Structural Plasticity and Functional Dynamics of Pigeon Cryptochrome 4 as Avian Magnetoreceptor ☆

Chuanye Xiong, Palanisamy Kandhan, Brian D. Zoltowski * and Peng Tao *

Department of Chemistry, *Southern Methodist University, Dallas, TX 75275, United States*
O'Donnell Data Science and Research Computing Institute, *Southern Methodist University, Dallas, TX 75275, United States*
Center for Drug Discovery, *Design, and Delivery (CD4), Southern Methodist University, Dallas, TX 75275, United States*

Correspondence to Brian D. Zoltowski and Peng Tao: Department of Chemistry, Southern Methodist University, Dallas, TX 75275, United States. *bzoltowski@smu.edu (B.D. Zoltowski), ptao@smu.edu (P. Tao)*
https://doi.org/10.1016/j.jmb.2025.169233
Edited by Ruth Nussinov

Abstract

Cryptochromes (CRYs) are key flavoproteins involved in biological processes such as circadian rhythm regulation and magnetoreception. Type IV CRYs have been identified as primary candidates for avian magnetoreception. However, their structural flexibility, particularly within the cryptochrome C-terminal extension (CCE) and phosphate-binding loop (PBL), remains poorly understood. In this study, we employed temperature replica exchange molecular dynamics (T-REMD) simulations combined with advanced dimensionality reduction techniques, including autoencoder and time-lagged independent component analysis (t-ICA), to explore the conformational space of *Columba livia* cryptochrome 4 (*Cl*CRY4), as the only available crystal structure of Type IV CRYs to date. By using *Drosophila* cryptochrome (*d*CRY) as a reference structure, we assessed the reliability of T-REMD sampling in capturing key states of *Cl*CRY4. Our results indicate that the CCE region of *Cl*CRY4 displays unique conformational dynamics and cooperative interactions with the PBL, highlighting the need for further investigation. The clustering analysis of *Cl*CRY4 conformations revealed multiple structural states, underscoring the functional significance of its intrinsically disordered regions (IDRs). This study provides a novel computational approach for studies of CRYs dynamics, through which the modeling of one CRY with full structure could be used to benchmark the computational study of another CRY only with partial structural information available.

Introduction

Intrinsically disordered regions (IDRs) of signaling proteins are increasingly recognized as essential regulators in signal transduction pathways, where they often impart dynamic, environmentally sensitive regulation of protein–protein interactions or recognition of small molecule effectors [1–6]. IDRs often mediate transient yet highly specific protein–protein interactions, making them central to signal integration, timing, and feedback in complex biological networks. However, due to their dynamic, stimuli-dependent conformational landscape, these regulatory regions remain enigmatic, due to the lack of experimental approaches capable of delineating their structure [2,7]. Currently, the lack of experimentally defined structures of IDRs also limits the effectiveness of computational approaches in developing small molecules that target specific conformational sub-states within their dynamic ensembles [4].

☆ This article is part of a special issue entitled: 'Allostery in disease (2025)' published in Journal of Molecular Biology.

The cryptochrome (CRY) proteins [8–10] of the CRY-photolyase family (CPF) offer a compelling case study for examining the role of IDRs in signal transduction. CRYs are evolutionarily derived from photolyases that employ a flavin adenine dinucleotide (FAD) cofactor and an antenna chromophore to mediate light-dependent DNA repair [11]. Structurally, CRYs retain a photolyase homology region (PHR), which consists of an N-terminal α/β domain and a C-terminal α-helical domain connected through a flexible linker [10]. During evolution, CPF members diversified into multiple CRY subtypes that are differentiated by their ability to bind FAD and through species and isoform specific cryptochrome C-terminal extensions (CCEs) [9]. These CCE domains vary between 50 and 250 amino acids, providing a diverse and complex conformational landscape, which has been implicated in protein–protein interactions, signal transduction, regulation of photochemical activity, and in mediating isoform selective drug design in non-photoactive CRYs [9,12,13].

It is generally accepted that CCE domains lack defined secondary structure and play an essential role in modulating interactions with the PHR through flexible loop regions termed the protrusion motif (PM) and phosphate binding loop (PBL) that cradle either FAD (Type I and IV) [10] or small molecule regulators as treatments for diabetes and cancer (Type II) [14]. Dynamic regulation of the flexible loop regions is coupled to CCE dynamics, which facilitates the regulation of protein–protein interactions involved in controlling circadian rhythms (Type I and II), magnetoreception (Type IV), and many others, highlighting their functional significance in CRY-mediated processes [9]. Despite their importance, the dynamic nature of these flexible loop regions and CCE makes their structural characterization challenging through experimental techniques [15]. Currently, only Type I *Drosophila* CRY (*d*CRY) has a full-length structure [16], where structures of Type II and IV CRYs are limited to the isolated PHR domains that often lack defined structures of the PM or PBL [17], thereby limiting our ability to decipher how changes in amino acid sequence dictates the conformational landscape of the CCE and flexible loop regions.

Recent drug discovery efforts targeting mammalian Type II CRYs [18–20] have provided new insights into this problem. Biophysical studies suggest that the conformational dynamics of apo-CRYs, particularly in the loop regions flanking the primary pocket, are not static but vary significantly between isoforms, even in the absence of cofactors or ligands [21,22]. These findings imply that the intrinsic sequence of CRY proteins can encode distinct dynamic landscapes that shape their interaction profiles and signaling outputs. However, such comparative apo-state studies across the broader CPF remain sparse. As a result, it is essential to develop improved computational approaches capable of predicting both the structure and dynamics of CCE regions, as well as their coupling to apo (Type II) and FAD-binding (Type I and IV) CRYs. In the present work, we propose to employ photoactive Type I (*Drosophila* CRY: *d*CRY) [23] and Type IV (*Columba livia* CRY4: *Cl*CRY4) [15] as model systems to examine whether the knowledge of *d*CRY can be leveraged to predict conformational dynamics in other CRYs and whether these predictions can be extended to apo-CRYs as models of Type II CRY regulation.

Notably, the CCE domains of *Cl*CRY4 and *d*CRY share similarities in length and structural responses to blue-light activation [9]. In *Cl*CRY4, the CCE undergoes structural rearrangements that may play a role in magnetoreception, whereas in *d*CRY, the CCE is released from the primary pocket upon FAD photoreduction to the flavin reduced anionic semiquinone state [9]. Given the importance of CCE dynamics in cryptochrome function, extensive sampling of *Cl*CRY4, including its CCE domain, is crucial for understanding its photoreception mechanism. The evolutionary correlation and structural similarity between *Cl*CRY4 and *d*CRY suggest that the modeling of *d*CRY including its CCE domain could serve as benchmark and guidance for the simulation of *Cl*CRY4 including its intrinsically disordered CCE region. Building on such structurally related and experimentally resolved full-length proteins, novel simulation strategies can be developed to better capture the behavior of highly dynamic and disordered regions.

Dimensionality reduction (DR) techniques have been increasingly employed to analyze molecular dynamics (MD) simulations of proteins [24–26]. Among these methods, principal component analysis (PCA) is a widely used linear DR method that reduces high-dimensional data while preserving the most significant variance in structural fluctuations [27]. In addition, time-lagged independent component analysis (t-ICA) is another linear DR method that treats protein simulation results as time-series data, effectively identifying slow conformational changes in protein dynamics [28]. To better capture the non-linearity inherent in protein conformational landscapes, non-linear DR techniques have been developed. Methods such as uniform manifold approximation and projection (UMAP) [25,29,30] and t-distributed stochastic neighbor embedding (t-SNE) [31–33] construct low-dimensional latent spaces that approximate the probability distributions of high-dimensional manifolds, providing more accurate representations of complex conformational transitions. Additionally, deep learning-based approaches, such as autoencoders [34,35], have proven particularly effective in the analysis of protein simulations.

In this study, we hypothesize that high-quality, crystal structure-based simulations of *d*CRY (Figure 1a), including its CCE domain, can serve as a reference for guiding and evaluating

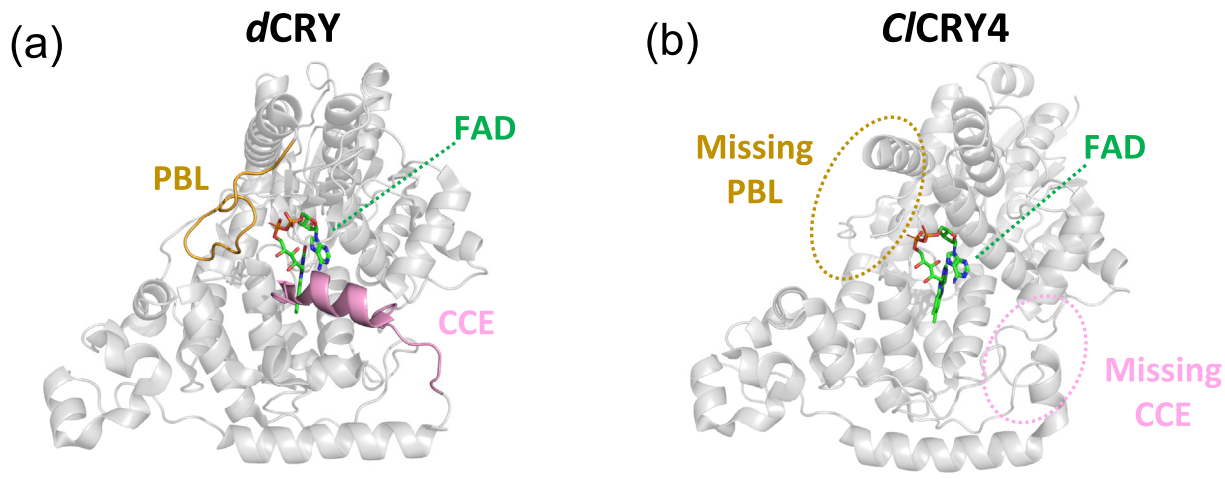

Figure 1. The structures of (a) *d*CRY (PDB ID: 4GU5), and (b) *Cl*CRY4 (PDB ID: 6PU0). The photolyase homology region (PHR) is shown in gray, the cryptochrome C-terminal extension (CCE) in pink, the phosphate binding loop (PBL) in brown, and the FAD in green.

simulations of the complete *Cl*CRY4 structure, including its CCE domain, which is currently unavailable. To test this approach, we employed both conventional MD simulations at ambient temperature and temperature replica exchange molecular dynamics (T-REMD) simulations [36] across a broad temperature range for *d*CRY and *Cl*CRY4. T-REMD simulations are particularly advantageous for overcoming local energy minima and exploring a wide conformational space in an unbiased manner. The conventional MD simulations of *d*CRY serve as a benchmark for T-REMD simulations of *d*CRY and as a reference template to filter T-REMD simulations of *Cl*CRY4 for further analysis. Using this novel computational protocol, we explored the conformational and dynamical landscapes of the complete *d*CRY and *Cl*CRY4 structures, including their CCE domains as IDRs, providing insights into the regulatory mechanisms of CCE domain flexibility in both CRYs.

Although FAD is essential for the functions of *d*CRY and *Cl*CRY4, herein we examine the apo states of these CRYs (without FAD) as a proof of principle on whether the structure of a full-length protein containing a defined conformation of an IDR can be used to inform studies of closely related proteins. Leveraging the apo-proteins, we provide an approach to interrogate whether there exist intrinsic differences in PHR dynamics that couple to the CCE through flexible loop regions near the empty primary pocket as has been observed in Type II mammalian CRYs [13,21,37]. Nevertheless, the redox state of FAD can significantly influence the conformational distribution of the CCE domain. In *d*CRY, the CCE domain docks into the FAD binding cleft in the dark state and undocks upon photoreduction [38]. In contrast, limited cryptochrome assay suggests that the CCE region of *Cl*CRY4 becomes more stable upon light exposure [15]. This simplified apo-proteins system provides a platform to establish a benchmark for future studies incorporating FAD in various oxidation states. Therefore, we excluded the FAD from the simulations in this study and focused on the correlation between the PHR and CCE domains, independent of cofactor binding.

