JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

Machine Learning Derived Collective Variables for the Study of Protein Homodimerization in Membrane

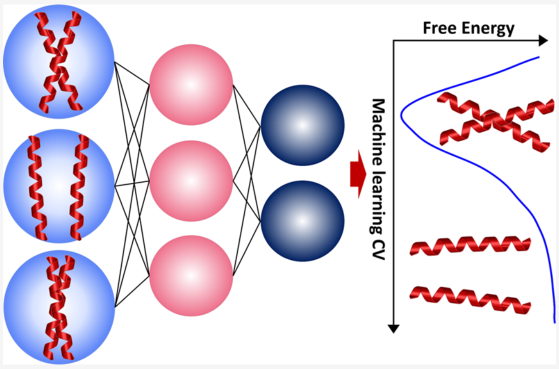
Published as part of Journal of Chemical Theory and Computation virtual special issue "Machine Learning and Statistical Mechanics: Shared Synergies for Next Generation of Chemical Theory and Computation".

Ayan Majumder and John E. Straub*

Cite This: *J. Chem. Theory Comput.* 2024, 20, 5774−5783 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The accurate calculation of equilibrium constants for protein−protein association is of fundamental importance to quantitative biology and remains an outstanding challenge for computational biophysics. Traditionally, equilibrium constants have been computed from one-dimensional free energy surfaces derived from sampling along a single collective variable. Importantly, recent advances in enhanced sampling methodology have facilitated the characterization of multidimensional free energy landscapes, often exposing multiple thermodynamically important minima missed by more restrictive sampling methods. A key to the effectiveness of this multidimensional sampling approach is the identification of collective variables that effectively define the configurational space of dissociated and associated states. Here we present the application of two machine learning methods for the unbiased determination of collective variables for enhanced sampling and analysis of protein−protein association. Our results both validate prior work, based on intuition derived collective variables, and demonstrate the effectiveness of the machine learning methods for the identification of collective variables for association reactions in complex biomolecular systems.



INTRODUCTION

Membrane proteins represent an important class of biomolecules that are essential to cellular organization and function. Most membrane proteins contain one or more transmembrane (TM) regions that span the membrane bilayer.[1,2] Membrane proteins also play crucial roles in cellular signaling with G-protein coupled receptors representing common therapeutic targets.[3] Association of membrane associated enzymes and their substrates through the interaction of the TM domains is known to play a key role in the biogenesis of a diverse array of proteins including the amyloid-β (Aβ) protein in Alzheimer's disease.[4,5] As a result, the association of TM regions of the protein in membrane has received significant attention in experiment, theory, and simulations for over four decades.[6−10] Nevertheless, there remain fundamental outstanding questions related to the nature of the associated state and the magnitude of related equilibrium association constants.

Molecular dynamics studies have been widely used to understand the structure and function of single-pass TM proteins. However, slow translational and rotational diffusion of even single-pass TM domains in membrane imposes a significant challenge to the effective sampling of thermodynamically relevant structures.[11] Enhanced sampling methods including umbrella sampling,[12,13] metadynamics,[14−16] and adaptive biasing forces[17] have been extensively used to explore the free energy landscape defining TM protein dimerization using both all-atom and coarse-grained models.[13,16,18−20] In addition, similar approaches have been used to understand protein−lipid interactions in complex membrane bilayers.[20] A converged free energy surface can provide insight into the nature of the dimer state ensemble and can be used to calculate the association binding constant, which can be quantitatively compared to experimental data. However, the effectiveness of these enhanced sampling methods depends on the choice of collective variables (CVs) that effectively differentiate not only monomer and dimer but also competing dimer substates.

In the study of protein homodimerization, commonly used CVs are the center-of-mass (COM) distance between the TM

Received: April 5, 2024
Revised: May 29, 2024
Accepted: June 14, 2024
Published: June 25, 2024

https://doi.org/10.1021/acs.jctc.4c00454

helices (D_{com}) and the distance root-mean-square displacement from a reference dimer structure (D_{rmsd}).[19−22] However, sampling over one-dimensional CVs can fail to sample all thermodynamically important structures in the dimer state ensemble. Sampling along D_{com} and D_{rmsd} using umbrella sampling can result in restrictive sampling that leads to overestimation of the binding free energy.[13] Two-dimensional umbrella sampling along the *xy*-transverse components of the center-of-mass distance between two TM helices has been shown to more effectively sample relevant dimer substates.[13] However, this approach can be computationally expensive and challenging to implement.

To address this shortcoming, we have recently proposed a metadynamics protocol to study the homodimerization of TM helices using three collective variables: the *x*-projection and *y*-projection of the center-of-mass distance between the two TM helices and the interhelical crossing angle.[16] Through comparison with the results of exhaustive unbiased sampling, we demonstrated that the proposed 3D metadynamics approach can effectively sample the free energy landscape characterizing the monomer and dimer state ensembles, leading to converged estimates of the dimerization free energy and associated equilibrium constant.

The choice of collective variables to study rare events in a complex biological system can be challenging. A large class of CVs can be defined as functions of atomic coordinates providing a coarse-grained representation of the slow modes and essential coordinates defining the biological process under study.[23,24] Recent application of machine learning methods to the identification of CVs has significantly improved the effectiveness of enhanced sampling methods.[25−29] Parrinello and co-workers employed the deep linear discriminant analysis (DeepLDA) approach to identify collective variables for sampling a rugged free energy surface connecting multiple metastable states using metadynamics.[26] The DeepLDA method uses a deep neural network to perform a nonlinear transformation of the input descriptors, optimized by linear discriminant analysis (LDA), to identify CVs.[30] The DeepLDA protocol offers an improvement to alternative LDA approaches such as harmonic linear discriminant analysis (HLDA), which relies on the identification of linearly separable input descriptors.[31] Tiwary and co-workers have developed the state predictive information bottleneck (SPIB) method to interpret high-dimensional molecular dynamics simulation data in the study of rare events.[27] The SPIB approach identifies collective variables of the system through the analysis of the input descriptors obtained from molecular dynamics trajectories using time delay as a hyperparameter. Collective variables obtained using the SPIB method have been used to effectively sample the left- to right-handed chirality transition of synthetic peptides and to study the permeation of small molecules through membrane bilayers.[32]

In this study, we employed the DeepLDA and SPIB methods to identify CVs for enhanced sampling simulations of TM protein homodimerization. Collective variables were derived from extensive unbiased simulations using the DeepLDA and SPIB methods. Well-tempered metadynamics was then applied using the learned collective variables identified through each approach to study the dimerization equilibrium. We compared the results obtained from the simulations using DeepLDA and SPIB derived CVs to those obtained from our 3D metadynamics protocol. Our results demonstrate that DeepLDA and SPIB methods are powerful tools for identifying collective variables for enhanced sampling simulation and analysis of complex biomolecular systems.

METHODS

Molecular Dynamics Simulation Model and Methods.

Two transmembrane proteins, glycophorin A (GpA) and WALP23, were studied. The structure of GpA was taken from the PDB ID 1AFO.[33] Residues 69−97 of GpA were placed in a membrane bilayer containing 404 POPC lipids. WALP23 was studied in a membrane bilayer consisting of 406 POPC lipids. All lipid bilayers were solvated with 12 MARTINI v3 water beads per lipid, and a 0.15 M NaCl concentration was used. The temperature of the simulations was maintained at 310 K using a velocity rescaling thermostat, and the pressure was maintained at 1 bar using the semi-isotropic Parrinello−Rahman barostat. The MARTINI v3 force field was used to simulate the protein-embedded membrane bilayers using the recommended sets of parameters.[34]

Defining the Input Descriptors.

