AlphaFold-SFA: accelerated sampling of cryptic pocket opening by fusing AlphaFold, slow feature analysis and metadynamics

Shray Vats[1], Pär Söderhjelm[2], Soumendranath Bhakat[3]

[1]Department of Computer Science, University of Texas at Austin, Austin, TX, 78705. Email: shrayvats@gmail.com
[2] Division of Biophysical Chemistry, Chemical Center, Lund University, P.O.B. 124, 22100, Lund, Sweden. Email: par.soderhjelm@bpc.lu.se
[3] AlloTec, 4059 Utah St. St. Louis, MO 63116. Email: bhakatsoumendranath@gmail.com

Abstract

Sampling *rare events* in proteins is crucial for comprehending complex phenomena like cryptic pocket opening, where transient structural changes expose new binding sites. Understanding these rare events also sheds light on protein-ligand binding and allosteric communications, where distant site interactions influence protein function. Traditional unbiased molecular dynamics simulations often fail to sample such rare events, as the free energy barrier between metastable states is large relative to the thermal energy. This renders these events inaccessible on the timescales typically simulated by standard molecular dynamics, limiting our understanding of these critical processes. In this working paper, we proposed a novel unsupervised learning approach termed as *slow feature analysis* (SFA) which aims to extract slowly varying features from high-dimensional temporal data. SFA trained on small unbiased molecular dynamics simulations launched from AlphaFold generated conformational ensembles manages to capture rare events governing cryptic pocket opening in plasmepsin II. Metadynamics simulations using SFA as collective variables manage to sample 'deep' cryptic pocket opening within a few hundreds of nanoseconds which was beyond the reach of microsecond long unbiased molecular dynamics simulations. Taken together, our results show how SFA acts as a dimensionality reduction tool which bridges the gap between AlphaFold, molecular dynamics simulation and metadynamics in context of capturing rare events in biomolecules, extending the scope of structure-based drug discovery in the era of AlphaFold.

Introduction

The challenge of accurately predicting the 3D structure of proteins based solely on their amino acid sequences has been a longstanding puzzle in the realm of structural biology. Traditionally, scientists relied heavily on experimental techniques like X-ray crystallography and cryogenic electron microscopy (Cryo-EM) to decipher these protein structures[1]. While these methods remain crucial for examining intricate biomolecules, the landscape underwent a transformative shift in 2021. This change was marked by the introduction of AlphaFold[2], an artificial intelligence (AI)-driven model that showcased its prowess in predicting protein 3D structures from their sequences. Building on this innovation, ColabFold[3] was subsequently developed, optimizing AlphaFold to

operate seamlessly on Google Colab, thereby democratizing AI-based protein structure prediction. However, it's essential to highlight a shared limitation across AlphaFold, X-ray crystallography, and Cryo-EM: while they excel at capturing a static representation or 'snapshot' of a protein, they fall short in depicting dynamic, biologically significant protein movements. Such movements, like the unveiling of cryptic pockets[4] or allosteric communication[5], often involve the observation of transient high-energy states, commonly referred to as *rare events*[6,7]

Molecular dynamics (MD) simulations[8,9] in theory, possess the capability to sample rare molecular events. Yet, a significant limitation emerges; the timescales that MD simulations can access are restricted. As a result, the molecular system frequently remains ensnared in a specific free energy trough. This confinement means it's challenging to capture these infrequent events with the desired level of detail. Enter a new trick: the stochastic subsampling of multiple sequence alignment (MSA). This technique recently enabled AlphaFold to sample a diverse conformational ensemble of 3D protein structures[10,11], even those in high energy states. Seeding MD simulations with these AlphaFold generated ensembles followed by Markov State Modelling (MSM) can provide Boltzmann-weighted probability distribution associated with cryptic pocket opening[12]. However, this approach isn't without its challenges. To gather sufficient samples to construct a reliable MSM[13–15], one would require a combined simulation time spanning tens of microseconds. Enhanced sampling techniques, especially metadynamics, offer a potent alternative to sample rare transitions within reasonable timescale. Metadynamics accelerates sampling by depositing Gaussian shaped bias along predefined reaction coordinates often known as collective variables[16,17]. Selecting the optimal collective variables[7], capable of capturing rare transitions, remains at the forefront of biomolecular simulation research.

In this working paper, we introduce slow feature analysis (SFA)[18], an unsupervised learning algorithm which can capture slowly varying features from high-dimensional temporal data generated by MD simulations. SFA trained on short MD simulations seeded from AlphaFold generated ensemble is used as collective variables in metadynamics to capture cryptic pocket opening in plasmepsin-II[19,20], a drug target for malaria. We propose slow feature analysis as a dimensionality reduction algorithm that bridges the gap between AlphaFold and metadynamics and manages to capture Boltzmann distribution associated with cryptic pocket opening.