Computational Methods

Initial structures

The *d*CRY crystal structure (Figure 1a) was obtained from the Protein Data Bank (PDB ID: 4GU5) and used for conventional MD simulations. To enhance the sampling efficiency in T-REMD simulations, we also constructed an alternative *d*CRY structure—featuring the CCE domain positioned away from the FAD-binding pocket—using AlphaFold [39] (Figure 2a). Our aim was to test whether initiating T-REMD simulations with the CCE domain outside of the pocket would allow the system to sample conformations similar to those observed in the crystal structure. Such a result could provide validation for the comprehensive sampling capability of T-REMD.

To construct the complete structure of *Cl*CRY4, we followed the approach as below:

(1) Using the SWISS-MODEL [40] and full *Cl*CRY4 sequence as the input, we obtained the full *Cl*CRY4 structure with the missing PBL (residues: 228–244) and CCE (residues: 498–525).
(2) We superimposed the SWISS-MODEL predicted full *Cl*CRY4 structure with the *Cl*CRY4 crystal structure (PDB ID: 6PU0) (Figure 1b) and attached the predicted PBL and CCE regions to the crystal structure.

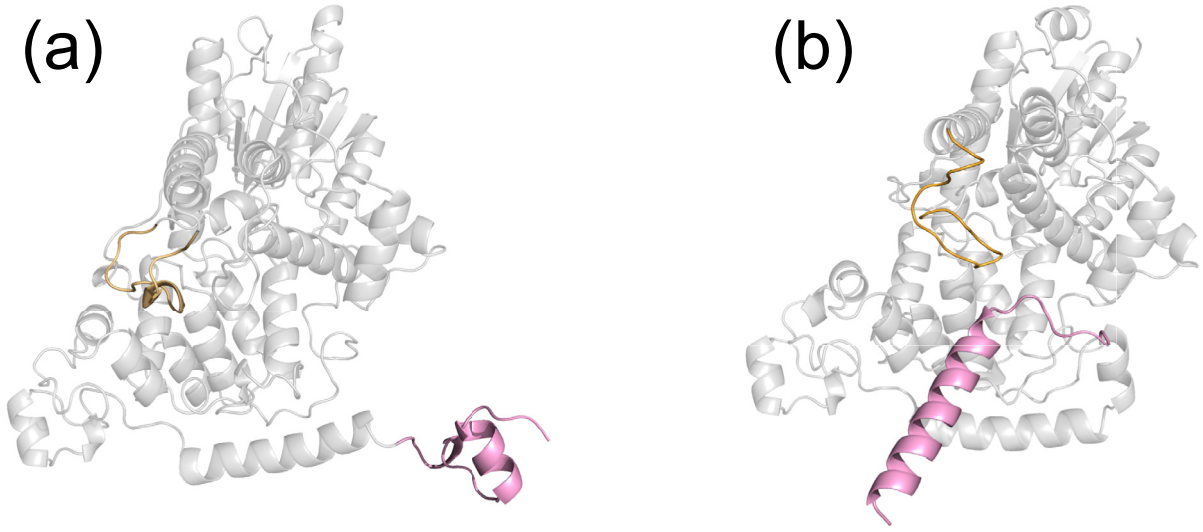

Figure 2. Initial structures used for the T-REMD simulations: (a) *d*CRY and (b) *Cl*CRY4. PBL and CCE domains are shown in brown and pink, respectively.

(3) To ensure structural plausibility and eliminate the steric clashes, sidechain conformations in the combined model were optimized using SCWRL4.0 [41]

SWISS-MODEL was selected for *Cl*CRY4 structure construction because it accurately reproduced the CCE domain in the *d*CRY crystal structure (PDB ID: 4GU5), which served as a benchmark, with the CCE domain correctly docked in the FAD pocket (Figure S5). In contrast, when using AlphaFold for prediction, the CCE domain was placed outside the pocket. The full-length amino acid sequence for *Cl*CRY4, including the previously unresolved PBL and CCE regions, was obtained from UniProt (ID: A0A2I0LZR8) [42]. It should be noted that although the *d*CRY structure generated by AlphaFold has a CCE domain conformation differing from that observed in the crystal structure, it is nevertheless ideally suited as the initial structure for T-REMD simulations. This approach can properly equilibrate the system, mitigating potential inaccuracies introduced by theoretical modeling while also benefiting from an alternative representation of the CCE domain as an intrinsically disordered region (IDR).

To reliably model the impact of FAD on the functions of cryptochrome, high quality force field parameters for FAD in different redox states are necessary but beyond the scope of the current study.

To minimize the impact of FAD force field parameter quality and to focus on the simulation protocol being tested with *d*CRY and *Cl*CRY4, FAD was excluded from both proteins in the simulations conducted in this study.

dCRY molecular dynamics simulations

(a) Conventional MD simulations

The *d*CRY crystal structure (PDB ID: 4GU5) [43] was used to set up conventional MD simulations. Explicit water molecules were employed to solvate the system. AMBER FF14SB force field [44] and TIP3P water model [45] were employed for the simulations. Covalent bonds involving hydrogen atoms were constrained using the SHAKE algorithm [46]. A rectangular simulation box was constructed by adding a minimum buffer of 15 Å between the cryptochrome proteins and the box boundaries in all directions, resulting in final box dimensions of 112.2 Å × 105.6 Å × 95.9 Å. 12 Na^+ ions were added to the simulation box to neutralize the net charge of the system. Energy minimization was performed in two stages: the first 20 steps used the steepest descent method, followed by the conjugate gradient method for the remaining steps, totaling 1000 minimization steps. During the heating phase, the system was gradually heated from 100 K to 300 K through 1 ns (ns) MD simulation, followed by a 100 ns equilibration simulation and a final 1 microsecond (μs) production simulation, both in the canonical (NVT) ensemble. A time step of 2 femtoseconds (fs) was used in all simulations in this study.

(b) Temperature replica exchange molecular dynamics (T-REMD) simulations

T-REMD simulations were performed to further explore the conformational space of *d*CRY. The force field parameters and SHAKE algorithm are the same to those described above for the

conventional MD simulations of *d*CRY. The starting *d*CRY structure for T-REMD simulations (Figure 2a) was immersed into a rectangular simulation box measuring 112.2 Å × 123.9 Å × 97.2 Å. It then underwent energy minimization and heating following the procedures previously described for conventional MD simulations. The same number of Na^+ ions as in the conventional MD simulation were added to achieve charge neutrality. To achieve sufficient sampling, 32 replicas were generated with simulation temperatures spanning 300 K to 500 K (see Supplementary Table S1 for specific values). Following an initial 10 ns equilibration, each replica was subjected to 40 ns of production simulation.

The exchange probability between two replicas at adjacent temperatures, $P_{exchange}$, was calculated using the following equation [47]:

$$P_{exchange} = \exp\left(-(E_j - E_i)\left(\frac{1}{k_B T_i} - \frac{1}{k_B T_j}\right)\right), \quad (1)$$

where E_j and E_i represent the total energy of the system in replicas j and i, respectively. T_j and T_i are the corresponding temperatures, and k_B is the Boltzmann constant.

*Cl*CRY4 molecular dynamics simulations

(a) T-REMD simulations

The T-REMD simulations of *Cl*CRY4 were conducted following the same protocol described for *d*CRY. The starting conformation (Figure 2b) was placed in a simulation box measuring 104.7 Å × 93.9 Å × 128.1 Å, with one Na^+ ion added to neutralize the system.

(b) Conventional MD simulations

A total of 15 conventional MD simulations were performed for *Cl*CRY4, using 15 representative structures shown in Figure 5 (See Results for details). These structures were selected as the frames closest to the average RMSD values within each of the 15 clusters identified through clustering analysis. Each simulation was conducted for 100 ns production run, resulting a cumulative simulation time of 1.5 μs.

Autoencoder

An autoencoder neural network architecture [48,49] was employed as a dimensionality reduction method for analyzing cryptochrome simulation data. The autoencoder is composed of two sequentially connected components: an encoder and a decoder (Figure 3). The encoder compresses high-dimensional input data ($x \in \mathbb{R}^d$) into a two-dimensional (2D) latent representation ($x \in \mathbb{R}^2$), while the decoder reconstructs data in high-dimensional space from the latent space. In this study, unless otherwise specified, the structural coordinates from the aligned PHR region across *d*CRY T-REMD, *d*CRY conventional MD, and *Cl*CRY4 T-REMD simulations were used to train the autoencoder model. The encoder is a fully connected feedforward network with two hidden layers containing 512 and 256 nodes, respectively. The decoder mirrors this architecture. The loss function ($\mathscr{L}$) was computed as:

$$\mathscr{L} = \mathscr{L}_{recon} + \lambda \mathscr{L}_{L2} \quad (2)$$

where $\mathscr{L}_{recon}$ and $\mathscr{L}_{L2}$ represent the reconstruction loss and L_2 regularization term, respectively. λ is a weight factor and was set to 0.001 in our case. $\mathscr{L}_{recon}$ and $\mathscr{L}_{L2}$ are computed through the following equations, respectively:

$$\mathscr{L}_{recon} = \frac{1}{d * N} \sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2 \quad (3)$$

$$\mathscr{L}_{L2} = \frac{1}{d * N} \sum_{i=1}^{N} \|z_i\|^2 \quad (4)$$

where N denotes the data size of mini-batch for training, and x_i, $\hat{x}_i$ refers to the original and

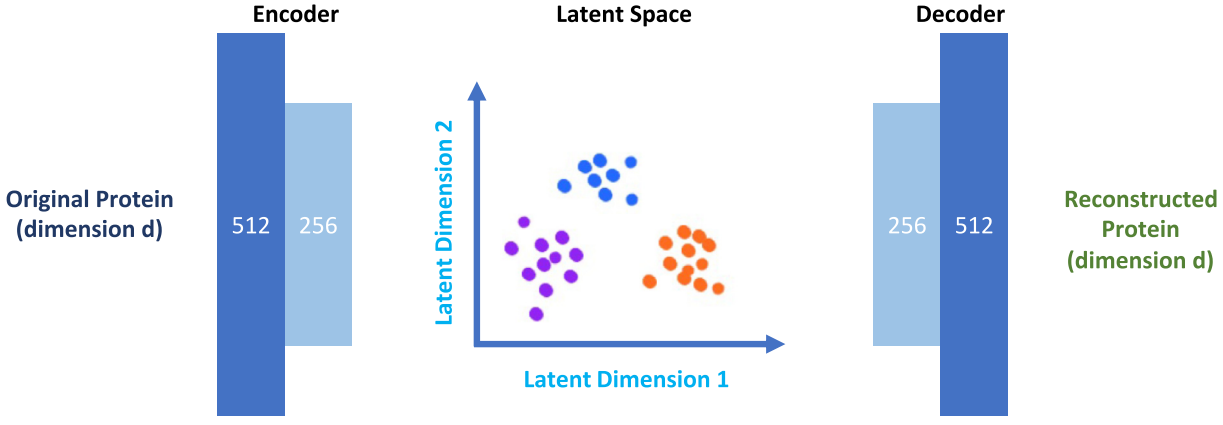

Figure 3. Architecture of Autoencoder.

reconstructed data, respectively. The term z_i represents the data in low dimensional latent space.

