COMPUTATIONAL METHODS TO DESIGN BIOPHYSICAL EXPERIMENTS FOR THE STUDY OF PROTEIN DYNAMICS

BY

SHRIYAA MITTAL

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biophysics and Quantitative Biology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2020

Urbana, Illinois

Doctoral Committee:

Assistant Professor Diwakar Shukla, Chair
Professor Martin Gruebele
Professor Zaida Luthey-Schulten
Professor Saurabh Sinha

Abstract

In recent years, new software and automated instruments have enabled us to imagine autonomous or "self-driving" laboratories of the future. However, ways to design new scientific studies remain unexplored due to challenges such as minimizing associated time, labor, and expense of sample preparation and data acquisition. In the field of protein biophysics, computational simulations such as molecular dynamics and spectroscopy-based experiments such as double electron-electron resonance and Fluorescence resonance energy transfer techniques have emerged as critical experimental tools to capture protein dynamic behavior, a change in protein structure as a function of time which is important for their cellular functions. These techniques can lead to the characterization of key protein conformations and can capture protein motions over a diverse range of timescales.

This work addresses the problem of the choice of probe positions in a protein, which residue-pairs should experimentalists choose for spectroscopy experiments. For this purpose, molecular dynamics simulations and Markov state models of protein conformational dynamics are utilized to rank sets of labeled residue-pairs in terms of their ability to capture the conformational dynamics of the protein. The applications of our experimental study design methodology called *OptimalProbes* on different types of proteins and experimental techniques are examined.

In order to utilize this method for a previously uncharacterized protein, atomistic molecular dynamics simulations are performed to study a bacterial di/tri-peptide transporter a typical representative of the Major Facilitator Superfamily of membrane proteins. This was followed by ideal double electron-electron resonance experimental choice predictions based on the simulation data. The predicted choices are superior to the residue-pair choices made by experimentalists which failed to capture the slowest dynamical processes in the conformational ensemble obtained from our long timescale simulations.

For molecular dynamics simulations based design of experimental studies to succeed both ensembles need to be comparable. Since this has not been the case for double electron-electron resonance distance distributions and molecular simulations, we explore possible reasons that can lead to mismatches in order to reconcile simulated ensembles with experimentally obtained distance traces.

This work is one of the first studies towards integrating spectroscopy experiment design into a computational method systematically based on molecular simulations.

To my parents.

Acknowledgments

This dissertation is only a part of my PhD. Here, I can thank only those who directly impacted the work included in this document, among numerous people who I have been fortunate to be associated with during my PhD, even if so briefly.

Thanks to Diwakar Shukla for being my research advisor and introducing me to various aspects of the academic endeavor. To my committee member and collaborator, Martin Gruebele, for being an excellent role model as a scientist. To members of my committee, Saurabh Sinha and Zaida Luthey-Schulten for their encouragement along the way.

Thanks to many labmates for their cooperation and discussions in helping me complete this work. To my friends, biophysics classmates, and many biophysics graduate students, with whom I have shared memorable moments. To my collaborator Drishti Guin with whom I briefly dabbled in biophysics experiments. To the biophysics program coordinators for their administrative assistance.

Thanks to Computational Science and Engineering and Graduate College at the University of Illinois for fellowships over the course of my PhD.

Finally but foremost, to my parents, my sister, and Alex Moffett. I am grateful for their unconditional love and support, without which I could not have completed my PhD.

Table of Contents

List of Abbreviations

BDDM	n-Dodecyl-β-D-Maltoside
BRET	Bioluminescence Resonance Energy Transfer
cryoEM	cryo-Electron Microscopy
DCC	Dynamic Cross Correlation
DEER	Double Electron-Electron Resonance
EPR	Electron Paramagnetic Resonance
FRET	Förster/Fluorescence Resonance Energy Transfer
GMRQ	Generalized matrix Rayleigh quotient
GPCR	G-Protein Coupled Receptors
HDX-MS	Hydrogen-Deuterium Exchange
IF	Inward Facing
LRET	Luminescence Resonance Energy Transfer
MD	Molecular Dynamics
MFS	Major Facilitator Superfamily
MS	Mass Spectrometry
MSM	Markov State Model
MTSSL	1-oxyl-2,2,5,5-tetramethyl-pyrroline-3-methyl)methanethiosulfonate
NMR	Nuclear Magnetic Resonance
NSS	Neurotransmitter: Sodium Symporter
OC	Occluded
OF	Outward Facing
PDB	Protein Data Bank
POT	Proton-dependent Oligopeptide Transporter
PTR	Peptide Transporter
RMSD	Root Mean Squared Deviation

SASA	Solvent Accessible Surface Area
SDSL	Site-Directed Spin Labeling
tIC	Time-structure based Independent Component
tICA	Time-structure based Independent Component Analysis
TM	Transmembrane
TPT	Transition Path Theory
TRAM	Transition-based Reweighting Analysis Method
TTET	Triplet-Triplet Energy Transfer

Chapter 1

Introduction

Proteins and their dynamics are crucial for all biological functions in cells [1]. Folding of proteins from large amino acid chains post-translation populating intermediate states and finally into their native structure that enables their cellular function is an example of protein dynamics. Protein mis-folding or malfunctions are culprits for onset of most diseases such as cancers, Alzheimer's disease, Parkinson's disease, Huntington's disease, cystic fibrosis among other degenerative disorders and metabolic disorders such as type 1 and type 2 diabetes [2, 3]. As a result, proteins are targets for typical drug-discovery efforts, specifically membrane proteins such as transporters that can carry drug molecules through cellular membranes into their acting sites inside cells or G-protein coupled receptors (GPCRs) that carry information into the cell about drug binding through their active and inactive conformations [4]. Membrane protein dynamics involve proteins to undergo conformational changes, that are more subtle as compared to protein folding but equivalently important, that enable proteins to open or close gates, allow alternate access to a substrate on either side of the cellular membrane, or binding of membrane-peripheral protein partners.

X-ray crystallography has led to tremendous advance in our understanding of proteins since 1958 when the first three-dimensional structure of sperm whale myoglobin was described [5]. In conjunction with X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy has also allowed researchers to visualize 3D protein structure. Unlike X-ray experiments, NMR is not limited by a crystal lattice, allowing researchers to capture some semblance of the dynamic regions of proteins. X-ray crystallography structures can only provide a window into a single conformation of the protein, sometimes where dynamic regions of proteins remain unresolved or the captured conformation is sparsely populated in solution. A downside with NMR is it's limited use for large proteins, especially membrane proteins which form a major focus of this work. Irrespective of the experimental method, membrane protein structures are sparse in the Protein Data Bank. Only 2% of all protein structures are those of membrane proteins and most of which have been resolved only recently. In the past few years, time-resolved serial femtosecond crystallography has allowed researchers to capture multiple snapshots of protein dynamics in the fs to ms timescale [6]. Recent methodological advances have also led to a membrane protein structure determination using cryo-electron microscopy (cryoEM) [7,8],

although at a lower resolution than X-ray crystallography structures.

While the above mentioned three-dimensional structure determinations techniques advance further, another source of unprecedented insights into membrane proteins has been through spectroscopy which provide an indirect measurement of the protein's structure as well as it's conformational dynamics and kinetics. For example, Fourier transform Infrared (FTIR) spectroscopy can be used to determine the secondary structure of proteins through the different hydrogen bonding patterns formed by α-helices, β-sheets, β-turns, or coils. The absorption of infrared radiation excites vibrational transitions of molecules, different hydrogen bonding patterns yield different vibration frequencies which can then be assigned to secondary structure content of the protein. In this work we refer to spectroscopy techniques such as electron paramagnetic resonance (EPR), double electron-electron resonance (DEER), triplet-triplet energy transfer (TTET), Förster resonance energy transfer or fluorescence resonance energy transfer (FRET), luminescence resonance energy transfer (LRET), bioluminescence resonance energy transfer (BRET) which are commonly used to measure distances and interactions in biomolecular systems including proteins.

These techniques are a mechanism to describe energy transfer or coupling between two probe molecules strategically placed on the proteins. The energy transfer or coupling is measured via a technique specific signal and it's strength or efficiency is dependent on the distance between the probe molecules. Signal obtained is typically sensitive to the changes in the distance between probes and hence, allows researchers to capture multiple distances between residue-pairs or secondary structure pairs in a protein. For example, in FRET experiments, two light sensitive molecules are attached to two residues of a protein. When the donor molecule is excited, it may transfer energy to the acceptor molecule and the energy transfer, E is related to the distance between donor and acceptor molecule, r as $E \propto \frac{1}{r^6}$. Each technique differs in it's theoretical basis, practical implementation by using different probe molecules, and analysis of the obtained energy transfer efficiency.

On the computational methods side, molecular dynamics (MD) simulations are a powerful tool to study biological systems, specifically proteins, and have been widely employed for understanding their dynamics in folding, ligand perception and binding, conformational changes, and disorder. Given the functional relevance of protein dynamics in both health and disease, in our works, we use unbiased MD simulations to sample the conformational dynamics of proteins and propose the use of MD simulation datasets as a predictive tool for the study of biological systems in conjunction with biophysical spectroscopy experiments. In particular, we employ MD simulations as a resource to design biophysical experiments to aid researchers in their choice of residue-pairs for energy transfer spectroscopy experiments. We use kinetic network models called Markov state models as a framework for experiment design to choose residue-pairs that maximize the number of

slow dynamical processes from experiments on a given protein.

The remainder of this chapter will introduce these methods which form the basis of our computational method to design experiments. First, MD simulations, challenges and recent advancements in the use of MD simulations for protein conformational changes. Second, Markov state models, a key idea to draw insights from MD simulations such as thermodynamic stability of different protein conformations and kinetics of transitions among protein conformations. Third, Generalized matrix Rayleigh quotient, a property derived based on the eigenvalues of an Markov state model's transitional probability matrix.

1.1 Molecular dynamics simulations: Challenges and advances[1]

Generally, an MD simulation investigation of proteins involves the following steps: (1) Setting up the MD system where the atomic coordinates are derived from experimental structural studies and sometimes homology-based computational models [9–12]. (2) Performing energy minimization and equilibration, followed by large time production runs, also referred to as sampling, using available MD engines such as NAMD [13], AMBER [14], GROMACS [15], OpenMM [16]. Typically, MD programs employ molecular mechanics force-fields parameterized based on previously evaluated and benchmarked datasets. Sampling can be performed in an unbiased manner without changing the underlying Hamiltonian or with a modified Hamiltonian, either with the aim to increase probability of exploring rare conformations or with a fixed end goal to discover the path from the starting to the end conformation [17–28]. (3) Analyzing the trajectory data obtained from sampling the conformational landscape, such as monitoring the fluctuations in the protein's observable over time, or more complex calculations. Atomistic MD simulations enable researchers to draw conclusions regarding the nanoscopic structural changes associated with the protein and the necessity of the structural changes for the physiological function of the protein. Often, results obtained from MD simulations are accompanied by validation with biophysical and biochemical experiments.

Recently, there has been considerable progress in bioinformatics based methods to predict contacts between co-evolving residues and using these evolutionary coupled residues to determine the structural model of proteins [29–34]. Tang et al. have proposed the combination of contacts inferred from evolutionary couplings and sparse NMR to build verifiable structural models [35]. Methods to use evolutionary couplings to predict structures have contributed to the pool of protein structures and in theory can also be employed for MD studies in the absence of alternatives.

MD simulations are heavily dependent upon the force field used while performing simulations. Various

[1]This section is adapted with permission from Mittal S and Shukla D. Molecular Simulation. 2018; 44(11):891-904. Copyright 2018 Taylor & Francis.

studies have determined that thermodynamic behavior of folded protein can be described with accuracy but the force fields can also be biased towards the folded structural states only [36–39]. The choice of force field is often subjective and MD studies may require simulations to be performed in multiple force-fields resulting in similar or contesting observations from the resulting datasets.

Moreover, sampling the conformational landscape for biologically relevant time periods, requires intense computations on specialized hardware [40, 41], distributed computing resources [42] or high performance machines [43]. An often exploited solution to prevent excessive exploration in low-energy regions is to provide the protein an opportunity to access the higher energy conformations via accelerated MD, replica exchange MD, self guided molecular/langevin dynamics [20,21,25,28] or perform simulations to drive sampling in a pre-chosen direction such as adaptive biasing force method, temperature accelerated MD, umbrella sampling and metadynamics [17,22,24,26]. Pathway generation methods such as string method, transition path sampling, steered MD, targeted MD [18, 19, 23, 27] and others attempt to mitigate the problem of inefficient sampling of the transitions across high energy barriers. These 'biased' sampling techniques have been instrumental in studying a diverse variety of proteins. Despite increasing thermodynamic accuracy, these techniques have limitations on account of losing out on kinetics information.

Once the simulation dataset is obtained, using the researcher's preferred choice of the sampling technique depending on the scientific purpose, simulation observables are matched with experimental results [44]. Simulations which are performed on distributed computing resources have mostly been in the form of short simulations [45–51], where iterative rounds are seeded from structures or conformational regions on the landscape which are deemed to be less sampled in the previously obtained data. The analysis of this data demands a sophisticated statistical method called Markov state models [52]. MSMs can also be employed to analyze simulation data obtained as single or few long trajectories and MSM based calculations have been employed to validate results from kinetic experiments [53, 54].

1.2 Markov state models

Markov state models (MSMs) are kinetic network models of protein conformations. An MSM consists of the state decomposition, their equilibrium populations and the transition probabilities among the states. The transition probability matrix (T_{ij}) described the probability of transitioning from discrete conformational state i to state j in some lag time τ. An MSM must ensure that that the timescale τ chosen for state transition determinations is approximately Markovian. This implies that the transitional probabilities should depend only on the last state visited, and not on the states visited before that. An MSM built in a memoryless

manner provides a skeleton upon which calculations such as mean first passage time, free-energy calculations and pathway determination of the protein's conformational changes can be performed. Steps involved in MSM construction are explored in this section.

Featurization

The state decomposition is obtained by clustering the trajectories into distinct states based on a structural metric called features as calculated for every frame in the collected simulation data. Some examples of these features include root mean squared deviation (RMSD) of the entire protein or certain dynamic regions of the protein, amino acid dihedral angles of which ones typically used are phi (ϕ), psi (ψ), and chi (χ), one or more distances between various entities in the simulation system. The choice of features has consequence on the conclusions and the significance of the results from the resulting MSM. Features are chosen based on the type of simulation system, such as in the case of ligand or drug binding studies to a protein, it is common practice to use distance between ligand atoms and atoms in the ligand binding pocket. In a recent study on abscisic acid binding to PYL receptor proteins, researchers used 4 distances between ligand and final binding site residues as well as 4 distances between ligand and residues in the binding pathway to capture the recognition and binding mechanism using MSMs [51]. Structural regions of the protein that are known to play an important role in it's function or activation are also included in the choice of features. For example, Shukla et al. included root mean squared fluctuations (RMSF) of heavy atoms from the N-terminal lobe residues and A-loop residues of c-src tyrosine kinase as features for the MSM since the flip of a DFG motif at the N-terminal end of the A-loop play is a known indicator of activation in kinases [45]. While the activation mechanism of kinases had been studied previously, studies of new proteins requires agnostic choice of features in order to capture the slowest physical processes involved in it's dynamics.

Dimensionality reduction[2]

Construction of MSMs can also be aided by dimensionality reduction, as it provides a reduced subspace to perform clustering and decompose the trajectory data into states which are geometrically and kinetically distinct from others, by weeding out statistical noise [52, 55]. The aim of dimensionality reduction is to transform a large set of features to a different basis set which maximizes a measurable quantity such as variance [56, 57] or Shannon information [58]. A statistical method called principal component analysis (PCA) attempts to transform the large dimension dataset, obtained after converting the trajectory snapshots to features, into a new basis set or a coordinate system which has the maximum uncorrelated variance values.

[2]This section is adapted with permission from Mittal S and Shukla D. Molecular Simulation. 2018; 44(11):891-904. Copyright 2018 Taylor & Francis.

However, the MD data have time stamps associated with them, which cannot be ignored at the risk of losing information and studies on the kinetics of the protein's dynamic processes; similar to the field of signal processing. Hence, time-structure based Independent Component Analysis (tICA) [59, 60] is a better suited dimensionality reduction method in MD simulations than PCA. tICA is inherently similar to PCA but takes into account the time progression of the MD simulation datasets. The use of tICA for MD simulation data analysis is indeed borrowed from signal processing [61, 62].

In tICA, the features are transformed to find linear combinations of them resulting in a Z matrix where the column vectors are the weights of the feature in the linear combination. Each transformed coordinate has two properties: (1) maximal autocorrelation to obtain the slow motions as observed in the data, and (2) uncorrelated to the previous linear combinations to ensure each vector describes an orthogonal slow process. The covariance matrix for a multidimensional time series, $r(t)$, with N snapshots or frames is given as,

$$C^r(0)_{ij} = \frac{1}{(N-1)} \sum_{t=1}^{N} r_i(t) r_j(t) \tag{1.1}$$

and the time-lagged covariance matrix, after a lag of ΔN snapshots,

$$C^r(\Delta N)_{ij} = \frac{1}{(N-\Delta N-1)} \sum_{t=1}^{N-\Delta N} r_i(t) r_j(t+\Delta N) \tag{1.2}$$

The covariance and the time-lagged covariance matrices can be transformed to the new transformation, Z, as $C^Z(0) = Z^T C^r(0) Z$ and $C^Z(\Delta N) = Z^T C^r(\Delta N) Z$, respectively. The time-lagged independent components (tICs) are then determined by solving the eigenvectors of the generalized eigenvalue problem, $C^r(\Delta N)Z = \Lambda C^r(0)Z$ where Λ is a diagonal matrix of the eigenvalues, λ_i, corresponding to the auto-correlations. If the solutions to this equation are ordered in the decreasing order of the eigenvalues such that $\lambda_0 > \lambda_1 > \cdots > \lambda_d$, then Z_0 is the slowest time-structure based Independent Component (tIC) and so on.

In recent years, tICs have been regularly used for post simulation analysis in order to determine the slowest modes of a protein's conformational dynamics [63–68]. tICA can be performed within python packages used to build MSMs, MSMBuilder [69] and pyEMMA [70].

State decomposition via clustering

Once these structural quantities are calculated for each frame, we cluster or combine different frames into microstates via standard clustering algorithms such as k-means, mini-batch k-means [71], k-mediods [72], k-centers [73] or hierarchical methods [74]. Protein conformations with similar values for the slowest tICs or features if tICA is not performed will be combined together leading to distinct states. Clustering is an ideal

choice as it enables researchers to define states without any bias. In our work we use mini-batch k-means to obtain microstates for all MSMs. The mini-batch k-means algorithm works as follows:

1. Assign a number of clusters, k. Randomly choose k observations from the dataset and use these as the initial means.

2. Randomly choose b datapoints, where b is a pre-decided batch size. Assign these chosen data points to one of the k clusters by associating every datapoint with the nearest mean. A common way to determine distances in high-dimensional data is Euclidean distance.

3. The centroid of each of the k clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

This method is computationally less expensive since Step 2 is only performed on a random sample of data as opposed to all data in a typical k-means implementation.

Transition matrix estimation

Next, the transitions between the clustered states are determined at a chosen lagtime τ to enforce the Markovian property. The Markovian property ensures that the probability of future states/conformations are dependent entirely on the present state of the protein and not on the states/conformations that preceded it. At the chosen τ a count matrix $C(\tau)$ can be defined from the data, where $C_ij(\tau)$ represents the number of transitions from state i to state j. $T(\tau)$ is the transition probability matrix, $T_{ij}(\tau)$ being the probability of transition between state i and state j. For an N state MSM, both $C(\tau)$ and $T(\tau)$ are $N \times N$ matrices, related as,

$$T_{ij}(\tau) = \frac{C_{ij}(\tau)}{C_i\tau} \tag{1.3}$$

$$\text{where } C_i(\tau) = \sum_j C_{ij}(\tau) \tag{1.4}$$

Rather than $T(\tau)$, it is often more useful to determine K, an $N \times N$ rate matrix for a continuous-time Markov process, which enables one to estimate decay or rates of conformational changes from one state to another. K can be estimated from discrete-time observations that is, $T(\tau)$ using a maximum likelihood estimator approach [75,76] or observable operator models [77]. At this stage microscopic reversibility among the states according to detail balance can also be introduced. A Markov process is reversible when the rate matrix, K, or the transition probability matrix, $T(\tau)$ satisfies the detailed balance condition with respect to

a stationary distribution, π, towards which the process relaxes over time,

$$\pi K = 0 \text{ or } \pi T(\tau) = 0 \tag{1.5}$$

$$\pi_i K_{ij} = \pi_j K_{ji} \text{ or } \pi_i T_{ij}(\tau) = \pi_j T_{ji}(\tau), \forall i \neq j \tag{1.6}$$

It follows that for an N state Markov state model,

$$p_i(t+\tau) = \sum_{j=1}^{N} p_j(t) p_{ji}(\tau) \tag{1.7}$$

where $p_i(t)$ is the probability that the protein is in state i at time t. To determine the state of the protein at time $t+\tau$, $p(t+\tau)$ which is a vector of probabilities of occupying all of the N states at time $t+\tau$ can be obtained using equation (1.8),

$$p(t+\tau) = p(t)T(\tau) \tag{1.8}$$

The transition probability matrix T(τ) can be decomposed into its eigenfunctions and eigenvalues as,

$$\phi_i T(\tau) = \lambda_i T(\tau) \tag{1.9}$$

If the eigenvalues, λ_i, are arranged in a descending order, the first (also the largest) eigenvalue, λ_1=1 and is the sum of the equilibrium probabilities of N states. The rest of the eigenvalues $\lambda_{i>1} < 1$ correspond to the relaxation time scales, t_i as follows,

$$t_i = -\frac{\tau}{\ln \lambda_i} \tag{1.10}$$

The top m-eigenvalues provide the best estimate of the m-slowest timescales for the protein's dynamics. A more detailed mathematical description of MSMs can be accessed from the referenced literature articles [52, 78, 79].

MSMs are now widely used frameworks for the analysis of MD datasets and multiple software suites provide MSM construction as well as analysis capabilities [69, 70].

1.3 Generalized matrix Rayleigh quotient

A Generalized matrix Rayleigh quotient (GMRQ) score is used to choose the best parameters for MSM construction during the many steps discussed in the previous section. Some of these parameters are the choice of featurization scheme, number of features used, number of dimensionality reduced components [80],

clustering algorithm and number of clusters among others. Once the simulation dataset is obtained, GMRQ scores for multiple MSMs with different hyper-parameters are determined and the set of hyper-parameters which maximize GMRQ are chosen to build the final MSM for analysis. This approach has been used by some of the recent protein conformational dynamics studies to choose the best parameters to build an MSM for analysis of MD simulation datasets [55, 65, 66, 81–83].

Before the construction of MSM, we do not know the best parameters which would provide an estimate of the slowest timescale dynamic modes of the protein. Thus, the true eigenfunctions of the transition probability matrix, $T(\tau)$, of the MSM are not known *a priori*. A trial MSM is built to obtain a guess and based on the variational principle of conformational dynamics [84, 85] a quantitative estimate (GMRQ) of the m-slow processes captured by the first m-eigenfunctions of the trial MSM is obtained. The GMRQ score is the sum of the m largest eigenvalues, $\hat{\lambda}_i$, GMRQ $= \sum_{i=1}^{m} \hat{\lambda}_i$. It has been proven that the upper limit for GMRQ is the score for MSM with the true eigenfunctions. In other words, the sum of the true eigenvalues GMRQ $\leq \sum_{i=1}^{m} \lambda_i$ [86, 87] is the upper limit of the GMRQ score.

In order to avoid over-fitting to the MD simulation dataset, a k-fold cross-validation approach is used. The dataset is split into the training and testing dataset. The GMRQ is first maximized on the training dataset by building an MSM on only this data, called the training score and then the coefficients are used to obtain the test GMRQ score over the testing data. This is repeated k times and a mean is reported as the final score. The hyper-parameters which yield the highest mean of k-fold cross validated testing data score can be chosen to construct the MSM. Usually, 5-fold cross validation is used and the dataset is split equally into training and testing data. For all future purposes, the mean of the 5-fold cross validated test data score is referred to as the GMRQ score for a given set of hyper-parameters.

Overall, a high GMRQ indicates an better model to describe the dynamics of the protein from the underlying conformational landscape.

1.4 Dissertation overview

The overarching objective of this work is to use MD simulations as a predictive and validation technique alongside biophysical spectroscopy experiments, especially DEER spectroscopy of membrane proteins.

- Chapter 2 describes the development of a methodology we call *OptimalProbes* for the prediction of ideal residue-pairs to probe in DEER experiments. This work was published in *The Journal of Physical Chemistry B* [88].

- Chapter 3 demonstrates how *OptimalProbes* can be extended to related experimental techniques such

LRET, TTET and Trp-Tyr fluorescence quenching experiments. Work presented in this chapter is published in *The Journal of Physical Chemistry B* [89].

- Chapter 4 utilizes *OptimalProbes* for predicting the conformational dynamics of a novel Multi-Facilitator Superfamily membrane protein, $PepT_{So}$, a bacterial di/tri-peptide transporter. This work was published in *ACS Central Science* [90].

- Chapter 5 examines many reasons for a mismatch between residue-pair distance distributions from MD simulations and DEER experiments, which might limit the integration of computational and experimental studies for the study of protein dynamics.

- Chapter 6 discusses some of the remaining challenges and opportunities in integrative modeling of protein conformational heterogeneity and our thoughts on the necessity for tools and platforms for scientists to share data and build models.

Chapter 2

Predicting Optimal DEER Label Positions to Study Protein Conformational Heterogeneity[1]

2.1 Overview

Double electron-electron resonance (DEER) spectroscopy is a powerful experimental technique for understanding the conformational heterogeneity of proteins. It involves attaching nitroxide spin labels to two residues in the protein to obtain a distance distribution between them. However, the choice of residue-pairs to label in the protein, requires cautious thought and experimentalists are required to pick label positions from a large set of all possible residue-pair combinations in the protein. In this paper, we address the problem of the choice of DEER spin label positions in a protein. For this purpose, we utilize all-atom molecular dynamics simulations of protein dynamics, to rank the sets of labeled residues pairs in terms of their ability to capture the conformational dynamics of the protein. Our design methodology is based on the following two criteria: 1) an ideal set of DEER spin label positions should capture the slowest conformational change processes observed in the protein dynamics and 2) any two sets of residue-pairs should describe orthogonal conformational change processes to maximize the overall information gain and reduce the number of labeled residues pairs. We utilize Markov state models of protein dynamics to identify slow dynamical processes and a genetic algorithm based approach to predict the optimal residue-pair choices with limited computational time requirement. We predict the optimal residue-pairs for DEER spectroscopy in β_2 Adrenergic Receptor, C-terminal domain of calmodulin and peptide transporter $PepT_{So}$. We find that our choices are ranked higher than those used to perform DEER experiments on the proteins investigated in this study. Hence, the predicted DEER residue-pair choices determined from our method provide maximum insight into the conformational heterogeneity of the protein while using the minimum number of labeled residues.

[1]This chapter is reproduced with permission from Mittal S, Shukla D. Journal of Physical Chemistry B. 2017; 121(42):9761-9770. Copyright 2017 American Chemical Society.

2.2 Introduction

Double Electron-Electron Resonance (DEER) is an experimental technique based on electron paramagnetic resonance (EPR) which has become a crucial resource for protein structure determination [91, 92]. Using Site-Directed Spin Labeling (SDSL) [93, 94], two nitroxide spin labels are attached to two cysteine mutated residues. These spin labels include 1-oxyl-2,2,5,5-tetramethyl-pyrroline-3-methyl)methanethiosulfonate (MTSSL), iodoacetamide-PROXYL (IA-PROXYL), unnatural amino acids p-acetyl-L-phenylalanine and 2,2,6,6-tetramethyl-piperidine-1-oxyl-4-amino-4-carboxylic acid [92] and a spin labeled lysine (SLK-1) [95,96], all of which possess an unpaired electron leading to the formation of a magnetic dipole to allow for dipole coupling measurements. The DEER experiment measures the decay of the dipolar coupling between the unpaired electrons of the attached spin labels, which is then processed to obtain the distance distributions between the labeled residues on the protein. Alexander et al. proposed that the spin label distances could be converted to the distance between the C_β-C_β atoms of the residues using a motion-on-a-code model [97]. These distances were then incorporated into the structure prediction suite Rosetta to generate low-RMSD protein structure predictions for T4-lysozyme and other proteins [98]. Roux and Islam have developed a restrained-ensemble molecular dynamics (MD) simulations method that incorporates DEER distance distributions to refine the structural predictions of proteins, with T4-lysozyme as the benchmark protein [99–101]. Hence, EPR/DEER has been a popular choice of technique to study a variety of biological processes such as folding of T4 lysozyme [91], conformational heterogeneity of enzyme HIV-1 protease [102], transitions in the intrinsically disordered protein IA_3 [103], allosteric affects in Hsp70 chaperone [104] and nucleic acids [105]. Since the technique is not limited by the size of the protein, it has also been employed to study conformational dynamics of membrane proteins [106] such as a lipid flippase from *Escherichia coli* MsbA [107], lactose permease LacY [108], GPCRs Rhodopsin [109] and β_2 Adrenergic Receptor [110], a peptide transporter $PepT_{So}$ [111], an ABC transporter [112] among others.

Prior to performing the DEER experiment, key challenges faced by the experimentalists include 1) the effect of the conformational dynamics of the spin labels on the obtained distance distributions and 2) the choice of residues for labeling or the residue-pair distances to measure. The introduction of spin labels makes the obtained distance distributions dependent on both the dynamics of the spin labels and the protein backbone [92]. Several studies have reported approaches that include a comprehensive rotamer orientation search of the spin label to incorporate the effect of the spin label conformations from the protein structural studies [108,113]. However, the question - which residue-pairs to label with the nitroxide molecules to obtain DEER distributions from the experiment - is not trivial. If we consider a protein of size R residues, there are $R(R-1)/2$ residues pair distances to choose from. This product is the number of all possible residue-pairs

in the protein. Further, DEER experiments usually entail measurement of multiple distance pairs, say k residue-pairs. This leads to ${}^{R(R-1)/2}C_k$ possibilities to choose k pairs from R residues. This number is usually too large for any protein of biological interest. Given the constraints of time and resources, it is not possible to try all possibilities. Thus, multiple residue-pairs have to be chosen before the experiment is performed. Additionally, an optimal choice of residue-pairs is one where the number of pairs are minimum and each pair captures structural transitions of the proteins which are different from others.

MD simulations have been used extensively to capture the long timescale conformational dynamics associated with protein folding [114,115] and conformational change [45,55,63,110,116–122]. MD Simulation datasets could be used to extract residue-pair distances which provide insight into distinct conformational states of the protein and identify residue-pairs that play a critical role in the conformational transitions of proteins. These residue-pairs are associated with the slow dynamical processes or the high free energy barriers observed in the conformational free energy landscape of the protein. Here, we hypothesize that residue-pairs involved in the slow functional dynamics could also be used to perform EPR/DEER experiments. The purpose of DEER spectroscopy or any other biophysical experiment is to describe the structural changes (such as the breaking and forming of residue contacts) associated with the protein function and these residue-pairs are associated with the functional dynamics of the protein. Therefore, MD simulations could be utilized to solve the problem experimentalists face before the DEER experiment is performed. It is also possible to use MD simulation datasets obtained via accelerated MD [25], steered MD [19], umbrella sampling [17], metadynamics [22] or replica exchange [21] techniques for this purpose.

However, the task of choosing spin label positions is not as simple as measuring the relaxation rate for all residue-pairs using simulation datasets and identifying the optimal set based on the residue-pairs with slowest relaxation kinetics. Imagine a scenario, where two chosen residue-pairs describe the same dynamical process (such as rocking of a membrane helix in a protein) or their motion is highly correlated. In this case, the information provided by the two DEER measurements would not be independent of each other and lead to redundancy. Therefore, the ideal set of residue-pairs for spin labeling should not only describe the slowest dynamical processes but also provide orthogonal information. In this paper, we have reported a protocol to predict optimal pair wise EPR/DEER experimental label positions from the MD simulation datasets using these two criteria.

Markov state models (MSMs) provide a natural framework for identifying the minimal set of orthogonal residue-pairs associated with slowest dynamical processes observed in simulation datasets. MSMs coarse grain the conformational dynamics of a protein by eliminating the fast conformational dynamics and retaining only the slowest dynamical processes. [55,79] The protocol constructs an MSM from the simulation dataset

using the distances between a set of residue-pairs as the geometric metric for state decomposition. Multiple such MSMs are constructed using different sets of residue-pairs as their metric. Then we assign a generalized matrix Rayleigh quotient (GMRQ) score [84, 85] to each MSM. A higher GMRQ score for an MSM indicates that the MSM is able to capture the slow timescale processes of the underlying protein dynamics. residue-pairs used to construct the MSM with high GMRQ score are chosen to be the optimal residue-pairs to perform DEER experiments. In addition to an exhaustive residue-pairs search protocol, a genetic algorithm based improvement over the exhaustive search method is demonstrated. This improved approach reduces the computation time requirement. A detailed description of the implementations are discussed in the Theory and Methods section.

2.3 Methods

Optimal DEER label positions prediction method

First, we need to determine whether a residue-pair distance can be measured using DEER spectroscopy, the simulation data is perused to determine if two residues are always within a predetermined range. The range can be obtained based on the instrumentation available to perform the DEER experiments. Measurement in the range of 18-60 Å for membrane proteins and upto 100 Å in cytoplasmic proteins is possible [92]. We use the C_α distances between two residues for all distance calculations. This may not be the best estimate of the distances obtained from DEER experiments as we neglect the effect of spin probe size on the measurement but it presents a rational choice for obtaining inter-residue distances from the simulation datasets. The distance distributions estimated from experiments depend on the choice of spin label, dynamics of the many dihedral angles in the spin label and the choice of the lipid bilayer mimetic viz. detergent micelles, lipid bicelles, nanodiscs or liposomes [92, 123, 124]. Simulations on the other hand are free from bias due to inclusion of spin labels or the bilayer mimetic. The datasets employed in this study represent simulations of membrane proteins embedded in a lipid layer or cytoplasmic protein in a water box. Hence, the C_α-C_α distances from the simulation datasets are a measure of the true dynamics of the protein.

Consider $G(V, E)$, an undirected graph with vertex set V and edge set E. V comprises of vertices v_i where $1 \leq i \leq R$ (R is the number of residues in the protein of interest) and an edge e_{ij} indicates the residue-pair distance between v_i and v_j is measurable through EPR/DEER experiment technique. If the C_α-C_α distance of the residue-pair i and j is within the specified range, an edge is added between vertices v_i and v_j in the graph G. This graph is stored as an $R \times R$ adjacency matrix, A whose elements A_{ij} are 1 if there is an edge between v_i and v_j; 0 otherwise. A is a symmetric matrix and only the upper-triangular or alternatively the

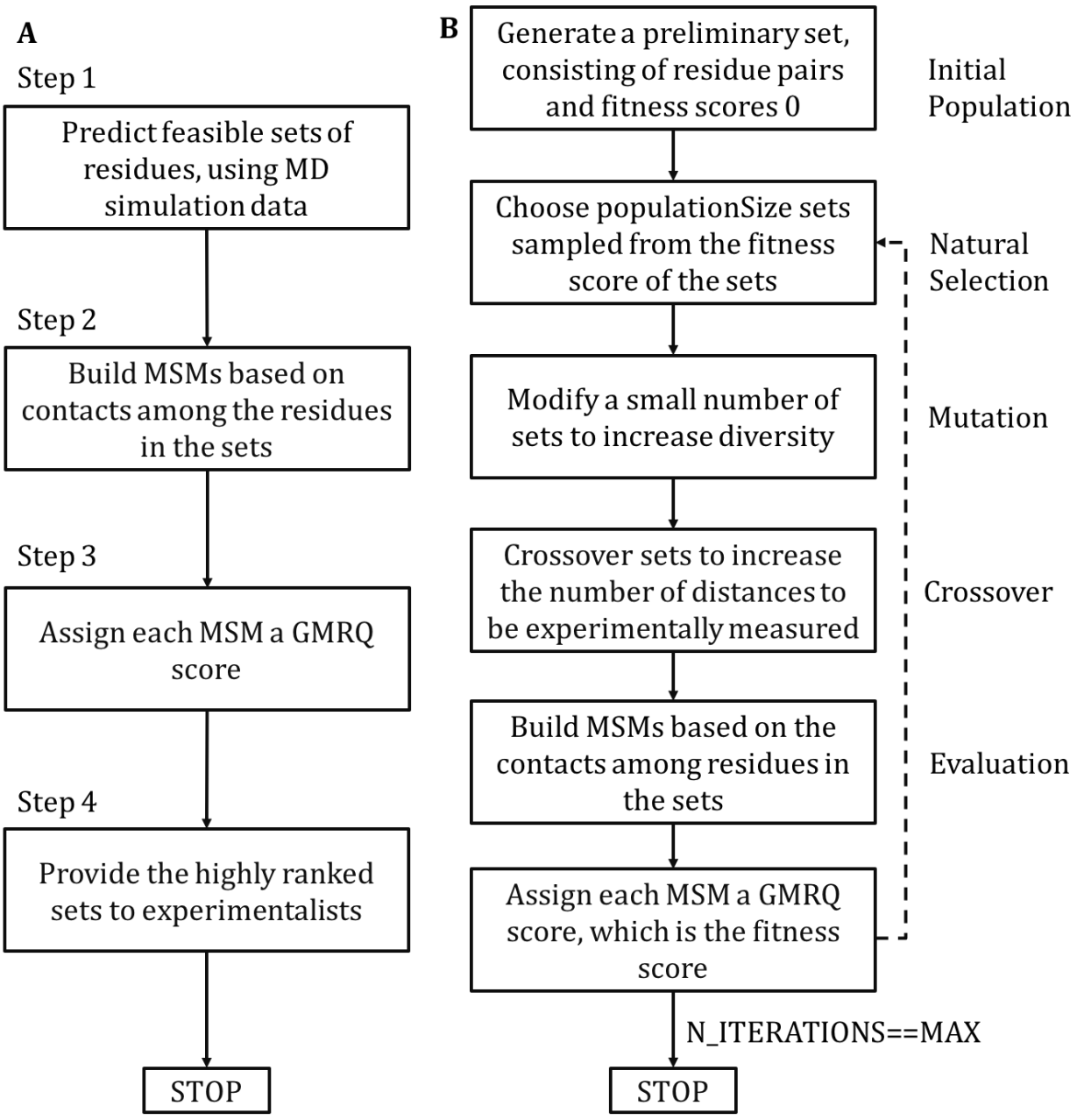


Figure 2.1: (A) Workflow for the optimal EPR/DEER label positions prediction protocol. (B) An improved workflow following the scheme of a genetic algorithm for optimal label positions prediction.

lower-triangular matrix is needed. At this stage the matrix A is densely populated and would yield too many residue-pair possibilities if scanned. Hence, we attempt to make A sparse by including topology information of the protein and some DEER experiment constraints.

1. Each immovable secondary structural element in the protein is allowed only one label position; for instance if v_i and v_j residues are part of the same coil, $e_{ij} = 0$.

2. Non-solvent exposed regions in the protein are usually inaccessible for spin label modification as they will introduce steric clashes with nearby residues and greatly alter the structural dynamics of the protein. Accordingly, edges which involve inaccessible residues are removed from the graph.

3. Edges corresponding to the residue-pairs in a membrane protein with one residue on the extracellular side of the protein and other on the intracellular side of the protein are eliminated from the graph G due to the large distances between them [111].

On the basis of above constraints, the adjacency matrix is then parsed to obtain M feasible sets, $S_1 \cdots S_M$, of positions which allow DEER distance measurements among them. If $S_1 = \{v_1, v_2 \cdots v_n\}$, it indicates that $n(n-1)/2$ DEER distance measurements occur, between $(v_1, v_2) \cdots (v_1, v_n), (v_2, v_3) \cdots (v_{n-1}, v_n)$. The number of residue-pairs in each set $S_1 \cdots S_M$ differs.

Finally, an MSM is constructed for each set of residue-pairs. The MD dataset is clustered into 200 states based on the C_α-C_α distances of the residue-pairs in the set. Hyper-parameters other than the choice of residue-pair distances for clustering are not varied to keep the dimensionality of the problem small. The choice of hyper-parameters are provided in Supplementary Table 2.1 for the proteins investigated in this paper. The MSMs are constructed using the MSMBuilder3.4 [69] package and GMRQ scores are obtained using the Osprey package [125]. When arranged in a descending order of the GMRQ score for each set, the set S_m $(1 \leq m \leq M)$, corresponding to the highest score provides the optimal choice of DEER label positions on the protein (Figure 2.1A).

This method has been demonstrated on three biological proteins, β_2 Adrenergic Receptor (β_2AR), C-terminal domain of calmodulin and bacterial peptide transporter $PepT_{So}$. MD simulation datasets for these three proteins are available and experimental observations are available for β_2AR [110] and $PepT_{So}$ [111]. The predicted residue-pair choices are reported in the Results and Discussion section.

Genetic algorithm for obtaining the optimal set

Observations from the above protocol indicate that MD simulation datasets can serve as a good resource to predict optimal DEER label positions. However, the protocol relies on an exhaustive search. It involves

extracting all possible sets of residue-pairs, building an MSM for each set and assigning a GMRQ score to each of the MSMs. In most cases, the number of predicted sets will be large. Therefore, significant computational resources are required to build MSMs for each set while identifying the hyper-parameters that maximize the GMRQ. The computational time requirement could be circumvented by using a genetic algorithm scheme.

Genetic algorithms are well established and widely used algorithms with multitude of applications. [126,127] A standard genetic algorithm scheme mimics the evolution of a species with the aim of maximizing its 'fitness' or survival probability. The scheme starts with an initial population of species. Each member of the population has an associated 'fitness score' which decides whether it will survive 'natural selection' to continue onto the next generation or perish. It is based on a pre-decided metric which is the desired trait in the resulting population. It is not necessary that all members of the population are propagated. A small number also undergo 'mutations' which introduce upward or downward change in the surviving capability of the species. In the 'crossover' step, a few selected members are joined with each other to finally obtain the new generation. During the evaluation step, each member of the present generation is assigned a new fitness score. Once the species have been assigned a fitness score the next iteration begins from 'natural selection' again. These steps are followed until a termination condition is reached.

The key idea is to begin with only a small number of sets of residue-pairs to construct MSMs. Based on the GMRQ scores for these MSMs, new sets are chosen for subsequent iterations. Here, we describe the series of steps of the improved method to predict the optimal residue-pairs for DEER experiments, via a genetic algorithm scheme (Figure 2.1B).

1. **Identify the set of all feasible residue-pairs** Q, from the MD simulation dataset, using the constraints listed in the previous section. Elements of set Q are $Q_i = (v_x, v_y)$ where $1 \leq i \leq \mid Q \mid$, v_x and v_y are residue numbers with $1 \leq x \neq y \leq R$, R is the number of residues in the protein.

2. **Assign each element of Q a fitness score**, $f_i = 0$ where $1 \leq i \leq \mid Q \mid$ for the first iteration.

3. **Initial Population**: Choose *populationSize* (a number chosen in advance) elements from Q randomly. This is the current generation set G_0, where $G_{0,i} = (v_x, v_y)$ and $1 \leq i \leq populationSize$. Build MSMs featurized on the C_α distance of the residue-pairs in each element of G_0 and assign a GMRQ score to these MSMs. The GMRQ score is now the fitness score assigned to the elements of G_0.

4. **Natural Selection**: For the new iteration $ITER$, choose *populationSize* elements from G proportional to their newly assigned fitness scores, for the new generation set G_{ITER}.

5. **Mutation**: Change $(mutationPercent * populationSize)/100$ items from set G_{ITER} and replace them with randomly chosen residue-pairs from Q not already present in G_{ITER}.

6. **Crossover**: Choose $(crossoverPercent * populationSize)/100$ pairs from the set G_{ITER} to combine with each other and add them to the current population. Crossovers are responsible for increasing the number of residue-pairs in the predicted sets.

7. **Evaluation**: Build MSMs featurized on the C_α distance of the elements of G_{ITER} and assign a GMRQ score (aka fitness score) to these MSMs. This step is the same as steps two and three in the exhaustive search method. Further, obtain a scaled fitness score based on the number of residue-pairs in each element. This allows to lower rank choices which predict larger number of distance measurements. This is required since one of the goals of the method is to minimize the number of distance measurements required in DEER experiments.

8. **Stop**: This process is continued starting from natural selection again, until the maximum number of iterations, *N_ITERATIONS* are achieved.

The algorithm parameters *populationSize*, *mutationPercent*, *crossoverPercent* and *N_ITERATIONS* are user defined quantities which will have impacts on the running time as well as convergence of the algorithm. Population size of 20, mutation and crossover rates of 50% and 20%, respectively, were found to converge faster and have been used in this study. The genetic algorithm approach is a heurestic based optimization, which means whether the global optimum is reached cannot be guaranteed and the program may make choices during its run which will cause it to retreat from an approaching optimal solution. If it is an unlucky choice, then a good solution may be reached after a large number of iterations.

These algorithms have been implemented in the Python language and use Numpy [128], MDTraj [129], MSMBuilder3.4 [69] and Osprey package [125] as dependencies. Various user inputs such as protein topology information, inaccessible residues, undesired residue-pairs, DEER distance measurement constraints based on the instrument and technique, lower and upper bound on number of measurements, genetic algorithm parameters and MSM construction hyper-parameters can be specified in the user's input data to the program. These details are provided along with the program source code and also included in the Supporting Information. The implemented program can run MSM construction in parallel, if multiple processors are available. Combined with automatic parallelization, the genetic algorithm is considerably fast in predicting optimal residue-pairs for DEER experiments.

2.4 Results

Predicting optimal label positions for β_2 Adrenergic Receptor

Previously published all-atom MD simulation dataset of β_2AR [130] was used for analysis in the current work. Specifically, twenty-four simulations of agonist-bound protein with the Nb80 nanobody removed initiated from the active crystal structure (PDB: 3P0G [131]) with protonated residue Asp 130 were used. The individual simulation time range from 2 μs up to 11.4 μs, with the cumulative dataset used being $\sim$127 μs.

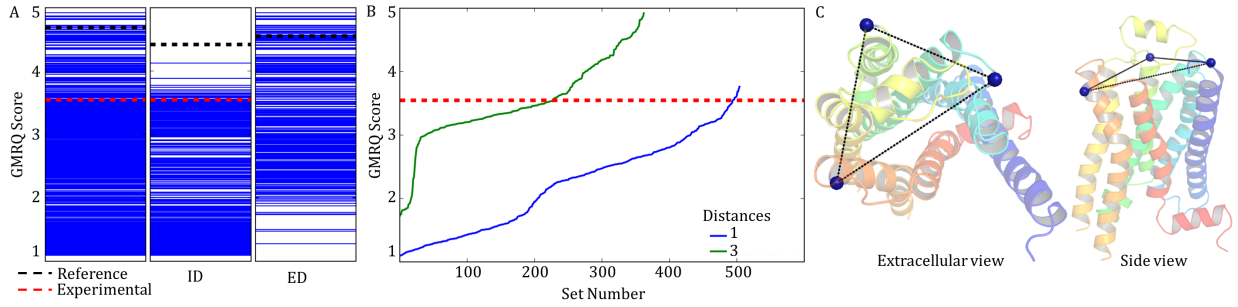


Figure 2.2: (A) GMRQ scores corresponding to MSMs constructed with the predicted sets of residue-pairs for DEER experiments in β_2AR. The first column indicates scores for 868 MSMs. The second and third columns indicate the MSM scores which have only intracellular distances (ID) and extracellular distances (ED), respectively. The horizontal red line is the GMRQ score referring to the MSM constructed using the contact distance of residues Asp149 and Leu266; the residue-pair for which DEER experimental distance distribution is available [110]. The horizontal black lines correspond to a reference MSM constructed using inter-residue contacts on both intracellular and extracellular, or either domains of the protein. (B) GMRQ scores of the 868 predicted sets of residue-pairs differentiating the number of residue-pairs (or distances) measured in the set. (C) A cartoon representation of β_2AR showing the set of residue-pairs ranked highest by our method.

First, we constructed and scored 868 MSMs and the obtained scores range between 1.09 and 4.92 (Figure 2.2A) for all possible probe sets. An MSM based on the distance chosen by Manglik et al. for DEER experiments on the β_2AR [110] was built and scored (red line in Figure 2.2A,B). The experimental residue-pair choice was optimal in choosing a single distance to measure. However, it can be seen in Figure 2.2B that several residue-pair sets that involve three distances as opposed to one distance measured in the DEER experiment have a higher score than the experimental residue-pair. We also built an MSM with all intracellular and extracellular distances as metric and its score (black line in Figure 2.2A) was used as a reference, it captures the maximum possible conformational heterogeneity in β_2AR among all possible residue-pairs accessible for DEER experiments. The maximum score for the three distances was comparable to the reference MSM indicating that additional distance measurements will not lead to an information gain.

Predicted choices with residue-pair distances on the intracellular side or the extracellular side, are shown separately in the second and third column of Figure 2.2A. It can be seen that the highest set of residue-pairs is on the extracellular side of the protein and comprises of 3 distances. These distances involve residues on

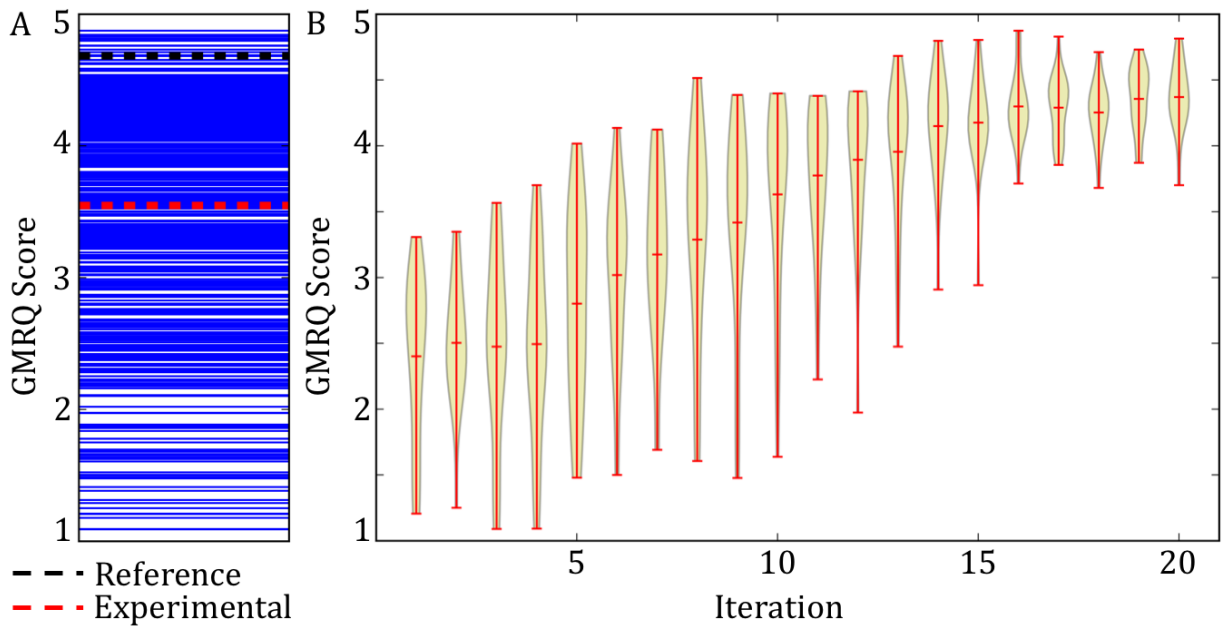


Figure 2.3: (A) GMRQ scores corresponding to MSMs corresponding to MSMs constructed with the predicted sets of residue-pairs for DEER experiments in β_2AR using the genetic algorithm approach. (B) Violin plot demonstrates the increase in GMRQ scores over 20 iterations of the genetic algorithm based method. The vertical red lines indicate the range of the GMRQ scores within each iteration and the horizontal mark in the middle is the mean of the GMRQ scores for the current iteration.

helices 4, 6 and on the loop joining helices 2 and 3 (Figure 2.2C). It has been reported by Kohlhoff et al. and Shukla et al. the activation of the protein occurs initially with helix 6 moving away from helix 3 [47, 119]. Further, since the experimental DEER residues are on the intracellular side of the protein (on helices 4 and 6 [110]), we extracted the highest ranked intracellular choice. This choice of residue-pairs involves residues on helices 3, 4 and 7 (Supplementary Figure 2.2). As reported in literature, these are the helices which are involved in the activation of β_2AR. RMSD of the NPxxY region on helix 7 is a distinct characteristic in the inactive and intermediate states [47, 119, 130]. Thus, the residue-pairs chosen from our method are involved in slow processes in the protein which is its functional switching between active and inactive states.

The exhaustive search method requires construction of 868 MSMs. The limiting step in this method is the MSM construction process that can take large amounts of computation time for long timescale MD simulation datasets. For the genetic algorithm scheme, we used a population size 20, 50% mutation rate and 20% crossover rate. Commensurate GMRQ scores were obtained with smaller number of MSM constructions (Figure 2.3A) and the scores obtained show consistent increase with each iteration. The maximum score converges over 20 iterations as shown via a violin plot in Figure 2.3B. The red vertical lines indicate the range of the score in the current iteration. As expected, some of the higher iterations have MSMs that are scored low, but they are eliminated as the number of iterations progress. We conclude that the genetic algorithm

based approach was an improvement over the previous method as it required 480 MSMs as opposed to 868 MSMs. This improved method is sufficient to predict the optimal set of residue-pairs.

Predicting optimal label positions for calmodulin

We have used the previously published MD simulation dataset from Shukla et al. on conformational dynamics of apo and holo C-terminal domain of calmodulin [63]. A similar analysis as for β_2AR in the previous section was performed for the apo-CaM dataset using 455 μs of simulation data. No residues were indicated as inaccessible regions. The lower limit of DEER distance measurements was kept at 5 Å. This value may be too close to perform an actual DEER experiment. This was done because there is no experimental data already available for CaM and we chose this example to demonstrate that our method can be generalized. The results with actual parameters 18-100 Å are provided in the Supplementary Figure 2.5. Thus, the developed protocol can point future experiments to identify structural changes and possible folding pathways of cytoplasmic proteins.

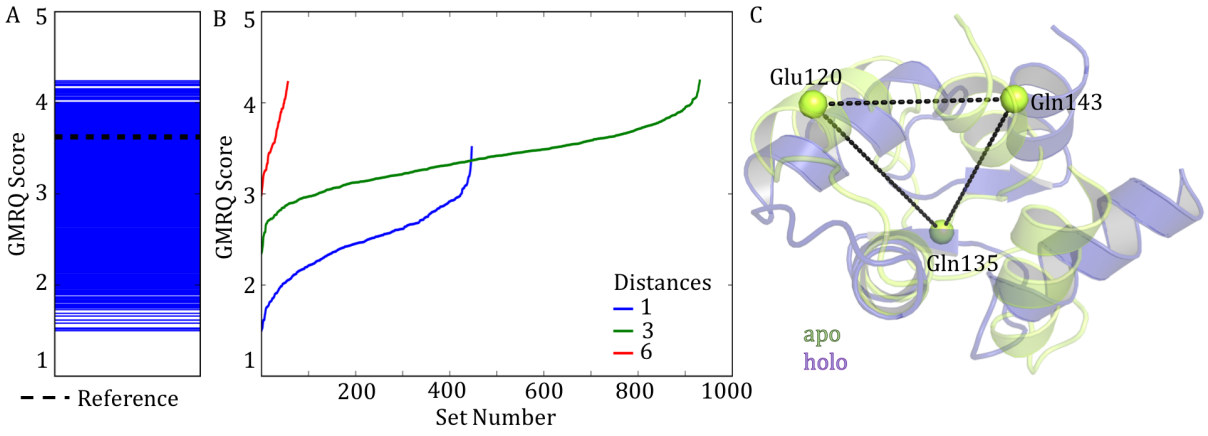


Figure 2.4: (A) GMRQ scores corresponding to MSMs constructed with the predicted sets of residue-pairs for DEER experiments in apo C-terminal domain of calmodulin. The horizontal black line corresponds to a reference MSM constructed using all inter-residue contacts of the protein. (B) GMRQ scores of the 1440 predicted sets of residue-pairs differentiating the number of residue-pairs (or distances) measured in the set. (C) A cartoon representation of the apo (green, PDB: 1CFD [132]) showing the set of residue-pairs ranked highest by our method. The holo (purple, PDB: 1CLL [133]) C-terminal domain of calmodulin is shown for comparison.

Figure 2.4A and Figure 2.4B show the scores for the MSMs based on the 1440 sets of residue-pairs identified using the exhaustive search method. The highest ranked choice of residue-pairs is shown in Figure 2.4C which involves residues Glu120, Gln135 and Gln143. Gln135 is close to one of the Calcium ion ligating residue Asp133 [63], and it may not be suitable for spin label placement due to steric clashes with other ion ligating residues. If due to experimental constraints, it is not possible to insert a spin label on some residues, then these residues can be eliminated appropriately. Our top predicted choice of residue-pairs is widely different from residues involved in the conformational dynamics. However, our second and fourth choices,

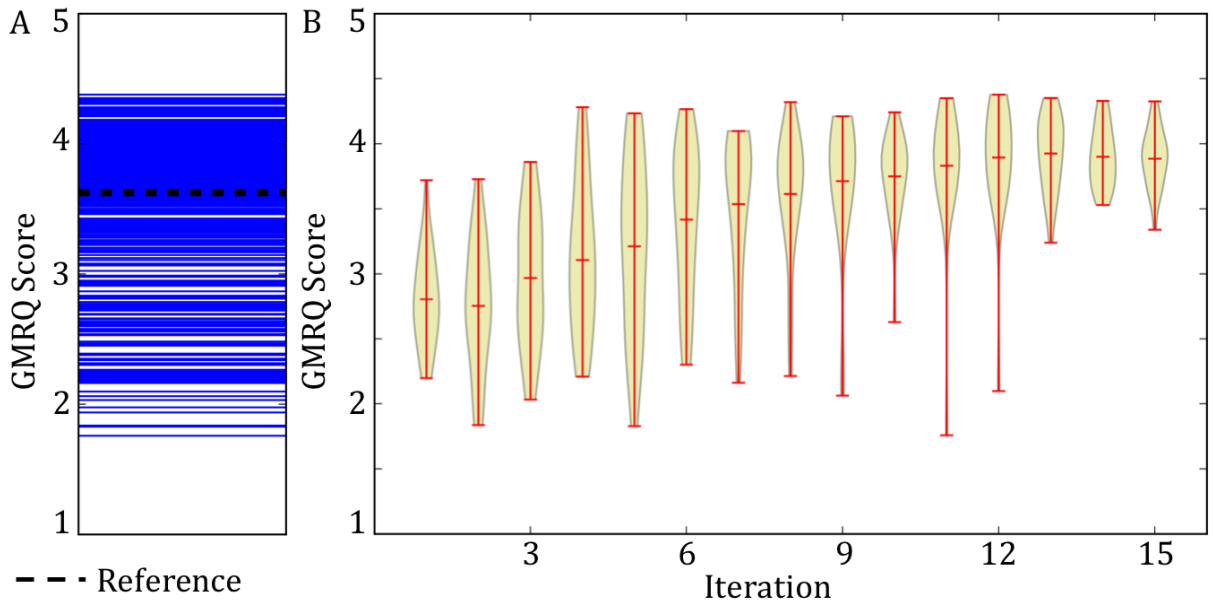


Figure 2.5: (A) GMRQ scores corresponding to MSMs corresponding to MSMs constructed with the predicted sets of residue-pairs for DEER experiments in the apo C-terminal domain of calmodulin using the genetic algorithm approach. (B) Violin plot demonstrates the increase in GMRQ scores over 15 iterations of the genetic algorithm based method. The vertical red lines indicate the range of the GMRQ scores within each iteration and the horizontal mark in the middle is the mean of the GMRQ scores for the current iteration.

including many others in the top rank choices, pick residues on the helix G which is distorted. Residue in its vicinity have been used by Shukla et al. to study the local unfolding of helix G in apo-C-CaM [63]. The genetic algorithm based approach was also used on the apo-C-CaM dataset with the same parameters as for β_2AR - population size 20, 50% mutation rate and 20% crossover rate. Here the GMRQ scores converge in only 15 rounds (Figure 2.5A,B) and required construction of only 360 MSMs.

Predicting optimal label positions for PepT$_{So}$

In this section, we apply our method on the bacterial peptide transporter PepT$_{So}$. In our work discussed in Chapter 4 we performed MD simulations on the protein for $\sim$55 μs using the inward facing crystal structure (PDB: 4UVM [111]) as the starting structure. Multiple conformational intermediate states which are involved in the transition from the inward facing to outward facing state were obtained.

Figure 2.6A shows the result of our exhaustive search method for 2023 MSMs. These residue sets include 1,3 or 6 residue-pairs on either the extracellular or the intracellular side the protein (Figure 2.6B). The top choice is on the intracellular side and is shown on the protein in Figure 2.6C. Proteins belonging to the MFS family have a common conformational change characteristic which is the alternating access mechanism of the protein involving helix motions on both sides of the protein [134, 135]. Hence, we performed a preliminary

analysis to obtain 2325 more sets of residue-pairs. These new sets include label positions on both intracellular and extracellular sides of $PepT_{So}$. This was done by combining the best choice on the intracellular side with all extracellular choices and vice versa. As expected, we observe that all of the mixed choices (Figure 2.7A,B) are usually higher ranked. The highest ranked choice is indicated on the protein from both intracellular and extracellular views in Figure 2.7C.