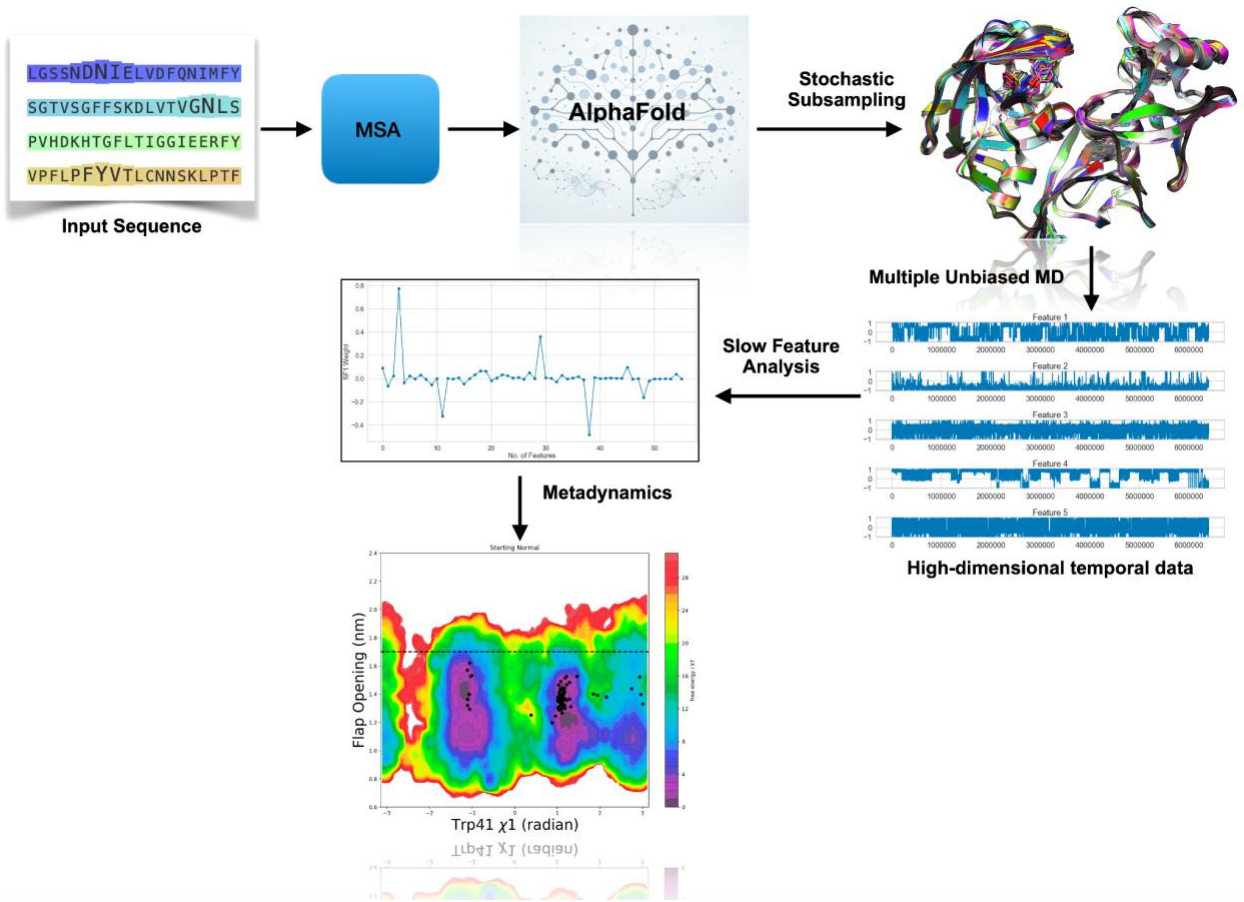

Figure 1. Sampling of deep cryptic pocket opening in plasmepsin II is a *rare event*. Multiple sequence alignment (MSA) of an input sequence followed by stochastic subsampling enabled AlphaFold to generate a conformational ensemble of plasmepsin II with structural diversity. AlphaFold generated structural ensemble was used as a starting point to run multiple short (~40 ns each) unbiased MD simulations which generated high-dimensional temporal data associated with protein dynamics. Slow feature analysis (SFA) is an unsupervised learning algorithm which captures *slowly varying* features from high-dimensional temporally evolving data. SFA as the collective variables in metadynamics simulations, efficiently sample *rare* deep pocket opening in plasmepsin II. Remarkably, these simulations manage to capture these elusive events within just a few hundred nanoseconds, showcasing the power of fusing SFA, AlphaFold and metadynamics to capture protein dynamics.

Methods

Structural Ensemble generation using AlphaFold

Structural ensemble for a given sequence was generated using ColabFold implementation of AlphaFold(link:https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/Alpha

Fold2.ipynb#scrollTo=KK7X9T44pWb7). We generated initial multiple sequence alignment (MSA) using the MMseqs2 method[21] implemented with ColabFold. We then stochastically subsampled the MSA to a maximum of 32 cluster centers and 64 extra sequences (noted as *max_msa* = 32:64). For the generation of the structural ensemble, we opted for the "*complete*" pairing strategy, which only pairs sequences with a full taxonomic match. We set the number of random seeds to 16 and enabled model dropout. Using dropout, combined with the increased number of random seeds, allows AlphaFold's neural network to tap into the model's inherent uncertainties leading to generation of structural ensemble (total 80 structures for plasmepsin-II) with conformational heterogeneity.

Molecular Dynamics Simulations

Each structure from the conformational ensemble generated by AlphaFold was prepared for molecular dynamics simulations using the *tleap* module from Amber2022[22], following the protocol outlined by *Meller et al*[12]. In brief, the proteins were parameterized with the AMBER FF14SB force field[23]. To achieve system neutrality, 17 Na^+ ions were added to each system. Systems were then solvated within a truncated octahedron box, ensuring a minimum of 10 Å between the protein and the edge of the box. The system underwent a two-phase minimization: (a) an initial phase where only the water and ions were minimized while the protein was held in place using a restraint potential of 100 kcal/mol^{-1} Å^2 (200 steepest descent steps followed by 200 conjugate gradient steps), and (b) an unrestrained minimization of the entire system over 500 steps.