The autoencoder was implemented using CUDA-supported PyTorch 2.2.2 [50], enabling GPU-accelerated deep learning computations. The dataset was split into training and evaluation sets, with 80% allocated for training and the remaining 20% for evaluation. A mini-batch is a subset of the entire training dataset used during one iteration of the gradient descent optimization, allowing efficient computation of gradients of autoencoder model parameters. During training, a mini-batch size of 200 frames was used. The learning rate was set to 10^{-4}. Model parameters were optimized using the Adam optimizer. For activation functions, the Rectified Linear Unit (ReLU) was applied to the hidden layers of both encoder and decoder, while a Sigmoid activation function was used in the decoder's output layer to constrain the reconstructed values within the range (0, 1).

Time-lagged independent component analysis (t-ICA)

The t-ICA method [28,51,52] was employed as a dimensionality reduction technique to transform and project protein simulation data onto a 2D latent space. Mathematically, t-ICA solves the generalized eigenvalue problem:

$$C_\tau r_i = C_0 \lambda_i r_i \tag{5}$$

where r_i represents the independent components, and λ_i denotes the corresponding normalized autocorrelations. C_0 and C_τ correspond to the mean-free covariance matrix and the time-lagged covariance matrix, respectively. The covariance matrixes are computed as:

$$C_0 = (X_t - \mu)^T (X_t - \mu) \tag{6}$$

$$C_\tau = (X_t - \mu)^T (X_{t+\tau} - \mu) \tag{7}$$

where X_t represents the sequence of time-dependent multivariate protein coordinate data, and $X_{t+\tau}$ denotes the protein trajectory data at a time lag τ. The term μ corresponds to the mean value of the entire dataset. In this study, the t-ICA method was implemented using the PyEMMA 2 package [53]

Clustering analysis

Protein simulation data were analyzed using the MDSCAN package [54], which provides high memory efficiency for large-scale clustering based on root-mean-square deviation (RMSD). MDSCAN encodes MD simulation trajectories into a vantage-point tree (vp-tree), which accelerates the *k*-nearest neighbor search process by reducing the computational cost of RMSD or Euclidean distance calculations. The high efficiency is achieved through triangle inequality for efficient pruning. Additionally, a quasi-minimum spanning tree (quasi-MST) is constructed using a dual-heap strategy, in which a primary heap prioritizes nodes with high core distances, while an auxiliary heap handles nodes that fail to satisfy the following equation:

$$d_{mr}(i,j) = \{\kappa(i) | \forall j \in \eta(i) : \kappa(j) \leq \kappa(i)\} \tag{8}$$

Term $d_{mr}(i,j)$ represents the mutual reachability distance,$\kappa(i)$ is the core distance from node i to its kth nearest neighbor. Term $\eta(i)$ is the set of nodes j satisfying $d(i,j) \leq \kappa(i)$. Here, $d(i,j)$ denotes the input RMSD or Euclidean distance. Overall, MDSCAN leverages the vp-tree structure and a dual-heap approach to construct the quasi-MST, enabling efficient clustering of large MD trajectories with minimal memory usage while preserving accuracy.

Root-mean-square deviation (RMSD)

The RMSD was calculated to quantify the conformational differences between two structures using the following equation [55]:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\boldsymbol{x}_i - \boldsymbol{x}_i^{ref}\right)^2} \tag{9}$$

where $\boldsymbol{x}_i$, $\boldsymbol{x}_i^{ref}$ represent the coordinates of the ith atom in the given structure $\boldsymbol{x}$ and reference structure $\boldsymbol{x}^{ref}$, respectively. Prior to RMSD calculation, structural superimposition of $\boldsymbol{x}$ onto $\boldsymbol{x}^{ref}$ was performed. n denotes the total number of atoms considered. Unless otherwise specified, alpha carbon atoms (C_α) along the protein backbone were used to calculate the RMSD in this study.

Root-mean-square fluctuation (RMSF)

The RMSF was calculated to assess the flexibility of each residue during the simulation [56]. The fluctuation of residue i, denoted as $RMSF_i$, is given by:

$$RMSF_i = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \left(\boldsymbol{x}_i(t) - \boldsymbol{x}_i^{average}\right)^2} \tag{10}$$

where n represents the total number of frames, and $\boldsymbol{x}_i(t)$ is the coordinate of the alpha carbon, C_α in residue i at time t, and $\boldsymbol{x}_i^{average}$ denotes the average C_α coordinate of residue i over all n frames.

Dynamic cross correlation matrix (DCCM)

To quantify dynamic couplings and collective motions in protein MD simulations, we calculated the DCCM [57] using the following equation:

$$C_{ij} = \frac{\langle \Delta \boldsymbol{r}_i \cdot \Delta \boldsymbol{r}_j \rangle}{\sqrt{\langle \Delta \boldsymbol{r}_i^2 \rangle \langle \Delta \boldsymbol{r}_j^2 \rangle}} \tag{11}$$

where C_{ij} represents the correlation coefficient between the motions of residues i and j. The terms $\Delta \boldsymbol{r}_i$ and $\Delta \boldsymbol{r}_j$ represent the displacements of the corresponding C_α atoms from their respective

average positions throughout the trajectory. The notation $\langle ... \rangle$ indicates a time average over the entire trajectory.

Results and Discussion

T-REMD simulations

In the T-REMD simulations, 40,000 exchange attempts were performed, with 500 simulation steps between each attempt. Frames were saved every 1,000 steps. The average exchange rate was 20.4% for *d*CRY and 20.9% for *Cl*CRY4. To ensure sufficient sampling from the T-REMD calculations, each of the 32 replicas underwent a 40 ns production MD simulation, resulting in a total of 1.28 μs of sampling (640,000 snapshots) for both *d*CRY and *Cl*CRY4.

Latent space from dimensionality reduction analysis

Due to its superior performance in analyzing complex high-dimensional data [34,58,59], the autoencoder was selected as the primary dimensionality reduction method in this study. For the convenience of comparison between *d*CRY and *Cl*CRY4, a set of residues within the PHR domains, aligned with both *d*CRY and *Cl*CRY4, were selected for analysis and referred to as the common core region. We combined the simulation data of both *d*CRY and *Cl*CRY4 for dimensionality reduction analysis using the coordinates of the common core region. The two main components from the autoencoder analysis were used to construct a 2D latent space, onto which the distribution of the production T-REMD simulations of both *d*CRY and *Cl*CRY4, as well as conventional MD simulations, were plotted (Figure 4). The t-ICA was applied in a similar manner for comparison.

*d*CRY conventional MD simulation

The *d*CRY MD simulations remain stable, with RMSD values consistently below 1.8 Å throughout the entire trajectory (Figure S1). The distribution of conventional MD simulations of *d*CRY is plotted in 2D latent spaces of both the autoencoder and t-ICA analyses (green regions in Figure 4a and b). In the autoencoder latent space (Figure 4a), the *d*CRY conventional MD simulations are entirely contained within the distribution of *d*CRY T-REMD simulations. However, in the t-ICA latent space (Figure 4b), the conventional MD simulations are positioned at the periphery of the *d*CRY T-REMD distribution. Consistent with its low RMSD values, the distribution of *d*CRY conventional MD simulations is narrow, indicating a limited sampling range within its conformational space.

*d*CRY T-REMD simulations

For comparison, the RMSD of *d*CRY T-REMD simulations was calculated using the same reference structure as in the RMSD analysis of the *d*CRY conventional MD simulations. Most *d*CRY T-REMD simulations maintained RMSD values below 7 Å, while some replicas remained under 7 Å for the majority of the simulation, with occasional fluctuations exceeding 8 Å (Figure S2) (Replica 4 in Figure S2a as red plot and Replica 5 in Figure S2b as blue plot). The consistency of RMSD values across the T-REMD simulations strongly supports the reliability of the *d*CRY sampling. The distribution of *d*CRY T-REMD simulations is plotted in the latent spaces derived

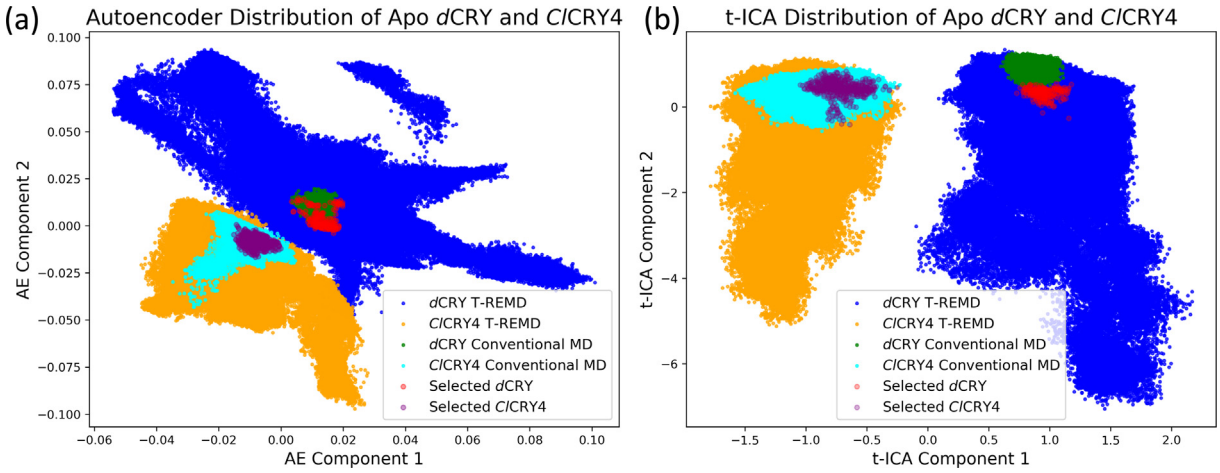

Figure 4. Distribution of *d*CRY and *Cl*CRY4 simulations in the latent spaces using (a) Autoencoder (AE) and (b) time-lagged independent component analysis (t-ICA). The latent spaces for both methods were constructed using *d*CRY T-REMD (blue), *Cl*CRY4 T-REMD (brown), and *d*CRY conventional MD (green). Selected *d*CRY (red) and *Cl*CRY4 (purple) T-REMD conformations are highlighted. The distribution of *Cl*CRY4 conventional MD simulation is plotted in cyan.

from both autoencoder and t-ICA analyses (blue regions in Figure 4a and b). As expected, the T-REMD simulations exhibit a significantly broader distribution comparing to the conventional MD simulations, reflecting enhanced conformational sampling. Moreover, the complete encapsulation of the *d*CRY conventional MD simulation within the T-REMD distribution further supports the reliability of T-REMD in accurately capturing the conformational landscape of cryptochromes.

The *d*CRY T-REMD simulations were further evaluated using *d*CRY conventional MD simulations by calculating the pairwise RMSD based on the complete *d*CRY structure. For each of the 50,000 snapshots from conventional MD simulations, the *d*CRY T-REMD snapshot with the lowest RMSD was selected (the algorithm for selecting the frames were shown in Table S3 in SI). A total of 227 snapshots were selected from T-REMD simulations through this process. The distribution of the 227 snapshots selected from the *d*CRY T-REMD simulations is plotted in both latent spaces (red regions in Figure 4a and b). The selected snapshots possess the conformation similar to the *d*CRY crystal structure, validating the thoroughness of the *d*CRY T-REMD simulations initiated from the structure with CCE positioned outside the FAD binding pocket.