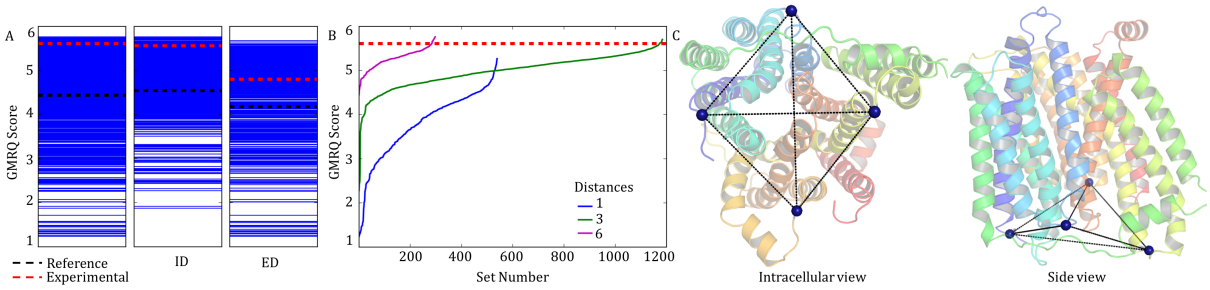


Figure 2.6: (A) GMRQ scores corresponding to MSMs constructed with the predicted sets of residue-pairs for DEER experiments in $PepT_{So}$. The first column indicates scores for 2023 MSMs. The second and third columns indicate the MSM scores which have only intracellular distances (ID) and extracellular distances (ED), respectively. The horizontal red lines are the GMRQ score referring to the MSM constructed using the 8 residue-pairs distances; the residue-pairs for which DEER experimental distance distribution is available [111]. The horizontal black lines correspond to a reference MSM constructed using inter-residue contacts on both intracellular and extracellular, or either domains of the protein. (B) GMRQ scores of the 2023 predicted sets of residue-pairs differentiating the number of residue-pairs (or distances) measured in the set. (C) A cartoon representation of $PepT_{So}$ (PDB: 4UVM [111]) showing the set of residue-pairs ranked highest by our method.

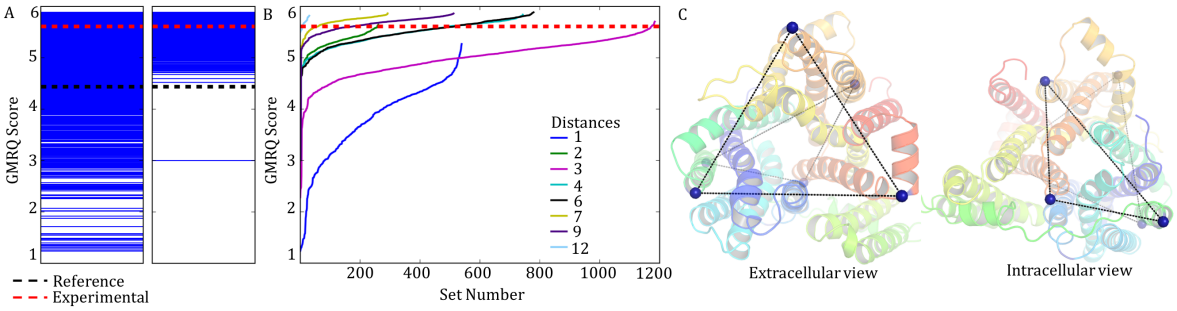


Figure 2.7: (A) GMRQ scores corresponding to MSMs constructed with the predicted sets of residue-pairs for DEER experiments in $PepT_{So}$. The first column indicates scores for 4348 MSMs. The second column indicates the MSM scores which have both intracellular distances and extracellular distances. The horizontal red lines are the GMRQ score referring to the MSM constructed using the 8 residue-pairs distances; the residue-pairs for which DEER experimental distance distribution is available [111]. The horizontal black lines correspond to a reference MSM constructed using inter-residue contacts on both intracellular and extracellular, or either domains of the protein. (B) GMRQ scores of the 4348 predicted sets of residue-pairs differentiating the number of residue-pairs (or distances) measured in the set. (C) A cartoon representation of $PepT_{So}$ (PDB: 4UVM [111]) showing the set of residue-pairs ranked highest by our method on both, extracellular (left) and intracellular (right) sides of the protein.

In Chapter 4 we determine that transmembrane helices 1, 2, 4, 7, 8, and 10 are involved in the transition from inward facing to outward facing states. As shown in Figure 2.7C, the probe positions identified are on helix 4, 10 and 11 on the extracellular side and helix 4, 9 and 11 on the intracellular side of the protein. A comparison of an inward facing structure (PDB: 4UVM [111]) and outward facing structure in Supplementary Figure 2.6 clearly illustrates how the predicted residues are on the crucial helices. It has also been reported that

2 helices in $PepT_{So}$, helix A and B are not present in other MFS family of proteins and do not contribute during the alternating access mechanism of conformational change. We too observe that none of our highly ranked choices predict residues on these two helices.

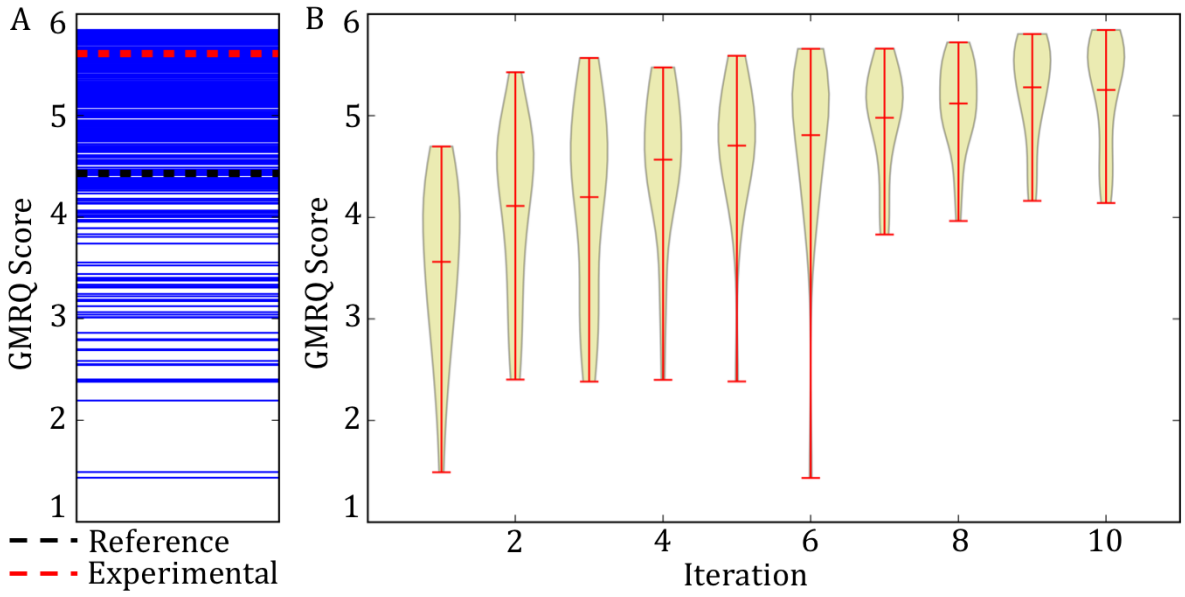


Figure 2.8: (A) GMRQ scores corresponding to MSMs corresponding to MSMs constructed with the predicted sets of residue-pairs for DEER experiments in $PepT_{So}$ using the genetic algorithm approach. (B) Violin plot demonstrates the increase in GMRQ scores over 10 iterations of the genetic algorithm based method. The vertical red lines indicate the range of the GMRQ scores within each iteration and the horizontal mark in the middle is the mean of the GMRQ scores for the current iteration.

The experimental DEER spectroscopy data for $PepT_{So}$ measures eight residue-pair distances. [111] The MSM corresponding to these 8 inter-residue distances is ranked high (red line in Figure 2.6A,B and Figure 2.7A,B). These eight residue contacts together can capture the slow timescale dynamics of the protein. However, the experimental choice involves many redundant residue-pairs which do not contribute any new information as compared to that captured by another distance measurement. Using our method, we have predicted sets of residue-pairs which are ranked higher than the experimental residue-pairs' MSM. The potential of our method is demonstrated by the fact that some of these higher ranked choices involve less than 8 inter-residue distances. Clearly, our choices are optimal and provide a comprehensive picture of the protein's structural changes with a minimal set of experimental spin labels. Finally, the DEER distance distribution for our top choice were obtained using RotamerConvolveMD [108] python library and the resulting histograms are included in the Supplementary Figure 2.7. We have also looked at some of the lowest ranked choices (Supplementary Figure 2.8); the lowest ranking choices on the extracellular and intracellular side of the protein pick a pair of residues only on the 6 helices which are part of the N-bundle of the protein. Hence,

they do not capture the relative motions of the helices from the C-bundle and the N-bundle that allow substrate transport in the MFS family of proteins.

The genetic algorithm based approach was also used on the $PepT_{So}$ dataset with the same parameters as for β_2AR and apo-C-CaM - population size 20, 50% mutation rate and 20% crossover rate. Here the GMRQ scores converge in 10 rounds and involves construction of only 240 MSMs (Figure 2.8A,B).

2.5 Discussion

The exhaustive search method is a proof of concept of the idea that residue-pairs involved in the slow functional dynamics can be used to perform DEER experiments. We have shown that the reported method is an effective way to not only choose DEER experiment label positions but also minimize resource utilization as it predicts the least number of distances to be measured. Our protocol is not specific to any protein and can be used to study conformational heterogeneity in different types of proteins and also provides experimentalists with multiple good choices. Thus, if a certain point mutation leads to loss of function of the protein, another choice can be used for DEER experiments. Furthermore, we have demonstrated that the genetic algorithm based optimization is efficient in picking residue-pairs for experiments.

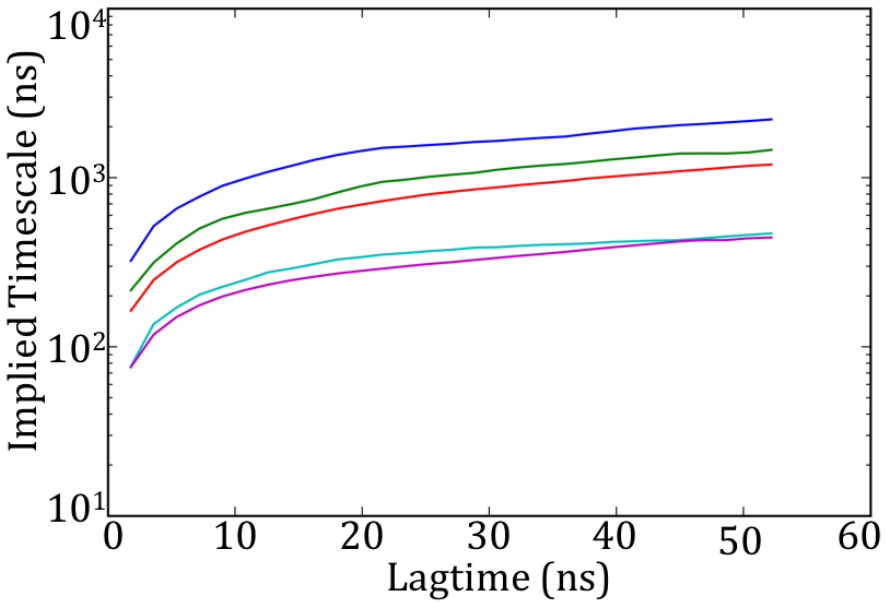
The novelty of our method is the use of a large amount of dynamic information as compared to the previous prediction methods based on static structure and sequence information [136]. In essence, MD simulation datasets provide us with the information of pairwise residue motions that describe the structural rearrangements observed during the conformational dynamics of the protein for the informed design of experiments. Our method relies on simulation datasets that explore such structural transitions in proteins. A future direction for our method is to include structural, thermodynamic or kinetic information available from other experimental techniques. In such cases, experiments which provide orthogonal information would enhance our understanding of protein dynamics. In this manner, the algorithm can be used as a tool to design experiments that would capture conformations which have not been observed so far - thus, determine the best set of experiments for a comprehensive study of protein conformational dynamics. Finally, the algorithm could also be combined with methods that provide microscopic kinetics from stationary state distributions to utilize simulation data from accelerated sampling methods which lack accurate kinetic information.

2.6 Supplementary Information

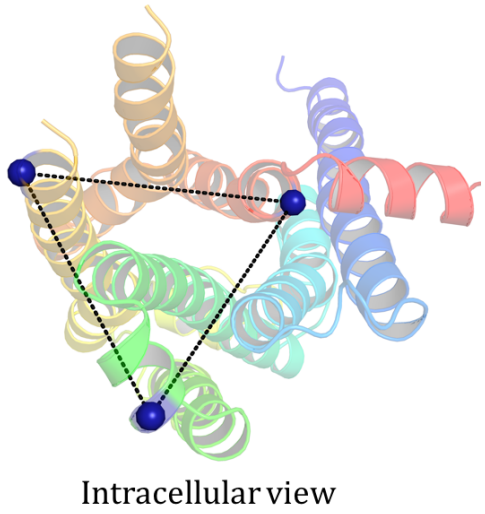
MSM construction. Hyper-parameters chosen for MSM construction for the three biological proteins analysis in the current work are mentioned in Supplementary Table 2.1. Parameters that are not indicated have the default values as implemented in MSMBuilder3.4 [69].

Supplementary Table 2.1: MSM hyper-parameters

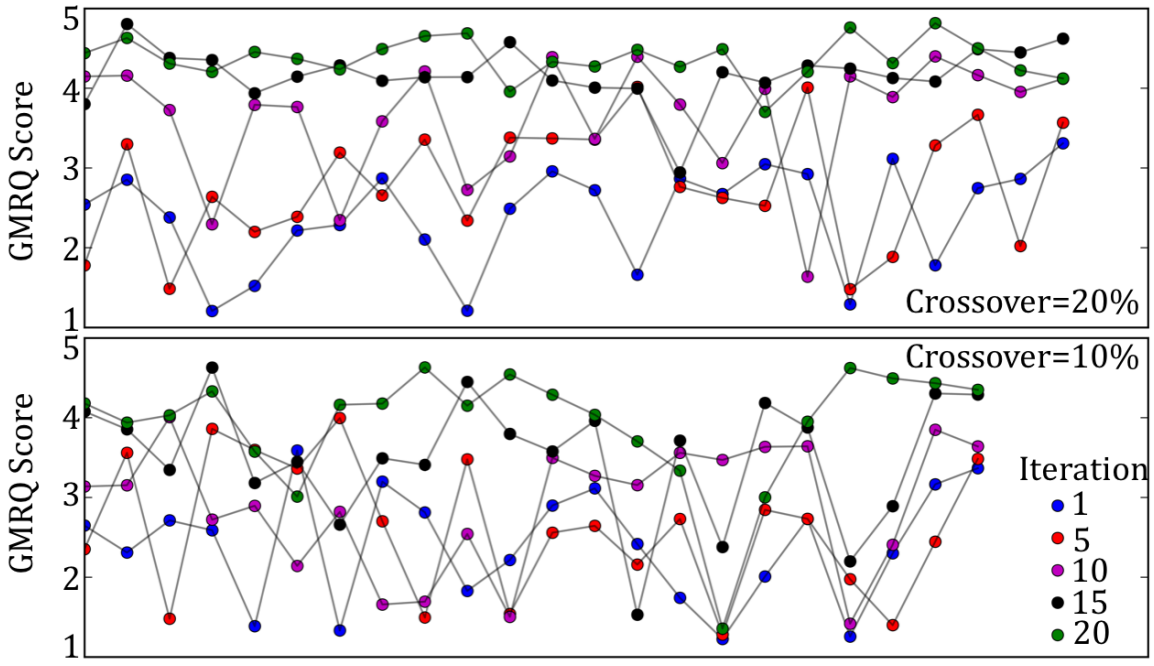
	β_2AR	Apo-C-CaM	PepT_{So}
Clustering	mini-batch k-means [71]	mini-batch k-means	mini-batch k-means
Clusters	200	200	200
MSM timescales	5	5	5
MSM lag time	50 ns (Supplementary Figure 2.1)	50 ns [63]	24 ns (see Chapter 4)



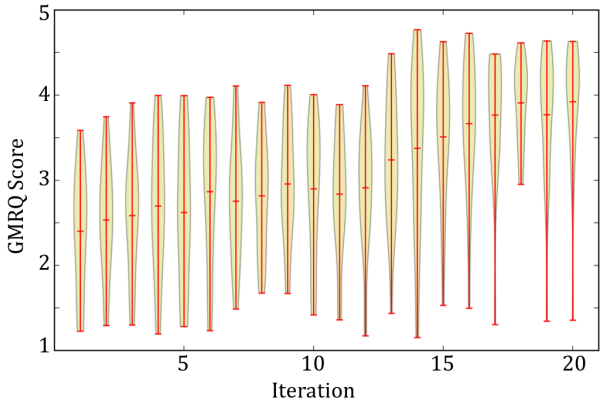
Supplementary Figure 2.1: Implied timescales plot from transition probability matrix of the MSM for β_2AR. Eigenvalues of the transition probability matrix correspond to the dominant rates of transition in the MSM. The top 5 eigenvalues for the MSM shown here converged at a lag time of ∼50 ns. The MSM was featurized based on all inter-residue contacts on the intracellular and extracellular sides of the protein and decomposed into 200 states.



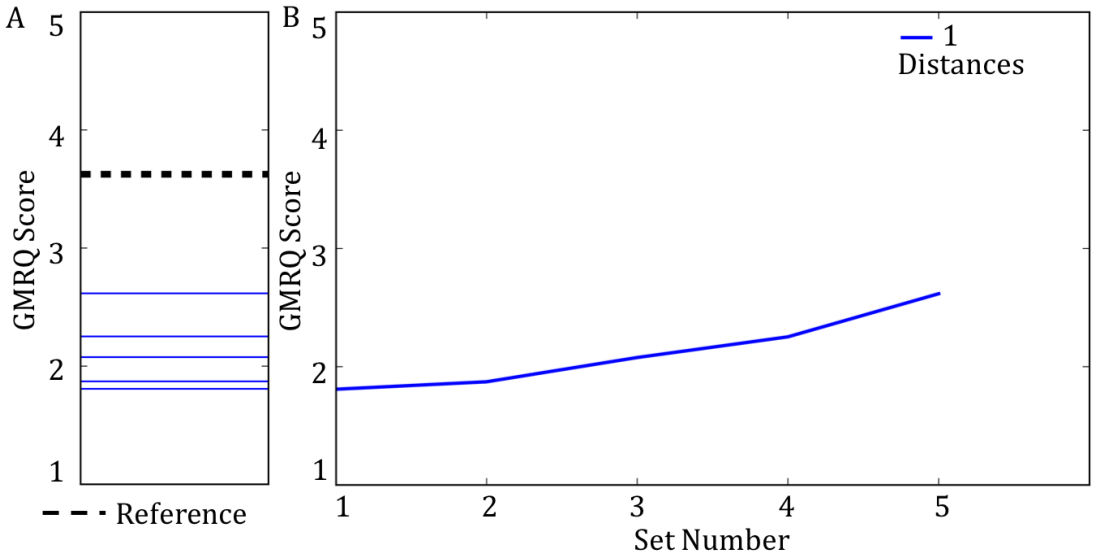
Supplementary Figure 2.2: Cartoon representation of β_2AR where the residues that were ranked highest on the intracellular side by our algorithm are indicated. The residues identified were on helix 3,4 and 6. The protein structure was obtained from the MD simulation data set [130].



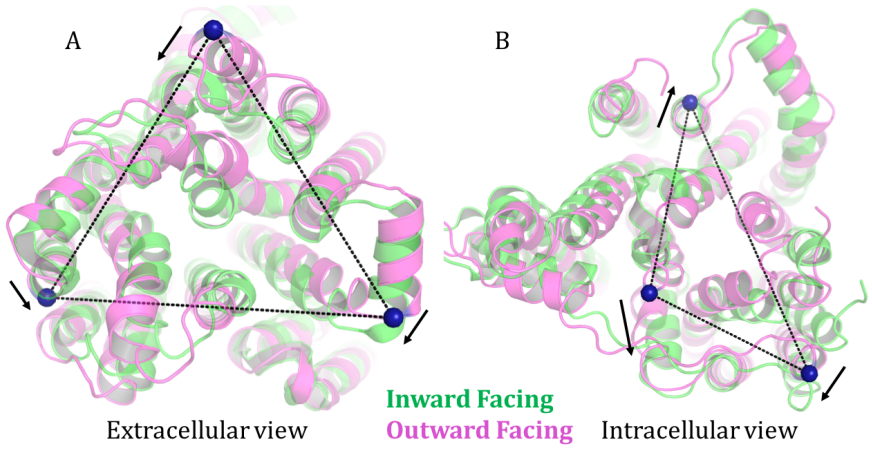
Supplementary Figure 2.3: A comparative analysis of two runs of the genetic-algorithm-based method on the protein β_2AR with different crossover rates. The scores converge and reach higher GMRQ scores by iteration 20 in the top plot, however the bottom case would require more number of iterations for the scores to converge.



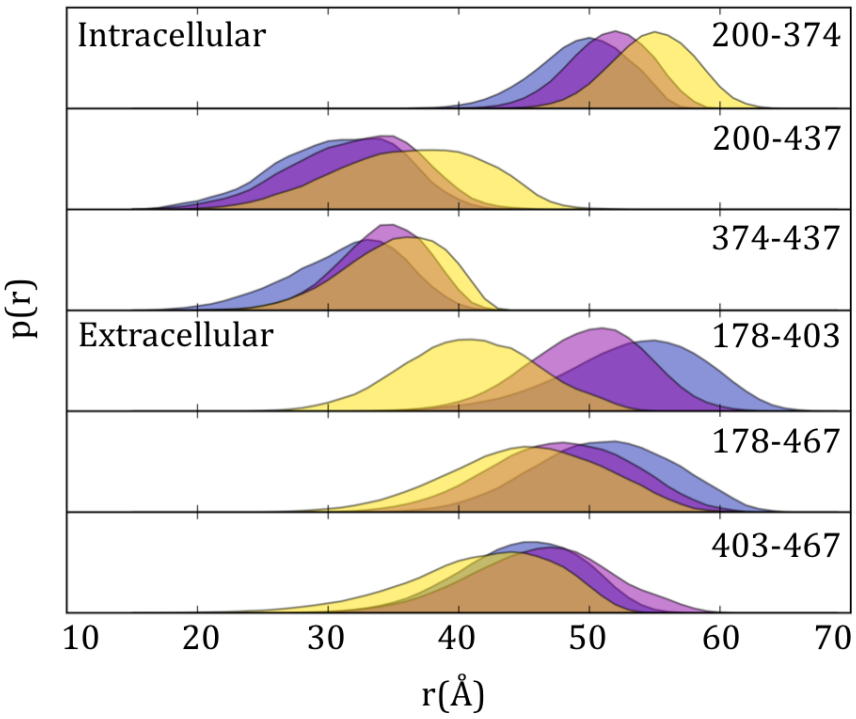
Supplementary Figure 2.4: Violin plot demonstrating the increase in the GMRQ score over 20 iterations of the genetic-algorithm-based method, with crossover rate of 10% on the protein β_2AR. The vertical red lines indicate the ranges of the GMRQ scores within the various iterations, and the horizontal mark in the middle of each line is the mean of the GMRQ scores for that iteration.



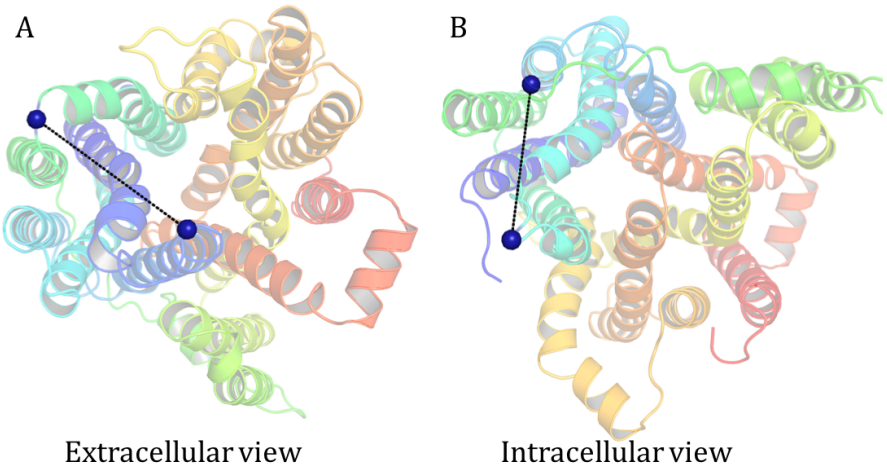
Supplementary Figure 2.5: (A) GMRQ scores corresponding to MSMs constructed with the predicted sets of residue-pairs for DEER experiments in the apo C-terminal domain of calmodulin using a DEER measurement range of 18-100 Å. The horizontal black line corresponds to a reference MSM constructed using all inter-residue contacts of the protein. (B) GMRQ scores of the predicted sets of residue-pairs differentiating the number of residue-pairs (or distances) measured in the set.



Supplementary Figure 2.6: Comparative analysis of the inward facing (green) and outward facing (pink) conformation of $PepT_{So}$. The top ranked residue-pairs chosen are shown on the (A) extracellular and (B) intracellular side of the protein.



Supplementary Figure 2.7: DEER distance distributions for the highest ranking choices in $PepT_{So}$. Inward-facing, occluded and outward-facing plots are represented in yellow, violet and blue, respectively. The distance distributions were obtained using RotamerConvolveMD [108].



Supplementary Figure 2.8: Low ranked choices are shown on the (A) extracellular and (B) intracellular side of $PepT_{So}$.

Chapter 3

Maximizing Kinetic Information Gain of Markov State Models for Optimal Design of Spectroscopy Experiments[1]

3.1 Overview

Spectroscopic techniques such as Trp-Tyr quenching, luminescence resonance energy transfer, and triplet-triplet energy transfer are widely used for understanding the dynamic behavior of proteins. These experiments measure relaxation of a particular labeled set of residue-pairs and the choice of residue-pairs requires careful thought. As a result, experimentalists must pick residue-pairs from a large pool of possibilities. In the current work, we show that molecular simulation datasets of protein dynamics can be used to systematically select an optimal set of residue positions to place probes for conducting spectroscopic experiments. The method described in this work, called Optimal Probes, can be used to rank trial sets of residue-pairs in terms of their ability to capture the conformational dynamics of the protein. Optimal Probes ensures two conditions, residue-pairs capture the slow dynamics of the protein and their dynamics is not correlated for maximum information gain, to score each trial set. Eventually, the highest scored set can be used for biophysical experiments to study kinetics of the protein. The scoring methodology is based on kinetic network models of protein dynamics and a variational principle for molecular kinetics to optimize the hyper-parameters used for the model. We also discuss that the scoring strategy used by Optimal Probes is the best possible way to ensure the ideal choice of residue-pairs for experiments. We predict the best experimental probe positions for proteins λ-repressor, β_2 Adrenergic Receptor, and villin headpiece domain. These proteins have been well-studied and allow for a rigorous comparison of Optimal Probes predictions with already available experiments. Additionally, we also illustrate that our method can be used to predict the best choice for experiments, by including any previous experiment choices available from other studies on the same protein. We consistently find that the best choice cannot be based on intuition or structural information such as distance difference between few known stable structures of the protein. Therefore, we show that incorporating protein dynamics could be used to maximize the information gain from experiments.

[1]This chapter is reproduced with permission from Mittal S, Shukla D. Journal of Physical Chemistry B. 2018; 122(48):10793-10805. Copyright 2018 American Chemical Society.

3.2 Introduction

Proteins' conformational diversity is important for the wide variety of functions they perform [1, 55]. This diversity is due to their dynamic behavior, a change in their structure as a function of time. It is yet not possible to directly look at the dynamics of a protein via structural experiments. Instead spectroscopy-based methods such as electron paramagnetic resonance (DEER/EPR) [106], fluorescence-quenching, Förster resonance energy transfer (FRET) and other energy transfer techniques [137, 138] have emerged as critical tools to capture conformational plasticity. Experiments can lead to characterization of key protein conformations and capturing protein motions over a diverse range of timescales [1]. However, spectroscopy-based methods are only able to characterize few inter-atomic or inter-residue distances in a protein. Most techniques involve labeling the target protein with a donor and an acceptor molecule at two chosen positions. Donor and acceptors are chemical entities that can act as probes for the local/global conformational change in the protein. The probes can be two different molecules fused to two residues in the protein. For instance, donor xanthone (Xan) and the acceptor naphthylalanine (Nal) for Triplet-Triplet Energy Transfer (TTET) experiments [139]; fluorophores Alexa Fluor 350, Alexa Fluor 488, Alexa Fluor 594 and others in the series are commonly used as donor and acceptor probes for FRET [140]. Sometimes, intrinsically fluorescent amino acids such as tryptophan is paired with a tyrosine or a cysteine, or other organic dye molecules, by generating a mutant protein, which will participate in proton or electron transfer due to close contact [140–142]. On the other hand, techniques such as DEER/EPR use chemically identical probes, most commonly a paramagnetic spin label MTSSL, on two residues of the protein [92]. This site specific labeling step is followed by monitoring the emitted fluorescence in FRET, luminescence in LRET, quenching or triplet absorbance in TTET and interspin dipolar interaction in DEER/EPR. Since a single distance probe on the protein will not be sufficient to characterize a protein's dynamics or kinetics via the experimental technique, multiple pairwise measurements need to be gathered. Multiple measurements can lead to a reliable observation of the overall dynamics of the protein [143, 144].

How do experimentalists choose the multiple residue-pairs to label in a protein? For a protein of R residues there are $R(R-1)/2$ residue-pairs. If one chooses to measure distances between all residues, it would require expressing and purifying huge quantities of the protein. This would also need large number of site specific mutations to introduce individual probe molecules. Clearly, it is impossible to probe all residue-pairs via an experimental technique due to resource and time limitations. If not all, but k residue-pairs are labeled there are ${}^{R(R-1)/2}C_k$ possible ways in which k pairs can be chosen for a protein with R residues. For a 100 residue protein ($R = 100$) with $k = 2$, this accounts to almost 12 million options. Without a systematic approach, it is unimaginable to sift through each possibility to determine the one that will be most useful

to characterize the protein's dynamics. As a result, experimentalists often choose residue-pairs based on human intuition from prior structural information such as few protein structures only. Such choices may work well for proteins which have been studied previously, but it is of no value in case the protein under investigation is novel.

In Chapter 2, we have proposed the use of MD simulation datasets to methodically pick residue-pairs for DEER/EPR spectroscopy measurements. DEER/EPR measures the dipolar interaction between two paramagnetic loci in the protein and provides a distance distribution. Hence, it is a rich source for structural information of proteins. In the current work, we extend the earlier method for MD simulation guided prediction of residue-pair choices for kinetics experiments. A rationally chosen set of residue-pairs for biophysical experiments should, firstly be able to characterize the dynamics that occur at a slow timescale. Slow kinetics between conformations of the proteins indicates regions on the protein's energy landscape that are stable and require a high energy barrier to transition. These are often the pathways that are critical for the biological function of proteins, such as for a protein to fold into its native state. Secondly, the choice of residue-pairs must be such it captures conformations that have not been observed by other residue-pair measurements. In such a manner, we aim to maximize information about protein dynamics with minimum distance measurements. We have already demonstrated such a prediction is possible for DEER spectroscopy experiments.

In this paper, we demonstrate the technique for three separate kinetics experiments, Trp-Tyr fluorescence quenching [138, 141, 145, 146], luminescence resonance energy transfer (LRET) [147–151] and triplet-triplet energy transfer (TTET) [53, 139, 152, 153]. For each experiment, we adapt the method depending on the experimental constraints and features, and demonstrate the method's applicability to pick ideal residue-pairs on a protein for which prior MD dataset is available. Starting with thousands of trial sets of residue-pairs, for each trial, we featurize the MD dataset based on the distances among those residues, decompose the data into clusters, and build a Markov state model (MSM) [55, 154] for the protein's dynamics. Using the variational principle of conformational dynamics [84, 85], a generalized matrix Rayleigh quotient (GMRQ) score is attributed to the MSM. A higher GMRQ score indicates an MSM that can estimate the slow modes in the protein's dynamic behavior, indicating a better MSM as opposed to another which may have a lower GMRQ score. Hence, the high GMRQ score can also indicate a better choice of residue-pair distances used to characterize the protein's underlying dynamics. After scoring every trial set, we propose the highest scoring choice of residue-pairs for experiments. MSMs provide a statistical network model to estimate the processes which exhibit the slow dynamics in the protein and the GMRQ score is a theoretical framework to score the extent of metastability of individual conformations discretized in the MSM. For each experimental

technique, we compare our predicted choice with residue-pairs picked by experimentalists from literature. We consistently find that our predicted choices are scored higher and capture the key processes in the dynamics of the protein.

We describe our results for the experimental technique Trp-Tyr fluorescence quenching and TTET using the folding trajectories of lambda (λ) repressor protein [155] and villin headpiece (double-norleucine HP35 mutant) [53, 156], respectively, and LRET via the activation conformational dynamics dataset for a GPCR β_2 Adrenergic Receptor (β_2AR) [130]. Although it is still challenging to sample the dynamics of proteins, long timescale MD simulations can now be performed routinely because of recent advance in computational resources [40, 42, 43]. For each case, we also identified experiments that have previously been done on the same protein to compare our results [146,150,153]. Since it is useful to make avail of all previous information accessible for a protein when designing future studies, we also demonstrate inclusion of prior experiment choices while predicting next best set of residue-pairs for kinetics experiments. An asset of our method is that we repurpose already generated MD simulation datasets to predict the optimal residue-pair choices for a range of experiments. Our proposed method is not restricted to a particular experimental technique. Despite the wealth of information available in MD simulations, their use for biophysical experiment design is not yet mainstream.

3.3 Methods

There are a large number of possibilities for choosing a set of k distances to measure using either of the three biophysical kinetics experimental techniques; Trp-Tyr fluorescence quenching, LRET or TTET. The goal of our method is to narrow down this search space and assign a score to each possibility. Once such a score is assigned, the highest scoring choice of residue-pairs is picked for experiments. In this section, we discuss the role of MD simulations, MSMs, and the variational principle in our proposed method. This is followed by experiment technique specific features in our algorithm and a step-by-step recipe for optimal residue-pairs prediction.

Using MD simulation datasets to obtain residue-pair distances. MD simulations of proteins can successfully capture the atomistic detail into a protein's conformational change [82,157] or folding mechanism [45,47,83,155,158]. Computational simulations can validate or strengthen the experimental observations and vice versa. MD simulations with complementary biophysical experiments, such as CryoEM [159], NMR [160, 161], small-angle X-ray scattering (SAXS) [162], DEER [110], single-molecule FRET [163], and hydrogen-deuterium exchange coupled with mass spectrometry [164] are typical. These studies focus on different

biomolecules and aim to address varied scientific questions. Physical observables for biophysical experiments can be approximated either by direct or indirect order parameters calculated from simulation datasets, albeit with many limitations [165]. For example, deuterium exchange can be estimated via a function of the amide contacts and hydrogen bonds in the protein [164]. Distance between amino acid side chains can be used to approximate the effect of a Tyr quenching a Trp residue [146]. Similarly, the observable for fluorescence quenching as well as LRET and TTET can be estimated via residue-residue distances (proxy for inter-probe distances) in the protein. These residue-pair distances are computed for every frame in the simulation data. For example, given a set of residues $\{r_1, r_2, r_3, r_4\}$, the simulation dataset can be "featurized" using the set all residue-pair distances$\{d_{r_1,r_2}, d_{r_1,r_3}, d_{r_1,r_4}, d_{r_2,r_3}, d_{r_2,r_4}, d_{r_3,r_4}\}$. Hence, we achieve a "dimensionality reduced" dataset where each observed conformation of the protein is represented by a 6 element vector of inter-residue distance values, which can be used to approximate the experimental observables from the simulation dataset.

Building Markov state models based on residue-pair distances. Once this featurization is achieved, the conformational landscape can be decomposed into states or clusters using standard clustering algorithms. This conformational decomposition is based on a structural characteristic which in our example case were the 6 distances among the chosen residues of the protein. Next, transitions between the clustered states can be recorded every τ ns where τ is called the lag time for an MSM. Given a state decomposition, we construct an MSM by ensuring that the transition probability matrix at the given lag time ($T(\tau)$) follows the Markovian property and reversibility among states. MSMs obtained this way are kinetic network models over the conformational landscape of the protein and provide an estimate of the relaxation timescales of the protein's dynamics via a master equation formalism [154]. The transition probability matrix $T(\tau)$ can also be decomposed into its eigenvectors and eigenvalues, λ_i. The largest eigenvalue is 1 and the eigenvector gives the equilibrium population of states in the MSM. The rest of the largest m eigenvalues correspond to the m slowest relaxation timescales, t_i, for the protein's dynamics as $t_i = -\frac{\tau}{\ln \lambda_i}$. If an MSM estimates slower relaxation times, clearly it is a better MSM. The reader must note that, any MSM we construct is dependent on the chosen distances, such as $\{d_{r_1,r_2}, d_{r_1,r_3}, d_{r_1,r_4}, d_{r_2,r_3}, d_{r_2,r_4}, d_{r_3,r_4}\}$, for featurization.

Using the variational principle to obtain GMRQ score. There are various hyper-parameters that have to be selected when building an MSM, (i) featurization metric, (ii) number of features, (iii) use of dimensionality reduction method, (iv) number of tICA dimensions or tICs, (v) tICA lagtime, (vi) number of clusters, and (vii) clustering algorithm. For a given set of residues, their associated residue-residue distances are fixed as the MSM featurization metric in our case. Thus, the number of distances is the size of the feature vector or the number of features. Since the number of experimental measurements that can be

performed cannot be large, our number of features are usually small. Hence, we do not use tICA or any other dimensionality reduction protocol. For every protein, we use a lag time pre-determined using the convergence of implied timescales criteria. We also fix the number of clusters to 200 and use mini-batch k-means clustering algorithm. Thus, the only hyper-parameter that varies is the choice of residue-pairs.

Using a variational principle approach, the hyper-parameters used for MSM building can be optimized in order to reach a "good" MSM which provides a reasonable estimate of the slow kinetic processes with large relaxation times [84,85]. Osprey [125] implements the variational principle, provides a GMRQ score for given hyper-parameters and has recently been used for MSM construction from MD simulation datasets [87, 166]. Since the GMRQ score indicates how well an MSM estimates the slow relaxation times for the protein's dynamics, it can be used as a measure of how good the choice of the given residue-pairs is for featurization. The GMRQ score calculation is based on k-fold cross-validation [86] of the featurized MD simulation data. We perform a five fold cross-validation, $k = 5$, where the simulation data is equally split into training and testing dataset. The eigenvectors of the MSM are estimated based on the training data, these eigenvectors are then used to estimate the eigenvalues on the testing data. The sum of the top m eigenvalues is essentially the score, i.e. GMRQ $= \sum_{i=1}^{m} \lambda_i$. The mean of 5 independent runs on the test data is the final GMRQ score. In this manner we can determine a GMRQ score for every set of residue-pairs. However, first we need to enumerate all possible sets of residues and the associated residue-pairs.

Reducing the number of feasible sets of residue-pairs

Residue-pair selection is dependent on experimental technique and protein dynamics. Each specific experiment has its instrument and technique related constraints. We take these into account when selecting a residue in the optimal set of residue-pairs for kinetics experiments. In addition, protein topology and protein dynamics can provide significant insight into optimal choices for experiments. Once the first residue choice is made, the choice for the second residue to probe can be informed based on the technique constraints and the protein dynamics. However, if protein dynamics information from MD simulations is not available, the probability of a good choice for the second residue will be low simply based on putative dynamic regions judged based on a single structure. Sometimes, experimentalists may choose label positions based on large change in a residue-pair distance between an active an inactive structure, or some other order parameter between two stable structures. However, this is insufficient to account for the protein's dynamics, and the procedure described below is quintessential in systematically listing feasible sets of residue-pairs.

Trp-Tyr fluorescence quenching. To pick the first residue for probing protein dynamics appears to be an easy choice. However, the choice of the first residue also determines the subsequent choices. Hence, it is

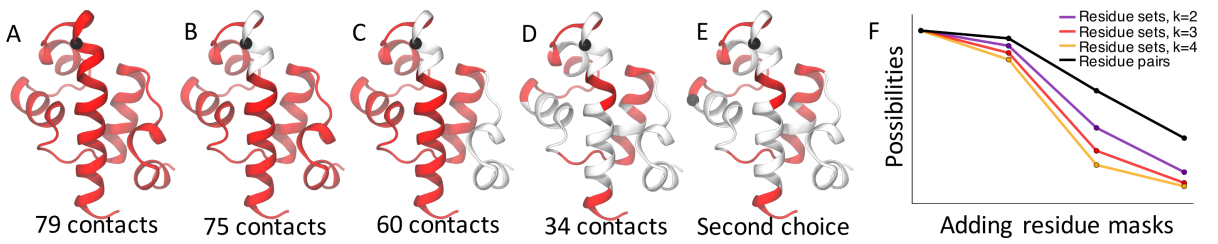


Figure 3.1: Possible residue-pairs and sets of residue-pair decrease exponentially as experimental technique and MD simulation dependent masks are incorporated in Optimal Probes. (A) In protein λ-repressor, once residue Lys26 (C_α atom shown in black) is chosen for site-directed fluorescence labeling, there are 79 possible ways to choose another residue to probe the corresponding distance pair. (B) Adding a constraint, to not probe residues within 2 residues of each other, the number of choices reduces to 75. Residues shown in white are now inaccessible. (C) Excluding the disordered helix 5 of λ-repressor decreases choices to 60. (D) Based on MD simulation data, Optimal Probes determines the residues which come within 7 Å of Lys26 to ensure fluorescence quenching. The residues highlighted in red are only for demonstration, not based on actual distance constraints. (E) Once the possibilities are narrowed down to 34 contacts, the second choice is easier, say Ser45 (C_α atom shown in black). (F) As the masks are added, the possible residue-pairs drop significantly (black line). Multiple residue-pair measurements, denoted by $k = 2$ (purple), $k = 3$ (red), and $k = 4$ (yellow), also decrease exponentially.

important to look at all possibilities rather than randomly choose few residue-pairs for experiments. Figure 3.1 demonstrates that upon selection of one residue, the options for the second, and further next choices decrease, using the protein λ-repressor. Similar masking concepts can be used for any cytoplasmic protein of interest to computational or experimental biophysicists.

In Figure 3.1A, we choose residue Lys26 on λ-repressor protein to demonstrate this concept. The C_α atom of Lys26 is shown using VDW representation in black. At this stage it appears that there are 79 possible choices for a residue-pair contact. To probe the slow dynamic modes in the protein, it is ideal not to probe two residues extremely close to each other, Thus, all residues within c positions, say $c = 2$ of Lys26 are removed (Figure 3.1B). The parameter c can be modified by the user in the user data provided to Optimal Probes. If the user can give secondary structure information, our method can exclude residues which are on the same secondary structure element. We also provide the user an option to exclude certain residues or a range of residues, which will not be chosen while predicting optimal residue-pair sets (Figure 3.1C). As an example, Prigozhin et al. avoided probe mutations on helix 5 (residues 71 to 86) of λ-repressor since it was previously shown to be unstructured in solution [146]. Other reasons for elimination could be low solvent exposure for residues within the protein core, conserved residue positions whose mutation will disrupt protein function or charged residue positions which form salt-bridge interactions with other residues important for the conformational switching mechanism. Further, we know that the distance between the amino acid side chains must be smaller than 7 Å for quenching of Trp fluorescence by Tyr [146]. Based on 647.1 μs MD simulation dataset [155] already available, the Optimal Probes algorithm masks residues which never come within 7 Å of Lys26 and chooses only those which can definitely lead to quenching. This is shown via the residues highlighted in red in Figure 3.1D. Reader must note that the residues shown for this masking step are only for visual example and not based on the actual distance value. Once the possibilities are narrowed

down to 34 contacts, the second choice is made, such as Ser45 in Figure 3.1E. This second choice is followed by similar masks as shown in Figure 3.1A-D, and so on for the subsequent residue choices. Thus, any residue choice made also affects the next residue choice that gets included for the trial set distances.

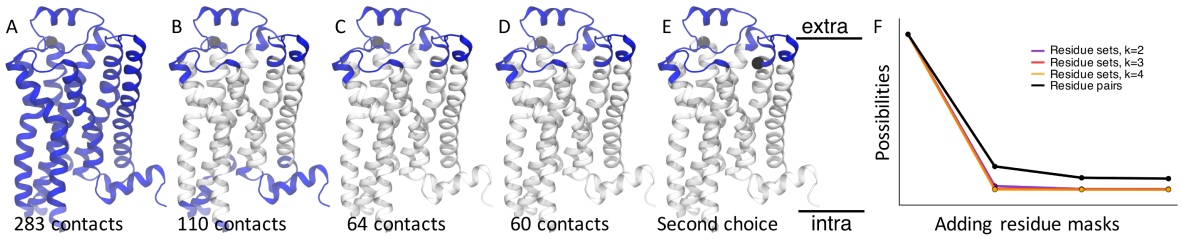


Figure 3.2: Possible residue-pairs and sets of residue-pair decrease exponentially as experimental technique and MD simulation dependent masks are incorporated in Optimal Probes for a membrane protein. (A) In protein β_2AR, once residue Met171 (C_α atom shown in black) is chosen for labeling, there are 283 possible ways to choose another residue to probe the corresponding distance pair. (B) TM region residues are susceptible to steric clashes and hence cannot be chosen for labeling, reducing the choices to 110 residues. Residues shown in white are now inaccessible. (C) Since the chosen residue Met171 is on the extracellular end of β_2AR there are 64 inter-residue contacts possible on this side. (D) Adding a constraint, to not probe residues within 2 residues of each other, the number of choices reduces to 60. (E) Once the possibilities are narrowed down, the second choice is made, say Ile94 (C_α atom shown in black). (F) As the masks are added, the possible residue-pairs drop significantly (black line). Multiple residue-pair measurements, denoted by $k = 2$ (purple), $k = 3$ (red), and $k = 4$ (yellow), also decrease exponentially.

We use this to our advantage in the proposed method. By reducing the number of pairs, we exponentially decrease the number of residue sets that need to be scored and can still provide the best possible choice for experimentalists. λ-repressor, an 80 residue protein, has 3160 residue-pairs; if we can reduce this to 1000 (black line in Figure 3.1F), $k = 2$ experimental measurements would lead to ${}^{1000}C_2$=499,500 residue sets, a decrease of 90% (purple line in Figure 3.1F). With $k = 3$ and $k = 4$, this drop is 97% and 99%, respectively. Further by introducing secondary structure information, we can decrease the set of possible residue-pairs significantly.

LRET. In addition to the residue masks discussed for Trp-Tyr fluorescence quenching, we add additional masks for LRET experiment choices on β_2AR originating from it being a membrane protein. In Figure 3.2A and Figure 3.2B, for demonstration, we pick residue Met171 among the 284 protein residues, followed by masking all residues in the transmembrane (TM) region. For β_2AR, 174 residues are embedded in the lipid bilayer, leaving 110 residues on either extracellular or intracellular end of the protein as potential candidates for attaching probes. Due to steric hindrance of bulky lipid molecules in membrane protein experiments, it is difficult to attach long FRET or LRET fluorophore probes in the TM region. However, experimentalists can perform Trp-Tyr fluorescence spectroscopy experiments using the inherent fluorescence of Tryptophans already present in the TM region [167]. Here we focus on describing the masks associated with the LRET experimental technique only. Next, since Met171 is on the extracellular side of β_2AR, we restrict the second probe choice only to the extracellular residues as well. This leaves us with 64 possible contacts with Met171 (Figure 3.2C). The other 46 residues are on the intracellular end of the protein. As discussed above, we

mask residues within c positions of Met171, here $c = 2$ residues are hidden (Figure 3.2D).

In LRET, the residue-pair to which acceptor and donor fluorophores are attached must show a distance change within the range of 10-100 Å [168]. We used 137 μs MD simulation dataset [130] to determine the pairs which lie within the experimental range. This is followed by making the second residue choice (Ile94 in Figure 3.2E). Also, it is useful to probe the motion of helixes with respect to each other, and hence, we can ensure the algorithm picks at most one residue on each helix while enumerating a list of all possible residue-pair sets. The number of pairs decrease from 40186 to 2835 (92% decrease) without incorporating any dynamics information to account for the distance change of 10-100 Å (black line in Figure 3.2F). Thus, once this information is also included, it will lead to a further drop in the possible residue-pair possibilities. The purple, red and yellow lines in Figure 3.2F indicate a huge decrease when listing combinations with $k = 2$, $k = 3$, and $k = 4$ among these residue-pairs.

TTET. In this experimental technique, the acceptor and donor are two different chemical moieties and they must come within the interaction range of their van der Waals radius for energy transfer to occur between the donor and acceptor. The typical value of the interacting range is 6 Å [53], which is dependent on the experimental instrument available. We demonstrate the predictions for TTET on a 35 residue protein fragment of villin. Since this is a cytoplasmic protein, we use the same masks as discussed for fluorescence quenching to reduce the subset of residues-pairs.

We have included the information necessary to incorporate the above described residue masks as user provided parameters. This can be fed to our program via an intuitive text file. A sample of this file is provided in the Supporting Information.

Prediction of a best residue-pair set to capture slow protein motions via experiments

In practice, we employ experimental technique constraints, protein topology information, and MD simulation dataset to construct a list of residue-pair sets. Each of these residue-pair sets can be viably measured by the experimental technique of choice. We refer to each set of residue-pairs listed so far as a "trial set". In the same example as before $\{r_1, r_2, r_3, r_4\}$ is a set of residues and the corresponding trial set $\{d_{r_1,r_2}, d_{r_1,r_3}, d_{r_1,r_4}, d_{r_2,r_3}, d_{r_2,r_4}, d_{r_3,r_4}\}$ where d_{r_1,r_2} is the inter-residue distance between residues r_1 and r_2. Another potential choice could be $\{r_1, r_3, r_4\}$, leading to residue-pairs $\{d_{r_1,r_3}, d_{r_1,r_4}, d_{r_3,r_4}\}$ as another trial set and so on. Accordingly, the residue-pairs among the residues will be different for every trial set.

In order to check which among the trials is the optimal residue-pair set, we use the exhaustive search strategy for all the discussed experimental techniques. In practice optimization techniques such as a genetic algorithm

approach can also be used to accelerate the search for the best choice. Finding an optimal set of residue-pairs for an experiment measurement requires the following steps for every trial set of distances.

1. Featurize the MD simulation dataset using distances among the residues pairs in the trial set. In our previous work, we use C_α-C_α distances since there is no consensus on which distance can ideally represent the interspin distance of the MTSSL spin labels. However, for kinetics experiments, the closest heavy atom distance is ideal as it is expected that the fluorescence, energy transfer or luminescence are primarily due to the side chain and not the backbone interaction. However, this can be trivially altered to calculate C_α-C_α distances or any closest atom distance. For instance, hydrogens atom distances may be the best choice for contact calculation in proton transfer experiments. This is a subjective choice, since probes can be variable in length, flexible and lead to some bias in the obtained experimental measurement [169].

2. Cluster the featurized data into 200 clusters using mini-batch k-means. Since our MD simulation data are in the range of hundreds of microseconds, 200 clusters will avoid fine partitions of the conformational space and not introduce any statistical error. Further, a smaller number of states can ensure that they correspond to suitably populated regions on the conformational landscape or free energy minima [170].

3. Determine a GMRQ score for the MSM corresponding to the state decomposition and the transitions among the states. We use the top 5 timescales to estimate the GMRQ score. As a result, the theoretical maximum is 6 [86, 87, 166], and the GMRQ scores can range between 0 and this upper limit.

Once all trial sets are scored, the one with the highest GMRQ score is chosen as the optimal set of residue-pairs for experiments. Details for MD simulation data, hyper-parameters for MSM construction and GMRQ score calculation are listed in Supplementary Table 3.1, Supplementary Table 3.2, and Supplementary Table 3.3. All codes have been implemented in Python, using Numpy, MDTraj 1.7.2 [129], MSMBuilder3.4 [171] and Osprey 1.0.0.dev0 [125] packages for specific functionalities. We use the Tcl scripting interface in VMD 1.9.3 [172] to visualize the predicted set of residue-pair choices.

Incorporating previous experiment choices when predicting next set of residue-pairs for experiments

In order to include any previous inter-residue relaxation kinetics measurements done beforehand, we include the corresponding residue-pairs in our trial set. If the distances available from a previous set of experiments involved residues $\{r_a, r_b\}$ and $\{r_c, r_d\}$, the trial set of distances would be, $\{d_{r_1,r_2}, d_{r_1,r_3}, d_{r_1,r_4}, d_{r_2,r_3}, d_{r_2,r_4}, d_{r_3,r_4}, d_{r_a,r_b}, d_{r_c,r_d}\}$. Similarly, we add the distances d_{r_a,r_b} and d_{r_c,r_d} to every trial set. Now the same procedure as described above is followed to score the trial sets. Finally the one with the highest GMRQ

score is chosen. The residue-pairs for which experimental information is available can be indicated in the user's input data. In this manner, we can still ensure that the proposed set of residue-pairs capture the slow processes in the underlying dynamics of the protein as well as ensure maximum overall information gain.

Checking for residue solvent exposure before predicting optimal residue-pairs for experiments

In the procedure described so far we used MD simulations to check whether the distance exhibited by the residue-pair is within the range necessary for an experimental technique and to build MSMs. Moving on, we can also use the dataset to check the extent of residue side chain accessibility for attaching probes. Accessibility is a required condition for site-directed fluorescence labeling, as well as to ensure that the probe will not introduce steric clashes and not alter the structural fold of the protein of interest. Previously, experimentalists may sometimes check the accessibility in two known states of a protein, such as an inactive and active conformation, or in protein homologs, and only within the native state dynamics of the folded state of the protein [173]. However, in a solution ensemble where proteins can undergo large scale dynamics, assuming native state dynamics or two state behavior can be grossly incorrect. Thus, it may be useful to consider the solvent exposure of a residue throughout the conformational landscape sampled via MD simulations. Readers will note that this might not be a concern in the case of Trp-Tyr fluorescence if no mutants are needed and these residues are already present in the sequence of the protein.

We used a solvent accessible surface area (SASA) implementation from MDTraj [129] to determine the SASA of each residue in the protein for every frame in the MD dataset. The algorithm provides the users with a list of residues that manifest large SASA values over the entire MD sampled conformational landscape. Tien et al. derived the amino acid solvent accessibility for all 20 residues and we used their theoretically derived values as the benchmark amino-acid solvent accessibility (Supplementary Table 3.4) [174]. Further, we used a cutoff of 80%, for example, the solvent accessibility upper bound for cysteine residue is 167 $Å^2$ and we check that the cysteine residue has at least 134 $Å^2$ SASA. If this condition is achieved, we can use this specific cysteine to attach kinetic probe molecules. The amino acid SASA benchmark values and the cutoff percentage can be modified by users via a text file provided as input to the code.

Computing decay rates from MD simulation for estimating kinetics experiment observables

There is often a disagreement while making quantitative comparison between experimentally derived observables and simulation [175]. This is expected since there are key differences between simulations and

experiments such as lipid or membrane composition, no fluorophores or probes or sequence mutations when the proteins are simulated using MD, force-field accuracy, and buffer solution conditions. There are various methods proposed which bias the simulation data to match with experiments, via ensemble reweighting [176–179] or by using biased MD [99, 180, 181]. However, our goal in this paper is to estimate the experimentally obtained observable from MD simulation data as accurately as possible. If we can estimate the experimental observable, we can compare whether the Optimal Probes predicted residue-pairs for the chosen experimental technique can indeed capture the slow dynamics of the protein. We want to make sure that that the experimental observable we predict must be a proxy for the kinetics as observed from fluorescence quenching, LRET or TTET experiments.

Fluorescence observable. In order to predict the relaxation time constant of distance pairs, we use a procedure similar to that used by Prigozhin et al. for comparing Trp-Tyr quenching kinetics traces with MD simulations [146]. For a chosen residue-pair (r_i,r_j), we compute the distance between the closest heavy atom of the residues as d_{r_i,r_j}. This value is scaled to obtain the Dexter energy transfer efficiency [141, 146] as,

$$\delta_{r_i,r_j} = e^{\frac{-d_{r_i,r_j}}{0.5}} \tag{3.1}$$

when the computed distances are in nanometers. We then calculate autocorrelation of δ_{r_i,r_j} as a function of time using Numpy function *correlate* and fit the resulting autocorrelation curves values with exponential functions using scipy's optimization function *curve_fit*. The 4 different functions used in this paper are summarized in Table 3.1.

Table 3.1: Fitting Functions for Decay Rates from MD Dataset

Fit	Fitted parameters (Number: Variables)	Equation
Single exponential	2 : A, τ	$y(t) = Ae^{(-t/\tau)}$
Double exponential	4 : A_1, τ_1, A_2, τ_2	$y(t) = A_1e^{(-t/\tau_1)} + A_2e^{(-t/\tau_2)}$
Stretched single exponential	3 : A, τ, β	$y(t) = Ae^{(-t/\tau)^\beta}$
Stretched double exponential	6 : A_1, τ_1, β_1, A_2, τ_2, β_2	$y(t) = A_1e^{(-t/\tau_1)^{\beta_1}} + A_2e^{(-t/\tau_2)^{\beta_2}}$

TTET observable. For predicting TTET physical observables from MD simulations, we use an MSM based prediction model [53]. TTET is also based on Dexter transfer mechanism and thus, in principle the method discussed for fluorescence could also be used to estimate the TTET relaxation timescales for a protein. For the MSM based TTET prediction model, first an MSM is constructed based on a geometric criterion. We build our MSM based on the ϕ, ψ, and χ_1 dihedral angles of each residue in the protein. The featurized dataset is decomposed into 200 clusters and the transition probability matrix is determined using 1 ns lag

time. This is followed by splitting each state into two, a "light" and a "dark" state. There is population in the light state before quenching occurs and in the dark state post quenching. Next, transfer coefficients and transition rates are estimated based on the contact distances. We used C_β distance among residues for contact calculation, except glycine for which the C_α atom was used. A cut-off of 4.4 Å for each residue was used to estimate the TTET active boundary in order to determine whether quenching occurred [182].

3.4 Results

Optimal Probes predictions for fluorescence (Trp-Tyr) quenching spectroscopy

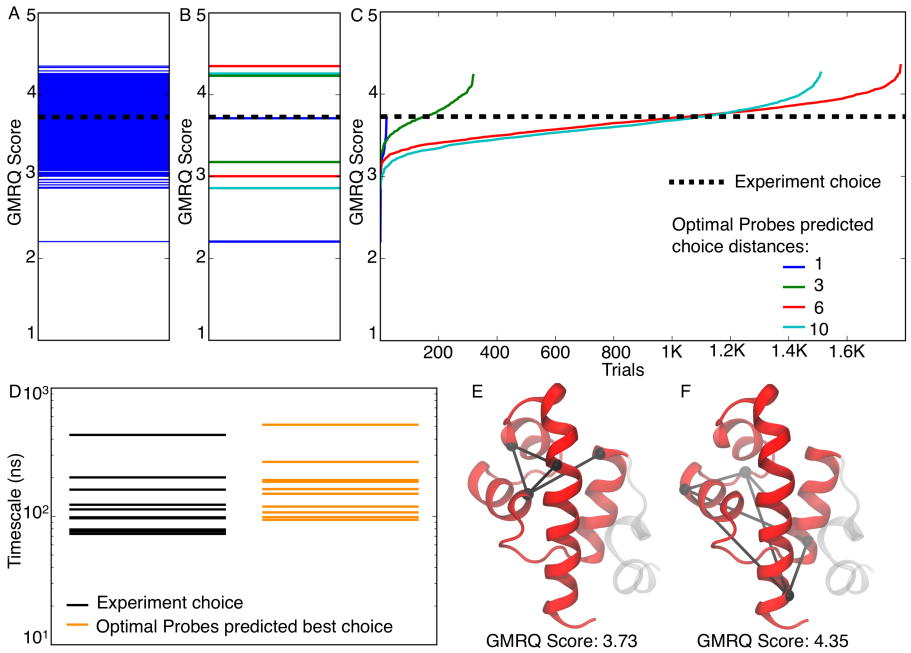


Figure 3.3: Optimal Probes predictions for Trp-Tyr fluorescence quenching on λ-repressor. (A) GMRQ scores for MSMs corresponding to 3,639 trial residue sets for λ-repressor. The experimental choice MSM score is shown in dashed black line. (B) Best and lowest GMRQ score with different number of residue-pairs (or distances) in the trial set, 10 (cyan), 6 (red), 3 (green) or 1 (blue). (C) GMRQ scores for the trial sets of residue-pairs differentiating the number of residue-pairs (or distances) measured in the set. (D) Comparing top 10 implied timescales for the experimental distances MSM (black) and Optimal Probes best choice distances MSM (orange) at lag time 60 ns. (E) Cartoon representation of λ-repressor showing the experimental choice residue-pairs. (F) Cartoon representation of λ-repressor showing the Optimal Probes predicted best set of residue-pairs.

We used Optimal Probes to predict residue-pairs for fluorescence quenching experiments. In this work, we focus on fluorescence quenching between two naturally occurring amino acids, Trp and Tyr. We use a primarily α-helical 80 residue fragment of λ-repressor protein to demonstrate the design of optimal fluorescence quenching experiments using MD simulation. We utilized 647.1 μs MD simulation dataset reported by LindorffLarsen et al. to observe folding and unfolding events of the protein [155].

Using the masks relevant to Trp-Tyr fluorescence quenching, we scored 3,639 trial sets. Scores for these

sets of residue-pairs range from 2.21 to 4.35 as shown in Figure 3.3A. The possible residue sets consist of 1 to 10 residue-pairs shown with different colors in Figure 3.3B,C. The highest scoring choice consists of 6 pairs with residues Glu10, Gly41, Ser45, and Tyr60. We also used the residue-pairs used by Prigozhin et al. for Trp-Tyr experiments on λ-repressor, Trp22-Tyr33, Trp22-Phe51, Tyr33-Phe51, and Phe51-Leu69 [146] to featurize the MD dataset, followed by assigning a GMRQ score for the corresponding MSM. We used the same hyper-parameters to score the experimental choice of residue-pairs. The GMRQ score associated with the experimental choice is 3.726 (dashed black line in Figure 3.3A,B and C). Many of the choices scored by our method ranked higher than the experimental choice. This indicates that the choice made by experimentalists may not always represent the optimal residue-pairs. Looking at the implied timescales for the MSMs, we see that not only the slowest timescale is larger for the highest ranking predicted choice, but the other timescales also correspond to kinetically slower processes in the protein (Figure 3.3D and Supplementary Figure 3.1). The experimental choice distances and Optimal Probes best choice distance are visualized in Figure 3.3E,F on the folded structure of the protein and in Supplementary Figure 3.2 on an unfolded snapshot.