After the minimization process in Amber2022, we transformed Amber topologies into Gromacs format via Acpype[24]. Each system was gradually heated from 0 to 300 K for 500 ps in an NVT ensemble, with harmonic restraints (500 kJ mol^{-1}nm^{-2}) on the backbone's heavy atoms. Subsequently, systems were equilibrated for 200 *ps* in an NPT ensemble at 300 K, devoid of restraints. The *Parrinello–Rahman* barostat[25] ensured a consistent pressure of 1 bar, while the v-rescale thermostat controlled the temperature. Production runs were performed in the NPT ensemble, maintaining conditions at 300 K and 1 bar. The leapfrog integrator and Parrinello–Rahman thermostat was employed with a 2 fs timestep. Nonbonded interactions had a cutoff of 1.0 nm and long-range electrostatic interactions were treated using the Particle Mesh Ewald (PME) method[26] with a 0.16 nm grid spacing. The LINCS algorithm[27] was used to constrain H-bonds. Heating, equilibration, and production runs were performed using Gromacs 2022[28].

For the 80 structures of plasmepsin-II produced by AlphaFold, we conducted two independent 40ns production runs, each with unique initial starting velocities. Trajectories were saved every ps.

Slow Feature Analysis (SFA)

Slow Feature Analysis (SFA) is a dimensionality reduction technique designed to process high-dimensional, temporally evolving data. Its primary goal is to transform a J-dimensional input signal, $c(t)$, using a set of nonlinear functions, $g_k(c)$, to produce output signals $y_k(t) = g_k\big(c(t)\big)$. These output signals are optimized to minimize $\Delta y_k = \langle \dot{y}_k^2 \rangle$, where $\dot{y}$ represents the derivative of y, and $\langle\rangle$ indicates temporal averaging. This minimization targets the extraction of features that vary slowly over time. SFA also imposes additional constraints as follows:

a) each output feature must have zero mean $\langle y_k \rangle = 0$,
b) unit variance $\langle y_k^2 \rangle = 1$, and
c) decorrelated from others $\langle y_k y_{k'} \rangle = 0$ for all $k' < k$

These constraints ensure that each extracted feature is scaled similarly, uncorrelated with others, and avoids the trivial solution, $y_k = c$ where c is a constant. When detailing the algorithm below, signals represented by capitals represent raw signals, while signals represented by lower-case represent normalized signals.

To perform the SFA algorithm, first start with a J-dimensional input signal $X(t)$. Next, normalize the input signal to get:

$$c(t) = \left[c_1(t) \dots c_j(t)\right]^T \tag{1}$$

where:

$$c_j(t) = \frac{C_j(t) - \langle C_j \rangle}{\sqrt{\langle \big(C_j(t) - \langle C_j \rangle\big)^2 \rangle}} \tag{2}$$

Such that:

$$\langle c_j \rangle = 0 \text{ and } \langle c_j^2 \rangle = 1$$

Next, perform a linear or a non-linear expansion using a set of functions $H(c)$ to produce an expanded signal $Z(t)$. In the case of a quadratic expansion this would include monomials of degree one and degree two including mixed terms as shown below, which would result in quadratic SFA:

$$H(c) = \left[c_1, \dots, c_j, c_1c_1, c_1c_2, \dots, c_jc_j\right] \tag{3}$$

and thus:

$$Z(t) = H\big(c(t)\big) = \big[c_1(t), \dots, c_j(t), c_1(t)c_1(t), c_1(t)c_2(t), \dots, c_j(t)c_j(t)\big] \tag{4}$$

The next phase involves sphering (or whitening) of $Z(t)$, generating a normalized signal $z(t)$ through the equation:

$$z(t) = S(Z(t) - \langle Z \rangle) \tag{5}$$

where:

$$\langle z \rangle = 0 \text{ and } \langle zz^T \rangle = I$$

In this case, I is the identity covariance matrix and S is a sphering matrix. Matrix S can be solved by performing PCA on matrix $(Z(t) - \langle Z \rangle)$.

Further, PCA is applied to the matrix $\langle \dot{z}\dot{z}^T \rangle$ where $\dot{z}$ is the time derivative of the sphered expanded signal $z(t)$. The K eigenvectors with the lowest eigenvalues λ_k result in the normalized weight vectors w_k satisfying:

$$w_k : \langle \dot{z}\dot{z}^T \rangle w_k = \lambda_k w_k \text{ where } \lambda_1 \leq \lambda_2 \dots \leq \lambda_k \tag{6}$$

This leads to the formulation of the desired set of real-valued functions $g(c) = [g_1(c), \dots, g_k(c)]^T$ where:

$$g_k(c) = w_k^T \cdot h(c) \tag{7}$$

and:

$$h(c) = S(H(c) - \langle Z \rangle) \tag{8}$$

Here, S is the sphering matrix that was earlier solved to normalize the expanded signal $Z(t)$. Now, define the output signal $y(t)$ as:

$$y(t) = g\big(c(t)\big) \tag{9}$$

where:

$$\langle y \rangle = 0, \langle yy^T \rangle = I, \text{and } \Delta y_k = \langle \dot{y}_k^2 \rangle = \lambda_k$$

In this formulation, the components of y(t), representing the extracted slow features, are characterized by zero mean, unit variance, and mutual decorrelation, thereby encapsulating the core principles of SFA.

