D RMSD plot shifts towards the Light state, while the distribution of the Light state trajectory remains close to the optimized Light state as reference structure (Fig. 3D). These cases represent significantly different protein responses under rigid body perturbations, which could shed light on individual key allosteric residues.

RRS 112, 119 and 124 simulations are not the only cases displaying distinctive perturbation results. Other RRS simulations displaying similar effect in the 2D RMSD distributions are listed in Table 1. For more than thirty residues, the rigid residue perturbations mainly affect Light state only. For only handful residues, the rigid residue perturbations exert significant impact on Dark state. Comparison with the literature reveals that ten residues among those selected to affect mainly Light state in Table 1 have been reported in previous VVD experimental studies. Four residues were reported in the first crystallographic study of VVD, and further confirmed by a crystallographic study of fully light-adapted VVD dimer. Pro30 was reported to display largest shift (2.0 Å) in
the Light state comparing to the Dark state. (In the literature of experimental study of VVD, the residue index is 36 larger than the index used in this study. Therefore Pro30 is referred as Pro66 in the cited articles. This also applies for the following residues in this paragraph.) It was suggested that a mutation at Ile16 prevents conformational switching associated with Light state. Another selected residue, Met19 is reported to stabilize packing of the N-terminal cap against the PAS \( \beta \)-sheet. Asp46 participates in a salt bridge, which was suggested to mediate conformational changes of related LOV domains. Five residues were reported in a mechanistic study of VVD dimerization\(^\text{31}\). Change of Pro6 orientation interrupts key backbone interactions and facilitates the projection of the N-terminus position in the VVD dimer. A mutation at Met19 prevents the forming of VVD cross-linked dimer. On the contrary to Met19, a mutation of Glu135 leads to enhanced dimerization of VVD, suggesting that this position is also critical to VVD function. Another two selected residues Val132 and Tyr141 are located at a hydrophobic pocket that close to N-terminus, which contributes to the propagation of conformational changes triggered by light activation of VVD. In another experimental study of VVD Light state dimer using hydrogen exchange mass spectrometry\(^\text{35}\), two residues, Tyr51 and Ile16, are identified to contribute to asymmetry of two VVD units in its Light state dimer, which may be related to allosteric interaction of two VVD units. In a recent experimental study\(^\text{36}\), mutation of Met129, also listed in Table 1, was reported to stabilize VVD Light state. The residues listed above are not directly located at the flavin binding site, and could allosterically contribute to the VVD function. This certainly enhances the confidence level of the remaining residues, which could be used as guidance for potential target of further mechanistic studies. Due to the relatively short history of VVD study, it is unsurprising that not all of the residues listed in Table 1 have been tested and reported in the literature. Intriguingly, all of the residues covered by experiments belong to the group mainly affects Light state in Table 1. There are much fewer residues in the other two groups, which mainly affect either Dark state only or both states. The reason that residues from these two groups have yet to be tested and reported is likely because it is difficult to intuitively identify and select these residues. The seven residues listed under these two groups in Table 1 actually provide such targets for further studies, which may lead to new direction of VVD studies.

### Clustering analysis

To explore the overall distribution of the simulations, all the RRS and unperturbed simulations of VVD Dark and Light states are plotted on a 2D RMSD plot against Dark and Light crystallographic structures. The overall distribution on the 2D RMSD plot is large and diverse (Fig. 4A). The unperturbed Dark and Light states distribute only on the small fraction of the overall distribution. To facilitate the analysis of distribution on 2D RMSD plot, clustering analysis was carried out for the 2D RMSD distribution, and resulted in a total of 15 clusters. The distributions of Dark and Light state simulations among 15 clusters are plotted in Fig. 4B. All the RRS and unperturbed Dark state simulations mainly dwell in the cluster 1, and with significant contribution to clusters 2 and 3 (Fig. 4B). The RRS and unperturbed Light state simulations contribute significantly to clusters 2 through 10. The diversity of Light state contributions to the different clusters is consistent with the Light state

---

**Figure 3.** 2D contour plots of Dark and Light states distributions. (A) Unperturbed simulations (30 ns to be consistent with RRS simulations); (B) RRS 112; (C) RRS 119; (D) RRS 124. RMSD values for X and Y axes are calculated with reference to the optimized Dark and Light structures, respectively.

**Table 1.** Selected RRS simulations with significant effect on the 2D RMSD distributions for either Dark or Light states or both.
being more dynamic, thereby enabling sampling of diverse states. Further, the overlap of contributions between clusters 2 and 3 in both the Light and the Dark states indicates fluidity in the conformational landscape of the two proteins. These states may contribute to residual activity of LOV proteins in the Dark state that inhibits the fidelity of optogenetic tools. 

The contributions from individual RRS simulations in each cluster are also calculated and plotted (Fig. 5). Most RRS Dark state simulations distribute in Cluster 1 (Fig. 5). But for some of Dark state RRS simulations, the contribution from Cluster 1 is below 50%, showing the significant impact on overall dynamics from perturbations on individual residues. It should be noted that the distributions of most RRS Dark state simulations in Cluster 1 are close to 100%. Therefore, the plot for RRS Dark state simulations (blue line in Fig. 5, Cluster 1) seems to be upside-down, which is not the case. All the RRS Light state, except for RRS119 Light state simulations, have almost no contribution to Cluster 1 (Fig. 5).

Cluster 2, adjacent to Cluster 1, has dominant RRS Dark state simulations distribution and significant contributions from numerous RRS Light state simulations. Cluster 3, adjacent to Cluster 2, has dominant RRS Light state simulations distribution and significant contributions from numerous RRS Dark state simulations. These results are consistent with the data from unperturbed states, where clusters 2 and 3 had contributions in both the Light and the Dark states, indicating these clusters may lie on the conformational trajectory between the Light and Dark states and thereby, contribute to the lack of a digital switch in optogenetic tools. Cluster 6 is also adjacent to Cluster 1, and has dominant RRS Light state simulations distribution and significant contributions from numerous RRS Dark state simulations, similar to Cluster 3. However, Cluster 6 has far less overall distribution from RRS simulations than Clusters 2 and 3.

On the contrary, Clusters 4, 5, 7, 8, and 9 only have significant contribution from RRS Light state simulations, without noticeable contribution from any Dark state simulations. These five clusters are aggregated together at the bottom of the distribution map (Fig. 4A). Clusters 10 through 15 only have significant contribution from a few RRS Light state simulations, and there is no noticeable contribution from any RRS Dark state simulations. These last six clusters cover the top right region of the cluster maps (Fig. 4A). Therefore, the 15 clusters could be divided into the following three groups. Group 1 comprising Clusters 1, 2, 3, and 6 can be considered as the main distribution area for the RRS Dark state. Group 2 comprising Clusters 4, 5, 7, 8, and 9 can be considered as the main distribution area for the RRS Light state. Group 3 comprising Clusters 10, 11, 12, 13, 14, and 15 are the rare distributions, which are accessed mainly by some RRS Light simulations. Representative structures were selected for each cluster and are illustrated for clusters in the above three groups (Fig. 6).

The distributions of unperturbed and RRS 112, 119, and 124 simulations among the above 15 clusters are listed in Table 2. The 210 ns unperturbed Dark state simulations have more than or close to 10% distribution in Clusters 1, 2, 3, 4, 5, and 9, and spans over Group 1 and 2 of clusters. It also has more than 1% distribution in Clusters 7, 8, and 10. On the other hand, the 210 ns unperturbed Light state simulations have more than 10% distribution in Clusters 9, 10, and 13. It also has more than 1% distribution in Clusters 7, 8, 14, and 15. Therefore, the distribution of the unperturbed Light state simulations spans over Groups 2 and 3 of clusters and cover the corridor of the cluster map located at the right hand side (Fig. 4A). It should be noted that both unperturbed simulations have nontrivial distributions in Clusters 7, 8, 9, and 10, indicating the overlap between unperturbed Dark and Light states. Distributions of all RRS simulations among 15 clusters are listed in Table S3.

The impacts from rigid residue constraints on the distribution of Dark and Light states are very different (Table 2). With a rigid residue 112, the Dark state mainly distributes in Clusters 1, 2, and 3 (Group 1) with more than 10% contribution from each cluster. However, with a rigid residue 112, the Light state mainly distributes in Clusters 3 and 4 (Groups 1 and 2). As comparison, the unperturbed Light state simulation has 0.0% contribution from Clusters 3 and 4. Therefore, the rigid residue 112 caused a complete switching of Light state from the unperturbed Light state distributions. With a rigid residue 119, 99.8% of Dark state mainly distributes in Cluster 1 (Group 1); and the Light state mainly distribute in Clusters 1 and 2 (Group 1), also a complete switching from

Figure 4. (A) Cluster analysis of 2D RMSD distribution of all simulations combined; (B) Distribution of all simulations among 15 clusters.
the unperturbed Light state distributions. With a rigid residue 124, the Dark state mainly distributes in Clusters 1, 2, and 3 (Group 1) with more than 10% contribution from each cluster; and the Light state mainly distributes in Cluster 7 (93.5%) and marginally in Cluster 8 (6.5%) (Group 2). Apparently, perturbations on different residues have drastically different and subtle impact on overall dynamics of both Dark and Light states.

**Correlation motion.** Cross-correlation matrix has been widely used to probe protein allosteric mechanisms. The heat maps of cross-correlation matrices for Dark and Light states from unperturbed and RRS simulations with residues 112, 119, 124 being held rigid are illustrated in Fig. 7. The heat maps and histograms of all RRS as well as 30 ns unperturbed simulations are listed in Table S4.
For the unperturbed simulations, although the basic features of heat maps are somewhat similar between Dark and Light states, the Light state has more prominent positive and negative correlations than the Dark state (Fig. 7B). These correlation features are even more prominent for RRS 112 simulations, especially the Light state (Fig. 7D). The correlation matrix heat maps for RRS119 simulations (Fig. 7E and F) are similar to the ones of their corresponding unperturbed simulations (Fig. 7A and B), respectively. However, the correlation matrix heat map for RRS124 Dark state simulation (Fig. 7G) is similar to the one of unperturbed Light state (Fig. 7B), while the one of RRS124 Light state simulation (Fig. 7H) is similar to the one of unperturbed Dark state (Fig. 7A), suggesting tendency of switching between Dark and Light states. It should be noted that many other RRS simulations also display similar correlations of VVD structure either more or less prominent in terms of correlations among different parts (Table S4).

The patterns of different correlations among different parts of VVD protein indicate that the protein structure could be divided into different regions with distinguished dynamical behavior. Using the correlation matrix heat map of RRS 112 Light state as guidance, VVD structure could be divided into seven regions (Fig. 8): Region 1 (residues 1 through 5) is the N-terminus, region 2 (residues 6 through 30, including α-helix and loop structure), region 4 (residues 35 through 83, including two β-sheets and two short α-helices and loops connecting them), region 5 (residues 84 through 108 including α-helix and loop structures), region 6 (residues 109 through 131 including two β-sheets and a loop in between), region 7 (residues 132 through 138 as a loop structure), and region 8 (residues 139 through 148) as the C-terminus. Region 1 as N-terminus displays clear negative correlation with regions 2, 4, 6, and 8, and positive correlation with regions 3 and 5. On the contrary, C-terminus region 8 with a β-sheet structure displays clear positive correlation with regions 2, 4, and 6, and clear negative correlation with regions 1, 3, and 5. The most striking feature is arguably region 4 as the largest self positively correlated region. Overall regions 2, 4, 6, and 8 are positively correlated with each other, but are somewhat negatively correlated with regions 1, 3, 5, and 7.

**Protein configurational entropy.** Configurational entropies (referred as entropy) of VVD are calculated through quasi-harmonic analysis for each simulation. The 30 ns trajectories of unperturbed simulations were used for the entropy calculations for the consistency with regard to RRS simulations. For the convenience of comparison, the entropy of the unperturbed Dark state simulation was used as reference for all other simulations, including both Dark and Light states. The relative VVD entropies are plotted in Fig. 9 and listed in Table S5 and with ascending order in Table S6.

For the unperturbed simulations, the Light state has entropy higher than the Dark state by 0.215 kcal/mol·K. For the RRS simulations of Dark state, the relative entropies range from ~0.122 kcal/mol·K (RRS 141) to 0.230 kcal/mol·K (RRS 109). For 123 Dark state RRS simulations, the entropy of VVD protein is higher than the one from the unperturbed simulation, while for the other 25 Dark state RRS simulations, the entropy of VVD is lower. The relative VVD entropies from the most RRS simulations of Light state (with reference to the unperturbed Dark state simulation) are positive. However, when comparing to the unperturbed Light state simulation, total of 118 RRS simulations of Light state have lower VVD entropies (horizontal line in Fig. 9). This shows that the entropic response from the Light state to the rigid residue perturbation is intrinsically different from the Dark state.

The VVD entropy differences between Dark and Light states when the same residue is held rigid are plotted in Fig. 10 and also listed in Tables S5 and S6. For both RRS 112 and 119 simulations, the differences between Dark and Light state entropies are rather small as 0.007 and 0.004 kcal/mol·K, respectively, and are among the smallest of all RRS simulations (Table S6). RRS124, on the other hand, is the simulation with the most entropy decrease of

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Unperturbed</th>
<th>RRS112</th>
<th>RRS119</th>
<th>RRS124</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dark% Light%</td>
<td>Dark% Light%</td>
<td>Dark% Light%</td>
<td>Dark% Light%</td>
</tr>
<tr>
<td>1</td>
<td>18.9 0.0</td>
<td>22.3 0.0</td>
<td>99.8 25.1</td>
<td>11.4 0.0</td>
</tr>
<tr>
<td>2</td>
<td>10.3 0.0</td>
<td>58.0 5.4</td>
<td>0.2 67.0</td>
<td>18.6 0.0</td>
</tr>
<tr>
<td>3</td>
<td>18.3 0.0</td>
<td>12.4 55.4</td>
<td>0.0 5.7</td>
<td>68.1 0.0</td>
</tr>
<tr>
<td>4</td>
<td>18.3 0.0</td>
<td>0.1 35.1</td>
<td>0.0 0.2</td>
<td>1.9 0.0</td>
</tr>
<tr>
<td>5</td>
<td>15.7 0.2</td>
<td>0.0 1.5</td>
<td>0.0 0.0</td>
<td>0.0 0.0</td>
</tr>
<tr>
<td>6</td>
<td>0.2 0.0</td>
<td>7.3 0.0</td>
<td>0.0 1.0</td>
<td>0.0 0.0</td>
</tr>
<tr>
<td>7</td>
<td>3.9 5.0</td>
<td>0.0 0.9</td>
<td>0.0 0.0</td>
<td>0.0 93.5</td>
</tr>
<tr>
<td>8</td>
<td>3.2 6.3</td>
<td>0.0 0.2</td>
<td>0.0 0.0</td>
<td>0.0 6.5</td>
</tr>
<tr>
<td>9</td>
<td>9.1 20.8</td>
<td>0.0 0.0</td>
<td>0.0 0.0</td>
<td>0.0 0.0</td>
</tr>
<tr>
<td>10</td>
<td>1.4 49.6</td>
<td>0.0 0.0</td>
<td>0.0 0.0</td>
<td>0.0 0.0</td>
</tr>
<tr>
<td>11</td>
<td>0.1 0.0</td>
<td>0.0 1.5</td>
<td>0.0 1.1</td>
<td>0.0 0.0</td>
</tr>
<tr>
<td>12</td>
<td>0.6 0.7</td>
<td>0.0 0.0</td>
<td>0.0 0.0</td>
<td>0.0 0.0</td>
</tr>
<tr>
<td>13</td>
<td>0.0 14.5</td>
<td>0.0 0.0</td>
<td>0.0 0.0</td>
<td>0.0 0.0</td>
</tr>
<tr>
<td>14</td>
<td>0.0 1.8</td>
<td>0.0 0.0</td>
<td>0.0 0.0</td>
<td>0.0 0.0</td>
</tr>
<tr>
<td>15</td>
<td>0.0 1.2</td>
<td>0.0 0.0</td>
<td>0.0 0.0</td>
<td>0.0 0.0</td>
</tr>
</tbody>
</table>

Table 2. Distribution of selected individual Dark and Light state simulations among 15 clusters. (Significant distributions are listed in bold).
Light state simulation comparing to its Dark state simulation ($-0.225$ kcal/mol · K). Clearly, rigid body perturbation on residue 124 has opposite effect on Dark and Light state entropies.

**Entropy contribution from individual residue.** The entropy contribution from each individual residue in each simulation was calculated through the analysis of the cross-correlation matrices. For each RRS Dark state simulation, the relative entropy contribution from each individual residue was calculated with reference to the entropy contribution from the same residue in the unperturbed simulation of the VVD Dark state. The same calculation was also carried out for the RRS Light state simulations. The relative individual residue entropies are plotted as heat maps for RRS simulations of Dark and Light states in Fig. 11. The blue diagonal lines in both plots are due to the fact that the entropies from residues as rigid body is significantly smaller than the reference state. In addition, the averaged relative entropies of individual residues from all the RRS simulations are plotted in Fig. 12 for Dark and Light states, respectively. Most residues do not show significant entropy change upon rigid residue perturbation. However, certain residues display rather consistent but different trends in RRS simulations of Dark

*Figure 7. Cross-correlation matrix heat maps for unperturbed, RRS 112 (Glu), RRS 119 (Asn), RRS 124 (Val) MD simulations of Dark and Light states.*
The most striking case is residue Lys85, which has significantly positive entropic response in RRS simulations of Dark state and significantly negative entropic response in RRS simulations of Light state.

Figure 8. Structural regions revealed by cross-correlation analysis. (A) Regions highlighted in matrix of RRS 112 of Light state. (B) Regions illustrated in crystallographic structure of Light state. Listed regions: Region number, residues (color): 1, 1–5 (Blue); 2, 6–30 (ice blue); 3: 31–34 (yellow); 4: 35–83 (red); 5: 84–108 (violet); 6: 109–131 (brown); 7: 132–138 (magenta); 8: 139–148 (green). Residues E112, N119, and V124 are also labeled.

Figure 9. Relative entropy of VVD in dark and light states from RRS simulations. The entropy from unperturbed dark simulation is used as reference. Red horizontal line in Light state plot marks the entropy level from the unperturbed Light state simulation.

Figure 10. Entropy difference between dark and light VVD states from unperturbed simulation and RRS simulations.
both Dark and Light states, Lys85 forms strong hydrogen bond with the phosphate group of FAD (Fig. 13). It is notable that N-terminus residues 1 through 10 have generally positive entropic response towards rigid residue simulations in Dark state, and these positive entropic responses are further enhanced in the Light state. On the contrary, the entropic responses from residues 28 through 37 change from positive in Dark state simulations to significant negative in the Light state simulations. The individual residue entropies were also normalized with regard to the atoms number in each residue and illustrated in Figure S1. The similarity between normalized and original residue entropies plots shows that the differences of residue responses are not scaled with residue size, and therefore inherent to protein structure.

Principal component analysis (PCA). PCA was carried out for all the simulations to project protein motions onto normal mode vectors. The vectors were used to explore and compare the global motions of VVD in each simulation. Because the main purpose of PCA is characterizing the global motions of the protein, only α carbons were included for this analysis. To uncover the correlation between VVD Dark and Light states dynamics, the optimized Dark and Light structures were used as reference structures for the PCA of unperturbed Light and Dark state simulations, respectively. The accumulative contributions of the first 20 principal components (PCs) of two VVD states are plotted (Figure S2). For both simulations, the PC1s are the dominant vectors with 82% and 92% contribution to the overall protein dynamics, respectively. The remaining contribution to protein dynamics is evenly distributed among other PCs. The dot product between two dominant PC1 vectors is 0.928, showing that both vectors reflect the change between the Dark and Light states. Therefore, PC1 from the Light state trajectory was used for the projection of RRS simulations.
Projection of RRS simulations onto principal component vectors. To compare the impact of rigid residue perturbations on VVD Dark and Light states distribution in the space of principal components, each pair of RRS simulations were projected onto the PC1 vector from unperturbed Light state simulations (Table S7). The density distribution of each trajectory projected onto each vector was calculated employing kernel density estimator (KDE). The density distributions for the unperturbed and RRS 112, 119, 124 simulations are shown in Fig. 14.

The unperturbed Dark and Light state simulations are well separated (Fig. 14A). The unperturbed Dark state distribution centering close to the origin (blue line in Fig. 14A) is more diffused than the unperturbed Light state centering close to far left (red line in Fig. 14A). For the RRS112 simulations, the distributions of both Dark and Light state simulations become more diffused and move to each other and with significant overlap. Both distributions of RRS112 simulations display multiple peaks, which correspond to two basins of attractions for both simulations on 2D RMSD plot (Fig. 2B). For the RRS119 simulations, the distribution of Dark state simulation remains close to origin (blue line in Fig. 14C) similar to unperturbed Dark state simulation. This similarity supports the observation that rigid residue 119 exerts very little impact on the VVD Dark state in the simulation (Fig. 3C). However the distribution of RRS119 Light state simulation (red line in Fig. 14C) displays significant shift and overlap with the Dark state. On the contrary, the distribution of RRS124 Light state simulation remains close to the region of far left (red line in Fig. 14D) similar to unperturbed Light state simulation. This similarity supports the observation that rigid residue 124 exerts very little impact on the VVD Light state in the simulation (Fig. 3D). But the distribution of RRS124 Dark state simulation (blue line in Fig. 14D) displays significant shift toward the Light state.

Discussion
In the design of optogenetic tools, a fundamental limitation of LOV-based systems is the lack of a digital switch from a non-functional off-state to a highly-function on-state. Rather, optogenetic tools have residual activity
in both the Light and Dark states thereby limiting their fidelity. To examine the origins of the lack of a digital LOV-based photoswitch, we wished to computationally examine the allosteric activity of LOV based proteins to extract information relevant to the conformational landscape and configurational entropy of Dark and Light states of VVD. In so doing, we discovered an inherent difference in correlating sampling of conformational space with calculations of configurational entropy. Specifically, where the VVD Dark state samples more conformational space, but the Light state has higher configurational entropy. These analyses shed light on the nature of allostery in LOV based photoswitches.

Examination of the unperturbed 210 ns simulations revealed profound differences between the overall dynamics of VVD Dark and Light states. Specifically, the Dark state samples more conformational space during the course of 210 ns simulations than the Light state (Fig. 1B). Further, the Dark state samples conformational space, which lie close to the Light state, thereby likely contributing to residual activity in both states and loss of a digital switch as observed in optogenetic tools. The counterintuitive nature of the conformational sampling of the Dark state, could lie in the need to facilitate a fast switch from the dormant Dark state to active Light state and to reduce barriers in the transition. In such a system, the Dark state samples conformational space poised for activity. Blue-light then populates the Light state that is “locked” into specific conformational space to better serve its signaling function.

MD simulations provide direct time series protein dynamical data. Conformational landscape is a comprehensive representation of intrinsic protein dynamics. The configurational entropy calculated based on MD simulations is a quantitative metric to measure the distribution of given protein on its conformational landscape. Because it is preferred to calculate configurational entropy around single attraction basin when using quasi-harmonic approximation, the first 30 ns trajectories of the unperturbed simulations, during which two states of VVD protein remain separate and close to their native crystallographic structures, were used to compute the configurational entropies for the sampling space. In addition, 30 ns trajectory was carried out for each RRS simulation to be consistent with unperturbed simulations. Our previous study also support that the 30 ns trajectories are sufficient to probe protein allosteric mechanisms.

The 2D RMSD plots of RRS simulations provide detailed information with regard to the comparison of impacts from rigid residue perturbation on Dark and Light state dynamics. The different impact on the VVD dynamics from the perturbation on residues 112, 119, and 124 could provide mutagenesis targets for different purposes, i.e. when different changes are desired on the Dark and Light states, or to improve the digital-nature of optogenetic tools. Interestingly, there are many more residues (more than 30) upon which the rigid body perturbation significantly affects the Light state (Table 1), than those affect either Dark state only (four residues) or both states (three residues). These disproportional numbers of residues in different groups demonstrate the potential value of this analysis as the guidance for further experimental study. These also suggest that the Light state is much more prone to perturbations than the Dark state, and may correlate with the narrower distribution of the Light state simulation on the 2D RMSD plot (Fig. 2B).

One interesting observation is that the side chains of selected key residues 112, 119, and 124 are located at the protein surface and point away from the binding site of flavin (Fig. 15). This certainly should not be considered as purely coincidental. These three residues are identified through their dramatic impact on overall protein dynamics. Therefore it is logical to hypothesize that these residues could carry allosteric function to propagate the signal triggered by light induced activation of VVD.

The clustering analysis of the distribution of all RRS and unperturbed simulations on 2D RMSD plots provides a comprehensive representation of the conformational space accessible to the VVD Dark and Light states. The functional conformations and structures of both Dark and Light states as well as potential intermediates all belong to this conformational space. In a previous crystallographic study of VVD mutant Cys71Val, two distinctive structures were identified for its Light state (Fig. 16. 3D72_A, blue, and 3D72_B, ice blue)11. The extended N-terminus in the Cys71Val mutation was also observed through small-angle X-ray scattering (SAXS) studies13,40. Correspondingly, representative conformations of Clusters 1 (red) and 13 (magenta) (Fig. 16) are similar to these two structures. The native Light state displays yet another conformation of N-terminus (3RH8, green structure in Fig. 16). It was also stated that electron density of residues 37–44 was weak or absent in the crystallographic
data of VVD Cys71Val mutant in Light state. The wide distribution of this region represented by all 15 clusters (Fig. 2B) also showed that this region is very flexible with access to large conformational space.

Another intriguing observation of clustering analysis is that in clusters 1 through 10, the Fα and Eα helices, which cradle the FAD cofactor, remain in an extremely similar orientation, leaving a narrow space in between (Fig. 6A and B). However, in clusters 11 through 15, both Fα and Eα helices display obvious deviations among these structures, creating more space probably responding to the photo-activation of FAD.

Both Pro42 and Cys71 (residue number 6 and 35 in the current study, respectively) are key residues projecting the N-terminus into different conformations (24,31). Correspondingly, both RRS6 and RRS35 simulations showed significant impact on VVD Light state dynamics (Fig. 17). In RRS6, the Light state distribution splits into two attraction basins. In RRS35, the Light state distribution shifts upward along Y-axis, farther away from crystallographic Light state conformation.

The other two key residues 55 and 171 (residue number 19 and 135 in the current study, respectively) for VVD dimerization process discussed in ref. 31 also have significant impact on Light state dynamics under constraint perturbation (Fig. 17). In RRS19, the Light state distribution shifts toward left hand side along X-axis, closer to crystallographic Dark state conformation. But in RRS135, the Light state distribution splits into two attraction basins, which are significantly separate from each other. For all four residues above, the cross-correlations are significantly enhanced in Light states, while the cross-correlations remains moderate similar to unperturbed Dark state (Fig. 18).

It is intriguing that rigid residue perturbations on different residues have drastically different impact on the overall distributions of VVD Dark and Light state simulations. The systematic probe of all residues through RRS simulations provides a comprehensive map of impact for every single residue on overall protein dynamics, which could serve as guidance to identify potential target for improving the fidelity of optogenetic tools. The fact that very few residues could serve as targets to affect Dark state dynamics signifies the potential usefulness of RRS as systematic probing tool to identify allosteric sites with specific effects.

