

Deciphering the Allosteric Process of *Phaeodactylum tricornutum* Aureochrome 1a LOV Domain

Hao Tian, Francesco Trozzi, Brian D. Zoltowski, and Peng Tao*

*Department of Chemistry, Center for Research Computing, Center for Drug Discovery,
Design, and Delivery (CD₄), Southern Methodist University, Dallas, Texas, United States
of America*

E-mail: ptao@smu.edu

Abstract

The conformational-driven allosteric protein diatom *Phaeodactylum tricornutum* aureochrome 1a (PtAu1a) differs from other light-oxygen-voltage (LOV) proteins for its uncommon structural topology. The mechanism of signaling transduction in PtAu1a LOV domain (AuLOV) including flanking helices remains unclear because of this dissimilarity, which hinders the study of PtAu1a as an optogenetic tool. To clarify this mechanism, we employed a combination of tree-based machine learning models, Markov state models, machine learning based community analysis and transition path theory to quantitatively analyze the allosteric process. Our results are in good agreement with reported experimental findings and revealed a previously overlooked C α helix and linkers as important in promoting the protein conformational change. This integrated approach can be considered as a general workflow and applied on other allosteric proteins to provide detailed information about their allosteric mechanisms.

Introduction

Light, oxygen or voltage (LOV) domains are a subdivision of the Per-Arnt-Sim (PAS) superfamily that are sensitive to blue light and undergo conformational as well as dynamical changes upon light activation.^{1,2} This activation begins with the formation of a covalent bond between a cofactor and a conserved cysteine residue. Possible cofactors include flavin adenine dinucleotide (FAD), flavin mononucleotide (FMN) or riboflavin.³ This covalent bond further promotes the overall structural changes, resulting in the alteration of the protein-protein interactions and thus signal transduction.⁴

Phaeodactylum tricornutum aureochrome 1a (PtAu1a) is a recently discovered LOV protein that consists of an unstructured N-terminal region, and a basic region leucine zipper (bZIP) DNA-binding domain connected to a C-terminal LOV core⁵. The LOV domain, together with two flanking helices (A' α and J α) is usually referred to AuLOV⁶. The protein is dynamically stable in the dark state due to the interaction between the LOV core and

bZIP.⁷ This interaction prohibits the protein binding with DNA.⁸ A photo-induced covalent bond is formed between the C4a position of the cofactor FMN and a nearby sulfur in Cys287. This covalent bond triggers a series of conformational changes, including the undocking and unfolding of J α helix from the LOV core surface, the release of A' α from the hydrophobic site on LOV domain surface, and dimerization of the LOV domains.⁷

These events lead to the increase of PtAul1a affinity for DNA binding and are proposed to be allosteric.⁹ Recent research has revealed that a combination of structural changes in the LOV core and the undocking of J α is essential for the release of A' α and further LOV domain dimerization.⁷ The allosteric mechanism in PtAul1a is considered to be different from other LOV proteins since the location of the LOV domain is in the C-terminus in PtAul1a, while in the N-terminus in others.^{10,11} This structural difference raises the question on allosteric transmission in PtAul1a.

Various computational methods have been applied to explore protein allosteric mechanisms at atomic level.¹²⁻¹⁴ Molecular dynamics (MD) simulations are capable of providing atomic-scale information, as well as structure-function relationships^{15,16}, and are widely used in sampling protein motions and structure landscapes.¹⁷ The significant computational power provided by graphical processing units (GPUs) has promoted the timescale of MD simulations from nanoseconds to milliseconds.^{18,19} To obtain more biologically meaningful information from trajectories, Markov state models (MSMs) are often used to extract asymptotic kinetic information based on limited simulations.^{20,21} Kinetically separate macrostates can be obtained from MSMs in the reduced dimension. Differences among these subspaces can be quantified to gain insight into protein structure and function relations.

The success of MSMs depends on appropriate dimensionality reduction methods that can preserve global distances while retaining the most structural information.²² New dimensionality reduction methods have been developed to project the high-dimensional trajectories to lower dimensions for thorough study. However, many methods, such as principal component analysis (PCA)²³, time-structure based independent component analysis (t-ICA)²⁴

and t-distributed stochastic neighbor embedding (t-SNE) method²⁵, suffer from problems including maintaining the similarity between high dimensional space and low dimensional space and not resistant to system noise²⁶. In the current study, MD simulations were projected to 2D space via the ivis framework²⁷, which is a nonlinear method based on Siamese neural networks (SNNs) and has been shown powerful in interpreting biological systems²⁸.

Machine learning has recently achieved great accomplishments in chemistry and biology. Raccuglia et al. applied machine learning algorithms trained on failed experimental data to predict reaction results with high accuracy.²⁹ Faber et al. employed machine learning techniques for feature vector representations of crystal structures.³⁰ Botu et al. integrated machine learning framework to accelerate *ab initio* molecular dynamics simulation.³¹ The broad applications of machine learning stem from the ability to process large datasets and, more importantly, provide explanatory details^{32,33}. These favorable metrics offer a new prospective direction for the research on protein allostery. In this study, two tree-based machine learning models, random forest (RF) and one-vs-one random forest (OvO RF), were used to study the structural differences between macrostates and determine the contribution of residues along the allosteric process. In combination with machine learning and dynamic community analysis, Zhou et al. developed a new approach, known as machine learning based community analysis³⁴, to identify important structures in dynamically-driven protein ensembles. Here we applied this method on AuLOV and proved the feasibility of this method in partitioning conformational-driven protein structures.

The AuLOV is investigated in this study through the MD simulations, tree-based machine learning models, machine learning based community analysis and transition path theory. Our results identified key residues that are consistent with experimental discoveries and suggested the importance of C α helix, overlooked thus far. Moreover, we quantified the important role of N- and C-terminal linkers in modulating AuLOV allostery. The integrated methods determined the importance of each residue in the allosteric process, and therefore provided new insights into the allosteric mechanisms, which may promote future research on

PtAula as an optogenetic tool.

Methods

Molecular Dynamics (MD) simulations

The initial structures of native dark state monomer and native light state of AuLOV dimer were taken from the Protein DataBank (PDB)³⁵ with the PDB ID being 5dkk for the native dark state and 5dkl for the native light state. To keep the same number of residues in all structures, the longest common residue sequences (from Ser240 to Glu367) were modeled. Both the native dark and light structures contain FMN as a cofactor. The force field for the cofactor FMN was used from a previous study³⁶. In order to fully explore the protein movements with regard to the formation of the covalent bond between cysteine 287 and FMN, two new transient states, referred to as transient dark state and transient light state, were generated. Transient states were generated by forcing the Cysteiny-Flavin C4a adduct into the native dark state structure (transient dark state) or breaking the Cysteiny-Flavin C4a adduct in the native light-state structure (transient light state). These transient structures facilitate analysis of allosteric interconversion between light- and dark-state structures.

The crystal structures were added with hydrogen atoms and were further solvated in a water box with the TIP3P water molecules³⁷. Sodium cations and chloride anions were added for neutralization. For each structure, energy minimization was done with the steep descent method and the adopted basis Newton-Raphson minimization. System temperature was raised to 300K through a 20 picoseconds (*ps*) MD simulations. Another 20*ps* simulations were done for equilibrium. 10 nanoseconds (*ns*) of isothermal-isobaric ensemble (NPT) followed by 1.1 microseconds (μs) of canonical ensemble (NVT) Langevin MD simulations were carried out at 300K. The first 0.1 μs NVT simulations was considered as an equilibration stage and was discarded. Three NVT MD simulations were conducted independently for each protein structure. Therefore, a total of 12 μs simulations were generated for analysis.

SHAKE method was used to constrain all bonds associated with hydrogen atoms. 2 femtoseconds (fs) step size was used for all MD simulations. Trajectories were saved for every 100 ps . Periodic boundary condition (PBC) was applied in simulations. Particle mesh Ewald (PME) algorithm³⁸ was used to calculate the electrostatic interactions. MD simulations were conducted using GPU accelerated OpenMM³⁹ and CHARMM27 force field⁴⁰.

Analysis of Simulation Trajectories

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

The dynamics stability of a MD simulation trajectory is measured by the root-mean-square deviation, which is calculated as:

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

where r_i^0 represents the coordinate of an atom i in Cartesian coordinate system and U is the most appropriate alignment transformation matrix between two structures. For each trajectory, the first frame was treated as the reference structure.