Application of SFA on molecular dynamics

Slow feature analysis was performed on training data (80 structures * 2 clones each * 40 ns = 6.4 microsecond) generated from unbiased molecular dynamics simulations launched from the conformational ensemble generated by AlphaFold.

In this study, we adopted linear Slow Feature Analysis (SFA) for our experimental analysis, implemented in the *sklearn SFA* package (https://sklearn-sfa.readthedocs.io/en/latest/). Our initial step involved the compilation of featurised trajectory data into a J-dimensional input signal, denoted as $c(t)$ and normalizing the data by solving for the whitening matrix S for the input signal using PCA and subsequently transforming it into c^{white}. Whitening can be expressed as the linear map $c^{white} = D^{-\frac{1}{2}}U^T c$ where the covariance matrix of $c(t)$ is decomposed as $C_c = UDU^T$.

We then do PCA on finite differences of the whitened input signal $\dot{c}_t^{white} = c_{t+1}^{white} - c_t^{white}$ to get the decomposition of the covariance matrix for the finite differences $\dot{c}$ such that $C_{\dot{c}} = V\Lambda V^T$.

The eigenvectors corresponding to the smallest eigenvalues in this decomposition represented the normalized weights of the slowest moving features. These extracted weights, derived from the training data, were then utilized to construct a new collective variable for subsequent metadynamics simulations. This approach leverages the strength of linear SFA in distilling critical dynamical features from high-dimensional data, enabling nuanced exploration and modeling of the underlying processes.

Markov State Model

Markov state model was performed on sin and cos transformed $\chi 1$ and $\chi 2$ angles of Trp41 extracted from unbiased molecular dynamics simulations (80 structures * 2 independent clones * 100ns each = total 16 μs) launched from AlphaFold generated conformational ensemble. *Kmeans* clustering was performed on the transformed dihedral space with *k=200*. Finally, maximum likelihood MSM was generated using a lag time of 6 ns (Figure S2 highlights the implied timescale plot). Equilibrium populations extracted from MSM is projected along different features to highlight conformational heterogeneity associated with plasmepsin II. PCCA+[29] was used to

generate macrostate definition which manages to distinguish conformational states associated with Trp41 (Figure S2). MSM generation was performed using PyEMMA 2.5.7[30].

Metadynamics

Well-tempered metadynamics simulations were performed using the first two slow features as collective variables (CVs) at 300K. Gaussians of height 1.50 kJ/mol were deposited at every 500 steps. The Gaussian widths for the first two slow features were set at 0.32 and 0.25, determined by taking approximately one-third of the standard deviation from the unbiased training data. The bias-factor was set at 20. We performed two distinct metadynamics simulations, each lasting around ~400 ns. One began from the closed state (PDB: 1LF4)[31] and the other from the cryptic pocket open state (PDB: 2BJU)[32]. Metadynamics simulations were performed using Gromacs 2022 patched with Plumed 2.7[33]. Unbiased free energy surfaces along different features were extracted using the reweighting protocol developed by *Tiwary and Parrinello*[34].

Results

SFA captures critical fluctuations necessary for cryptic pocket opening

SFA trained on small independent unbiased molecular dynamics simulation launched from a structural ensemble generated by AlphaFold managed to capture flipping of Trp41 in plasmepsin II, necessary for cryptic pocket opening. It also captured flipping of Tyr77 along $\chi 1$ angle as another key feature (Figure S3). Flipping of Tyr77 in conjunction with flipping of Trp41 exposes a fully 'open' cryptic pocket primed for ligand binding.

Metadynamics using SFA accelerates sampling of cryptic pocket opening

A recent study highlighted how microsecond long unbiased MD simulation launched from apo plasmepsin II failed to capture cryptic pocket opening[12]. To test the effectiveness of SFA as CVs with metadynamics we performed two independent simulations, one launched from 'closed' state (PDB: 1LF4) and the other from the cryptic pocket 'open' state (PDB: 2BJU, removing the ligand to make it apo). It is key to highlight unbiased MD simulations launched from *apo-like* 'open' state failed to sample open ↔ closed transitions (Figure S1, Supporting information).

Metadynamics simulations using first two slow features as CVs managed to sample multiple flipping events along Trp41 $\chi 1$ and $\chi 2$ angles leading to close ↔ open transitions in plasmepsin II within ~350-400 ns of simulation time. Reweighted free energy surfaces from metadynamics simulations agreed extremely well with MSM suggesting that we reached convergence (Figure 2) for the choice of force field and water model (see Figure S4 and S5 in Supporting Information regarding convergence of metadynamics simulations).

