THE UNIVERSITY OF CHICAGO

FAST, ATOMIC-LEVEL SIMULATIONS OF THE FORCED UNFOLDING OF PROTEINS USING A NEW MEMBRANE BURIAL POTENTIAL

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY
ZONGAN WANG

CHICAGO, ILLINOIS
MARCH 2019

To my grandparents,

WANG De Shun and *MA Dong E*,

my parents,

WANG Xin Jian and *GAO Qin*,

and my fiancée

GUO Ni Ning.

Song of the Grand River sung,

I head resolute for the east,

Having vainly delved in all schools

For clues to a better world.

Ten years face to wall,

I shall make a break-through,

Or die an avowed rebel

Daring to tread the sea.

– *ZHOU En Lai*, 1917 (translated by *Nancy T. Lin*)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This thesis would be impossible without the full support from my family. I want to thank my parents for putting my education in the first place when I was young and always urging me to go further. I am grateful to my fiancée Nining that her love and trust makes home a forever warm place to rest in. I wish my grandparents were here to see that their youngest grandson becomes the first doctor in the family.

I owe a great intellectual debt to my advisors, Tobin Sosnick and Karl Freed. It is mutually advantageous for theoretical work to crosstalk with experiment, and Tobin is excellent at interpreting these two cultures of science. Tobin's affablility and approachableness allow students in our group to express bravely and work on difficult problems in the long term. Every student in our group has a long-term independent project that is aimed at exploring an unspoiled territory of science, which process often takes a full five or six years. However, never has the publication list become the paramount factor in evaluating that student's effort. Moreover, Tobin finds every opportunity to send students to conferences so that students can expose themselves to novel ideas and reflect on their own projects through the communications with colleagues from other places. For me, it was the Gorden Conference of membrane proteins at Boston in 2017 where I was impressed by studies of membrane protein unfolding using AFM-based single-molecule force spectroscopy and came forth the idea of simulaing the experiment with our own tool, which later directly leads to my thesis project.

Karl is a consummate theorist and is an inspiration to a lifetime of tackling hard problems with insightful techniques. Karl and Prof. WU Chi at CUHK proposed an exchange program for undergraduate students from USTC to come to UChicago and do research. I was selected as one of the nine students to visit UChicago in the summer of 2011 and to learn how to conduct research under Karl's supervision. Later, it was Karl's recommendation that led to my admission to the graduate program at UChicago. Karl's encouraging words, sense of humer, and infinite wisdom always come to my rescue when it seemed like I have exhausted all possible ways of looking at a problem.

I also want to thank my *pro forma* advisor Prof. Aaron Dinner for being part of my thesis committee and providing helpful suggestions regarding my career development regularly.

Joining a wrong group could make the whole PhD life miserable, but joining a right one is something you want to brag about to others. I want to thank my lab mates for vigorous discussion and a willingness to provide help patiently whenever needed. In particular, Esmael Haddadian taught me to perform MD simulations almost hand-by-hand in the first place in the summer of 2011. Together, we studied the aggregation of beta-amyloid. John Jumper is probably one of the most intellegent and resourceful students I've met at UChicago. His thesis project, *Upside*, a versatile extensible coarse-grained package that can perform fast but accurate MD simulations, paves the groundwork for studying protein dynamics in a lot of areas of protein science. Besides, the ingenious methods that John employed to parametrize the program shed light upon future projects. As an old Chinese proverb goes, even a good wife cannot cook without rice. Without *Upside*, my study of the forced unfolding of membrane proteins wouldn't be possible. Aashish Adhikari and Yu Wookyung helped me start my project when I joined the group in the January of 2014, teaching me how to run coase-grained simulations using OOPS and how to make plots with GNUplot. Michael Baxa and Adam Zmyslowski are never impatient in advising me on how to articulate the information in paper and make figures that best convey clear information. I've also learnt a lot from Kevin Song and Nabil Faruk on analyzing simulation results. I also thank Jack Skinner, Josh Riback, Alex French, Isabelle Gagnon, and James Hinshaw for helpful discussions. We come from different corners of the world with different backgrounds and faiths. We are like a big warm family.

Finally, I want to thank my friends here, especially XU Ruoyu and MA Xiufeng, for making Chicago an enjoyable place to stay despite the freezing cold in winter and the dautingly high criminal rate around Hyde Park, and LUO Zhiqiang, XU Wensheng, DENG Zhiyou, WANG Sheng, WANG Lei, LI Wenxin, LI Jingxian, LI Tian, and BAO Ying for caring me as a little brother; my friends in other places, especially GUO Zhi'ang, WANG Genqi, LÜ Kongpeng, LIU Wenhao, LI Mingyun, HUANG Yun, WU Boning, WANG Liang, WEI Bingbing, BAI Shirong, WANG Zhefan, ZHAO Xuefang, YUAN Shuai, ZHU Linchen, ZHOU Ang, DONG Xialing, PAN Yujing, SHANG Qichao, LUO Jiaze, TANG Jun, ZHANG Jun, TAO Yu, and GAO Xu, when I was frustrated and needed them most; my friends who ran marathons with me for enduring the pain and enjoying the gain together, especially LIU Di, CAI Zhengxu, LI Gang, LUAN Gan, WANG Xin, and HUANG Zili; my friends in the Northwestern University Reading Club, especially PAN Saijie, CHEN Yaoyao, LIU Zhenyao, WU Wenhui, PENG Zeyang, CHEN Xiao, LIU Yilan, REN Baixin for sharing excellent books and provoking new thoughts; my friend, CHEN Huangrui, for standing firmly for justice and not being scared by calumny; and my friends TONG Yan, WANG Shuxuan, SHAN Danni, and LIN Jing, for being cat buddies together.

ABSTRACT

Membrane proteins carry great importance in cellular functions, such as nutrient uptake, transport of membrane-impermeable molecules, ion balance, etc. These make membrane proteins the prime drug targets. In fact, 50% of modern medicines target helical membrane proteins. However, despite the biological importance, membrane proteins are notoriously difficult to study. Because it is difficult to obtain high-resolution structures, membrane proteins are greatly under-represented in Protein Data Bank. The scarcity of available structures makes it even harder to obtain information of the dynamic behaviors of membrane proteins. The topic lying at the heart of the problem is how transmembrane proteins fold.

Observation of how bacteriorhodopsin (bR) folds *in vitro* leads to a two-stage thermodynamic model, which was brought up in 1990. In the first stage, unfolded chain forms helices on the surface of the membrane and the helices spontaneously insert into the bilayer. In the second stage, the transmembrane helices assemble into a well-functional tertiary structure. To study the folding in terms of a thermodynamic model is meaningful because *in vivo* folding must proceed within the thermodynamic context. The importance of understanding the two-stage model is that the two stages are controlled by different forces. The first stage is driven mainly by the hydrophobic effect, whereas the second stage is mediated by various weak interactions.

There are several factors that add to the difficulty of studying the folding of transmembrane proteins computationally. First, the structures of transmembrane proteins are in nature complex. Ideally, the transmembrane protein consists of several transmembrane helices, each of which is hydrophobically stabe and spans the lipid bilayer, such as bR. However, exceptions happen all the time. Re-entrant helices enter and exit the bilayer on the same side; interfacial helices lie on the interface of the membrane; and there are kinks and coils in the middle of a transmembrane helix. Moreover, sometimes the transmembrane helix is not hydrophobically stable by itself but via the association with neighboring helices. Second, the folding and stability of transmembrane proteins are dictated by a delicate balance of

various weak interactions, such as van der Waals forces, H-bonding, salt-bridge, and weakly polar interactions. The interplay between the protein and the lipid bilayer also plays an important role. Besides, multiple functional conformations exist. The energy minimum may only correlate to one of them. Thus, developing a force field, which describes the balance of those weak interactions well and is able to distinguish native-like structures from non-native structures, is the prerequisite for computational study. Third, it is hard to mimic the realistic protein/membrane complex, as the complex is highly heterogeneous and composed of a variety of biomolecules at different concentrations. Lastly, it is usually very expensive to simulate membrane proteins. The size of the system is typically larger than 100,000 atoms, including the protein, lipids, ligands, water molecules and ions, and the timescale required for obtaining physically meaningful results is usually microsecond or longer.

To circumvent these problems or to look at the problem from a different angle, people unfold membrane proteins by force and obtain the folding energetics by extrapolating the applied force back to zero. Experimentally, single-molecule force spectroscopy (SMFS), such as atomic force spectroscopy (AFM) and magnetic tweezers, allows scientists to manipulate biomolecules on the single-molecule level. SMFS has proven beneficial in detecting sparsely populated intermediates and yielding kinetic insights into the unfolding pathways of membrane proteins.

To put it all together, before my study, Dr. John Jumper in our group has developed a fast, atomic-level coarse-grained model, *Upside*, which is capable of *de novo* folding of proteins shorter than 100 residues in cpu-hours. *Upside* is a non-Gõ, physics based model with five atoms per residue (N, Cα, C, H, O), a side chain bead and with residue- and neighbor-dependent Ramachandran maps. The energy function includes H-bonds, side chain-side chain and side chain-backbone interactions (including helix capping), and a solvation term. At each step, the side chain bead is first decorated to each of the residues. The positions of the side chain beads are determined based on the joint probability of all side chain beads which gives the lowest global free energy for all side chains. The force is computed using the joint probability. Then, the side chain beads are undecorated while the forces are applied to the backbone atoms.

I have incorporated *Upside* with my new knowledge-based membrane burial potential as the implicit solvent for membrane proteins, which dynamically calculates the degree of side chain exposure to lipids during the simulations and includes energies for unsatisfied H-bond donors and acceptors in the membrane. Hence, I am able to perform fast, atomic-level simulations on membrane proteins. In specific, I choose to study the forced unfolding of membrane proteins by SMFS.

I have developed an accurate and fast atomic-level simulation that allows me to conduct hundreds of unfolding simulations to characterize the folding energy surface under force. The algorithm reproduces many of the experiment features of SMFS studies for the unfolding of bR and GlpG. I find that the mode of force application alters the perception of the folding landscape. For GlpG unfolding using a weaker spring to mimic a magnetic tweezers measurement, the force remains nearly constant after the initial unfolding event and few if any intermediates are observed, as found in experiment. With a stiff cantilever, however, the force drops to near-zero after each major unfolding event and numerous intermediates are observed. Notably, the application of constant force, whether intentionally or as a result of force being applied through a weak spring constant, as with magnetic tweezers or after a substantial portion of the protein is unfolded, reduces the probability of observing intermediates while increasing the apparent unfolding cooperativity as compared to the use of stiffer cantilevers with short attachment handles. Our method can assist experimental studies by helping convert force extension curves to structures, pathways, and energies, which can be challenging. Moreover, our method can be employed to simulate complicated thought experiments beyond current experimental capabilities such as pulling on multiple sites in multiple directions with different strength springs with either membrane or soluble proteins.

CHAPTER 1
INTRODUCTION

I studied the dynamics of transmembrane proteins using a coarse-grained model with an implicit membrane force field. There are two main types of architecture of transmembrane proteins: α-helical bundles and β-barrels. In this thesis, I limit the scope of my study to transmembrane α-helical proteins. In particular, I am focused on simulating the forced unfolding of transmembrane helical proteins.

1.1 Motivation of the study

Transmembrane helical proteins play a pivotal role in cell biology, including nutrient uptake, transport of membrane-impermeable molecules, ion balance, signal transduction, intercellular communication, immune response. Therefore, they are prime drug targets [1]. Specifically, more than 50% of modern medicinal drugs target only four key gene families: class I GPCRs, nuclear receptors, ligand-gated ion channels and voltage-gated ion channels [2].

Despite their biological importance, membrane proteins are greatly under-represented in structural databases that only ∼2% deposited structures in Protein Data Bank are membrane proteins [3], among which ∼85% are integral membrane proteins [4, 5] (the rest are β-barrels). As a comparison, ∼30% of human genome encodes membrane proteins [6, 7]. The number of resolved structures has risen from marginally over 100 at the turn of the millennia [8] to ∼200 at 2009 [9] and to over 4000 in Dec. 2018 [10].

The scarcity of available structures stems from the difficulty in obtainning high-resolution structures experimentally that expression of sufficient amount of membrane proteins for crystallization is often impeded by toxicity to the host cells or misfolding of the protein [3]. The majority of membrane protein structures is determined by X-ray diffraction, solution nuclear magnetic resonance (NMR), or more recently, single-particle electron cryomicroscopy (cryo-EM). However, the proteins are often not in a membrane, but rather, in a crystal lattice or in a membrane-mimetic, such as a micelle or bicelle [11], which necessitates the knowledge of protein-lipid bilayer interaction in order to attain correct comformation of protein in its native membrane. Moreover, the lack of atomic coordinates makes it even harder to obtain information of dynamic behaviors, for instance the conformational transitions between outward-facing and inward-facing states of membrane transporters [3, 12].

Continual efforts have been put into studying transmembrane proteins computationally to aid the interpretation of experiment and to provide testable predictions for experimentalists. In terms of predicting the tertiary structures, the computational methods can be categorized into two complementary classes: (i) static methods to predict 3D structure from primary sequence, such as homology modeling [13], fold recognition [14, 15], and *de novo* prediction [16–21], and (ii) dynamic methods studing comformational change of a protein along a time trajectory, including all-atom dynamic simulations and coarse-grained models [22–27].

1.1.1 To fold or to unfold?

The central topic of studying transmembrane proteins is how the proteins fold. However, it is formidable to study the folding of transmembrane proteins computationally.

How do transmembrane proteins fold?

Evidences have shown that the folding of bacteriorhodopsin (bR) is thermodynamically controlled and the lateral interactions among transmembrane helices provide the driving forces for reaching the native state [28, 29]. Studying how transmembrane proteins fold in terms of thermodynamic models is of great importance for understanding the *in vivo* folding process aided by translocons [30], because the biological process must proceed within the thermodynamic context [31].

A well accepted two-stage thermodynamic model for the folding of transmembrane proteins *in vitro* was first proposed by Popot and Engelman in 1990 [29], albeit admittedly oversimplified, describing the assembly of isolated fragments of bR in lipid bilayers into functional form [28]. The model states that stable transmembrane helices form independently in stage I, principally due to the hydrophobic effect and the formation of main-chain hydrogen bonds in the non-aqueous environment, and that transmembrane helices associate and assemble to form the tertiary structure in stage II [29].

Almost at the same time, Jacobs and White proposed a three-step model [32] based on observations of the partitioning of small hydrophobic peptides and the spontaneous insertion of helical hairpins into and across membranes [33]. The three steps are I. the binding of peptide chain to the membrane interface and the formation of main-chain hydrogen bonds (interfacial partitioning), II. formation of helices at the interface (interfacial folding), and III. spontaneous insertion of helices as well as the charged and polar residues connecting the helices [32]. Later in 1999, White and Wimley combined the abovementioned two models and proposed a four-step thermodynamic model, describing the folding proceed as: partitioning, folding, insertion, and association [31].

What makes studying the folding of transmembrane proteins computationally so difficult?

Besides the aforementioned lack of available atomic structures, there are several factors that add to the difficulty of studying the folding of transmembrane proteins computationally. First, the structure and topology of transmembrane proteins are in nature complex. Canonically, multi-span transmembrane proteins consist of a number of transmembrane helices, each of which is hydrophobically stable and extends over the membrane bilayer. bR is a good example of such transmembrane protein. However, more complex crystal structures have reveald a number of structure features that do not follow the cannonical view. For example, re-entrant helices (short helices that enter and exit the membrane on the same side), amphipathic helices (surface-bound), and kinks and coils in the membrane region [1].

Second, the folding of transmembrane proteins is governed by a delicate balance of various weak interactions [34, 35]. The hydrophobic effect cannot act as the main driving force for the transmembrane helices to associate, distinct to soluble proteins in an isotropic aqueous environment. In the two-stage model, while the hydrophobic effect largely leads to the spontaneous insertion in stage I, it is the weak interactions, such as van der Waals forces (eg. GXXXG motif [36]), hydrogen bond (eg. Cα hydrogen bond [37]), salt bridge, weakly polar interactions (eg. cation-π interaction, aromatic-aromatic interaction), that are the main contributors in stage II [38]. Deriving a force field to precisely balance those weak interactions and to be able to distinguish native-like from non-native structures is challenging.

Third, to accurately mimic a realistic protein/membrane complex is not easy, as biological membranes are highly heterogenous and composed of a variety of biomolecules with different concentrations [39]. For all-atom MD simulations, dedicated efforts have been put into standardizing the preparation for building the protein/membrane complex as well as continuously adding new types of lipid and ligand molecules into the library [39–41]. However, the knowledge of membrane composition is required *a priori* to build such a realistic model.

To unfold in experiment

Experimentally, single-molecule force spectroscopy (SMFS), such as atomic force spectroscopy (AFM) and magnetic tweezers, allows scientists to manipulate biomolecules on the single-molecule level [42]. SMFS has proven beneficial in detecting sparsely populated intermediates and yielding kinetic insights into the unfolding pathways of membrane proteins [43]. Although SMFS unfolds membrane protein under non-equilibrium conditions, the energetics of membrane protein folding can be extrapolated back to zero force [44, 45]. This gives people an opportunity to circumvent the difficulties of folding a membrane protein and to look at the problem from another perspective.

1.1.2 All-atom or coarse-grained?

The state-of-the-art MD simulations are capable of exploring membrane protein conformations and protein-lipid interactions in native-like bilayer environments [39]. Unfortunately, the computational resources required for performing all-atom MD simulations of transmembrane proteins are often prohibitively expensive. The size of the system is usually very large. The total atoms of the system, including the protein, lipid bilayer, pore water, bulk water, and counter ions, are typically over 100,000 [40]. On the other hand, the timescale required for acquiring physically meaningful results is microsecond or longer [3].

Coarse-grained (CG) models address the both the size and timescale issues via simplification of the representation of the constituent molecules of biomolecular systems [46]. In specific, there are few conversations between SMFS experiment and computation, partly due to the demanding computational resources. For example, to fully unfold a bR molecule, which extended chain is $\sim$ 80 nm, at an pulling rate of 300 nm/s [43], needs $\sim$ 0.26 s. This timescale for simulating a membrane protein is daunting even using CG-MD simulation. CG models expedite the MD simulations and allow for slower and more realistic extraction velocities for simulating SMFS experiments [47], therefore increasing the likelihood of observing transient intermediates.

1.1.3 Concerns of CG models with implicit solvent

To further accelerate the simulation, an implicit solvent model is preferred over an explicit solvent model with coarse-grained solvent molecules. However, when an implicit solvent model is used, there are several issues that should be born in mind. First, the distinct physical ambience of membrane bilayer to the aqueous phase demands the knowledge of the exposure status of residues, namely to what extent a residue is in contact with the surrounding lipid molecules. Naturally, only those residues exposed to the lipids should interact with them. Moreover, charged or polar residues in the cavity or pore of ion channels or transporters add to the complexity of the problem, because they interact with the water molecules rather than lipids. Unfortunately, to determine a residue's exposure status (in a coarse-grained manner) is not a readily solvable problem. I will elaborate this point in detail in **Chapter 2**.

The second issue is the depiction of the atomic details of the membrane bilayer, eg. the composition of lipids or dynamic changes in membrane curvature due to the interplay of membrane and transmembrane protein. A full treatment including the effect of membrane deformation and structure perturbation of the protein would compromise the speed of the simulation gained by CG model. Sacrificing the atomic details of the bilayer would prevent us from studying problems such as the mechanisms of ion selectivity, conduction, etc.

1.1.4 Putting it all together

Before my study, Dr. John Jumper in our group has developed a fast, atomic-level CG model, *Upside*, which is capable of *de novo* folding of proteins shorter than 100 residues in cpu-hours. *Upside* is a non-Gõ, physics-based model with five atoms per residue (N, Cα, C, H, O), a side chain bead and with residue- and neighbor-dependent Ramachandran maps. The energy function includes H-bonds, side chain-side chain and side chain-backbone interactions (including helix capping), and a solvation term. At each step, the side chain bead is first decorated to each of the residues. The positions of side chain beads are determined based on the joint probability of all side chain beads which gives the lowest global free energy for all side chains. The force is computed using the joint probability. Then, the side chain beads are undecorated while the forces are applied to the backbone atoms.

Using *Upside* as the simulation engine, I choose an statistical approach to derive a potential that can be integrated with *Upside* as the implicit solvent for transmembrane proteins. Statistical approaches have been demonstrated successful in assessing the residue-specific energies of insertion of helical transmembrane proteins into lipid bilayers. The potential serves as an membrane bilayer averaged over lipid membranes from different species and organelle locations. The potential is determined by the position of a residue embedded in the bilayer and its exposure status with respect to the surrounding environment.

Then, to simulate the forced unfolding of transmembrane proteins, I implement a force function in *Upside* that can apply force. Because *Upside* does not have an explicit time (i.e. the timescale has to be estimated via comparison of the folding rate in *Upside* folding simulations against that of all-atom MD simulations), I set a recorder for integration steps in order to excute the constant velocity pulling.

1.2 Contributions of this thesis

This dissertation contributes to studying the dynamics of transmembrane helical proteins using a coarse-grained model with an improved implicit membrane burial potential.

In **Chapter 2**, we advanced a new statistical membrane burial potential that incorporates some new features. Our potential is referenced to the interface region, rather than the bilayer center, to better capture the physiochemical properties of this region. In addition, we include a penalty for the presence of unsatisfied hydrogen bond donors and accepters in the lipid bilayer. Third, we address challenges inherent in devising implicit solvent models of properly accounting for the exchange of protein-lipid interactions for protein-protein interactions upon protein contact. Lastly, we integrated the membrane burial potential with our *Upside* algorithm. The combination offers a computationally efficient, flexible tool that lays the groundwork for a variety of simulations of membrane protein dynamics.

In **Chapter 3**, we developed an accurate and fast atomic-level simulation that allows us to simulate single-molecule force spectroscopy (SMFS) experiments to characterize the unfolding energy surface. We first conducted a variety of tests of our implementation of force in the *Upside* simulations before moving on to membrane protein systems. Then, we simulated hundreds of unfolding events of two model membrane proteins: bacteriorhodopsin (bR) and GlpG. In the simulations of bR, We reproduced the characteristic experimental features, including even the back-and-forth unfolding of single helical turns, and identified similar intermediate states. For GlpG, we explored different modes of force application and found that how force is applied alters the perception of the energy landscape. We observed multiple-step unfolding in diverse pathways starting from either terminus or the center of GlpG, when pulling laterally on the protein using a stiff cantilever which allows for restoration of springs after an unfolding event, emulating an AFM experiment. We also observed cooperative unfolding in a few or single all-or-nothing step of GlpG, when pulling laterally with a constant force, mimicking a magnetic tweezers measurement. Hence, the mode of force application strongly affects the observed unfolding cooperativity and identification of intermediates, important issues that should be considered when interpreting unfolding data and designing experiments. Our method can be employed to devise complicated *gedanken* pulling experiments that are as-yet impossible.

Chapters 2 and 3 are intended for publication, so they contain redundancies with other parts of the thesis.

Finally, **Chapter 4** concludes with future works that build on the methods and perspectives developed in this work. Future applications to membrane protein dynamics and SMFS are presented.

CHAPTER 2

A MEMBRANE BURIAL POTENTIAL WITH H-BONDS AND APPLICATIONS TO CURVED MEMBRANES AND FAST SIMULATIONS

This material presented in this chapter has appeared in Z. Wang, J. M. Jumper, S. Wang, K. F. Freed, and T. R. Sosnick, *A membrane burial potential with h-bonds and applications to curved membranes and fast simulations*, **Biophys. J.**, 115:1872-84, 2018 [48]. Copyright from *Biophysical Journal*, doi:10.1016/j.bpj.2018.10.012 (2018).

We use the statistics of a large and curated training set of transmembrane helical proteins (TMH) to develop a knowledge-based potential that accounts for the dependence on both the depth of burial of the protein in the membrane and the degree of side chain exposure. Additionally, the statistical potential includes depth-dependent energies for unsatisfied backbone hydrogen bond donors and acceptors, which are found to be relatively small, ∼2 RT. Our potential accurately places known proteins within the bilayer. The potential is applied to the mechanosensing MscL channel in membranes of varying thickness and curvature, as well as to the prediction of protein structure. The potential is incorporated into our new *Upside* molecular dynamics (MD) algorithm. Notably, we account for the exchange of protein-lipid interactions for protein-protein interactions as helices contact each other, thereby avoiding overestimating the energetics of helix association within the membrane. Simulations of most multimeric complexes find that isolated monomers and the oligomers retain the same orientation in the membrane, suggesting that the assembly of pre-positioned monomers presents a viable mechanism of oligomerization.

2.1 Introduction

Helical transmembrane proteins perform numerous cellular functions, including uptake of nutrients, transport of membrane-impermeable molecules, ion balance, signal transduction, intercellular communication, immune response, making helical transmembrane proteins prime targets for drug activity [1]. Despite their biological importance, membrane proteins are greatly under-represented in structural databases. Only ∼2% of the deposited structures in the Protein Data Bank (PDB) are membrane proteins [3], for which ∼85% are helical transmembrane proteins [4, 5]. By comparison, ∼30% of the human genome encodes membrane proteins [6, 7]. This scarcity results from the relative difficulty in obtaining high-resolution structures experimentally, making it even more challenging to obtain information regarding dynamic behaviors, for instance, of conformational transitions between outward-facing and inward-facing states of membrane transporters [3, 12].

Considerable effort has been devoted to studying helical transmembrane proteins computationally [3] using static methods for predicting 3D structures from sequence, e.g. homology modeling [13], fold recognition [14, 15], and *de novo* prediction [16–21]. Dynamical methods including all-atom simulations and coarse grained (CG) models have been applied as well [22–27]. Energy functions are essential prerequisites for properly discriminating native from non-native models and for characterizing the interactions between proteins and their surrounding environment **(Fig. 2.1)**. Scoring functions may be developed either on physics- [16, 24, 49–55], learning- [25, 56–58], or knowledge-based considerations [59–72]..

Statistical approaches enjoy success in assessing the residue-specific energies of insertion of helical transmembrane proteins into lipid bilayers. Their use traces back to the first study in the 1980s of the distributions of positively charged residues in bacterial inner membranes [61]. Later, with more available structures, statistical potentials derived from empirical amino acid (AA) distributions have described experimentally observed trends, for example, the preference of positively charged residues in the cytoplasm [73], the snorkeling of Trp and Tyr to the membrane surface [74], and the bias of small side chains to reside at the helix-helix interfaces [75]. Statistical potentials have been used in analyzing for hydropathy [65], orienting proteins in membrane [65, 68–71], discriminating protein-protein interaction sites [69], characterizing topology of different types of membrane-associated peptides [70], recapitulating

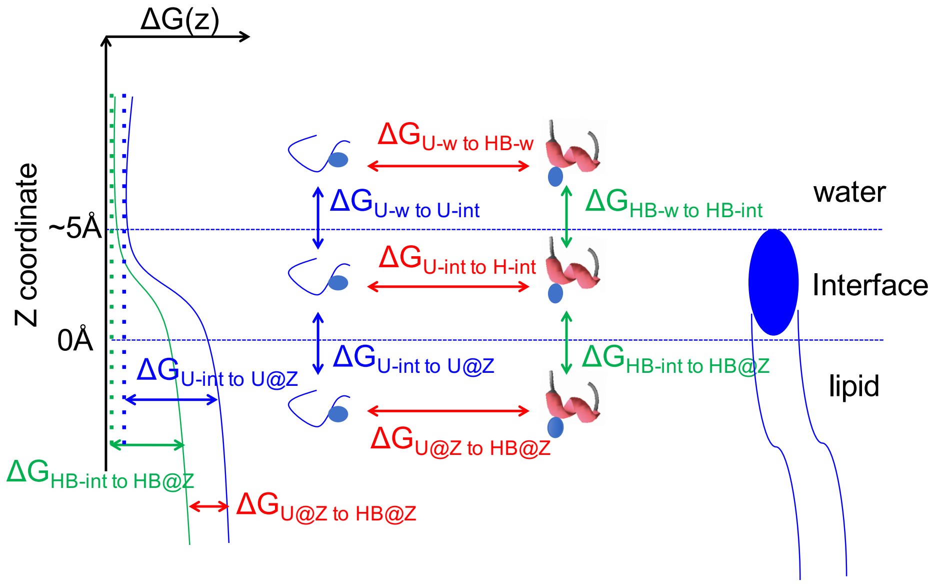
ΔG(z)
Z coordinate
~5Å
0Å
ΔG$_{\text{U-w to HB-w}}$
ΔG$_{\text{U-w to U-int}}$
ΔG$_{\text{HB-w to HB-int}}$
water
ΔG$_{\text{U-int to H-int}}$
Interface
ΔG$_{\text{U-int to U@Z}}$
ΔG$_{\text{U-int to U@Z}}$
ΔG$_{\text{HB-int to HB@Z}}$
lipid
ΔG$_{\text{U@Z to HB@Z}}$
ΔG$_{\text{HB-int to HB@Z}}$
ΔG$_{\text{U@Z to HB@Z}}$

Figure 2.1: **The major steps for the transfer of a residue into the lipid bilayer.** The U, HB in the subscripts of the ΔG's refer to unfolded, and H-bonded, respectively; w, int, @Z stand for water, interface, and at depth Z, respectively. Copyright from *Biophysical Journal*.

experimental data for the effects of mutation on dual-topology of proteins [70], estimating membrane bilayer thickness [71], and refining models for helical transmembrane proteins [71].

The present study focuses on deriving an improved knowledge-based potential for burial of proteins and on employing this potential for curved membranes and for molecular dynamics simulations. Energies are determined from the statistics of a large training set of proteins (curated for lipid exposure) and account for both the depth in the membrane and the level of side chain exposure to the lipid, i.e. E(Z, exposure) $\propto$ ln(frequency) **(Fig. 2.2AB)**. We also incorporate depth-dependent energies for unsatisfied backbone H-bond donors and acceptors within the bilayer **(Figs. 2.1 and 2.2C)**. Validation of the potential is based on its ability to locate proteins within the bilayer at the same position as that determined using the Orientations of Proteins in Membranes (OPM) approach, which employs an implicit solvation model of the lipid bilayer [76]. Our new statistical potential, named *UChiMemPot*, outperforms prior statistical potentials because of the use a better curated training set, inclusion of a dependence on membrane exposure and thickness, and the introduction of energies for the burial of unsatisfied hydrogen bonding groups. *UChiMemPot* may be used to rank and place homology models, to identify the membrane thickness and curvature, and thereby to provide the lowest insertion energy for a given protein conformation. For example, we find that the closed state of the MscL mechanosensitive channel [77] always has a lower insertion energy than the open state, but the difference is reduced in thinner membranes.

We conclude with the incorporation of the membrane burial potential in our new molecular dynamics (MD) algorithm *Upside*, that can reversibly fold soluble proteins containing up to 100 residues without the use of fragments or homology [78, 79]. In order to apply *Upside* to helical transmembrane proteins, the lipid-protein interactions within the membrane are updated dynamically to avoid the inclusion of protein-lipid interactions when these interactions are displaced by protein-protein interactions **(Fig. 2.2E)**. An analysis of the position of isolated monomers from multimeric complexes suggests that their docking provides a viable assembly strategy for many helical transmembrane proteins. However, monomers from some ion channels experience larger movement, and their assembly may involve conformational changes, possibly by an induced-fit mechanism.

2.2 Derivation

2.2.1 Training set & test set

Structures and bilayer thickness are obtained from the OPM website [76] and used in fitting data and training parameters. The dataset includes all proteins in the helical polytopic superfamilies as of January 2017 [4] and is culled by resolution (≤ 3 Å) and sequence identity ($\leq 25\%$) using *PISCES* [80], resulting in the retention of 171 helical transmembrane proteins. The proteins are ordered according to size and partitioned into five equal groups, with four used for training (140 proteins) **(Fig. 2.3A)** and one as the test set (31 proteins). The chain length distributions in the test and the training sets are similar, as are the number of residues

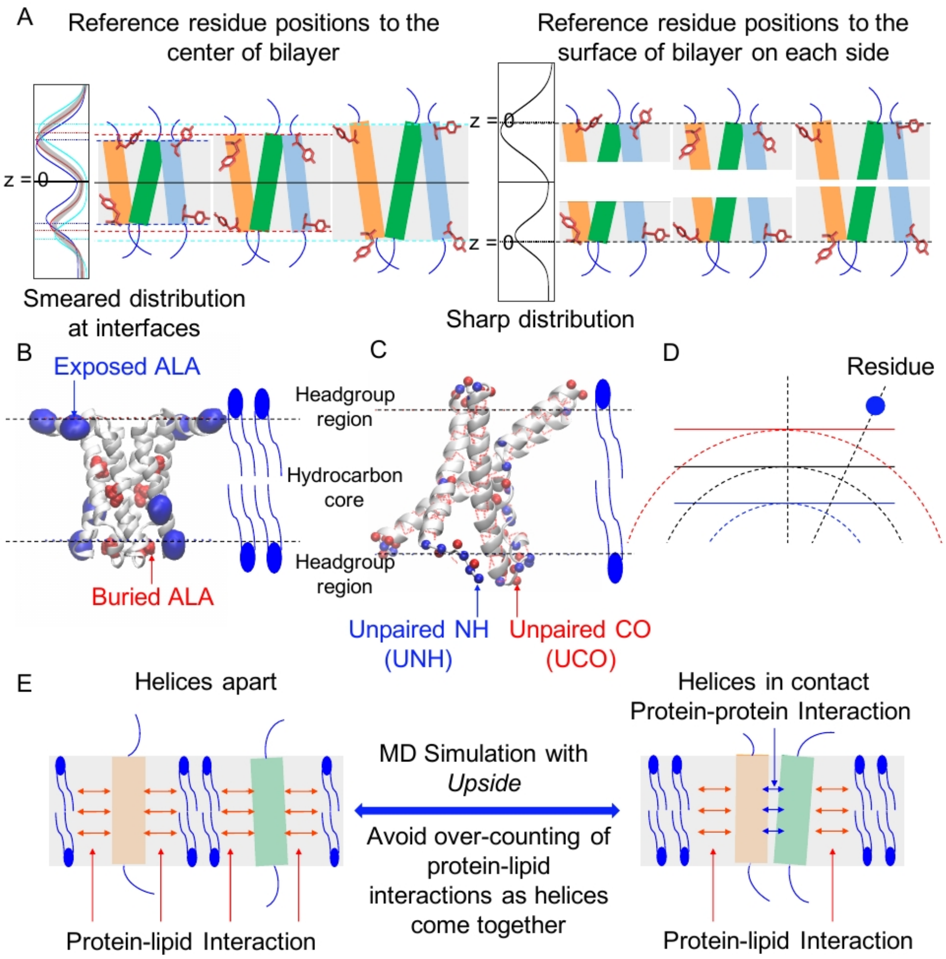
A
Reference residue positions to the center of bilayer
Reference residue positions to the surface of bilayer on each side
z = 0
z = 0
z = 0
Smeared distribution at interfaces
Sharp distribution
B
Exposed ALA
Buried ALA
C
Headgroup region
Hydrocarbon core
Headgroup region
Unpaired NH (UNH)
Unpaired CO (UCO)
D
Residue
E
Helices apart
Helices in contact
Protein-protein Interaction
MD Simulation with *Upside*
Avoid over-counting of protein-lipid interactions as helices come together
Protein-lipid Interaction
Protein-lipid Interaction

Figure 2.2: **Features of the *UChiMemPot* potential.**
(A) Our statistical potential (left) is referenced to lipid headgroup interface (z=0) rather than the center of the membrane bilayer (left). We separate the proteins halves and moved them apart in order to align to the interfaces (right). Our profiles are derived separately for the outer and the inner leaflets and concatenated in the middle to ensure continuity. **(B)** Protein burial status illustrated with the SWEET transporter 4qnd.pdb). The lipid-exposed and protein-buried alanines are depicted with larger blue spheres and smaller red van der Waals spheres, respectively. **(C)** We include unsatisfied backbone H-bond donors and acceptors in the potential, highlighted with the sensor domain of potassium channel KvAP (1ors.pdb) using the VMD definition. Hydrogen bonds are shown using dashed red lines. Unsatisfied hydrogen bond donor and accepters, NH (UNH, blue) and CO atoms (UCO, red), often are in loops or kinks. **(D)** A schematic of planar and spherical bilayers. **(E)** To apply the Upside MD package, we include lipid-protein interactions and energies of backbone exposure within the membrane, and do so in a dynamic manner to remove the energies of protein-lipid interactions as helices come together. The residue indices of UNHs and UCOs are as follows. UNH, 22 residues: 0,1, 2, 4, 5, 16, 29, 31, 34, 61, 63, 64, 65, 67, 71, 72, 75, 79, 81, 97, 130, 131. UCO, 27 residues: 0, 2, 3, 15, 25, 26, 27, 29, 30, 31, 57, 60, 61, 62, 65, 68, 69, 72, 73, 77, 92, 93, 126, 127, 128, 129, 130.
Copyright from *Biophysical Journal*.

in each of the 5 groups.

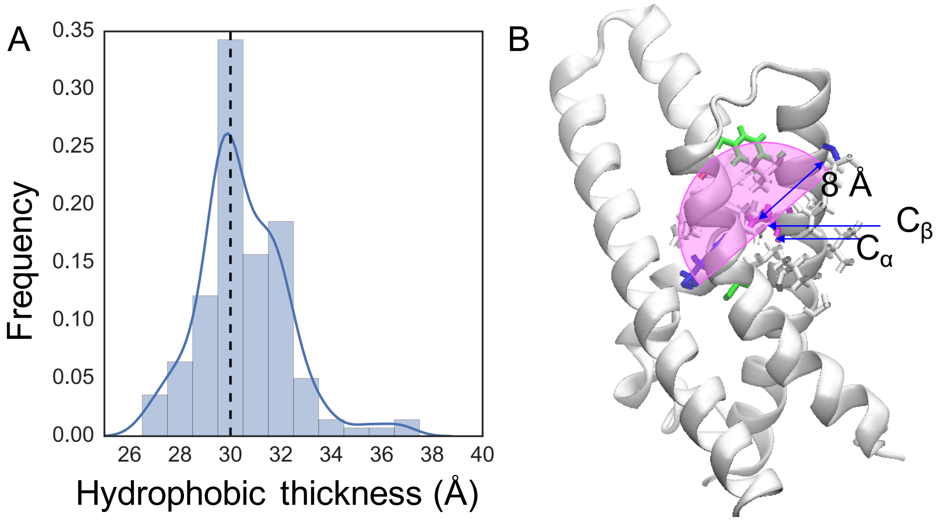

Figure 2.3: **Bilayer thickness and side chain burial status employed in our statistical potential. (A)** Hydrophobic thickness distribution of 140 TMHs in the training set. 30 Å is the most probable hydrophobic thickness, supporting the previous choice of 30 Å as the invariant hydrophobic thickness for referencing residue positions to the bilayer center. However, the variance argues that the statistical potential should allow for a variable hydrophobic thickness, using the bilayer surface as the reference depth. **(B)** The burial status of a given residue is calculated as the number of heavy atoms (from the non-neighboring residues) buried in the hemisphere located at the C_β atom and is pointing along C_α-C_β (illustrated with 1ors.pdb). Copyright from *Biophysical Journal*.

2.2.2 Feature extraction

Incomplete and missing side chains are added to the backbone using the VMD plugin, *Psfgen* [81]. The residue properties that are extracted from the protein structure include the residue type, secondary structure (SS), residue burials, solvent accessible surface area (SASA), H-bonds, and atomic coordinates of C_α, C_β, and backbone N and O, features that are included in data fitting and training. The computation of SS and SASA use methods in the Python library, *MDTraj* [82].

Residue burial

The residue burial is defined in terms of the number of heavy atoms in a hemisphere with an 8 Å radius, centered at the C$_\beta$ and directed along the C$_\alpha$-C$_\beta$ vector [83] (**Fig. 2.3B**). *Upside* provides the computing engine for calculate residue burial. The strong correlation between our metric and the standardly used SASA, supports our procedure for identifying exposure levels (**Fig. 2.4**). A residue is considered to be buried in the protein core when its burial exceeds a residue-determined value, chosen empirically, such that ∼25% of the Arg, Asn, Asp, Gln, Glu, and Lys residues and ∼50% of the other residues are assigned as buried. The high correlation between the residue burial method we employ (counting nearby heavy atoms) and SASA validates our procedure for identifying whether a residue is exposed to the surround solvent or buried in the protein interior.

We used *Upside* as the computing engine to calculate the residue burial, which precise definition is excerpted directly from Dr. John M. Jumper's PhD thesis as follows.

We define a count of surrounding residues in a similar manner to [83] but modify the construction so that it is differentiable.

The main component of the environment interaction is to count the number of side chains beads within a fixed radius of the C_β atom in a hemisphere above the atom. We define

$$b_i = \sum_j p_{\chi_j}(\phi_j, \psi_j) sigmoid\left(\frac{|x_{ji}| - 8\text{Å}}{1\text{Å}}\right) sigmoid\left(\frac{d_i^{C_\beta} \cdot \hat{x}_{ji} + 0.1}{1.0}\right) \tag{2.1}$$

$$x_{ji} = x_j^{SC} - x_i^{C_\beta} \tag{2.2}$$

Where $x_i^{C_\beta}$ is the position of the C_β on residue i and $d_i^{C_\beta}$ is the corresponding C_α-C_β bond vector. The sigmoid parameters are chosen to approximately maximize distinctiveness of the b_i burial distributions for different residue types. In a well-formed definition of burial, the burial distribution for a hydrophobic residue like valine should be very distinct from the burial distribution of a charged residue like aspartic acid.

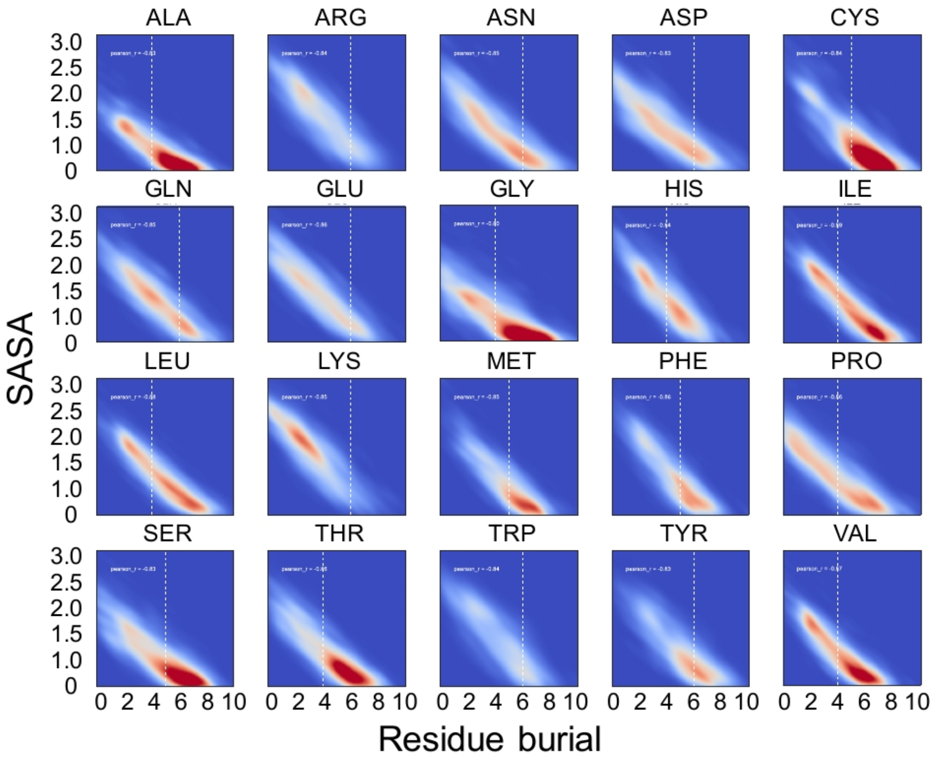

Figure 2.4: **Correlation between SASA and residue burial for each type of amino acid.** For each subplot, the x-axis and the y-axis stand for the residue burial and the SASA, respectively. The white dashed line marks the empirically chosen threshold for residue burial, above which the residue is considered buried in the protein core. Copyright from *Biophysical Journal*.