Although the experimental MSM GMRQ score is lower than the Optimal Probes best choice, the latter also has 6 distances as opposed to 4 in the other. Hence, we also looked at the highest scoring choice with 3 distances (upper green line in Figure 3.3B), this choice has a GMRQ score of 4.23 and is still higher than the experimental choice score. Even the top choice with a single residue-pair, Ala15-Ala66, has a score 3.71 which is very close to the four residue-pair experimental choice score of 3.726. But we also see that not all choices with 10, 6 or 3 distances lie in the higher range and the lowest scoring choice, has a single distance, scores at 2.21 (Supplementary Figure 3.3).

Optimal Probes predictions for fluorescence spectroscopy with prior experimental choices

We demonstrate the inclusion of previously probed residue-pairs when predicting next set of residue-pairs for experiments using λ-repressor as an example protein. The experiment of choice is Fluorescence (Trp-Tyr) quenching spectroscopy, and thus, the residue masks are same as used in the previous section. Since λ-repressor protein is a well-studied system, the experimentalists were able to make informed choices for probe positions. This could be a possible reason for the high score associated with their choice. For optimal residue-pairs predictions with prior experiment choices, we chose two among those four residue-pairs used by experimentalists. We used residue-pairs Trp22-Tyr33 and Trp22-Phe51 as previously available information (see inset in Figure 3.4A). The score for an MSM constructed using these choices is 4.428 (dashed black line

in Figure 3.4A). Optimal Probes predictions for GMRQ scores range from 3.56 to 4.97. The Optimal Probes predicted highest scoring choice has 6 distances, shown on the λ-repressor folded structure in Figure 3.4B and unfolded structure in Supplementary Figure 3.4. The top choice with 3 distances and 2 distances yield a GMRQ score of 4.89 and 4.48, respectively. In all, 897 sets of residue-pairs were scored to obtain the optimal choice which improved the score associated with a two residue-pair choice used as previous information. In this manner, Optimal Probes can design experiments on proteins studied using the same or other kinetics experimental techniques that can generate incremental knowledge into the protein's dynamics.

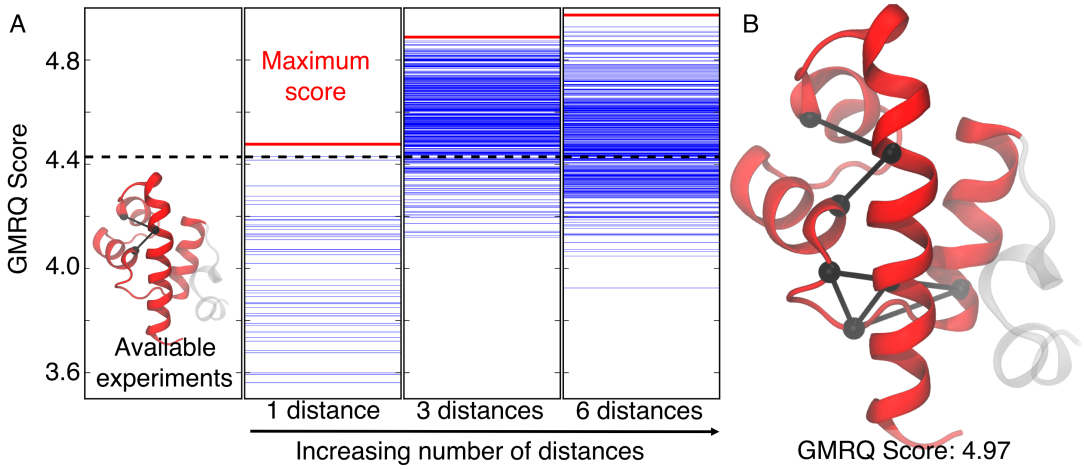


Figure 3.4: Optimal Probes predictions for fluorescence quenching experiments with previous information on λ-repressor. (A) residue-pairs Trp22-Tyr33 and Trp22-Phe51, used as previous experiment choices for predictions using Optimal Probes, are shown on the folded structure of the protein. The corresponding GMRQ score is 4.428 (dashed black line). Optimal Probes predicted sets with different number of residue-pairs (or distances) in the trial sets are shown separately for 1, 3, and 6 distances. (B) Cartoon representation of λ-repressor showing the Optimal Probes predicted best choice of distances. Residues are labeled in Supplementary Figure 3.4.

Optimal Probes predictions for LRET spectroscopy

In this section, we demonstrate that Optimal Probes successfully predicts residue-pairs for LRET experiments. In LRET, the probes are a lanthanide series cation (Terbium or Europium) and a fluorophore which are fused to two residues in the protein, followed by measuring the luminescence energy transfer from the donor to the acceptor. We show our results on the β_2AR protein, a G-protein coupled receptor (GPCR). The activation mechanism of this GPCR was extensively studied via long timescale MD simulations by Dror et al. of which we used 137 μs spread over 24 trajectories ranging from 2 to 15.42 μs [130]. To prepare viable sets of residue-pairs before GMRQ scoring, we use the residue masks as discussed before and shown in Figure 3.2A-D.

LRET experiments have only been recently used to study conformational changes in membrane proteins. Thus, this experimental technique has not yet been used on β_2AR. However, DEER and FRET experiments

have been performed on β_2AR previously [110, 163] and it is a well-characterized protein. Another similar GPCR protein, arginine-vasopressin type 2 receptor (V2R) was studied using LRET experiments [150]. Activation mechanism of V2R was observed to be consistent with observation for β_2AR and rhodopsin GPCRs. Rahmeh et al. used two residue-pairs in V2R to probe using LRET with donor fluorophore Fluorescein Arsenical Helix binder (FlAsH) and the acceptor Lumi4-terbium maleimide (Lumi4-Tb) [150]. Luminescence emission between the probes was monitored to estimate the decay rates timescales associated with inter-residue contact formation. Since we want to be able to use transferable insights between similar proteins, we use equivalent positions on the β_2AR protein Leu266-Arg344 and Ser329-Arg344 as a proxy for LRET experiments on V2R (see Supporting Information for details). In this work, all further references to experiment choices on β_2AR protein refer to these two residue-pairs.

Scores for 13,323 sets of residue-pairs range from 1.08 to 5.52 (Figure 3.5A). We can also look at the choices separately on either the intracellular or extracellular side of the protein. The possible residue sets consist of either 1, 3, 6, 10, or 15 residue-pairs (Figure 3.5B,C). The highest scoring choice consists of 10 residue-pairs. The experimental choice MSM scored at 3.726 and has residues pairs on the intracellular end (dashed black line in Figure 3.5A,B and C). The implied timescale plots (Figure 3.5D, Supplementary Figure 3.5) show that the Optimal Probes predicted best choice is able to capture slowest modes of the protein by a difference of greater than an order of magnitude. The experimental choice distances and Optimal Probes best choice distance are visualized in Figure 3.5E,F. The Optimal Probes predicted top choices with 10 and 6 distances (Figure 3.5G) have very similar scores. Thus the choice with 6 distances would be sufficient.

In order to obtain optimal residue-pairs for LRET experiments on both the intracellular and extracellular side of the β_2AR protein together, we generated a list of mixed trial sets of residue-pairs. We chose the highest scoring set each with 1, 3, 6, 10, and 15 residue-pairs. If the highest choice was on the intracellular side, we combined it with all possible choices on the extracellular side and vice-versa. We then ranked the mixed trial sets using the same Optimal Probes protocol. The scores ranging from 2.69 to 5.66 are shown in Supplementary Figure 3.6. The scored choices consist of a minimum of 2 residue-pairs to a maximum of 25 residue-pairs. The highest scoring choice consists of 13 residue-pairs, 10 on the extracellular side and 3 on the intracellular side (Supplementary Figure 3.7).

Optimal probes predictions for TTET

In this section, we demonstrate that Optimal Probes can successfully predict residue-pairs for TTET experiments. We use 354.9 μs simulations of villin headpiece double-norleucine HP35 mutant protein [53, 156]. The MD dataset consists of large number of small trajectories for this α-helical 35 residue protein fragment.

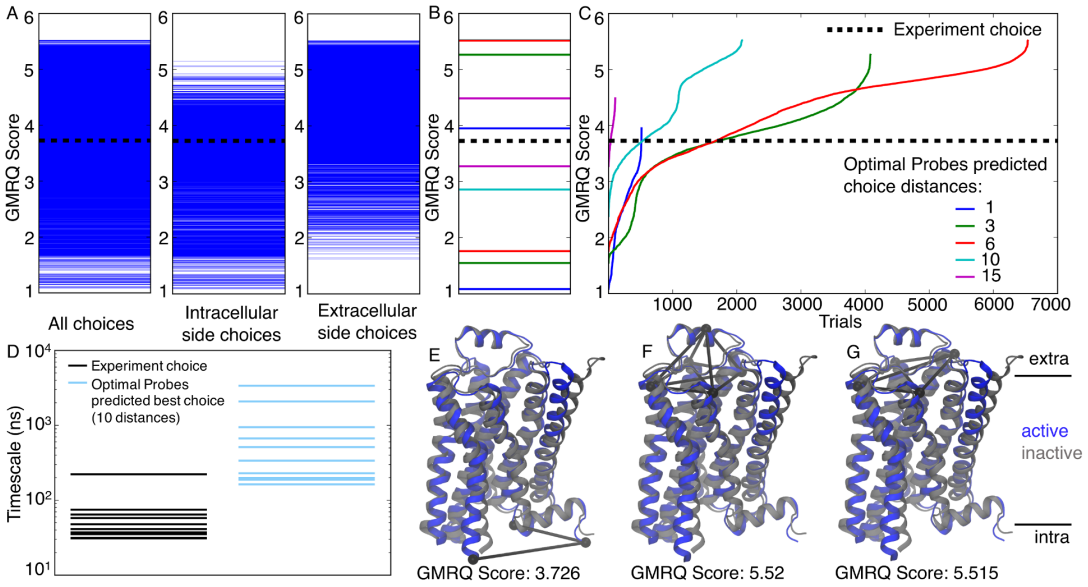


Figure 3.5: Optimal Probes predictions for LRET on β_2AR. (A) GMRQ scores for MSMs corresponding to 13,323 trial residue sets for β_2AR. The second and third columns indicate the GMRQ scores for trial sets that have only intracellular side distances and extracellular side distances, respectively. The experimental choice MSM score is shown in dashed black line. (B) Best and lowest GMRQ score with different number of residue-pairs (or distances) in the trial set, 15 (magenta), 10 (cyan), 6 (red), 3 (green) or 1 (blue). (C) GMRQ scores for the trial sets of residue-pairs differentiating the number of residue-pairs (or distances) measured in the set. (D) Comparing top 10 implied timescales for the experimental distances MSM (black) and Optimal Probes best choice distances MSM (skyblue) at lag time 50.4 ns. (E) Cartoon representation of β_2AR showing the experimental choice residue-pairs. (F) Cartoon representation of β_2AR showing the Optimal Probes predicted best set of residue-pairs with 10 distances. (G) Cartoon representation of β_2AR showing the Optimal Probes predicted best set of residue-pairs with 6 distances.

The masks used to reduce the initial set of choice are similar to fluorescence quenching. We use closest heavy atom scheme to calculate all distances to featurize the MD simulations.

Upon scoring, 10,964 trials sets in all, we obtained scores for these sets ranging from 0.997 to 5.91 as shown in Figure 3.6A. The possible residue sets consist of 1, 3, 6 or 10 residue-pairs as shown in different colors in Figure 3.6B,C and the highest scoring choice consists of 10 residue-pairs (cyan). TTET experiments are available for the residue-pairs Lys6-Trp22, Leu0-Trp22, Trp22-Phe34, and Leu0-Phe24 [153]. An MSM was constructed using these inter-residue distances and the GMRQ score for the same was 5.39 (dashed black line in Figure 3.6A,B and C). As in the previous examples for λ-repressor and β_2AR, we see that there are many choices scored higher than the experimental choice. Although, the experimental choice is not optimal, but it can be called a good choice. Any randomly chosen set of residue-pairs would not have ranked as high. We postulate that the experimentalists were able to make a good choice because villin has been studied extensively in literature as a model protein for folding studies.

Moreover, we observe that Optimal Probes scored all of the trial set choices with 10 or 6 residue-pairs higher than the experimental choice which had 4 residue-pairs (cyan and red line in Figure 3.6B). Infact, of all the trial sets, 38% of the choices included 5 residues and the 10 distances among them. Only some of the

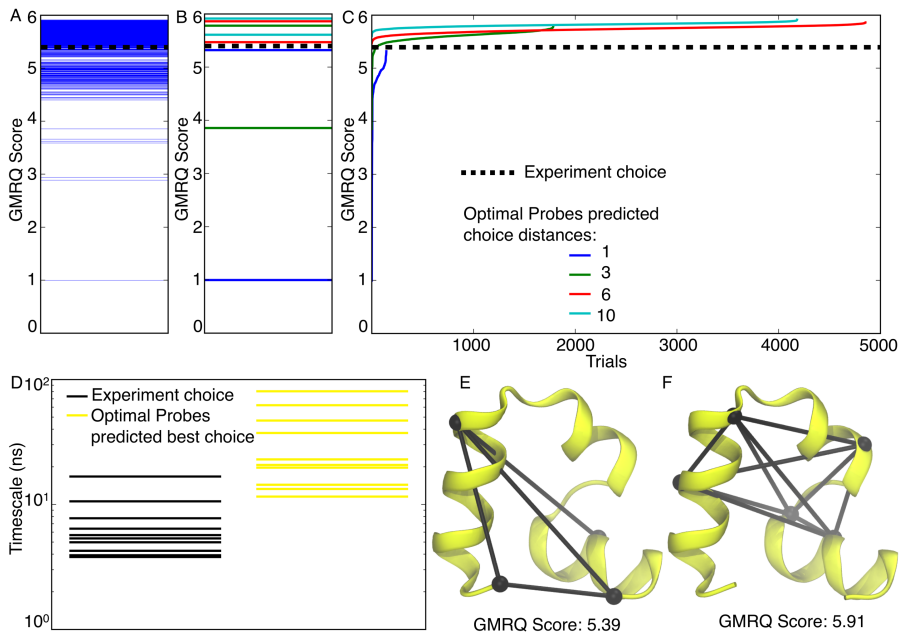


Figure 3.6: Optimal Probes predictions for TTET on villin. (A) GMRQ scores for MSMs corresponding to 10,964 trial residue sets for villin. The experimental choice MSM score is shown in dashed black line. (B) Best and lowest GMRQ score with different number of residue-pairs (or distances) in the trial set, 10 (cyan), 6 (red), 3 (green) or 1 (blue). (C) GMRQ scores for the trial sets of residue-pairs differentiating the number of residue-pairs (or distances) measured in the set. (D) Comparing top 10 implied timescales for the experimental distances MSM (black) and Optimal Probes best choice distances MSM (yellow) at lag time 1 ns. (E) Cartoon representation of villin showing the experimental choice residue-pairs. (F) Cartoon representation of villin showing the Optimal Probes predicted best set of residue-pairs with 10 distances.

choices with 3 residues pairs are lower ranked than the experimental choice. But, all choices with a single residue-pair are considerably lower ranked, the highest one scoring at 5.32. Just by a difference of one residue Leu21 instead of Pro20, the score changed drastically to 2.89. Beauchamp et al. also observe significant differences in the kinetics of residue-pairs just by changing a single residue in villin [53]. The implied timescales plot (Figure 3.6D and Supplementary Figure 3.8) show the top 10 timescales for the experimental distances MSM and Optimal Probes best choice distances MSM in black and yellow, respectively. Most of the estimated timescales for the protein's dynamics are larger for the Optimal Probes highest ranked choice. The experimental choice distances and Optimal Probes best choice distance are visualized in Figure 3.6E,F on the folded structure of the protein and in Supplementary Figure 3.9 on an unfolded snapshot. The highest ranking choice with 6 distances scored at 5.86 (Supplementary Figure 3.10), which is very close to the highest scoring choice overall at 5.91.

We also looked at the distance distribution of residue-pairs in the experimental set (Supplementary Figure 3.11) which usually showed a single peak. But, this is not seen in the distance distribution for the best ranking choices with 10 or 6 distances. This bodes well since, Optimal Probes is able to pick residue-pairs which have different values in metastable states because of structural differences and hence important in

the protein's unfolding/folding. Atleast in the experimental choice for TTET, the distances in the native state span over 1 nm (dotted lines in Supplementary Figure 3.11A). However, for λ-repressor, most distances picked by experimentalists have native state values which are <5 Å, whereas the distance choices picked by Optimal Probes have varying values (dotted lines in Supplementary Figure 3.12 and Supplementary Figure 3.13). TTET relaxation times are obtained using the methodology proposed by Beauchamp et al. as described in the Methods section and is shown in Supplementary Figure 3.14 [53]. The residue-pairs which are dark either cannot exhibit TTET (as derived from the constructed MSM) or have extremely slow relaxation times. We also highlight the experimental residue-pairs in the lower triangle and the Optimal Probes predicted best choice residue-pairs (in the upper triangle) through yellow stars.

3.5 Discussion

In common practice, it can seem obvious to assume that distances that show most change between the folded and the unfolded structure would be a good choice for experiments. A distance difference plot, between the folded and unfolded structures of λ-repressor (Supplementary Figure 3.15) can show the residue-pairs that undergo large distance changes. Similarly, we can obtain the distance difference plots for villin (Supplementary Figure 3.16) and between the active and inactive structures of β_2AR (Supplementary Figure 3.17). But, based on these plots it appears that many residue-pair possibilities can be chosen. Potentially, accessibility or solvent exposure of residues could be used to narrow down choices. We observe that accessibility of residues cannot help experimentalists choose the optimal residues for experiments, it can only reduce the possible choices. This functionality is included in Optimal Probes as discussed in the Methods section. Also, some choices that appear to be non-optimal from these criterion may be kinetically important and hence provide information about the slow dynamics of the protein. This is evident from the distance difference maps where the distances picked by experimentalists and by Optimal Probes are highlighted with yellow stars. The choices that are found kinetically relevant by experimentalists and by Optimal Probes do not necessarily show a large distance change in the structures compared.

In order to take the kinetics into account, can experimentalists use decay rates for each individual distance? We can estimate the kinetic decay rates for each of the distance pairs from the MD simulation data using the procedure detailed in Methods. The τ_{slow} values from residue-pair autocorrelation for λ-repressor and β_2AR residue-pairs are shown in the Supporting Information (Supplementary Figure 3.18, Supplementary Figure 3.19, and Supplementary Figure 3.20). The estimated TTET relaxation times for villin are shown in Supplementary Figure 3.14. In Supplementary Figure 3.18A, among the four distances, only three dis-

tances could be fit using the double exponential fit function in Table 3.1. This is consistent with previous observations [146]. Supplementary Table 3.5 and Supplementary Table 3.6 list the functions used to fit autocorrelation curves for every distance referenced in this work. As seen from the autocorrelation decay curve figures not all distances could be fit, but it appears (visually) that some of these distances correspond to the slow dynamics of the protein. If we can determine the decay rates from MD datasets, why are the estimated slow timescale values not a sufficient condition for optimal experiment choice? This is because orthogonality among residue-pairs is also necessary. If the dynamics of two residue-pairs are correlated, then the protein dynamics information captured by one among the chosen pairs is redundant and does not add any new information.

One way to look at the correlation among residue-pairs (or distances) in MD simulations is through the dynamic cross correlation (DCC). The procedure for DCC calculation is described in the Supporting Information. As seen in the DCC maps in Supplementary Figure 3.21, Supplementary Figure 3.22, and Supplementary Figure 3.23 for λ-repressor, β_2AR, and villin, respectively, many of the pairwise distances show no correlation (DCC values close to 0). In Supplementary Figure 3.21A, the 4 experimental choice distances of λ-repressor show high correlation values which means they do not capture most of the processes that are necessary in the folding of the protein. We also observe the DCC values for residue-pairs on the intracellular side and the ones on the extracellular side of protein β_2AR show less correlation as compared to residue-pairs on the same side. This is seen by the formation of a distinct pattern of two squares, a 10×10 and another of 3×3, in Supplementary Figure 3.20C for 10 extracellular distances and 3 intracellular distances. However, DCC values do not seem to provide a clear cut-off that can indicate a good set of residue-pairs. This leaves out correlation as a bad measure for choosing optimal residue-pair choices. Moreover, pairwise DCC calculations are non-trivial to perform for long timescale simulation datasets.

Previously, in Chapter 2 we had shown that MD simulations can aid in experiment design for DEER spectroscopy by maximizing the information gain from the experimental observations. Recently, Hays et al. iteratively performed DEER restrained simulations for DEER experiment design [183]. They maximize mutual information among residue-pairs to ensure maximal information gain. The iterative scheme only guarantees that the chosen residue-pairs will lie on the parts of the protein that do not undergo simultaneous conformational change. It is expected that the initial rounds of iterative design will select for fast moving but likely functionally irrelevant conformational change processes in the protein. Finally, DEER or other EPR spectroscopy methods are primarily focused towards structural information. In this work, we argue that Optimal Probes is the method of choice for simulation guided design of kinetics experiments. Searching for the best choices based on τ_{slow} and DCC is not straightforward, and will require some trade-offs on both

ends. In contrast, GMRQ provides a framework which assigns a high score to an MSM that can capture the kinetically slowest motions in the protein, as well as motions that are independent or not-correlated to each other. Evidently a scoring method based on GMRQ, as used in Optimal Probes is necessary for the design of spectroscopy experiments.

Through the application of Optimal Probes on three different experimental techniques and different proteins, we illustrate that the method can be used for most types of energy transfer spectroscopy methods. Depending on the natural amino acids used as acceptor or quencher, or use of external fluorophores, the user can modify the distance constraint for quenching and adapt the Optimal Probes for their choice of fluorescence spectroscopy. For example, recently Watson et al. have proposed the use of selenomethionine (M_{Se}) for Trp quenching to probe protein dynamics. This technique requires a van der Waals interaction between the Se-atom and the Trp residue for quenching to occur via electron transfer [184]. Instead of a mask of 7 Å for Trp-Tyr quenching, a value of 3-4 Å can be used to ensure that the residues will be in the van der Waals contact range. Thus, optimal residue-pairs can be chosen for experiments employing M_{Se} quenching of Trp Fluorescence. We also propose that Optimal Probes can help design experiments for live cells or in vivo studies. A direct application is to employ some of the recently collected crowding simulation datasets [185,186] to predict residue-pairs for fast relaxation imaging (FReI) [187] experiments to understand the protein dynamics in vivo. This can also be key to understand the difference in protein kinetics in vitro versus in vivo.

As readers would note, extensive MD simulations datasets are the basis for probes prediction using our method to determine the inter-residue distances in the dynamics datasets. Moreover, the experimentally obtained measurement is between the chemical probes and hence not exactly between the protein's backbone residues or amino acid side chains. Oftentimes, the sequence of the protein on which experiments are performed are not exactly the same as for MD, since site-specific mutations are required to perform quenching experiments. Thus, all conclusions drawn provide a qualitative and global picture for the folding or conformational change of proteins. For instance, the predictions for TTET relaxation times of distances can also be dependent on the hyper-parameters used for MSM construction among other factors [53]. These differences exist in most interpretations that combine experiments and simulation based studies. Moreover, prior to spectroscopic experiments, it may be necessary to test that the function of the protein is not lost entirely. This is not yet possible using computational predictions. If a potential method can predict functional loss from mutagenesis, it can be included within the framework of Optimal Probes such that those residues are not selected for optimal residue-pair predictions. Web servers such as DynaMut [188], Site Directed Mutator [189], and I-Mutant 2 [190] can predict the effects of point mutations on protein stability

which could be used to asses protein function. We report the difference in free energy values ($\Delta\Delta G$) upon site-specific mutations in Supplementary Table 3.7, Supplementary Table 3.8, and Supplementary Table 3.9 using the recent web server DynaMut.

Further, we also made predictions for LRET experiments on β_2AR using two residue-pair choices as previous information (data not shown). Since V2R is a considerably less studied protein, these positions can be mapped back onto V2R residue choices for future experiments. Hence, our method can potentially also be used for making residue-pair predictions for kinetic experiments on a protein whose X-ray/NMR/Cryo-EM structure is not available by using MD simulation dataset of a structurally similar protein.

Atomistic simulations have been a means to rationalize conclusions post experiments and to enhance low-resolution experimental information [191]. However, to the best of our knowledge Optimal Probes protocol is the first method to utilize detailed atomistic simulation data to design experiments in a standardized manner. In our work, we have provided the first steps towards a feedback like cascade for simulations and experiments, where information can flow both ways. We envision a situation where both simulations and experiments stimulate each other and enable scientists to understand protein dynamics.

3.6 Supplementary Information

Calculating dynamic cross correlation. Dynamic cross correlation (DCC) is defined as

$$\mathrm{DCC}(d_{r_1,r_2}, d_{r_3,r_4}) = \frac{\langle \Delta d_{r_1,r_2}(t) \cdot \Delta d_{r_3,r_4}(t) \rangle_t}{\sqrt{\langle ||\Delta d_{r_1,r_2}(t)||^2 \rangle_t} \sqrt{\langle ||\Delta d_{r_3,r_4}(t)||^2 \rangle_t}} \tag{3.2}$$

$$\Delta d_{r_1,r_2}(t) = d_{r_1,r_2}(t) - \langle d_{r_1,r_2}(t) \rangle_t$$

$$\Delta d_{r_3,r_4}(t) = d_{r_3,r_4}(t) - \langle d_{r_3,r_4}(t) \rangle_t$$

where $\langle A \rangle_t$ means the ensemble average of the quantity A. This formulation of DCC among distances is based on the DCC analysis for the motions between atoms from MD simulations [192]. DCC values for residues pairs with themselves are expected to show correlation value of 1.

Extracting unfolded structures from MD simulation dataset. The unfolded structure chosen for visualization is the MD simulation data frame with the highest radius of gyration. For λ-repressor, the data frame with R_g = 24.71 Å is chosen. In contrast, the R_g for the folded structure is ∼12 Å. The unfolded structure chosen for villin is the MD simulation data frame with ∼22 Å, whereas the R_g for the folded structure is ∼10.02 Å. All R_g values are determined using MDTraj [129].

Choosing equivalent positions for LRET experiment residues on β_2AR. Rahmeh et al. used residue-pairs Ala267-Cys358 and Ser330-Cys358 in V2R for LRET probes [150]. In both distance measurements, Cys358 was labeled with the donor Fluorescein Arsenical Helix binder (FlAsH). Ala267 is the last residue on TM helix 6 in V2R at the intracellular side, the corresponding residue in β_2AR is Leu266 on TM helix 6. Similarly, Ser330 is the last residue on TM helix 7 in V2R at the intracellular side, corresponding to residue Ser329 for β_2AR. A sequence alignment for the two proteins, human V2R and β_2AR, was generated using T-Coffee web server [193] and is shown in Supplementary Figure 3.24. Experimentalists chose Cys358 as a probe for the TM helix 8 in V2R. For this, we use the last residue on TM helix 8 of β_2AR which is Arg344. Thus, the equivalent residue-pairs on β_2AR are Leu266-Arg344 and Ser329-Arg344.

Using stretched exponential fitting functions. As listed in Supplementary Table 3.5 and Supplementary Table 3.6, some of the distances are fitted to a single or double stretched exponential function. The physical consequence of this choice could be that there are multiple pathways adopted by the protein for the given distance choice. As a result, the β, β_1 or β_2 factors in the stretched exponential can distribute the timescales of all the pathways in order to obtain the final distance decay rate. Stretched exponentials for proteins motions have been used previously [194] and in particular for luminescence emission decay rates.

Supplementary Table 3.1: Simulation datasets and MSM hyper-parameters (λ-repressor)

Experiment	**Trp-Tyr fluorescence spectroscopy**
Protein	λ-repressor, λ_{6-85}
Residues	80
Dataset	647.1 μs from LindorffLarsen et al. [155]
Trajectory frames timestep	200 ps
Analysis stride	10 frames
Clustering algorithm	mini-batch k-means
Number of clusters	200
MSM timescales	5
MSM lag time	60 ns, 30 ns #1
Number of MSMs scored	3,639 & 897 #1
Experiments previously reported	Yes [146]

#1: Only for residue-pair predictions with previous available information.

Supplementary Table 3.2: Simulation datasets and MSM hyper-parameters (β_2AR)

Experiment	**Luminescence resonance energy transfer**
Protein	β_2 Adrenergic Receptor, β_2AR
Residues	284#1
Dataset	137 μs from Dror et al. [130] #2
Trajectory frames timestep	180 ps
Analysis stride	10 frames
Clustering algorithm	mini-batch k-means
Number of clusters	200
MSM timescales	5
MSM lag time	50.4
Number of MSMs scored	13,323 (one-side) & 1963 (two-side)
Experiments previously reported	Yes [150] #3

#1: β_2AR actual protein is larger but 41 residues on the intracellular side, between TM helixes 6 and 7, are missing from the crystal structure (PDB: 3P0G [131]) and in the MD simulations [130].

#2: Only protein backbone coordinates of system A are extracted from the published dataset for the current analysis.

#3: Available experiments are for a different GPCR, V2R.

Supplementary Table 3.3: Simulation datasets and MSM hyper-parameters (villin)

Experiment	**Triplet-triplet energy transfer**
Protein	Villin headpiece double-norleucine HP35 mutant
Residues	35
Dataset	354.9 μs from Beauchamp et al. [53, 156]
Trajectory frames timestep	250 ps
Analysis stride	4 frames
Clustering algorithm	mini-batch k-means
Number of clusters	200
MSM timescales	5
MSM lag time	1 ns
Number of MSMs scored	10,964
Experiments previously reported	Yes [153]

Supplementary Table 3.4: Solvent accessibility of protein residues

Residue (3 Letter code)	Solvent accessibility ($Å^2$) #1
Ala	129
Arg	274
Asn	195
Asp	193
Cys	167
Glu	223
Gln	225
Gly	104
His	224
Ile	197
Leu	201
Lys	236
Met	224
Phe	240
Pro	159
Ser	155
Thr	172
Trp	285
Tyr	263
Val	174

#1: From Tien et al. [174]

Supplementary Table 3.5: Fitting function for residue-pair choices (λ-repressor)

Experiment	Choice	residue-pair	Fit
Fluorescence spectroscopy	Experiment	22, 33	Double exponential
		22, 51	Double exponential
		33, 51	Double exponential
		51, 69	None
	Optimal Probes Best	10, 41	Stretched double exponential
		10, 45	Double exponential
		10, 60	Double exponential
		41, 45	Double exponential
		41, 60	Stretched single exponential
		45, 60	Stretched single exponential

Supplementary Table 3.6: Fitting function for residue-pair choices (β_2AR)

Experiment	Choice	residue-pair	Fit
LRET	Experiment	266, 344	Stretched single exponential
		329, 344	Stretched single exponential
	Optimal Probes Best (10 distances)	105, 171	Stretched double exponential
		105, 183	Stretched double exponential
		105, 297	Stretched double exponential
		105, 306	Stretched single exponential
		171, 183	Stretched double exponential
		171, 297	None
		171, 306	Stretched double exponential
		183, 297	Stretched single exponential
		183, 306	Stretched single exponential
		297, 306	Stretched single exponential
	Optimal Probes Best (6 distances)	102, 171	None
		102, 297	Double exponential
		102, 306	Stretched single exponential
		171, 297	None
		171, 306	Stretched double exponential
		297, 306	Stretched single exponential
	Optimal Probes Best (Two-sided) (13 distances)	105, 172	None
		105, 183	Stretched double exponential
		105 297	Stretched double exponential
		105, 306	Stretched single exponential
		172, 183	Stretched double exponential
		172, 297	Stretched single exponential
		172, 306	Single exponential
		183, 297	Stretched single exponential
		183, 306	Stretched single exponential
		297, 306	Stretched single exponential
		147, 268	Stretched double exponential
		147, 333	Double exponential
		268, 333	None

Supplementary Table 3.7: Effect on protein stability (β_2AR)

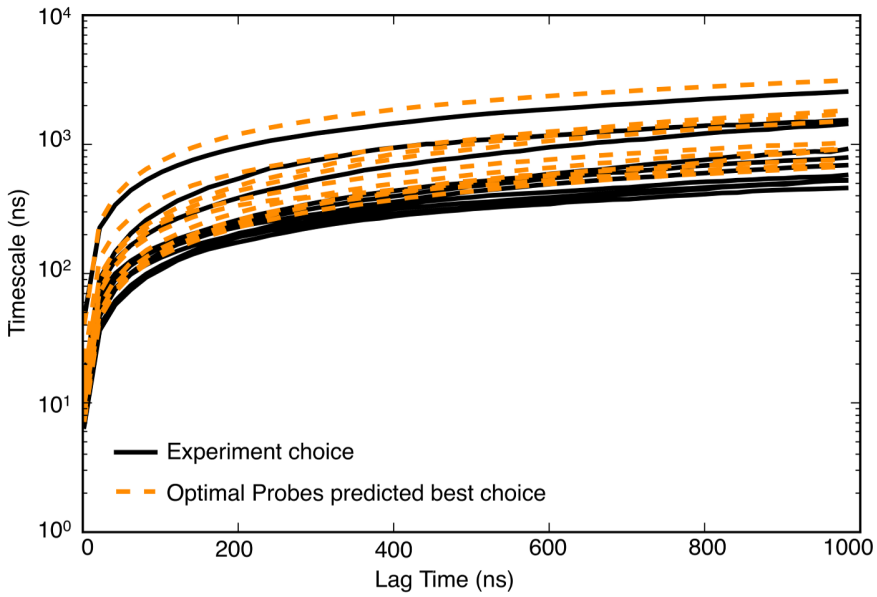
Residue positions choice	Mutation	$\Delta\Delta G$ (kcal/mol)
Optimal Probes Best (5 residues)	W105C	1.841
	M171C	1.366
	N183C	1.733
	V297C	1.451
	E306C	1.872
	M171W	1.493
	N183W	1.574
	V297W	1.464
	E306W	1.85
	W105Y	1.887
	M171Y	1.529
	N183Y	1.536
	V297Y	1.438
	E306Y	1.872
Experiments (Manglik et al. [110])	N148C	1.745
	L266C	1.225

Supplementary Table 3.8: Effect on protein stability (λ-repressor)

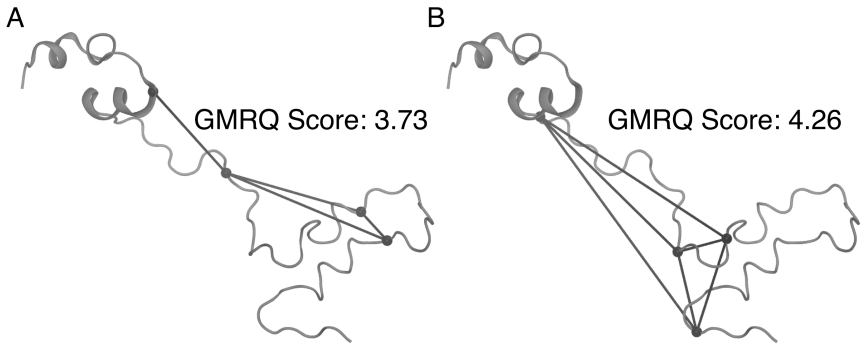
Residue positions choice	Mutation	$\Delta\Delta G$ (kcal/mol)
Optimal Probes Best (4 residues)	E10C	-0.144
	G41C	-0.677
	S45C	-0.043
	Y60C	-0.03
	E10W	0.913
	G41W	-0.131
	S45W	0.097
	Y60W	0.699
	E10Y	0.179
	G41Y	-0.865
	S45Y	0.051
Experiments (Prigozhin et al. [146])	F51Y	-0.391
	L69W	0.75

Supplementary Table 3.9: Effect on protein stability (villin)

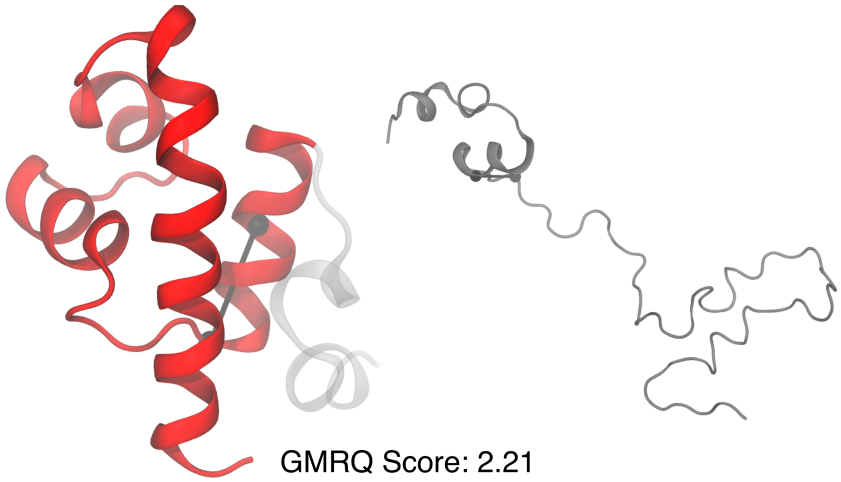
Residue positions choice	Mutation	$\Delta\Delta G$ (kcal/mol)
Optimal Probes Best (5 residues)	K6C	-0.716
	G10C	0.916
	S14C	0.101
	L21C	-0.011
	H26C	0.515
	K6W	0.575
	G10W	-0.314
	S14W	0.967
	L21W	-0.268
	H26W	0.827
	K6Y	0.089
	G10Y	1.347
	S14Y	0.055
	L21Y	-0.202
	H26Y	2.751



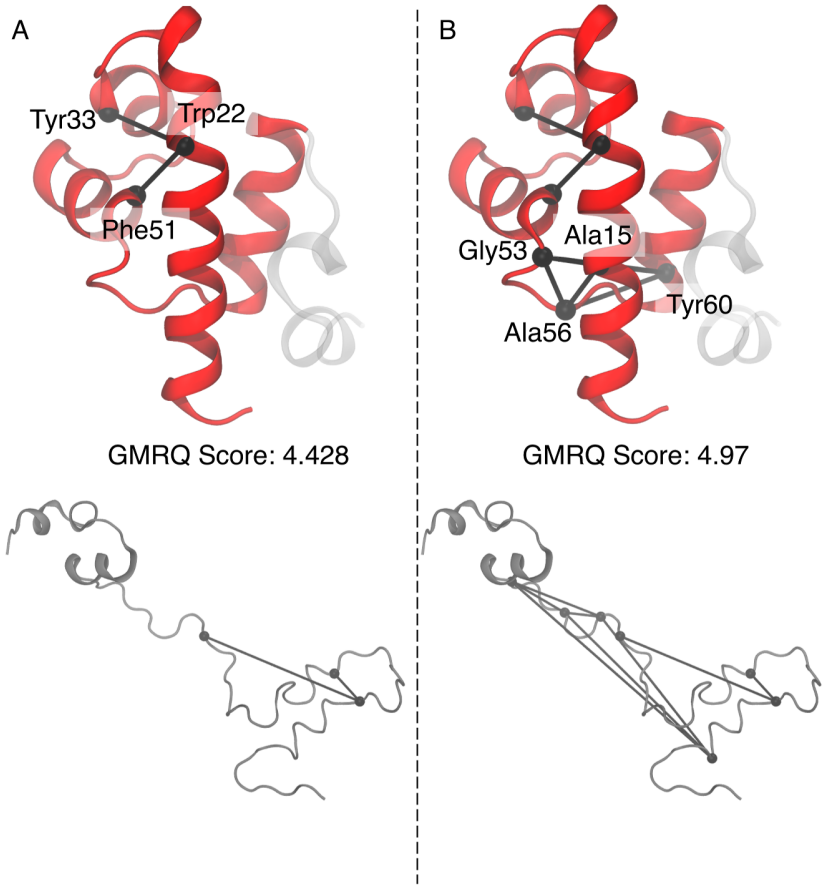
Supplementary Figure 3.1: Comparing top 10 implied timescales for the experimental distances MSM and Optimal Probes best choice distances MSM as a function of lag time on λ-repressor MD simulation dataset.



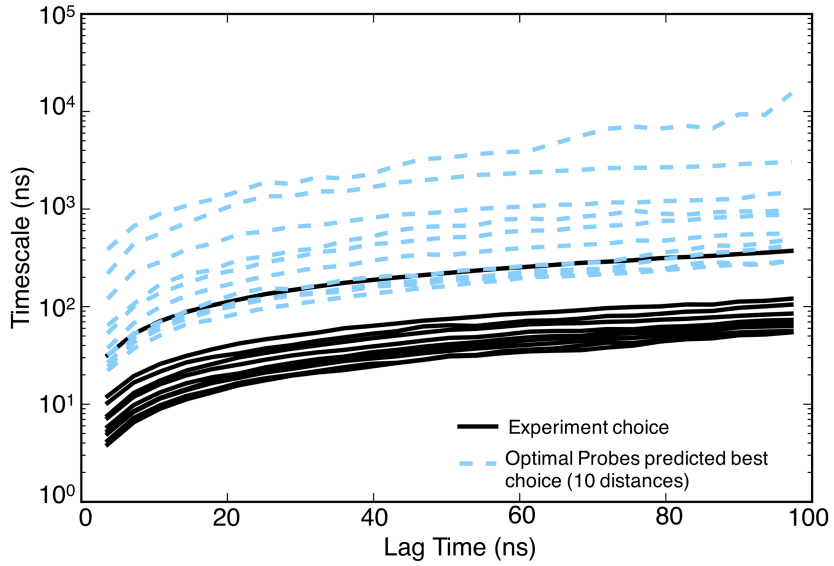
Supplementary Figure 3.2: (A) Cartoon representation of λ-repressor (unfolded structure) showing the experimental choice residue-pairs. (B) Cartoon representation of λ-repressor (unfolded structure) showing the Optimal Probes predicted best set of residue-pairs.



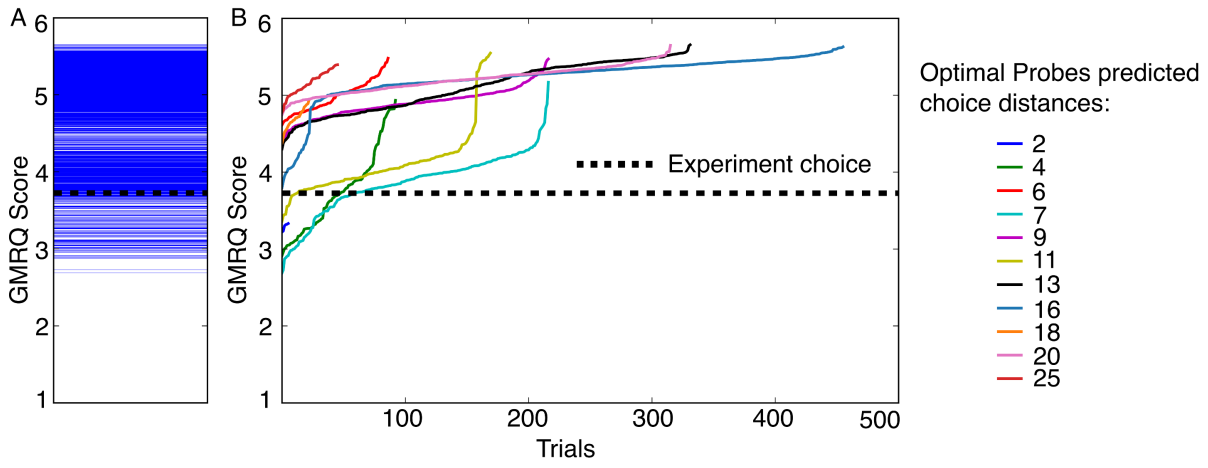
Supplementary Figure 3.3: Cartoon representation of λ-repressor showing the Optimal Probes predicted lowest scoring choice residue-pairs.



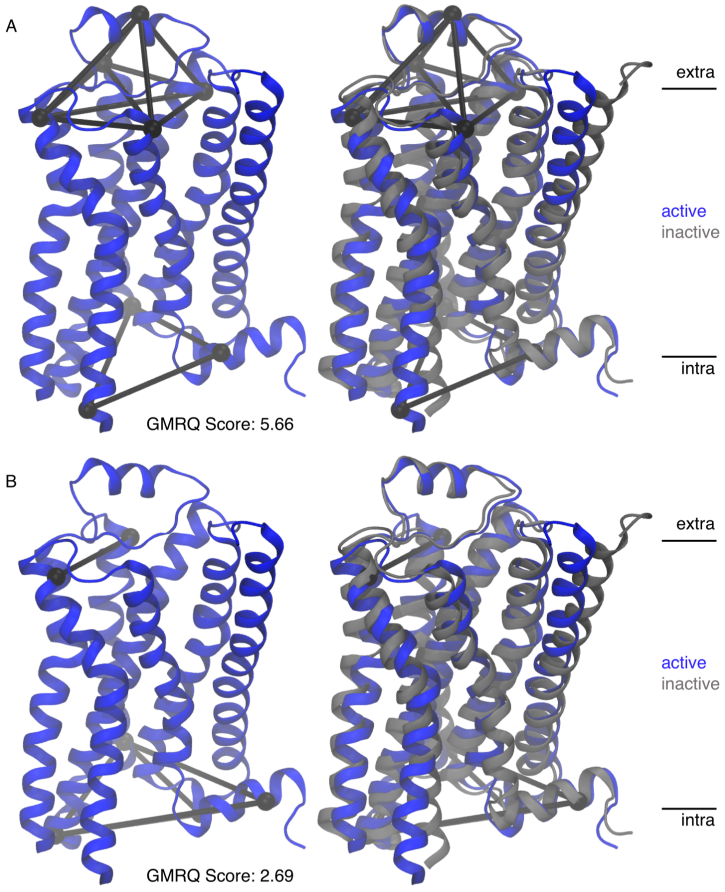
Supplementary Figure 3.4: Optimal Probes predictions for fluorescence quenching experiments with previous information on λ-repressor. (A) residue-pairs Trp22-Tyr33 and Trp22-Phe51 are used as previous experiment choices for predictions using Optimal Probes. (B) Predicted best choice of experiments includes 6 distances, among residues Ala15, Gly53, Ala56, and Tyr60.



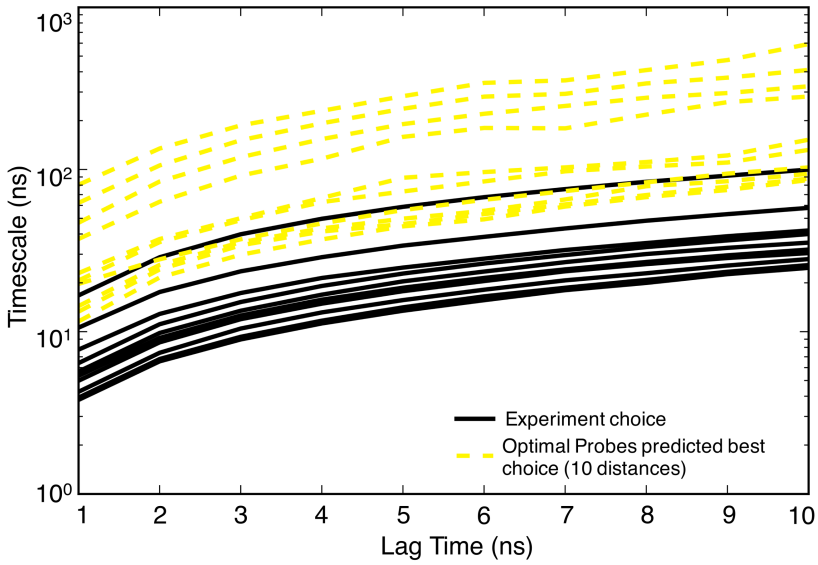
Supplementary Figure 3.5: Comparing top 10 implied timescales for the experimental distances MSM and Optimal Probes best choice distances MSM as a function of lag time on β_2AR MD simulation dataset.



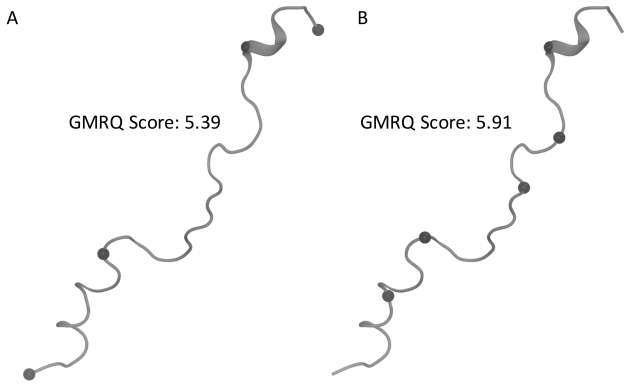
Supplementary Figure 3.6: Optimal Probes predictions for LRET on both, extracellular and intracellular sides of β_2AR. (A) GMRQ scores for MSMs corresponding to 1963 trial residue sets for β_2AR. The experimental choice MSM score is shown in dashed black line. (B) GMRQ scores for the trial sets of residue-pairs differentiating the number of residue-pairs (or distances) measured in the set.



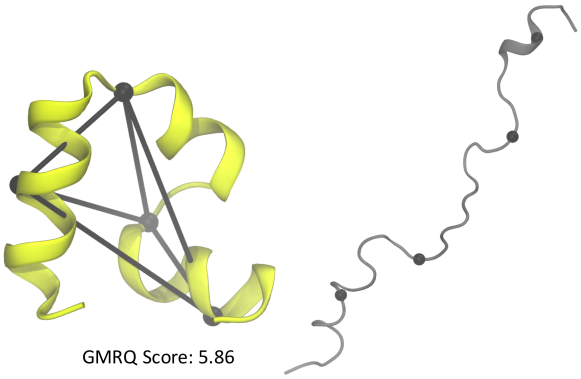
Supplementary Figure 3.7: Cartoon representation of β_2AR showing the Optimal Probes predicted best and lowest choice residue-pairs, for probes on both extracellular and intracellular side.



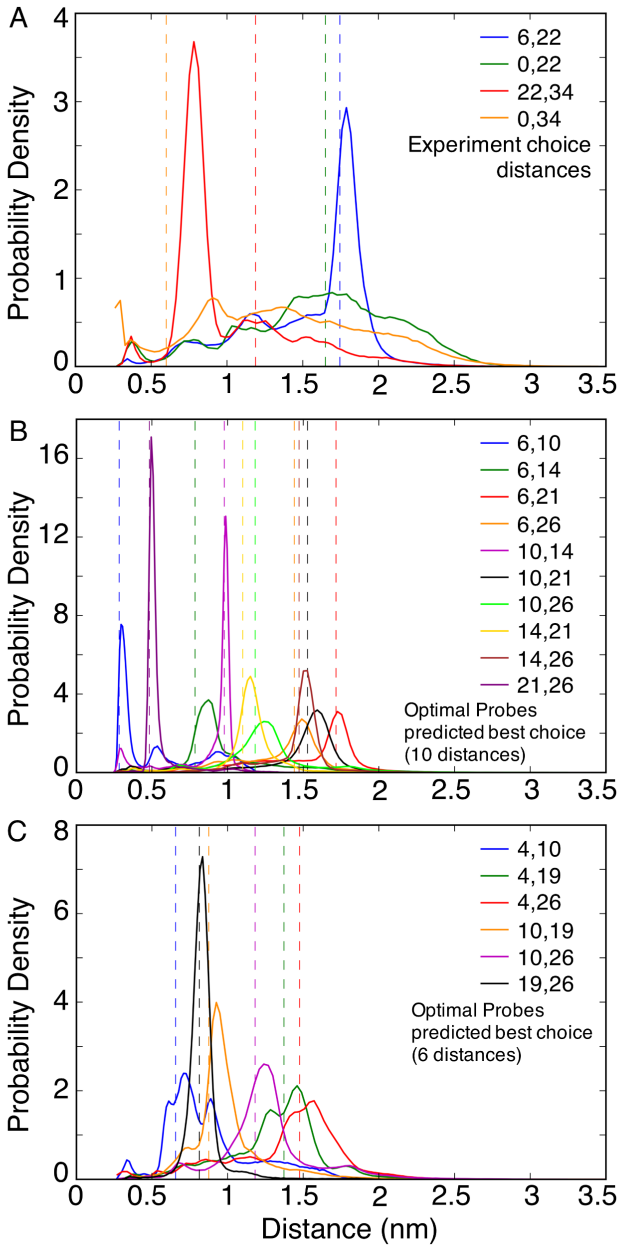
Supplementary Figure 3.8: Comparing top 10 implied timescales for the experimental distances MSM and Optimal Probes best choice distances MSM as a function of lag time on villin MD simulation dataset.



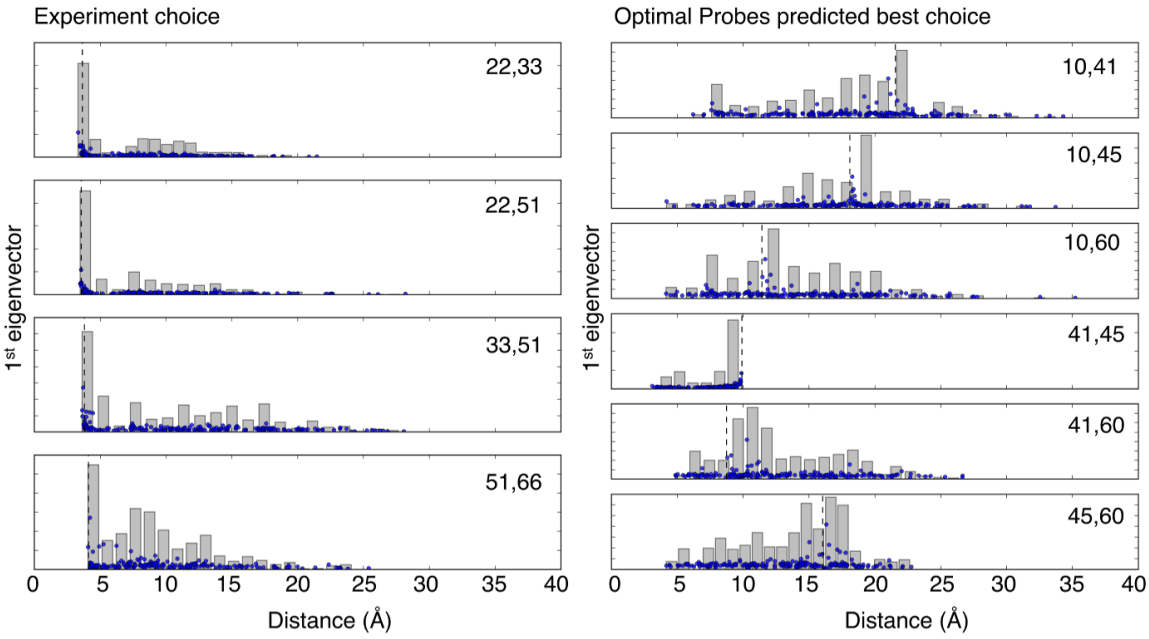
Supplementary Figure 3.9: (A) Cartoon representation of villin (unfolded structure) showing the experimental choice residue-pairs. (B) Cartoon representation of villin (unfolded structure) showing the Optimal Probes predicted best set of residue-pairs.



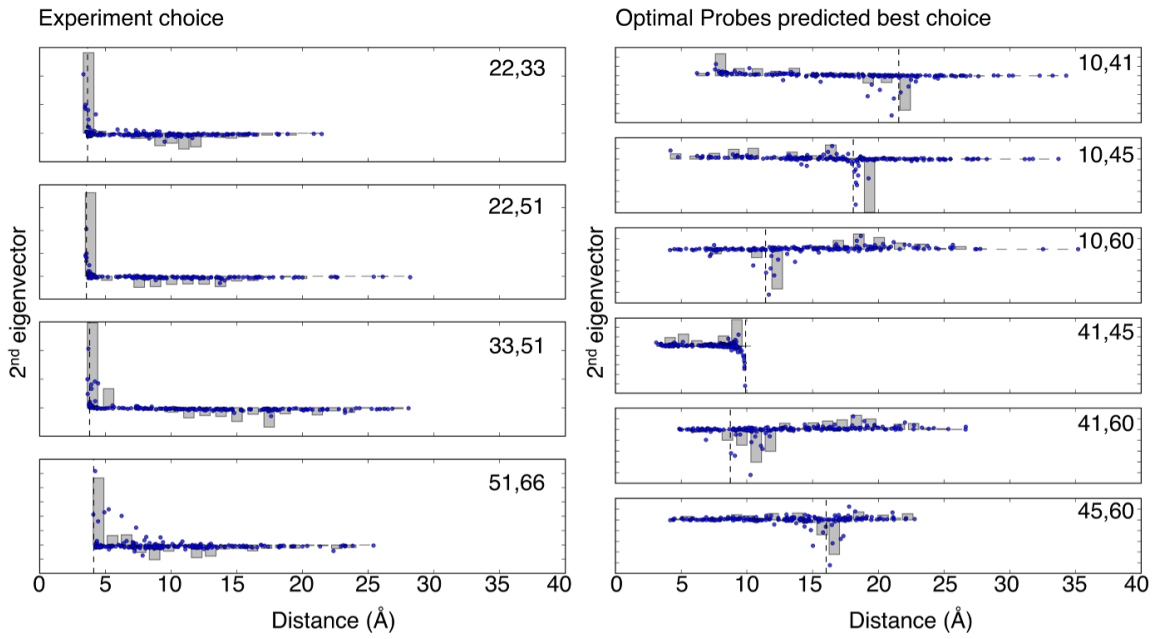
Supplementary Figure 3.10: Cartoon representation of villin showing the Optimal Probes predicted best set of residue-pairs with 6 distances.



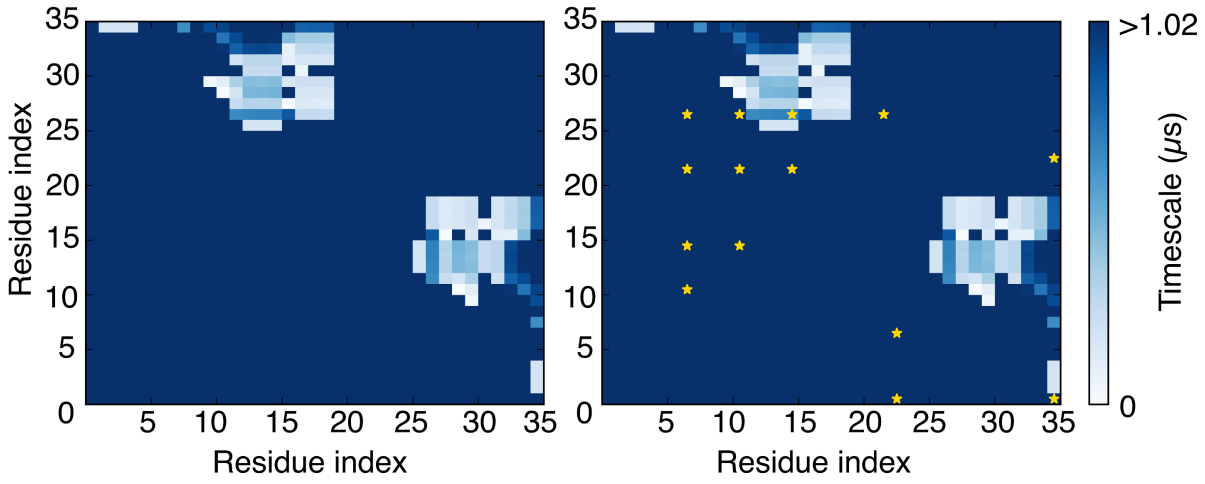
Supplementary Figure 3.11: Distance distribution for villin (A) experimental choice, (B) Optimal Probes predicted best choice (10 distances), and (C) Optimal Probes predicted best choice (6 distances) residue-pairs. The dotted lines are the distance values are observed in the native (folded) structure.



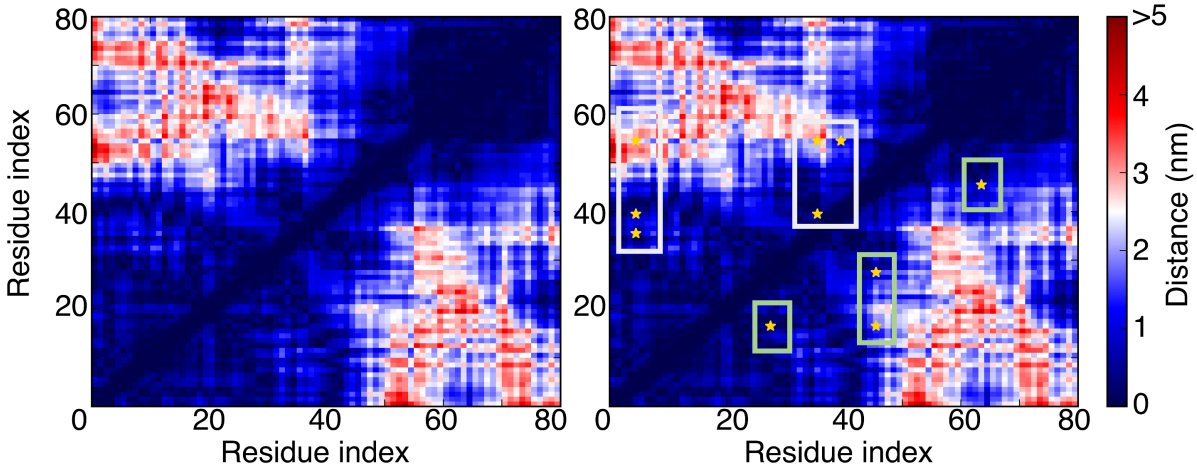
Supplementary Figure 3.12: Equilibrium population distribution of states in the λ-repressor MSMs. First eigenvector of the transition probability matrix shows the equilibrium population of all conformational states. Plot of the equilibrium populations projected onto distances in the experimental residue-pairs MSM (left) and Optimal Probes predicted best choice residue-pairs MSM (right). Blue dots are individual states. Histogram in gray is the binned sum of all states with a particular value of distance (x-axis). The dotted lines are the distance values are oberved in the native (folded) structure.