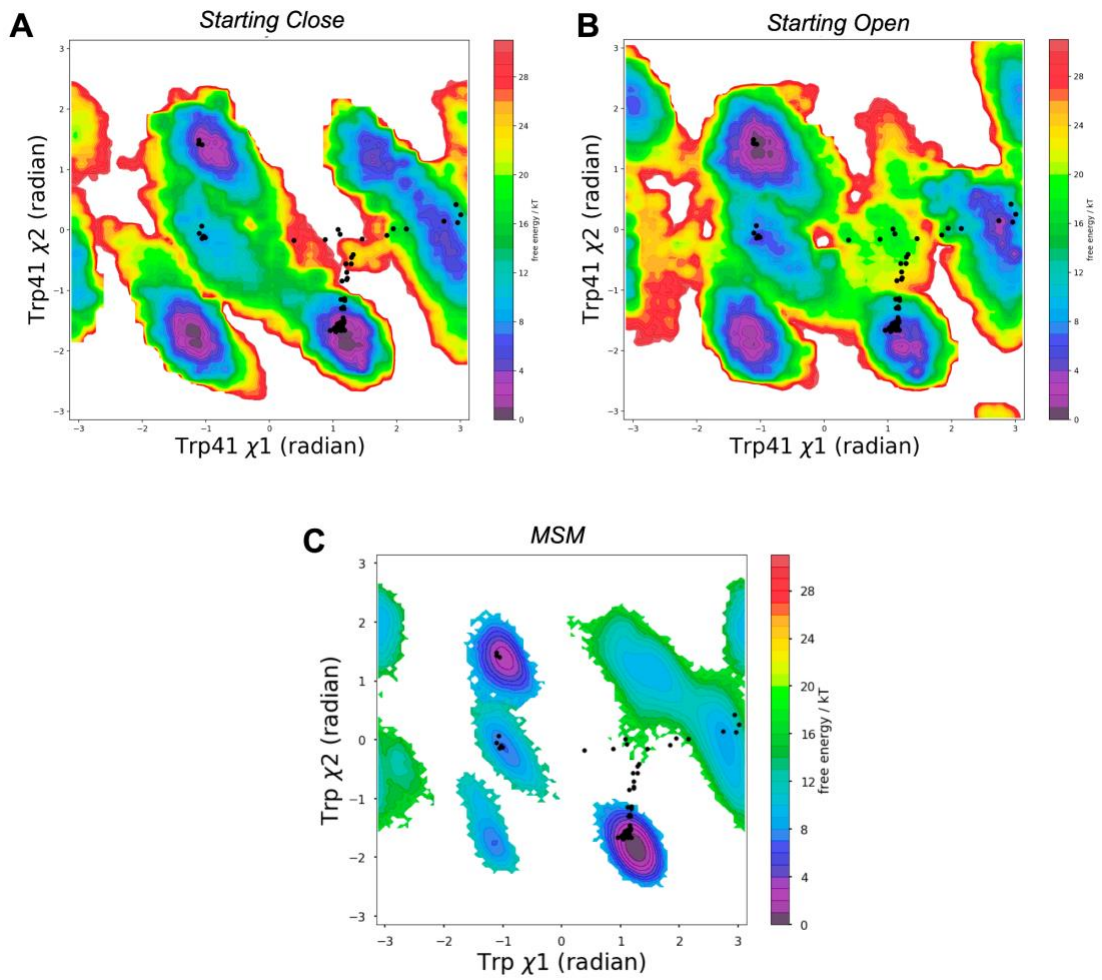

Figure 2. Metadynamics simulations using first two slow features as CVs managed to capture Trp41 flipping necessary for cryptic pocket opening in plasmepsin II. Reweighted free energy surface from metadynamics, starting from close (A) and open (B) states of plasmepsin II yields a similar landscape when compared with Markov state model approach (C). It is important to note that the simulation length for the metadynamics simulations is ~400ns each whereas the Markov state model was performed using an aggregate of 16 μs of simulation data. AlphaFold generated structural ensembles are highlighted in *black dots*.

Metadynamics also captured flipping of Tyr77 (Figure S6 in Supporting Information) which is a key residue governing the flap 'opening' in plasmepsin II. Opening of the flap in conjunction with Trp41 flipping exposes the 'deep' cryptic pocket (Figure 3).