The root-mean-square fluctuation is used to measure the fluctuation of atoms in each frame with regard to the first frame in a MD simulation trajectory. Specifically, C α atoms were considered important in representing the protein motions and the corresponding RMSFs of each C α were calculated as:

$$\text{RMSF}_i = \sqrt{\frac{1}{T} \sum_{j=1}^T (r_i(t) - \bar{r}_i)^2} \quad (2)$$

where T is the number of frames and \bar{r}_i is the averaged Cartesian coordinate of the i^{th} C α in the given trajectory.

Feature Processing

The $3N$ degrees of freedom in the Cartesian coordinate system hinders a thorough analysis of MD simulations in biological systems. Pair-wised $C\alpha$ distances is usually extracted to represent the structural characteristics of protein configurations⁴¹. In the current study, a feature vector of each structure was constructed by calculating the distance pairs between one α carbon atom and another α carbon atoms in amino acids following the order of residue sequence. This feature vector was further encoded by a previously proposed transformation method⁴² with a cutoff of 10\AA .

ivis Dimensionality Reduction Method

ivis is a machine learning based dimensionality reduction method that is originally developed for single cell technology²⁷. The ivis framework applies the Siamese neural networks architectures that is composed of three identical base neural networks. For each base neural network, there are three dense layers consisting of 500, 500, and 2,000 neurons with a final embedding layer of 2 neurons. A novel triplet loss function is implemented in the training process:

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

The symbol a represents the point of interest, often referred to as anchor point. The symbol p represents a positive point that is selected based on the k -nearest neighbors (KNNs) algorithm. The symbol n represents a negative point that is randomly selected from the rest of data samples. The similarity between two points is calculated as the Euclidean distance (D). The margin (m) is defined as the minimum distance between any pair of points and was set to default value of one. The advantage of ivis method lies in the triplet loss function, which aims to minimize the distance between anchor points and positive points while maximizing the distance between anchor points and negative points.

Adam optimizer with a learning rate of 0.001 was applied to train the neural network. To prevent overfitting, early stopping of 5 was used to terminate the training iteration if the triplet loss function does not decrease with 5 consecutive epochs.

Machine Learning Methods

Random Forest (RF) and One-vs-one (OvO) Random Forest

Random forest is a tree-based machine learning technique^{43,44} that was applied to learn the structural differences among macrostates. In this study, each random forest model is composed of 50 decision trees. Decision trees were trained individually and the final result of a random forest model is formed by a voting algorithm. Scikit-learn⁴⁵ version 0.20.1 was used to implement the random forest model.

Random forest model overcomes the problem of overfitting by employing several decision trees. However, in multi-task classification jobs, one-vs-one random forest model is more common and superior than random forest model by constructing one classifier for each pair of classes.⁴⁶ The overall output is the weighted sum of all base classifiers. In the current study, 10 macrostates were trained with 45 random forest models. One-vs-one random forest model provides weighted sum of overall feature importance with specific feature importance regarding two given classes.

Feature Importance

The feature importance in a random forest model is calculated using the Gini impurity, which is calculated as:

$$\text{Gini impurity} = \sum_{i=1}^C -f_i(1 - f_i) \quad (4)$$

where f_i and C are the frequency of one label at a node that are chosen to divide the data set and the number of labels, respectively. A random forest model consists of multiple decision

tree models. The importance of feature i in each decision tree is calculated as:

$$f_i = \frac{\sum_l^s n_j}{\sum_{k \in \text{all nodes}} n_k} \quad (5)$$

where s is the frequency of node j split on feature i . The importance of feature i in a random forest model is calculated by averaging its importances among decision tree models:

$$F_i = \frac{\sum_{j \in \text{all decision trees}} \text{norm} f_i}{N} \quad (6)$$

where $\text{norm } f_i$ and N are the normalized feature importance of one decision tree and the number of decision trees, respectively.⁴⁷

Pair-wised $C\alpha$ distances were extracted as the input features and the corresponding feature importances were calculated. For each $C\alpha$ distance, the importance was added to the related two residues. The accumulated feature importance of residues implies their contributions in the allosteric process.

Markov State Model

The long timescale protein dynamics is tracked by construction of the Markov state model⁴⁸. Each simulation frame is assigned to different microstates through MiniBatch k-means clustering method. Compared with microstates, macrostates are more biologically meaningful as they are considered as kinetically-separate equilibrium states. 10 macrostates were generated by Perron-cluster cluster analysis (PCCA)⁴⁹. Lagtime is needed to build a MSM and was determined as $40ns$ based on the implied relaxation timescale. Transition matrix and corresponding transition probabilities were estimated based on this MSM. MSMBuilder⁵⁰ package (version 3.8.0) was used to build MSMs.

Machine Learning based Community Analysis

Machine learning based community analysis³⁴ is a newly proposed method by Zhou et al., which groups residues into communities. The main idea of this analysis is maximizing the overall feature importances across different communities while minimizing the total feature importances within each community. For an undirected graph characterizing the protein, nodes can be used to represent residues, and edges can be used to represent weighted C α distances. For node i in community C_m , the inner edges of i are defined as the summation of edge values between node i and any other node in C_m , whereas the external edges of i are defined as the summation of edge values between node i and any other node in other communities. For each iteration of ML communities partition, node i can be moved to another community or swapped with another node in different communities. The benefit of these two explorative moves can be calculated as the external edges subtracted by the inner edges.

1. ML communities are randomly partitioned;
2. The benefits of moving one node into another community and swapping one node with another between different communities are estimated to search for the maximum moving and swapping strategy, respectively;
3. One moving or swapping strategy with the highest benefit is chosen;
4. Repeat step 2 and 3 with new ML community configuration until the highest benefit of all moving and swapping strategy is less than 0;
5. ML communities construction is completed if any strategy will increase the number of inner edges for each ML community.

The Kernighan-Lin algorithm⁵¹ has been implemented⁵² to search for local minimum values in graph theory. In the current research, the feature importances of C α distances

from the one-vs-one random forest model based on AuLOV dimer simulations were used. In order to apply ML based community analysis on monomer, averaged importance for each C α distance in monomer was calculated based on the dimer feature importance results.

Transition Path Theory

Transition path theory (TPT)^{53,54} is used to identify the most probable routes from one macrostate to another. Dark state and Light state were chosen based on the transition probability estimated in MSMs as initial and final states, respectively. All other states are considered as transition states. Possible transition paths from the dark state to the light state were calculated. The definition of the committor probability q_i^+ is the probability from one state to a target state. Based on this definition, q_i^+ is equal to zero for all microstates in initial state and q_i^+ is equal to one for all microstates in final state. The committor probability of other microstates is calculated as:

$$-q_i^+ + \sum_{k \in I} T_{ik} q_k^+ = - \sum_{k \in \text{target state}} T_{ik} \quad (7)$$

where T_{ik} is the transition probability from state i to state k .

$$f_{ij} = \pi_i q_i^- T_{ij} q_j^+ \quad (8)$$

where π_i and q_i^- are the equilibrium probability of T_π and the backward-committor probability, calculated as $q_i^- = 1 - q_i^+$, respectively. The backward flux f_{ji} were also considered and subtracted in calculating the net flux $f_{ij}^+ = \max(0, f_{ij} - f_{ji})$.

The flux from initial state to final state can be decomposed to individual pathways p_i , which can be calculated as:

$$p_i = \frac{f_i}{\sum_j f_j} \quad (9)$$

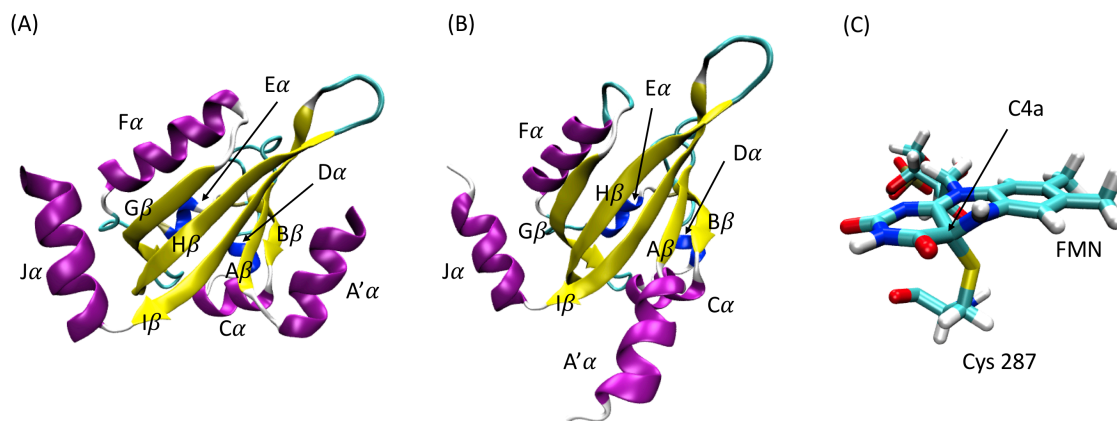


Figure 1: The native dark and light structures of AuLOV monomer. (A) Native dark structure; (B) Native light structure; (C) A covalent bond is formed between C4a position of FMN and sulfur atom in Cys 287 upon light perturbation.