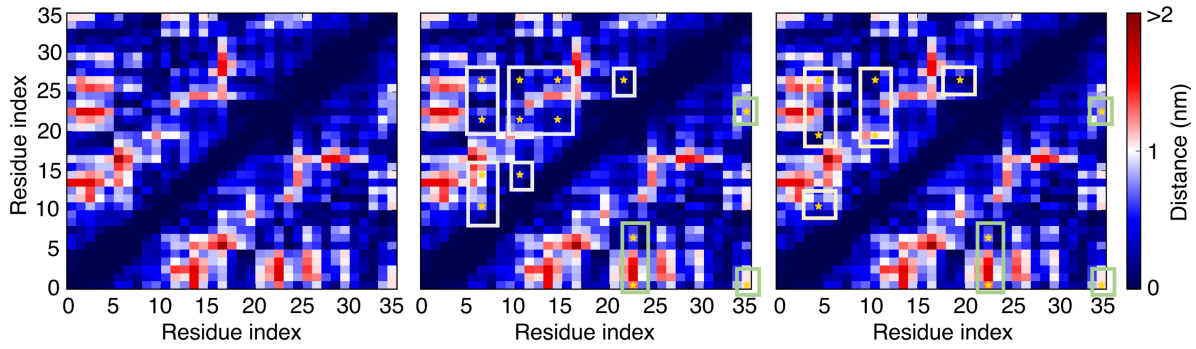
Supplementary Figure 3.13: Slowest dynamical process in the λ-repressor MSMs. Second eigenvector of the transition probability matrix shows the slowest process in the protein's dynamics. Plot of the second eigenvector projected onto distances in the experimental residue-pairs MSM (left) and Optimal Probes predicted best choice residue-pairs MSM (right). Blue dots are individual states. Histogram in gray is the binned sum of all states with a particular value of distance (x-axis). The dotted lines are the distance values are observed in the native (folded) structure.



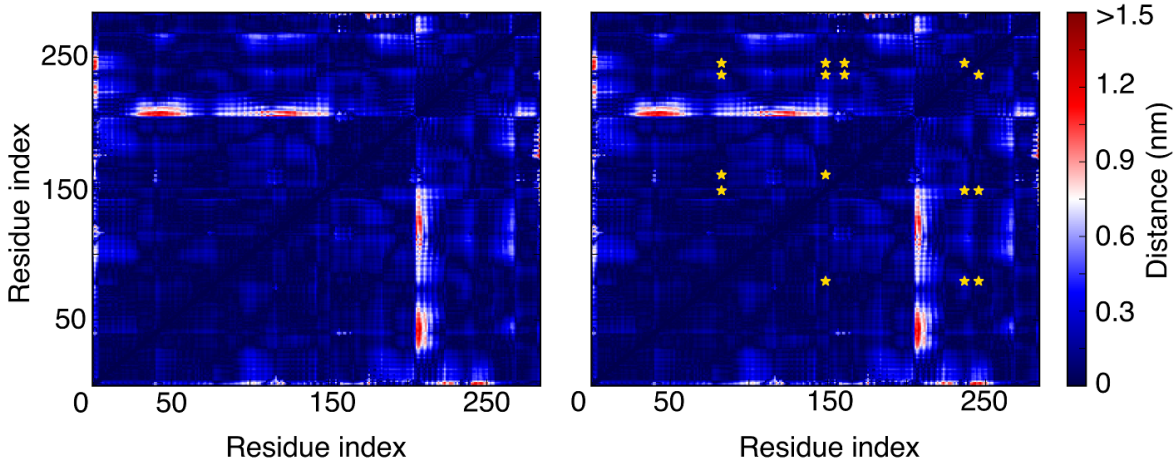
Supplementary Figure 3.14: TTET relaxation times for villin residue-pairs. The experimental residue-pairs in the lower triangle and the Optimal Probes predicted best choice residue-pairs (in the upper triangle) through yellow stars



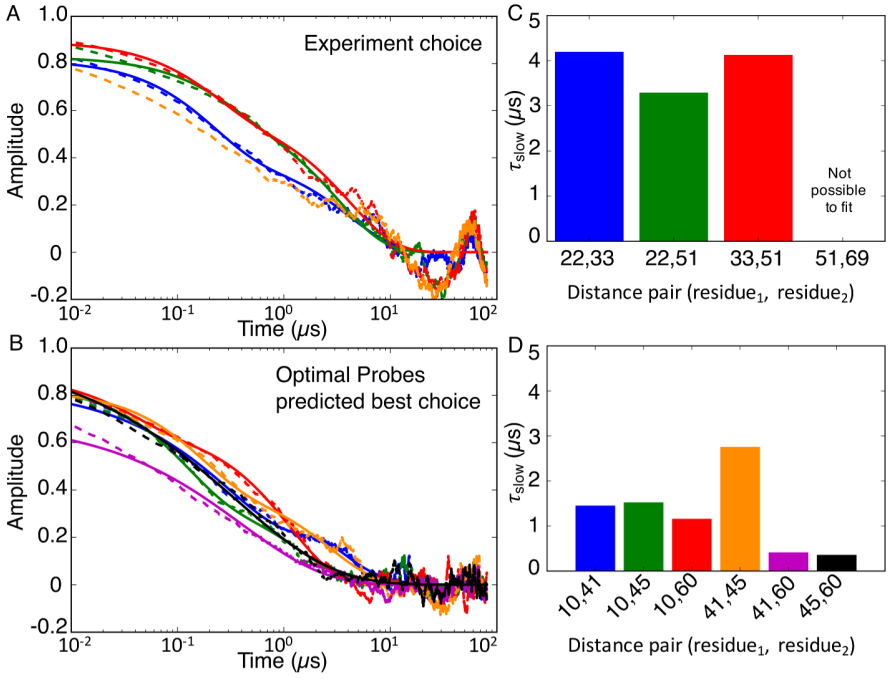
Supplementary Figure 3.15: Distance difference plot between λ-repressor folded and unfolded structure. Yellow stars marked on the distance difference plot on the right show the distances picked by experimentalists [146] and by Optimal Probes' best choice in the lower triangle (also shown in green boxes) and upper triangle (also shown in gray boxes), respectively.



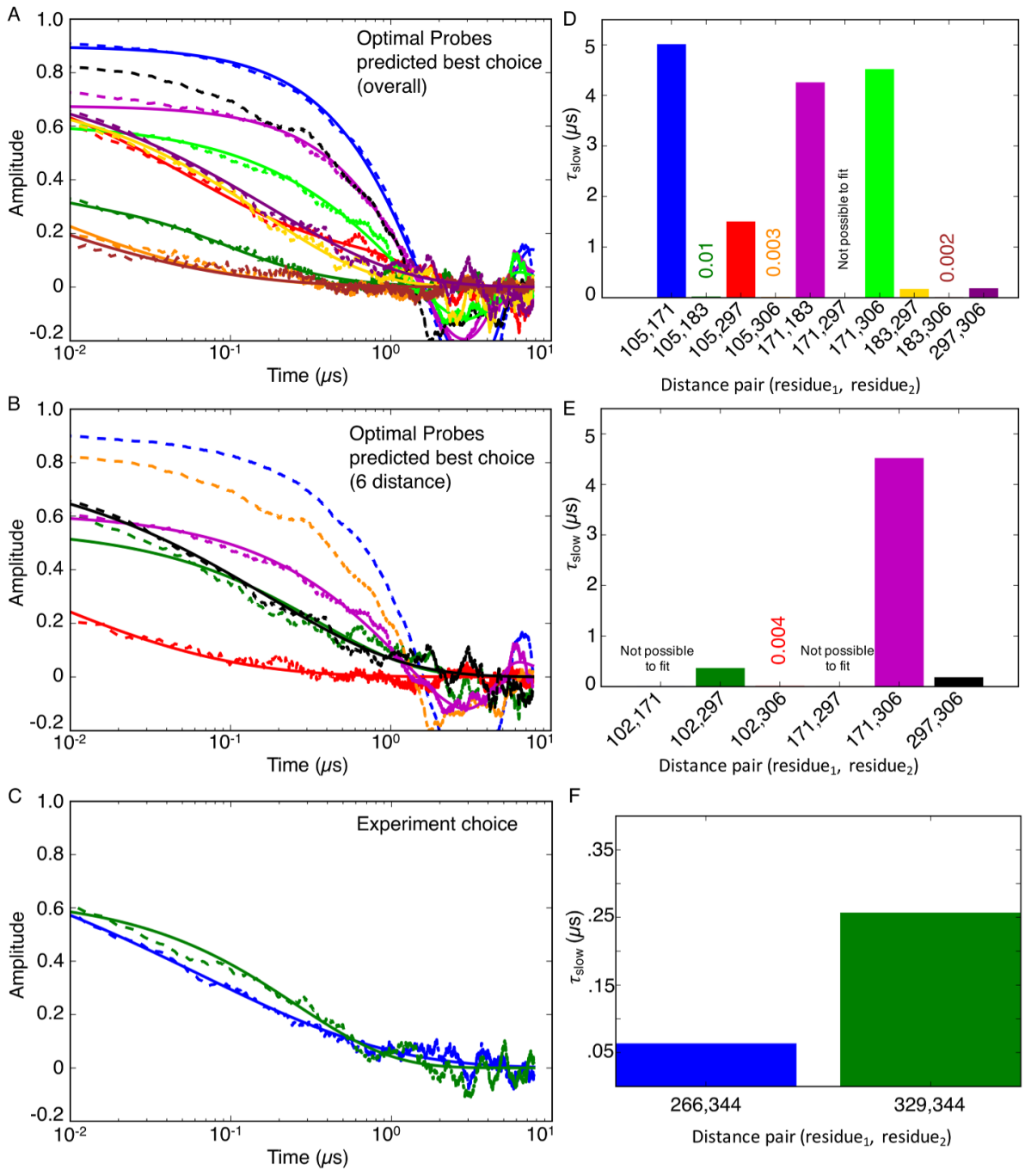
Supplementary Figure 3.16: Distance difference plot between villin folded and unfolded structure. Yellow stars marked on the distance difference plot show the distances picked by experimentalists [153] in the lower triangle (also shown in green boxes). The Optimal Probes' best choice is shown in the upper triangle (also marked in gray boxes) of the center color-map. The Optimal Probes' best choice with 6 distances is shown in the upper triangle (also marked in yellow boxes) of the right color-map.



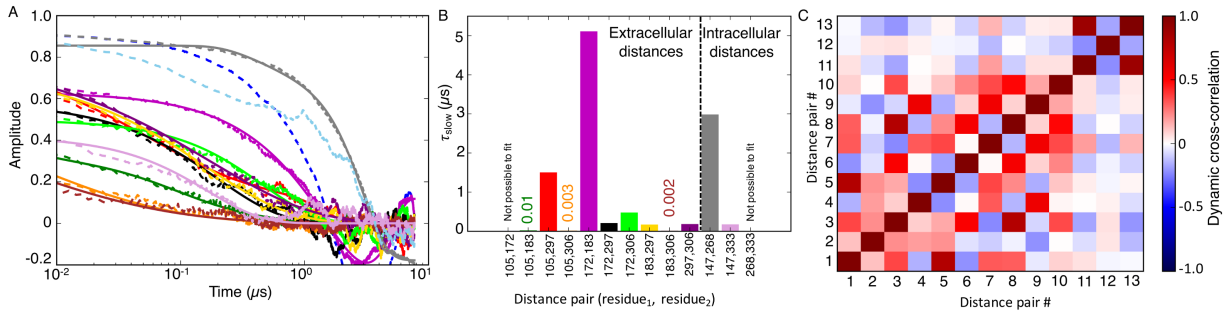
Supplementary Figure 3.17: Distance difference plot between β_2AR active and inactive structure. Yellow stars marked on the distance difference plot on the right show the distances picked by Optimal Probes' best choice (upper triangle) and Optimal Probes' best choice with 6 distances (lower triangle).



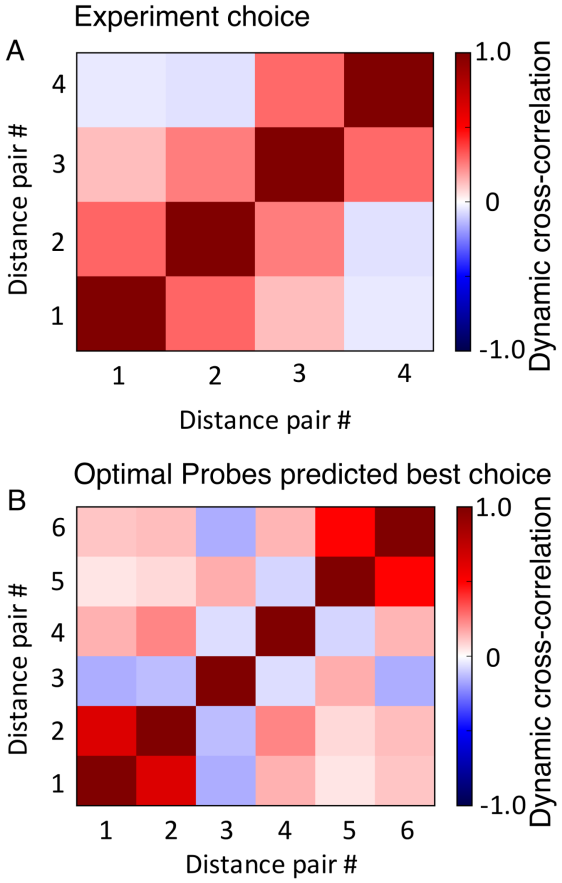
Supplementary Figure 3.18: Kinetic decay rates of residue-pairs for fluorescence experiments on λ-repressor. Autocorrelations from MD data (dashed lines) and the fits (solid lines) of the Dexter-weighted distance for residue-pairs in (A) experimental choice, and (B) Optimal Probes predicted best choice. τ_{slow} obtained using the fit of the Dexter-weighted distance for residue-pairs in (C) experimental choice, and (D) Optimal Probes predicted best choice.



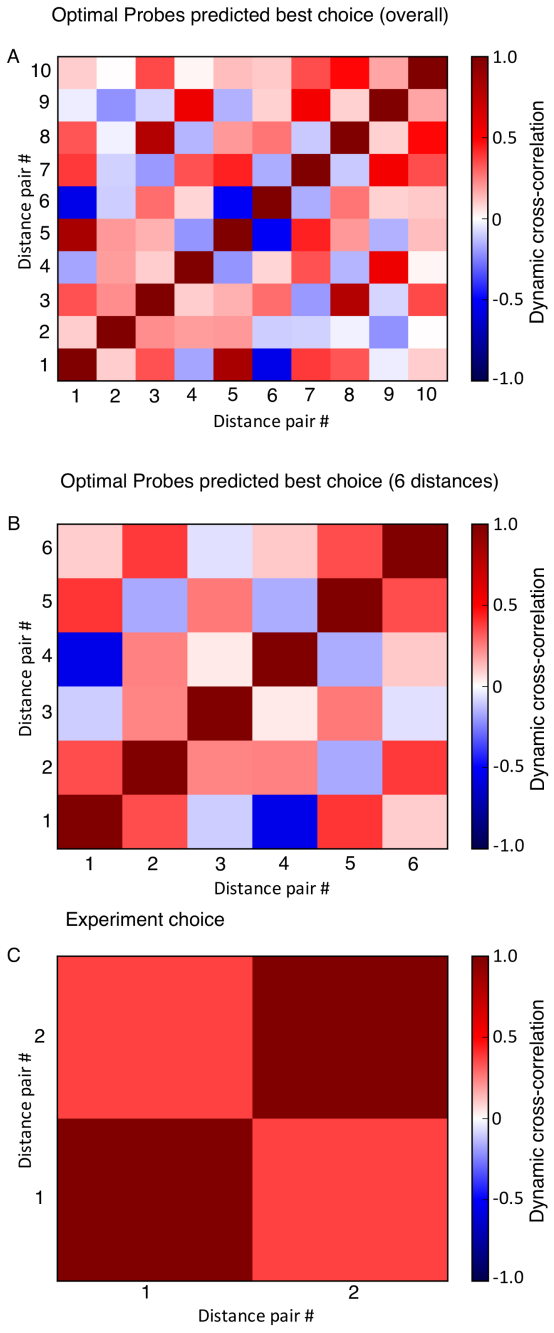
Supplementary Figure 3.19: Kinetic decay rates of residue-pairs for LRET experiments on β_2AR. Autocorrelations from MD data (dashed lines) and the fits (solid lines) of the Dexter-weighted distance for residue-pairs in (A) experimental choice, (B) Optimal Probes predicted best choice (overall), and (C) Optimal Probes predicted best choice with 6 distances. τ_{slow} obtained using the fit of the Dexter-weighted distance for residue-pairs in (D) experimental choice, (E) Optimal Probes predicted best choice (overall), and (C) Optimal Probes predicted best choice with 6 distances.



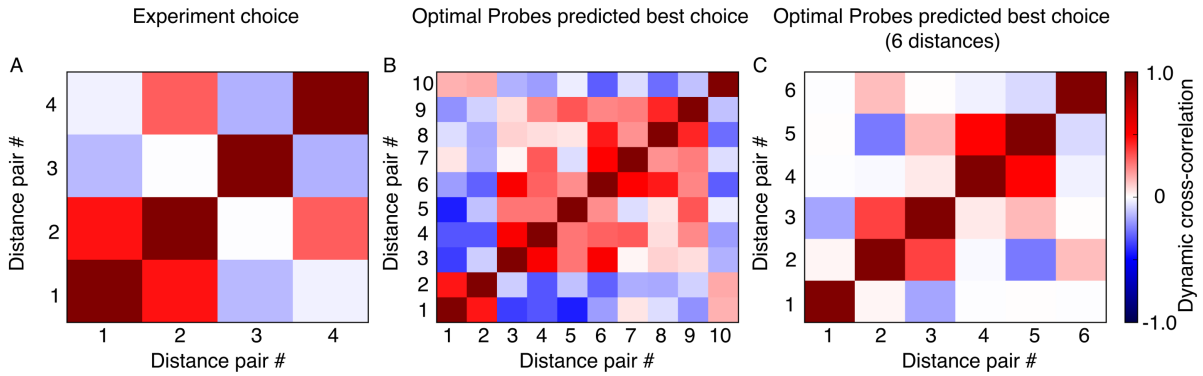
Supplementary Figure 3.20: Kinetic decay rates of residue-pairs and dynamic cross correlation among residue-pairs for LRET experiments on both, extracellular and intracellular sides of β_2AR. (A) Autocorrelations from MD data (dashed lines) and the fits (solid lines) of the Dexter-weighted distance for residue-pairs in the Optimal Probes predicted best choice. (B) τ_{slow} obtained using the fit of the Dexter-weighted distance for the 13 residue-pairs, 10 on extracellular side and 3 on intracellular side. (C) Pairwise DCC among the 13 residue-pairs.



Supplementary Figure 3.21: Pairwise DCC among residue-pairs in (A) experimental choice, and (B) Optimal Probes best predicted choice for λ-repressor.



Supplementary Figure 3.22: Pairwise DCC among residue-pairs in (A) experimental choice, (B) Optimal Probes best predicted choice (overall), and (C) Optimal Probes best predicted choice (6 distances) for β_2AR.



Supplementary Figure 3.23: Pairwise DCC among the residue-pairs in experiments, best predicted choice (overall), and best predicted choice with 6 distances for villin.

```
B2AR     1 MGQPGNGSAFLLAPNGSHAPDH--------DVTQE-----RDEVWVVGMGIVMSLIVLAI
V2R      1 ---------MLMASTTSAVPGHPSLPSLPSNSSQERPLDTRDPLLARAELALLSIVFVAV

B2AR    48 VFGNVLVITAIAKFERL--QTVTNYFITSLACADLVMGL-AVVPFGAAHILMKMWTFGNF
V2R     52 ALSNGLVLAALARRGRRGHWAPIHVFIGHLCLADLAVALFQVLPQLA-------------

B2AR   105 W--CEFWTSIDVLC----------VTASIETLCVIAVDRYFAITSP---FKYQSLLTKNK
V2R     99 WKATDRFRGPDALCRAVKYLQMVGMYASSYMILAMTLDRHRAICRPMLAYRHGSGAHWNR

B2AR   150 ARVILLMVWIVSGLTSFLPIQMHWYRATHQE----AINCYANETC-CDFFTNQAYAIASS
V2R    159 P---VLVAWAFSLLLS-LP-QLFIFAQRNVEGGSGVTDCWA---CFAEPWGRRTYVTWIA

B2AR   205 IVSFYVPLVIMVFVYSRVFQEAKRQL--QKIDKSEGRFHVQNLSQVEQDGRTGHGLRRSS
V2R    211 LMVFVAPTLGIAACQVLIFREIHASLVPGPSERPGGR------RRGRRTGSPGEGAHVSA

B2AR   263 KFCLKEHKALKTLGIIMGTFTLCWLPFFIVNIVHVIQDNL-IRKEVYILLNWIGYVNSGF
V2R    265 AVA----KTVRMTLVIVVVYVLCWAPFFLVQLWAAWDPEAPLEGAPFVLLMLLASLNSCT

B2AR   322 NPLIYCRSPDFRIAFQELLCLRRSSLKAYGNGYSSNGNTGEQSGYHVEQEKENKLLCEDL
V2R    321 NPWIYA-------------------------SFSSS----------VSSELRSLLCCA--

B2AR   382 PGTEDFVGHQGTVPSDNIDSQGRNCSTNDSLL------
V2R    344 ---------RGRTP-PSLGPQDESCTTASSSLAKDTSS
```

Supplementary Figure 3.24: A sequence alignment of the human V2R and β_2AR proteins obtained using T-Coffee [193].

Chapter 4

Free Energy Landscape of the Complete Transport Cycle in a Key Bacterial Transporter[1]

4.1 Overview

$PepT_{So}$ is a proton-coupled bacterial symporter, from the major facilitator superfamily (MFS), which transports di/tri-peptide molecules. Recently obtained crystal structure of $PepT_{So}$ provides unprecedented opportunity to gain an understanding of functional insights of substrate transport mechanism. Binding of proton and peptide molecule induces conformational changes into occluded (OC) and outward-facing (OF) states, which we are able to characterize using molecular dynamics simulations. The structural knowledge of OC and OF state are important to fully understand the major energy barrier associated with the transport cycle. In order to gain functional insight into the interstate dynamics, we performed extensive all atom molecular dynamics simulations. Markov state model (MSM) was constructed to identify the free energy barriers between the states and kinetic information on intermediate pathways was obtained using transition path theory (TPT). TPT shows that OF state is obtained by the movement of TM1 and TM7 at the periplasmic side approximately 12-16 Å away from each other and the inward movement of TM4 and TM10 at the cytoplasmic halves to 3-4 Å characterizes the OC state. Helix distances distributions obtained from MD simulations were compared with experimental double electron-electron resonance (DEER) spectroscopy. Our finding sheds light on the conformational cycle of this key membrane transporter and the functional relationships between the multiple intermediate states.

4.2 Introduction

Understanding the exchange of biological molecules across the phospholipid bilayer is of fundamental interest to the scientific community. The transport of a large variety of molecules is essential for the pathological physiological functions of the cell. In particular, peptide transporter proteins act as carriers that facilitate the transport of small groups amino acids across the cell membrane. These transporters belong to the peptide

[1]This chapter is adapted reproduced with permission from Selvam B, Mittal S, Shukla D. ACS Central Science. 2018; 4(9):1146-1154. Copyright 2018 American Chemical Society. BS and SM contributed equally to this work.

transporter (PTR) family which are members of the major facilitator superfamily (MFS) of secondary active membrane transporters [195]. To understand the functional mechanism of MFS transporters, Kaback et al. proposed an alternate access model which involves alternate accessibility to either side of the membrane via distinct inward facing (IF), occluded (OC) and outward facing (OF) conformational states [196]. Another model is the rocker-switch model where the substrate is posited to bind to the center of the transporter, causing rigid body motion of the N and C domains to alternate access between the intracellular and extracellular sides [197]. The third hypothesis is the elevator-like model where the substrate binds to a single domain, locks the pore channel, slides downwards and opens up to release the substrate [198–200]. Even with progress in X-ray crystallographic techniques the atomic level structural information of the distinct states and other intermediate states are still unknown. In humans, peptide transporters, $PepT_1$ and $PepT_2$ are expressed in the intestine and kidney and are actively involved in the uptake of dietary peptide molecules [201]. Hence, peptide transporters are considered as crucial targets for drug delivery and a means to improve the pharmacokinetics of drug molecules [202]. Despite the known importance of peptide transporters, the structural changes involved in the overall functional dynamics have not been studied comprehensively. In this work, we have chosen a bacterial peptide transporter from *Shewanella oneidensis*, $PepT_{So}$, to understand the conformational dynamics and mechanistic transport in this peptide transporter and by extension of the POT family.

$PepT_{So}$ is a bacterial proton coupled oligopeptide transporter (POT) consisting of 12 transmembrane helices (TMs) divided into N (TM1-TM6) and C (TM7-TM12) terminal domains each with six helices [111,203]. The N and C domains are connected by two short helices (SHs) which are closely packed to the C-terminal domain. $PepT_{So}$ uses an inward electrochemical potential as the driving force to transport di/tri-peptide molecules into the cell against their concentration gradient. $PepT_{So}$ shares close sequence similarity and structural homology with other peptide transporters $PepT_{St}$ [204–206], GkPOT [207], $PepT_{So2}$ [208,209], YePEPT [210] and $PepT_{Xc}$ [211]. $PepT_{So}$ and YePEPT were crystallized in the apo form (other peptide transporters were crystallized as holo structures) in the IF state. Biochemical and biophysical experiments provide insights into functional behavior of the transporter proteins, dynamics information on conformational transition between different structural states is missing. The repeat swap method (RSM) was used as an alternative protocol to obtain the OF conformational state of $PepT_{So}$ [212]. The method involves swapping the N and C domains to create a template, aligning the swaps to the template coordinate file and finally constructing the homology model to obtain the RSM model. However, the some of the limitations of this technique are that it requires careful assessment of the sequences for structural alignments of the internal repeats, cannot provide an array of intermediate states and no information is gained regarding the dynamics of the

process of peptide transport. Limited computational methods have been employed to study the functional mechanisms and conformational dynamics of transporters [116, 213–215]. These methods contributed to the investigation of substrate driven structural changes of transporters, but failed to obtain the kinetics of the crucial conformational transitions which would be critical to identify the transitions among key intermediates states which are essential to the functional mechanism. Evidently, these transporters exhibit large conformational changes, which cannot be understood using the static snapshots provided by X-ray crystallographic studies, and hence the mechanism of peptide transport is still unknown.

MD simulations are an appealing methodology to determine the conformational changes and a mechanistic insight into $PepT_{So}$. In our study, we performed atomistic MD simulations over a duration of $\sim$54 μs using an adaptive sampling approach (see Methods). Further we constructed a Markov state model (MSM), a technique that has been used extensively to study conformational diversity in biological systems [45, 47, 55, 82, 216]. Using MSMs we were able to combine a large number of shorter simulations and perform efficient analysis on the huge amount of the data. To estimate the timescales of the transitions between the intermediates in the transport cycle, trajectories were reconstructed using the kinetic Monte Carlo approach. MD simulations have unraveled the OC and OF states of $PepT_{So}$ along with other intermediate states that the protein can adopt to enable transport. We compared the helix distance distributions from our MD simulation ensemble to the experimental double electron-electron resonance (DEER) spectroscopy results. To our knowledge, this is the first long timescale MD simulations based thermodynamics and kinetics study that captures key intermediate states and the functional landscape of $PepT_{So}$ using unbiased computational methods.

4.3 Methods

Molecular dynamics simulations

The crystal structure of $PepT_{So}$ was used as a starting structure for MD simulation. The 3D coordinates (PDB: 4UVM [111]) were obtained from Protein Data Bank. The tleap program in AmberTools14 [14] was used to build the MD system. The protein was solvated in a phospholipid bilayer (POPC) in an orthorhombic box containing TIP3P water molecules [217] in a periodic box size 98×98×119 $Å^3$. A salt (NaCl) concentration of 0.15M was used to neutralize the MD system. All chain termini were capped with neutral acetyl and methylamide groups. The standard protonation states was used for the titratable groups and the final MD system contained approximately 110,000 atoms. The MD system was energy minimized for 20,000 steps using the conjugate gradient method, slowly heated from 0 to 300 K and equilibrated for 40 ns.

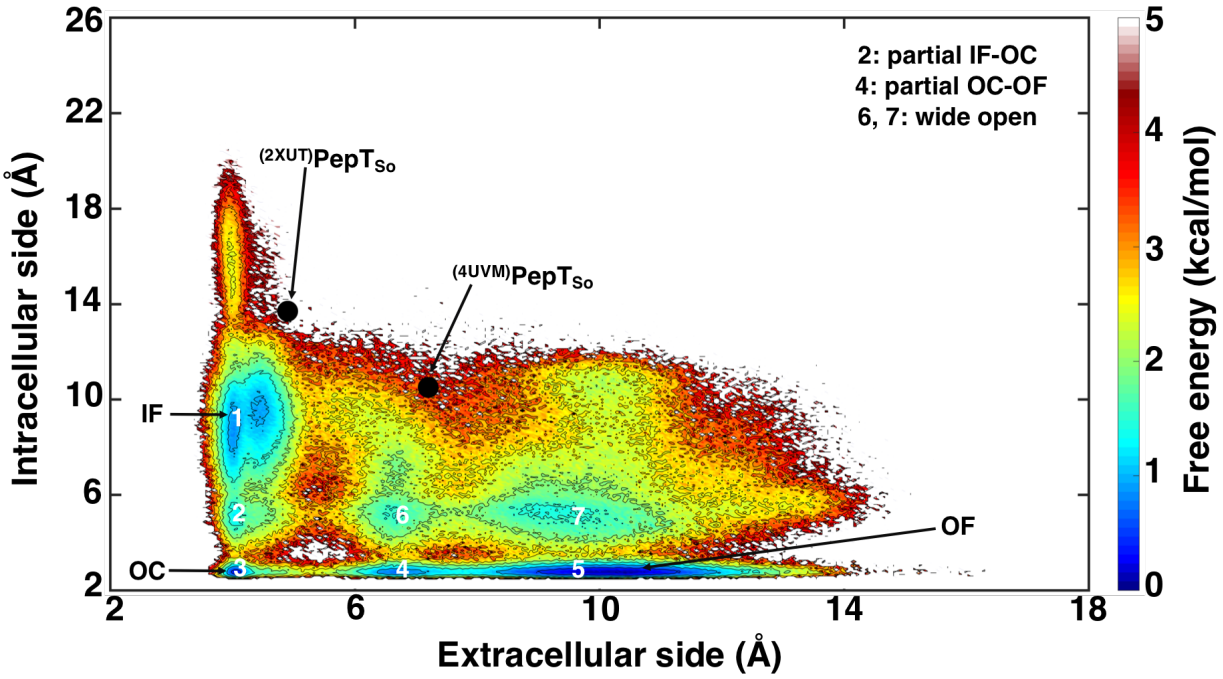


Figure 4.1: Conformational landscape of $PepT_{So}$. The conformational landscape is generated using the extracellular and intracellular side distances measured between atom pairs Arg32-CZ (TM1)-Asp316-CG (TM7) and Ser131-CO (TM4)-Tyr431-OH (TM10), respectively. The conformational states are depicted as IF (1), partial IF-OC (2), OC (3), partial OC-OF (4), OF (5) and wide open states (6 and 7). The black dots indicate the $PepT_{So}$ crystal structures available in the protein data bank.

The MD simulations were performed in constant NPT conditions at 300 K and 1 atm. The temperature was controlled using a Berendsen thermostat and the pressure was maintained using Berendsen barostat [218]. Long range electrostatic interaction was treated with the Particle Mesh Ewald method [219] and bonds involving hydrogens were constrained using the SHAKE algorithm [220]. The non-bonded distance cutoff was set to 10 Å and an integration step of 2 fs was used. All simulations were performed using the AMBER FF14SB force field [221].

An adaptive sampling approach was used to select the new starting structures for the subsequent MD runs to enhance the conformational sampling of the free energy landscape. For each round, the previous sampled data was clustered using the k-means algorithm based on the extracellular and intracellular experimental DEER residue-pair distances, and the least populated states were chosen to conduct the next round of simulations. The sampling bias introduced in the dataset from seeding new trajectories in this manner is eliminated in the way an MSM (discussed below) is constructed on the data [222]. Adaptive sampling is a widely used sampling methodology, and has been used to predict novel conformations of the proteins, pathways of conformational change, protein folding, and even protein-protein association [45,47,82,223–225]. In addition to unbiased MD simulation data obtained as above, 5 μs of accelerated MD (aMD) simulation were also performed also using the adaptive sampling protocol (Supplementary Table 4.1). For aMD, a boost

potential (4 kcal/mol) was added to the dihedrals of the protein and a further boost potential (0.2 kcal/mol) was added to the entire MD system. The integration step chosen is 3 fs [226]. The free energy landscape in shown in Supplementary Figure 4.21. The aMD simulation data was clustered and the starting structures were chosen for classical MD (cMD) to sample the conformational landscape efficiently (Supplementary Table 4.2). The final cMD was performed for a total duration of $\sim$54 μs. Each individual MD trajectory is of $\sim$34 ns.

Markov state models

The MSMBuilder3.4 Python package [69] was used to build the MSM on the $PepT_{So}$ trajectory data. The seven transmembrane helical distances and three extracellular and intracellular residue-pair distances were chosen as featurization metrics to construct an MSM (Supplementary Figure 4.22). 24 ns was determined to be a Markovian lag time from the implied time-scales plot (Supplementary Figure 4.23). The number of clusters was chosen to be 200 as it yielded the highest generalized matrix Rayleigh quotient (GMRQ) score while building multiple MSMs on varying this hyper-parameter (Supplementary Figure 4.24) [125]. TPT analysis was performed to obtain the top flux pathway (Supplementary Figure 4.25 and Supplementary Figure 4.26).

Kinetic Monte Carlo simulations

Kinetic Monte Carlo is a method for sampling from a kinetic model, which can be used to create trajectories of state-to-state dynamics. For any chosen initial state i, a transition to any state j from the set of all states in the MSM occurs with probability p_{ij} from the MSM's reversible maximum-likelihood transition matrix. This is implemented as, (1) generate a pseudo-random number between 0 and 1, (2) take a cumulative sum of p_{ij} values over all possible j ($S_n = \sum_i^n p_{ij}$), and if the pseudo-random number lies between S_n and S_{n+1}, (3) transition to state $j = n + 1$. This state, j, is added to the trajectory and the process is repeated for the desired number of steps.

DEER distance distribution analysis

To validate our predicted structures with the DEER experiments, we constructed an augmented Markov state model (AMM) [178] for each experimental DEER distribution as constraints using pyEMMA v2.4+936.g26d8e55 [70]. The list of constraints is provided in Supplementary Table 4.3. We extracted 50 structures from the highest weighted clusters (in the MSM or the AMM) for each of the following conformations: IF, OC, OF, partial IF-OC, partial OC-OF, and two wide-open states. Using the Python library RotamerConvolveMD [108],

we calculate the distances between the MTSSL probe conformations, mapped onto the rotamer library, for each pair of residues on all the structures. The raw distance information is then plotted as a histogram and the distance range spanned by the distance distribution in order to compare to the experimental information.

4.4 Results

MD simulations reveal the conformational sub-states of $PepT_{So}$

The IF state crystal structure of $PepT_{So}$ (PDB: 4UVM [111]) was used as the starting conformation for simulations. The MD simulations were conducted using an adaptive sampling approach (see Methods). From our simulation data, we were successfully able to identify the functionally important intermediate states. The free energy landscape projected onto the extracellular and intracellular distances, weighted by the MSM equilibrium probability distribution, reveals the energy minima corresponding to conformations states, IF (1), partial IF-OC (2), OC (3), partial OC-OF (4), OF (5) and wide open intermediate states (6 and 7) (Figure 4.1). The extracellular and intracellular side distances determined between the residue-pairs Arg32-CZ (TM1)-Asp316-CG (TM7) and Ser131-CO (TM4)-Tyr431-OH (TM10), respectively, show that the helices of the N and C domains are ∼7.0 and ∼10.5 Å apart in the $PepT_{So}$ crystallized IF state (PDB: 4UVM). We observe that the intracellular distance between TM4 and TM10 may increase up to ∼20 Å albeit resulting in a high energy unstable state. $PepT_{So}$ adopts a partial IF-OC state as the intracellular helical distance between TM4-TM10 reduces to 5-6 Å by a movement of both helices inwards towards each other. Further, this distance reduces to 3 Å leading to the OC state. The transition from OC to OF involves two stages; first, the moving apart of TM1 and TM7 at the extracellular side to a distance of 7-8 Å by forming an intermediate partial OC-OF state, and second, extending the distance of the extracellular vestibule of TM1 and TM7 up to 12-16 Å (Supplementary Figure 4.1, Supplementary Figure 4.2). The IF, OC and OF states can be distinguished by passing a spherical probe from one side of the protein to the other, calculated using the HOLE program [227]. The probe radius through the different states is visualized in Figure 4.2 and the residues used to determine intracellular and extracellular distances in our study are indicated, as they constrict the IF, OC and OF states.

To determine the conformational exchange between these intermediate states, we performed transition path theory (TPT) analysis on MSM (see Methods) [228]. All high-ranked paths undergo a transition from the IF to OF via the OC state and other intermediate states. The free energy barrier for the transition from IF to OC via the partial IF-OC state is ∼2-2.5 kcal/mol and for subsequent transition to OF through states partial OC-OF or a wide open state is ∼1.5-2 kcal/mol. Hence, the total free energy barrier of ∼4 kcal/mol

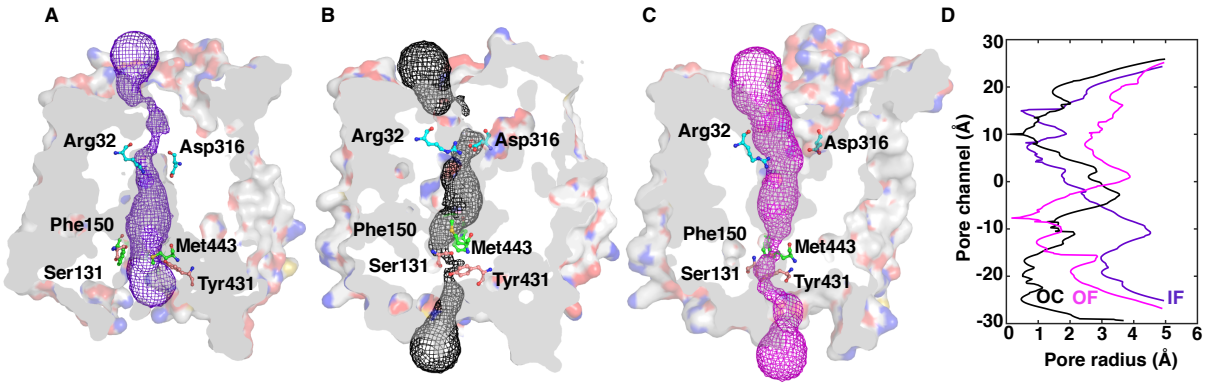


Figure 4.2: The distinct conformational states of $PepT_{So}$ are visualized by passing a spherical probe from one side of the protein to the other through (*A*) the crystal structure IF state, and for the MD simulations predicted structures for (*B*) OC and (*C*) OF states, calculated using the HOLE program [227]. The gating residues Arg32 (TM1)-Asp316 (TM7) and Ser131 (TM4)-Tyr431 (TM10) that act as bottlenecks for the conformational transition are indicated. (*D*) The pore radius along the protein for the three conformational states.

was determined for one complete cycle from IF to OF at equilibrium.

Structural characteristics of the OC state

The predicted OC state shows large deviations in the intracellular side of N and C domains as compared to the IF state. The intracellular halves of the helices rearrange in several positions as compared to the crystal structure - TM1 moves inwards and closely packs with TM3 and TM4; TM10 and TM11 undergo inward movement and interact with TM2 and TM4, respectively; TM4 and TM5 are tilted ∼5° and ∼10°, respectively and move inwards closer to the center of the transporter. TM2 becomes more straight compared to the kinked helix in the IF crystal structure. On the extracellular half - TM7 and TM8 are rotated by ∼15° and ∼8°, respectively, and move closer to TM1; TM9 and TM10 move up to 6 Å and 11 Å outwards and form extensive contacts with the loops joining TM7 and TM8 (Supplementary Figure 4.3).

The OC state is stabilized by extensive intramolecular hydrogen bonds between the N and C domains. At the extracellular side, the residues Asn33 (TM1), Ser165 (TM5), Ser320 (TM7) and Gln341 (TM8) form a hydrogen bond network that acts as the lid by packing the helices close to each other (Supplementary Figure 4.4). The closure of the pore channel is further stabilized by Arg32 (TM1)-Asp316 (TM7) salt bridge interaction on this end of the transporter. Asp316 of this ionic residue-pair is known to play a major role in proton driven peptide transport [203, 205, 212, 229]. The Glu419 (TM10) interaction with Lys318 (TM7), Thr416 (TM10) and Asn344 (TM8) stabilizes the conformation of TM7 and 8. The Glu419 (TM10) is conserved residue in POT family and its mutation to Gln results in loss of proton driven peptide transport [205]. Another conserved motif, ExxERxxY on TM1 (Supplementary Figure 4.5) and its interaction with Lys127 (TM4) plays a crucial role in peptide transport and its mutation abolishes the transport function

[205].

On the intracellular side, the Pro127 (TM4) introduces helix kink which facilitates the inward movement of TM4 and TM5. The glycine residues (Gly418 (TM10), Gly426 (TM10) and Gly440 (TM11)) increase the structural flexibility allowing the helices to twist. The Tyr431-OH (TM10) establishes a hydrogen bond contact with Ser131-CO (TM4) and stabilizes the OC state (Supplementary Figure 4.4). Our predictions conform with the hypothesis of Stelzl et al. for the extracellular and intracellular gating residues that are critical for the conformational switch and functional mechanism of transporters [108]. A comparison of the predicted OC state has been made with the multidrug transporter EmrD (PDB: 2GFP [230], Supplementary Figure 4.6, Supplementary Figure 4.7) and xylose transporter XylE (PDB: 4GBY [231], Supplementary Figure 4.8, Supplementary Figure 4.9), MFS family OC structures in the protein data bank. The observed gating residues Thr119 (TM4)-Phe311 (TM10) in EmrD OC and Met149 (TM4)-Ser396 (TM10) in XylE OC state are comparable to $PepT_{So}$.

Structural characteristics of the OF state

The predicted OF structure reveals dramatic rearrangements at the extracellular side compared to IF (crystal structure) and OC (predicted structure) whereas only subtle changes at the intracellular side compared to OC (Supplementary Figure 4.10). TM7 rotates $\sim 7^{\circ}$ as compared to IF and Asp316 (TM10) moves $\sim$12 Å away from Arg32 (TM1) forming a new contact with Asn454 (TM11) which then interacts with His61 (TM2). Our predictions are consistent with the structure proposed by Parker et al. [211]. His57 (His61 in $PepT_{So}$) is involved in proton driven peptide transport in $PepT_1$ [232–234]. Thus, the extracellular binding partners move away from each other and increases the channel viability and hence adopts the OF state. TM7 is stabilized by an interaction between Lys318 (TM7) and Glu419 (TM10) at the extracellular part of $PepT_{So}$. The ExxERxxxY motif forms similar contacts to those seen in the OC state. The kink cause by Pro71 in TM2 results in slight bending of the helix allowing Trp76-O (TM2) to form a contact with Thr441 (TM11). The interaction of Tyr431-OH (TM10) with Ser131-CO and Gly135-O on TM4 stabilizes the intracellular half of the OF state.

Our predicted OF structure from MD simulations shows good agreement with the fucose transporter (FucP, PDB: 3O7Q), which was crystallized in OF state [235], with an RMSD of 2.9 Å (Supplementary Figure 4.11, Supplementary Figure 4.12). The intermediate states 6 and 7 are predicted to be wide open states that may also lead to transitions to OF state. The $PepT_{So}$ OF state predicted using RSM [111] was compared to predicted OF structure and has an RMSD of 3.6 Å (Supplementary Figure 4.13, Supplementary Figure 4.14).

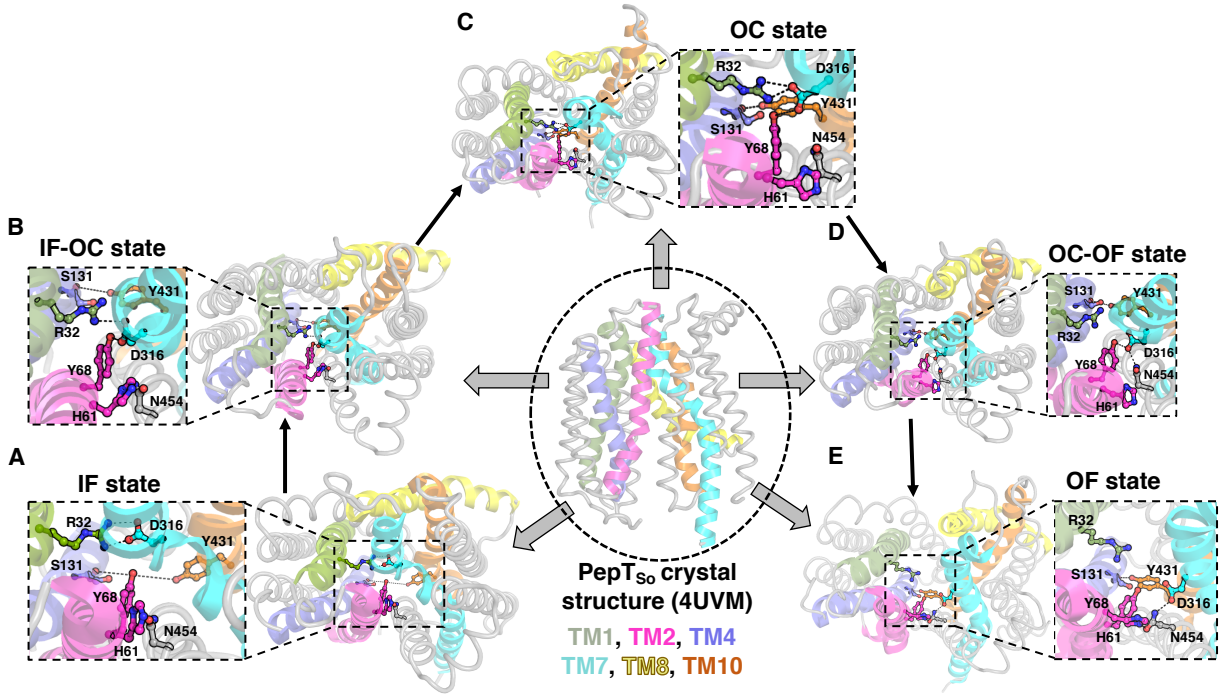


Figure 4.3: $PepT_{So}$ is shown in the center with TM 1, 2, 4, 7, 8, 10 in green, magenta, blue, cyan, yellow, and orange respectively. The short helices (SH) which join N and C domains are not shown for clarity. Remaining 6 helices are in gray. All state conformations are the extracellular view of the protein. (*A*) IF state is stabilized by ion-lock at the periplasmic side and the cytoplasmic half is wide open. (*B*) Inward movement of TM4 and TM10 determines the partial IF-OC state. (*C*) Further inward movement leads to formation of hydrogen bond interaction between Tyr431-Ser131 in OC state. (*D*) Gating residues at the cytoplasmic side weaken the extracellular interaction to form partial OC-OF state. (*E*) Helices TM1 and TM7 move far away to 15 Å in OF state.

Switching of gating residues determines conformational changes in $PepT_{So}$

From TPT, we identified that the interaction between Arg32-CZ (TM1)-Asp316-CG (TM7) acts as the gating bottleneck on the extracellular side of the transporter. In the IF crystal structure (PDB: 4UVM), the distance between these atoms is 7.2 Å. However, our simulations reveal that these residues come closer and form a salt bridge interaction locking TM1 and TM7 to characterize an IF state (Figure 4.3A). Previous studies have also shown that the IF state is stabilized by the formation of this ionic-lock [205]. On the intracellular side, Newstead et al. observe that the hydrophobic gate between Phe150-CB (TM5) and Met443-CB (TM11) form the intracellular gate to characterize the OC state [203]. We find these residues determine only the partial IF-OC state (Figure 4.3B), Supplementary Figure 4.15, Supplementary Figure 4.16). After $PepT_{So}$ adopts the partial IF-OC state, an additional hydrogen bond between Ser131-CO (TM4) and Tyr431-OH (TM10) results in complete transition to OC state (Figure 4.3C).

Next, the partial OC-OF state is obtained as the distance between Arg32-CZ (TM1) and Asp316-CG (TM7) increases up to 7-8 Å. The weakening of ionic interaction between Arg32 (TM1) and Asp316 (TM7) leads Asp316 to form a new polar contact with Tyr68 (TM2). Further, it establishes a contact with Asn454 (TM11) which results in complete loss of interaction with Arg32 (Figure 4.3D). The periplasmic halves of

the N and C terminal domains start moving further away from each other (on the extracellular side) and results in loss of Tyr68 interaction with Asp316. His61 (TM2), Asn454 (TM11) and Asp316 (TM7) form a hydrogen bond triad, thus adopting the final OF state (Figure 4.3E). The predicted OF state reveals that the distance between the residue-pairs Arg32 (TM1) and Asp316 (TM7) may increase up to 12-16 Å to recognize the substrate molecule and initiate the transport cycle.

Using a kinetic Monte Carlo reconstructed trajectory of length 25 μs from the constructed dataset, the mean passage time for the transition from IF to OF, via OC was found to be $\sim$1 μs (Supplementary Figure 4.17). This synthetic trajectory reveals that the rates of transitions between alternate opening of extracellular and intracellular gates shows faster dynamics as compared to previous studies where they have obtained the transition from IF to partial IF-OC and to OF in $\sim$0.6 μs [111, 211].

Comparison of predicted OC and OF structures with experiments

DEER spectroscopy is a biophysical technique that allows to determine residue-pair distance distributions between two cysteins that have been modified via site-directed spin labeling (SDSL). The technique has been used for widely used for the study of membrane proteins [106] where multiple peaks in the distributions indicate a diverse conformational composition of the protein during the experiment. Atom-pair distances can also be obtained from the ensemble information from MD simulations and hence, provide an avenue for comparison with DEER spectroscopy distance distributions.

We compared the helix distance distribution measurements obtained from our simulations with experimental data (Figure 4.4, Supplementary Figure 4.18) by constructing augmented Markov models (AMMs) [70, 178] and the RotamerConvolveMD Python package [108] that maps a rotamer library of the spin labels on the residues to estimate the distribution (see Methods). The overall distributions were found to be in good agreement with SDSL DEER experiments [111]. For the residues pairs Ser141-Ser432, Ser141-Met438, Ile47-Val330 and Arg201-Glu364 the distance distribution ranges are larger for the simulation data than the experimental observations. This increase in distance distribution shows that the transporter may adopt a wide range of flexibility in order to transport diverse substrate molecules. Further, we compared the experimental data with distance distributions obtained from representative structures of IF, OC, and OF states individually (Supplementary Figure 4.19) as well as other intermediate states (Supplementary Figure 4.20) in our simulations data, to characterize the distance ranges spanned by the predicted free energy minima.

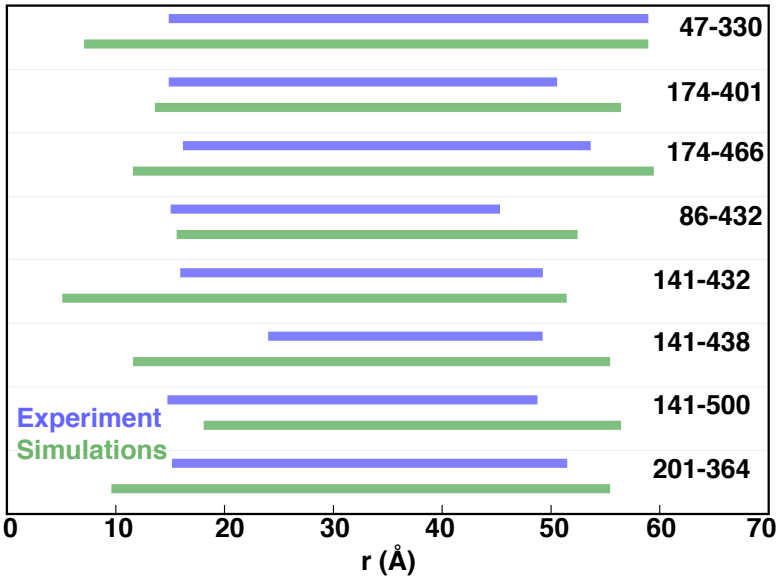


Figure 4.4: MD simulation predicted DEER distance distribution ranges (green) are compared to the experimental DEER distance distribution range (blue).

4.5 Discussion

Our results reveal the conformational changes that characterize various states of $PepT_{So}$ and transitions between them. Our analysis indicates that hydrogen bonds, hydrophobic and aromatic interactions act as gating mechanisms to stabilize the key functional conformation states. The residue-pair Arg32 (TM1)-Asp316 (TM7) forms the salt bridge interaction at the extracellular side and locks the $PepT_{So}$ in the IF state. We show that the formation of OC state involves two steps, i) the helices involving residues Phe150 (TM5) and Met443 (TM11) come closer to ∼5-6 Å, and ii) the following residues from helices TM4 and TM10 form additional hydrogen bond between Ser131-CO (TM4) and Tyr431-OH (TM10). The OF state is obtained by breakage of ionic lock between Arg32 (TM1)-Asp316 (TM7) and movement of TM1 and TM7 to ∼12-16 Å away from each other.

Water molecules around the lipid bilayer and the transporter protein could also play an important role in the conformational change. Supplementary Figure 4.27 shows that water molecules have a higher preference for the phosphate group in the POPC lipid head groups as compared to the nitrogen due to the hydrophobicity posed by 3 methyl groups around it [236]. Lipid bilayers properties in our simulations, area per lipid and membrane thickness (Supplementary Figure 4.28) indicate that the lipid bilayer remains in the same configuration throughout the simulation time. Overall, our predicted values agree well with experimental studies [237] where area per lipid value is 68.3 ± 1.5 $Å^2$ and membrane thickness value of 37 Å although a

slight variation in the computed values compared to experimental observations could be due to differences in temperature and physiological conditions. Moreover, there is usually atleast one water molecule between the lipid molecules that prevents them from coming close to each other (Supplementary Figure 4.29) and hence, stabilizing the lipid bilayer.

Water molecules are co-transported along with substrate molecules and this property has been well studied for membrane transporters [238]. We calculated the hydration level in three states and observed the fluctuation of number of permeating water molecules (Supplementary Figure 4.30). Large number of water molecules were noticed for the OF state compared to OC and IF states. We also compared the water molecules in the IF state with other available crystal structures from the POT transporter family (Supplementary Figure 4.31). The water permeation could also be associated with the transitions among the different conformational states of the protein.

$PepT_{So}$ and other POT family members $PepT_{St}$, $PepT_{So2}$, $PepT_{Xc}$ and GkPOT have conserved sequence motifs and structural folds suggesting that the mechanistic basis of substrate transport will be universal in this family. We posit in the OF state, the binding of a proton and a peptide molecule should increase the structural plasticity of the extracellular side of the transporter and initiate structural rearrangements of helices. The movements of helices are driven through a network of hydrogen bonding interactions involving TM1, TM2, TM7 and TM11 resulting in the OC state. Tyr68 (TM2) and Asn454 (TM11) act as a key residues that drives Asp316 (TM7) to form a salt bridge with Arg32 (TM1). Biophysical studies also show that Tyr68 is critical for affinity and specificity for peptides [205]. The closure of extracellular part results in formation of the OC state and the peptide molecule move into the central cavity. The increase in strength of the salt bridge at the extracellular side results in weakening of the intracellular gating residues. Helices TM4, TM5, TM10 and TM11 increase in structural flexibility and thereby determine the functionally important conformational states to allow substrate transport. The proton and substrate translocates to the conserved ExxERxxxY motif and finally leaves the transporter into the cytoplasm of the cell.

Our study is a first large-scale simulation of a member of the POT family. It is a first extensive analysis of the diversity of the conformational states of the protein and the many rare intermediate states. However, our study is based on equilibrium simulations in the apo state of $PepT_{So}$. The varying dynamics in the presence of the proton and peptide are yet to be demonstrated and understood in great detail. Our study opens new dimensions to obtain a mechanistic understanding into the POT family of proteins and to enable design and transport of peptido-mimetic drugs.