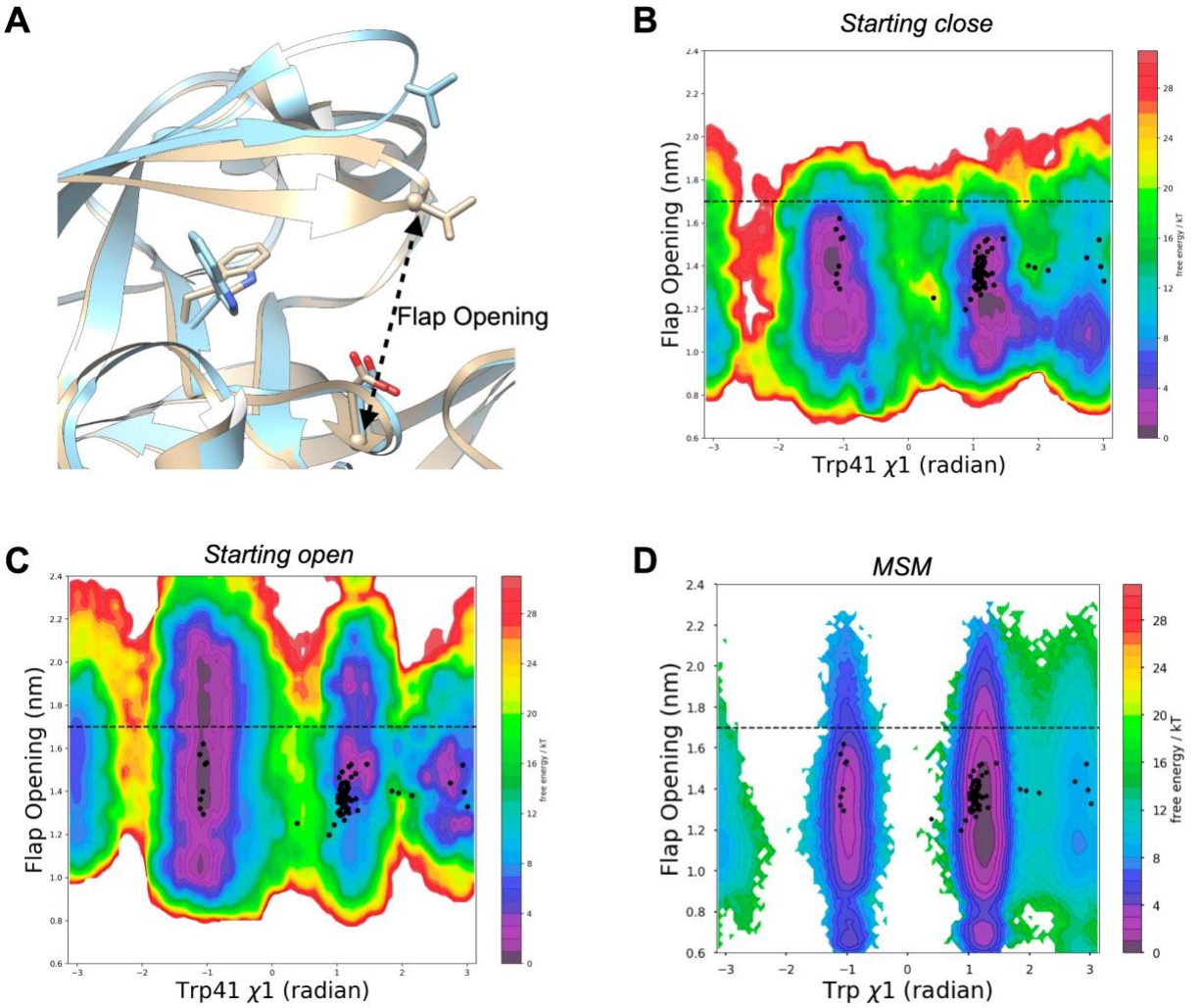

Figure 3. SFA-metadynamics managed to capture a deep cryptic pocket opening in plasmepsin II. (A) Flap opening is defined by the *Cα*--*Cα* distance between Asp34 and Val78. Flap opening in conjunction with Trp41 flipping (blue) exposes a deep cryptic pocket in plasmepsin II when compared with close apo conformation (silver). (B, C) Metadynamics simulations starting from close and open apo-like states managed to sample 'deep' cryptic pocket opening (Flap opening distance > 1.7 nm) within ~400ns when compared with MSM (total ~16μs of aggregate simulation). It is important to highlight that the AlphaFold generated ensemble (black dots) failed to sample a deep cryptic pocket opening in plasmepsin II.

Metadynamics using SFA CVs also managed to capture an alternate flipped state of Trp41 which has not been sampled by AlphaFold. Such a state has been captured by holo crystal structure of plasmepsin II, PDB: 4Z22[35] (Figure S7 in Supporting information).

Discussion

Stochastic subsampling of MSA allowed AlphaFold to sample a structural ensemble of plasmepsin II with conformational diversity; however, it failed to sample the 'deep' cryptic pocket opening (Figure 3). Well-tempered metadynamics simulations with first two slow features as CVs managed to capture 'deep' cryptic pocket opening in plasmepsin II within a factor of total simulation length when compared to MSM based approach. Deep cryptic pocket opening is a rare event in plasmepsin II which is a combination of flipping of Trp41 and flap opening. Recent works highlighted how we can combine structural ensembles generated by AlphaFold with autoencoder framework[36] and MSM[12] to predict Boltzmann distribution and capture rare events. The enhanced sampling strategy we propose, termed SFA-metadynamics, augments the toolkit of approaches aiming to extract the Boltzmann distribution from structural ensembles generated by AlphaFold. This strategy is particularly effective to capture rare transitions, such as opening of the 'deep' cryptic pocket in plasmepsin II. SFA in principle is similar to time-structure based independent component analysis (tICA)[37]. However, SFA is based on a straightforward idea: extract uncorrelated output signals that are ordered by slowness. This allows SFA to capture slowly varying features from high-dimensional and noisy temporal data generated by AlphaFold seed molecular dynamics simulations. SFA can also act as a reduced space on which one can build MSM. Further, conformations which contribute to the first few slow features can be used as seeds to launch small MD simulations. These MD simulations can be stitched together using MSM to predict thermodynamics and kinetics associated with conformational transitions. Kinetics extracted from MSM built on top of SFA can be directly compared with infrequent metadynamics[38] with slow features as CVs. In essence, SFA emerges as an innovative technique that bridges the gap in the estimation of kinetics linked to rare events, whether one uses MSM or metadynamics.

Conclusions

In this working paper, we innovatively integrate the Slow Feature Analysis (SFA) learned from simulations seeded by AlphaFold, together with metadynamics, to explore the cryptic pocket opening in plasmepsin II. Through the synergistic application of SFA-metadynamics, we were able to detect multiple transitions between the closed and open states of plasmepsin II in just a few hundred nanoseconds. In contrast, traditional methods necessitate several microseconds of simulation data to construct a fully connected Markov state model. This work underscores a novel approach that merges the capabilities of AlphaFold, SFA, and metadynamics. It offers a robust framework to predict the Boltzmann distribution corresponding to conformational shifts necessary to capture rare events such as cryptic pocket opening and allosteric modulation in proteins. It is important to highlight that in order for methods such as SFA, deep-TICA[39], RAVE[40] to be useful, the training data should transiently sample conformational heterogeneity. Short molecular dynamics simulations launched from AlphaFold generated ensemble encompasses necessary conformational heterogeneity which can be extracted by SFA and used as CVs within enhanced

sampling framework such as metadynamics to efficiently sample rare conformational transitions which plays a key role in molecular recognition and conformational dynamics of biomolecules. We will enhance this working paper by illustrating how SFA trained on AlphaFold seeded molecular simulations in tandem with metadynamics can accelerate protein-ligand binding and elucidate novel allosteric mechanisms involved in protein-protein interactions.

