Sec-Facilitated Protein Translocation and Membrane Integration

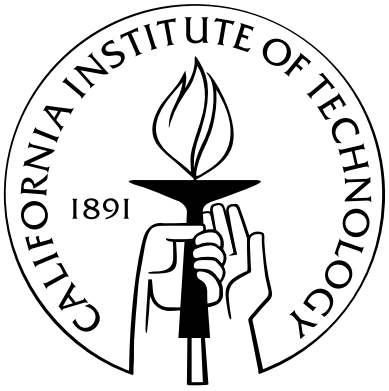
Thesis by

Bin Zhang

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2013

(Defended October 3, 2012)

Acknowledgments

First, I owe my deepest gratitude to my thesis advisor Professor Thomas F. Miller. It has been his stimulating discussions, constant support and encouragement, and endless patience that made everything possible to complete my degree. His breadth of knowledge, scientific insight and perpetual passion filled the journey of my PhD study full of inspiration. I could not thank him enough for his dedication and committment. My five-year experience with him is an invaluable fortune from which I will benefit for the rest of my career.

I would also like to thank my committee members: Professor William Clemons, Professor Zhen-Gang Wang, and Professor Dennis Dougherty for their guidance and feedback on my research project, and for their help in my job applications. I truly appreciate their valuable time, and it is a great honor to have them on my committee.

Additionally, I am grateful to all the members of the Miller group for making every day at work full of joy and excitement. I am lucky to work with these many talented colleagues, and my conversations with them have always been eye-opening. I am especially indebted to their help in improving my presentation skills and in proofreading my manuscripts.

Last but not least, I would like to thank my family and my friends. Living in a less-developed small town, my parents have tried to save every penny they can so that I will have

the opportunity to go to graduate school and receive a world-class education. Even though it means extra working hours and no vacations, they never have complaints and are always supportive in pursuing my dreams. For that, I owe all the credits of my accomplishments to them. I feel extremely blessed to meet my fiancée, Na Hu during my PhD study, and words are not enough to praise her understanding, support, and encouragement. I also thank all the friends I met at Caltech, espeically Xin Hu, Mulin Cheng, Bolin Lin, and Yao Sha for smoothing my transition to the USA, and for making my graduate school experience more enjoyable.

Abstract

The Sec translocon is a central component of the cellular machinery for targeting and delivering nascent proteins. Ubiquitous across all kingdoms of life, it is a protein-conducting channel that facilitates recognition of integral membrane protein domains and the establishment of integral membrane protein topology. Structural, biochemical, and biophysical studies have illuminated the role of the Sec translocon in both cotranslational and posttranslational protein targeting. In particular, quantitative assays have established the dependence of transmembrane domain (TM) stop-transfer efficiency and integral membrane protein topogenesis on the physicochemical properties of the translocon and protein nascent chain. These studies provide a valuable starting point for understanding the molecular regulation of the translocon and its sensitivity to mutations in protein sequence and external driving forces; however, complexities associated with the Sec machinery, including the role of collaborating molecular motors, the importance of large-scale conformational changes in the translocon, and the crowded molecular environment of the channel interior, obscure the mechanistic basis for many experimentally observed trends. My PhD research has focused on the development of a unified, mechanistic understanding of Sec-facilitated protein targeting.

Using both atomistic and coarse-grained molecular simulations, we have investigated

the conformational landscape for the Sec translocon. We found that inclusion of a hydrophobic peptide substrate in the translocon stabilizes an open conformation of the lateral gate (LG) that is necessary for membrane integration, whereas inclusion of a hydrophilic peptide substrate favors only the closed LG conformation. We demonstrated that the translocon plug moiety adopts markedly different conformations in the channel, depending on whether the substrate peptide is hydrophobic or hydrophilic in character. Finally, we showed that the energetics of the translocon LG opening in the presence of the substrate peptides can be modeled in terms of the energetics of the peptide interface with the membrane. The manuscript associated with this study is published in *PNAS*, 107, 5399 (2010).

We further developed a novel computational protocol that combines nonequilibrium growth of the nascent protein with microsecond-timescale molecular dynamics trajectories. Analysis of multiple, long-timescale simulations elucidated molecular features of protein insertion into the translocon, including signal-peptide docking at the translocon LG, large-lengthscale conformational rearrangement of the translocon LG helices, and partial membrane integration of hydrophobic nascent-protein sequences. Furthermore, the simulations demonstrated the role of specific molecular interactions in the regulation of protein secretion, membrane integration, and integral membrane protein topology. Salt-bridge contacts between the nascent-protein N-terminus, cytosolic translocon residues, and phospholipid head groups were shown to favor conformations of the nascent protein upon early-stage insertion that are consistent with the Type II (N_{cyt}/C_{exo}) integral membrane protein topology; and extended hydrophobic contacts between the nascent protein and the membrane

lipid bilayer were shown to stabilize configurations that are consistent with the Type III (N_{exo}/C_{cyt}) topology. These results provide a detailed, mechanistic basis for understanding experimentally observed correlations between integral membrane protein topology, translocon mutagenesis, and nascent-protein sequence. The manuscript associated with this study is published in *J. Am. Chem. Soc.*, 134, 13700 (2012).

Finally, we introduced a coarse-grained modeling approach that spans the nanosecond- to minute-timescale dynamics of cotranslational protein translocation. The method enabled direct simulation of both integral membrane protein topogenesis and TM stop-transfer efficiency. Simulations revealed multiple kinetic pathways for protein integration, including a mechanism in which the nascent protein undergoes slow-timescale reorientation, or flipping, in the confined environment of the translocon channel. Competition among these pathways gives rise to the experimentally observed dependence of protein topology on ribosomal translation rate and protein length. We further demonstrated that sigmoidal dependence of stop-transfer efficiency on TM hydrophobicity arises from local equilibration of the TM across the translocon LG, and it was predicted that slowing ribosomal translation yields decreased stop-transfer efficiency in long proteins. This work reveals the balance between equilibrium and nonequilibrium processes in protein targeting, and it provids new insight into the molecular regulation of the Sec translocon. The manuscript associated with this study is published in *Cell Reports*, in press.

This research has significantly enriched the mechanistic understanding of Sec-facilitated protein translocation and membrane integration with ample molecular details. The unifying picture that we propose establishes fundamental connections between previously disparate

experimental studies, and it lays down the foundation for future verification and refinement.

Contents

Abbreviations

CG	–	coarse-grained
CTL	–	C-terminal tail length
ER	–	endoplasmic reticulum
FE	–	free energy
LG	–	lateral gate
MD	–	molecular dynamics
MDL	–	mature domain length
PME	–	particle mesh Ewald
POPC	–	palmitoyloleoylphosphatidylcholine
PP	–	pore-plug
RMS	–	root mean square
SP	–	signal peptide
TM	–	transmembrane
WHAM	–	weighted histogram analysis method

Chapter 1

Introduction

Most proteins are synthesized in the cytosolic region of the cell by ribosomes, while many of them function at distinct compartments including the membrane bilayer, the interior of endoplasmic reticulum (ER), and even the exterior of the cell. Cell has evolved a sophisticated pathway for delivering proteins from cytosol to their destinations, and the first step of this delivery process occurs as proteins translocate across the eukaryotic ER membrane, or the bacterial plasma membrane. Targeting protein to a particular membrane is achieved via a short stretch of hydrophobic residues located at its N-terminus, the signal peptide. Protein translocation across the membrane can proceed either while it is being actively translated via the ribosome (cotranslational) or after the translation is finished (posttranslational). In both situations, a protein conduction channel, a multispanning membrane protein termed the Sec translocon, facilitates the translocation.

As both soluble proteins and integral membrane proteins share the same targeting pathway, the Sec translocon is expected to allow both translocation and integration to occur (Figure 1.1). Soluble proteins that function in a more hydrophilic environment, such as the ER lumen, will translocate completely across the membrane from its cis side (cytosol) to the trans side (ER lumen). The translocon in this case functions as a hydrophilic chan-

nel that shields the protein from the surrounding hydrophobic lipid molecules. Membrane proteins, on the other hand, are composed of transmembrane (TM) segments that span the membrane lipid bilayer. The translocon-assisted integration of TM helix into the membrane helps to bypass the energetic barrier imposed by polar lipid head groups. What determines the final destination of a given protein and whether the translocon plays an active role in sorting proteins, however, are unclear.

The anchoring of membrane proteins into the lipid bilayer is further complicated with topological constraints. In many cases, the solvent-exposed group of a membrane protein plays important functional roles , and which side of the membrane it stays has direct consequences on the productivity of the protein. The molecular mechanism for the establishment of integral membrane protein topology via Sec translocon remains elusive.

Studies of the cotranslational membrane protein integration suggest the interaction between the nascent protein and the membrane lipids to play an important role in directing the integration of TM helices [1–3]. Evidence leading to this proposition first comes from cross-linking experiment, which revealed that nascent protein TM domain is exposed to surrounding lipid molecules as soon as it arrives in the channel, and introducing charged hydrophilic residues into the TM segment abolishes its membrane integration [1]. This view is further supported by the measurement of striking correlations between a “biological hydrophobicity scale" for peptides and the relative fraction of peptides that undergo Sec-mediated integration vs. translocation from the von Heine and White lab [4–6]; and it has been justified in terms of an effective thermodynamic partitioning for peptide substrates between the largely hydrophilic interior of the channel and the hydrophobic interior of the

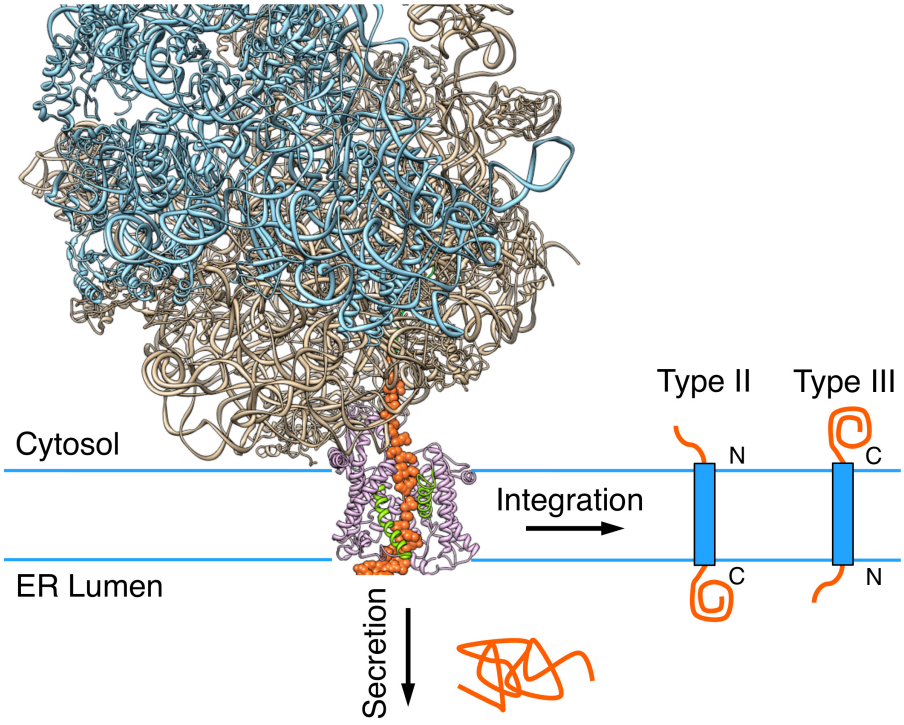


Figure 1.1: Illustration of the cotranslational protein translocation pathway. **(center)** Atomistic representation of the cellular machinery involved in the cotranslational protein translocation. The ribosome (blue and yellow) carrying the nascent protein (orange) is shown in complex with the Sec translocon (violet). The two helices of the translocon that allows the lateral partition of TM segments are colored in green. **(right)** The two distinct orientations adopted by integral membrane proteins. In the Type II orientation, the protein N-terminus rests in the cis side (cytosol) of the membrane bilayer, while the C-terminus is translocated to the trans side (ER lumen). In the Type III orientation, the protein adopts the completely opposite topology. **(bottom)** The secreted product for a soluble protein.

membrane [1, 4]. Consistent with the partition model, the derived biological hydrophobicity scale succeeds in distinguishing single spanning TM domains from the soluble counterparts [6].

This thermodynamic equilibrium model of local TM segment for Sec-facilitated membrane protein integration, however, is in apparent contrast with kinetic effects observed in membrane protein topogenesis studies conducted by Spiess and co-workers. In a set of experiments, they measured the fraction of proteins that adopt the Type II orientation (i.e., with N terminus in cytosol and C terminus in ER lumen, Figure 1.1) as a function of

nascent protein length and ribosomal translation rate [7, 8]. An unexpected rise of the Type II integration was observed both when the nascent protein was lengthened and when the protein translation rate was slowed down. These pronounced kinetic effects argue strongly for existence of long-ranged and long-timescale regulation during the membrane protein integration process.

The development of a unified, mechanistic understanding of Sec-facilitated protein targeting is hindered by the complex and important roles of collaborating molecular motors, large-scale conformational changes in the translocon, and the crowded molecular environment of the channel interior. Currently, there is no coherent approach to explore the mechanistic basis for these observed kinetic effects. Nor is it clear how to reconcile the apparent role of equilibrium partitioning in the work of von Heine and White with the effects of nonequilibrium (i.e., kinetic) regulation in the work of Spiess and co-workers. It is thus the goal of my PhD research to address these challenges and to establish fundamental connections between previously disparate experimental studies of Sec-facilitated protein translocation and integral membrane protein topogenesis using computer simulation.

The rest of the thesis is organized as follows. Chapter 2 presents our investigation of the conformational landscape for the Sec translocon using both atomistic and coarse-grained molecular simulations. In particular, using enhanced thermodynamic sampling methods, we calculate the energetic cost of translocon lateral opening that is necessary for protein integration, and we study its regulation with the inclusion of peptide substrates. We find that the energetics of the translocon lateral gate (LG) opening in the presence of a peptide substrate is governed by the energetics of the peptide interface with the membrane,

and we discuss its implication for the regulation of protein secretion and membrane integration. To gain further dynamical insight on the role of large lengthscale translocon conformational changes, in chapter 3, we directly simulate the early-stage Sec-facilitated protein translocation and membrane integration with atomistic resolution. This is achieved via a novel computational protocol that combines nonequilibrium growth of the nascent protein with microsecond-timescale molecular dynamics trajectories. Results from these simulations help to elucidate the role of specific molecular interactions in the regulation of protein secretion, membrane integration, and integral membrane protein topology. Finally, in chapter 4, we introduce a coarse-grained modeling approach that spans the nanosecond- to minute-timescale dynamics of cotranslational protein translocation. The method enables direct simulation of both integral membrane protein topogenesis and TM stop-transfer efficiency and allows straightforward comparison between simulated results and experimental observations. Mechanistic analysis of simulated trajectories reveal the molecular basis of Sec-facilitated integral membrane protein integration, and it reconciles the conflicting experimental evidences for both the thermodynamic and kinetic interpretations.

Chapter 2

Hydrophobically Stabilized Open State for the Lateral Gate of the Sec Translocon

2.1 Introduction

The Sec translocon is a heterotrimeric complex of membrane-bound proteins that forms a passive channel for posttranslational and cotranslational protein translocation, as well as the cotranslational integration of proteins into the phospholipid bilayer [9]. Structural [10–14], biochemical [15, 16], and genetic [17] studies indicate that the translocon undergoes large-scale conformational changes during both the translocation and integration pathways. The translocon channel exhibits a ring, or pore, of hydrophobic amino acid residues, as well as an α-helical plug moiety that rests against the pore to occlude the channel; secretion of protein domains via the translocation pathway requires displacement of the plug with respect to the pore (Figure 2.1, left) [11, 15, 17]. Furthermore, a pair of TM helices in the translocon forms a LG that opens to expose the interior of the channel to the membrane bilayer (Figure 2.1, right) and facilitates membrane integration [1, 16, 18]. However, the detailed mechanism for membrane integration via the LG and the role of translocon con-

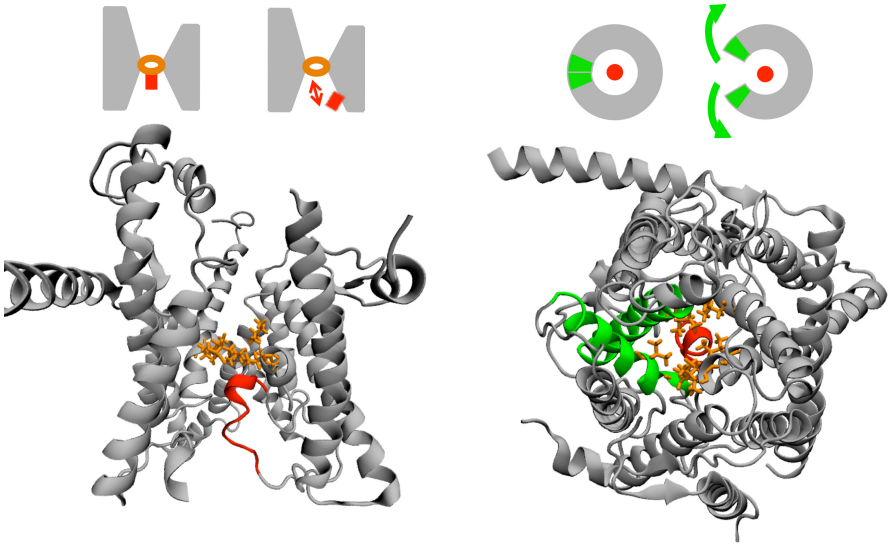


Figure 2.1: Structural features of the Sec translocon. **(left)** The translocon is viewed from within the plane of the membrane, with the pore residues shown in orange and the plug moiety shown in red. The schematic illustrates the PP displacement that is needed to allow for protein translocation via the channel. **(right)** The translocon is viewed from outside the membrane on the cytosolic side, with the pore and plug colored as before and with the TM2b and TM7 helices that form the LG shown in green. The schematic illustrates the LG motion that opens the interior of the translocon to the membrane.

formational changes in gating between the protein translocation and membrane integration pathways remain unclear.

In this chapter, the conformational landscape for the Sec translocon is investigated using atomistic and coarse-grained (CG) molecular simulations. We find that inclusion of a hydrophobic peptide substrate in the translocon stabilizes an open conformation of the LG that is necessary for membrane integration, whereas inclusion of a hydrophilic peptide substrate favors only the closed LG conformation. We demonstrate that the translocon plug moiety adopts markedly different conformations in the channel, depending on whether the substrate peptide is hydrophobic or hydrophilic in character. Finally, we show that the energetics of the translocon LG opening in the presence of the substrate peptides can be modeled in terms of the energetics of the peptide interface with the membrane. These re-

sults are consistent with an alternative interpretation of the biological hydrophobicity scale in terms of the free energy (FE) cost for opening the LG of the translocon, which suggests a refinement of the hydrophobic partitioning model in which substrate-controlled conformational gating of the translocon LG leads to regulation of the protein translocation and integration pathways.