4.6 Supplementary Information

Supplementary Table 4.1: Adaptive sampling rounds for accelerated MD simulations

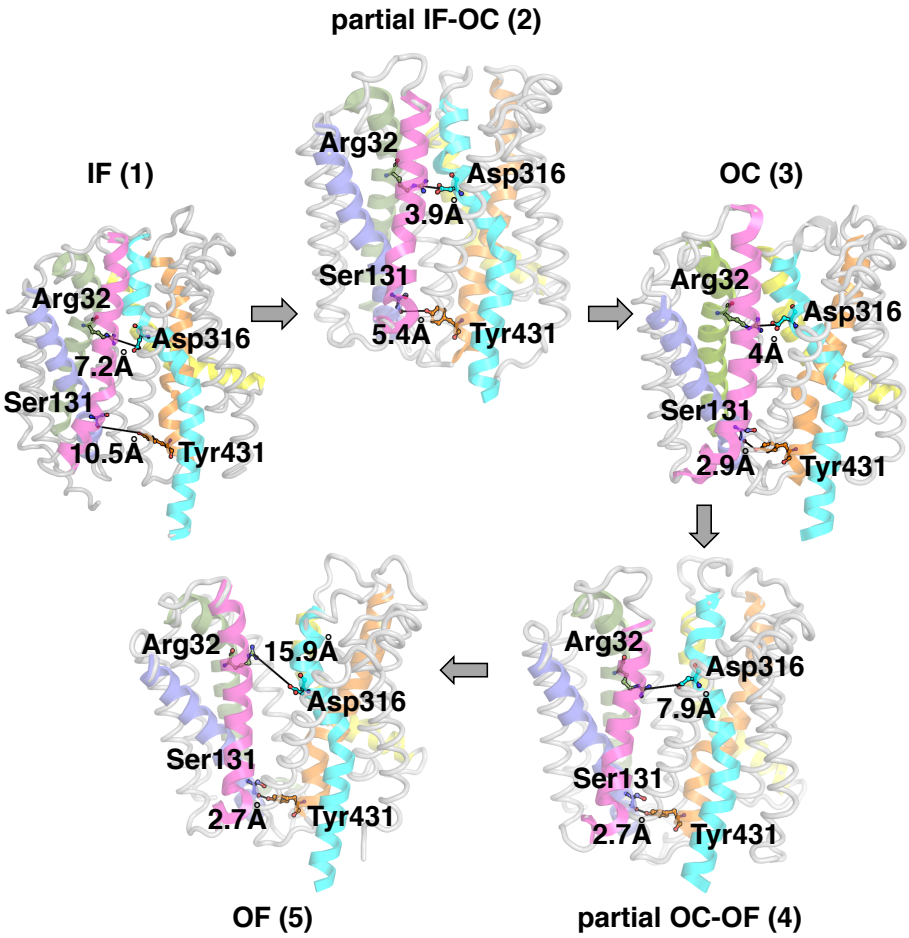
Number of rounds	Time (ns)
Round 1	135
Round 2	135
Round 3	108
Round 4	107
Round 5	97
Round 6	87
Round 7	66
Round 8	54
Round 9	135
Round 10	108
Round 11	81
Round 12	68
Round 13	95
Round 14	169
Round 15	164
Round 16	139
Round 17	90
Round 18	157
Round 19	136
Round 20	145
Round 21	118
Round 22	110
Round 23	2171

Supplementary Table 4.2: Adaptive sampling rounds for classical MD simulations

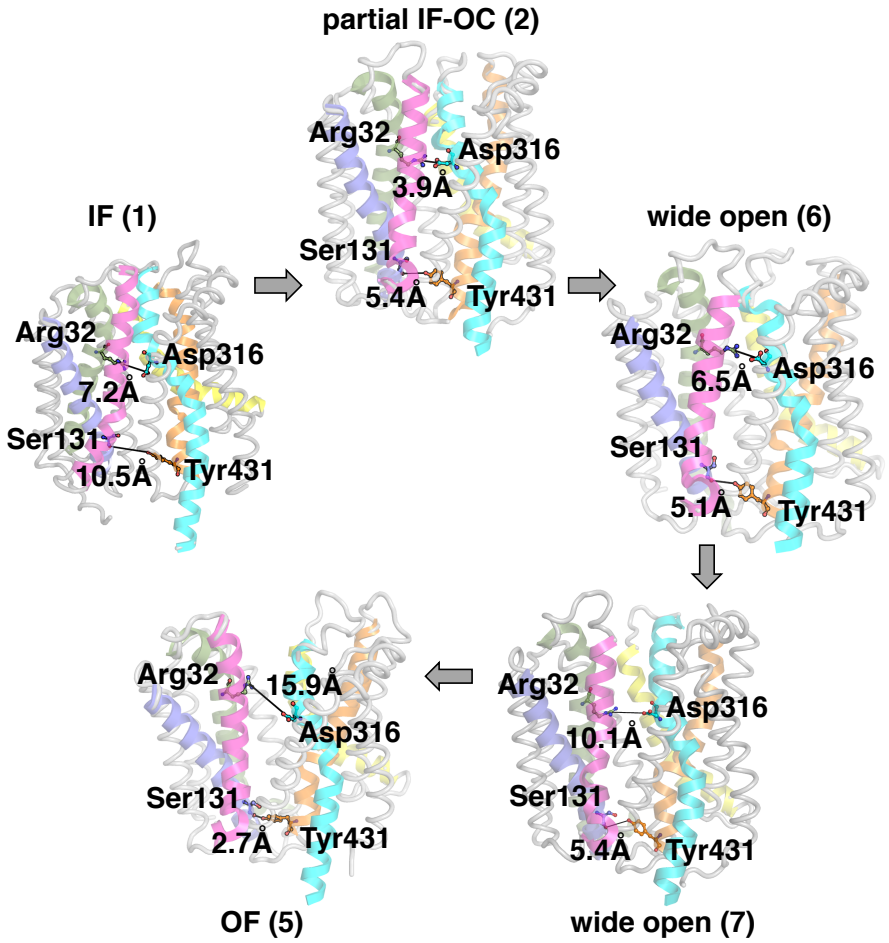
Number of rounds	Time (μs)
Round 1	6.4
Round 2	6.8
Round 3	40.5

Supplementary Table 4.3: Constraints used for augmented Markov models (σ values are fixed at 0.1)

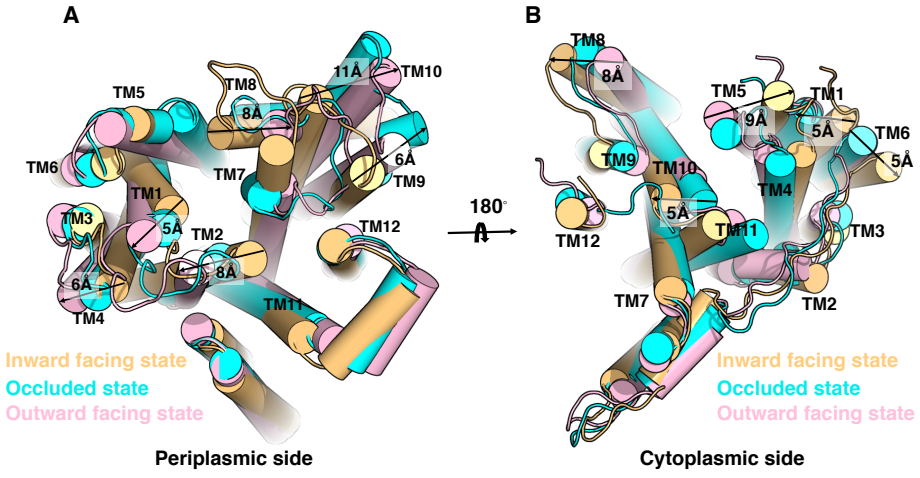
Distance (k)	Residue numbers	m_k (Å)
1	47, 330	39.875078
2	174, 401	41.78273
3	174, 466	37.278798
4	86, 432	41.883655
5	141, 432	36.666667
6	141, 438	28.412256
7	141, 500	43.81818
8	201, 364	44.585633



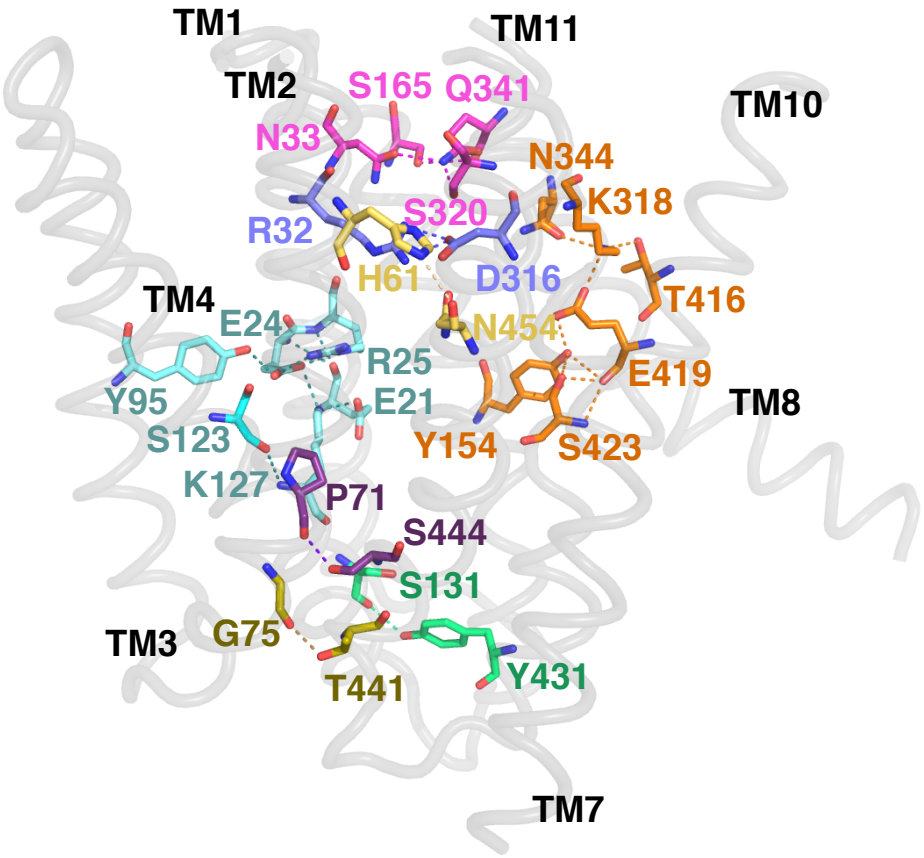
Supplementary Figure 4.1: Most dominant conformational transition of $PepT_{So}$ from IF to OF, via partial IF-OC, OC and partial OC-OF states. The extracellular and intracellular gating residues and their distances in each state have been indicated. The numbers in brackets refer to the energy minima in the free-energy landscape.



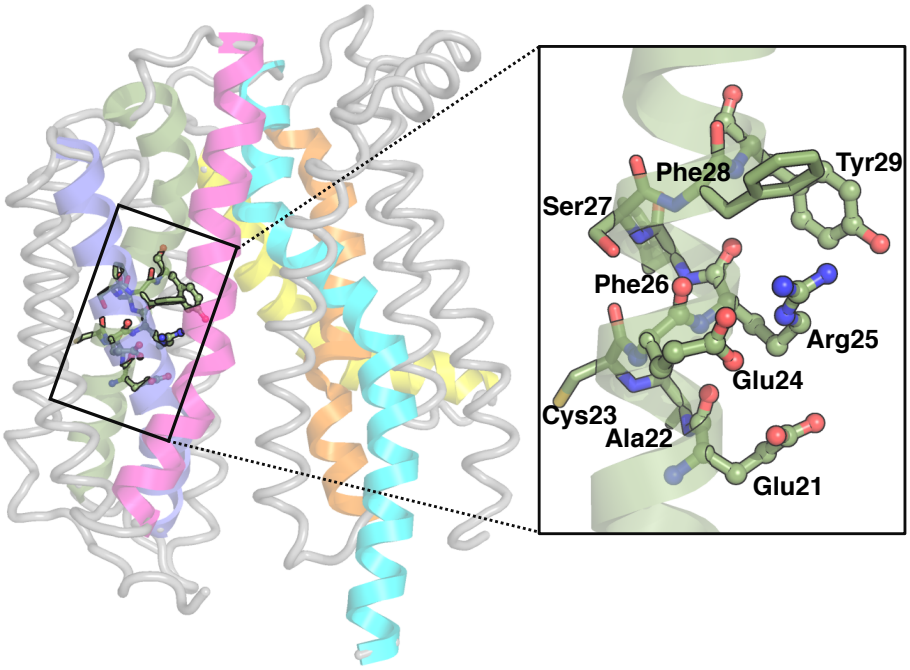
Supplementary Figure 4.2: Another possible conformational transition of PepT$_{So}$ from IF to OF, via partial IF-OC and wide open states. The extracellular and intracellular gating residues and their distances in each state have been indicated. The numbers in brackets refer to the energy minima in the free-energy landscape.



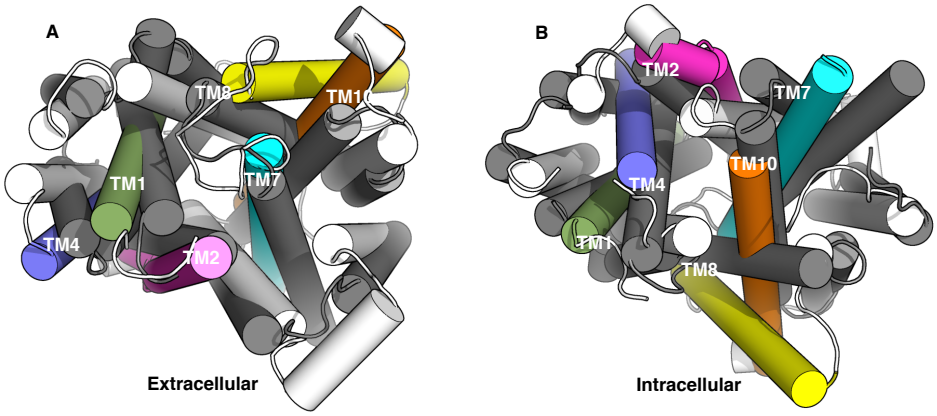
Supplementary Figure 4.3: Conformational transition of helical tips of PepT$_{So}$ IF (yellow), OC (cyan) and OF (magenta) states. Helical tips of three states show drastic changes in TM1, 2, 4, 7, 8, 9 and 10 at both (A) extracellular and (B) intracellular sides.



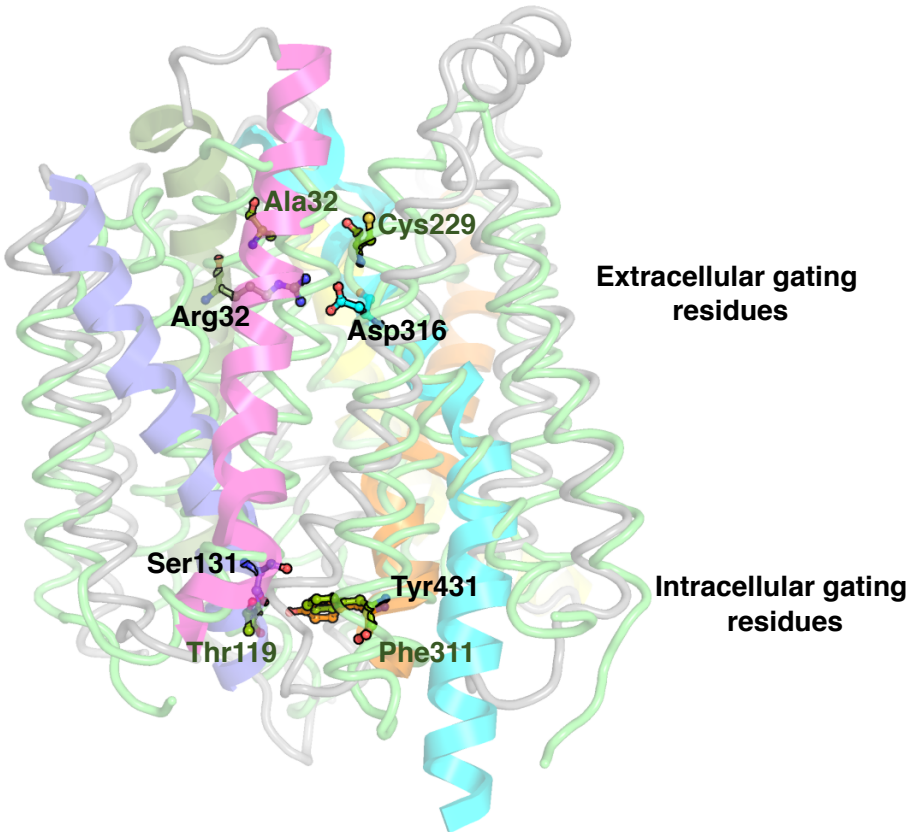
Supplementary Figure 4.4: MD predicted $PepT_{So}$ OC state. Detailed description of residues involved in the interactions between the helices that stabilize the OC state of PepT*So* are shown. Asn33 (TM1), Ser320 (TM7), Gln341 (TM8), Arg32 (TM1), Asp310 (TM7), His61 (TM2) and Asn454 (TM11) form a hydrogen bond network and lock the extracellular side. The conserved residue Glu419 (TM10) form an extensive network of polar interaction and stabilize the conformation of C domain. The ExxERxxxY motif on TM1 forms ionic interaction with Lys127 and neighboring residues. The intracellular side of the transporter is locked by hydrogen bond interaction between Ser131 (TM4)-Tyr431 (TM10). Pro71 (TM2)-Ser444 (TM11) and Gly75 (TM2)-Thr441 (TM11) interactions also favor the conformation of the OC state.



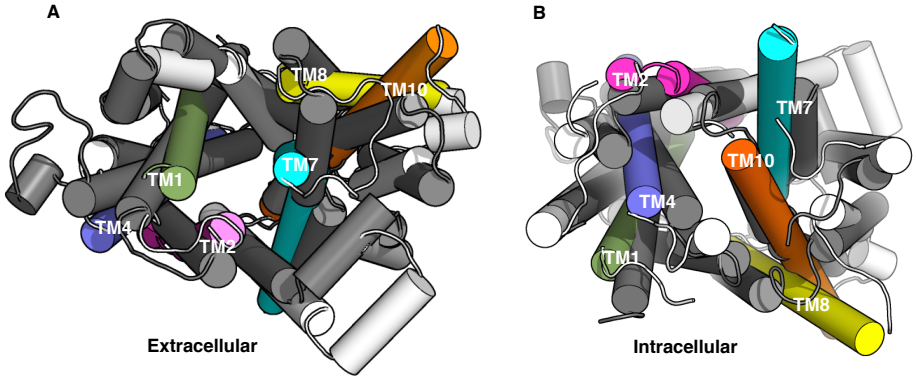
Supplementary Figure 4.5: The motif ExxERxxxY is conserved in the POT family of transporter on TM1.



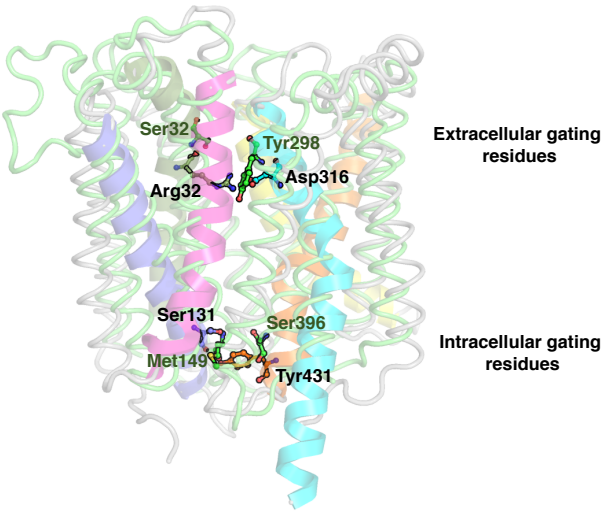
Supplementary Figure 4.6: Comparison of predicted $PepT_{So}$ OC state (colored) with the EmrD OC crystal structure (PDB: 2GFP [230], black) viewed on the (A) extracellular and (B) intracellular side.



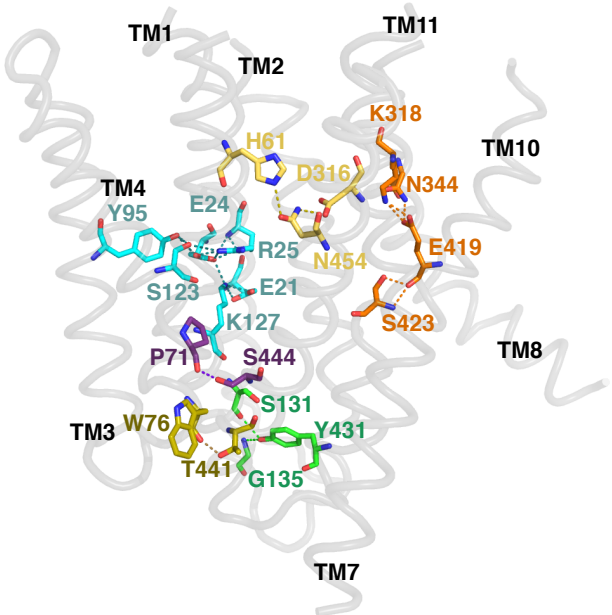
Supplementary Figure 4.7: Gating residues for the predicted $PepT_{So}$ OC state (colored) and EmrD OC crystal structure (PDB: 2GFP [230], green) are indicated.



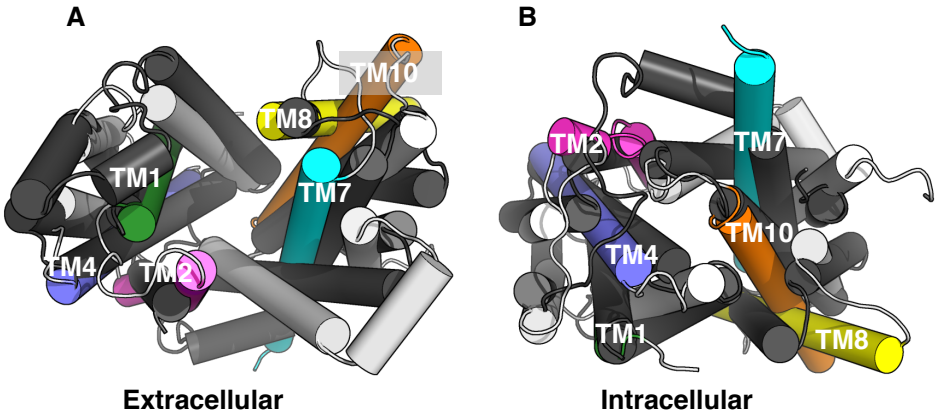
Supplementary Figure 4.8: Comparison of predicted $PepT_{So}$ OC state (colored) with the XylE OC crystal structure (PDB: 4GBY [231], black) viewed on the (A) extracellular and (B) intracellular side.



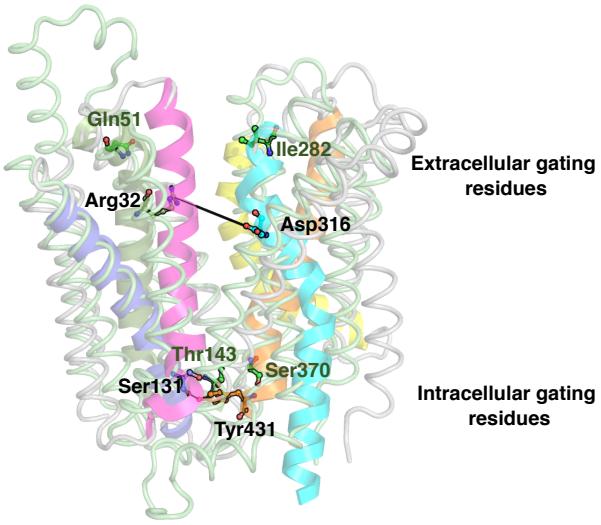
Supplementary Figure 4.9: Gating residues for the predicted $PepT_{So}$ OC state (colored) and XylE OC crystal structure (PDB: 4GBY [231], green) are indicated.



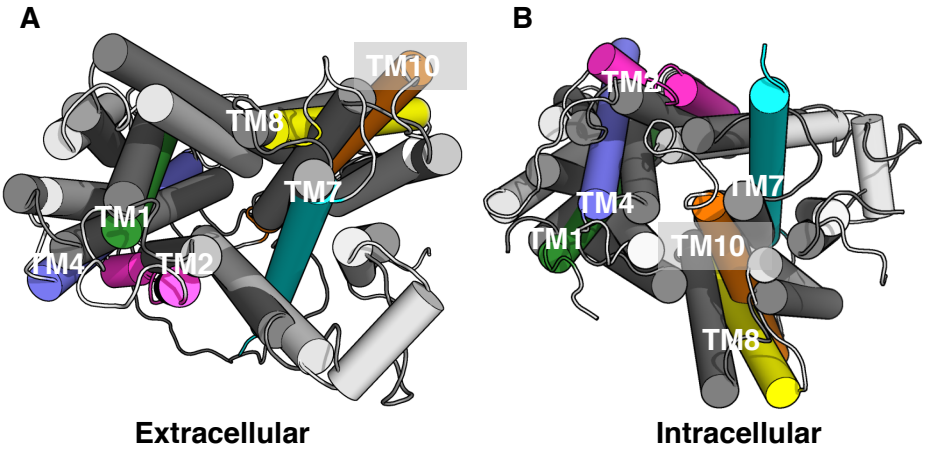
Supplementary Figure 4.10: MD predicted $PepT_{So}$ OF state. Detailed description of residues involved in the interaction between the helices that stabilizes the OF state of $PepT_{So}$ are shown. The residues His61 (TM2), Asn454 (TM11) and Asp316 form a hydrogen bond network and stabilize the extracellular part of OF state. The conserved residues Gln419 (TM10) and ExxERxxxY (TM1) contacts are similar as in the predicted OC state. The residues Ser131 (TM4), Gly134 (TM4), Tyr431 (TM10), Trp76 (TM2) and Thr441 (TM11) form interactions that lock the OF state and close the pore channel on the intracellular side.



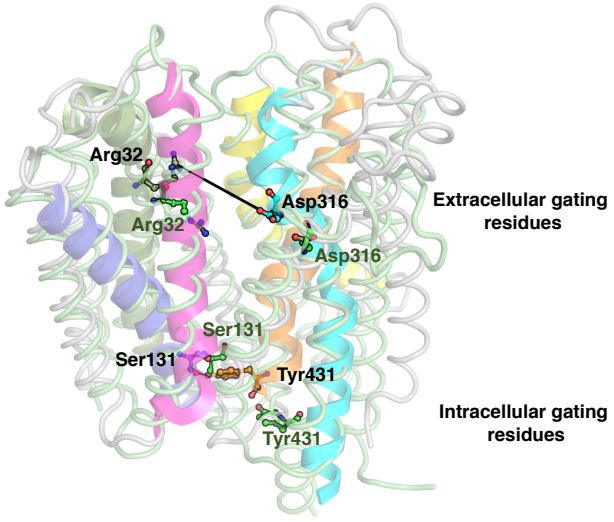
Supplementary Figure 4.11: Comparison of FucP OF (PDB: 3O7P [235], black) with predicted OF structure viewed on the (A) extracellular and (B) intracellular side.



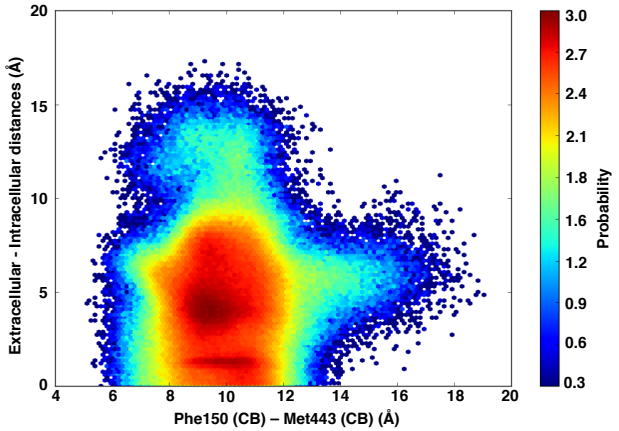
Supplementary Figure 4.12: Gating residues for the predicted $PepT_{So}$ OF state (colored) and FucP OF crystal structure (PDB: 3O7Q [235], green) are indicated.



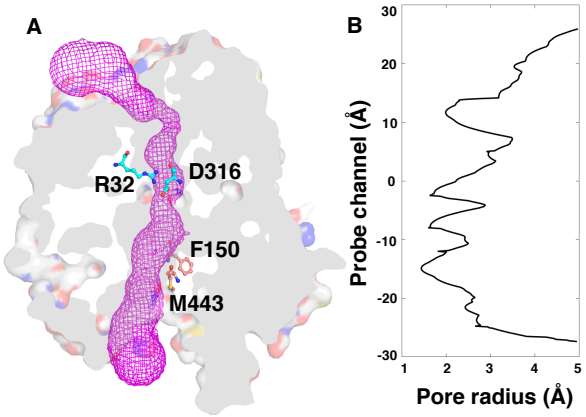
Supplementary Figure 4.13: Comparison of $^{\mathrm{RSM}}\mathrm{PepT}_{\mathrm{So}}$ with OF predicted structure. The RSM modeled $\mathrm{PepT}_{\mathrm{So}}$ [111] (black) OF structure was compared with $\mathrm{PepT}_{\mathrm{So}}$ OF MD predicted structure. Transmembrane helices 1, 2, 4, 7, 8 and 10 align well with RSM modeled $\mathrm{PepT}_{\mathrm{So}}$ structure at both (A) extracellular and (B) intracellular ends.



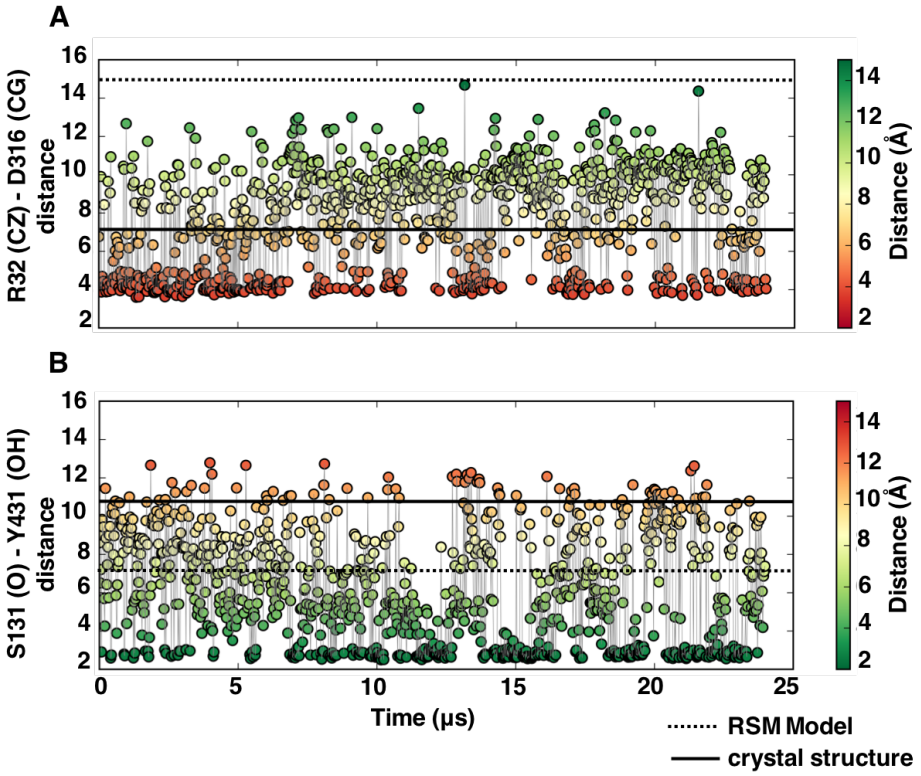
Supplementary Figure 4.14: Gating residues for the predicted $\mathrm{PepT}_{\mathrm{So}}$ OF state (colored) and RSM modeled $\mathrm{PepT}_{\mathrm{So}}$ [111] (green) are indicated.



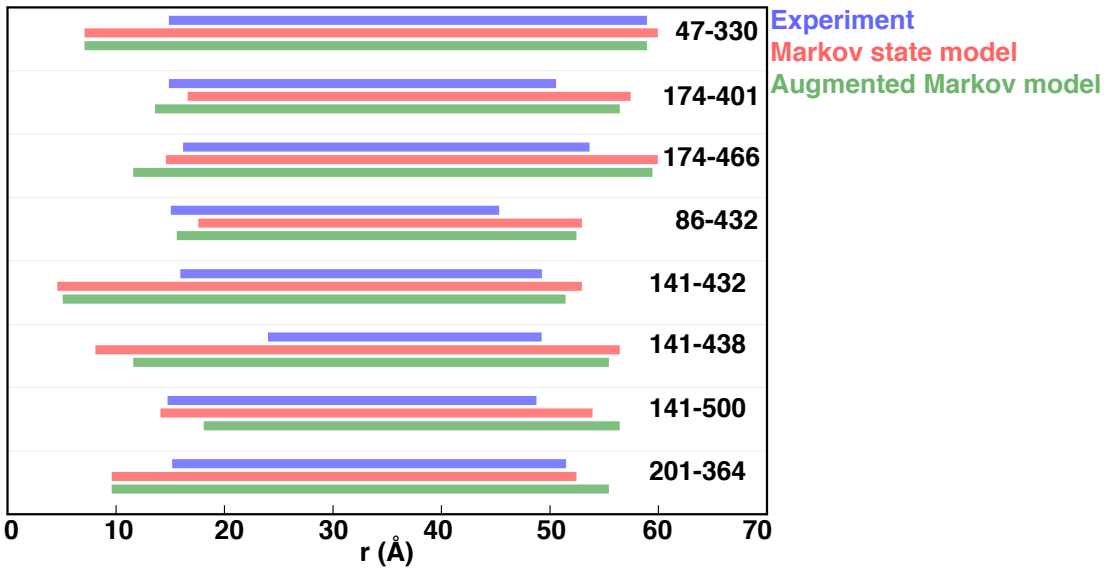
Supplementary Figure 4.15: Raw MD simulation data was projected on the difference between the extracellular and intracellular residue-pairs and the distance between the residue-pairs Phe150-CB (TM5)-Met443-CB (TM11).



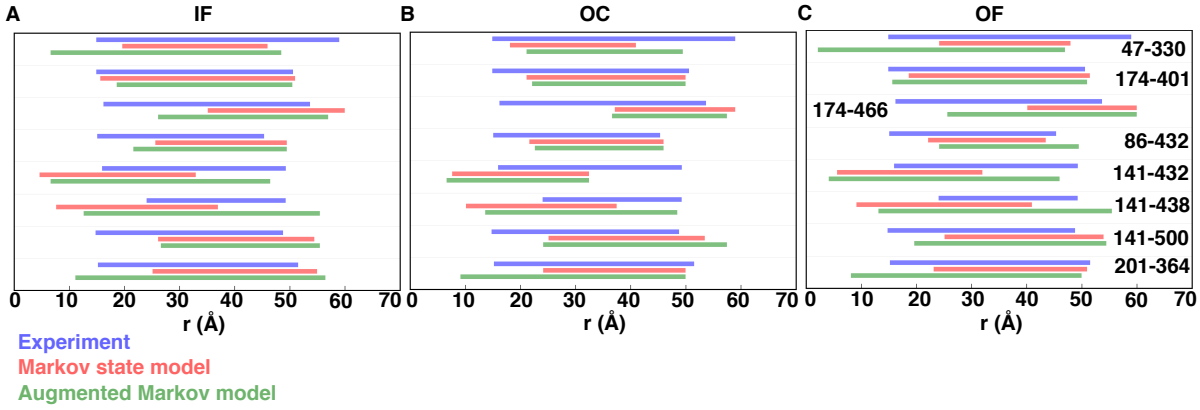
Supplementary Figure 4.16: (A) The residue-pair Arg32-CZ (TM1)-Asp310-CG (TM7) and Phe150-CB (TM5)-Met443-CB (TM11) interactions are indicated, which characterize the partial IF-OC state. (B) The channel pore radius for the partial IF-OC state.



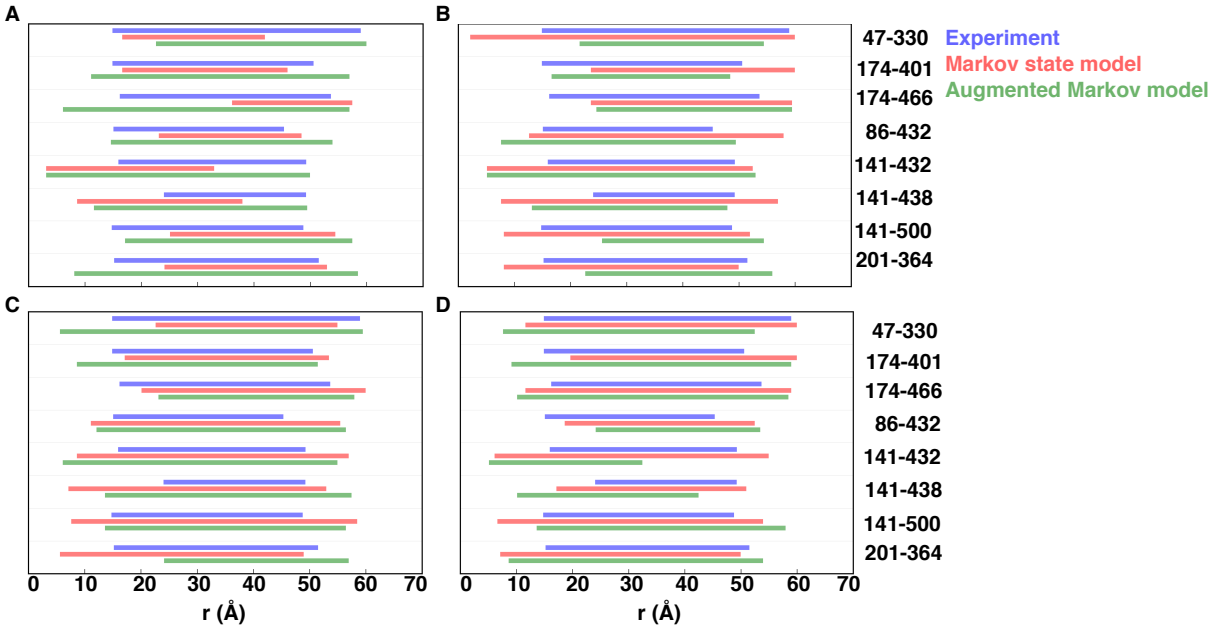
Supplementary Figure 4.17: Kinetics of the conformational changes and their timescales. (*A*) The extracellular and (*B*) intracellular distances are shown as a function of time. The color bar indicates the extent of opening and closing of the gating residues of $PepT_{So}$. The distances in the crystal structure and the RSM model are indicated in black and dotted lines, respectively.



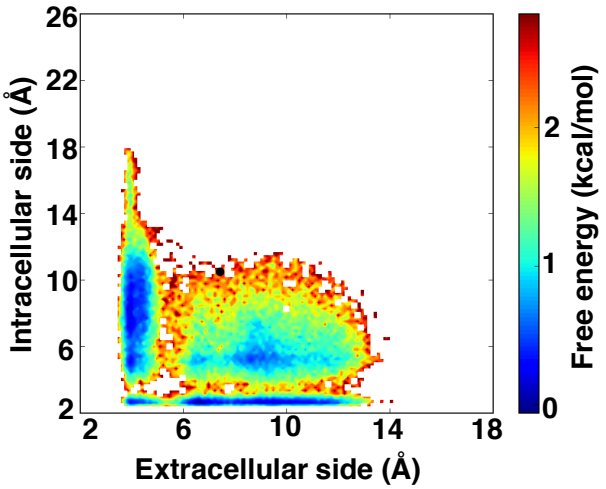
Supplementary Figure 4.18: MD simulation predicted DEER distance distribution ranges (green and red) are compared to the experimental DEER distance distribution range (blue). Red and green simulation predictions are based on Markov state model and augmented Markov model, respectively.



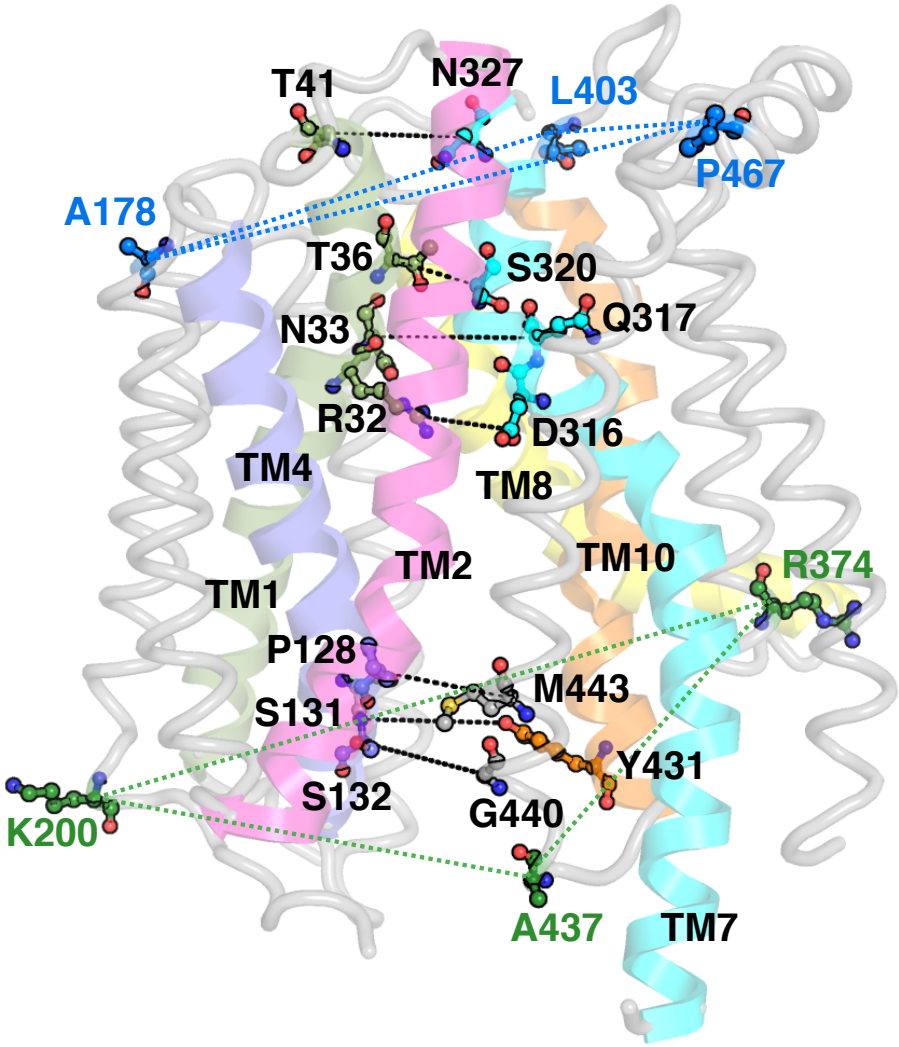
Supplementary Figure 4.19: MD simulation predicted DEER distance distribution ranges (green and red) are compared to the experimental DEER distance distribution range (blue) for the (*A*) IF, (*B*) OC, and (*C*) OF states from the PepT_{So} conformational landscape. Red and green simulation predictions are based on Markov state model and augmented Markov model, respectively.



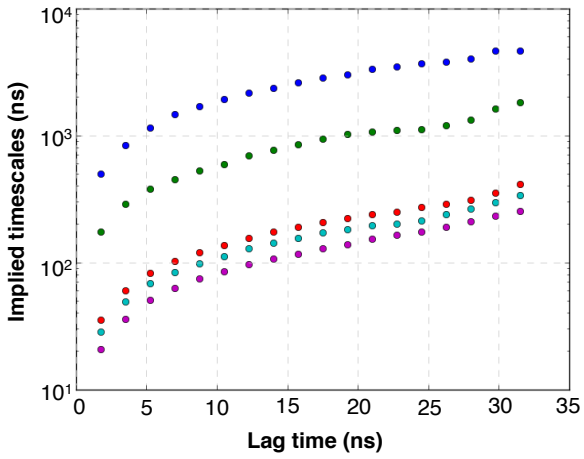
Supplementary Figure 4.20: MD simulation predicted DEER distance distribution ranges (green and red) are compared to the experimental DEER distance distribution range (blue) for the (*A*) partial IF-OC, (*B*) partial OC-OF, and (*C*,*D*) wide-open states from the PepT_{So} conformational landscape. Red and green simulation predictions are based on Markov state model and augmented Markov model, respectively.



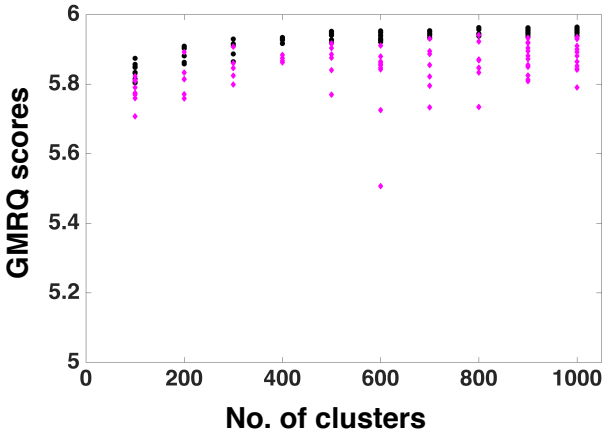
Supplementary Figure 4.21: $PepT_{So}$ two dimensional free energy plot for the raw data obtained using 5 μs accelerated MD simulations. The black dot indicates the crystal structure of $PepT_{So}$ (PDB: 4UVM [111]) in the IF state.



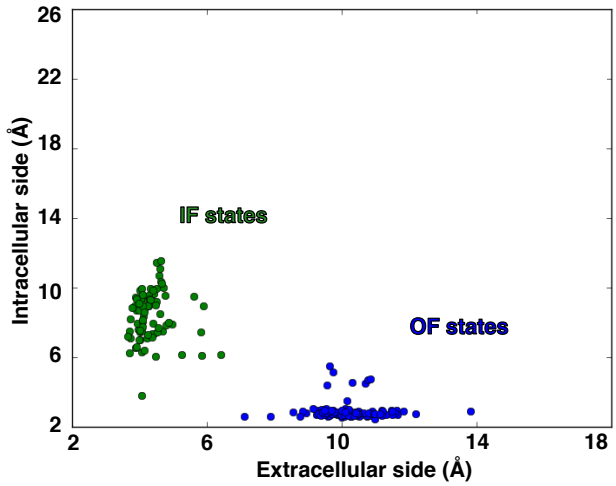
Supplementary Figure 4.22: Extracellular, transmembrane and intracellular residue-pair distances used for MSM construction.



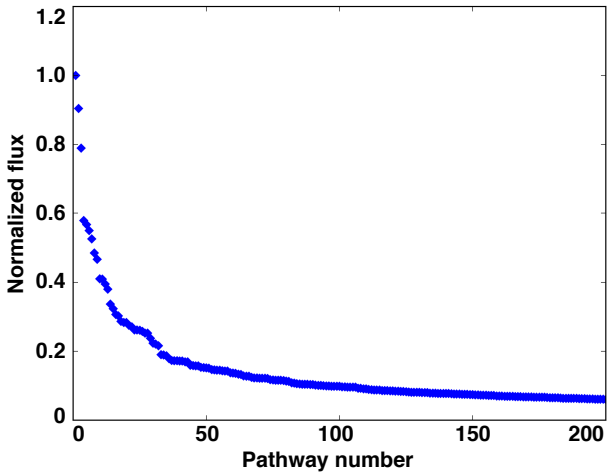
Supplementary Figure 4.23: Implied timescales from transition probability matrix of the MSM. Eigenvalues of the transition probability matrix correspond to the dominant rates of transition in the 200 state model. The top 5 eigenvalues for the MSM are shown here which converged at a lag time of 24 ns.



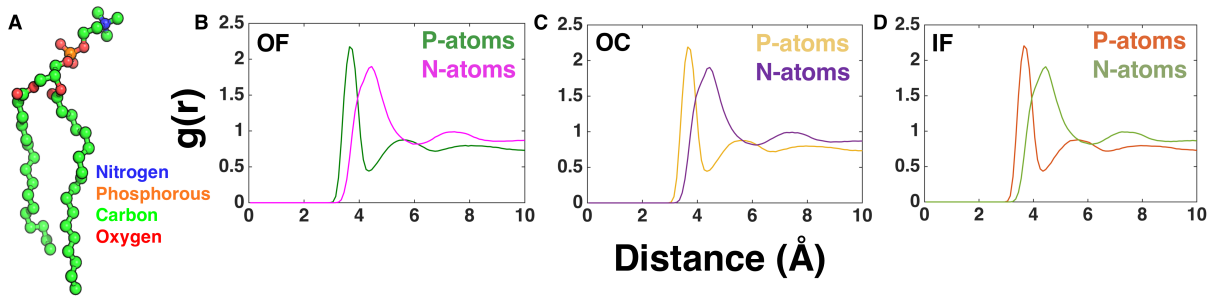
Supplementary Figure 4.24: Comparison of the maximum GMRQ scores of MSM built using variable cluster numbers. 200 clusters yields the highest GMRQ score and hence was used for all MSM construction and analysis. The black and pink dots correspond to the scores for training and testing datasets to calculate GMRQ, respectively.



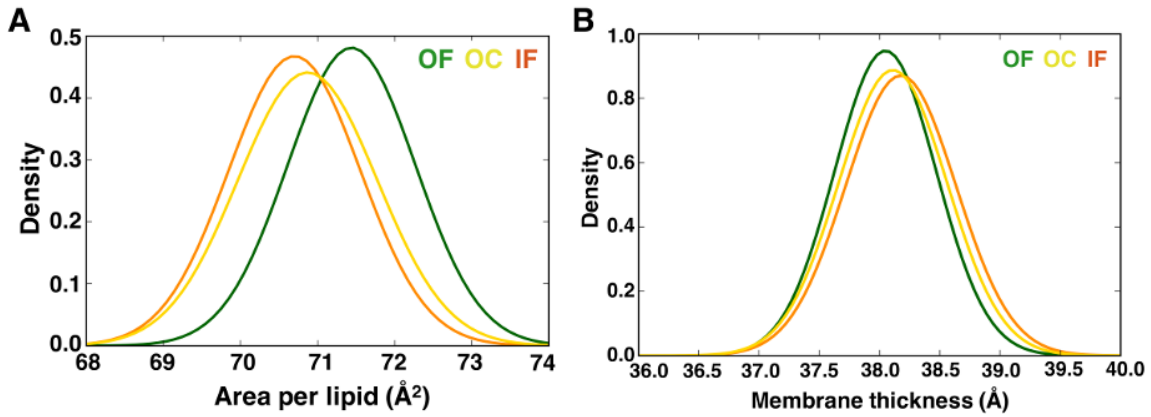
Supplementary Figure 4.25: Sampled conformations from the IF and OF microstates from the MSM.



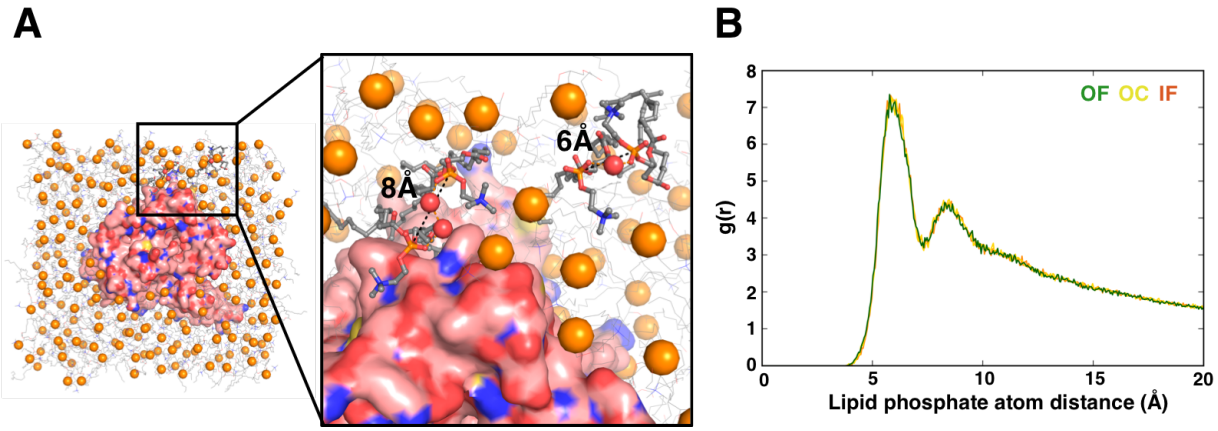
Supplementary Figure 4.26: Normalized flux values for the top 200 reactive paths between IF and OF microstates in the MSM. There are several paths with high flux and large number of pathways with lower flux values.



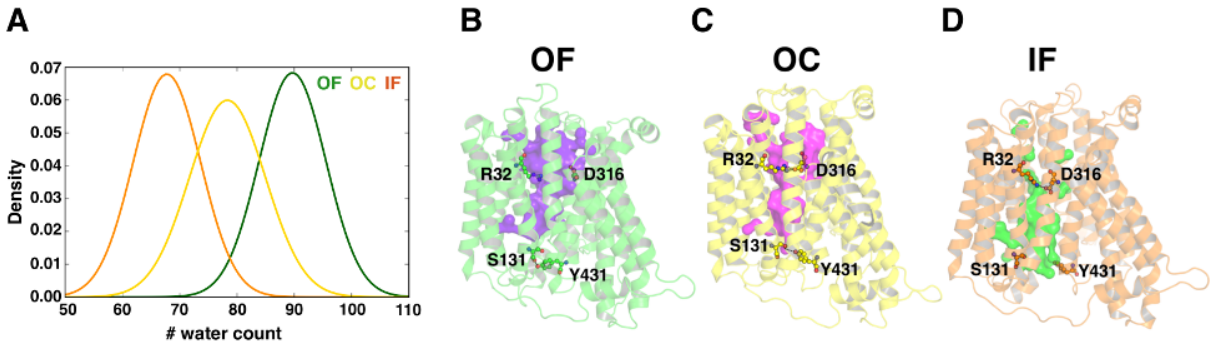
Supplementary Figure 4.27: (A) A single POPC lipid molecule indicating positions of nitrogen (blue), phosphorous (orange), carbon (green), and oxygen (red) atoms. Radial distribution of water around lipid bilayer head group atoms in (B) OF, (C) OC, and (D) IF state. The orientation of water molecules around the phosphate and nitrate groups are calculated using VMD 1.9.2.



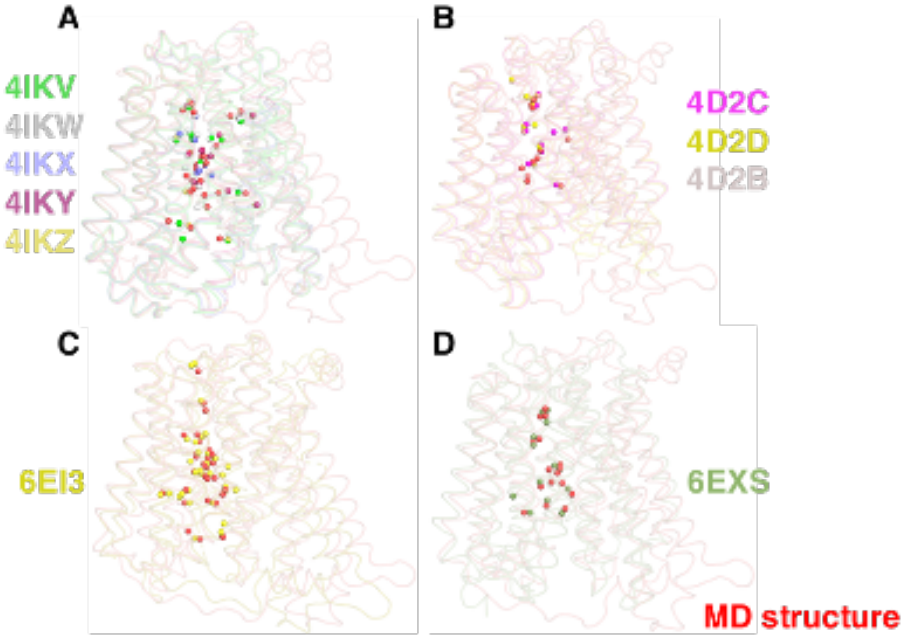
Supplementary Figure 4.28: (A) Probability distribution of area per lipid. To obtain the probability distribution, a normal distribution is fitted to the histogram obtained for 500 structures for each state, OF (μ=71.4;σ=0.83), OC (70.9;0.9) and IF (70.7;0.85). (B) Probability distribution of membrane thickness of the POPC lipid bilayer for OF (green), OC (yellow), and IF (orange) states. A normal distribution is fitted to the histogram obtained for 500 structures for each state, OF (μ=38.05;σ=0.42), OC (38.11;0.45) and IF (38.18;0.46).



Supplementary Figure 4.29: (A) Lipid bilayer molecules in an MD snapshot shows water mediated lipid molecule stabilization. B) The radial distribution plots of phosphate atoms distances in the head group of lipid molecules.



Supplementary Figure 4.30: (A) Probability distribution of water molecules in the protein tunnel. To obtain the probability distribution, a normal distribution is fitted to the histogram obtained for 500 structures for each state, OF (μ=89.7;σ=5.84), OC (78.3;6.65) and IF (67.8;5.87). The water conducting channels are visualized for (B) OF, (C) OC, and (D) IF states.



Supplementary Figure 4.31: Comparison of water molecules inside the MD structure (red) protein pore channel with water molecules in the crystal structures of (A) GkPOT [207], (B) PepT$_{\mathbf{St}}$ [204], (C) PepT$_{\mathbf{Xc}}$ [211] and (D) PepT$_{\mathbf{Sh}}$ [239].

Chapter 5

Reconciling Membrane Protein Simulations with Experimental DEER Spectroscopy Data

5.1 Overview

Spectroscopy experiments are crucial to study membrane proteins for which traditional structure determination methods still prove challenging. Double electron-electron resonance (DEER) spectroscopy experiments provide protein residue-pair distance distributions that are indicative of their conformational heterogeneity. Atomistic molecular dynamics (MD) simulations are another tool that have proved vital to study the structural dynamics of membrane proteins such as to identify inward-open, occluded, and outward-open conformations of transporter membrane proteins, among other partially open/closed states of the protein. Yet, studies have reported that there is no direct consensus between distributional data from DEER experiments and MD simulations, which has challenged validation of structures obtained from long timescale simulations and using simulations to design experiments. Current coping strategies for comparisons rely on heuristics, such as mapping nearest matching peaks between two ensembles or biased simulations. Here we examine the differences in residue-pair distance distributions arising due to choice of membrane around the protein and covalent modification of a pair of residues to nitroxide spin labels in DEER experiments. Through comparing MD simulations of two proteins, $PepT_{So}$ and LeuT - both of which have been characterized using DEER experiments previously - we show that the proteins' dynamics are similar despite the choice of the detergent micelle as a membrane mimetic in DEER experiments. On the other hand, covalently modified residues show slight local differences in their dynamics and a huge divergence when the spin labels' anointed oxygen atom pair distances are measured rather than protein backbone distances. Given the computational expense associated with pairwise MTSSL labeled MD simulations, we examine the use of biased simulations to explore the conformational dynamics of the spin labels only to reveal that such simulations alter the underlying protein dynamics. Our study identifies the main cause for the mismatch between DEER experiments and MD simulations and will accelerate developing potential mitigation strategies to improve simulation observables match with DEER spectroscopy experiments.

5.2 Introduction

Double electron-electron resonance (DEER) spectroscopy has made incredible progress in the study of biomolecules such as cytoplasmic and membrane proteins and nucleic acids [92, 240], including experiments *in vitro* and *in vivo* [241–243]. In DEER experiments, a spin probe is covalently attached to two residues on the biomolecules. Distances between these two spin probes can be determined by measuring the dipolar coupling between an electron pair, one unpaired electron on each of the spin probes. The interaction between electrons is measured in the time domain and then mathematically transformed to distance distributions. Methodological developments have made it possible to obtain distance distributions upto 10 nm in cytoplasmic proteins and 8 nm in membrane proteins [91, 92, 106, 244].

DEER spectroscopy experiments are key for structural insights into membrane proteins for which structure determination methods such as X-ray crystallography and NMR have proved challenging. Given the advance in computational resources, there are numerous extensive MD simulation studies of membrane proteins including GPCRs, transporters, ion channels, integrins, and transmembrane receptor kinases. The observable from DEER experiments, residue pair distance distributions can be directly compared to dynamics informations from molecular dynamics (MD) simulations in order to characterize the structural consequences of the obtained distance distributions. Yet, there is often no direct consensus between distributional data from DEER experiments and MD simulations, which has challenged validation of structures obtained from long timescale simulations. Several methods have been introduced to reconcile experimentally characterized distance distributions with simulations such as restrained ensemble MD (reMD) [100, 245] and ensemble-biased metadynamics (EBMetaD) [246] simulations, both methods employed the experimentally obtained distance distribution to bias a simulation ensemble. Another method to syncretize unbiased MD simulations with experiments is labeling a residue with a spin probe whose conformational orientations are sampled using a spin probe rotamer library [108, 113]. This method is independent from any experiment data bias and relatively computationally inexpensive since no additional simulations are required, but is unable to consider the protein's conformational dynamics.

Typically, we observe mis-matches in terms of relative peak heights when there are multiple peaks in the experiment and unbiased simulated distributions, peak positions, and lower and higher extremes of the distance values. Commonly we observe that experimental distributions exhibit larger distance values, which are not sampled in any of the MD simulation ensemble. These differences can be visualized in Supplementary Figure 5.1A where we compare distance distributions from our previous PepT_{So} simulations with experimental DEER distributional data. Most potential for mismatch between experiments and simulation distance distributions stems from differences in experiment conditions and standard simulations protocols.

Since membrane proteins are embedded in lipid bilayers in physiological conditions, simulations are typically performed in lipid bilayers. These lipid bilayer can be homogeneous or heterogeneous with different types of lipid molecules [247]. On the other hand, biophysical experiments are more amenable in bilayer mimetics. While most of these bilayer mimetics such as nanodisc [248], lipodisq nanoparticles [249], bicelle [250], liposome [251], micelle [111] have been used for DEER spectroscopy studies of membrane proteins, detergent micelles are most commonly used. While there are many different detergents molecules that can form micelles, the most popularly used is n-Dodecyl-β-D-Maltoside (BDDM) detergents.

Another significant basis for mismatch between observed peaks in experiments and simulations is the use of spin probes in DEER experiments, which is absent in wild-type protein simulations. Using site-directed spin labeling (SDSL), two nitroxide spin labels are attached to two cysteine mutated residues. These spin labels can be of different types such as 1-oxyl-2,2,5,5-tetramethyl-pyrroline-3-methyl)methanethiosulfonate (MTSSL), iodoacetamide-PROXYL (IA-PROXYL), unnatural amino acids p-acetyl-l-phenylalanine and 2,2,6,6-tetramethylpiperidine-1-oxyl-4-amino-4-carboxylic acid, and a spin-labeled lysine (SLK-1). DEER experimental measurements among two spin labels are a proxy to explain the protein's residue pair distances. Relying on cysteine modifications and addition of flexible spin probe molecules pose a possibility of modifying the observed protein's dynamics from DEER experiments. For example, MTSSL spin probe has 5 linker dihedrals attributing large rotational flexibility to the protein residue [108]. Recently metal cations such as Gd^{3+}, Cu^{2+} and Mn^{2+} based spin labels which are more rigid have been used [252–254] but their applications in the study of membrane proteins is limited [255].

Based on the above discussed modifications in DEER experiments as compared to physiological conditions, we propose five potential impacts on a protein, it's dynamics and hence the observed DEER experimental observables. Since DEER experiments are typically performed with proteins embedded in bilayer mimetics, such as detergent micelles rather than lipid bilayer, membrane diffusion, packing flexibility and interactions can (1) allow for shifts in DEER distributions and peaks and (2) alter the secondary structure and accessibility of various helices and loops in the protein. Previous studies that draw comparisons between micelle and bilayer environments on membrane proteins have been limited to either small peptides, such as single transmembrane helices or are based on ns-timescale simulations which do not provide a realistic picture of a protein's conformational dynamics. (3) Since DEER measurements require a covalent modification on atleast two sites of the protein, we evaluate whether this modulates the proteins underlying free energy landscape by biasing it to adopt only a subset of the available conformations. (4) Not only can the MTSSL probes have local structural affects on the protein, to what extent can the distance between unpaired electrons on oxygen atom of MTSSL spin probes provide an approximation of the wild-type protein's dynamics. (5) Multiple

flexible bonds of nitroxide spin probes [108] such as MTSSL spin probes may have different timescales than those from the wild-type residue which will equilibrate at a different timescale than the protein changing the experimentally observed dipolar couplings. We evaluate the perturbations and these impacts in this work and discern which among these is the main cause for mismatch between experiments and simulations.

Here, we directly compare the biophysical effect of different experiment and simulation conditions by performing MD simulations in conditions similar to experiments. To evaluate the effect of membrane environment on protein structure and dynamics, we compare long timescale simulations of two proteins in a BDDM micelle and a more typical lipid bilayer. Specifically we perform simulations of two proteins, $PepT_{So}$ and LeuT, which are biologically important representative proteins of two different membrane protein families, Major Facilitator Superfamily (MFS) and Neurotransmitter: Sodium Symporter (NSS), respectively. Residue pairs in both protein have been previously characterized using DEER experiments [106, 111, 256, 257]. LeuT has many three-dimensional structures determined through X-ray crystallography and has been investigated using computational simulations. Recently, two crystal structures of $PepT_{So}$ were resolved [111, 258] and we have examined this protein using long timescale MD simulations in our previous work in Chapter 4. We follow our micelle and bilayer simulations by introducing nitroxide spin labels MTSSL on a pair of residues in $PepT_{So}$ in order to examine the perturbations caused by the probe's site specific mutations during DEER spectroscopy experiments. We then perform restrained ensemble molecular dynamics (reMD) simulations on order to evaluate the spin pair equilibration and it's impact on the protein's conformational landscape and residue pair distance distributions.

5.3 Methods

MD simulations

All simulations were setup up using CHARMM-GUI [259–263], built with a rectangular box and a minimum water height of 10 Å above and below the membrane. System specific details are provided below.

All simulations were run using NAMD 2.13 MD package [13] and the CHARMM 36 force field [264–266] on the Blue Waters petascale computing facility. We used the NAMD inputs provided by CHARMM-GUI for minimization and equilibration in six consecutive steps followed by production runs in the NPT ensemble where constant temperature was maintained by employing Langevin dynamics with a damping coefficient of 1 ps^{-1}. The Langevin piston method was employed to maintain a constant pressure of 1.0 atm with a piston period of 50 fs. Nonbonded interactions were smoothly switched off at 810 Å and long-range electrostatic interactions were calculated using the particle mesh Ewald (PME) method. For all simulation steps, bond

distances involving hydrogen atoms were fixed using the SHAKE algorithm.