Supporting Information

Supporting information contains sampling of unbiased molecular dynamics simulations, implied timescale plot and macrostate definition generated by PCCA+ clustering, time-trace of slow features during metadynamics, convergence plots for metadynamics simulations and free energy surface highlighting alternate Trp41 conformation.

Competing Interest Statement

Authors declare no competing interests

References

(1) Shoemaker, S. C.; Ando, N. X-Rays in the Cryo-Electron Microscopy Era: Structural Biology's Dynamic Future. *Biochemistry* **2018**, *57* (3), 277–285. https://doi.org/10.1021/acs.biochem.7b01031.

(2) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(3) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making Protein Folding Accessible to All. *Nat Methods* **2022**, *19* (6), 679–682. https://doi.org/10.1038/s41592-022-01488-1.

(4) Kuzmanic, A.; Bowman, G. R.; Juarez-Jimenez, J.; Michel, J.; Gervasio, F. L. Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. *Acc Chem Res* **2020**, *53* (3), 654–661. https://doi.org/10.1021/acs.accounts.9b00613.

(5) Kornev, A. P.; Taylor, S. S. Dynamics-Driven Allostery in Protein Kinases. *Trends Biochem Sci* **2015**, *40* (11), 628–647. https://doi.org/10.1016/j.tibs.2015.09.002.

(6) Valsson, O.; Tiwary, P.; Parrinello, M. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu Rev Phys Chem* **2016**, *67* (1), 159–184. https://doi.org/10.1146/annurev-physchem-040215-112229.

(7) Bhakat, S. Collective Variable Discovery in the Age of Machine Learning: Reality, Hype and Everything in Between. *RSC Adv* **2022**, *12* (38), 25010–25024. https://doi.org/10.1039/D2RA03660F.

(8) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99* (6), 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011.

(9) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-Timescale Molecular Dynamics Simulations of Protein Structure and Function. *Curr Opin Struct Biol* **2009**, *19* (2), 120–127. https://doi.org/https://doi.org/10.1016/j.sbi.2009.03.004.

(10) del Alamo, D.; Sala, D.; Mchaourab, H. S.; Meiler, J. Sampling Alternative Conformational States of Transporters and Receptors with AlphaFold2. *Elife* **2022**, *11*, e75751. https://doi.org/10.7554/eLife.75751.

(11) Stein, R. A.; Mchaourab, H. S. SPEACH_AF: Sampling Protein Ensembles and Conformational Heterogeneity with Alphafold2. *PLoS Comput Biol* **2022**, *18* (8), e1010483.

(12) Meller, A.; Bhakat, S.; Solieva, S.; R. Bowman, G. Accelerating Cryptic Pocket Discovery Using AlphaFold. *J Chem Theory Comput* **2023**, *19* (14), 4355–4363. https://doi.org/10.1021/acs.jctc.2c01189.

(13) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J Am Chem Soc* **2018**, *140* (7), 2386–2396. https://doi.org/10.1021/jacs.7b12191.

(14) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know about Markov State Models but Were Afraid to Ask. *Methods* **2010**, *52* (1), 99–105. https://doi.org/https://doi.org/10.1016/j.ymeth.2010.06.002.

(15) Suárez, E.; P. Wiewiora, R.; Wehmeyer, C.; Noé, F.; D. Chodera, J.; M. Zuckerman, D. What Markov State Models Can and Cannot Do: Correlation versus Path-Based Observables in Protein-Folding Models. *J Chem Theory Comput* **2021**, *17* (5), 3119–3133. https://doi.org/10.1021/acs.jctc.0c01154.

(16) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *WIREs Computational Molecular Science* **2011**, *1* (5), 826–843. https://doi.org/https://doi.org/10.1002/wcms.31.

(17) Bussi, G.; Laio, A. Using Metadynamics to Explore Complex Free-Energy Landscapes. *Nature Reviews Physics* **2020**, *2* (4), 200–212. https://doi.org/10.1038/s42254-020-0153-0.

(18) Wiskott, L.; Sejnowski, T. J. Slow Feature Analysis: Unsupervised Learning of Invariances. *Neural Comput* **2002**, *14* (4), 715–770. https://doi.org/10.1162/089976602317318938.

(19) Bhakat, S. Pepsin-like Aspartic Proteases (PAPs) as Model Systems for Combining Biomolecular Simulation with Biophysical Experiments. *RSC Adv* **2021**, *11* (18), 11026–11047. https://doi.org/10.1039/D0RA10359D.

(20) Bhakat, S.; Söderhjelm, P. Flap Dynamics in Pepsin-Like Aspartic Proteases: A Computational Perspective Using Plasmepsin-II and BACE-1 as Model Systems. *J Chem Inf Model* **2022**, *62* (4), 914–926. https://doi.org/10.1021/acs.jcim.1c00840.

(21) Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat Biotechnol* **2017**, *35* (11), 1026–1028. https://doi.org/10.1038/nbt.3988.