To train the neural network (NN) models, input descriptors were chosen based on the TM homodimer conformations. In an α-helix, any pair of residues R and $R + 4$ represent the same face of interaction. For example, it is well-known that the GXXXG motif of GpA plays an important role in governing dimer structures where the two Gly residues appear on a common face of the helix, facilitating helix−helix interactions.[13] Building on this insight, we defined four points on the α-helix as the center of mass of the backbone beads of residues R and $R + 4$, $R + 1$ and $R + 5$, $R + 2$ and $R + 6$, and $R + 3$ and $R + 7$. For GpA and WALP23, the residue R was chosen to be 79 and 6, respectively. A total of 16 interhelical distances were defined by these four points on each helix, and the crossing angle between two helices was used as input descriptors to represent each frame of the trajectory. A pictorial representation of all 17 descriptors is shown in Supplementary Figure 1.

Identifying Collective Variables Using the DeepLDA Protocol.

LDA is widely used as a dimensionality reduction method for a classification problem, which seeks to find a linear combination of the input descriptors that separates the data into different classes. We take $x_1, x_2, ..., x_N = X$ to be a set of N samples containing d descriptors values. The objective of LDA is to find a linear projection of the descriptors A, where XA^T is maximally separable. The projection matrix A is chosen to maximize the ratio

$$\frac{AS_{\text{b}}A^T}{AS_{\text{w}}A^T} \tag{1}$$

where S_{w} is the within class scatter matrix and S_{b} is the between class scatter matrix defined through the total scatter matrix S_{t}.

$$S_{\text{t}} = \frac{1}{N-1}\bar{X}^T\bar{X} \tag{2}$$

where $S_{\text{t}} = S_{\text{b}} + S_{\text{w}}$. A generalized eigenvalue problem can be solved to find the solution for eq 1 in the form

$$S_{\text{b}}e_i = \nu_i S_{\text{w}} e_i \tag{3}$$

for $i = 1, ..., d$, the eigenvectors e_i form the desired projection matrix A.

We studied the thermodynamic properties of two classes, namely, bound and unbound states of TM homodimers, using the DeepLDA method. The input descriptors of the system were fed into a neural network model with θ parameters to obtain a nonlinear transformation represented by $\mathcal{N}_\theta(X)$. The representation obtained from the topmost hidden layer $h = \mathcal{N}_\theta(X)$ is then used to perform LDA (Figure 1a). The loss function for the two-class classification problem is defined as

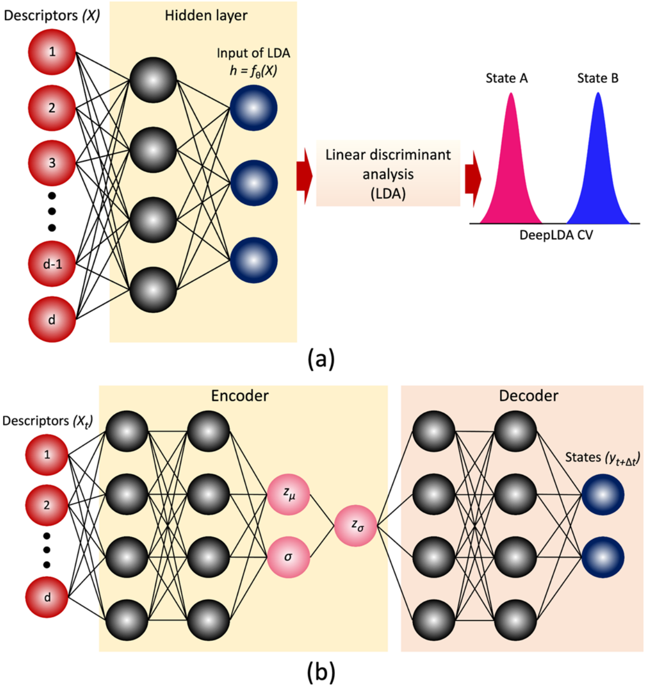


Figure 1. Schematic representation of the construction of collective variables using the (a) DeepLDA and (b) SPIB protocols. Seventeen descriptors, represented by the red circles, were used to define a homodimer configuration and to train the NN models.

$$\mathcal{L}_{\text{LDA}} = -\nu_1 \tag{4}$$

where the eigenvalue ν_1 corresponds to the amount of discriminative separation along the eigenvector e_1. Another term was added to the loss function to regulate the separation between two classes (see the Supporting Information). The resulting DeepLDA CV can be calculated as $S = he_1^T$. The DeepLDA code was adapted from the work of Bonati, Rizzi, and Parrinello.[26]

Identifying Collective Variables Using the SPIB Protocol. SPIB uses an information bottleneck approach that seeks a concise representation z based on the input descriptors X providing maximum information about the target y (Figure 1b). A model can be trained from the initial state labels by using the SPIB protocol, which is iteratively refined to obtain the final state representation. In practice, a nonlinear NN model is trained by feeding input descriptors to predict the information bottleneck z, which can be used to identify the collective variables for use in enhanced sampling simulations and analysis. Then, another NN model is used as a decoder, predicting the future state label y based on the value of z. The goal of the information bottleneck approach is to predict the maximum information about the target while retaining the minimum information about the initial descriptors.

An unbiased trajectory can be represented by descriptors x_1, x_2, ..., x_{M+s} and initial state labels y_1, y_2, ..., y_{M+s}, where x and x_s are the values of descriptors at an arbitrary time t and $t + \Delta t$. For large trajectory data, a model can be trained using the SPIB protocol by maximizing the objective function

$$\mathcal{L}_{\text{SPIB}} = \frac{1}{M}\sum_{n=1}^{M}\left[\log q(y^{n+s}|z^n) - \beta \log \frac{P(z^n|x^n)}{P(z^n)}\right] \tag{5}$$

where the first term of the objective function reflects the prediction capability of the desired output state and the second term can be interpreted as a regularizer of the information from x^n to z^n. The trade-off between these two terms can be controlled by the parameter β. The terms $q(y^{n+s}|z^n)$, $P(z|x)$, and $P(z)$ are calculated through deep NN models. The SPIB code employed in this study was adapted from the work of Wang and Tiwary.[27]

3D Metadynamics Enhanced Sampling Simulations. Well-tempered metadynamics was employed to study the homodimerization of TM helices using three collective variables (3D metadynamics), defined as the x-projection and y-projection of the center-of-mass distance between two helices and the crossing angle between two helices. A detailed description of the 3D metadynamics protocol was provided in a previous study.[16]

Well-tempered metadynamics[35] was also performed along the collective variables developed using DeepLDA and SPIB protocols. A Gaussian of height 0.1 kJ/mol was added along the CVs after every 2000 steps using a multiple walker approach. A biasfactor value of 10 was used. All simulations were performed using GROMACS v2021,[36] patched with PLUMED v2.9.[37] The NN models were developed using PyTorch,[38] and metadynamics along the CVs developed by the NN models was performed using the PLUMED interface for PyTorch. A flowchart illustrating the workflow used in this study is shown in Supplementary Figure 2.