Minimization was done for 10,000 steps, total equilibration and production run time for individual simulations are noted in Supplementary Table 5.1. Production simulations were run at 303.15 K, trajectory parameters were determined every 2 fs, and coordinates were saved every 100 ps.

All trajectory analysis was done using MDTraj 1.7 [129] except where otherwise noted. Analysis methods and workflows are explained in the Supplementary Information.

Determining a micelle size for membrane protein simulations

Previous work on simulating protein-micelle complexes [267] posit the use of number of detergents more than the aggregation number of a detergent-only micelle which is 135-145 for the n-Dodecyl-β-D-Maltoside (BDDM) detergent [268, 269]. Moreover, the BDDM micelle size was determined to be 72 kDa [270]; with a 510.621 g/mol molecular weight of a BDDM molecule this yields $\sim$141 detergent molecules in the micelle. To test the stability of the protein-micelle complexes, we took a single structure of the PepT_{So} protein from our previous simulations in Chapter 4 and embedded it in 150, 180 and 200 BDDM detergent molecules. The three simulation setups with 145668, 145847, and 146061 atoms respectively comprised of protein, detergents, waters, and 0.15 M KCl ions. Simulation seetup with 150 detergent molecules was minimized for 10,000 steps and the other two were minimized for 20,000 steps. We ran each of these simulations for 60 ns each post-equilibration and only last 50 ns were used for analysis (Supplementary Table 5.2).

Our goal was to identify a suitable micelle size for protein-micelle complex simulations and use this size for all simulations in this work. We performed short simulations of PepT_{So} protein in three difference micelle sizes, with 150, 180, and 200 BDDM detergent molecules and assessed the protein's structure and dynamics in all three micelle sizes. RMSD of the protein converges to values between 0.28-0.32 nm within 50 ns, and these values are lower when only the transmembrane region of the protein is included (Supplementary Figure 5.2A). We do not expect sampling any conformational change in the protein's structure in such short trajectories.

We then evaluated the extent of sphericity of the micelle, measured by calculating it's eccentricity where a perfectly spherical object has eccentricity 0. We find that in all three cases, micelles in our simulations are mostly-spherical with an average eccentricity of 0.23 ($\pm$0.02) for micelle with 150 and 200 detergents and 0.22 ($\pm$0.02) for micelle formed by 180 detergent molecules (Supplementary Figure 5.2C). Radius of the micelles do not show much variation, indicating that the micelles do not distort (Supplementary Figure 5.2D). As expected, micelle with more detergents have a larger average radius - 4.4 nm, 4.58 nm, and 4.68 nm for 150, 180, and 200 detergent micelles, respectively. Supplementary Figure 5.2E shows a radial distribution of

distances between BDDM detergent molecule headgroups. Since the distribution is same for all three micelle sizes, we conclude that detergent packing is similar in all three micelles.

Our preliminary simulations indicated that protein dynamics and shape of detergent micelle do not vary with number of detergents in the micelle and we chose 150 detergent micelle for rest of our simulations to keep the system sizes smaller and conserve computational resources. We were also able to confirm that the simulation setup is stable for all three micelle size choices.

Setting up LeuT simulations in bilayer and detergent micelle

We compiled 28 LeuT crystal structures of which all but 2 have no mutations in the proteins sequence. Structures 3TT1, an Outward Facing (OF) structure with 2 mutations, and 3TT3, an Inward Facing (IF) structure with 4 mutations [271], were modeled on the wild-type LeuT sequence using Modeller [272] interface in Chimera [273]. Based on these PDBs, we identified 36 unique structural models for the LeuT protein from residue Arg5 to Ala513 (509 residues in all). Most of these 36 structures were missing residues either on EL2, EL3 or EL6, the size of the largest missing region in any structure is 6 residues. These missing regions were modelled to yield 72 LeuT structures as a starting point for our simulations. These structures were aligned in VMD [172] using orient and a linear algebra Tcl package, La.

During setup in CHARMM-GUI, LeuT protein was capped with ACE and CT3 residues. For protein-bilayer complexes, the structures were embedded in a POPE bilayer of 150 lipid molecules equally distributed in the upper and lower leaflet using the Insertion method. For protein-micelle complexes, protein structures were embedded in 150 BDMM detergent molecules. We only added 3 Cl^- ions to neutralize the system. Since we are only interested in the equilibrium conformational changes of *apo* protein, we did not want to introduce Na^+ ions which are known to play an important role in the transport mechanism of LeuT. Ion binding and substrate transport are coupled and ions can be considered as substrate. Details for system size and simulation time are provided in Supplementary Table 5.1.

Setting up PepT$_{So}$ simulations in bilayer and detergent micelle

We used 42 structures extracted from our previous simulations of PepT$_{So}$ in Chapter 4 from residue Pro8 to Tyr512 (505 residues in all). During setup in CHARMM-GUI, the protein was capped with ACP and CT3 residues. For protein-bilayer complexes, the structures were embedded in a heterogeneous POPE/POPG (3:1 ratio) bilayer of 200 lipid molecules equally distributed in the upper and lower leaflet using the Replacement method. For protein-micelle complexes, protein structures were embedded in 150 BDMM detergent molecules. We added 0.15 M NaCl ions in addition to neutralizing the system. Details for system size and

simulation time are provided in Supplementary Table 5.1.

Setting up $PepT_{So}$ Simulations with MTSSL probes on residue-pair

The same method as for $PepT_{So}$ in detergent micelle was followed, and residues Asn174 and Ser466 were mutated to MTSSL (1-oxyl-2,2,5,5-tetramethylpyrroline-3-methyl methanethiosulfonate) [274] probes. This corresponds to one of the extracellular residue pairs chosen by Fowler et al. for DEER experiments [111]. Here, we have used WYF parameter for cation pi interactions as available in CHARMM-GUI.

For all simulations described above we examine convergence by randomly sampling 25%, 50%, and 75% of the trajectories and graphing experimental residue pair distance distributions shown in Supplementary Figure 5.3,Supplementary Figure 5.4,Supplementary Figure 5.5,Supplementary Figure 5.6,and Supplementary Figure 5.7. We choose to look at these residue pair distance distributions, as a check for convergence, because these will be the main focus of most results in this work. We see that for all systems, error bars are small even with 25% simulation data and they decrease as we include more data. We can conclude that multiple trajectories sample each region of the conformational ensemble and no single trajectory shifts the distance distributions completely.

Restrained-ensemble molecular dynamics (reMD) simulations for $PepT_{So}$

We used 42 different protein conformations as starting point for reMD simulations in vacuum which means the protein was not surrounded by lipids, water or ions. CHARMM-GUI default value of 25 spin labels copies were attached to each labeled protein residue. Experimental distance distributions from Fowler et al. were provided as target histograms [111]. We also used default values for force constants, bin widths and Gaussian natural spread [263]. Simulations were run using a special version of NAMD 2 [13, 245].

For system reMD (1 dist), we attached MTSSL probes on residues Asn174 and Ser466, and restrained this distance. For system reMD (2 dist), MTSSL probes were places on four residue and two distances were restrained, Asn174-Ser466 and Arg201-Glu364. These residue pairs are on opposite side of the protein. For system reMD (8 dist), MTSSL probes were places on 12 residues, and 8 experimentally studied residue pairs were restrained. Details for system size and simulation time are provided in Supplementary Table 5.3. We used an integration timestep of 1 fs and saved trajectory coordinates at a frequency of 50 ps.

Since reMD simulations are biased simulations, where the distance between the probe molecules is restrained to a targeted distribution using harmonic forces we use these simulations as an opportunity to explore the protein and the MTSSL probe dynamics when the experiment and experiment distance distributions show a perfect match. This is also why reMD simulations in vacuum are efficient and sufficient to sample the spin

probe dynamics.

5.4 Results

Residue pair distances from micelle-embedded proteins resemble trends in bilayer-embedded proteins

$PepT_{So}$ is a proton-coupled bacterial symporter for which, recently, researchers characterized eight inter-residue distance distributions using DEER [111]. There are two known crystal structures for this protein found in the bacteria *Shewanella oneidensis*, 2XUT [258] and 4UVM [111], both in the inward-facing conformation of the protein. $PepT_{So}$ is a promiscuous transporter for most di/tri-peptides, but it belongs to the Proton-dependent oligopeptide transporter (POT) family and the Major Facilitator Superfamily (MFS) whose members have a wide variety of functions and are found in many different organisms including humans. According to the OPM database, there are crystal structures available for 24 proteins from the POT family and 65 from MFS, which is also one of the largest superfamilies of membrane proteins. Like most MFS family transporters, $PepT_{So}$ has 14 transmembrane helixes, divided into N terminal bundle (TMs 1 to 6), C terminal bundle (TMs 7 to 12), and two helixes A and B between the two domains packed closely with the C terminal bundle.

LeuT, a leucine transporter, has many high-resolution crystal structures and been extensively characterized using DEER experiments [106, 256, 257]. LeuT belongs to the Neurotransmitter: Sodium Symporter (NSS) family whose other members include Dopamine, noradrenaline, GABA, glycine, and serotonin transporters. LeuT was the first structure resolved using X-ray crystallography from the NSS family and of the many structures resolved since there, only one structure is in the inward-facing state as a quadruple mutant (3TT1 [271]). The LeuT-fold consists of 12 transmembrane helixes, of which TMs 1 to 5 and TMs 6 to 10 are inverted repeats of each other.

$PepT_{So}$ and LeuT, both are important model proteins from two different families of membrane proteins and yet, LeuT has been well studied using computational simulations with both unbiased and biased protocols, whereas there are only a few short-timescale computational studies focused on $PepT_{So}$. Our previous work in Chapter 4 sampled the conformational dynamics of $PepT_{So}$ using long timescale 54 μs MD simulations and analyzed it's equilibrium dynamics using Markov state model based analysis. These simulations were carried out in a POPC bilayer using the AMBER FF14SB force field. In order to compare dynamics of $PepT_{So}$ protein in detergent micelle and bilayer and capture solely the effect of the membrane environment, we replicated our simulations in a POPE/POPG bilayer in CHARMM 36 force field. Simulations from our

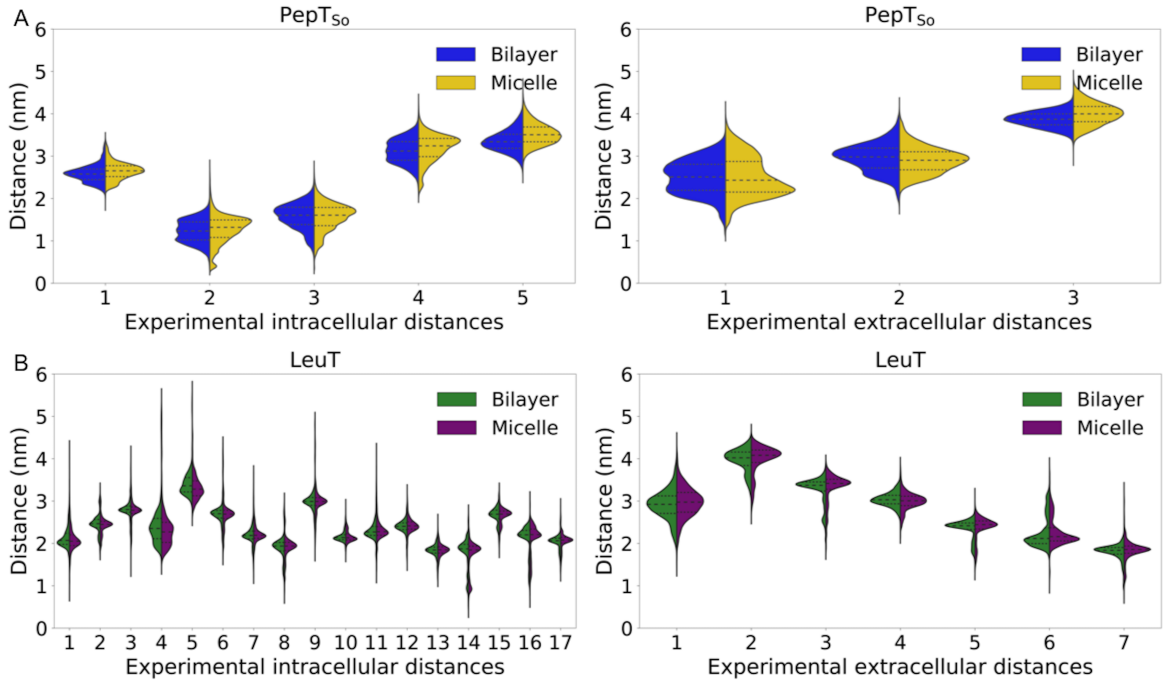


Figure 5.1: (A) Violin plot shows distance distributions for 5 intracellular residue pair distances and 3 extracellular residue pair distances measured by Fowler et al. as observed from MD simulations of $PepT_{So}$ protein in micelle (yellow, right) and bilayer (blue, left) [111]. (B) Violin plot shows distance distributions for 17 intracellular residue pair distances and 7 extracellular residue pair distances measured by Kazmier et al. as observed from MD simulations of LeuT protein in micelle (purple, right) and bilayer (green, left) [257]. Distance distributions among C_α atoms and sidechain atoms is shown in Supplementary Figure 5.21 and Supplementary Figure 5.22.

previous work provide a benchmark for sufficient conformational sampling since we were able to sample IF, OC, OF and multiple other intermediate states (Supplementary Figure 5.8). Here, we compare our atomistic molecular dynamics simulations of $PepT_{So}$ and LeuT in BDDM micelles and in lipid bilayers.

In Supplementary Figure 5.8A,B, we project our $PepT_{So}$ simulation datasets on gating residue pairs, Ser131-Tyr431 on the intracellular side and Arg32-Asp316 on the extracellular distance. We compare the sampled regions with our previous simulations (Supplementary Figure 5.8D) to conclude that all physical conformations of the protein have been well sampled. Similarly, in Supplementary Figure 5.9 we project our LeuT simulation datasets on one residue pair each on the intracellular and extracellular side of the protein, Arg5-Asp369 and Arg30-Asp404, respectively. These residue pairs are based on gating residues identified in hSERT [275] which also has a typical LeuT-fold and shares 35.5% sequence similarity with LeuT protein. Of the four residues involved in the gating residues, Asp404 from LeuT is homologous to Glu493 in hSERT and the three others are arginines.

Upon comparing simulated and experimental distance distributions from our micelle and bilayer simulations (Supplementary Figure 5.1B,C) we clearly see that distance distributions obtained from micelle simulations

are no better at matching with experiments. However, by comparing distance distributions as determined from our MD simulations for micelle and bilayer environment simulations for both proteins in Figure 5.1 we examine the impact of the choice of membrane on the protein's dynamics, residue distances and secondary structure. For $PepT_{So}$ protein, five out of a total of eight distance distributions show a higher median value (middle horizontal line on violin plots in Figure 5.1A) in micelles as compared to bilayer. For instance, residue pair 174-466 shows a single peak in the distributions for both micelle and bilayer, but the data has a median value of 3.87 nm mean in bilayer whereas this value is 4.00 nm in micelle. On the other hand, two distances distributions for residue pairs 47-330 and 174-401 show lower median values in the micelle than in the bilayer. One distance distribution for residue pair 141-438 show about the same value 1.6 nm in both micelle and bilayer. In Supplementary Figure 5.10, we show that most of the mean or median values lie along the black dotted line, indicating that they are similar for micelle and bilayer and the differences are minimal. Mean and median values for all inter-helix distances also fall along the dotted line.

Based on visual inspection, not only the positions of the peaks but the number of peaks in the distance distributions can differ such as 3 peaks in bilayer versus 2 in micelle for residue pairs 141-432 and 141-500. Interestingly, these two distance distributions show new peaks in micelle where little or no data is seen in those regions in bilayer. For LeuT, although the distance distributions differ, the variation is much less (Figure 5.1B and Supplementary Figure 5.10), for example none of the 24 experimental distance distributions show a peak in bilayer which is not there in micelle or vice versa.

For $PepT_{So}$, five of the experimental residue-pair distance distributions also show slightly broader distributions. For inter-helix distributions (Supplementary Figure 5.10) we see that few upper values and lower values lie below the dotted line meaning that the distributions move towards larger values in micelles. Does this mean that micellar environments shift the distributions to larger values? This is unlikely because for LeuT we see values that are both above and below the black dotted line in experimental distances as well as inter-helix distances.

In order to conclude that the reason for mismatch between DEER experimental observables and MD simulations distance distributions is due to the use of detergent micelle in MD simulations, distance distributions should have exhibited a consistent behavior such as micellar distributions are always larger, smaller, wider or narrower. However, in our simulations there are no dramatic or homogeneous shifts in the distance distributions.

Proteins in micelles and bilayers show structural similarity

For both proteins, we measure helicity of transmembrane helices and Supplementary Figure 5.11 shows the distribution of helicity values. Values closer to 1 indicate helical nature and decreasing values show loss of helicity. TMs 7 and 10 exhibit a wider range of helicity in $PepT_{So}$ which indicates their dynamic nature. In Chapter 4 we report that one of the extracellular gating residue is on TM7 and one of the intracellular gating residues is on TM10. Given that the median of TM7 helicity is 0.76 in both micelle and bilayer, lowest among all other transmembrane helices, none of the helices lose their entire alpha-helical nature. Moreover broader distributions for TMs 7 and 10 are seen in both micellar and bilayer environments.

In LeuT TMs 1 and 6 show wide helicity ranges, this is the case in both environments. Readers must note that TM1 here indicates residues of TM1a, the first half of TM1 helix. TM1a is of particular interest in LeuT and other NSS family transporters [271, 276, 277] because in IF structures this region is away from the bundle as shown in Supplementary Figure 5.12C. In Supplementary Figure 5.11B these low values of TM1 helicity arise from IF trajectories and other trajectories that transition to IF like states. Our simulation ensemble includes two independent trajectories based on the IF structure 3TT3 [271]. A recent study by Gotryd et al. reports a LeuT IF structure but the structure is unreleased as of the writing of this work [277]. LeuT TM1a dynamics show a significant distinction in OF and IF states in our MD simulation trajectories (Supplementary Figure 5.12), TM1a helicity drops to 20-30% in IF trajectory whereas this is 50-80% in OF trajectories. Due to the dynamic nature of this region, it follows that one of the gating residues on both intracellular and extracellular side of the protein are also positioned on TM1. This distinct behavior of TM1a is also seen in Supplementary Figure 5.9A,B where LeuT is open on both extracellular and intracellular sides. Other studies on transporter proteins using extensive MD simulations [275, 278, 279] have also reported observing this hourglass-like state of the transporter. Terry et al. have reported evidence for this conformation in LeuT which is due to a weaker coupling between extracellular and intracellular side of LeuT [280]. We suggest that this weaker coupling allows LeuT to explore a large range of intracellular gating distance while the extracellular side of the protein is also open.

TM regions of $PepT_{So}$ and LeuT show structural similarity in both micelle and bilayer, but could the choice of the membrane milieu affect the intracellular and extracellular flanking regions of our proteins? We compare the distributions for these regions such as helicity of two short helices in $PepT_{So}$ one on each side. For LeuT we compare the helical content of the loop regions which connect the TM helices. While their might be molecular level differences in protein residue interactions with lipids in bilayer or detergents in micelle, Supplementary Figure 5.13 and Supplementary Figure 5.14 show that distance distributions are similar and not impacted by the choice of membrane environment.

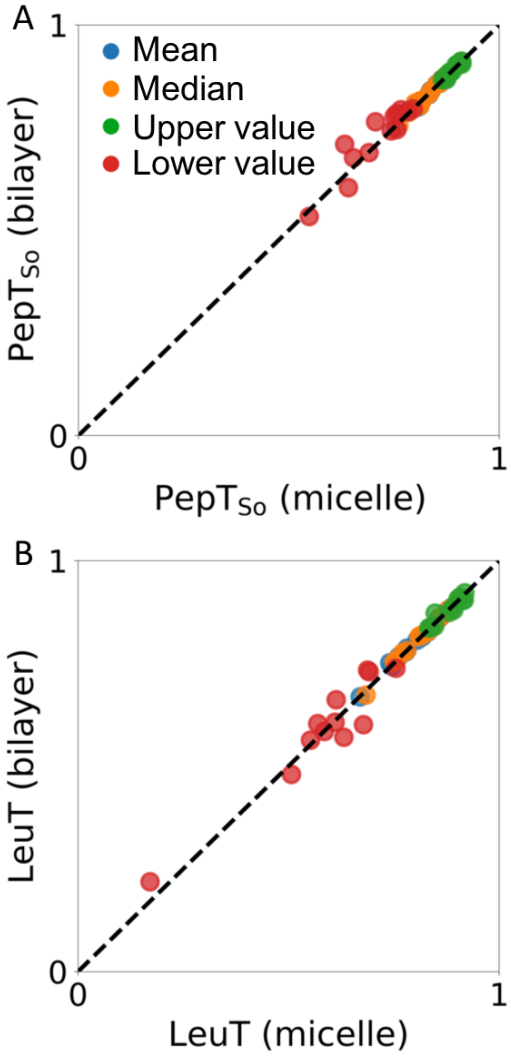


Figure 5.2: Comparing mean (blue), median (orange), upper value (green), and lower value (red) of alpha-helical content of (A) 14 TM helices in PepT$_{So}$, and (B) 12 TM helices in LeuT. Markers below the black dotted line indicate larger values observed in micelle environment. Markers above the black dotted line indicate larger values observed in bilayer environment. Markers along the black dotted line indicate similar observations in micelle and bilayer simulations.

Figure 5.2 strikingly shows that TM helicity median and mean values lie along the black-dotted line, and in most cases lower and upper values also don't deviate much in micelle and bilayer. In general, helicity values or distributions are not different which means that micelles do not impact the structure of the protein.

Covalent modification due to MTSSL probes cause small local structural perturbations on the protein

We used Kullback-Leibler (KL) divergence to quantify mis-match and differences among the distance distributions. KL Divergence for two distributions P and Q is 0 if and only if P and Q are equal almost surely. We calculated KL divergence among distance distributions from MD simulations and micelle and MD simulations in bilayer discussed above. Among 8 experimentally characterized distances in $PepT_{So}$, we found that residue pair Asn174-Ser466 has the highest KL divergence value. Hence, we chose this residue pair for further study, specifically to perform simulations with realistic nitroxide DEER labels. We attached an MTSSL DEER probe on Asn174 and Ser466 after mutating them to cystines via CHARMM-GUI these residues and simulated our protein in a BDDM micelle for $\sim$19 μs. To our knowledge, this is the first study of the impact of MTSSL spin labels on a protein and the resulting DEER distance distributions using unbiased simulations.

Figure 5.3A shows the simulated conformational ensemble projected on the intracellular and extracellular gating residues. Comparing this landscape to those for the $PepT_{So}$ simulations without probes in micelle shows that both ensembles capture all conformational states of the protein. This follows that probe molecules do not seem to interfere with the conformational dynamics of $PepT_{So}$ protein in any way that could hinder it's transport function.

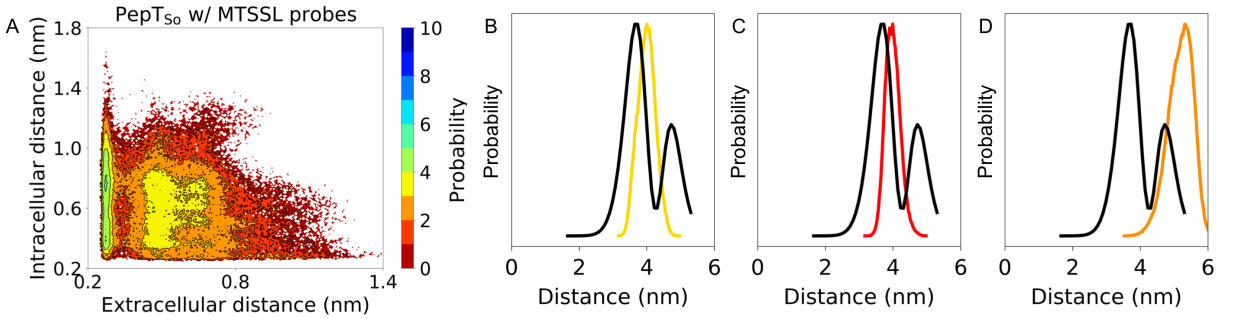


Figure 5.3: (A) The conformational landscapes of $PepT_{So}$ protein are generated by projecting all simulation data on the chosen extracellular and intracellular side distances measured between Arg32-Asp316 and Ser131-Tyr431, respectively.Conformational landscape for $PepT_{So}$ MD simulations in BDDM micelle with an MTSSL labeled residue pair. (B) Distance distribution for MTSSL labeled residue pair in $PepT_{So}$, 174-466, from simulations in BDDM micelle without probes (yellow), and (C) with probes (red). (D) Distance distribution for MTSSL labeled residue pair in $PepT_{So}$, 174-466, from simulations in BDDM micelle without probes (orange) where distances is measured between ON atoms on MTSSL labels. Black lines show DEER experiment distance distributions.

In order to understand the local effects of the MTSSL probes on the protein, we calculate Phi & Psi dihedral angles and generate Ramachandran plots for the mutated residues 174 and 466. We see a slightly larger coverage for residue 174 with MTSSL probe (Supplementary Figure 5.15B) as opposed to when it is an Asn residue (Supplementary Figure 5.15A), while there is no difference for residue 466. Similarly, when we look

at the regions surrounding the labeled residues, specifically two residues both before and after the labeled residues, we see larger distribution for residue 174 (Supplementary Figure 5.15E,F). Hence, we conclude that while mutants created for DEER spectroscopy experiments slightly impact the local dynamics and secondary structure of the protein, this affect is not significant and does not alter the overall conformational dynamics of the protein.

We suggest that any alteration seen in transport activity could be due to the kinetic rates of the transport function which would not affect the DEER observations unless functional interactions are mutated. Fowler et al. tested the transport activity of their $PepT_{So}$ double cysteine mutants and 174-466 mutant although decreased activity, did not abolish AlaAla transport entirely [111]. Kazmier et al. also tested binding of Leu to spin-labeled LeuT pairs and most double mutants retained more than 50% binding affinity as the wild type protein.

MTSSL probe distances are vastly different when compared to distances from wild-type protein simulations

We examine the impact of a spin-probe labeled residue pair on the resulting distance distributions (Supplementary Figure 5.16) by comparing micelle simulations with and without MTSSL probes. Since the probe molecules are on the extracellular side of the $PepT_{So}$ protein, we observe that the intracellular side distances show no differences and two extracellular side distance distributions do appear slightly perturbed. This affect is likely because of the two MTSSL molecules on this side of the protein. A closer look at the distribution for the residue pair labeled with MTSSL probes, 174-466, shows that the median values as well as 25% and 75% quartile values are conserved. Overall, we don't see any significant changes in the distance distributions for all 8 experimental distances as compared to the distance distributions obtained from simulations in BDDM micelle. This is expected, if there is no overall difference in the underlying conformational landscape as we discussed above, individual residue pair distances also would not deviate. Quantitatively, symmetrised divergence values indicated that distance distributions from MD simulations in micelle with and without probes were less divergent as compared to distance distributions from MD simulations in bilayer and micelle. Most nitroxide spin labels such as MTSSL consist of an unpaired electron on an oxygen atom which we will refer to as the ON atom. Dipolar coupling in DEER experiments is measured between two ON atoms, our simulations with two MTSSL labeled residues now allow us to obtain ON-ON atom distance distributions. Supplementary Figure 5.17 shows the ON-ON atom distance distributions as compared to the C_α and sidechain atom distances as observed in these probe simulations. Given the long length of the MTSSL molecule it is expected that ON-ON atom distance distributions are upward shifted with a median value of

5.22 nm. On the other hand the median value of the closest heavy atom based distance distribution is ∼0.9 nm lower than the ON-ON atom distance median.

Comparing these distributions to the experimental DEER distribution, we see single peaks from MD simulations whereas two peaks are seen in the experimental distribution, black lines in Figure 5.3. From Figure 5.3B and C, it appears that the MD simulation data captures conformations corresponding to the first peak, but Figure 5.3D shows that the ON-ON atom distance distribution points to conformations captured corresponding to the second peak with larger distance value. Evidently, the choice of atom for distance calculations is imperative and may significantly impact structural inferences from MD simulations. It is also important to note that while the ON-ON atom distance distribution changes our view of which peak our data corresponds to, we still do not match the experimental data and nor do we obtain multi-modal distributions as seen in experiments.

Restrained-ensemble MD simulations sample spin probe dynamics, but alter protein dynamics

Our results above elucidate that MTSSL probes modulate the distance distributions obtained from DEER experiments and the experimentally characterized distance distributions are a function of both the protein's dynamics as well as the probe's dynamics. MTSSL spin labels are long and flexible molecules and their dynamics have not been examined previously over a long time. We believe that our previous simulations are not sufficient, making unbiased simulations intractable to explore MTSSL probe dynamics. Restrained-ensemble MD (reMD) simulations have been used previously to restrain MTSSL probe's dynamics to the experimentally obtained DEER distributions and we explore this avenue to deconvolute the effect of MTSSL probe's dynamics from the experimental distributional data.

For our reMD simulations we first restrained residue pair 174-466. Since this residue pair is on the extracellular side of the $PepT_{So}$ protein, we chose another pair, 201-364, which had the highest KL divergence on the intracellular side. Hence, our next set of reMD simulations involved two restrained pairs one on each side of the protein. We dubbed these set of simulations as reMD (1 dist) and reMD (2 dist).

While in Figure 5.4A, B and Supplementary Figure 5.18A,B, the distance distributions between the ON-ON atoms of the MTSSL probes show a match with the experimental distribution, the closet-heavy atom distances don't. In Figure 5.4A, residue pair 174-466 distribution in teal violin plot has a single dominant peak with a median value of 3.22 nm, whereas the experimental distribution has two peaks. Moreover the same peak as seen in unbiased BDDM micelle simulations distribution shown in yellow violin plot is 4 nm. For comparison, this value is 3.98 nm for our unbiased simulations with a labeled residue pair. In general

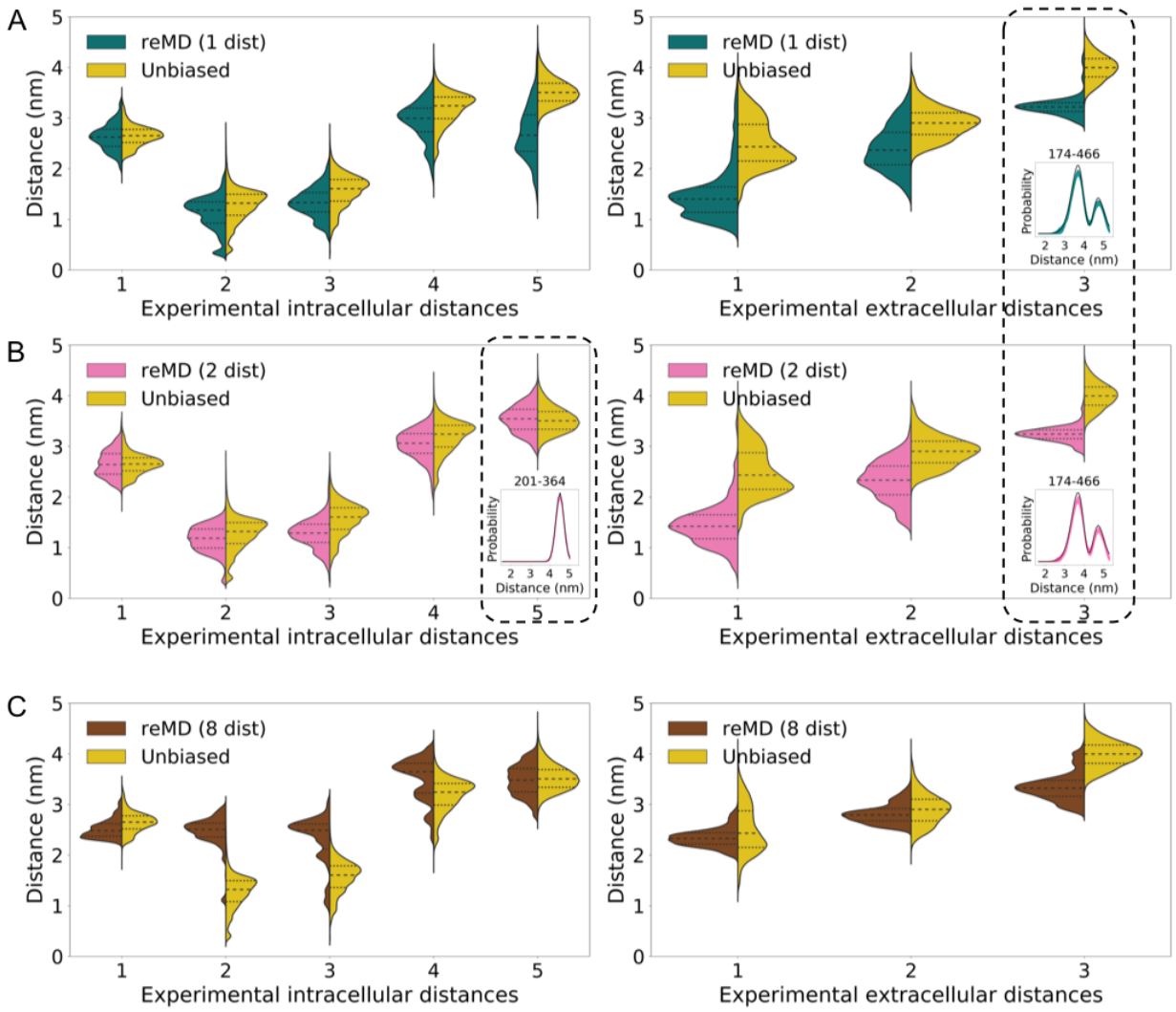


Figure 5.4: (A) Violin plot shows distance distributions for 5 intracellular residue pair distances and 3 extracellular residue pair distances as observed from (A) reMD (1 dist) simulations where residue pair 174-466 is restrained, teal violin plots, (B) reMD (2 dist) where residue pairs 174-466 and 201-364 are restrained, pink violin plots, and (C) reMD (8 dist) where all 8 residue pairs are restrained, brown violin plots. Yellow violin plots correspond to unbiased simulations of $PepT_{So}$ protein in micelle. Black dotted outlined residues pairs in (A) and (B) are restrained pairs and probe distances are shown to match with experimental DEER distance distributions.

all three extracellular distances in Figure 5.4A are lower shifted in reMD simulations. This shift is also seen in Figure 5.4B for the three extracellular distances. reMD (1 dist) and reMD (2 dist) systems both have a labeled residue pair on the extracellular side which would explain lower shifts for all distance shown on this side of the protein. While the distributions are lower shifted in reMD for the extracellular distances, neither a lower shift not an upward shift is seen in the five intracellular distances in Figure 5.4A or B. Comparing the residue pair 201-364 in Figure 5.4A and B, we note that when this residue pair is not restrained (teal violin plot) its mean value is 2.66 nm and when it is restrained this value is 3.54 nm, very close to unbiased simulation value of 3.51 nm.

What happens when we restrain all 8 residue pairs in system reMD (8 dist)? Three out of five distances

- distance #2, 3, and 4 - on the intracellular side show an upward shift, the median value of the brown violin plots is higher than the median value of the yellow violin plot distributions. Distance #1 and 2 on the extracellular side also are shifted up as compared to systems reMD (1 dist) and reMD (2 dist), although their median values are still lower than those in unbiased simulations.

An upward shift in distance distributions is similar to what we observe in Figure 5.3 where the ON-ON atom based distances shifted the distribution upwards by ∼0.9 nm. However, the origins of these shift are different. In particular, considering the distance distribution for residue pair 174-466 which is the third distance on the extracellular side, a lower shift in all reMD simulations compared to unbiased simulations without probes (yellow violin plots in Figure 5.4) and with probes (red violin plots in Supplementary Figure 5.18) indicates that reMD simulations alter the backbone dynamics in a way MTSLL probe labeled simulations did not. Vast differences in backbone dihedral angles of the relevant residues in reMD simulations support this observation (Supplementary Figure 5.19). These drastic shifts in distance distributions are mirrored in the underlying conformational landscapes (Supplementary Figure 5.20). Hence, bias introduced in reMD simulations via additional energetic terms for force calculations affect the protein structure differently than the modulation caused when MTSSL probes are attached to residues but simulated with unbiased MD simulations.

Similar to our MTSSL-labeled simulations, reMD simulations also suggest that the DEER experiment distance distributions are a convolution of both the spin probe distances as well as the inherent protein dynamics based distances. The impact of spin labels is not straightforward, while the ON-ON atom distance will always be larger than residue backbone or closest-heavy atom distances, the spin labels may modulate the protein's conformational dynamics. Unbiased simulations are not ideal to capture this impact, in our ∼19 μs simulations we did not see this difference, however reMD simulations proved otherwise. Moreover, limited by computational resources it is not feasible to perform long timescale residue pairwise simulations with MTSSL probes. At the same time, while pairwise reMD simulations are also not cheaper, multiple residue pairs can be combined together as we demonstrated for our systems reMD (2 dist) and reMD (8 dist). While this may make them computationally tractable, this still does not solve the problem of an unbiased match with MD simulations from long timescale MD simulations. reMD simulations with multiple restrained residue pairs also raise the unexplored concern that what number of restraints in reMD simulations would be adequate to capture an MD ensemble where all residue pair distance distributions can correspond to their DEER experiments observables without perturbing the protein's conformational dynamics.

5.5 Discussion

This work highlights the necessity for careful interpretation of DEER spectroscopy and MD simulations in membrane protein biophysics. Scarcity of membrane protein biophysical characterization necessitates that we salvage all information available from laboratory experiments and computational simulations. Hence DEER spectroscopy and MD simulations will continue to be important techniques in progressing our understanding of protein dynamics. It is, therefore, imperative to understand how to best compare data obtained from both techniques, not only to show a validation of MD simulations with experiments, but also to avoid misleading conclusions and to draw predictive conclusions. Previous work in Chapter 2 has proposed optimization protocols to choose ideal choice of residue pairs for DEER experiments from already performed MD simulations. These protocols can also be used iteratively, performing simulations followed by experiments and then more simulations to update our understanding of a protein's conformational changes, as shown in Chapter 3. Such methods can be used to their full potential once we can decipher structural characterization of different protein modes identified via multiple peaks in DEER distance distributions. Hence, in this work we carry out a comprehensive study of potential reasons for discrepancy between DEER experiment distributional data and residue pair distributions from atomistic MD simulations.

We show that major reason for the difference between experiments and simulation distributions is due to the long length of the MTSSL label and it's slow dynamics. The slow dynamics of the flexible MTSSL probes could not be captured in unbiased MD simulations and we examined this using biased simulation methods. While reMD simulations and other biased simulation strategies can reconcile experiments and simulations for the restrained residue pairs, reMD yielded significant changes in the protein's conformational dynamics itself including, both residue-level and global perturbations. It is also not feasible for researchers to perform DEER experiments on all residue pairs of a protein which can be followed by multiple residue pair biased reMD simulations. On the other hand, while unbiased MD simulations do not cause any unphysical perturbations in the protein, it is computationally expensive to perform long timescale MD simulations with MTSSL probes. We surmise that when using methods such as *OptimalProbes* it would be sufficient to perform MD simulations for the top predicted choices for DEER experiments .

Another strategy is for experiments to use alternative probe molecules, such as metal cation based probes which are more rigid [253] or use biophysical experimental methods that do not require any changes to the covalent structure of the target protein that affect the protein's dynamics and sometimes function. One such technique is Hydrogen Deuterium Exchange Mass Spectrometry which has been used for membrane proteins including for MFS and NSS family proteins [281–283]. While these experiments are performed in micelle and nanodisc, our work shows that membrane environment does not influence protein dynamics while covalently

linking MTSSL spin probes do alter the obtained measurements.

5.6 Supplementary Information

Experimental DEER distances

Experimental DEER distances and distance distributions were extracted from previous experiments published in ref. [111] and [257] using Plot Digitizer Java program.

For $PepT_{So}$, 8 DEER distributions are available:

- 5 intracellular distances (86-432,141-432,141-438,141-500,201-364)
- 3 extracellular distances (47-330,174-401,174-466)

For LeuT protein, we have examined 24 distance distributions because these distributions have data available for Apo system in ref. [257].

- 17 intracellular distances (185-271,79-277,184-277,7-86,12-86,12-377,71-193,193-377,12-371,71-89,71-184,71-377,79-377,71-425,71-455,277-425,277-455)
- 7 extracellular distances (309-480,123-240,208-240,37-123,37-208,123-306,208-306)

$PepT_{So}$ experimental distributions were fitted to multiple Gaussian distributions in order to get an equal-sized-bin distribution for KL divergence calculations and restrained MD simulations. Comparisons shown in Supplementary Figure 5.23.

MD simulations

List of LeuT structural models

2A65 [284], 2Q6H [285], 2Q72 [285], 2QB4 [285], 2QEI [285], 3F3A [286], 3F3C [286], 3F3D [286], 3F3E [286], 3F48 [286], 3F4I [286], 3F4J [286], 3GJD [287], 3GWU [288], 3GWV [288], 3GWW [288], 3TT1 (chains A & B) [271], 3TT3 (chain A) [271], 3USG [289], 3USI (chains A & B) [289], 3USJ (chains A & B) [289], 3USK (chains A, B, C, & D) [289], 3USL [289], 3USM [289], 3USO (chains A & B) [289], 3USP [289], 5JAE (chains A & B) [290], 5JAF [290].

Data analysis

Micelle radius. First, we compute the radius of gyration (R_g) of the micelle using *compute_rg* in MDTraj 1.7 [129], which is related to the micelle radius (R) as, $R = \sqrt{\frac{5}{3}} R_g$ [291]. This formula hold when the micelle is assumed to be spherical in shape.

Eccentricity. The shape of the micelle and the protein-micelle complex is determined using the ratio between moments of inertia I_1, I_2, and I_3 Supplementary Table 5.2. Eccentricity is calculated as $1-I_{min}/I_{avg}$ [291, 292]. The moments are inertia are defined as the eigenvalues of a moment of inertia tensor calculated using *compute_inertia_tensor* in MDTraj 1.7 [129].

Distance distributions. All inter-residue distance distributions are estimated as the distance between closest heavy atoms between the two residues, unless otherwise mentioned.

Inter-helix distances Transmembrane helix ends for proteins are defined based on OPM database [293] numbering for 14 helixes in PepT_{So} and 12 in LeuT. We determine inter-helix distances among all helixes on intracellular and extracellular side of the proteins. For PepT_{So} these are $2 \times (14)(13)/2 = 182$ distances and for LeuT these are $2 \times (12)(11)/2 = 132$ distances.

Kullback-Leibler (KL) divergence. KL divergence (also called relative entropy) is a measure of how one probability distribution is different from a second, reference probability distribution. KL divergence for two distributions P and Q is 0 if and only if P and Q are equal almost surely. For two discrete probability distributions P and Q defined on the same probability space, X, the KL divergence of Q from P is defined to be,

$$KL(P|Q) = -\sum_{x \epsilon X} P(x) log(\frac{Q(x)}{P(x)}) \tag{5.1}$$

KL divergence is an assymmetric measure by definition, and wherever possible we have used this measure both ways to validate our conclusions regarding similarity and difference among probability distribution. We used *scipy.stats.entropy* routine to calculate KL divergence values. Another useful measure of divergence between probability distribution we use is Symmetrised Divergence, which is symmetric and non-nagative defined as,

$$\text{Divergence} = KL(P|Q) + KL(Q|P) \tag{5.2}$$

When calculating frequencies used for the KL divergence we corrected for the presence of frequencies of zero by adding a very small value to the probability distribution.

Helical content. The helical content of the all TM helixes are calculated as defined in the NAMD 2.11 manual [13]. The python implementation is taken from `https://github.com/amoffett/alpha_helical_content` as used in ref. [294]. The individual helixes in this work are as defined by the OPM database web server [293] for PepT_{So} PBD 4UVM [111] and LeuT PDB 2A65 [284]. Specifically, TM1 of LeuT refers to residues in TM1a only, which are residues 15 to 25 whereas the helix ends at residue 35.

Supplementary Table 5.1: List of MD simulations.

Complex	Components[#1]	# Trajectories	# Atoms	Equilibration run (ps)	Simulation time (for analysis[#2], μs)
LeuT-bilayer	150 POPE, Cl^-	72	56,707 - 66,784	675	32.18
LeuT-micelle	150 BDDM, Cl^-	72	107,521 - 145,589	450	28.73
$PepT_{So}$-bilayer	150 POPE, 50 POPG, NaCl	42	66,045 - 74,700	675	27.3
$PepT_{So}$-micelle	150 BDDM, NaCl	42	130,506 - 195,681	750	20.42
$PepT_{So}$-micelle-MTSSL probes	150 BDDM, NaCl	42	128404 - 181800	750	18.78

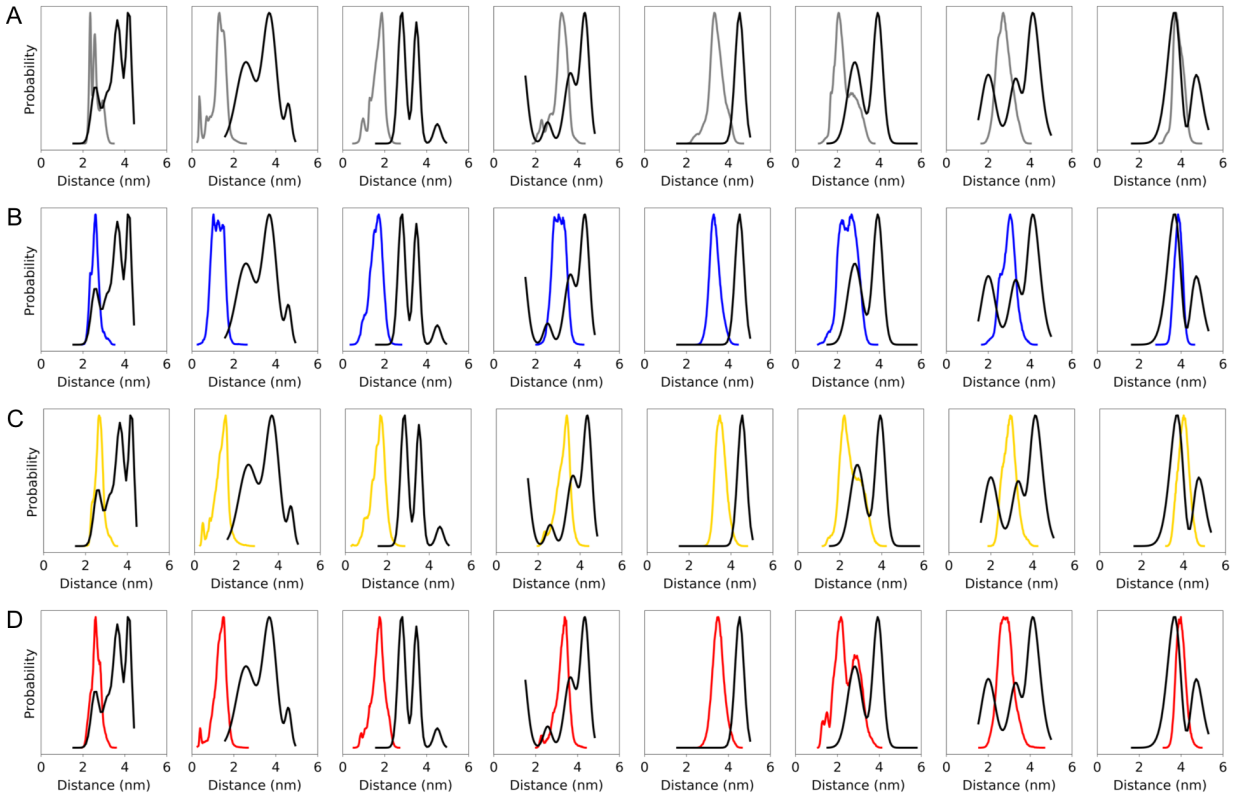
#1: All systems contain protein and TIP3P water.
#2: For all trajectories, we eliminate the first 10 ns of the production run from analysis.

Supplementary Table 5.2: Geometry of protein-micelle complexes with varied micelle sizes.

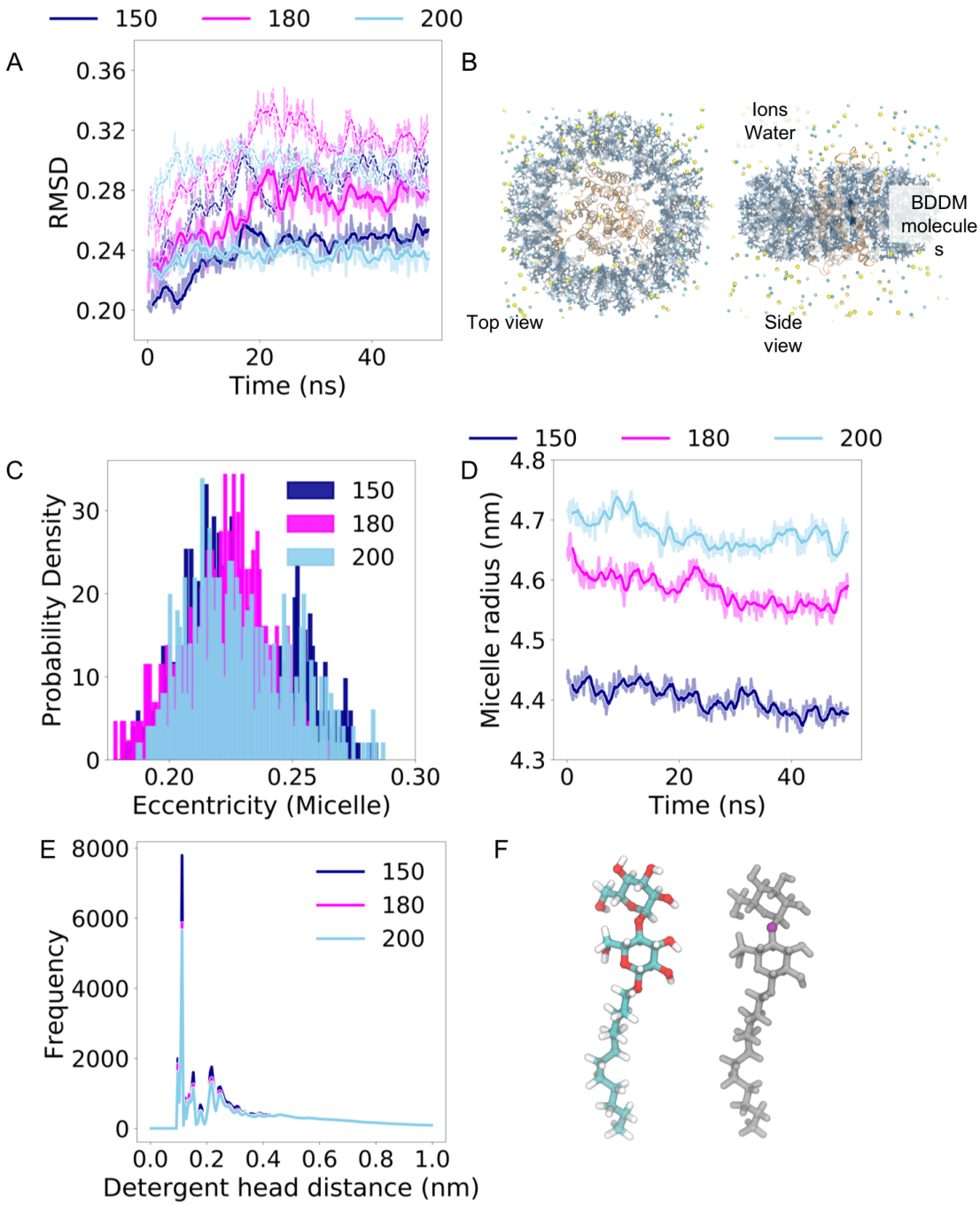
	Complex PepTSo w/ detergents	Micelle Radius (nm)	I1 : I2 : I3	Eccentricity
Micelle	150 detergents	4.4 ± 0.02	1.46 : 1 : 1.24	0.23 ± 0.02
	180 detergents	4.58 ± 0.03	1.42 : 1 : 1.25	0.22 ± 0.02
	200 detergents	4.68 ± 0.02	1.53 : 1 : 1.16	0.22 ± 0.02
Protein+Micelle	150 detergents	-	1.26 : 1 : 1.16	0.16 ± 0.02
	180 detergents	-	1.23 : 1 : 1.18	0.16 ± 0.01
	200 detergents	-	1.38 : 1 : 1.13	0.17 ± 0.02

Supplementary Table 5.3: List of reMD simulations.

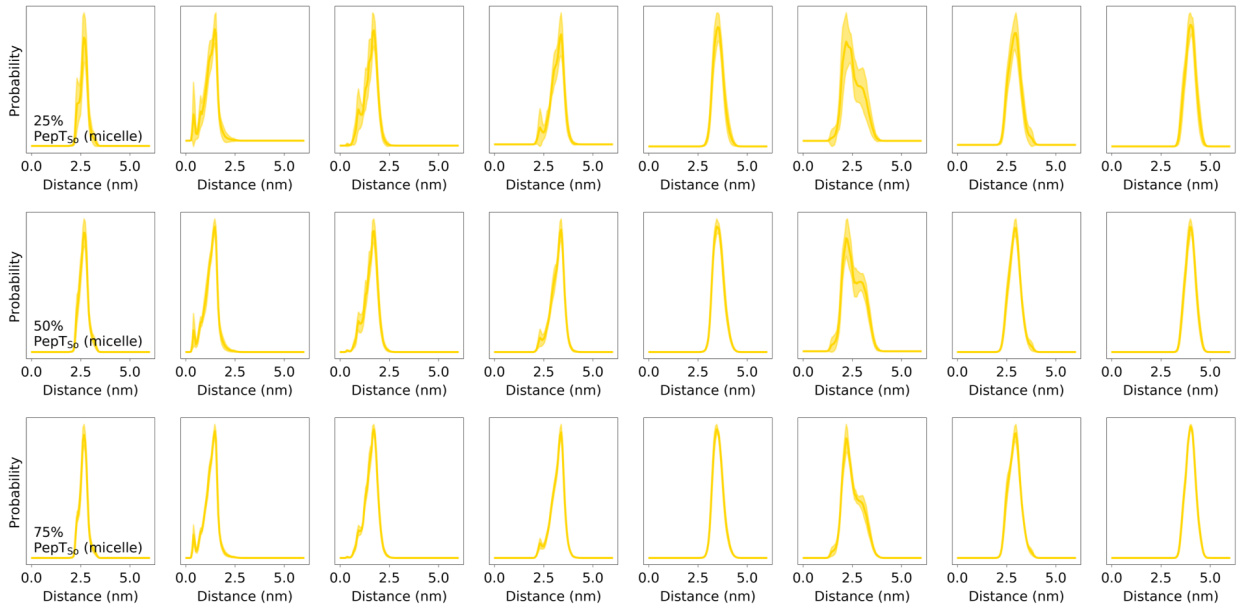
System	# Trajectories	# Atoms	Equilibration run (ps)	Production run (ns)	Simulation time (μs)
reMD (1 dist)	42	9795	25	∼95	3.98
reMD (2 dist)	42	11645	25 (2 setups required 50 ps)	∼95	3.88
reMD (8 dist)	42	19049	25 (8 setups required longer)	∼65	2.67



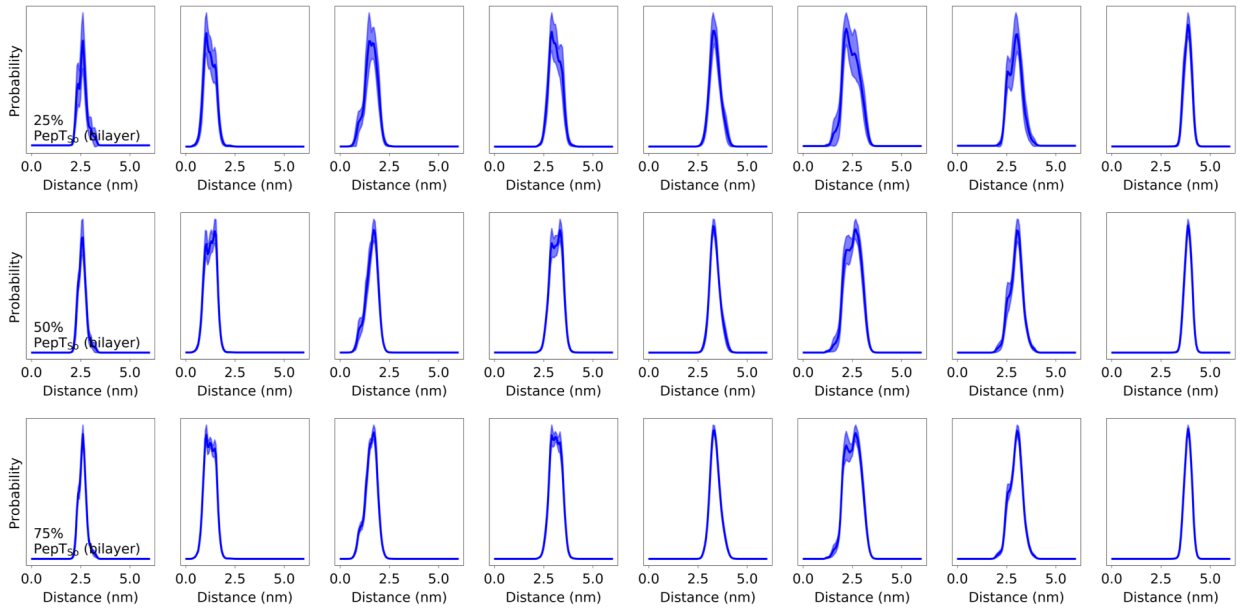
Supplementary Figure 5.1: Experimentally characterized residue pair distance distributions as observed in our MD simulations in (A) POPC bilayer (simulations previously performed in Chapter 4), (B) POPE/POPG (3:1 ratio) bilayer, (C) BDDM micelle, and (D) BDDM micelle with MTSSL labeled residue pair. Black lines show experimental DEER distance distributions obtained from Fowler et al. as discussed above [111].



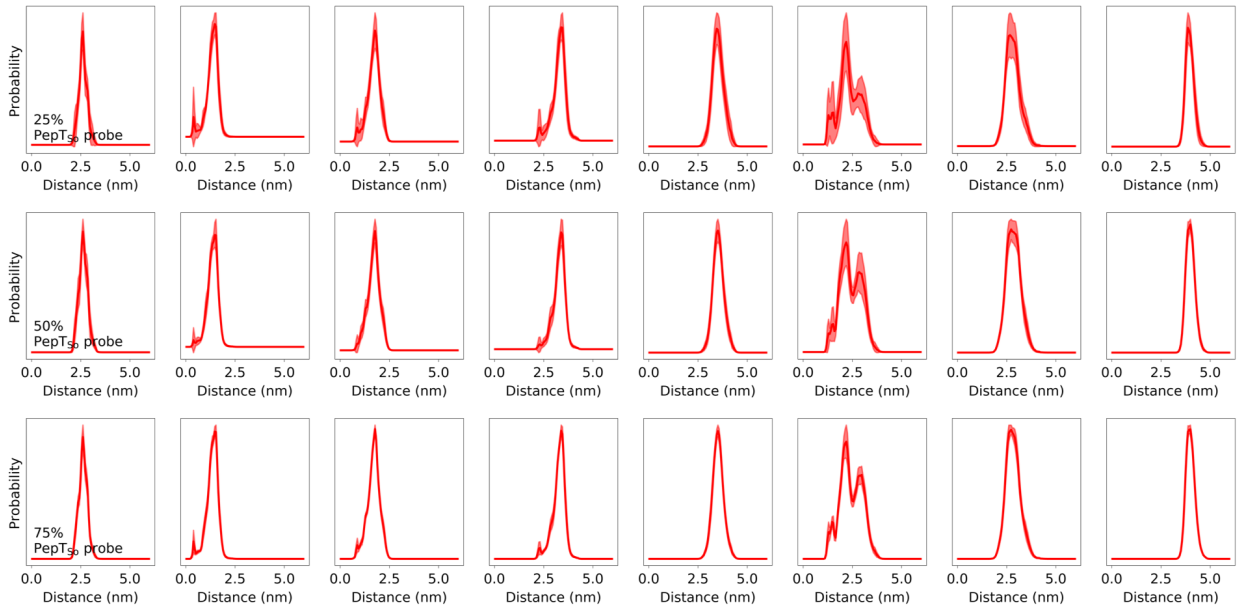
Supplementary Figure 5.2: (A) RMSD of protein with respect to the starting frame is shown with time. Dotted lines show RMSD of the full protein while the bold lines show RMSD of transmembrane region of the protein. Shaded regions show instantaneous values while the lines show a running time average RMSD over a 1 ns time window. (B) An example protein-micelle setup top and side view including BDDM detergent molecules and ions. (C) Probability distribution of micelle eccentricity values. (D) Micelle radius with time is shown. Shaded regions show instantaneous values while the lines show a running time average radius over a 1 ns time window. (E) Radial distribution of distances between BDDM detergent molecule headgroups. Headgroup distances are estimated using the distance among oxygen atoms highlighted in magenta in (F). Colors indicate three micelle sizes, micelle with 150 (blue), 180 (magenta), 200 (skyblue) detergent molecules.