2.2 Conformational Landscape of the Sec Translocon

To investigate the conformational flexibility of the translocon in the absence of peptide substrates, we calculate its two-dimensional FE surface in the LG and pore-plug (PP) motions using both atomistic and CG molecular dynamics (MD) simulations.

2.2.1 Atomistic Simulations

The archaeal Sec translocon [10] was studied using MD simulations with over $115,000$ atoms. The channel is modeled in a membrane composed of 254 palmitoyloleoylphosphatidylcholine (POPC) lipid molecules and with 24296 explicit water molecules. Atomistic interactions were described using the CHARMM27 force field with the TIP3P water model [19]. Counterions were included to achieve electroneutrality at a salt concentration of approximately 50 mM. MD trajectories were performed at constant temperature and pressure using orthorhombic periodic boundary conditions. Long-range electrostatics were calculated using the particle mesh Ewald (PME) technique [20]. Details of the atomistic simulations and initialization protocol are described in appendix A section: *Atomistic Simulations*.

2.2.2 Coarse-grained Simulations

Simulations were also performed using a residue-based coarse-grained (RBCG) representation for the system. Each amino acid in the channel was represented with one particle to describe the backbone group containing the α-carbon and, for residues other than glycine, a second particle to describe the side chain group [21]; the lipid molecules, counterions, and solvent are similarly coarsened using the Martini potential [22]. Following the atomistic simulations, the CG simulations were performed at constant temperature and pressure using orthorhombic periodic boundary conditions, as is detailed in appendix A section: *Coarse-Grained Simulations*.

Although the RBCG potential is parameterized to reproduce pairwise interactions for amino acid side chain and backbone groups, it has been found to poorly preserve protein tertiary structure for long MD simulations [23, 24]. As shown in the blue curves in Figure 2.2C, this issue also arises in our simulations for the Sec translocon. The radius of gyration of the channel, defined as the root mean square (RMS) distance between CG particles in the translocon and its center of mass, drifts downward as the channel deforms with increasing simulation time. Similarly, the RMS displacement of the translocon backbone CG particles following best-fit rigid-body alignment [25] drifts upwards. To stabilize the CG simulations, we thus introduce scaffolding for sections of the translocon by adding weak interactions between pairs of the CG particles. Pairs of CG particles that are included in the scaffolding share an auxiliary harmonic bond with an optimal distance equal to the separation of the particles in the crystal structure and with a force constant equal to 0.2 kcal/mol/Å^2. Scaffolding interactions are included for a pair of CG particles if both

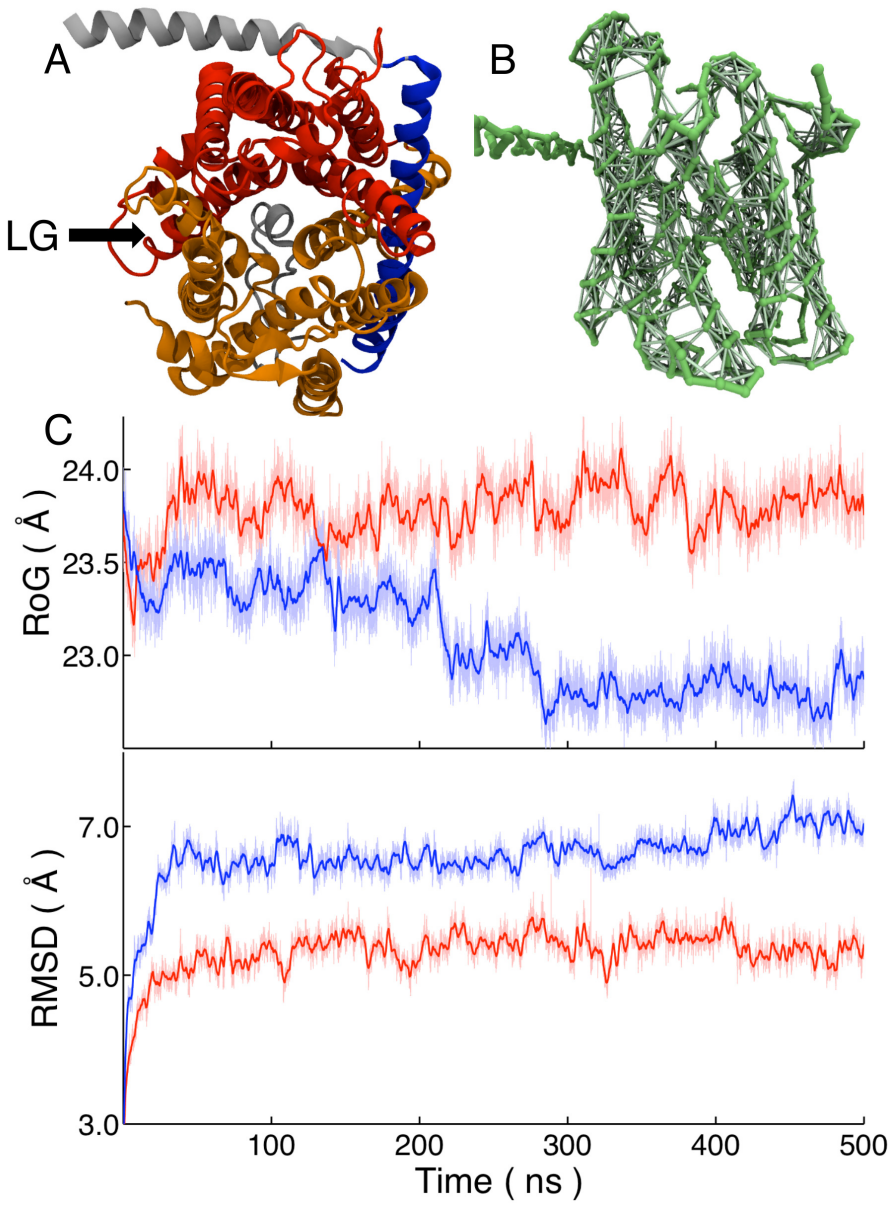


Figure 2.2: Stabilizing the CG model. **(A)** Subsets of the translocon, viewed from top, are used in the CG scaffolding protocol described in the text. **(B)** The auxiliary scaffolding interactions among CG particles for the translocon backbone, viewed from the side, are shown explicitly. **(C)** Without scaffolding, the CG model does not preserve the structural integrity of the translocon in long simulations, as is demonstrated for the translocon radius of gyration (RoG) along an MD trajectory (blue). Inclusion of the pairwise scaffolding interactions stabilizes CG MD simulations of the translocon (red). The RMS displacements for the translocon backbone CG particles are also included. The heavier lines indicate the 1 ns rolling averages.

are contained in one of the following subsets of the translocon: (1) residues Lys^{2}-Val^{45} and Ile^{71}-Pro^{205} in the α-subunit, and the entire β-subunit (Figure 2.2A, gold), (2) residues Trp^{29}-Arg^{66} in the γ-subunit, which include the domain that forms the hinge for the translocon (Figure 2.2A, blue), and (3) residues Pro^{205}-Leu^{433} in the α-subunit, which include the TM6-10 (Figure 2.2A, red). Scaffolding interactions are also included between particles in subsets in 1 and 2 and between particles in subsets 2 and 3. However, they are not included between particles in subsets 1 and 3, and all scaffolding interactions are restricted to pairs of CG particles that are within 7 Å in the original mapping from the crystal structure [10]. The translocon scaffolding is designed to stabilize the CG simulations without biasing or hindering the LG or PP motions. The red curves of Figure 2.2C demonstrate that the scaffolding succeeds in stabilizing the structure of the translocon in long-timescale CG simulations, and the results presented in appendix A sections: *Scaffolding Contribution to the FE Profile* and *Trajectories Without Scaffolding* indicate that the scaffolding does not significantly alter the conformational landscape of the translocon.

2.2.3 Collective Variables and Free-Energy Calculations

The FE surface for the translocon is calculated as a function of collective variables that quantify opening of the LG, d_{LG}, and the displacement of the plug moiety from the channel pore, d_{PP},

$$F(d_{\mathrm{LG}}, d_{\mathrm{PP}}) = -k_{\mathrm{B}}T \ln P(d_{\mathrm{LG}}, d_{\mathrm{PP}}), \tag{2.1}$$

where k_{B} is Boltzmann's constant and $P(d_{\mathrm{LG}}, d_{\mathrm{PP}})$ is the equilibrium probability distribution for the collective variables at temperature T. The LG distance d_{LG} is defined as the

distance of minimum approach between the line of least-squares fitting for the α-carbons of the residues in the TM2b helix and the corresponding line for the TM7 helix. The PP distance collective variable d_{PP} is defined as the distance between the center of mass of the α-carbons for the residues that comprise the isoleucine ring of the channel and the center of mass of the α-carbons for the residues of the plug domain. Full details and illustrations of the collective variables are provided in appendix A section: *Collective Variables*. The crystal structure reported in Ref.10 exhibits collective variable values of $(d_{\mathrm{LG}}, d_{\mathrm{PP}})$ =(5.99 Å, 10.64 Å) in the atomistic representation.

The weighted histogram analysis method (WHAM) [26] in two dimensions was used to construct the FE surface from over 80 independent MD trajectories that were restrained to different reference values for the collective variables. These trajectories included the auxiliary restraining potential $\frac{1}{2}\kappa_{\mathrm{LG}}(d_{\mathrm{LG}}(\mathbf{x}) - d^{\circ}_{\mathrm{LG}})^2 + \frac{1}{2}\kappa_{\mathrm{PP}}(d_{\mathrm{PP}}(\mathbf{x}) - d^{\circ}_{\mathrm{PP}})^2$, where $\mathbf{x}$ is the set of Cartesian positions for the atoms, and where $\kappa_{\mathrm{LG}} = 15.0$ kcal/mol/Å^2 and $\kappa_{\mathrm{PP}} =$ 10.0 kcal/mol/Å^2. The restraint values for the collective variables formed a uniform 8×10 grid spanning $d^{\circ}_{\mathrm{LG}}/\text{Å} \in [6, 13]$ and $d^{\circ}_{\mathrm{PP}}/\text{Å} \in [11, 20]$. To achieve adequate sampling in the atomistic simulations, additional trajectories were performed with restraints of $(d^{\circ}_{\mathrm{LG}}/\text{Å}, d^{\circ}_{\mathrm{PP}}/\text{Å})$ = (8.5, 12), (8.5, 13), (8.5, 14), (8.5, 15), (8.5, 17), (8.5, 18), and (8.5, 19). Each restrained MD trajectory was run for a length of 2 ns in the atomistic model and 20 ns in the CG model. To minimize the equilibration time, restrained trajectories were initialized from trajectories performed with neighboring values of the restraint. A modified ridge estimator was used to smooth the calculated FE surfaces. Error estimates for the atomistic and CG free-energy profiles are provided in appendix A section: *Scaffolding Contribution to the*

Free-Energy Profile.

2.2.4 Atomistic and CG Free-Energy Surfaces

Figure 2.3A presents the FE surface calculated from the atomistic simulations of the Sec translocon. It reveals a simple conformational landscape with a single minimum located around the values for the collective variables corresponding to the experimental crystal structure. No metastable open conformations for the channel are found with regard to displacements in either the LG or PP distances. The FE surface supports the conclusion that the crystal structure captures the relevant conformation for the membrane-bound translocon, in agreement with previous MD simulations [27–30]. However, it also indicates that structural fluctuations in the translocon that are large enough to allow for either protein translocation or membrane integration are thermodynamically unfavorable. Given that an α-helical peptide is approximately 10-12 Å in diameter, Figure 2.3A suggests that a FE penalty in excess of 20 kcal/mol must be incurred for either the protein translocation or the membrane integration pathways in the absence of other facilitating interactions. Below, we consider the role of the substrate in shifting this FE landscape.

Figure 2.3A also reveals very little correlation between the opening of the LG and the displacement of the plug moiety. Indeed, as is shown appendix A section: *Free-Energy Surface Cross Sections*, cross sections of the FE surface at different fixed values of the PP distance are essentially identical. This suggests that the stabilization of the LG does not explicitly depend on the displacement of the plug moiety [31], at least according to this measure.

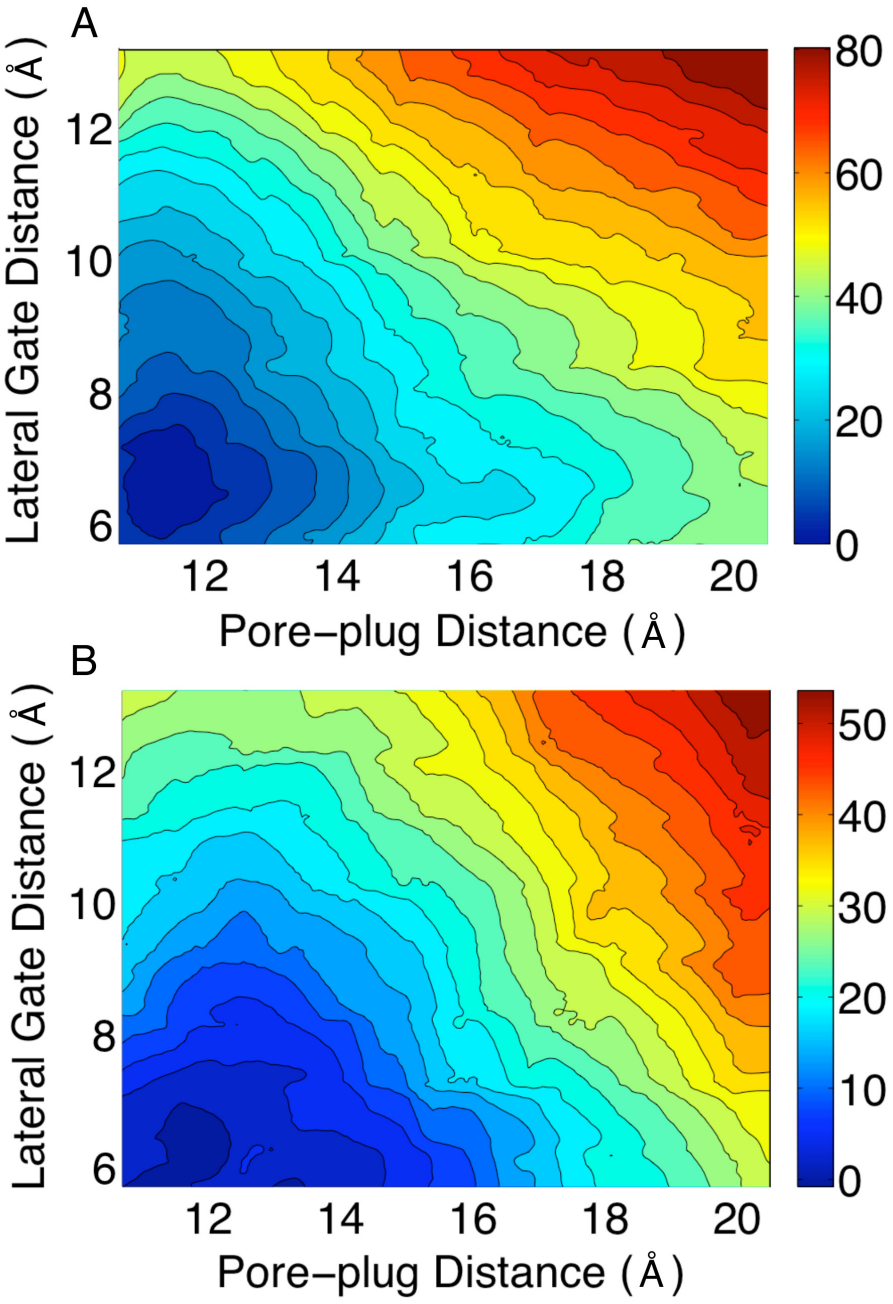


Figure 2.3: Free energy profiles for the Sec translocon from atomistic **(top)** and CG **(bottom)** simulations. Energies in kcal/mol.

Figure 2.3B presents the corresponding FE surface for the translocon from our CG simulations. Although the thermodynamic penalty for displacing the LG and the PP is reduced in the CG model, these results compare closely with those in Figure 2.3A, suggesting that the CG model and the scaffolding protocol reproduce the conformational landscape from the atomistic simulations. The effect of the scaffolding interactions on this calculation are discussed in appendix A section: *Scaffolding Contribution to the FE Profile*. Given the agreement between the atomistic and CG models, as well as the fact that the CG simulations increase the computational speed of the simulations by more than an order of magnitude, we employ the CG model for the remainder of the study.

2.3 Substrate Peptides Alter Translocon Conformation

2.3.1 Hydrophobic versus Hydrophilic Peptide Insertion

To investigate the influence of substrate peptides on the conformational landscape of the translocon, we consider the one-dimensional FE profile along the LG distance for the translocon containing either a hydrophobic polyleucine (Leu_{30}) peptide substrate or a hydrophilic polyglutamine (Gln_{30}) peptide substrate. The side chains for the leucine and glutamine residues occupy similar steric volumes [21], allowing the simulations to isolate the role of peptide hydrophobicity. The entire system, including the inserted peptides, are simulated using the CG protocol described previously. To prevent the diffusion of the peptides into the membrane, and thus to ensure a well-defined FE profile for the translocon containing the substrate peptide, a weak restraint potential was used to tether the center of

mass of the peptide to the center of mass of the channel pore residues. The details of the initialization protocol and simulations for the translocon-substrate system are provided in appendix A section: *Initializing the Peptide Substrate*. The WHAM algorithm was again employed to construct the FE profile from 9 independent trajectories for the translocon-substrate system that are harmonically restrained to different values for the LG distance on a uniform grid in the range $d^{\circ}_{\mathrm{LG}}/\text{Å} \in [7, 15]$ using $\kappa_{\mathrm{LG}} = 5.0$ kcal/mol/Å^2; to achieve adequate sampling, an additional trajectory was performed with the hydrophobic substrate at $d^{\circ}_{\mathrm{LG}} = 10.5$ Å and two additional trajectories were performed with the hydrophilic substrate at $d^{\circ}_{\mathrm{LG}} = 6$ and 9.5 Å. Each of the 21 sampling trajectories was run for a simulation time of 1.5-1.6 μs, where all but the last 800 ns was discarded as equilibration.

Figure 2.4 presents FE profiles calculated for the translocon peptide substrate. The black curve, for reference, presents the result for the translocon without peptide substrate and is consistent with the data presented in Figure 2.3B. These results demonstrate that the hydrophilic peptide shares the same basin of stability as the translocon in the absence of substrate, while an open conformation for the LG motion is stabilized for the translocon containing the hydrophobic substrate.