RESULTS AND DISCUSSION

We studied the homodimerization of GpA (SEPEITLIIFG-VMAGVIGTILLISYGIRR) and WALP23 (GWWLALALALA-LALALALALWWA) using metadynamics using the collective variables derived from the machine learning based protocols. Dimerization kinetics of GpA and WALP23 have been extensively characterized by using computational and experimental studies. These systems have been frequently used to assess new enhanced sampling approaches because of the wealth of data that is available. The effectiveness of the enhanced sampling methods is dependent on the choice of collective variables. A proper choice of collective variables should facilitate the sampling of both native and non-native configurations of the homodimer, leading to a converged estimation of the association free energy.

Generating the Training Data Sets. To use machine learning based collective variables, it is important to train a NN model by including both native and non-native dimer structures in the training data sets. Unbiased simulations of GpA and WALP23 in POPC were performed to build the training data sets. A total of 25 and 20 independent trajectories of length 5 μs were performed for GpA and WALP23, respectively. All simulations were started by placing the helices at least 2.5 nm apart to study the spontaneous aggregation of the homodimer. The population density obtained from the unbiased simulations was projected onto the crossing angle and D_{com} space (Supplementary Figure 3). Many of the trajectories were found to be stuck in a single energy minimum, so the potential of mean force (PMF) associated

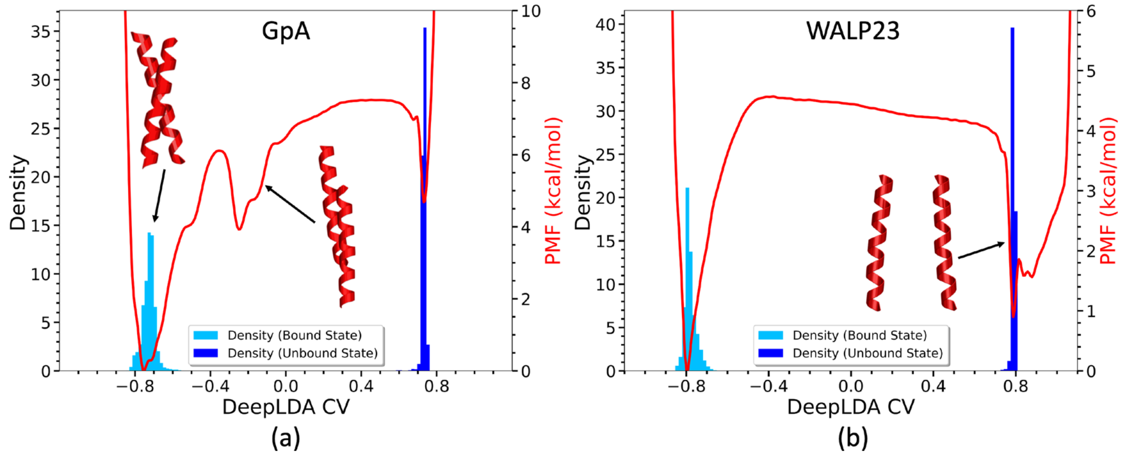


Figure 2. Potential of mean force (PMF) for the association of (a) GpA and (b) WALP23 as a function of DeepLDA CV. The minimum value of the PMF was set to zero. Density distributions of the bound and unbound state training data points along the DeepLDA CV are shown as a histogram.

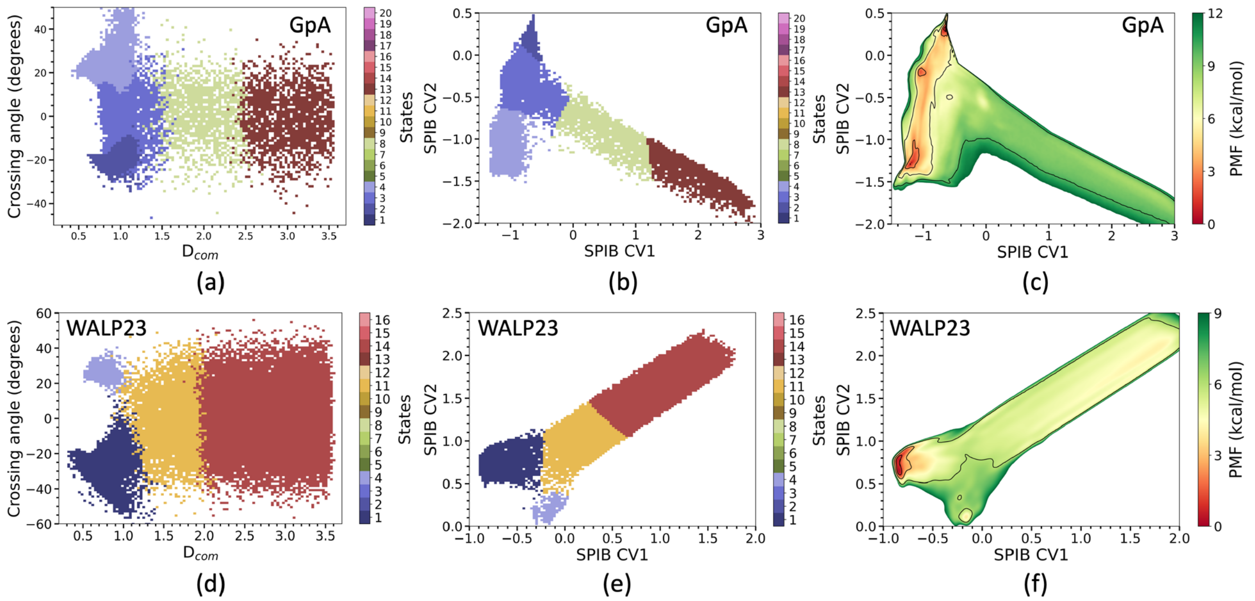


Figure 3. SPIB derived converged state representations for the homodimerization of GpA and WALP23 projected onto the (a, d) crossing angle and D_{com} and (b, e) CV space. Potential of mean force along the 2D CV representing the association of (c) GpA and (f) WALP23. The minimum value of the PMFs was set to zero to guide the comparison.

with the dimerization of the homodimers was not calculated. However, we characterized the bound state structural ensembles for both GpA and WALP23. The structural ensemble obtained from the unbiased simulations was found to be similar to that obtained from the 3D metadynamics simulations.[16] This finding suggests that the concatenated trajectory acquired from the independent, unbiased simulations contains information about those metastable states essential to the description of the association equilibrium for GpA and WALP23.

One of the major characteristics of the homodimer structures is the crossing angle distribution between two helices (Supplementary Figure 3). Previous studies have shown that the thermodynamic origin of different crossing angle distributions can be identified as a specific interaction hotspot on the *xy*-projection of the COM−COM distances between two helices. We chose four points that can be associated with different interaction sites on the helix. Sixteen interhelical distances were defined by the four points on each helix, and the interhelical crossing angle was chosen as descriptors for each homodimer configuration (Methods).