*Cl*CRY4 T-REMD simulations

The RMSD values of *Cl*CRY4 T-REMD simulations are also consistent among all replicas. Most *Cl*CRY4 T-REMD simulations (26 replicas) have RMSD values remain under 5 Å. Among the remaining six replicas, three replicas have RMSD values fluctuating around 5 Å and remain under 6 Å during the entire simulation (Replica 5 in Figure S3b as blue plot, Replica 20 in Figure S3e as red plot, and Replica 27 in Figure S3g as green plot), two replicas have RMSD fluctuating up to 6 Å (Replica 15 in Figure S3d as green plot, and Replica 32 in Figure S3h as red plot). Only one replica has small portion of the simulation with RMSD values exceeding 7 Å (Replica 23 in Figure S3f as green plot).

The distribution of *Cl*CRY4 T-REMD simulations (brown area in Figure 4a) is positioned adjacent to that of *d*CRY T-REMD simulations in the autoencoder latent space, based on the common core region. Interestingly, the *Cl*CRY4 T-REMD distribution appears more compact than the *d*CRY T-REMD distribution (blue), despite *Cl*CRY4 possessing a larger CCE domain and potentially greater flexibility. This observation may seem counterintuitive but could be attributed to specific structural constraints or stabilization effects in *Cl*CRY4. A similar trend is observed in the t-ICA latent space, where the *Cl*CRY4 T-REMD distribution is distinct from that of *d*CRY T-REMD, with a noticeable gap between the two (Figure 4b). In both latent spaces, the distribution of *Cl*CRY4 T-REMD simulations is comparable in contrast to that of *d*CRY T-REMD simulations. This comparison further supports the reliability of *Cl*CRY4 T-REMD simulations for subsequent sampling and analysis.

*Cl*CRY4 conventional MD simulations

The *d*CRY conventional MD simulations were used as reference to select *Cl*CRY4 T-REMD snapshots for further analysis. RMSD values were calculated using the common core region shared between *d*CRY and *Cl*CRY4 to guide the selection process. Specifically, for each of the 50,000 snapshots from *d*CRY conventional MD simulations, the *Cl*CRY4 T-REMD structure with the lowest RMSD, based on the common core region, was chosen. This selection strategy aligns with the central hypothesis of this study—that high-quality, crystal structure-based simulations of *d*CRY can serve as a reliable reference for guiding and evaluating simulations of the complete *Cl*CRY4 structure. A total of 529 snapshots were selected from the *Cl*CRY4 T-REMD simulations with the smallest RMSD values relative to the 50,000 *d*CRY conventional MD simulation frames via the algorithm listed in Table S3. The distribution of these selected *Cl*CRY4 T-REMD snapshots (purple regions in Figure 4) appears largely symmetrical in relation to the distribution of the selected *d*CRY T-REMD snapshots (red regions in Figure 4), further supporting the validity of this comparative approach.

To further investigate the structural features of the selected conformations, we performed clustering analysis on the 529 chosen structures using MDSCAN. A total of 15 clusters were identified (see MDSCAN Clustering Analysis of *Cl*CRY4 T-REMD Simulations section in Supplementary Materials for details). For each cluster, the structure with the lowest RMSD relative to the cluster's average conformation was selected as the representative structure (Figure 5). The clusters were ranked in descending order based on the number of snapshots assigned to each cluster. The population size of each cluster is provided in Table S2 in Supplementary Materials. Notably, the PBL and CCE regions exhibited the highest flexibility among the selected structures, highlighting their dynamic nature.

To investigate the conformational space and dynamics of *Cl*CRY4 at ambient temperature, we conducted 100 ns conventional MD simulations at 300 K for each representative structure from the 15 identified clusters. The RMSD values of these 15 *Cl*CRY4 conventional MD simulations (Figure S4) exhibit greater consistency than those from *Cl*CRY4 T-REMD simulations. Among these simulations, 13 maintained RMSD values below 4 Å throughout their entire trajectories. The simulation based on Cluster 8 (Figure S4c, brown plot) briefly exceeded 4 Å, but the RMSD

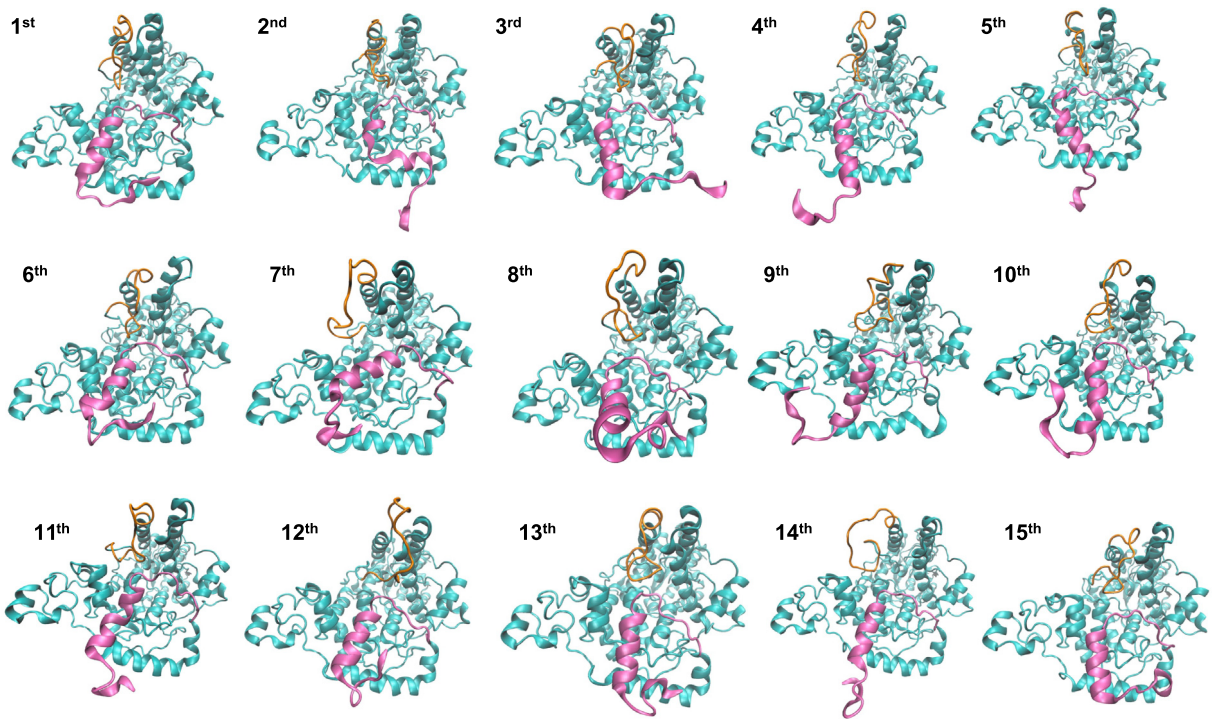

Figure 5. Representative structure of each cluster from the clustering analysis of the selected 529 *Cl*CRY4 T-REMD snapshots. The cluster population decreases as the cluster index increases. The PHR is shown in brown and the CCE domain is shown in pink.

remained below 3.5 Å for most of the simulation. The only exception was the simulation based on Cluster 7 (Figure S4c, blue plot), where RMSD values fluctuated around 4 Å during the first 60% of the trajectory before increasing above 4.5 Å, though remaining below 5 Å for the entire duration. This high level of consistency, combined with an appropriable range of fluctuation, strongly supports the reliability and sampling efficiency of the *Cl*CRY4 conventional MD simulations. These simulations provide valuable insights into the dynamic properties of *Cl*CRY4, including its CCE domain.

The distribution of the 1.5 μs *Cl*CRY4 conventional simulations is plotted in both autoencoder and t-ICA latent spaces (cyan regions in Figure 4), alongside *d*CRY T-REMD and conventional MD simulations, and *Cl*CRY4 T-REMD simulations. Notably, the distribution of the 529 selected snapshots from *Cl*CRY4 T-REMD simulations (purple regions in Figure 4) fully overlaps with that of the *Cl*CRY4 conventional MD simulations. Significantly, the distribution of *Cl*CRY4 conventional MD simulations entirely overlaps with the broader distribution of *Cl*CRY4 T-REMD simulations (brown regions in Figure 4). This hierarchical inclusiveness provides strong validation for both the T-REMD and conventional MD simulations of *Cl*CRY4. The extensive coverage of *Cl*CRY4 conventional MD simulations within the latent spaces suggests that these simulations further explore the conformational landscape of *Cl*CRY4 at a biologically relevant temperature, underscoring their efficiency in capturing protein functional dynamics.

To further compare *Cl*CRY4 T-REMD and conventional MD simulations, we performed dimensionality reduction analyses using both autoencoder and t-ICA methods, incorporating the full protein structure, including both the PHR and CCE domains of *Cl*CRY4 (Figure 6). The hierarchical inclusiveness relationships among the distributions of *Cl*CRY4 T-REMD simulations, conventional MD simulations, and the 529 selected snapshots are preserved in the latent spaces based on the complete *Cl*CRY4 structure. Consistent with the analyses based on the common core region, the *Cl*CRY4 T-REMD simulations display the broadest distribution (brown regions in Figure 6), indicating extensive conformational sampling. Interestingly, the distribution of the 529 selected snapshots from the *Cl*CRY4 T-REMD simulations is more dispersed than that observed in the common core region-based analyses, likely due to substantial structural variations in the CCE domain. This observation highlights the influence of the CCE IDR, which was not included in the common core region, underscoring its role in contributing to *Cl*CRY4 conformational diversity.

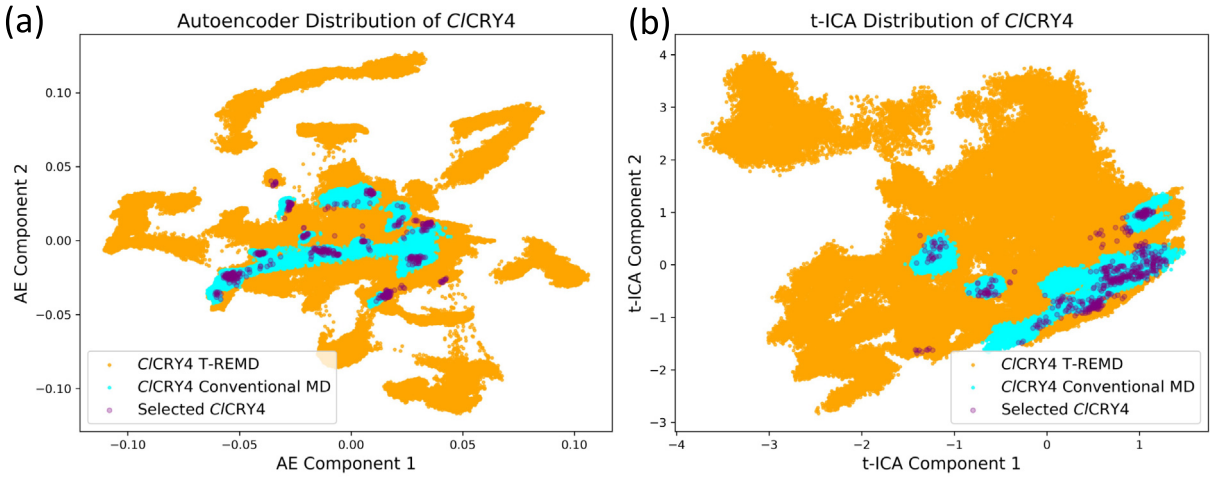

Figure 6. Distribution of *Cl*CRY4 simulations in the latent spaces using full *Cl*CRY4 structure including both PHR and CCE domains: (a) Latent space based on autoencoder (AE) analysis; (b) Latent space based on t-ICA. The complete *Cl*CRY4 structural coordinates from both T-REMD and conventional MD simulations were used for the autoencoder and t-ICA analyses.