Recent structural studies have considered the role that translocon-docking macromolecules play in stabilizing the open LG; crystal structures with the Sec translocon in complex with SecA [12] or a Fab fragment [13] exhibit partial opening of the LG, whereas a recent, subnanometer-resolution microscopy study finds no such opening of the LG for the translocon docked with the ribosome [14]. The results in Figure 2.4 predict the hydrophobic substrate to stabilize the open LG, even in the absence of such complexation events.

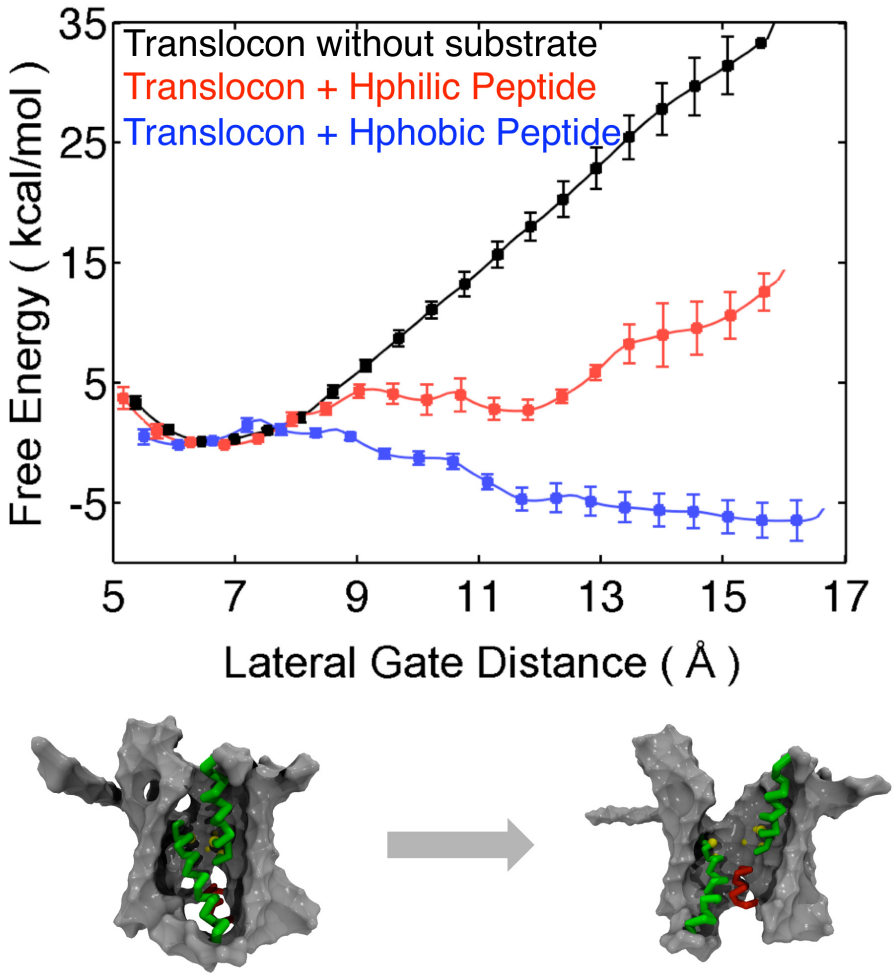


Figure 2.4: Free energy profiles along the LG distance for the translocon, with and without peptide substrates. Below, snapshots showing the translocon in closed versus open configurations of the LG distance.

To investigate the metastability of the LG conformations for the translocon containing the substrate peptides, long CG MD trajectories that were not restrained with respect to the LG distance nor the center of mass of the peptide substrate were initialized from open ($d_{\mathrm{LG}} = 15$ Å), partially open ($d_{\mathrm{LG}} = 11$ Å), and closed ($d_{\mathrm{LG}} = 6$ Å) configurations for the LG. They are plotted as a function of the LG distance in Figure 2.5A. For the trajectories initialized from the closed LG with either the hydrophobic and hydrophilic peptide substrate, the LG remained closed on the timescale of the simulations. However, for the

trajectories initialized from the open LG, the simulation with the hydrophobic substrate remains open, while the simulation with the hydrophilic substrate relaxes toward smaller values of d_{LG} on the timescale of hundreds of nanoseconds. Similarly, for the trajectories initialized from the partially open LG, the simulation with the hydrophobic substrate relaxes toward larger values of d_{LG}, while the simulation with the hydrophilic substrate exhibits gradual closure of the LG over the course of the trajectory. Additional trajectories performed without scaffolding interactions are reported in appendix A section: *Trajectories Without Scaffolding*.

The trajectories in Figure 2.5A for the initially open and partially open LG with the hydrophilic peptide relax toward the closed configurations, but they do not fully close the LG distance within the 500 ns of simulation time. This slow timescale for relaxation is related to the conformation of the plug moiety for the translocon. As is illustrated in Figure 2.5B, the translocon in these simulations has in fact eliminated the open surface area of the LG (defined in appendix A section: *Collective Variables*) , but this is not captured by the LG distance collective variable that is plotted in Figure 2.5A. It is not clear whether the ability of the plug to partially prop open the region between TM2b and TM7 at the bottom of the translocon channel is functionally relevant, although it is thought that a peptide signal sequence performs this function at the top of the channel [10, 32].

The results in Figure 2.5B indicates that a metastable closed state for the LG of the translocon is supported by both the hydrophilic and hydrophobic substrate with a surface area of approximately 400-450 $\mathrm{\AA}^2$, whereas a metastable open state for the LG is supported only by the hydrophobic substrate with a surface area of 600-650 $\mathrm{\AA}^2$. The closed state al-

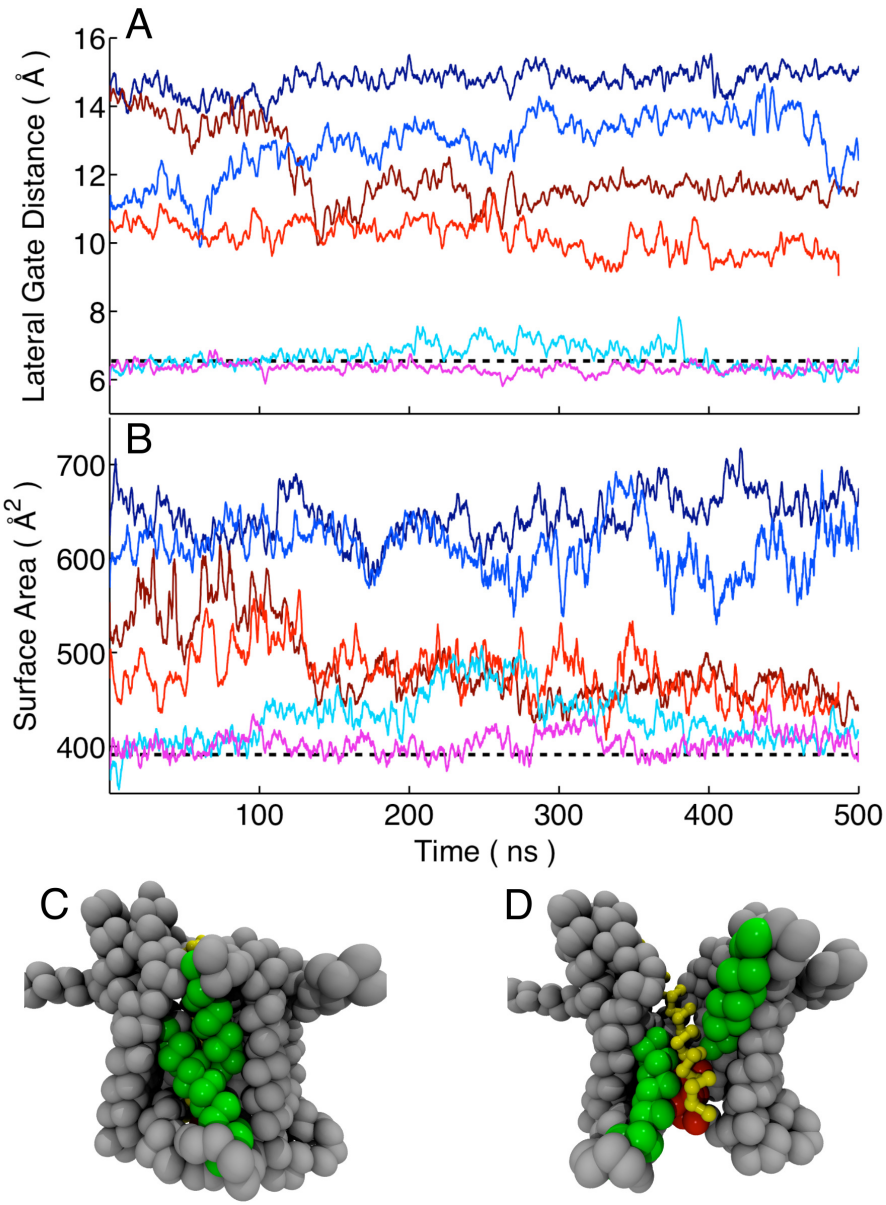


Figure 2.5: CG MD trajectories for the translocon containing either the hydrophobic (blue-shaded) or hydrophilic (red-shaded) substrate are initialized from open, partially open, and closed configurations of the LG. **(A)** The LG distance d_{LG} for the trajectories is plotted as a function of simulation time. **(B)** The LG surface area for the trajectories is plotted as a function of simulation time. Heavy lines indicate 1 ns rolling averages. Also shown are snapshots of the translocon at the end of the initially closed trajectory with hydrophilic substrate **(C)** and the initially open trajectory with hydrophobic substrate **(D)**.

lows for little contact of the peptide substrate with the hydrophobic membrane interior and exhibits values for the LG surface area and distance values that are typical of the translocon without substrate, whereas the open state allows for extensive exposure of the substrate to the membrane and provides space for the exit of the substrate from the channel (Figures 2.5C and D). Although the fact that the closed state is metastable for both the strongly hydrophobic and strongly hydrophilic substrates suggests that the closed state would also be metastable for substrates of intermediate hydrophobicity, additional trajectories provided in appendix A section: *Trajectories with Substrate of Intermediate Hydrophobicity* show this explicitly.

2.3.2 Orientation of the Substrate Peptide and the Translocon Plug

The reason that long ($>$1.5 μs) sampling trajectories were needed to equilibrate the FE profile in Figure 2.4 is due to the slow reorientation of the channel plug moiety with respect to the peptide substrate (Section *Initializing the Peptide Substrate* in appendix A). Figure 2.6A illustrates that for the hydrophilic substrate, the plug is preferentially positioned between the peptide substrate and the LG, whereas for the hydrophobic substrate, the orientation is reversed such that the plug is behind the substrate with respect to the LG opening.

To quantify this effect, we introduce an order parameter for the relative orientation of the pore and the plug residue. We define θ to be the angle between a vector v_1 that points from the peptide substrate to the plug moiety and a vector v_2 that points outward from the opening of the LG. If $\cos(\theta) > 0$, then the plug is between peptide and the LG, as is shown for the snapshot of the hydrophilic peptide in Figure 2.6A. For $\cos(\theta) < 0$, the reverse

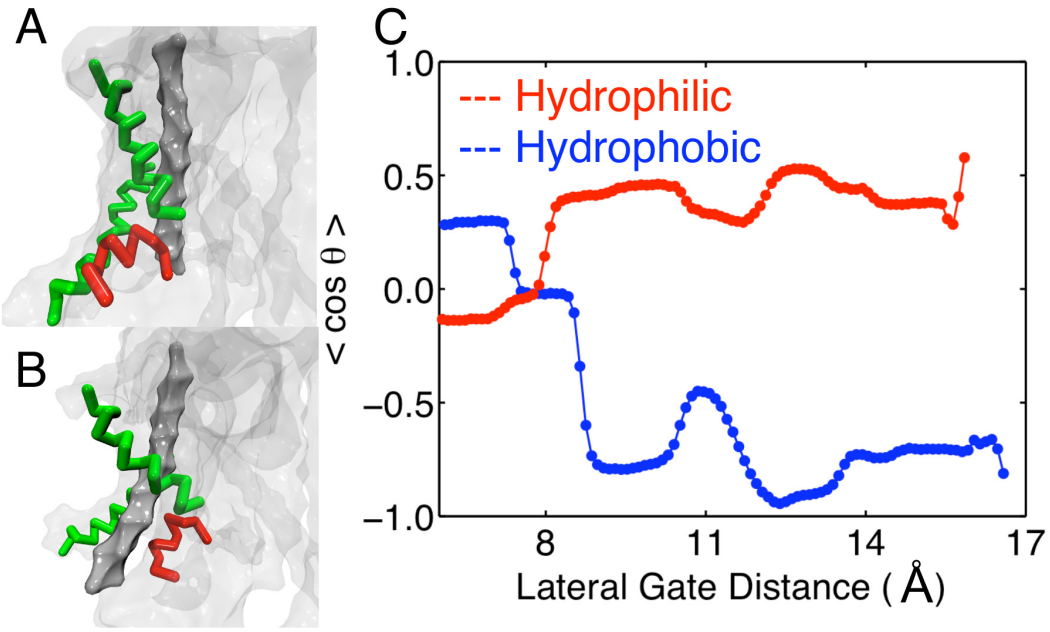


Figure 2.6: Relative orientation of the peptide substrate (dark grey), the plug moiety (red), and the LG helices (green). **(A)** and **(B)** Snapshots illustrating that the hydrophilic peptide (A) is behind the plug residue with respect to the LG, whereas the hydrophobic peptide (B) is in front of the plug residue and closer to the interior of the membrane. **(C)** The ensemble average for the order parameter describing the relative orientation of the substrate and plug.

orientation is observed. A detailed and illustrated definition of θ is provided in appendix A section: *Collective Variables*.

Figure 2.6C presents the equilibrium expectation value for the orientational order parameter $\cos(\theta)$ as a function of the LG distance, again obtained using the WHAM algorithm and the simulation data corresponding to Figure 2.4. Indeed, this plot reveals that for open configurations of the LG, the relative orientation of the plug moiety and the peptide substrate is strongly dependent on the nature of the peptide. The trend observed in Figure 2.6C indicates that the hydrophobic substrate assumes an orientation that achieves greater exposure to the hydrophobic lipids of the membrane interior, whereas the hydrophilic substrate favors the orientation in which it remains more fully in the channel and shielded from the membrane by the plug. This result may suggest that the plug (or its replacement moiety

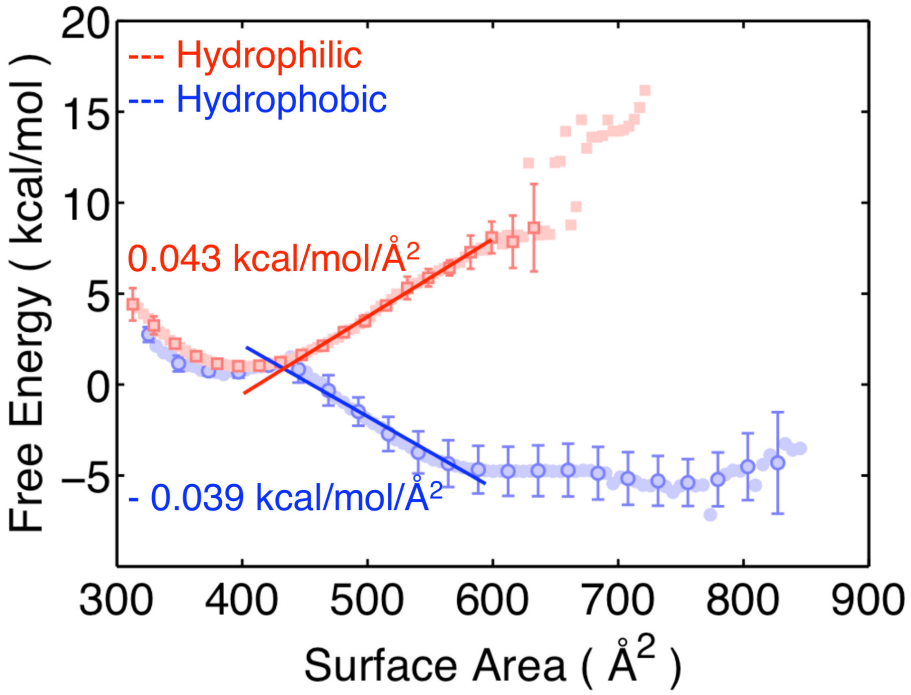


Figure 2.7: Free energy profiles for the translocon with peptide substrates as a function of the LG surface area. For the case of the hydrophilic peptide, the slope for the linear fit of the data is shown.

in a plug-deletion mutant of the translocon [11]) plays a role in guiding the substrate toward either the translocation or membrane integration pathways.

2.3.3 Hydrophobicity and the Energetics of the Lateral Gate

To analyze the energetics of the LG motion for the translocon including peptide substrates, we calculate the FE profile for these systems as a function of the LG surface area. This calculation again employs the WHAM algorithm and the simulation data corresponding to Figure 2.4. A detailed definition of the LG surface area collective variable, which quantifies the area between the TM2b and TM7 helices that comprise the LG, is provided in appendix A section: *Collective Variables*. The FE profiles calculated as a function of the LG surface area are presented in Figure 2.7. Closed configurations for the LG correspond to a surface

area of approximately 400-450 Å^2 (Figure 2.5B). As the LG opens, the linear behavior for the FE profiles is consistent with a model, $F = \sigma A$, in which the FE of opening the LG, F, is equal to the product of the LG surface area, A, and a constant marginal FE, σ. Linear fits to the FE profiles in the range of 450-600 Å^2 and the resulting estimates for σ are also included in the figure; the fitting range is chosen based on the characteristic values for the LG surface area for the closed and open states of the LG observed in Figure 2.5B.

Figure 2.7 suggests that the energetics of the LG conformation is governed by a simple balance between hydrophobic and hydrophilic contacts in the system. For the case of the hydrophilic substrate, the opening of the translocon LG corresponds to the formation of an interface between the hydrophobic interior of the membrane and the hydrophilic substrate in the channel. It is thus reasonable that the calculated value of $\sigma = 0.04$ kcal/mol/Å^2 for this case is similar to the range of values (0.025-0.035 kcal/mol/Å^2) that have been estimated for the surface tension between hydrophilic residues and the lipid bilayer [33, 34]. On the other hand, the opening of the LG for the case of the hydrophobic substrate thus corresponds to the *removal* of a hydrophobic-hydrophilic interface from the system. The interior of the channel for the Sec translocon is a largely hydrophilic environment [10, 35, 36], such that opening the LG replaces an area of hydrophobic-hydrophilic contacts between the substrate and the channel interior with more favorable hydrophobic-hydrophobic contacts between the substrate and the membrane.

For larger values of the LG surface area, the FE profile for the hydrophobic substrate deviates from the linear fit as expected. Once the surface area is sufficiently large to allow for full contact between the hydrophobic substrate and the membrane (600-650 Å^2, Figure

2.5B), further opening of the LG does not allow for any additional favorable hydrophobic contacts; it instead introduces contacts between the membrane and the hydrophilic interior of the channel, leading to a change in the marginal FE and the calculated turnover in the FE profile.