The side chain probability $p_{\chi_j}(\phi_j, \psi_j)$ *appearing in the definition Eq. 2.1 is the prior probability of the side chain bead, not the marginal probability from belief propagation. The derivative of the marginal probabilities with respect to the side chain positions are complex, much moreso than the derivative of the free energy with respect to the side chain positions. Furthermore, for intellectual self-consistency, the side chain rotamers ensemble should account for burial interactions. Unfortunately, belief propagation is not defined for many-body interactions, such as would arise from a nonlinear function of the burial bi. It would be possible to extend belief propagation to handle many-body terms at a lower level of approximation, although it is not clear how accurate such an approximation would be and whether it would be guaranteed to converge. Due to the derivative difficulties of using the marginal probabilities* *without perturbing for burial and the intellectual difficulties of extending belief propagation to handle many-body terms, we simply avoid the issues by using only the prior probabilities to define the environment interactions. The precise probabilities used to define burial are unlikely to compromise the ability of the b_i to distinguish hydrophobic and hydrophilic residues.*

The burial calculation is implemented in the EnvironmentCoverage *class of Upside.*

H-bond

The energy function includes a contribution from the energetics of burying unsatisfied hydrogen bond donors and acceptors in the membrane. The hydrogen bonds (hydrogen donor-acceptor contacts) are classified into four categories:
(i) backbone-backbone -NH. . .O=C-,
(ii) backbone-side chain -NH. . .O-,
(iii) sidechain-backbone -H. . .O=C-,and
(v) side chain-side chain -H. . .O- (**Table 2.1**).

Two common definitions of the H-bonding geometry are compared (**Fig. 2.5**), namely, the Baker-Hubbard definition (distance(H-A) < 3.5 Å and $\angle$DHA $> 90.0°$) [84] and the VMD definition (distance(D-A) < 4.0 Å and $\angle$DHA $> 90.0°$) [81], where H, D, and A denote the hydrogen, the donor, and the acceptor, respectively. Both methods assign NH and OH as donors, and the acceptors are O and N. The distributions of distance(H-A), distance(D-A), and $\angle$DHA for the two definitions are similar, therefore we choose the VMD definition for the convenience in visualizing the H-bonds in VMD. Based on the distributions obtained at loose distance and angle cutoffs, we reinforce strict cutoffs: dist(D-A) < 3.5 Å and $\angle$DHA $> 105.0°$ (**Figs. 2.5, 2.6, 2.7, 2.8**).

Table 2.1: Possible donors and acceptors in each type of amino acid.

aa	donor	acceptor
ALA	H	O
ARG	H, HE, HH11, HH12, HH21, HH22	O
ASN	H, HD21, HD22	O, OD1
ASP	H	O, OD1, OD2
CYS	H	O
GLN	H, HE21, HE22	O, OE1
GLU	H	O, OE1, OE2
GLY	H	O
HIS	H, HD1, HE2	O, ND1, NE2
ILE	H	O
LEU	H	O
LYS	H, HZ1, HZ2, HZ3	O
MET	H	O
PHE	H	O
PRO		O
SER	H, HG	O, OG
THR	H, HG1	O, OG1
TRP	H, HE1	O
TYR	H, HH	O, OH
VAL	H	O

Atoms are selected following the topology file definition of the CHARMM force field top_all27_prot_lipid.inp [85].

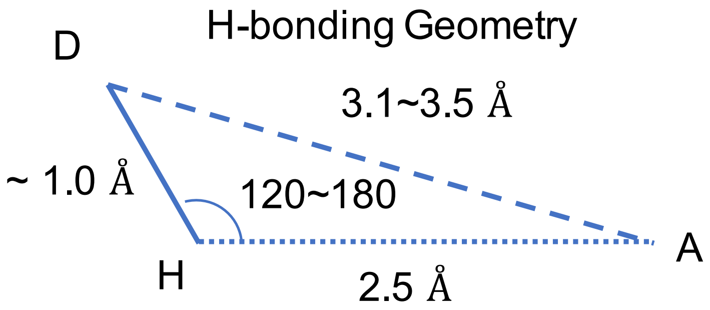

Figure 2.5: H-bond definitions. The Baker-Hubbard definition accepts an H-bond when dist(H-A) < distance cutoff and ∠DHA > angle cutoff and that VMD accepts an H-bond when dist(D-A) < distance cutoff and ∠DHA > angle cutoff. The donors considered by both methods are NH and OH, and the acceptors considered are O and N. Note: VMD uses hydrogens other than NH and OH as the donors. Copyright from *Biophysical Journal.*

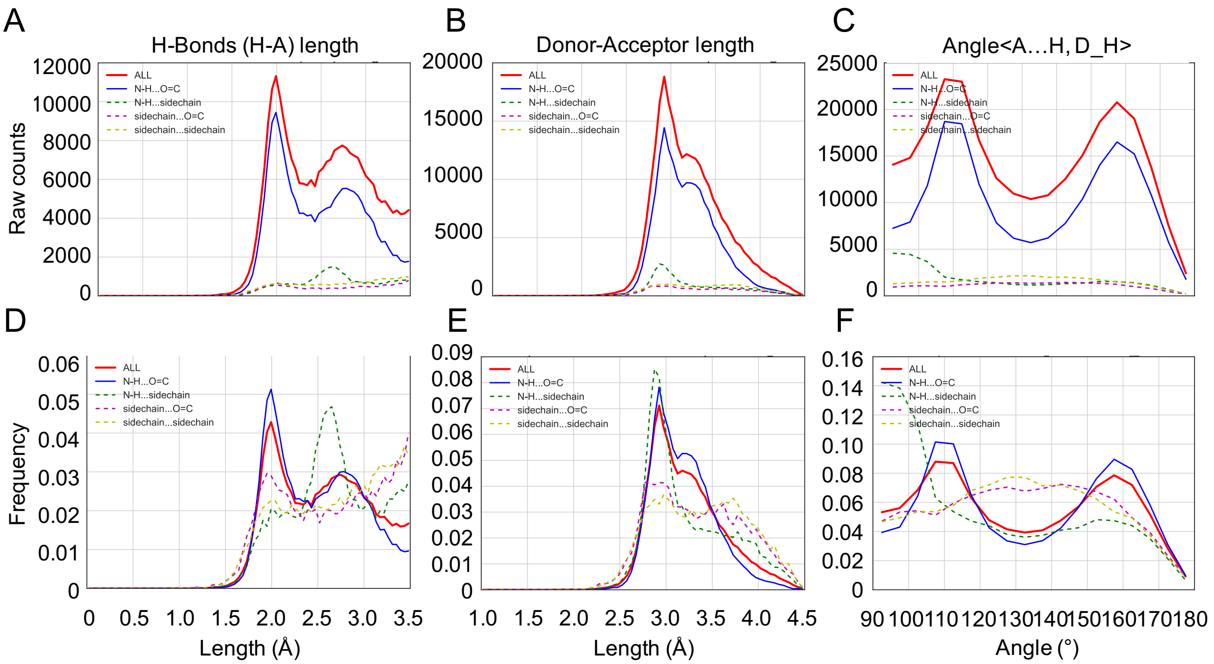

Figure 2.6: **Distributions of dist(H-A), dist(D-A), and ∠DHA of donor-acceptor contact pairs recognized by the Baker-Hubbard definition**. Here, we use loose distance and angle cutoffs: dist(H-A) $<$ 0.35 nm and ∠DHA $>$ 90.0°. Copyright from *Biophysical Journal*.

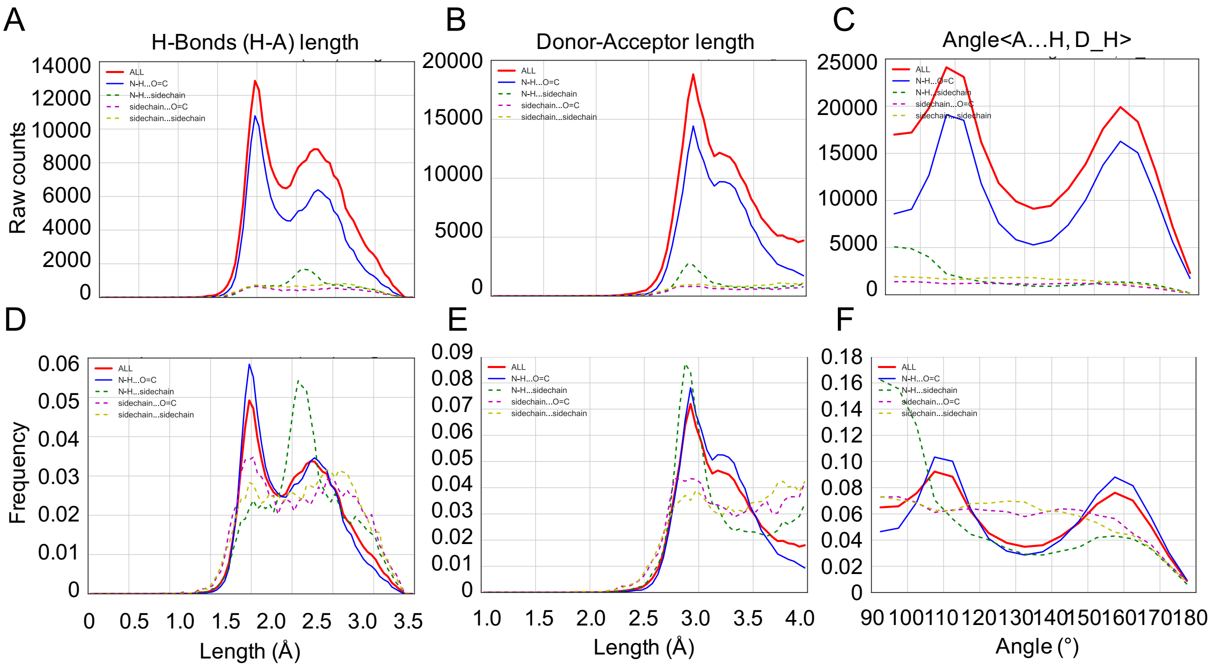

Figure 2.7: **Distributions of dist(H-A), dist(D-A), and ∠DHA of donor-acceptor contact pairs recognized by the VMD definition**. Here, we use loose distance and angle cutoffs: dist(D-A) $<$ 0.40 nm and ∠DHA $>$ 90.0°. Copyright from *Biophysical Journal*.

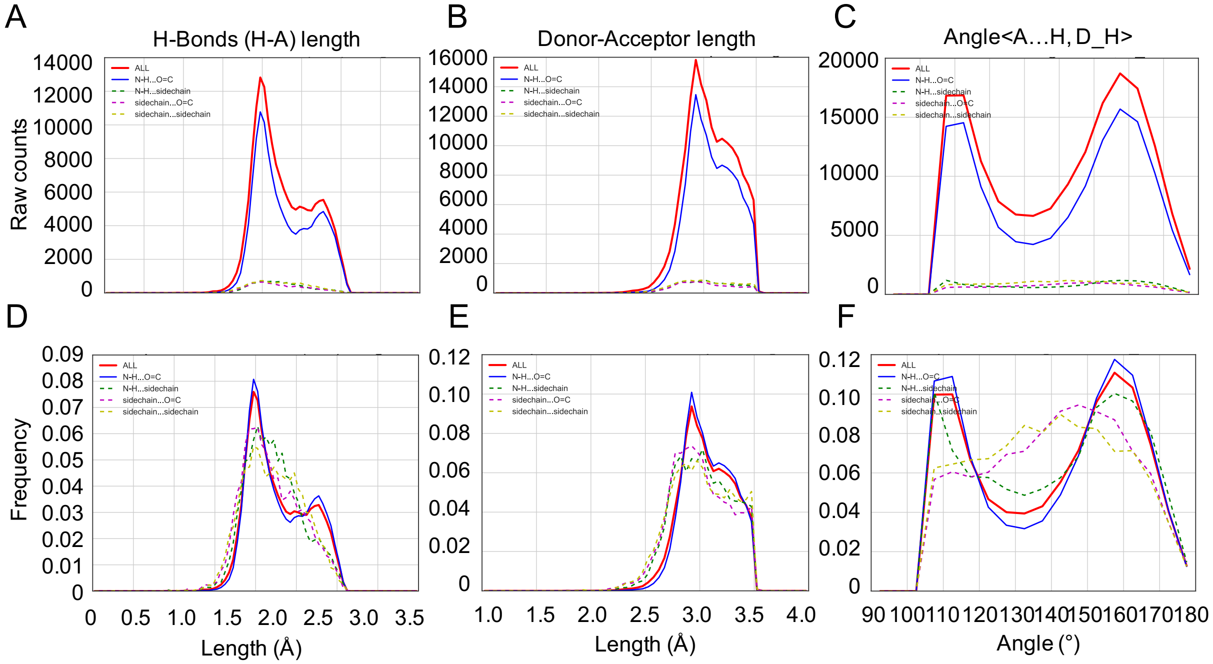

Figure 2.8: **Distributions of dist(H-A), dist(D-A), and ∠DHA of donor-acceptor contact pairs recognized by the VMD definition with strict cutoff**. Here, we use strict distance and angle cutoffs: dist(D-A) $<$ 0.35 nm and ∠DHA $>$ 105.0°. Copyright from *Biophysical Journal*.

2.2.3 Energy form & data fitting

The residue locations are defined in terms of the positions of the C_β (HA2 for Gly) atoms and are collected into 2 Å bins along the membrane normal z. Because individual helical transmembrane proteins may be embedded in bilayers with different hydrophobic thicknesses, residues with the C_β atoms lying in the range $[0, +\infty]$ and in $(-\infty, 0)$ are referenced to the interface of the outer leaflet and the inner leaflet, respectively **(Fig. 2.2A)**.

A depth-dependent potential profile for each type of AA with the residue burial $\leq$ the according threshold (**Table 2.2**) is derived first through **Eq. 2.3** to **Eq. 2.6**, in which $n^{aa}(z)$ is the number of a given type of amino acid in the bin [z-1, z+1] Å. In case there is no data point in the bin, value is extrapolated from neighboring bins. The raw histogram is smoothed first by cubic spline interpolation and then filtered with Savitzky-Golay filter [86]. In addition to standard AA interactions, we incorporate z-dependent energies for unpaired

NH and CO, which are termed as UNH and UCO. The 20 standard AA terms and the 2 unpaired H-bond terms are normalized separately.

Table 2.2: Residue burial threshold for each type of amino acid.

aa	Residue burial threshold	aa	Residue burial threshold
ALA	4.0	LEU	4.0
ARG	6.0	LYS	6.0
ASN	6.0	MET	5.0
ASP	6.0	PHE	5.0
CYS	5.0	PRO	4.0
GLN	6.0	SER	5.0
GLU	6.0	THR	4.0
GLY	4.0	TRP	6.0
HIS	4.0	TYR	6.0
ILE	4.0	VAL	4.0

$$P^{aa}(z) = n^{aa}(z) / \sum_{z} n^{aa}(z) \tag{2.3}$$

$$P^{ref}(z) = \sum_{aa} n^{aa}(z) / \sum_{aa} \sum_{z} n^{aa}(z) \tag{2.4}$$

$$Prop^{aa}(z) = P^{aa}(z) / P^{ref}(z) \tag{2.5}$$

$$E^{aa}_{histogram}(z) = -ln[Prop^{aa}(z)] = -ln\left[\frac{n^{aa}(z) \cdot \left(\sum_{aa} \sum_{z} n^{aa}(z)\right)}{\left(\sum_{z} n^{aa}(z)\right) \cdot \left(\sum_{aa} n^{aa}(z)\right)}\right] \tag{2.6}$$

Then, one Gaussian and one sigmoid are employed to fit the potential on each side. As the potential indicates the free energy difference in transferring a residue from the aqueous phase to the lipid bilayer, the potential profile is forced to be 0 at |z| = thickness/2 + 20 Å, where it is considered far away enough from the membrane.

Lastly, the potential profiles on two sides are concatenated at z = 0 Å to ensure continuity at the origin (**Figs. 2.9, 2.10, 2.11**). The potential profile of each AA is described by 7 parameters on each side through **Eq. 2.7** to **Eq. 2.13**. We used the limited-memory BFGS method [87] provided by the Python library *Scipy* [88] to fit the parameters and L2 regularization [89] to prevent overfitting. The derivative calculation in fitting was handled with Theano framework [90].

The inner and outer energy profiles are merged at z = 0 Å to ensure continuity at the origin. The number of residues of each type of AAs are listed in **Table 2.3**.

$$sigmoid(z; E, m, s) = E/\big[1 + exp\big(|s|(z - m)\big)\big] \tag{2.7}$$

$$gaussian(z; E, m, s) = E{\cdot}exp\big(- |s|(z - m)^2\big) \tag{2.8}$$

$$sig(z; p, forcezero) = sigmoid(z; p[0], p[1], p[2]) - sigmoid(forcezero; p[0], p[1], p[2]) \tag{2.9}$$

$$gau(z; p, forcezero) = gaussian(z; p[3], p[4], p[5]) - gaussian(forcezero; p[3], p[4], p[5]) \tag{2.10}$$

$$pot\_oneside(z; p, forcezero) = sig(z; p, forcezero) + gau(z; p, forcezero) - ln|p[6]| \tag{2.11}$$

$$dpot = pot\_oneside(-0.5{\cdot}thickness; p_{out}, 20) - pot\_oneside(0.5{\cdot}thickness; p_{in}, -20) \tag{2.12}$$

$$pot\_twoside(z; p_{out}, p_{in}, thickness) = \begin{cases} pot\_oneside(z - 0.5{\cdot}thickness; p_{out}, 20), z{\geq}0 \\ pot\_oneside(z + 0.5{\cdot}thickness; p_{in}, -20) + dpot, \\ z < 0 \end{cases} \tag{2.13}$$

Table 2.3: Number of residues of each type of AA involved in the potential derivation.

aa	In. exp.	In. bur.	In. total	Out. exp.	Out. bur.	Out. total
ALA	1533	3193	4726	1569	2919	4488
ARG	2040	150	2190	830	104	934
ASN	1070	285	1355	1103	239	1342
ASP	1057	189	1246	1055	102	1157
CYS	151	295	446	150	296	446
GLN	902	184	1086	908	206	1114
GLU	1360	190	1550	1066	141	1207
GLY	1502	2585	4087	1633	2505	4138
HIS	398	453	851	477	445	922
ILE	1333	2351	3684	1461	2192	3653
LEU	2439	3718	6157	2306	3269	5575
LYS	1780	100	1880	748	49	797
MET	657	821	1478	616	703	1319
PHE	1862	1273	3135	1778	1493	3271
PRO	1112	742	1854	1112	935	2047
SER	1478	1205	2683	1370	1195	2565
THR	876	1554	2430	1040	1537	2577
TRP	990	152	1142	1057	155	1212
TYR	1132	386	1518	1382	312	1694
VAL	1455	2351	3806	1461	2174	3635
total	25127			23122		

In. and Out. stand for the inner leaflet and outer leaflet of the membrane bilayer, respectively. Exp. and bur. are short for exposed to the surround solvent and buried in the protein core, respectively.

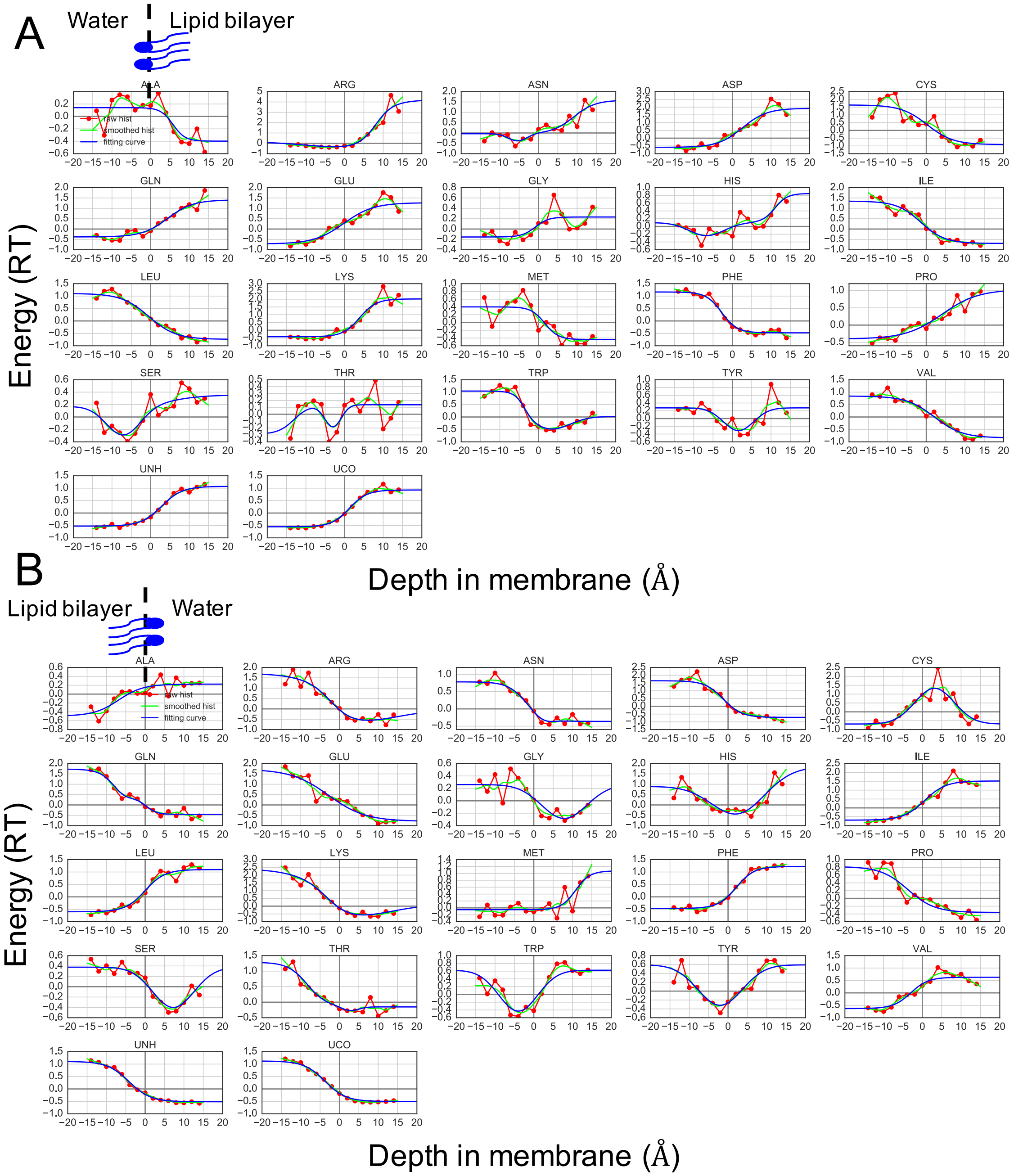

Figure 2.9: **Comparisons of raw histogram, smoothed histogram, and fitting curve of statistical potential profiles. (A-B)** For each amino acid and unsatisfied H-bond term, the raw histogram, smoothed histogram, and the fitting curve of potential profile on the cytoplasmic and periplasmic side are plotted. Copyright from *Biophysical Journal*.

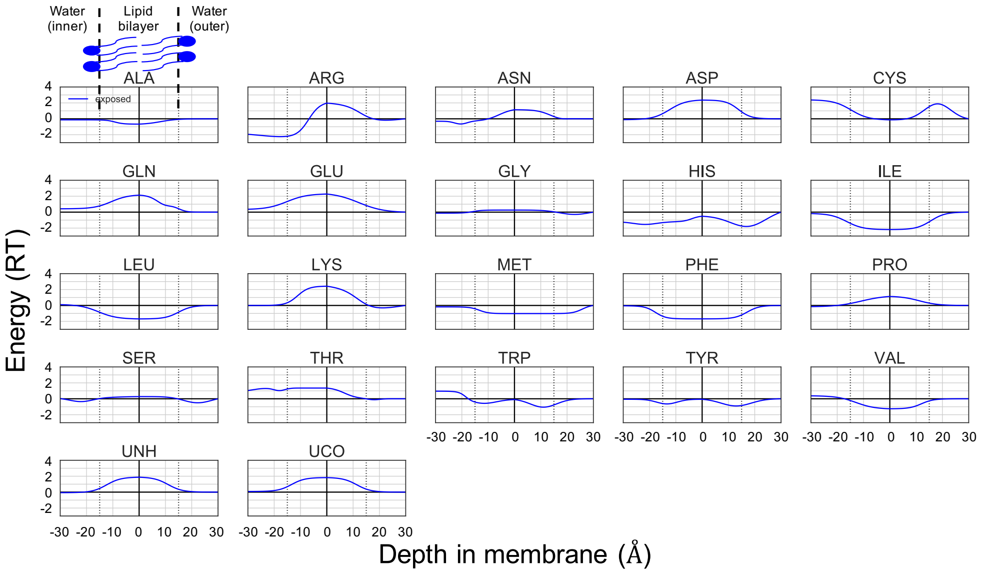

Figure 2.10: **Membrane burial potential profiles of the 20 amino acids**. Potential profiles of 20 amino acids and unsatisfied backbone H-bond donors (UNH) and acceptors (UCO). The potential profile at thickness = 30.0 Å is plotted. Dashed vertical lines delineate the inner and outer water-lipid interfaces. Copyright from *Biophysical Journal*.

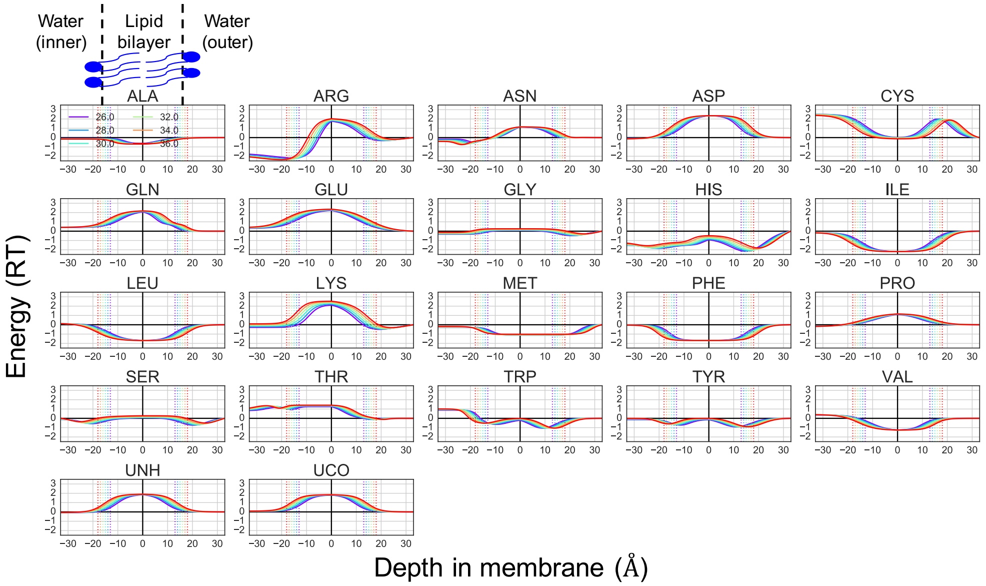

Figure 2.11: **Thickness dependence of potential profiles**. For each amino acid, potential profile at thickness = 26.0 Å, 28.0 Å, 30.0 Å, 32.0 Å, 34.0 Å, and 36.0 Å are plotted. Dashed vertical lines delineate the water-lipid interfaces on the two sides, which colors are in line with those of the potential curves. We can see how the potential profiles of 20 amino acids and unsatisfied backbone H-bond donor and acceptor change in accordance to the change of bilayer thickness. Copyright from *Biophysical Journal*.

2.3 Potential profiles

The bilayer thickness and position of most membrane protein structures are unknown because the structures are not determined in a native, biologically relevant bilayer. To obtain this information, we adopt the common practice of specifying both the protein position and the bilayer thickness with values providing the lowest insertion energy from the OPM website [76]. The method has been validated experimentally [91] and serves as our standard of nativeness for these two quantities. The OPM database treats the lipid bilayer as an anisotropic solvent in which the rigid protein floats in a hydrophobic slab of adjustable thickness [52]. The calculated lipid boundaries in OPM are located ~5 Å from the phosphate groups at the level of the carbonyl groups of the lipid molecules (**Fig. 2.2B**).

We adopt the convention where the lipid head group boundary is used as z=0 reference point to define the burial depth **(Figs. 2.2A, 2.12)**. This convention differs from the standard definition where the bilayer center defines the z = 0 point. We believe referencing to the bilayer boundary better accounts for variable bilayer thickness, as side chains are more sensitive to the distance from this boundary, where the physiochemical properties rapidly change, than to the distance from the relatively homogeneous center of the bilayer (e.g., placing a residue 1 or 5 Å from the boundary often produces a larger energetic difference than placing the residue 1 or 5 Å from the center of the bilayer). As a result, our potential exhibits sharper features near the boundary **(Fig. 2.2A)**. Because the potential is a measure of the energy difference in transferring a residue from the aqueous phase to the lipid bilayer, the potential is defined with E = 0 at |z| = thickness/2 + 20 Å, which is considered to be far enough from the bilayer to be considered to be in bulk solvent.

The frequency of occurrence of each residue type as a function of depth in the bilayer and exposure to lipids are used to calculate the 2D cross-membrane distribution of C_β atoms **(Fig. 2.12)**. The potential energy is calculated according to E = -RTln(frequency) using only the lipid-exposed side chains. The potentials for the inner and outer leaflets are fit separately to the sum of a Gaussian and a sigmoid (four functions in total). Additionally,

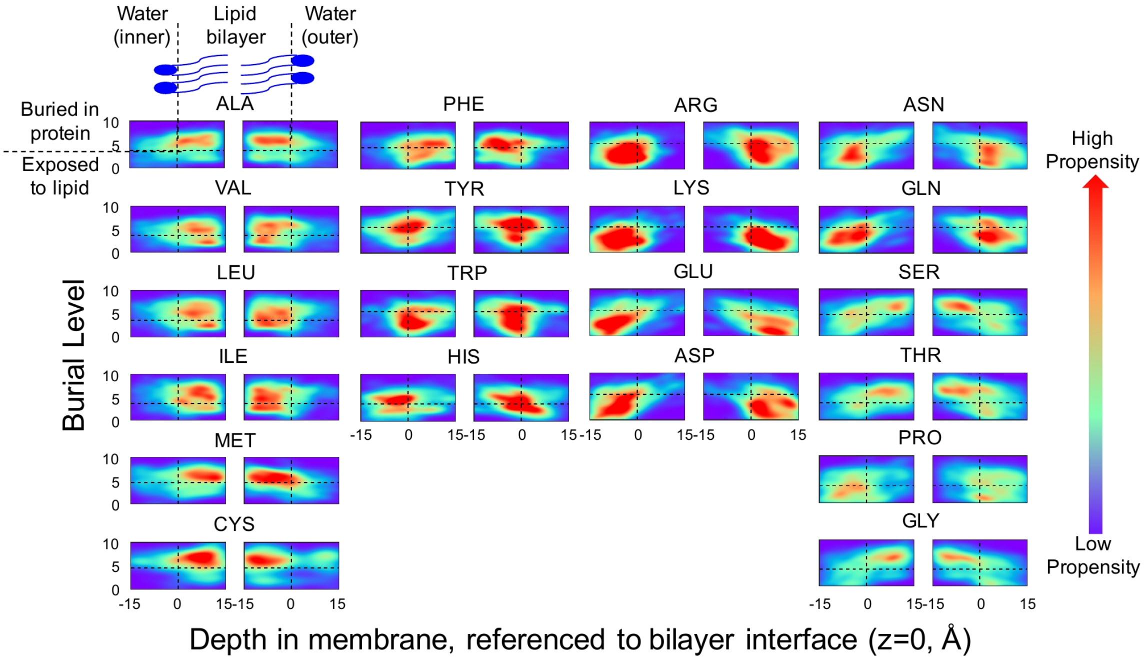

Figure 2.12: **Distributions of 20 amino acids**. The distribution of each amino acid (C_β position, HA2 for Gly) is calculated separately for the periplasmic and cytoplasmic leaflets. Dashed vertical line in each subplot marks the water-lipid interface; dashed horizontal line in each subplot separates the exposed from the buried residues, a difference used in the derivation of the potential. Copyright from *Biophysical Journal*.

the frequency distributions of unpaired amide and carbonyl groups are converted to energies irrespective of whether they are buried inside the protein or exposed to lipid.

The overall trends in the amino acid distribution are similar to previously observed distributions. Hydrophobic residues, such as Ala, Gly, Ile, Leu, Met, Phe, and Val, have a strong preference to be buried in the lipid bilayer. When Gly, Ala, Ser (and Thr to a lesser degree) are located below the bilayer boundary, they are more likely than other amino acids to be buried inside the protein than exposed to lipids. The bias for Gly and Ala may be due to their small size which facilitates more intimate contact between transmembrane helices. For instance, GxxxG motifs often play an important role in mediating interhelical interactions in helical transmembrane proteins [36]. Arg and Lys have higher tendencies than other hydrophilic residues to be situated near the membrane surface. This tendency reflects the ability of longer normally-charged side chains to snorkel towards the head group and aqueous phase [92].

The charged states of the acidic and basic residues (Arg, Asp, Glu, and Lys) are generally unspecified from the structure alone. Inherently, residue-level statistical potentials are agnostic on this issue, being only sensitive to the location of the residues. When the two acidic residues (pKa ∼4 in solution) are deeply buried in the membrane, charge neutralization by protonation is likely, as the energetic cost of an upward pKa shift and protonation is less than the cost of burying a charged group [93]. A similar argument holds for Lys [94], but the high pKa of Arg (∼14) may result in this side group remaining charged [95]. According to our statistical potential, it costs only 2-2.4 RT to transfer any of these four normally-charged residues from water to the center of a lipid bilayer. These values are much lower than those calculated even using sophisticated treatments of the cost of charge burial [96], and they are similar in magnitude. These observations support the proposal that these four amino acids have substantial pKa shifts and are charge neutral when buried deep in the membrane.

The transfer energy for an Asn is less than half that for an Asp (∼1 RT versus ∼2 RT). The transfer of a Gln is nearly the same as Glu (∼2 RT). Because few charged and polar residues lie within the lipid bilayer, the statistical error is higher for these residues. However, their burial energies do not contribute much to the overall burial energy because they are infrequently buried. The potential profiles indicate that each unpaired NH (UNH) and unpaired CO (UCO) in the middle of lipid bilayers has a burial penalty of ∼2 RT **(Fig. 2.10)**.

Both Trp and Tyr exhibit pronounced snorkeling in which a side chain extends towards the solvent, although Trp penetrates deeper into the lipid bilayer than Tyr. Exposed phenylalanines exhibit a mild asymmetric preference for the inner leaflet. The majority of cysteines are buried in the protein core forming disulfide bonds. Histidines are asymmetrically distributed, being depleted on the cytoplasmic side but enhanced on the periplasmic side. This difference is mainly due to metal ligation in respiratory and photosynthetic proteins [70].

2.4 Positioning into a flat membrane

To assess the ability of *UChiMemPot* to reproduce the orientations in OPM, grid searches in tilt and burial depth are performed to find the lowest energy orientations for the 31 proteins in the test set **(Fig. 2.13)**. Although a variety of knowledge-based burial potentials exist, we only compare ours to the relatively recent asymmetric Ez-3D C_β statistical potential [70] because other potentials are derived in a similar manner and the increase in the size of the training set likely accounts for much of the improvement in accuracy. *UChiMemPot* on average predicts a very similar orientation as OPM **(Fig. 2.14A, B)**. The standard deviations from the OPM orientation for the tilt and shift for 31 test cases are 5.3° and 1.2 Å, respectively, which are similar to the precisions of OPM (4.0° and 2 Å, respectively) [97]. When the unsatisfied H-bond (UHB) terms are included in *UChiMemPot*, the standard deviations are 3.0° and 1.4 Å, respectively. As a comparison, the standard deviations of tilt and shift for using the asymmetric Ez-3D potential are larger, 6.5° and 2.0 Å, respectively.

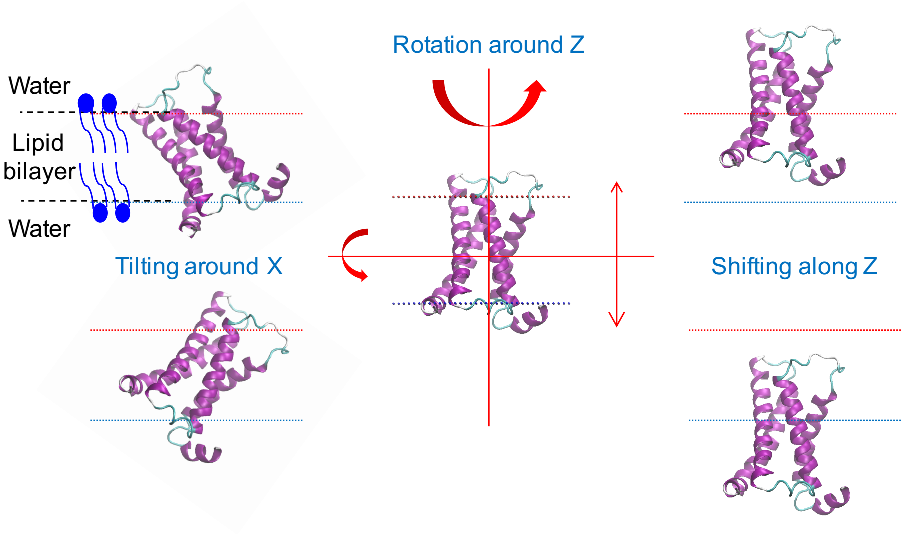

Figure 2.13: **Protocol of performing grid search**. A schematic of grid search protocol using voltage-sensor domain of phosphatase (PDB: 4g80). To conduct the grid search, the helical transmembrane protein is sequentially rotated around the z-axis, tilted away the x-axis, and shifted along the z-axis. The search is carried out exhaustively in a discrete manner and the insertion energy is calculated at each position with the energy denoted as E(rotation, tilt, shift). The global energy minimum is chosen as best orientation. The OPM orientation is used as the initial position. Copyright from *Biophysical Journal*.

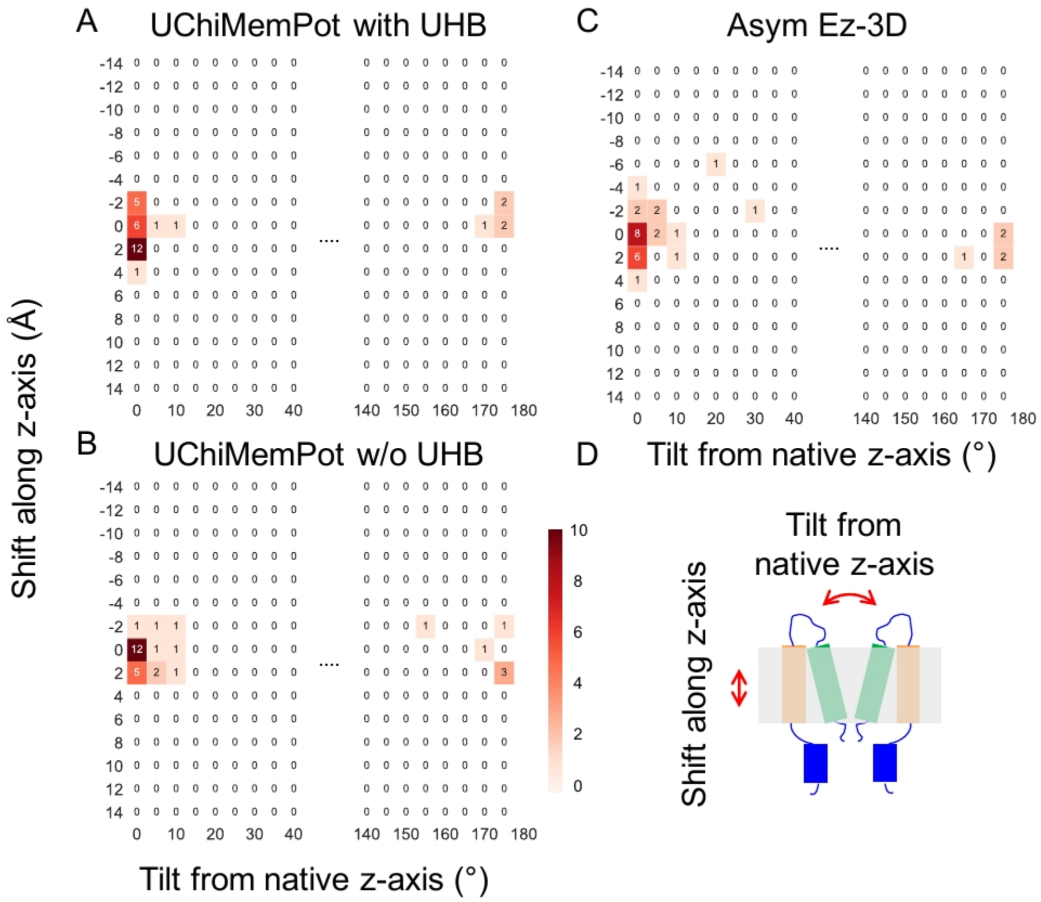

Figure 2.14: **Protein positions with the lowest membrane insertion energies**. Tilt and shift distributions of the lowest energy orientations for the 31 proteins in the test set. The three potentials applied are: the *UChiMemPot* potential **(A)** with or **(B)** without UHB penalties, and the **(C)** asymmetric Ez-3D C_β potential. Non-zero values are colored for visualization purposes. **(D)** An illustration of the relative shift and tilts from the native z-axis. Copyright from *Biophysical Journal*.

Because *UChiMemPot* uses a training set nearly twice as large as that of the asymmetric Ez-3D C_β potential, it is expected to outperform the older potential **(Fig. 2.14C)**. The asymmetric Ez-3D C_β potential predicts five inverted orientations, whereas *UChiMemPot* without the UHB terms predicts six. The asymmetric Ez-3D potential does not fare as well for tilt angles; two cases yield deviations in excess of 20°, compared to none from *UChiMemPot*. Upon addition of the UHB terms, the quality of *UChiMemPot* further improves the prediction of the tilt and asymmetry, with only two cases having a deviation in the tilt angle exceeding 5°, although the number of inverted orientations remains at five. An examination of the energy landscape indicates that the flipped conformations are nearly isoenergetic with the correct conformation **(Fig. 2.15)**.