Application of the DeepLDA Protocol. The DeepLDA protocol has been successfully used to derive collective variables to study rare events in biomolecules.[39,40] We used the DeepLDA protocol to study the dimerization of proteins in the membrane. Data obtained from the unbiased simulations were used to train the NN models. Two classes, namely, bound and unbound states, were defined to derive a one-dimensional collective variable connecting the monomeric and dimeric states of the protein. For GpA and WALP23, the bound state structures were chosen if the D_{com} value was less than 1.2 and 1 nm, respectively, and the unbound state structures were chosen if the D_{com} value was greater than 2.5 nm. The histogram of the training data points projected onto the DeepLDA derived CV is shown in Figure 2. Well-tempered metadynamics was performed using the DeepLDA derived CV to study the

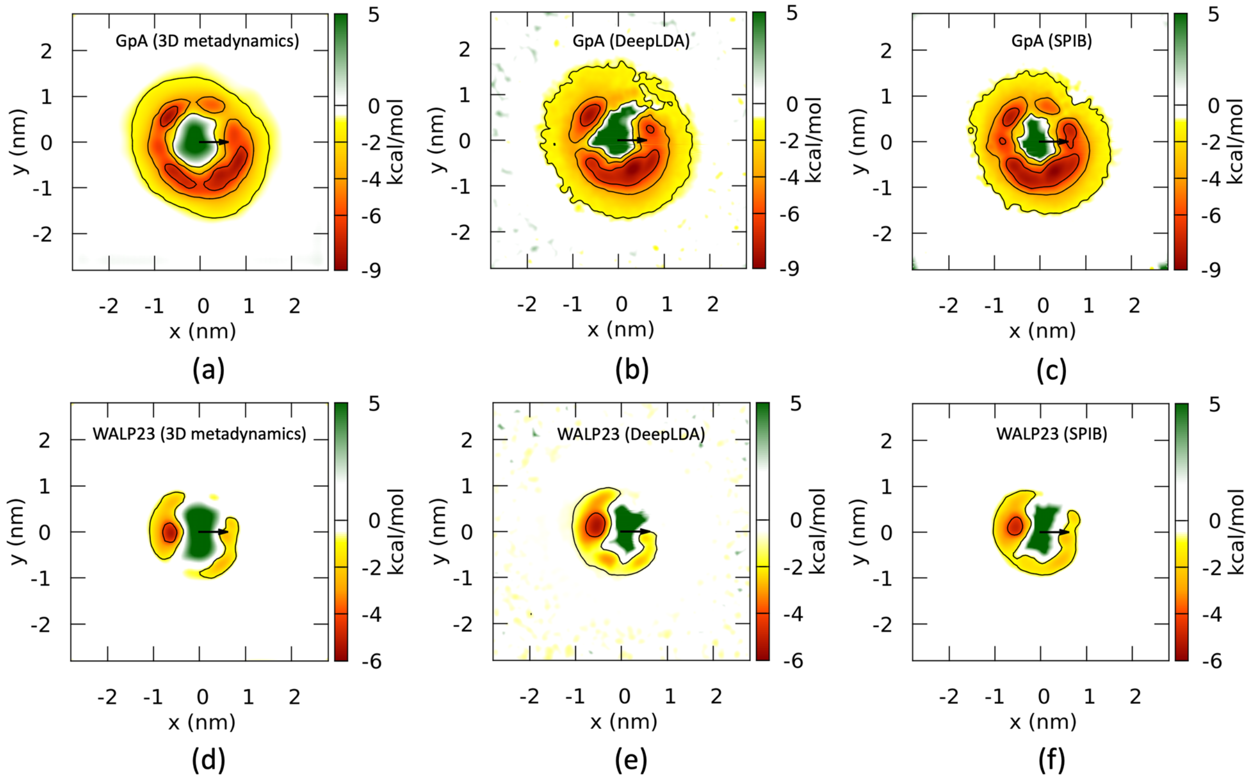


Figure 4. Potential of mean force along the x-projection and y-projection of the COM−COM distances between two helices of GpA and WALP23 obtained from (a, d) the 3D metadynamics and simulations along the CVs derived from the (b, e) DeepLDA and (c, f) SPIB protocols. The COM of one helix was centered on the xy-plane. COM^{79}-to-Gly^{79} vector of GpA and COM^{11}-to-Ala^{11} of WALP23 are represented by the black arrow. COM^{X} of a helix was defined as the center of mass of $X - 2$ to $X + 2$ residues. Panels (a) and (d) were derived from the previous study.[16]

dimerization free energy surfaces of GpA and WALP23. The PMFs obtained from the metadynamics simulations are shown in Figure 2. For GpA, three wells in the PMF were found along the CV, and in the case of WALP23, two wells were observed. The well near the bound state population of the protein was observed due to the stable dimeric forms of the homodimers. The well near the unbound state population of the homodimers arises due to the excess area available to the dimer (Supplementary Figure 4). For GpA, the well that is observed at −0.2 along the CV cannot be separated from the bound state well along the lateral COM−COM distance (D_{com}) of the helices, but it is separable if we consider the z movement of the helices (Supplementary Figure 4). Thus, the origin of that well can be referred to as a metastable state of GpA that arises due to the movement of the helices along the membrane normal.

Application of the SPIB Protocol. The SPIB protocol was used to derive a 2D collective variable describing the dimerization of GpA and WALP23. Dimer configurations obtained from the unbiased simulations were used to train neural network (NN) models. The SPIB protocol uses a time delay (Δt) as a hyperparameter to incorporate information from states at time t in predicting the state at time $t + \Delta t$. We utilized all independent unbiased trajectories separately to extract state information after a time delay of Δt. Based on the values of the crossing angle and D_{com} of the homodimer configurations, we defined 20 and 16 initial states for GpA and WALP23, respectively (Supplementary Figure 5). The encoder and decoder models were then trained to obtain the converged populations of the final states (Figure 3). The projection of the initial data points onto the 2D collective variables derived from the SPIB protocol is shown in Figure 3. The final state populations suggest that the SPIB protocol can retain complete information about the stable dimeric states observed in the unbiased simulations. Well-tempered metadynamics simulations were performed using the two collective variables derived from the SPIB protocol. The PMFs along the SPIB derived CVs obtained for the homodimerization of GpA and WALP23 are shown in Figure 3. The convergence of PMFs along the DeepLDA and SPIB derived CVs was assessed by projecting the PMF onto D_{com} (Supplementary Figure 6).

Extending CVs Using DeepLDA and SPIB. We further characterized the importance of all descriptors in deriving collective variables (CVs) using DeepLDA and SPIB protocols (Supplementary Figures 7 and 8). The importance of a descriptor was calculated by summing the weight of that descriptor in the first hidden layer of the NN models. The results indicate that the crossing angle distribution plays a dominant role in defining a collective variable using the SPIB protocol. This trend is predictable, as many final state representations obtained from the SPIB protocol are separable along the crossing angle space. In contrast, weights of the distance descriptors were found to be similar. In the case of CVs derived from the DeepLDA protocol, no specific trends in the descriptor weights were observed.

Density along the x-projection and y-projection of the COM−COM distances between two helices carries important information about the interaction sites on the helix. All of the dimers exhibit a unique hotspot on the xy-plane. We have previously demonstrated that it is possible to refine a PMF along the xy-plane, leading to different handedness properties of the homodimer.[16]