As demonstrated in this research, the RRS method could be employed as an effective computational tool to systematically perturb the protein structure through molecular dynamics simulations to screen key residues potentially important for protein allosterity. This method has several advantages. First, no a priori knowledge about the target protein is necessary, which makes it easy to implement. Second, no mutations are generated to screen residues, because rigid body constraints are used as perturbation instead of mutations. Therefore, the chemical
The nature of molecular interaction between each target residue and its local environment is reserved, and the main effect being observed from the simulation would be impact on overall protein dynamics from removing the internal degrees of freedom of target residue. Because no mutations are necessary, the difficulty of choosing somewhat arbitrary amino acids for mutation could be avoided. Third, extensive molecular simulations could probe the change of real protein dynamics, and are more reliable than approximation methods such as elastic network models. Fourth, the scan provides complete information of all residues. Therefore any previously unnoticed but potentially important residues could be identified for further investigation. This method does have certain limitations. For example, the rigid body constraint is an unphysical perturbation. Therefore, the observation from using this method could not be rigorously verified by experiments. However, this could be overcome by additional simulations of mutants on the key residues selected using RRS method. In addition, the computational cost of proteins containing large number of residues could be prohibitively expensive for long molecular dynamics simulations. This could be overcome by running relatively short simulations as prescreening to select residues for longer simulations. Considering all the factors above, the RRS method is an effective and affordable computational tool.

Figure 18. Cross-correlation matrix heat maps for RRS 6 (Pro), RRS 19 (Met), RRS 35 (Cys), and RRS 135 (Glu) MD simulations of Dark and Light states.
could be valuable and compatible with other methods for many systematic and comprehensive studies of protein structures and functions.

Materials and Methods

Molecular dynamics simulations. The initial structures for Dark and Light states of VVD were obtained from the Protein Data Bank (PDB) with IDs as 2PD7 and 3RH8, respectively. The Dark state in 2PD7 contains residues 36 through 184, and the Light state in 3RH8 contains residues 37 through 184. For consistency and convenience, residue 36 in 2PD7 was removed and residues 37 through 184 in both states were numbered as residues 1 through 148. Both structures include protein and ligand flavin adenine dinucleotide (FAD). Because flavin mononucleotide (FMN) and FAD carry similar biological role, adenosine monophosphate (AMP) moiety was removed from FAD to form FMN, which was parameterized with force field from a previous study. Hydrogen atoms were added to the VVD Dark and Light states, which were subsequently solvated using explicit water model (TIP3P) and neutralized and balanced by sodium cations and chloride anions. The ionic strength of the simulation box is 0.11 M for both Dark and Light states. The simulation systems were subjected to energy minimization with 200 steepest descent steps (for both), 1709 adopted basis Newton-Raphson (ABNR) minimization steps for Dark state and 1540ABNR steps for Light state, yielding a total gradient of less than 0.03 kcal/mol/Å. The optimized simulation boxes were subjected to 12 picoseconds (ps) equilibrium simulations raising the temperature from 100 K to 300 K. 10 nanoseconds (ns) of isothermal-isobaric ensemble (NPT) MD simulations were carried out for both states. Appropriate simulation boxes with size closest to the average NPT simulation boxes were selected for two VVD states for production canonical ensemble (NVT) Langevin MD simulations at 300 K.

We applied rigid residue scan (RRS) method developed in our group to systematically detect key allosteric residues related to Dark and Light states. In RRS method, rigid body constraint using an efficient integrator was applied on each individual residue in separate simulations (referred as RRS simulations). Therefore, there are 148 rigid residue simulations for each state. The simulations without rigid residue constraints were referred as unperturbed simulations. For both unperturbed and RRS simulations, initial 4 ns NVT simulations were considered as equilibrium stage and not counted as production run. Besides these 4 ns equilibration, total of 210 ns unperturbed MD simulations were carried out for Dark and Light state, respectively, and total of 30 ns RRS MD simulations were carried out for each residue in each state. Overall we performed 9,300 ns production NVT simulations.

For all simulations, bonds associated with hydrogen atoms were constrained, and step size of 2 femtosecond (fs) was used. Simulation trajectories were saved every 500 MD steps. Cubic simulation box and periodic boundary condition were applied for all MD simulations. Electrostatic interactions were calculated using particle mesh Ewald (PME) method. All simulations were carried out using CHARMM simulation package version 39b1 and the CHARMM22 force field with CMAP correction.

Analysis of MD trajectory. Least-square-fitting alignment root-mean-square deviation (RMSD). The minimized structures for the Dark and Light states were used as reference structures for RMSD analyses, which is defined as:

\[ R = \sqrt{\frac{\sum_{i=1}^{N} (r_i^0 - U r_i)^2}{N}}. \]

N is the number of atoms, \( r_i^0 \) is the Cartesian coordinate vector for atom \( i \), and \( U \) is the best-fit alignment transformation matrix between a given structure and its reference structure. For every trajectory, RMSD values were calculated using both Dark and Light optimized structures as references. Two RMSD values were used to plot 2D RMSD distribution.

Cross-correlation (normalized covariance) matrix. The cross-correlation matrix was calculated to detect the correlation among atomic motions of protein. The matrix element \( C_{ij} \) measuring correlation between atoms \( i \) and \( j \) is defined as

\[ C_{ij} = \frac{c_{ij}}{c_{ii}^{1/2} c_{jj}^{1/2}} = \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\left( \langle r_i^2 \rangle - \langle r_i \rangle^2 \right)^{1/2} \left( \langle r_j^2 \rangle - \langle r_j \rangle^2 \right)^{1/2}}, \]

where \( c_{ii} \), \( c_{jj} \), and \( c_{ij} \) are the atomic coordinate covariance matrix elements, and \( r_i \) and \( r_j \) are Cartesian coordinate vectors from the least-square fitted structures. The translational and rotational motions were projected out through the least-square fitting procedures. The normalized matrix elements range from -1 to 1, with positive value indicating positive correlation and negative values indicating negative correlations.

Quasi-harmonic analysis and configurational entropy calculation. Quasi-harmonic analysis was carried out through the inversion of cross-correlation matrix \( C \)

\[ F_{ij} = k_B T [C^{-1}]_{ij}. \]

\( F \) is the force constant matrix for the effective quasi-harmonic potential, \( k_B \) is the Boltzmann constant, and \( T \) is the temperature. The normal modes and corresponding frequency \( \omega \) of the molecule on the effective quasi-harmonic potential can be calculated through the solution of secular equation.
\[ \det(F - \omega^2 \mathbf{M}) = 0, \]  
where \( \mathbf{M} \) is the mass matrix of protein. Subsequently, protein configurational entropy \( S_{\text{conf}} \) within harmonic limit was calculated using the frequency from above quasi-harmonic analysis,

\[ S_{\text{conf}} = k_B \sum_{i} \frac{\hbar \omega_i/kT}{e^{\hbar \omega_i/kT} - 1} - \ln(1 - e^{-\hbar \omega_i/kT}). \]

The configurational entropy for each individual residue could be calculated in the same way using sub-correlation-matrix for atoms within the same residue (including side chain and backbone). The configurational entropy for each individual residue calculated in this way does not include the correlation between the target residue and the rest of the system. Therefore, the sum of individual residue configurational entropies is not expected to match with the one of whole protein.

**Principal component analysis (PCA).** PCA was carried out for each simulation to examine overall protein motions. The only difference between PCA and quasi-harmonic analysis is that PCA does not use mass weighting. Because the focus of PCA is overall protein motions, only protein \( \alpha \) carbons from each trajectory were used for this purpose. It was occurred to us that the different choices of reference structures could reveal different prospects of protein dynamics. Therefore, both optimized Dark and Light structures were used as reference structures for PCA of each simulation. For each RRS simulation, the residue being held rigid was included in the PCA just as other residues. All the analyses described above were carried out using CHARMM version 39b1.

**References**

How to cite this article: Zhou, H. et al. Revealing Hidden Conformational Space of LOV Protein VIVID Through Rigid Residue Scan Simulations. Sci. Rep. 7, 46626; doi: 10.1038/srep46626 (2017).

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2017

Acknowledgements
This work was partially supported by the Edward R. Biehl Graduate Fellowship (HZ), the National Institutes of Health (R15GM109282: BDZ), and Ralph E. Powe Junior Faculty Enhancement Award (PT). Computational time was provided by Southern Methodist University’s Center for Scientific Computation and Texas Advanced Computing Center (TACC) at the University of Texas at Austin.

Author Contributions
H.Z. and P.T. designed the project. H.Z. performed and analyzed the simulations. H.Z., B.D.Z., and P.T. interpreted the data and wrote the manuscript.

Additional Information
Supplementary information accompanies this paper at http://www.nature.com/srep

Competing Interests: The authors declare no competing financial interests.
PAPER III

Dynamics Sampling in Transition Pathway Space

Hongyu Zhou and Peng Tao *

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

Supporting Information

ABSTRACT: The minimum energy pathway contains important information describing the transition between two states on a potential energy surface (PES). Chain-of-states methods were developed to efficiently calculate minimum energy pathways connecting two stable states. In the chain-of-states framework, a series of structures are generated and optimized to represent the minimum energy pathway connecting two states. However, multiple pathways may exist connecting two existing states and should be identified to obtain a full view of the transitions. Therefore, we developed an enhanced sampling method, named the direct pathway dynamics sampling (DPDS) method, to facilitate exploration of a PES for multiple pathways connecting two stable states as well as addition minima and their associated transition pathways. In the DPDS method, molecular dynamics simulations are carried out on the targeting PES within a chain-of-states framework to directly sample the transition pathway space. The simulations of DPDS could be regulated by two parameters controlling distance among states along the pathway and smoothness of the pathway. One advantage of the chain-of-states framework is that no specific reaction coordinates are necessary to generate the reaction pathway, because such information is implicitly represented by the coordinates along the pathway. The chain-of-states setup in a DPDS method greatly enhances the sufficient sampling in high-energy space between two end states, such as transition states. By removing the constraint on the end states of the pathway, DPDS will also sample pathways connecting minima on a PES in addition to the end points of the starting pathway. This feature makes DPDS an ideal method to directly explore transition pathway space. Three examples demonstrate the efficiency of DPDS methods in sampling the high-energy area important for reactions on the PES.

1. INTRODUCTION

Sampling transition pathways that connect two states for a given system are an important and active area of methodology development in computational chemistry. Chain-of-states methods were developed to obtain a minimum energy pathway connecting two predefined states. Many chain-of-states methods were developed to obtain minimum energy pathways. Elber and Karplus first used a line integral representation of a discretized path for optimization. Following the same line of thinking, nudged elastic band (NEB) methods were developed by projecting out perpendicular components of elastic forces and parallel components of the true force with respect to the path under minimization. In zero temperature string (ZTS) methods, the states along the transition pathway are evenly distributed and re-distributed along the fitted path after each step of minimization or simulation. It should be noted that the ZTS method could automatically optimize end structures to local minima, regardless of whether the starting pathway connects two minima or not. Many variations of chain-of-states methods with improved efficiency and accuracy have been developed subsequently, but will not be covered in detail.

Due to the high number of degrees of freedom, the convergence of chain-of-states methods can be slow in some applications, especially for macromolecules. Therefore, specialized chain-of-states methods were developed to generate reference reaction paths as a good approximation to the true minimum energy pathway (MEP) with fast convergence rates. For example, the replica path (RPATH) method implemented in the CHARMM program package utilizes harmonic restraints of both best-fit root-mean-square distances (RMSD) and metallines defined by RMSD to control the distribution of replicas along the pathway and the smoothness of the pathway. In addition to harmonic restraints, equal distance holonomic constraints have also been implemented in CHARMM to maintain an even distribution of replicas along the pathway during the pathway optimization. Based on a recent benchmark study, the MEP calculated by these reference pathway methods are comparable with NEB results, but with a superior convergence rate. Because of low computational cost, such reference pathway methods could be used for extensive sampling of pathways.

In addition to determining the MEP, sampling along the pathway to estimate the transition free energy is also an active area of methodology development. To obtain a sufficient sampling around high-energy states for the accurate free energy estimation, enhanced sampling methods were developed. While not typically considered as an enhanced sampling method,
umbrella sampling methods are arguably the most widely applied methods to gain sufficient sampling in high-energy regions to calculate a free energy profile with regard to the predefined order parameters.\textsuperscript{16–19} Metadynamics methods do not sample toward a specific transition pathway but, instead, discourage revisiting already sampled states to promote exploration of other parts of a potential energy surface (PES), including high-energy states.\textsuperscript{20–23} Yang and co-workers proposed a minimum free energy pathway method combined with hybrid quantum mechanical and molecular mechanical free energy (QM/MM-FE) methods to couple the optimization of the MEP of an enzymatic reaction pathway using QM/MM methods with sampling of the environment to obtain the most probable reaction pathway.\textsuperscript{24,25} Transition path sampling (TPS) methods were first proposed by Pratt\textsuperscript{26} and much further developed by Chandler and his collaborators.\textsuperscript{27–31} Different from other enhanced sampling methods, no specific reaction pathway defined by a certain reaction coordinate is required in TPS methods. In TPS methods, a large number of simulations are carried out to sample the transitions between two target stable states. To enhance the success rate of transitions in the simulations, the TPS simulations often start from high-energy states, which are preferably close to the transition state regions between two states. Roux and co-workers combined string methods with a large number of short simulations, which were referred to as a swarm of trajectories, to identify the most probable transition pathways between two stable states.\textsuperscript{32} Vanden-Eijnden and his co-workers further advanced the reaction pathway sampling by proposing transition path theory (TPT).\textsuperscript{33–40} In TPT, all the transition pathways connecting two states are viewed as a part of a long trajectory, which samples both states numerous times. The probability density function and probability current function were proposed to measure the transition between two end states in the TPT. Dominant reaction path (DRP) methods were developed by Faccioli and co-workers based on the Fokker–Planck equation to search for the transition pathway with a minimum action potential.\textsuperscript{41–43}

In addition to the development of the transition pathway optimization and enhanced sampling methods either targeting a specific transition pathway or between two predefined states, further methodology development is still necessary to explore potential multiple transition pathways connecting two states and detect additional minima on the PES of systems of interest and their associated transition pathways. The chain-of-states framework has been applied to sample the potentials of the mean force along a reference pathway showing certain advantages over related methods.\textsuperscript{44} In a similar vein, we proposed a new method to combine the chain-of-states framework with molecular dynamics (MD) simulations to directly sample the target PES in the transition pathway space. We refer to this approach as the direct pathway dynamics sampling (DPDS) method. The remaining part of this work is organized as the following. Computational framework and methods are described in section 2. Test of the DPDS method using three model systems is presented in section 3. Both advantages and limitations of this method are discussed in section 4. The contribution is concluded in section 5.

2. COMPUTATIONAL METHOD

2.1. Chain-of-States Representation of Reaction Pathway. A transition pathway represented by a series of structures on a PES is the basic representation of a simulation system. The original chain-of-states algorithm was developed for energy optimization on a PES to characterize the MEP. The final MEP from any given representation of a transition pathway is likely a local MEP instead of the global one. Therefore, by combining the chain-of-states framework with molecular dynamics simulations, the sampling efficiency on a PES could be greatly enhanced. In a chain-of-states framework, a series of structures described by certain coordinates are constructed to represent a transition pathway from one state to another. Cartesian coordinates are a convenient option to generate and maintain replicas within a chain-of-states framework. The distance between two states (replicas) \(i\) and \(j\), \(d_{ij}\), can be defined as the following:\textsuperscript{14,15}

\[
 d_{ij} = \sqrt{\frac{\sum_{i=1}^{N} w_i (r_i^j - U^0 r_i^j)^2}{\sum_{i=1}^{N} w_i}}
\]

where \(N\) is the total number of atoms in the system, \(r_j^i\) is the Cartesian coordinate of atom \(i\) in the structure \(j\), \(U^0\) is the rotation matrix to superimpose the structure \(j\) over the structure \(i\) for least-squares fitting between two structures, \(w_i\) is the weight factor for atom \(i\), which includes atomic mass and additional factors.

In addition, any other appropriate reaction coordinates (RCs) or collective variables (CVs) can also be employed to maintain these replicas. The initial pathway can start with any state, and the MEP is a reasonable starting point. The end replicas are preferred to be two stable states as minima on the PES.

2.2. Restraints To Control Chain-of-States Pathways. Restraints are used to maintain the distance between adjacent states and prevent collapse of the pathway into a local minimum. The harmonic potential using best-fit RMSD has been shown as an effective metric to control the interstructure distance in the optimization of the MEP within the chain-of-states framework. The total potential energy associated with these added harmonic potentials to control RMSD between adjacent replicas is defined as the following:\textsuperscript{13,15}

\[
 E_{\text{rms}} = \sum_{i=1}^{n} \frac{1}{2} k_{\text{rms}} (d_{i,i+1} - \bar{d})^2
\]

where \(k_{\text{rms}}\) is the harmonic force constant to restrain RMSD distances between adjacent replicas along the reaction pathway, \(d_{i,i+1}\) is the best-fit RMSD between replica \(i\) and \(i + 1\), \(\bar{d}\) is the average RMSD between adjacent replicas, and \(n\) is the total number of replicas of the given pathway.

To control the smoothness of the sampling pathway, an additional potential could be added to restrain the pseudoangle between structures along the pathway:\textsuperscript{13,15}

\[
 E_{\text{angle}} = \sum_{i=1}^{n} \frac{1}{2} k_{\text{angle}} (\text{COSMAX} - \cos(\theta))^2 \quad \text{COSMAX} > \cos(\theta)
\]

\[
 E_{\text{angle}} = 0 \quad \text{COSMAX} \leq \cos(\theta)
\]

The pseudoangle \(\theta\), illustrated in Figure 1, describes the smoothness of the pathway. The force constant \(k_{\text{angle}}\) controls the smoothness of the pathway by keeping \(\theta\) from getting too small. COSMAX is a cutoff value determining the value of \(\cos(\theta)\) subjected to the control of pathway smoothness. Both pathway RMSD and pseudoangle \(\theta\) controlling are implemented in the RPATH module of the CHARMM program package.\textsuperscript{13}

2.3. Molecular Dynamics Simulation within Chain-of-States Formalism. The conventional application of chain-of-states is to minimize the MEP, with the change of structures along the pathway only to reduce either the total energy of the
pathway when using restraints or forces perpendicular to the pathway. To achieve the goal of enhancing sampling efficiency on a PES, the dynamical propagation needs to be integrated with the chain-of-states control. This integration is implemented through adding pathway control forces, including distance potential and angle potential, into the potential functions for dynamical propagation. The force originating from the harmonic potential based on the RMSD restraints applied on atom \( i \) in structure \( i \) with Cartesian coordinates as \( r_i \) is defined as

\[
f_{\text{force}}(r_i) = \frac{\partial E_{\text{ms}}}{\partial r_i}
\]

Similarly, the force originating from the harmonic potential applied to pseudo-angle \( \theta \) associated with structures, \( i, i+1, \) and \( i+2 \), is calculated as

\[
f_{\text{angle}}^i = \frac{\partial E_{\text{ang}}}{\partial r_i}
\]

Both \( f_{\text{force}} \) and \( f_{\text{angle}} \) are added to the force used for the MD simulations integration scheme for atom \( i \) in structure \( i \). This integration scheme was implemented in CHARMM associated with the RPATH module.

2.4. Nudged Elastic Band. Although NEB methods were originally developed to optimize the transition pathway on a PES to obtain the MEP, it is also suitable for the current framework. In the NEB methods, the forces along the pathway are projected using the pathway tangent vector (\( \tau_i \)).

\[
F_i = F_i^+ + F_i^-
\]

\[
F_i^+ = -\nabla V(r_i)(1 - \tau_i)
\]

\[
F_i^- = -V\left(\frac{1}{2}k\sum_{j=1}^{N}(\Delta l^i - \Delta l_j)^2\right)(\tau_i)
\]

where \( F_i^+ \) and \( F_i^- \) are the components of force that are perpendicular and parallel to the tangent vector of replica \( i \) (\( \tau_i \)), respectively. \( V \) is the potential energy function, and \( \Delta l \) is the distance between adjacent replicas.

2.5. Control of the Two End Replicas. The main goal of conventional chain-of-states methods is to obtain an MEP connecting two minima. In most cases, two minima are characterized before obtaining the MEP connecting them. In such situations, two predefined minima are set as two end replicas on the initial chain-of-states setup and are constrained by removing all its degrees of freedom during the optimization process. Although restraint as harmonic potential can also be employed to control the end replicas, constraint is used for evaluating sampling efficiency of DPDS under the strict condition.

In DPDS applications, the end replicas could be handled differently to enhance the sampling efficiency. Three different ways to control the end replicas of the reaction pathway in DPDS methods were tested. (1) Controlling two ends: two end replicas in a chain-of-states setup are constrained similarly to the conventional application of the chain-of-states optimization toward the MEP. Therefore, the DPDS method will sample the reaction pathway space specifically connecting two predefined minima. (2) Controlling one end: only one end replica in a chain-of-states setup is constrained during the DPDS. In this way, the other end replica can potentially sample different space other than the initial structure. This can provide sufficient flexibility to the sampling system and allow additional minima to be identified during the simulations. (3) Releasing two ends: none of the end replicas in a chain-of-states setup is controlled by constraint during the DPDS, giving the most flexibility in terms of sampling the PES. In this way, the DPDS method can sample many stationary structures on a PES and identify transition pathways connecting these structures. In general, the DPDS method can be considered as a four-dimensional (4D) sampling method with the fourth dimension as transition pathway.

2.6. Clustering Analysis of Pathway Simulations. Similar to the clustering analysis of different conformers of flexible molecules, the transition pathways can also be subjected to clustering analysis to identify distinctive pathways. Generalized coordinates depending on the systems can be used for such clustering analysis. For a nonlinear system with \( m \) atoms, there can be \( l \) \((l \in \{1, 3m-6\})\) generalized coordinates \( \{g_1(R), g_2(R), \ldots, g_l(R)\} \) to be utilized to describe the transitions associated with conformational and structural changes. \( R \) is the set of \( 3m \) Cartesian coordinates \((x_i, i = 1, 3m)\) of all \( m \) atoms in the system. For meaningful comparison, the translational and rotational degrees of freedom should be projected out in most cases. For pathways comprising \( n \) structures with each structure described as \( i \) generalized coordinates, each structure \( i \) can be described by its \( l \) generalized coordinates \( \{g_{i1}(R), g_{i2}(R), g_{i3}(R), \ldots, g_{il}(R)\} \); \( R_i \), is the set of \( 3m \) Cartesian coordinates of structure \( i \). Therefore, a given pathway can be described by its pathway generalized coordinates matrix:

\[
\begin{bmatrix}
    g_{i1}(R) & \ldots & g_{il}(R) \\
    \vdots & \ddots & \vdots \\
    g_{n1}(R) & \ldots & g_{nl}(R)
\end{bmatrix}
\]

For multiple pathways with the same number of structures, these pathway generalized coordinate matrices can be collected and subjected to vector quantization methods, such as k-means clustering analysis. For the three testing models in this study, different generalized coordinates were used to describe their potential energy surfaces and transitions. The clustering analyses were carried out using pathway generalized coordinate matrices for each testing system.

2.7. Implementation of DPDS Method. The basic steps of the DPDS method are outlined as a flowchart in Figure 2, comprising three parts: initial pathway preparation, dynamics pathway simulation, and final analysis. The goal of the preparation part is to generate an appropriate transition path representation as a starting point for the pathway dynamics simulations. To achieve this goal, one can adapt a conventional chain-of-states scheme to generate an MEP. Following this approach, as illustrated in the flowchart, two structures representing the start and end (for example reactant and product) were carried out using pathway generalized coordinate matrices...
product) states of a transition process should be identified initially. Then a series of structures connecting these two states can be constructed through linear interpolation or other means and be subjected to the chain-of-states pathway optimization to obtain an approximate MEP.

The optimized pathway will then be subjected to the dynamics simulations outlined in part two. If there is special interest in the transition between two end replicas of the initial pathway, either restraints (such as harmonic potential) or constraints (with these two structures being treated as rigid bodies) could be applied on these two end replicas during the simulations to prevent the sampling from escaping these states. Using $k_{\text{trans}}$ and $k_{\text{ang}}$ as controlling factors, molecular dynamics simulations can be applied to the initial pathway to sample the transitions connecting the initial two end states. Various analyses, including using order parameters, $k$-means clustering, or advanced time-structure independent component analysis, can be applied to monitor the coverage of simulations and evaluate the convergence of the simulations. If exploration of the potential energy surface is desired to identify additional stationary states and their associated transition pathways, the restraints or constraints on either one or both end replicas can be removed during the simulations to allow the sampling of the different area on the potential energy surface. In addition, some adjustment of $k_{\text{trans}}$ and $k_{\text{ang}}$ can also be applied during this part of the simulations. The coverage of simulations should also be monitored to estimate convergence of the simulations.

As the final analysis in part three, the clustering analysis should be carried out on all the pathway simulations or some selected part. $k$-means clustering algorithms can be applied for this purpose. The clustering analysis will reveal all the distinctive transition pathways sampled in the simulations that connect either predefined or newly discovered stationary structures on the target potential energy surface. If necessary, further exploration or sampling can be carried out for any newly identified transition pathway.

3. RESULTS

Three systems with different complexity and levels of theory were employed to test the DPDS method.

3.1. Isomerization of Alanine Dipeptide. The first test case is the isomerization of the alanine dipeptide ($N$-acetylalanyl-$N$-methylamide). Two backbone dihedral angles ($\phi$ and $\psi$) can be used to describe the isomerization process of this molecule (Figure 3A). An MEP was generated connecting two conformers $C_{\text{eq}}$ and $C_{\text{ax}}$ corresponding to two minima on the PES (Figure 3B). The CHARMM 22 force field with CMAP backbone dihedral angle corrections was used for the calculation. No solvent molecule is present in the model system. The chain-of-states calculations of the alanine dipeptide were carried out using 25 replicas. The barrier for the isomerization with reference to conformer $C_{\text{ax}}$ is 8.74 kcal/mol. This MEP was used as the start pathway in the following DPDS simulations.

The distribution of a 1 $\mu$s conventional MD simulation of the alanine dipeptide is projected on the two-dimensional (2D) surface defined by $\phi$ and $\psi$ (Figure 3C). The sampling covers two main attraction basins $a$ and $b$ circled by dashed lines in Figure 3C (basin $b$ spreads to four corners of the plot as one single basin). The term attraction basin refers to the region with minimum free energy compared to adjacent regions. Apparently, the area between two basins was not sampled efficiently at all.

![Figure 3. Alanine dipeptide as test case: (A) alanine dipeptide structure and two dihedral angles ($\phi$ and $\psi$) as reaction coordinates; (B) initial minimum energy pathway for alanine dipeptide isomerization; (C) distribution of conventional molecular dynamics simulations of alanine dipeptide on its potential energy surface with reference to $\phi$ and $\psi$. Two main attraction basins are labeled as $a$ and $b$.](image-url)
because every time the sampling changes from one basin to the other basin, the transition process is transient without significant sampling of the pathway connecting two basins.

3.1.1. RPATH/Restraints Using $k_{\text{rms}}$ Only. To explore the effect of $k_{\text{rms}}$ on the sampling efficiency, $k_{\text{ang}}$ was set to zero in the simulations of this section. Using a chain-of-states pathway with 25 structures, with restraint of distance between adjacent structures, the sampling was enhanced in various ways. With a rather small $k_{\text{rms}}$ at 0.1 kcal/(mol·Å²), the sampling of the smaller attraction basin was significantly enhanced (Figure 4A). To save space, the unit for $k_{\text{rms}}$ (kcal/(mol·Å²)) may not be presented in the remaining of the text. Various transition regions were sampled. However, the transition region between two basins was not sampled sufficiently due to the small $k_{\text{rms}}$. With larger $k_{\text{rms}}$ of 100 and 1000, the sampling between basins including the transition from the larger basin to itself was much enhanced (Figure 4B,C). For the comparison of sampling efficiency, the PES of alanine dipeptide was divided into 40,000 grid squares by dividing both $\phi$ and $\psi$ into 200 bins. The coverage of the surface was estimated based on the distribution of sampling trajectories among grid squares and plotted in Figure 4D. With $k_{\text{rms}}$ ranging between 0.1 and 10000, the coverage of the PES varies significantly with $k_{\text{rms}}$ = 10 showing the most efficient sampling (magenta line in Figure 4D). The samplings of $k_{\text{rms}}$ as 1, 10, and 10000 did not show significant improvement of coverage on the PES and are listed in Table S1 in the Supporting Information. It should be noted that the two end replicas are constrained during the direct pathway simulations described here and in sections 3.1.2 and 3.1.3.

