Title

Machine Learning–Assisted Analysis of Conformational Dynamics in a Photosynthetic Protein Using Molecular Dynamics Simulations

By

YASH SINGH SENGAR

M.Sc. Chemistry, NIT Calicut
Computational Biophysics Lab, IISc Bangalore

DEC.2025

Abstract

Understanding the conformational dynamics of photosynthetic proteins is essential for elucidating their functional mechanisms. In this study, molecular dynamics (MD) simulations combined with unsupervised machine learning were employed to investigate the structural flexibility and dominant collective motions of a photosynthetic protein (PDB ID: 1RWT). Classical MD simulations were used to generate atomistic trajectories, from which structural descriptors such as RMSD, RMSF, and solvent-accessible surface area (SASA) were extracted. Principal component analysis (PCA) was applied as an unsupervised machine learning method to reduce the dimensionality of the conformational space and identify dominant motions. The resulting conformational ensemble was further analyzed through free energy landscape (FEL) reconstruction. The results reveal multiple metastable conformational states and highlight the ability of machine learning techniques to extract physically meaningful insights from simulation data. This work demonstrates a robust framework for integrating molecular simulations with data-driven approaches to study protein dynamics.

1. Introduction

Photosynthetic proteins play a crucial role in converting light energy into chemical energy and are inherently dynamic in nature. Their function is often governed not by a single static structure but by an ensemble of conformations that interconvert over time. Capturing these conformational changes is therefore essential for understanding structure–function relationships in photosynthesis.

Molecular dynamics (MD) simulations provide a powerful computational tool to explore protein dynamics at an atomistic resolution. However, MD trajectories are high-dimensional and complex, making it challenging to extract meaningful collective motions directly. Machine learning (ML) approaches, particularly unsupervised methods, have emerged as effective tools for identifying dominant patterns and reducing the dimensionality of simulation data.

In this project, we combine MD simulations with unsupervised machine learning techniques to study the conformational dynamics of a photosynthetic protein. Principal component analysis (PCA) is employed to identify collective motions, and the free energy landscape (FEL) is constructed to characterize the stability of different conformational states.

2. Methods

2.1 Molecular Dynamics Simulations

Classical molecular dynamics simulations were performed for the photosynthetic protein with PDB ID **1RWT**. The protein structure was prepared, solvated, and subjected to energy minimization, equilibration, and production MD runs using standard protocols. The resulting trajectory provides a time-resolved description of the protein's conformational dynamics under physiological conditions.

2.2 Trajectory Analysis

To assess the structural stability and flexibility of the protein, several standard MD descriptors were computed:

- **Root Mean Square Deviation (RMSD):** Used to monitor overall structural stability relative to the initial structure.
- **Root Mean Square Fluctuation (RMSF):** Calculated per residue to identify flexible and rigid regions.
- **Solvent Accessible Surface Area (SASA):** Evaluated to quantify changes in solvent exposure during the simulation.

These descriptors serve as both physical observables and input features for downstream machine learning analysis.

2.3 Machine Learning Analysis: Principal Component Analysis

Principal component analysis (PCA), an unsupervised machine learning technique, was applied to the MD trajectory to reduce the dimensionality of the conformational space. PCA identifies orthogonal collective variables (principal components) that capture the largest variance in atomic motions.

The first two principal components (PC1 and PC2) were used to project the trajectory onto a low-dimensional space, enabling visualization and identification of dominant conformational states.