Results

MD simulations analysis

The native dark and light structures of AuLOV are illustrated in Figure 1. In the native light state, a covalent bond is formed between C4a in FMN and sulfur in residue Cys287 upon light perturbation (Figure 1C). This covalent bond triggers a global conformational change and protein dimerization. To explore this effect and the protein response, the covalent bond between FMN and Cys287 is constructed in the dark state structure to construct a transient dark state. On the other hand, the covalent bond between FMN and Cys287 is removed in the light state structure to construct the transient light state. Both transient dark and transient light states are subjected to the simulation and analysis to aid in mapping allosteric trajectories in response to blue-light activation and thermal reversion to the dark-state.

The time evolution of the RMSD in four trajectories is plotted in Figure 2. All RMSD values were calculated with reference to the first frame of each trajectory. The average RMSDs in native dark, native light, transient dark and transient light states are 1.75Å, 2.04Å, 2.39Å and 2.08Å, respectively. The plots show that each simulation is stable with

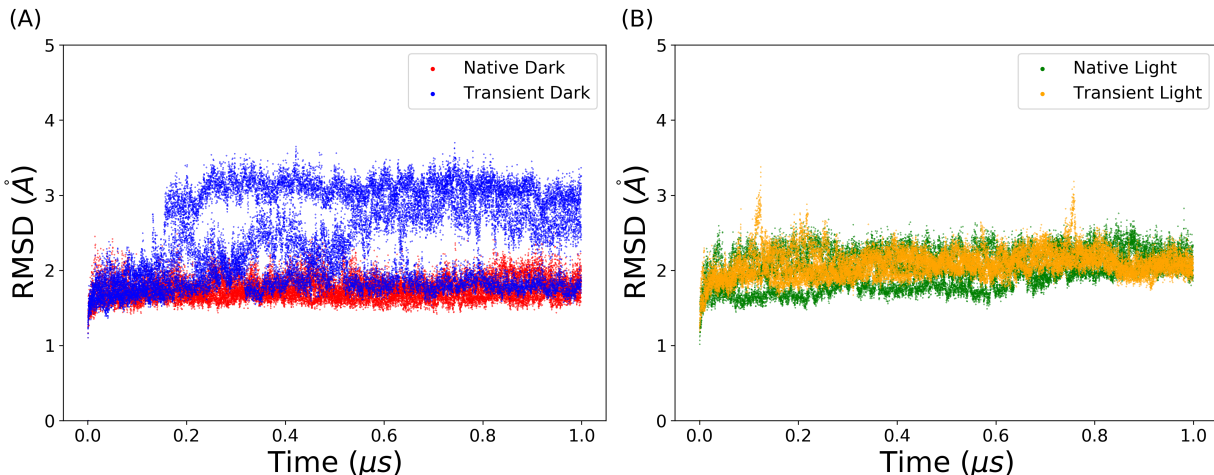


Figure 2: The RMSDs of AuLOV MD simulation trajectories. (A) Native dark and transient dark states; (B) Native light and transient light states.

low RMSD fluctuation values. The transient dark state is more dynamically active than the native dark state, indicating that the formation of covalent bond increases the flexibility of protein. The RMSD results also imply the stability of the native dark state compared with the native light state.

As the allosteric process of AuLOV is characterized by conformational changes in the secondary structures, the backbone $C\alpha$ is selected to measure the influence of light absorption on the protein structure. The RMSFs of the $C\alpha$ atoms in AuLOV simulations are calculated and plotted in Figure 3. Both $A'\alpha$ and $J\alpha$ helices were found to be dynamically active in all four states with increased dynamics in the two transient states. Differences between the two chains can be further quantified by comparing the RMSF values. In the native dark state, chains A and B showed no difference. In the native light state, the $A'\alpha$ in chain A is more flexible than that in chain B. Through the formation of the covalent bond in the transient dark state, both $A'\alpha$ and $J\alpha$ in chain A showed enhanced flexibility.

Markov state model partitions kinetically separate macrostates

To represent the protein structure and movements, pair-wised $C\alpha$ distances were calculated as the representation of protein configurations. A total of 32,131 $C\alpha$ distances were

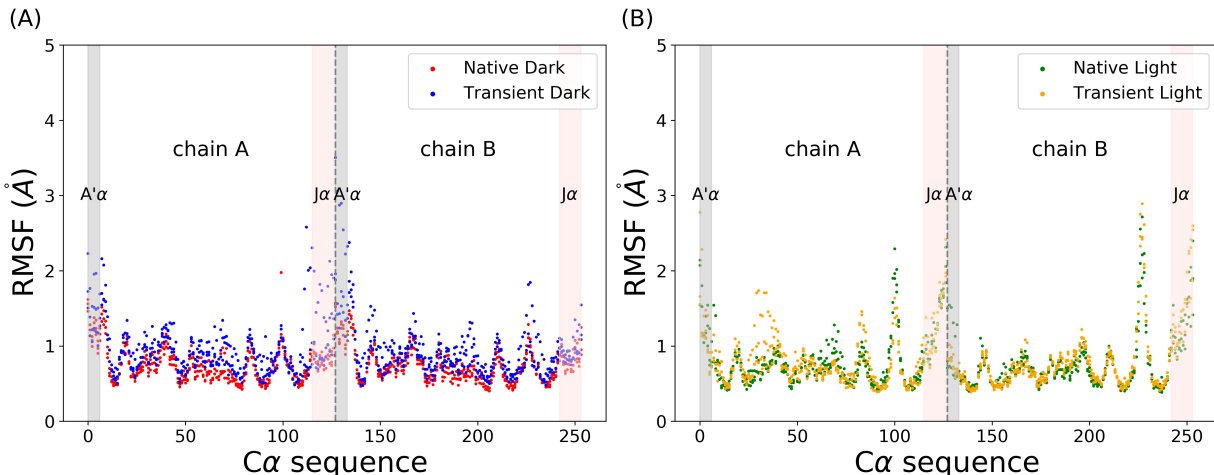


Figure 3: The RMSFs of AuLOV $C\alpha$ atoms in simulations. (A) Native dark and transient dark states; (B) Native light and transient light states. Grey dashed lines separate two chains.

extracted from the AuLOV dimer, composed of 254 residues. For each $C\alpha$ distance, the value was further encoded through the feature preprocessing method outlined in the methods. For feature transformation, 10.0\AA was chosen as the threshold for feature transformation. The ivis dimensionality reduction method was applied to extract the collective variables and project the embedding layer into a 2D map. The distribution of four states in the ivis result is plotted in Figure 4A. The plot revealed that the transient dark state partially overlaps with the native dark state and the transient light state. The large region in transient dark state is mainly because of the enhanced dynamics brought by the formation of the covalent bond. For light states, there are two major regions of native light state, being connected with the transient light state.