The agreement of the simulation data in Figure 2.7 with the expression $F = \sigma A$ suggests that the relative FE between the metastable LG closed state (approximately 400 Å^2) and the open configurations that allow for membrane integration (approximately 600 Å^2) is governed by the interfacial energy between the peptide substrate and the membrane interior. This FE relationship depends linearly on both the LG surface area and the hydrophobicity of the peptide substrate, such that if it is assumed that the detailed sequence of residues in the substrate can be ignored [6], then the relative FE between the closed and open states depends simply on the number of hydrophobic and hydrophilic peptides in the substrate. Given that the simulation data indicates that changing from a completely hydrophobic 30-residue peptide to one that is completely hydrophilic alters the relative FE of LG opening by approximately 12 kcal/mol, this analysis suggests that replacing a single hydrophobic residue in the substrate with a hydrophobic residue will lead to a change in the FE of LG opening of approximately 0.6 k_BT. It follows that for substrates of intermediate hydrophobicity, the thermodynamic balance between open and closed LG states can be significantly shifted by changing only a small number of substrate residues.

Naturally, the simplicity of the CG model employed should discourage overinterpretation of the quantitative details of the simulation results, and a more extended discussion of the accuracy of the CG model is provided in appendix A section: *Side-Chain Transfer Free*

Energies for the CG Residues. However, the energy scales obtained from the simulations reported here and the linear dependence of the calculated LG FE on the LG surface area and support a simple and intuitive analysis of the energetics of the translocon LG in the presence of a peptide substrate. Section *Mutations in the Translocon Pore Residues* in appendix A further discusses how the free energy of LG opening depends on the hydrophobicity and bulkiness of the translocon pore residues.

2.4 Implications for Translocon Regulatory Function

Efforts to understand the regulatory function of the Sec translocon have focused on the strong correlation between the water/octanol transfer FE for a TM peptide domain and the relative fraction of substrate peptides that undergo membrane integration vs. translocation [4–6]. This correlation has been interpreted in terms of a two-state model in which the peptide equilibrates between the hydrophobic membrane environment and the hydrophilic channel environment [1, 4]. Assuming that this equilibrium is genuinely realized (and assuming that the solvation FE for the peptide in the channel does not change with the predominant LG conformation), then the partitioning of the peptide substrate between the translocon and membrane environments would be independent of the LG conformation. That is, switching of the predominant LG conformation between open and closed states under the control of the substrate hydrophobicity (Figure 2.7) would not affect the regulation of the substrate peptides between the Sec-mediated membrane insertion and translocation pathways.

However, if instead of being completely reversible, the exit of the peptide substrate

from the translocon is irreversible, then the results presented here suggest an alternative interpretation of the data of Hessa *et al.* [5, 6]. Assuming that only the open LG allows for membrane integration, and utilizing the separation between the timescale within which substrates are driven into the channel by either the ribosome or another molecular motor ($\sim$milliseconds) and the timescale within which the LG undergoes conformational rearrangements ($\sim$100 ns), then the rate at which the integration product is formed is proportional to the population of the open LG conformation, $k_{integ} \propto P_{open}$. Similarly, the rate for the translocation is proportional to the population of the closed LG conformation, $k_{trans} \propto P_{closed}$. The balance of this conformational partitioning of the LG, as we have argued in connection with Figure 2.7, depends primarily on the effective hydrophobicity of the substrate peptide.

This model describes a regulation between the translocation and integration pathways that is controlled by substrate-sensitive conformational gating of the translocon. It is consistent with the experimental observation of a two-state balance between translocation and integration, and it predicts the experimentally observed correlation of that balance with peptide hydrophobicity. The model is based on a nonequilibrium description of the slow substrate insertion dynamics and an equilibrium description of the faster conformational motions of the translocon; since both open and closed states have finite equilibrium populations, all substrates will experience at least fleeting exposure to the interior of the membrane [1]. Direct, nonequilibrium simulations of protein translocation and membrane integration will yield further insights into this possible mechanism of regulation.

Chapter 3

Direct Simulation of Early Stage Sec-Facilitated Protein Translocation

3.1 Introduction

In the previous chapter, we investigated the conformational landscape of the Sec translocon, and the role of peptide hydrophobicity in regulating protein translocation versus membrane integration using enhanced thermodynamic sampling methods. However, many important dynamical aspects of this cellular machinery remain elusive. Particularly little is known about the dynamics of the translocon and nascent protein during the earliest stages of protein translocation, a critical period in the regulation of protein secretion, membrane integration, and membrane protein topology. Outstanding questions relate to nascent-protein conformations that are visited in the early stages of insertion, molecular mechanisms that connect features of the translocon and nascent-protein residues to its targeted destination and topology, and the initiation of nascent-protein secondary structure. These issues are difficult to experimentally probe because they involve transient interactions and processes, confined molecular environments, and membrane-bound complexes that create challenges for high-resolution approaches. Traditional atomistic simulation techniques employed in

the previous chapter also fail in addressing these issues because of the long timescale and large lengthscale involved in the underlying processes.

In this chapter, we introduce a protocol for directly modeling the dynamics of nascent-protein insertion into the translocon, and we leverage the specialized Anton computing system [37, 38] to perform microsecond-timescale simulations of early-stage protein translocation and membrane integration. The reported simulations, although short in comparison to second-minute timescales of the biological process, are nonetheless extremely long by the standards of state-of-the-art MD studies and provide a powerful exploratory tool for investigating the early-stage dynamics of nascent-protein insertion into the translocon. Insertion of both hydrophobic and hydrophilic nascent-protein domains is modeled, and quantitative metrics are employed to characterize nascent-protein and translocon conformational changes, the formation of salt bridges and specific interactions, and the development of large-lengthscale hydrophobic contacts. These simulations, when interpreted in combination with experimental studies and previous nanosecond-timescale MD simulation studies of the translocon [27–30, 39, 40], offer new insights into the mechanistic details of Sec-facilitated protein translocation and membrane integration.

3.2 Methods

Using microsecond-timescale MD trajectories with more than 120,000 atoms, we explicitly model the insertion of nascent-protein residues into the *Thermotoga maritima* SecYEG channel via the SecA ATPase molecular motor (Figure 3.1). The structure of the SecA-SecYEG complex has been obtained via crystallography [12]. The all-atom simulation

cell (Figure 3.1, middle panel) includes explicit solvent, counterions, and 222 POPC lipid molecules. Long MD trajectories are performed using the special-purpose Anton computing system [37, 38]; to meet the system-size limitations of the Anton hardware, SecA is truncated at a distance of 15 Å from the translocon, and the heavy atoms of SecA are harmonically restrained to their corresponding positions in the crystal structure. The residues of SecYEG and all other atoms in the system are unrestrained. Full details of the simulations are described in appendix B.

We introduce a nonequilibrium simulation protocol to describe the SecA-driven insertion of nascent-protein residues into the translocon channel. The novel protocol includes nanosecond-timescale growth of the nascent amino acid chain at the cytosolic mouth of the translocon followed by microsecond-timescale evolution of the system (Figure 3.1, bottom panel). Long MD trajectories are performed using the special-purpose Anton computing system [37, 38] under conditions of constant pressure and temperature. As is emphasized below, the presented insertion protocol describes necessary features of SecA-driven nascent-protein insertion, including the sequential introduction of nascent-protein residues and molecular confinement at the cytosolic mouth of the translocon channel; however, the detailed mechanism of the SecA driving force remains an open question [3, 41–43].

The key features of the insertion protocol are as follows. After initial equilibration of the SecA-SecYEG complex, a nascent protein composed of $n+4$ amino acid residues is introduced, with the four residues at the N-terminus aligned with the axis of the translocon channel and with the center-of-mass position of the remaining n residues placed at an “insertion point" at the cytosolic mouth of the translocon channel. Each of the n residues

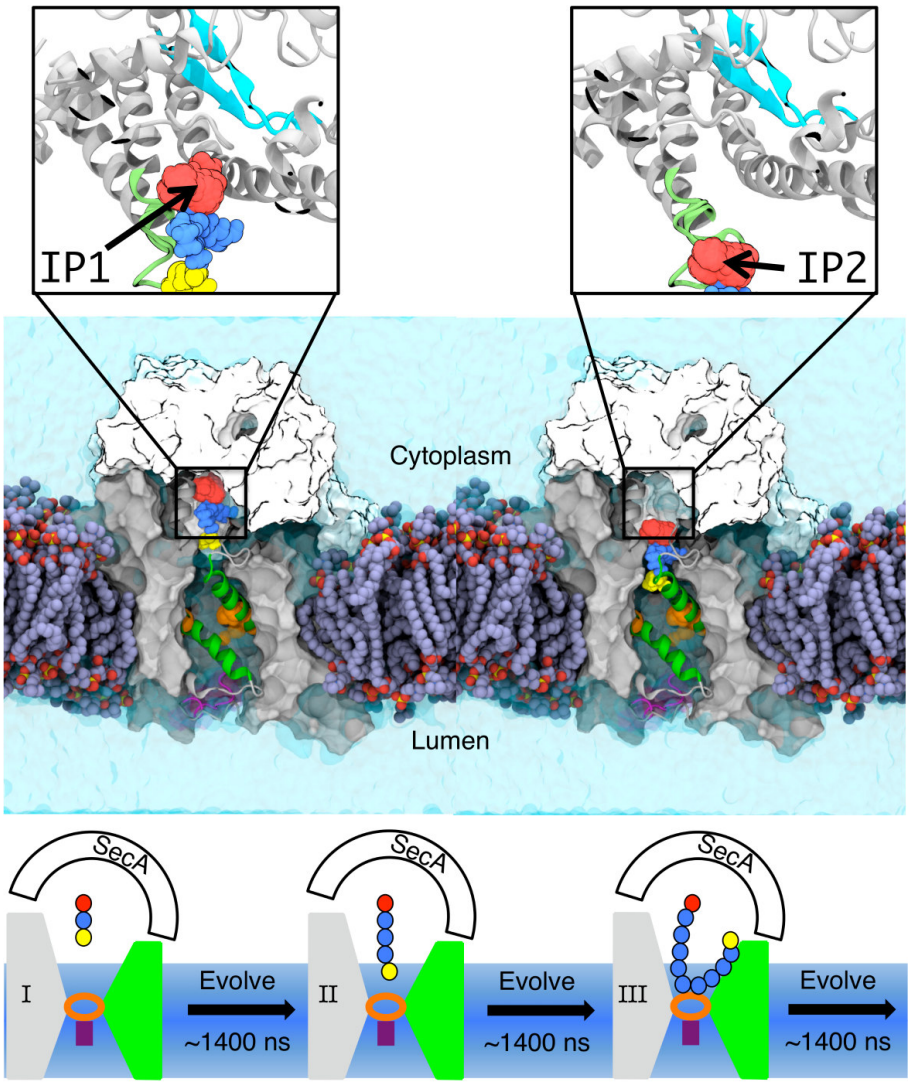


Figure 3.1: Early-stage protein insertion into the Sec translocon. **(Middle Panel)** The all-atom system employed in the MD simulations, including the translocon (gray surface, with the helices TM2b and TM7 in green, the pore residues in orange, and the plug moiety in violet), the truncated SecA protein (white surface), and the nascent protein undergoing insertion (yellow, blue, red). **(Top Panel)** Expanded view of the interface region between SecA and the translocon, with the two nascent-protein insertion points (IP1 and IP2) indicated. The SecA two-helix-finger and β-sheet domains are indicated in light green and light blue, respectively, and the nascent-protein residues are presented with same color scheme as in the middle panel. **(Bottom Panel)** Schematic illustration of the simulation protocol used to model nonequilibrium protein insertion. The translocon is shown in gray, with the LG region indicated in green, the pore residues in orange, and the plug moiety in violet. SecA is shown in white, and the nascent-protein sequence is shown using the same coloring scheme as in the middle panel. Each period of nascent-protein growth is followed by a microsecond-timescale trajectory at fixed protein length.

Table 3.1: Summary of the insertion trajectories. [a]

Trajectory	IP	Mature Domain	Evolution period (μs) #1	#2	#3	Total Length (μs)
T_1	1	L_{30}	0.90	1.54	1.41	4.05
T_2	1	Q_{30}	0.90	1.67	1.41	4.18
T_3	2	L_{30}	1.38	1.50	1.80	4.88
T_4	2	Q_{30}	1.38	1.50	1.80	4.88

[a] In each trajectory, the first growth period spans 0.055 μs, whereas the latter two span 0.075 μs.

on the C-terminal end of the nascent protein exists in either an off-state, in which non-bonding interactions between each residue and the rest of the system are excluded, or an on-state, in which all interactions are included; residues in the off-state are tethered to the insertion point via harmonic restraints. We model the SecA-driven protein insertion using a repeated two-step cycle composed of *(i)* growth, in which residues at the C-terminal end of the nascent protein are sequentially switched from the off-state to the on-state at a pace of one residue per five nanoseconds, and *(ii)* evolution, in which the system is evolved without growth using standard MD. The simulations presented in this chapter include three growth/evolution cycles (Figure 3.1, bottom panel), with each growth period leading to the insertion of fifteen new protein residues, followed by an evolution period of $0.9-1.8$ μs in time (Table 3.1). The appendix B provides full details of the simulation protocol employed for the nascent-protein growth period.

This insertion protocol is used to obtain four microsecond-timescale simulations of early-stage protein translocation and membrane integration via the translocon. Each of the simulated insertion trajectories (T_1-T_4) employs one of two different insertion points and one of two different nascent-protein sequences (Table 3.1). The nascent-protein sequence is composed of a hydrophobic, N-terminal signal peptide (SP) and a C-terminal mature

domain sequence. In all cases, the SP sequence ($MGPRL_{11}$, with residues listed from the N-terminus) matches the 15-residue synthetic SP that was employed by Spiess and co-workers for the investigation of integral membrane protein topogenesis [7]; the C-terminal mature domain for the nascent protein is comprised of either a purely hydrophilic 30-mer of glutamine (Q_{30}) or a purely hydrophobic 30-mer of leucine (L_{30}). The first insertion point (IP1) employed in these simulations is positioned close to the highly conserved β-sheet that connects the NBD1 and PPXD domains of SecA and that is thought to be the binding site for the nascent protein [44] (Figure 3.1, top panel); the second insertion point (IP2) is positioned close to the loop of the two-helix-finger domain of SecA, which has been suggested to mechanically push the nascent protein through the translocon [12, 42]. Exact coordinates for the insertion points are provided in the appendix B.

Although it is generally agreed that SecA utilizes ATP hydrolysis to drive the nascent protein across the translocon channel [45, 46], questions remain regarding the details of this process, including the oligomeric state of SecA [47–49], the nature of SecA conformational changes that generate the driving force for nascent-protein insertion [12, 44, 50, 51], and the exact roles of ATP binding and hydrolysis events [41, 43, 52, 53]. The goal of the current study is not to investigate the detailed mechanism of the SecA motor action; rather, we aim to characterize the conformational dynamics and mechanisms associated with nascent-protein insertion into the translocon. We thus model only the most fundamental roles of SecA in the insertion process: providing confinement of the nascent protein at the cytosolic mouth of the translocon channel and enforcing sequential insertion of the nascent protein into the translocon. Although the extent to which this simplification impacts any

conclusions about posttranslational protein translocation is difficult to assess without a better understanding of the SecA mechanism, we note that cotranslational (ribosome-driven) nascent-protein insertion does not involve explicit coupling of a molecular motor to conformational changes in the translocon [54]; the insertion protocol employed here is thus at least relevant for the cotranslational pathway. Furthermore, the fundamental issues that are the focus of this study, including the conformational dynamics of the nascent protein and translocon, are expected to arise in all biological pathways for Sec-facilitated protein translocation [55, 56].

3.3 Results and Discussion

3.3.1 Translocon Conformational Response

The insertion simulations reveal mechanistic features of both early-stage protein translocation and membrane integration. Figure 3.2 presents snapshots of trajectories T_1 and T_2, respectively, at various times during protein insertion. The system is viewed from the perspective of the lipid bilayer, with the translocon LG helices (TM2b and TM7) in green, the pore residues in orange, and the plug moiety in red. Prior to the introduction of the mature domain residues at 0.9 μs, the trajectories are identical, exhibiting configurations for which the two LG helices are in close proximity. At longer insertion times, it is seen in both trajectories that hydrophobic residues (blue) of the nascent protein localize at the translocon LG helices, which undergo significant separation.

Figure 3.3 quantifies the LG conformational changes as a function of simulation time.

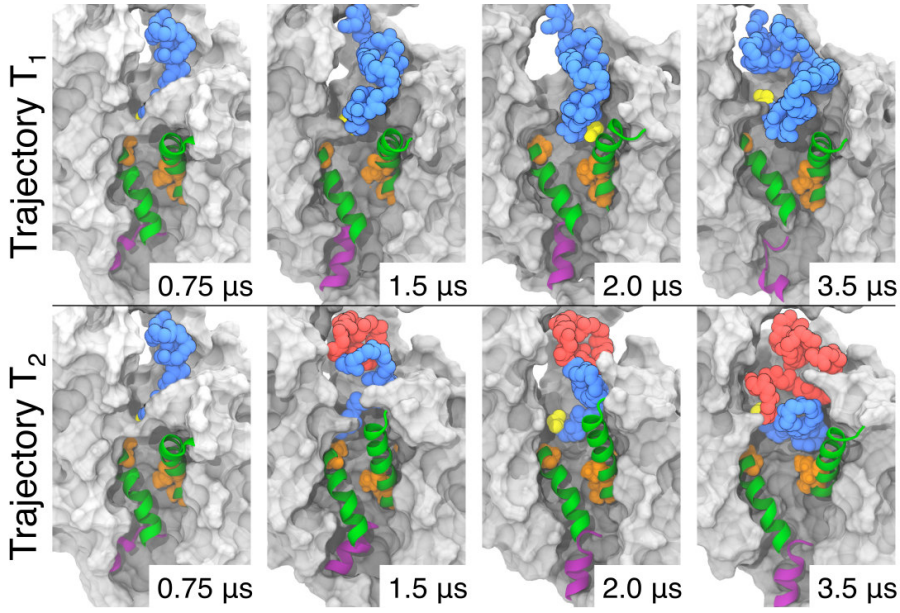


Figure 3.2: Structural features of the nascent protein and translocon at various times along the insertion trajectories T_1 and T_2. The translocon is shown in gray surface, with the two LG helices in green, the pore residues in orange and the plug moiety in violet. The nascent-protein SP and the hydrophobic mature domain of the nascent protein are colored in blue, while the hydrophilic mature domain is colored in red.