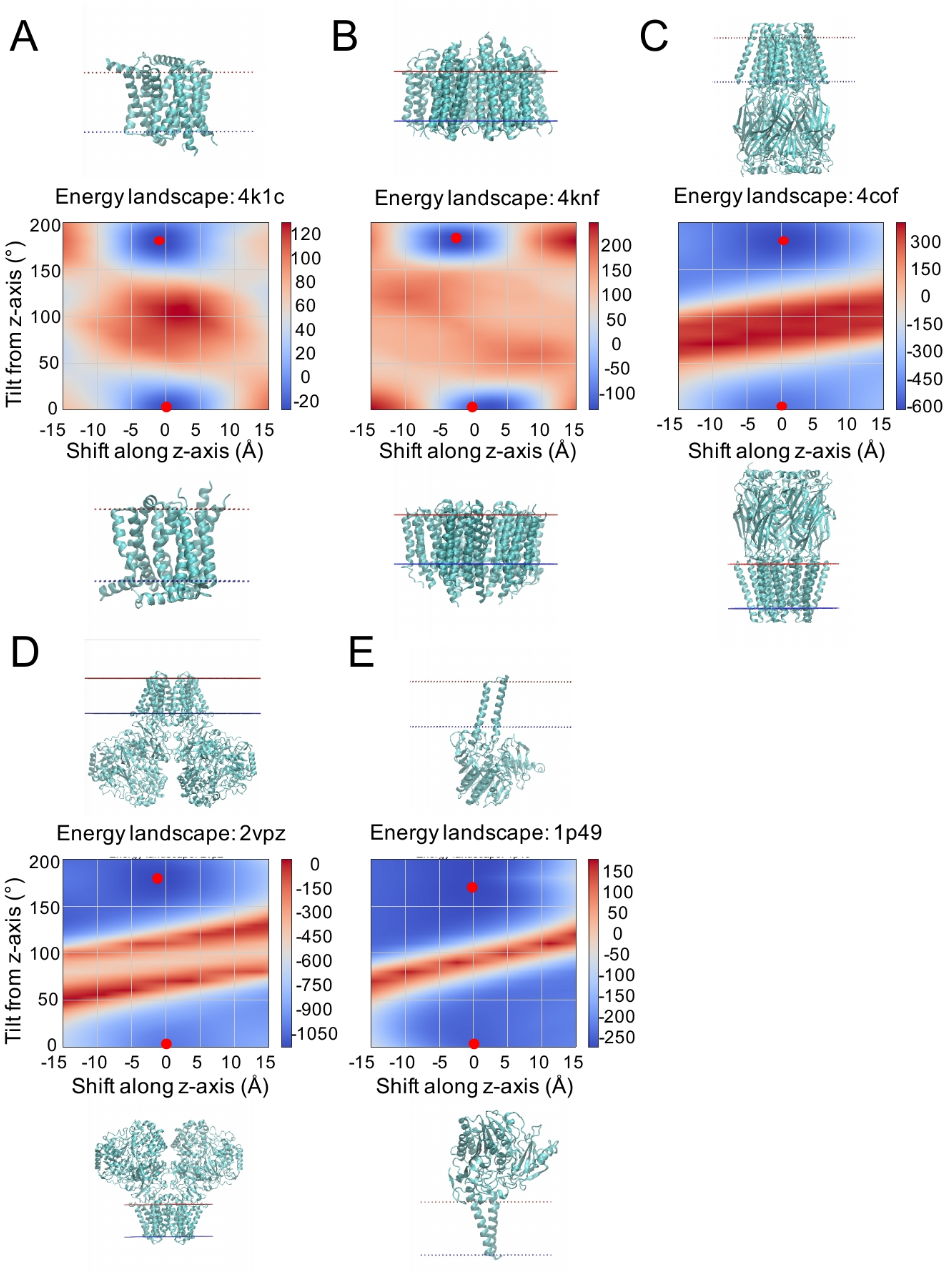

Figure 2.15: **Energy landscape produced by grid search**. **(A-E)** Energy surfaces with two basins representing up and down orientations calculated with our potential plus the UHB effect. Two energetic basins are observed for each of the pseudo energy landscape, one located at the OPM-defined orientation and the other at the flipped orientation. The global minimum energy state is represented by the red dot in the upper basin; the red dot in the lower basin representing the local minimum energy state. Copyright from *Biophysical Journal*.

2.5 Bilayer thickness with lowest insertion energy

The lowest insertion energy is used to distinguish the optimal hydrophobic thickness of the bilayer for a given protein. As a first example, *UChiMemPot* identifies the lowest energy of the 518-residue GLUT3 glucose transporter (4zw9A) occurring at a bilayer thickness of 33.8 Å, similar to OPMs value of 32.8±1.4 Å. **Table 2.4** presents ten more comparisons which generally agree with OPM predictions (average difference is 1.9 Å, with only one protein (4od4A) differing by more than 3 Å from OPMs value). Our identification of the membrane thickness producing the lowest insertion energy does not imply that any given membrane/lipid composition actually adopts this thickness.

Table 2.4: Prediction of optimal bilayer thickness based only on insertion energy.

PDB name	PDB id	*UChiMemPot* (Å)	OPM(Å)
GLUT3 glucose transporter	4zw9A	33.8	32.8±1.4
Protease GlpG (*Haemophilus influenzae*)	2nr9A	25.4	28.2±1.6
Bile acid sodium symporter ASBT	3zuxA	25.8	27.8
Biotin transporter BioY	4dveA	29.6	31.8±1.2
Nickel/cobalt transporter CbiM	4m58A	28.4	29.8±0.9
C-C chemokine receptor type 5	4mbsA	34.0	34.4±1.5
Protease GlpG (*E. coli*)	4njnA	25.6	28.6
4-hydroxybenzoate octaprenyltransferase	4od4A	30.8	27.4
Protein YetJ	4pgrA	31.4	29.8±2.0
Rhodopsin I	5awzA	30.2	32.4±2.2
ECF transporter, S-component	5d0yA	30.4	29.8±1.8

Error bars shown when provided by OPM website.

2.6 Identification of decoys in structure prediction

Here we test whether *UChiMemPot*s insertion energy can be used to select more native-like decoys (lower C_α-RMSD) generated by modern structure prediction methods. We begin with the example of the GLUT3 glucose transporter (4zw9A), containing 12 transmembrane helices. The recent Area-Under-Curve or AUC-maximized DeepCNF method utilizes evolutionarily-determined contacts that are identified from sequence covariation [98, 99] and produces a best model that is quite good with a TM-score = 0.74 [100], a C_α-RMSD = 6.0 Å **(Fig. 2.16)**, and the lowest membrane insertion energy, although higher than the native structure. Individual residues have similar burial energies in the best model as in the native protein.

To broaden the test, we generated additional decoys of poorer quality by randomly removing a fraction of the pairwise contacts (between 10 and 100%) and replacing them with an equal number of randomly chosen contacts. This decoy set yields a strong correlation between RMSD and burial energy (Pearson correlation coefficient = 0.944). Ten additional cases are summarized in **Figs. 2.17, 2.18, 2.19, 2.20, 2.21, 2.22, 2.23, 2.24, 2.25, 2.26**. In general, the native structure has a low or the lowest insertion energy compared to predicted models, and lower C_α-RMSD correlates with lower insertion energy. These promising results suggest that *UChiMemPot* (or another burial potential) could be incorporated in the future as part of the DeepCNF energy function in order to produce better models.

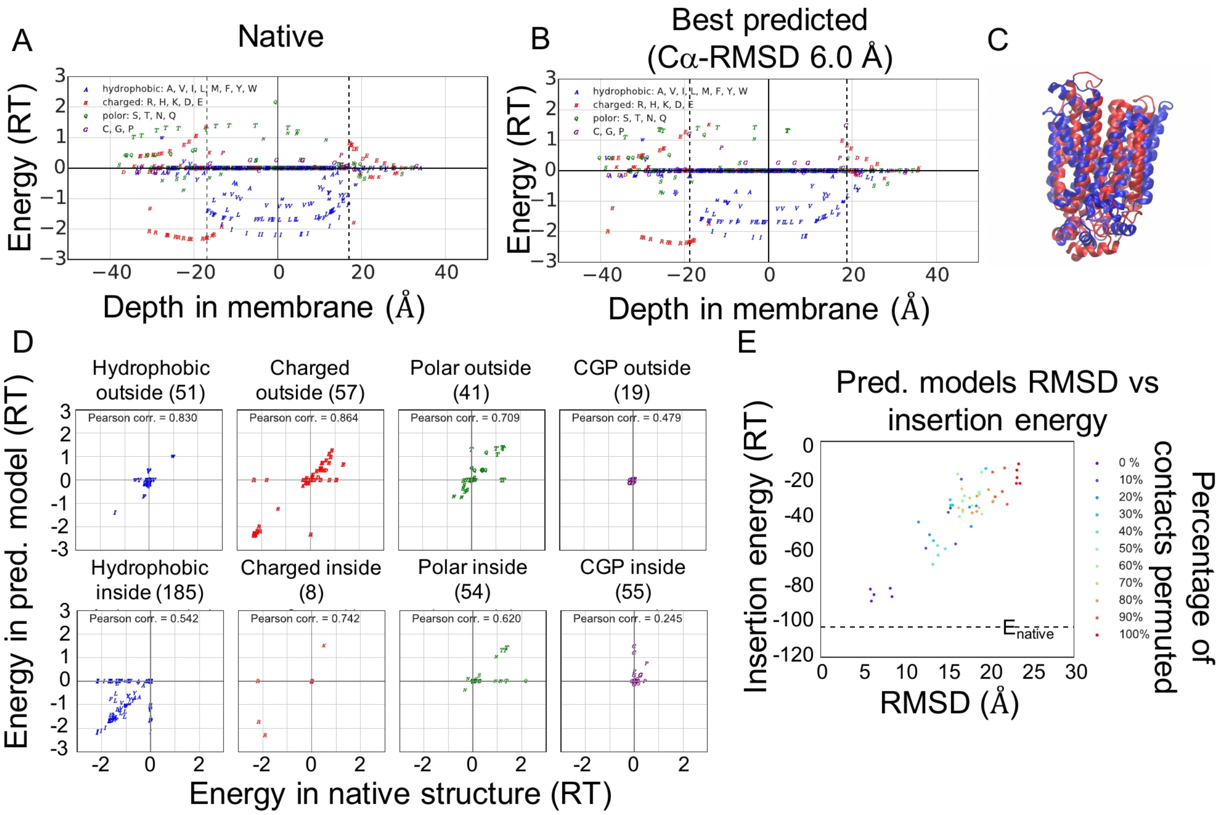

Figure 2.16: **Decomposition of residue burial energies for the known 518-residue human GLUT3 glucose transporter structure (4zw9A) and models predicted by DeepCNF**. **(A, B)** Plots of residue position versus membrane burial energy using *UChiMemPot*. Residues buried in the protein interior have zero value and are plotted on the x-axis. Vertical dashed lines delineate the boundary of membrane bilayer after insertion. **(C)** Structural comparison between the native structure (in blue) and the best predicted model (in red). **(D)** Correlation in the burial energy between residues in the native structure and the best predicted model. Numbers in the parenthesis are the numbers of residues in that category (e.g., there are 51 hydrophobic residues located outside the membrane). **(E)** Total insertion energy of predicted models versus C_α-RMSD. The horizontal dashed line is the insertion energy of the native structure. The contact map is predicted using AUC-maximized DeepCNF and used to generate models. Fifty additional decoys for testing are obtained by removing a random fraction from 10-100% of the pair-wise contacts and replacing them with an equal number of randomly chosen contacts (five models are generated at every 10% increment). Copyright from *Biophysical Journal*.

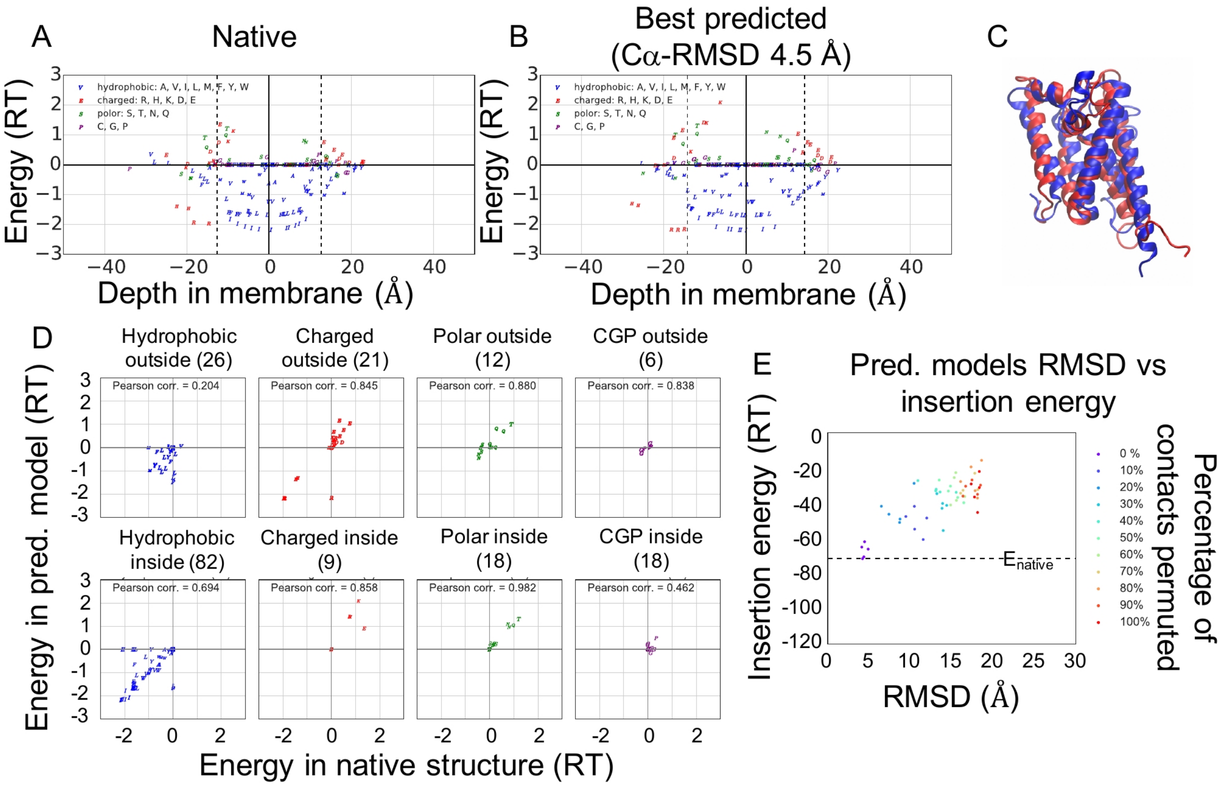

Figure 2.17: **Energetic decomposition for the known native 2nr9A and models predicted by DeepCNF**. 2nr9A is Protease GlpG that has 192 residues and 6 transmembrane helices. The best model predicted by AUC-maximized DeepCNF has TM-score 0.770 and C_α-RMSD 4.47. The Pearson correlation coefficient between the RMSD and the insertion energy is 0.906. Copyright from *Biophysical Journal*.

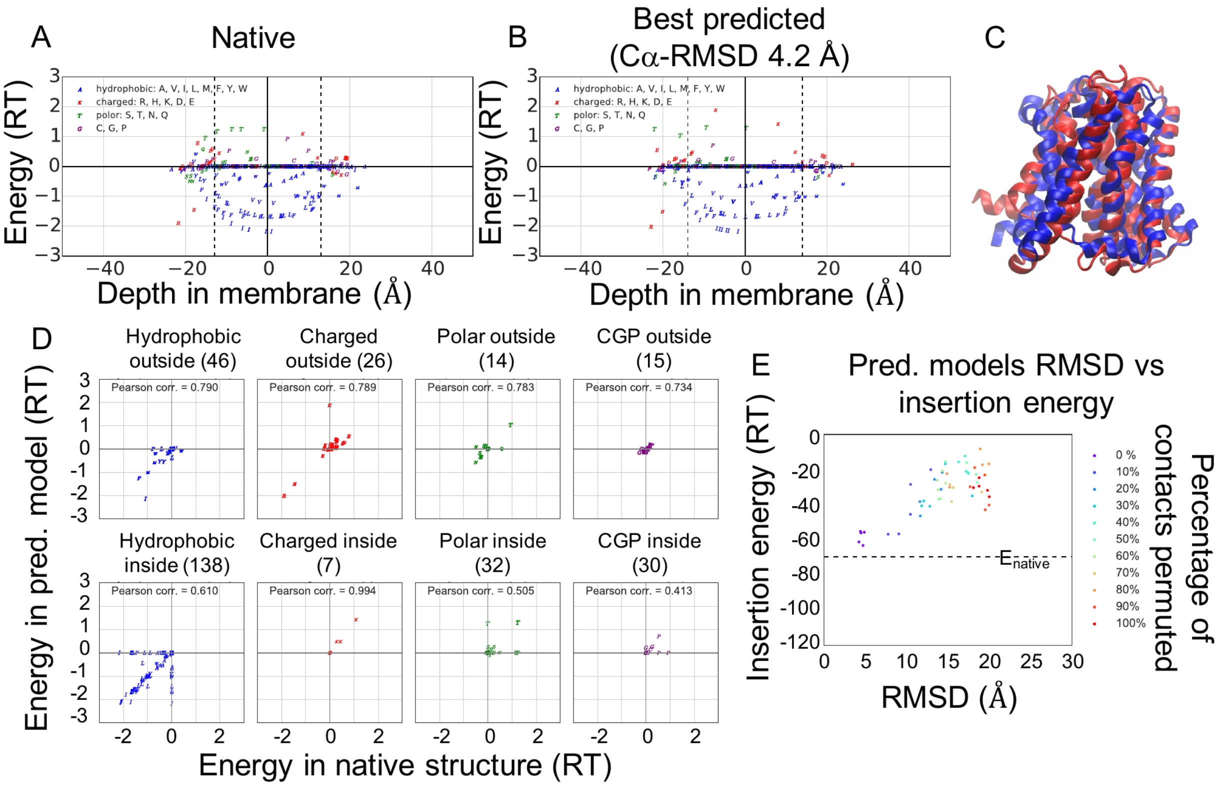

Figure 2.18: **Energetic decomposition for the known native 3zuxA and models predicted by DeepCNF**. 3zuxA is Bile acid sodium symporter ASBT that has 308 residues and 10 transmembrane helices. The best model predicted by AUC-maximized DeepCNF has TM-score 0.731 and C_α-RMSD 4.22. The Pearson correlation coefficient between the RMSD and the insertion energy is 0.858. Copyright from *Biophysical Journal*.

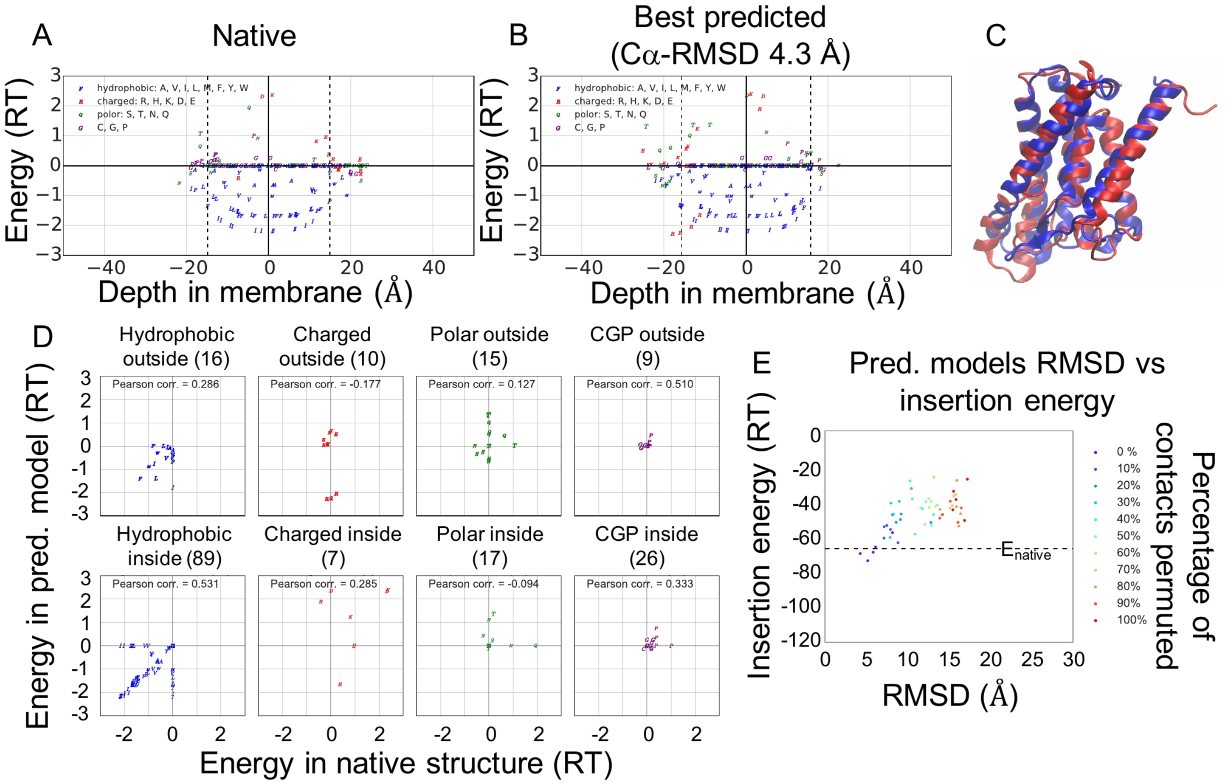

Figure 2.19: **Energetic decomposition for the known native 4dveA and models predicted by DeepCNF**. 4dveA is Biotin transporter BioY that has 189 residues and 6 transmembrane helices. The best model predicted by AUC-maximized DeepCNF has TM-score 0.505 and C_α-RMSD 4.26. The Pearson correlation coefficient between the RMSD and the insertion energy is 0.813. Copyright from *Biophysical Journal*.

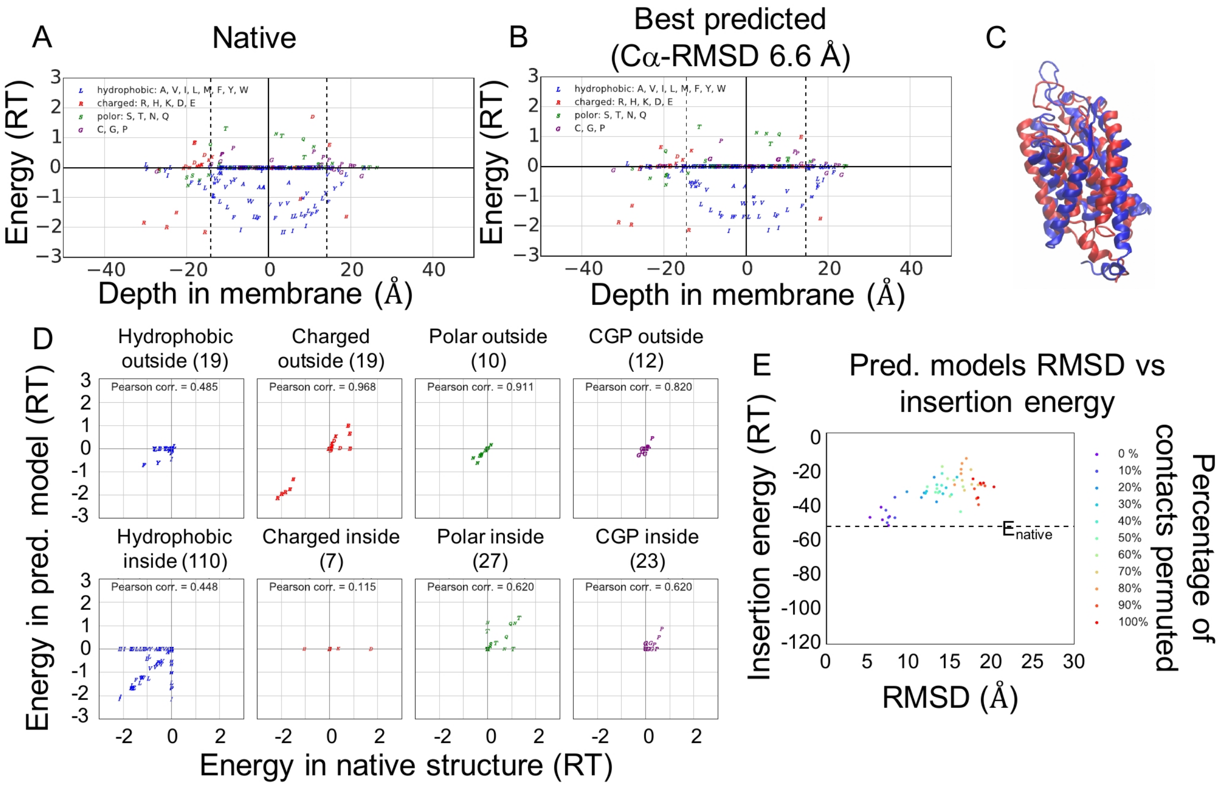

Figure 2.20: **Energetic decomposition for the known native 4m58A and models predicted by DeepCNF**. 4m58A is Nickel/cobalt transporter CbiM that has 227 residues and 7 transmembrane helices. The best model predicted by AUC-maximized DeepCNF has TM-score 0.529 and C_α-RMSD 6.62. The Pearson correlation coefficient between the RMSD and the insertion energy is 0.810. Copyright from *Biophysical Journal*.

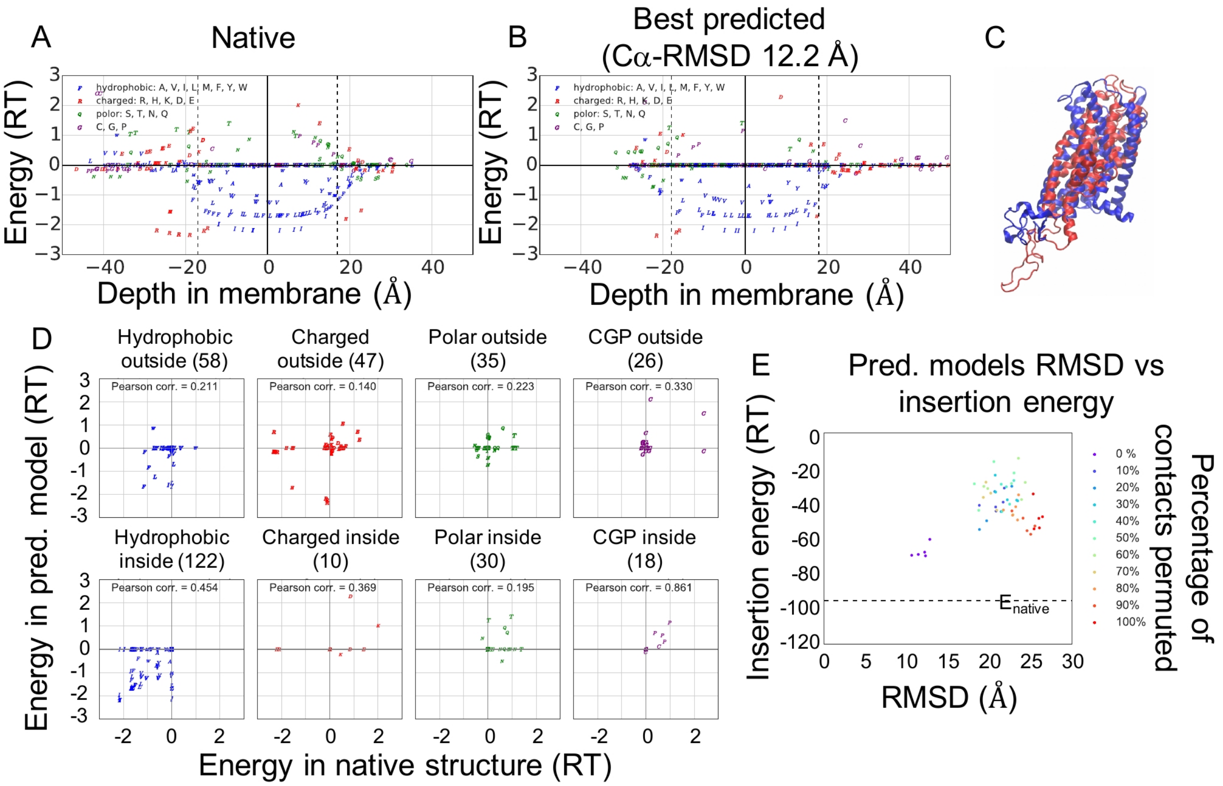

Figure 2.21: **Energetic decomposition for the known native 4mbsA and models predicted by DeepCNF**. 4mbsA is C-C chemokine receptor type 5 that has 346 residues and transmembrane helices. The best model predicted by AUC-maximized DeepCNF has TM-score 0.435 and C_α-RMSD 12.18. The Pearson correlation coefficient between the RMSD and the insertion energy is 0.476. Copyright from *Biophysical Journal*.

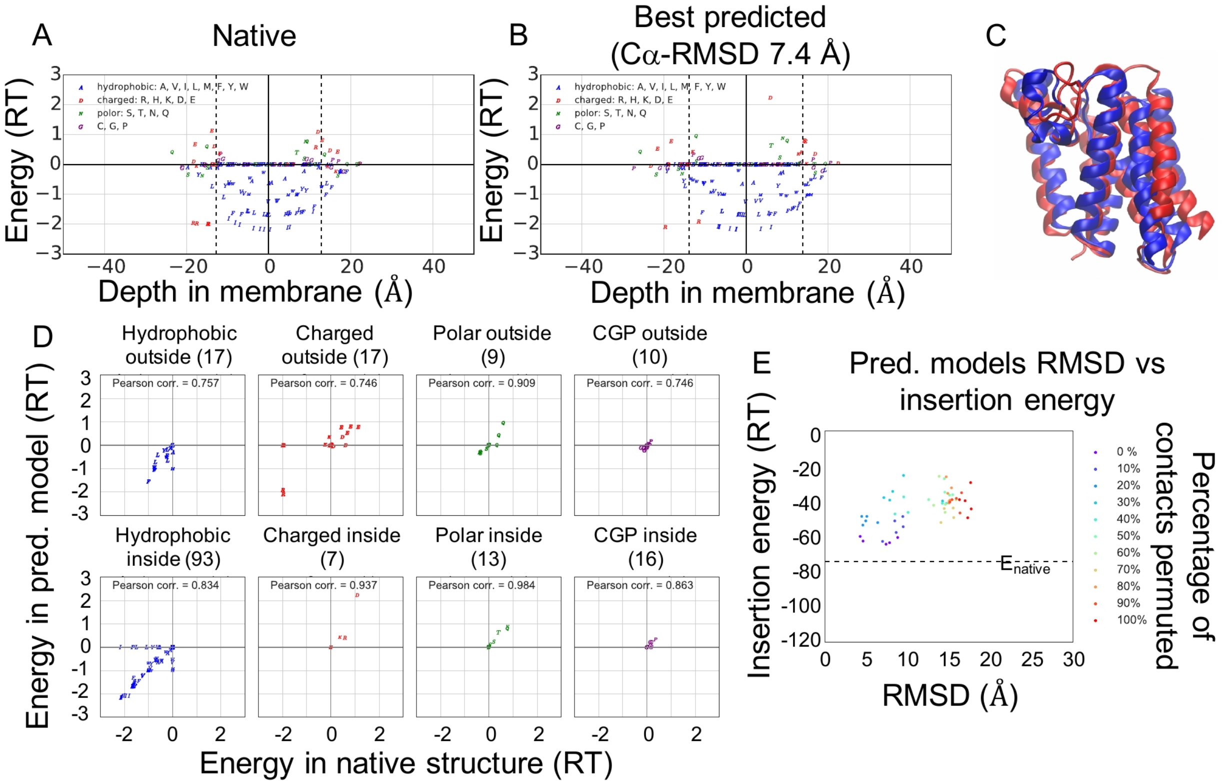

Figure 2.22: **Energetic decomposition for the known native 4njnA and models predicted by DeepCNF**. 4njnA is Protease GlpG that has 182 residues and 6 transmembrane helices. The best model predicted by AUC-maximized DeepCNF has TM-score 0.579 and C_α-RMSD 7.38. The Pearson correlation coefficient between the RMSD and the insertion energy is 0.685. Copyright from *Biophysical Journal*.

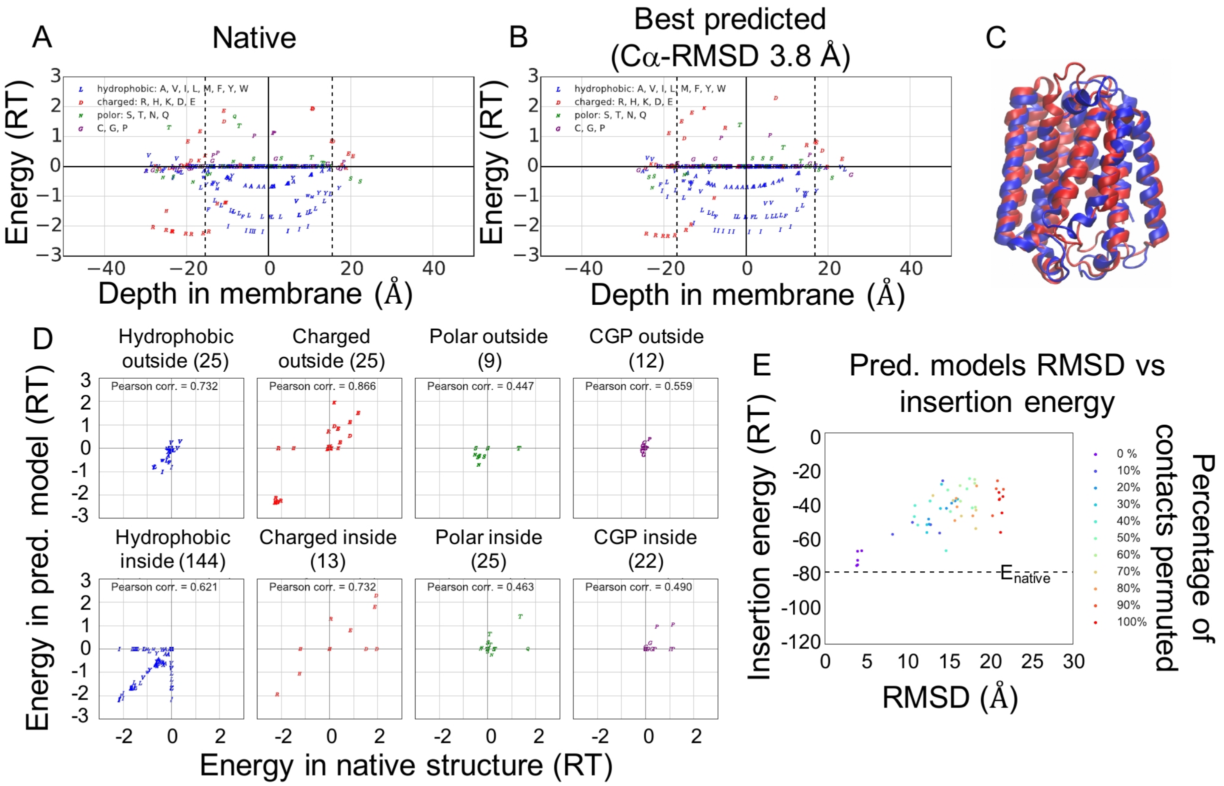

Figure 2.23: **Energetic decomposition for the known native 4od4A and models predicted by DeepCNF**. 4od4A is 4-hydroxybenzoate octaprenyltransferase that has 275 residues and 9 transmembrane helices. The best model predicted by AUC-maximized DeepCNF has TM-score 0.748 and C$_\alpha$-RMSD 3.83. The Pearson correlation coefficient between the RMSD and the insertion energy is 0.888. Copyright from *Biophysical Journal*.

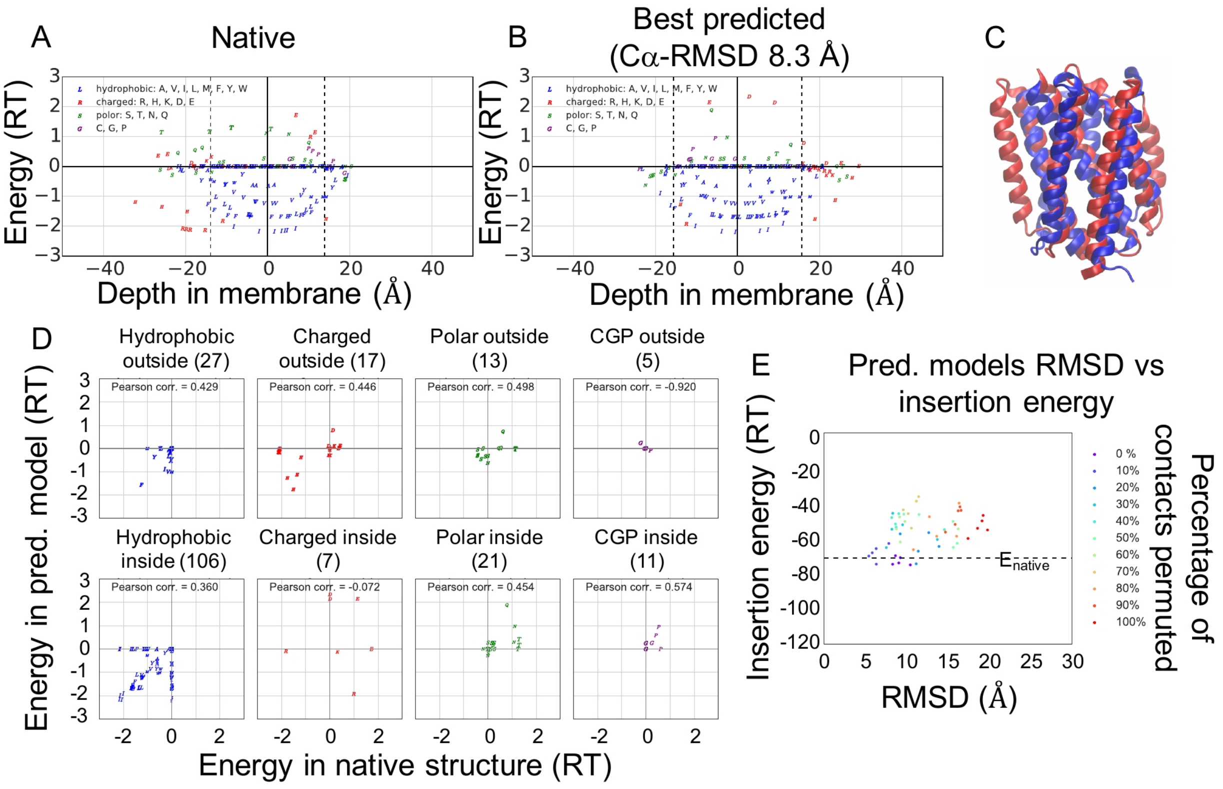

Figure 2.24: **Energetic decomposition for the known native 4pgrA and models predicted by DeepCNF**. 4pgrA is Protein YetJ in the closed conformation that has 207 residues and 7 transmembrane helices. The best model predicted by AUC-maximized DeepCNF has TM-score 0.552 and C$_\alpha$-RMSD 8.30. The Pearson correlation coefficient between the RMSD and the insertion energy is 0.474. Copyright from *Biophysical Journal*.

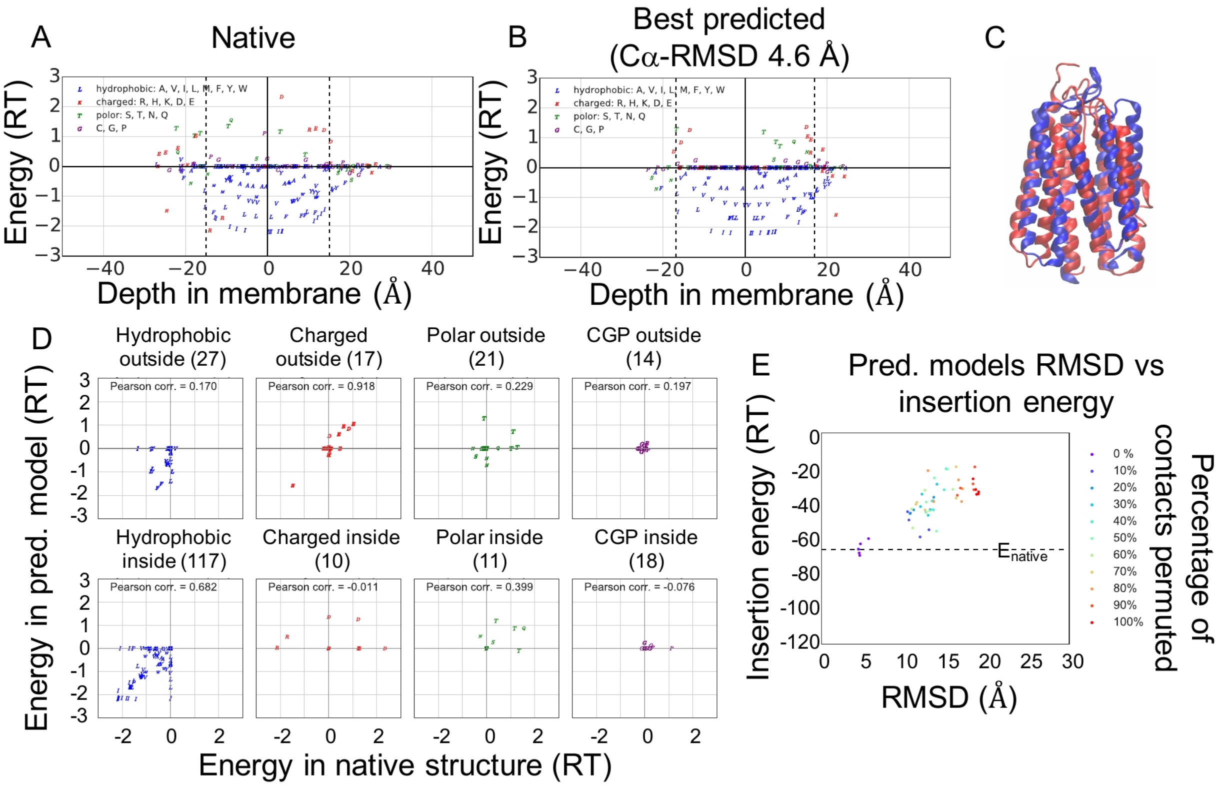

Figure 2.25: **Energetic decomposition for the known native 5awzA and models predicted by DeepCNF**. 5awzA is Rhodopsin I that has 235 residues and 7 transmembrane helices. The best model predicted by AUC-maximized DeepCNF has TM-score 0.723 and C_α-RMSD 4.63. The Pearson correlation coefficient between the RMSD and the insertion energy is 0.943. Copyright from *Biophysical Journal*.

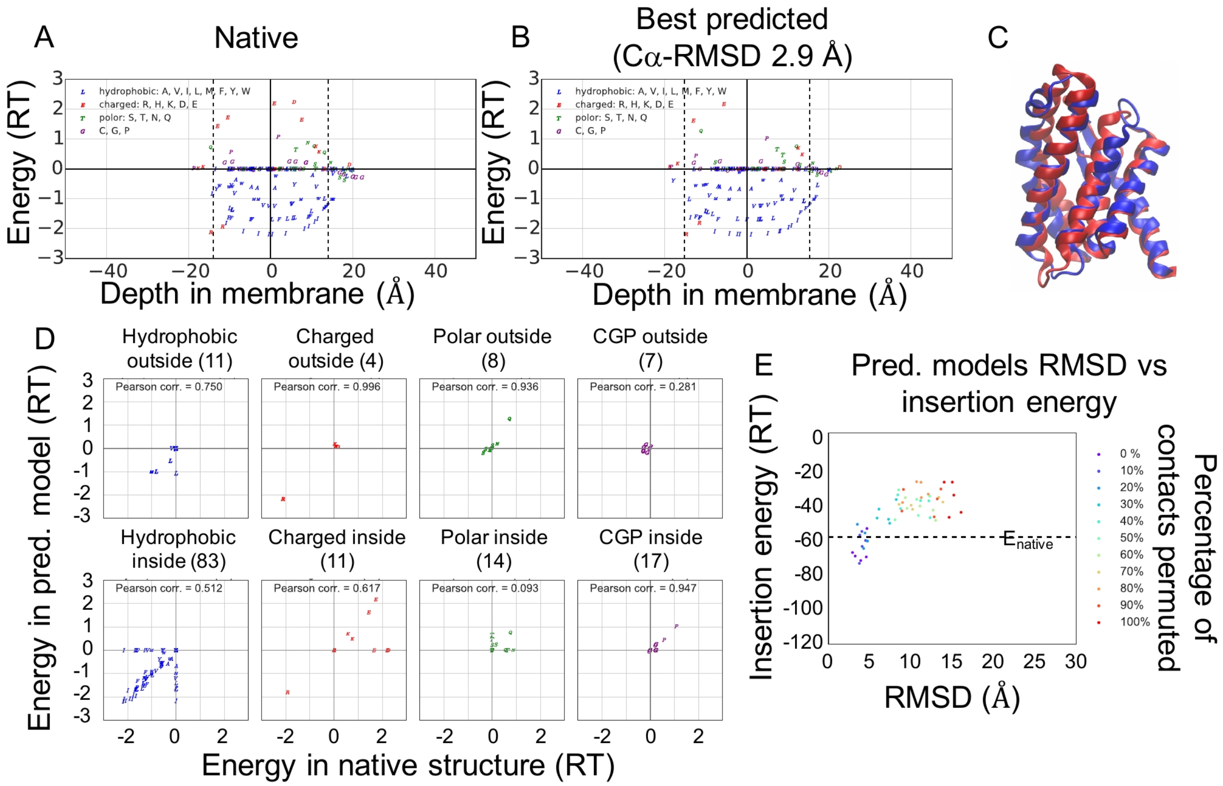

Figure 2.26: **Energetic decomposition for the known native 5d0yA and models predicted by DeepCNF**. 5d0yA is ECF transporter that has 155 residues and 6 transmembrane helices. The best model predicted by AUC-maximized DeepCNF has TM-score 0.639 and C_α-RMSD 2.93. The Pearson correlation coefficient between the RMSD and the insertion energy is 0.915. Copyright from *Biophysical Journal*.