Comparison of dCRY and ClCRY4 conventional simulations

The RMSF of each residue's alpha carbon, Cα, was calculated for both *d*CRY and *Cl*CRY4 based on their conventional MD simulations (Figure 7). For comparison, RMSF values were also computed based on the T-REMD simulations of both systems. In *d*CRY conventional MD simulations, the CCE domain and PBL exhibit minimal fluctuations (Figure 7a). In T-REMD simulations, the *d*CRY CCE domain displays significantly higher fluctuations, with peak values approaching 16 Å. Two regions within the *d*CRY PHR domain, including the PBL (residues 249–263) and the protrusion motif (residues 288–306), display increased flexibility, with peak RMSF values ranging between 7 and 8 Å. These differences may result from a combination of factors, including the limited sampling efficiency of *d*CRY conventional MD simulations and the enhanced conformational exploration facilitated by higher temperatures in T-REMD simulations.

The RMSF profiles of *Cl*CRY4 remain consistent between conventional MD simulations and T-REMD simulations (Figure 7b). This result is expected, as the *Cl*CRY4 conventional MD simulations were initiated from 15 representative structures selected from the T-REMD simulations. Considering that the distribution of *d*CRY conventional MD simulations is significantly narrower than that of the *d*CRY T-REMD simulations (Figure 7a), the close agreement in RMSF values further underscores the sampling efficiency and reliability of our protocol. This finding validates the approach of using T-REMD

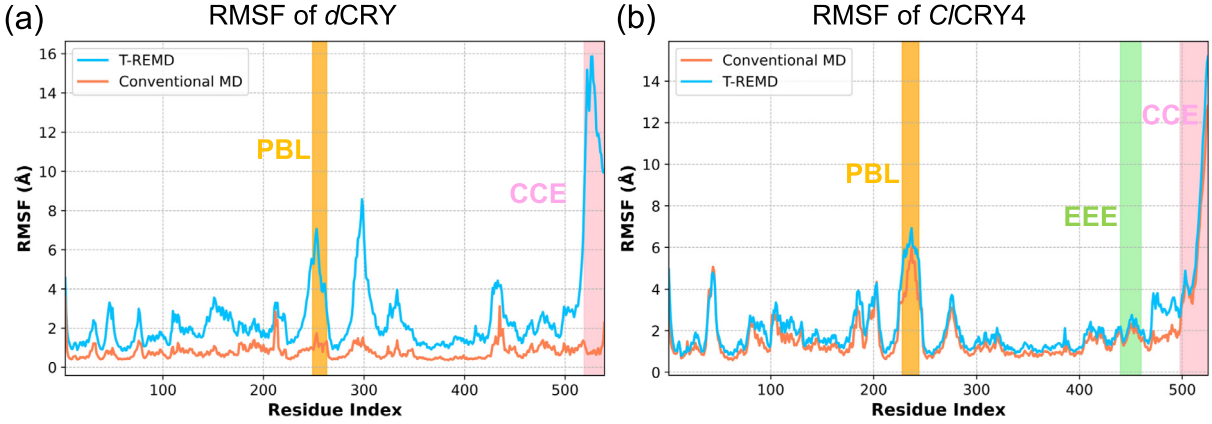

Figure 7. RMSF plots of conventional MD and T-REMD simulations. (a) *d*CRY; (b) *Cl*CRY4. PBL region is labeled in orange, EEE region is labeled in green, and CCE region is labeled in pink.

simulations to guide the selection of initial conformations for conventional MD simulations, ensuring robust and representative sampling of *Cl*CRY4 conformational dynamics.

The dynamic couplings and correlations among residues in *d*CRY and *Cl*CRY4 were assessed using the DCCM based on conventional MD simulations for each system. The residue correlation maps derived from DCCM are plotted for both *d*CRY and *Cl*CRY4 (Figure 8), in which blue indicates negative correlations between corresponding residues, and red indicates positive correlations. In *d*CRY, the correlation between the CCE and the PHR domains, particularly the PBL region in the PHR domain, is not pronounced. A closer examination reveals that interactions between the CCE domain and PBL in *d*CRY display a mix of positive and negative correlations (Figure 8a), suggesting weak and heterogeneous coupling between these regions. In contrast, *Cl*CRY4 displays a stronger correlation between its CCE and PHR domains. Notably, the first half of the *Cl*CRY4 CCE domain shows a predominantly positive correlation with the PBL, while the second half shows negative correlations (Figure 8b). Among the residues in the *Cl*CRY4 PBL domain, Arg236, Thr237, and Pro244 show the strongest positive correlation with the CCE, whereas Asn240, Ser241, and Leu242 show the strongest negative correlation.

Based on the DCCM, we calculated the average absolute correlation coefficients between the PBL and CCE domains, which are 0.15 and 0.06 for *Cl*CRY4 and *d*CRY, respectively, showing a stronger coupling between the two domains in *Cl*CRY4. To probe the origins of such correlation, we analyzed the distribution of the closest residue pair distance between the PBL and CCE domains from conventional MD simulations of both *Cl*CRY4 and *d*CRY (Figure S7). Overall, the distance distribution of *Cl*CRY4 is closer than that of *d*CRY. This observation aligns with the stronger dynamic correlation between the CCE and PBL domains in *Cl*CRY4 than in *d*CRY. These differences suggest distinct modes of intramolecular communication between the CCE and PHR domains in *d*CRY and *Cl*CRY4, potentially influencing their regulatory mechanisms.

To examine the conformational space of the CCE domain, dimensionality reduction analyses were performed separately for *d*CRY and *Cl*CRY4 using the full region, including PHR and CCE domains, from their respective conventional MD simulations. The distribution and population density for *d*CRY and *Cl*CRY4 are plotted in Figures 9 and 11, respectively. In the latent space generated from the autoencoder analysis, a single population density maximum is observed for *d*CRY (Figure 9b), whereas t-ICA analysis reveals three distinct maxima (Figure 9d). These local maxima in the population density distribution correspond to local minima in the free-energy landscape, which we refer to as attraction basins. These basins represent regions where the system has a non-trivial probability of remaining before transitioning to other states.

The representative *d*CRY structures corresponding to these attraction basins were identified separately within the t-ICA latent space and autoencoder latent space. Notably, these structures display minimal structural differences, including the CCE and PBL domains, as reflected by the low RMSD values between the structures derived from the t-ICA and autoencoder analyses (Table S5). Structural comparisons further confirm the high similarity among these conformations (Figure 10). Across all identified structures, the helical orientation of the CCE domain remains approximately perpendicular to the PBL backbone, indicating a conserved structural arrangement.

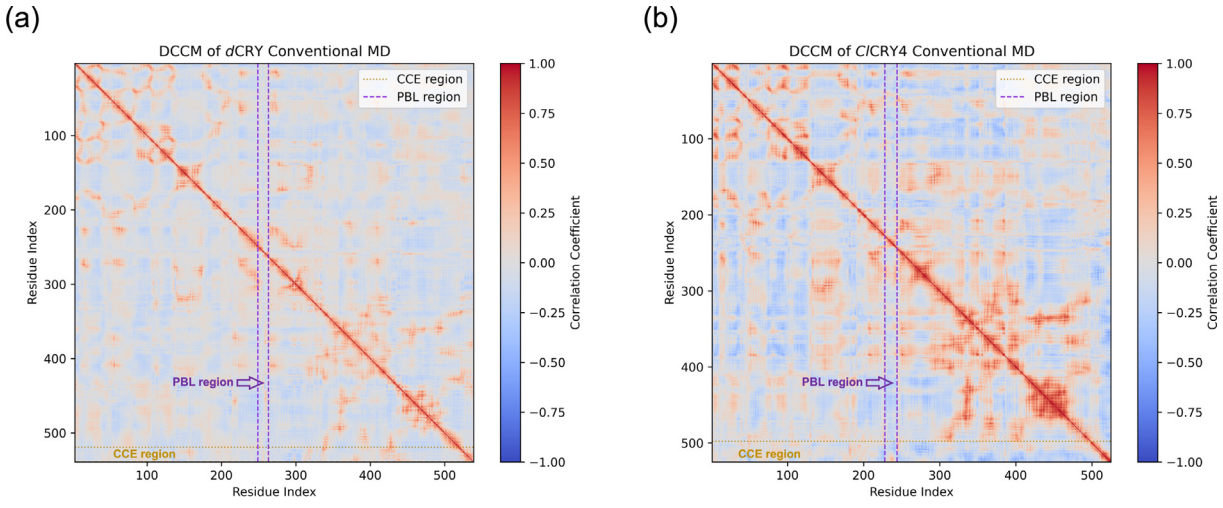

Figure 8. Dynamic cross correlation matrix (DCCM) plots of conventional MD simulations: (a) *d*CRY; (b) *Cl*CRY4.

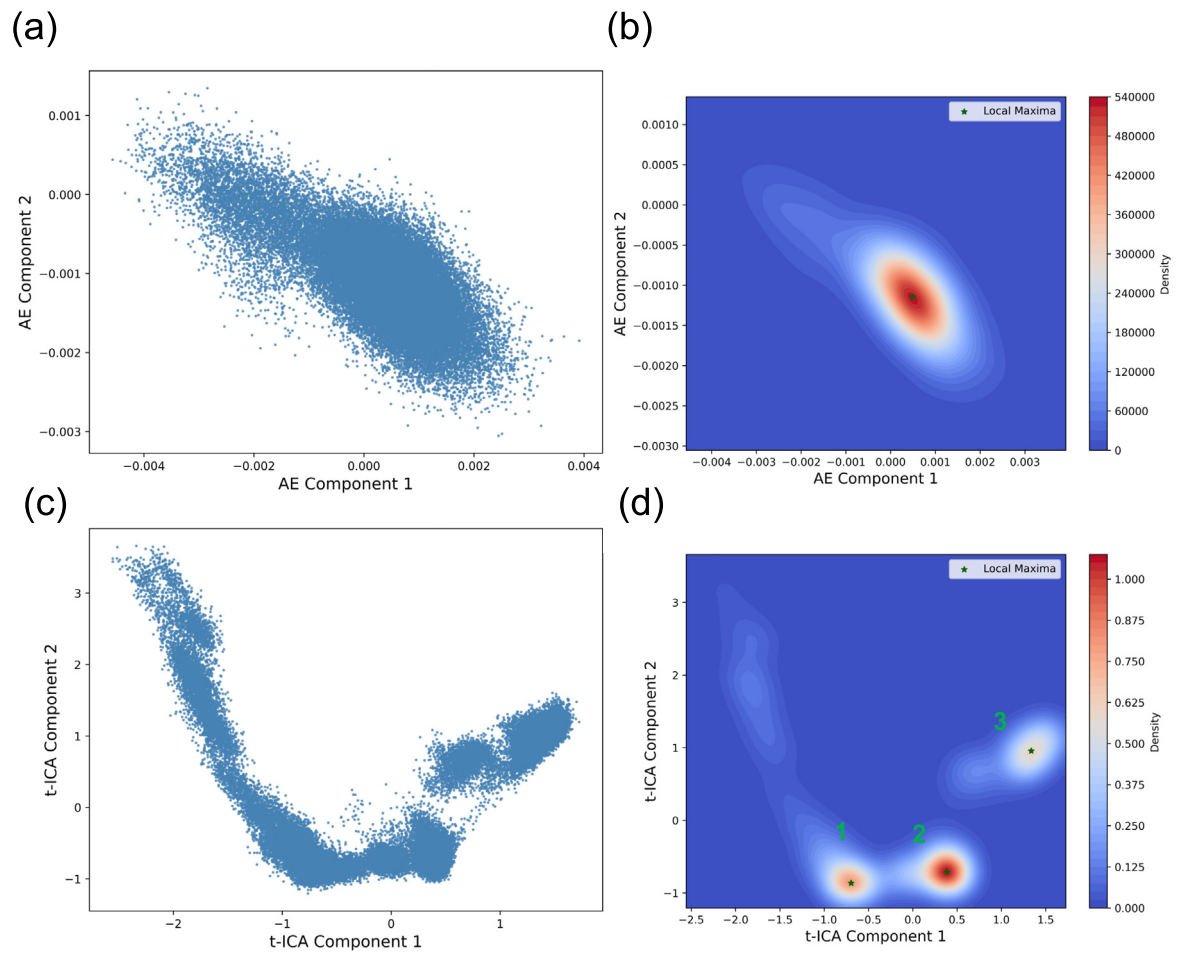

Figure 9. Dimensionality reduction analyses of *d*CRY conventional simulations. (a) Distribution and (b) population density on the autoencoder latent space. (c) Distribution and (d) population density on the t-ICA latent space. The local maxima are marked as green stars. The *d*CRY structural coordinates from the conventional MD simulations were used for the autoencoder and t-ICA analyses.