Markov state model is based on clustering results on the reduced dimension projected by ivis framework. To construct MSMs, MiniBatch k-means clustering method was applied to partition protein spaces in the 2D region into 300 microstates. The top 20 relaxation timescales calculated by different MSMs with different lagtimes are shown in Figure 4B. The implied timescale converges after $40ns$, which was chosen as the lagtime for MSM. The number of macrostates depends on the gap between the timescales, and a total of

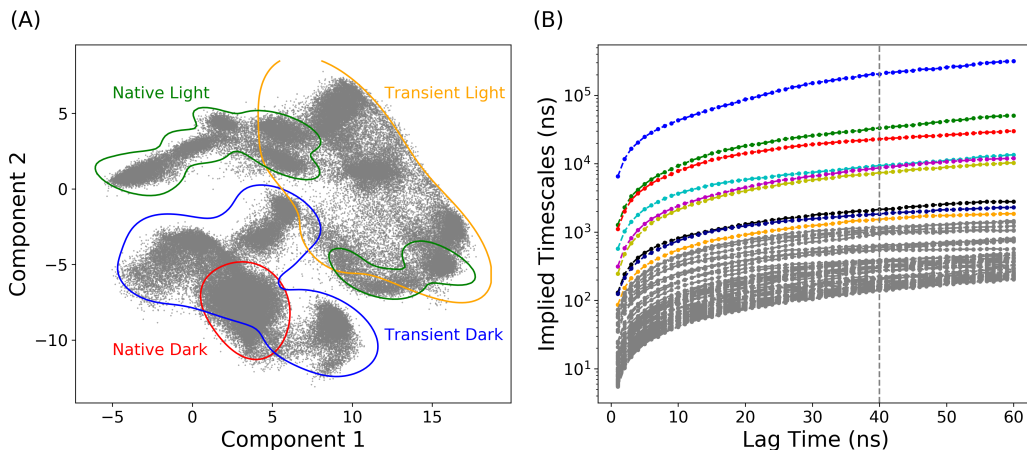


Figure 4: ivis dimensionality reduction result and the implied timescales with regard to different lagtimes. (A) The distribution of four protein states onto the 2-dimensional space; (B) Estimated relaxation timescale with different lag times calculated by MSMs.

10 macrostates were chosen to divide the reduced protein spaces into kinetically separate macrospaces. For each microstate, the corresponding labels of macrostates were determined by the PCCA method, which is based on the eigenfunction of the transition probability matrix in MSM. The resulting macrostates with their associated transition probabilities are illustrated in Figure 5. Two dark states and two light states were divided into 4 and 6 macrostates, respectively. For macrostates (state 1, 2, 3 and 10) that are in the area of the native dark state and the transient dark state, based on the similarity of crystal dark structure, state 2 was treated as the native dark state. State 9 was recognized as the native light state using the same method while other macrostates were considered as transition states.

One-vs-one random forest model extracts key residues

In order to extract the key residues that play a vital role in AuLOV allostery, supervised machine learning models were applied to explore the structural differences among macrostates. Here, pair-wised $C\alpha$ distances were chosen as the translation and rotation invariant collective variables for the description of protein structures in the simulations. For

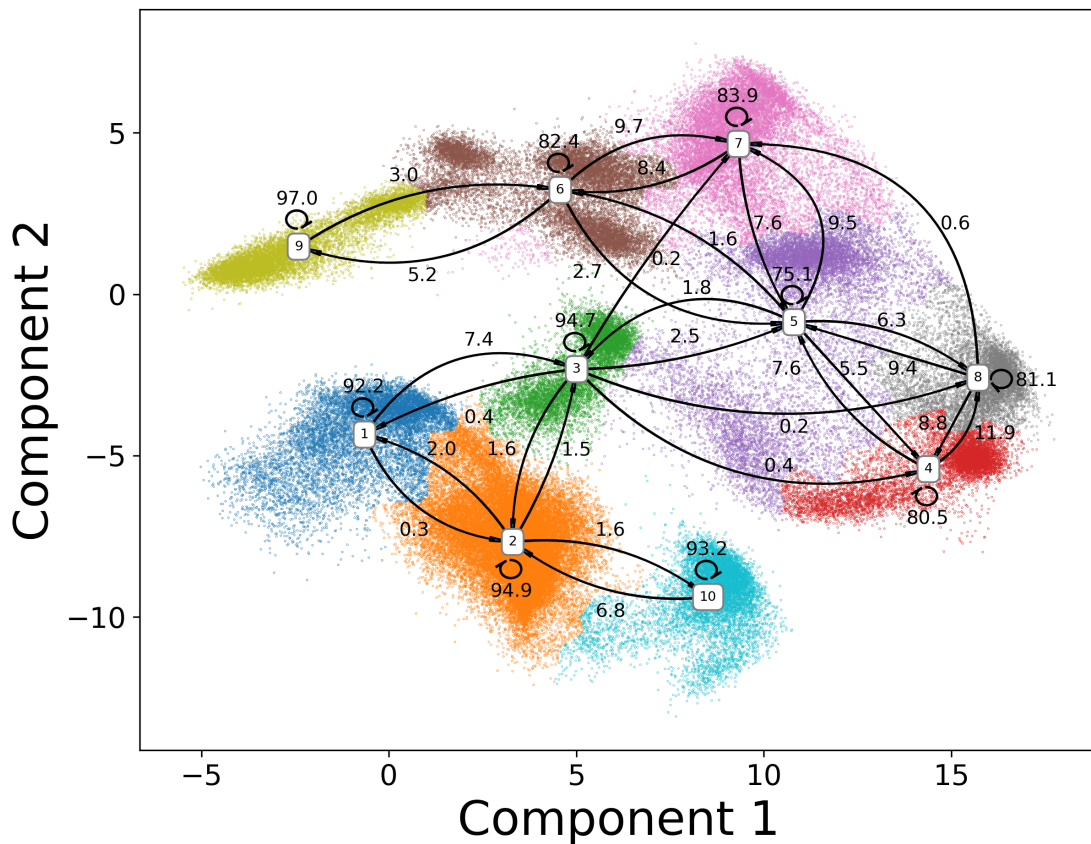


Figure 5: Macrostates in MSM with transition probability. Based on the transition probabilities, states 2 and 9 were considered as the native dark and native light states among macrostates, respectively. Other macrostates were treated as transition states.

each simulation, frames were saved for every 100 picoseconds (ps), resulting in 10,000 frames for every $1 \mu s$ MD trajectory. Accordingly, 120,000 samples with 32,131 features were extracted from the simulated trajectories. Each frame was labeled based on the macrostate results. Random forest and one-vs-one random forest models were applied to distinguish the intrinsic conformational differences among macrostates. Training scores and testing scores were plotted in Figure 6. The testing accuracy was 93.5% in random forest model at depth 9 and 94.5% in OvO random forest model at depth 8. The high classification accuracy indicated that the two tree-based models were able to capture the characteristics of protein configuration of each macrostate using pairwise $C\alpha$ distances.

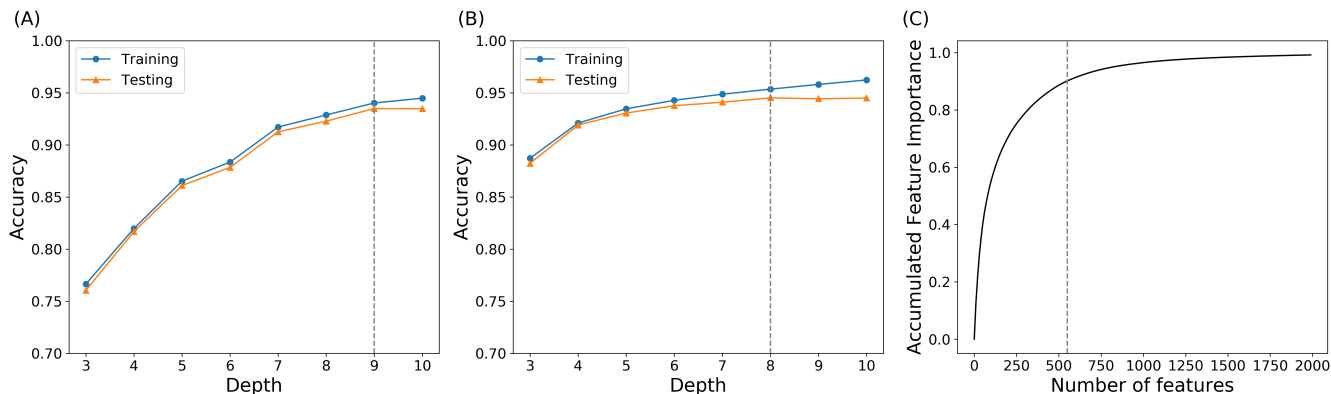


Figure 6: Tree-based models for macrostate classification. (A) Prediction accuracy of random forest model with different tree depths; (B) Prediction accuracy of one-vs-one random forest model with different tree depths; (C) Accumulated explained importance of OvO random forest model in 8 tree depth with regard to the number of features. The top 550 features account for 90.2% of the overall importance.