3.1.2. RPATH/Restraints Using $k_{\text{ang}}$. Another controlling factor, $k_{\text{ang}}$, is also applied. To test effectiveness of $k_{\text{ang}}$ a rather small $k_{\text{rms}}$ of 0.1 kcal/(mol·Å²) is used. The most effective $k_{\text{ang}}$ values are 10, 100, and 1000 kcal/mol (Figure 5). To save space, the unit for $k_{\text{ang}}$ (kcal/mol) may not be presented in the remainder of the text. Interestingly, different $k_{\text{ang}}$ values led to the significant enhancement of sampling in different transition regions. With a $k_{\text{ang}}$ value of 10 (Figure 5A), the transition path region connecting the same attraction basin $b$ was extensively sampled (circled by blue dashed line), and a novel transition path region connecting two main attraction basins $a$ and $b$ was also revealed (circled by red dashed line). With a $k_{\text{ang}}$ value of 100, the transition path region covering the original MEP of alanine dipeptide as the starting pathway of DPDS was extensively sampled (Figure 5B, circled by red dashed line). With a larger $k_{\text{ang}}$ value of 1000, a second transition path region parallel to the starting pathway for DPDS (Figure 5B) was reached and extensively sampled (Figure 5C, circled by a red dashed line). Using $k_{\text{ang}}$ as the other controlling factor, different valleys for transition pathways on the PES could be reached and sampled extensively using the DPDS method. The samplings of $k_{\text{ang}}$ as 0.1 and 1 and did not show significant improvement of coverage on the PES and are listed in Table S1 in the Supporting Information.

3.1.3. RPATH/Restraints Using $k_{\text{rms}}$ and $k_{\text{ang}}$. Based on the outcome of simulations testing $k_{\text{rms}}$ and $k_{\text{ang}}$ individually, different combinations of $k_{\text{rms}}$ and $k_{\text{ang}}$ were used for alanine dipeptide simulations to gain a better understanding of the efficiency of the DPDS method. With $k_{\text{rms}}$ as 10 and $k_{\text{ang}}$ as 10, both transition path regions connecting attraction basin $b$ to itself and the one connecting basins $a$ and $b$ were extensively sampled (Figure 6A, B).
circled by dashed lines). With $k_{\text{rms}}$ as 10 and $k_{\text{ang}}$ as 100, the transition path region covering the original MEP as the starting point of the simulation was extensively sampled (Figure 6B, circled by a red dashed line), similar to the previous simulation with the same $k_{\text{rms}}$ but $k_{\text{rms}}$ as 0.1 (Figure 5B). However, different from that case, the transition path region connecting the attraction basin $b$ to itself was also extensively sampled with $k_{\text{rms}}$ as 10 (Figure 6B, circled by a blue dashed line). Strikingly, with $k_{\text{rms}}$ as 10 and $k_{\text{ang}}$ as 1000, two transition pathway regions parallel to each other and connecting attraction basins $a$ and $b$ were both extensively sampled (Figure 6C, circled by dashed line). This observation demonstrates DPDS as a powerful method to explore and sample multiple transition pathway regions connecting two states. Another interesting observation of the sampling efficiency is that the PES coverage of the simulation with $k_{\text{ang}}$ as 100 (cyan line in Figure 6D) or 1000 (magenta line in Figure 6D) have dramatic increases around 10 and 16 ns, respectively. These sudden increases of coverage occurred because pathway sampling switched to a different region on the PES. To illustrate these jumps in the DPDS simulation from one transition pathway region to a different region, the distributions before and after the jumping are illustrated in Figures 7 and 8 for these two simulations. The samplings with $k_{\text{ang}}$ as 0.1 and 1 did not show significant improvement of the coverage on the PES and are listed in Table S1 in the Supporting Information.

### 3.1.4. RPATH/Restraints with Removing Constraints on Either One or Both Ends

For many systems, not only are transition pathways unknown but also the minimum on the PES or stable states are also unknown. It will be beneficial if DPDS could help to search new minima while directly sampling the transition pathway space. To use the DPDS method as a potential tool to explore the PES, the constraints on either one or both of the end replicas could be removed to allow the sampling of the different states on the PES.

First, the constraint on replica number 25 (A25) was removed. Replica A25 is a minimum energy structure in basin $b$ and A1 is a minimum energy structure in basin $a$ (Figure 3B). The coverage and sampling efficiency with $k_{\text{rms}}$ as 0.1 and $k_{\text{ang}}$ as 10 are similar to the simulation with constraints on both end replicas but with higher coverage (34.5%; Figure 9A). However, with $k_{\text{rms}}$ as 10 and $k_{\text{ang}}$ as 10, DPDS did not show improvement in either minimum sampling or PES coverage (Figure 9B). With $k_{\text{rms}}$ as 10 and $k_{\text{ang}}$ as 100, transition pathway regions that connect basin $b$ to itself or basins $a$ and $b$ were sampled extensively (Figure 9C). Interestingly, a new transition pathway region connecting attraction basins $a$ and $b$ through the top right corner of the PES was sampled with $k_{\text{rms}}$ as 0.1 and $k_{\text{ang}}$ as 1000 simulation (Figure 9D, circled by red dashed line). Various combinations of $k_{\text{rms}}$ and $k_{\text{ang}}$ were also tested with constraint only on A25 with similar sampling coverage, and are listed in Table S1 in the Supporting Information.

Second, the constraint on A1 was removed for alanine dipeptide DPDS. Therefore, only A25 was constrained. The coverage and sampling efficiency with constrained A25 combination of $k_{\text{rms}}$ as 0.1 and $k_{\text{ang}}$ as 10 (Figure 10A) and $k_{\text{rms}}$ as 10 and $k_{\text{ang}}$ as 10 (Figure 10B) are similar to the simulation with constraint only on A1 (Figure 9A,B). With $k_{\text{rms}}$ as 10 and $k_{\text{ang}}$ as 100, a new sampling pattern on the PES was observed (Figure 10C). However, with strong $k_{\text{rms}}$ (1000), the DPDS simulation could be trapped in a local attraction basin when the constraint on the A1 replica is removed, ultimately leading to less efficient sampling (Figure 10D). Various combinations of $k_{\text{rms}}$ and $k_{\text{ang}}$ were also tested with constraint only on A1 with similar sampling coverage, and are listed in Table S1 of the Supporting Information.

Third, the constraints on both end replicas were removed for alanine dipeptide DPDS. The coverage and sampling efficiency with a combination of $k_{\text{rms}}$ as 0.1 and $k_{\text{ang}}$ as 10 (Figure 11A) are similar to the previous simulations without constraint on either end replica, indicating that a weak $k_{\text{rms}}$ does not pull either end replica out of attraction basins after removing the constraint forces. With a combination of $k_{\text{rms}}$ as 10 and $k_{\text{ang}}$ as 100, both minimum states and PES coverage are improved (Figure 11B). Interestingly, with $k_{\text{rms}}$ as 10 and $k_{\text{ang}}$ as 100, basin $a$ is much less sampled compared to other samplings (Figure 11C).
because the pathway was trapped within basin $b$ during this simulation.

The above simulations show that, with a proper setup, the sampling efficiency could be much improved. New transition pathways could be discovered and sampled for further analysis.

3.1.5. NEB Results. Another popular chain-of-states method, the NEB method, was also applied within the DPDS framework. Both $k_{\text{rms}}$ and $k_{\text{ang}}$ were applied to control the sampling efficiency. With $k_{\text{rms}} = 0.1$ and $k_{\text{ang}} = 100$, the sampling showed the most effective results. The sampling with constraints on both ends was the most effective in finding the minimum and covering the PES (Figure 12A). Two different transition pathway regions connecting two attraction basins were detected and extensively sampled. With constraint on either end replicas, the simulations are similar to each other. The transition pathway region connecting the attraction basin $b$ to itself was detected and extensively sampled in both simulations with constraints on either end replica (Figure 12B,C). In the simulations without constraints on either end replica, all the transition pathway regions detected in the above three simulations were also detected and extensively sampled (Figure 12D). All other DPDS results using the NEB method are illustrated in Table S2 in the Supporting Information.

3.1.6. Distinguished Pathways Identified through DPDS. All the DPDS simulations carried out for the alanine dipeptide were combined and subjected to clustering analysis to identify distinguished transition pathways. The clustering analysis was carried out using a density-based spatial clustering of application with noise (DBSCAN) algorithm. For each pathway snapshot containing 25 replicas sampled using the DPDS method, two dihedral angles of alanine dipeptide in each replica leading to a total of 50 dihedral angles for each pathway were used for the clustering analysis. A cutoff of 200° was chosen for the ideal results, in which a total of 17 clusters were generated. Among these 17 clusters, nine clusters representing unique transition pathways were identified and are listed in Figure 13. All 17 clusters are presented in Table S3 in the Supporting Information. Pathway 1 represents transitions from attraction basin $a$ to $b$. It also has a loop structure within basin $b$ (Figure 13A). Pathway 2 has a shape similar to that of pathway 1, but with significant shifting toward attraction basin $a$ (Figure 13B). In addition to the transition pathways connecting attraction basins $a$ and $b$, pathway 3 presents another transition pathway that comes from basin $b$ going through basin $a$ before going back to basin $b$ (Figure 13C). Pathway 4 represents a straight transition pathway going through both attraction basins, with a loop within attraction basin $b$ (Figure 13D). Pathway 5 represents two new transition pathways; one connects two attraction basins, and the other one connects attraction basin $b$ to itself (Figure 13E). The major part of pathway 6 resides in attraction basin $b$ with the remaining part going through attraction basin $a$ (Figure 13F). However, the transition between the two attraction basins is actually a false pathway, because it goes through an energy peak by having two adjacent replicas residing on the two sides of the peak with no replica in the high-energy area. This resulted from the simulations with small $k_{\text{rms}}$. Pathway 7 resides in attraction basin $b$ with a shape similar to that of pathway 1, but with significant shifting toward attraction basin $a$.
basin \( b \) and with the transition pathway connecting basin \( b \) itself (Figure 13G). Pathway 8 (Figure 13H) represents a transition pathway connecting basin \( b \), which is also represented in pathways 5, 6, and 7, and the transition pathway connecting two attraction basins, also seen in pathways 1 and 3. Pathway 9 (Figure 13I) represents an alternative pathway connecting attraction basins \( a \) and \( b \), which is similar to the one represented in pathway 4. The aggregation of replicas was due to the small \( k_{\text{rms}} \) value used to carry out simulations that contribute significantly to this cluster.

The clustering analyses demonstrate the effectiveness of using the DPDS method as a tool to sample multiple pathways on the PES. With different combinations of controlling factors \( k_{\text{rms}} \) and \( k_{\text{ang}} \) as well as constraints on the end replicas, the DPDS could automatically sample multiple pathways, which provides valuable information for complicated processes. Careful analyses are also essential to obtain information about unique reaction pathways sampled in the simulations.

### 3.2. SCC-DFTB Reaction Pathway Sampling

The second test case is the intramolecular condensation reaction of \( \beta \)-alanine, a simple organic reaction (Scheme 1). Despite its apparent chemical simplicity, this reaction presents a real challenge for pathway sampling with two chemical bonds forming and two chemical bonds breaking simultaneously. The MEP connecting the reactant and product was constructed using the intrinsic reaction coordinate (IRC) method. Its distribution on the PES is illustrated in Figure 14A. The lengths of two key bonds, carbon–oxygen (C1–O3) and carbon–nitrogen (C1–N11), are used as reaction coordinates to construct the PES. The actual transition state identified from a separate quantum mechanical calculation was also plotted as a red dot along the MEP. The MEPs connecting reactant and product were also constructed using RPATH, the chain-of-states method implemented in

![Figure 13](image)

Figure 13. Unique transition pathways identified through clustering analysis of alanine dipeptide DPDS simulations. Nine among a total of 17 pathways are listed. All 17 pathways are illustrated in Table S3 in the Supporting Information.

#### Scheme 1

![Scheme 1](image)

Figure 14. Minimum energy pathways of intramolecular condensation of \( \beta \)-alanine reaction: (A) using intrinsic reaction coordinates; (B) using RPATH method in CHARMM \( (k_{\text{rms}} = 1000, k_{\text{ang}} = 100) \); (C) using RPATH method in CHARMM \( (k_{\text{rms}} = 10000, k_{\text{ang}} = 1000) \).
CHARMM (Figure 14B,C). The energy barrier from the IRC calculation is around 47.268 kcal/mol for this reaction.

3.2.1. DPDS with Combining $k_{rms}$ and $k_{avg}$. Using chain-of-states with 20 structures along the pathway, different combinations of $k_{rms}$ and $k_{avg}$ were used to carry out DPDS for this reaction. Self-consistent charge density functional tight binding (SCC-DFTB) was used to reproduce energetic results similar to those obtained from B3LYP/6-31+G(d,p) level of theory. Therefore, the SCC-DFTB method was applied for all DPDS simulations of this reaction. DFTB3 mparameters with third order correction were used for the calculation. No dispersion correction was employed in the current study. The Anderson mixing scheme was used in SCC iterations. The convergence criterion for the SCF cycle was set to be 10^-7. For each combination of $k_{rms}$ and $k_{avg}$ values, DPDS were carried out at 300 K for 1 ns using a 1 fs time step.

First, the constraints were applied on both end replicas during the simulations. The simulations demonstrated that large $k_{rms}$ and particularly large $k_{avg}$ values were necessary to enhance the sampling along the pathway, especially the region close to the transition state region (Figure 15). With $k_{avg}$ smaller than 100, no apparent enhanced sampling was observed. With larger $k_{rms}$ and $k_{avg}$ the sampling along the pathway was much enhanced (Figure 15A–C). The most effective combination to sample the transition pathways of this system is $k_{rms}=10000$ and $k_{avg}=1000$ (Figure 15D). The DPDS is evenly distributed along the pathway with extensive sampling throughout the pathway connecting two minimum states.

3.2.2. DPDS with Removing Constraint on Either One or Both Ends. When the constraint on A20 (product; see Figure 14) was removed leaving constraint only on A1, with $k_{rms}=100$ and $k_{avg}=1000$, the sampling in the product region was extended toward large C1–O3 distances significantly (Figure 16A). The reason for this is due to the escaping of the water molecule as a product away from the β-lactam ring product. With $k_{rms}=1000$ and $k_{avg}=1000$, the sampling is better controlled with enhanced sampling along the pathway (Figure 16B).

When the constraint on A1 (reactant; see Figure 14) was removed leaving constraint only on A20, the sampling in the reactant region was much extended (Figure 16C). With combinations of large $k_{rms}$ and $k_{avg}$ ($k_{rms}=1000$ and $k_{avg}=1000$) and ($k_{rms}=10000$ and $k_{avg}=1000$), the sampling along the pathway was also enhanced (Figure 16C,D). When the constraints on both A1 and A20 were removed, the sampling in the reactant region was also much extended with a combination of large $k_{rms}=1000$ and $k_{avg}=1000$ (Figure 16E). With $k_{rms}=10000$ and $k_{avg}=1000$, the sampling along the pathway is much enhanced without constraints on either end replicas (Figure 16F).

3.3. β-Hairpin. As a third example, a more complex system of β-hairpin peptide folding was used to test sampling efficiency of DPDS method. The sequence of the peptide is GEWTYD-DATKTFTVTE. The fold structure was obtained from a crystallographic structure available from the protein data bank (PDB code 3GB1, residues 41–56). An extended structure of this peptide was generated using CHARMM. To better evaluate and compare the sampling efficiency, the overall distribution of peptide conformation was generated based on all the obtained sampling and used as a background for distribution plots in this section. To build an overall distribution, a Markov state method, MSMbuilder, was used to calculate dihedral angles along the backbone of the peptide for each structure. Then, time-structure independent component analysis (tICA) implemented in MSMBuilder was applied on all simulations of the peptide. The first two dominant components tICA1 and tICA2 were used to plot the overall distribution of β-hairpin simulations. A k-means clustering method was used to cluster the overall distribution into 11 clusters (Figure 17). Based on the nature of these clusters, we divided those clusters into three groups: β-hairpin (fold region), unfold region, and misfold region (Figure 17).

Two conventional MD simulations starting from either the fold structure or an extended structure were carried out for 1 μs. The simulation starting from the extended structure was trapped in cluster 4 with presence in clusters 2 and 6 in the misfold region (Figure 18A). On the other hand, the simulation starting from the fold structure was mainly distributed in clusters 3 and 9 in both the fold and unfold regions (Figure 18B).

An MEP was generated with 25 replicas connecting the extended and fold conformations (Figure 18C). The DPDS was carried out starting from this MEP with different $k_{rms}$ and $k_{avg}$ combinations and two end replicas being constrained. The distribution of each simulation is illustrated in Table S4 in the Supporting Information. Clustering analysis using k-means clustering algorithm was carried out, and divided simulations with the same control of end replicas together into 10 clusters. For the DPDS with two end replicas being constrained, four out of 10 pathways (PA1 to PA4) are illustrated in Figure 19. Pathway PA1 demonstrates a smooth transition through a series of unfold clusters and fold clusters (5, 7, 10, 11, 9, and 1) (Figure 19A). Another slightly different pathway PA2 goes through clusters 2 and 8 instead of cluster 7 before cluster 3 (5, 2, 8, 3, 11, 9, and 1) (Figure 19B). Pathway PA3 actually starts in cluster 2 with extensive sampling and goes through a misfold cluster 6 (2, 6, 8, 3, 9, and 1; Figure 19C). These three clusters represent smooth transition channels between two end states through different clusters. Pathway PA4 represents a stochastic transition between two end states. With switching back and forth, this pathway samples clusters 5, 2, 6, 8, 7, 10, 11, 3, 9, and 1 (Figure 19D). Several other pathways are similar to pathway PA4 (Table S5 in Supporting Information). The uneven distribution shown in pathway PA4 could result from a weak $k_{avg}$ that contributes to this cluster.

![Figure 15. Simulation of intramolecular condensation of β-alanine reaction using DPDS method. Constraints were applied on both end replicas: (A) $k_{rms}=1000$, $k_{avg}=100$; (B) $k_{rms}=10000$, $k_{avg}=100$; (C) $k_{rms}=1000$, $k_{avg}=1000$; (D) $k_{rms}=10000$, $k_{avg}=1000$.](image-url)
Four representative pathways (PB1 to PB4) are illustrated in Figure 20 among 10 clusters generated from simulations without constraining the fold end replica (labeled as A25). In pathway PB1, end replica A25 remains in cluster 1. This pathway represents transition through clusters 5, 2, 6, 8, 3, 9, and 1 (Figure 20A). Pathway PB2 (Figure 20B) is similar to pathway PA2 (Figure 19B) with end replica A25 remaining in cluster 1. In pathway PB3, end replica 25 migrates to cluster 9. Pathway PB3 has uneven distribution and goes through clusters 5, 2, 7, 10, 11, 9, and 1 and ends within cluster 9 (Figure 20C). End replica 25 also migrates to cluster 9 in pathway PB4, which goes through clusters 5, 2, 10, 7, 8, 6, 3, 11, 1, and 9 in a zigzag pattern (Figure 20D).

When the constraint on replica 1 is removed, DPDS displays much more diversity than the above samplings. Pathway PC1 (Figure 21A) starts with replica 1 in cluster 2 as a misfold structure and goes through clusters 8, 3, 11, 9, and 1 and ends within cluster 9 (Figure 20C). End replica 25 also migrates to cluster 9 in pathway PB4, which goes through clusters 5, 2, 10, 7, 8, 6, 3, 11, 1, and 9 in a zigzag pattern (Figure 20D).

When the constraint on replica 1 is removed, DPDS displays much more diversity than the above samplings. Pathway PC1 (Figure 21A) starts with replica 1 in cluster 2 as a misfold structure and goes through clusters 8, 3, 11, 9, and 1, similar to part of pathway PA2 (Figure 19B). Pathway PC2 (Figure 21B) has replica 1 remaining in cluster 5 and half of it similar to pathway PC1. Pathway PC3 (Figure 21C) starts with replica 1 in cluster 2, goes through clusters 5 and 6, and leads to a second half similar to pathway PC1. In pathways PC4, PC5, and PC6, replica 1 samples different part of clusters 6 and 4 as misfold regions (Figure 21D–F). All three pathways have zigzag patterns, and go through the unfold region before reaching the fold region.

When constraints on both end replicas were removed, DPDS did not present much more diverse pathways (Figure 22).
Pathway PD1 (Figure 22A) is similar to previous pathways, such as PB2. Pathway PD2 (Figure 22B) is very similar to pathway PC1. Pathway PD3 (Figure 22C) starts in cluster 2 and goes through cluster 6, both in the misfold region, before going through the unfold region and reaching the fold region. The end replica 25 in pathway PD3 is rather close to cluster 3 in the unfold region. Pathways PD4, PD5, and PD6 are zigzag pathways with replica 1 heavily sampling the misfold region (Figure 22D–F).

From the sampling, it seems to be extremely unlikely that misfold states of this β-hairpin can directly change to the fold states. All the simulations presented in this study indicate that the misfold states need to unfold first before reaching the fold states.

3.4. Comparison with the Transition Path Sampling Method

For comparison, another pathway sampling method, the TPS method, implemented in CHARMM was applied on alanine dipeptide isomerization and β-hairpin peptide folding. To apply the TPS method, two attraction basins and an appropriate start geometry for successful shooting trajectories are required. In TPS method, new transition pathways are generated from old pathways through shooting and shifting algorithms. In the shooting algorithm, a random frame in the old pathway is selected and modified by adding a random momentum perturbation. MD simulations are carried out from the selected frame along the new velocities in both forward and backward directions in time to generate a new transition pathway. In the shifting algorithm, a random frame and its atomic velocities in the old pathway are selected. MD simulations are carried out from the selected frame and velocities in either forward or backward directions in time to generate a new transition pathway with the same length as the old one, effectively shifting the old pathway either forward or backward in time. The shifting algorithm is complementary to the shooting algorithm. Although the new pathways overlap with the old pathway, shifting moves will improve the convergence of transition path sampling. Successful shooting trajectories are those simulations starting from the start geometry that connects two attraction basins. Normally structures similar to the transition state could serve as the start geometry leading to successful shooting trajectories.

For alanine dipeptide, two attraction basins a and b are defined based on the dihedral angles used as reaction coordinates. To ensure sufficient TPS sampling, these two basins need to be defined relatively large (for a, ϕ(40°,100°) and ψ(−150°,25°), blue rectangle in Figure 23A; for b, ϕ(−170°,−50°) and ψ(−50°,170°), red rectangle in Figure 23A). After some initial tests, an appropriate start geometry of alanine dipeptide was identified. Starting from this initial geometry, a total of 1,000,000 shooting pathways each as 2 ps long were carried out. The shooting and shifting movement ratio was chosen as 9:1 for optimal sampling efficiency. The acceptance rate is about 9% for shooting movement and 90% for pathway shifting movement. The TPS simulations of alanine dipeptide lead to good coverage of the potential energy surface with 72% coverage (Figure 23A). Although most major transition pathways connecting basins a and b were sampled, one transition pathway between two basins was not detected by these otherwise comprehensive simulations.

For the β-hairpin, the attraction basins were defined using two structural order parameters, one as distance d between α carbons of two terminal residues (Gly1 and Glu16) and the other one as angle θ defined by Gly1, Ala8, and Glu16 α carbons (with Ala8 as the vertex). The two attraction basins were defined as fold
Although there are different means to define attraction basins more relevant to folding, the above order parameters were chosen for the purpose of interpretation and implementation. The replica number 13 from the minimum energy pathway of $\beta$-hairpin folding was chosen as the initial geometry for TPS simulations. Starting from this geometry, a total of 1,000,000 shooting pathways each as 10 ps long were carried out. Longer trajectories are necessary for shooting simulations to reach both attraction basins. The shooting and pathway shifting movement ratio was chosen as 9:1 for optimal sampling efficiency. Similar to the alanine dipeptide case, the acceptance rate is about 9% for shooting movement and 90% for pathway shifting movement. The TPS simulations mainly cover unfold regions, part of the misfold region, and a very small part of the fold region (Figure 23B). The lack of coverage on the folded structures is due to the quick termination of the shooting trajectories when reaching an attraction basin representing folded structures.
In the above TPS simulations, the shooting trajectories can be up to a certain length (2 ps for alanine dipeptide and 10 ps for β-hairpin), but the length of each trajectory varies significantly. Therefore, it is not convenient to carry out the pathway clustering analysis. The plots of the TPS simulations on the potential energy surfaces of these two systems could be compared with the DPDS. For the simple alanine dipeptide case, the TPS simulations provided sufficient coverage on the PES, but missed a key transition pathway. For the β-hairpin, the TPS simulations only cover a portion of the regions sampled by the DPDS simulations.

4. DISCUSSION

In this study, we developed and tested the DPDS method as a new approach for direct sampling in the transition pathway space. Through the analyses of three test cases, the DPDS method demonstrates certain advantages. The chain-of-states setup of the method ensures the sampling of transition pathway on the PES of the target systems. This could greatly enhance the sampling efficiency when the main goal of sampling is exploring feasible transition pathways connecting multiple minimum states. The DPDS method is compatible with most chain-of-states methods, such as nudged elastic band and string methods. Another advantage of the DPDS method is that minimum knowledge about the pathways on the target PES is required. In addition, the dynamical sampling increases the probability of the sampling to reach high-energy barrier transition pathways even starting with the global minimum energy pathway. By removing constraints or restraints on either or both end replicas, the DPDS could reach any transition pathway on a PES.

However, it should be emphasized that, like all other enhanced sampling methods, it is not guaranteed that DPDS can exhaust all the feasible pathways even with long simulation time. Therefore, the use of parameters $k_{rms}$ and $k_{ang}$ provides additional control of DPDS to reach different transition pathways. With small $k_{rms}$ and $k_{ang}$ the sampling will favor low-energy space. By increasing $k_{rms}$ one could obtain the sampling more faithful to the actual transition pathways. On the other hand, by increasing $k_{ang}$ the DPDS is more likely to switch to and detect multiple transition pathways. When testing the different combinations of $k_{rms}$ and $k_{ang}$ the DPDS of alanine dipeptide isomerization exhausts all the major transition pathways identified on its PES. Analyzing DPDS results to obtain transition pathway information is relatively easy, because each snapshot of the sampling is a transition pathway itself. Clustering analysis of DPDS results will directly lead to multiple transition pathways connecting multiple minima on the targeting PES. However, one should be cautious about applying high $k_{ang}$ in DPDS, because high $k_{ang}$ will force smoothing of the transition pathway and lead to high-energy barriers.

One may be concerned about the dependence of the DPDS sampling on the initial pathway. Therefore, an MEP (Figure 24A) different from the one used in the DPDS simulation of alanine dipeptide was employed as the initial pathway and subjected to three DPDS simulations using optimal combinations of $k_{rms}$ and $k_{ang}$ parameters. The simulation with $k_{rms} = 10$ and $k_{ang} = 10$ (Figure 24B) starting from this MEP is similar to the simulation starting from the original MEP with the same $k_{rms}$ and $k_{ang}$ (Figure 6A). The simulation with $k_{rms} = 10$ and $k_{ang} = 100$ does sample the transition region between attraction basins a and b (circled region in Figure 24C) where the original MEP is located. Although the simulation with $k_{rms} = 10$ and $k_{ang} = 1000$ presents unique coverage on the PES (Figure 24D), the pathways sampled in this simulation were also sampled in the DPDS with constraint on replica A (Figure 8C,D). This suggests that the dependence of DPDS on the start pathway is not significant.