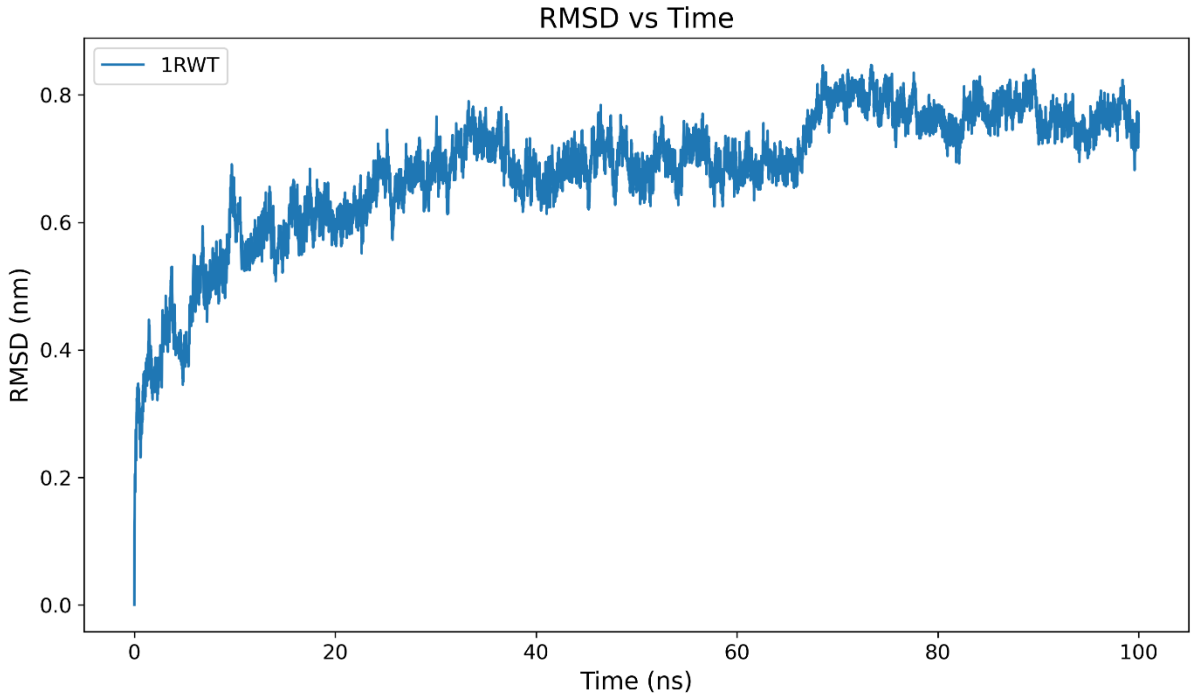
2.4 Free Energy Landscape Construction

The free energy landscape (FEL) was reconstructed as a function of the first two principal components using a Boltzmann inversion of the probability distribution. This approach allows identification of low-energy basins corresponding to metastable conformational states and provides an energetic interpretation of the PCA results.

3. Results

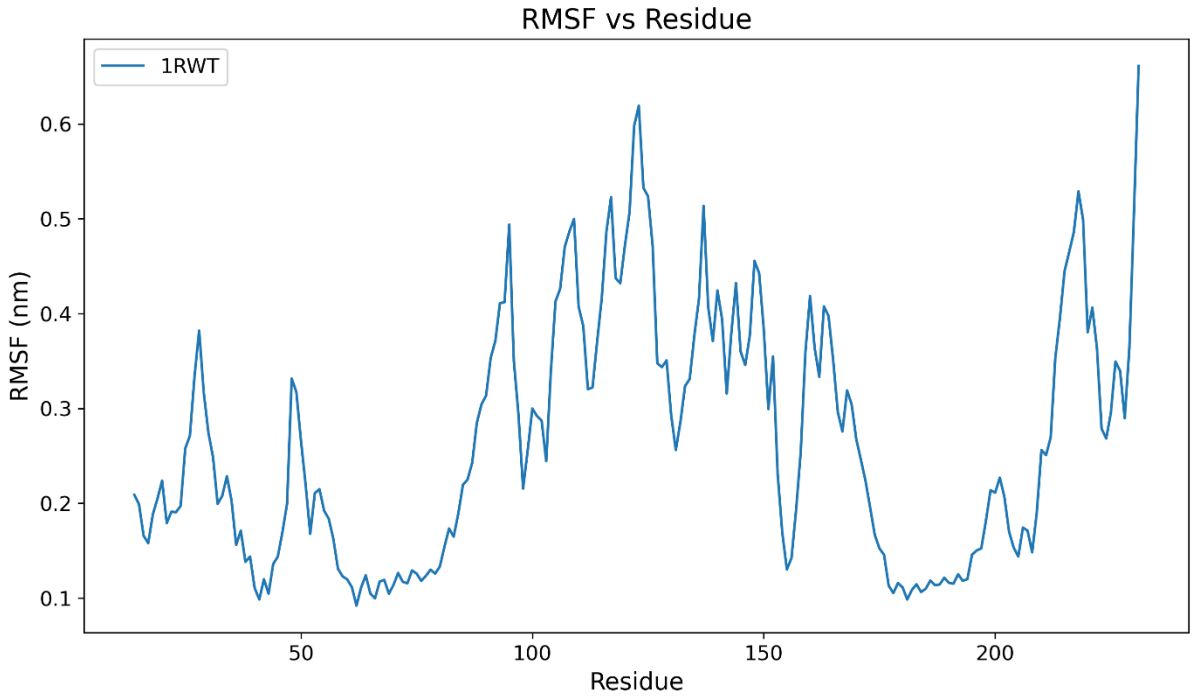
3.1 Structural Stability and Flexibility