The advantage of tree-based models comes from the ability to quantitatively evaluate the contribution of each feature in classification model through the value of feature importance. Superior than random forest model, one-vs-one random forest model was applied to compute the feature importance for any two different macrostate pairs by conducting a random forest classifier between these two specific macrostates. Therefore, for any two different macrostates, one distinct random forest estimator was built. A combination of $10 * 9 / 2 = 45$ basic random forest classifiers were constructed for pairwise macrostates classification. Accumulated feature importance of one-vs-one random forest at depth 8 was plotted in Figure 6C. Overall, this method was an effective model, in which the top 550 features out of 32,131 features accounted for 90.2% of overall feature importance.

The feature importance of $C\alpha$ distances that are related with two residues located on different chains, named as cross-monomer features, accounts for 59.77% of the overall importance. Therefore, the $C\alpha$ distances within one chain accounts for 40.23%. This shows that the OvO random forest can capture the structural changes within each monomer, as well as the relative motions between monomers.

In order to identify key residues based on the results of the OvO random forest model,

Table 1: Top 20 residues identified by OvO random forest.

Residue ID	Residue Type	Residue ID	Residue Type
250 ^a	GLN	349	VAL
245	LEU	247	THR
252	PHE	248	ALA
351	CYS	249	GLN
244	ALA	331	PHE
312	ASP	314	SER
246	GLN	334	ALA
268	SER	313	MET
350	GLN	336	LEU
335	ALA	251	ASN

^a Experimentally confirmed important residues are shown in bolded font.

Table 2: Accumulated feature importance of secondary structures in AuLOV.

Secondary structure	Importance percentage
A' α	15.17%
A β	7.12%
B β	2.28%
C α	9.12%
D α	0.70%
E α	0.05%
F α	2.14%
G β	6.43%
H β	14.13%
I β	13.46%
J α	8.04%
Linkers	21.36%

the feature importance value of each C α distance was added and accumulated to the two related residues. The top 20 residues were listed in Table 1. Among the identified residues, several have been experimentally confirmed to be important to allostery and are depicted in bold font. Met313, Phe331 and Cys351 are found to undergo changes in orientation. Ala248, Gln249, Gln250 and Asn251 are residues linking the A' α helix to the A β strand that are important for signal transduction⁸. Gln350 was also identified as essential for signal transduction in LOV domains, where it either undergoes a Gln-flip process in response to N5 protonation⁵⁵ or undergoes rotation between exposed and buried conformations⁵⁶ to relay signal transduction from the flavin active site to N- or C-terminal components. We also

identify Phe252 as important for allostery. Notably, Phe252 was found by HDX-MS to be important in destabilization of A' α that is coupled to conformational changes in B β and C α ⁸. Therefore, the OvO random forest can successfully identify important residues reported in experimental results. The residue importance can be accumulated to the protein's secondary structures and the results were shown in Table 2. A' α and J α account for 15.17% and 8.04% of overall importance, respectively. It is also worth noting that the importance of C α helix and linkers in AuLOV are also significant at 9.12% and 21.36%, respectively.

Machine learning-driven community analysis identifies C α helix and linkers important in promoting conformational change

To explore the significance of different protein secondary structures, machine learning-driven community analysis was applied to split the protein structure into communities. The machine learning based community (referred to as ML community) analysis was developed to divide residues into several communities so that the feature importance for pairwise C α distances across different communities is maximized, while the feature importance within each community is minimized.

The relationship between the feature importance for pairwise C α distances within ML communities and the number of ML communities were plotted in Figure 7A. Applying an elbow criterion, four ML communities were selected with the total feature importance within each ML community accounting for 0.50% and the total feature importance among ML communities accounting for 99.50%. Therefore, the changes among ML communities accounted for the dominant majority of the overall feature importance and were able to explain the changes between different communities. The changes within each ML community were discarded due to the negligible importance. By applying ML based community analysis, dynamics in each protein structure can be attributed to the changes within partitioned ML communities.

The distribution of different communities, with a complete partition result corresponding

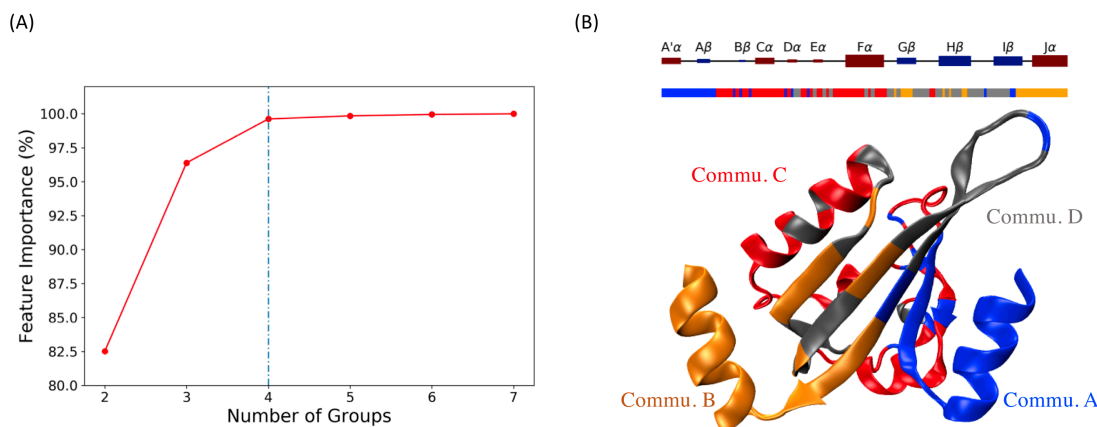


Figure 7: ML based Community analysis results of AuLOV. (A) Total feature importance among ML communities with regard to different number of communities; (B) Four ML communities named Commu. A, Commu. B, Commu. C and Commu. D as blue, orange, red and grey colors, respectively.

Table 3: Accumulated feature importance between each ML community pair.

Features	Commu. A	Commu. B	Commu. C	Commu. D
Commu. A	0.12%	13.58%	25.87%	13.87%
Commu. B		0.03%	13.17%	16.44%
Commu. C			0.15%	16.57%
Commu. D				0.20%

to protein secondary structure, is shown in Figure 7B. Commu. A (blue) includes most of A'α helix and Aβ sheet, Commu. B (orange) includes Jα helix with part of Gβ and Hβ sheet on the LOV core, Commu. C (red) includes Cα helix, part of Fα and linkers while Commu. D (gray) includes part of Fα helix, Gβ, Hβ and Iβ sheet.

The machine learning based community analysis offered additional information based on the selected four ML communities and the corresponding different regions in the protein structure during simulation. The accumulated overall feature importance among each ML community pair was listed in Table 3. Correlations between Commu. A, Commu. B and the rest of the protein accounted for 82.99% of total feature importance. This was not surprising since the A'α in Commu. A and Jα in Commu. B are the most distinguishing parts, which undergo significant conformational changes from native dark to native light state. Through

Table 4: The changes of Commu. A, Commu. B and Commu. C during transitions between states.

Adjacent macrostates	Commu. A	Commu. B	Commu. C
State 2 (Dark) \rightarrow State 3 ^a	7.26%	26.69%	4.01%
State 2 \rightarrow State 10	6.05%	18.51%	3.03%
State 3 \rightarrow State 5	16.91%	12.74%	7.68%
State 5 \rightarrow State 7	7.44%	2.04%	18.41%
State 5 \rightarrow State 4	9.55%	6.40%	3.64%
State 5 \rightarrow State 6	10.96%	6.98%	16.27%
State 7 \rightarrow State 6	11.22%	3.97%	8.86%
State 6 \rightarrow State 9 (Light)	12.35%	10.56%	19.13%
Non-Adjacent macrostates			
State 10 \rightarrow State 9	21.39%	23.63%	0.23%
State 10 \rightarrow State 6	26.97%	15.11%	2.26%

^a Bolded is state-transitions with large Commu. C component.

the accumulated feature importance of ML communities, $A'\alpha$ and $J\alpha$ were confirmed to convey significant allosteric characteristics. However, it should be noted that after excluding the importance of Commu. A and B, the feature importance of Commu. C still accounts for 16.57% of total feature importance. Several transitions between adjacent macrostate pairs were found to be related with large Commu. C component, shown in Table 4. However, for transitions between non-adjacent macrostates, the position of Commu. C accounts for less importance explaining the difference between macrostate pairs.