Because of the convenience of the setup, DPDS can be easily applied with QM/MM methods to sample chemical reaction pathways. The sampling of intramolecular condensation reaction of β-alanine using the DPDS method indicates that high $k_{rms}$ and $k_{ang}$ are necessary for sufficient sampling of chemical reactions. This is due to the inherent high-energy barrier of chemical reactions, which requires large force to maintain even distribution along the pathway. Another popular sampling method along the chemical reaction pathway is the umbrella sampling. In an umbrella sampling method, harmonic potentials using order parameters or collective variables are implemented for sampling windows to force the simulations of certain regions. Using DPDS method, one has the flexibility through different weighting factors to control the RMSD distance between replicas. The umbrella sampling is often combined with the weighted histogram analysis method (WHAM) to obtain free energy information along the sampling windows. By obtaining enhanced sampling along multiple pathways, the DPDS method can also be used to estimate free energy information on the targeting PES. This is under development for future publication.

The DPDS method could be the most effective to search for transition pathways of a complicated system as demonstrated by the β-hairpin peptide in this study. Many complicated biomolecular processes, such as protein folding, ligand binding, and protein–protein binding, do not have convenient order parameters or collective variables to describe the processes as transition pathways. This could be resolved by the DPDS method through sampling transition pathways connecting two end states, which can be unfold/fold proteins or protein unbound/bound with its binding partners. Without defining specific order parameters, DPDS could potentially detect the most probable transition pathways connecting two end states and explore multiple transition pathways. Again, different combinations of $k_{rms}$ and $k_{ang}$ can be an effective means to drive sampling among different transition pathways.

Based on the flexibility from the combinations of $k_{rms}$ and $k_{ang}$, DPDS simulations could be carried out with different emphases.
On one hand, with low $k_{\text{rms}}$ and $k_{\text{ang}}$ values, the simulations will sample the attraction basins with moderate increasing coverage of the transition region as shown in Figure 4. These simulations could be utilized to detect and sample attraction basins. On the other hand, larger $k_{\text{rms}}$ and $k_{\text{ang}}$ values will enhance the detection and sampling of transition pathway regions as shown in Figures 5 and 6. For complex systems, as demonstrated by the $\beta$-hairpin in this study, the DPDS simulations with different combinations of $k_{\text{rms}}$ and $k_{\text{ang}}$ are useful to explore the potential energy surface closely related to the transition processes of interest. These simulations will provide not only the information about potential attraction basins but the transition pathways connecting these basins as well. Although transition pathways sampled in DPDS simulations are not minimum energy pathways, the representative pathways generated from clustering analyses can be subjected to further optimization to obtain minimum energy pathways.

Although the TPS method could be powerful to sample the transition pathway space for a simple system as demonstrated by alanine dipeptide, the efficient application of this method for complex systems is challenging and requires the careful setup of the simulations. In the $\beta$-hairpin peptide case, the TPS simulations only sampled a portion of the overall sampling space from the DPDS simulations. This is because TPS is limited to two predefined attraction basins, which are used as criteria to terminate the simulations. Although the shifting movement of TPS could help to enhance the sampling efficiency, these termination criteria still limit the sampling efficiency of new minima or attraction basins. The DPDS simulations, especially when removing restraints or constraints on both or either end replicas, could greatly enhance the detection and the sampling of new minimum or attraction basins for complex systems.

There is also a benefit for the computational cost of DPDS method in terms of sampling the high-energy regions along transition pathways. Comparing to the TPS method, it is not necessary to choose an appropriate initial geometry a priori for shooting simulations in DPDS method. Because of the chain-of-states framework, the “acceptance” rate for transition pathways in DPDS simulations is 100% compared to the 9% acceptance rate of shooting trajectories in the TPS simulations for the two test cases in this study, even after our best effort to fine-tune the TPS simulations. It should be noted that our experience of using the TPS method is limited, and it is likely that a better setup could lead to a higher acceptance rate. Nevertheless, the setup for high sampling efficiency of transition pathways could prevent effective application of the TPS method especially for complex systems. In general, the convenience of the setup within the chain-of-states framework in DPDS will help to generate sufficient coverage of transition pathway regions and to explore the potential energy surface for additional attraction basins.

Similar to many other enhanced simulation methods, there is no foolproof way to guarantee the convergence of the DPDS simulations. However, one still could have a good idea about the converging trend of the simulations. To evaluate the convergence of DPDS simulations, one of the key factors is selecting appropriate generalized coordinates for distribution plots. This would be relatively easy for simple cases, such as the alanine dipeptide isomerization and the $\beta$-alanine intramolecular condensation reaction tested in this study. For complex systems with high degrees of freedom, appropriate dimension reduction will be necessary and critical. It was shown in this and several other studies that time–structure independent component analysis can be very effective to construct generalized coordinates to be used to plot the distribution of dynamics simulations. The distribution of accumulated DPDS simulations projected on the suitable generalized coordinates will be an effective tool for not only monitoring the convergence of the simulations but also identifying unique transition pathways through clustering analysis.

5. CONCLUSION

In this study, we developed a direct pathway dynamics sampling (DPDS) method for efficient sampling of a potential energy surface and exploring transition pathways. Two parameters $k_{\text{rms}}$ and $k_{\text{ang}}$ could be implemented for effective controlling of pathway sampling. Sampling with small $k_{\text{rms}}$ and $k_{\text{ang}}$ will favor a low-energy space, similar to conventional molecular dynamics sampling. Higher $k_{\text{rms}}$ will lead to enhanced sampling along the pathway with decreasing distances among replicas. Higher $k_{\text{ang}}$ will smoothen the pathway and increase the likelihood for the simulation to switch to new transition pathways. Using different combinations of $k_{\text{rms}}$ and $k_{\text{ang}}$ the DPDS method can sample the majority of the PES important for transitions and detect multiple pathways, which would not be easily obtained and analyzed otherwise. The convenience of setup and analysis of DPDS results for transition pathway sampling related to biomolecules, such as proteins. Using RMSD combining with weight factors, complex processes such as $\beta$-hairpin peptide folding were efficiently sampled and multiple folding pathways were identified. In summary, the DPDS method provides a simple and effective means to directly sample transition pathways for complex systems and can be easily combined with various levels of theory.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b00606.

Distributions of DPDS simulations and clustering analysis of alanine dipeptide isomerization and $\beta$-hairpin peptide folding (PDF)

AUTHOR INFORMATION

Corresponding Author

*Email: ptao@smu.edu.

ORCID

Peng Tao: 0000-0002-2488-0239

Funding

This work was partially supported by the Edward R. Biehl Graduate Fellowship (H.Z.), SMU Dean’s Research Council, and American Chemical Society Petroleum Research Fund (Grant 57521-DNI6).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Computational time was provided by Southern Methodist University’s Center for Scientific Computation and Texas Advanced Computing Center (TACC) at the University of Texas at Austin. We thank Dr. Shuanghong (Sharon) Hou for critical reading of the manuscript and fruitful discussions and Dr. Alex Lippert for help with the manuscript preparation.


PAPER IV

Recognition of Protein Allosteric States and Residues: Machine Learning Approaches

Hongyu Zhou, Zheng Dong, and Peng Tao

Allostery is a process by which proteins transmit the effect of perturbation at one site to a distal functional site upon certain perturbation. As an intrinsically global effect of protein dynamics, it is difficult to associate protein allostery with individual residues, hindering effective selection of key residues for mutagenesis studies. The machine learning models including decision tree (DT) and artificial neural network (ANN) models were applied to develop classification model for a cell signaling allosteric protein with two states showing extremely similar tertiary structures in both crystallographic structures and molecular dynamics simulations. Both DT and ANN models were developed with 75% and 80% of predicting accuracy, respectively. Good agreement between machine learning models and previous experimental as well as computational studies of the same protein validates this approach as an alternative way to analyze protein dynamics simulations and allostery. In addition, the difference of distributions of key features in two allosteric states also underlies the population shift hypothesis of dynamics-driven allostery model. © 2018 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.25218

Introduction

Allostery, which is referred to as a process by which proteins transmit the effect of perturbation at one site to a distal functional site, is fundamental to many biological regulations. Numerous studies have been conducted in the past half centuries. In the early 60s, two theoretical models, Monod–Wyman–Changeux (MWC) and Koshland–Némethy–Filmer (KNF) models, were proposed to explain significant conformational change observed in protein hemoglobin upon binding with oxygen molecules as concerted or sequential processes, respectively. Since then, protein allostery was commonly considered as the significant conformational change observed in protein structure upon local perturbation. However, there are many allosteric proteins being identified without significant conformational change upon perturbation. In contrast to the conformation-driven allostery observed in hemoglobin, new theoretical models were proposed as dynamics-driven allostery or population shift among different states to explain protein allostery without significant conformational changes. In these models, it was proposed that the external perturbations cause significant changes in the distribution of protein in different states, and lead to the change of free energy landscape related to protein allosteric functions. Various studies were carried out to distinguish different states through simulations using principal component analysis based on the cross correlation matrix of protein simulations. Despite the progress made in these studies, further development is still necessary for better recognition of the different states of dynamics-driven allosteric proteins.

Identifying allostery-related residues and the pathways responsible for allosteric transformation is another challenge for the protein allostery studies. The theory for allosteric information transduction within the proteins has evolved from single pathway formed by residues into allosteric information transduction network model. Numerous methods for identifying key allosteric residues from simulations have been developed recently. These computational methods focus on correlation analysis related to protein dynamics. Potential contribution from simple geometric parameters, such as distances between residues or dihedral angles to allostery, has not been explored extensively.

In computer science, machine learning (ML) methods were developed for many purpose including pattern classification. Due to their various advantages, ML methods have also been applied in computational biology. Many ML methods are specialized in classification with high accuracy, and can also provide insights into the intrinsic differences in classification model. Therefore, ML methods are applied in this study to develop classification model with regard to protein allostery. Specifically, two widely applied ML methods, neural networks and decision tree models, are used to analyze geometric parameters including distances among residues and backbone dihedral angles, and develop prediction models to differentiate states of dynamics-driven allosteric proteins.

Neural network, also named as artificial neural network, was first proposed in the 1960s to mimic the biological neural networks in animal brains. Recently, being developed as deep learning methods, the artificial neural network model has been widely used in many applications, including artificial intelligence and image recognition. Since its initial application...
in computational chemistry in the 1990s,\textsuperscript{[29,30]} artificial neural network model has been applied in rational drug design.\textsuperscript{[31–33]} Being a nonlinear activation function, artificial neural network method is particularly suitable for modeling nonlinear relationships.\textsuperscript{[34]}

Decision tree model, as another ML method, is widely used to identify key factors that contribute the most to the target states. In general, decision tree model is easy to apply on large amount of data with high dimensions. The resulted classification model based on the decision tree method is also easy to interpret related to the nature of the systems being studied.\textsuperscript{[35]} Due to these advantages, decision tree model is often used to preprocess raw data in combination with other ML methods. Therefore, both artificial neural network and decision tree methods were applied in this study to develop prediction models for protein allostery.

The second PDZ domain (PDZ2) in the human PTP1E protein is a typical dynamics-driven allosteric protein upon binding with its allosteric effectors, and has been subjected to both experimental and computational investigations. Therefore, it is used as model system in this study and subjected to above two ML methods to develop classification models associated with its allosteric states. There are two goals to achieve in this study: developing theoretical prediction models to recognize two allosteric states of PDZ2 (unbound and bound) and identifying key geometric features that potentially drive allostery of this protein. It is expected that the selected ML methods could facilitate to reveal key features to influence the overall protein allosteric processes.

**Methods**

**Molecular dynamics simulations**

The initial structures of PDZ2 protein were obtained from Protein DataBank (PDB)\textsuperscript{[36]} with codes as 3LNX and 3LNY, for the unbound and bound states, respectively (Fig. 1). These PDB structures were processed with hydrogen atoms added and solvated in a cubic water box as TIP3P model\textsuperscript{[37]} with charge balancing ions as sodium and chlorine added. The systems were then subjected to energy minimization. Consequently, the systems were subjected to 12 picoseconds (ps) molecular dynamics (MD) simulations to gradually raise the temperature to 300K before being equilibrated via 10 nanoseconds (ns) isothermal-isobaric ensemble (NPT) MD simulations at 300 K and 1 atm. Afterwards, canonical ensemble (NVT) Langevin MD simulations were carried out as the production runs. For all above simulations, 2 femtoseconds (fs) step size was used. The chemical bonds associated with hydrogen were fixed using SHAKE method.\textsuperscript{[38]} Cubic periodic boundary condition (PBC) was applied in these simulations. The long-range electrostatic interactions were modeled using the particle mesh Ewald algorithm.\textsuperscript{[39]} All simulations were carried out using CHARMM simulation package\textsuperscript{[40]} version 40b1 and the CHARMM22 force field.\textsuperscript{[41]} For both unbound and bound states of PDZ2, total of 13 simulations of 34 ns in length were carried out. For all trajectories, the initial 4 ns were discarded as equilibrium phase. Frames were saved every 10 ps. Therefore, 3000 frames were extracted from each 30 ns trajectory and subjected to the ML model analysis. Among 13 simulations of each state, 10 simulations were randomly selected as training set, and remaining three simulations were used as testing set. Cross-validation on training set was used to optimize the classification models and tested on test sets.

**Machine learning methods**

The machine learning methods applied in this study include the artificial neural network (ANN) model, and the decision tree (DT) model. A typical ANN model consists of input layer, hidden layers, and output layer. Each layer consists a set of “nodes” interconnected with other nodes in the adjacent layer(s). These nodes contain activation functions. The connections among nodes are weighted by additional factors. During the training process of an ANN model, the original data from the training set were entered to the input layer and went through the hidden layer(s) before reaching the output layer. A feedback process called “back propagation” was employed to minimize the error at the output layer. The purpose of the back propagation is optimizing the activation functions and weights on internode connections to achieve the minimum prediction error at the output layer upon convergence.\textsuperscript{[42]} When there is more than one hidden layer, ANN is also referred to as deep neural network model, which usually requires much higher computational cost in training process.\textsuperscript{[43]} Therefore, only one hidden layer was used in the initial ANN model setup, and was shown to be sufficient. An additional regularization including L2 penalty term was used to avoid over-fitting problem in the training process. L2 penalty term was added to the ANN model when updating the weight of each node. This penalty term limits the changes of weights during each iteration to avoid over-fitting. Overall, the number of nodes in hidden layer and L2 penalty term was refined to achieve the highest accuracy.\textsuperscript{[44–46]}

DT method has been widely used in strategy determination and identification of important factors. Combined with
chemical descriptors, DT method was also applied to predict chemical activities. The DT method was also applied in this study to develop a classification model. It provides an efficient algorithm to identify how the results can be predicted from individual features based on the information entropy gain. The DT model implemented in scikit-learn package was employed and refined to achieve the best predictive model in this study. Comparing to the ANN model, the classification or prediction model resulted from DT model is easier to interpret and understand.

Both pairwise distances for alpha carbons ($C\alpha$) and backbone dihedral angles ($\psi$ and $\phi$) were used as features to train the ANN and DT models. MSMbuilder\cite{49} package was employed to extract full $C\alpha$ pairwise distances and dihedral angles from simulation trajectories. For better performance of these two ML models, prescreening all features is necessary. Tree-based feature selection methods implemented in scikit-learn package\cite{48} were applied to prescreen important features for the ML analyses presented in this study.

To assess the performance of each classification model, we calculated four summary metrics including accuracy, recall, precision, and F1 score, which are defined as

$$\text{Accuracy} = \frac{TP + TN}{all} ; \quad \text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN} ; \quad F_1 = \frac{2 \cdot \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where true positive (TP) and true negative (TN) are defined as the number of structures that are classified correctly into unbound and bound state. False positive (FP) and false negative (FN) are defined as the number of structures that are misclassified into the other states.

Analysis of MD trajectories

Root-Mean-Square Deviation (RMSD) and Root-Mean-Square Fluctuation (RMSF). The RMSD is used to measure the overall conformational change during the MD simulations with regard to a reference structure. For a molecular structure represented by Cartesian coordinate vector $r_i$ ($i$ = 1 to $N$) of $N$ atoms, the RMSD is calculated as the following:

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} (r_i^2 - U_{r_i})^2}{N}}$$

where $T$ is the total number of frames in the given MD trajectory, $r_i^j$ is the coordinate atom $i$ in the frame $j$, and $U_{r_i}$ is the averaged coordinate of atom $i$ in the given trajectory. This analysis is based on the simulation frames superimposed to the averaged structure of the given trajectory.

Principal Component Analysis (PCA). By applying quasi-harmonic analysis implemented in the CHARMM program, PCA was performed on the unbound and bound state simulations to obtain dominant modes in each state. Translational and rotational components were projected out for each frame. All analyses were carried out using CHARMM simulation package version 40b1.

Cross-correlation matrix is a measurement of the correlated movement of a set of atoms. Each matrix element is defined as

$$C_{ij} = \frac{c_{ij}}{c_{i}^{1/2} c_{j}^{1/2}} = \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\left(\langle r_i^2 \rangle - \langle r_i \rangle^2 \right) \left(\langle r_j^2 \rangle - \langle r_j \rangle^2 \right)}^{1/2} ,$$

where $C_{ij}$ is the measurement of the correlated movement between atoms $i$ and $j$, $c_{ij}$, $c_{ii}$, and $c_{jj}$ are the covariance matrix elements, and $r_i$ and $r_j$ are Cartesian coordinate vectors from the least-square fitted structures, hence with translation and rotation projected out. Matrix elements $C_{ij}$ are between $-1$ and 1 with negative values indicating negative correlation and positive values indicating positive correlation between the motions of atoms $i$ and $j$. It should be noted that the correlation is defined as related movement along the line between two points. Correlated movement along orthogonal paths yields a cross-correlation matrix element of zero.

Dynamical Network Analysis. Potential allosteric pathways consisting residues identified by machine learning models were examined through dynamical network analysis using the NetworkView plugin implemented in VMD program.\cite{51,52} In the dynamical network analysis, if the backbone alpha carbons of any residue pairs are within 4.5 Å for more than 75% of simulation time, these two residues are considered as being connected. The connection strength for each connected residue pair is weighted by correlation value of these two residues in the cross-correlation matrix. For any two residues not connected, optimal pathways may be identified through other connected residues and the connections among them.

Results

Prescreening features for further analysis

The pairwise distances for $C\alpha$ and backbone dihedral angles were subjected to a prescreening process using DT model to select features for efficient machine learning analysis. All 26 trajectories for both unbound and bound PDZ2 states were used for the prescreening purpose. Total of 4371 $C\alpha$ pair

$$\text{RMSF}_i = \sqrt{\frac{1}{T} \sum_{j=1}^{T} (\mathbf{r}_i^j - \overline{\mathbf{r}_i})^2} ,$$

where $T$ is the total number of frames in the given MD trajectory, $\mathbf{r}_i^j$ is the coordinate atom $i$ in the frame $j$, and $\overline{\mathbf{r}_i}$ is the averaged coordinate of atom $i$ in the given trajectory. This analysis is based on the simulation frames superimposed to the averaged structure of the given trajectory.
distances and backbone dihedral angles were subjected to the prescreening process. The number of important features that could be selected depends on the depth of DT model. With the depth \( n \), the maximum number of features that can be covered in the model is \( 2^n - 1 \). For feature prescreening purpose, to ensure that the DT model covers all the possible features in the affordable computational costs, the depth of DT model was set as 20. After training this DT model, total of 289 features each with importance greater than 0.1% were selected for the following analysis. Combined together, these 289 features contribute 90.0% as total importance to the model.

**PDZ2 state classification by DT and ANN models**

Using the preselected 289 features, the DT model was further refined through the following training procedure. Ten trajectories were randomly selected among 13 independent simulation trajectories as training set for the unbound and bound states of PDZ2, respectively. For each state, 10 selected trajectories were randomly divided into five groups each with two trajectories. For each 30 ns trajectory, 3000 frames evenly distributed along the trajectory were selected for the training and testing purpose. The five groups of trajectories of both unbound and bound states were subjected to five rounds of cross-validation process described as the following. In each round of the validation process, one group of both unbound and bound states trajectories was selected as the test set for validation purpose with the remaining four groups as the training set.

For the DT model, depths of the tree ranging from 3 to 12 were tested in the cross-validation process. With depths as 4 and 5, the best performance is achieved to avoid potential over-fitting problem (Fig. 2a). The DT model with depth 4 showed higher prediction power for the additional six simulations of unbound and bound states than the one with depth 5. Therefore, the DT model with depth 4 was selected as the final model. For ANN model, six different values of a parameter alpha, also referred to as learning rate, were tested for the best performance, with alpha as 1 (\( \log(\alpha) = 0 \)) leading to the best prediction model (Fig. 2b). For the best DT model with depth 4 and ANN model with alpha as 1, the prediction accuracy for the six testing trajectories is 75% and 80%, respectively (Figs. 2c and 2d). In addition, one dummy classifier was built to generate random predictions as a baseline comparison for the ANN and DT classifiers. Random dummy predictions were repeated 100 times, and the metrics calculated by averaging these 100 dummy classifications is 0.5 with standard deviation as 0.0034 (Fig. 2e). The differences between the baseline dummy classifier and the ANN or DT classifier suggest that, although the unbound and bound states have similar structure with less than 2 Å RMSD differences, these two states are clearly differentiable using machine learning methods.

One of the advantages about the two prediction models using machine learning methods is that they could calculate the probability of any given structure that belongs to either unbound or bound state. The distribution of this probability was calculated for all the testing trajectories using both DT and ANN models, and is plotted in Figure 3. In the distributions calculated using DT model, there are five peaks in each state. Each peak from one state overlaps with a corresponding peak from the other state. The major difference between each peak from two states is the height (Fig. 3a). For example, the unbound state simulations have the highest peak close to the unbound state end of \( x \)-axis. For the bound state, the highest peak is the closest to the bound state end of \( x \)-axis. However,
the second highest peak of the bound state is close to the unbound state end. In the ANN prediction model, the probability distribution of each state has only one major peak very close to each end of the $x$-axis, reflecting the high prediction accuracy of this model. In addition to the differentiation between two states, the calculated probabilities could also be utilized to select representative structures for various states, especially those different from both unbound and bound states, which are referred to as intermediate states. Using the probabilities calculated by the ANN model, the representative structures were selected for the unbound, bound, and intermediate states (Fig. 4). The colored arrows in unbound and bound states provide structural information differentiating these states from the intermediate state.

Identifying key residues

Another important implication of machine learning models is identifying the important features strongly correlated with allosteric states. In both DT and ANN models, the contribution from each feature to differentiate two states is calculated and can be used to rank the features. In the two models of this study, both C$_x$ distances and backbone dihedral angles are used and ranked together based on their contributions. The top 10 features with the highest contributions are listed in Table 1 for the DT and ANN models, respectively. In the DT model, eight top features are C$_x$ distances, while five of top ten features are C$_x$ distances in the ANN model. Among the top 10 features, two models share three features (C$_x$ distance between residues 38 and 71, backbone dihedral angle $\psi$ connecting residues 1 and 2, backbone dihedral angle $\phi$ connecting residues 22 and 23). Among the top 10 features reported from the DT and ANN models, there are 19 different residues involved. Total of 16 among these 19 residues have been identified as related to PDZ2 allostery upon binding with the same peptide in several studies. The top three features listed in Table 1 from the DT and ANN models are subjected to further analysis described as the following.

Further analysis of the key residues

To illustrate the difference between the distributions of the unbound and bound states of PDZ2, a 2D-RMSD plot with reference to the crystal unbound and bound structures is shown in Figure 5a. The distribution plot shows that the bound state simulations sampled a region similar to the unbound state simulation, but covered larger conformational space. To further compare the simulations of the two states, distributions of three key features identified in the DT and ANN models (C$_x$ distances between residues Lys38 and His71 and between residues Asn16 and Arg31, backbone dihedral angle $\psi$ connecting residues Pro1 and Lys2) are plotted in Figures 5b–5d. C$_x$ distance between Lys38 and Thr70 was not plotted because residue Thr70 is adjacent to residue His71. Interestingly, although the unbound and bound states have similar structures with low RMSD difference, the distributions of these three key features are significantly different between the two states. For the dihedral angle between residues Pro1 and Lys2, which
appeared as the top feature in ANN model and the second most important feature in the DT model, the relative heights of two peaks are switched in the bound state compared with the unbound state. This observation is consistent with the population shift hypothesis, that the free energy landscapes of two allosteric states are different upon perturbations despite the similarity of their structures. The distribution of the Cα distance between Lys38 and His71 is also significantly different between the two states. The most probable value of this distance in the bound state is larger than the one in the unbound state (Fig. 5b). The distribution of the Cα distance between residues Asn16 and Arg31 is peaked around 29 Å in both states. But the probability at the peak is much higher in the unbound state than in the bound state (Fig. 5d). Interestingly, the pairing residues for key Cα distances, Lys38:His71 and Asn16:Arg31 are far from each other and across the protein structure, as they are located either on or close to distal loop structures (Fig. 6). These results suggest that the correlated fluctuation of Lys38:His71 and Asn16:Arg31 or their associated secondary structures play a critical role to differentiate the unbound and bound states, and hence, serve as key factors related to the PDZ2 allostery.

In addition to the distribution analysis, the fluctuations of the key residues are another comparison between the different simulations. RMSF analysis could be used to measure the averaged structural fluctuations of each residue in dynamics simulations. PCA is a widely applied method to analyze the global motion of protein structures based on dynamics simulations. Therefore, we applied RMSF and PCA on the simulations of both unbound and bound states of PDZ2. In the RMSF plot (Fig. 7a), the four key residues Asn16, Arg31, Lys38, and His71, display rather high fluctuations. The cumulative contributions from PCA modes are plotted for both states in Figure 7b. For both states, the 20 modes with lowest frequencies account for more than 50% of the total variances. Therefore, the average of these modes was used to measure the fluctuation of each residue in principal component (PC)

Table 1. Top 10 important features identified by decision tree and artificial neural networks models.

<table>
<thead>
<tr>
<th>Decision tree</th>
<th>Neural networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Residues</td>
</tr>
<tr>
<td>Cα distance</td>
<td>38(b), 71(a,b)</td>
</tr>
<tr>
<td>Cα distance</td>
<td>15(a,b), 31(a,b)</td>
</tr>
<tr>
<td>Cα distance</td>
<td>31(a,b), 69(a,b)</td>
</tr>
<tr>
<td>Cα distance</td>
<td>18(a,b), 28(b)</td>
</tr>
<tr>
<td>Cα distance</td>
<td>23(b), 31(a,b)</td>
</tr>
<tr>
<td>Cα distance</td>
<td>31(a,b), 71(a,b)</td>
</tr>
<tr>
<td>Cα distance</td>
<td>23(a,b), 23(b)</td>
</tr>
<tr>
<td>Cα distance</td>
<td>31(a,b), 52(b)</td>
</tr>
</tbody>
</table>

[a] Residue has already been identified by NMR studies. [b] Residue has already been identified by other computational studies.
vector space (Fig. 7c). Three residues, Asn16, Arg31, and His71 also display high fluctuations in the PC vector space.

PC1, the most dominant PC modes of unbound and bound states simulations, are illustrated in Figure 8. The loop between residues Val26 to Gly33 in both states displays higher fluctuation comparing with other part of the protein. Also, the bound state has a higher fluctuation than the unbound state. As shown in Figure 8, the fluctuation in that loop shows a trend to change the shape of the protein, which could be one of the reasons for the fluctuation difference between Asn16 and Arg31 as shown in Figure 5. Those differences between the PC1 modes in the two states could account for the allosteric effects.