2.7 Insertion into intrinsically curved membrane

2.7.1 Introduction

The calculation of protein-membrane interactions and the determination of the lowest energy conformation of the combined bilayer-protein system is a very challenging task [101–103]. The shape of the membrane affects the position and conformation of the embedded membrane proteins [101, 102] and conversely, the membrane proteins can play a pivotal role in altering membrane shape and local properties [102, 104]. Furthermore, membrane thickness can vary around the protein as the lipids adjust to the hydrophobic pattern of the proteins surface [27, 101, 102]. A complete solution to this problem requires considering all these factors, which is beyond the scope of the present study.

Statistical potentials can provide an estimate of the differences in insertion energy into bilayers of different thicknesses or geometries for a given protein conformation, or of the insertion of proteins with different conformations into a bilayer of given thickness, or of a combination of the two. Although these calculations do not include the energy required to curve the membrane, change its thickness, or the proteins conformation, they can be used to identify whether bilayer thickness or curvature can have a significant effect on the relative insertion energies.

Applications to curved membranes introduce the added complexity of calculating the effective depth of a residue. The depth is the shortest distance between the residue and a middle hypersurface, halfway between the inner and outer surfaces **(Fig. 2.2D)**. Because we reference to the bilayer interface, determining the burial depth is trivial.

As an illustration of our ability to study curved membranes of variable thicknesses, we examine the opening of the mechanosensitive channel MscL. This channel adopts open and closed conformations depending on multiple factors, including the membrane composition which affects thickness, curvature, line tension and lateral pressure [77, 105–107]. Experiments find that a hydrophobic mismatch and a thinner membrane lowers the activation free energy of MscL and promotes the open state. By itself, however, the mismatch is insufficient to open the channel because the open state has a larger cross-sectional area, which is normally disfavored due to lateral membrane pressure by 19 RT [105]. However, an increase in membrane curvature in a manner that reduces the pressure helps trigger the gating and open the channel.

2.7.2 Mathematical models

For a curved membrane, the bilayer thickness and the embedded depth of residues can be different from those in a planar bilayer. The bilayer thickness can be considered as the distance between the interfaces of two monolayers. Assuming a middle hypersurface in between the two monolayers, the embedded depth is the normal distance of the residue to that middle hypersurface, and the bilayer thickness is the distance between the two monolayers. Using the basic knowledge from basic differential geometry, we can calculate the normal distance to an arbitrary surface analytically for any point in the space [108].

Consider an isosurface in $\mathbb{E}^3$:

$$S_{const} = \{(x,y,z)|F(x,y,z) = const\} \neq \emptyset \tag{2.14}$$

$$\nabla F(P) = (F_x, F_y, F_z)(P) \neq \vec{0}(\forall P \in S_{const}) \tag{2.15}$$

$\rightarrow S_{const}$ is a surface.

Let $(x(t), y(t), z(t))$ be a curve on S_{const}, we have:

$$F(x(t), y(t), z(t)) = const \tag{2.16}$$

$$F_x \cdot x'(t) + F_y \cdot y'(t) + F_z \cdot z'(t) = \big\langle (F_x, F_y, F_z), (x'(t), y'(t), z'(t)) \big\rangle \tag{2.17}$$

Therefore, ∇F is S_{const}'s normal vector and $T_{P_0(x_0,y_0,z_0)}S$ is the tangent plane of S that passes P_0 on S (**Eq. 2.18**).

$$T_{P_0(x_0,y_0,z_0)}S : F_x(P_0)\cdot(x-x_0) + F_y(P_0)\cdot(y-y_0) + F_z(P_0)\cdot(z-z_0) = 0 \tag{2.18}$$

Consider the middle hypersurface is an isosurface in $\mathbb{E}^3$ (**Eq. 2.19**), and there are N residues, $P_0 \backsim P_{N-1}$, in the system, the problem can be phrased as $\forall i \in \{0, 1, \dots, N-1\}$, find $Q_i \in S_0$ such that $\overrightarrow{Q_iP_i} \perp T_{Q_i}S_0$, where $T_{Q_i}S_0$ is the tangent plane of S_0 that passes Q_i. From **Eq. 2.14** to **Eq. 2.18**, Q_i can be solved from **Eq. 2.20**.

$$S_0 = \{(x,y,z)|F(x,y,z) = 0\} \tag{2.19}$$

$$\begin{cases} F(Q_i) = 0 \\ \nabla F(Q_i) \parallel \overrightarrow{Q_iP_i} \end{cases}, \forall i \in \{0, 1, \dots, N-1\} \tag{2.20}$$

We calculated the insertion energy of MscL into the following shapes of bilayers. Here, $r = \sqrt{x^2+y^2}$; t is the bilayer thickness at $(x,y) = (0,0)$; R is the radius of the sphere.

Concentric circular sphere, **fig. 2.27**

$$\text{Outer sphere, } r^2 + \left[z + \left(R + \frac{t}{2}\right)\right]^2 = (R+t)^2 \tag{2.21}$$

$$\text{Inner sphere, } r^2 + \left[z + \left(R + \frac{t}{2}\right)\right]^2 = R^2 \tag{2.22}$$

$$\text{Middle hypersurface, } r^2 + \left[z + \left(R + \frac{t}{2}\right)\right]^2 = \left(R + \frac{t}{2}\right)^2 \tag{2.23}$$

$$\begin{aligned} \text{Embedment depth} &= \left\|(r_0, z_0) - \left(0, -\left(R + \frac{t}{2}\right)\right)\right\| - \left(R + \frac{t}{2}\right) \\ &= \left\|(r_0, z_0 + R + \frac{t}{2})\right\| - \left(R + \frac{t}{2}\right) \end{aligned} \tag{2.24}$$

$$\text{Bilayer thickness} = t \tag{2.25}$$

Symmetric convex sphere, **fig. 2.28**

$$\text{Outer sphere, } r^2 + \left[z + \left(R - \frac{t}{2}\right)\right]^2 = R^2, z + \left(R - \frac{t}{2}\right) \geq 0 \tag{2.26}$$

$$\text{Inner sphere, } r^2 + \left[z - \left(R - \frac{t}{2}\right)\right]^2 = R^2, z - \left(R - \frac{t}{2}\right) \leq 0 \tag{2.27}$$

$$\text{Middle hypersurface, } z = 0 \tag{2.28}$$

$$\text{Embedment depth} = z \tag{2.29}$$

$$\text{Bilayer thickness} = 2 \cdot \left|\sqrt{R^2 - r^2} - \left(R - \frac{t}{2}\right)\right| \tag{2.30}$$

Symmetric concave sphere, **fig. 2.29**

$$\text{Outer sphere, } r^2 + \left[z - \left(R + \frac{t}{2}\right)\right]^2 = R^2, z - \left(R + \frac{t}{2}\right) \leq 0 \tag{2.31}$$

$$\text{Inner sphere, } r^2 + \left[z + \left(R + \frac{t}{2}\right)\right]^2 = R^2, z + \left(R + \frac{t}{2}\right) \geq 0 \tag{2.32}$$

$$\text{Middle hypersurface, } z = 0 \tag{2.33}$$

$$\text{Embedment depth} = z \tag{2.34}$$

$$\text{Bilayer thickness} = 2 \cdot \left|-\sqrt{R^2 - r^2} + \left(R + \frac{t}{2}\right)\right| \tag{2.35}$$

Concentric cylindrical sphere, **fig. 2.30**

$$\text{Outer sphere, } y^2 + \left[z + \left(R + \frac{t}{2}\right)\right]^2 = (R + t)^2 \tag{2.36}$$

$$\text{Inner sphere, } y^2 + \left[z + \left(R + \frac{t}{2}\right)\right]^2 = R^2 \tag{2.37}$$

$$\text{Middle hypersurface, } y^2 + \left[z + \left(R + \frac{t}{2}\right)\right]^2 = \left(R + \frac{t}{2}\right)^2 \tag{2.38}$$

$$\begin{aligned}\text{Embedment depth} &= \left\|(y_0, z_0) - \left(0, -\left(R + \frac{t}{2}\right)\right)\right\| - \left(R + \frac{t}{2}\right) \\ &= \left\|(y_0, z_0 + R + \frac{t}{2})\right\| - \left(R + \frac{t}{2}\right)\end{aligned} \tag{2.39}$$

$$\text{Bilayer thickness} = t \tag{2.40}$$

External convex sphere + internal plane, **fig. 2.31**

$$\text{Outer sphere, } r^2 + \left[z + \left(R - \frac{t}{2}\right)\right]^2 = R^2, z + \left(R - \frac{t}{2}\right) \geq 0 \tag{2.41}$$

$$\text{Inner sphere, } z = -\frac{t}{2} \tag{2.42}$$

$$\text{Middle hypersurface, } r^2 + (2z + R)^2 = R^2 \tag{2.43}$$

$$\text{Embedment depth} = sign\left(\overrightarrow{QP}\right) \cdot dist, dist = argmin\left(\left\|\overrightarrow{QP}\right\|\right), \exists Q \in S \tag{2.44}$$

$$\text{Bilayer thickness} = \left(\sqrt{R^2 - r^2} - R + \frac{t}{2}\right) + \frac{t}{2} = \sqrt{R^2 - r^2} - R + t \tag{2.45}$$

External concave sphere + internal plane

$$\text{Outer sphere, } r^2 + \left[z - \left(R + \frac{t}{2}\right)\right]^2 = R^2, z - \left(R + \frac{t}{2}\right) \leq 0 \tag{2.46}$$

$$\text{Inner sphere, } z = -\frac{t}{2} \tag{2.47}$$

$$\text{Middle hypersurface, } r^2 + (2z - R)^2 = R^2 \tag{2.48}$$

$$\text{Embedment depth} = sign\left(\overrightarrow{QP}\right) \cdot dist, dist = argmin\left(\left\|\overrightarrow{QP}\right\|\right), \exists Q \in S \tag{2.49}$$

$$\text{Bilayer thickness} = \left(-\sqrt{R^2 - r^2} + R + \frac{t}{2}\right) + \frac{t}{2} = -\sqrt{R^2 - r^2} + R + t \tag{2.50}$$

2.7.3 Results

We have calculated the difference in energies for insertion of the closed and open states of MscL for several different classes of curved membranes and varying radii **(Fig. 2.27, 2.28, 2.29, 2.30, 2.31)**. The thickness of a planar bilayer with the lowest insertion energy is ∼28 Å for the closed state but is below 22 Å for the open state (the lower limit of our thickness calculation). However, the insertion energy for the closed state always is lower (preferred) than the insertion energy for the open state at all thicknesses (e.g., $E_{insert}^{open} > E_{insert}^{closed}$ at 22 Å thickness). A membrane thickness of 22 Å provides the best option for stabilizing the open state relative to the closed state and minimizes the penalty for hydrophobic mismatch in adopting the open state **(Fig. 2.27G)**. This finding supports the experimental proposal that hydrophobic mismatch (e.g., by altering the lipid composition to change the bilayer thickness) and an increase in curvature combine to promote conformational transitions to the open state of the channel (the curvature also helps reduce line tension, which also promotes the open state [77]).

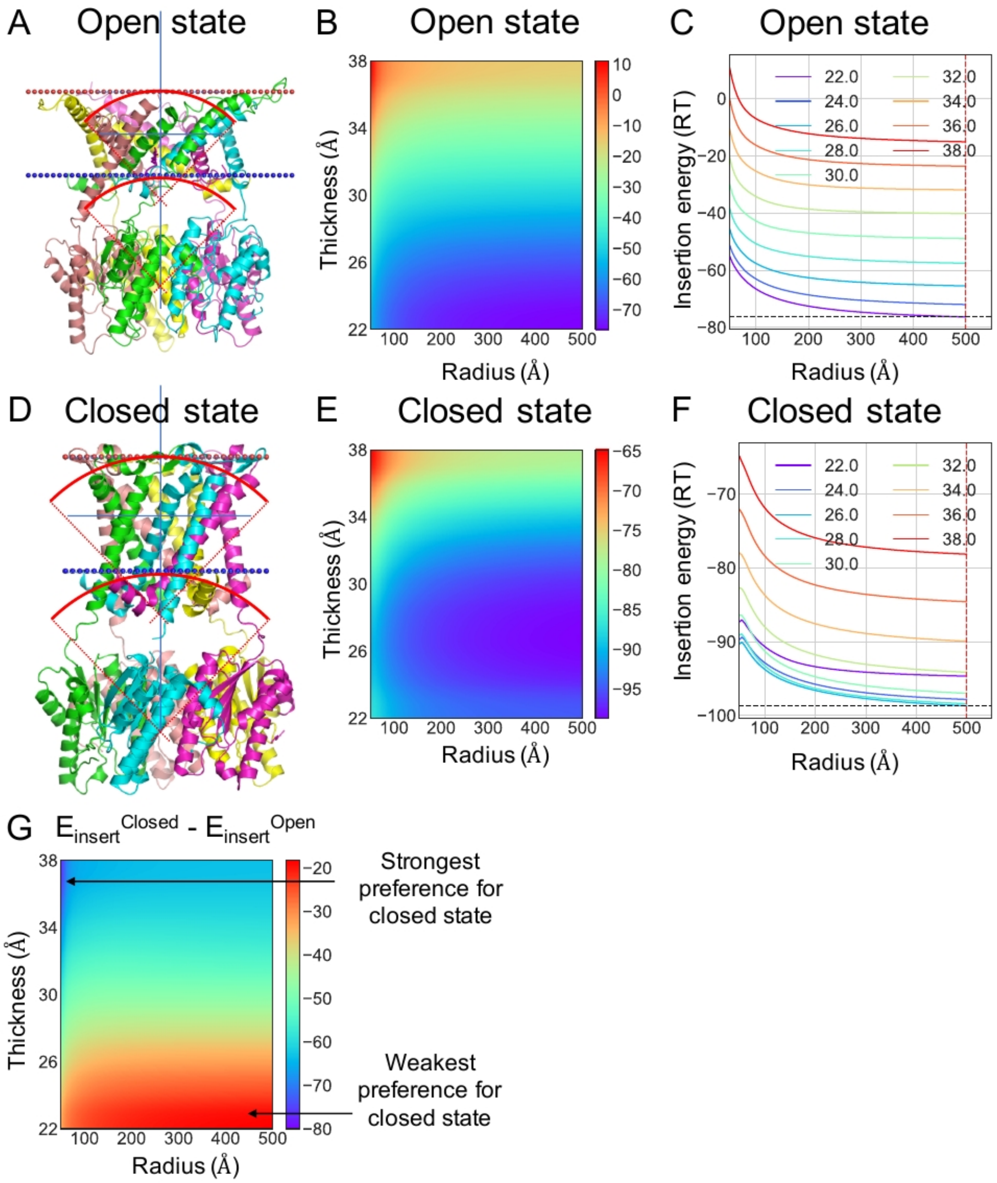

Figure 2.27: **Insertion of MscL into flat or spherical membranes**. The insertion energy is calculated at different bilayer thicknesses and curvatures. The thickness refers to the distance between the cytoplasmic and periplasmic interfaces at (x, y) = (0, 0). **(A, D)** Illustration of the insertion of the open state and closed state. The curvature of the bilayer is exaggerated for illustration purposes. **(B, E)** Insertion energy profile as a function of bilayer thickness and radius. **(C, F)** Insertion energy profile as a function of sphere radius at different bilayer thicknesses. **(G)** Difference of insertion energies between the closed and the open states indicates that thicker bilayers are more likely to trap the protein in the closed state. Note the cost of altering the membrane curvature is not included in the calculation but must be accounted for when determining the total energy of the protein/bilayer system. Copyright from *Biophysical Journal*.

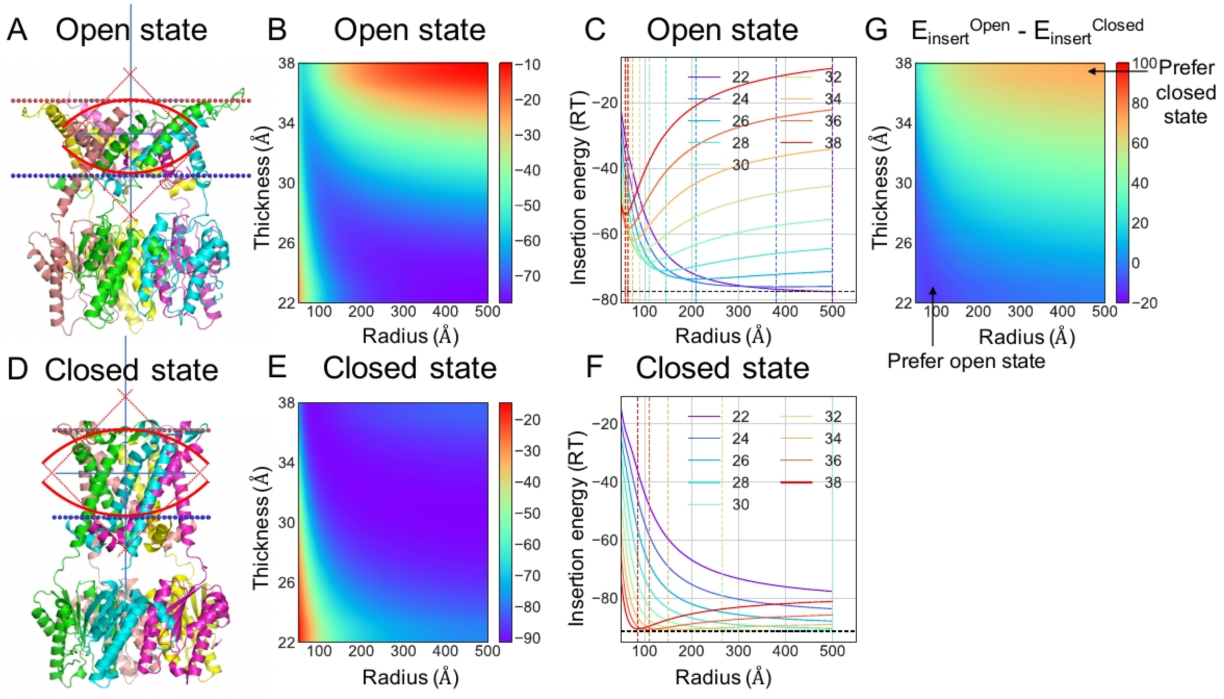

Figure 2.28: **Insertion into curved membranes in the shape of convex sphere**. When embedded in convex bilayers, the insertion energy is lower for open state in thinner bilayers with minimal curvature, which resemble a planar bilayer, as well as by thick bilayers with high curvature (**B-C**). From another perspective, thick bilayers must be bent in order to reduce the bilayer thickness at the edge of the protein to better accommodate the hydrophobic mismatch. On the other hand, thin bilayers already hydrophobically match the hydrophobic surface of the protein, and bending is counterproductive. Similar to the open state in bilayers with a convex shape, the closed state can be stabilized by a thin bilayer with small curvature and thick bilayers with high curvature, though the stabilizing effect of the thick curved bilayers is less pronounced with the closed state (**E-F**). For the closed state, the lowest insertion energy is similar with a medium thick bilayer (∼32 Å) having a small curvature and a thick bilayer (∼38 Å) having high curvature (**F**). Thus, the closed state is stabilized relative to the open state for a medium thick flat bilayer or a thick bilayer with a convex shape and high curvature (**G**).

When considering the equilibrium between the open and closed states, we find that thin convex bilayers reduce the insertion energy of the open state relative to the closed state with an increasing curvature enhances this stabilization of the open state (**G**).
Copyright from *Biophysical Journal*.

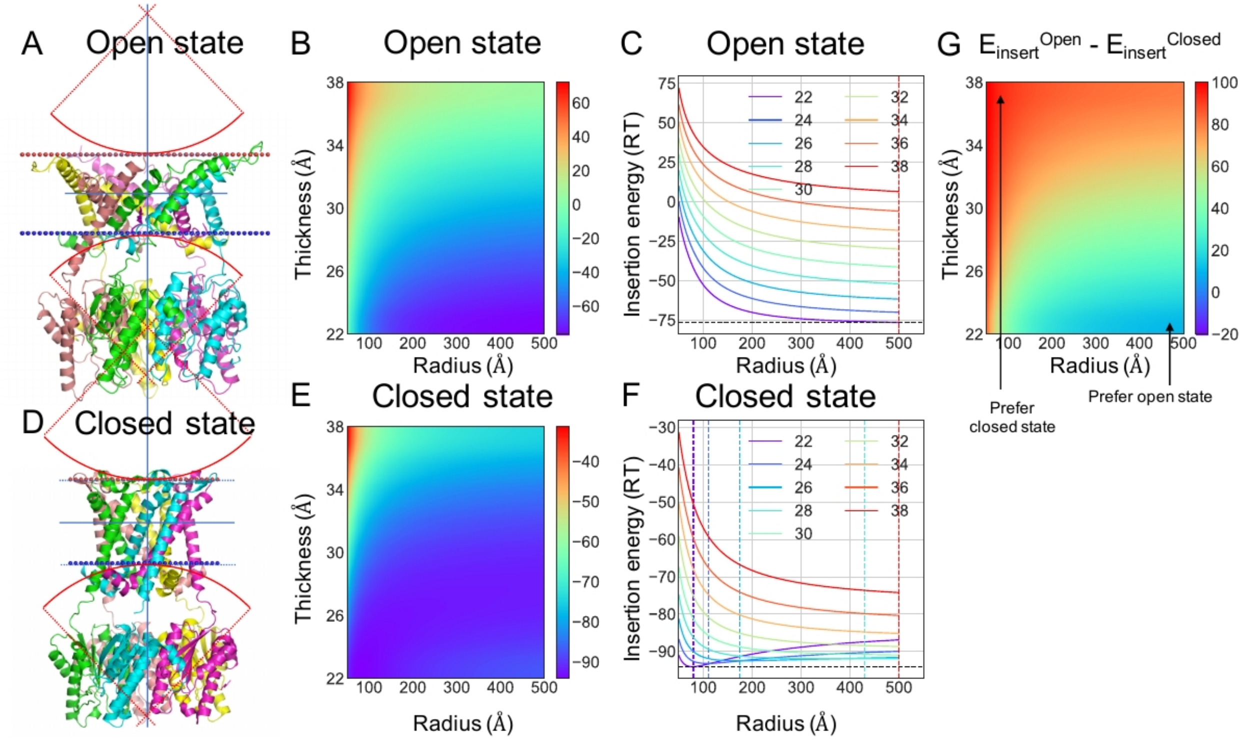

Figure 2.29: **Insertion into curved membranes in the shape of concave sphere**. The difference of the insertion energy profiles implies that it is energetically unfavorable to open the channel in the bilayer of concave shape compared to the convex shape. Because higher curvature results in thickened bilayer at the edge of the protein for the membrane bilayer in a concave shape, the open state tends to be stabilized at decreasing curvature. When embedded in concave bilayers, optimal curvatures exist in thin bilayers for the closed state but the minimum is a flat bilayer for the open state. The minimum insertion energy obtained for thin bilayer (∼22 Å) with high curvature is lower than the energy obtained at a flat bilayer. Copyright from *Biophysical Journal*.

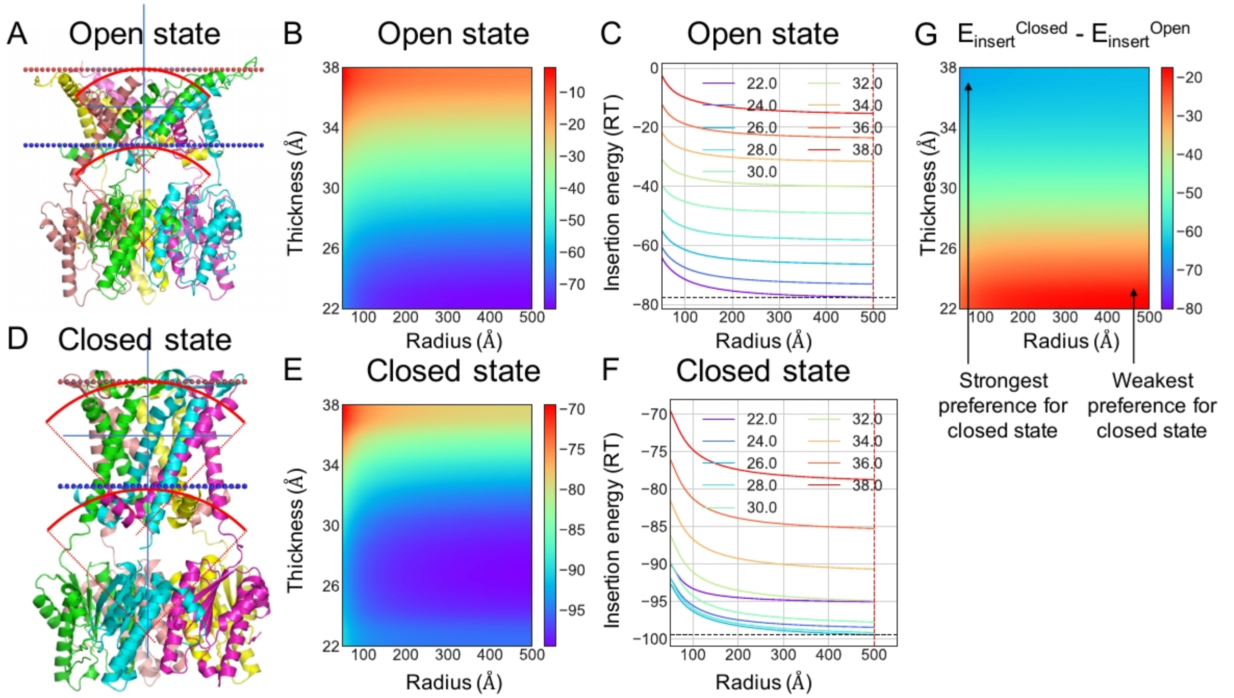

Figure 2.30: **Insertion into curved membranes in the concentric cylindrical shape**. Similar to the concentric circular shape, insertions into curved membranes in a concentric cylindrical shape are universally unfavorable than flat membranes for the open and closed states. Copyright from *Biophysical Journal*.

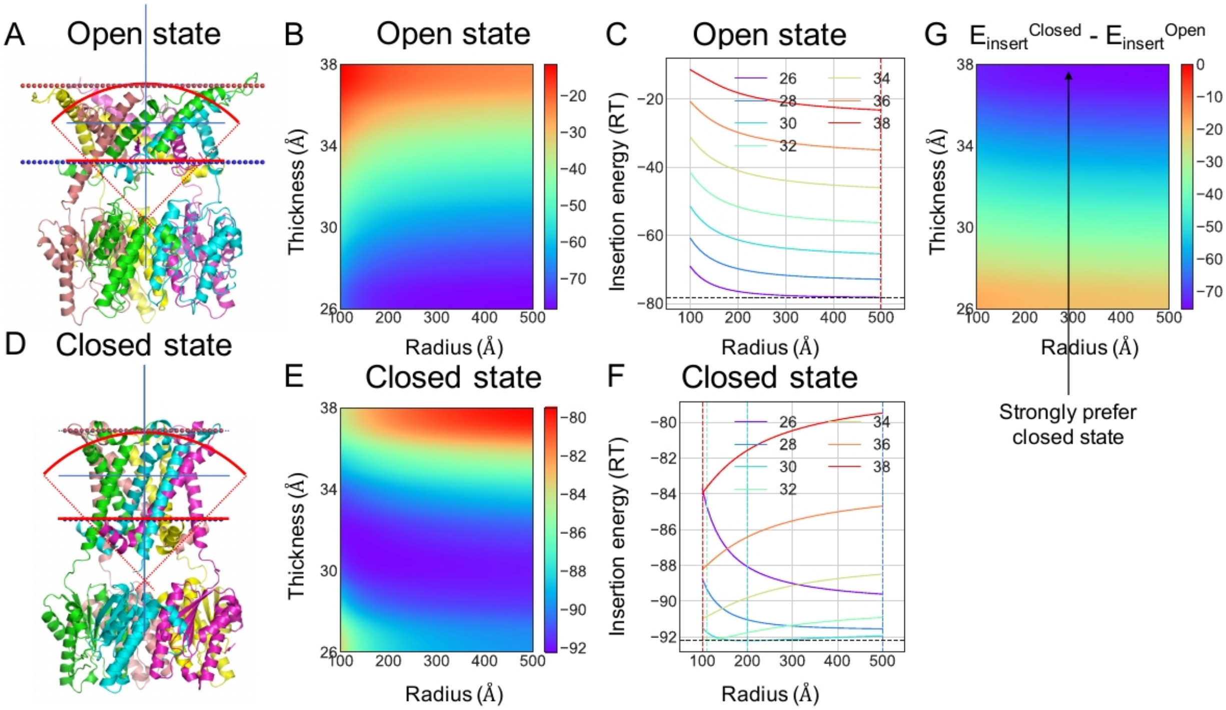

Figure 2.31: **Insertion into curved membranes in the shape of external convex sphere and internal plane**. Calculations of asymmetric bilayers were also carried out. The stabilizing or destabilizing effect is not as pronounced as that of a symmetric bilayer due to the curvature of membrane on only one side. Copyright from *Biophysical Journal*.

2.8 Integration with *Upside*

2.8.1 Implementation

The environment coverage and H-bond are evalutated for each residue.

C_β positions are used to compute the side chain energy for 20 normal type of amino acids. A sigmoid function of the environment coverage, defined by a midpoint and sharpness, is used to modulate the C_β energy. Fully-exposed residues have full side chain energies. Positions of backbone H-bond donors and acceptors are used to compute the UHB terms. The H-bond is given by a probabilty inferred by *Upside*. Therefore, the UHB energy is computed by multiplying with one minus that probability.

The membrane burial potential is implemented in the MembranePotential class in *Upside*.

2.8.2 Dynamic orientaions

To illustrate the integration of *UChiMemPot* with the *Upside* algorithm, we have run MD simulations for a variety of multimeric helical transmembrane proteins as well as for their monomeric units using the *UChiMemPot* and the asymmetric Ez-3D method applied only to the lipid-exposed residues. Energies for the contributions from unsatisfied hydrogen bonds (UHB) are assigned to all residues with an unpaired backbone H-bond partner, regardless of the exposure status for residues located within the bilayer. Proteins are restrained as quasi-rigid bodies such that only minimal internal movements are allowed (short strong springs placed between backbone atoms allow atomic vibrations).

The simulations begin from the OPM position and bilayer thickness, and the distribution of conformations in the trajectory is analyzed (**Fig. 2.32**). Regardless of the applied potential, the proteins, on average, execute small deviations from the orientation of the native state (the standard deviation of the relative shift and tilt on average are ∼1 Å and ∼2°, respectively, for the multimers). When UHB energies are incorporated, *UChiMemPot* produces a tighter angular distribution than asymmetric Ez-3D.

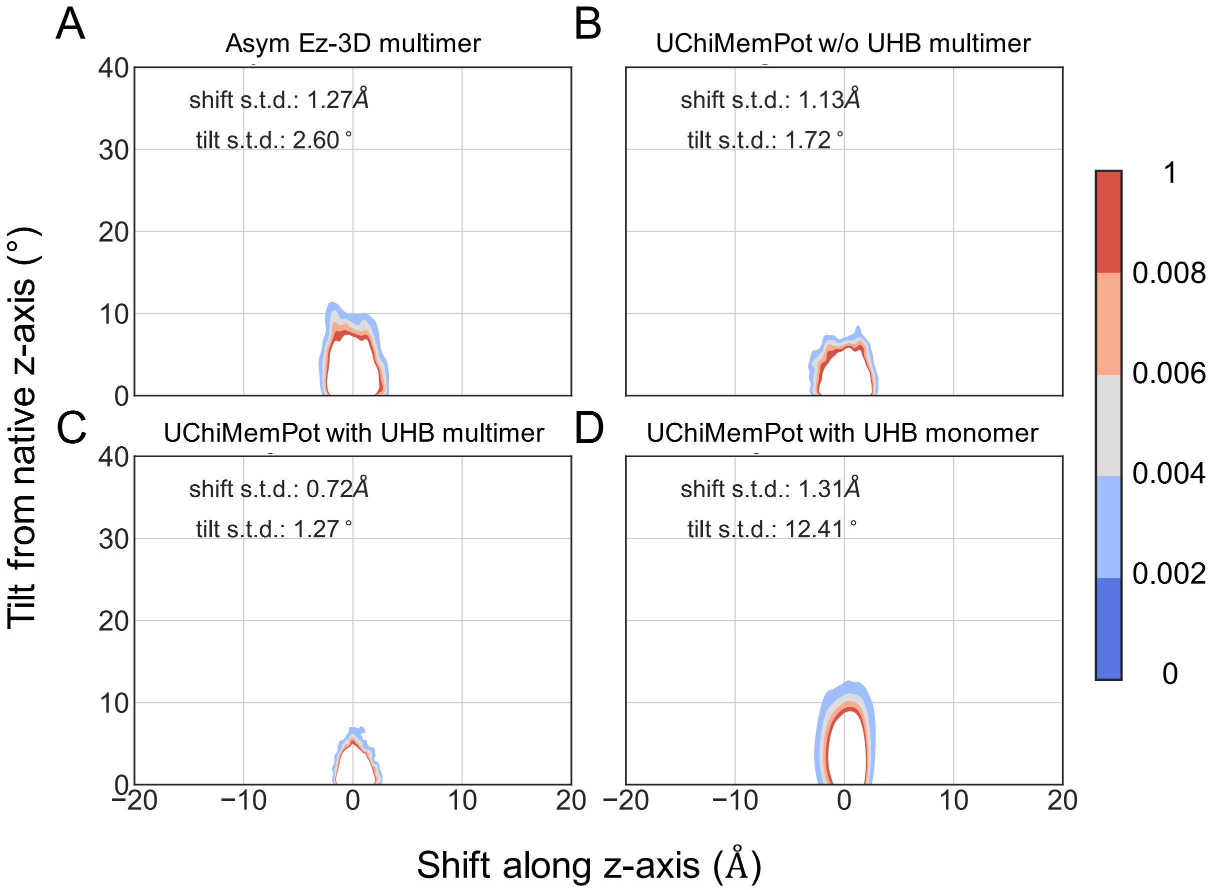

Figure 2.32: **Comparison of the distributions of the shift and tilt from *Upside* simulations for multimers and monomers for 3 potentials**. Averaged distributions of relative shift and tilt from the initial orientation for 23 multimers / monomers with different membrane potential applied. The three potentials applied are: Asymmetric Ez-3D C_β potential (denoted as Asym Ez-3D), our potential without UHB terms (*UChiMemPot* w/o UHB), and our potential with UHB terms (*UChiMemPot* with UHB). The x-axis stands for the deviation of shift and the y-axis the deviation of tilt from the native z-axis. Copyright from *Biophysical Journal*.

We have investigated the effects of oligomerization on the position within the membrane. Specifically, we compared the positions of the constituent monomers in the complex to the position they would have if they were isolated (but with the same protein conformation). For most complexes, we find the isolated monomers retain a very native-like orientation in the membrane, both for depth and tilt. This observation suggests that docking of properly positioned monomers is a viable assembly strategy (**Fig. 2.33**).

However, monomers from some ion channels experience larger movements due to the exposure of charged and polar residues normally solvated and lining the channel cavity. These channels include the potassium channel KcsA (tetramer, 1r3j), formate transporter FocA (pentamer, 3kcu), calcium-gated potassium channel MthK (tetramer, 3ldc), NaK potassium channel (tetramer, 3ouf), sodium pumping rhodopsin NaR (dimer, 3x3b), and nitrite transporter NirC (tetramer, 4fc4). This implies that the assembly of these helical transmembrane proteins may involve a partial induced-fit mechanism. We are unaware of experimental studies of isolated subunits, so these results should be considered testable predictions.

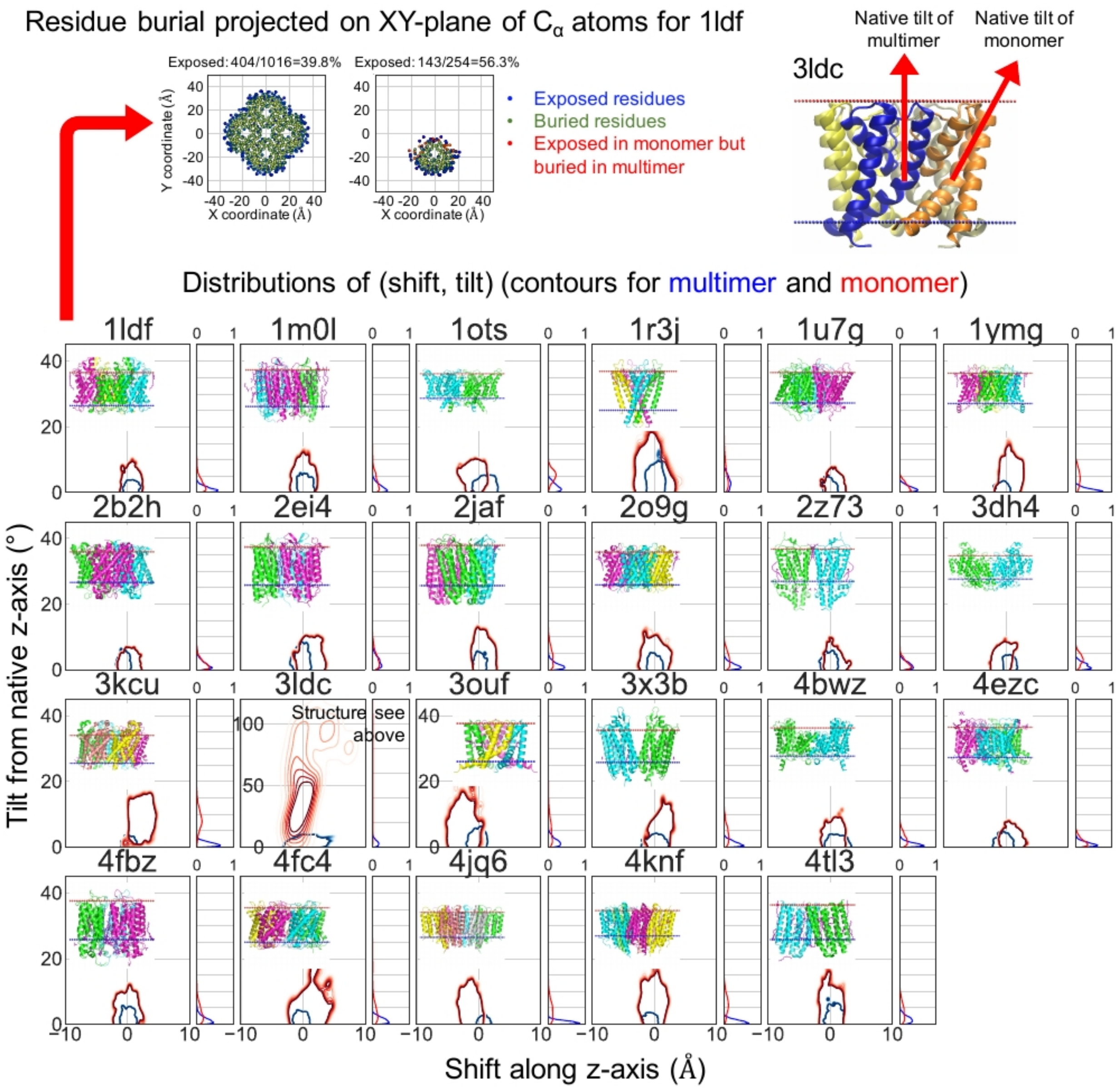

Figure 2.33: **Distributions of relative shift and tilt from *Upside* simulations for multimers and monomers**. For each of the oligomers, the contour maps of the distribution of the relative shift and tilt of the multimer (blue) and monomer (red) are plotted. The 1D projections of the tilt angles are plotted to the right of each plot on a scale normalized from 0 to 1. Proteins shown in VMD's New Cartoon representation are obtained from the website of OPM. Monomeric units have different colors. Multimeric complexes in general have small movements in the simulation, retaining their native orientations, as opposed to some of the monomeric units. In the 2D projection of the residue burial of C_α atoms on the XY-plane (top), some residues in or near the cavity in the tetramer become exposed to lipid in the monomer (green circle becomes a red circle). The 3D structure illustrates that the tilt angle of the monomer and multimer can be different. Copyright from *Biophysical Journal*.

2.9 Discussions

We advance a new statistical potential for the burial into membranes of amino acids and hydrogen bond donors and acceptors that incorporates several new features. First, positions within bilayer are referenced to the lipid head group interface, which more accurately integrates data from systems with variable bilayer thicknesses (**Figs. 2.2A, 2.11**). This referencing is important since helical transmembrane proteins come from different species and organelle locations, and the thicknesses vary widely (e.g., from 26 to 38 Å for the 9 organelle locations in 74 species for the 140 helical transmembrane proteins in the training set). Nevertheless, the use of a thickness of 30 Å suffices as a reasonable compromise for most purposes (**Fig. 2.3A**). Our ability to employ bilayers with a range of different thicknesses also permits calculating the membrane thickness and curvature that provides the lowest insertion energy.

Second, our statistical potential excludes contributions from the residues buried within the protein, in contrast to most previous methods that retain all residues located within the bilayer region [65, 66, 68–71]. A very recently published method [72] takes advantage of a recent automated method implemented in Rosetta, *mp_lipid_acc* [109], to distinguish lipid-accessible and lipid-inaccessible residues in the derivation of the potential for helical transmembrane proteins and β-barrels. Unlike our procedure, a potential profile for lipid-inaccessible residues is considered as part of the total insertion energy.

Third, we take into consideration the presence of unpaired backbone H-bonding groups. The free energy of transferring a peptide group containing an amide proton and carbonyl oxygen from water to liquid alkane is estimated to be 13 RT [31, 38, 110]. This magnitude is considerably larger than our statistical potentials value of $\sim$4 RT. Support for the lower cost is the consistent with our values for the transfer of Gln, Arg, and Asn side chains from solvent to inside the bilayer, 2.0, 2.0 and 1.0 RT, respectively, or 3.5, 3.5 and 2.5 RT higher than the value for transferring an alanine. Fleming and coworkers obtained similar values for the change in folding free energy at the center of the bilayer for a β-barrel upon substitution of a Gln, Arg or Asn with an alanine, 2.4, 3.6, or 2.3 RT, respectively [111, 112]. The smaller energies in the statistical potential and experiment for burying an amide side chain, compared to transfer studies, suggests that other factors stabilize the burial of peptide groups in membrane proteins. These factors could include residual solvent in the bilayer and interactions with nearby backbone and other side chain atoms, resulting in only partial exposure to the lipids [113]. As a result of carefully treating the bilayer boundary position, lipid exposure and inclusion of hydrogen bonding, our statistical potential is able to accurately reproduce the optimal thickness and position determined recently by the OPM implicit bilayer method [59, 76].