Table 1. Dimerization Free Energy of GpA and WALP23 Homodimers Reported in Previous Studies

protein/sequence	methods/force field	collective variables	free energy of dimerization (kcal/mol)	medium
GpA				
residues 73−95	CHARMM27	D_{com}	11.5[47]	dodecane
residues 69−97	CHARMM36m	geometrical route	10.7[18]	POPC bilayer
residues 69−97	CHARMM36 (rescaled)	D_{rmsd}	3.0−3.8[19]	POPC bilayer
full TM domain	MARTINI v2	D_{com}	9.1[22]	DPPC bilayer
residues 69−97	MARTINI v2	D_{rmsd}	8.4[20]	POPC bilayer
residues 69−97	MARTINI v2	D_{rmsd}	9.3[13]	POPC bilayer
residues 69−97	MARTINI v2	2D US	7.5[13]	POPC bilayer
residues 69−97	MARTINI v3	D_{rmsd}	5.9[34]	POPC bilayer
TM domain	MARTINI v3	D_{com}	3.1[51]	DLPC bilayer
residues 69−97	MARTINI v3	3D metadynamics	7.0[16]	POPC bilayer
	experimental		3.4−12.1[41−46]	
WALP23				
	MARTINI v2	D_{com}	4.8[48]	DOPC bilayer
	MARTINI		2.9[49]	di-C18:2PC
	MARTINI v3	D_{com}	2.9[34]	POPC bilayer
	MARTINI v3	D_{com}	2.6[34]	DOPC bilayer
	MARTINI v3	3D metadynamics	3.2[16]	POPC bilayer
	MARTINI v3	3D metadynamics	2.6[16]	DOPC bilayer
$(AALALAA)_3$	experimental		3.0[50]	di-C18:1PC

We calculated a PMF as a function of the *xy*-projection of the COM−COM distances obtained from the simulations along CVs derived from the DeepLDA and SPIB protocols (Figure 4). One of the helices was centered on the *xy*-plane, and the COM^X-to-Res^X vector of the same helix was aligned with the positive *x*-axis. The population density of the other helix with respect to that of the centered helix is represented as a PMF on the *xy*-plane. Results obtained from the simulations along DeepLDA and SPIB derived CVs were found to be similar to those of the 3D metadynamics simulation. GpA showed multiple minima on the *xy*-plane, representing the presence of multiple interaction sites. Meanwhile, a predominant minimum was observed for WALP23, demonstrating a specific interaction pattern of the helix.

Comparison of Computed Association Free Energies with Experiment. Several experimental and computational studies have evaluated the dimerization free energies of GpA and WALP23. The experimental results for the free energy of dimerization of GpA vary from 3.4 to 12.1 kcal/mol.[41−46] Chipot and co-workers reported the free energy of dimerization of GpA to be 11.5 kcal/mol in dodecane using the all-atom CHARMM27 force field[47] and 10.7 kcal/mol in a POPC bilayer using the CHARMM36m force fiield.[18] Best and co-workers reparameterized the protein−lipid interaction of the CHARMM36 force field and reported the dimerization free energy of GpA to be 3.0−3.8 kcal/mol in a POPC bilayer.[19] The MARTINI coarse-grained force field has also been extensively used to study protein homodimerization in membrane bilayers. Using the MARTINI force field, Marrink and co-workers evaluated the dimerization free energy of GpA to be 9.1 kcal/mol in a DPPC bilayer.[22] Sansom and co-workers[20] and Straub and co-workers[13] used the D_{rmsd} collective variable to study GpA and reported dimerization free energies of 8.4 and 9.3 kcal/mol, respectively. However, a two-dimensional umbrella sampling simulation predicted the association free energy of the GpA homodimer to be 7.5 kcal/mol. Castillo et al. studied the WALP23 homodimer in DOPC and reported a dissociation free energy of 4.8 kcal/mol.[48] Marrink and co-workers studied the equilibrium association constant of WALP23 and found the associated dimerization free energy to be 2.9 kcal/mol.[49] An experimental study of $(AALALAA)_3$ using FRET spectroscopy reported the dimerization free energy to be 3 kcal/mol.[50] In the recent development of the MARTINI force field, Souza et al. reported the dimerization free energy of GpA in POPC and that of WALP23 in POPC to be 5.9 and 2.9 kcal/mol, respectively.[34] Using the 3D metadynamics protocol, we reported the free energy of dimerization of GpA and WALP23 in POPC to be 7.0 and 3.2 kcal/mol, respectively.[16] The dimerization free energies of GpA and WALP23 homodimers obtained from previous studies are presented in Table 1.

We calculated the association free energies of GpA and WALP23 from the simulations along DeepLDA and SPIB derived CVs. The probability densities of homodimer conformations were projected onto the lateral COM−COM distance of the helices (D_{com}) and are shown in Figure 5. An entropic factor $k_BT\ \ln(r)$ was added to the PMF to account for the excess area available to the dimer at larger D_{com} values. The binding constant was then calculated by integrating the PMF using[47,52−54]

$$K_D = \frac{1}{2} \times \frac{2\pi}{A_0} \int_0^{r_c} r\ e^{-\Delta W(r)/k_BT}\ dr$$

$$\Delta G = -k_BT\ \ln(K_D) \tag{6}$$

The cutoff distance r_c was chosen to be 2.3 nm to separate the dimeric and monomeric states of the homodimer. The PMFs were found to plateau after r_c and were set to zero. A reference area A_0 was chosen to be 1 nm^2. The dimerization free energies of the GpA and WALP23 homodimers obtained from this study are presented in Table 2. The free energy of dimerization obtained from the simulations along DeepLDA and SPIB derived CVs was found to be comparable to that obtained from the 3D metadynamics simulations.

Comparison of Computed Dimer Structures with Experiment. Structures of transmembrane protein homodimers have been extensively characterized by using experimental and computational studies. GpA structures derived

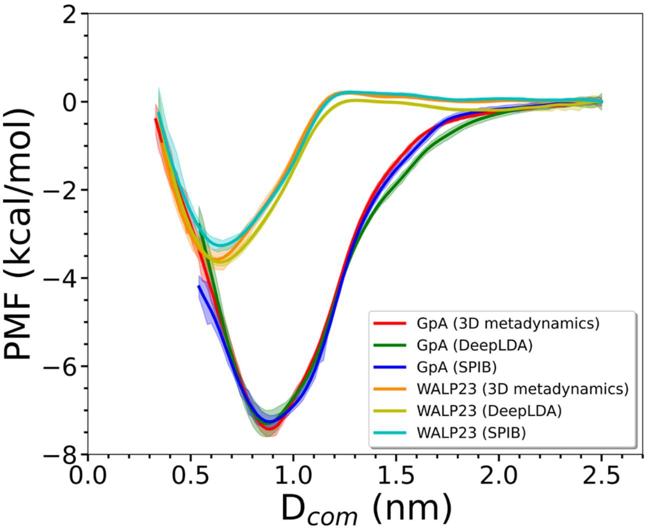


Figure 5. Potential of mean force (PMF) for the association of GpA and WALP23 as a function of the lateral center-of-mass distances between two helices (D_{com}), obtained from the 3D metadynamics and simulations along the CVs derived from the DeepLDA and SPIB protocols. The PMFs were set to plateau at zero to guide the comparison. Error bars associated with each PMF were calculated using block average analysis.

from NMR in detergent micelle,[33] solid state NMR,[55] and crystallographic analysis in lipid cubic phase bilayers[56] provide consistent information. The GpA homodimer was found to be stabilized by interhelical interactions between residues of the glycine zipper (GXXXG) motifs, forming a right-handed helical structure with a crossing angle of −22°. Mutations of the GXXXG motif[57,58] and T87[59,60] were found to disrupt the dimerization propensity of GpA. Sengupta et al. studied triple mutations of the GXXXG motif and T87F mutant of GpA using the MARTINI v2 force field.[22] Although the mutants were found to dimerize, their dimerization tendency was reduced compared to that of the wild-type sequence. Previous studies using one-dimensional collective variables showed that the GpA homodimer is stabilized by interactions conveyed through the $G_{79}XXXG_{83}XXXT_{87}$ motif, forming a right-handed helix with a crossing angle of −26°.[13] However, in many cases, the simulations were found to be trapped in one energy minimum, restricting the sampling of important thermodynamically relevant conformations. A study using a two-dimensional collective variable showed the importance of non-native dimer structures as well as native interaction through the GXXXG motif in the calculation of the association constant of the GpA homodimer.[13] A complete estimation of the stable structural ensemble can be compared with the experimental results to calibrate the computational models.