Thus far, we focus on the distance distributions and residues fluctuations of key features identified by the machine learning models. The mechanisms how the key residue pairs are correlated with each other still remain unclear. Therefore, dynamical networks analysis,\[51,52\] a correlation-matrix-based method, was applied to identify potential allosteric pathways for Lys38:His71 and Asn16:Arg31 as the key residue pairs (Fig. 9). The analysis reveals that Val22 serves as one key residue involving correlation between Lys38 and His71. Although Val22 is not close to either Lys38 or His71 in sequence, it is located at the middle of these two residues in space and closer to Lys38 than to His71 (blue pathway in Fig. 9). In addition, four residues (Val22, Ile20, Leu18, Ser17) from the loop containing Asn16 and four residues (His32, Gly33, Gly34, and Tyr36) from the loop containing Arg31 form a communication pathway involved with the correlation between Asn16 and Arg31 (red pathway in Fig. 9). It is interesting that both pathways share the same residue Val22, which is also associated with multiple key features selected from the two machine learning models (Table 1) and other experimental and computational studies.\[53–56\]

Discussion

In this study, the decision tree and artificial neural networks models were applied to develop classification models of two allosterically related states of PDZ2 domain from PSD-95 protein. Principal component analysis of protein dynamics and RMS fluctuation analysis of individual residues were carried out to further evaluate the machine learning models. Dynamical network analysis was used to identify potential pathways accounting for the correlations among the key residues.

Classification of two states

In addition to the conformation-driven allostery, dynamics-driven allostery model plays increasingly important role in protein allostery from dynamical ensemble point of view.\[8,57–59\] In dynamics-driven allostery model, it is likely entropy instead of enthalpy that drives protein allostery because of the absence of significant conformational changes. There are some studies utilizing parameters associated with whole proteins instead of individual residues, such as RMSD, principal component analysis, and correlation matrix, to investigate protein allostery.\[6,11,51,52\] But there is still need for methods that differentiate allosteric states of proteins, build connections with individual residues, and provide guidance for mutagenesis studies to control protein allostery. Machine learning models have been widely used in information technology classification applications,\[60–62\] and are regaining popularity in computational chemistry and biology.\[63–65\] One of the goals of this study is exploring new ways to differentiate protein allosteric states and build connections between protein allostery and individual residues. Therefore, in this study, the DT and ANN models were built to achieve more than 75% and 80% prediction accuracy to differentiate the unbound and bound allosteric states of PDZ2, respectively. More importantly, both models provide quantitative evaluation of features, which are associated with specific residues. The good agreement between the residues identified in both models and the previous experimental as well as
computational studies strongly suggests that the machine learning models could provide insight into protein allostery as complement to other widely used analyses of protein simulations. The distinct distributions of key features in the unbound and bound states plotted in Figure 3 provide an alternative picture of population shift hypothesis underlying dynamics-driven protein allostery.

One difficulty in protein allostery study is finding appropriate transition state with reference to distinct allosteric states. Allosteric processes, especially dynamics-driven allostery, usually occur in short time scales, and are difficult to be characterized experimentally. Using the quantitative machine learning models developed in this study, the distributions of simulations with regard to two allosteric states could be plotted (Fig. 4). The sampling located at the middle of two states could be considered as intermediate states and subjected to further analyses.

Given the effectiveness of the machine learning models presented in this study, one would logically expect that many other machine learning models could also be useful for analyzing simulations of protein allosteric states. Therefore comparison among different machine learning models for protein simulations will be the focus of future studies. In this study, the ANN model has better training and prediction accuracy than the DT model. However to develop the accurate ANN model, the DT model is necessary to prescreen the potentially important features. This is mainly due to the different characteristics of these two methods. The DT model focuses on individual parameters as features for classification. As contrast, the ANN model uses the combination of all features with different weights for classification purpose. In the future applications on different systems, caution should be used with regard to the choices and usage of machine learning models.

Identifying important features

An important strength displayed by machine learning models in this study is identifying key features for protein allostery. Both Cα distances and backbone dihedral angles can be easily used simultaneously for the development of accurate classification models. The fact that both distances and dihedral angles are among the top features suggests that many other order parameters of molecular simulation systems could be utilized for machine learning models for either allostery or other purposes such as computer-aided molecular design. The distributions of the selected individual features demonstrate significant difference between structurally similar allosteric states, and provide an alternative way of analyzing population shift of simulations upon allosteric or other perturbations on proteins. In addition, the key features specifically associated with individual residues provide unambiguous candidates for mutagenesis studies of proteins comparing to other studies using global descriptors of protein dynamics.

The top 10 features identified in the DT and ANN models comprise 19 residues. It is significant that 16 of these 19 residues have been identified in one experimental NMR study and several computational studies. For the top three features identified in this study, residues Pro1 and Lys2 serve as part of the allostery communication network identified in a protein network model. Lys38 is regarded as one of the “hot residues” in another simulation study of PDZ2 as well as part of the communication network. Asn16, Arg31, and His71 were identified as key allosteric residues in both NMR study and other computational studies. From biological point of view, Asn16 and His71 are located in the binding pocket, and could stabilize the binding peptide. Residue Arg31 displayed a significant relaxation contribution value in a conformation exchange (Rex) study. Some experimental studies

Figure 8. PC1 modes illustrated as porcupine plot: a) unbound state, b) bound state. Colored arrows indicate the direction and magnitude of movement. Val26-Gly33 loop is highlighted in blue color. [Color figure can be viewed at wileyonlinelibrary.com]

Figure 9. Network pathways for Lys38-His71 (blue) and Arg31-Asn16 (red). [Color figure can be viewed at wileyonlinelibrary.com]
also pointed out the residues located in \( \beta 1/\beta 2 \) loop including Asn16, and residues located in \( \beta 2/\beta 3 \) loop including Arg31 were important for peptide binding.\(^{[56]}\) Considering that a large number of features associated with all the residues in the protein were treated equally to develop the classification models in this study, the overwhelming agreement with other studies strongly support the effectiveness of machine learning models for protein dynamics analysis.

The differences in the distributions of these key features between two states (Fig. 5) provide not only mechanistic insight into the machine learning models, but also a more quantitative view of population shift hypothesis of protein allostery. The different distribution of \( C_x \) distance between Lys38 and His71 may suggest the importance of secondary structures (a loop structure containing Lys38 and a helix structure containing His71, see Fig. 6) for the protein allostery. The key residue selections also agree with the residue fluctuation features containing His71, see Fig. 6) for the protein allostery. The dynamical network analysis\(^{[56]}\) identified two pathways containing additional residues which may play an important role for the communication between two key residue pairs (Fig. 9). The fact that residue Val22 being part of both pathways and also associated with several top key features identified in this study further support the notion that the machine learning models could be complimentary to the existing analysis tools of protein simulations by providing more insights related to individual residues. In general, these machine learning models could be applied to investigate the distribution differences between different states of dynamics-driven allosteric proteins, for which the conformational changes are not significant and the differences are difficult to be described by other analysis methods.

There might be concern that the important residues identified using ML methods are the outcome instead of the cause of allostery. According to the population shift hypothesis, distribution differences are essential for investigating the mechanism of allostery. Although not determined as either the cause or the outcome of allostery, it is an important step to identify the residues displaying distribution differences between two allosteric states. According to most experimental and computational studies about protein allostery, important residues behave differently through allosteric processes but are fundamental to allostery effect. However probing these unlikely events is beyond the scope of this study. Because the ML methods in this study do not require any a priori knowledge to identify most important residue pairs to differentiate two allosteric states, these models could serve as a complimentary method for the dynamical network analysis, which requires a priori knowledge about the source and target residues for the investigation of the potential allosteric pathways in the proteins of interest.

Conclusions

In this study, both decision tree and artificial neural network as machine learning models were applied to systematically investigate allosteric mechanism of PDZ2 protein upon binding with a peptide. Although there is no significant conformational change displayed between the unbound and bound states of PDZ2, two classification models developed in this study provide more than 75% of accuracy to differentiate these two states. Both models also provide a quantitative evaluation of the contributions from individual features to overall difference between the two states. Most residues associated with the important features including \( C_x \) distances and backbone dihedral angles have also been reported as key allosteric residues in both experimental and computational studies. Furthermore, the distributions of key features in different states provide alternative ways to analyze the population shift of protein ensemble upon allosteric perturbations. Additional analyses were also carried out for PDZ2 simulations using widely applied approaches including principal component analysis, RMS fluctuation analysis and dynamical network analysis, and showed good agreement with the machine learning results. Overall, the adopted machine learning methods on molecular dynamics simulations of protein in this study showed promise as a systematic and unbiased means to gain insight into protein allostery, especially the specific contribution from individual residues.

Acknowledgment

Computational time was provided by Southern Methodist University’s Center for Scientific Computation and Texas Advanced Computing Center (TACC) at the University of Texas at Austin. The authors thank Drs. Gennady Verkhivker and Shouyi Wang for critical reading of the manuscript and fruitful discussions.

Keywords: allostery  machine learning  molecular dynamics  classification  protein


[34] S. Grossberg, Neural Netw. 1988, 1, 17.
PAPER V

REDAN: relative entropy-based dynamical allosteric network model

Hongyu Zhou & Peng Tao

To cite this article: Hongyu Zhou & Peng Tao (2018): REDAN: relative entropy-based dynamical allosteric network model, Molecular Physics, DOI: 10.1080/00268976.2018.1543904

To link to this article: https://doi.org/10.1080/00268976.2018.1543904
RESEARCH ARTICLE

REDAN: relative entropy-based dynamical allostERIC network model

Hongyu Zhou and Peng Tao

Department of Chemistry, Southern Methodist University, Dallas, TX, USA

ABSTRACT

Protein allostery is ubiquitous phenomena that are important for cellular signalling processes. Despite extensive methodology development, a quantitative model is still needed to accurately measure protein allosteric response upon external perturbation. Here, we introduced the relative entropy concept from information theory as a quantitative metric to develop a method for measurement of the population shift with regard to protein structure during allosteric transition. This method is referred to as relative entropy-based dynamical allosteric network (REDAN) model. Using this method, protein allostery could be evaluated at three mutually dependent structural levels: allostERIC residues, allostERIC pathways, and allostERIC communities. All three levels are carried out using rigorous searching algorithms based on relative entropy. Application of the REDAN model on the second PDZ domain (PDZ2) in the human PTP1E protein provided metric-based insight into its allostery upon peptide binding.

Introduction

Molecular dynamics (MD) simulations have been widely applied to investigate protein structures and functions [1]. Function regulations of many proteins involve external or internal perturbations including light stimulation [2], ligand or peptide binding [3], stress activation [4], pH activation [5], etc., which are essential for protein regulations. In general, the regulations of protein function due to external perturbations are referred to as allostery [6], which are ubiquitous molecular processes in biological systems. Recently, a population shift model was proposed that different function-related protein conformations could coexist [7,8]. Upon external perturbation, the free energy landscape of a target system could change significantly whereas the populations of different states are shifted. These changes of free energy landscape are essential for so-called dynamics-driven allostery [6,9–13].

Dimensionality reduction methods could be applied to investigate the distribution changes using only limited number (usually up to three) of collective variables [14]. Due to unavoidable structural information loss, it is difficult to investigate how the dynamics lead to the distribution changes, and how the perturbation information propagates inside protein. To address this difficulty, an improved method is needed to accurately compare the distribution between the simulations of two allosteric states to offset the structural information loss due to dimensionality reduction analysis.

One metric to quantitatively measure the difference between two probability distributions is relative entropy [15]. Relative entropy, also known as Kullback–Leibler divergence, is a concept in statistics to measure how one probability distribution diverges from the expected distribution with broad application in many fields [16–20].
By adapting this metric into MD simulation analyses, one would be able to qualitatively describe how one simulation diverges from other simulations. This measurement could be applied on many distributions. The distribution differences measured by relative entropy are equivalent to the free energy changes upon external perturbations, and can be considered as one of the allostery effects.

To analyse protein structure–function relations and quantify the communication among residues inside protein, a group of approaches referred to as protein structure network methods were developed to identify network of residues to model residue communication based on protein structural dynamics. In protein structure network analysis, each amino acid residue is considered as a node, and edges are built to connect nodes to obtain different network representation of a protein. Specifically, protein contact network (PCN) and residue interaction network (RIN) models were developed and applied to reveal the residues crucial for protein stability, and identify domains, hubs, and clusters of residues correlated with protein functions [21–23]. Elastic network models were developed to investigate the interactions among residues through approximating inter-residue interactions by harmonic elastic restraints [24,25]. The network analysis has also been adapted broadly to analyse MD simulations. Dynamics network analysis method models the residue interaction in the network using the correlation matrix based on MD simulations [26]. These network analyses have been widely applied to investigate the communication among residues in proteins [27]. However, no method has been developed to utilise simulation distribution information, which closely correlates with the functions, and is readily available from the MD simulations of macromolecules. In addition, few methods could quantitatively characterise the allosteric effects of proteins upon external perturbations. Here, we developed a novel quantitative network analysis method utilising distribution information from MD simulations specifically targeting protein allostery. This method is referred to as relative entropy-based dynamical allosteric network (REDAN) model and could be applied to compare distribution differences of two allosteric states upon perturbation and build quantitative network model.

In REDAN model, each amino acid residue is considered as a node, and connection between any node pair is considered as an edge. The change of distance distribution between any node pair can be calculated using relative entropy method and used as the weight for the corresponding edge. These weights quantitatively measure the response of protein dynamics upon perturbation and could be used to characterise allostery induced by the same perturbation. Therefore, this network model could quantitatively describe protein allosteric effects from the perspective of structural biology and population shifting. Higher relative entropy indicates significant allosteric effect or larger distribution shift due to perturbations. Using this allosteric network model, we can quantitatively compare allosteric effects upon perturbation with minimum structural information loss.

Similar with other network models [21,26], the pathway and community analyses could also be conducted in this allosteric network model. A typical allosteric pathway consists a series of edges connecting two distal residues to exhibit the potential communication between residues leading to the allosteric effects. An allosteric community represents a group of residues with minimum allosteric effects upon perturbation. The second PDZ domain (PDZ2) in the human PTP1E protein [28] is an allosteric protein which could propagate signals to other part of molecular complex upon peptide binding [28,29], and is subjected to the allosteric pathway and community analysis using REDAN method to reveal potential allosteric mechanism and identify allostery-related residues.

Materials and methods

MD simulation

For PDZ2 system, the initial structures were obtained from the PDB [30] with the ID as 3LNX (unbound state) and 3LNY (peptide bound state), respectively. After adding hydrogen atoms, PDZ2 is solvated using explicit water model (TIP3P) [31] and neutralised with sodium cations and chloride anions to maintain 0.1 M ionic strength. The simulation system was then subjected to the adopted basis Newton–Raphson energy minimisation, which yielded a total gradient of less than 0.001 kcal/(molÅ). After the minimisation, 10 nanoseconds (ns) of isothermal-isobaric ensemble (NPT) MD simulations followed by 100 ns of canonical ensemble (NVT) Langevin MD simulation at 300 K were conducted for both PDZ2 domain unbound and bound states. For all simulations, SHAKE constraint was applied to constrain all bonds associated with hydrogen atoms. Step size of 2 femtosecond (fs) was used and simulation trajectories were saved every 100 picosecond (ps). Cubic simulation box and periodic boundary condition were applied for all MD simulations. Electrostatic interactions were calculated using particle mesh Ewald method [32]. All simulations were carried out using CHARMM [33] simulation package version 41b1 with the support of GPU calculations based on OpenMM [34].
Relative entropy

Relative entropy method was applied to calculate the difference between the distributions of the distance between the alpha carbon (Cα) of two residues upon perturbation. The probability distributions of the Cα distance before and after allosteric perturbation are represented as $P$ and $Q$, respectively, with $p(x)$ and $q(x)$ as the distribution density at distance $x$. The relative entropy $D_{KL}$ between $P$ and $Q$ is calculated as the following:

$$D_{KL}(P||Q) = \int p(x) \ln \frac{p(x)}{q(x)} \, dx.$$  \hspace{1cm} (1)

Because the above equation is not symmetrical measurement for $P$ and $Q$, we symmetrise the relative entropy between $P$ and $Q$ by taking the average of $D_{KL}(P||Q)$ and $D_{KL}(Q||P)$. This averaged relative entropy is referred to as the perturbation relative entropy (PRE) between two distributions of the same distance in different allosteric states upon perturbation (Equation (2)).

$$\text{PRE}(P||Q) = \frac{D_{KL}(P||Q) + D_{KL}(Q||P)}{2} = \frac{1}{2} \int (p(x) - q(x)) \ln \frac{p(x)}{q(x)} \, dx.$$  \hspace{1cm} (2)

In any distribution, e.g., $P$, the free energy at distance $x$ ($A_{P_x}$) can be estimated from the distribution probability at $x$ as the following:

$$A_{P_x} = -k_B T \ln p(x),$$  \hspace{1cm} (3)

where $k_B$ is the Boltzmann constant, and $T$ is the temperature. Combining Equations (2) and (3), the PRE between distributions $P$ and $Q$ is a direct measurement of the free energy difference for the given order parameter between two states:

$$\text{PRE}(P||Q) = \frac{1}{2} \int (p(x) - q(x)) \ln \frac{p(x)}{q(x)} \, dx = -\frac{1}{2k_B T} \int (p(x) - q(x)) (A_{P_x} - A_{Q_x}) \, dx.$$  \hspace{1cm} (4)

Allosteric pathways

The allosteric networks can be built based on PRE matrix. PRE value measures the magnitude of the distribution shifting upon perturbations, and can be considered to indicate the significance of the allosteric effects. To identify potential allosteric pathways between two distant residues with large PRE, a cutoff value to control the edge length is necessary to facilitate the analysis. An edge between any residue pairs will be chosen if the most probable distance between the Cα of these two residues is smaller than the given cutoff value. For each chosen edge, a weight is defined as $1/\text{PRE}$. Therefore, the pathway with the smallest overall weight implies the propagation channel with the largest allosteric effect. The shortest pathway was identified by the Dijkstra’s algorithm [35], which is the most common pathway searching algorithm. Using Dijkstra’s algorithm, the search starts with the starting node, and iteratively loops all the available nodes until reaching the destination node to identify the shortest path connecting two nodes. More details could be obtained by referring to the literature [35].

Allosteric communities

The main objective for community analysis is dividing the residues into different communities, so that the total PRE associated with residue pairs within each community is a minimum, and the total PRE associated with residue pairs across different communities is a maximum. Therefore, the overall allosteric effects upon perturbation could be projected onto the correlation among communities. Both Girvan–Newman [36] and Kernighan–Lin [37] algorithms are implemented in this study to construct communities.

Girvan–Newman (GN) algorithm

The GN algorithm is a top-down community detection approach, which removes the ‘most valuable edge’ in each iteration, and recalculates the betweenness of all remaining edges until no edge remains. This algorithm depends on the graph construction and cutoff values. The optimal communities are determined by modularity value [36], which is the measurement of the strength of the community separation. Better community structure is indicated by larger modularity value. Final communities are selected with the highest modularity during iteration.

Kernighan–Lin (KL) algorithm

The KL algorithm [37] is a heuristic algorithm for finding the partition of graphs. The algorithm is independent to the graph construction and cutoff value, and only depends on the relative entropy matrix. Multiple random initializations are carried out in KL algorithm to search for the lowest possible relative entropy value within each community. The KL algorithm is outlined as the following.

Assuming $n$ communities labelled as $C_1$ through $C_n$, the total PRE inside communities are defined as

$$T = \sum_{i=1}^{n} \sum_{ij \in C_i} \text{PRE}_{ij}.$$  \hspace{1cm} (5)
where \( i \) and \( j \) are the residues in Community \( C_i \), and \( \text{PRE}_{ij} \) is the perturbation relative entropy between distance distribution of residues \( i \) and \( j \) upon perturbation.

Assuming that node \( i \) belongs to Community \( C_m \), the internal PRE of node \( i \) in community \( C_m \) is defined as Equation (6), and the external PRE of node \( i \) with reference to community \( C_q \) is defined as Equation (7):

\[
\text{In}_i = \sum_{j \in C_m} \text{PRE}_{ij},
\]

\[
\text{Ex}_{i,C_q} = \sum_{j \in C_q} \text{PRE}_{ij},
\]

The allosteric communities can be optimised by inserting node \( i \) from \( C_m \) into \( C_k \) or swapping node \( i \) from \( C_m \) with node \( j \) from \( C_k \). The benefits of the total PRE inside communities are calculated as Equations (8) and (9) for inserting and swapping operations, respectively:

\[
\text{Benefit} = T_{\text{new}} - T_{\text{old}} = \text{Ex}_{i,C_k} - \text{In}_i,
\]

\[
\text{Benefit} = T_{\text{new}} - T_{\text{old}} = (\text{Ex}_{i,C_k} + \text{Ex}_{j,C_m}) - (\text{In}_i + \text{In}_j) - 2 \times \text{PRE}_{ij}.
\]

Therefore, the optimal KL communities can be computed by selecting maximum benefit operation during each iteration until converging to a minimum total PRE value inside communities. However, the KL algorithm can only achieve a solution as a local minimum. In the current study, we repeat the KL algorithm until the lowest PRE value in communities remains unchanged for more than 1000 times, then the current partition is selected as the final community configuration. In addition, the KL algorithm could be applied on the GN searching results to further optimise the communities until convergence. This combination of GN and KL methods is referred to as a hybrid GN–KL algorithm.

**Results**

Although PDZ2 exhibits signal propagation upon ligand binding, the structures of the PDZ2 unbound state (3LNX [38]) and bound state (3LNY [38]) are very similar. It was shown that the distributions between the unbound and bound simulations are significantly different [39]. REDAN model is built based on those differences. Considering each residue as a node, the significance of allosteric effects for any node pair is measured as the relative entropy divergences between its distributions in two states and treated as the weight of the edge connecting these two nodes. These weights could reflect the allosteric response of the corresponding edges upon peptide binding and are referred to as PRE. It is worth to mention that because the free energy can be calculated based on probability distribution \( P \) as \( A = -k_B T \text{ln} P \), the PRE measures the change of free energy upon peptide binding. Therefore, the edges along with their weights can be used to model the direction of free energy propagation upon perturbation.

The PRE values of all edges in PDZ2 are calculated and illustrated in Figure S1. For most residue pairs, the PRE values upon peptide binding are close to zero, and are significant for only part of the residue pairs, making it a sparse matrix. The sparsity of the PRE matrix makes it suitable for a sparse protein network as illustrated in Figure S2. Comparing with other network methods including PCN, RIN, and dynamical networks analysis [21,26], the REDAN method could identify key allosteric edges between the residues far from each other rather than adjacent residues.

The distributions of edges with the highest and lowest PRE values are illustrated in Figure 1, respectively. Clearly, the peptide binding does not equally influence the distance distributions of different residue pairs. For the residues pair N14:A74 with the highest PRE, the unbound state has the distance around 19 Å with the peak density. Upon peptide binding, the distribution is broadened with a new peak appearing around 21 Å (Figure 1(a)), leading to the PRE of this edge upon peptide binding as 2.019. As a comparison, for the residue pair D56:V64, the peptide binding does not lead to observable distribution changes, which results in the PRE of this distribution close to zero (Figure 1(c)). The probability distribution was closely related to the free energy. The free energy profiles with reference to the edge distance between residue pairs N14:A74 and D56:V64 are plotted in Figures 1(b,d), respectively. With the large PRE value, the change of the free energy profile upon perturbation is more significant for the N14:A74 pair than the D56:V64 pair. Therefore, the PRE can be used as an adequate metric to measure the free energy changes upon external perturbations. These calculated PRE values are used in REDAN model to identify allostery-related residues, residue pairs, allosteric pathways, and allosteric communities.

**Identification of allosteric effects and allostery-related residues**

The REDAN model provides a tool to easily detect the residues and residue pairs that are more responsive to allosteric perturbations. For PDZ2, residue pair N14:A74 has the highest PRE upon peptide binding. The top five residue pairs with the highest PRE value are listed in Table S1. The residue pairs with the highest PRE are all
correlated with $\beta1/\beta2$ loop with $\alpha3$ helix (Figure 1(e)). Interestingly, the peptide binding site is formed between $\beta2$ strands and $\alpha3$ helix.

For each residue, the PRE associated with all edges which include that specific residue could be summed together as residue specific total PRE. This total PRE may reflect the significance of allosteric effects between each individual residue and the rest of protein upon perturbations. All residues in PDZ2 are sorted using their total PRE with the top 15 residues listed in Table 1 and the complete list provided in Table S2. Because the edge can be considered as the direction of free energy propagation, the total PRE could reflect the magnitude of free energy passing through that residue as a node upon perturbation. The top 15 residues cover exactly the residues from G68 to V75 and V26 to H32 (Figure S3). Comparing with a previous network analysis and an NMR study related to PDZ2 bound with the same peptide [29,40,41], 12 out of these 15 residues have been identified as allosterically or functionally related residues (Table 1). The residues V26 to H32 form $\beta2/\alpha1$ loop and the residues G68 to V75 form $\beta5/\alpha3$ loop and part of $\alpha3$ helix. Those regions are highlighted as allostery-related structures in many studies [28,29,40,41].

**Allosteric pathways**

The residue pairs identified above with significant allosteric effects usually are not adjacent with each other. For example, the distance between N14:A74 residue pair is around 20 Å. The significant allosteric effect between these two residues could not be fully accounted for by non-bonded interactions between them, because the non-bonded interactions are too small at this distance to exert any significant impact. Alternatively, significant distribution changes correlated with large allosteric effect could stem from the accumulation of shorter-range allosteric effects. In REDAN model, the decomposition analysis of the long-range allosteric effect into sequential short-range and smaller allosteric effects is carried out using the shortest pathway searching algorithm. For example, in the PDZ2 protein, the large allosteric effect displayed by N14:A74 residue pair (Figure 1(a)) is decomposed into a series of sequential residue pairs.

---

**Figure 1.** The significance of distribution changes and free energy surface changes quantified by PRE. (a) Residue pair (N14:A74) with the highest PRE in the protein; (b) The free energy surface of the N14:A74 distance distribution; (c) The residue pair (D56:V64) with the lowest PRE; (d) The free energy surface of the D56:V64 distance distribution; (e) Residues N14 and A74 illustrated in PDZ2; (f) Pathway decomposition: the distributions for decomposed residue pairs; (g) Pathway decomposition analysis of N14:A74 pair with cutoff value as 12 Å; (h) Pathway decomposition analysis of N14:A74 pair with cutoff value as 5 Å. These results demonstrate that the PRE is an effective measurement to quantify allosteric effect at residue pair level.

**Table 1.** Top 15 residues with the highest residue specific PRE.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Residue</th>
<th>Total PRE</th>
<th>Rank</th>
<th>Residue</th>
<th>Total PRE</th>
<th>Rank</th>
<th>Residue</th>
<th>Total PRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T70$^b$</td>
<td>54.99</td>
<td>6</td>
<td>Q73$^b$</td>
<td>46.70</td>
<td>11</td>
<td>V75$^b$</td>
<td>37.07</td>
</tr>
<tr>
<td>2</td>
<td>V26$^{ab}$</td>
<td>54.86</td>
<td>7</td>
<td>A69$^{ab}$</td>
<td>46.67</td>
<td>12</td>
<td>K72</td>
<td>35.88</td>
</tr>
<tr>
<td>3</td>
<td>N27$^{ab}$</td>
<td>52.62</td>
<td>8</td>
<td>R31$^a$</td>
<td>45.32</td>
<td>13</td>
<td>H32</td>
<td>34.75</td>
</tr>
<tr>
<td>4</td>
<td>H71$^{ab}$</td>
<td>50.49</td>
<td>9</td>
<td>T28$^{ab}$</td>
<td>43.28</td>
<td>14</td>
<td>V30$^{ab}$</td>
<td>32.03</td>
</tr>
<tr>
<td>5</td>
<td>A74$^b$</td>
<td>46.76</td>
<td>10</td>
<td>S29$^b$</td>
<td>38.14</td>
<td>15</td>
<td>G68</td>
<td>30.76</td>
</tr>
</tbody>
</table>

$^a$Residues identified through an NMR study [40].