The LG is characterized in terms of its width profile along the channel axis, η, which lies perpendicular to the plane of the membrane. As is illustrated in Figure 3.3A, the LG width profile measures the minimum horizontal distance between helices TM7 and TM8 (green) on one side of the translocon LG and helices TM2b and TM3 (yellow) on the other; detailed definitions for both the channel axis and the LG width profile in terms of the molecular coordinates are provided in appendix B. Figure 3.3B presents the LG width profile obtained at various times during trajectories T_1 and T_2. The two profiles coincide for the initial stages of insertion ($t = 0.5\ \mu$s), with the LG width narrowing in the region of the translocon pore residues ($\eta \approx -18$) and widening at the cytosolic ($\eta > -15$) and lumenal ($\eta < -20$) openings. Significant changes in the width profiles for these two trajectories emerge at longer times. Comparison of the width profiles for trajectory T_2 (red)

at $t = 2.0$ μs and $t = 3.5$ μs reveals nearly uniform widening of the LG along the channel axis, including the region of the pore residues. Trajectory T_1 (blue) also shows extensive widening of the translocon channel at the cytosolic opening, although it is accompanied by contraction of the LG width in the regions of the pore residues and the lumenal opening. For all insertion times, Figure 3.3C presents the difference between the channel width profile for trajectories T_1 and T_2, which emphasizes the differing extent to which the LG opens during nascent-protein insertion.

Figure 3.4 provides insight into the mechanistic basis for LG opening in Figure 3.3. Figures 3.4A and 3.4B present snapshots from insertion trajectories T_1 and T_2, respectively, after $t = 3.5$ μs of simulation time; the simulation cell is viewed along the channel axis from the cytosolic side of the membrane, and the density field of the membrane lipid tails is shown in grayscale. The density field represents the number density of heavy atoms in the hydrophobic lipid tails projected onto the *x*-*y* plane of the simulation cell; it is plotted with Gaussian smoothing on a lengthscale of 2 Å. For clarity, only the LG residues of the translocon (green) and the residues of the nascent protein (blue, red) are shown explicitly, and the set of points at a distance of 18 Å from the channel axis are indicated (orange). In both trajectories, the nascent protein is localized in the region of the LG helices, with hydrophobic residues (blue) in close contact with the hydrophobic lipid tails. The more hydrophobic nascent protein (Figure 3.4A) partially exits the translocon channel in favor of the membrane interior.

To quantify the relative degrees to which trajectories T_1 and T_2 exhibit membrane integration, Figure 3.4C plots the number of nascent-protein residues, $\mathcal{N}$, that exit the translo-

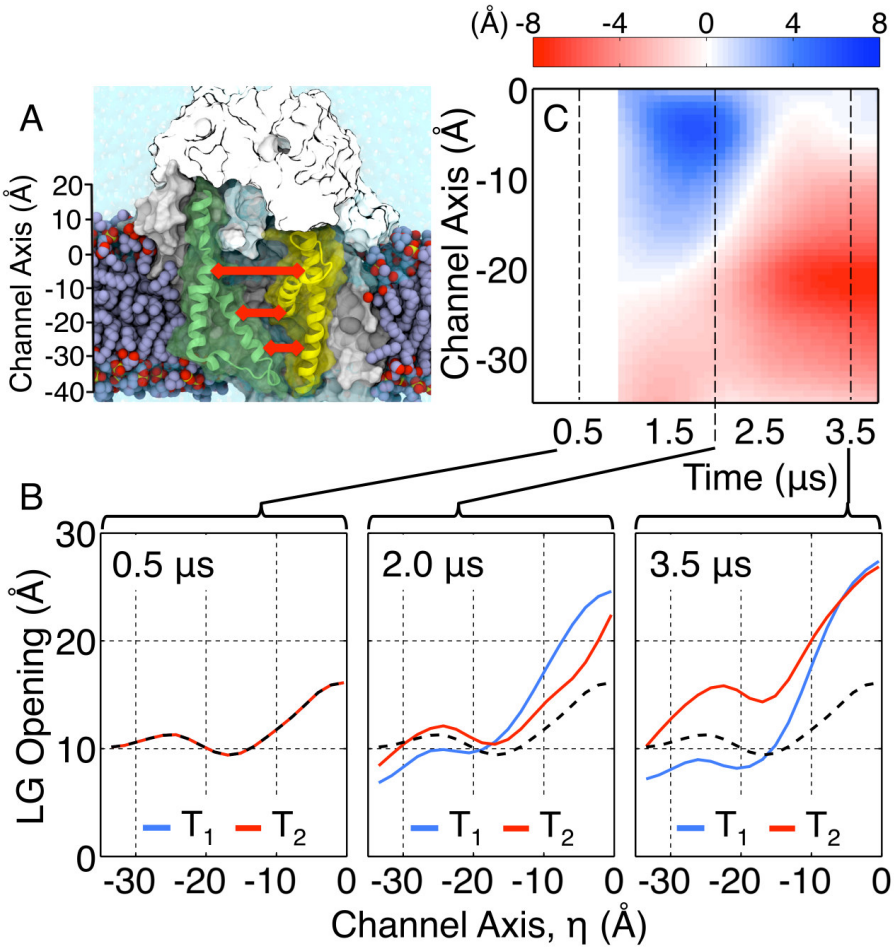


Figure 3.3: Translocon LG width profiles along trajectories T_1 and T_2. **(A)** Illustration of the LG width profile, which is indicated with red arrows. The coordinate associated with the channel axis is indicated at left. **(B)** The LG width profiles for trajectories T_1 (blue) and T_2 (red) at various times. The data at time $0.5\,\mu s$ is repeated in the dashed black curve. **(C)** The difference in the LG width profiles between trajectories T_1 and T_2.

con channel as a function of simulation time; specifically, the figure reports the number of residues for which the corresponding α-carbon lies beyond 18 Å from the channel axis (indicated in orange in Figures 3.4A and 3.4B) and falls between -30 and 0 Å along the channel axis (indicated in Figure 3.3A). Markedly different behaviors are seen for the two trajectories, with insertion of the more hydrophobic peptide leading to a much greater degree of membrane integration.

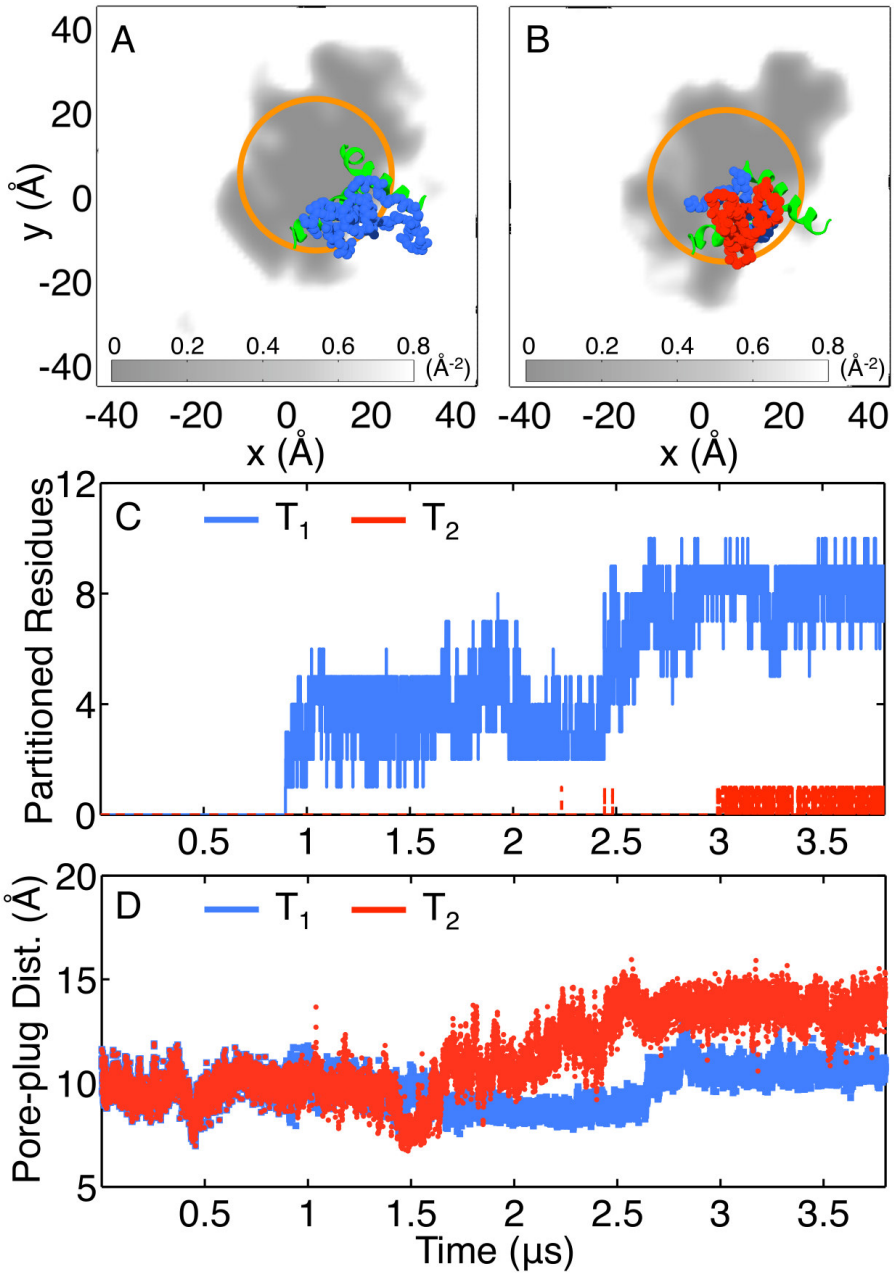


Figure 3.4: Early stage membrane integration. **(A, B)** Representative configurations from trajectories T_1 and T_2 after $t = 3.5$ μs of simulation time. The nascent-protein residues (hydrophobic in blue, hydrophilic in red) and the translocon LG helices (green) are shown in atomistic detail. The density field for the hydrophobic lipid tails is projected onto the x-y plane, with gray indicating low density and white for high density. The orange circles indicate positions that are 18 Å from the center of the channel axis. **(C)** The time evolution of the number of nascent-protein residues, $\mathcal{N}$, that partition into the membrane during the insertion simulations. **(D)** The time evolution of the pore-plug distance in the insertion simulations.

For trajectories T_1 and T_2, the differences in the channel width profiles seen in Figure 3.3 correlate with the differing degree of membrane integration in Figure 3.4. For both trajectories, the hydrophobic SP binds at the LG, leading to partial opening, as was predicted in earlier free energy calculations of the translocon conformational landscape in the presence of a hydrophobic substrate [40]. In trajectory T_1, the more hydrophobic nascent protein partitions directly from the cytosolic region of the channel, without inducing any widening of the LG in the channel pore or lumenal regions (Figure 3.3, blue). In trajectory T_2, the hydrophilic nascent protein does not partition into the membrane interior and instead remains localized in the translocon channel; to accommodate the volume of the growing protein, the LG widens along the entire channel axis, including the region of the pore residues (Figure 3.3, red).

Finally, Figure 3.4D plots the structural response of the translocon plug moiety during the insertion trajectories, with the distance between the pore and plug residues plotted as a function of simulation time. This distance measures the minimal separation between the α-carbon atoms of the six residues of the translocon pore (residues 82, 86, 187, 191, 274, 396) and the α-carbon atoms for the residues in the plug moiety (residues 65-74). For trajectory T_1, the results show little change in the pore-plug distance, despite the significant degree of membrane integration observed.

The insertion trajectories presented in Figures 3.3 and 3.4 exhibit important mechanistic features of early stage membrane integration (trajectory T_1) and protein translocation (trajectory T_2). The corresponding analysis of trajectories T_3 and T_4 reveals similar mechanistic features (Figures B.6, B.8 and B.9). Although it is important to avoid overinterpret-

ing the small number of illustrative MD trajectories presented here, these long-timescale simulations nonetheless reveal details of the conformational changes that are central to regulation of stop-transfer efficiency in Sec-facilitated protein translocation. In particular, we note that the results in Figures 3.3 and 3.4 are consistent with the observation that cross-linking of the LG helices inhibits the protein translocation pathway [16], since trajectory T_2 exhibits significant opening of the LG. Figure 3.4D is also consistent with experimental evidence that the conformation of the plug moiety is not significantly altered upon membrane integration [57], as well as the observation that deletion of the plug moiety has little effect on stop-transfer efficiency [58]. In addition to finding little movement of the plug upon membrane integration, Ref. 57 reports more significant displacement of the plug during protein translocation; we note that Figure 3.4D also shows a greater degree of displacement of the plug moiety for trajectory T_2 than for trajectory T_1, although our simulations probe stages of the protein translocation that are too early to exhibit the full degree of plug displacement.

3.3.2 Nascent-Protein Hydrophobic Contacts

In addition to its role in the regulation of stop-transfer efficiency, the Sec translocon influences the orientation, or topology, of integral membrane proteins. One such effect is that increasing SP hydrophobicity leads to a diminished fraction of proteins that undergo integration in the Type II orientation [7, 59–61], suggesting that hydrophobic contacts involving the nascent protein play a role in regulating integral membrane protein topogenesis [8, 18]. Here, we explore this effect by characterizing the degree to which the insertion

simulations exhibit hydrophobic contacts that stabilize nascent-protein configurations that are consistent with the early stages of either Type II or Type III membrane integration.

Figure 3.5 illustrates the nascent-protein conformational dynamics that accompany early-stage membrane integration. Figures 3.5A-D present snapshots of the trajectories T_1 and T_3 after 3.5 μs of simulation, with parts A and B showing the configuration of the SP relative to the translocon LG and parts C and D characterizing the solvation environment of the SP. The corresponding results for trajectories T_2 and T_4 lead to similar conclusions and are presented in Figure B.10.

Figures 3.5A and 3.5B illustrate strikingly different configurations for the nascent protein following early-stage insertion into the translocon. In both cases, the SP intercalates between the two LG helices. However, in part A, the SP adopts a partially helical conformation with the N-terminus buried inside the translocon channel. This orientation of the nascent protein enables the hydrophobic residues of the SP to extend across the LG and to make contact with the membrane hydrophobic lipid tails. In part B, the SP remains disordered, with the charged N-terminus exposed to the lipid phosphate head groups; the remainder of the SP occupies the interior of the translocon, and the LG helices widen to a lesser degree than in part A. The nascent protein in part B adopts a looped configuration that has been anticipated for early-stage Type II membrane integration [62]; in contrast, the buried N-terminal configuration for the nascent protein in part A is more consistent with the early stages of Type III membrane integration [18].

Figures 3.5C and 3.5D present the nascent-protein solvation environment for these two configurations, including the density of water molecules (light blue) within 8 Å of the SP

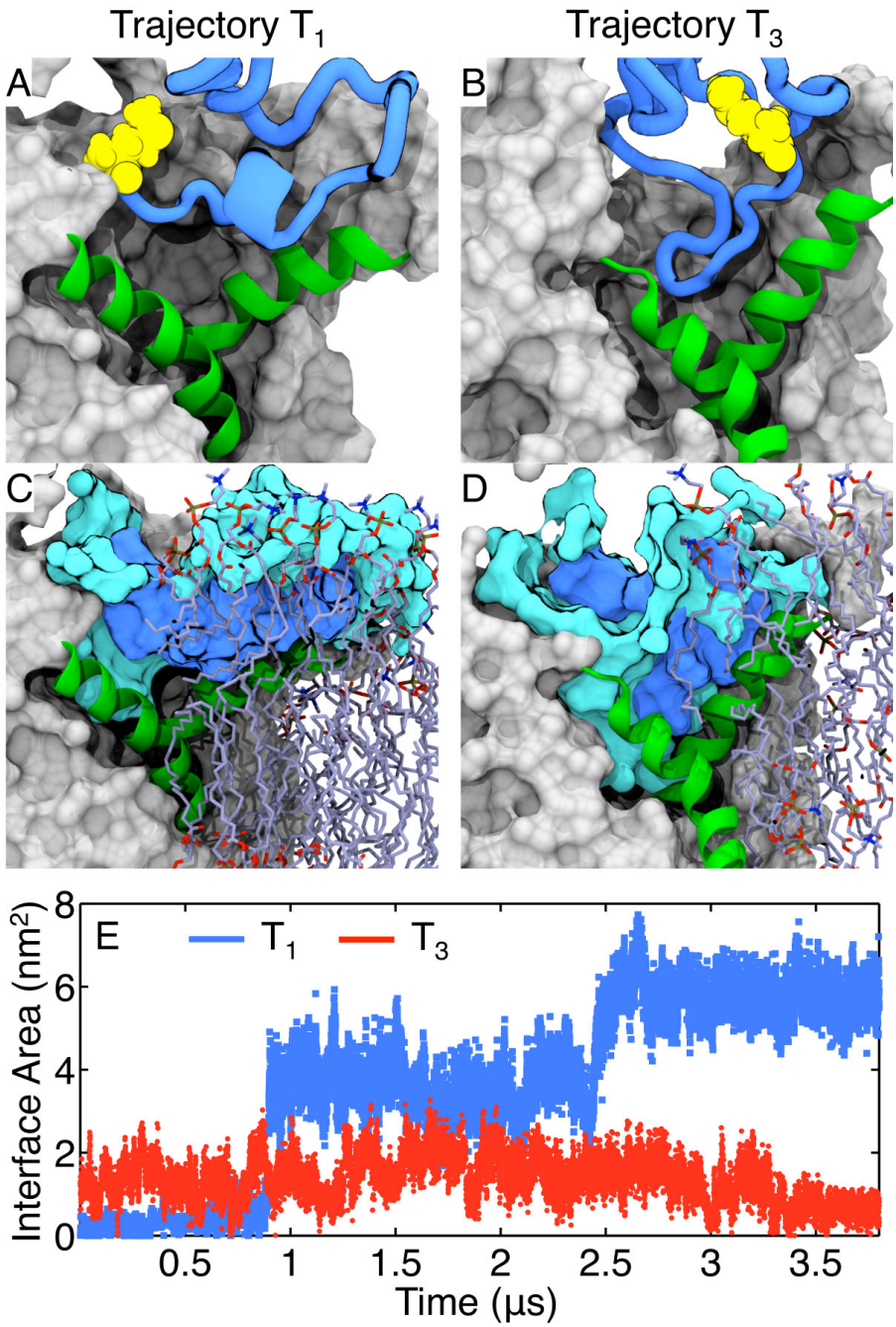


Figure 3.5: The SP adopts configurations that differ with respect to orientation, secondary structure, and solvation environment along the simulated insertion trajectories T_1 and T_3. **(A, B)** The conformation of the nascent protein in trajectories T_1 (part A) and in trajectory T_3 (part B). The nascent protein is presented in blue, with the N-terminal residue highlighted in yellow. The translocon is shown as a gray surface. **(C, D)** Formation of the hydrophobic interface between the SP (blue) and the lipid bilayer. Water within 8 Å of the SP is shown as a light blue surface. **(E)** The hydrophobic contact area between the SP and the surrounding lipid molecules, plotted as a function of time in trajectories T_1 (blue) and T_3 (red).