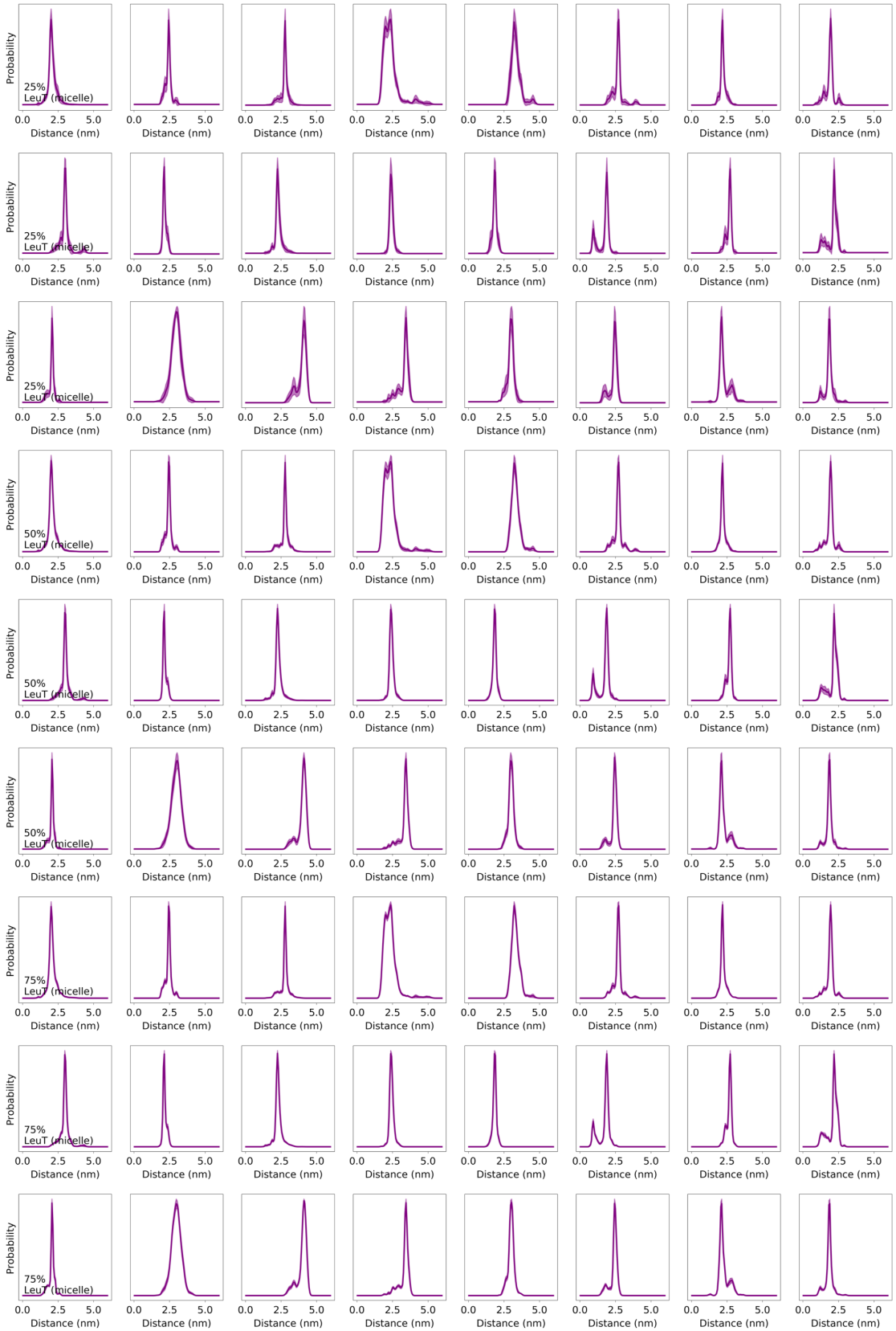
Supplementary Figure 5.3: Residue pair distance distributions for $PepT_{So}$ simulations in BDDM micelle averaged over 25%, 50%, and 75% of the collected trajectories. Filled regions show error bars in the distance distribution as obtained from 10 iterations where a subset of the trajectories are selected randomly.



Supplementary Figure 5.4: Residue pair distance distributions for $PepT_{So}$ simulations in bilayer averaged over 25%, 50%, and 75% of the collected trajectories. Filled regions show error bars in the distance distribution as obtained from 10 iterations where a subset of the trajectories are selected randomly.

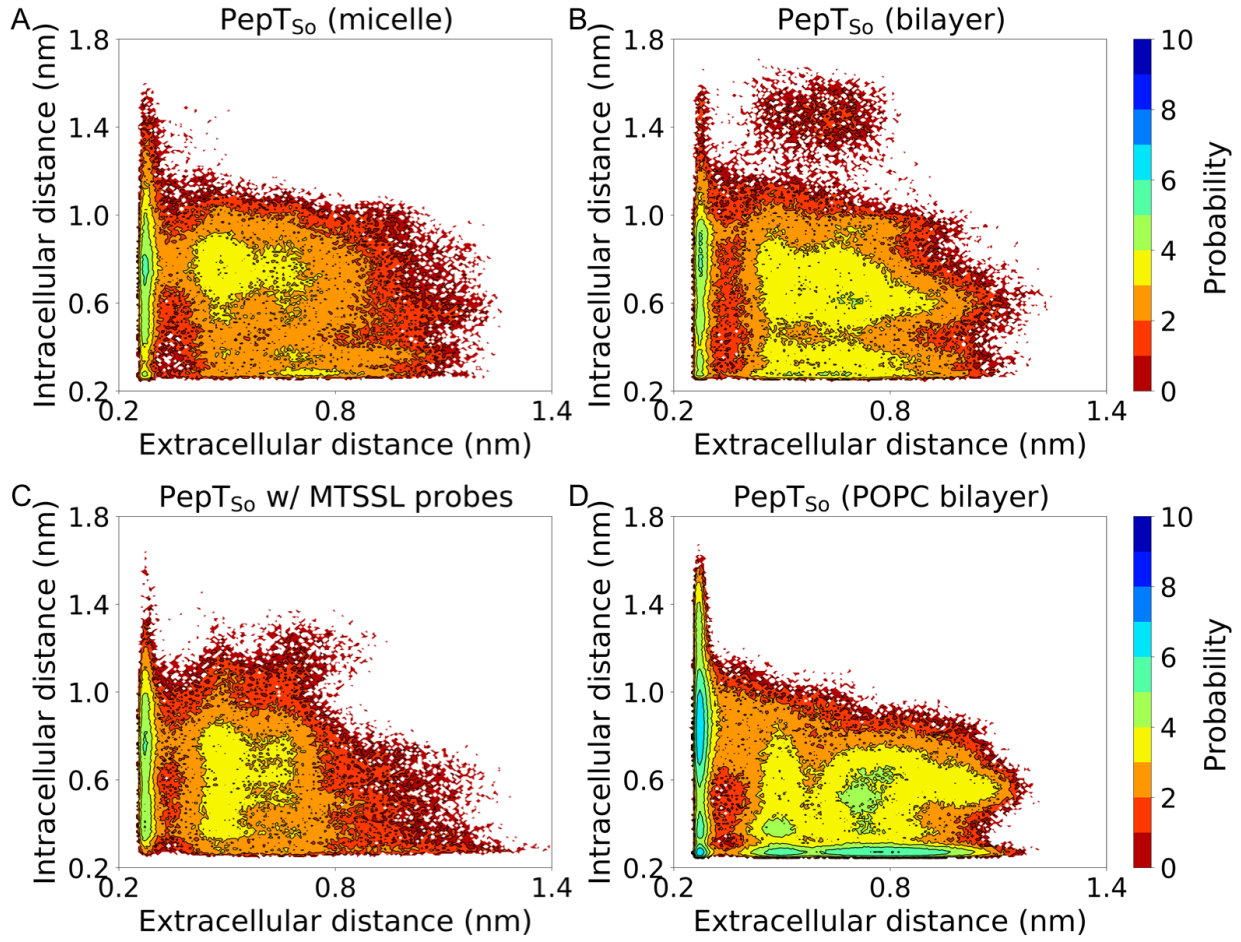


Supplementary Figure 5.5: Residue pair distance distributions for $PepT_{So}$ simulations in BDDM micelle with a residue pair labeled residue pair averaged over 25%, 50%, and 75% of the collected trajectories. Filled regions show error bars in the distance distribution as obtained from 10 iterations where a subset of the trajectories are selected randomly.

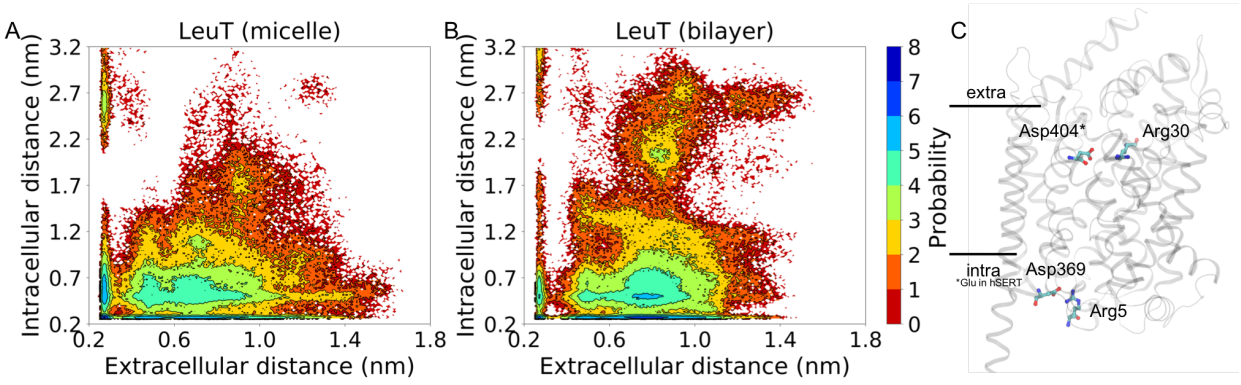


Supplementary Figure 5.6: Residue pair distance distributions for LeuT simulations in BDDM micelle averaged over 25%, 50%, and 75% of the collected trajectories. Filled regions show error bars in the distance distribution as obtained from 10 iterations where a subset of the trajectories are selected randomly.

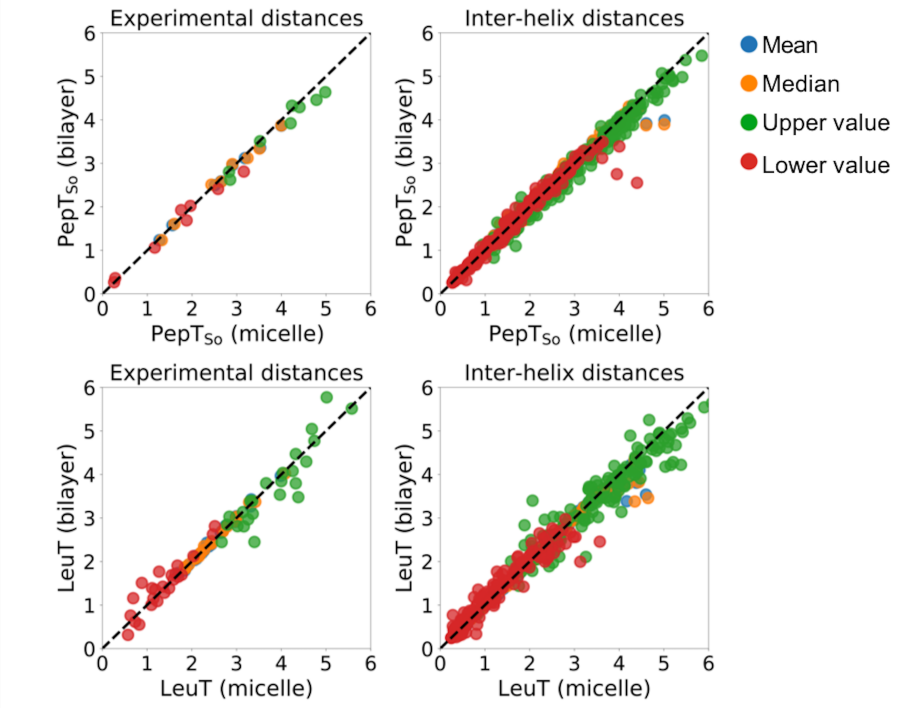
Supplementary Figure 5.7: Residue pair distance distributions for LeuT simulations in bilayer averaged over 25%, 50%, and 75% of the collected trajectories. Filled regions show error bars in the distance distribution as obtained from 10 iterations where a subset of the trajectories are selected randomly.



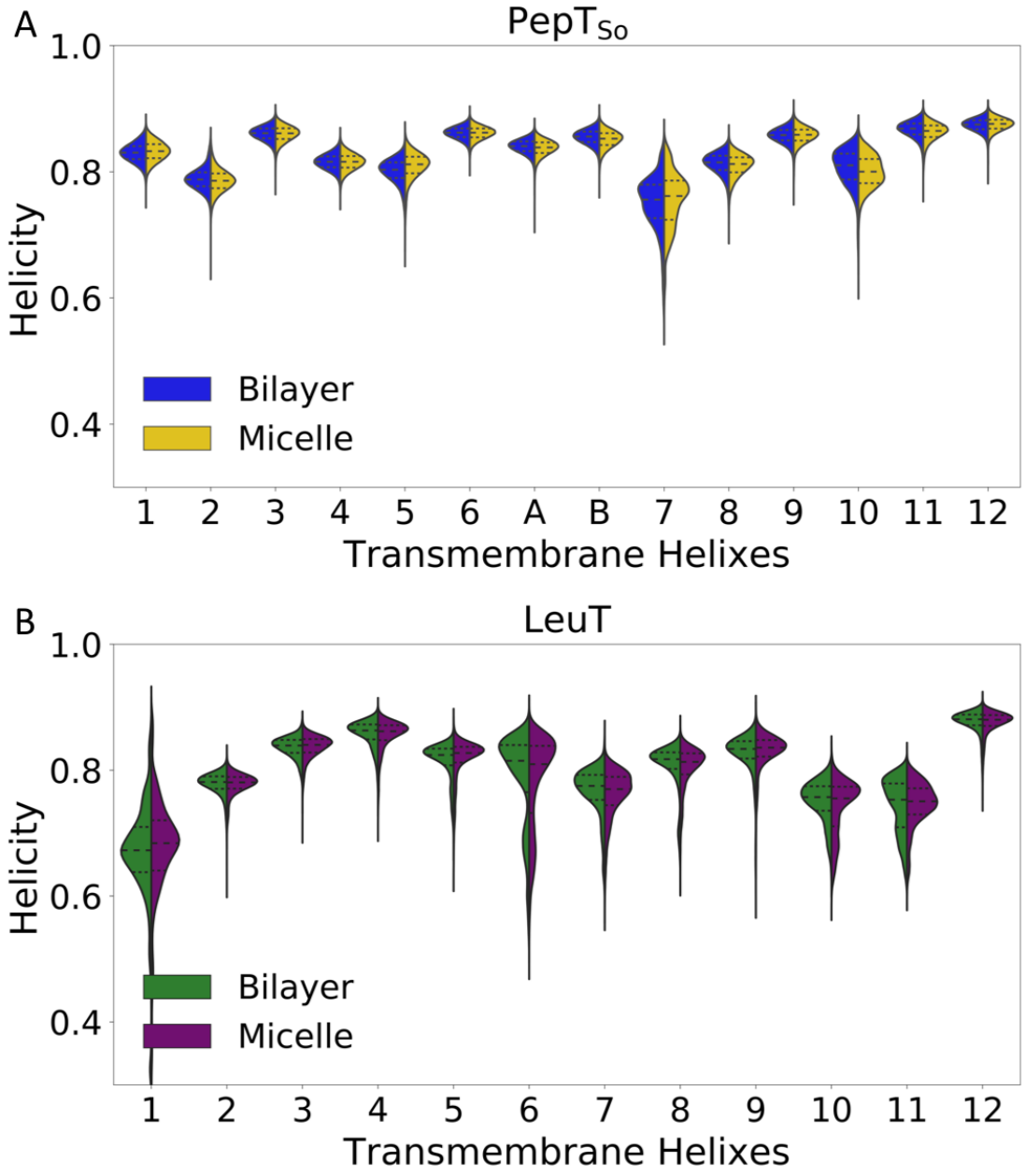
Supplementary Figure 5.8: The conformational landscapes of $PepT_{So}$ protein are generated by projecting all simulation data on the chosen extracellular and intracellular side distances measured between Arg32-Asp316 and Ser131-Tyr431, respectively. (A) Conformational landscape for $PepT_{So}$ MD simulations in BDDM micelle. (B) Conformational landscape for $PepT_{So}$ MD simulations in POPE/POPG (3:1 ratio) bilayer. (C) Conformational landscape for $PepT_{So}$ MD simulations in BDDM micelle with an MTSSL labeled residue pair. (D) Conformational landscape from our previous simulations in a POPC bilayer and using an AMBER FF14SB force field.



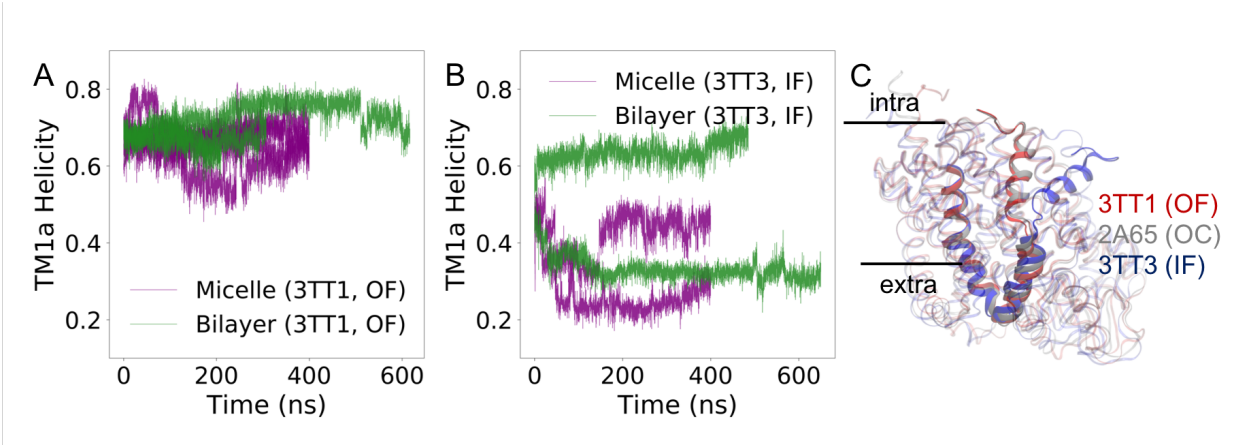
Supplementary Figure 5.9: The conformational landscapes of LeuT protein are generated by projecting all simulation data on the chosen extracellular and intracellular side distances measured between Arg30-Asp404 and Arg5-Asp369, respectively. (A) Conformational landscape for LeuT MD simulations in BDDM micelle. (B) Conformational landscape for LeuT MD simulations in a bilayer. (C) Gating residues used to determine extracellular and intracellular distances are shown on a cartoon representation of a three-dimensional LeuT structure.



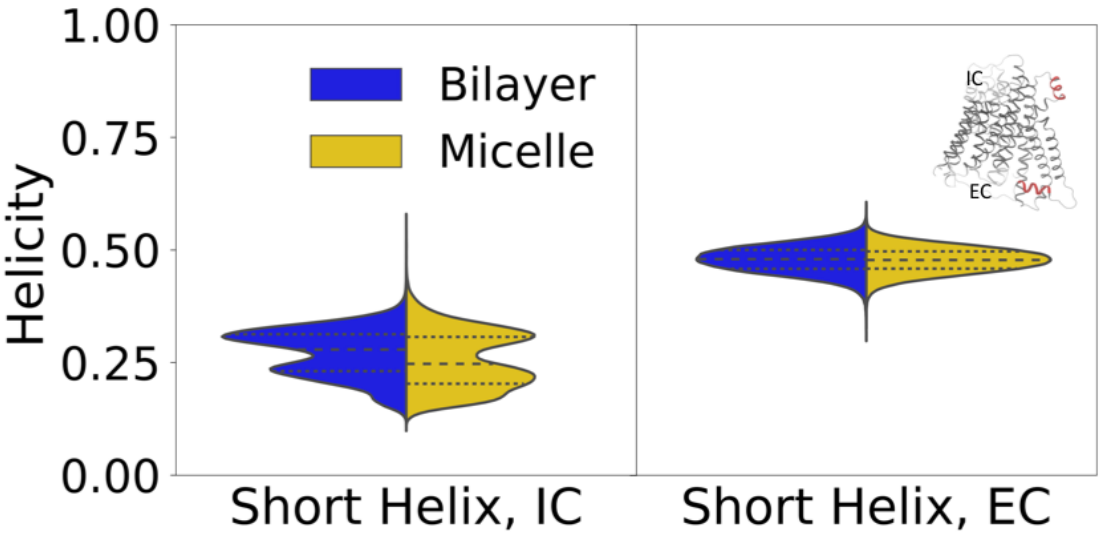
Supplementary Figure 5.10: Comparing mean (blue), median (orange), upper value (green), and lower value (red) of distance distributions of experimental residue pair distances and all inter-helix residue pair distances. Markers below the black dotted line indicate larger values observed in micelle environment. Markers above the black dotted line indicate larger values observed in bilayer environment. Markers along the black dotted line indicate similar observations in micelle and bilayer simulations.



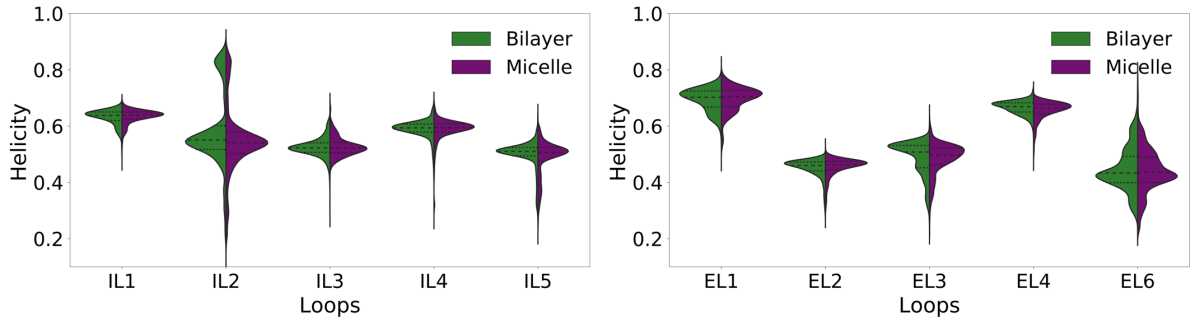
Supplementary Figure 5.11: (A) Violin plot shows alpha-helical content for 14 TM helices as observed from MD simulations of $PepT_{So}$ protein in micelle (yellow, right) and bilayer (blue, left). (B) Violin plot shows alpha-helical content for 12 TM helices as observed from MD simulations of LeuT protein in micelle (purple, right) and bilayer (green, left).



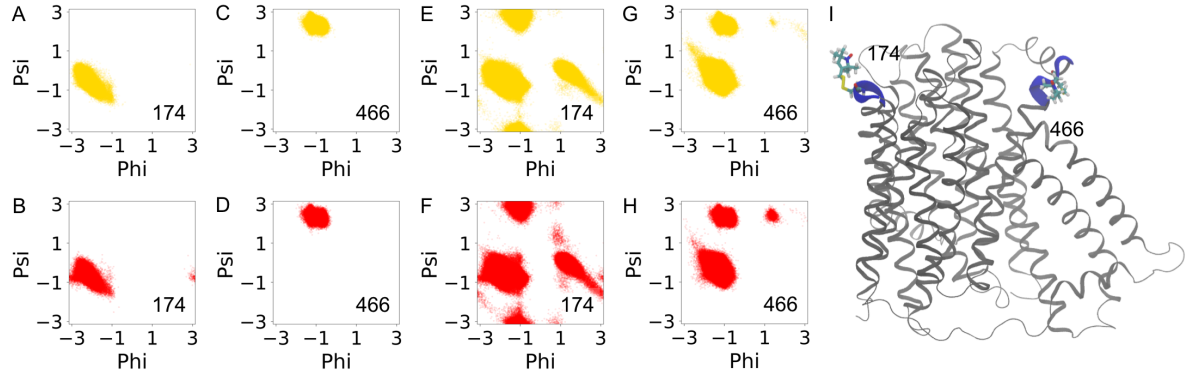
Supplementary Figure 5.12: (A) TM1a alpha-helical content of trajectories started from OF structure of LeuT in micelle (purple) and bilayer (green). (B) TM1a alpha-helical content of trajectories started from IF structure of LeuT in micelle (purple) and bilayer (green). (C) Superposed structures of LeuT's OF, OC and IF structures.



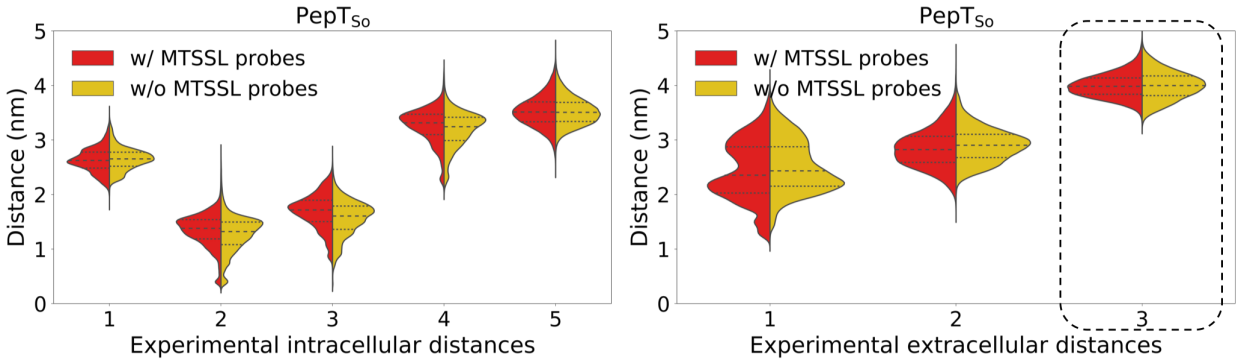
Supplementary Figure 5.13: Violin plot shows alpha-helical content of a short helix on the intracellular (IC) side and another of the extracellular (EC) side of $PepT_{So}$ protein in micelle (yellow, right) and bilayer (blue, left). Inset shows two short helices in red on the $PepT_{So}$ protein structure in grey.



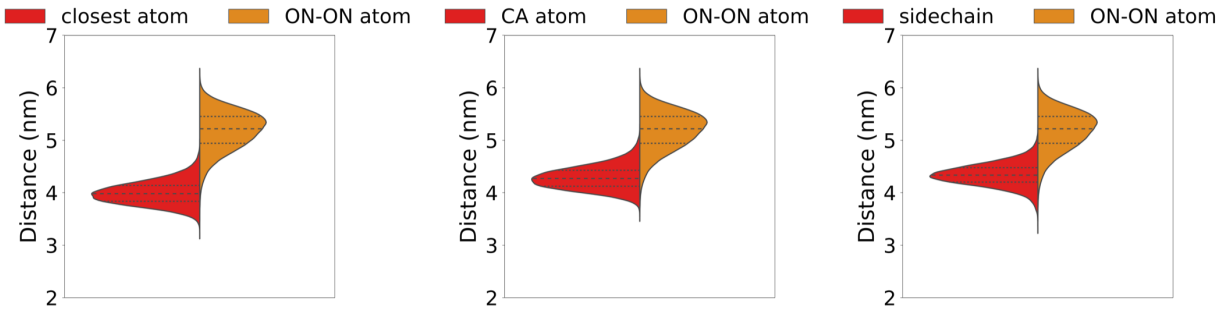
Supplementary Figure 5.14: Violin plot shows alpha-helical content of intracellular loops (ILs) and extracellular loops (ELs) in LeuT protein in micelle (purple, right) and bilayer (green, left). Loop EL5 is only 4 residues long and too short to determine it's alpha-helical content.



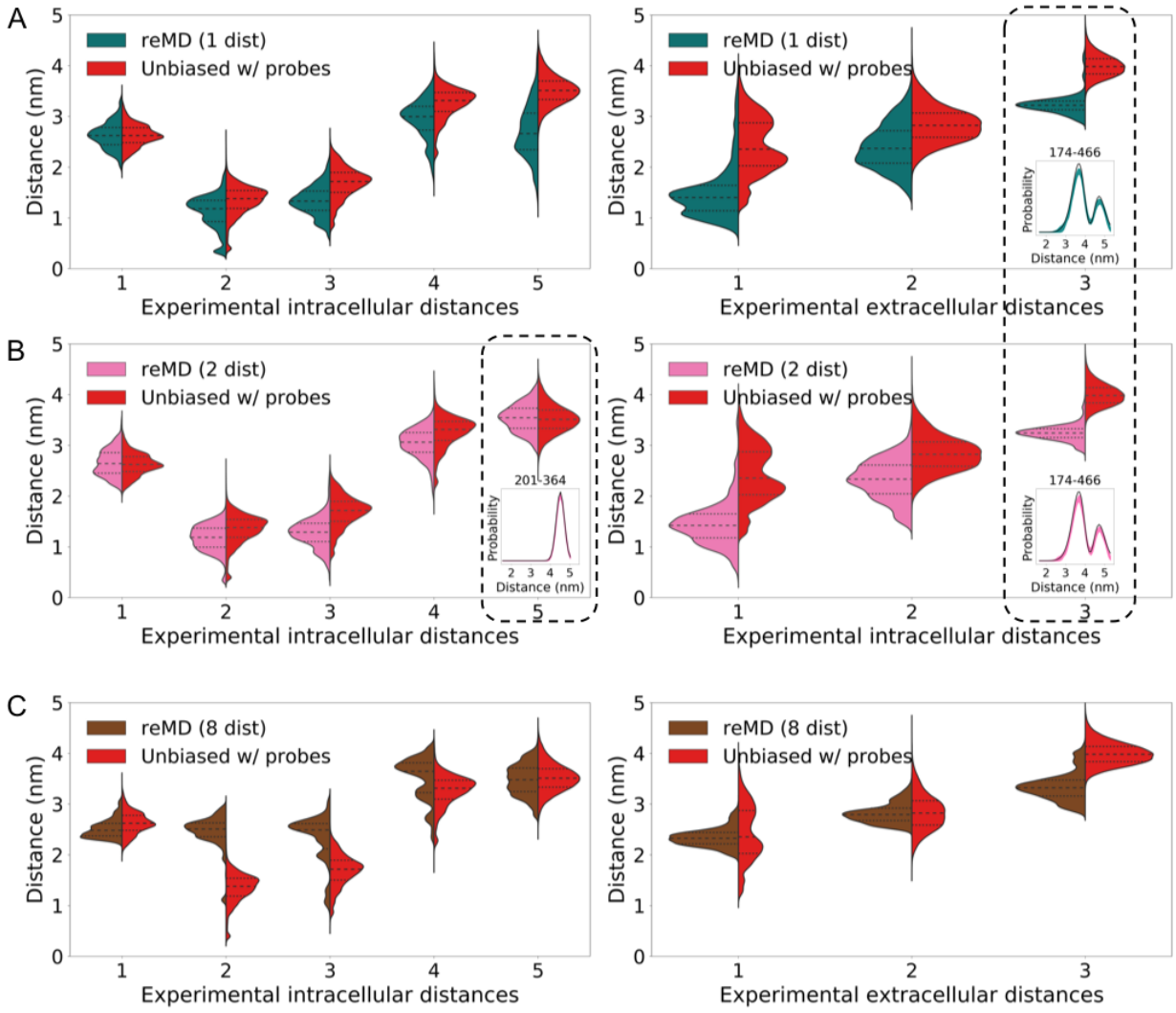
Supplementary Figure 5.15: (A-D) Ramachandran plots for residues 174 and 466. Yellow and red colors indicate residue dihedral angle distribution in micelle and bilayer MD simulations, respectively. (E-F) Ramachandran plots for regions surrounding residues 174 and 466. (I) Residues 174 and 466 are shown on a cartoon representation of a three-dimensional $PepT_{So}$ structure.



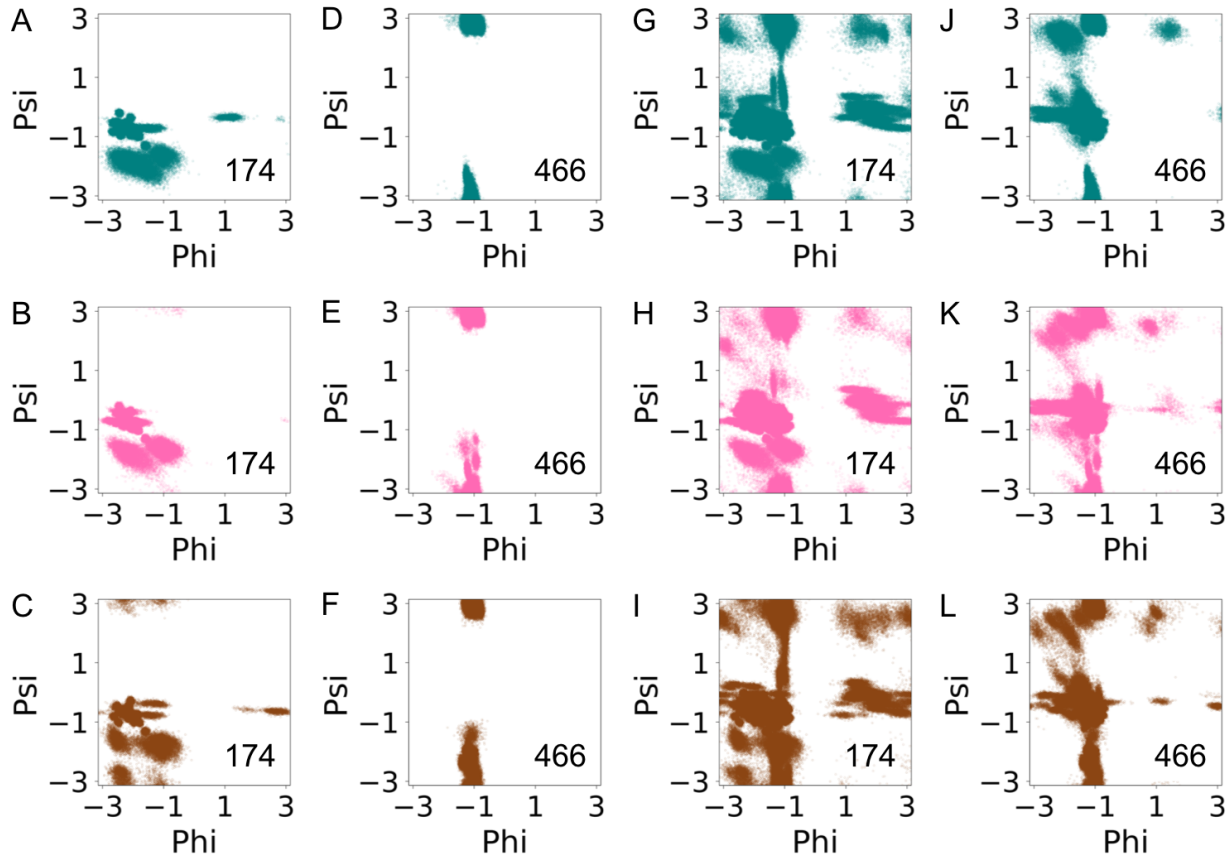
Supplementary Figure 5.16: (A) Violin plot shows distance distributions for 5 intracellular residue pair distances and 3 extracellular residue pair distances measured by Fowler et al. as observed from MD simulations of $PepT_{So}$ protein in micelle without MTSSL probes (yellow, right) and with an MTSSL probe labeled residue pair (red, left). Black dotted outlined residues pair is the labeled residue pair.



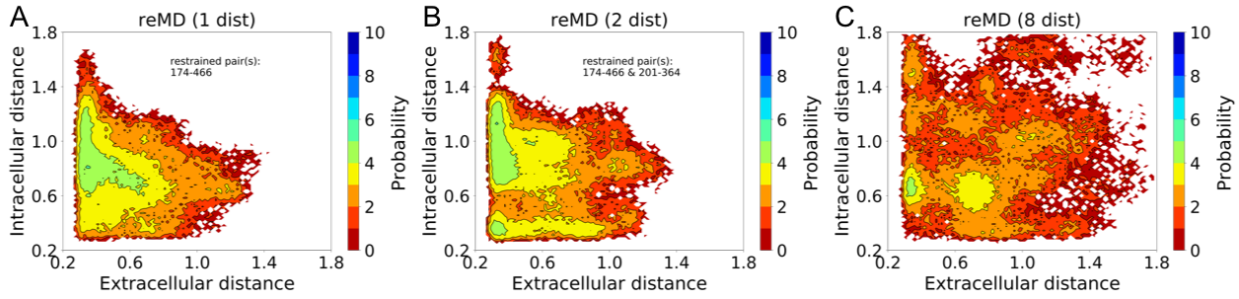
Supplementary Figure 5.17: Violin plots compare distance distributions for simulations with MTSSL probes as measured between the ON atom with the closest heavy atom, C_α atom, and the closest sidechain atom of the labeled residues. ON-ON atom distance distributions are shown in orange and the backbone atom distance distributions are shown in red.



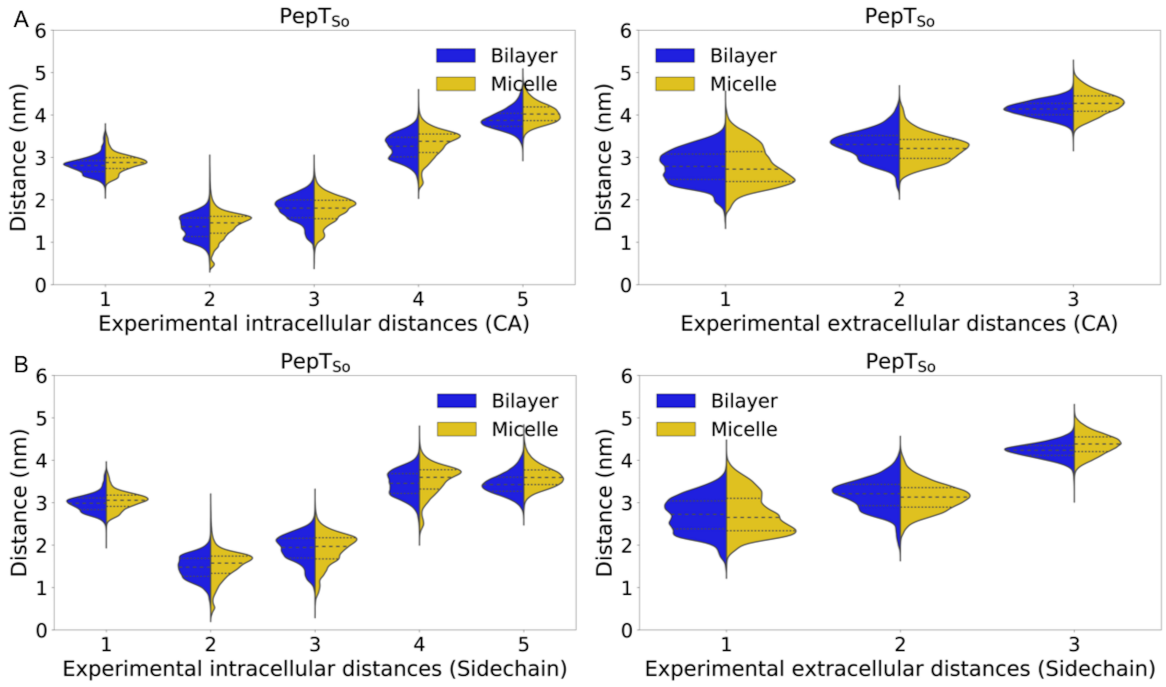
Supplementary Figure 5.18: (A) Violin plot shows distance distributions for 5 intracellular residue pair distances and 3 extracellular residue pair distances as observed from (A) reMD (1 dist) simulations where residue pair 174-466 is restrained, teal violin plots, (B) reMD (2 dist) where residue pairs 174-466 and 201-364 are restrained, pink violin plots, and (C) reMD (8 dist) where all 8 residue pairs are restrained, brown violin plots. Yellow violin plots correspond to unbiased simulations of $PepT_{So}$ protein in micelle with MTSSL molecules on residues 174 and 466. Black dotted outlined residues pairs in (A) and (B) are restrained pairs and probe distances are shown to match with experimental DEER distance distributions.



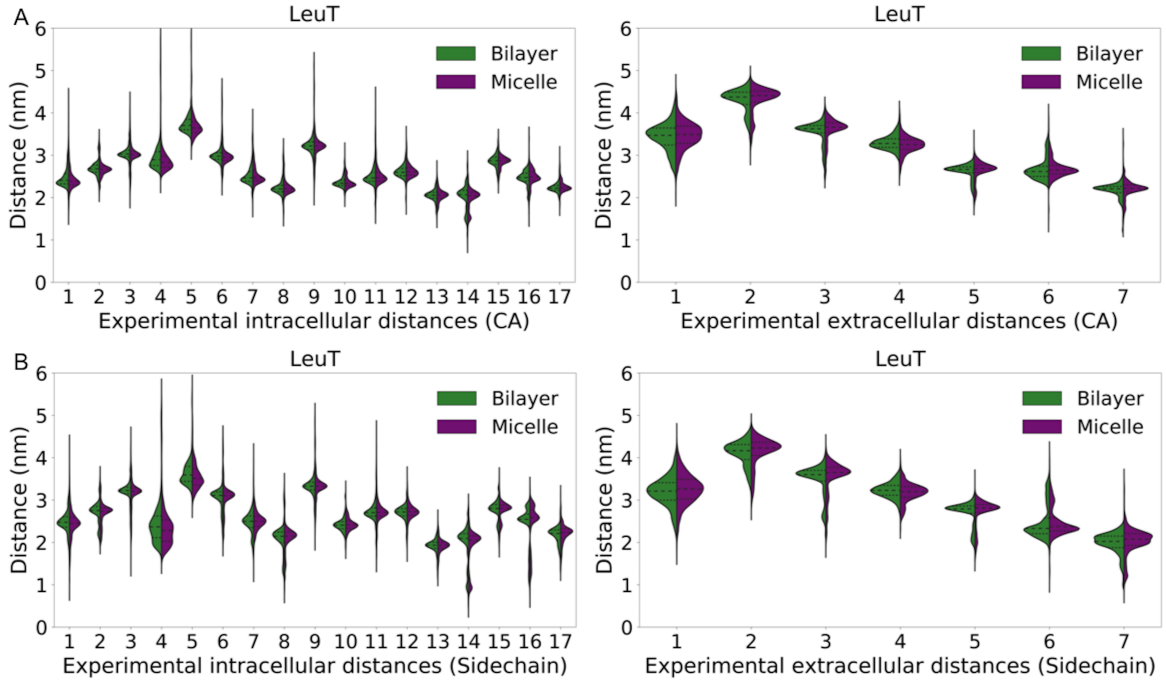
Supplementary Figure 5.19: (A-F) Ramachandran plots for residues 174 and 466. Teal, pink, and brown colors indicate residue dihedral angle distribution in reMD (1 dist), reMD (2 dist), and reMD (8 dist) MD simulations, respectively. (G-L) Ramachandran plots for regions surrounding residues 174 and 466.



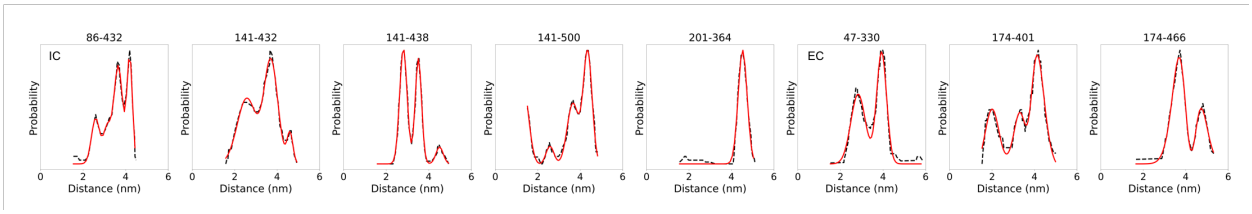
Supplementary Figure 5.20: Conformational landscape for $PepT_{So}$ (A) reMD (1 dist), (B) reMD (2 dist), and (C) reMD (8 dist) simulations. The conformational landscapes are generated using the same residue pairs as in Supplementary Figure 5.8.



Supplementary Figure 5.21: Violin plot shows distance distributions for 5 intracellular residue pair distances and 3 extracellular residue pair distances measured by Fowler et al. as observed from MD simulations of PepT$_{So}$ protein in micelle (yellow, right) and bilayer (blue, left) [111]. (A) Residue pair backbone distances as measured between C_α atom of residues. (B) Residue pair sidechain distances i.e. closest distance between any two non-hydrogen atoms in residue sidechains.



Supplementary Figure 5.22: Violin plot shows distance distributions for 17 intracellular residue pair distances and 7 extracellular residue pair distances measured by Kazmier et al. as observed from MD simulations of LeuT protein in micelle (purple, right) and bilayer (green, left) [257]. (A) Residue pair backbone distances as measured between C_α atom of residues. (B) Residue pair sidechain distances i.e. closest distance between any two non-hydrogen atoms in residue sidechains.



Supplementary Figure 5.23: Black dotted lines indicate experimental distributions obtained by tracking data from Fowler et al. and red lines indicate multiple Gaussian fitted to the experimental traces [111].

Chapter 6

Conclusion and Future Directions

The work in this dissertation describes the integration of MD simulations and DEER spectroscopy experiments. Designing experiments based on MD simulations is routine, some examples for such studies on membrane proteins are discussed here. Selvam et al. perform MD simulations of glucose transport through a rice SWEET2b transporter and validate their results using site-directed mutagenesis experiments [278]. Researchers identified seven high contact residues with glucose in the transport tunnel, five of which revealed significant effects in glucose transport from experiments. In another study that used adaptive sampling based simulation strategies, followed by MSM bases analysis of human serotonin transporter (hSERT) researchers identified a novel sodium binding site which was previously not known for NSS family transporters [275]. Two glutamic acid residues that form this binding site were then mutated to alanine, validating the loss of serotonin transport experimentally.

Using simulations as a means to validate and mechanistically elucidate observations from experiments is also routine. For example, Adhikary et al. study LeuT protein constituted inside phospholipid bilayer (POPC:POPG in 3:2 ratio) nanodiscs using hydrogen-deuterium exchange coupled with mass spectrometry (HDX-MS) [164]. Both HDX-MS experiments and lipid bilayer simulations show significant deuteration differences in the similar regions of the protein. Through residue-level protection factors calculation from short MD simulation trajectories, researchers were able to narrow down the segment responsible for the measured changes from peptides to individual residues. Another study combined HDX-MS experiments and MD simulations to study the effect of substrate and inhibitor binding to an MFS family xylose transporter, XylE [295].

In our work, we go a step further and show as a proof of concept that MD simulations and experiments can be used together to obtain full understanding of a proteins dynamics. Our method *OptimalProbes* is one of the first examples where MD simulations are used to predict spectroscopy experiments in a systematic manner. Moreover, as we demonstrated in Chapter 3 the method can also be used in an iterative manner, where simulations are succeeded by experiments, followed by updated predictions from experiments. Hays et al. use a similar methodology rooted in mutual information, instead of GMRQ as the objective function,

to predict residue-pairs for DEER experiments in an iterative manner [183]. Their method picks residue-pairs that minimize the mutual information among residue-pair distance distributions so that the final set of residue-pairs provide orthogonal information in terms of the conformational dynamics. In our method, *OptimalProbes* this condition is inherent in the definition of GMRQ which we aim to maximize for our top predictions of choice of residue-pairs for spectroscopy experiments.

The methods discussed in Chapters 2 and 3 are extremely general and may be used for mostly any experimental technique that yields residue-pair dynamics data. Our computational tool *OptimalProbes* is experimental technique agnostic and can predict optimal choice of residue-pairs for a diverse range of experimental techniques and a variety of proteins. Spectroscopy techniques such as DEER have also been applied on nucleic acids DNA and RNA using a nitroxide spin label 3-iodomethyl-1-oxy-2,2,5,5-tetramethylpyrroline that can be attached to a phosphorothioate group on a nucleotide [296] or a Ç (C-spin) spin label which is less flexible, is a 2′-deoxycytidine analogue, and base pairs with guanine [297, 298]. With improved force-fields [299–303], nucleic acid MD simulations are becoming increasingly accurate and amenable. As computational simulations become more tractable, they will be indispensable in the study of protein-complex structures [304], and the association and dissociation of protein-protein systems [305–307] or protein-nucleic acids systems [308, 309]. *OptimalProbes* thus has a potential to predict ideal choice of probe positions for the study of conformational heterogeneity in nucleic acids.

MD simulations also provide an attractive means to compare change in protein conformational dynamics due to post-translational modifications [51, 294, 310] and macromolecular crowding in-cell that resemble in vivo conditions [311, 312]. The versatility of DEER spectroscopy experiments as well as other spectroscopy experiments is also well suited to comparatively study the structural dynamics of proteins in different states. For example, Shi et al. used DEER experiments on two sets of residue-pairs to compare the conformations of an F/G loop in CYP119 protein in an apo state and when bound to a substrate and separately to two inhibitor molecules [313]. The F/G loop exhibits an open conformation in apo protein and slightly closed form when bound to inhibitor molecules, but is disordered when bound to substrate lauric acid. The disorder is characterized by DEER experiments since the relevant distance distribution ranges 2.2 nm without a predominant peak. Guin et al. use FRET labels mCherry and mEGFP to compare the binding of heat shock protein Hsp70 and heat shock cognate protein Hsc70 to PGK protein in living cells [314, 315]. While this study used fluorescent proteins as FRET probes, we have shown that Trp-Tyr quenching fluorescence experiments can be used to identify protein folding and unfolding such as in the presence of denaturants [316]. Hence, we envision that scientific studies that compare systems in different equilibrium conditions and systems involving multiple proteins/nucleic acids entities will benefit significantly from using protocols that

use MD simulations to design experimental studies.

A major requirement of the *OptimalProbes* methodology is long timescale MD simulations that are able to sample rare conformational change events reversibly. As system sizes get larger, enhanced MD simulation strategies such as accelerated MD, metadynamics, replica exchange, and umbrella sampling and coarse-grained MD simulations where multiple atoms are grouped into a single bead, are better suited to address scientific inquiries into conformational dynamics. In order to utilize these simulations, the core idea behind *OptimalProbes* would still be valid, however a novel objective function is needed to assign order to all the possible residue-pair choices for experiments. Wu et al. have proposed multiensemble Markov models using transition-based reweighting analysis method (TRAM) as an approach to combined unbiased and biased MD simulations expanding the power of the MSM framework [317]. TRAM is included in pyEMMA [70]. Since TRAM can estimate the thermodynamics and kinetics of the conformational transitions, a way to design experiments experiments using TRAM would be to (1) perform biased MD simulations to accelerate sampling of rare events, (2) use biased simulations as starting structures to perform unbiased adaptively sampled MD simulations, (3) build TRAM based multiensemble Markov models, and finally (4) score multiensemble Markov models.

Finally, one of the challenges in the field of molecular simulations is sharing simulation data. While there is a need for analysis and method development in the field of biophysical dynamics to use MD simulations and biophysical experiments in conjunction with each other, these techniques cannot be used by experimentalists if simulation datasets are not available in the public domain. Unavailability of datasets also leads to duplicate simulations which is inefficient utilization of computational resources for the entire scientific community. At the same time, method developments are bounded when experimental datasets are not available. In chapter 5 we describe a cumbersome manner in which we extracted DEER spectroscopy distance distributions from published work where the authors did not provide access to the datasets.

Major hurdles for sharing datasets is the large size of such data and lack of standardized protocols and formats to share data. In order to overcome this issue, we envision a platform based on cloud-based services which can store datasets and serve as a work engine for researchers to run analysis such as *OptimalProbes* on shared simulation datasets. There are existing initiatives to share biomolecular simulation data but they serve only as centralized deposit locations. A common platform to share data, run analysis, and store analysis would allow users to (1) request data when necessary, (2) analyze existing or new simulation data without demanding local need for large computational resources, (3) update protein dynamics models as they collect new simulation or experimental data, and (4) visualize and work via an intuitive graphical user interface for non-intensive tasks. Such a platform would democratize the study of dynamics through computational

and experimental methods, both of which provide a wealth of information and insights into biomolecular structural heterogeneity and biomolecular function.

References

[1] Henzler-Wildman K, Kern D. Dynamic Personalities of Proteins. Nature. 2007;450(7172):964–972.

[2] Chaudhuri TK, Paul S. Protein-Misfolding Diseases and Chaperone-Based Therapeutic Approaches. FEBS Journal. 2006;273(7):1331–1349.

[3] Mukherjee A, Morales-Scheihing D, Butler PC, Soto C. Type 2 Diabetes as a Protein Misfolding Disease. Trends in Molecular Medicine. 2015;21(7):439–449.

[4] Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, Karlsson A, Al-Lazikani B, Hersey A, Oprea TI, Overington JP. A Comprehensive Map of Molecular Drug Targets. Nature Reviews Drug Discovery. 2016;16(1):19–34.

[5] Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A Three-dimensional Model of the Myoglobin Molecule Obtained by X-ray Analysis. Nature. 1958;181(4610):662–666.

[6] Standfuss J. Membrane Protein Dynamics Studied by X-ray Lasers – or Why Only Time Will Tell. Current Opinion in Structural Biology. 2019;57:63–71.

[7] Goldie KN, Abeyrathne P, Kebbel F, Chami M, Ringler P, Stahlberg H. Cryo-electron Microscopy of Membrane Proteins. In: Methods in Molecular Biology. Humana Press; 2013. p. 325–341.

[8] Cheng Y. Membrane Protein Structural Biology in the Era of Single Particle Cryo-EM. Current Opinion in Structural Biology. 2018;52:58–63.

[9] Carlsson J, Coleman RG, Setola V, Irwin JJ, Fan H, Schlessinger A, Sali A, Roth BL, Shoichet BK. Ligand Discovery From a Dopamine D3 Receptor Homology Model and Crystal Structure. Nature Chemical Biology. 2011;7(11):769–778.

[10] Colas C, Grewer C, Otte NJ, Gameiro A, Albers T, Singh K, Shere H, Bonomi M, Holst J, Schlessinger A. Ligand Discovery for the Alanine-Serine-Cysteine Transporter (ASCT2, SLC1A5) From Homology Modeling and Virtual Screening. PLoS Computational Biology. 2015;11(10):e1004477.

[11] Ung PMU, Song W, Cheng L, Zhao X, Hu H, Chen L, Schlessinger A. Inhibitor Discovery for the Human GLUT1 From Homology Modeling and Virtual Screening. ACS Chemical Biology. 2016;11(7):1908–1916.

[12] Kim SK, Chen Y, Abrol R, Goddard WA, Guthrie B. Activation Mechanism of the G Protein-Coupled Sweet Receptor Heterodimer With Sweeteners and Allosteric Agonist. Proceedings of the National Academy of Sciences. 2017;114(10):2568–2573.

[13] Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. Scalable Molecular Dynamics with NAMD. Journal of Computational Chemistry. 2005;26(16):1781–1802.

[14] Case D, Babin V, Berryman J, Betz R, Cai Q, Cerutti D, Cheatham III T, Darden T, Duke R, Gohlke H, Goetz A, S G, N H, Janowski P, Kaus J, Kolossvry I, Kovalenko A, Lee T, LeGrand S, Luchko T, Luo R, Madej B, Merz K, Paesani F, Roe D, Roitberg A, Sagui C, Salomon-Ferrer R, Seabra

G, Simmerling C, Smith W, Swails J, Walker R, Wang J, Wolf R, Wu X, Kollman P. Amber 14. 2014;University of California.

[15] Hess B, Kutzner C, Van Der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. Journal of Chemical Theory and Computation. 2008;4(3):435–447.

[16] Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang LP, Simmonett AC, Harrigan MP, Stern CD, Wiewiora RP, Brooks BR, Pande VS. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. PLoS Computational Biology. 2017;13(7):e1005659.

[17] Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. Journal of Computational Chemistry. 1992;13(8):1011–1021.

[18] Schlitter J, Engels M, Krüger P. Targeted Molecular Dynamics: A New Approach for Searching Pathways of Conformational Transitions. Journal of Molecular Graphics. 1994;12(2):84–89.

[19] Izrailev S, Stepaniants S, Balsera M, Oono Y, Schulten K. Molecular Dynamics Study of Unbinding of the Avidin-Biotin Complex. Biophysical Journal. 1997;72(4):1568–1581.

[20] Wu X, Wang S. Self-Guided Molecular Dynamics Simulation for Efficient Conformational Search. The Journal of Physical Chemistry B. 1998;102(37):7238–7250.

[21] Sugita Y, Okamoto Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. Chemical Physics Letters. 1999;314(1-2):141–151.

[22] Laio A, Parrinello M. Escaping Free-Energy Minima. Proceedings of the National Academy of Sciences. 2002;99(20):12562–12566.

[23] Bolhuis PG, Chandler D, Dellago C, Geissler PL. Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. Annual Review of Physical Chemistry. 2002;53(1):291–318.

[24] Maragliano L, Vanden-Eijnden E. A Temperature Accelerated Method for Sampling Free Energy and Determining Reaction Pathways in Rare Events Simulations. Chemical Physics Letters. 2006;426(1-3):168–175.

[25] Hamelberg D, de Oliveira CAF, McCammon JA. Sampling of Slow Diffusive Conformational Transitions With Accelerated Molecular Dynamics. The Journal of Chemical Physics. 2007;127(15):155102.

[26] Darve E, Rodríguez-Gómez D, Pohorille A. Adaptive Biasing Force Method for Scalar and Vector Free Energy Calculations. The Journal of Chemical Physics. 2008;128(14):144120.

[27] Vanden-Eijnden E, Venturoli M. Revisiting the Finite Temperature String Method for the Calculation of Reaction Tubes and Free Energies. The Journal of Chemical Physics. 2009;130(19):194103.

[28] Miao Y, McCammon JA. Unconstrained Enhanced Sampling for Free Energy Calculations of Biomolecules: A Review. Molecular Simulation. 2016;42(13):1046–1055.

[29] Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3D Structure Computed from Evolutionary Sequence Variation. PLoS ONE. 2011;6(12):e28766.

[30] Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. Cell. 2012;149(7):1607–1621.

[31] Marks DS, Hopf TA, Sander C. Protein Structure Prediction From Sequence Variation. Nature Biotechnology. 2012;30(11):1072–1080.

[32] Ovchinnikov S, Kamisetty H, Baker D. Robust and Accurate Prediction of ResidueResidue Interactions Across Protein Interfaces Using Evolutionary Information. eLife. 2014;3:e02030.

[33] Hopf TA, Schärfe CP, Rodrigues JP, Green AG, Kohlbacher O, Sander C, Bonvin AM, Marks DS. Sequence Co-Evolution Gives 3D Contacts and Structures of Protein Complexes. eLife. 2014;3:e03430.

[34] Liu Y, Palmedo P, Ye Q, Berger B, Peng J. Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. Cell Systems. 2018;6(1):65–74.e3.

[35] Tang Y, Huang YJ, Hopf TA, Sander C, Marks DS, Montelione GT. Protein Structure Determination by Combining Sparse NMR Data With Evolutionary Couplings. Nature Methods. 2015;12(8):751–754.

[36] Mittal J, Best RB. Tackling Force-Field Bias in Protein Folding Simulations: Folding of Villin HP35 and Pin WW Domains in Explicit Water. Biophysical Journal. 2010;99(3):L26–L28.

[37] Piana S, Lindorff-Larsen K, Shaw DE. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? Biophysical Journal. 2011;100(9):L47–L49.

[38] Kulik HJ, Luehr N, Ufimtsev IS, Martinez TJ. Ab Initio Quantum Chemistry for Protein Structures. The Journal of Physical Chemistry B. 2012;116(41):12501–12509.

[39] Piana S, Klepeis JL, Shaw DE. Assessing the Accuracy of Physical Models Used in Protein-Folding Simulations: Quantitative Evidence From Long Molecular Dynamics Simulations. Current Opinion in Structural Biology. 2014;24:98–105.

[40] Shaw DE, Chao JC, Eastwood MP, Gagliardo J, Grossman JP, Ho CR, Lerardi DJ, Kolossváry I, Klepeis JL, Layman T, McLeavey C, Deneroff MM, Moraes MA, Mueller R, Priest EC, Shan Y, Spengler J, Theobald M, Towles B, Wang SC, Dror RO, Kuskin JS, Larson RH, Salmon JK, Young C, Batson B, Bowers KJ. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. Communications of the ACM. 2008;51(7):91.

[41] Shaw DE, Grossman JP, Bank JA, Batson B, Butts JA, Chao JC, Deneroff MM, Dror RO, Even A, Fenton CH, Forte A, Gagliardo J, Gill G, Greskamp B, Ho CR, Ierardi DJ, Iserovich L, Kuskin JS, Larson RH, Layman T, Lee LS, Lerer AK, Li C, Killebrew D, Mackenzie KM, Mok SYH, Moraes MA, Mueller R, Nociolo LJ, Peticolas JL, Quan T, Ramot D, Salmon JK, Scarpazza DP, Schafer UB, Siddique N, Snyder CW, Spengler J, Tang PTP, Theobald M, Toma H, Towles B, Vitale B, Wang SC, Young C. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE Press; 2014. p. 41–53.

[42] Shirts M, Pande VS. Screen Savers of the World Unite! Science. 2000;290(5498):1903–1904.

[43] Bode B, Butler M, Dunning T, Gropp W, Hoe-fler T, mei Hwu W, Kramer W. The Blue Waters Super-System for Super-Science. Contemporary HPC Architectures. In: Contemporary HPC Architectures. Sitka Publications; 2012. p. ISBN: 978–1–4665–6835–8.

[44] Allison JR. Using Simulation to Interpret Experimental Data in Terms of Protein Conformational Ensembles. Current Opinion in Structural Biology. 2017;43:79–87.

[45] Shukla D, Meng Y, Roux B, Pande VS. Activation Pathway of Src Kinase Reveals Intermediate States As Targets for Drug Design. Nature Communications. 2014;5:3397.

[46] Lapidus LJ, Acharya S, Schwantes CR, Wu L, Shukla D, King M, DeCamp SJ, Pande VS. Complex Pathways in Folding of Protein G Explored by Simulation and Experiment. Biophysical Journal. 2014;107(4):947–955.