For those transitions between states where Commu. C accounts for a large component, two promising routes from dark state to light state can be formed as: 1) State 2 \rightarrow 3 \rightarrow 5 \rightarrow 7 \rightarrow 6 \rightarrow 9 and 2) State 2 \rightarrow 3 \rightarrow 5 \rightarrow 6 \rightarrow 9. The low transition possibilities starting from state 2 and 9 to adjacent states indicated the stability of both the dark and light states. On the contrary, it is more likely for protein to shift between transition states. These two proposed pathways led a hypothesis that Commu. C is important in propagating allosteric perturbations.

To estimate the probability of the two identified channels which present a large Comm. C component, the transition pathway theory was employed to generate an ensemble of pathways to calculate the probability of every pathway from state 2 (dark region) to state 9 (light

Table 5: The probability of top 10 channels calculated by transition path theory.

Channels	Probability
2, 3, 5, 7, 6, 9	28.8%
2, 3, 5, 6, 9	25.6%
2, 10, 5, 7, 6, 9	5.4%
2, 3, 7, 6, 9	4.7%
2, 3, 5, 8, 5, 7, 6, 9	3.9%
2, 3, 4, 8, 5, 7, 6, 9	3.0%
2, 3, 8, 5, 7, 6, 9	2.9%
2, 10, 5, 6, 9	2.3%
2, 1, 3, 5, 7, 6, 9	2.1%
2, 10, 5, 8, 5, 7, 6, 9	1.3%
Top 10 channels	80.0%

region). A total of 3,151 pathways were generated and divided as 212 distinct channels floating from these two states. The probability of each channel was calculated based on net flux from the initial state to the target state. Overall, the probability of top 10 channels was listed in the Table 5, with the contributions from these 10 channels accounting for 80.0% of total population.

Among all 212 channels, the two proposed channels 2-3-5-7-6-9 and 2-3-5-6-9 are the top 2 populated channels with 28.8% and 25.6% of overall probability, respectively. The sum of contributions from two proposed channels accounts for 54.4% contributions, which is significant compared with the rest pathways, suggesting the importance of Commu. C movement during the allosteric process. The first channel is more probable than the second channel. This agrees with the observation that the transition probability from 5 to 7 (9.5%) as one step in the first channel is greater than that from 5 to 6 (1.6%) as one step in the second channel. Interestingly, the ML based community analysis shows higher contribution from the Commu. C to the transition between state 5 and 7 than that between state 5 and 6 .

Different communities account for different importance in each macrostate transition. To better show the trend of components in Commu. A, Commu. B and Commu. C with regard to Commu. D, the change of importance along the two proposed paths was plotted in Figure

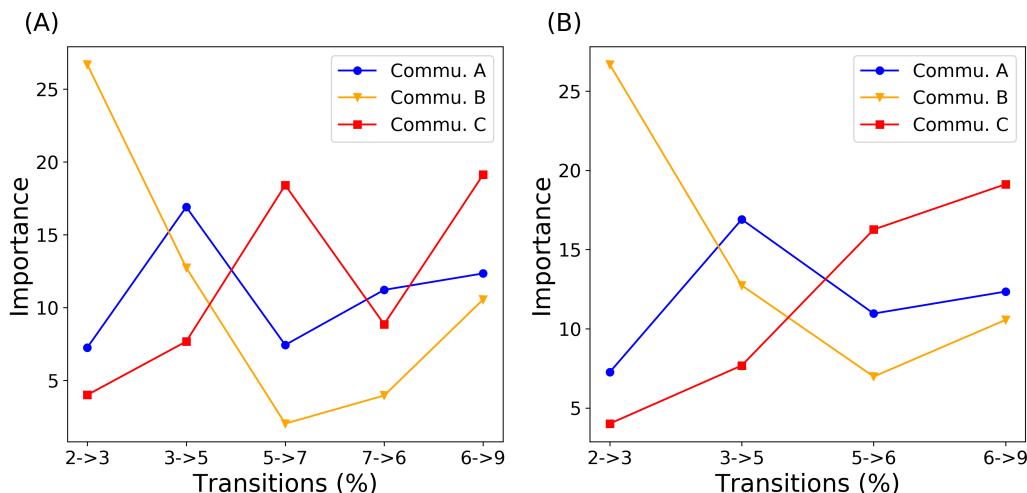


Figure 8: Change of community importance along transitions. Community importance change along two identified paths from dark state to light state. (A) 2 (dark) - 3 - 5 - 7 - 6 - 9 (light); (B) 2 (dark) - 3 - 5 - 6 - 9 (light).

8. Two paths shared similar characteristics: 1) Commu. A accounts for little importance at the beginning of allosteric process while the contribution goes up in later transitions; 2) Commu. B starts with high importance and decreases drastically after the first transition; 3) Commu. C is more important at the end of allosteric process.

Discussion

PtAula is an allosteric protein that undergoes conformational change upon light activation beginning with the formation of covalent bond between Cys287 and FMN.⁵⁷ This computational study of AuLOV is integrated with MD simulations and other computational methods to provide quantitative analysis of the dynamics and importance of residues with regard to the overall allosteric process. While there is extensive research on the regulatory role of $J\alpha$ and dimerization controlling helix $A'\alpha$, a detailed mechanism of allostery with signal transmission route still needs scrutinization.

Signal transduction in LOV domain containing proteins typically involves coupling of adduct formation to conformational changes in the N- and C-termini via propagation across

a central β -sheet⁵⁸⁻⁶¹. Central to this signal transduction are key residues within the I β strand that enables its coupling with the J α helix and interaction with A' α in the dark state, specifically the residue equivalent to Gln350 that is essential for LOV signal transduction^{55,56}. In AuLOV, several additional light-induced rotamers (Met313, Leu317, Leu331, Leu333, and Cys351) were observed on the β -sheet surface⁸. Here, through the accumulated residue importance in the one-vs-one random forest model, we successfully identify Met313, Leu331 and Leu351 as being important in differentiating allosteric changes in AuLOV. In our models, these residues contribute to conformational changes linking the β -sheet surface to A' α through Gln350. We note that our computational methods mirror those identified experimentally where A' α contributes to the dynamic stability of the dark state by the interaction with LOV core through a hinge region. The hinge region consists of four conserved residues (Ala248, Gln249, Gln250 and Asn251), which were also found to be important via our approach (Table 1). Overall, the strong correlations between previous experimental results, and our Markov state model and OvO random forest analysis, confirms our methodology as being able to discern allosteric pathways in AuLOV.

Although chain A and chain B in AuLOV are dynamically identical in the dark state, the A' α helices of the two chains differ in conformations upon dimerization⁸. Our simulation results confirmed the differences between these two chains through a comparison of RMSF values. The RMSF results reflect that A' α in chain A is more dynamically active than that in chain B. The asymmetry property in A' α could originate from either the interaction between A' α and J α on different chains or the asymmetrical conformational change⁸, thus requiring further detailed study.

ML based community analysis used in this study provided an approach to partition protein conformation into communities based on the feature importance of pairwise C α distances. Through this analysis, three important communities were identified. Commu. A containing A' α and Commu. B containing J α were expected to account for great contribution, since these two helices undergo notably conformational changes upon light activation

(Table 2). The $C\alpha$ and $F\alpha$ helices stand out as Commu. C surprisingly provided additional information for allosteric process. Commu. C accounts for great importance in adjacent transitions between macrostates and accounts for less importance in nonadjacent transitions compared with Commu. A and B.

Transition path theory calculations further validated the important allosteric function of Commu. C. For all possible transition paths found by TPT, the top 2 channels are those with large Commu. C components and together constitute over 50% of the overall possibility. Although Commu. C consists of two helices as $C\alpha$ and $F\alpha$, these two helices are not equally important. The allosteric role of $F\alpha$ should be evaluated with caution since its accumulated feature importance is relatively low (Table 2), and the importance in Commu. D, which also includes part of $F\alpha$, is the least important community. Because $C\alpha$ helix is important in both OvO random forest result and ML based community analysis, it is reasonable to conclude that $C\alpha$ may play an important role in controlling AuLOV allostery. Moreover, Commu. C also includes several linking residues that account for a large portion of the overall importance, indicating the indispensable role of linkers in the allosteric process as reported in previous studies^{24,62}.