(dark blue). In part C, the absence of solvent density at the interface between the SP and the lipid tails is clear; the LG helices separate to make room for this hydrophobic contact, and water molecules at the cytosolic mouth of the translocon evacuate the space between the hydrophobic residues of the SP and the interior of the lipid membrane. Part D reveals a different solvation environment for the SP, with water molecules solvating the LG region due to the presence of the charged SP N-terminus and the hydrophilic lipid heads.

Figure 3.5E quantifies the magnitude and time dependence of the hydrophobic contact between the SP and the membrane interior. For both trajectories, the contact surface area between the SP residues and the lipid molecules are plotted as a function of simulation time; details of the surface area calculation are provided in appendix B. For trajectory T_1, in which the N-terminus of the SP is buried in the channel interior, the hydrophobic contact area increases markedly with simulation time; sharp increases in the curve correspond to the periods of peptide growth in the insertion simulation protocol. In contrast, the loop configuration adopted by the SP in trajectory T_3 leads to consistently small hydrophobic contact area at all times.

These results suggest that nascent-protein configurations that are on-pathway for Type III integration exhibit significant hydrophobic contact between the SP and the membrane interior, whereas configurations that are consistent with early-stage Type II integration exhibit aqueous solvation of the LG region. It follows that increased hydrophobicity of the SP residues will preferentially stabilize configurations of the kind shown in Figures 3.5A and 3.5C, enhancing the degree to which the nascent proteins undergo Type III integration. Similarly, mutation of positively charged residues on the N-terminus of the nascent-protein

SP will destabilize configurations of the kind shown in Figures 3.5B and 3.5D, decreasing the degree to which nascent proteins undergo Type II integration. Both of these trends have been experimentally observed [7, 59–61]. The simulation results presented here suggest a simple mechanistic basis for understanding the sensitivity of integral membrane protein topology to hydrophobic residues in the nascent-protein SP sequence.

3.3.3 Nascent-Protein Salt-Bridge Formation

Finally, we investigate the mechanism by which nascent-protein salt-bridge formation influences the topology of integral membrane protein TM domains. The mutation of negatively-charged residues at the cytosolic mouth of the translocon alters observed fractions of Type II and Type III integral membrane proteins, suggesting that electrostatic interactions involving the nascent protein play a role in conferring integral membrane protein topology [63]. Furthermore, favorable interactions involving the translocon are thought to facilitate the translocation of Arg-containing peptide sequences [64, 65] and to reconcile large discrepancies between the experimentally observed stop-transfer efficiency of Arg-containing peptides and computed water/membrane transfer free energies [5, 58, 66–69]. We explore these effects by characterizing the interactions of the translocon with positively charged residues in the nascent protein during insertion.

Figure 3.6A presents representative configurations from the insertion trajectories, viewed along the channel axis from the cytosolic side of the membrane. These snapshots reveal salt-bridge contacts that are formed between the Arg residue of the nascent-protein SP (blue) and either negatively charged residues on the translocon (E330, E110, and D404;

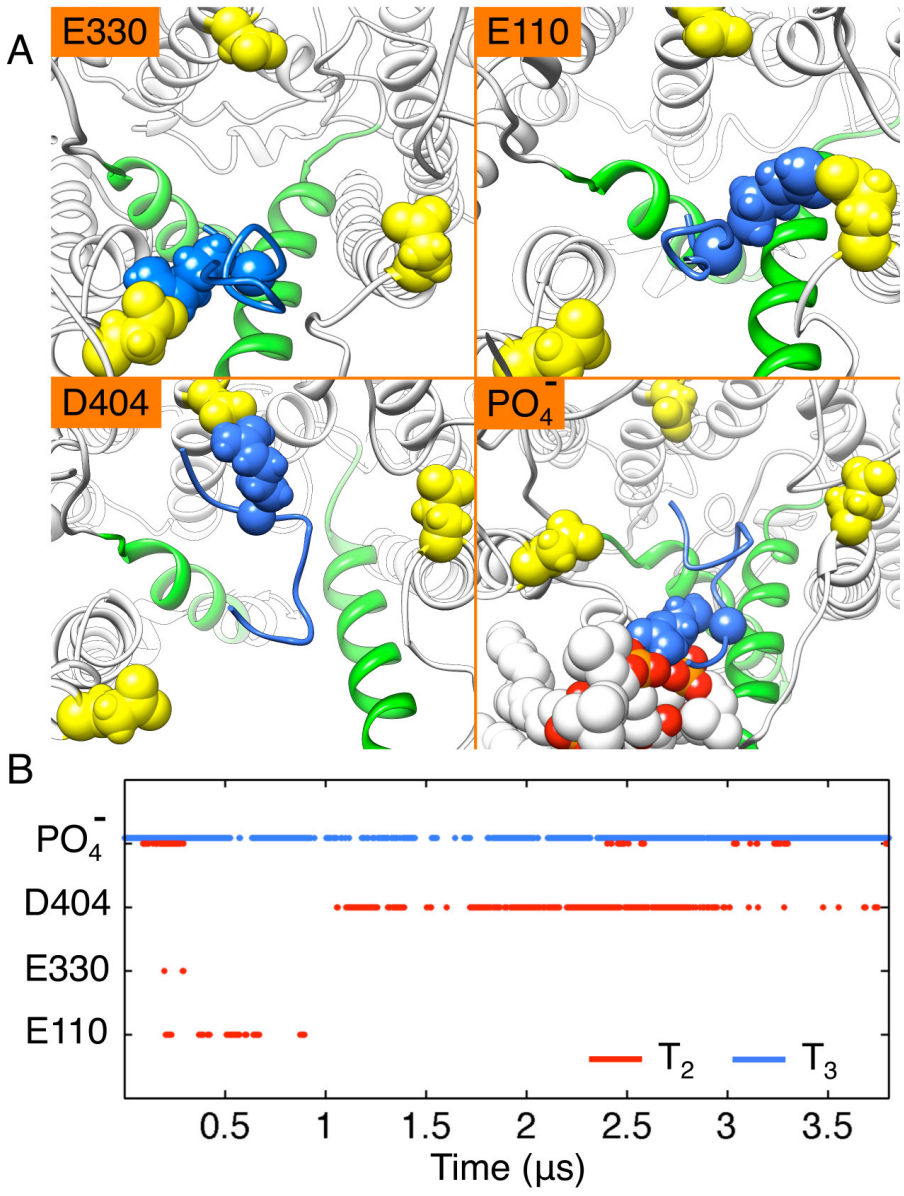


Figure 3.6: Formation of salt bridges involving the N-terminus of the nascent protein. **(A)** Representative configurations associated with salt bridges that are observed in the insertion trajectories. The SP is shown in blue, with its Arg residue shown in the space-filling representation. The translocon is shown in white ribbon, with the two LG helices in green. The negatively charged residues on the translocon are shown in yellow, and the lipid head groups are shown in orange and red. **(B)** The time evolution of the salt bridges formed during trajectories T_2 (red) and T_3 (blue).

yellow) or the phosphate head groups of the lipid bilayer (red). The configurations in panels E330, E110, and D404 are obtained from trajectory T_2 after 0.5, 1.4 and 2.6 μs of simulation time, respectively, whereas panel PO_4^- corresponds to trajectory T_3 after 2.6 μs. Figure 3.6B presents the time dependence of salt-bridge contacts in the simulations. The contacts are defined to include configurations for which the protonated nitrogen atom of either the Arg residue or the N-terminus of the SP is within 4 Å of the anionic oxygen atom of the corresponding translocon residue or phosphate group. The corresponding time-series plots for trajectories T_1 and T_4 are provided in Figure B.7. The structural alignment used to determine the eukaryotic homologues of residues E330, E110 and D404 is presented in Figure B.2.

It is clear from Figure 3.6B that salt-bridge contacts form almost immediately upon nascent-protein insertion trajectories and persist over microsecond timescales. In the first microsecond of trajectory T_2, which corresponds to the initial translation of the hydrophobic SP sequence, the nascent protein forms transient contact with the lipid head groups, as well as residues E330 and E110 of the translocon. At longer times, following translation of mature-domain residues, trajectory T_2 exhibits extended salt-bridge contact with residue D404. In contrast, trajectory T_3 immediately forms salt-bridge contact with the phosphate head groups of the membrane lipid molecules that persist throughout the length of the simulation.

The observed salt-bridge contacts involving the nascent protein are consistent with experimental observations that negatively charged translocon residues play a role in establishing the orientation of integral membrane protein TM domains. In particular, we see

the N-terminus of the nascent protein interacting at the earliest stages of insertion with the homologue of residue E110; mutation of this residue was experimentally found to decrease membrane integration for TM domains in the Type III orientation and increase membrane integration of TM domains in the Type II orientation [63]. Furthermore, long-lived salt-bridge contacts between the nascent protein and residue D404 are observed in trajectory T_2; mutation of the homologues of residue D404 and E330 is experimentally found to increase membrane integration of TM domains in the Type III configuration and decrease integration of TM domains in the Type II orientation [63]. The position of residue D404 and E330 at the cytosolic mouth of the translocon (Figure B.2) suggests that these residues favor configurations that are consistent with Type II membrane integration; the results for trajectory T_3 in Figure 3.6B, as well as in Figures 3.5B and 3.5D, suggest that a similar effect may arise from interactions of the N-terminus with the phosphate head groups of the membrane bilayer [18, 70].

3.4 Conclusions

We have introduced a simulation protocol for modeling the nonequilibrium dynamics of nascent-protein insertion into the Sec translocon. The approach is employed in combination with microsecond-timescale MD trajectories to investigate early-stage Sec-facilitated protein translocation and membrane integration. Analysis of multiple, long-timescale simulations reveals important molecular features of protein insertion into the translocon, including SP docking at the translocon LG, large-lengthscale conformational rearrangement of the translocon LG helices, and partial membrane integration of hydrophobic nascent-

protein sequences.

All-atom simulations reveal the role of specific molecular interactions in the regulation of protein secretion, membrane integration, and integral membrane protein topology. In particular, it is shown that hydrophobic nascent-protein domains stabilize open configurations of the translocon LG and facilitate partitioning of the nascent protein into the membrane lipid bilayer. Furthermore, we find that particular salt-bridge contacts between the nascent-protein N-terminus, cytosolic translocon residues, and phospholipid head groups favor conformations of the nascent protein that are consistent with the Type II topology, whereas increased SP hydrophobicity stabilizes nascent-protein configurations that are consistent with the Type III topology.

This work reports new insights obtained from detailed, microsecond-timescale MD simulations, and it provides a mechanistic basis for understanding experimentally observed correlations between integral membrane protein topology, translocon mutagenesis, and nascent-protein primary sequence. However, we also emphasize the limitations of all-atom MD trajectories for studying slower (i.e., second- to minute-timescale) features of Sec-facilitated protein translocation that give rise to experimentally observed kinetic effects [7, 71], and more generally, the biological timescales of protein biogenesis and transport. Regardless of important recent advances in the computation of MD trajectories, the accessible timescales for atomistic simulations will remain many orders of magnitude shorter than biologically relevant timescales for the foreseeable future. Ongoing efforts to understand protein biogenesis and transport must also involve the development of new methods and strategies for coarse-grained theoretical descriptions of the protein translocation ma-

chinery.

Chapter 4

Long-Timescale Dynamics and the Regulation of Sec-Facilitated Protein Translocation

4.1 Introduction

So far, we have shown that computer simulation studies provide a useful approach to understanding the translocon by connecting high-resolution structures to its detailed molecular interactions and dynamics. Yet the biological timescales for cotranslational protein translocation (i.e., minutes) vastly exceed the reach of atomistic MD simulations, and the large number of trajectories needed to explore the parameter space of protein sequence and translation rate with statistical significance ($\sim 10^5$ in the current study) dramatically constrains the computational cost of applicable simulation methods. As alluded to at the end of last chapter, new approaches are needed to bridge the hierarchy of timescales in Sec-facilitated protein translocation and membrane integration and to identify the mechanisms that govern these fundamental cellular processes.

In this chapter, we develop a CG model that enables simulation of the translocon and its associated macromolecular components on timescales beyond the scope of previously em-

ployed methodologies. The model explicitly describes the configurational dynamics of the nascent protein chain, conformational gating in the Sec translocon, and the slow dynamics of ribosomal translation (Figure 4.1). We use the model to perform minute-timescale CG trajectories to investigate the role of the Sec translocon in governing both stop-transfer efficiency (i.e., propensity of TM to undergo integration into the cell membrane versus secretion across the membrane) and integral membrane protein topogenesis (i.e., the propensity of TM to undergo membrane integration in the N_{cyt}/C_{exo} orientation versus the N_{exo}/C_{cyt} orientation). These simulations provide a direct probe of the mechanisms, kinetics, and regulation of Sec-facilitated protein translocation and membrane integration. Analysis of the full ensemble of nonequilibrium CG trajectories reveals the molecular basis for experimentally observed trends in integral membrane protein topogenesis and TM stop-transfer efficiency; it demonstrates the role of competing kinetic pathways and slow conformational dynamics in Sec-facilitated protein targeting; and it provides experimentally testable predictions regarding the long-timescale dynamics of the Sec translocon.

4.2 Signal Orientation and Protein Topogenesis

SP orientation is a determining factor in integral membrane protein topogenesis [72]. The orientation of N-terminal signals help to establish the topology of multidomain integral membrane proteins and to dictate whether N-terminal or C-terminal domains undergo translocation across the membrane. Biochemical studies have established the dependence of SP orientation upon a range of factors, including SP flanking charges [73, 74], SP hydrophobicity [59–61], protein mature domain length (MDL) [7], and the ribosomal transla-

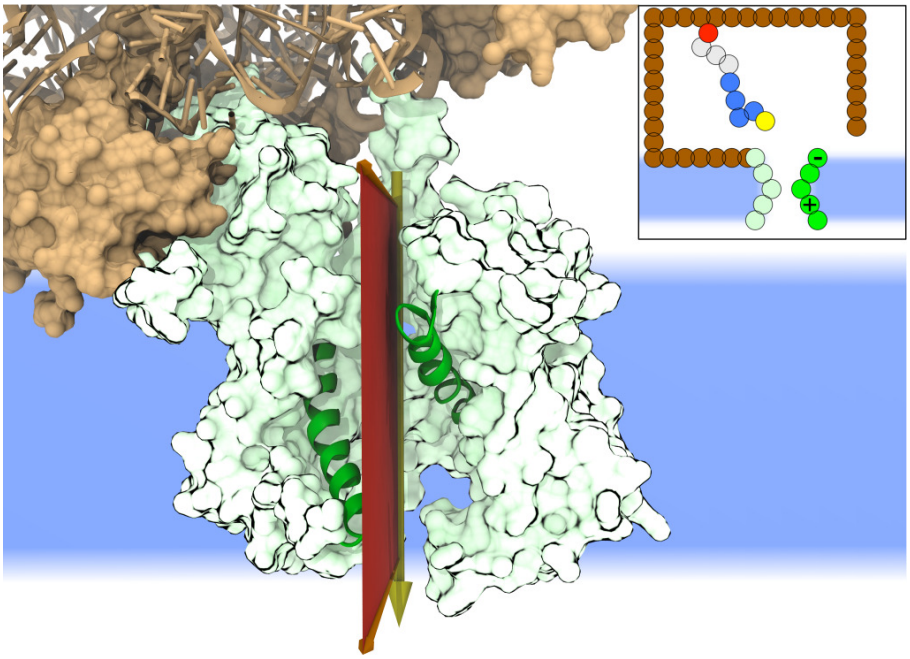


Figure 4.1: Structural features of the cotranslational Sec machinery. The ribosome (brown) is shown in complex with the Sec translocon (green). The CG model projects the protein nascent chain dynamics onto the plane (red) that intersects the translocon channel axis and that bisects the LG helices (dark green). **(Inset)** The CG model includes beads for the translocon (green), the ribosome (brown), and the protein nascent chain. The LG helices are shown in dark green, the ribosome exit channel is shown in red, and the lipid membrane is shown in blue. The nascent chain is composed of beads for the SP (yellow and blue) and the mature domain (gray).

tion rate [7]. In this section, we employ the CG model to directly simulate cotranslational protein integration and to determine the molecular mechanisms that give rise to these experimentally observed relations.

4.2.1 Direct Simulation of Cotranslational Protein Integration

We consider the process in which cotranslational integration of a signal anchor protein yields either the Type II (N_{cyt}/C_{exo}) or Type III (N_{exo}/C_{cyt}) orientation of the uncleaved SP domain; this nomenclature for the orientation of single-spanning membrane proteins follows earlier work [72]. Figure 4.2 illustrates the simulation protocol, with the N-terminal

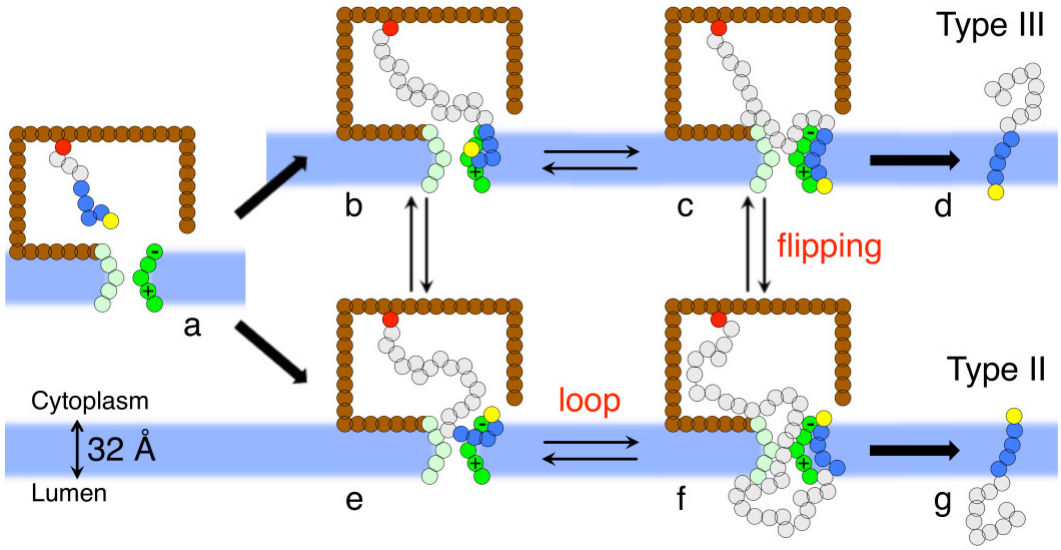


Figure 4.2: Kinetic pathways for Type II and Type III membrane integration of signal anchor proteins obtained from direct CG simulations. The coloring scheme is described in Figure 4.1.

SP domain shown in blue and yellow.