$^b$Residues identified through two network analyses [29,41].
with short-range allosteric effect using a cutoff value as 12 Å: N14:R79, R79:S17, S17:V75, V75:S21 and S21:A74 (Figure 1(g)). Comparing distributions in Figure 1(a,f), it is clear that the decomposed residue pairs have a smaller shift of distribution upon peptide binding but all in the same direction to the larger allosteric effect displayed by N14:A74 residue pair. This series of short-range edges with significant PRE values may contribute to the large allosteric effect between N14:A74 as one important pathway consisting of N14, S17, S21, V75 and R79. It should be noted that the potential allosteric communication between residues N14 and A74 does not necessarily propagate only through this identified pathway. However, all five residue-pairs as part of this pathway have increasing distance distribution upon peptide binding, which is consistent with the target N14:A74 edge, making it likely that this pathway correlates with the overall allosteric effect.

The PRE values of the short-range residue pairs listed above are 1.385 (N14:R79), 0.660 (R79:S17), 1.337 (S17:V75), 0.815 (V75:S21), and 1.045 (S21:A74) as shown in Figure 1(f) and individually in Figure S4. Residues N14 and S17 belong to β1/β2 loop (covering residues 13 through 19), and residue S21 belongs to β2 strand. Residues A74, V75, and R79 belongs to α3 helix. N14:R79 pair has the highest PRE along this pathway. Comparing with the β1/β2 loop region, the α3 helix as a stable secondary structure could be more stable. Therefore, this pathway decomposition may reveal that the large PRE between N14:A74 may stem from the fluctuation of β1/β2 loop. Among A74, V75, and R79 residues, R79 is the closest residue in α3 helix structure with regard to the β1/β2 loop. Therefore, to further evaluate allosteric response from the β1/β2 loop, the distribution of residue pair distances and corresponding PRE values between R79 and all β1/β2 loop residues (10–21) are plotted in Figure S5. Among these residue pairs, the PRE values increase from the lowest one between E10:R79 with 0.020 to the highest one between N14:R79 with 1.385, and sequentially decrease to 0.182 as the one between S21:R79. Central three residues N14, D15, and N16 have PRE values higher than 1, suggesting that this loop region significantly changes the conformation upon peptide binding.

It has been suggested that allostery was a complex biological function, and multiple pathways could coexist and lead to the allosteric effects, ranging from long-range global pathways to short-range local pathways [42]. Although some pathways may be more dominant than other pathways for propagation purpose, the allosteric effect should be considered as the result of cooperation among multiple pathways [42]. To identify potential multiple pathways, a cutoff value was applied to differentiate allosteric pathways with different interaction ranges. This cutoff value is used as the upper bound to search for the shortest allosteric pathway connecting the target residue pair. This gives flexibility of this model to survey important allosteric pathways at any distance range. To evaluate the impact of different cutoff values on allosteric pathways, 16 different cutoff values ranging from 5 to 20 Å are used for allosteric pathway identification (Table 2). Cutoff values shorter than 5 Å do not lead to any allosteric pathways. Different cutoff values do lead to different allosteric pathways. But for each specific cutoff value, unique allosteric pathway could be determined. For the cutoff value of 5 Å, the adjacent residues as N14-K13-A12 and residues from 83 through 74 are identified as the shortest allosteric pathway (Figure 1(h)), highlighting the importance of the local interaction for the allosteric effect. The allosteric pathway identified using the cutoff value as 12 Å is illustrated in Figure 1(g), because this value was used in another allosteric pathway analysis [26] and also used as the cutoff value for non-bonded interaction in the MD simulations. Overall, different cutoff values leading to different allosteric pathways provide the flexibility to identify pathways targeting the interactions within different ranges, and could provide insights into allosteric effects from different aspects.

### Allosteric communities

The allostery could be referred to as the distribution changes related to protein conformation upon perturbations. The influence of perturbation is not equally exerted on each residue. Some residue pairs could be affected more than others upon perturbations as demonstrated in Figure 1. Using the PRE values of the different residue pairs, the residues can be divided into different groups, with which the total PRE value within each group is minimised, and the total PRE values across different groups are maximised. These groups are named as “allosteric communities” as domains that are less affected by the perturbations.

To construct communities through the minimisation of total PRE value within each community, both GN and KL algorithms as well as the hybrid GN–KL algorithm are implemented in this study. GN algorithm [36,43] has been widely applied in biological and social network community analyses. As described in the methodology section, GN algorithm iteratively removes the most valuable edge in the network to identify the community without minimising the PRE inside the community. As comparison, the KL algorithm[37] is a minimisation algorithm which iteratively reaches local minimum. The total PRE values inside communities using these algorithms are plotted in Figure 2(b). Apparently, the KL algorithm is much better than the GN algorithm.
Table 2. Different pathways with different cutoff values

<table>
<thead>
<tr>
<th>Cutoff value (Å)</th>
<th>Shortest allosteric pathway</th>
<th>Cutoff value (Å)</th>
<th>Shortest allosteric pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>N14, K12, A12, Q83, G82, T81, N80, R79, L78, T77, E76, V75, A74</td>
<td>12</td>
<td>N14, R79, N16, V75, S21, A74</td>
</tr>
<tr>
<td>6</td>
<td>N14, N16, K13, Q83, G82, T81, N80, T77, A74</td>
<td>13</td>
<td>N14, R79, S17, A74</td>
</tr>
<tr>
<td>7</td>
<td>N14, A45, K13, Q83, N16, R79, T81, L78, A74</td>
<td>14</td>
<td>N14, R79, S17, A74</td>
</tr>
<tr>
<td>8</td>
<td>N14, A45, K13, G44, S17, R79, T81, T77, Q73, T70, G25, A74</td>
<td>15</td>
<td>N14, R79, S17, A74</td>
</tr>
<tr>
<td>9</td>
<td>N14, G44, I20, V22, H71, G25, A74</td>
<td>16</td>
<td>N14, L78, K13, A74</td>
</tr>
<tr>
<td>10</td>
<td>N14, G44, S17, V75, G25, A74</td>
<td>17</td>
<td>N14, L78, K13, A74</td>
</tr>
<tr>
<td>11</td>
<td>N14, G44, S21, H71, V22, A74</td>
<td>18</td>
<td>N14, L78, K13, A74</td>
</tr>
<tr>
<td>12</td>
<td>N14, R79, S17, V75, S21, A74</td>
<td>19</td>
<td>N14, L78, K13, A74</td>
</tr>
</tbody>
</table>

Figure 2. Comparison of different community detection algorithms. (a) Total PRE within and between communities using Kernighan–Lin (KL) algorithm; (b) Minimisation of total PRE within allosteric communities using different algorithms; (c) Communities constructed using KL algorithm; (d) Communities constructed using Girvan–Newman (GN) algorithm; (e) Communities constructed using the hybrid GN–KL algorithm. The GN algorithm is effective to determine the suitable number of communities, but could be trapped in local minimum. The KL algorithm could optimise the communities significantly with high computational cost. The hybrid GN–KL algorithm is both computationally efficient and rigorous with the results similar to the KL algorithm.

to identify communities with the minimum PRE values. However, the computational cost of KL algorithm is much higher than the GN algorithm. Overall, the hybrid GN–KL algorithm could produce comparable results to the KL algorithm with much lower computational cost.

As one of its advantages, the GN algorithm is parameter-free, and could be used to determine the optimal number of allosteric communities with maximum modularity of the network [36]. Applying the GN algorithm, it was determined that five communities are the most suitable for PDZ2. Community analysis using GN, KL, and the hybrid GN–KL algorithms are illustrated in Figure 2(c–e), respectively. Usually, the allosteric effects induced by external perturbations alter the protein conformation without changing the secondary structure. Therefore, stable secondary structures including α-helices and β-strands likely belong to same community. Overall, most α-helix and β-strand secondary structures are conserved in the community analyses.

For five communities in PDZ2 domain using KL algorithm (Figure 2(a)), the percentage of total PRE values of all residues pairs within each community are only 0.8%, 1.2%, 0.9%, 1.0% and 1.3% of the overall total PRE values of PDZ2 upon peptide binding as allosteric perturbation, respectively. Therefore, the PRE values among these communities account for 94.8% of total PRE values related to protein allostery. The total PRE value between communities 2 and 4 accounting for 19.0% and the one between communities 4 and 5 accounting for 18.0%. Actual total PRE value for each community pair is listed in Table S3, and the residues in each community are listed in Table S4. The community 4 (residues 66–80) is potentially the most important, and contains β5/α3 loop (66–70) and entire α3 helix (71–79).

The community analysis is further evaluated through comparison with the principal component analysis (PCA). First, the simulations of the PDZ2 unbound and bound states are projected onto the two main
components (PC1 and PC2) from PCA (Figure 3(a)). Clustering analysis reveals that two states of PDZ2 are significantly different in the PC1/PC2 space. Consequently, all community pairs (including self-pair) from different states are also projected onto the principal component space (Figure 3(b–p)). It should be noted that for each community pair including self-pairs, PCA was carried out separately to construct PC1/PC2 surface for projection specifically for that community pair. All community self-pairs do not show significant distribution changes between two states (Figure 3(b,g,k,n,p)). Other community pairs generally show significant differences between two states with the most significant changes coming from pairs including 1:3, 2:3, 2:4, 3:5, and 4:5.

Through this community analysis, the distribution shifting upon peptide binding as PDZ2 allostery can be quantified as the correlation among the allosteric communities. This community analysis provides a quantitative tool with statistical significance to quantify the distribution changes induced by allosteric perturbation from different regions in the protein.

Discussion

The REDAN model approaches protein allostery based on the population shift concept through relative entropy measurement, and can quantitatively measure the difference between two probability distributions [15]. Based on MD simulations, a distribution could be obtained for many collective variables to represent their free energy profile. Relative entropy could be calculated to measure the response of any collective variables with regard to allosteric perturbations. Higher relative entropy indicates a larger change of distributions upon perturbations and could be closely related to allostery. Therefore, the relative entropy could be considered as the amplitude of allosteric effect.

The REDAN method could be used to identify the most affected residues and residue pairs upon allosteric perturbations. In PDZ2 domain, the $\alpha_C$ pair distance with the highest PRE reveals that the distance distribution between $\beta_1/\beta_2$ loop and $\alpha_3$ helix is significantly affected by the peptide binding. The significance of $\beta_1/\beta_2$ loop has been identified in many studies related to PDZ2 allostery [28,41]. In a dynamical interaction correlation analysis conducted by Karplus and coworker [28], the loop $\beta_1/\beta_2$ is referred to as a key part in the allosteric pathway. Another study also emphasised the importance of $\beta_1/\beta_2$ loop through structural network and elastic network analysis [41]. For each individual residue, the summation of all PRE values between this particular residue and all other residues can be considered as a metric to measure the total amount of information passing through this residue upon perturbation. The residues with the highest total PRE values also have a significant
agreement with those network or experimental studies [28,29,40,41].

Comparing with individual residues, potential allosteric pathways are more informative to demonstrate the allosteric mechanisms. The shortest pathway algorithms were applied to identify the pathways between two distal residues with significant PRE. Through pathway decomposition analysis, the large allostatic effect between two distal residues could be decomposed into several short-range residue pairs with smaller PRE values. These short-range residue pairs may provide structural information important for allostery. This is also supported by other studies, which indicate that multiple pathways may coexist and be responsible for the allosteric effect between two distal residues [42]. Using cutoff value for pathway searching, the REDAN model provides flexibility to explore the allosteric pathways at different scales.

The distribution shift upon allosteric perturbation can be represented as the allosteric communities in the REDAN model. The allosteric communities are constructed through the minimisation of total PRE values within each community. As shown in Figure 3, the distribution changes within each community are insignificant, and the majority of distribution differences come from across communities. Therefore, the amount of distribution changes upon allosteric perturbation is quantified as the interactions among different communities.

The construction of allosteric communities is not a trivial task since searching communities with minimum total relative entropy is known as an NP-Hard problem. In this study, widely applied GN and KL algorithms are shown to be suitable for the purpose of allosteric community analysis. The GN algorithm [36] can determine the optimal number of communities based on the modularity of remaining network after decomposition, without explicitly minimising the total PRE in each community. As comparison, the KL algorithm is an explicit minimisation algorithm, which can obtain a local minimum value of the total relative entropy within each community. But the computational cost of the KL algorithm is significantly higher than the one of the GN algorithm, and the number of communities needs to be pre-determined. The hybrid GN–KL algorithm was developed to take advantage of both algorithms by applying GN algorithm to select communities as an initial guess, and KL algorithm to optimise the partitions. The detailed comparison of these three algorithms is provided in Table S5. Among five allosteric communities identified for PDZ2, the community 4 has a total PRE correlated to the rest of protein as more than 50%, indicating that the peptide binding can significantly alter the interaction of the residues in community 4 (L66 to N80) with the rest of protein.

Community 4 also includes all the residues in β5/α3 loop and α3 helix, which consists of the binding pocket of peptide. This highlights the importance of the peptide binding pocket in the allosteric processes. In general, allosteric community analysis could be utilised to divide the protein residues into different allosteric communities to investigate the allosteric mechanism from a global point of view.

**Conclusion**

The current study introduced a new method named related entropy-based dynamical allosteric network (REDAN) model to quantitatively characterise protein allosteric effects upon external perturbations. Relative entropy was applied to quantify the allosteric effects for pair-wised residues based on the distribution differences. Because the population distribution is directly linked to the free energy, any changes in population distributions essentially reflect the changes of free energy surface due to external perturbations. Adapting the shortest pathway searching algorithms, multiple potential allosteric pathways connecting two distal allosteric residues could be identified. The flexibility of using different cutoff values and identifying multiple allosteric pathways could provide deep insight into protein allostery. The allosteric community analysis could further identify the communities, which hold significant contribution to overall relative entropy among them but have minimum relative entropy within each community. Both GN and KL algorithms, and the hybrid GN–KL algorithm were implemented for community identification. The application of the REDAN model on allosteric PDZ2 protein demonstrates its effectiveness and efficiency for protein allostery analysis. Overall, this method could be applied on any two different protein states upon perturbations, and quantify the impacts from the perturbation on the internal dynamics and function-related residues.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Funding**

This work was supported by National Institutes of Health [grant number GM122013]. Computational time was generously provided by Southern Methodist University’s Center for Scientific Computation.

**ORCID**

Peng Tao © http://orcid.org/0000-0002-2488-0239
References

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. 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. 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. However, biomacromolecules including proteins normally have hundreds to thousands of degrees of freedom. The curse of dimensionality induces the difficulties for many analyses of long-time MD simulations, including extracting the important essential motions, clustering different states based on kinetics or structures, visualization of the free energy landscape, etc. 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. 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. In many cases, describing important dynamics using several predefined collective variables (CVs) is an efficient approach. 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. Compared with small molecules, proteins have higher dimensionality, and inappropriate CVs could disguise protein kinetic barriers. Valid CVs should be suitable to capture key dynamical events in simulations in order to obtain meaningful insight. Natural contacts, root-mean-square deviations (RMSDs), radius of gyration (Rg), and structural reaction coordinates including P and Q values 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 and are normally categorized as either linear or nonlinear. The coordinates employed in the linear dimensionality reduction methods are linear combinations of

Received: June 27, 2018
Published: September 25, 2018
input variables, including principal component analysis (PCA), also referred to as quasi-harmonic analysis in MD simulations,\textsuperscript{21} and time-structure based independent component analysis (t-ICA).\textsuperscript{4} Nonlinear dimensionality reduction methods construct coordinates as a nonlinear function of the input variables, including diffusion map,\textsuperscript{22} isomap,\textsuperscript{20} autoencode neural networks,\textsuperscript{23} etc. A thorough comparison for diffusion map,\textsuperscript{22} isomap,\textsuperscript{20} the locally linear embedding (LLE) method,\textsuperscript{19} and PCA was reported in a previous study.\textsuperscript{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.\textsuperscript{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.\textsuperscript{26} Previous studies suggest that t-ICA has better performance than PCA for extracting the slowest dynamical modes.\textsuperscript{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\textsuperscript{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.\textsuperscript{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.\textsuperscript{51} 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\textsuperscript{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,\textsuperscript{53} such as gene expression analysis,\textsuperscript{54} single-cell visualization\textsuperscript{55} and cell types detections.\textsuperscript{36}

With some promising development,\textsuperscript{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.\textsuperscript{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.\textsuperscript{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\textsuperscript{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)\textsuperscript{43} with the IDs as 2PD7 and 3R8H, 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 adenose 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.\textsuperscript{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)\textsuperscript{45} and neutralized with sodium cations and chloride anions. Initially, 10 ns of isobaric–isotropic 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.\textsuperscript{46} All simulations were carried out using the CHARMM\textsuperscript{47} simulation package version 41b1 with the support of GPU calculations based on OpenMM.\textsuperscript{48}

#### Relaxation Time Scale

MSMBuilder\textsuperscript{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,\textsuperscript{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

$$t(r) = -\frac{r}{\ln \lambda_2}$$  \hspace{1cm} (1)

where $\lambda_2$ is the second eigenvalue, and $r$ 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:

$$\text{RMSD} = \frac{1}{N} \sqrt{\sum_{i=1}^{N} (r_i^0 - r_i)^2}$$  \hspace{1cm} (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 quantify the sampling following a previous study.\textsuperscript{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

$$C_{ij} = \frac{c_{ij}}{c_{ii}^{1/2} c_{jj}^{1/2}} = \frac{\langle \tau_i \rangle_{\tau_j} - \langle \tau_i \rangle_{\tau_j}}{[\langle \tau_i^2 \rangle_{\tau_j} - \langle \tau_i \rangle_{\tau_j} - \langle \tau_i \rangle_{\tau_j}^{1/2}]}$$  \hspace{1cm} (3)

where $C_{ij}$ is the measure of correlated movement between atoms $i$ and $j$, $c_{ij}$, $c_{ii}$, $c_{jj}$ and $c_{ij}$ are the correlation matrix elements, and $\tau_i$ and $\tau_j$ 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 $x(t) = (x_1(t),...,x_n(t))$, t-ICA is performed by solving the following generalized eigenvalue problem

$$\mathbf{C} = \mathbf{CF}$$  \hspace{1cm} (4)

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

$$\mathbf{C} = \langle (\mathbf{x}(t) - \langle \mathbf{x}(t) \rangle)(\mathbf{x}(t) - \langle \mathbf{x}(t) \rangle) \rangle$$  \hspace{1cm} (5)

$$\mathbf{\bar{C}} = \langle (\mathbf{x}(t) - \langle \mathbf{x}(t) \rangle)(\mathbf{x}(t+\tau) - \langle \mathbf{x}(t) \rangle) \rangle$$  \hspace{1cm} (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,\textsuperscript{51} 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 $p_{i|j}$. 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 $\sigma_i$. The probability of $x_i$ to be selected as the neighbor of $x_j$ is a conditional probability calculated as

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_j^2)}{\sum_{k\neq i} \exp(-\|x_i - x_k\|^2/2\sigma_k^2)}$$  \hspace{1cm} (7)

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

$$q_{ij} = \frac{(1 + \|x_i - y_j\|^2)^{-1}}{\sum_{k\neq i} (1 + \|x_i - y_k\|^2)^{-1}}$$  \hspace{1cm} (8)

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

$$C = KL(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$  \hspace{1cm} (9)

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

One remaining parameter to be selected is the bandwidth of Gaussian distribution $\sigma_i$ 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 $\sigma_i$ could be used for all data points. A binary search of $\sigma_i$ is carried out for each data point to match a fixed hyperparameter "perplexity" that is specified by users. This perplexity is defined as

$$\text{Perp}(P) = 2^{-\sum_i \log_2 p_i}$$  \hspace{1cm} (10)
For each data point $x_i$, $\sigma_i$ 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.\(^{51}\)

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 $\mu$ as an integer number, only the $3 \mu$ nearest neighbors for each data point $x_i$ are considered. For the structures not belonging to the $3 \mu$ 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 $\mu$s 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 $> 2D$ t-ICA $> 2D$ 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
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

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

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

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.
The number of clusters is chosen as ten, because it is the smallest number of clusters to achieve RMDSdiff 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 RMDSs for all cluster pairs for these ten states are shown in Figure 2a. It is clear that RMDSdiff (1.869 Å) is significantly larger than RMDSsame (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,30 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 slowdowns 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.

**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, 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
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 5. t-SNE 2D projection free energy surface using different perplexity \( \mu \) values. The joint probability calculated for 3 \( \mu \) 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.
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 $k_B T$. In the bonded configuration, the sampling of this region is increased, leading to a lower free energy barrier around 4 $k_B T$. 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.
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.53–56 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 3N-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...
maximum structural and dynamical information. Recently, some nonlinear dimensionality reduction methods including diffusion map, isomap, autoencode, and time-lagged autoencoder 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. 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. 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. 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. 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). 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 allosteric 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 $\sim 1 k_B T$. 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 an 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.
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


(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.


PAPER VII

RESEARCH ARTICLE

Allosteric mechanism of the circadian protein Vivid resolved through Markov state model and machine learning analysis

Hongyu Zhou¹, Zheng Dong¹, Gennady Verkhivker²,³, Brian D. Zoltowski¹, Peng Tao¹*¹

1 Department of Chemistry, Center for Scientific Computation, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas, United States of America, 2 Graduate Program in Computational and Data Sciences, Schmid College of Science and Technology, Chapman University, Orange, California, United States of America, 3 Chapman University School of Pharmacy, Irvine, California, United States of America

* ptao@smu.edu

Abstract

The fungal circadian clock photoreceptor Vivid (VVD) contains a photosensitive allosteric light, oxygen, voltage (LOV) domain that undergoes a large N-terminal conformational change. The mechanism by which a blue-light driven covalent bond formation leads to a global conformational change remains unclear, which hinders the further development of VVD as an optogenetic tool. We answered this question through a novel computational platform integrating Markov state models, machine learning methods, and newly developed community analysis algorithms. Applying this new integrative approach, we provided a quantitative evaluation of the contribution from the covalent bond to the protein global conformational change, and proposed an atomistic allosteric mechanism leading to the discovery of the unexpected importance of A’α/Aβ and previously overlooked Eα/Fα loops in the conformational change. This approach could be applicable to other allosteric proteins in general to provide interpretable atomistic representations of their otherwise elusive allosteric mechanisms.

Author summary

Allostery is an important but elusive property that governs critical functionality of many proteins. Quantitative analysis is needed to provide significant insight into protein allostery and lead to better prediction power of this ubiquitous phenomenon. We developed machine learning methods based on robust Markov state model to delineate allosteric mechanism of Vivid as an allosteric protein in the filamentous fungus Neurospora crassa, regulating circadian rhythm of this organism. We accurately reconstructed the equilibrium distributions for two allosteric configurations of Vivid, and determined structural differences among these states. Intriguingly, the novel community analysis derived from machine learning methods reveals the importance of two loop regions for Vivid allostery through quantitative evaluations with statistical significance.
Introduction

Light, oxygen, or voltage (LOV) domains are small, commutable elements that couple blue-light activation to protein conformational changes for blue-light responses in bacteria, archaea, fungi and plants. One common feature shared by all LOV domains is that a cofactor, flavin adenine dinucleotide (FAD), flavin mononucleotide (FMN) or riboflavin,[1] forms a covalent bond with a conserved cysteine residue upon external light activation. Covalent adduct formation, subsequently facilitates a large conformational change in the protein leading to alteration in enzyme activity and/or protein–protein interactions.[2, 3] The mechanism of LOV domain global conformational change induced by covalent bond is widely accepted as an allosteric process and remains as a focal point of LOV domain studies with an aim of developing novel optogenetic tool through manipulating LOV domain allostery.

Vivid (VVD) is a LOV-domain containing photoreceptor from the filamentous fungus Neurospora crassa that modulates circadian rhythms in this organism. In Neurospora, circadian-clock regulated gene expression is dictated by a heterodimeric complex involving the photosensitive protein White Collar-1 (WC-1) and the non-photosensitive protein WC-2. Upon blue-light exposure it is believed that an additional copy of WC-1 is recruited to light-responsive elements (LRE’s) to form a hetero-trimeric complex involving a WC-1 homodimer (WCC). One of the blue-light induced gene products is VVD, which competes for binding to WC-1 to disrupt the WCC and modulate light-induced gene expression.[4] VVD activity is dependent upon activation through blue light, where Cys108 of VVD forms a covalent bond between its sulfur and C4a position of co-factor FAD, which in-turn induces N terminus conformational change of VVD necessary for WCC regulation. Based on previous experimental and computational studies, VVD serves as a good candidate for optogenetic tool developing. However the mechanism of VVD allostery correlated with its global conformational change upon covalent bond formation is still an open question.

Investigating protein allostery through computational means is effective and is under constant development. Various studies were conducted to reveal the underlining mechanisms for different allosteric proteins.[5–7] One method developed, named rigid residue scan (RRS), has been applied in several systems including PDZ[8] and PDZ2 domains.[9] Other methods including dynamical network analysis[10], elastic network model[11], relative entropy based allostery network[12], sequence and structural analysis[13], correlation based analysis[14] etc. have been widely applied on many allosteric systems. Markov state model (MSM) of molecular kinetics has been widely used in recent years to estimate long-time kinetic information from short trajectories.[15–17] Molecular dynamics (MD) simulations often involve long time samplings or enhanced sampling to detect rare events with statistical significance.[17] Implementing predefined order parameters as reaction coordinates is a useful means to analyze protein simulations. However, the possibilities for neglecting the true kinetic information underlying the simulations with hidden important barriers remain as one intrinsic limitation of the predefined order parameters approaches.[18, 19] Although Markov state analysis could be applied to separate the structures based on their kinetic information, a quantitative strategy to measure the primary differences among different states is still absent. Long-time scale molecular dynamics simulations could provide sufficient sampling of the conformational landscape of proteins. But obtaining statistically significant insights into protein dynamics from massive simulation datasets presents a major challenge[20–22].

Artificial neural networks (ANN)[23, 24], tree based model including decision tree (DT)[25], and random forest (RT)[26] are widely applied as classification methods in machine learning. ANN mimics neural networks consisting of neurons in the brain, and has been applied in many classification problems to achieve high accuracy[27]. Decision tree was
constructed to quantify the importance for making decisions or predictions of each dimension in the input data that is statistically relevant to relative entropy metrics in distinguishing between different distributions.[28–30] Recently, we demonstrated significant effectiveness of DT and ANN methods to build allostery classification models and identify allosterically important residues.[31, 32]

To gain further insight into VVD allostery mechanism, more quantitative description would be necessary in addition to normal qualitative analysis of protein allostery. In this work, we developed a novel computational framework that can significantly boost the applicability of molecular simulation techniques to probe dynamic allostery in protein systems, and applied this approach on VVD. Specifically, we combined machine learning and dynamic community analysis of the residue interaction networks to obtain robust quantitative descriptions of conformational ensembles and protein states, and to rigorously correlate variations in conformational ensembles to underlying allosteric mechanisms. Both methods are enabled by a new application of machine learning and network modeling to the analysis of thermodynamic and kinetic information from MSM. The proposed models are applied to (a) rapidly recognize and identify structural and dynamic patterns of complex conformational ensembles; (b) identify key functional states that are hidden in the conformational ensembles, and (c) reconstruct the mechanisms of dynamics driven allostery through integration of machine learning and network analysis.