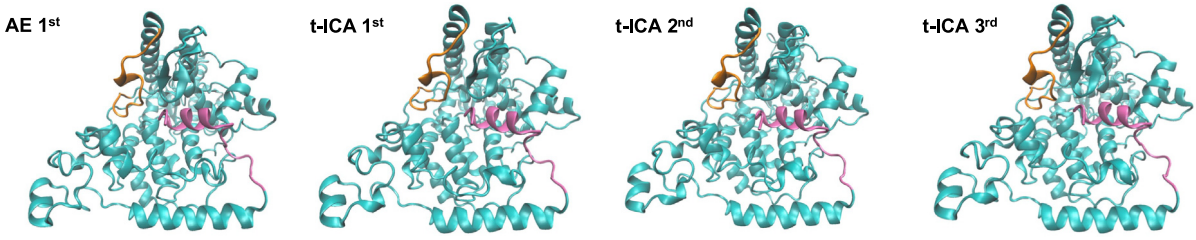

Figure 10. Representative structures of *d*CRY in the attraction basins in distribution density map based on autoencoder and t-ICA methods. The PBL is shown in brown and the CCE domain is shown in pink.

Around ten local attraction basins are identified in both autoencoder and t-ICA analyses for *Cl*CRY4. Because the attraction basins in the autoencoder latent space are more uniformly distributed comparing to those in the t-ICA latent space and considering that the autoencoder projection method displays a lower reconstruction error (Table S6), representative *Cl*CRY4 structures were selected based on the autoencoder latent space (Figure 12). The CCE domain conformation varies significantly among these structures. In five attraction basins (1, 2, 4, 5, and 7 in Figure 11b), the *Cl*CRY4 CCE domain adopts an orientation with its C-terminal pointing inward, and the helical structure of the CCE domain remains stable (structures 1, 2, 4, 5, and 7 in Figure 12). In contrast, in attraction basin 3, the CCE domain helix undergoes significant unfolding, although the

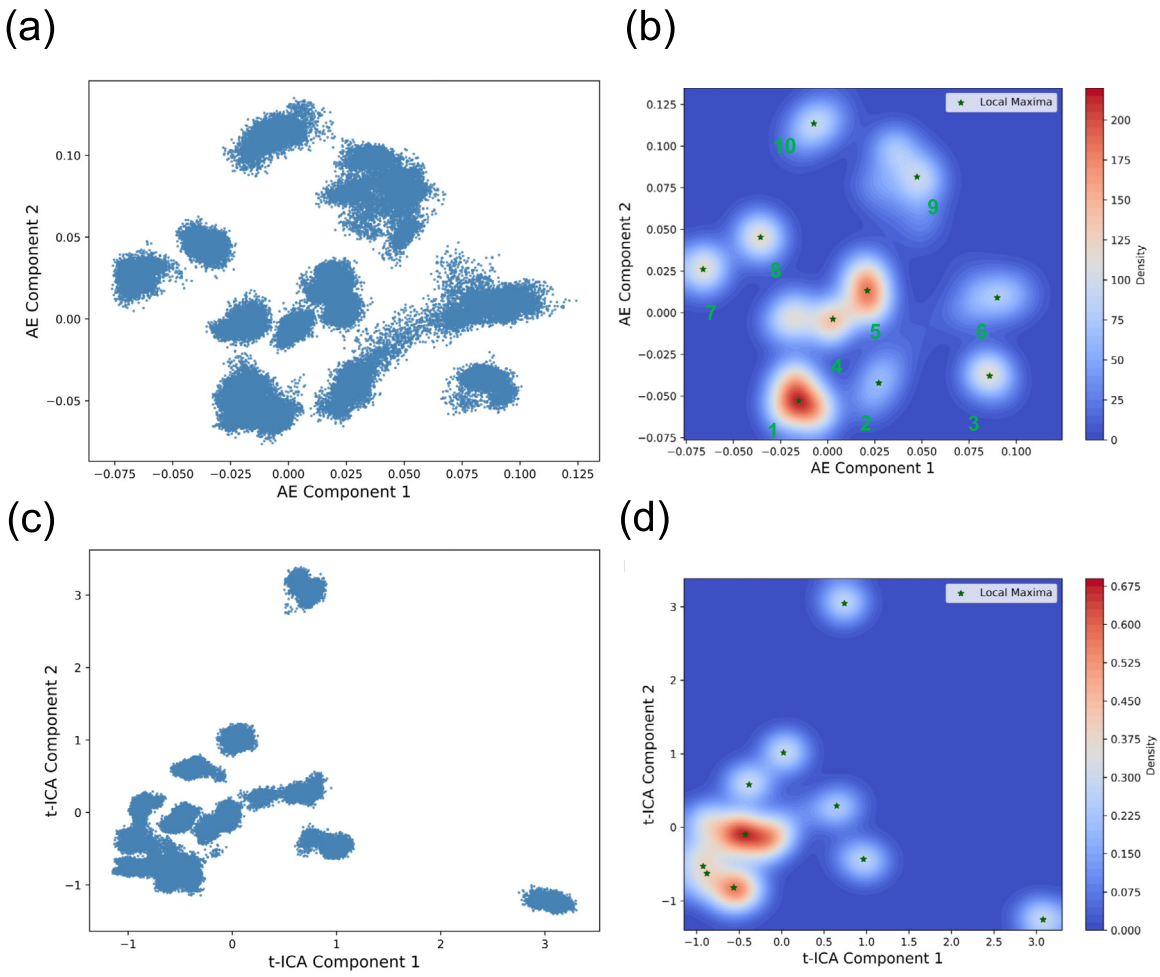

Figure 11. Dimensionality reduction analyses of *Cl*CRY4 conventional simulations. (a) Distribution and (b) population density on the autoencoder latent space. (c) Distribution and (d) population density on the t-ICA latent space. The local maxima are marked as green stars. The complete *Cl*CRY4 structural coordinates from conventional MD simulations were used for the autoencoder and t-ICA analyses.

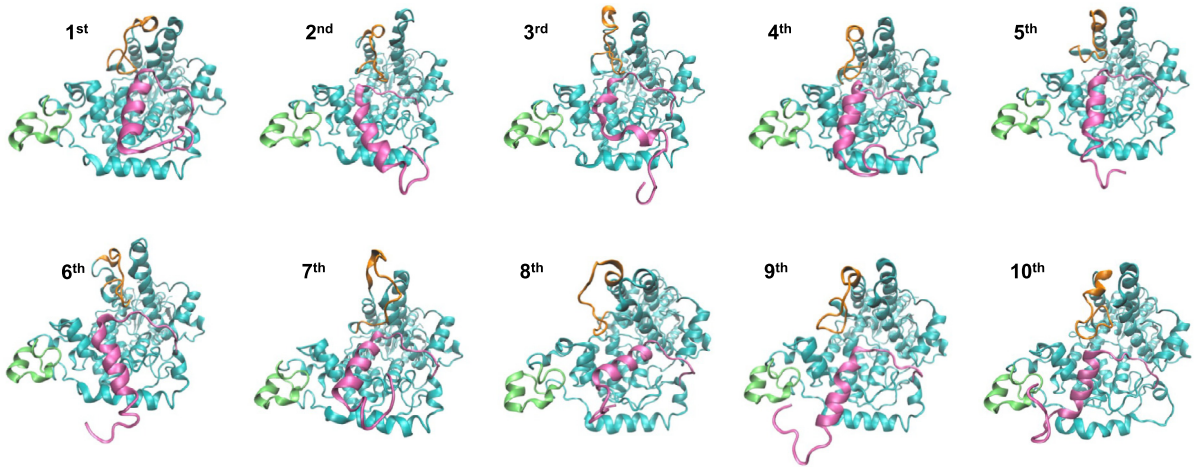

Figure 12. Representative structures of *Cl*CRY4 in the attraction basins on the distribution density map based on autoencoder method. The PBL is shown in brown and the CCE domain is shown in pink.

C-terminal still points in the same direction as in the above five basins. In attraction basins 6, 9, and 10, the CCE domain adopts an orientation opposite to that observed in the six basins described above, yet its helical structure remains stable. Notably, in attraction basin 8, the CCE domain helix adopts a unique position distinct from all other states, suggesting an alternative conformational state that may play a role in *Cl*CRY4 dynamics. As for basins 9 and 10, the CCE helix orients close to so-called EEE region, which is considered as active region promoting the signal transduction [60,61]. The EEE region contains a motif of three consecutive glutamic acids (E) as residues 450 to 452 and it was observed as an IDR in previous study [62]. The RMSF values for the EEE region form a peak close to the CCE domain (Figure 7b). The PBL regions for the selected 10 conformations in Figure 12, only 7 and 8 display open-gate conformation while others display close-gate conformation. Among the 10 representative *Cl*CRY4 structures, only conformations 7 and 8 exhibit an open-gate conformation in the PBL regions, while the others display a closed-gate conformation. This observation is consistent with the previous finding that in dark condition the PBL is in close-gate shape [62].

Given the fundamental methodological differences between the autoencoder and t-ICA approaches, the distributions of the simulation on these latent spaces differ. Specifically, t-ICA is a linear dimensionality reduction technique that explicitly leverages temporal correlations to identify slow collective motions, while the autoencoder utilizes a nonlinear neural-network-based approach to learn a data-driven representation without explicitly incorporating temporal sequence information. Consequently, the distributions of conformations and the positions of attraction basins (distinct maxima of density distribution) identified by the two methods may differ significantly.

A 2D latent space was used as a balance between interpretability and fidelity. Although higher-dimensional embeddings can retain more structural information, clearer and more intuitive visualization of protein conformational distributions is provided by 2D representations, which are commonly adopted in related studies [25,63,64]. To assess information loss, reconstruction accuracy was evaluated across different latent dimensions, revealing that while mean squared error (MSE) decreases with increasing dimensionality of latent space, the improvement from 2D to 3D was relatively modest (Table S6).