RMSD of the protein backbone as a function of simulation time, indicating structural equilibration and stability.



The RMSD profile shows an initial equilibration phase followed by a stable plateau, indicating that the protein reaches a stable conformational ensemble during the simulation. This stability confirms the suitability of the trajectory for further analysis.

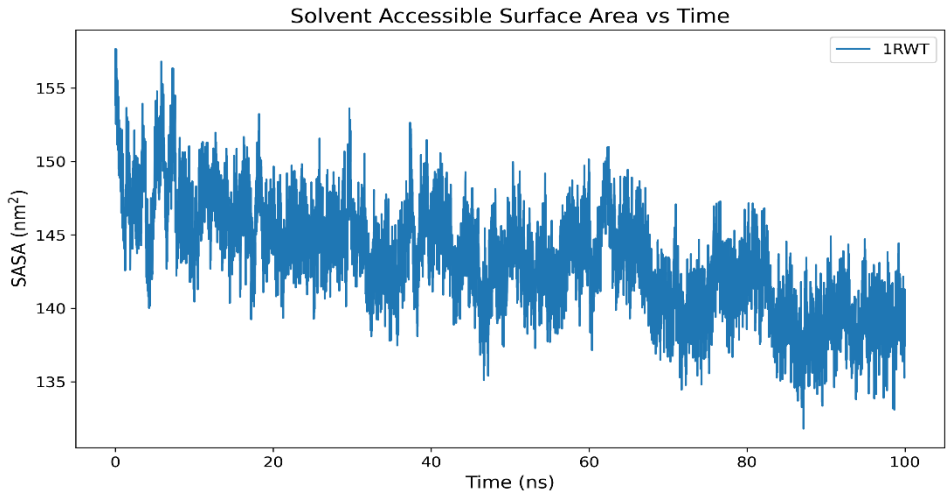
RMSF per residue showing localized flexibility across the protein structure.



Residue-wise RMSF analysis reveals localized regions of enhanced flexibility, likely corresponding to loop regions or functionally important segments. These flexible regions may play a role in accommodating conformational changes required for photosynthetic activity.