Using the proposed computational framework, we examined allosteric mechanisms of VVD and verified the impacts of the key local covalent bond upon photo excitation to global motions of VVD, and revealed the importance of A'α/Aβ and Eα/Fα loops in the conformational change. A good agreement between our analysis and experimental observations of VVD validated the applicability of the proposed approach, and provided structural insights into mechanism of conformational changes and allostery in allosteric proteins. Our methodology could facilitate the usage of VVD as an optogenetic tool by providing quantitative measurement of individual residues’ contribution to protein allostery.

Results

Markov states analysis of VVD simulation

There are two native crystallographic structures of VVD: dark structure (without covalent bond formed between FAD and VVD residue Cys108) and light structure (with such covalent bond). We referred to these two states as native dark (non-bonded) and native light (bonded) configurations (Fig 1). To probe the response from protein with regard to the covalent bond between FAD and VVD, two new configurations were constructed: dark structure with the above covalent bond, and light state without the above covalent bond. We referred to these two states as transient dark (bonded) and transient light (non-bonded) configurations, respectively. Three independent 1 μs simulations were carried out for each configuration, leading to 12 μs production trajectories. All 12 μs trajectories of VVD are projected onto a two dimensional (2D) plot of root-mean-square deviation (RMSD) of VVD backbone alpha carbon atoms (Ca) with reference to the native dark and light structures, respectively (Fig 2A). The plot reveals that the simulations of light state configurations may reach the native dark state structure. On the contrary, the simulations of dark state configurations show less fluctuation than the light state configurations in simulations.

In order to apply MSM analysis, k-means clustering analysis was applied to divide the sampling space into 300 microstates based on structural differences (Fig 3A). The transition probabilities were estimated among microstates at a specified interval of time named as lag time. An adequate lag time should be selected based on the convergence of the estimated relaxation
The data plotted in Fig 2B suggest that the estimated timescale is converged after 30 nanoseconds (dashed grey line), which is chosen as the lag time for the MSM analysis. The number of macrostates should be rigorously chosen to better represent the free energy landscape. Overall, having eight macrostates will result in the best separation to represent kinetically meaningful states on the free energy surface as shown in Fig 3A.

Perron-cluster cluster analysis (PCCA) was applied to map microstates onto macrostates based on the eigenfunction structure of transition probability matrix (Fig 3A and 3B). The representative structure for each macrostate is illustrated in Fig 3C. The averaged RMSDs of the macrostates 2, 3 and 7 with reference to the crystal dark and light structures of VVD are 2.84Å and 4.38Å, respectively. Similarly, the averaged RMSDs for the macrostates 1, 4, 5 and 8 are 4.69Å and 3.31Å with reference to the crystal dark and light structures, respectively. Locating at the top right corner of the 2D RMSD plot, macrostate 6 is far away from both crystal dark and light conformations of VVD, with the averaged RMSD values as 6.47Å and 4.62Å.
respectively. As the sampling of macrostate 6 only occurs in transient light configuration simulations, the macrostate 6 is referred to as a “hidden” state, which cannot be reached in the simulation of native states.

To assess the effectiveness of the above analysis for VVD, we carried out time-structure independent components analysis (t-ICA) and principal component analysis (PCA) to verify that the markovian property is well maintained using MSM. For comparison purpose, 30ns as the lag time and total of eight macrostates are used for both t-ICA and PCA. The results for the comparison shown in S1(A)–S1(C) Fig represent the projection of VVD simulations onto the surfaces of 2D-RMSD, t-ICA and PCA, respectively. The relaxation timescales of these MSM models are shown in S1(E)–S1(G) Fig. The relaxation timescale estimated from t-ICA is significantly higher than the ones with 2D-RMSD and PCA, which indicates that t-ICA may capture slow kinetic components better than 2D-RMSD and PCA. However, the connectivity of microstates on the projection surface of t-ICA is lower than the one on 2D-RMSD or PCA surfaces. Thus the identified “strongest connected subgraph”[34] on the t-ICA surface does not contain all microstates. Based on the ergodic cutoff criterion, during the construction phase of MSM, 173 out of 300 microstates on t-ICA surface were discarded because they are weakly connected to other microstates. The highly disconnected communities indicate that the major t-ICs could be the spurious collective variables due to high dimensionality. Therefore, we selected top five features identified in Table 1 (which will be discussed later) and construct t-ICA surface only on those features. The projection surface is shown in S1(D) Fig and the relaxation timescale is shown in S1(H) Fig. Microstates grouped in eight macrostates on 2D-RMSD, t-ICA, PCA, and t-ICA with five features are illustrated in S1(I)–S1(L) Fig, respectively. With only five selected features, the t-ICA results are much improved, suggesting that t-ICA method works better with reduced dimensionality. The results of 2D-RMSD, PCA, and t-ICA with five selected features are similar with each other. However the axis of 2D-RMSD could better represent structural information than PCA or t-ICA with direct measurement of difference from the two key structures of VVD.

### Table 1. Top 5 features in overall importance to distinguish difference states.

<table>
<thead>
<tr>
<th>features</th>
<th>OVERALL</th>
<th>Rank: 1</th>
<th>Rank: 2</th>
<th>Rank: 3</th>
<th>Rank: 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>213 (T38 –G105)</td>
<td>2.83%</td>
<td>S3-S5: 4.28%</td>
<td>S1-S7: 4.03%</td>
<td>S5-S6: 4.00%</td>
<td>S2-S6: 4.00%</td>
</tr>
<tr>
<td>227 (T38 –K119)</td>
<td>2.38%</td>
<td>S2-S4: 4.55%</td>
<td>S3-S4: 4.00%</td>
<td>S1-S3: 4.00%</td>
<td>S2-S6: 4.00%</td>
</tr>
<tr>
<td>189 (T38 –K81)</td>
<td>1.75%</td>
<td>S3-S7: 4.28%</td>
<td>S2-S5: 4.08%</td>
<td>S3-S5: 4.05%</td>
<td>S1-S7: 4.02%</td>
</tr>
<tr>
<td>29 (H37 –V67)</td>
<td>1.52%</td>
<td>S1-S7: 4.05%</td>
<td>S3-S6: 4.00%</td>
<td>S2-S8: 3.99%</td>
<td>S3-S8: 3.99%</td>
</tr>
<tr>
<td>355 (L39 –E102)</td>
<td>1.51%</td>
<td>S5-S6: 5.72%</td>
<td>S4-S7: 5.48%</td>
<td>S1-S6: 5.31%</td>
<td>S4-S5: 4.44%</td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pcbi.1006801.t001
To ensure the kinetic similarity within the microstates during the clustering, the averaged RMSD in each microstate is plotted for 2D-RMSD, t-ICA, PCA and t-ICA with five selected features models in **S2 Fig**, respectively. It is assumed that the conformations with small RMSDs may interchanged quickly. Practically, averaged RMSD inside each microstate smaller than 2.0Å is sufficient to imply the kinetic similarity within that microstate.[33] The averaged RMSDs are smaller than 2.0Å for all microstates using three models, indicating that the kinetic similarity within each microstate is well maintained. In addition, the markovian property using 2D-RMSD was also tested using Chapman-Kolmogorov test [17] by comparing the probability directly observed in the simulation with the estimated probability using lag time as 30ns (**S3 Fig**). To avoid spurious large error bar due to the difference of saving coordinates frequency between the reference study (0.2ps) and the current study (100ps), the denominator in the reference paper [17] equation 66 was replaced by the ratio of all transition count to the actual transition counts in the simulation. Therefore, the error bar is less dependent on the saving frequency of the simulation. The similarity between these two probabilities shown in **S3 Fig** suggests that the markovian property using 2D-RMSD as reaction coordinates for MSM model is well maintained.

After the construction of MSM, the transition probabilities estimated among adjacent macrostates are shown in **Fig 4**. For each state, the probability to remain in the current state is higher than switching to other states, which suggests that each macrostate is a minimum on free energy surface, and the kinetic barriers prevent the switching to other states. The above transition probability matrix was calculated based on all 12 μs MD trajectories. To further explore the cofactor covalent bond effect, the transition probability matrices were calculated separately for six non-bonded (for native dark and transient light configurations) and six bonded trajectories (for native light and transient dark configurations). As shown in **Fig 4B and 4C**, forming of covalent bond has significant impact on the transition probabilities among macrostates, which suggests that the covalent bond could alter the free energy surface and energy barriers among different states.

The steady state distribution that the system may reach at the infinite time could be estimated based on the calculated transition probabilities. The eigenvector associated with the eigenvalue 1.0 for the transition probability matrix is the stationary distribution for each state. This is only an approximation, because after discretizing the phase space into microstates, the markovian properties may not hold precisely.[35] However, it is still valuable to investigate the distribution at infinite time (referred to as steady state thereafter) to obtain an overall picture regarding to the long time behavior. The steady state distributions based on the non-bonded

---

**Fig 4. The transition probabilities estimated based on Markov state model among macrostates.** (a) The transition probabilities estimated using all trajectories; (b) The transition probabilities estimated using non-bonded trajectories; (c) The transition probabilities estimated using bonded trajectories. State 6 only appears in the non-bonded configurations. State 8 only appears in the bonded configurations. The transition probability among macrostates are significantly different, showing that the formation of photo-induced covalent bond could mediate the kinetics of the system.

https://doi.org/10.1371/journal.pcbi.1006801.g004
(Fig 5A) and bonded trajectories (Fig 5B) are illustrated separately. For comparison purpose, the distributions based on non-bonded and bonded MD trajectories, which are referred to as ensemble distributions, are illustrated in S4 Fig. Overall, the steady state distribution differences between non-bonded and bonded configurations are significant (Fig 5C). The light state sampling is significantly enhanced in the bonded configuration, primarily in the state 8. In the non-bonded configuration, the samplings of dark state conformations including states 2, 3, and 7 are more extensive than the bonded configuration (Fig 5C). The hidden state (state 6) is only sampled in the bonded configurations. Similar with the Fig 4B and 4C comparison, these results indicate that the bonded configurations favor the native light state structure and the non-bonded configurations favor the native dark state structure. The convergence of simulations is verified by RMSD and configurational entropy in S5 Fig. The plot of configurational entropy (S5B Fig) indicates that the simulations are well converged after ~600ns samplings.

The sampled conformational space in the 12µs MD simulations has a remarkable agreement with our previous study of VVD through perturbed MD simulations. Our analyses show that the MD simulations of transient light and dark configurations sampled larger conformational space than the native light and dark configurations. These results conform that the covalent bond could facilitate the conversion of dark state to the light state, agreeing with the experimental observations. Markov state analysis provides more quantitative descriptions than intuitive interpretation based on sampling space. The covalent bond affects the transition probabilities among macrostates and the steady states/equilibrium distributions significantly. The steady states distribution can be considered as the free energy for each state. Therefore, the changes of steady states distribution can be regarded as the changes of free energy surface of the protein dynamics due to the formation of covalent bond. The transition probabilities between state3-state2, state7-state5, state5-state4 increase from (3%, 4%, 11%) to (14%, 10%, 14%), respectively, comparing the non-bonded and bonded configurations. The increase of the transition probability is significant, which lead to the estimated steady states distribution differences in dark and light states. These differences suggest that the transitions from dark state to light state could be triggered mainly by the formation of covalent bond without excitation energy dissipation.

Another difference is the behavior of conformations dwelling in state 1. Based on the non-bonded configuration simulations, starting from state 1, the probability for protein directly changing to state 8 is 0%, and to state 6 is 16%. In the bonded configuration simulations, these probabilities are 6% and 0%, respectively. Meanwhile, state 6 is regarded as a hidden state as it was not sampled in either native light or native dark configurations, and is structurally

---

**Fig 5. The estimated steady state distributions for each state based on Markov state model.** (a) Steady state distribution based on non-bonded configuration simulations; (b) Steady state distribution based on bonded configuration simulations; (c) Steady state distribution differences between non-bonded and bonded configurations. States 2, 3, and 7 are considered as dark states region. States 1, 4, 5, and 8 are considered as light state region. State 6 is considered as hidden region.

https://doi.org/10.1371/journal.pcbi.1006801.g005
different from both crystal dark and light states. The different behavior of state 1 in non-bonded and bonded configurations could be interpreted as a stabilization effects to the light states from the covalent bond. With the covalent bond, the light state can be stabilized to the state 8 conformations without reaching hidden state 6.

Based on the above observations, we can hypothesize that the covalent bond has a significant role in light state conformation. With this covalent bond, the light state conformation could be stabilized. Otherwise it would be trapped in the hidden state conformations, which cannot be sampled in the native states. Our results demonstrated that even without energy available in an excited state upon blue light excitation, the single covalent bond could trigger the global conformational change of VVD.

The MSM analysis reveals the impacts of covalent bond between cofactor and protein in conformational distribution and free energy barriers among macrostates. However, the conformational characteristic for each state and the mechanism for the conformational changes are still unclear. Therefore, several supervised machine-learning models were applied to study the intrinsic structural properties of macrostates.

**Machine learning identifies key structural features and interactions that characterize allosterically important protein macrostates**

To apply the supervised machine learning models and study the structural differences, appropriate collective variables are needed to describe a protein structure. For the small organic molecules, several descriptors including the topological torsion[39], reduced graph descriptor[40] have been developed, and widely used in quantitative structure–activity relationship (QSAR) and docking studies.[41, 42] Here, we chose the pair-wised distances of alpha carbon (Cα) of amino acids as translation and rotation invariant collective variables for protein structures in our simulations. Total 10,878 pair-wised distances were constructed based on 148 residues. For each simulation, frames are saved for every 100 picoseconds (ps), resulting in 10,000 frames for every 1 µs MD trajectory. Therefore, 120,000 "data points" with 10,878 features were extracted from 12 µs MD trajectories. Above macrostate analysis were used to label each frame. After the preparation of data, decision tree, random forest, one-vs-one random forest and artificial neural networks models were applied to distinguish the intrinsic conformational differences among macrostates. Dimensionality deduction was done by one-vs-one random forest before applying artificial neural networks model. Each machine learning model is described in Methodology section. The cross-validation was applied to refine the parameters of these models. The training and testing error of 12-fold cross-validation are plotted in Fig 6. The final selected parameters are indicated by dashed vertical line in each subplot.

The results for optimized machine learning models and a dummy classifier are shown in Fig 7A. Dummy classifier was generated based on random guesses.[43] The training accuracy for neural networks, decision tree, random forest and one-vs-one random forest models are 95.0%, 98.3%, 98.1%, and 99.1%, respectively. The validation accuracy for the artificial neural networks is the highest, having a mean value of 90.1%. Two random forest classifiers and decision tree classifier have relatively lower performances but still significantly higher than the dummy classifier as control. These indicate that the models are able to catch the structural characteristic of each Markov state using the pair-wised Cα distances.

Although artificial neural network model provides the highest classification accuracy, tree-based methods were chosen for further analyses, because these methods could evaluate the contribution from each pair-wised Cα distance. Especially, one-vs-one random forest was applied to compute the feature importance for any two different states pair by performing a random forest classification just between these two states. Therefore, for any two different
macrostates, one distinct random forest classifier was built. The combination of 28 basic random forest classifiers, which were calculated as $N^2(N-1)/2$ for eight ($N=8$) macrostates, were constructed for pair-wised macrostates classification. Overall, this method provides a very effective model, in which 367 features out of 10,878 features account for 90% distinguishability (Fig 7B).

Top five ranked features computed by the one-vs-one random forest classifier are listed in Table 1. The overall importance of features was calculated by the average of the 28 selected random forest classifiers feature importance. The Cα distances between T38 and G105 is identified as the top feature with the averaged importance as 2.83%. Specifically, it has 4.28%, 4.03%, 4.00% and 4.00% feature importance in distinguishing between States 3 and 5, States 1 and 7, States 5 and 6, States 2 and 6, respectively. The distributions in eight macrostates of top two features are plotted in Fig 8A and 8B, respectively. The most states distributions are well separated based on these two top features. States 2, 3, and 7, which are regarded as ‘dark state’ regions, have shorter distance distributions in both T38-G105, and T38-K119 pairs. States 5, 8, 1 and 4, which are regarded as ‘light state’ regions, have much longer distance distributions.
than the ‘dark state’ regions. The ‘hidden’ State 6 has the largest distance in both features. For comparison purpose, one of the lowest importance (0.00%) features as feature 4976 between L75 and R169 is also plotted (Fig 8C). For the low ranked features, the distributions for all macrostate are very similar, indicating that those distances are not affected by covalent bond formation as intrinsic allosteric effect.

Comparing with decision tree and random forest, one-vs-one random forests model has at least two advantages. One-vs-one random forest model could provide feature weights specifically for any two different states and is unbiased for features. Overall, only 367 out of 10,878 features have more than 90% distinguishability in one-vs-one random forest model (Fig 7B). The feature importance in one-vs-one random forest model could directly represent the distribution differences between any state pairs for a particular distance. Top ranked features in this model have distinctive distributions in different states (Fig 8A and 8B), while the low ranked features have indistinguishable distributions in all states (Fig 8C), even though those residues are rather far away from each other. This demonstrates the effectiveness of machine learning

**Fig 8. Distributions of selected features (pair-wised Cα distances) in different states.** (a) Distribution of Feature 213 as T38/G105 distance in all eight macrostates. This is the top one with overall 2.83% importance. (b) Feature 227 as T38/K119 distance in all eight macrostates. This is the second most important feature with overall 2.38% importance. (c) Feature 4976 L75-R169 with overall 0.0% importance. This unimportant feature has the similar distribution in all eight macrostates.

https://doi.org/10.1371/journal.pcbi.1006801.g008
methods to select features and residues closely related to allosterically important macrostates of VVD.

Although the above structural differences among macrostates revealed through machine learning analysis are informative, some disadvantages do exist. First, the top features could be correlated with each other as the top two features sharing the same residue T38. Second, it could be misleading if only a limited number of top ranked features are selected for investigation since the feature importance between the top ranked and low ranked features are insignificant. Even for top ranked feature 4 and 5 in Table 1, the differences of importance are less than 0.01%. Third, the residues associated with the top features intend to be far away from each other. It is difficult to differentiate these long-distance distribution differences as either directly being correlated with key residues interactions or the result from accumulation of some function related short-range interactions. In addition, the important short-range interactions would have low feature importance, because their distinguishability may not be as significant as the long distance distributions. Therefore, instead of focusing on the residues associated with the top ranked features, we further developed community analysis with more statistical significance.

**Machine learning-driven community analysis specifies a’α/aβ and εα/βα loops as allosteric molecular switches between dark and light states**

Inspired by dynamics network analysis[44], the machine learning based community (referred to as ML community) analysis was developed to divide residues into several groups so that the feature importance for pair-wised Cα distances among groups is maximized, while the feature importance within each group is minimized. The detailed algorithm to construct ML communities is described in the Methodology section. As shown in Fig 9A, with the number of ML communities increasing, the feature importance for pair-wised Cα distances within ML communities increases. Applying an elbow criterion, four ML communities were selected with the total feature importance within each ML community accounting for 0.56% and total feature

![Fig 9. Community analysis of VVD based on one-vs-one random forest model. (a) Total feature importance among ML communities with regard to different number of communities; (b) Four ML communities named Commu. A, Commu. B, Commu. C and Commu. D as blue, red, yellow and grey colors, respectively. The ultimate number of ML communities is determined as four based on the elbow criterion. The feature importance across different ML communities accounts for 99.44% of total importance. The feature importance between Commu. A (which is mainly N-terminus) and the rest of protein accounts for 89.415%. The feature importance between Commu. B and Commu. C as well as D accounts for 9.103% of total feature importance.](https://doi.org/10.1371/journal.pcbi.1006801.g009)
importance among ML communities accounting for 99.44%. Therefore, the further analysis focuses on the distribution changes among ML communities, neglecting the distribution differences within each ML community, to reveal the overall dynamics associated with ML communities in each configuration. All residues belonging to each ML community are listed in S1 Table in Supporting Information, and plotted in Fig 9(B) named as Commu. A, B, C and D.

Comparing with the secondary structures shown in Fig 9B, Commu. A includes N-terminus from H37 to G43. Commu. B includes the loops in A’α/Aβ and Eα/Fα. It should also be noted that the residue C108 that is bonded to the cofactor also belongs to the Commu. B. Commu. C and D comprise the majority of VVD. Commu. C includes the majority of Aβ strand, Bβ strand, Dα helix, Fα helix and Gβ strand, and part of cofactor binding sites. Commu. D contains the rest of protein, including A’α helix, Cα helix, Eα helix, Hβ strand, and Iβ strand. The N-terminus and loops are well preserved in Commu. A and B, suggesting specific roles of these two secondary structures.

The accumulated overall feature importance between each ML community pair is listed in Table 2. The correlation between Commu. A (which is mainly N-terminus) and the rest of protein accounts for 89.415% total of feature importance in the one-vs-one random forest classifier. This is not surprising since the N-terminus is the most flexible and distinguishable part between the native dark and light states of VVD. However, it should be noted that after excluding Commu. A, the feature importance between Commu. B and Commu. C as well as D still accounts for 9.103% of total feature importance. This suggests that the position of two loops (A’α/Aβ and Eα/Fα) in Commu. B could play an important role to distinguish each macrostate. Although the Commu. C and D comprise the majority of proteins, the accumulated feature importance between them is less than 1%.

The machine learning based community analysis provides additional information with regard to the different parts of structures during the simulations. In addition to N-terminus, the motions of Commu. B (loops in A’α/Aβ and Eα/Fα) are also significant in distinguishing between the light and dark states. The relative distinguishability of four ML communities associated with key macrostate pairs are listed in Table 3. Besides the N-terminus, the relative position of Commu. B with Commu. C/D is also important to distinguish between the adjacent macrostates (bold in Table 3) including the transition from State 3 to State 2, from State 2 to State 7, from State 7 to State 5, from State 5 to State 4, and from State 4 to State 1. However, for two non-adjacent macrostates, the position of N-terminus (Commu. A) is determinative to the states as shown in Table 3.

Based on the above results, we hypothesize that when the photo-induced covalent bond is being formed in the dark state, one possible mechanism for protein going through conformational changes from the dark to light state is that the position of Commu. B changes first and subsequently facilitates the conformational change of N-terminus as Commu. A. This transition sequence may have a higher probability than for N-terminus directly changing to another state as shown in Fig 4(A), as well as in Fig 4(B) and 4(C). Overall, VVD has higher probability to switch to the adjacent macrostate with significant changes in Commu. B and little changes of Commu. A. For example, given a structure as dark state conformation starting from state 3,
Table 3. The changes of Commu. A and Commu. B during transitions between states. (Bolded is state-transitions with large Commu B component).

<table>
<thead>
<tr>
<th>Adjacent macrostates</th>
<th>A with C and D</th>
<th>B with C and D</th>
</tr>
</thead>
<tbody>
<tr>
<td>State 3 (crystal dark) → State 2</td>
<td>38.35%</td>
<td>46.38%</td>
</tr>
<tr>
<td>State 2 → State 7</td>
<td>50.12%</td>
<td>30.22%</td>
</tr>
<tr>
<td>State 3 → State 7</td>
<td>92.65%</td>
<td>0.87%</td>
</tr>
<tr>
<td>State 7 → State 5</td>
<td>59.85%</td>
<td>19.37%</td>
</tr>
<tr>
<td>State 5 → State 4 (crystal light)</td>
<td>71.15%</td>
<td>17.60%</td>
</tr>
<tr>
<td>State 4 → State 1</td>
<td>50.38%</td>
<td>38.01%</td>
</tr>
<tr>
<td>State 7 → State 4</td>
<td>85.09%</td>
<td>0.32%</td>
</tr>
<tr>
<td>State 1 → State 8</td>
<td>94.97%</td>
<td>0.52%</td>
</tr>
<tr>
<td>State 1 → State 6</td>
<td>74.41%</td>
<td>3.72%</td>
</tr>
</tbody>
</table>

Non-Adjacent macro-states

<table>
<thead>
<tr>
<th></th>
<th>A with C and D</th>
<th>B with C and D</th>
</tr>
</thead>
<tbody>
<tr>
<td>State 2 → State 8</td>
<td>79.71%</td>
<td>0.01%</td>
</tr>
<tr>
<td>State 3 → State 4</td>
<td>91.27%</td>
<td>0.01%</td>
</tr>
<tr>
<td>State 3 → State 6</td>
<td>81.98%</td>
<td>0.01%</td>
</tr>
<tr>
<td>State 3 → State 1</td>
<td>78.13%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pcbi.1006801.t003

The most likely route to go to light state conformation in state 4 is State 3 → State 2 → State 7 → State 5 → State 4 with the probability as 0.15, 0.23, 0.1 and 0.14 in bonded configuration as shown in Fig 4(C). These transitions are shown as bolded state transitions in Table 3 with the highest Commu. B component. Meanwhile, the probability of state 3 directly changing to state 7 or state 7 directly changing to state 4 is much lower as 0.08 and 0.05 in Fig 4(C), and those transitions have larger Commu. A changes as shown in Table 3. These observations indicate that the transition mechanism from dark state to light state with the highest probability is changing the relative position of Commu. B first, instead of changing N-terminus as Commu. A directly. Meanwhile, as shown in Fig 4(B) and 4(C), the bonded configuration has a higher probability to change from the dark to the light conformation than in the non-bonded configuration. Therefore, we hypothesize that the photo-induced covalent bond increases the flexibility of Commu. B comparing to the non-bonded configurations.

To test this hypothesis, the transition pathway theory (TPT) [45] was employed to generate an ensemble of pathways to verify the transition pathway from state 3 (crystal dark conformation) to state 4 (crystal light conformation). Total of 10,017 pathways were generated, and could be grouped as 111 distinct channels floating from state 3 to state 4. The probability of each channel is proportional of the flux through this channel with reference to all channels flux. [45] Overall, the probability for top 20 channels are listed in the Table 4, with the contribution from these channels accounting for more than 98% of total population.

Among all 111 channels, the proposed channel 3–7–5–4 is the third most populated channels with around 15% contribution (red pathways in Fig 10). Only 3–7–5–4 and 3–7–4 channels have higher contribution. The contribution is significant compared with many other pathways, suggesting the importance of the loop movement during the transition between dark and light states. Besides, the RMSF analysis was also conducted. The results shown in S6 Fig suggest that the photo-induced covalent bond could enhance the fluctuation of A’α/Aβ loop, which may facilitate the transition.

To summarize, a general goal of this new community analysis is to minimize the feature weights within each community while maximizing the feature weights among communities. Therefore, we can ignore the internal difference inside each community in different macrostates, and focus on the global differences among communities associated with macrostates.
Table 4. The probability of top 20 channels.