In this context, it is important to derive an enhanced sampling technique that exhaustively samples all possible dimer configurations. Sahoo et al. studied the GpA homodimer using the MARTINI v3 force field and reported three conformational clusters characterized by different handedness properties.[51] We have studied GpA and WALP23 homodimerization using the MARTINI v3 force field, employing our 3D metadynamics protocol.[16] Five different stable homodimer configurations of GpA were identified through comparison to unbiased simulations. GXXXG motif interactions were found to play an important role in the formation of a right-handed helical dimer. WALP23 forms a right-handed helical dimer stabilized by the A_9XXXA_{13} motif, whereas the left-handed helix formation was found to be facilitated by the A_7XXXA_{11} motif.

To characterize the structural ensembles of GpA and WALP23 obtained from the simulations along DeepLDA and SPIB derived CVs, we projected the PMF onto the crossing angle and D_{com} (Figure 6). A similar crossing angle distribution of WALP23 was observed from the 3D metadynamics and simulations along DeepLDA and SPIB derived CVs. A stable right-handed structure (cluster R) of WALP23 with an average crossing angle of −22° was found, and a shallow minimum at a crossing angle of 22° predicts a stable left-handed dimer conformation (cluster L). The probability distribution of GpA along the crossing angle predicts three major conformational ensembles (clusters L1, L2, and R). A similar qualitative pattern was observed using all three protocols. However, the global minimum of PMF was observed to be located at different positions. The need to sample multiple states across the crossing angle distribution demonstrates the importance of the proper choice of collective variables to facilitate enhanced sampling.

■ CONCLUSIONS

The calculation of association constants for transmembrane protein helices is of fundamental importance in biophysics and our understanding of biomolecular organization and cellular signaling. Converged estimations of the dimerization free energy can be compared to experimental results to calibrate computational models. Previous studies showed that the effectiveness of an enhanced sampling method in capturing relevant dimer conformations and determining the equilibrium association constant depends on the choice of the collective variables. Sampling along one-dimensional collective variables D_{com} and D_{rmsd} was found to lead to frustrated sampling, leading to an overestimation of the dimerization free energy. Recently, we proposed a metadynamics protocol to investigate the homodimerization of transmembrane proteins using three collective variables: the *x*- and *y*-projections of the relative center-of-mass distances between two helices and the interhelical crossing angle. Comparisons between 3D metadynamics simulations and extensive unbiased simulations demonstrate that the 3D metadynamics approach comprehensively samples thermodynamically relevant dimer conformations and provides a converged estimation of the dimerization free energy.

Table 2. Minimum of PMFs along D_{com} (ΔW_{min}) and Dimerization Free Energy (ΔG) of GpA and WALP23 Homodimers Obtained from the 3D Metadynamics and the Simulations along the DeepLDA and SPIB Derived CVs

	3D metadynamics		DeepLDA		SPIB	
protein	ΔW_{min} (kcal/mol)	ΔG (kcal/mol)	ΔW_{min} (kcal/mol)	ΔG (kcal/mol)	ΔW_{min} (kcal/mol)	ΔG (kcal/mol)
GpA	−7.3 ± 0.1	−7.0 ± 0.1	−7.3 ± 0.2	−7.0 ± 0.2	−7.2 ± 0.1	−7.0 ± 0.1
WALP23	−3.6 ± 0.2	−3.2 ± 0.2	−3.6 ± 0.1	−3.2 ± 0.1	−3.3 ± 0.1	−3.0 ± 0.1

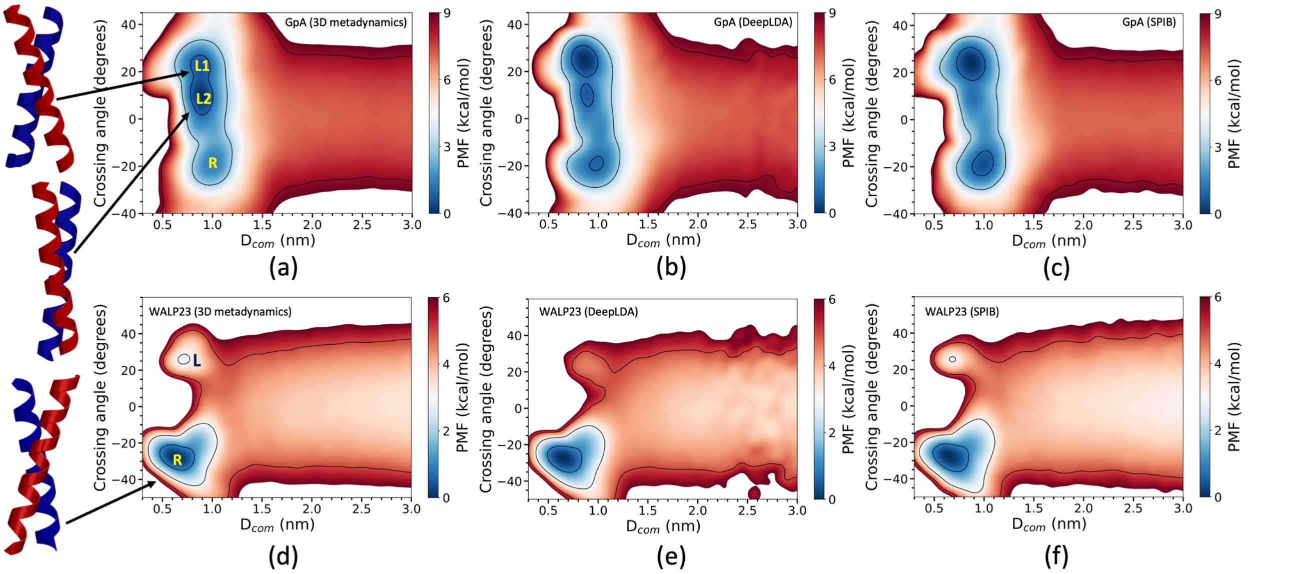


Figure 6. Potential of mean force (PMF) for the association of GpA and WALP23 as a function of crossing angle and D_{com} obtained from (a, d) the 3D metadynamics and simulations along the (b, e) DeepLDA and (c, f) SPIB derived CVs. The minimum value of the PMFs was set to zero to guide the comparison. Clusters were labeled as L or R based on the handedness properties of the helix.

In this work, we derived collective variables using two machine learning based protocols, DeepLDA and SPIB, to investigate the association/dissociation equilibrium of GpA and WALP23 homodimers. To ensure optimal performance of the neural network based collective variables in studying transmembrane protein dimerization, it is essential that the training data set contains information on both native and non-native dimer structures. To derive these collective variables, a deep neural network model was trained by using data obtained from extensive unbiased simulations. Following the identification of the essential collective variables, metadynamics simulations can be initiated from any equilibrium configuration of the transmembrane homodimer embedded in a membrane bilayer.