Concluding Remark

In this study, we combined conventional molecular dynamics (MD) simulations at ambient temperature and temperature replica exchange molecular dynamics (T-REMD) simulations with advanced dimensionality reduction techniques to investigate the conformational landscape of *Columba livia* cryptochrome 4 (*Cl*CRY4). Particular attention was given to its cryptochrome C-terminal extension (CCE) domain and its interactions with the phosphate binding loop (PBL). By using the high-resolution *Drosophila* cryptochrome (*d*CRY) structure as a reference, we validated the sampling efficiency of T-REMD and demonstrated the robustness of our computational framework in capturing wide range of conformational states of CRYs.

The comparative analysis of *d*CRY and *Cl*CRY4 presented in this study revealed key structural and dynamic differences, particularly in the CCE domain, which may be critical for their functional divergence. Clustering analysis of selected *Cl*CRY4 T-REMD snapshots, using *d*CRY conventional MD simulation as a reference, identified multiple conformational states, highlighting the intrinsic flexibility of the CCE domain. This proof-of-concept study offers a comprehensive computational perspective on sampling the structure of the full *Cl*CRY4, revealing that the CCE domain displays more prominent correlation with PBL than in *d*CRY. In addition, *Cl*CRY4 CCE domain displaces wide range of distribution in its conformational space with significantly different orientations which are likely key to *Cl*CRY4 function.

It is important to note that our current simulations do not account for the presence of FAD, whose redox state can significantly influence the conformational distribution of the CCE domain. Given the intrinsically disordered nature of the CCE domain, further computational studies incorporating various redox states of FAD are necessary to elucidate the behavior of the CCE domain and to understand the underlying mechanisms of FAD-mediated photoreception in CRYs. It should also be noted that the current study is limited to comparing structural relationships among CRYs with similarly sized CCE domains. Its applicability to CRYs with significantly longer CCE regions—such as mammalian CRY1 [65] or *Platynereis dumerilii* L-CRY [66]—remains to be evaluated.

Overall, the proposed simulation procedure could serve as novel computational protocol for further study of *Cl*CRY4 and other CRYs by leveraging the conformational space of available high-resolution structure of CRYs. This could pave the road for studies to provide insights into CRY structural dynamics and potential functional implications in magnetoreception. Furthermore, it holds broader applicability in exploring intrinsically disordered regions in proteins, enhancing our understanding of their dynamic behavior and elucidating their roles in biological signaling and functional regulation.

Funding information

Research reported in this article was supported by the National Institute of General Medical Sciences of the National Institutes of Health under awards no. R15GM122013 (PT), R15GM109282 (BDZ), and the Welch Foundation (BDZ).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to ensure grammatical correctness. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Chuanye Xiong: Writing – original draft, Validation, Resources, Methodology, Investigation, Data curation, Conceptualization. **Palanisamy Kandhan:** Validation, Resources. **Brian Zoltowski:** Writing – review & editing, Supervision. **Peng Tao:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Formal analysis, Data curation, Conceptualization.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Computational time was provided by the Southern Methodist University's O'Donnell Data Science and Research Computing Institute (ODSRCI). We would like to thank Ms. Anna Elise Lester for insightful discussions related to this work.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.jmb.2025.169233.

Received 10 March 2025;
Accepted 23 May 2025;
Available online 27 May 2025

Keywords:
T-REMD;
Autoencoder;
t-ICA;
*d*CRY;
*Cl*CRY4

References

[1]. Brodsky, S., Jana, T., Mittelman, K., Chapal, M., Kumar, D.K., Carmi, M., Barkai, N., (2020). Intrinsically disordered regions direct transcription factor in vivo binding specificity. *Mol. Cell* **79** (3), 459–471.e4. https://doi.org/10.1016/j.molcel.2020.05.032.

[2]. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., Kim, P.M., Kriwacki, R.W., Oldfield, C.J., Pappu, R.V., Tompa, P., Uversky, V.N., Wright, P.E., Babu, M.M., (2014). Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114** (13), 6589–6631. https://doi.org/10.1021/cr400525m.

[3]. Babu, M.M., (2016). The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.* **44** (5), 1185–1200. https://doi.org/10.1042/BST20160172.

[4]. Maiti, S., Singh, A., Maji, T., Saibo, N.V., De, S., (2024). Experimental methods to study the structure and dynamics of intrinsically disordered regions in proteins. *Curr. Res. Struct. Biol.* **7**, 100138. https://doi.org/10.1016/j.crstbi.2024.100138.

[5]. Holehouse, A.S., Kragelund, B.B., (2024). The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Rev. Mol. Cell Biol.* **25** (3), 187–211. https://doi.org/10.1038/s41580-023-00673-0.

[6]. Wright, P.E., Dyson, H.J., (2015). Intrinsically Disordered proteins in Cellular signalling and regulation. *Nature Rev. Mol. Cell Biol.* **16** (1), 18–29. https://doi.org/10.1038/nrm3920.

[7]. Martínez de Paz, A., Khajavi, L., Martin, H., Claveria-Gimeno, R., Tom Dieck, S., Cheema, M.S., Sanchez-Mut, J.V., Moksa, M.M., Carles, A., Brodie, N.I., Sheikh, T.I., Freeman, M.E., Petrotchenko, E.V., Borchers, C.H., Schuman, E.M., Zytnicki, M., Velazquez-Campoy, A., Abian, O., Hirst, M., Esteller, M., Vincent, J.B., Malnou, C.E., Ausió, J., (2019). MeCP2-E1 isoform is a dynamically expressed, weakly DNA-bound protein with different protein and DNA interactions compared to MeCP2-E2. *Epigenet. Chromat.* **12** (1), 63. https://doi.org/10.1186/s13072-019-0298-1.

[8]. Lin, C., Todo, T., (2005). The cryptochromes. *Genome Biol.* **6** (5), 220. https://doi.org/10.1186/gb-2005-6-5-220.

[9]. DeOliveira, C.C., Crane, B.R., (2024). A structural decryption of cryptochromes. *Front. Chem.* **12** https://doi.org/10.3389/fchem.2024.1436322.

[10]. Karki, N., Vergish, S., Zoltowski, B.D., (2021). Cryptochromes: photochemical and structural insight into magnetoreception. *Protein Sci.* **30** (8), 1521–1534. https://doi.org/10.1002/pro.4124.

[11]. Wang, J., Du, X., Pan, W., Wang, X., Wu, W., (2015). Photoactivation of the cryptochrome/photolyase

superfamily. *J. Photochem. Photobiol. C Photochem. Rev.* **22**, 84–102. https://doi.org/10.1016/j.jphotochemrev.2014.12.001.

[12]. Partch, C.L., Clarkson, M.W., Özgür, S., Lee, A.L., Sancar, A., (2005). Role of structural plasticity in signal transduction by the cryptochrome blue-light photoreceptor. *Biochemistry* **44** (10), 3795–3805. https://doi.org/10.1021/bi047545g.

[13]. Lara, J., Zoltowski, B.D., (2020). A tail of CRY selectivity. *Nature Chem. Biol.* **16** (6), 608–609. https://doi.org/10.1038/s41589-020-0531-z.

[14]. Surme, S., Ergun, C., Gul, S., Akyel, Y.K., Gul, Z.M., Ozcan, O., Ipek, O.S., Akarlar, B.A., Ozlu, N., Taskin, A. C., Turkay, M., Gören, A.C., Baris, I., Ozturk, N., Guzel, M., Aydin, C., Okyar, A., Kavakli, I.H., (2023). TW68, cryptochromes stabilizer, regulates fasting blood glucose levels in diabetic *ob/ob* and high fat-diet-induced obese mice. *Biochem. Pharmacol.* **218**, 115896. https://doi.org/10.1016/j.bcp.2023.115896.

[15]. Zoltowski, B.D., Chelliah, Y., Wickramaratne, A., Jarocha, L., Karki, N., Xu, W., Mouritsen, H., Hore, P.J., Hibbs, R. E., Green, C.B., Takahashi, J.S., (2019). Chemical and structural analysis of a photoactive vertebrate cryptochrome from pigeon. *Proc. Natl. Acad. Sci.* **116** (39), 19449–19457. https://doi.org/10.1073/pnas.1907875116.

[16]. Lin, C., Top, D., Manahan, C.C., Young, M.W., Crane, B. R., (2018). Circadian clock activity of cryptochrome relies on tryptophan-mediated photoreduction. *Proc. Natl. Acad. Sci.* **115** (15), 3822–3827. https://doi.org/10.1073/pnas.1719376115.

[17]. Wang, X., Jing, C., Selby, C.P., Chiou, Y.-Y., Yang, Y., Wu, W., Sancar, A., Wang, J., (2018). Comparative properties and functions of type 2 and type 4 pigeon cryptochromes. *Cell. Mol. Life Sci.* **75** (24), 4629–4641. https://doi.org/10.1007/s00018-018-2920-y.

[18]. Kriebs, A., Jordan, S.D., Soto, E., Henriksson, E., Sandate, C.R., Vaughan, M.E., Chan, A.B., Duglan, D., Papp, S.J., Huber, A.-L., Afetian, M.E., Yu, R.T., Zhao, X., Downes, M., Evans, R.M., Lamia, K.A., (2017). Circadian repressors CRY1 and CRY2 broadly interact with nuclear receptors and modulate transcriptional activity. *Proc. Natl. Acad. Sci.* **114** (33), 8776–8781. https://doi.org/10.1073/pnas.1704955114.

[19]. Kavakli, I.H., Gul, S., Turkay, M., (2022). Identification of novel small molecules targeting core clock proteins to regulate circadian rhythm. *Curr. Opin. Chem. Eng.* **35**, 100730. https://doi.org/10.1016/j.coche.2021.100730.

[20]. Rasmussen, E.S., Takahashi, J.S., Green, C.B., (2022). Time to target the circadian clock for drug discovery. *Trends Biochem. Sci.* **47** (9), 745–758. https://doi.org/10.1016/j.tibs.2022.04.009.

[21]. Miller, S., Srivastava, A., Nagai, Y., Aikawa, Y., Tama, F., Hirota, T., (2021). Structural differences in the FAD-binding pockets and lid loops of mammalian CRY1 and CRY2 for isoform-selective regulation. *Proc. Natl. Acad. Sci.* **118**, (26)e2026191118 https://doi.org/10.1073/pnas.2026191118.

[22]. Pu, H., Bailey, L.C., Bauer, L.G., Voronkov, M., Baxter, M., Huber, K.V.M., Khorasanizadeh, S., Ray, D., Rastinejad, F., (2025). Pharmacological targeting of BMAL1 modulates circadian and immune pathways. *Nature Chem. Biol.* **21** (5), 736–745. https://doi.org/10.1038/s41589-025-01863-x.

[23]. Czarna, A., Berndt, A., Singh, H.R., Grudziecki, A., Ladurner, A.G., Timinszky, G., Kramer, A., Wolf, E., (2013). Structures of *drosophila* cryptochrome and mouse cryptochrome1 provide insight into circadian function. *Cell* **153** (6), 1394–1405. https://doi.org/10.1016/j.cell.2013.05.011.

[24]. Moradi, S., Nowroozi, A., Aryaei Nezhad, M., Jalali, P., Khosravi, R., Shahlaei, M., (2024). A review on description dynamics and conformational changes of proteins using combination of principal component analysis and molecular dynamics simulation. *Comput. Biol. Med.* **183**, 109245. https://doi.org/10.1016/j.compbiomed.2024.109245.

[25]. Trozzi, F., Wang, X., Tao, P., (2021). UMAP as a dimensionality reduction tool for molecular dynamics simulations of biomacromolecules: a comparison study. *J. Phys. Chem. B* **125** (19), 5022–5034. https://doi.org/10.1021/acs.jpcb.1c02081.