None of the current statistical potentials have been employed in MD simulations to the best of our knowledge. Likewise, all coarse-grained models employing implicit lipids typically over count protein-lipid and protein-protein interactions. This is partly due to the difficulty of determining the burial status of residues during simulation with implicit solvent methods. For example, the AWSEM-membrane model (a learned potential) can fold some helical transmembrane proteins at modest resolution [25], although a residue embedded in the lipid bilayer always retains the membrane burial potential irrespective of its exposure to lipids.

2.9.1 Use of statistical potentials and other limitations of the approach

Although widely used, the interpretation of a statistical potential obtained using the Boltzmann relationship has been debated [70, 114, 115]. Some believe that the use of statistical potentials implies the presumption that the observed distributions reflect the average of individual free energies for each element. However, context along with structural and functional requirements also can contribute to the preference for a location, thereby reducing the validity of the connection between energy and frequency of each component. Generally, the relationship between statistical potentials and basic statistical mechanics can be uncertain [115–117]. Some statistical potentials have been derived using only probability theory, but the defining relations are generally the same [118–121].

Statistical potentials typically neglect explicitly non-additive many-body interactions both to greatly facilitate practical computations and because of a gross paucity of training data [115, 118]. Studies using continuum/implicit membrane models [103] as well as atomistic work [122, 123] have found that non-pairwise terms are important for multiple charged or polar residues in the membrane. For example, the translocation of Arg through the hydrocarbon core of a lipid bilayer proceeds by the formation of a water-filled defect that keeps the Arg hydrated even at the center of the bilayer. In this case, adding additional Arg residues to the water defect causes only a small change in free energy [122, 123].

Errors associated with neglecting many-body interactions in statistical potentials can be minimized by using conditional probabilities. In membrane burial potentials, lipid exposure and burial depth implicitly include the impact of some many-body effects while retaining the benefits of additivity. In the end, the utility of a statistical burial potential for membrane proteins should be judged by its the correlation with experimental hydrophobicity scales [68] and success in a number of applications [65, 68–71].

Our method contains some issues. First, we deliberately include the exposed polar or charged residues in the pores or cavities, so as to be compatible with the implementation of the statistical potential in our MD simulations. This deficiency arises because the identification of whether a residue is lipid exposed is not a readily solvable problem (**Fig. 2.34**). A proper evaluation requires the complete knowledge of the entire structure to determine whether a residue is in contact with lipids or lines an internal cavity or channel. Consequently, results for systems with charged or polar residues in solvated pores are compromised in this regard. Generally, however, there are few solvent-exposed polar or charged residues in pores in the training set so their exclusion would have minimal effect on the derivation of our potential. For simulation purposes, this problem can be resolved when the solvated residues are known *a priori*, as the *Upside* code has the option of excluding their burial energies.

Our model is not suitable for studies that require atomic details of the lipid bilayer, such as the composition of lipids or dynamic changes in the membrane curvature. As discussed above, we are limited to considering insertion energies that are evaluated assuming a specific membrane curvature or protein conformation and ignoring the energy required to alter the membranes geometry or the proteins conformation. A full treatment should include both the energy of deforming a membrane and perturbing the protein structure [102]. A good example of this inclusion is in work by Panahi and Feig who have developed an implicit membrane model allowing for local bilayer deformation in response to the insertion of transmembrane proteins and have employed the potential in MD simulations [27].

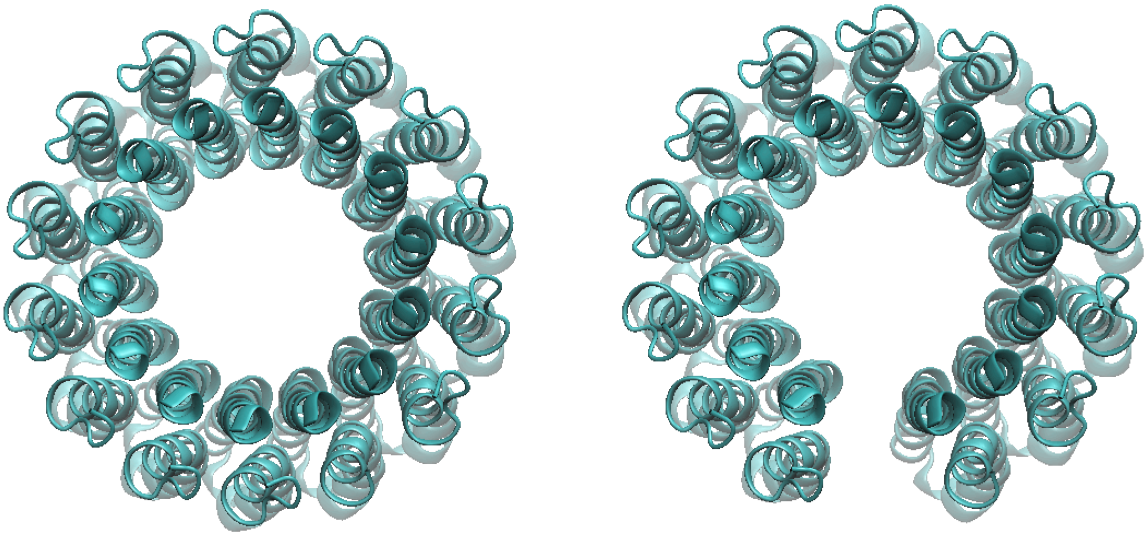

Figure 2.34: **Top view of ATP synthase (PDB id: 3v3c) and a model lacking one of the chains**. The native version has a aqueous pore whereas the defect allows for the entry of lipids. We do not exclude the exposed polar or charged residues in the pores or cavities on purpose so as to be compatible with employing the statistical potential in our MD simulations. Furthermore, determining whether a residue is lipid exposed is not a readily solvable problem with an implicit membrane. As shown in the figure below, one must know the entire structure to know whether a channel contains lipids. Currently, we do not know how to incorporate this complexity into a fast MD algorithm. In the end, since there are few exposed polar or charged residues in pores in the training set, their exclusion doesnt make much difference.

2.10 Conclusions

We provide an updated statistical potential for the burial of proteins in membranes. Our potential is referenced to the interface region, rather than the bilayer center, to better capture the physiochemical properties of this region. In addition, we include a penalty for the presence unsatisfied hydrogen bond donor and accepters, finding that the penalty for each is about 2 RT, which is lower than many prior estimates. We address challenges inherent in devising implicit bilayer models of properly accounting for the exchange of protein-lipid interactions for protein-protein interactions upon protein contact. Our model cannot yet fold membrane proteins from an extended solvated chain, especially for those requiring insertion by the translocation (very hydrophobic proteins tend to form collapsed globules in solution). Nonetheless, when our membrane burial potential is integrated with our *Upside* algorithm, the combination offers a computationally efficient, flexible tool that lays the groundwork for a variety of simulations of membrane protein dynamics.

A web server is available to insert helical transmembrane protein models into membranes and run *Upside* simulations of the inserted model as rigid body to generate an ensemble of orientations: http://sosnick.uchicago.edu/serverlinks.html.

2.11 Aknowledgement

Work in this chapter is supported by The National Institute of General Medical Sciences Grants GM055694 (to T.R.S. and K.F.F.), GM087519 (to E. Perozo), and T32GM008720 (to J. Piccirilli). Computations are produced using the Midway resources of the Research Computing Center (RCC) at the University of Chicago.

CHAPTER 3

FAST SIMULATIONS AND THE INTERPRETATION OF UNFOLDING MEASUREMENTS OF MEMBRANE PROTEINS UNDER FORCE

This material presented in this chapter has appeared in [124].

We simulate 100's of unfolding events using our new *Upside* MD algorithm. For bacteriorhodopsin, the major experimental features are reproduced down to the level of the back-and-forth unfolding of single helical turns. When pulling laterally on GlpG, a variety of pathways are seen with multiple unfolding steps starting from either terminus. The mode of application of force alters the perception of the folding landscape. For GlpG unfolding using a weaker spring to mimic a magnetic tweezers measurement, the force remains nearly constant after the initial unfolding event and few if any intermediates are observed, as found in experiment. In contrast when using a stiff cantilever, the force drops to near-zero after each major unfolding event and numerous intermediates are observed. Hence, the mode of application of force strongly affects the observed unfolding cooperativity and identification of intermediates, important issues that should be considered when designing experiments and interpreting unfolding data.

3.1 Introduction

Single-molecule force spectroscopy (SMFS) is a powerful tool to investigate the dynamics of biomolecules. Among the ever-expanding repertoire of single-molecule manipulation techniques, atomic force microscope (AFM), optical tweezers (OT), and magnetic tweezers (MT) are the most common [125]. They have been proven beneficial in detecting sparsely populated intermediates and elucidating unfolding pathways of soluble [126–128] and membrane proteins [43, 129–131], including bacteriorhodopsin and GlpG.

Bacteriorhodopsin (bR), a light-driven proton pump with seven transmembrane (TM) helices, is a paradigm for membrane proteins, which forced unfolding has been extensively studied by both experiments [43, 129, 132–135] and MD [47, 136]. GlpG, a rhomboid intramembrane protease from *E. coli*, comprises six TM helices and cleaves a specific peptide bond near the membrane. Because of its functional importance and detailed structural information available in the Protein Data Bank (PDB), GlpG has come forth as essential model for studying the folding and stability of helical transmembrane proteins by chemical biology [34, 137, 138], magnetic tweezers [130], and MD simulations [139, 140].

Simulations have aided the experimental SMFS studies by revealing the complexity of the unfolding and refolding processes [47, 136, 141–143]. Despite increased computing capacity, simulating the forced unfolding of macromolecules with all-atom methods remains difficult [136]. Coarse-grained (CG) models enable more extensive sampling and allow for slower, more realistic pulling velocities and lower forces [47], both better matching the experimental studies and increasing the likelihood of observing transient intermediates.

A major challenge in coarse-graining is to find the right balance between accelerating the simulations and retaining the critical features of the system. We have addressed this challenge with our new *Upside* model, which is capable of *de novo* folding of proteins shorter than 100 residues in cpu-hours [78, 79]. To this physics-based model which has six atoms per residue and realistic Ramachandran maps, we have incorporated our new knowledge-based membrane burial potential that dynamically calculates the degree of side chain exposure to lipids during the simulations (i.e., we correct for the loss of lipid-protein interactions as helices come together) [48]. The membrane potential also includes energies for unsatisfied H-bond donors and acceptors in the membrane. By allowing for these unsatisfied H-bonding groups, helices fold and unfold within the bilayer in an energetically plausible manner during the unfolding simulations.

We perform 100's of forced unfolding simulations of bR and GlpG, first to test our ability to reproduce the experimental data and then to investigate the effects of pulling under various protocols, including different cantilever (spring) stiffness, and operating in constant velocity

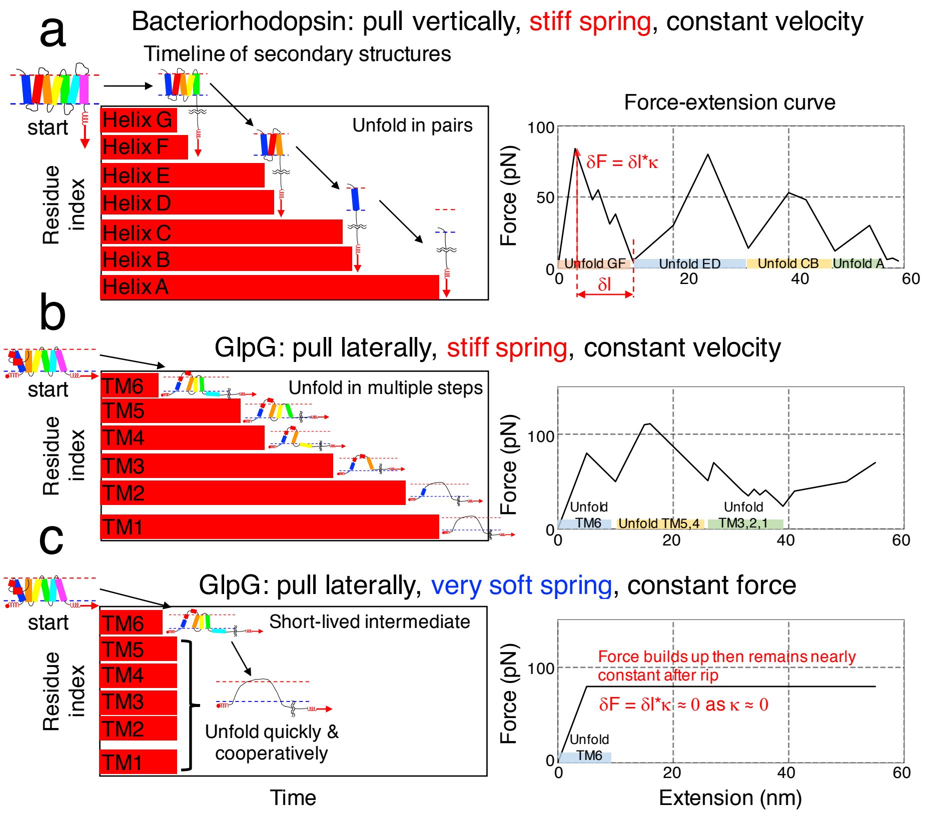

Figure 3.1: **Idealized forced unfolding results for bR and GlpG highlighting how different pulling protocols influence the observation of intermediates**. The red virtual springs apply force by moving perpendicular or parallel to the bilayer surface at a constant velocity. Panel **a, b**. A drop in force, δF, occurs with a stiff spring when the unfolded chain length increases by δl increasing the probability that an intermediate will be observed. Panel **c**. When the transmission of force is applied with a very soft spring, force is maintained as δForce = - δl·k$_{cantilever}$ $\sim$ 0, effectively resulting in a force clamp. The timeline of secondary structures (TSS) depicts the change of secondary structure of all residues versus time. GlpG's two small interfacial helices are not shown.

or force mode **(Fig. 3.1)**. For bR, our simulations largely match experimental AFM data, for example, observing many of the same intermediates [43]. For GlpG, we conduct a principle component analysis [144] that identifies that unfolding can occur from either terminus or the middle of the protein. This observation is in apparent disagreement with the experimental observation where only the C- to N-terminal route was identified [130]. We propose that this experiment did not observe the other routes because of the experimental protocol. We conclude with an analysis of how different SMFS modes explore different parts of the energy surface and how this property can influence the interpretation of the folding behavior.

3.2 Implementation

Upside is a non-Gõ, physics-based model with five atoms per residue (N, Cα, C, H, O), a side chain bead and with residue- and neighbor-dependent Ramachandran maps [145]. The energy function includes H-bonds, side chain-side chain and side chain-backbone interactions (including helix capping), and a solvation term. The energy function is trained using contrastive divergence. The side chains are represented by multi-position, amino acid- and directional-dependent beads. Their positional probabilities are given by the probability distribution having the lowest global free energy for all side chains (minimize G = E - TS). The use of an instantly equilibrated probability distribution calculated at every MD step is novel and greatly smooths the energy surface and enables *Upside* simulations to be extremely fast.

Force is applied to the chosen Cα, based on the virtual cantilevers spring constant (κ) and position, which may be moved with at pulling velocity v. The applied force is computed by $\kappa \cdot$(tip position - Cα position).

An initial time is needed in the input. A time step counter is set in the function in order to record the integration times and thus to compute the position of the tip and the applied force. The tip position is computed by initial tip position + v $\cdot$ time, where the time is given by initial time + *Upside* time step $\cdot$ counter. All heavy atoms in *Upside* have a mass of 1.

The AFM function is implemented in the AFMPotential class in *Upside* (in bond.cpp).

The tension function used in the force clamp simulation is implemented in the TensionPotential class in bond.cpp. Its implementation is simpler because there is no need to estimate the time and store the tip position and residue position.

3.3 Usage

In addition to the standard configuration for running *Upside* folding simulation of small soluble proteins [78, 79], we implement our membrane burial potential, which dynamically accounts for the degree of side chain exposure to lipids [48] and pulling function in the simulations.

3.3.1 Configurations of gradual pulling simulations

(1) Prepare the initial protein structure in a pickle file format as the input.

```
  python PDB_to_initial_structure.py \
  pdbname.pdb pdbname                \
  --allow-unexpected-chain-breaks    \
  --record-chain-breaks              \
  --disable-recentering
```

(2) Prepare the H5 file for the simulation, which includes all the simulation parameters.

```
python upside_config.py                                           \
--output               pdbname.h5                                 \
--fasta                pdbname.fasta                              \
--initial-structure    pdbname.initial.pkl                        \
--hbond-energy         $(cat UPSIDE_param_dir/ff_1/hbond)         \
--dynamic-rotamer-1body                                           \
--rotamer-placement    UPSIDE_param_dir /ff_1/sidechain.h5        \
--rotamer-interaction  UPSIDE_param_dir ff_1/sidechain.h5         \
--environment          UPSIDE_param_dir/ff_1/environment.h5       \
--rama-sheet-mixing-energy $(cat UPSIDE_param_dir/ff_1/sheet)     \
--rama-library         UPSIDE_param_dir/common/rama.dat           \
--reference-state-rama UPSIDE_param_dir/common/rama_reference.pkl \
--membrane-thickness   membrane_thickness                         \
--membrane-potential   membrane_potential_fpath                   \
--ask-before-using-AFM AFM_fpath                                  \
--AFM-time-initial     0
```

AFM_fpath is the path to the file that defines the pulling residue, tip position, spring constant and pulling velocity. One or more residues can be pulled. In *Upside*, the unit of the energy is k_BT: 1 k_BT ≈ 4.114 pN·nm at T = 1.0 (≈ 298 K). The unit of the spring constant is k_BT/Å^2: 1 k_BT/Å^2 ≈ 41.14 pN/Å. 1 *Upside* time step ≈ 0.1 ns, so the pulling velocity 0.001 Å/*Upside* time step ≈ 10^6 nm/s, the same as the extraction velocity in the CG-MD simulations [47].

(3) Run *Upside*.

```
upside pdbname.h5            \
--seed           random_seed \
--temperature    temperature \
--frame-interval frame_intvl \
--duration       duration    \
--disable-recentering
```

3.3.2 Configurations of force clamp simulations

The only difference with the configuration above is in the preparation of the H5 file. A tension file is supplied to *Upside* instead of an AFM file, which defines the pulling residue and pulling force. One or more residues can be pulled.

```
python upside_config.py                                           \
--output               pdbname.h5                                 \
--fasta                pdbname.fasta                              \
--initial-structure    pdbname.initial.pkl                        \
--hbond-energy         $(cat UPSIDE_param_dir/ff_1/hbond)         \
--dynamic-rotamer-1body                                           \
--rotamer-placement    UPSIDE_param_dir /ff_1/sidechain.h5        \
--rotamer-interaction  UPSIDE_param_dir ff_1/sidechain.h5         \
--environment          UPSIDE_param_dir/ff_1/environment.h5       \
--rama-sheet-mixing-energy $(cat UPSIDE_param_dir/ff_1/sheet)     \
--rama-library         UPSIDE_param_dir/common/rama.dat           \
--reference-state-rama UPSIDE_param_dir/common/rama_reference.pkl \
--membrane-thickness   membrane_thickness                         \
--membrane-potential   membrane_potential_fpath                   \
--tension              tension_fpath
```

3.4 Validating correct physics

3.4.1 Worm-like chain model and the analytical solution of contour length

Protein and DNA behavior under force is usually described by the worm-like chain (WLC) model and its variants for polymer elasticity [130, 146, 147]. According to the WLC model [146], the force (F) and the extension (x) of the unfolded protein has the following relation, where k_B is the Boltzmann constant, T is the temperature, $L_p = 0.4$ nm is the persistent length of unfolded polypeptide (which corresponds to chain spatial memory), and Lc is the contour length (total length) of the unfolded polypeptide.

$$F = \frac{k_B T}{L_p}\left[\frac{1}{4}\left(1 - \frac{x}{L_c}\right)^{-2} + \frac{x}{L_c} - \frac{1}{4}\right] \tag{3.1}$$

Let $\alpha = \frac{k_B T}{L_p}$, $\lambda = 1 - \frac{x}{L_c}$, $\omega = \frac{4F}{\alpha} - 3$ substitute them into **Eq. 1**, we have

$$4\lambda^3 + \omega\lambda^2 - 1 = 0 \tag{3.2}$$

.

According to Cardanos method [148], any cubic equation can be solved analytically. $ax^3 + bx^2 + cx + d = 0 (a \neq 0, a, b, c, d \in \mathbb{R})$

Let $x = y - \frac{b}{3a} \Rightarrow y^3 + \left(-\frac{b^2}{3a^2} + \frac{c}{a}\right) y + \left(\frac{2b^3}{27a^3} - \frac{bc}{3a^2} + \frac{d}{a}\right) = 0$

Let $\begin{cases} P = -\frac{b^2}{3a^2} + \frac{c}{a} \\ Q = \frac{2b^3}{27a^3} - \frac{bc}{3a^2} + \frac{d}{a} \end{cases} \Rightarrow y^3 + Py + Q = 0.$

Let $\Delta = \left(\frac{P}{3}\right)^3 + \left(\frac{Q}{2}\right)^2$ and $\begin{cases} S = \left(-\frac{Q}{2} + \sqrt{\Delta}\right)^{\frac{1}{3}} \\ T = \left(-\frac{Q}{2} - \sqrt{\Delta}\right)^{\frac{1}{3}} \end{cases}$,

we have three roots: $\begin{cases} y_1 = S + T \\ y_2 = \beta S + \beta^2 T \\ y_3 = \beta^2 S + \beta T \end{cases}$,

where $\beta = \frac{-1+i\sqrt{3}}{2}$ and $\beta^2 = \frac{-1-i\sqrt{3}}{2}$ are the two complex cubic roots of -1.

Here, Δ is the discriminant of the cubic equation.

If $\Delta > 0$, there is only one real root y_1 and two complex roots y_2 and y_3.

If $\Delta = 0$, if $P = Q = 0$ all three roots are equal to 0, otherwise there are three real roots and two of them are equal.

If $\Delta > 0$, there are three unequal real roots with the following relation:

$$\begin{cases} x_1 + x_2 + x_3 = -\frac{b}{a} \\ 1/x_1 + 1/x_2 + 1/x_3 = -\frac{c}{d} \\ x_1 \cdot x_2 \cdot x_3 = -\frac{d}{a} \end{cases} \text{, where } x_i = y_i - \tfrac{b}{3a}, \text{ i} = 1, 2, 3.$$

Now, back to **Eq. 3.2**, let

$$\lambda = y - \frac{\omega}{12} \tag{3.3}$$

, we have

$$y^3 + \frac{-\omega^2}{48} y + \left(\frac{\omega^3}{27 \cdot 32} - \frac{1}{4} \right) = 0 \tag{3.4}$$

and

$$\Delta = \left(\frac{-\omega^2}{48 \cdot 3} \right)^3 + \left(\frac{\omega^3}{27 \cdot 64} - \frac{1}{8} \right)^2 = \frac{1}{64} \left(1 - \frac{\omega^3}{27 \cdot 4} \right) \tag{3.5}$$

In a normal SMFS experiment of unfolding bR, the force F is between 0 and 500 pN. At T = 298 K, $k_B T = 4.114\ pN \cdot nm$. Only when $F < 19.96$ pN is $\Delta > 0$; otherwise $\Delta \leq 0$. Therefore, **Eq. 3.2** has only one real root mostly ($F \geq 20$ pN, $\Delta > 0$), which is the solution to our problem. When $\Delta < 0$, we have

$$\begin{cases} \lambda_1 + \lambda_2 + \lambda_3 = -\frac{\omega}{4} < 0 \\ 1/\lambda_1 + 1/\lambda_2 + 1/\lambda_3 = 0 \\ \lambda_1 \cdot \lambda_2 \cdot \lambda_3 = \frac{1}{4} > 0 \end{cases} \tag{3.6}$$

Assuming $\lambda_1 \leq \lambda_2 \leq \lambda_3$, we have $\lambda_1 < \lambda_2 < 0 < \lambda_3$ and λ_3 is the root we want.

In summary, L_c can be solved analytically given force and extension.

3.4.2 Calibration of virtual cantilever using thermal fluctuations.

Knowledge of the interaction forces between surfaces gained through AFM is crucial in a variety of applications and necessitates a precise knowledge of the cantilever spring constant. Thermal fluctuation measurement is a good way to validate the cantilever spring constant in experiment [149].

To test whether our spring constant, κ, functions as intended, we compared the observed thermal fluctuations of the cantilever to those expected from the equipartition theorem, $\langle z^2 \rangle = k_b T/\kappa$ [149]. We used the first 3, 10, 20, 50 residues of bR and ran simulations with the first residue attached to the virtual cantilever and the rest of the segment restrained as a rigid body. In this case, we can measure the thermal fluctuation of the tip of the cantilever via the fluctuation of the residue (**Fig. 3.2a**). The square root of the mean fluctuations has a linear relation with the reverse of the square root of the spring constant (**Fig. 3.2b**) [149].

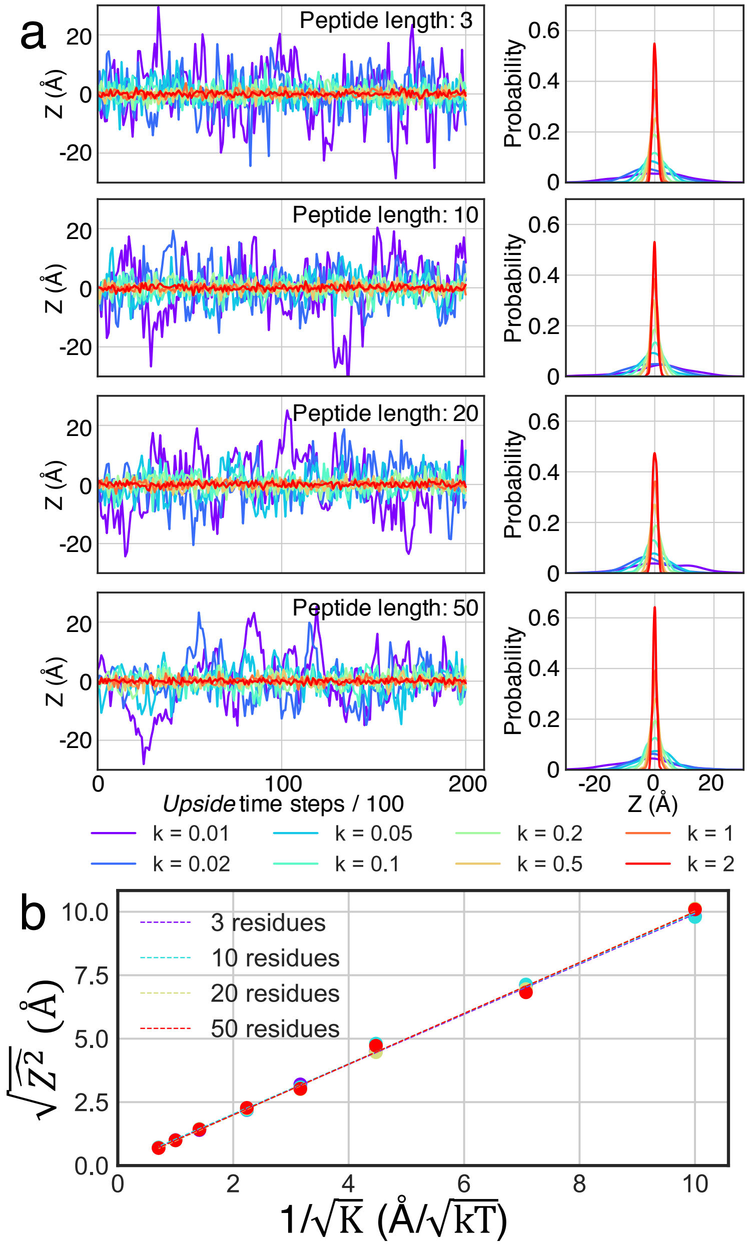

Figure 3.2: **Calibrating our virtual cantilever: Stiffness, thermal fluctuations, and the equipartition theorem**. **a**. Thermal fluctuations and their distributions. A 3-50 residue segment of bR is attached to the tip of the cantilever and the fluctuations of the residue attached at the end to the cantilever are measured (i.e., same location, but with varying mass only). **b**. According to the equipartition theorem, the square root of the mean fluctuations should be equal to the inverse of the square root of the spring constant. This is observed in our simulations.

3.4.3 Calibration of the contour length per amino acid in Upside simulations

Earlier experimental [150] and simulation [142] studies found that Lc per amino acid may be different. The discrepancy may affect the identification of the structural content of the intermediate state.

We ran the simulation of the truncated bR molecules (which are in native orientations as in the whole bR) (**Table 3.4**). The truncation points were chosen to match the experimental intermediates [43]. For example, the truncated bR-A160 has 72 residues, from the C-terminus to the residue before A160. This truncated version has residues 161-232 unfolded to match the intermediates where residues 1-160 are folded while 161-232 are unfolded. We can obtain the Lc of the unfolded segment between the C-terminus and A160 when this truncated bR is fully extended under force.

For each of the truncated bR species, we fit the FEC with a WLC model **Eq. 3.1** of the end-to-end distance and the applied force using a fixed persistence length (Lp) of 0.40 nm (**Fig. 3.3a, b**), to determine Lc values shown as a function of the number of residues (**Fig. 3.3c and tables 3.1, 3.2**).

We obtained a slope of 0.390 nm in agreement with the experimental estimate of 0.40±0.02 nm from experiment [150]. Note that the average distance between consecutive C_α is 0.38 nm from protein structures in the PDB [142]. This value of 0.390 nm is $\sim 7\%$ larger than 0.364 nm, a value recently obtained by a high precision measurement [43]. Remarkably, for the truncated bR molecules, our Lc values exhibit the same minor deviations from linearity as those observed experimentally. The reproduction of these small deviations implies that they are real. The only reasonable source of the variability is a sequence dependent for Lc, consistent with experimental [150] and simulation [142] findings. Beyond providing support for the accuracy of our simulations, the residue dependence should be a useful diagnostic in identifying the sequence of the segment that is unfolded for a given Lc value (**Fig. 3.3c**).

Table 3.1: Contour length (Lc) of bR intermediates.

Truncated bR	Lc, simulation (nm) (Fitted with Lp = 0.4 nm)	Lc predicted by $\Sigma_{sequence}$ Lp(AA) (nm)	Lc, experiment (nm)	Description
A160 (72)	28.4	27.1	26.9	Top of helix E
T157 (75)	29.2	28.4	6.0	
F154 (78)	30.1	29.6	5.0	
V151 (81)	30.9	30.6	5.0	
I148 (84)	31.5	31.5	4.0	
L146 (86)	32.0	32.2	5.0	
A143 (89)	32.7	33.5	4.0	
A139 (93)	33.7	35.0	6.0	
V136 (96)	34.4	36.3	6.0	
S132 (100)	35.2	37.6	4.0	
V130 (102)	35.7	38.4	4.0	Bottom of helix E
K129 (103)	36.1	38.6	4.0	
L127 (105)	37.5	39.4	4.0	Top of helix D
V124 (108)	39.2	40.7	4.0	
I119 (113)	42.1	42.7	4.0	
V101 (131)	52.0	50.0	4.0	Top of helix C
D96 (136)	53.2	51.8	4.0	
P91 (141)	54.3	53.7	4.0	
Y83 (149)	56.1	56.7	4.0	Bottom of helix C
P77 (155)	57.0	59.0	4.0	
F71 (161)	59.0	61.7	4.0	
G63 (169)	62.8	64.6	4.0	Top of helix B
F54 (178)	68.3	68.2	4.0	
V29 (203)	80.1	77.8	4.0	Top of helix A
G16 (216)	83.2	82.8	4.0	
P8 (224)	84.5	86.1	4.0	Bottom of helix A

The truncated bR is named after the structural position (i.e. residue index) of the intermediate, as defined by the last folded residue. Numbers in parentheses indicate the number of residues of the truncated bR molecules. For example, A160 has 72 residues, from residue 161 to residue 232, the C-terminus. The Lc of the truncated bR in its fully extended state in simulation is in the 2nd colunm; the predicted Lc calculated by summing the Lp of each type of amino acids over the sequence is in the 3rd column; and the Lc of unfolded segment of the corresponding intermediate in experiment is in the 4th column.

Table 3.2: Inferred Lc values (in nm) associated with each residue in helices E to A of bR.

Helix E	**A160**	K159	S158	**T157**	F156	G155	**F154**	F153	L152	V151
	28.4	28.7	28.9	29.2	29.5	29.8	30.1	30.4	30.6	30.9
	Y150	L149	**I148**	Y147	**L146**	M145	A144	**A143**	T142	S141
	31.1	31.3	31.5	31.8	32.0	32.2	32.5	32.7	33.0	33.2
	I140	**A139**	W138	W137	**V136**	F135	R134	Y133	**S132**	Y131
	33.4	33.7	33.9	34.2	34.4	34.6	34.8	35.0	35.2	35.4
	V130	K129	T128							
	35.7	36.1	36.8							
Helix D	**L127**	A126	G125	**V124**	L123	G122	T121	G120	**I119**	M118
	37.5	38.1	38.6	39.2	39.8	40.4	40.9	41.5	42.1	42.6
	I117	G116	D115	A114	G113	V112	L111	A110	L109	I108
	43.2	43.8	44.3	44.8	45.4	46.0	46.5	47.0	47.6	48.2
	T107	G106	Q105	D104	A103	D102				
	48.7	49.2	49.8	50.4	50.9	51.4				
Helix C	**V101**	L100	L99	A98	L97	**D96**	L95	L94	L93	L92
	52.0	52.2	52.5	52.7	53.0	53.2	53.4	53.6	53.9	54.1
	P91	T90	T89	F88	L87	W86	D85	A84	Y83	R82
	54.3	54.5	54.8	55.0	55.2	55.4	55.6	55.9	56.1	56.2
Linker	Y79	I78	**P77**	N76	Q75	E74	G73	G72	**F71**	P70
	56.7	56.8	57.0	57.3	57.7	58.0	58.3	58.7	59.0	59.5
	V69	M68	T67	L66	G65	Y64				
	60.0	60.4	60.9	61.4	61.8	62.3				
Helix B	**G63**	L62	L61	M60	S59	L58	Y57	M56	T55	**F54**
	62.8	63.4	64.0	64.6	65.2	65.9	66.5	67.1	67.7	68.3
	A53	I52	A51	P50	V49	L48	T47	T46	I45	A44
	68.8	69.2	69.7	70.2	70.7	71.1	71.6	72.1	72.5	73.0
	Y43	F42	K41	K40	A39	D38	P37			
	73.5	74.0	74.4	74.9	75.4	75.9	76.3			
Linker	D36	S35	V34	G33	M32	G31	K30			
	76.8	77.3	77.7	78.2	78.7	79.2	79.6			
Helix A	**V29**	L28	F27	Y26	L25	T24	G23	L22	G21	M20
	80.1	80.3	80.6	80.8	81.1	81.3	81.5	81.8	82.0	82.2
	L19	A18	T17	**G16**	L15	A14	L13	W12	I11	W10
	82.5	82.7	83.0	83.2	83.4	83.5	83.7	83.8	84.0	84.2
	E9	**P8**	R7	G6	T5	I4	Q3	A2	E1	
	84.3	84.5	84.9	85.3	85.7					

Lc values obtained directly from the simulations of truncated bR are in red. The Lc value is for the unfolded segment C-terminal to a residue. For example, the contour length for the unfolded segment from the C-terminus to K159 (having residues 160 to 232) is 28.7 nm. The last 4 residues (I4, Q3, A2, E1) are not in 1qhj.pdb.

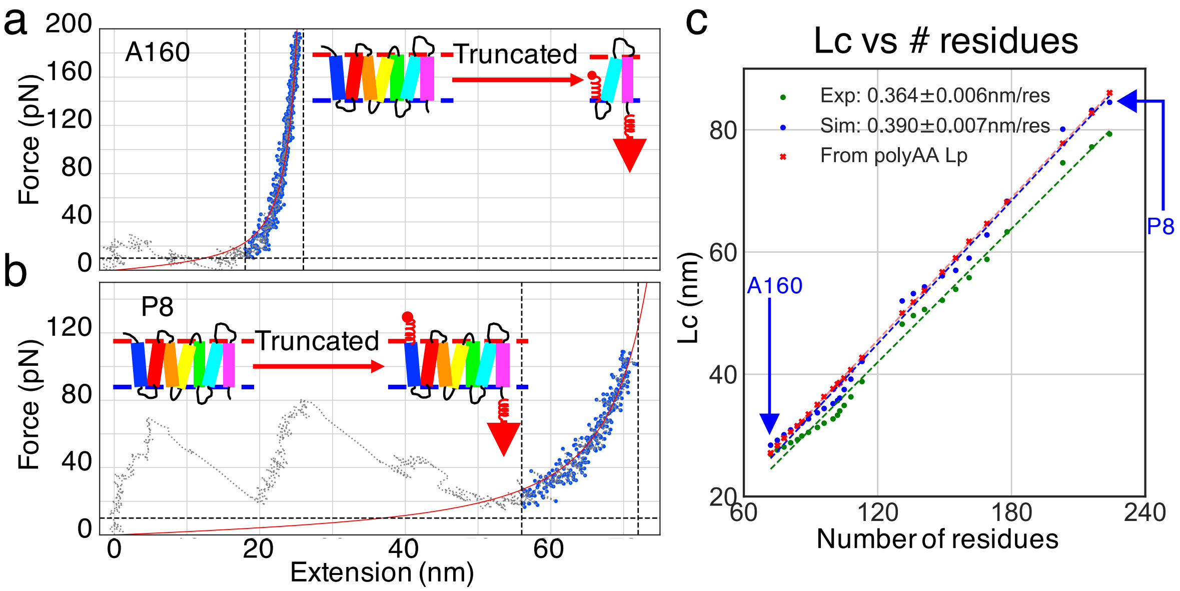

Figure 3.3: **Calibration of contour length (Lc) per amino acid**. **a, b**. WLC fitting of the end-to-end distance (extension) and the force using a fixed Lp of 0.4 nm of truncated bR species A160 and P8, respectively. WLC fitting (curve in red) was performed on the data points in blue between the vertical black dashed lines. The elasticity of the unfolded segment is well described by the WLC model. **c**. Lc of unfolded segment as a function of number of residues from simulations compared to experiment [43]. In the experiment, the number of unfolded amino acids is calculated based on $n_{aa} = (\Delta L_0 + \Delta d)/L_0^{aa}$, where Δd is the vertical distance of the folded structure along the pulling axis in native bR [151] and $L_0^{aa} = 0.366$ nm is the Lc per amino acid based the distance between the first intermediate in the helix pair ED (A160) and the first in the helix A (V29). The same deviations from linearity are observed in the simulations and experiment. This similarity implies that there is a similar sequence dependence that both highlights the accuracy of the simulations, and that the residue dependence could be useful in identifying the sequence of the segment that is unfolded for a given Lc value.

3.4.4 Obtain the Lc value for unfolded segment of bR associated with each residue in the sequence.

Lc values associated with each residue were interpolated based on **Table 3.1**. Therefore, given the structural position (i.e. residue index), we can infer the Lc value of that unfolded segment of bR associated with that residue.

3.4.5 Force clamp simulations of ubiquitin.

We unfolded ubiquitin (1ubq.pdb) to its fully extended state under high force (800 pN) and ran force clamp simulations with a constant force applied to both ends of the protein (procedure described in **Table 3.7**). We replicated all-atom MD results [142] in a few cpu-hours. Without force, the highly stretched polypeptide contracts considerably but remains extended under force as low as 10 pN (**Fig. 3.4a**). The distributions of (ϕ, ψ) angles and end-to-end distances at different forces were reproduced those of the all-atom MD study [142] (**Fig. 3.4b, c**). Also, we obtained good fitting of the average end-to-end distances and the applied forces according to the WLC model (**Fig. 3.4e**). The Lc was determined at the minimum fitting error (**Fig. 3.4d**).

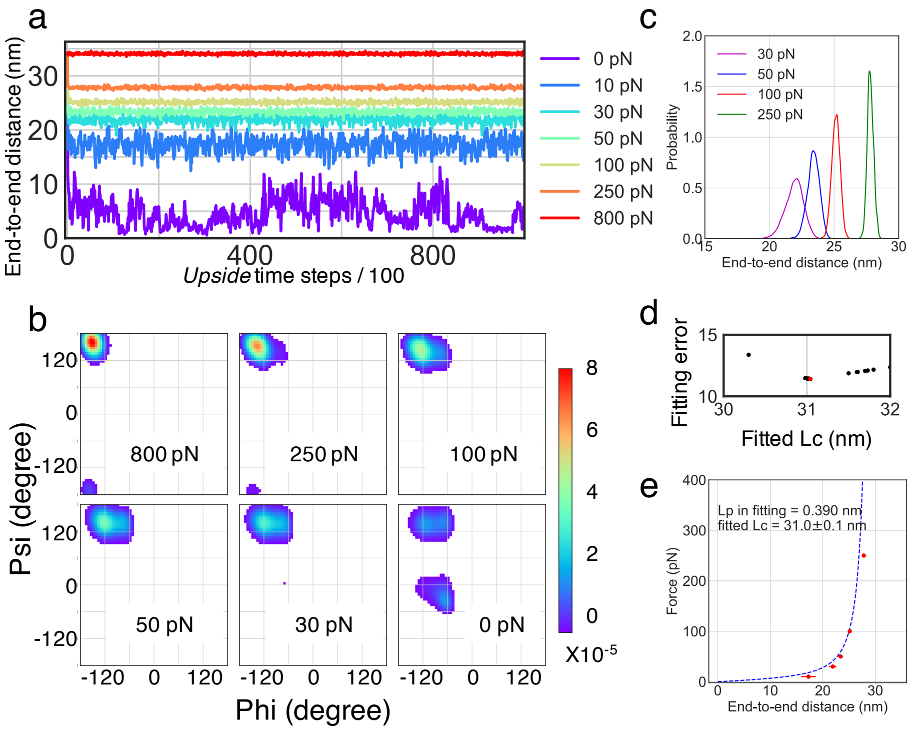

Figure 3.4: **Reproduction of the all-atom MD of unfolded ubiquitin**. **a**. End-to-end distances of the protein under force. **b**. Distribution of ϕ, ψ angles of all 76 residues of the protein over time. **c**. Distribution of end-to-end distances. **d**. Fitting error versus fitted Lc. The minimum fitting error was obtained at Lc = 31.0 nm (red dot). **e**. WLC fitting of the average end-to-end distance and the applied force. The error bars show the standard deviation of the end-to-end distances. When fitting the data to obtain the Lc value, the value of Lp was fixed at 0.39 nm.

3.4.6 Force clamp simulations of 76-residue homopolymers.

First, we emphasize that it is the "*chicken or the egg*" dilemma for knowing Lp and Lc using the WLC model (**Eq. 3.1**) given the applied force and protein extension. Second, We notice that experimentally measured Lp (0.40 nm) [132] and Lc (0.40±0.02 nm/residue) [150] are very close (under this condition). Therefore, we assume that the Lp value averaged by the protein sequence can be approximated by Lc = Lp·N, where N is the number of residues. Lc = Lp·N is equal to be expressed as Lc = $\Sigma_{sequence}$Lp, which means the change of Lc is due to the removal or induction of residues.

Because we have calibrated the Lc per amino acid in previous section and obtain Lc per amino acid as 0.390 nm in our simulations, we approximate the Lp value as 0.390 nm in our simulation. Knowing Lp enables us to obtain Lc of systems other than bR using the WLC model, given force and extension.

The "*chicken or the egg*" dilemma for knowing Lp and Lc can be illustrated by the following 2 steps.

Step 1.

WLC model + experiment Lp (0.40 nm)
Lc per amino acid (0.390 nm) (obtained by simulations of truncated bR molecules)
updated Lp (0.390 nm)

Step 2.