Following previous experimental work [7], we consider the integration of proteins that vary with respect to both SP sequence and MDL. The SP is composed of either a canonical sequence of CG beads (RL_4E), a sequence in which the positive charge on the N-terminal group is eliminated (QL_4E), or a sequence with enhanced SP hydrophobicity (RL_6E). To model the hydropathy profile of the engineered protein H1ΔLeu22 studied by Goder and Spiess [7] (Figure C.20), we consider proteins that include a hydrophilic mature domain with a hydrophobic patch near the SP; specifically, we model the protein mature domain using the Q_5LQ_n sequence of CG beads, such that the total peptide length ranges from 30 to 80 beads (90 to 240 residues). The sensitivity of protein topology to hydrophobic patches on the mature domain is examined in Figure C.13.

CG trajectories are continued until the protein nascent chain reaches either Type II or Type III integration. Depending upon the rate of ribosomal translation and the MDL,

each CG trajectory thus ranges from 2 to 20 s of simulation time; the corresponding CPU time required to perform each trajectory is approximately 0.2-10 hours. Each data point in Figures 4.3A-C is obtained by averaging the results of at least 600 independent CG trajectories. Full details of the simulation protocol are provided in Appendix C section: *Simulation Protocol*.

Figures 4.3A-C present the fraction of peptides that are calculated to undergo Type II integration as a function of protein MDL. In each case, the CG model predicts a strong dependence of SP topology on the length of the protein mature domain, with a fast rise in the Type II integration fraction at short lengths plateauing to a fixed value at longer MDL. The CG model also finds significant dependence of signal topology on the SP charge distribution (Figure 4.3A), SP hydrophobicity (Figure 4.3B), and ribosomal translation rate (Figure 4.3C). Each of these trends is in striking agreement with the findings of Goder and Spiess [7]; in addition to the crossover from strong to weak dependence of the signal topology with increasing MDL, the experimental study likewise reports Type II integration to be reduced with the removal of positively charged N-terminal groups, more hydrophobic SP sequences, and faster protein insertion. (See also Figure C.21). Figures C.14-C.16, and C.22 provide additional tests and comparisons of the CG model against protein topogenesis experiments, analyzing factors that include negative N-terminal charges, elongated N-terminal domains, charge mutations on the translocon, and charged patches on the nascent-protein mature domain. In the following, we use the CG simulations to enable the detailed analysis of the insertion dynamics and to determine the mechanistic origin of these various trends.

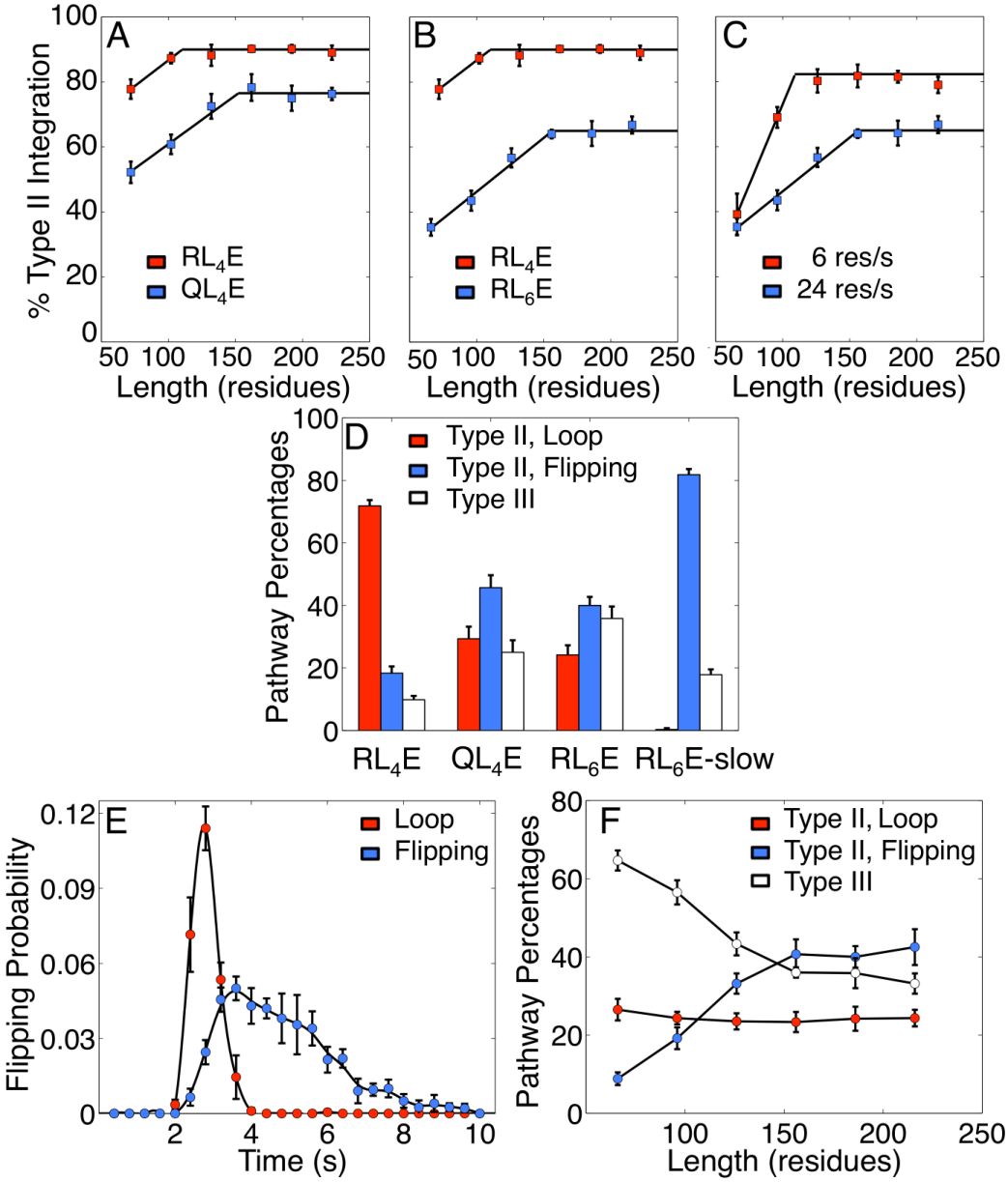


Figure 4.3: CG simulation results for integral membrane protein topogenesis. **(A-C)** Fraction of Type II integration as a function of protein MDL, with data sets that vary with respect to (A) SP charge distribution, (B) SP hydrophobicity, and (C) ribosomal translation rate. **(D)** Fraction of CG trajectories that follow the Type II loop pathway (red), Type II flipping pathway (blue), and the Type III pathway for membrane integration (white). **(E)** The distribution of arrival times for CG trajectories at state f of Type II integration via the loop pathway (red) and the flipping pathway (blue). **(F)** MDL-dependence of the fraction of CG trajectories that follow each integration mechanism.

4.2.2 Competition Between Kinetic Pathways Governs Topogenesis

Inspection of the ensemble of CG trajectories reveals multiple kinetic pathways by which the protein nascent chain achieves Type II or Type III integration (Figure 4.2). During early-stage protein insertion, the SP typically binds at the LG in one of two conformations, either with its N-terminus buried inside the translocon (state *b*) or exposed to the membrane (state *e*); similar conformations have been observed in microsecond-timescale, all-atom MD simulations of early-stage peptide insertion (Figure 3.5). From state *e*, further insertion of the nascent chain yields state *f*, in which the SP assumes the N_{cyt}/C_{exo} orientation; continued translocation of the mature domain in this orientation eventually leads to the Type II integration. From state *b*, further insertion leads to state *c*, in which the SP assumes the N_{exo}/C_{cyt} orientation; this orientation does not directly facilitate mature domain translocation, without which the protein assumes Type III integration. Slow transitions between states *c* and *f* are also observed in many trajectories; this conformational change, in which the SP "flips" between Type III and Type II integration topologies, is found to lie at the heart of many of the trends in Figures 4.3A-C.

To analyze the flow of trajectories among these competing mechanisms, the CG trajectories are categorized according to the chronology with which they pass through the states *a*-*g* in Figure 4.2. Each trajectory is associated with either the Type III mechanism ($a-b-c-d$), the Type II loop mechanism ($a-e-f-g$), or the Type II flipping mechanism ($a-b-c-f-g$). We emphasize that trajectories need not pass irreversibly through these states. Trajectories that visit state *c* prior to Type II integration are associated with the flipping mechanism, whereas any other trajectory that reaches Type II integration is asso-

ciated with the loop mechanism; all remaining trajectories are associated with the Type III mechanism. The definition for state *c* in terms of the coordinates of the model is presented in appendix C section : Simulation Protocols. Figure 4.3D presents the fraction of trajectories passing through each of these competing mechanisms, and it compares the effect of SP sequence and translation rate on the mechanism of integration. A total protein nascent chain length of 210 residues is considered for all cases in this figure.

Differences between the RL_4E and QL_4E data sets in Figure 4.3D help to explain the shift between the two corresponding data sets in Figure 4.3A. For the canonical SP sequence (RL_4E), Figure 4.3D shows that CG trajectories predominantly follow the Type II loop mechanism for integration. However, upon mutating the SP sequence with respect to the number of charged residues (QL_4E), the Type II flipping mechanism and the Type III mechanism become more prevalent. Removal of the N-terminal charge group diminishes the electrostatic stabilization of the SP in the N_{cyt}/C_{exo} orientation. The CG trajectories are thus less likely to visit states *e* and *f*, which are on-pathway for Type II loop integration, in favor of states *b* and *c*, which are on-pathway for both Type II flipping and Type III integration. Interestingly, the flipping mechanism allows for significant compensation of the Type II integration fraction upon mutation of the charge group; the effect of the SP sequence mutation on the flow of CG trajectories (Figure 4.3D) is thus much greater than the corresponding effect on the final branching ratio between Type II and Type III integration (Figure 4.3A). The simulations reveal a competition between electrostatic stabilization and SP reorientation kinetics that contributes to the well-known “positive-inside rule" for integral membrane protein topology [7, 75]. Furthermore, these results suggest that hinder-

ing the $c \rightarrow f$ flipping transition, perhaps via small molecule binding [76, 77], may lead to a larger effect on the Type II integration fraction than is observed with N-terminal charge mutation.

Comparison of the data for the RL_4E and RL_6E sequences in Figure 4.3D explains the shift between the two corresponding data sets in Figure 4.3B. Figure 4.3D shows that increasing the hydrophobicity of the SP reduces the flow of integration trajectories through the Type II loop mechanism. As before, this can be attributed to changes in the stability of states along the competing kinetic pathways. Increasing the hydrophobicity of the SP sequence significantly stabilizes SP configurations in state *b*, which favorably expose the hydrophobic segment to the membrane, instead of configurations in state *e*, which bury the hydrophobic segment inside the translocon. This effect draws trajectories away from the loop mechanism (Figure 4.3D) and leads to decreased Type II integration (Figure 4.3B).

Differences between the RL_6E and RL_6E-slow data sets in Figure 4.3D help to explain the shift between the two corresponding data sets in Figure 4.3C. Slowing the rate of ribosomal translation in proteins from 24 res/s to 6 res/s causes the CG trajectories to shift almost entirely to a Type II flipping mechanism. These differences are remarkable since they involve no change in the interactions of the system; the shifts in SP topology (Figure 4.3C) and integration mechanism (Figure 4.3D) with protein translation rate are purely kinetic effects. With slower translation, partially translated protein nascent chains have more time to undergo conformational sampling and are more likely to visit state *c*; it is therefore expected that Figure 4.3D shows Type II loop integration decreases in favor of combined Type II flipping integration and Type III integration. However, the corresponding decrease

in Type III integration is more surprising.

The decrease in Type III integration upon slowing translation arises from the important role of the flipping transition from state c to state f, which enables the nascent chain to reach the more thermodynamically favorable configurations associated with the N_{cyt}/C_{exo} SP orientation. Figure 4.3E plots the distribution of arrival times at state f for trajectories that follow either the Type II loop mechanism (red) or the Type II flipping mechanism (blue). Trajectories complete the loop mechanism relatively quickly, whereas the timescale for flipping persists as long as 10 s. The flipping transition thus introduces a slow timescale for conformational dynamics that couples to the dynamics of ribosomal translation. Slowing ribosomal translation provides more time for the nascent chain to undergo flipping; this purely kinetic effect enhances Type II integration in Figure 4.3C.

The final trend left to explain in Figures 4.3A-C is the dependence of the Type II integration fraction on the MDL. For every data set, the Type II integration fraction increases with MDL before plateauing to a constant value. Figure 4.3F elucidates this trend by presenting how the insertion mechanism varies with MDL; the percentage of CG trajectories following each mechanism is calculated as in Figure 4.3D.

With increasing MDL (Figure 4.3F), the fraction of trajectories following the Type II loop mechanism remains relatively unchanged, whereas the prevalence of Type II flipping increases at the expense of the Type III mechanism. As was seen from Figure 4.3E, trajectories commit to the Type II loop mechanism relatively early during insertion, prior to the full completion of ribosomal translation; it follows that increasing the MDL will have little effect on the fraction of trajectories following this mechanism. Furthermore, the trade-off

in Figure 4.3F between the Type II flipping and Type III mechanisms occurs for the same reason as was discussed for slowed ribosomal translation; increasing the MDL in Figure 4.3F provides more time for the tethered nascent chain to undergo the slow flipping transition from state c to the thermodynamically favored state f. At long MDL, the crowded environment in the ribosome-translocon junction causes nascent chain configurations in state c to be driven into state d before they can undergo the flipping transition; this causes the fraction of Type II flipping trajectories to cease rising in Figure 4.3F, such that relative fraction of Type II flipping and Type III trajectories approach a constant value. The results in Figure 4.3F correspond to the particular case of the RL_6E SP sequence and the 24 res/s translation rate; however, the trends are general and explain the MDL dependence of the Type II integration fraction in Figures 4.3A-C.

4.2.3 Loop versus Flipping Mechanisms

Observation of competing pathways for Type II integration is an unexpected and significant feature of the CG simulations presented here. Both the loop and flipping mechanisms for SP integration have been proposed in previous experimental studies [7, 18, 62, 78], although the possible role of peptide sequence and ribosomal translation rate in converting between these mechanisms has not been emphasized. Experimental support for the loop mechanism includes evidence that the protein nascent chain remains enclosed within the ribosome-translocon junction during the establishment of SP orientation [79]. Indeed, nascent proteins are found to be protected from cytosolic fluorescent quenching agents [36, 80] or proteases [81, 82] in some systems, although proteins with more hydrophobic SP sequences

are found to exhibit protease degradation in translation-stalled intermediates [82]. The loop mechanism is also consistent with observations that Type II integration is uninhibited by inclusion of bulky N-terminal domains in the protein nascent chain sequence [62, 83]. On the other hand, Spiess, Rapoport, and co-workers have proposed the flipping mechanism for Type II integration to explain observed trends in SP topogenesis [7, 8, 18]; and direct evidence in support of the flipping mechanism has recently been reported [78] under the assumption that translation-stalled intermediates of the ribosome/translocon/nascent-chain complex reflect the kinetic pathway for membrane integration. The observed co-existence of the loop and flipping mechanisms in our CG simulations helps to reconcile these experimental findings, and it provides a basis for understanding the competing influences of SP hydrophobicity, SP charge distribution, MDL, and ribosomal translation rate in regulating Sec-facilitated Type II and Type III protein integration.

In assessing the role of the Type II flipping mechanism in physiological systems, we note that many naturally occurring proteins exhibit longer N-terminal domains and less hydrophobic SP than the protein sequences considered in both here and in the work of Goder and Spiess [7]. As discussed previously, Figure 4.3D reveals that decreasing SP hydrophobicity leads to a decrease in the fraction of undergoing the Type II flipping mechanism. Furthermore, CG simulations performed using protein nascent chain sequences with longer N-terminal domains (Figure C.22), reveal a corresponding decrease in the fraction of trajectories that exhibit the Type II flipping mechanism.

4.3 Regulation of Stop-Transfer Efficiency

In addition to facilitating the translocation of proteins across the phospholipid membrane, the Sec translocon plays a key role in determining whether nascent protein chains become laterally integrated into the membrane [18]. Strong correlations between the hydrophobicity of a TM and its stop-transfer efficiency have led to the suggestion of an effective two-state partitioning of the TM between the membrane interior and a more aqueous region [1, 5]. However, models for this process based purely on the thermodynamic partitioning of the TM do not account for the experimentally observed dependence of stop-transfer efficiency on the length of the protein nascent chain [71], nor would such models anticipate any change in TM partitioning upon slowing ribosomal translation. Furthermore, recent theoretical [40] and experimental work [58] point out that the observed correlations between stop-transfer efficiency and substrate hydrophobicity can also be explained in terms of a kinetic competition between the secretion and integration pathways under the substrate-controlled conformational gating of the translocon. To further elucidate the mechanism of Sec-facilitated regulation of protein translcocation and membrane integration, we employ the CG model to directly simulate cotranslational stop-transfer regulation and to analyze the role of competing kinetic and energetic effects.

4.3.1 Direct Simulation of Cotranslational TM Partitioning

Following recent experimental studies [5, 6, 58, 84], we consider the cotranslational partitioning of a stop-transfer TM (i.e., the H-domain) where the protein nascent chain topology is established by an N-terminal anchor domain. Stop-transfer efficiency is defined as the

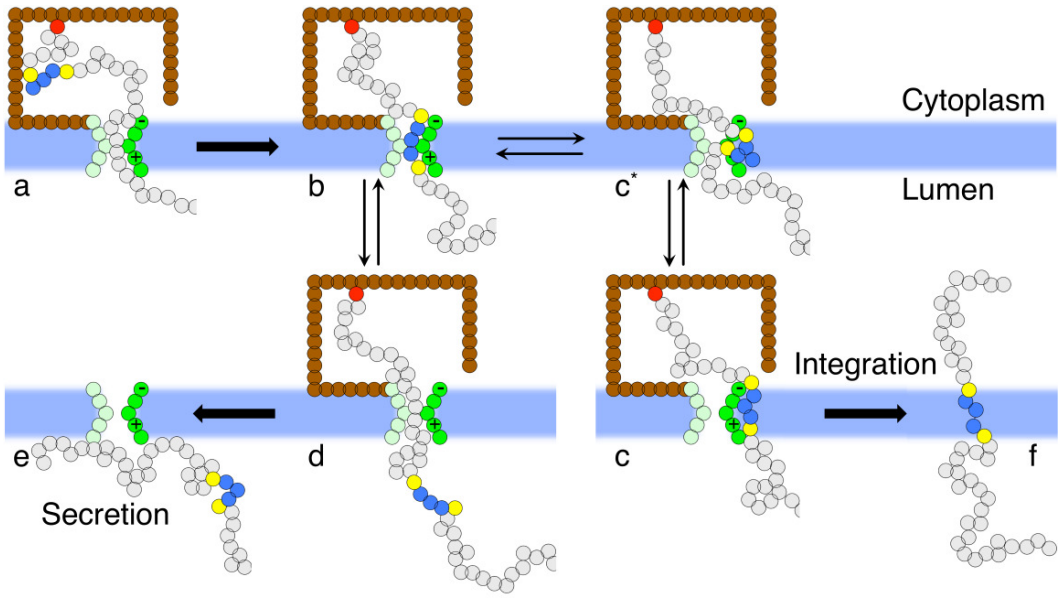


Figure 4.4: Kinetic pathways for cotranslational protein translocation and membrane integration obtained from direct CG simulations. The H-domain of the protein nascent chain is shown in blue and yellow. The full N-terminal anchor domain of the protein nascent chain is not shown here.

fraction of translated proteins that undergo H-domain membrane integration, rather than translocation. Figure 4.4 illustrates the simulation protocol, with the H-domain shown in blue.