Examination of the two most probable channels linking conformational changes through the identified communities can allow construction of allosteric paths (Figure 8). In this study we identify that the $J\alpha$ helix is fundamental in the early stage of AuLOV allostery, followed by changes in the $A'\alpha$ helix in later stages. In the first transition step from state 2 \rightarrow state 3, Commu. B accounts for a large component compared with Commu. A, indicating the importance of $J\alpha$ in the initial stage of allostery. As the allosteric perturbation propagates, the importance of Commu. B decreases and Commu. A becomes the more significant region. This important shift implied and confirmed the experimental finding that, after initial Cys287-FMN covalent bond formation, the first response of the protein structure is the undocking of the $J\alpha$, which is essential to the release of $A'\alpha$ ^{56,63,64}. The rising importance of Commu. C, together with the transition path theory results, suggests that Commu. C,

especially C α helix and linkers, is vital in the allosteric process and should be investigated further.

Conclusion

The LOV protein PtAu1a is a member of Aureochrome family that binds DNA upon blue light activation⁵. Studies of the LOV domain with N- and C-terminal helices indicate that in the absence of light it exists as monomeric units; upon blue light absorption, cysteinyl-flavin bond formation triggers a global conformational change which ultimately results in the dimerization of the LOV domains. In the present study, the protein dynamics of AuLOV with N- and C-terminal helices were simulated using MD simulations and analyzed using a series of computational methods. We quantified the differences of A' α and J α helices dynamics in four functional states and the importance of each residue in the two chains with regard to the protein allosteric process. Key residues in overall structural changes identified by one-vs-one random forest agree with the results reported in other experimental work. Markov state model, combined with transition path theory, studied the importance of protein structures by a machine learning based community analysis. The functional role of key Community C, which includes the C α helix and linkers, is revealed through in-depth analysis as propagating the allosteric perturbation. Overall, this study quantitatively analyzed the allostery process of AuLOV and linked the macroscopic conformational change to residue level importance. Our results provided new opportunities for a detailed mechanism explanation and offered further opportunities for the research of PtAu1a as an optogenetic tool. Future studies can facilitate our understanding of global protein conformational changes in the context of full-length PtAu1a.

Acknowledgement

We gratefully acknowledge funding sources, including NIH research grant R15GM122013 to PT and NIH research grant R15GM109282 to BDZ. Computational time was generously provided by Southern Methodist University’s Center for Research Computing. The authors thank Ms. Xi Jiang from the Biostatistics Ph.D. program in the Statistics department of SMU for fruitful discussions.

References

- (1) Herrou, J.; Crosson, S. Function, structure and mechanism of bacterial photosensory LOV proteins. *Nature reviews microbiology* **2011**, *9*, 713–723.
- (2) Christie, J. M.; Salomon, M.; Nozue, K.; Wada, M.; Briggs, W. R. LOV (light, oxygen, or voltage) domains of the blue-light photoreceptor phototropin (nph1): binding sites for the chromophore flavin mononucleotide. *Proceedings of the National Academy of Sciences* **1999**, *96*, 8779–8783.
- (3) Pudasaini, A.; El-Arab, K. K.; Zoltowski, B. D. LOV-based optogenetic devices: light-driven modules to impart photoregulated control of cellular signaling. *Frontiers in molecular biosciences* **2015**, *2*, 18.
- (4) Loros, J. J.; Dunlap, J. C. Genetic and molecular analysis of circadian rhythms in *neurospora*. *Annual review of physiology* **2001**, *63*, 757–794.
- (5) Takahashi, F.; Yamagata, D.; Ishikawa, M.; Fukamatsu, Y.; Ogura, Y.; Kasahara, M.; Kiyosue, T.; Kikuyama, M.; Wada, M.; Kataoka, H. AUREOCHROME, a photoreceptor required for photomorphogenesis in stramenopiles. *Proceedings of the National Academy of Sciences* **2007**, *104*, 19625–19630.

- (6) Hepp, S.; Trauth, J.; Hasenjäger, S.; Bezold, F.; Essen, L.-O.; Taxis, C. An optogenetic tool for induced protein stabilization based on the *Phaeodactylum tricornutum* aureochrome 1a LOV domain. *Journal of Molecular Biology* **2020**,
- (7) Herman, E.; Sachse, M.; Kroth, P. G.; Kottke, T. Blue-light-induced unfolding of the J α helix allows for the dimerization of aureochrome-LOV from the diatom *Phaeodactylum tricornutum*. *Biochemistry* **2013**, *52*, 3094–3101.
- (8) Heintz, U.; Schlichting, I. Blue light-induced LOV domain dimerization enhances the affinity of Aureochrome 1a for its target DNA sequence. *Elife* **2016**, *5*, e11860.
- (9) Hisatomi, O.; Nakatani, Y.; Takeuchi, K.; Takahashi, F.; Kataoka, H. Blue light-induced dimerization of monomeric aureochrome-1 enhances its affinity for the target sequence. *Journal of Biological Chemistry* **2014**, *289*, 17379–17391.
- (10) Losi, A.; Gärtner, W. Bacterial bilin-and flavin-binding photoreceptors. *Photochemical & photobiological sciences* **2008**, *7*, 1168–1178.
- (11) Crosson, S.; Rajagopal, S.; Moffat, K. The LOV domain family: photoresponsive signaling modules coupled to diverse output domains. *Biochemistry* **2003**, *42*, 2–10.
- (12) Malmstrom, R. D.; Kornev, A. P.; Taylor, S. S.; Amaro, R. E. Allostery through the computational microscope: cAMP activation of a canonical signalling domain. *Nature communications* **2015**, *6*, 1–11.
- (13) Ruschak, A. M.; Kay, L. E. Proteasome allostery as a population shift between interchanging conformers. *Proceedings of the National Academy of Sciences* **2012**, *109*, E3454–E3462.
- (14) Weinkam, P.; Pons, J.; Sali, A. Structure-based model of allostery predicts coupling between distant sites. *Proceedings of the National Academy of Sciences* **2012**, *109*, 4875–4880.

- (15) Li, H.; Wu, H.; Li, B.; Gao, Y.; Zhao, X.; Zhang, L. Molecular dynamics simulation of fracture mechanism in the double interpenetrated cross-linked polymer. *Polymer* **2020**, 122571.
- (16) Zhang, H.; Li, H.; Hu, F.; Wang, W.; Zhao, X.; Gao, Y.; Zhang, L. Cavitation, crazing and bond scission in chemically cross-linked polymer nanocomposites. *Soft matter* **2019**, 15, 9195–9204.
- (17) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics* **2011**, 134, 174105.
- (18) Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L.-P.; Shukla, D. et al. OpenMM 4: a reusable, extensible, hardware independent library for high performance molecular simulation. *Journal of chemical theory and computation* **2013**, 9, 461–469.
- (19) Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J. et al. Millisecond-scale molecular dynamics simulations on Anton. Proceedings of the conference on high performance computing networking, storage and analysis. 2009; pp 1–11.
- (20) Suárez, E.; Adelman, J. L.; Zuckerman, D. M. Accurate estimation of protein folding and unfolding times: beyond Markov state models. *Journal of chemical theory and computation* **2016**, 12, 3473–3481.
- (21) Adelman, J. L.; Ghezzi, C.; Bisignano, P.; Loo, D. D.; Choe, S.; Abramson, J.; Rosenberg, J. M.; Wright, E. M.; Grabe, M. Stochastic steps in secondary active sugar transport. *Proceedings of the National Academy of Sciences* **2016**, 113, E3960–E3966.
- (22) Zhou, H.; Wang, F.; Tao, P. t-Distributed Stochastic Neighbor Embedding Method