<table>
<thead>
<tr>
<th>Channels</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>State 3–7–5–4</td>
<td>33.099%</td>
</tr>
<tr>
<td>State 3–7–4</td>
<td>21.809%</td>
</tr>
<tr>
<td><strong>State 3–2–7–5–4</strong></td>
<td>14.819%</td>
</tr>
<tr>
<td>State 3–2–7–4</td>
<td>10.369%</td>
</tr>
<tr>
<td>State 3–2–4</td>
<td>4.35%</td>
</tr>
<tr>
<td>State 3–2–5–4</td>
<td>3.567%</td>
</tr>
<tr>
<td>State 3–7–6–1–4</td>
<td>3.008%</td>
</tr>
<tr>
<td>State 3–7–1–4</td>
<td>1.257%</td>
</tr>
<tr>
<td>State 3–5–4</td>
<td>0.983%</td>
</tr>
<tr>
<td>State 3–2–7–6–1–4</td>
<td>0.857%</td>
</tr>
<tr>
<td>State 3–7–6–4</td>
<td>0.694%</td>
</tr>
<tr>
<td>State 3–2–7–1–4</td>
<td>0.492%</td>
</tr>
<tr>
<td>State 3–2–7–6–4</td>
<td>0.464%</td>
</tr>
<tr>
<td>State 3–7–3–1–4</td>
<td>0.443%</td>
</tr>
<tr>
<td>State 3–2–7–5–1–4</td>
<td>0.387%</td>
</tr>
<tr>
<td>State 3–7–5–6–1–4</td>
<td>0.326%</td>
</tr>
<tr>
<td>State 3–2–7–5–8–4</td>
<td>0.307%</td>
</tr>
<tr>
<td>State 3–2–8–4</td>
<td>0.281%</td>
</tr>
<tr>
<td>State 3–7–5–8–4</td>
<td>0.262%</td>
</tr>
<tr>
<td>State 3–2–7–5–8–1–4</td>
<td>0.231%</td>
</tr>
<tr>
<td><strong>Total 20 channels</strong></td>
<td><strong>98.004%</strong></td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pcbi.1006801.t004

N-terminus standing out as Commu. A is expected, as this is the most distinguishable part between dark and light states. The loops between A’α/Aβ and Eα/Fα standing out as Commu. B provides additional important information to distinguish between the dark and light states. The two loops in Commu. B are far away from each other, but the distance distributions within Commu. B are consistent throughout the dark and light states, with only 0.107% accumulative feature importance. Due to the significant feature importance of Commu. B correlated with the rest of protein, we propose that the two loops in Commu. B mediate the transition from the dark state to the light state (Fig 11). At initial stage of the transition from the dark to the light state, the Commu. B may function as a switch to be turned on first (States 3 → 2) before the N-terminus (Commu. A) undergoes significant conformational change to reach an intermediate state (States 2 → 7). This route is more likely than the change from State 3 directly to the intermediate state (States 3 → 7), in which Commu. A and B undergo the conformational change concomitantly. To further verify the low importance of Commu. C and D during the conformational change, the structural comparison between different macrostates for the proposed pathway is illustrated in S7 Fig and listed in S2 Table. The results clearly suggest that the Commu. C and D do not have significant structural differences among different macrostates, and highlight the conformational changes of Commu. B and A.

Discussion

More about community analysis

Some key functional positions have been revealed to control LOV allostery without affecting LOV photocycle kinetics.[4, 46–48] The photo-induced covalent bond between the conserved Cys residue and flavin cofactor initiates conformational changes within N- or C-terminal extensions (Ncap/Ccap) to the LOV core.[49] It was proposed that the conformational change
of these N/Ccap elements regulates activity of LOV domains or recruit proteins to Ncap, Ccap or β-sheet surfaces. [4, 50–52]

The results of simulations and machine learning-driven community decomposition allowed to quantify the role of specific regions in allosteric conformational changes and led to major steps of the allosteric mechanism. We found that, in addition to the primary large-scale conformational changes cluster to the N-terminus, the structural changes differentiating N-terminal states are coupled to the rearrangement of two loop regions (A’α/Aβ and Eα/Fα). Therefore, there is a higher probability for covalent bond formation to induce conformational changes in the loop structures first, than to induce reorientation of the N-terminus directly. Notably, although the covalent bond is formed due to the external blue light stimulation, the subsequent conformational changes can be attributed to the existence of the covalent bond without the activation energy dissipations in the protein. Such findings are consistent with recent reports indicating that chemical reduction of the flavin cofactor to form the neutral semiquinone is sufficient to induce a conformational response in VVD, independent of photo-excitations. [53]
Delineating the atomistic details of the allosteric mechanism revealed that focusing on individual residues was insufficient to illustrate a global conformational response. Rather, community analysis presented in this study specified coupling between Commu. B and the N-terminus as Commu. A, where Commu. B is impacted first before significant conformational change occurs at the N-terminus. As a result, residues stabilizing this community could play a central role in switching the conformation of protein from dark to light state. These findings not only reveal significant agreement with experimental observations, but also identify unexpected regions that may play a substantial role in modulating LOV conformational dynamics.

To highlight these findings, we divide Commu. B into three characteristic regions based on experimental data: the molecular swivel (PGG motif residues 42–44), the N-terminal hinge (residues 65–72), and the FAD-binding loop (Eβ/Fα loop). Previous experimental studies have identified the first two regions as essential for mediating conformational changes in VVD.\[4, 38\] Namely, the PGG motif was identified as a molecular swivel that is essential for conformational changes in the N-terminus (N-latch; Commu. A), which distinguishes light- and dark-state conformations. Similarly, mutations in the N-terminal hinge abrogate structural changes.
The N-terminal hinge loop between L64-S70 is the last part of N-terminal cap (H37-S70) which is different from other PAS (Per-Arnt-Sim) proteins including photoactive yellow protein (PYP)[54], Avena sativa LOV2 domain (AsLOV2) phototropin[51] and Drosophila clock protein Period[55]. The mutagenesis studies revealed that the substitution of Cys71, which is the next residues of the identified loop, either enhance conformational changes (C71V) or abrogate the conformational response (C71S) both in vitro and in vivo.[4] Further experimental structural analysis revealed that the hydrogen bond between Asp68 in the identified A′α/Aβ loop with Cys71 could be crucial for N-terminal conformational changes. In addition, the experimental observations show that Pro66, from the identified A′α/Aβ loop, undergoes the largest shifts (2.0Å) in the light state versus dark state[4], which also has agreement with RMSF plot in S6 Fig. Notably, recent studies of other LOV proteins, as well as VVD homologs, identify the N-terminal hinge as a hot spot for evolutionary adaptation, where residues within the core loop facilitate integration of an oxidative stress sensing mechanism into VVD-like proteins by modifying the initial conformational response[38, 56], or aid in differentiation of signaling mechanism by regulating the location of a key evolutionarily selected residue in the adjacent Aβ strand.[57]

### Relation to experimental studies

The synergy between the experimental and theoretical studies not only validates the importance of these structural communities in LOV signal transduction, but also highlights how these communities signal through each other. Here, we show that the photo-induced covalent bond formation first initiates a conformational change in Commu. B, consisting of the swivel and hinge regions. These propagate to the N-latch (Commu. A) to differentiate light- and dark-state conformations. Another question that may arise is if the key conformational changes occur within Commu. A, and B, what are the fundamental function of Commu. C and D. A careful examination of the methodology conducted here and existing experimental studies can shed light on the role of these distinct communities. The current study identified the hinge loop (residues 64–70) as key components of Commu. B, but it did not include residues 71–74 which have been identified as either essential for function (Cys71)[4] or aid in evolutionary adaptation of signaling mechanisms (Ala72 and Ile74)[57]. Rather these residues belong to Commu. C (Cys71) and D (Ala72 and Ile74), as is a key signaling residue in the PAS protein CLOCK (Trp362).[58] None of these residues undergo large conformational changes at the Cα position, rather side chain reorganizations occur due to steric constraints or H-bond changes. Combining these experimental observations indicates that the approach outlined here keenly identifies communities dictating global conformational changes (changes in Cα), but may not include residues near community junctures that enable adaptation in function (Ala72, Ile74) or relay the initial chemical event via a subtle conformational change (Cys71, Ile74). These residues cannot be identified easily by existing computational techniques, because examining every rotamer/H-bond change for its contribution to a global conformational response is not feasible due to the computational time necessary to complete such a task. However, our study indicates these residues likely will reside at the junctures between communities, thus our approach can narrow down candidate residues that may be subtly important for the conformational change, or that are excellent targets for mutagenesis to fine tune signaling mechanisms.

A second unexpected observation of the current studies was the inclusion of the Ea/Fα loop in Commu. B. Currently, the function of the Ea/Fα loop in VVD/LOV signal transduction is largely unknown. It was initially identified as the “FAD insertion loop” due to its presence in fungal proteins VVD and WC1, which were found to bind FAD instead of FMN.
Crystal structures of VVD confirmed contacts between the loop and the adenine moiety of FAD [4], however, plant photoreceptors Zeitlupe (ZTL), Flavin-Kelch-Fbox-1 (FKF1) and Leucine-Kelch-Repeat protein 2 (LKP2) all contain a Ea/Fα loop, but selectively bind only FMN.[57] Similarly, experimental studies of VVD homologs in Trichoderma reesei and Botrytis cinerea confirmed that these proteins bind FMN, despite the presence of the Ea/Fα loop.[56, 59] Thus, the purpose of the Ea/Fα loop remains elusive. Given the unstructured and dynamic nature of the loop, it is particularly challenging to study using traditional experimental approaches.

Here, we identify the Ea/Fα loop as contributing to the initial conformational changes driving rearrangement of the N-terminus, thus the Ea/Fα loop may be a hidden and largely unexplored region to modulate signal transduction in LOV proteins. Indeed, there is some experimental evidence to support such an assertion. Namely, deletions of the Ea/Fα loop were shown to dampen conformational changes in FKF1 that were observed using Small-Angle-X-ray scattering.[60] Furthermore, a recent study identified a mutation in the FKF1 Ea/Fα loop (H105L) that enhanced light-driven activity in designed optogenetic tools.[61] Finally, a possible role of the Ea/Fα loop was also proposed but not confirmed for the VVD homolog in B. cinerea, where the primary signaling mechanisms were found to diverge from that in VVD.[59]

Based on our results, computational approaches to identify how the Ea/Fα loop may modulate signal transduction in LOV proteins, could lead to a new avenue to tune LOV optogenetic tools.

Concluding remarks

In this work, by using a novel computational framework for dissecting protein allostery, we examined and reconstructed molecular mechanism of Vivid (VVD) protein, which forms a covalent bond between cofactor and a cysteine residue upon blue light activation, and facilitates a large conformational change on N-terminus for circadian signal transduction. By integrating Markov state model, machine learning classification models, and a newly developed community analysis, we accurately reconstructed the equilibrium distributions for bonded and non-bonded configurations, and determined structural differences among these states. A machine learning-based community analysis provided atomistic details of coordinated global motions of functional regions with statistical significance. We systematically verified the impacts of the key local covalent bond upon photo excitation to global motions of VVD, and revealed the importance of A’α/Aβ and Ea/Fα loops in conformational change. The results of this analysis are consistent with the experiments and validated the robustness of the proposed approach in identifying functionally relevant molecular switches of allosteric changes. Overall, this study reveals the detailed mechanism of conformational changes from the dark state to the light state, and the central role of covalent bond in the VVD protein. Our findings also suggested how manipulating these elements with light in LOV proteins can link chemistry with modulation of allosteric changes, thereby providing a path for rational engineering of LOV ontogenetic tools.[1]

Materials and methods

Molecular dynamics simulation

The initial structures of dark and light states of VVD were obtained from the Protein Data Bank (PDB)[62] with the ID as 2PD7 and 3RH8, respectively. The dark and light state sequences start from Met36 and His37, respectively. For consistency, residue 36 from the dark state was removed to maintain the same number of residues in both states. Both structures include the flavin adenine dinucleotide (FAD) as cofactor. FAD and flavin mononucleotide (FMN) are two types of cofactors commonly existing in the PAS (Per-Arnt-Sim) domain family, and the difference between FMN and FAD comes from the adenosine monophosphate
(AMP) moiety. Because FMN and FAD carry similar biological role, the AMP moiety was removed from FAD to construct FMN, and the FMN force field from a previous study was used.\cite{63} Total of four simulation systems were constructed based on crystal dark state structure with or without the photo-induced covalent bond, and crystal light state structure with or without this bond. The crystal dark state structure without the photo-induced covalent bond is referred to as native dark configuration; the crystal light state structure with this bond is referred to as native light configuration. As comparison, the crystal dark state structure with the photo-induced covalent bond is referred to as transient dark configuration, and the crystal light state structure without this bond is referred to as transient light configuration.

Hydrogen atoms were added to the crystal structures, which are subsequently solvated using explicit water model (TIP3P)\cite{64} and neutralized with sodium cations and chloride anions. Initially, 10 nanoseconds (ns) of isothermal-isobaric ensemble (NPT) molecular dynamics (MD) simulations were carried out, and then 1.1 microseconds (\(\mu s\)) of canonical ensemble (NVT) Langevin MD simulation at 300K were conducted. First 100 ns simulations were discarded as equilibration, and the following 1 \(\mu s\) simulation was used in the analysis. Three independent simulations with 1.1 \(\mu s\) length were carried out for each configuration, and total of 12 \(\mu s\) simulations were applied in the analysis. After solvation of the simulation systems, the numbers of TIP3P water molecules added are 7240, 7239, 9430, 9429 for native dark, transient dark, native light, and transient light configurations, respectively. For all simulations, SHAKE method was applied to constrain all bonds associated with hydrogen atoms. Step size of 2 femtosecond (fs) was used and simulation trajectories were saved every 100 picoseconds (ps). Cubic simulation box and periodic boundary condition were applied for all MD simulations. Electrostatic interactions were calculated using particle mesh Ewald (PME) method.\cite{65} All simulations were carried out using CHARMM\cite{66} simulation package version 41b1 with the support of graphics processing unit (GPU) calculations based on OpenMM.\cite{67}

**Markov state model**

MSMBuilder\cite{68} was employed to build Markov state model (MSM). To apply MSM, each frame needs to be assigned to a microstate, and transition probability was estimated between different states. To fulfill the “memoryless” assumption underlining MSM, transitions among microstates need to be faster than transitions among macrostates to avoid disguising important kinetic barriers. Therefore, constructing appropriate collective variables (CV) to describe a microstate is critical.\cite{69, 70} Common methods to generate CVs include time structure based independent analysis (t-ICA)\cite{71} and principal component analysis (PCA)\cite{72}. In the current study, the RMSD values calculated with reference to crystal dark and light structures were used as CVs to describe the microstates. 30 ns were chosen as the lag time, and eight macrostates were chosen based on the ‘gaps’ in the implied timescale plot. Perron-cluster cluster analysis (PCCA)\cite{73} implemented in MSMBuilder\cite{68} was applied to cluster the microstate into the macrostates. All the equilibrium or steady state distribution was estimated from the transition probability among different macrostates. In building the MSM, the hyperparameters in MSMBuilder remained as the default setting, including ergodic cutoff being turned on, the reversibility of transition matrix being enforced using maximum likelihood method (MLE), the prior counts for the transition between states being set as zero, and the sliding window setting being turned on. The MSMBuilder used in current study is version 3.8.0.

**Machine learning**

Supervised machine learning model including artificial neural network and tree based models were used in the current research. A typical artificial neural network model consists of input
layer, hidden layer and output layer with a number of nodes connected with each other. The training processes of artificial neural network model is a back propagation processes implemented in scikit-learn as a Python package.\[43\] The input data are extracted from the featureization results for each saved simulation frame from trajectories. The target label for each date point is the sequential number of each macrostate. In the artificial neural network model, starting with a random weight assigned to each node, each cycle of training is to minimize the total loss regarding with target label using stochastic gradient descent (SGD) algorithm until weight on each node converges to a minimum. The loss function is defined as

\[
\text{Loss}(\hat{y}, y, W) = -y \ln \hat{y} - (1 - y) \ln (1 - \hat{y}) + a\| W \|^2,
\]  
(Eq 1)

where \( y \) is the label predict by the model, \( \hat{y} \) is the correct label, \( \| W \| \) is the weights on the nodes, and \( a \) is named as L2 regulation term to regulate the model to avoid overfitting the weights.

Other supervised models applied in the current study are tree-based machine learning models, including Decision Tree[25], Random Forest[26] and One-vs-one Random Forest. The decision tree is a recursive partition algorithm that groups the samples with the same label together. For a given data set \( Q \), the algorithm selects the parameter \( \theta = (j, t) \) consists feature \( j \) and a threshold \( t \) to divide the data into two parts \( Q_{\text{left}} \) and \( Q_{\text{right}} \) as the following:

\[
Q_{\text{left}}(\theta) = (x, y) | x_j \leq t,
\]  
(Eq 2)

\[
Q_{\text{right}}(\theta) = (x, y) | x_j > t,
\]  
(Eq 3)

where \( x \) is the training data, \( y \) is the training label. The selection of parameter will minimize the total “impurity” as the following

\[
Q^* = \arg \min_\theta \left( \frac{n_{\text{left}}}{N} \cdot H(Q_{\text{left}}(\theta)) + \frac{n_{\text{right}}}{N} \cdot H(Q_{\text{right}}(\theta)) \right),
\]  
(Eq 4)

where \( H() \) is the impurity measurement function. Common measurements of the impurity for a given dataset include cross-entropy measurement \(-\Sigma_k p_k \log p_k \) and Gini impurity \( \Sigma_k p_k (1 - p_k) \), where \( p_k \) represents distribution of certain class within total dataset. The scikit-learn package employed in the current study used the Gini impurity for training purpose. Therefore, the feature importance is calculated as the sum of all Gini impurity decreasing for all nodes based on the particular feature. However, the algorithm implemented in decision tree models is deterministic with the best splitting of input data, which might be biased towards some features and input conditions.[26] To overcome this, random forest model consisting of multiple decision tree models was applied. In random forest model, each tree classifies the input data with different random seeds, and the final prediction is the average of all single decision tree models. The feature importance from random forest has more statistical significance than single decision tree model. One-vs-one random forest model is a further improvement than the random forest model in multi-class classification task. The one-vs-one classifier is a common strategy in the multi-class classification task.[74, 75] Instead of training only one classifier to classify all classes, one classifier was trained specifically for any two classes pair, and the overall prediction model is weighted by the prediction of all classifiers.[75] Although computational costs are higher than the single classifier, the statistical significance of this model is much higher, and overfitting is less likely. In the current study, for the eight meta-stable states, total 28 random forest classifiers for state pairs among 1 through 8 were trained. Compared with single random forest model, one-vs-one random forest provides not only the
overall feature importance, but also feature weights specifically to distinguish any two particular states.

Pairwise distances for alpha carbons (Cα) were used as features to train the supervised machine learning models. Pairwise distances are invariant with regard to translation and rotation motions of whole molecule. MSMbuilder package was employed to extract Cα pairwise distances from the trajectories. All the machine learning models were implemented using scikit-learn package [43] in python. The performance of machine learning model is assessed by the accuracy of classifier, which is defined as the fraction of the number of the correct classified data with reference to the number of whole input data.

Machine learning based community analysis

Based on the network and community analysis described in the previous studies,[44, 76] focusing on the community of residues rather than single residues could have more statistical significance. In this study, we propose to group residues into communities, so that the impacts of external perturbations on the distribution differences within the same community are minimum. We refer to these communities as machine learning based communities or simply as ML communities. Therefore, the change of protein motion upon perturbation could be characterized as the relative motion among ML communities related to different states. The feature weights calculated by the machine learning models were applied to construct ML communities. The feature weights indicate the distinguishability between the different states distributions for that specific residue distance. Lower feature weights represent that the specific distance distribution is less distinguishable between different states, and vice versa. Therefore, the community analysis is transformed into a local minimum search problem based on machine learning weights. The Kernighan–Lin algorithm in graph partition problem[77] was implemented to search the local minimum value.

The protein can be modeled as an undirected graph with nodes represented by the residues, and edges represented by the pairwise Cα feature weights. The goal of ML community analysis is to partition the protein graph into several communities and maintain that the total feature importance in each community is minimized. To apply Kernighan–Lin algorithm,[77] we assume that there are n communities labeled as C_j through C_m. The total feature importance for any partition of communities T is defined as the total edge inside each community as the following equation.

\[ T = \sum_i \left( \sum_{j \in C_j} E_{ij} \right), \]  

(Eq 5)

where i, j are the residues in Community C_j, and E_{ij} is the feature importance between residues i and j. The internal edges and external edges for node i are defined as the following. Assume that node i belongs to Community C_m, internal edges of node i, \( I_{in} \), is defined as the total edge value between each node in C_m with node i, and the external edges of node i, \( E_{x_i} \), regarding to any other Community C_q are defined as the total edges of node in C_q with node i.

\[ I_{in} = \sum_{j \in C_m} E_{ij}, \]  

(Eq 6)

\[ E_{x_i,C_q} = \sum_{j \in C_q} E_{ij}. \]  

(Eq 7)

For each iteration in the algorithm, the ML community partitions can be improved by inserting node i into other community or swapping node i with node j from any different community. For inserting node i into community C_q, the benefit of total edge in communities is calculated as

\[ T_{now} - T_{old} = E_{x_i,C_q} - I_{in}. \]  

(Eq 8)
For swapping node $i$ from community $C_m$ and node $j$ from community $C_k$, the benefit of total edge in communities is calculated as

$$T_{\text{new}} - T_{\text{old}} = (E_{i,C_m} + E_{j,C_k}) - (In_i + In_j) - 2 * E_{ij}.$$  

(Eq 9)

After defining insertion and swapping operations, the ML community construction algorithm is described as the following:

1. The ML communities are first initialized with random partitions.
2. For each node, the benefits of moving into another ML community are calculated to identify the insertion operation with the maximum benefit.
3. For any pairs of nodes from different ML communities, the benefits of swapping those two nodes are calculated to identify the swapping operation with the maximum benefit.
4. Either swapping or insertion operation with a higher benefit is chosen.
5. For the new community configuration, steps 2 through 4 are repeated until the benefits of insertion or swapping are less than 0.
6. The ML community configuration is final when any insertion or swapping operations will increase the total internal edges within each ML community.

The above algorithm can only reach a local minimum as final solution. Some algorithms like Simulated Annealing[78] could improve the searching for global minimum. In the current study, we repeat 10,000 times with different random starting configurations, and the lowest value was chosen as the final solution.

**Root-Mean-Square Deviation (RMSD) and Fluctuation (RMSF)**

The conformational change during the MD simulations can be measured by RMSD regard to a reference structure. For a molecular structure represented by Cartesian coordinate, the RMSD is defined as the following:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} (r_{ij}^2 - U r_j)^2}{N}}.$$  

(Eq 10)

The Cartesian coordinate vector $r_{ij}^2$ is for the $i^{th}$ atom in the reference structure, $r_j$ is the $j^{th}$ atom in a given structure. $U$ is the rotation matrix to superimpose the given structure with the reference structure. $N$ is the total number of atoms in the structure. For each simulation, the RMSD values with reference to the crystal dark and light structures were calculated to quantify the sampling following a previous study.[36]

Similarly, the fluctuation of atoms during MD simulation with reference to the averaged structure can be measured by RMSF. The $RMSF_i$ of atom $i$ for a given MD trajectory is defined as

$$RMSF_i = \sqrt{\frac{1}{T} \sum_{j=1}^{T} (v_{ij} - \bar{v}_j)^2},$$  

(Eq 11)

where $T$ is the total number of frames in the given MD trajectory, $v_{ij}$ is the coordinate atom $i$ in the frame $j$, and $\bar{v}_j$ is the averaged coordinate of atom $i$ in the given trajectory.
Transition path theory

After the MSM is established, the transition path theory (TPT) [45, 79] can be applied to estimate the potential transition path related to the conformational changes. Applying TPT for VVD, the target transition paths should connect an initial state A including the native dark macrostate (state 3) and a target state B including native light macrostate (state 4). All other states are considered as the intermediate states (1). In TPT, the essential concept is "committor probability" $q^+_i$, which is defined as the probability from any state $i$ to reach the target state $B$ rather than initial state $A$. By definition, all the microstates $i$ belonging to state $A$ have $q^+_i = 0$. Meanwhile, all the microstates $i$ belonging to state $B$ have $q^-_i = 1$. The committor probabilities for any other microstates can be calculated by solving the following linear equation:

$$-q^-_i + \sum_{k \neq i} T_{ik}q^+_k = -\sum_{k \neq B} T_{ik},$$

(Eq 12)

where $T_{ik}$ is the transition probability from state $i$ to state $k$. The backward-committor probability $q^-_i$ is simply calculated as $q^-_i = 1 - q^+_i$. After the committor probability is calculated, the effective flux from microstate $i$ to $j$, which is determined by the transitions from $A$ to $B$ passing through these states, can be calculated as Eq 13

$$f^+_ij = \pi_i q^-_i T_{ij} q^+_j,$$

(Eq 13)

where $\pi_i$ is the equilibrium distribution for state $i$. The above definition does not consider the backward flux $f^-_ij$. Therefore, the net flux from $A$ to $B$ transition at edge $i, j$ can be calculated as $f^+_ij = \max(0, f^+_ij - f^-_ij)$. The net flux $f^+_ij$ is essentially the fluxes leaving state $A$ and reaching state $B$. Meanwhile, total flux for the transition from $A$ to $B$ per lag time $\tau$ can be calculated as the following

$$F = \sum_{i \in A} \sum_{j \in B} \pi_i T_{ij} q^+_j.$$

(Eq 14)

The flux from state $A$ to state $B$ can be decomposed into distinct individual pathway $P_i$. The pathway decomposition algorithm implemented in MSMBuilder is Dijkstra algorithm, which searches for the highest flux pathway first, then removes the pathway from net flux matrix by subtracting the flux of the path from every edge in the path, and continues search until all possible pathways are identified.

Supporting information

S1 Fig. Comparison between 2D-RMSD, time-structure independent components analysis (t-ICA), and principal component analysis (PCA) models to construct Markov state model (MSM). Projection and grouping of VVD simulations as microstates on the surfaces of (a) 2D-RMSD, (b) t-ICA and (c) PCA and (d) t-ICA with five selected features, respectively. Scanning of lag time for the estimation of relaxation timescales for (e) 2D-RMSD, (f) t-ICA and (g) PCA, and (h) t-ICA with five selected features, respectively. Microstates grouped in eight macrostates on (i) 2D-RMSD, (j) t-ICA, and (k) PCA and (l) t-ICA with five selected features surfaces, respectively.

(TIF)

S2 Fig. Averaged Cα RMSD within each microstate in 2D-RMSD, t-ICA, PCA, and t-ICA with five selected features, respectively.

(TIF)

S3 Fig. Testing the markovian property using Chapman-Kolmogorov method to compare the probability directly observed in the simulation and the estimated probability using lag
time as 30ns. (a) VVD dark state 3; (b) VVD light state 4.

S4 Fig. Ensemble distributions based on MD trajectories in (a) non-bonded configurations and (b) bonded configurations.

S5 Fig. Convergence test of VVD simulations using RMSD and configurational entropy. (a) RMSD fluctuation along each trajectory; (b) The accumulative configurational entropy along each trajectory. The configurational entropy plot indicates that the simulations are well converged after 600ns samplings.

S6 Fig. RMS fluctuation analysis of each residue for bonded and non-bonded configurations. The flexibility of A’α/β loop is enhanced upon formation of photo-induced covalent bond between cofactor and VVD.

S7 Fig. The representative structures for macrostates 2, 3, 4, 5, and 7, as important for conversion between the dark and light states of VVD. The structure alignment reveals the significant conformational changes of Commu. A and B among different macrostates, and shows that the Commu. C and D do not have significant conformational differences in these macrostates.

S8 Fig. Flowchart summarizing the general analysis procedure presented in this study.

S1 Table. List of residues in each ML community.

S2 Table. Structural comparison (RMSD in Å) among different macrostates. Communities C and D are combined for the analysis.

S1 Dataset. Python scripts implemented in this study with sample data.

Acknowledgments

Computational time was generously provided by Southern Methodist University’s Center for Scientific Computation.

Author Contributions


Data curation: Hongyu Zhou, Peng Tao.


Funding acquisition: Peng Tao.

Investigation: Peng Tao.

Project administration: Peng Tao.

Resources: Peng Tao.

Software: Peng Tao.

Supervision: Peng Tao.

Validation: Peng Tao.

Visualization: Hongyu Zhou, Peng Tao.

Writing – original draft: Hongyu Zhou, Gennady Verkhivker, Brian D. Zolotowski, Peng Tao.

Writing – review & editing: Hongyu Zhou, Gennady Verkhivker, Brian D. Zolotowski, Peng Tao.

References