WLC model + updated Lp (0.390 nm)
Lc of other systems, such as 76-residue poly-alanine

Now, we hypothesized that the strong correlation between the Lc values obtained in simulation and experiment (**Fig. 3.3c**) may result from the dependence on sequence. Accordingly, for each type of amino acid, we pulled on 76-residue homopolymer to its fully extended state and then reduced the force to a constant value (procedure described in **Table 3.2**). Similar to ubiquitin (Ub), we obtained good WLC fitting of the average end-to-end distances and the applied force. We can compute the Lp value for each type of amino acid, denoted as Lp(AA), using **Eq. 3.1** with F = 200 pN, x = end-to-end distance under 200 pN, and the Lc value of that 76-residue homopolymer.

If Lp is independent of residue type, Lp(AA) should be 0.390 nm regardless of the residue type. However, we found that Lp ranges from 0.24 nm for Val to 0.567 nm for Asp. Although Pro76 has the smallest Lc (26.6 nm), Pros Lp is relatively large (0.353 nm). The results are listed in **Table 3.7**.

If we add up the Lp(AA) values for the Ub sequence, we obtain 29.3 nm, different from the Lc(Ub) (31.0 nm) obtained from our simulation (**Fig. 3.4e**). Therefore, Lc(Ub) $\neq \Sigma_{Ub\ sequence}$ Lp(AA). On the other hand, we can predict the Lc values of truncated bR species by summing the Lp(AA) values. The predicted Lc values have in a linear relationship with the number of residues (**Fig. 3.3c and table 3.3**).Therefore, Lc(truncated bR)= $\Sigma_{sequence\ of\ truncated\ bR}$ Lp(AA).

We can see that Lc = $\Sigma_{sequence}$ Lp or equally Lc = Lp·N only works for bR but not for Ub. The is because the Lp value (0.390 nm) obtained in Step 1 is averaged by the sequence of bR not by Ub. In other words, if the Lp value updated in Step 1 is obtained by simulations of truncated Ub molecules, Lc = $\Sigma_{sequence}$ Lp will work for Ub but not bR.

The mild dependence of Lp on amino acid type in a manner that matches experiment (**Fig. 3.3c and table 3.3**) suggests that the calibration of Lp should be carried out for each system if the analysis solely relies on the WLC model. Not calibrating the Lp value or the Lc value per amino acid may lead to identifying intermediates incorrectly using the WLC model, because the same force, extension but different Lp will result in different Lc, which in turn leads to different structural position of the intermediate. The residue-dependence of the Lp will help identify the boundaries of the unfolded chain segment.

Table 3.3: Persistence length (Lp) of 76-residue homopolymers.

76-residue homopolymers	Lc (nm)	Average end-to-end distance (nm) under 200 pN	Lp(AA) (nm)
Ala^{76}	31.7	28.3±0.2	0.454
Val^{76}	31.4	26.7±0.2	0.240
Ile^{76}	29.9	26.2±0.2	0.345
Leu^{76}	30.1	26.5±0.2	0.369
Met^{76}	30.9	27.3±0.3	0.397
Phe^{76}	31.6	27.5±0.2	0.322
Tyr^{76}	31.7	27.8±0.2	0.352
Trp^{76}	31.5	27.9±0.2	0.414
Arg^{76}	31.3	27.6±0.2	0.384
His^{76}	30.8	27.6±0.3	0.488
Lys^{76}	31.1	27.4±0.3	0.381
Asp^{76}	31.0	27.5±0.3	0.407
Glu^{76}	31.0	27.4±0.3	0.386
Ser^{76}	31.5	27.9±0.2	0.416
Thr^{76}	31.4	27.5±0.2	0.354
Asn^{76}	30.4	27.5±0.3	0.567
Gln^{76}	31.0	27.2±0.3	0.364
Cys^{76}	31.1	27.5±0.3	0.389
Gly^{76}	31.0	27.9±0.2	0.525
Pro^{76}	26.6	23.3±0.2	0.353

3.5 Unfold Bacteriorhodopsin (bR)

3.5.1 Unfold monomeric bR in an accuracy comparable to experimental results

Monomeric bR is placed within an implicit membrane bilayer modeled using our new membrane burial potential [48]. Force is applied with a virtual cantilever spring attached at the C-terminus and is increased by moving the spring at a constant velocity normal to the bilayer (z-direction) (**Fig. 3.5a**). Force typically accumulates to ∼100 pN at which point an unfolding event occurs that allows the spring to return towards its equilibrium position. The magnitude of the force change, δF, is proportional both to the length δl of the newly unfolded segment and the stiffness of the cantilever kcantilver, δForce = $-\delta l \cdot \kappa_{cantilever}$. Each such unfolding event signals the presence of an intermediate or the release of the entire protein from the bilayer. Repetition of these unfolding events produces a force-extension curve, FEC, with a sawtooth pattern that recapitulates key features of the experiments, including the extension of the unfold segments being well described by the WLC model [43, 132, 134].

Because the conformation of the protein is known at every time point in simulation, we can identify the FEC's sawtooth pattern as reflecting the sequential unfolding of pairs of TM helices in the order GF, ED, and CB. The pairwise unfolding of TM helices is a well-known consequence of the up-down topology of the protein (pulling a single helix out would yield an energetically unfavorable conformation with an unfolded segment traversing the bilayer). The first, GF helix pair unfolds relatively quickly because it is connected directly to the cantilever and because force rapidly accumulates upon movement of the spring. The FEC for the rest of the trajectory is dominated by the build-up of force as the unfolded segments are stretched (an entropic tension), punctuated drops in force reflect the unfolding of the pairs of helices. The final step arises when the A helix is extracted from the bilayer.

Our study sets the velocities of the cantilever and spring constant to 10^6 nm/s and 0.05 kT/Å^2 (0.2 pN/nm at 298 K), respectively, chosen partly to match the experimental

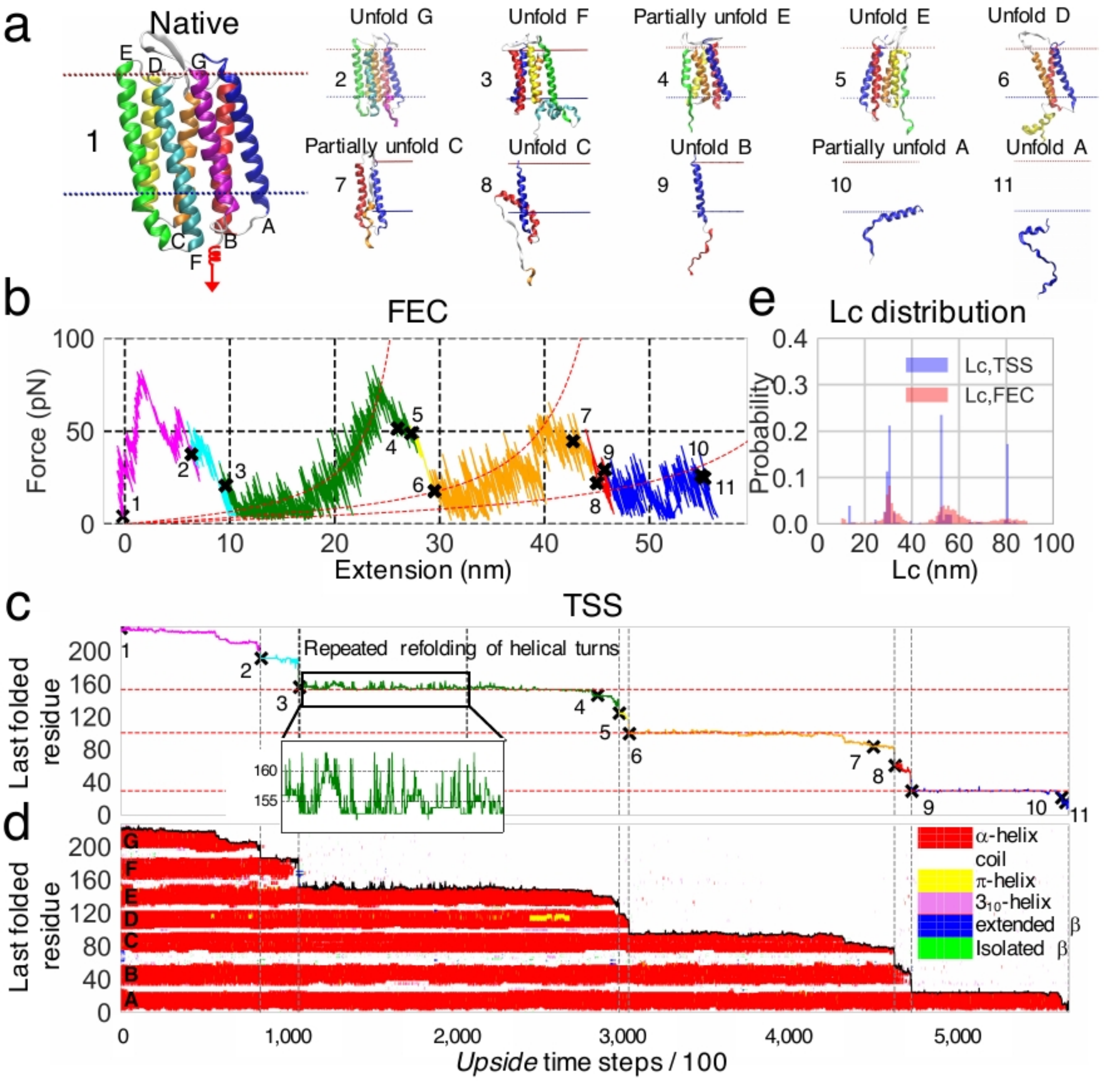

Figure 3.5: **A representative unfolding trajectory of bR**. **a**. Typical species along the unfolding trajectory. **b**. The FEC (multi-color trace). The red dashed curves are fit to the WLC model, using the contour lengths (Lc) of the most populated states identified in panel **c**. **c**. The index of the most C-terminal residue that remains folded obtained from the time course of secondary structure formation (TSS, panel **d**). This index is used to identify the folded regions in helices G to A (black labels at time step zero). The "X" and associated number in panels **b** and **c** correspond to reference points in the trajectory in panel **a** that serve to connect the FEC to the TSS. The red horizontal dashed lines in panel c identify the most populated intermediates during the unfolding of the ED and CB helix pairs, and helix A. Computations of secondary structure follow the DSSP convention [152]: coil refers to either H-bonded turn, bend or loops and irregular elements. Grey vertical dashed lines in panels **c** and **d** define the time points where a given TM helix has completely unfolded. **e**. Probability distributions of the Lc obtained from the FEC of the unfolding of bR (as in panel **b**) fit with the WLC (**Eq. 3.1**), and TSS (as in panel **c**) from direct identification of number of unfolded residues in a trajectory, followed by simulations of these shorter segments and fitting to a WLC model (**Table 3.2**).

sawtooth pattern obtained by Perkins and coworkers [43]. Their FEC traces are similar over the wide range of velocities and spring constants of 30-3000 nm/s and 13-58 pN/nm, respectively. The major intermediate of ED helices exhibits an average unfolding force of 84±3 pN, close to the experimental value of 94±1 pN [43]. Generally, the use of either a faster pulling speed or weaker spring constant produces a FEC with a shallower sawtooth pattern as the force has insufficient time or distance, respectively, to relax back to a zero force condition. For example, altering our speed by a factor of 2 or the spring constant by a factor of 5 has minimal effect on the depth of our sawtooth pattern (**Table 3.4**), thereby assuring the important conditions that the simulations employ sufficiently slow speed and/or low pulling force to be able to match experiments.

Table 3.4: Comparison of the mean unfolding force (in pN) and the s.e.m. for bR intermediates between experiment and simulation.

	Major intermediate in ED helix pair	Major intermediate in CB helix pair	Major intermediate in helix A
Experiment	94±1 (A160)	49±2 (V101)	62.0±0.6 (V29)
k = 0.05, v = 0.001	83.8±2.6 (F153) [a]	43.3±2.1 (L100)	22.6±1.6 (V29)
k = 0.01, v = 0.001	69.7±1.4 (F153)	41.2±1.0 (L100)	22.5±0.9 (V29)
k = 0.05, v = 0.0005	76.9±2.6 (F153)	40.4±2.5 (L100)	18.8±1.9 (V29)

The spring constant (κ) is in kT/Å^2; the pulling velocity (v) is in Å/*Upside* time step; and the temperature (T) is in *Upside* temperature unit (1 ≈ 300 K). The major intermediate identified in each region is put in the parentheses, as indexed by the last folded residue.
a. The comparison between experimental [43] and simulations is conducted for the most populated intermediate, which is given in the parentheses; however, we also observe a K159 intermediate in the simulations which corresponds to the major experimental intermediate.

The simulations are conducted without the retinal, which is attached to bR's G helix [153] and stabilizes the protein [154]. Hence, presumably more force would be required to remove the G helix in the holoprotein. However, the G helix is removed first, and, hence, the rest of the trajectory should be unaffected. As just noted, the average force for the next unfolding event, the unfolding of the ED helical intermediate, agrees well with the experiment. In addition to calculating the unfolded lengths using the FEC, the position of the last remaining folded residue is identified directly from the simulations and used to construct a plot highlighting the structure of the intermediates, their boundaries, and the lengths of the unfolded segments (**Fig. 3.5c**). The structured regions are plotted as a function of time to generate a timeline of secondary structure (TSS, **Fig. 3.5d**).

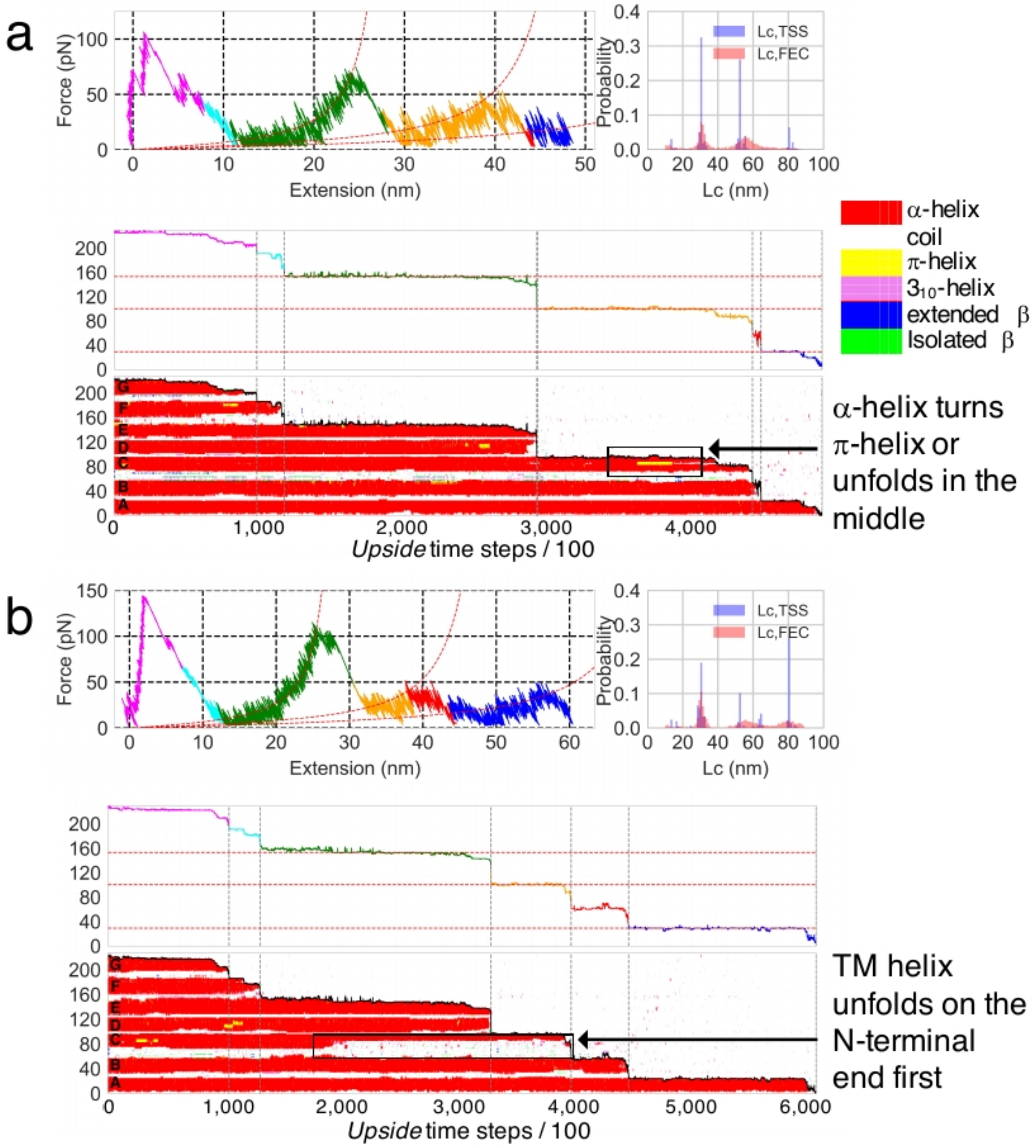

Figure 3.6: **Unfolding trajectories of bR**. The trajectories largely support the common assumption that secondary structures remain intact within the membrane bilayer during the unfolding process. However, exceptions can be seen in these two TSS: **a**. Part of a TM helix may turn into p-helix or unfold in the middle (e.g., Helix C, black box in the TSS plot). **b**. A TM helix can unfold from the N-terminal end rather than the C-terminal end (e.g., Helix C, black box in the TSS plot).

The contour lengths (Lc) and folded regions are inferred from the FEC. These regions and those explicitly identified by the TSS are similar (**Fig. 3.5e**). This agreement supports the standard assumption that the whole length of the unfolded segment is at the C-terminus of the protein, while the remaining portion of the protein stays intact, with the helices remaining stationary within the bilayer [43]. However, the agreement is not absolute as our simulations find that partially unfolded helices can translate vertically in the bilayer (**Fig. 3.5a10**), or change from α-helix to 3_{10} helix or π-helix (**Fig. 3.6a**), and even unfold at the amino terminus (**Fig. 3.6b**). These events are likely to be missed in an experiment.

To further examine the agreement between our simulations and experiment, we compare the populations and structures of the intermediates. Following the procedure employed for soluble proteins [127], the population distribution of intermediates obtained from the TSS are fit with multiple Gaussian functions, assuming an uncertainty of one amino acid (**Fig. 3.8**). We identify 29 intermediates with 15, 11 and 3 having folded-unfolded boundaries in the ED, CB, and A helices, respectively (**Fig. 3.7 and table 3.5**). Among this group of 29 intermediates, 11 correspond exactly to one of the 26 experimental intermediates [43] and another 10 are within one residue of an experimental intermediate. We fail to identify 5 intermediates (3 near the bottom of the E helix, 1 in the middle of the loop connecting the CB helices, and 1 at the bottom of the A helix) while identifying 8 that are not observed experimentally (**Table 3.5**).

The disparity between simulations and experiment in identifying intermediates may reflect real differences, such as errors in our energy function, pulling speed, or effective temperature. However, the different protocols - using either experimental FEC or the simulated TSS - for identifying intermediates, can also affect the determination of the intermediates. For example, we observe unfolding occurring N-terminal to last folded helix (**Fig. 3.6b**), a possibility that is not allowed in the experimental analysis.

An impressive feature of the recent experimental AFM study [43] is the ability to observe back-and-forth unfolding and refolding of small 2-4 residues units representing half to a full

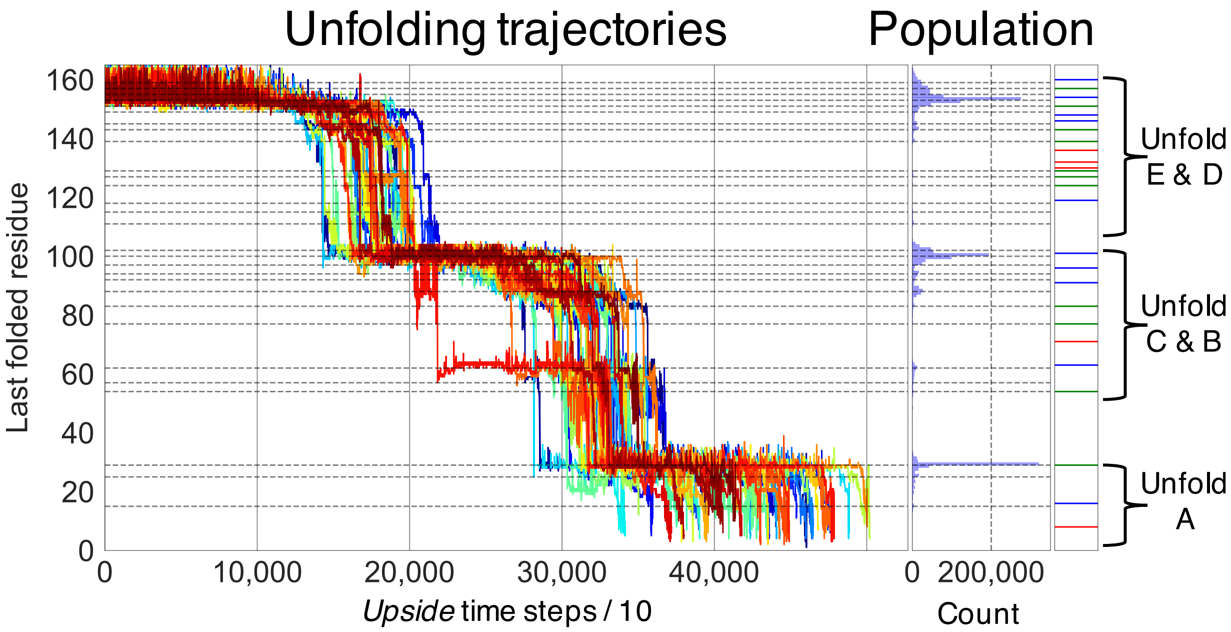

Figure 3.7: **Unfolding trajectories and intermediates of bR**. Time dependent unfolding trajectories plotted according to the index of the last folded residue (left), the corresponding population distribution (middle), and the intermediates found in experiment [43] (right). Of the 90 total trajectories, only the 48 where helix A unfolds by itself are presented (in the rest of the rest of trajectories, more than 2 TM helices are pulled out of the membrane in the last observed unfolding event). For the 48 trajectories, the time spent at each position is histogrammed (blue bars) and fit using multiple Gaussians with standard deviation (σ) of one residue to identify the number and position of simulated intermediates. Black dashed lines (left and middle) designate the intermediates found in our simulations. The blue, green and red solid lines (right) indicate intermediates identified in experiment and also found within one residue, but not observed in our simulations, respectively.

helical turn. We likewise observe these back- and-forth transitions between micro-states in all three major regions (ED, CB and A helices) (**Figs. 3.5c and 3.7**). That *Upside* captures these events and identifies most intermediates is a positive indication that we have achieved a relatively accurate representation of the system, especially in terms of the amount of force and the loading rate, as well as in the balance of forces such as in having an appropriate energetic penalty for unfolding portions of helices within the bilayer.

Table 3.5: Comparison of bR intermediates identified in the 2017 experiment, our simulation and a 2016 CG study.

2017 experiment [43]	Our simulation[a]	2016 CG study[b] [47]	Description
160	159		Top of helix E
157	157	156.8	
154	155, 153		
151	151	150.8	
148	149		
146	145		
143	143		
139	139	140.7	
136			
132			
130			Bottom of helix E
129	129		
127	127		Top of helix D
124	124		
119	118		
	115		
	111		
101	102, 100	101.4	Top of helix C
96	97	95.4	
	94		
91	92		
	88	89.0	
83	83		Bottom of helix C
77	77	75.7	
71			
63	62		Top of helix B
	57		
54	54		
		33.8	
29	29	29.6	Top of helix A
	25		
		21.2, 19.1	
16	15		
8			Bottom of helix A

a. k = 0.05 kT/Å^2, v = 0.001 Å/*Upside* time step, T = 1.0 ≈ 300 K.
b. See Fig. 5A in ref. [47], the intermediates are taken from the analysis of force peak groups, which were compared to previous experiments [135, 155, 156]. The position of an intermediate is given by the last folded residue of that intermediate in the protein.

3.5.2 Methods

Structure and sequence of bR

The bR structure (1qhj.pdb) and orientation within the lipid bilayer was obtained from the OPM database [76]. The membrane thickness was set to 30.0 nm as identified by OPM. Truncated versions of bR (used in the calibration of the contour length per amino acid) were made from the native structure of bR, which are in the native orientations as in the whole bR.

Unfolding pathway analysis

For every frame in trajectory, the Lc of already unfolded segment can be determined through FEC or TSS. Assuming intact secondary structure remains unchanged within the bilayer, Lc is uniquely determined (labeled as Lc,FEC) given a force and an extension, from which we can infer how many residues have unfolded. Force was measured and recorded into the H5 file during the simulation, while extension was calculated as the distance that the Cα atom of the C-terminus has moved.

On the other hand, if the number of unfolded residues is known, Lc can be determined by mapping the number of unfolded residues to pre-determined Lc value (labeled as Lc,TSS). Secondary structures were computed by the `compute_dssp` function in *MDTraj* [82], which follows the DSSP definition [52]. Then Lc,TSS is obtained after identifying the most C-terminal residue which remained folded.

Trajectories plotted according to the last (C-terminal) folded residue were smoothed by a Savitzky-Golay filter [86] in *Scipy* [88], in which the `window_length` was set to 11, `polyorder` 3, `mode` nearest. Then, the population distribution was histogrammed and fitted with multiple Gaussian functions to identify the number position of the simulated intermediates. Amplitudes and positions were fit assuming a width (standard deviation) of one amino acid, i.e. the positional uncertainty is assumed to be ±1 amino acid. Three major unfolding

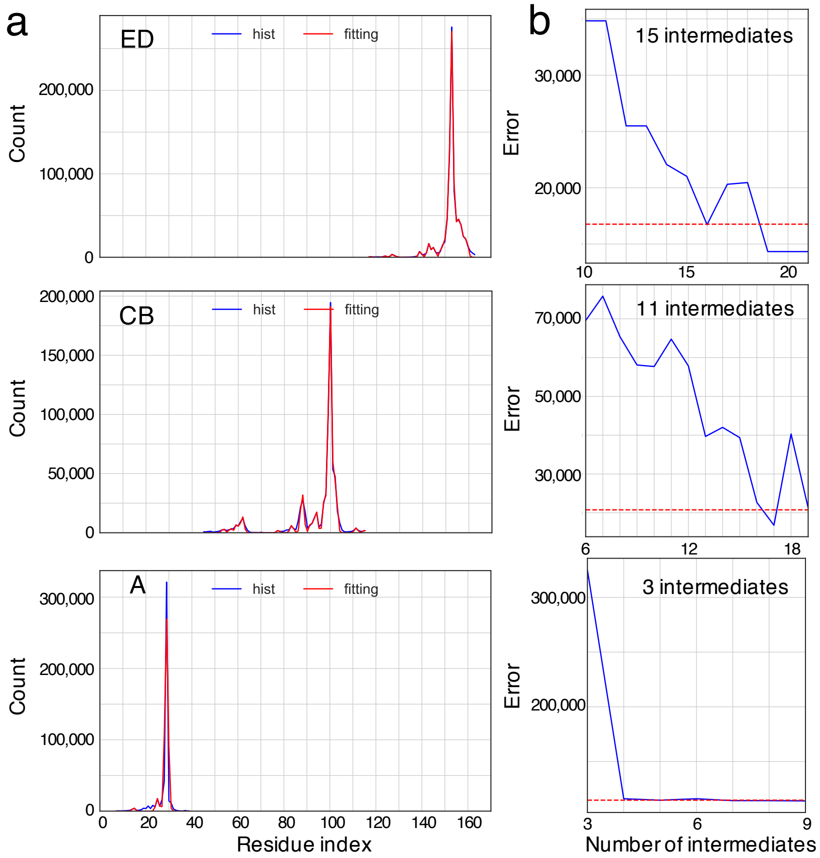

Figure 3.8: **Identifying intermediates by fitting with multiple Gaussian functions**. **a**. For the 48 trajectories shown in **Fig. 3.7**, the time spent at each position is histogrammed (blue lines) and fit using multiple Gaussians (red) with standard deviation (σ) of one residue to identify the number and position of intermediates. The number and initial position of the Gaussians was manually adjusted to minimize the fitting error; additional Gaussians were added until the error plateaued. The upper, middle and lower panels refer to unfolding occurring within the ED, CB or A helices, respectively. The index refers to the last residue that remains folded, as identified in the TSS, and is listed in **Table 3.5**. **b**. Fitting error as a function of number of intermediates provided the intermediates are evenly distributed within each unfolding region (blue). The fitting error after manual adjustment of the number of intermediates and their positions (red dashed).

regions, denoted ED, CB, and A, were fit separately. In general, the more intermediates, the smaller the fitting error. To prevent over-fitting, we initially assume that intermediates are evenly distributed within each major unfolding region and obtained the fitting error as a function of the number of intermediates. Later, by adjusting the number of intermediates and their initial positions manually, we acquired fewer intermediates with a relatively low fitting error (**Fig. 3.8**).

3.6 Unfold GlpG

3.6.1 GlpG can unfold along multiple routes with well-populated intermediates.

We repeated the pulling simulations on GlpG using the same pulling velocity and spring stiffness as used in the bR simulations. Experiments with GlpG [130] can be replicated by attaching two virtual springs to the N- and C- termini that are on the same side of the bilayer (**Fig. 3.9a**). The C-terminal spring is translated horizontally to the membrane surface to generate force parallel to the surface (the other spring is held fixed, but similar outcomes are produced when the N-terminal spring moves at the same net velocity and the C-terminal spring is fixed, **Fig. 3.17 and table 3.6**).

Figs. 3.9e and 3.12 illustrate the diverse set of pathways that emerge for unfolding beginning at either terminus or, more rarely, starting at the central helices. Unfolding from the N-terminus (on the N→C Pathway) typically proceeds sequentially for 3 helices, TM1→TM2→TM3 followed by the unfolding of TM4-6 in any order (**Figs. 3.9c, 3.9e left**, 29 of 50 trajectories in **3.12a, and 3.10**). Unfolding from the C-terminus (C→N Pathway) typically proceeds sequentially TM6→TM5→TM4→TM3→TM2→TM1 (**Figs. 3.9e middle**, 3 of 50 trajectories in **3.12d, and 3.11**).

In the FEC plots **Fig. 3.9d**, the forces generated by both springs are very similar, indicating that the force had time to equilibrate across the protein, a necessary condition for making comparisons to minute long experiments where the force loading rate is below 1 pN/s [130]. The measurement of force at both ends track with each other except an occasional small lag at one end of the protein when an unfolding event occurs at the other end of the protein. For instance, when the TM1 helix unfolds first, the force measured at the N-terminus drops faster than that at the C-terminus (**Fig. 3.9d left**, at extension ~5 nm), whereas the force at the C-terminus drops faster when TM helices close to that end unfold first (**Fig. 3.9d middle**, at extension ~10 nm and 20 nm).

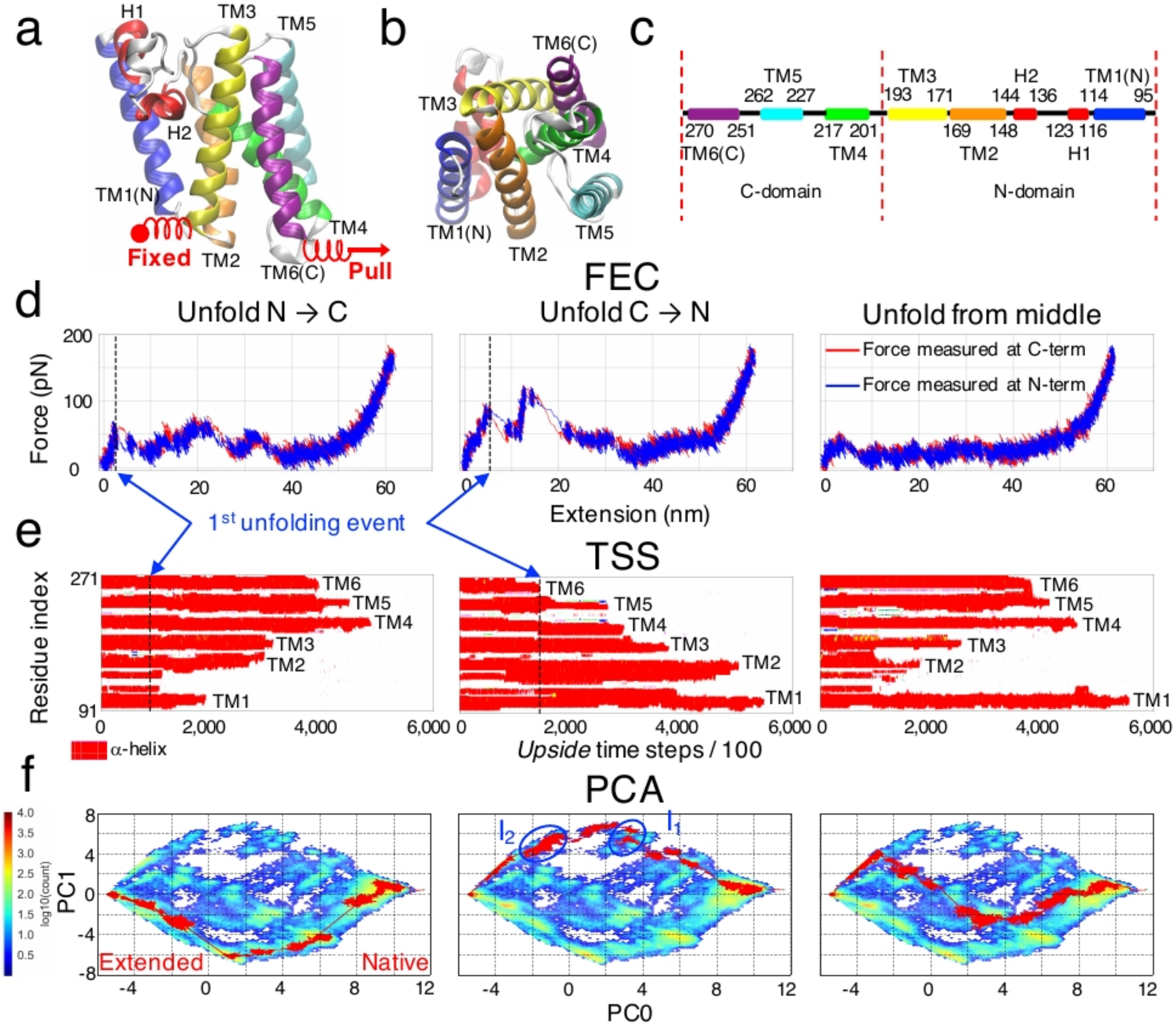

Figure 3.9: **Diversity in unfolding pathways and intermediates of GlpG**. **a, b**. Side and bottom views of GlpG (2xov.pdb) and **c**. secondary structure and definition of N- and C-domain. **d, e, f**. FEC, TSS and PCA plots illustrating unfolding beginning at the N-terminus (left; 14^{th} trajectory shown in **Fig. 3.12a**), at the C-terminus (middle, 1^{st} trajectory in **Fig. 3.12d**), and at the middle of the protein (right, 41^{st} trajectory in **Fig. 3.12c**). In panel **e**, each red strip in the TSS plot represents one helix. Unfolding pathways are defined by the order of TM helix unfolding. In panel **f**, the PCA heat map is evaluated from 50 trajectories, while the red curve depicts the trajectory for a single unfolding pathway from the native state to the fully extended state. The two blue circles in the middle subplot of panel **f** replicate the two experimentally observed unfolding intermediates I_1 and I_2, formed by the unfolding of the TM6 and TM5 helices, followed by the TM4 and TM3 helices, respectively, with the final step being the unfolding of TM2 and TM1 helices [130].

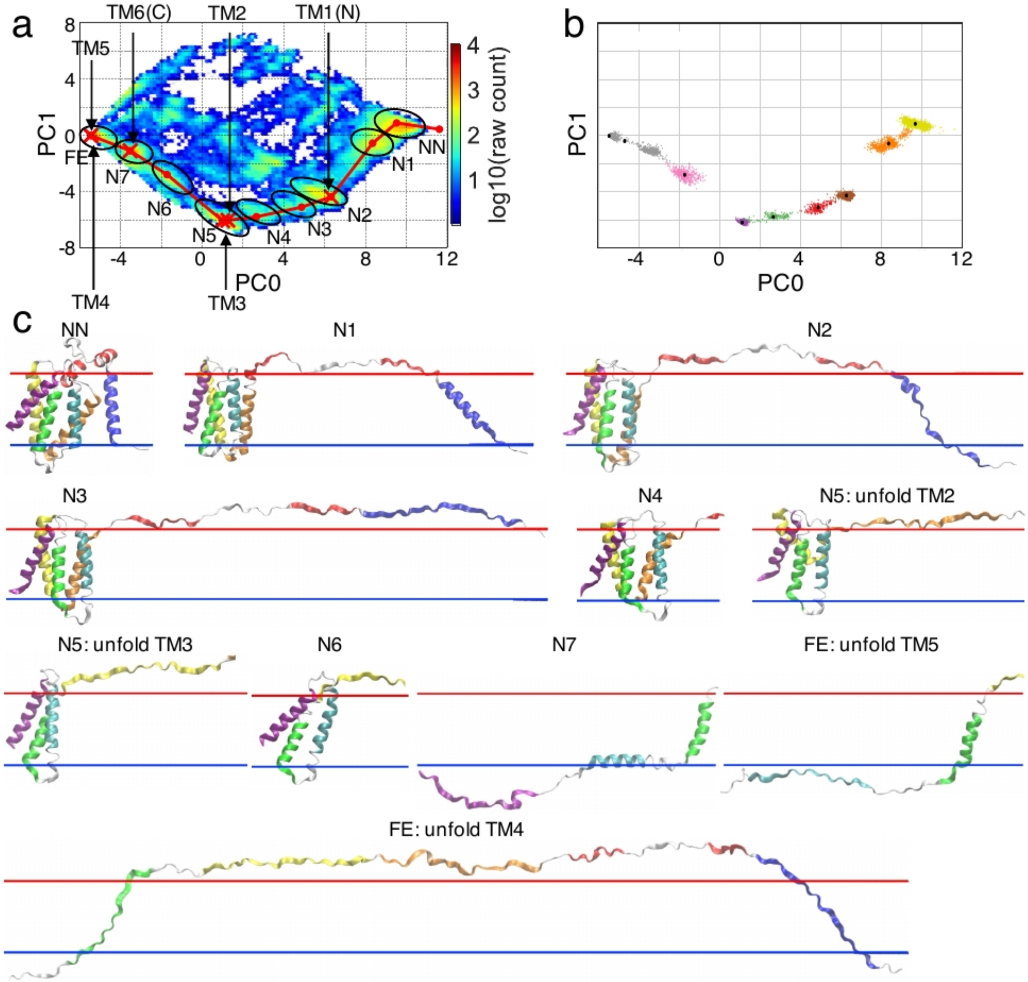

Figure 3.10: **Examples of an N→C unfolding pathway of GlpG. a. Unfolding pathway connected by representative structures on the PCA plot**. The representative intermediates are chosen such that they are either the cluster center or the structure when a TM helix unfolds. These structures are considered as the intermediates. We use NN, N1, . . ., N7, and FE to denote the clusters as well as the intermediates. NN is short for near-native, and FE fully-extended. **b. Clustering analysis of the trajectory**. Nine clusters are identified. **c. Snapshots of the representative structures**. For the illustrative reasons, unfolded segments sometimes are not shown in the snapshots when there is no significant conformational change.

In the NN state, helices rearrange.
In N1, the two interfacial helices H1, H2 unfold and separate.
In N2, TM1(N) unfolds.
In N3, TM1 flips to the other side of bilayer.
In N4, TM4 partially unfolds in its C-terminal.
In N5, TM2 and TM3 unfold.
In N6, the helices in the C-domain rearrange.
In N7, TM6 (C) unfolds, and the C- terminus of TM4 refolds.
In FE, TM5 and TM4 unfold, unfolded TM1 may re-enter the bilayer.

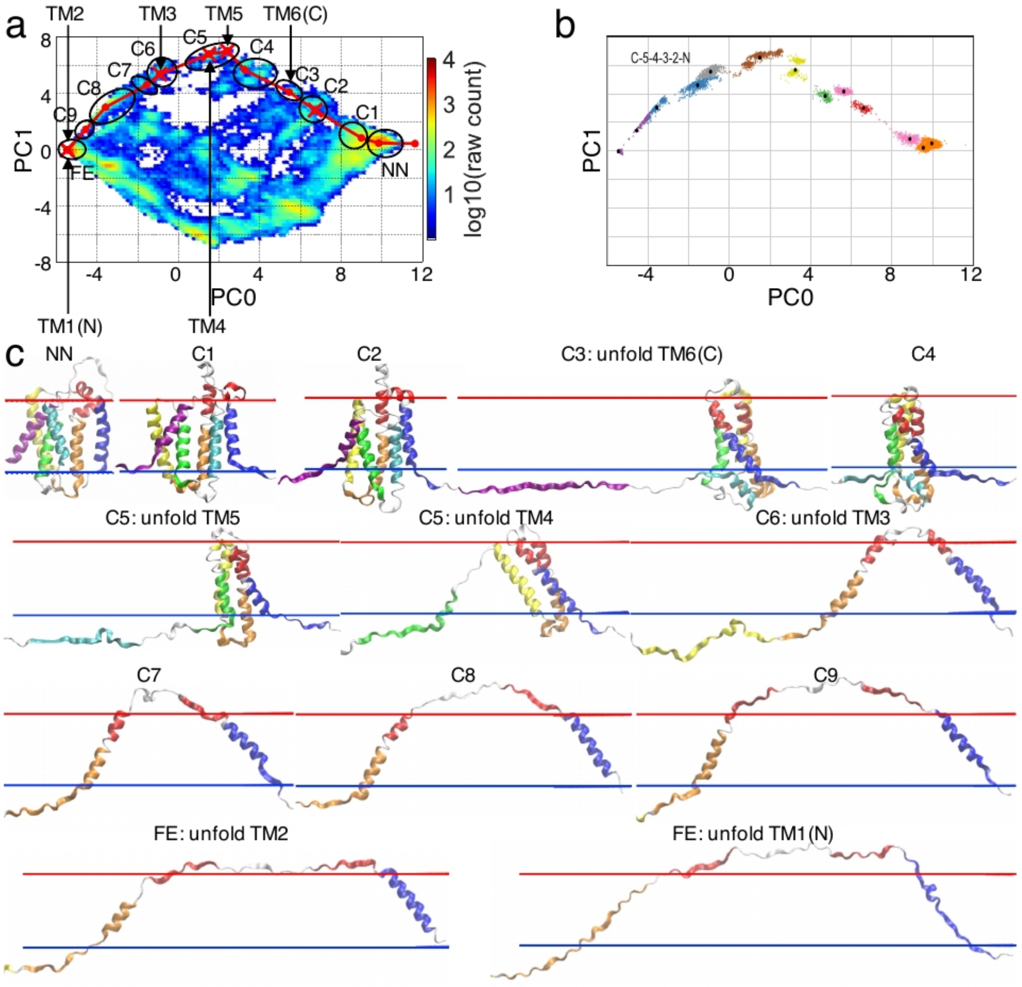

Figure 3.11: **Example of a C→N unfolding pathway of GlpG. a. Unfolding pathway connected by representative structures on the PCA plot**. We use NN, C1, . . ., C9, and FE to denote the clusters as well as the intermediates. **b. Clustering analysis.** Thirteen clusters are identified. **c. Snapshots of the representative structures**.

In the NN state, helices rearrange.

In C1, the interfacial helix H2 aligns with TM2 and pushes part of TM2 out of the bilayer, TM2 bends, TM1 (N) and TM6 (C) partially unfold, and H1 unfolds.

In C2, TM6 unfolds one more helical turn, TM2 partially unfolds in its C-term and H1 refolds.