3.2 Solvent Exposure Dynamics

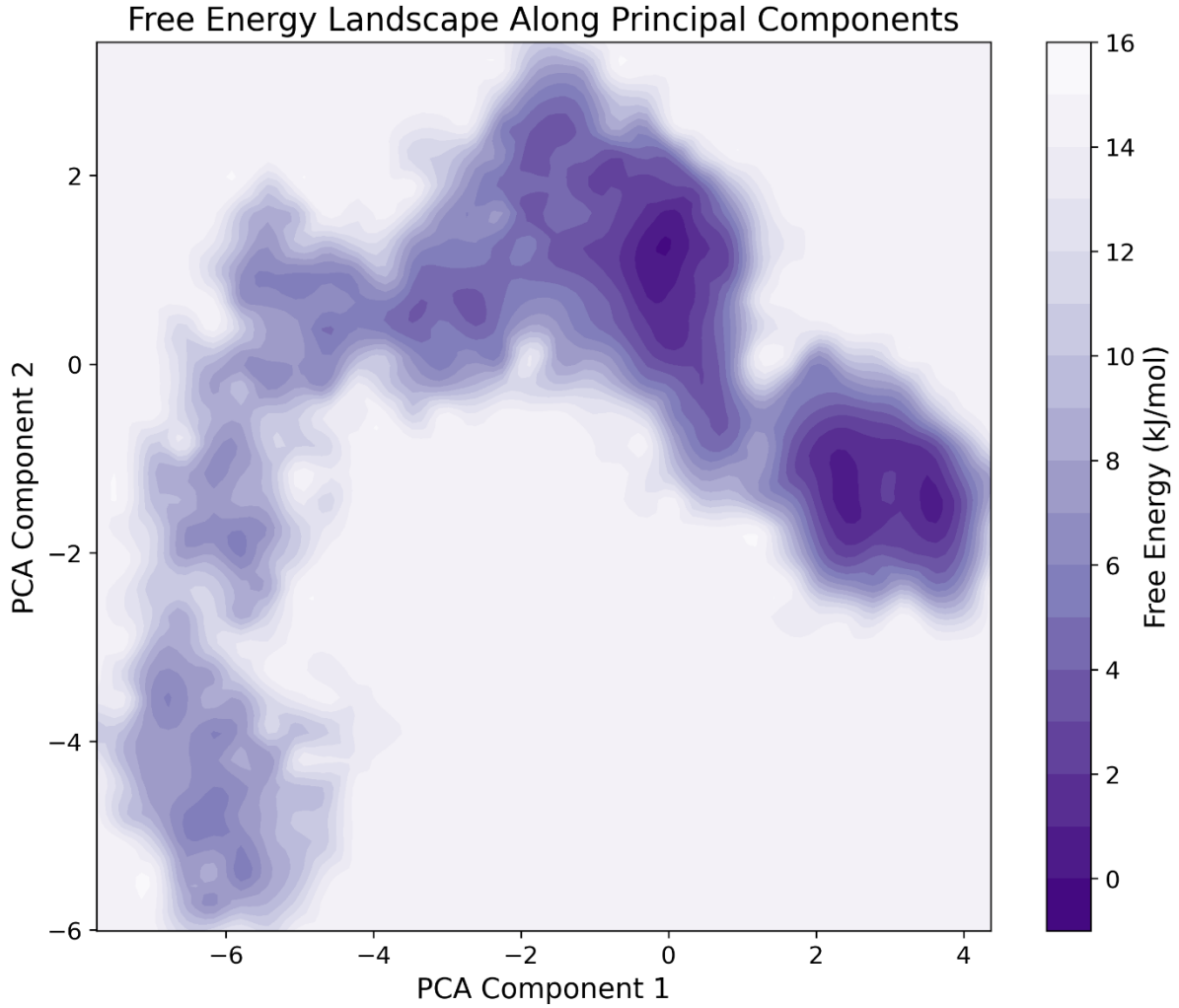
Solvent Accessible Surface Area (SASA) as a function of time, reflecting dynamic solvent exposure.



The SASA profile exhibits moderate fluctuations over time, reflecting dynamic solvent exposure of the protein surface. These fluctuations suggest conformational breathing motions rather than large-scale unfolding events.