We compared the results obtained from simulations characterizing the dimerization of TM proteins using different collective variables. The dimerization free energies obtained from simulations using CVs derived from DeepLDA and SPIB protocols were found to be similar to those obtained through the 3D metadynamics simulations. Projecting the probability density for the dimer conformational ensembles of GpA and WALP23 onto the *xy*-plane and onto the crossing angle and D_{com}, similar hotspots on the *xy*-plane and along the crossing angle were observed using all three approaches. Simulations using CVs derived from DeepLDA and SPIB, as well as 3D metadynamics simulations, sample similar homodimer conformational ensembles. Our results demonstrate the broad applicability of the 3D metadynamics approach in studying transmembrane protein homodimerization and suggest that the DeepLDA and SPIB protocols can be employed to effectively derive CVs that inform and enhance the simulation and analysis of protein−protein association in complex biomolecular systems.

■ ASSOCIATED CONTENT

Data Availability Statement

Initial structures of membrane bilayers containing the GpA and WALP23 homodimers and the scripts used to perform well-tempered metadynamics along the collective variables derived using the DeepLDA and SPIB protocols are freely available at https://github.com/ayan-majumder95/tm_ml_cv.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c00454.

> Detailed description of training data sets and the DeepLDA[61] and SPIB protocols, pictorial representation of descriptors, workflow of the study, results obtained from the unbiased simulations and those obtained from the simulation along DeepLDA derived CV, initial state descriptions used to derive CVs using the SPIB protocol, convergence analysis of the PMFs, and importance of the descriptors (PDF)

■ AUTHOR INFORMATION

Corresponding Author

John E. Straub − *Department of Chemistry, Boston University, Boston, Massachusetts 02215, United States;* orcid.org/0000-0002-2355-3316; Email: straub@bu.edu

Author

Ayan Majumder − *Department of Chemistry, Boston University, Boston, Massachusetts 02215, United States;* orcid.org/0009-0005-4218-3276

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.4c00454

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors gratefully acknowledge the generous support of the National Institutes of Health (grant no. R01 GM107703), the National Science Foundation (grant no. CHE1900416), and the high-performance computing resources of the Boston University Shared Computing Cluster (SCC). We thank the reviewers for valuable comments that improved our work.

■ REFERENCES

(1) Engel, A.; Gaub, H. E. Structure and mechanics of membrane proteins. *Annu. Rev. Biochem.* **2008**, *77*, 127−148.

(2) Majumder, A.; Vuksanovic, N.; Ray, L. C.; Bernstein, H. M.; Allen, K. N.; Imperiali, B.; Straub, J. E. Synergistic computational and experimental studies of a phosphoglycosyltransferase membrane/ligand ensemble. *J. Biol. Chem.* **2023**, *299*, 105194.

(3) Gomes, I.; Ayoub, M. A.; Fujita, W.; Jaeger, W. C.; Pfleger, K. D.; Devi, L. A. G protein-coupled receptor heteromers. *Annual review of pharmacology and toxicology* **2016**, *56*, 403−425.

(4) Irvine, G. B.; El-Agnaf, O. M.; Shankar, G. M.; Walsh, D. M. Protein aggregation in the brain: the molecular basis for Alzheimer's and Parkinson's diseases. *Molecular medicine* **2008**, *14*, 451−464.

(5) Nguyen, P. H.; Ramamoorthy, A.; Sahoo, B. R.; Zheng, J.; Faller, P.; Straub, J. E.; Dominguez, L.; Shea, J.-E.; Dokholyan, N. V.; De Simone, A.; et al. Amyloid oligomers: A joint experimental/computational perspective on Alzheimer's disease, Parkinson's disease, type II diabetes, and amyotrophic lateral sclerosis. *Chem. Rev.* **2021**, *121*, 2545−2647.

(6) Mori, T.; Miyashita, N.; Im, W.; Feig, M.; Sugita, Y. Molecular dynamics simulations of biological membranes and membrane proteins using enhanced conformational sampling algorithms. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2016**, *1858*, 1635−1651.

(7) Dominguez, L.; Foster, L.; Straub, J. E.; Thirumalai, D. Impact of membrane lipid composition on the structure and stability of the transmembrane domain of amyloid precursor protein. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, E5281−E5287.

(8) Dominguez, L.; Foster, L.; Meredith, S. C.; Straub, J. E.; Thirumalai, D. Structural heterogeneity in transmembrane amyloid precursor protein homodimer is a consequence of environmental selection. *J. Am. Chem. Soc.* **2014**, *136*, 9619−9626.

(9) Baaden, M.; Marrink, S. J. Coarse-grain modelling of protein-protein interactions. *Curr. Opin. Struct. Biol.* **2013**, *23*, 878−886.

(10) Ito, S.; Sugita, Y. Free-energy landscapes of transmembrane homodimers by bias-exchange adaptively biased molecular dynamics. *Biophys. Chem.* **2024**, *307*, 107190.

(11) Filippov, A.; Orädd, G.; Lindblom, G. The effect of cholesterol on the lateral diffusion of phospholipids in oriented bilayers. *Biophysical journal* **2003**, *84*, 3079−3086.

(12) Roux, B. The calculation of the potential of mean force using computer simulations. *Computer physics communications* **1995**, *91*, 275−282.

(13) Majumder, A.; Kwon, S.; Straub, J. E. On Computing Equilibrium Binding Constants for Protein-Protein Association in Membranes. *J. Chem. Theory Comput.* **2022**, *18*, 3961−3971.

(14) Lelimousin, M.; Limongelli, V.; Sansom, M. S. Conformational changes in the epidermal growth factor receptor: Role of the transmembrane domain investigated by coarse-grained metadynamics free energy calculations. *J. Am. Chem. Soc.* **2016**, *138*, 10611−10622.

(15) Leone, V.; Marinelli, F.; Carloni, P.; Parrinello, M. Targeting biomolecular flexibility with metadynamics. *Curr. Opin. Struct. Biol.* **2010**, *20*, 148−154.

(16) Majumder, A.; Straub, J. E. The role of structural heterogeneity in the homodimerization of transmembrane proteins. *J. Chem. Phys.* **2023**, *159*, 134101.

(17) Comer, J.; Gumbart, J. C.; Hénin, J.; Lelièvre, T.; Pohorille, A.; Chipot, C. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. *J. Phys. Chem. B* **2015**, *119*, 1129−1151.

(18) Blazhynska, M.; Gumbart, J. C.; Chen, H.; Tajkhorshid, E.; Roux, B.; Chipot, C. A Rigorous Framework for Calculating Protein-Protein Binding Affinities in Membranes. *J. Chem. Theory Comput.* **2023**, *19*, 9077−9092.

(19) Domański, J.; Sansom, M. S.; Stansfeld, P. J.; Best, R. B. Balancing force field protein-lipid interactions to capture transmembrane helix-helix association. *J. Chem. Theory Comput.* **2018**, *14*, 1706−1715.

(20) Domański, J.; Hedger, G.; Best, R. B.; Stansfeld, P. J.; Sansom, M. S. Convergence and sampling in determining free energy landscapes for membrane protein association. *J. Phys. Chem. B* **2017**, *121*, 3364−3375.

(21) Majumder, A.; Straub, J. E. Addressing the excessive aggregation of membrane proteins in the MARTINI model. *J. Chem. Theory Comput.* **2021**, *17*, 2513−2521.

(22) Sengupta, D.; Marrink, S. J. Lipid-mediated interactions tune the association of glycophorin A helix and its disruptive mutants in membranes. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12987−12996.

(23) Chipot, C. Free energy methods for the description of molecular processes. *Annual Review of Biophysics* **2023**, *52*, 113−138.

(24) Noé, F.; Clementi, C. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Curr. Opin. Struct. Biol.* **2017**, *43*, 141−147.