- with the Least Information Loss for Macromolecular Simulations. *Journal of chemical theory and computation* **2018**, *14*, 5499–5510.
- (23) Levy, R.; Srinivasan, A.; Olson, W.; McCammon, J. Quasi-harmonic method for studying very low frequency modes in proteins. *Biopolymers: Original Research on Biomolecules* **1984**, *23*, 1099–1112.
- (24) Naritomi, Y.; Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *The Journal of chemical physics* **2011**, *134*, 02B617.
- (25) Maaten, L. v. d.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*, 2579–2605.
- (26) Amid, E.; Warmuth, M. K. A more globally accurate dimensionality reduction method using triplets. *arXiv preprint arXiv:1803.00854* **2018**,
- (27) Szubert, B.; Cole, J. E.; Monaco, C.; Drozdov, I. Structure-preserving visualisation of high dimensional single-cell datasets. *Scientific reports* **2019**, *9*, 1–10.
- (28) Tian, H.; Tao, P. ivis Dimensionality Reduction Framework for Biomacromolecular Simulations. *arXiv preprint arXiv:2004.10718* **2020**,
- (29) Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73–76.
- (30) Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry* **2015**, *115*, 1094–1101.
- (31) Botu, V.; Ramprasad, R. Adaptive machine learning framework to accelerate ab initio

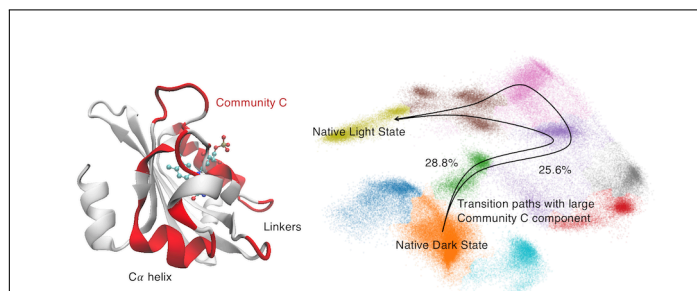
- molecular dynamics. *International Journal of Quantum Chemistry* **2015**, *115*, 1074–1083.
- (32) Kotsiantis, S. B.; Zaharakis, I.; Pintelas, P. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering* **2007**, *160*, 3–24.
- (33) Jing, Y.; Bian, Y.; Hu, Z.; Wang, L.; Xie, X.-Q. S. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. *The AAPS journal* **2018**, *20*, 58.
- (34) Zhou, H.; Dong, Z.; Verkhivker, G.; Zoltowski, B. D.; Tao, P. Allosteric mechanism of the circadian protein Vivid resolved through Markov state model and machine learning analysis. *PLoS computational biology* **2019**, *15*, e1006801.
- (35) Berman, H. M.; Bhat, T. N.; Bourne, P. E.; Feng, Z.; Gilliland, G.; Weissig, H.; Westbrook, J. The Protein Data Bank and the challenge of structural genomics. *Nature structural biology* **2000**, *7*, 957–959.
- (36) Freddolino, P. L.; Gardner, K. H.; Schulten, K. Signaling mechanisms of LOV domains: new insights from molecular dynamics studies. *Photochemical & Photobiological Sciences* **2013**, *12*, 1158–1170.
- (37) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983**, *79*, 926–935.
- (38) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *The Journal of chemical physics* **1995**, *103*, 8577–8593.

- (39) Eastman, P.; Pande, V. OpenMM: A hardware-independent framework for molecular simulations. *Computing in science & engineering* **2010**, *12*, 34–39.
- (40) Foloppe, N.; MacKerell, A. D., Jr All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of computational chemistry* **2000**, *21*, 86–104.
- (41) Wang, F.; Zhou, H.; Olademehin, O. P.; Kim, S. J.; Tao, P. Insights into key interactions between vancomycin and bacterial cell wall structures. *ACS omega* **2018**, *3*, 37–45.
- (42) Tian, H.; Tao, P. Deciphering the Protein Motion of S1 Subunit in SARS-CoV-2 Spike Glycoprotein Through Integrated Computational Methods. *arXiv preprint arXiv:2004.05256* **2020**,
- (43) Liaw, A.; Wiener, M., et al. Classification and regression by randomForest. *R news* **2002**, *2*, 18–22.
- (44) Wang, F.; Shen, L.; Zhou, H.; Wang, S.; Wang, X.; Tao, P. Machine Learning Classification Model for Functional Binding Modes of TEM-1 β -Lactamase. *Frontiers in molecular biosciences* **2019**, *6*, 47.
- (45) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research* **2011**, *12*, 2825–2830.
- (46) Zhou, H.; Dong, Z.; Tao, P. Recognition of protein allosteric states and residues: Machine learning approaches. *Journal of computational chemistry* **2018**, *39*, 1481–1490.
- (47) Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.
- (48) Wang, F.; Zhou, H.; Wang, X.; Tao, P. Dynamical Behavior of β -Lactamases and Penicillin-Binding Proteins in Different Functional States and Its Potential Role in Evolution. *Entropy* **2019**, *21*, 1130.

- (49) Deuffhard, P.; Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear algebra and its applications* **2005**, *398*, 161–184.
- (50) Harrigan, M. P.; Sultan, M. M.; Hernández, C. X.; Husic, B. E.; Eastman, P.; Schwantes, C. R.; Beauchamp, K. A.; McGibbon, R. T.; Pande, V. S. MSMBuilder: statistical models for biomolecular dynamics. *Biophysical journal* **2017**, *112*, 10–15.
- (51) Lin, S.; Kernighan, B. W. An effective heuristic algorithm for the traveling-salesman problem. *Operations research* **1973**, *21*, 498–516.
- (52) Zhou, H.; Tao, P. REDAN: relative entropy-based dynamical allosteric network model. *Molecular physics* **2019**, *117*, 1334–1343.
- (53) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences* **2009**, *106*, 19011–19016.
- (54) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. Transition path theory for Markov jump processes. *Multiscale Modeling & Simulation* **2009**, *7*, 1192–1219.
- (55) Yee, E. F.; Diensthuber, R. P.; Vaidya, A. T.; Borbat, P. P.; Engelhard, C.; Freed, J. H.; Bittl, R.; Möglich, A.; Crane, B. R. Signal transduction in light–oxygen–voltage receptors lacking the adduct-forming cysteine residue. *Nature communications* **2015**, *6*, 1–10.
- (56) Ashutosh, P.; Shim, J. S.; Song, Y. H.; Shi, H.; Takatoshi, K.; Somers, D. E.; Takato, I.; Zoltowski, B. D. Kinetics of the LOV domain of ZEITLUPE determine its circadian function in Arabidopsis. *eLife* **2017**, *6*.
- (57) Salomon, M.; Christie, J. M.; Knieb, E.; Lempert, U.; Briggs, W. R. Photochemical and mutational analysis of the FMN-binding domains of the plant blue light receptor, phototropin. *Biochemistry* **2000**, *39*, 9401–9410.

- (58) Zayner, J. P.; Antoniou, C.; Sosnick, T. R. The amino-terminal helix modulates light-activated conformational changes in AsLOV2. *Journal of molecular biology* **2012**, *419*, 61–74.
- (59) Zoltowski, B. D.; Schwerdtfeger, C.; Widom, J.; Loros, J. J.; Bilwes, A. M.; Dunlap, J. C.; Crane, B. R. Conformational switching in the fungal light sensor Vivid. *Science* **2007**, *316*, 1054–1057.
- (60) Harper, S. M.; Neil, L. C.; Gardner, K. H. Structural basis of a phototropin light switch. *Science* **2003**, *301*, 1541–1544.
- (61) Halavaty, A. S.; Moffat, K. N-and C-terminal flanking regions modulate light-induced signal transduction in the LOV2 domain of the blue light sensor phototropin 1 from *Avena sativa*. *Biochemistry* **2007**, *46*, 14001–14009.
- (62) Papaleo, E.; Saladino, G.; Lambrughi, M.; Lindorff-Larsen, K.; Gervasio, F. L.; Nussinov, R. The role of protein loops and linkers in conformational dynamics and allostery. *Chemical reviews* **2016**, *116*, 6391–6423.
- (63) Harper, S. M.; Christie, J. M.; Gardner, K. H. Disruption of the LOV- J α Helix Interaction Activates Phototropin Kinase Activity. *Biochemistry* **2004**, *43*, 16184–16192.
- (64) Herman, E.; Kottke, T. Allosterically regulated unfolding of the A' α helix exposes the dimerization site of the blue-light-sensing aureochrome-LOV domain. *Biochemistry* **2015**, *54*, 1484–1492.

Graphical TOC Entry



Title: Deciphering the allosteric process of
Phaeodactylum tricornutum Aureochrome 1a LOV
Domain

Authors: Hao Tian, Francesco Trozzi, Brian D.
Zoltowski and Peng Tao