In C3, TM6 unfolds, TM5 comes out of the bilayer, H1 aligns with TM1, and TM1 tilts in order to accommodate the elongation in its C-term due to the alignment of H1.

In C4, two more helical turns of TM1 unfold, and TM4 and TM5 partially unfold.

In C5, TM5 and TM4 unfold.

In C6, TM3 unfolds, TM1 and TM2 come apart.

In C7, H1 unfolds.

In C8, TM1 and TM2 come further apart.

In C9, H2 unfolds.

In FE, TM2 and TM1 unfold.

Unexpectedly, the more stable N-terminal domain [137] is more likely to unfold before the C-terminal domain under force. This counterintuitive observation arises because of the differential hydrophobicity of the TM helices [157]. The TM2, TM5 and TM1 helices are the most hydrophobic, while the TM6 helix is the least. Although hydrophobicity promotes helix insertion into the bilayer, it has the complementary effect of promoting dissociation of helices from the other hydrophobic helices. In fact, the TM5 helix is completely dissociated from the other TM helices in the crystal structure of GlpG [34]. The dissociation of the TM1 and TM2 helices is energetically less costly than the dissociation of the TM6 and TM4 helices, a process that exposes polar and charged groups (**Fig. 3.13a**, and the near-native and N1 intermediates in **Fig. 3.10c**). In addition, the 34-residue segment between the TM1 and TM2 helices is of sufficient length to allow the TM1 helix to dissociate and remain upright in the bilayer. However, the linker between TM5 and TM6 has only 10 residues, so TM6 must tilt into the bilayer or the GlpG structure must be distorted for TM6 to dissociate. As a result, the C→N pathway does not involve the TM6 helix dissociation; rather the ends of the TM1 and TM6 helices unfold (**Fig. 3.13b**, C2 intermediate in the C→N pathway in **Fig. 3.11c**). These multiple factors explain the preference for unfolding to occur along the N→C pathway.

A principal component analysis (PCA) is performed to interpret high-variance collective protein motions observed along the simulated unfolding pathways [144]. Internal coordinates, such as inter-residue separations, are a poor choice for analysis as very different conformations can have the same distance, e.g., the N-to-C distance can be the same but very different structurally depending on whether unfolding begins at the N- or C-terminus. Hence, the native Cα-Cα contacts are used to derive the principal components (**Fig. 3.9f**). As observed in the PCA heat maps, the unfolding pathway can begin from either the N- (**Figs. 3.9f left and 3.12a**) or C-terminus (**Figs. 3.9f middle and 3.12d**) and proceeds along the lower or upper edge of the heat map, respectively. For unfolding beginning in the middle of the protein, the pathway traverses the center of the map (**Figs. 3.9f right and 3.12c**). Beyond highlighting the pathway multiplicity, the PCA heat maps also emphasize that unfolding occurs through about 10 microstates (**Figs. 3.10 and 3.11**).

3.6.2 Cooperative unfolding of GlpG by force clamping with a weak spring (MT mode)

The degree of pathway diversity deduced from the simulations discussed in the previous section departs with that emerging from magnetic tweezers (MT) measurements that concluded unfolding initiates from the C-terminus [130]. This conclusion emerged from the identification of two intermediates lacking the C-terminal helices. The difference between our simulations and the MT experiments is a result of a difference in force application accentuated by limited experimental time-resolution. Our unfolding simulations mimic an AFM measurement in that force builds up as the cantilever is translated, followed by a rapid drop in the force after each unfolding event as the newly unfolded region provides sufficient slack to allow the cantilever to relax back towards its equilibrium position: $\text{Force}_{after} = \text{Force}_{before}$ - $\delta l_{cantilever} \sim 0$. The rapid relaxation of force reduces the probability that any other part of the protein unfolds in the same kinetic event. Consequently, unfolding occurs in multiple distinguishable steps and the FEC has multiple corresponding peaks.

The behavior observed with MT produces a different picture. The withdrawal of the magnets increases the pulling force on the protein until a portion of the protein unfolds. However, unlike the simulations using a stiff cantilever, the MT measurement maintains the force for the remaining duration of the trajectory because the magnetic field varies on micron scale. Even after the bead has relaxed 10's of nanometers as the unfolded protein segments extend, the bead still resides in nearly the same magnetic field as before, and hence at the same force level. Effectively, the MT mode equates to operating with a very weak spring constant so that $\text{Force}_{after} = \text{Force}_{before}$ - $\delta l_{cantilever} \sim \text{Force}_{before}$. The next effect is the force level present at the beginning of the unfolding event remains nearly constant for the remainder of the measurement (see Fig. 3a in ref. [130]). At this elevated force, the protein

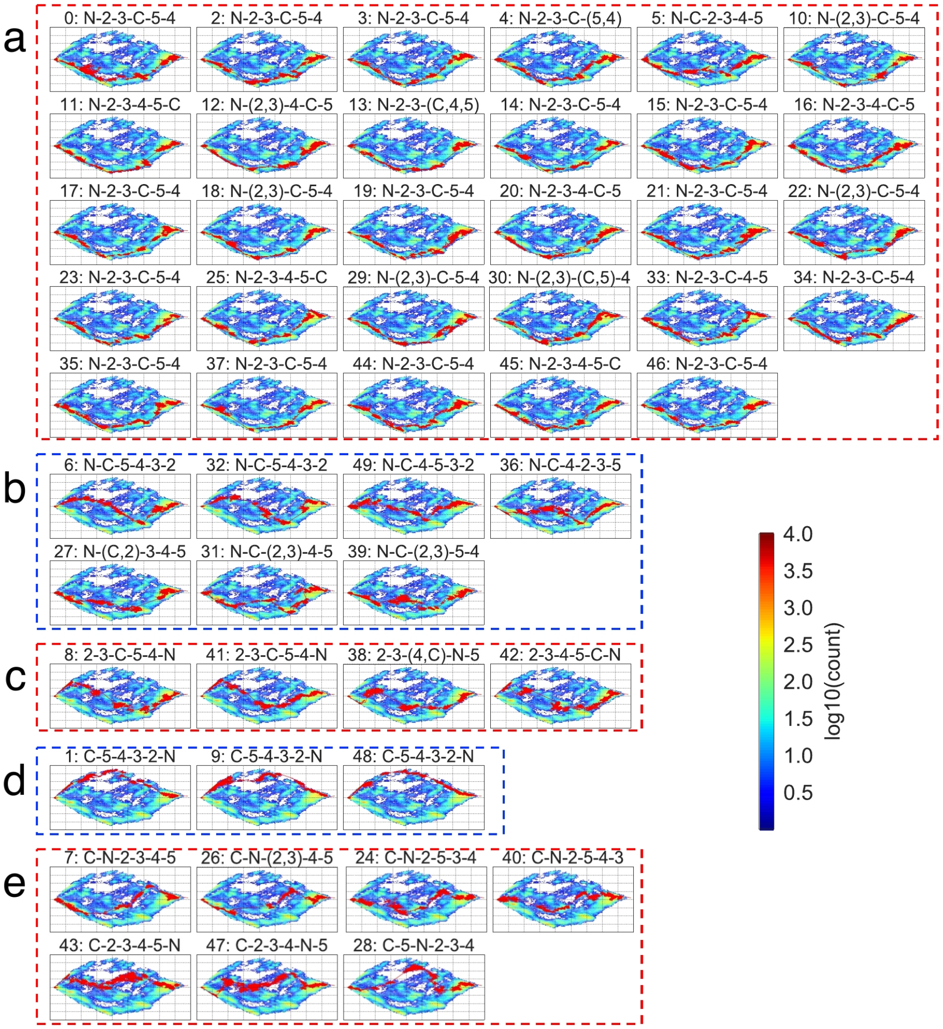

Figure 3.12: **Unfolding trajectories of GlpG obtained with a stiff cantilever** (T = 0.9, $\kappa = 0.05$ kT/Å, pulling velocity = 0.001 Å/*Upside* time step; pulling the C-terminus and fixing the N-terminus with an equal spring). The heat map is obtained from 50 trajectories. The red curve of each subplot is the unfolding pathway from the native state to the fully extended (FE) state for a given trajectory. The trajectories are categorized based on their unfolding pathways. The title of each subplot indicates the index of the trajectory and its unfolding pathway. For example, 4: N-2-3-C-(5,4) means that the unfolding pathway of the fourth trajectory is N→2→C→(5, 4), in which TM5 and TM4 unfold nearly simultaneously and therefore are put in parentheses. **a**. Unfolding starts from TM1 (N) and proceeds to the C-domain when all the TM helices in the N-terminal domain unfold. The pathways traverse the lower edge of the PCA plots. **b**. Unfolding starts from TM1, followed by the unfolding of TM6 (C), leading to zigzag pathways across the PCA plots. **c**. Unfolding starts from the middle of GlpG. **d**. Unfolding starts from TM6 and proceeds to the N-domain when all the TM helices in the C-domain unfold. In contrast to **a** the pathways flank the upper edge of the PCA plots. **e**. Similar to **b**, with unfolding starting from TM6, followed by the unfolding of a TM helix in the N-domain, which results in zigzag pattern through the middle on the PCA plot. The ratio of unfolding from the N-domain first to unfolding from the C-domain first is 40:10 (number of panels in **a**+**b**+**c** : number of panels in **d**+**e**).

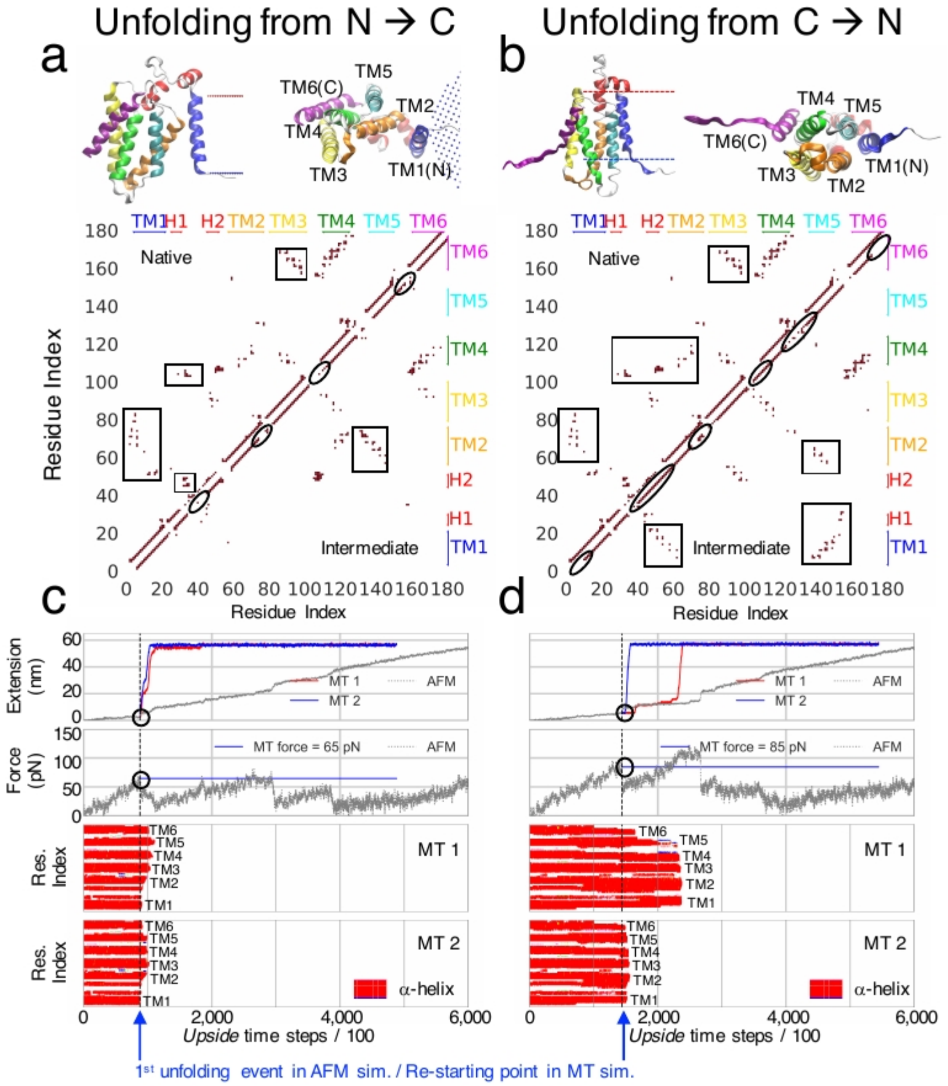

Figure 3.13: **GlpG simulations mimicking a magnetic tweezer measurement. a, b**. Structures and contact maps of the major unfolding intermediate on the N→C and the C→N pathways. The intermediates selected from the N→C and the C→N pathways correspond to the near-native (NN) intermediate in **Fig. 3.10c** and the C2 intermediate in **Fig. 3.11c**, respectively. Differences in contacts between the native state and the intermediate used as the re-starting point are marked by black rectangles and ellipses (along the main diagonal). The length of the lines near the main diagonal identifies the length of the folded portion of the helices (e.g., the TM6 helix in the N→C pathway is present in the first intermediate but is partially unfolded in its counterpart along the C→N pathway). **c, d**. Extension and force profiles, along with TSS plots for two of the 20 N→C generated pathways (denoted MT1 and MT2) and for two of the 20 C→N pathways. After an initial force loading period, the protein begins to unfold (blue arrows), and the force is held constant for the rest of the trajectory. For comparison, the extension and force profiles (grey) for the N→C and C→N trajectories are shown for an AFM- style simulation conducted using a stiff cantilever where the force is allowed to relax after an unfolding event (**Fig. 3.9**).

often is pulled apart in a few or even a single all-or-none process. Hence, at most, only a few intermediates are observed, and folding appears more cooperative in the MT measurements than in our simulations that model an experiment with a stiff cantilever. This effect is most pronounced for "brittle" proteins where the first unfolding event requires higher force than the subsequent unfolding events.

Appreciating this difference between the two modes, we mimic the MT experiment by employing a modified force clamp protocol (**Fig. 3.1c**). Force is gradually increased until the first unfolding event occurs, whereupon the force is held constant for the remainder of the trajectory (**Fig. 3.13**). The N→C and C→N pathways are investigated in detail by re-starting 20 simulations from the structure present right at the first rupture point on each route, which occurs at a force of 65 or 85 pN, respectively. As anticipated, the unfolding of GlpG appears more cooperative along both unfolding pathways than if the cantilever was allowed to relax. All the helices unfold almost immediately and together after the first unfolding event (**Fig. 3.13c, d**, all red bars disappear at the same time, unlike the behavior in **Fig. 3.9e**). Additionally, fewer intermediates are seen and they are more transient especially along the N→C unfolding pathway. For 20 trajectories conducted on each of the N→C and C→N pathways, we observe one major intermediate (at $\sim$ 20.5 nm) and two major intermediates (at extensions of $\sim$ 5.5 nm and $\sim$ 12.5 nm), respectively (**Figs. 3.13c, d, 3.14b**). This difference between the two styles of experiments are readily apparent in the PCA heat maps where the MT-style (weak spring) measurement yields only one well populated intermediate on either the N→C or C→N pathway, and little population appears elsewhere on the PCA heat map (**Fig. 3.17c**). In contrast, the AFM-style (stiff spring) measurement populates dozens of intermediates across the entire map.

The behaviors found for trajectories along the two pathways differ in a manner consistent with experiments. Intermediates formed on the C→N pathway live longer than those on the N→C route (**Fig. 3.14**). The protein unfolds in less than 20k time steps for 19 of 20 simulations along the N→C unfolding pathway, whereas 7 of 20 simulations along the C→N unfolding pathway, the protein does not unfold within 50k steps despite the elevated applied force (**Fig. 3.14a**). Consequently, the observation of intermediates on the C→N pathway is more probable despite the fact that the dominant pathway (higher flux) passes though the N→C route.

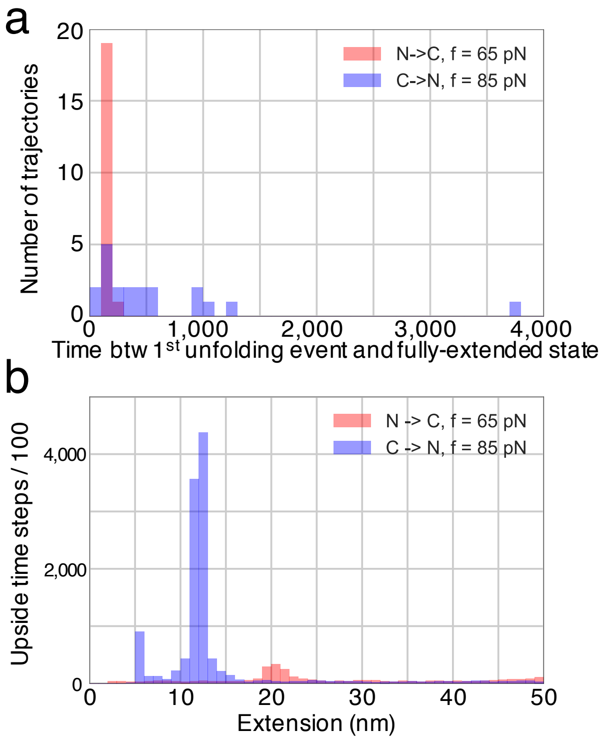

Figure 3.14: **Lifetime and position of intermediates in MT mode for 40 simulations**. **a**. Number of trajectories having different lifetimes between the first unfolding event and the fully-extended state for the 20 unfolding trajectories occurring from either end. **b**. Corresponding aggregate time spent at a given extension (intermediate state). Each peak represents an intermediate (one for N→C, two for C→N). The fully unfolded state has an extension above 50 nm, which defines the upper limit of the x-axis.

Another consideration in the MT experiment [130] that could result in the unfolding of GlpG being cooperative with a preference for unfolding beginning from the C-terminus is the use of a 60 Hz CCD camera [130]. At this relatively slow frame rate, intermediates populated for less than 16 ms could elude detection. Since the intermediates are longer lived when unfolding begins from the C-terminus, they are more likely to be detected than those on the N→C pathway. Thus, the use of a slow camera introduces a bias to observe intermediates on the C→N route and increases the apparent degree of folding cooperativity.

3.6.3 Other SMFS modes applied to GlpG

To further explore SMFS modes of unfolding, We performed standard force clamp simulations where the force is rapidly set and held at a constant value throughout the unfolding trajectory. The values generally are set at a force substantially less than the level where the first unfolding event occurs when operating under pulling mode with increasing force. We find that multi-step sequential unfolding from both N→C and C→N are more likely to be seen under lower force (e.g., 40 versus 60 pN), though the protein tends to unfold more cooperatively and more quickly at either force compared to the pulling with the stiff cantilever (**Figs. 3.17d, 3.15**).

We also pulled in the same manner as in the AFM simulations on bR. Force was applied with a stiff cantilever orthogonal to the bilayer to either the N- or the C-terminus. As with bR, we observed the characteristic sawtooth patterns (**Fig. 3.16**). However, as previously noted [157], the TM helices of GlpG are not as hydrophobic as bRs, so GlpGs helices were pulled out at lower force ($\sim$ 60 pN for the first pair of helices versus $\sim$ 94 pN for bR). The last TM helix was observed to be pulled out in a distinct event in 3 of 20 and 4 of 20 GlpG trajectories when pulling on the N- and C-terminus, respectively. In the rest of the trajectories, more than one TM helix is pulled out in the same unfolding event. Moreover, we observe that the TM6 helix unfolds before the TM5 helix when pulling on the N-terminus (**Fig. 3.16b**), a behavior consistent with the TM6 helix being intrinsically unstable within the bilayer. This inversion in the order of unfolding would be hard to infer from the FEC as it violates the assumption that the helices unfold according to their sequence order. The difference between this SMFS mode and the others is well illustrated in the PCA plot, which displays a series of intermediates going along the edges of the heat map (**Fig. 3.17b**). Since no corresponding experiment has yet been conducted with GlpG, these simulations provide testable predictions. Studies of other SMFS modes and the effects of mutations and temperature can be found in the supplement.

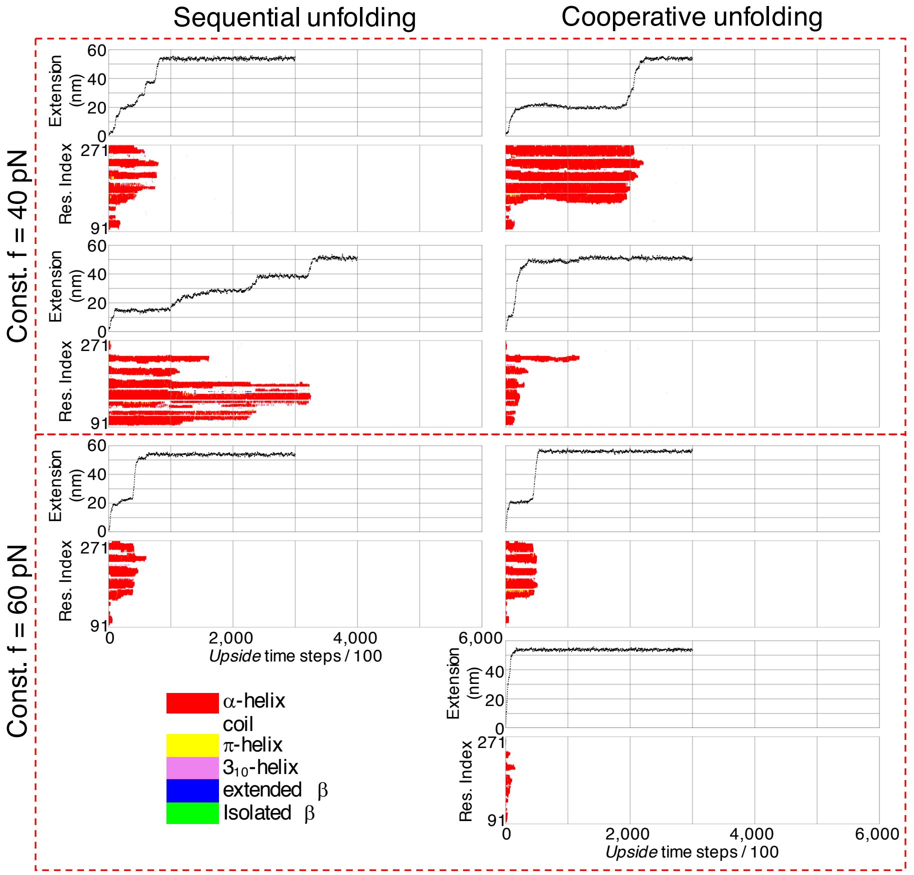

Figure 3.15: **Sequential and cooperative unfolding pathways of GlpG in force clamp simulations started from the native structures**. Four trajectories under a constant force 40 pN (upper panel) and three trajectories under a constant force 60 pN (lower panel) are shown. Every trajectory is presented by an extension plot and a TSS plot. Trajectories in the left column are examples of sequential unfolding pathways (with at least 3 intermediates that can be identified on the extension plot); those in the right column are examples of cooperative unfolding pathways (with no more than 2 intermediates identified from the extension plot). In the 20 simulations under 60 pN, we did not observe any trajectory unfolding from C- to N-domain sequentially.

3.6.4 Altering the pathway fluxes using mutation, temperature and spring constant

We also performed unfolding simulations on destabilizing GlpG mutants to examine the effects the unfolding pathways. Of the investigated residues having an H-bonding side chain in the N-domain, the E166A mutation is the most destabilizing [34]. This residue is located near the bottom of the TM2 helix and forms two H-bonds to the backbone nitrogens of Val96

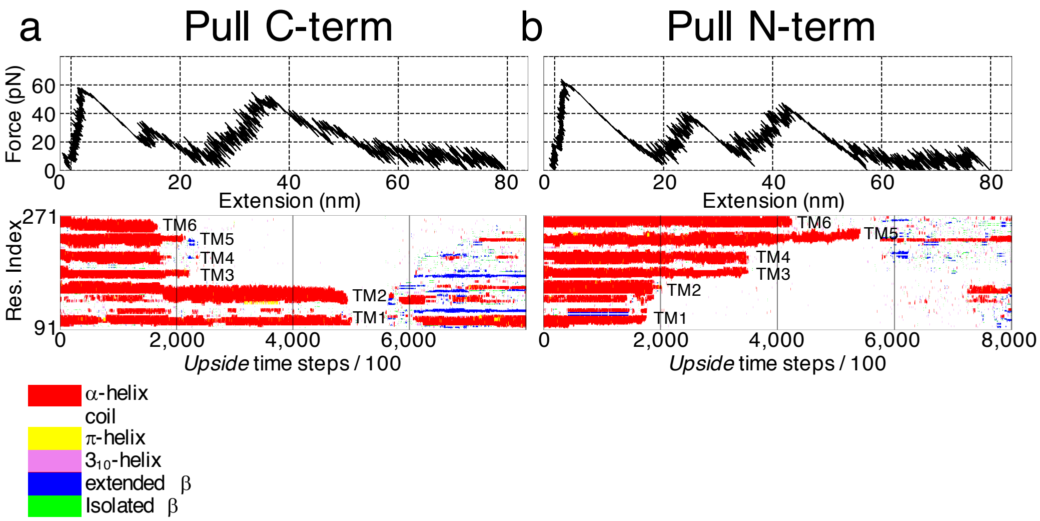

Figure 3.16: **Unfolding GlpG by pulling vertically in AFM, stiff cantilever mode. a, b**. FEC and TSS plots of an example trajectory pulling from the C- and N-terminus, respectively. 20 simulations were performed in each case. We observed that all TM helices become completely unfolded in 4 and 3 trajectories of pulling C- and N-terminus, respectively. After all the protein is pulled out, the extended chain starts collapsing and forms H-bonds again. This suggests that the pulling rate maybe slow enough. Notably, TM6 unfolds before TM5 **(b)**, implying that TM6 is not very stable by itself in the lipid bilayer.

and Thr97 on the TM1 helix and two to the side chains of Thr97 on the TM1 and Ser171 on the TM3 helix [34]. The G261V mutation on helix TM6 is at the center of the GxxxGxxxA motif that enables the close backbone-backbone association of the TM4 and TM6 helices. This mutation decreases the T_m by 28.1±0.08 $^\circ C$ [34] and increases the probability of unfolding from the C-domain by 50% (10 to 15 events, of a total of 50) (**Fig. 3.17 and table 3.6**). To our surprise, the disruption of the H-bond network at the bottom of the triad of the three TM helices in the N-domain barely changes the probability of initiating the unfolding from this end (40 versus 41 events of a total of 50) (**Fig. 3.17 and table 3.6)**.

Unfolding from the N- rather than the C-terminus is 4-fold more probable at 270 K. At 300 K, the ratio is reduced to 1.2. And weakening the spring constant by a factor of 5 further reduces the ratio to 0.7 (**Fig. 3.12 and table 3.6**). Even though there are differences, the fundamental heterogeneous pathway behavior remains (**Fig. 3.17a**).

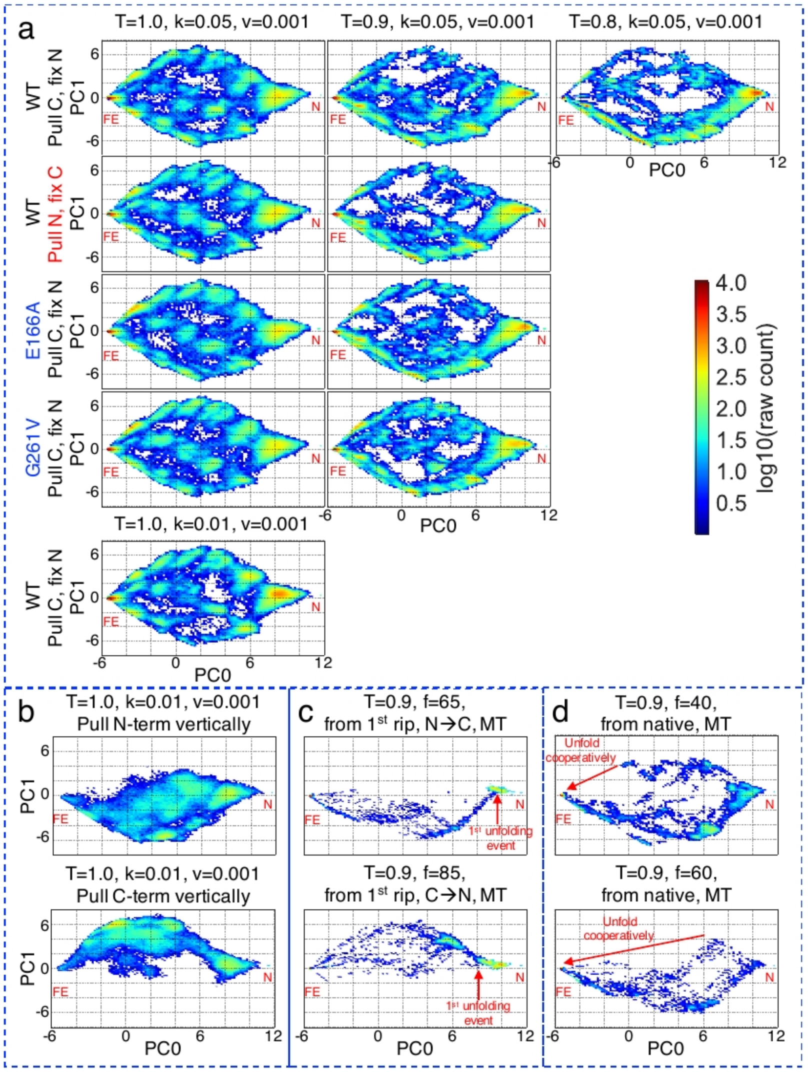

Figure 3.17: **Principal component analysis of unfolding trajectories of GlpG under various simulation protocols**. **a. Stiff cantilever mode, pulling laterally**. Each of the PCA plots is comprised of 50 trajectories. Despite differences, the fundamental heterogeneous pathway behavior remains. **b. Stiff cantilever mode, force is applied to either the N- or C-terminus vertically**. The PCA plots for pulling the N- and C-terminus contains 20 and 19 trajectories, respectively (the output file of one of the trajectories was corrupted). Notably, the PCA heat maps obtained in this mode fill in the blanks in the middle of the heat maps obtained in **a**. Those may represent structures that largely maintain the tertiary structure for the region embedded in the membrane, which would be difficult to observe in mode **a** because the tertiary structure is disrupted. Besides, pulling the N-terminus produces deterministically N→C pathways as expected, and vice versa. **c. Modified MT mode, pulling laterally, simulations were re-started at the 1st unfolding event in the N→C or C→N pathway in a (T=0.9, k=0.05, v=0.001). d. MT mode, pulling laterally, simulations were started from the native structure**. The unfolding is more cooperative under higher force and in the C→N pathway than the reverse. 20 trajectories are included in each PCA plot in **c** and **d**. N and FE stand for native and fully-extended in each subplot, respectively.

Table 3.6: Summary of unfolding pathways of GlpG.

Pulling scheme	T	κ of F [a]	WT/mutant	N-→C-domain[b]	C-→N-domain[c]
Pull C-term, fix N-term	1	$\kappa =$ 0.05	WT	27	23
Pull C-term, fix N-term	1	$\kappa =$ 0.05	E166A	27	23
Pull C-term, fix N-term	1	$\kappa =$ 0.05	G261V	22	28
Pull C-term, fix N-term	1	$\kappa =$ 0.01	WT	21	29
[d]Pull C-term, fix N-term	0.9	$\kappa =$ 0.05	WT	40	10
Pull C-term, fix N-term	0.9	$\kappa =$ 0.05	E166A	41	9
Pull C-term, fix N-term	0.9	$\kappa =$ 0.05	G261V	35	15
Pull C-term, fix N-term	0.8	$\kappa =$ 0.05	WT	42	8
Pull N-term, fix C-term	1.0	$\kappa =$ 0.05	WT	32	18
Pull N-term, fix C-term	0.9	$\kappa =$ 0.05	WT	43	7
Pull C-term, fix N-term	0.9	F = 40 pN	WT	16	4
Pull C-term, fix N-term	0.9	F = 60 pN	WT	13	7
Pull C-term, fix N-term	0.9	F = 80 pN	WT	13	7

Units for κ and v are the same as in **Table 3.5**.

a. κ or F is listed as relevant to the mode of applying force, gradual pulling or force, respectively.

b. The number of trajectories with an unfolding pathway that is initiated from the N-domain. For example, in **Fig. 3.12**, **a, b and c** are all counted as N- to C-domain unfolding pathways.

c. The number of trajectories with an unfolding pathway that is initiated from the C-domain. For example, in **Fig. 3.12**, **d and e** are both counted as C- to N-domain unfolding pathways.

d. The primary data set that is shown in the main text.

e. From the native structure refers to starting the simulation from the native structure instead of an intermediate at the first unfolding event.

3.6.5 Methods

Structure and sequence of GlpG and the mutations

The native structure and orientation within the lipid bilayer of GlpG (2xov.pdb) was taken from the OPM database [76]. The membrane thickness was set to 28.8 nm as predicted by OPM. Two GlpG mutants were made from the native structure using *SWISS-pdbviewer* [158]: E166A and G261V.

Principal component analysis of unfolding trajectories of GlpG

The programs *MDTraj* [82] and *scikit-learn* [88, 159] was used to perform the PCA using the Cα-Cα distances below 8 Å in the native state to define contacts. Structures from all trajectories under the same set of simulation conditions were included in the PCA. To derive the principal components, we used the Cα-Cα distances obtained at T = 1.0, spring constant = 0.05 kT/Å^2, and a pulling velocity = 0.001 Å/*Upside* time step. These principal components are used for the projection at the other conditions for comparison purposes.

Clustering analysis of unfolding trajectories of GlpG

We chose the Gaussian mixture algorithm implemented in *scikit-learn* [88, 159] to perform the clustering analysis performed on the structures in each trajectory. A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions, which is good for density estimation. The number of clusters was estimated and supplied to the program. The corresponding density for each training point was measured and the point with the maximal density was chosen as the center to represent the cluster. The cluster centers were considered as the intermediate states along the unfolding pathway.

3.7 Summary of simulation details in this chapter

Table 3.7: Simulation details.

System	Cantilever[a]	Attachments	κ	v	F	T	Number of sim.
Ubiquitin (FE)[b]	Soft	Pull both termini in opposite direction			0, 10, 30, 50, 100, 250, 800	1.0	1 per F
Truncated bR species	Stiff	Pull C-term vertically[c], fix N-term[d]	0.05	0.001		1.0	1 per species
bR	Stiff	Pull C-term vertically	0.05	0.001		1.0	90
bR	Stiff	Pull C-term vertically	0.01	0.001		1.0	80
bR	Stiff	Pull C-term vertically	0.05	0.0005		1.0	60
GlpG	Stiff	Pull C-term laterally[c], fix N-term	0.05	0.001		0.8, 0.9, 1.0	50 per T
GlpG	Stiff	Pull N-term laterally, fix C-term	0.05	0.001		0.9, 1.0	50 per T
GlpG E166A	Stiff	Pull C-term laterally, fix N-term	0.05	0.001		0.9, 1.0	50 per T
GlpG G261V	Stiff	Pull C-term laterally, fix N-term	0.05	0.001		0.9, 1.0	50 per T
GlpG	Stiff	Pull C-term laterally, fix N-term	0.01	0.001		1.0	50
GlpG before 1st rip	Soft	Pull C-term laterally, fix N-term			64.6, 84.5	0.9	20 per F[e]
GlpG	Soft	Pull C-term laterally, fix N-term			40, 60, 80	0.9	20 per F
GlpG	Stiff	Pull C-term vertically	0.05	0.001		1.0	20
GlpG	Stiff	Pull N-term vertically	0.05	0.001		1.0	20

Parameters used in the *Upside* simulation are summarized in the table. Units for κ, v, F, and T are the same as in **Table 3.6**.

a. Soft mode refers to the use of a very soft cantilever to mimic a magnetic tweezers measurement where the force is held essentially constant force after the 1st unfolding event occurs as the magnetic field varies slowly, on micron length scale, longer than the unfolded segments.

b. The simulations were started from a fully-extended state.

c. The direction is relative to the membrane bilayer.

d. Held with an equally stiff spring.

e. One of the output file is corrupted, so there are only 19 trajectories useful for analysis.

3.8 Discussion

We have modified our fast *Upside* MD algorithm to study the forced unfolding of proteins in a variety of different modes and on membrane proteins by including our new statistical potential [48]. *Upside* models the polypeptide backbone with 5 atoms with residue-dependent Ramachandran maps, while the multi-position side chain beads are repacked at every MD step. The great enhancement in speed afforded by this method allows us to run 100's of simulations and conduct principal component analysis to map out the energy surface. Unfolding of membrane proteins as large as bR are conducted in a single cpu-day with results that are in good agreement with experimental studies (**Figs. 3.5, 3.7**) [43].

The mode of force application including the strength of the spring constant strongly affects the observed detail and apparent unfolding cooperativity. Using a force clamp, irrespective of whether the force is gradually increased or rapidly set to a constant value, the protein tends to unfold more cooperatively (**Fig. 3.17c, d**). The force in the constant velocity mode is maintained to some degree after an unfolding event if the associated chain extension is less than that needed to bring the system back to its equilibrium position, $\delta l \cdot \kappa \ll Force_{before}$. In addition, the applied force can remain high after a rupture if the length of the newly unfolded segment is sufficiently long that the chain (and whatever handles are used to connect the protein to the instrument) can act as entropic springs thereby reducing the effective spring constant of the entire system. The effect becomes more pronounced in the later stages of unfolding, as the length of the unfolded regions becomes increasingly longer with each unfolding event, thereby reducing the probability of observing late intermediates.

Membrane proteins can be unfolded by either pulling vertically at one terminus or laterally at both ends. Each mode explores different regions of the energy surface. Pulling vertically produces a more deterministic unfolding route as TM helices unfold sequentially and in pairs, while the remaining portion of the protein largely remains intact. Pulling laterally tends to break lateral interhelical interactions and can lead to gross structural rearrangements even before any TM helix unfolds.

Few SMFS simulations have been conducted for membrane proteins, partly due to the computation resources required. One all-atom MD study [136] identifies a number of key residues that resist mechanical unfolding in the intermediate states probed by the experiment [43], although the pulling speed (1-50 m/s) precluded observing WLC behavior for the unfolded segments. A 2016 CG study uses the same pulling rate as ours ($\sim 10^6$ nm/s) and finds FECs with WLC behavior [47]. Although many features are similar between this and our studies, we observed more intermediates (**Table 3.5**).

We concur with a recent study of GlpG force-induced unfolding [140] whose major conclusion was that the two-stage model [29] is overly simplistic as isolated helices can co-exist with a multi-helix folded region, and all the helices do not have to be in the bilayer prior to the initiation of folding. Nevertheless, some technical differences are worth noting:

(**i**) We explore different parts of the landscape as we explicitly simulate the experimental pulling protocol by applying force with a moving springs, whereas the other study conducts umbrella sampling with an energetic bias determined by the N-to-C separation distance.

(**ii**) We do not restrain the secondary structure and allow the native helices to unfold.

(**iii**) We observe TM helices go surface-bound and quickly break and become extended.

(**iv**) We correct for loss of protein-lipid interactions as helices come together in the bilayer.

(**v**) we do not employ a Go model (we only study unfolding) nor stabilize the N-terminus to promote the C→N pathway. Consequently, we observe unfolding beginning from other regions than just the C-terminus.

(**vi**) We use principal components as the reaction coordinates to depict the pathways instead of the end-to-end distance and the average z-value of all Cα atoms to manifest the collective conformational change.

Interestingly, the groups earlier study found that folding could occur in either the N→C or C→N pathways [139] as we observe.

While our method has widespread applications, several issues exist. First, we cannot refold the membrane proteins by relaxing the force, as found experimentally [130, 131, 134]. Improvements in our energy function are in progress to address this issue. Also, we assume an infinite flat membrane bilayer, which is valid for bR, but experimentally, GlpG [130] is embedded in bicelles, which may not be large enough to accommodate all the states we generate in our simulations.

3.9 Conclusion

We have developed an accurate and fast near-atomic level method to conduct 100s of unfolding simulations to characterize the energy surface for force-induced unfolding. The method reproduces many of the experimental features of SMFS studies for the unfolding of bR [43] and GlpG [130]. The simulations can assist experimental studies by helping convert force-extension curves to structures, pathways, and energies, which can be challenging. For example, we identified the more stable amino-terminal domain of GlpG as the more likely terminus to unfold, but it escaped detection due to the all-or-none unfolding behavior along this route. The counterintuitive unfolding of this more stable end [137] arises in part from higher hydrophobicity of the amino terminal TM helices, which highlights a general folding property for membrane helices: While increased hydrophobicity promotes insertion into the bilayer, it also enhances dissociation as the lipid bilayer is a good solvent for isolated hydrophobic helices. In contrast TM helices that bury polar or charged groups upon association tend to remain associated as the cost of exposing their non-hydrophobic moieties to the bilayer is high.

We find that method of applying the force can significantly alter the region of the energy surface that is probed. Notably, the application of constant force reduces the probability of observing intermediates and increasing the apparent unfolding cooperativity, as compared to the use of stiffer cantilevers that can relax to lower force after an unfolding event which increases the probability of observing intermediates. The application of constant force can be intentional, but also can be a consequence of force effectively being applied through a weak spring constant, such as that inherent in magnetic tweezers measurements where the magnetic field varies slowly with distance, or after a substantial portion of the protein is unfolded and as it acts a weak entropic spring. Our method can be employed to simulate complicated *gedanken* pulling experiments beyond current experimental capabilities, such as pulling on multiple sites in multiple directions with different strength of the springs, and with either membrane or soluble proteins.

3.10 Aknowledgement

We thank members in our group for helpful discussions. Work in this chapter is supported by NIH/NIGMS Grants GM055694 (to TR Sosnick and KF Freed) and GM087519 (to E Perozo). Computations are produced using the Midway resources of the Research Computing Center (RCC) at the University of Chicago.

CHAPTER 4

FUTURE PERSPECTIVES

The previous chapters described a tool for simulating SMFS of transmembrane proteins. With coarse-grained model *Upside* and a proper energy function for transmembrane proteins *UChiMemPot* [48], I showed that it is possible to obtain useful information about unfolding pathways of transmembrane proteins [124], often consistent with various experimental and computational unfolding studies. Additionally, in some cases, unfolding simulations can help explicate and interpret results from such studies.

Notwithstanding the successful demonstrations of acquiring results with an accuracy comparable to experiments, the work described here still remains to be implemented on larger and more complex systems (eg. multiple domain transmembrane proteins, β-barrels). Although the topics described henceforth in this chapter is under development and unpublished, I will present some ideas to extend my work described in this thesis. First, I will present some ideas for improving the membrane burial potential for broader application. Next, I will show opportunities for applying our tool to more systems and obtaining more quantitative data regarding the unfolding kinetics. Last, I will discuss directions for simulating transmembrane proteins in general.

4.1 Improve the membrane burial potential

4.1.1 Optimize for folding transmembrane proteins

Because the membrane burial potential is only applied to residues exposed to the surrounding lipid molecules, to which extent a residue is in contact with the lipids is critical in balancing residue-residue and residue-lipid interactions. In **Chapter 2**, the midpoints for exposed and buried were chosen empirically (see **Fig. 2.4** and **Table 2.2**) to position transmembrane proteins in lipid bilayer while accounting for the paucity of charged residues in the middle of bilayers. Briefly, a residue is considered exposed to the surround solvent when the residue burial ≤ the midpoint; otherwise, it is buried in the protein core and thus is excluded from the potential application. Because it needs to assure the continuity of all potential functions in *Upside*, a modified sigmoid function is used as the coefficient of the membrane burial potential (**Fig. 4.1**).