(22) Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I. Y.; Berryman, J. T.; Brozell, S. R.; Cerutti, D. S.; Cheatham, T. E. I.; Cisneros, G. A.; Cruzeiro, V. W. D.; Darden, T. A.; Forouzesh, N.; Giambaşu, G.; Giese, T.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Harris, J.; Izadi, S.; Izmailov, S. A.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kovalenko, A.; Kurtzman, T.; Lee, T.

S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Machado, M.; Man, V.; Manathunga, M.; Merz, K. M.; Miao, Y.; Mikhailovskii, O.; Monard, G.; Nguyen, H.; O'Hearn, K. A.; Onufriev, A.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shajan, A.; Shen, J.; Simmerling, C. L.; Skrynnikov, N. R.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wang, J.; Wei, H.; Wu, X.; Wu, Y.; Xiong, Y.; Xue, Y.; York, D. M.; Zhao, S.; Zhu, Q.; Kollman, P. A. Amber 2022, University of California, San Francisco. **2022**.

(23) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J Chem Theory Comput* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

(24) Sousa da Silva, A. W.; Vranken, W. F. ACPYPE - AnteChamber PYthon Parser InterfacE. *BMC Res Notes* **2012**, *5* (1), 367. https://doi.org/10.1186/1756-0500-5-367.

(25) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J Appl Phys* **1981**, *52* (12), 7182–7190. https://doi.org/10.1063/1.328693.

(26) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J Chem Phys* **1993**, *98* (12), 10089–10092. https://doi.org/10.1063/1.464397.

(27) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J Comput Chem* **1997**, *18* (12), 1463–1472. https://doi.org/https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

(28) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/https://doi.org/10.1016/j.softx.2015.06.001.

(29) Röblitz, S.; Weber, M. Fuzzy Spectral Clustering by PCCA+: Application to Markov State Models and Data Classification. *Adv Data Anal Classif* **2013**, *7* (2), 147–179. https://doi.org/10.1007/s11634-013-0134-6.

(30) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J Chem Theory Comput* **2015**, *11* (11), 5525–5542. https://doi.org/10.1021/acs.jctc.5b00743.

(31) Asojo, O. A.; Gulnik, S. V; Afonina, E.; Yu, B.; Ellman, J. A.; Haque, T. S.; Silva, A. M. Novel Uncomplexed and Complexed Structures of Plasmepsin II, an Aspartic Protease from Plasmodium Falciparum. *J Mol Biol* **2003**, *327* (1), 173–181. https://doi.org/https://doi.org/10.1016/S0022-2836(03)00036-6.

(32) Prade, L.; Jones, A. F.; Boss, C.; Richard-Bildstein, S.; Meyer, S.; Binkert, C.; Bur, D. X-Ray Structure of Plasmepsin II Complexed with a Potent Achiral Inhibitor *. *Journal of Biological Chemistry* **2005**, *280* (25), 23837–23843. https://doi.org/10.1074/jbc.M501519200.

(33) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. PLUMED: A Portable Plugin for Free-Energy Calculations with Molecular Dynamics. *Comput Phys Commun* **2009**, *180* (10), 1961–1972. https://doi.org/https://doi.org/10.1016/j.cpc.2009.05.011.

(34) Tiwary, P.; Parrinello, M. From Metadynamics to Dynamics. *Phys Rev Lett* **2013**, *111* (23), 230602. https://doi.org/10.1103/PhysRevLett.111.230602.

(35) Rasina, D.; Otikovs, M.; Leitans, J.; Recacha, R.; Borysov, O. V; Kanepe-Lapsa, I.; Domraceva, I.; Pantelejevs, T.; Tars, K.; Blackman, M. J.; Jaudzems, K.; Jirgensons, A. Fragment-Based Discovery of 2-Aminoquinazolin-4(3H)-Ones As Novel Class Nonpeptidomimetic Inhibitors of the Plasmepsins I, II, and IV. *J Med Chem* **2016**, *59* (1), 374–387. https://doi.org/10.1021/acs.jmedchem.5b01558.

(36) P. Vani, B.; Aranganathan, A.; Wang, D.; Tiwary, P. AlphaFold2-RAVE: From Sequence to Boltzmann Ranking. *J Chem Theory Comput* **2023**, *19* (14), 4351–4354. https://doi.org/10.1021/acs.jctc.3c00290.

(37) Blaschke, T.; Berkes, P.; Wiskott, L. What Is the Relation Between Slow Feature Analysis and Independent Component Analysis? *Neural Comput* **2006**, *18* (10), 2495–2508. https://doi.org/10.1162/neco.2006.18.10.2495.

(38) Ray, D.; Parrinello, M. Kinetics from Metadynamics: Principles, Applications, and Outlook. *J Chem Theory Comput* **2023**, *19* (17), 5649–5670. https://doi.org/10.1021/acs.jctc.3c00660.

(39) Bonati, L.; Piccini, G.; Parrinello, M. Deep Learning the Slow Modes for Rare Events Sampling. *Proceedings of the National Academy of Sciences* **2021**, *118* (44), e2113533118. https://doi.org/10.1073/pnas.2113533118.

(40) Ribeiro, J. M. L.; Bravo, P.; Wang, Y.; Tiwary, P. Reweighted Autoencoded Variational Bayes for Enhanced Sampling (RAVE). *J Chem Phys* **2018**, *149* (7), 72301. https://doi.org/10.1063/1.5025487.