[47] Kohlhoff KJ, Shukla D, Lawrenz M, Bowman GR, Konerding DE, Belov D, Altman RB, Pande VS. Cloud-Based Simulations on Google Exacycle Reveal Ligand Modulation of GPCR Activation Pathways. Nature Chemistry. 2014;6(1):15.

[48] Lawrenz M, Shukla D, Pande VS. Cloud Computing Approaches for Prediction of Ligand Binding Poses and Pathways. Scientific Reports. 2015;5(1).

[49] Feinberg EN, Farimani AB, Hernandez CX, Pande VS. Kinetic Machine Learning Unravels Ligand-Directed Conformational Change of μ Opioid Receptor. bioRxiv. 2017; p. 170886.

[50] Sultan MM, Denny RA, Unwalla R, Lovering F, Pande VS. Millisecond Dynamics of BTK Reveal Kinome-wide Conformational Plasticity Within the Apo Kinase Domain. Scientific Reports. 2017;7(1).

[51] Shukla S, Zhao C, Shukla D. Dewetting Controls Plant Hormone Perception and Initiation of Drought Resistance Signaling. Structure. 2019;27(4):692–702.e3.

[52] Pande VS, Beauchamp K, Bowman GR. Everything You Wanted to Know About Markov State Models but Were Afraid to Ask. Methods. 2010;52(1):99–105.

[53] Beauchamp KA, Ensign DL, Das R, Pande VS. Quantitative Comparison of Villin Headpiece Subdomain Simulations and Triplet-Triplet Energy Transfer Experiments. Proceedings of the National Academy of Sciences. 2011;108(31):12734–12739.

[54] Lane TJ, Bowman GR, Beauchamp K, Voelz VA, Pande VS. Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. Journal of the American Chemical Society. 2011;133(45):18413–18419.

[55] Shukla D, Hernández CX, Weber JK, Pande VS. Markov State Models Provide Insights into Dynamic Modulation of Protein Function. Accounts of Chemical Research. 2015;48(2):414–422.

[56] Jolliffe IT. Principal Component Analysis and Factor Analysis. In: Principal component analysis. Springer; 1986. p. 115–128.

[57] Abdi H, Williams LJ. Principal Component Analysis. WIREs Computational Statistics. 2010;2(4):433–459.

[58] Sharpee T, Rust NC, Bialek W. Maximally Informative Dimensions: Analyzing Neural Responses to Natural Signals. In: Advances in Neural Information Processing Systems; 2003. p. 277–284.

[59] Schwantes CR, Pande VS. Improvements in MArkov State Model Construction Reveal Many Non-native Interactions in the Folding of NTL9. Journal of Chemical Theory and Computation. 2013;9(4):2000–2009.

[60] Pérez-Hernández G, Paul F, Giorgino T, Fabritiis GD, Noé F. Identification of Slow Molecular Order Parameters for Markov Model Construction. The Journal of Chemical Physics. 2013;139(1):015102.

[61] Molgedey L, Schuster HG. Separation of a Mixture of Independent Signals Using Time Delayed Correlations. Physical Review Letters. 1994;72(23):3634–3637.

[62] Naritomi Y, Fuchigami S. Slow Dynamics in Protein Fluctuations Revealed by Time-Structure Based Independent Component Analysis: The Case of Domain Motions. The Journal of Chemical Physics. 2011;134(6):065101.

[63] Shukla D, Peck A, Pande VS. Conformational Heterogeneity of the Calmodulin Binding Interface. Nature Communications. 2016;7:10910.

[64] Schwantes CR, Shukla D, Pande VS. Markov State Models and tICA Reveal a Nonnative Folding Nucleus in Simulations of NuG2. Biophysical Journal. 2016;110(8):1716–1719.

[65] Mukherjee S, Pantelopulos GA, Voelz VA. Markov Models of the apo-MDM2 Lid Region Reveal Diffuse yet Two-state Binding Dynamics and Receptor Poses for Computational Docking. Scientific Reports. 2016;6(1).

[66] Razavi AM, Khelashvili G, Weinstein H. A Markov State-based Quantitative Kinetic Model of Sodium Release from the Dopamine Transporter. Scientific Reports. 2017;7:40076.

[67] Ryckbosch SM, Wender PA, Pande VS. Molecular Dynamics Simulations Reveal Ligand-controlled Positioning of a Peripheral Protein Complex in Membranes. Nature Communications. 2017;8(1).

[68] Abramyan AM, Stolzenberg S, Li Z, Loland CJ, Noé F, Shi L. The Isomeric Preference of an Atypical Dopamine Transporter Inhibitor Contributes to Its Selection of the Transporter Conformation. ACS Chemical Neuroscience. 2017;8(8):1735–1746.

[69] Beauchamp KA, Bowman GR, Lane TJ, Maibaum L, Haque IS, Pande VS. MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. Journal of Chemical Theory and Computation. 2011;7(10):3412–3419.

[70] Scherer MK, Trendelkamp-Schroer B, Paul F, Prez-Hernndez G, Hoffmann M, Plattner N, Wehmeyer C, Prinz JH, No F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. Journal of Chemical Theory and Computation. 2015;11:5525–5542.

[71] Sculley D. Web-Scale k-means Clustering. In: Proc. 19th WWW. ACM; 2010. p. 1177–1178.

[72] de Hoon MJL, Imoto S, Nolan J, Miyano S. Open Source Clustering Software. Bioinformatics. 2004;20(9):1453–1454.

[73] Gonzalez TF. Clustering to Minimize the Maximum Intercluster Distance. Theoretical Computer Science. 1985;38:293–306.

[74] Müllner D. Modern Hierarchical, Agglomerative Clustering Algorithms. arXiv. 2011;1109.2378.

[75] McGibbon RT, Pande VS. Efficient Maximum Likelihood Parameterization of Continuous-time Markov Processes. The Journal of Chemical Physics. 2015;143(3):034109.

[76] Trendelkamp-Schroer B, Noé F. Efficient Estimation of Rare-Event Kinetics. Physical Review X. 2016;6(1).

[77] Wu H, Prinz JH, Noé F. Projected Metastable Markov Processes and Their Estimation with Observable Operator Models. The Journal of Chemical Physics. 2015;143(14):144101.

[78] Noé F, Fischer S. Transition Networks for Modeling the Kinetics of Conformational Change in Macromolecules. Current Opinion in Structural Biology. 2008;18(2):154–162.

[79] Prinz JH, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera JD, Schütte C, Noé F. Markov Models of Molecular Kinetics: Generation and Validation. The Journal of Chemical Physics. 2011;134(17):174105.

[80] McGibbon RT, Husic BE, Pande VS. Identification of Simple Reaction Coordinates From Complex Dynamics. The Journal of Chemical Physics. 2017;146(4):044109.

[81] Adelman JL, Ghezzi C, Bisignano P, Loo DDF, Choe S, Abramson J, Rosenberg JM, Wright EM, Grabe M. Stochastic Steps in Secondary Active Sugar Transport. Proceedings of the National Academy of Sciences. 2016;113(27):E3960–E3966.

[82] Moffett AS, Bender KW, Huber SC, Shukla D. Molecular Dynamics Simulations Reveal the Conformational Dynamics of Arabidopsis thaliana BRI1 and BAK1 Receptor-Like Kinases. Journal of Biological Chemistry. 2017;292(30):12643–12652.

[83] Lane TJ, Shukla D, Beauchamp KA, Pande VS. To Milliseconds and Beyond: Challenges in the Simulation of Protein Folding. Current Opinion in Structural Biology. 2013;23(1):58–65.

[84] Noé F, Nüske F. A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems. Multiscale Modeling and Simulation. 2013;11(2):635–655.

[85] Nüske F, Keller BG, Pérez-Hernández G, Mey ASJS, Noé F. Variational Approach to Molecular Kinetics. Journal of Chemical Theory and Computation. 2014;10(4):1739–1752.

[86] McGibbon RT, Pande VS. Variational Cross-Validation of Slow Dynamical Modes in Molecular Kinetics. The Journal of Chemical Physics. 2015;142(12):124105.

[87] Husic BE, McGibbon RT, Sultan MM, Pande VS. Optimized Parameter Selection Reveals Trends in Markov State Models for Protein Folding. The Journal of Chemical Physics. 2016;145(19):194103.

[88] Mittal S, Shukla D. Predicting Optimal DEER Label Positions to Study Protein Conformational Heterogeneity. The Journal of Physical Chemistry B. 2017;121(42):9761–9770.

[89] Mittal S, Shukla D. Maximizing Kinetic Information Gain of Markov State Models for Optimal Design of Spectroscopy Experiments. The Journal of Physical Chemistry B. 2018;122(48):10793–10805.

[90] Selvam B, Mittal S, Shukla D. Free Energy Landscape of the Complete Transport Cycle in a Key Bacterial Transporter. ACS central science. 2018;4(9):1146–1154.

[91] Borbat PP, Mchaourab HS, Freed JH. Protein Structure Determination Using Long-Distance Constraints From Double-Quantum Coherence ESR: Study of T4 Lysozyme. Journal of the American Chemical Society. 2002;124(19):5304–5314.

[92] Jeschke G. DEER Distance Measurements on Proteins. Annual Review of Physical Chemistry. 2012;63(1):419–446.

[93] Hubbell WL, Mchaourab HS, Altenbach C, Lietzow MA. Watching Proteins Move Using Site-Directed Spin Labeling. Structure. 1996;4(7):779–783.

[94] Hubbell WL, López CJ, Altenbach C, Yang Z. Technological Advances in Site-Directed Spin Labeling of Proteins. Current Opinion in Structural Biology. 2013;23(5):725–733.

[95] Schmidt MJ, Borbas J, Drescher M, Summerer D. A Genetically Encoded Spin Label for Electron Paramagnetic Resonance Distance Measurements. Journal of the American Chemical Society. 2014;136(4):1238–1241.

[96] Schmidt MJ, Fedoseev A, Bücker D, Borbas J, Peter C, Drescher M, Summerer D. EPR Distance Measurements in Native Proteins With Genetically Encoded Spin Labels. ACS Chemical Biology. 2015;10(12):2764–2771.

[97] Alexander N, Al-Mestarihi A, Bortolus M, Mchaourab H, Meiler J. *De novo* high-resolution protein structure determination from sparse spin-labeling EPR data. Structure. 2008;16(2):181–195.

[98] Hirst SJ, Alexander N, Mchaourab HS, Meiler J. RosettaEPR: An Integrated Tool for Protein Structure Determination From Sparse EPR Data. Journal of Structural Biology. 2011;173(3):506–514.

[99] Roux B, Weare J. On the Statistical Equivalence of Restrained-Ensemble Simulations With the Maximum Entropy Method. The Journal of Chemical Physics. 2013;138(8):084107.

[100] Islam SM, Stein RA, Mchaourab HS, Roux B. Structural Refinement From Restrained-Ensemble Simulations Based on EPR/DEER Data: Application to T4 Lysozyme. The Journal of Physical Chemistry B. 2013;117(17):4740–4754.

[101] Islam SM, Roux B. Simulating the Distance Distribution Between Spin-Labels Attached to Proteins. The Journal of Physical Chemistry B. 2015;119(10):3901–3911.

[102] Liu Z, Casey TM, Blackburn ME, Huang X, Pham L, de Vera IMS, Carter JD, Kear-Scott JL, Veloro AM, Galiano L, Fanucci GE. Pulsed EPR Characterization of HIV-1 Protease Conformational Sampling and Inhibitor-Induced Population Shifts. Phys Chem Chem Phys. 2016;18(8):5819–5831.

[103] Casey TM, Liu Z, Esquiaqui JM, Pirman NL, Milshteyn E, Fanucci GE. Continuous Wave W- and D-Band EPR Spectroscopy Offer "Sweet-Spots" for Characterizing Conformational Changes and Dynamics in Intrinsically Disordered Proteins. Biochem Biophys Res Commun. 2014;450(1):723–728.

[104] Lai AL, Clerico EM, Blackburn ME, Patel NA, Robinson CV, Borbat PP, Freed JH, Gierasch LM. Key Features of an Hsp70 Chaperone Allosteric Landscape Revealed by Ion-Mobility Native Mass Spectrometry and Double Electron-Electron Resonance. Journal of Biological Chemistry. 2017;292(21):8773–8785.

[105] Esquiaqui JM, Sherman EM, Ye JD, Fanucci GE. Conformational Flexibility and Dynamics of the Internal Loop and Helical Regions of the kink–Turn Motif in the Glycine Riboswitch by Site-Directed Spin-Labeling. Biochemistry. 2016;55(31):4295–4305.

[106] Mchaourab HS, Steed PR, Kazmier K. Toward the Fourth Dimension of Membrane Protein Structure: Insight into Dynamics from Spin-Labeling EPR Spectroscopy. Structure. 2011;19(11):1549–1561.

[107] Dong J. Structural Basis of Energy Transduction in the Transport Cycle of MsbA. Science. 2005;308(5724):1023–1028.

[108] Stelzl LS, Fowler PW, Sansom MS, Beckstein O. Flexible Gates Generate Occluded Intermediates in the Transport Cycle of LacY. Journal of Molecular Biology. 2014;426(3):735–751.

[109] Altenbach C, Kusnetzow AK, Ernst OP, Hofmann KP, Hubbell WL. High-Resolution Distance Mapping in Rhodopsin Reveals the Pattern of Helix Movement Due to Activation. Proceedings of the National Academy of Sciences. 2008;105(21):7439–7444.

[110] Manglik A, Kim TH, Masureel M, Altenbach C, Yang Z, Hilger D, Lerch MT, Kobilka TS, Thian FS, Hubbell WL, Prosser RS, Kobilka BK. Structural Insights into the Dynamic Process of β_2-Adrenergic Receptor Signaling. Cell. 2015;161(5):1101–1111.

[111] Fowler PW, Orwick-Rydmark M, Radestock S, Solcan N, Dijkman PM, Lyons JA, Kwok J, Caffrey M, Watts A, Forrest LR, Newstead S. Gating Topology of the Proton-Coupled Oligopeptide Symporters. Structure. 2015;23(2):290–301.

[112] Mishra S, Verhalen B, Stein RA, Wen PC, Tajkhorshid E, Mchaourab HS. Conformational Dynamics of the Nucleotide Binding Domains and the Power Stroke of a Heterodimeric ABC Transporter. eLife. 2014;3.

[113] Polyhach Y, Bordignon E, Jeschke G. Rotamer Libraries of Spin Labelled Cysteines for Protein Studies. Physical Chemistry Chemical Physics. 2011;13(6):2356–2366.

[114] Kubelka J, Chiu TK, Davies DR, Eaton WA, Hofrichter J. Sub-Microsecond Protein Folding. Journal of Molecular Biology. 2006;359(3):546–553.

[115] Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W. Atomic-Level Characterization of the Structural Dynamics of Proteins. Science. 2010;330(6002):341–346.

[116] Moradi M, Tajkhorshid E. Mechanistic Picture for Conformational Transition of a Membrane Transporter at Atomic Resolution. Proceedings of the National Academy of Sciences. 2013;110(47):18916–18921.

[117] Hulse RE, Sachleben JR, Wen PC, Moradi M, Tajkhorshid E, Perozo E. Conformational Dynamics at the Inner Gate of KcsA During Activation. Biochemistry. 2014;53(16):2557–2559.

[118] Dror RO, Mildorf TJ, Hilger D, Manglik A, Borhani DW, Arlow DH, Philippsen A, Villanueva N, Yang Z, Lerch MT, Hubbell WL, Kobilka BK, Sunahara RK, Shaw DE. Structural Basis for Nucleotide Exchange in Heterotrimeric G Proteins. Science. 2015;348(6241):1361–1365.

[119] Shukla D, Lawrenz M, Pande VS. Elucidating Ligand-Modulated Conformational Landscape of GPCRs Using Cloud-Computing Approaches. Methods Enzymology. 2015;557:551–572.

[120] Huang YM, Miao Y, Munguia J, Lin L, Nizet V, McCammon JA. Molecular Dynamic Study of MlaC Protein in Gram-Negative Bacteria: Conformational Flexibility, Solvent Effect and Protein-Phospholipid Binding. Protein Sci. 2016;25(8):1430–1437.

[121] Needham SR, Roberts SK, Arkhipov A, Mysore VP, Tynan CJ, Zanetti-Domingues LC, Kim ET, Losasso V, Korovesis D, Hirsch M, Rolfe DJ, Clarke DT, Winn MD, Lajevardipour A, Clayton AHA, Pike LJ, Perani M, Parker PJ, Shan Y, Shaw DE, Martin-Fernandez ML. EGFR Oligomerization Organizes Kinase-Active Dimers Into Competent Signalling Platforms. Nature Communications. 2016;7:13307.

[122] Vanatta DK, Shukla D, Lawrenz M, Pande VS. A Network of Molecular Switches Controls the Activation of the Two-Component Response Regulator NtrC. Nature Communications. 2015;6:7283.

[123] Georgieva ER, Ramlall TF, Borbat PP, Freed JH, Eliezer D. Membrane-Bound Alpha-Synuclein Forms an Extended Helix: Long-Distance Pulsed ESR Measurements Using Vesicles, Bicelles, and Rodlike Micelles. Journal of the American Chemical Society. 2008;130(39):12856–12857.

[124] Georgieva ER, Roy AS, Grigoryants VM, Borbat PP, Earle KA, Scholes CP, Freed JH. Effect of Freezing Conditions on Distances and Their Distributions Derived From Double Electron Electron Resonance (DEER): A Study of Doubly-Spin-Labeled T4 Lysozyme. Journal of Magnetic Resonance. 2012;216:69–77.

[125] McGibbon RT, Hernández CX, Harrigan MP, Kearnes S, Sultan MM, Jastrzebski S, Husic BE, Pande VS. Osprey: Hyperparameter Optimization for Machine Learning. Journal of Open Source Software. 2016;1(5).

[126] Winston PH. Artificial Intelligence. Addison-Wesley Pub Co; 1992.

[127] Mitchell M. An Introduction to Genetic Algorithms. MIT Press: Cambridge, MA, USA; 1998.

[128] Ascher D, Dubois PF, Hinsen K, Hugunin J, Oliphant T. Numerical Python; 1999.

[129] McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, Schwantes CR, Wang LP, Lane TJ, Pande VS. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophysical Journal. 2015;109(8):1528–1532.

[130] Dror RO, Arlow DH, Maragakis P, Mildorf TJ, Pan AC, Xu H, Borhani DW, Shaw DE. Activation Mechanism of the β_2-Adrenergic Receptor. Proceedings of the National Academy of Sciences. 2011;108(46):18684–18689.

[131] Rasmussen SGF, Choi HJ, Fung JJ, Pardon E, Casarosa P, Chae PS, DeVree BT, Rosenbaum DM, Thian FS, Kobilka TS, Schnapp A, Konetzki I, Sunahara RK, Gellman SH, Pautsch A, Steyaert J, Weis WI, Kobilka BK. Structure of a Nanobody-Stabilized Active State of the β2 Adrenoceptor. Nature. 2011;469(7329):175–180.

[132] Kuboniwa H, Tjandra N, Grzesiek S, Ren H, Klee CB, Bax A. Solution Structure of Calcium-Free Calmodulin. Nature Structural Biology. 1995;2(9):768–776.

[133] Chattopadhyaya R, Meador WE, Means AR, Quiocho FA. Calmodulin Structure Refined at 1.7 Å Resolution. Journal of Molecular Biology. 1992;228(4):1177–1192.

[134] Newstead S. Molecular Insights Into Proton Coupled Peptide Transport in the PTR Family of Oligopeptide Transporters. Biochimica et Biophysica Acta. 2015;1850(3):488–499.

[135] Quistgaard EM, Löw C, Guettou F, Nordlund P. Understanding Transport by the Major Facilitator Superfamily (MFS): Structures Pave the Way. Nature Reviews Molecular Cell Biology. 2016;17(2):123–132.

[136] Kazmier K, Alexander NS, Meiler J, Mchaourab HS. Algorithm for Selection of Optimized EPR Distance Restraints for *de Novo* Protein Structure Determination. Journal of Structural Biology. 2011;173(3):549–557.

[137] Ding B, Hilaire MR, Gai F. Infrared and Fluorescence Assessment of Protein Dynamics: From Folding to Function. The Journal of Physical Chemistry B. 2016;120(23):5103–5113.

[138] Davis CM, Gruebele M. Labeling for Quantitative Comparison of Imaging Measurements in Vitro and in Cells. Biochemistry. 2018;57(13):1929–1938.

[139] Bieri O, Wirz J, Hellrung B, Schutkowski M, Drewello M, Kiefhaber T. The Speed Limit for Protein Folding Measured by Triplet-Triplet Energy Transfer. Proceedings of the National Academy of Sciences. 1999;96(17):9597–9601.

[140] Chen H, Ahsan SS, Santiago-Berrios MB, Abruña HD, Webb WW. Mechanisms of Quenching of Alexa Fluorophores by Natural Amino Acids. Journal of the American Chemical Society. 2010;132(21):7244–7245.

[141] Chen Y, Barkley MD. Toward Understanding Tryptophan Fluorescence in Proteins. Biochemistry. 1998;37(28):9976–9982.

[142] Toseland CP. Fluorescent Labeling and Modification of Proteins. Journal of Chemical Biology. 2013;6(3):85–95.

[143] Cooke JA, Brown LJ. Distance Measurements by Continuous Wave EPR Spectroscopy to Monitor Protein Folding. In: Methods in Molecular Biology. Humana Press; 2011. p. 73–96.

[144] Altenbach C, Cai K, Klein-Seetharaman J, Khorana HG, Hubbell WL. Structure and Function in Rhodopsin: Mapping Light-Dependent Changes in Distance between Residue 65 in Helix TM1 and Residues in the Sequence 306-319 at the Cytoplasmic End of Helix TM7 and in Helix H8. Biochemistry. 2001;40(51):15483–15492.

[145] Ma H, Gruebele M. Kinetics Are Probe-Dependent During Downhill Folding of an Engineered λ_{6-85} Protein. Proceedings of the National Academy of Sciences. 2005;102(7):2283–2287.

[146] Prigozhin MB, Chao SH, Sukenik S, Pogorelov TV, Gruebele M. Mapping Fast Protein Folding With Multiple-Site Fluorescent Probes. Proceedings of the National Academy of Sciences. 2015;112(26):7966–7971.

[147] Cha A, Snyder GE, Selvin PR, Bezanilla F. Atomic Scale Movement of the Voltage-Sensing Region in a Potassium Channel Measured via Spectroscopy. Nature. 1999;402(6763):809–813.

[148] Kapanidis AN, Ebright YW, Ludescher RD, Chan S, Ebright RH. Mean DNA Bend Angle and Distribution of DNA Bend Angles in the CAP-DNA Complex in Solution. Journal of Molecular Biology. 2001;312(3):453–468.

[149] Selvin PR. Principles and Biophysical Applications of Lanthanide-Based Probes. Annual Review of Biophysics and Biomolecular Structure. 2002;31(1):275–302.

[150] Rahmeh R, Damian M, Cottet M, Orcel H, Mendre C, Durroux T, Sharma KS, Durand G, Pucci B, Trinquet E, Zwier JM, Deupi X, Bron P, Baneres JL, Mouillac B, Granier S. Structural Insights Into Biased G Protein-Coupled Receptor Signaling Revealed by Fluorescence Spectroscopy. Proceedings of the National Academy of Sciences. 2012;109(17):6733–6738.

[151] Zoghbi ME, Altenberg GA. Luminescence Resonance Energy Transfer Spectroscopy of ATP-Binding Cassette Proteins. BBA-Biomembranes. 2018;1860(4):854–867.

[152] Fierz B, Reiner A, Kiefhaber T. Local Conformational Dynamics in α-Helices Measured by Fast Triplet Transfer. Proceedings of the National Academy of Sciences. 2009;106(4):1057–1062.

[153] Reiner A, Henklein P, Kiefhaber T. An Unlocking/Relocking Barrier in Conformational Fluctuations of Villin Headpiece Subdomain. Proceedings of the National Academy of Sciences. 2010;107(11):4955–4960.

[154] Husic BE, Pande VS. Markov State Models: From an Art to a Science. Journal of the American Chemical Society. 2018;140(7):2386–2396.

[155] Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How Fast-Folding Proteins Fold. Science. 2011;334(6055):517–520.

[156] Beauchamp KA, McGibbon R, Lin YS, Pande VS. Simple Few-State Models Reveal Hidden Complexity in Protein Folding. Proceedings of the National Academy of Sciences. 2012;109(44):17807–17813.

[157] Karplus M, Kuriyan J. Molecular Dynamics and Protein Function. Proceedings of the National Academy of Sciences. 2005;102(19):6679–6685.

[158] Moffett AS, Shukla D. Using Molecular Simulation to Explore the Nanoscale Dynamics of the Plant Kinome. Biochemical Journal. 2018;475(5):905–921.

[159] Chen S, Zhao Y, Wang Y, Shekhar M, Tajkhorshid E, Gouaux E. Activation and Desensitization Mechanism of AMPA Receptor-TARP Complex by Cryo-EM. Cell. 2017;170(6):1234–1246.e14.

[160] Gardino AK, Villali J, Kivenson A, Lei M, Liu CF, Steindel P, Eisenmesser EZ, Labeikovsky W, Wolf-Watz M, Clarkson MW, Kern D. Transient Non-native Hydrogen Bonds Promote Activation of a Signaling Protein. Cell. 2009;139(6):1109–1118.

[161] Pontiggia F, Pachov DV, Clarkson MW, Villali J, Hagan MF, Pande VS, Kern D. Free Energy Landscape of Activation in a Signalling Protein at Atomic Resolution. Nature Communications. 2015;6:7284.

[162] Jones CP, Cantara WA, Olson ED, Musier-Forsyth K. Small-angle X-ray scattering-derived structure of the HIV-1 5′ UTR reveals 3D tRNA mimicry. Proceedings of the National Academy of Sciences. 2014;111(9):3395–3400.

[163] Gregorio GG, Masureel M, Hilger D, Terry DS, Juette M, Zhao H, Zhou Z, Perez-Aguilar JM, Hauge M, Mathiasen S, Javitch JA, Weinstein H, Kobilka BK, Blanchard SC. Single-Molecule Analysis of Ligand Efficacy in β_2AR-G-Protein Activation. Nature. 2017;547(7661):68–73.

[164] Adhikary S, Deredge DJ, Nagarajan A, Forrest LR, Wintrode PL, Singh SK. Conformational Dynamics of a Neurotransmitter:Sodium Symporter in a Lipid Bilayer. Proceedings of the National Academy of Sciences. 2017;114(10):E1786–E1795.

[165] van Gunsteren WF, Daura X, Hansen N, Mark AE, Oostenbrink C, Riniker S, Smith LJ. Validation of Molecular Simulation: An Overview of Issues. Angewandte Chemie International Edition. 2017;57(4):884–902.

[166] Feng J, Shukla D. Characterizing Conformational Dynamics of Proteins Using Evolutionary Couplings. The Journal of Physical Chemistry B. 2018;122(3):1017–1025.

[167] Alexiev U, Farrens DL. Fluorescence Spectroscopy of Rhodopsins: Insights and Approaches. BBA-Bioenergetics. 2014;1837(5):694–709.

[168] Dolino DM, Ramaswamy SS, Jayaraman V. Luminescence Resonance Energy Transfer to Study Conformational Changes in Membrane Proteins Expressed in Mammalian Cells. Journal of Visualized Experiments. 2014;91:e51895.

[169] Klose D, Klare JP, Grohmann D, Kay CWM, Werner F, Steinhoff HJ. Simulation vs. Reality: A Comparison of In Silico Distance Predictions with DEER and FRET Measurements. PLoS ONE. 2012;7(6):e39492.

[170] Bowman GR, Pande VS. Protein Folded States Are Kinetic Hubs. Proceedings of the National Academy of Sciences. 2010;107(24):10890–10895.

[171] Harrigan MP, Sultan MM, Hernández CX, Husic BE, Eastman P, Schwantes CR, Beauchamp KA, McGibbon RT, Pande VS. MSMBuilder: Statistical Models for Biomolecular Dynamics. Biophysical Journal. 2017;112(1):10–15.

[172] Humphrey W, Dalke A, Schulten K. VMD: Visual Molecular Dynamics. Journal of Molecular Graphics. 1996;14(1):33–38.

[173] Glaenzer J, Peter MF, Hagelueken G. Studying Structure and Function of Membrane Proteins With PELDOR/DEER Spectroscopy - A Crystallographers' Perspective. Methods. 2018;.

[174] Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum Allowed Solvent Accessibilites of Residues in Proteins. PLoS ONE. 2013;8(11):e80635.

[175] Prigozhin MB, Gruebele M. Microsecond Folding Experiments and Simulations: A Match Is Made. Physical Chemistry Chemical Physics. 2013;15(10):3372–3388.

[176] Boomsma W, Ferkinghoff-Borg J, Lindorff-Larsen K. Combining Experiments and Simulations Using the Maximum Entropy Principle. PLoS Comput Biol. 2014;10(2):e1003406.

[177] Beauchamp KA, Pande VS, Das R. Bayesian Energy Landscape Tilting: Towards Concordant Models of Molecular Ensembles. Biophysical Journal. 2014;106(6):1381–1390.

[178] Olsson S, Wu H, Paul F, Clementi C, Noé F. Combining Experimental and Simulation Data of Molecular Processes via Augmented Markov Models. Proceedings of the National Academy of Sciences. 2017;114(31):8265–8270.

[179] Matsunaga Y, Sugita Y. Linking Time-Series of Single-Molecule Experiments With Molecular Dynamics Simulations by Machine Learning. eLife. 2018;7:e32668.

[180] Roux B, Islam SM. Restrained-Ensemble Molecular Dynamics Simulations Based on Distance Histograms from Double Electron–Electron Resonance Spectroscopy. The Journal of Physical Chemistry B. 2013;117(17):4733–4739.

[181] Hustedt EJ, Marinelli F, Stein RA, Faraldo-Gómez JD, Mchaourab HS. Confidence Analysis of DEER Data and Its Structural Interpretation with Ensemble-Biased Metadynamics. Biophysical Journal. 2018;115(7):1200–1216.

[182] Moglich A, Joder K, Kiefhaber T. End-to-End Distance Distributions and Intrachain Diffusion Constants in Unfolded Polypeptide Chains Indicate Intramolecular Hydrogen Bond Formation. Proceedings of the National Academy of Sciences. 2006;103(33):12394–12399.

[183] Hays JM, Kieber MK, Li JZ, Han JI, Columbus L, Kasson PM. Refinement of Highly Flexible Protein Structures using Simulation-Guided Spectroscopy. Angewandte Chemie International Edition. 2018;57(52):17110–17114.

[184] Watson MD, Peran I, Zou J, Bilsel O, Raleigh DP. Selenomethionine Quenching of Tryptophan Fluorescence Provides a Simple Probe of Protein Structure. Biochemistry. 2017;56(8):1085–1094.

[185] Frembgen-Kesner T, Elcock AH. Computer Simulations of the Bacterial Cytoplasm. Biophys Rev. 2013;5(2):109–119.

[186] Yu I, Mori T, Ando T, Harada R, Jung J, Sugita Y, Feig M. Biomolecular Interactions Modulate Macromolecular Structure and Dynamics in Atomistic Model of a Bacterial Cytoplasm. eLife. 2016;5:e19274.

[187] Guzman I, Gruebele M. Protein Folding Dynamics in the Cell. The Journal of Physical Chemistry B. 2014;118(29):8459–8470.

[188] Rodrigues CH, Pires DE, Ascher DB. DynaMut: Predicting the Impact of Mutations on Protein Conformation, Flexibility and Stability. Nucleic Acids Research. 2018;46(W1):W350–W355.

[189] Pandurangan AP, Ochoa-Montaño B, Ascher DB, Blundell TL. SDM: A Server for Predicting Effects of Mutations on Protein Stability. Nucleic Acids Research. 2017;45(W1):W229–W235.

[190] Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: Predicting Stability Changes Upon Mutation From the Protein Sequence or Structure. Nucleic Acids Research. 2005;33(Web Server):W306–W310.

[191] Bonomi M, Pellarin R, Vendruscolo M. Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy. Biophysical Journal. 2018;114(7):1604–1613.

[192] Kasahara K, Fukuda I, Nakamura H. A Novel Approach of Dynamic Cross Correlation Analysis on Molecular Dynamics Simulations and Its Application to Ets1 Dimer–DNA Complex. PLoS ONE. 2014;9(11):e112419.

[193] Notredame C, Higgins DG, Heringa J. T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. Journal of Molecular Biology. 2000;302(1):205–217.

[194] Okan OB, Atilgan AR, Atilgan C. Nanosecond Motions in Proteins Impose Bounds on the Timescale Distributions of Local Dynamics. Biophysical Journal. 2009;97(7):2080–2088.

[195] Pao SS, Paulsen IT, Saier MH. Major Facilitator Superfamily. Microbiology and Molecular Biology Reviews. 1998;62(1):1–34.

[196] Kaback HR, Dunten R, Frillingos S, Venkatesan P, Kwaw I, Zhang W, Ermolova N. Site-Directed Alkylation and the Alternating Access Model For LacY. Proceedings of the National Academy of Sciences. 2007;104(2):491–494.

[197] Law CJ, Maloney PC, Wang DN. Ins and Outs of Major Facilitator Superfamily Antiporters. Annual Review of Microbiology. 2008;62:289.

[198] Coincon M, Uzdavinys P, Nji E, Dotson DL, Winkelmann I, Abdul-Hussein S, Cameron AD, Beckstein O, Drew D. Crystal Structures Reveal the Molecular Basis of Ion Translocation in Sodium/proton Antiporters. Nature Structural & Molecular Biology. 2016; p. 248–255.

[199] Ryan RM, Vandenberg RJ. Elevating the Alternating-Access Model. Nature Structural & Molecular Biology. 2016;23(3):187–189.

[200] Colas C, Ung PMU, Schlessinger A. SLC Transporters: Structure, Function, and Drug Discovery. MedChemComm. 2016;7(6):1069–1081.

[201] Daniel H, Kottra G. The Proton Oligopeptide Cotransporter Family SLC15 in Physiology and Pharmacology. Pflügers Archiv. 2004;447(5):610–618.

[202] Nielsen C, Brodin B. Di/tri-Peptide Transporters as Drug Delivery Targets: Regulation of Transport Under Physiological and Patho-Physiological Conditions. Current Drug Targets. 2003;4(5):373–388.

[203] Newstead S, Drew D, Cameron AD, Postis VLG, Xia X, Fowler PW, Ingram JC, Carpenter EP, Sansom MSP, McPherson MJ, Baldwin SA, Iwata S. Crystal Structure of a Prokaryotic Homologue of the Mammalian Oligopeptide-Proton Symporters, PepT1 and PepT2. EMBO Journal. 2010;30(2):417–426.

[204] Lyons JA, Parker JL, Solcan N, Brinth A, Li D, Shah ST, Caffrey M, Newstead S. Structural Basis for Polyspecificity in the POT Family of Proton-Coupled Oligopeptide Transporters. EMBO Reports. 2014;15(8):886–893.

[205] Solcan N, Kwok J, Fowler PW, Cameron AD, Drew D, Iwata S, Newstead S. Alternating Access Mechanism in the POT Family of Oligopeptide Transporters. EMBO Journal. 2012;31(16):3411–3421.

[206] Molledo MM, Quistgaard EM, Flayhan A, Pieprzyk J, Löw C. Multispecific Substrate Recognition in a Proton-Dependent Oligopeptide Transporter. Structure. 2018;26(3):467–476.e4.

[207] Doki S, Kato HE, Solcan N, Iwaki M, Koyama M, Hattori M, Iwase N, Tsukazaki T, Sugita Y, Kandori H, Newstead S, Ishitani R, Nureki O. Structural Basis for Dynamic Mechanism of Proton-Coupled Symport by the Peptide Transporter POT. Proceedings of the National Academy of Sciences. 2013;110(28):11343–11348.

[208] Guettou F, Quistgaard EM, Trésaugues L, Moberg P, Jegerschöld C, Zhu L, Jong AJO, Nordlund P, Löw C. Structural Insights into Substrate Recognition in Proton-Dependent Oligopeptide Transporters. EMBO Reports. 2013;14(9):804–810.

[209] Guettou F, Quistgaard EM, Raba M, Moberg P, Löw C, Nordlund P. Selectivity Mechanism of a Bacterial Homolog of the Human Drug-Peptide Transporters PepT1 and PepT2. Nature Structural & Molecular Biology. 2014;21(8):728–731.

[210] Boggavarapu R, Jeckelmann JM, Harder D, Ucurum Z, Fotiadis D. Role of Electrostatic Interactions for Ligand Recognition and Specificity of Peptide Transporters. BMC Biology. 2015;13(1):1.

[211] Parker JL, Li C, Brinth A, Wang Z, Vogeley L, Solcan N, Ledderboge-Vucinic G, Swanson JMJ, Caffrey M, Voth GA, Newstead S. Proton Movement and Coupling in the POT Family of Peptide Transporters. Proceedings of the National Academy of Sciences. 2017;114(50):13182–13187.

[212] Vergara-Jaque A, Fenollar-Ferrer C, Kaufmann D, Forrest LR. Repeat-Swap Homology Modeling of Secondary Active Transporters: Updated Protocol and Prediction of Elevator-Type Mechanisms. Frontiers in Pharmacology. 2015;6.

[213] Faham S, Watanabe A, Besserer GM, Cascio D, Specht A, Hirayama BA, Wright EM, Abramson J. The Crystal Structure of a Sodium Galactose Transporter Reveals Mechanistic Insights into Na^+/Sugar Symport. Science. 2008;321(5890):810–814.

[214] Park MS. Molecular Dynamics Simulations of the Human Glucose Transporter GLUT1. PLoS ONE. 2015;10(4):e0125361.

[215] Jewel Y, Dutta P, Liu J. Coarse-Grained Simulations of Proton-Dependent Conformational Changes in Lactose Permease. Proteins: Structure, Function, and Bioinformatics. 2016;84(8):1067–1074.

[216] Plattner N, Noé F. Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models. Nature Communications. 2015;6.

[217] Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of Simple Potential Functions for Simulating Liquid Water. The Journal of Chemical Physics. 1983;79(2):926–935.

[218] Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular Dynamics with Coupling to an External Bath. The Journal of Chemical Physics. 1984;81(8):3684.

[219] Darden T, York D, Pedersen L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. The Journal of Chemical Physics. 1993;98(12):10089.

[220] Kräutler V, Van Gunsteren WF, Hünenberger PH. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. Journal of Computational Chemistry. 2001;22(5):501–508.

[221] Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters From ff99SB. Journal of Chemical Theory and Computation. 2015;11(8):3696–3713.

[222] Huang X, Bowman GR, Bacallado S, Pande VS. Rapid Equilibrium Sampling Initiated From Nonequilibrium Data. Proceedings of the National Academy of Sciences. 2009;106(47):19765–19769.

[223] Zimmerman MI, Hart KM, Sibbald CA, Frederick TE, Jimah JR, Knoverek CR, Tolia NH, Bowman GR. Prediction of New Stabilizing Mutations Based on Mechanistic Insights from Markov State Models. ACS Central Science. 2017;3(12):1311–1321.

[224] Zhou G, Pantelopulos GA, Mukherjee S, Voelz VA. Bridging Microscopic and Macroscopic Mechanisms of p53-MDM2 Binding with Kinetic Network Models. Biophysical Journal. 2017;113(4):785–793.

[225] Paul F, Wehmeyer C, Abualrous ET, Wu H, Crabtree MD, Schöneberg J, Clarke J, Freund C, Weikl TR, Noé F. PRotein-Peptide Association Kinetics Beyond the Seconds Timescale From Atomistic Simulations. Nature Communications. 2017;8(1).

[226] Hopkins CW, Grand SL, Walker RC, Roitberg AE. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. Journal of Chemical Theory and Computation. 2015;11(4):1864–1874.

[227] Smart OS, Neduvelil JG, Wang X, Wallace B, Sansom MS. HOLE: A Program for the Analysis of the Pore Dimensions of Ion Channel Structural Models. Journal of Molecular Graphics. 1996;14(6):354–360.

[228] Metzner P, Schütte C, Vanden-Eijnden E. Transition Path Theory for Markov Jump Processes. Multiscale Modeling and Simulation. 2009;7(3):1192–1219.

[229] Colas C, Masuda M, Sugio K, Miyauchi S, Hu Y, Smith DE, Schlessinger A. Chemical Modulation of the Human Oligopeptide Transporter 1, hPepT1. Molecular Pharmaceutics. 2017;14(12):4685–4693.

[230] Yin Y, He X, Szewczyk P, Chang G. Structure of the Multidrug Transporter EmrD from Escherichia coli. Science. 2006;312(5774):741–744.

[231] Sun L, Zeng X, Yan C, Sun X, Gong X, Rao Y, Yan N. Crystal Structure of a Bacterial Homologue of Glucose Transporters GLUT1-4. Nature. 2012;490(7420):361–366.

[232] Uchiyama T, Kulkarni AA, Davies DL, Lee VH. Biophysical Evidence for His57 as a Proton-Binding Site in the Mammalian Intestinal Transporter hPepT1. Pharmaceutical Research. 2003;20(12):1911–1916.

[233] Chen XZ, Steel A, Hediger MA. Functional Roles of Histidine and Tyrosine Residues in the H^+-Peptide Transporter PepT1. Biochemical and Biophysical Research Communications. 2000;272(3):726–730.

[234] Terada T, Saito H, Mukai M, Inui KI. Identification of the Histidine Residues Involved in Substrate Recognition by a Rat H^+/peptide Cotransporter, PEPT1. FEBS Letters. 1996;394(2):196–200.

[235] Dang S, Sun L, Huang Y, Lu F, Liu Y, Gong H, Wang J, Yan N. Structure of a Fucose Transporter in an Outward-Open Conformation. Nature. 2010;467(7316):734–738.

[236] Hansen FY, Peters GH, Taub H, Miskowiec A. Diffusion of Water and Selected Atoms in DMPC Lipid Bilayer Membranes. The Journal of Chemical Physics. 2012;137(20):204910.

[237] Kučerka N, Tristram-Nagle S, Nagle JF. Structure of Fully Hydrated Fluid Phase Lipid Bilayers with Monounsaturated Chains. The Journal of Membrane Biology. 2006;208(3):193–202.

[238] Zeuthen T. Water-Transporting Proteins. The Journal of Membrane Biology. 2009;234(2):57–73.

[239] Minhas GS, Bawdon D, Herman R, Rudden M, Stone AP, James AG, Thomas GH, Newstead S. Structural Basis of Malodour Precursor Transport in the Human Axilla. eLife. 2018;7.

[240] Sowa GZ, Qin PZ. Site-directed Spin Labeling Studies on Nucleic Acid Structure and Dynamics. In: Progress in Nucleic Acid Research and Molecular Biology. Elsevier; 2008. p. 147–197.

[241] Joseph B, Sikora A, Bordignon E, Jeschke G, Cafiso DS, Prisner TF. Distance Measurement on an Endogenous Membrane Transporter in E. Coli Cells and Native Membranes Using EPR Spectroscopy. Angewandte Chemie International Edition. 2015;54(21):6196–6199.

[242] Singewald K, Lawless MJ, Saxena S. Increasing Nitroxide Lifetime in Cells to Enable In-Cell Protein Structure and Dynamics Measurements by Electron Spin Resonance Spectroscopy. Journal of Magnetic Resonance. 2019;299:21–27.

[243] Widder P, Schuck J, Summerer D, Drescher M. Combining Site-Directed Spin Labeling In Vivo and In-Cell EPR Distance Determination. Physical Chemistry Chemical Physics. 2020;22(9):4875–4879.

[244] Jeschke G, Polyhach Y. Distance Measurements on Spin-Labelled Biomacromolecules by Pulsed Electron Paramagnetic Resonance. Physical Chemistry Chemical Physics. 2007;9(16):1895.

[245] Shen R, Han W, Fiorin G, Islam SM, Schulten K, Roux B. Structural Refinement of Proteins by Restrained Molecular Dynamics Simulations with Non-interacting Molecular Fragments. PLoS Computational Biology. 2015;11(10):e1004368.

[246] Marinelli F, Faraldo-Gómez JD. Ensemble-Biased Metadynamics: A Molecular Simulation Method to Sample Experimental Distributions. Biophysical Journal. 2015;108(12):2779–2782.

[247] Marrink SJ, Corradi V, Souza PCT, Ingólfsson HI, Tieleman DP, Sansom MSP. Computational Modeling of Realistic Cell Membranes. Chemical Reviews. 2019;119(9):6184–6226.

[248] Martens C, Stein RA, Masureel M, Roth A, Mishra S, Dawaliby R, Konijnenberg A, Sobott F, Govaerts C, Mchaourab HS. Lipids Modulate the Conformational Dynamics of a Secondary Multidrug Transporter. Nature Structural & Molecular Biology. 2016;23(8):744–751.

[249] Sahu ID, McCarrick RM, Troxel KR, Zhang R, Smith HJ, Dunagan MM, Swartz MS, Rajan PV, Kroncke BM, Sanders CR, Lorigan GA. DEER EPR Measurements for Membrane Protein Structures via Bifunctional Spin Labels and Lipodisq Nanoparticles. Biochemistry. 2013;52(38):6627–6632.

[250] Ward R, Pliotas C, Branigan E, Hacker C, Rasmussen A, Hagelueken G, Booth IR, Miller S, Lucocq J, Naismith JH, Schiemann O. Probing the Structure of the Mechanosensitive Channel of Small Conductance in Lipid Bilayers with Pulsed Electron-Electron Double Resonance. Biophysical Journal. 2014;106(4):834–842.

[251] Teucher M, Zhang H, Bader V, Winklhofer KF, García-Sáez AJ, Rajca A, Bleicken S, Bordignon E. A New Perspective on Membrane-Embedded Bax Oligomers Using DEER and Bioresistant Orthogonal Spin Labels. Scientific Reports. 2019;9(1).

[252] Potapov A, Yagi H, Huber T, Jergic S, Dixon NE, Otting G, Goldfarb D. Nanometer-Scale Distance Measurements in Proteins Using Gd^{3+} Spin Labeling. Journal of the American Chemical Society. 2010;132(26):9040–9048.

[253] Bogetti X, Ghosh S, Jarvi AG, Wang J, Saxena S. Molecular Dynamics Simulations Based on Newly Developed Force Field Parameters for Cu^{2+} Spin Labels Provide Insights into Double-Histidine-Based Double Electron–Electron Resonance. The Journal of Physical Chemistry B. 2020;.

[254] Banerjee D, Yagi H, Huber T, Otting G, Goldfarb D. Nanometer-Range Distance Measurement in a Protein Using Mn^{2+} Tags. The Journal of Physical Chemistry Letters. 2012;3(2):157–160.

[255] Matalon E, Huber T, Hagelueken G, Graham B, Frydman V, Feintuch A, Otting G, Goldfarb D. Gadolinium(III) Spin Labels for High-Sensitivity Distance Measurements in Transmembrane Helices. Angewandte Chemie International Edition. 2013;52(45):11831–11834.

[256] Claxton DP, Quick M, Shi L, de Carvalho FD, Weinstein H, Javitch JA, Mchaourab HS. Ion/Substrate-Dependent Conformational Dynamics of a Bacterial Homolog of Neurotransmitter:sodium Symporters. Nature Structural & Molecular Biology. 2010;17(7):822–829.

[257] Kazmier K, Sharma S, Quick M, Islam SM, Roux B, Weinstein H, Javitch JA, Mchaourab HS. Conformational Dynamics of Ligand-Dependent Alternating Access in LeuT. Nature Structural & Molecular Biology. 2014;21(5):472.

[258] Newstead S, Drew D, Cameron AD, Postis VLG, Xia X, Fowler PW, Ingram JC, Carpenter EP, Sansom MSP, McPherson MJ, Baldwin SA, Iwata S. Crystal Structure of a Prokaryotic Homologue of the Mammalian Oligopeptide-Proton Symporters, PepT1 and PepT2. The EMBO Journal. 2010;30(2):417–426.

[259] Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: The Biomolecular Simulation Program. Journal of Computational Chemistry. 2009;30(10):1545–1614.

[260] Lee J, Cheng X, Swails JM, Yeom MS, Eastman PK, Lemkul JA, Wei S, Buckner J, Jeong JC, Qi Y, Jo S, Pande VS, Case DA, Brooks CL, MacKerell AD, Klauda JB, Im W. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. Journal of Chemical Theory and Computation. 2015;12(1):405–413.

[261] Wu EL, Cheng X, Jo S, Rui H, Song KC, Dávila-Contreras EM, Qi Y, Lee J, Monje-Galvan V, Venable RM, Klauda JB, Im W. CHARMM-GUI Membrane Builder Toward Realistic Biological Membrane Simulations. Journal of Computational Chemistry. 2014;35(27):1997–2004.

[262] Cheng X, Jo S, Lee HS, Klauda JB, Im W. CHARMM-GUI Micelle Builder for Pure/Mixed Micelle and Protein/Micelle Complex Systems. Journal of Chemical Information and Modeling. 2013;53(8):2171–2180.

[263] Qi Y, Lee J, Cheng X, Shen R, Islam SM, Roux B, Im W. CHARMM-GUI DEER Facilitator for Spin-Pair Distance Distribution Calculations and Preparation of Restrained-Ensemble Molecular Dynamics Simulations. Journal of Computational Chemistry. 2019;41(5):415–420.

[264] MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D, Karplus M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†. The Journal of Physical Chemistry B. 1998;102(18):3586–3616.

[265] Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, Mackerell AD. CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible With the CHARMM All-Atom Additive Biological Force Fields. Journal of Computational Chemistry. 2009; p. NA–NA.

[266] Best RB, Zhu X, Shim J, Lopes PEM, Mittal J, Feig M, MacKerell AD. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain $\chi 1$ and $\chi 2$ Dihedral Angles. Journal of Chemical Theory and Computation. 2012;8(9):3257–3273.

[267] Cheng X, Kim JK, Kim Y, Bowie JU, Im W. Molecular Dynamics Simulation Strategies for Protein–micelle Complexes. Biochimica et Biophysica Acta (BBA) - Biomembranes. 2016;1858(7):1566–1572.

[268] Lipfert J, Columbus L, Chu VB, Lesley SA, Doniach S. Size and Shape of Detergent Micelles Determined by Small-Angle X-ray Scattering. The Journal of Physical Chemistry B. 2007;111(43):12427–12438.

[269] Columbus L, Lipfert J, Jambunathan K, Fox DA, Sim AYL, Doniach S, Lesley SA. Mixing and Matching Detergents for Membrane Protein NMR Structure Determination. Journal of the American Chemical Society. 2009;131(21):7320–7326.

[270] Strop P, Brunger AT. Refractive Index-Based Determination of Detergent Concentration and Its Application to the Study of Membrane Proteins. Protein Science. 2005;14(8):2207–2211.

[271] Krishnamurthy H, Gouaux E. X-Ray Structures of LeuT in Substrate-Free Outward-Open and Apo Inward-Open States. Nature. 2012;481(7382):469–474.

[272] Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A. Comparative Protein Structure Modeling Using Modeller. Current Protocols in Bioinformatics. 2006;15(1):5.6.1–5.6.30.

[273] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera-A Visualization System for Exploratory Research and Analysis. Journal of Computational Chemistry. 2004;25(13):1605–1612.

[274] Jo S, Cheng X, Islam SM, Huang L, Rui H, Zhu A, Lee HS, Qi Y, Han W, Vanommeslaeghe K, MacKerell AD, Roux B, Im W. CHARMM-GUI PDB Manipulator for Advanced Modeling and Simulations of Proteins Containing Nonstandard Residues. In: Advances in Protein Chemistry and Structural Biology. Elsevier; 2014. p. 235–265.

[275] Chan MC, Selvam B, Young HJ, Procko E, Shukla D. The Substrate Import Mechanism of the Human Serotonin Transporter. ChemRxiv. 2019; p. 9922301.

[276] Coleman JA, Yang D, Zhao Z, Wen PC, Yoshioka C, Tajkhorshid E, Gouaux E. Serotonin Transporter–ibogaine Complexes Illuminate Mechanisms of Inhibition and Transport. Nature. 2019;569(7754):141–145.

[277] Gotfryd K, Boesen T, Mortensen JS, Khelashvili G, Quick M, Terry DS, Missel JW, LeVine MV, Gourdon P, Blanchard SC, Javitch JA, Weinstein H, Loland CJ, Nissen P, Gether U. X-Ray Structure of LeuT in an Inward-Facing Occluded Conformation Reveals Mechanism of Substrate Release. Nature Communications. 2020;11(1).

[278] Selvam B, Yu YC, Chen LQ, Shukla D. Molecular Basis of the Glucose Transport Mechanism in Plants. ACS Central Science. 2019;5(6):1085–1096.

[279] Cheng KJ, Selvam B, Chen LQ, Shukla D. Distinct Substrate Transport Mechanism Identified in Homologous Sugar Transporters. The Journal of Physical Chemistry B. 2019;123(40):8411–8418.

[280] Terry DS, Kolster RA, Quick M, LeVine MV, Khelashvili G, Zhou Z, Weinstein H, Javitch JA, Blanchard SC. A Partially-Open Inward-Facing Intermediate Conformation of LeuT Is Associated With Na+ Release and Substrate Transport. Nature Communications. 2018;9(1).

[281] Möller IR, Slivacka M, Nielsen AK, Rasmussen SGF, Gether U, Loland CJ, Rand KD. Conformational Dynamics of the Human Serotonin Transporter During Substrate and Drug Binding. Nature Communications. 2019;10(1).

[282] Martens C, Shekhar M, Borysik AJ, Lau AM, Reading E, Tajkhorshid E, Booth PJ, Politis A. Direct Protein-Lipid Interactions Shape the Conformational Landscape of Secondary Transporters. Nature Communications. 2018;9(1).

[283] Giladi M, Khananshvili D. Hydrogen-Deuterium Exchange Mass-Spectrometry of Secondary Active Transporters: From Structural Dynamics to Molecular Mechanisms. Frontiers in Pharmacology. 2020;11.

[284] Yamashita A, Singh SK, Kawate T, Jin Y, Gouaux E. Crystal Structure of a Bacterial Homologue of Na^+/Cl^-Dependent Neurotransmitter Transporters. Nature. 2005;437(7056):215–223.

[285] Singh SK, Yamashita A, Gouaux E. Antidepressant Binding Site in a Bacterial Homologue of Neurotransmitter Transporters. Nature. 2007;448(7156):952–956.

[286] Singh SK, Piscitelli CL, Yamashita A, Gouaux E. A Competitive Inhibitor Traps LeuT in an Open-to-Out Conformation. Science. 2008;322(5908):1655–1661.

[287] Quick M, Winther AML, Shi L, Nissen P, Weinstein H, Javitch JA. Binding of an Octylglucoside Detergent Molecule in the Second Substrate (S2) Site of LeuT Establishes an Inhibitor-Bound Conformation. Proceedings of the National Academy of Sciences. 2009;106(14):5563–5568.

[288] Zhou Z, Zhen J, Karpowich NK, Law CJ, Reith MEA, Wang DN. Antidepressant Specificity of Serotonin Transporter Suggested by Three LeuT–SSRI Structures. Nature Structural & Molecular Biology. 2009;16(6):652–657.

[289] Wang H, Elferich J, Gouaux E. Structures of LeuT in Bicelles Define Conformation and Substrate Binding in a Membrane-Like Context. Nature Structural & Molecular Biology. 2012;19(2):212–219.

[290] Malinauskaite L, Said S, Sahin C, Grouleff J, Shahsavar A, Bjerregaard H, Noer P, Severinsen K, Boesen T, Schiøtt B, Sinning S, Nissen P. A Conserved Leucine Occupies the Empty Substrate Site of LeuT in the Na^+-Free Return State. Nature Communications. 2016;7(1).

[291] Bruce CD, Berkowitz ML, Perera L, Forbes MDE. Molecular Dynamics Simulation of Sodium Dodecyl Sulfate Micelle in Water: Micellar Structural Characteristics and Counterion Distribution. The Journal of Physical Chemistry B. 2002;106(15):3788–3793.

[292] Bond PJ, Sansom MSP. Membrane Protein Dynamics versus Environment: Simulations of OmpA in a Micelle and in a Bilayer. Journal of Molecular Biology. 2003;329(5):1035–1053.

[293] Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM Database and PPM Web Server: Resources for Positioning of Proteins in Membranes. Nucleic Acids Research. 2011;40(D1):D370–D376.

[294] Moffett AS, Bender KW, Huber SC, Shukla D. Allosteric Control of a Plant Receptor Kinase Through S-Glutathionylation. Biophysical Journal. 2017;113(11):2354–2363.

[295] Jia R, Martens C, Shekhar M, Pant S, Pellowe GA, Lau AM, Findlay HE, Harris NJ, Tajkhorshid E, Booth PJ, Politis A. Hydrogen-deuterium Exchange Mass Spectrometry Captures Distinct Dynamics upon Substrate and Inhibitor Binding to a Transporter. bioRxiv. 2020; p. 23564.

[296] Cai Q, Kusnetzow AK, Hubbell WL, Haworth IS, Gacho GPC, Eps NV, Hideg K, Chambers EJ, Qin PZ. Site-directed Spin Labeling Measurements of Nanometer Distances in Nucleic Acids Using a Sequence-independent Nitroxide Probe. Nucleic Acids Research. 2006;34(17):4722–4730.

[297] Barhate N, Cekan P, Massey A, Sigurdsson S. A Nucleoside That Contains a Rigid Nitroxide Spin Label: A Fluorophore in Disguise. Angewandte Chemie International Edition. 2007;46(15):2655–2658.

[298] Cekan P, Smith AL, Barhate N, Robinson BH, Sigurdsson ST. Rigid spin-labeled nucleoside Ç: a nonperturbing EPR probe of nucleic acid conformation. Nucleic Acids Research. 2008;36(18):5946–5954.

[299] Ivani I, Dans PD, Noy A, Pérez A, Faustino I, Hospital A, Walther J, Andrio P, Goñi R, Balaceanu A, Portella G, Battistini F, Gelpí JL, González C, Vendruscolo M, Laughton CA, Harris SA, Case DA, Orozco M. PARMBSC1: A Refined Force Field for DNA Simulations. Nature Methods. 2015;13(1):55–58.

[300] Bergonzo C, Cheatham TE. Improved Force Field Parameters Lead to a Better Description of RNA Structure. Journal of Chemical Theory and Computation. 2015;11(9):3969–3972.

[301] Tan D, Piana S, Dirks RM, Shaw DE. RNA Force Field With Accuracy Comparable to State-of-the-Art Protein Force Fields. Proceedings of the National Academy of Sciences. 2018;115(7):E1346–E1355.

[302] Minhas V, Sun T, Mirzoev A, Korolev N, Lyubartsev AP, Nordenskiöld L. Modeling DNA Flexibility: Comparison of Force Fields from Atomistic to Multiscale Levels. The Journal of Physical Chemistry B. 2019;124(1):38–49.

[303] Kührová P, Mlýnský V, Zgarbová M, Krepl M, Bussi G, Best RB, Otyepka M, Šponer J, Banáš P. Improving the Performance of the Amber RNA Force Field by Tuning the Hydrogen-Bonding Interactions. Journal of Chemical Theory and Computation. 2019;15(5):3288–3305.

[304] lu Li Z, Buck M. Modified Potential Functions Result in Enhanced Predictions of a Protein Complex by All-Atom Molecular Dynamics Simulations, Confirming a Stepwise Association Process for Native Protein–Protein Interactions. Journal of Chemical Theory and Computation. 2019;15(8):4318–4331.

[305] Plattner N, Doerr S, Fabritiis GD, Noé F. Complete Protein–protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling. Nature Chemistry. 2017;9(10):1005–1011.

[306] Pan AC, Jacobson D, Yatsenko K, Sritharan D, Weinreich TM, Shaw DE. Atomic-Level Characterization of Protein–Protein Association. Proceedings of the National Academy of Sciences. 2019;116(10):4244–4249.

[307] Moffett AS, Shukla D. How Do Brassinosteroids Activate Their Receptors? bioRxiv. 2019; p. 630640.

[308] MacKerell AD, Nilsson L. Molecular Dynamics Simulations of Nucleic Acid–protein Complexes. Current Opinion in Structural Biology. 2008;18(2):194–199.

[309] Etheve L, Martin J, Lavery R. Dynamics and Recognition Within a protein–DNA Complex: A Molecular Dynamics Study of the SKN-1/DNA Interaction. Nucleic Acids Research. 2015;44(3):1440–1448.

[310] Moffett AS, Shukla D. Structural Consequences of Multisite Phosphorylation in the BAK1 Kinase Domain. Biophysical Journal. 2020;118(11):698–707.

[311] Feig M, Yu I, hung Wang P, Nawrocki G, Sugita Y. Crowding in Cellular Environments at an Atomistic Level from Computer Simulations. The Journal of Physical Chemistry B. 2017;121(34):8009–8025.

[312] Rickard MM, Zhang Y, Gruebele M, Pogorelov TV. In-Cell Protein–Protein Contacts: Transient Interactions in the Crowd. The Journal of Physical Chemistry Letters. 2019;10(18):5667–5673.

[313] Shi X, Chuo SW, Liou SH, Goodin DB. Double Electron-Electron Resonance Shows That Substrate but Not Inhibitors Cause Disorder in the F/G Loop of CYP119 in Solution. Biochemistry. 2020;.

[314] Guin D, Gelman H, Wang Y, Gruebele M. Heat Shock-Induced Chaperoning by Hsp70 Is Enabled In-Cell. PLoS ONE. 2019;14(9):e0222990.

[315] Guin D, Gruebele M. Chaperones Hsc70 and Hsp70 Bind to the Protein PGK Differently inside Living Cells. The Journal of Physical Chemistry B. 2020;.

[316] Guin D, Mittal S, Bozymski B, Shukla D, Gruebele M. Dodine as a Kosmo-Chaotropic Agent. The Journal of Physical Chemistry Letters. 2019;10(10):2600–2605.

[317] Wu H, Paul F, Wehmeyer C, Noé F. Multiensemble Markov Models of Molecular Thermodynamics and Kinetics. Proceedings of the National Academy of Sciences. 2016;113(23):E3221–E3230.