3.3 Identification of Collective Motions via PCA

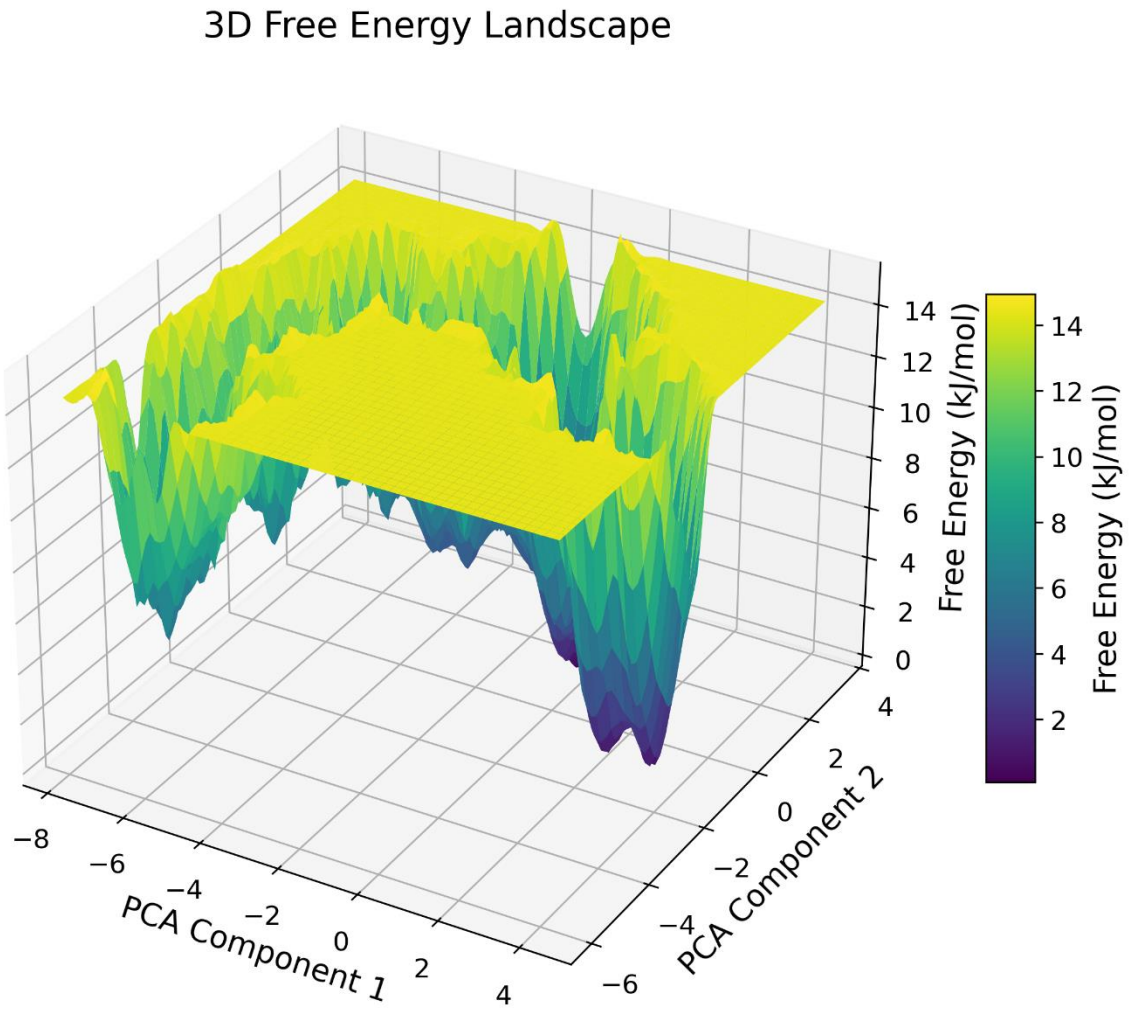
Projection of MD trajectory onto the first two principal components (PC1 and PC2) obtained from PCA.



Projection of the MD trajectory onto the first two principal components reveals clustering of conformations into distinct regions of conformational space. This demonstrates that a small number of collective variables capture the dominant motions of the protein.

3.4 Free Energy Landscape and Metastable States

Free energy landscape reconstructed along PC1 and PC2, highlighting metastable conformational states.



The free energy landscape constructed along PC1 and PC2 displays multiple low-energy basins separated by energy barriers. These basins correspond to metastable conformational states sampled during the simulation and align with the clusters observed in PCA space.

4. Discussion

The combination of MD simulations with unsupervised machine learning provides a powerful framework for analyzing protein dynamics. RMSD and RMSF analyses confirm the structural stability of the protein while highlighting

flexible regions that may be functionally relevant. PCA successfully reduces the complexity of the MD data and reveals dominant collective motions governing the conformational behaviour of the protein.

The free energy landscape offers an energetic perspective, validating the PCA-derived clusters as physically meaningful metastable states. Together, these results demonstrate how machine learning methods can complement traditional simulation analysis by uncovering patterns that are not immediately apparent from raw trajectories.

5. Conclusion

This study demonstrates the successful integration of molecular dynamics simulations with unsupervised machine learning techniques to analyze the conformational dynamics of a photosynthetic protein. Principal component analysis and free energy landscape reconstruction reveal multiple metastable states and dominant collective motions. The approach presented here provides a general and extensible framework for studying complex biomolecular systems using data-driven methods.