(25) Bonati, L.; Piccini, G.; Parrinello, M. Deep learning the slow modes for rare events sampling. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, e2113533118.

(26) Bonati, L.; Rizzi, V.; Parrinello, M. Data-driven collective variables for enhanced sampling. *journal of physical chemistry letters* **2020**, *11*, 2998−3004.

(27) Wang, D.; Tiwary, P. State predictive information bottleneck. *J. Chem. Phys.* **2021**, *154*, 134111.

(28) Das, S.; Raucci, U.; Neves, R. P.; Ramos, M. J.; Parrinello, M. How and when does an enzyme react? Unraveling α-Amylase catalytic activity with enhanced sampling techniques. *ACS Catal.* **2023**, *13*, 8092−8098.

(29) Chen, W.; Ferguson, A. L. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *Journal of computational chemistry* **2018**, *39*, 2079−2102.

(30) Dorfer, M.; Kelz, R.; Widmer, G. Deep linear discriminant analysis. *arXiv* **2015**, 1.

(31) Mendels, D.; Piccini, G.; Parrinello, M. Collective variables from local fluctuations. *journal of physical chemistry letters* **2018**, *9*, 2776−2781.

(32) Mehdi, S.; Wang, D.; Pant, S.; Tiwary, P. Accelerating all-atom simulations and gaining mechanistic understanding of biophysical systems through state predictive information bottleneck. *J. Chem. Theory Comput.* **2022**, *18*, 3231−3238.

(33) MacKenzie, K. R.; Prestegard, J. H.; Engelman, D. M. A transmembrane helix dimer: structure and implications. *Science* **1997**, *276*, 131−133.

(34) Souza, P. C.; Alessandri, R.; Barnoud, J.; Thallmair, S.; Faustino, I.; Grünewald, F.; Patmanidis, I.; Abdizadeh, H.; Bruininks, B. M.; Wassenaar, T. A.; et al. Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nat. Methods* **2021**, *18*, 382−388.

(35) Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters* **2008**, *100*, 020603.

(36) 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*, 19−25.

(37) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Computer physics communications* **2014**, *185*, 604−613.

(38) Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*; 2017.

(39) Ruiz Munevar, M. J.; Rizzi, V.; Portioli, C.; Vidossich, P.; Cao, E.; Parrinello, M.; Cancedda, L.; De Vivo, M. Cation Chloride Cotransporter NKCC1 Operates through a Rocking-Bundle Mechanism. *J. Am. Chem. Soc.* **2024**, *146*, 552−566.

(40) Ansari, N.; Rizzi, V.; Parrinello, M. Water regulates the residence time of Benzamidine in Trypsin. *Nat. Commun.* **2022**, *13*, 5438.

(41) Fleming, K. G.; Ackerman, A. L.; Engelman, D. M. The effect of point mutations on the free energy of transmembrane α-helix dimerization. *J. Mol. Biol.* **1997**, *272*, 266−275.

(42) Fleming, K. G. Standardizing the free energy change of transmembrane helix-helix interactions. *J. Mol. Biol.* **2002**, *323*, 563−571.

(43) Nash, A.; Notman, R.; Dixon, A. M. De novo design of transmembrane helix-helix interactions and measurement of stability in a biological membrane. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2015**, *1848*, 1248−1257.

(44) Hong, H.; Blois, T. M.; Cao, Z.; Bowie, J. U. Method to measure strong protein-protein interactions in lipid bilayers using a steric trap. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 19802−19807.

(45) Sarabipour, S.; Hristova, K. Glycophorin A transmembrane domain dimerization in plasma membrane vesicles derived from CHO, HEK 293T, and A431 cells. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2013**, *1828*, 1829−1833.

(46) Chen, L.; Novicky, L.; Merzlyakov, M.; Hristov, T.; Hristova, K. Measuring the energetics of membrane protein dimerization in mammalian membranes. *J. Am. Chem. Soc.* **2010**, *132*, 3628−3635.

(47) Henin, J.; Pohorille, A.; Chipot, C. Insights into the recognition and association of transmembrane α-helices. The free energy of α-helix dimerization in glycophorin A. *J. Am. Chem. Soc.* **2005**, *127*, 8478−8484.

(48) Castillo, N.; Monticelli, L.; Barnoud, J.; Tieleman, D. P. Free energy of WALP23 dimer association in DMPC, DPPC, and DOPC bilayers. *Chemistry and physics of lipids* **2013**, *169*, 95−105.

(49) Schäfer, L. V.; de Jong, D. H.; Holt, A.; Rzepiela, A. J.; de Vries, A. H.; Poolman, B.; Killian, J. A.; Marrink, S. J. Lipid packing drives the segregation of transmembrane helices into disordered lipid domains in model membranes. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 1343−1348.

(50) Yano, Y.; Matsuzaki, K. Measurement of thermodynamic parameters for hydrophobic mismatch 1: self-association of a transmembrane helix. *Biochemistry* **2006**, *45*, 3370−3378.

(51) Sahoo, A. R.; Souza, P. C.; Meng, Z.; Buck, M. Transmembrane dimers of type 1 receptors sample alternate configurations: MD simulations using coarse grain Martini 3 versus AlphaFold2Multimer. *Structure* **2023**, *31*, 735−745.

(52) Luo, H.; Sharp, K. On the calculation of absolute macromolecular binding free energies. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 10399−10404.

(53) Duboué-Dijon, E.; Hénin, J. Building intuition for binding free energy calculations: Bound state definition, restraints, and symmetry. *J. Chem. Phys.* **2021**, *154*, 204101.

(54) Roux, B. In *Computational Modeling and Simulations of Biomolecular Systems*; World Scientific, 2021.

(55) Smith, S. O.; Eilers, M.; Song, D.; Crocker, E.; Ying, W.; Groesbeek, M.; Metz, G.; Ziliox, M.; Aimoto, S. Implications of threonine hydrogen bonding in the glycophorin A transmembrane helix dimer. *Biophys. J.* **2002**, *82*, 2476−2486.

(56) Trenker, R.; Call, M. E.; Call, M. J. Crystal structure of the glycophorin A transmembrane dimer in lipidic cubic phase. *J. Am. Chem. Soc.* **2015**, *137*, 15676−15679.

(57) Brosig, B.; Langosch, D. The dimerization motif of the glycophorin A transmembrane segment in membranes: importance of glycine residues. *Protein Sci.* **1998**, *7*, 1052−1056.

(58) Langosch, D.; Brosig, B.; Kolmar, H.; Fritz, H.-J. Dimerisation of the glycophorin A transmembrane segment in membranes probed with the ToxR transcription activator. *J. Mol. Biol.* **1996**, *263*, 525−530.

(59) Doura, A. K.; Kobus, F. J.; Dubrovsky, L.; Hibbard, E.; Fleming, K. G. Sequence context modulates the stability of a GxxxG-mediated transmembrane helix-helix dimer. *Journal of molecular biology* **2004**, *341*, 991−998.

(60) Lemmon, M. A.; Flanagan, J. M.; Treutlein, H. R.; Zhang, J.; Engelman, D. M. Sequence specificity in the dimerization of transmembrane. alpha.-helixes. *Biochemistry* **1992**, *31*, 12719−12725.

(61) Raucci, U.; Rizzi, V.; Parrinello, M. Discover, sample, and refine: Exploring chemistry with enhanced sampling techniques. *J. Phys. Chem. Lett.* **2022**, *13*, 1424−1430.