[26]. Tian, H., Tao, P., (2020). Ivis dimensionality reduction framework for biomacromolecular simulations. *J. Chem. Inf. Model.* **60** (10), 4569–4581. https://doi.org/10.1021/acs.jcim.0c00485.

[27]. Babini, E., Bertini, I., Capozzi, F., Luchinat, C., Quattrone, A., Turano, M., (2005). Principal component analysis of the conformational freedom within the EF-hand superfamily. *J. Proteome Res.* **4** (6), 1961–1971. https://doi.org/10.1021/pr050148n.

[28]. Naritomi, Y., Fuchigami, S., (2013). Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis. *J. Chem. Phys.* **139**, (21)215102 https://doi.org/10.1063/1.4834695.

[29]. Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W.H., Ng, L.G., Ginhoux, F., Newell, E.W., (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnol.* **37** (1), 38–44. https://doi.org/10.1038/nbt.4314.

[30]. Sainburg, T., McInnes, L., Gentner, T.Q., (2021). Parametric UMAP embeddings for representation and semisupervised learning. *Neural Comput.* **33** (11), 2881–2907. https://doi.org/10.1162/neco_a_01434.

[31]. Appadurai, R., Koneru, J.K., Bonomi, M., Robustelli, P., Srivastava, A., (2023). Clustering heterogeneous conformational ensembles of intrinsically disordered proteins with T-distributed stochastic neighbor embedding. *J. Chem. Theory Comput.* **19** (14), 4711–4727. https://doi.org/10.1021/acs.jctc.3c00224.

[32]. Cao, C., Zhang, Z., Zhao, X., Zhang, T., (2020). Terahertz spectroscopy and machine learning algorithm for non-destructive evaluation of protein conformation. *Opt. Quantum Electron.* **52** (4), 225. https://doi.org/10.1007/s11082-020-02345-1.

[33]. Bhowmik, D., Gao, S., Young, M.T., Ramanathan, A., (2018). Deep clustering of protein folding simulations. *BMC Bioinf.* **19** (18), 484. https://doi.org/10.1186/s12859-018-2507-5.

[34]. Bandyopadhyay, S., Mondal, J., (2021). A deep autoencoder framework for discovery of metastable ensembles in biomacromolecules. *J. Chem. Phys.* **155**, (11)114106 https://doi.org/10.1063/5.0059965.

[35]. Zeng, J., Yang, Z., Tang, Y., Wei, G., (2024). Emerging Frontiers in conformational exploration of disordered proteins: integrating autoencoder and molecular

simulations. *ACS Chem. Neurosci.* **15** (23), 4241–4244. https://doi.org/10.1021/acschemneuro.4c00670.

[36]. Sugita, Y., Okamoto, Y., (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314** (1), 141–151. https://doi.org/10.1016/S0009-2614(99)01123-9.

[37]. Xing, W., Busino, L., Hinds, T.R., Marionni, S.T., Saifee, N.H., Bush, M.F., Pagano, M., Zheng, N., (2013). SCFFBXL3 ubiquitin ligase targets cryptochromes at their cofactor pocket. *Nature* **496** (7443), 64–68. https://doi.org/10.1038/nature11964.

[38]. Wang, Y., Veglia, G., Zhong, D., Gao, J., (2021). Activation mechanism of drosophila cryptochrome through an allosteric switch. *Sci. Adv.* **7**, (25)eabg3815 https://doi.org/10.1126/sciadv.abg3815.

[39]. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., Bodenstein, S.W., Evans, D.A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A.I., Cowie, A., Figurnov, M., Fuchs, F.B., Gladman, H., Jain, R., Khan, Y.A., Low, C.M.R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E.D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., Jumper, J.M., (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630** (8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w.

[40]. Schwede, T., Kopp, J., Guex, N., Peitsch, M.C., (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* **31** (13), 3381–3385. https://doi.org/10.1093/nar/gkg520.

[41]. Krivov, G.G., Shapovalov, M.V., Dunbrack Jr., R.L., (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins Struct. Funct. Bioinform.* **77** (4), 778–795. https://doi.org/10.1002/prot.22488.

[42]. The UniProt Consortium, (2025). UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Res.* **53** (D1), D609–D617. https://doi.org/10.1093/nar/gkae1010.

[43]. Levy, C., Zoltowski, B.D., Jones, A.R., Vaidya, A.T., Top, D., Widom, J., Young, M.W., Scrutton, N.S., Crane, B.R., Leys, D., (2013). Updated structure of drosophila cryptochrome. *Nature* **495** (7441), E3–E4. https://doi.org/10.1038/nature11995.

[44]. Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E., Simmerling, C., (2015). ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11** (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

[45]. Mark, P., Nilsson, L., (2001). Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *J. Phys. Chem. A* **105** (43), 9954–9960. https://doi.org/10.1021/jp003020w.

[46]. Ryckaert, J.-P., Ciccotti, G., Berendsen, H.J.C., (1977). Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **23** (3), 327–341. https://doi.org/10.1016/0021-9991(77)90098-5.

[47]. Sindhikara, D.J., Emerson, D.J., Roitberg, A.E., (2010). Exchange often and properly in replica exchange molecular dynamics. *J. Chem. Theory Comput.* **6** (9), 2804–2808.

[48]. Zhai, J., Zhang, S., Chen, J., He, Q., (2018). Autoencoder and its various variants. *In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 415–419. https://doi.org/10.1109/SMC.2018.00080.

[49]. Wang, Y.-B., You, Z.-H., Li, X., Jiang, T.-H., Chen, X., Zhou, X., Wang, L., (2017). Predicting protein–protein Interactions from protein sequences by a stacked sparse autoencoder deep neural network. *Mol. Biosyst.* **13** (7), 1336–1344. https://doi.org/10.1039/C7MB00188F.

[50]. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., (2019). PyTorch: an imperative style, high-performance deep learning library. *arXiv*. https://doi.org/10.48550/arXiv.1912.01703.

[51]. Schwantes, C.R., Pande, V.S., (2013). Improvements in Markov state model construction reveal many non-native Interactions in the folding of NTL9. *J. Chem. Theory Comput.* **9** (4), 2000–2009. https://doi.org/10.1021/ct300878a.

[52]. Wu, H., Nüske, F., Paul, F., Klus, S., Koltai, P., Noé, F., (2017). Variational Koopman models: slow collective variables and molecular kinetics from short off-equilibrium simulations. *J. Chem. Phys.* **146**, (15)154104 https://doi.org/10.1063/1.4979344.

[53]. Scherer, M.K., Trendelkamp-Schroer, B., Paul, F., Pérez-Hernández, G., Hoffmann, M., Plattner, N., Wehmeyer, C., Prinz, J.-H., Noé, F., (2015). PyEMMA 2: a software package for estimation, validation, and analysis of Markov models. *J. Chem. Theory Comput.* **11** (11), 5525–5542. https://doi.org/10.1021/acs.jctc.5b00743.

[54]. González-Alemán, R., Platero-Rochart, D., Rodríguez-Serradet, A., Hernández-Rodríguez, E.W., Caballero, J., Leclerc, F., Montero-Cabrera, L., (2022). MDSCAN: RMSD-based HDBSCAN clustering of long molecular dynamics. *Bioinformatics* **38** (23), 5191–5198. https://doi.org/10.1093/bioinformatics/btac666.

[55]. Verkhivker, G., Alshahrani, M., Gupta, G., Xiao, S., Tao, P., (2023). Probing conformational landscapes of binding and allostery in the SARS-CoV-2 omicron variant complexes using Microsecond atomistic simulations and perturbation-based profiling approaches: hidden role of omicron mutations as modulators of allosteric signaling and epistatic relationships. *Phys. Chem. Chem. Phys.* **25** (32), 21245–21266. https://doi.org/10.1039/D3CP02042H.

[56]. Ibrahim, M.T., Trozzi, F., Tao, P., (2022). Dynamics of hydrogen bonds in the secondary structures of allosteric protein *Avena sativa* phototropin 1. *Comput. Struct. Biotechnol. J.* **20**, 50–64. https://doi.org/10.1016/j.csbj.2021.11.038.

[57]. dos Santos Nascimento, I.J., De Souza, M., Medeiros, D. C., de Moura, R.O., (2022). Dynamic cross-correlation matrix (DCCM) reveals new insights to discover new NLRP3 inhibitors useful as anti-inflammatory drugs. *Med. Sci. Forum* **14** (1), 84. https://doi.org/10.3390/ECMC2022-13306.

[58]. Qian, J., Song, Z., Yao, Y., Zhu, Z., Zhang, X., (2022). A review on autoencoder based representation learning for fault detection and diagnosis in industrial processes.

Chemom. Intell. Lab. Syst. **231**, 104711. https://doi.org/10.1016/j.chemolab.2022.104711.

[59]. Li, H., Lyu, Q., Cheng, J., (2016). A template-based protein structure reconstruction method using deep autoencoder learning. *J. Proteomics Bioinform.* **9** (12), 306. https://doi.org/10.4172/jpb.1000419.

[60]. Bradlaugh, A.A., Fedele, G., Munro, A.L., Hansen, C.N., Hares, J.M., Patel, S., Kyriacou, C.P., Jones, A.R., Rosato, E., Baines, R.A., (2023). Essential elements of radical pair magnetosensitivity in drosophila. *Nature* **615** (7950), 111–116. https://doi.org/10.1038/s41586-023-05735-z.

[61]. Ramsay, J.L., Schuhmann, F., Solov'yov, I.A., Kattnig, D. R., (2024). Cryptochrome magnetoreception: time course of photoactivation from non-equilibrium coarse-grained molecular dynamics. *Comput. Struct. Biotechnol. J.* **26**, 58–69. https://doi.org/10.1016/j.csbj.2024.11.001.

[62]. Schuhmann, F., Ramsay, J.L., Kattnig, D.R., Solov'yov, I. A., (2024). Structural rearrangements of pigeon cryptochrome 4 undergoing a complete redox cycle. *J. Phys. Chem. B* **128** (16), 3844–3855. https://doi.org/10.1021/acs.jpcb.4c00424.

[63]. Wu, K.E., Yang, K.K., van den Berg, R., Alamdari, S., Zou, J.Y., Lu, A.X., Amini, A.P., (2024). Protein structure generation via folding diffusion. *Nature Commun.* **15** (1), 1059. https://doi.org/10.1038/s41467-024-45051-2.

[64]. Tian, H., Jiang, X., Xiao, S., La Force, H., Larson, E.C., Tao, P., (2023). LAST: latent space-assisted adaptive sampling for protein trajectories. *J. Chem. Inf. Model.* **63** (1), 67–75. https://doi.org/10.1021/acs.jcim.2c01213.

[65]. Kolarski, D., Miller, S., Oshima, T., Nagai, Y., Aoki, Y., Kobauri, P., Srivastava, A., Sugiyama, A., Amaike, K., Sato, A., Tama, F., Szymanski, W., Feringa, B.L., Itami, K., Hirota, T., (2021). Photopharmacological manipulation of mammalian CRY1 for regulation of the circadian clock. *J. Am. Chem. Soc.* **143** (4), 2078–2087. https://doi.org/10.1021/jacs.0c12280.

[66]. Poehn, B., Krishnan, S., Zurl, M., Coric, A., Rokvic, D., Häfker, N.S., Jaenicke, E., Arboleda, E., Orel, L., Raible, F., Wolf, E., Tessmar-Raible, K., (2022). A cryptochrome adopts distinct moon- and sunlight states and functions as sun- versus moonlight interpreter in monthly oscillator entrainment. *Nature Commun.* **13** (1), 5220. https://doi.org/10.1038/s41467-022-32562-z.