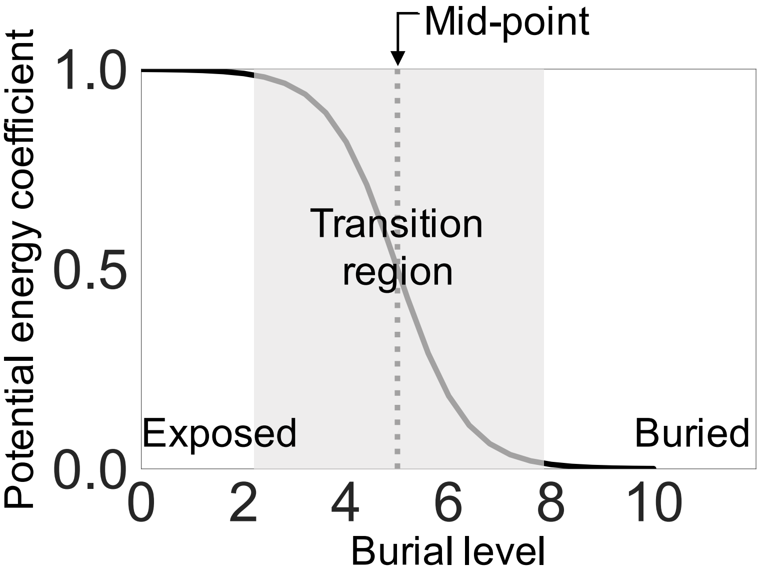

Figure 4.1: **Coefficient of membrane burial potential**. The sigmoid function enables continuous smooth transition from the fully exposed state to the fully buried state.

In this case, the potential profile as well as the choice of the midpoint may not be suitable for folding transmembrane proteins. Residue-lipid interaction could be traded off for residue-residue interaction as two or more helices come together (see **Fig. 2.2E**). In an extreme case in which all residues are considered exposed (the midpoint is very large), helices will naturally tend to align in parallel in order to maximize the residue-residue interactions (**Fig. 4.2**). The GxxxG motif [36] (in cross shape) would therefore be unattainable.

The contrast divergence technique from machine learning [78] can be used to optimize the membrane burial potential for folding transmembrane proteins. Because lateral association of proteins with a single-span TM helix can be treated as a constrained docking of rigid bodies whose backbone geometry requires mere minor adjustments [160], the dimerization of those proteins provide training cases. However, there is a caveat that the training set should

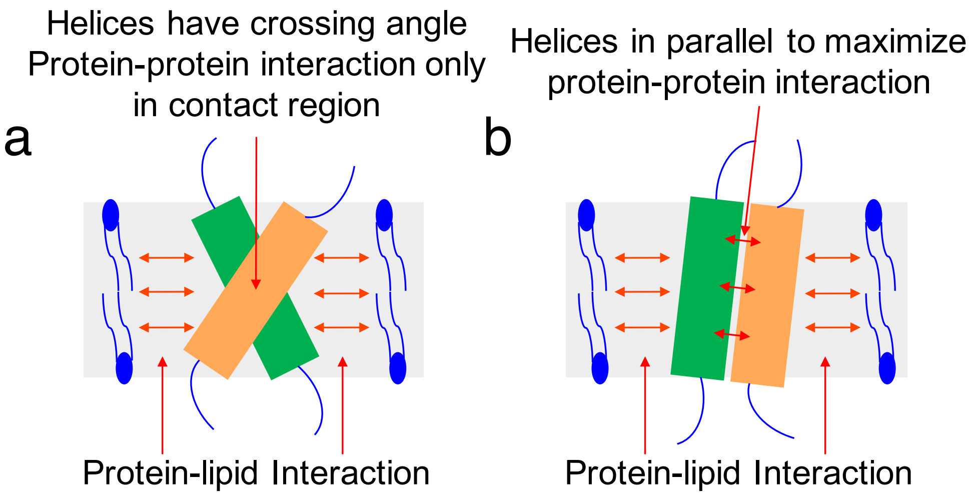
a
Helices have crossing angle
Protein-protein interaction only
in contact region
Protein-lipid Interaction
b
Helices in parallel to maximize
protein-protein interaction
Protein-lipid Interaction

Figure 4.2: **Balance between residue-lipid and residue-residue interactions**. If losing the residue-lipid interaction and gaining the residue-residue interaction is more favorable, the helices will tend to align in parallel to maximize the interhelical contact and the residue-residue interactions. In this case, the structure in scissors shape is not attainable.

include higher ordered oligomeric TM structures, such as bR and GlpG. Becasue residues in dimeric single-span TM helices are mostly fully-exposed even in the interhelical interface, the midpoint for recognizing exposed residues may be over-estimated. In addition, as structural rearrangements are of functional importance for TM proteins with more helices, it is not certain that those fold into a single (low-energy) structure (i.e., different conformations for one protein exist) [161].

4.1.2 Derive new potential for β-barrels

Previous study has shown that amino acid distributions of transmembrane α-helical proteins are highly correlated with those of the lipid-facing residues in β-barrels [69]. This correlation suggests that our potential with *Upside* could be suitable for β-barrels but further testing is required. Nonetheless, new potential profiles specific for β-barrels would be benificial, after all the lipid composition of the membranes in which β-barrels are embedded are different from those of transmembrane α-helical proteins [69, 162].

It may be difficult for our model to simulate the folding or unfolding of β-barrels because when the barrel opens it will be hard to differentiate the inward-facing residues from the outward-facing residues and thus to apply the membrane burial potential. However, it is possible for us to simulate (i) the dimerization or oligomerization of β-barrels to study the association and interaction of monomeric barrels, and (ii) the folding, unfolding, and refolding of the plug domain inside β-barrels under force, for example the mechanical unfolding of the first 161 residues inside FhuA by AFM [163]. In both cases, structures of the β-barrels need to be assumed to remain unchanged in the simulation. In other words, the barrels are rigid.

4.2 Apply forced unfolding to more systems

4.2.1 Unfold SOD, a small soluble protein

Our tool for simulating forced unfolding can be easily applied to soluble proteins. The membrane burial potential is not needed and the protein is usually smaller than transmembrane proteins, so more realistic pulling velocity is allowed. Besides, since *Upside* is capable of rapid *de novo* folding of soluble proteins shorter than 100 residues, we can refold the protein.

For instance, how Cu/Zu-superoxide dismutase 1 (SOD1) folds is of particular interest because its prion-like misfolding is linked to the disease ALS [127], which is a progressive neurodegenerative disease. SMFS (optical tweezers) has been employed to study the unfolding and refolding of the most misfolding-prone form of SOD1 [127] to illuminate the mechanism of misfolding. Our preliminary results show that unfolding SOD1 at lower temperature produces distribution of cumulative contour lengths similar to that obtained in experiment [127].

4.2.2 Unfold bR from the trimeric state

Preliminary results of the forced unfolding of bR from the trimeric state have shown that more and longer lived intermediates exist than in the unfolding from the monomeric state. It would be interesting to investigate the difference that roots in the interaction among the monomers of bR, which may shed light on the trimerization of bR molecules. Because we are able to perform thought experiment that are not yet possible practically, we can remove one or two monomers from the trimer or mutate certain residues on the protein during the simulation to detect key interactions responsible for that difference.

4.2.3 Unfold ClC transporter, a two-domain transmembrane protein

The ClC family includes a large number of passive channels and active transporters, vital for cellular functions such as the maintenance of membrane potential and volume homeostasis [164]. In particular, the *E. coli* ClC antiporter, ClC-ec1, is a dimer with a single transport pathway per subunit [164]. Compared to bR and GlpG, ClC-ec1 is structurally more complicated as each ClC-ec1 subunit is internally pseudo-symmetric, dividing the protein into two domains with inverted topology [164].

An unfolding study of ClC-ec1 by magnetic tweezers [164] with a protocol similar to unfolding GlpG [130] shows that the protein can be separated into two stable halves that unfold independently, in line with an evolutionary model in which the two halves arose from independent folding subunits that fused together later. We expect to observe more intermediates during the unfolding process with the AFM-style unfolding and hence to characterize the unfolding under close scrutiny.

4.2.4 Construct the energy surface revealed by unfolding simulations

Equilibrium free energy profile can be derived rigorously [45] from repeated non-equilibrium force measurements on the basis of an extension of Jarzynski's identity [165] between free energies and the irreversible work for titin I27 [166] and bR [167, 168]. Particularly, it would be interesting to construct the energy surface of GlpG revealed by our unfolding simulations, which will assist understanding of the folding energetics. Standard protocol of constructing the energy surface is expected as part of the workflow in conducting forced unfolding simulations.

4.3 Prediction for conformational chage of transmembrane proteins

Prediction of pathway between stable states can be performed using a variety of techniques, such as simple morphing methods based on interpolations in Cartesian [169] or internal coordinates [170, 171], algorithms built on an elastic network model [172] (eg. eBDIMS [173], ANMPathway [174]), and MD-based approaches (eg. string method with swarm-of-trajectories [175, 176]). The efficiency and accuracy of these methods are typically strongly tied to whether it can propose a reasonable initial trajectory for atomic MD. We expect that *Upside* with the membrane burial potential will be useful for studying conformational changes of transmembrane proteins, in elucidating conformational transition pathways difficult or expensive to sample with all-atom MD, exploring transient intermediate states which are hard to obtain in experiment, and providing physical meaningful trajectories that help generate experimentally testable hypotheses.

REFERENCES

[1] T. Nugent and D. T. Jones. Membrane protein structural bioinformatics. *J Struct Biol*, 179(3):327–37, 2012.

[2] John P. Overington, Bissan Al-Lazikani, and Andrew L. Hopkins. How many drug targets are there? *Nat Rev Drug Discovery*, 5:993–6, 2006.

[3] J. Koehler Leman, M. B. Ulmschneider, and J. J. Gray. Computational modeling of membrane proteins. *Proteins*, 83(1):1–24, 2015.

[4] A. L. Lomize and I. D. Pogozheva. Statistics from OPM as of Jan. 2017: 2505 integral membrane proteins out of 2914 membrane protein entries, Jan. 2017. URL `http://opm.phar.umich.edu/about.php`.

[5] G. E. Tusnády and D. Kozma. Statistics from PDBTM as of Jan. 2017: 2638 integral membrane proteins out of 3006 membrane protein entries, Jan. 2017. URL `http://pdbtm.enzim.hu/`.

[6] E. Wallin and G. von Heijne. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci*, 7(4):1029–38, 1998.

[7] M. S. Almen, K. J. Nordstrom, R. Fredriksson, and H. B. Schiöth. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol*, 7:50, 2009.

[8] Phillip J. Stansfeld. Computational studies of membrane proteins: from sequence to structure to simulation. *Curr Opin Struct Biol*, 45:133–41, 2017.

[9] Stephen H. White. Biophysical dissection of membrane proteins. *Nature*, 459:344–6, 2009.

[10] A. L. Lomize and I. D. Pogozheva. Statistics from OPM as of Dec. 2018: 4179 membrane protein entries, Dec. 2018. URL `http://opm.phar.umich.edu/about.php`.

[11] P. J. Stansfeld, E. P. Carpenter J. L. Parker S. Newstead J. E. Goose, M. Caffrey, and M. S. P. Sansom. Memprotmd: Automated insertion of membrane protein structures into explicit lipid membranes. *Structure*, 23:1350–1361, 2015.

[12] S. A. Shaikh, J. Li, G. Enkavi, P. C. Wen, Z. Huang, and E. Tajkhorshid. Visualizing functional motions of membrane transporters with molecular dynamics simulations. *Biochem*, 52(4):569–87, 2013.

[13] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, and A. Sali. Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci*, Chapter 2:Unit 2 9, 2007.

[14] J. Shi, T. L. Blundell, and K. Mizuguchi. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, 310(1):243–57, 2001.

[15] A. Roy, A. Kucukural, and Y. Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc*, 5(4):725–38, 2010.

[16] V. Yarov-Yarovoy, J. Schonbrun, and D. Baker. Multipass membrane protein structure prediction using rosetta. *Proteins*, 62(4):1010–25, 2006.

[17] P. Barth, J. Schonbrun, and D. Baker. Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc Natl Acad Sci U S A*, 104(40):15682–7, 2007.

[18] B. E. Weiner, N. Woetzel, M. Karakas, N. Alexander, and J. Meiler. BCL::MP-fold: folding membrane proteins through assembly of transmembrane helices. *Structure*, 21 (7):1107–17, 2013.

[19] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7): 1607–21, 2012.

[20] S. Wang, J. Peng, J. Ma, and J. Xu. Protein secondary structure prediction using deep convolutional neural fields. *Sci Rep*, 6:18962, 2016.

[21] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol*, 13(1):e1005324, 2017.

[22] E. Lindahl and M. S. Sansom. Membrane proteins: molecular dynamics simulations. *Curr Opin Struct Biol*, 18(4):425–31, 2008.

[23] F. Khalili-Araghi, J. Gumbart, P. C. Wen, M. Sotomayor, E. Tajkhorshid, and K. Schulten. Molecular dynamics simulations of membrane channels and transporters. *Curr Opin Struct Biol*, 19(2):128–37, 2009.

[24] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S. J. Marrink. The MARTINI coarse-grained force field: Extension to proteins. *J Chem Theory Comput*, 4(5):819–34, 2008.

[25] N. P. Schafer Kim, B. L. and P. G. Wolynes. Predictive energy landscapes for folding alpha-helical transmembrane proteins. *Proc Natl Acad Sci U S A*, 111:11031–6, 2014.

[26] P. J. Bond, J. Holyoake, A. Ivetac, S. Khalid, and M. S. Sansom. Coarse-grained molecular dynamics simulations of membrane proteins and peptides. *J Struct Biol*, 157:593–605, 2007.

[27] A. Panahi and M. Feig. Dynamic heterogeneous dielectric generalized born (dhdgb): An implicit membrane model with a dynamically varying bilayer thickness. *J Chem Theory Comput*, 9:1709–19, 2013.

[28] Sue-Ellen Gerchman Jean-Luc Popot and Donald M.Engelman. Refolding of bacteriorhodopsin in lipid bilayers: A thermodynamically controlled two-stage process. *J Mol Biol*, 198(4):655–76, 1987.

[29] J. L. Popot and D. M. Engelman. Membrane-protein folding and oligomerization - the two-stage model. *Biochem*, 29(17):4031–7, 1990.

[30] Florian Cymer, Gunnar von Heijne, and Stephen H. White. Mechanisms of integral membrane protein insertion and folding. *J Mol Biol*, 427:999–1022, 2015.

[31] S. H. White and W. C. Wimley. Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct*, 28:319–65, 1999.

[32] Russell E. Jacobs and Stephen H. White. The nature of the hydrophobic binding of small peptides at the bilayer interface: implications for the insertion of transbilayer helices. *Biochem*, 28(8):342137, 1989.

[33] D. M. Engelman and T. A. Steitz. The spontaneous insertion of proteins into and across membranes: the helical hairpin hypothesis. *Cell*, 23(2):411–22, 1981.

[34] R. P. Baker and S. Urban. Architectural and thermodynamic principles underlying intramembrane protease function. *Nat Chem Biol*, 8(9):759–68, 2012.

[35] Susan E. Harrington and Nir Ben-Tal. Structural determinants of transmembrane helical proteins. *Structure*, 17:1092103, 2009.

[36] A. Senes, D. E. Engel, and W. F. DeGrado. Folding of helical membrane proteins: the role of polar, GXXXG-like and proline motifs. *Current Opinion in Structural Biology*, 14(4):465–79, 2004.

[37] Alessandro Senes, Iban Ubarretxena-Belandia, and Donald M. Engelman. The CαH$\cdots$O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. *Proc Natl Acad Sci U S A*, 98(16):9056–61, 2001.

[38] Heedeok Hong. Toward understanding driving forces in membrane protein folding. *Arch Biochem Biophys*, 564:297–313, 2014.

[39] Emilia L. Wu, Xi Cheng, Sunhwan Jo, Huan Rui, Kevin C. Song, Eder M. Dávila-Contreras, Yifei Qi, Jumin Lee, Viviana Monje-Galvan, Richard M. Venable, Jeffery B. Klauda, and Wonpil Im. CHARMM-GUI membrane builder toward realistic biological membrane simulations. *J Comput Chem*, 35:1997–2004, 2014.

[40] Sunhwan Jo, Taehoon Kim, and Wonpil Im. Automated builder and database of protein/membrane complexes for molecular dynamics simulations. *PLoS ONE*, 2(9): e880, 2007.

[41] Sunhwan Jo, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J Comput Chem*, 29:1859–65, 2008.

[42] Harald Janovjak Alexej Kedrov, K. Tanuj Sapra, and Daniel J. Müller. Deciphering molecular interactions of native membrane proteins by single-molecule force spectroscopy. *Annu Rev Biophys Biomol Struct*, 36:233–60, 2007.

[43] H. Yu, M. G. Siewny, D. T. Edwards, A. W. Sanders, and T. T. Perkins. Hidden dynamics in the unfolding of individual bacteriorhodopsin proteins. *Science*, 355(6328): 945–50, 2017.

[44] Robert E. Jefferson, Duyoung Min, Karolina Corin, Jing Yang Wang, and James U. Bowie. Applications of single-molecule methods to membrane protein folding studies. *J Mol Biol*, 430:424–37, 2018.

[45] Gerhard Hummer and Attila Szabo. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proc Natl Acad Sci U S A*, 98(7):365861, 2001.

[46] Matthieu Chavent, Anna L Duncan, and Mark SP Sansom. Molecular dynamics simulations of membrane proteins and their interactions: from nanoscale to mesoscale. *Curr Opin Struct Biol*, 40:8–16, 2016.

[47] T. Yamada, T. Yamato, and S. Mitaku. Forced unfolding mechanism of bacteriorhodopsin as revealed by coarse-grained molecular dynamics. *Biophys J*, 111(10): 2086–98, 2016.

[48] Zongan Wang, John M. Jumper, Sheng Wang, Karl F. Freed, and Tobin R. Sosnick. A membrane burial potential with h-bonds and applications to curved membranes and fast simulations. *Biophys J*, 115:1872–84, 2018.

[49] B. R. Brooks, 3rd C. L. Brooks, Jr. A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: the biomolecular simulation program. *J Comput Chem*, 30(10):1545–614, 2009.

[50] S. Jo, T. Kim, V. G. Iyer, and W. Im. Charmm-gui: a web-based graphical user interface for charmm. *J Comput Chem*, 29(11):1859–65, 2008.

[51] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J Comput Chem*, 25(9):1157–74, 2004.

[52] A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg. Anisotropic solvent model of the lipid bilayer. 2. energetics of insertion of small molecules, peptides, and proteins in membranes. *J Chem Inf Model*, 51(4):930–46, 2011.

[53] Y. Tian, C. D. Schwieters, S. J. Opella, and F. M. Marassi. A practical implicit membrane potential for nmr structure calculations of membrane proteins. *Biophys J*, 109(3):574–85, 2015.

[54] T. Lazaridis. Effective energy function for proteins in lipid membranes. *Proteins*, 52 (2):176–92, 2003.

[55] V. Z. Spassov, L. Yan, and S. Szalma. Introducing an implicit membrane in generalized born/solvent accessibility continuum solvent models. *J Phys Chem B*, 106(34):8726–38, 2002.

[56] K. Illergard, S. Callegari, and A. Elofsson. MPRAP: an accessibility predictor for a-helical transmembrane proteins that performs well inside and outside the membrane. *BMC Bioinformatics*, 11:333, 2010.

[57] A. Ray, E. Lindahl, and B. Wallner. Model quality assessment for membrane proteins. *Bioinformatics*, 26(24):3067–74, 2010.

[58] A. J. Heim and Z. Li. Developing a high-quality scoring function for membrane protein structures based on specific inter-residue interactions. *J Comput Aided Mol Des*, 26 (3):301–9, 2012.

[59] L. Adamian and J. Liang. Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. *J Mol Biol*, 311(4):891–907, 2001.

[60] C. Wendel and H. Gohlke. Predicting transmembrane helix pair configurations with knowledge-based distance-dependent pair potentials. *Proteins-Structure Function and Bioinformatics*, 70(3):984–99, 2008.

[61] G. von Heijne. The distribution of positively charged residues in bacterial inner membrane-proteins correlates with the trans-membrane topology. *EMBO Journal*, 5(11):3021–7, 1986.

[62] I. T. Arkin and A. T. Brunger. Statistical analysis of predicted transmembrane alpha-helices. *Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology*, 1429(1):113–28, 1998.

[63] M. B. Ulmschneider and M. S. P. Sansom. Amino acid distributions in integral membrane protein structures. *Biochimica Et Biophysica Acta-Biomembranes*, 1512(1):1–14, 2001.

[64] T. Beuming and H. Weinstein. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. *Bioinformatics*, 20 (12):1822–35, 2004.

[65] M. B. Ulmschneider, M. S. P. Sansom, and A. Di Nola. Properties of integral membrane protein structures: Derivation of an implicit membrane potential. *Proteins-Structure Function and Bioinformatics*, 59(2):252–65, 2005.

[66] L. Adamian, V. Nanda, W. F. DeGrado, and J. Liang. Empirical lipid propensities of amino acid residues in multispan alpha helical membrane proteins. *Proteins-Structure Function and Bioinformatics*, 59(3):496–509, 2005.

[67] L. Adamian and J. Liang. Prediction of buried helices in multispan alpha helical membrane proteins. *Proteins-Structure Function and Bioinformatics*, 63(1):1–5, 2006.

[68] A. Senes, D. C. Chadi, P. B. Law, R. F. S. Walters, V. Nanda, and W. F. DeGrado. E-z, a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: Derivation and applications to determining the orientation of transmembrane and interfacial helices. *J Mol Biol*, 366(2):436–48, 2007.

[69] D. Hsieh, A. Davis, and V. Nanda. A knowledge-based potential highlights unique features of membrane alpha-helical and beta-barrel protein insertion and folding. *Protein Science*, 21(1):50–62, 2012.

[70] C. A. Schramm, B. T. Hannigan, J. E. Donald, C. Keasar, J. G. Saven, W. F. DeGrado, and I. Samish. Knowledge-based potential for positioning membrane-associated structures and assessing residue-specific energetic contributions. *Structure*, 20(5):924–35, 2012.

[71] T. Nugent and D. T. Jones. Membrane protein orientation and refinement using a knowledge-based statistical potential. *BMC Bioinformatics*, 14, 2013.

[72] J. Koehler Leman, R. Bonneau, and M. B. Ulmschneider. Statistically derived asymmetric membrane potentials from alpha-helical and beta-barrel membrane proteins. *Sci Rep*, 8:4446, 2018.

[73] G. von Heijne. Analysis of the Distribution of Charged Residues in the N-Terminal Region of Signal Sequences - Implications for Protein Export in Prokaryotic and Eukaryotic Cells. *EMBO Journal*, 3(10):2315–8, 1984.

[74] W. M. Yau, W. C. Wimley, K. Gawrisch, and S. H. White. The preference of tryptophan for membrane interfaces. *Biochem*, 37(42):14713–8, 1998.

[75] R. F. S. Walters and W. F. DeGrado. Helix-packing motifs in membrane proteins. *Proc Natl Acad Sci U S A*, 103(37):13658–63, 2006.

[76] M. A. Lomize, I. D. Pogozheva, H. Joo, H. I. Mosberg, and A. L. Lomize. Opm database and ppm web server: resources for positioning of proteins in membranes. *Nucleic Acids Research*, 40(D1):D370–D6, 2012.

[77] E. Perozo, A. Kloda, D. M. Cortes, and B. Martinac. Physical principles underlying the transduction of bilayer deformation forces during mechanosensitive channel gating. *Nat Struct Biol*, 9(9):696–703, 2002.

[78] J. M. Jumper, N. F. Faruk, K. F. Freed, and T. R. Sosnick. Trajectory-based training enables protein simulations with accurate folding and Boltzmann ensembles in cpu-hours. *PLOS Comp Biology*, 14:e1006578, 2018.

[79] J. M. Jumper, N. F. Faruk, K. F. Freed, and T. R. Sosnick. Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics. *PLOS Comp Biology*, 14:e1006342, 2018.

[80] G. Wang and Jr. Dunbrack, R. L. Pisces: a protein sequence culling server. *Bioinformatics*, 19(12):1589–91, 2003.

[81] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *J Mol Graph*, 14(1):33–8, 27–8, 1996.

[82] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernandez, C. R. Schwantes, L. P. Wang, T. J. Lane, and V. S. Pande. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophys J*, 109(8):1528–32, 2015.

[83] A. N. Adhikari, K. F. Freed, and T. R. Sosnick. De novo prediction of protein folding pathways and structure using the principle of sequential stabilization. *Proc Natl Acad Sci U S A*, 109(43):17442–7, 2012.

[84] E. N. Baker and R. E. Hubbard. Hydrogen bonding in globular proteins. *Prog Biophys Mol Biol*, 44(2):97–179, 1984.

[85] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, 102(18):3586–616, 1998.

[86] A. Savitzky and M. J. E. Golay. Smoothing + differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627, 1964.

[87] D. C. Liu and J. Nocedal. On the Limited Memory BFGS Method for Large-Scale Optimization. *Mathematical Programming*, 45(3):503–28, 1989.

[88] T. E. Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9(3):10–20, 2007.

[89] Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006.

[90] Team The Theano Development, R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, A. Belopolsky, Y. Bengio, A. Bergeron, J. Bergstra, V. Bisson, J. Bleecher Snyder, N. Bouchard, N. Boulanger-Lewandowski, X. Bouthillier, A. de Brébisson, O. Breuleux, P. Carrier, K. Cho, J. Chorowski, P. Christiano, T. Cooijmans, M.-A. Côté, M. Côté, A. Courville, Y. N. Dauphin, O. Delalleau, J. Demouth, G. Desjardins, S. Dieleman, L. Dinh, M. Ducoffe, V. Dumoulin, S. Ebrahimi Kahou, D. Erhan, Z. Fan, O. Firat, M. Germain, X. Glorot, I. Goodfellow, M. Graham, C. Gulcehre, P. Hamel, I. Harlouchet, J.-P. Heng, B. Hidasi, S. Honari, A. Jain, S. Jean, K. Jia, M. Korobov, V. Kulkarni, A. Lamb, P. Lamblin, E. Larsen, C. Laurent, S. Lee, S. Lefrancois, S. Lemieux, N. Léonard, Z. Lin, J. A. Livezey, C. Lorenz, J. Lowin, Q. Ma, P.-A. Manzagol, O. Mastropietro, R. T. McGibbon, R. Memisevic, B. van Merriënboer, V. Michalski, M. Mirza, A. Orlandi, C. Pal, R. Pascanu, M. Pezeshki, C. Raffel, D. Renshaw, M. Rocklin, A. Romero, M. Roth,

P. Sadowski, J. Salvatier, F. Savard, J. Schlüter, J. Schulman, G. Schwartz, I Vlad Serban, D. Serdyuk, S. Shabanian, É Simon, S. Spieckermann, S. Ramana Subramanyam, J. Sygnowski, J. Tanguay, G. van Tulder, et al. Theano: A python framework for fast computation of mathematical expressions. *ArXiv e-prints*, 1605, 2016.

[91] A. L. Lomize, I. D. Pogozheva, M. A. Lomize, and H. I. Mosberg. Positioning of proteins in membranes: a computational approach. *Protein Sci*, 9:1318–33, 2006.

[92] A. K. Chamberlain, Y. Lee, S. Kim, and J. U. Bowie. Snorkeling preferences foster an amino acid composition bias in transmembrane helices. *Journal of Molecular Biology*, 339(2):471–9, 2004.

[93] D. G. Isom, B. R. Cannon, C. A. Castaneda, A. Robinson, and B. Garcia-Moreno. High tolerance for ionizable residues in the hydrophobic interior of proteins. *Proc Natl Acad Sci U S A*, 105:17784–8, 2008.

[94] D. G. Isom, C. A. Castaneda, B. R. Cannon, and B. Garcia-Moreno. Large shifts in pKa values of lysine residues buried inside a protein. *Proc Natl Acad Sci U S A*, 108: 5260–5, 2011.

[95] C. A. Fitch, G. Platzer, M. Okon, B. E. Garcia-Moreno, and L. P. McIntosh. Arginine: Its pKa value revisited. *Protein Sci*, 24:752–61, 2015.

[96] H. Gong, G. Hocky, and K. F. Freed. Influence of nonlinear electrostatics on transfer energies between liquid phases: charge burial is far less expensive than born model. *Proc Natl Acad Sci U S A*, 105:11146–51, 2008.

[97] M. A. Lomize, A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg. Opm: Orientations of proteins in membranes database. *Bioinformatics*, 22(5):623–5, 2006.

[98] S. Wang, J. Ma, and J. Xu. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. *Bioinformatics*, 32(17):i672–i9, 2016.

[99] S. Wang, S. Sun, and J. Xu. AUC-Maximized Deep Convolutional Neural Fields for Protein Sequence Labeling. *Mach Learn Knowl Discov Databases*, 9852:1–16, 2016.

[100] Y. Zhang and J. Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*, 33:2302–9, 2005.

[101] O. S. Andersen and 2nd. R. E. Koeppe. Bilayer thickness and membrane protein function: an energetic perspective. *Annu Rev Biophys Biomol Struct*, 36:107–30, 2007.

[102] D. Argudo, N. P. Bethel, F. V. Marcoline, and M. Grabe. Continuum descriptions of membranes and their interaction with proteins: Towards chemically accurate models. *BBA-Biomembranes*, 1858:1619–34, 2016.

[103] K. M. Callenberg, N. R. Latorraca, and M. Grabe. Membrane bending is critical for the stability of voltage sensor segments in the membrane. *J Gen Physiol*, 140:55–68, 2012.

[104] H. T. McMahon and J. L. Gallop. Membrane curvature and mechanisms of dynamic cell membrane remodelling. *Nature*, 438(7068):590–6, 2005.

[105] S. I. Sukharev, W. J. Sigurdson, C. Kung, and F. Sachs. Energetic and spatial parameters for gating of the bacterial large conductance mechanosensitive channel, MscL. *J Gen Physiol*, 113:525–40, 1999.

[106] P. Wiggins and R. Phillips. Analytic models for mechanotransduction: gating a mechanosensitive channel. *Proc Natl Acad Sci U S A*, 101:4071–6, 2004.

[107] R. Phillips, T. Ursell, P. Wiggins, and P. Sens. Emerging roles for lipids in shaping membrane-protein function. *Nature*, 459:379–85, 2009.

[108] A. Pressley. *Elementary differential geometry.* Information science and statistics. Springer undergraduate mathematics series, London; New York, 2010.

[109] J. Koehler Leman, S. Lyskov, and R. Bonneau. Computing structure-based lipid accessibility of membrane proteins with texttmp_lipid_acc in RosettaMP. *BMC Bioinformatics*, 18:115, 2017.

[110] N. BenTal, D. Sitkoff, I. A. Topol, A. S. Yang, S. K. Burt, and B. Honig. Free energy of amide hydrogen bond formation in vacuum, in water, and in liquid alkane solution. *J Phys Chem B*, 101(3):450–7, 1997.

[111] D. C. Marx and K. G. Fleming. Influence of protein scaffold on side-chain transfer free energies. *Biophys J*, 113:597–604, 2017.

[112] C. P. Moon and K. G. Fleming. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proc Natl Acad Sci U S A*, 108:10174–7, 2011.

[113] J. U. Bowie. Membrane protein folding: how important are hydrogen bonds? *Curr Opin Struct Biol*, 21:42–9, 2011.

[114] P. D. Thomas and K. A. Dill. Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol*, 257:457–69, 1996.

[115] A. BenNaim. Statistical potentials extracted from protein structures: Are these meaningful potentials? *J Chem Phys*, 107:3698–706, 1997.

[116] A. V. Finkelstein, A. Y. Badretdinov, and A. M. Gutin. Why do protein architectures have boltzmann-like statistics. *Proteins*, 23:142–150, 1995.

[117] J. Skolnick, L. Jaroszewski, A. Kolinski, and A. Godzik. Derivation and testing of pair potentials for protein folding: when is the quasichemical approximation correct? *Protein Sci*, 6:676–88, 1997.

[118] M. Y. Shen and A. Sali. Statistical potential for assessment and prediction of protein structures. *Protein Sci*, 15:2507–24, 2006.

[119] K. T. Simons, I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, and D. Baker. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, 34:82–95, 1999.

[120] R. Samudrala and J. Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol*, 275:895–916, 1998.

[121] A. D. Solis and S. Rackovsky. Improvement of statistical potentials and threading score functions using information maximization. *Proteins*, 62:892–908, 2006.

[122] J. L. MacCallum, W. F. D. Bennett, and D. P. Tieleman. Transfer of arginine into lipid bilayers is nonadditive. *Biophy J*, 101:110–7, 2011.

[123] Dorairaj, S., and T. W. Allen. On the thermodynamic stability of a charged arginine side chain in a transmembrane helix. *Proc Natl Acad Sci U S A*, 104:4943–8, 2007.

[124] Zongan Wang, John M. Jumper, Karl F. Freed, and Tobin R. Sosnick. On the interpretation of unfolding measurements of membrane proteins under force using fast simulations. submitted, 2019.

[125] K. C. Neuman and A. Nagy. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nat Methods*, 5(6):491–505, 2008.

[126] J. Stigler, F. Ziegler, A. Gieseke, J. C. Gebhardt, and M. Rief. The complex folding network of single calmodulin molecules. *Science*, 334(6055):512–6, 2011.

[127] S. Sen Mojumdar, N. Scholl Z, D. R. Dee, L. Rouleau, U. Anand, C. Garen, and M. T. Woodside. Partially native intermediates mediate misfolding of sod1 in single-molecule folding trajectories. *Nat Commun*, 8(1):1881, 2017.

[128] P. E. Marszalek, H. Lu, H. Li, M. Carrion-Vazquez, A. F. Oberhauser, K. Schulten, and J. M. Fernandez. Mechanical unfolding intermediates in titin modules. *Nature*, 402(6757):100–3, 1999.

[129] F. Oesterhelt, D. Oesterhelt, M. Pfeiffer, A. Engel, H. E. Gaub, and D. J. Müller. Unfolding pathways of individual bacteriorhodopsins. *Science*, 288(5463):143–6, 2000.

[130] D. Min, R. E. Jefferson, J. U. Bowie, and T. Y. Yoon. Mapping the energy landscape for second-stage folding of a single membrane protein. *Nat Chem Biol*, 11(12):981–7, 2015.

[131] T. Serdiuk, D. Balasubramaniam, J. Sugihara, S. A. Mari, H. R. Kaback, and D. J. Muller. Yidc assists the stepwise and stochastic folding of membrane proteins. *Nat Chem Biol*, 12(11):911–7, 2016.

[132] D. J. Müller, M. Kessler, F. Oesterhelt, C. Moller, D. Oesterhelt, and H. Gaub. Stability of bacteriorhodopsin alpha-helices and loops analyzed by single-molecule force spectroscopy. *Biophys J*, 83(6):3578–88, 2002.

[133] C. A. Bippes and D. J. Müller. High-resolution atomic force microscopy and spectroscopy of native membrane proteins. *Reports on Progress in Physics*, 74(8), 2011.

[134] M. Kessler, K. E. Gottschalk, H. Janovjak, D. J. Müller, and H. E. Gaub. Bacteriorhodopsin folds into the membrane against an external force. *J Mol Biol*, 357(2): 644–54, 2006.

[135] M. Kessler and H. E. Gaub. Unfolding barriers in bacteriorhodopsin probed from the cytoplasmic and the extracellular side by afm. *Structure*, 14(3):521–7, 2006.

[136] C. Kappel and H. Grubmuller. Velocity-dependent mechanical unfolding of bacteriorhodopsin is governed by a dynamic interaction network. *Biophys J*, 100(4):1109–19, 2011.

[137] R. Guo, K. Gaffney, Z. Yang, M. Kim, S. Sungsuwan, X. Huang, W. L. Hubbell, and H. Hong. Steric trapping reveals a cooperativity network in the intramembrane protease glpg. *Nat Chem Biol*, 12(5):353–60, 2016.

[138] W. Paslawski, O. K. Lillelund, J. V. Kristensen, N. P. Schafer, R. P. Baker, S. Urban, and D. E. Otzen. Cooperative folding of a polytopic alpha-helical membrane protein involves a compact n-terminal nucleus and nonnative loops. *Proc Natl Acad Sci U S A*, 112(26):7978–83, 2015.

[139] N. P. Schafer, H. H. Truong, D. E. Otzen, K. Lindorff-Larsen, and P. G. Wolynes. Topological constraints and modular structure in the folding and functional motions of glpg, an intramembrane protease. *Proc Natl Acad Sci U S A*, 113(8):2098–103, 2016.

[140] W. Lu, N. P. Schafer, and P. G. Wolynes. Energy landscape underlying spontaneous insertion and folding of an alpha-helical transmembrane protein into a bilayer. *Nat Commun*, 9(1):4949, 2018.

[141] H. Lu, B. Isralewitz, A. Krammer, V. Vogel, and K. Schulten. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophys J*, 75(2): 662–71, 1998.

[142] G. Stirnemann, D. Giganti, J. M. Fernandez, and B. J. Berne. Elasticity, structure, and relaxation of extended proteins under force. *Proc Natl Acad Sci U S A*, 110(10): 3847–52, 2013.

[143] P. I. Zhuravlev, M. Hinczewski, S. Chakrabarti, S. Marqusee, and D. Thirumalai. Force-dependent switch in protein unfolding pathways and transition-state movements. *Proc Natl Acad Sci U S A*, 113(6):E715–24, 2016.

[144] M. Ernst, F. Sittel, and G. Stock. Contact- and distance-based principal component analysis of protein dynamics. *J Chem Phys*, 143(24):244114, 2015.

[145] D. Ting, G. Wang, M. Shapovalov, R. Mitra, M. I. Jordan, and Jr. Roland L. Dunbrack. Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. *PLoS Comput Biol*, 6(4):e1000763, 2010.

[146] C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith. Entropic elasticity of lambda-phage dna. *Science*, 265(5178):1599–600, 1994.

[147] T. Strick, J. F. Allemand, V. Croquette, and D. Bensimon. Twisting and stretching single dna molecules. *Progress in Biophysics & Molecular Biology*, 74(1-2):115–40, 2000.

[148] R. A. Evans. Abramowitz M - Handbook of Mathematical Functions with Formulas Graphs and Mathematical Tables Nbs Applied Mathematics Series 55. *Ieee Spectrum*, 3(7):161–, 1966.

[149] R. Levy and M. Maaloum. Measuring the spring constant of atomic force microscope cantilevers: thermal fluctuations and other methods. *Nanotechnology*, 13(1): 33–7, 2002.

[150] S. R. Ainavarapu, J. Brujic, H. H. Huang, A. P. Wiita, H. Lu, L. Li, K. A. Walther, M. Carrion-Vazquez, H. Li, and J. M. Fernandez. Contour length and refolding rate of a small protein controlled by engineered disulfide bonds. *Biophys J*, 92(1):225–33, 2007.

[151] S. Subramaniam and R. Henderson. Molecular mechanism of vectorial proton translocation by bacteriorhodopsin. *Nature*, 406:653–7, 2000.

[152] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637, 1983.

[153] P. Curnow, N. D. Di Bartolo, K. M. Moreton, O. O. Ajoje, N. P. Saggese, and P. J. Booth. Stable folding core in the folding transition state of an alpha-helical integral membrane protein. *Proc Natl Acad Sci U S A*, 108(34):14133–8, 2011.

[154] J. Cladera, J. Torres, and E. Padros. Analysis of conformational changes in bacteriorhodopsin upon retinal removal. *Biophys J*, 70(6):2882–7, 1996.

[155] K. Voitchovsky, S. A. Contera, and J. F. Ryan. Electrostatic and steric interactions determine bacteriorhodopsin single-molecule biomechanics. *Biophys J*, 93:2024–37, 2007.

[156] K. T. Sapra, J. Doehner, V. Renugopalakrishnan, E. Padros, and D. J. Müller. Role of extracellular glutamic acids in the stability and energy landscape of bacteriorhodopsin. *Biophys J*, 95:3407–18, 2008.

[157] Y. Yang, R. Guo, K. Gaffney, M. Kim, S. Muhammednazaar, W. Tian, B. Wang, J. Liang, and H. Hong. Folding-degradation relationship of a membrane protein mediated by the universally conserved atp-dependent protease ftsh. *J Am Chem Soc*, 140 (13):4656–65, 2018.

[158] N. Guex and M. C. Peitsch. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, 18(15):2714–23, 1997.

[159] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–30, 2011.

[160] Andrei L. Lomize and Irina D. Pogozheva. TMDOCK: an energy-based method for modeling α-helical dimers in membranes. *J Mol Biol*, 429:3908, 2017.

[161] Florian Cymer, Anbazhagan Veerappan, and Dirk Schneider. Transmembrane helixhelix interactions are modulated by the sequence context and by lipid bilayer properties. *Biochim. Biophys. Acta*, 1818(4):96373, 2012.

[162] Drake C. Mitchell. Progress in understanding the role of lipids in membrane protein folding. *Biochim. Biophys. Acta*, 1818(4):9516, 2012.

[163] Johannes Thoma, Bjorn M Burmann, Sebastian Hiller, and Daniel J Müller. Impact of holdase chaperones Skp and SurA on the folding of β-barrel outer-membrane proteins. *Nat Struct Mol Biol*, 22(10):795–804, 2015.

[164] Duyoung Min, Robert E. Jefferson, Yifei Qi, Jing Yang Wang, Mark A. Arbing, Wonpil Im, and James U. Bowie. Unfolding of a clc chloride transporter retains memory of its evolutionary history. *Nat Chem Biol*, 14:48996, 2018.

[165] C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys Rev Lett*, 78 (14):26903, 1997.

[166] Nolan C. Harris, Yang Song, and Ching-Hwa Kiang. Experimental free energy surface reconstruction from single-molecule force spectroscopy using jarzynskis equality. *Phys Rev Lett*, 99:0681014, 2007.

[167] Johannes Preiner, Harald Janovjak, Christian Rankl, Helene Knaus, David A. Cisneros, Alexej Kedrov, Ferry Kienberger, Daniel J. Muller, and Peter Hinterdorfer. Free energy of membrane protein unfolding derived from single-molecule force measurements. *Biophys J*, 93:9307, 2007.

[168] Patrick R. Heenan, Hao Yu, Matthew G. W. Siewny, and Thomas T. Perkins. Improved free-energy landscape reconstruction of bacteriorhodopsin highlights local variations in unfolding energy. *J Chem Phys*, 148:1233137, 2018.

[169] Werner G. Krebs and Mark Gerstein. The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res*, 28(8):166575, 2000.

[170] Dahlia R. Weiss and Michael Levitt. Can morphing methods predict intermediate structures? *J Mol Biol*, 385:66574, 2009.

[171] José Ramón López-Blanco, José I. Aliaga, Enrique S. Quintana-Orti, and Pablo Chacón. iMODS: internal coordinates normal mode analysis server. *Nucleic Acids Res*, 42(W1):W2716, 2014.

[172] Turkan Haliloglu, Ivet Bahar, and Burak Erman. Gaussian dynamics of folded proteins. *Phys Rev Lett*, 79(16):30903, 1997.

[173] Laura Orellana, Ozge Yoluk, Oliver Carrillo, Modesto Orozco, and Erik Lindah. Prediction and validation of protein intermediate states from structurally rich ensembles and coarse-grained simulations. *Nat Comm*, 7:12575, 2016.

[174] Avisek Das, Mert Gur, Mary Hongying Cheng, Sunhwan Jo, Ivet Bahar, and Benoit Roux. Exploring the conformational transitions of biomolecular systems using a simple two-state anisotropic network model. *PLoS Comput Biol*, 10(4):e1003521, 2014.

[175] Albert C. Pan, Deniz Sezer, and Benoit Roux. Finding transition pathways using the string method with swarms of trajectories. *J Phys Chem B*, 112:3432–40, 2008.

[176] Wenxun Gan, Sichun Yang, and Benoit Roux. Atomistic view of the conformational activation of src kinase using the string method with swarms-of-trajectories. *Biophys J*, 97(4):L8–L10, 2009.