The translated protein sequence is comprised of three components, including the N-terminal anchor domain, the H-domain, and the C-terminal tail domain. In all simulations, the N-terminal anchor domain includes 44 type-Q CG beads that link the H-domain to an anchor TM that is fixed in the N_{cyt}/C_{exo} orientation. The H-domain is comprised of the sequence P$\mathscr{L}_3$P, where the $\mathscr{L}$-type CG beads have variable hydrophobicity. The C-terminal domain includes a hydrophilic sequence of CG beads with periodic hydrophobic patches (poly-Q_5V), following the hydrophobicity profile of the dipeptidyl aminopeptidase B (DPAPB) protein studied by Junne and co-workers (Figure C.20) [58].

Stop-transfer efficiency is studied as a function of the hydrophobicity of the H-domain,

the C-terminal tail length (CTL), and the ribosomal translation rate. We consider CTL in the range of 5-45 beads (15-135 residues), and we consider water-membrane transfer free energies for the H-domain in the range of $\Delta G/k_{\mathrm{B}}T = [-5, 5]$, where ΔG corresponds to the sum over the individual transfer free energies of the CG beads in the H-domain.

CG trajectories are initialized with the H-domain occupying the ribosome-translocon junction, prior to translation of the C-terminal domain (Figure 4.4, state *a*). Each CG trajectory is terminated after full translation of the protein C-terminal domain, either when the H-domain integrates into the membrane and diffuses a distance of 16 nm from the translocon or when both the H-domain and the C-terminal domain fully translocate into the lumenal region. The N-terminal anchor TM of the protein nascent chain is fixed at a distance of 20 nm from the translocon; the simulations thus assume that the H-domain membrane integration mechanism does not involve direct helix-helix contacts with the N-terminal anchor TM [85]. Full details of the simulation protocol are provided in appendix C section: *Simulation Protocols*.

Figure 4.5 presents the calculated dependence of stop-transfer efficiency on the hydrophobicity of the H-domain, the length and hydrophobicity of the protein C-terminal domain, and the ribosomal translation rate. Each data point in Figures 4.5A, 4.5B and 4.5D is obtained from over 600 independent nonequilibrium CG trajectories; the simulation times for these trajectories span the range of 3-100 s. Figures C.17-C.19 provide additional tests and comparisons of the CG model against stop-transfer experiments, analyzing factors that include charged residues flanking the H-domain, hydrophobic patches on the C-terminal domain, and changes in protein translocation time.

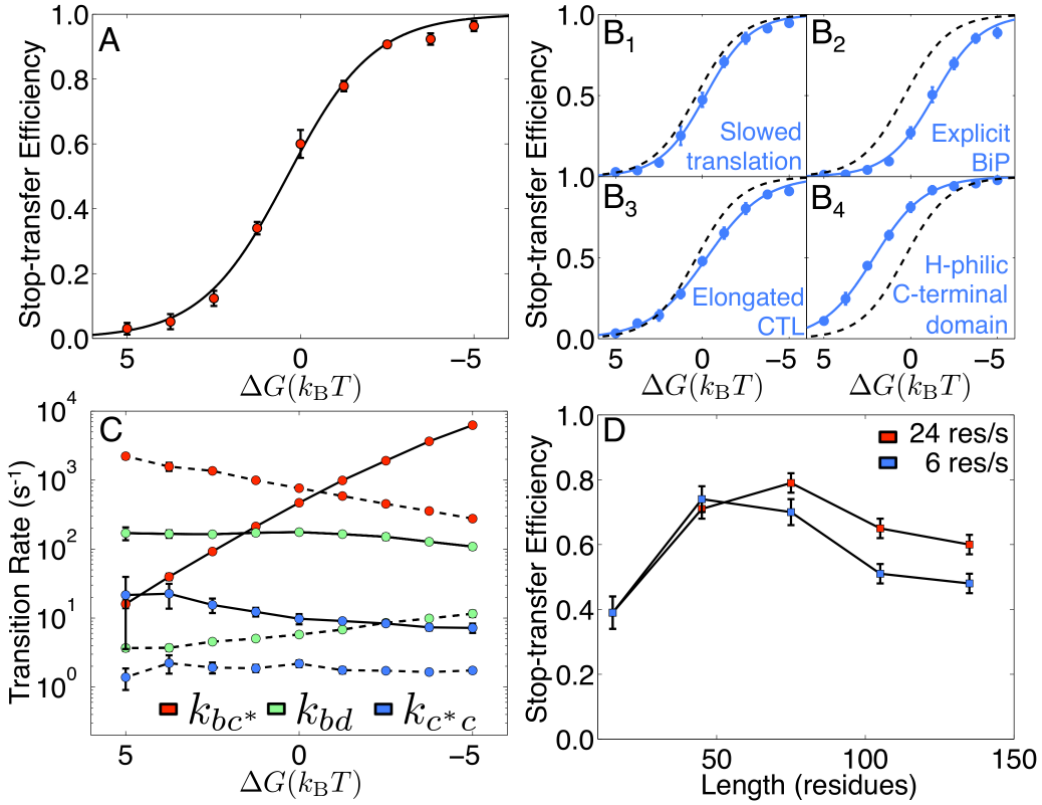


Figure 4.5: CG simulation results for TM partitioning. **(A)** Stop-transfer efficiency as a function of H-domain hydrophobicity. **(B)** Dependence of stop-transfer efficiency upon (B_1) slowing ribosomal translation rate from 24 to 6 res/s, (B_2) including explicit lumenal BiP binding, (B_3) increasing the CTL from 75 residues to 105 residues, and (B_4) replacing the hydrophobic beads in the protein C-terminal domain with hydrophilic beads; in each subpanel, the dashed line corresponds to the sigmoidal fit of the data in (A). **(C)** Equilibrium transition rates between the states in Figure 4.4 as a function of H-domain hydrophobicity. For each color, the forward rate is indicated with the solid line, and the reverse rate is indicated with dashed line. **(D)** Dependence of stop-transfer efficiency on CTL and the ribosomal translation rate, obtained for protein sequences with H-domain transfer FE of $\Delta G = -1.25 k_{\mathrm{B}}T$.

In Figure 4.5A, the stop-transfer efficiency is plotted as a function of the H-domain transfer FE, ΔG, for proteins with a CTL of 75 residues. The CG model recovers the experimentally observed [5] sigmoidal dependence of stop-transfer efficiency on H-domain hydrophobicity. The black curve in the figure corresponds to the state population for a system in apparent two-state thermal equilibrium,

$$P_{\mathrm{I}}(\Delta G) = (1 + \exp[-\beta\alpha\Delta G + \gamma])^{-1}, \tag{4.1}$$

where $\alpha = -0.80$, $\gamma = 0.29$, and $\beta = (k_{\mathrm{B}}T)^{-1}$ is the reciprocal temperature. The physical origin of this sigmoidal dependence of the stop-transfer efficiency, as well as the physical interpretation of the parameters α and γ, are a focus of the following analysis.

Figure 4.5B presents the calculated relationship between stop-transfer efficiency and H-domain hydrophobicity in systems for which either the ribosomal translation rate is slowed from 24 to 6 res/s (B_1), back-sliding of the protein nascent chain is inhibited to explicitly model the effect of the lumenal BiP binding (B_2), the CTL is increased from 75 to 105 residues (B_3), or the hydrophobic patches (V-type beads) in the C-terminal domain are replaced with hydrophilic, Q-type beads (B_4). In each case, the integration probability preserves the sigmoidal dependence on ΔG, and the best-fit value for the parameter α in each case is remarkably unchanged from the case in Figure 4.5A; for the four cases presented in Figure 4.5B, fitting the simulation data to equation (4.1) yields $\alpha = \{-0.77 \pm 0.08, -0.74 \pm 0.09, -0.60 \pm 0.06, -0.68 \pm 0.05\}$ and $\gamma = \{0.14 \pm 0.11, 1.0 \pm 0.19, -0.15 \pm 0.09, -1.44 \pm 0.13\}$; in each case the 95% certainty threshold for the sigmoidal fit is also indicated [86]. Cases B_1-B_3 each lead to a decrease in the stop-transfer efficiency for a

given value of ΔG (i.e., a rightward shift of the sigmoidal curve with respect to that obtained in Figure 4.5A), whereas decreasing the hydrophobicity of the C-terminal domain residues in case B_4 leads to an increase in stop-transfer efficiency.

4.3.2 The Origin of Hydrophobicity Dependence in TM Partitioning

Figure 4.4 introduces the primary mechanisms that the ensemble of CG trajectories are observed to follow in the simulations. Along the pathway to membrane integration, trajectories pass through configurations for which the H-domain occupies the translocon channel (Figure 4.4, state *b*), the membrane-channel interface across the open LG (state c^*), and the membrane region outside of the translocon with the LG closed (state *c*); upon completion of translation and release of the protein nascent chain, it diffuses into the membrane to reach the integration product (state *f*). Along the pathway to protein translocation, trajectories also pass through state *b*, before proceeding to configurations in which the H-domain occupies the lumen with the C-terminal domain threaded through the channel (state *d*); upon completion of translation, the C-terminal domain is secreted through the channel, yielding the translocation product (state *e*). In addition to the dominant pathways depicted in Figure 4.4, minor pathways for translocation and integration are observed for very short and very long CTL (Figure C.24). Complete definitions for the states in Figure 4.4 in terms of the coordinates of the CG model are provided in Figure C.10. We emphasize that trajectories do not irreversibly pass through the intermediate states in Figure 4.4; many trajectories backtrack repeatedly, starting down one pathway before finally proceeding down the other.

Given the observed mechanisms in Figure 4.4, we can derive and numerically test an

analytical kinetic model that explains the observed sigmoidal dependence of TM partitioning on H-domain hydrophobicity in Figures 4.5A and 4.5B. The analytical model assumes that *(i)* partitioning of the H-domain across the LG (i.e., transitions between states b and c^*) occurs on a faster timescale than all other transitions in the system, such that these states are always in equilibrium, *(ii)* the populations of states b and c^* are slowly varying on the timescale of translocation and integration (i.e., these populations satisfy a steady-state approximation), and *(iii)* only the rate of transitions between states b and c^* depend on the H-domain hydrophobicity. From the first two assumptions, it follows that the nonequilibrium populations of state b and c^* exhibit the functional form

$$\ln[P_b(t;\Delta G)/P_{c^*}(t;\Delta G)] = -\beta\alpha\Delta G + \text{const.}, \tag{4.2}$$

where $P_b(t;\Delta G)$ and $P_{c^*}(t;\Delta G)$ are the nonequilibrium populations for protein nascent chains of H-domain hydrophobicity ΔG at time t after the start of ribosomal translation. It then follows from the third assumption that

$$\ln[P_d(t;\Delta G)/P_c(t;\Delta G)] = -\beta\alpha\Delta G + \delta(t), \tag{4.3}$$

where $P_d(t;\Delta G)$ and $P_c(t;\Delta G)$ are the nonequilibrium populations for states c and d, and the function $\delta(t)$ is independent of ΔG. Since trajectories arrive irreversibly at states e and f, the fraction that undergo membrane integration is thus

$$P_{\mathrm{I}}(\Delta G) = \left(1 + \frac{\int_\tau^\infty dt\, P_d(t;\Delta G)k_{de}}{\int_\tau^\infty dt\, P_c(t;\Delta G)k_{cf}}\right)^{-1}, \tag{4.4}$$

where τ is the time at which translation completes and the protein is released from the ribosome. Finally, inserting equation (4.3) into equation (4.4) and using that the model assumptions imply that the nonequilibrium populations are separable functions of t and ΔG (i.e., $P_c(t;\Delta G) = \phi_1(t)\phi_2(\Delta G)$), we arrive at equation (4.1), with the ΔG-independent constant

$$\gamma = \ln\left[\frac{\int_\tau^\infty dt\, \phi_1(t) e^{\delta(t)} k_{de}}{\int_\tau^\infty dt\, \phi_1(t) k_{cf}}\right]. \tag{4.5}$$

We can use the CG simulations to numerically test the assumptions of this analytical model. Figure 4.5C presents the equilibrium transition rates among the states in Figure 4.4, which are obtained from the frequency of inter-state transitions in long CG trajectories of a protein nascent chain with a 75-residue C-terminal domain tethered at its C-terminus to the ribosome exit channel. The calculation is repeated for proteins with a range of values for the H-domain hydrophobicity, ΔG. Indeed, the figure confirms that partitioning of the H-domain across the LG of the translocon (i.e., forward and reverse transitions between states b and c^*) occurs on a faster timescale than most other transitions in the system. Furthermore, it is clear that the rates k_{bc^*} and k_{c^*b} are strongly dependent on the hydrophobicity of the H-domain, whereas the other transition rates are only weakly dependent on ΔG. These results are thus consistent with assumptions *(i)* and *(iii)* of the analytical model.

A more stringent numerical test of the analytical model is presented in Figure 4.6. From the ensemble of CG trajectories used to construct Figure 4.5A, we examine whether the nonequilibrium state populations are consistent with equation (4.3). In Figure 4.6A, the left-hand side of equation (4.3) is plotted at various times t during the TM partitioning and for proteins with a range of H-domain hydrophobicity, ΔG. The set of data points

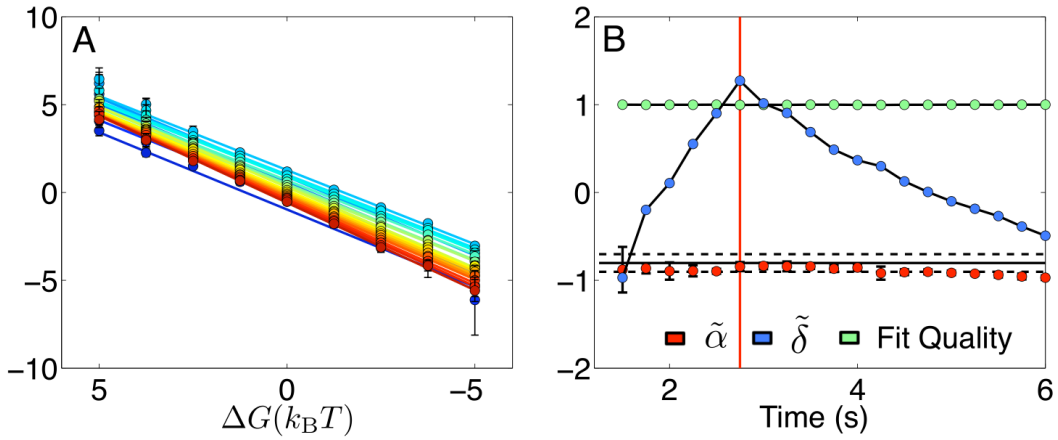


Figure 4.6: Numerical validation of the analytical model for TM partitioning, obtained from the ensemble of CG trajectories used to obtain Figure 4.5A. **(A)** The left-hand side of equation (4.3) is plotted at various times t during the TM partitioning and for proteins with a range of H-domain hydrophobicity, ΔG. The set of data points that correspond to each time t (indicated by color) is then fit to the linear function $-\beta\tilde{\alpha}\Delta G+\tilde{\delta}$. **(B)** The linear fit parameters $\tilde{\alpha}$ and $\tilde{\delta}$ (red and blue, respectively) obtained at each time t, as well as the R-squared measure of quality of the linear fit (green), are plotted. The solid line at -0.80 corresponds to the value for α obtained by directly fitting the data in Figure 4.5A with equation (4.1); dashed lines indicate the threshold of 95% certainty in this direct fit. The vertical red line at $t = 2.75$ s corresponds to the time at which translation of the protein nascent chain completes.

corresponding to each time t is then fit to a linear function of the form $-\beta\tilde{\alpha}\Delta G+\tilde{\delta}$. Figure 4.6B presents the linear-fit parameters $\tilde{\alpha}$ and $\tilde{\delta}$ (red and blue, respectively) obtained at each time t, as well as the R-squared measure of quality of the linear fit (green) [86]. Confirmation that the CG simulation data obeys equation (4.3) follows from the fact that that $\tilde{\alpha}$ is independent of t and the R-squared measure is near-unity for all t. Furthermore, the value $\tilde{\alpha} \approx -0.85$ obtained in this analysis of the nonequilibrium state populations precisely matches the value $\alpha = -0.80$ obtained from directly fitting the data in Figure 4.5A with equation (4.1), thus providing numerical support for the derivation of equations (4.1) and (4.5) from equation (4.3) under the assumptions of the analytical model.

The simple analytical model presented in this section provides a mechanistic basis for understanding the sigmoidal relationship between stop-transfer efficiency and H-domain

hydrophobicity that is observed in both simulations (Figures 4.5A and 4.5B) and experiment [5]. The H-domain achieves rapid, local equilibration across the translocon LG; this partitioning is highly sensitive to the hydrophobicity of the H-domain, which gives rise to the characteristic sigmoidal dependence of the curves in Figures 4.5A and 4.5B, and it is kinetically uncoupled from slower steps in the mechanisms of integration and translocation, which explains the robustness of parameter α in fitting the various sets of data in Figures 4.5A and 4.5B. We note that this mechanism involving local equilibration of the H-domain between the translocon and membrane interiors is consistent with recent experimental studies of stop-transfer efficiency [58, 69]. Kinetic and CTL effects in TM partitioning arise from competition among slower timescale processes in the secretion and integration pathways; these effects are manifest in parameter γ (equation (4.5)) and lead to lateral shifts of the sigmoidal curves in Figure 4.5B (equation (4.1)).

4.3.3 Kinetic and CTL Effects in TM Partitioning

The direction of the lateral shifts of the curves in Figure 4.5B can also be understood from analysis of the CG trajectories. In part B_1, slowing the translation rate allows for better equilibration among the states *d* and *c* prior to release of the protein from the ribosome, leading to increased population of the thermodynamically favored state *d* and enhancement of the secretion product; Figure C.23 demonstrates the relative increase of the nonequilibrium population in state *d* upon slowed ribosomal translation. In part B_2, the BiP motor enhances the secretion product by biasing against trajectories that back-slide from state *d*. Part B_3 exhibits a combination of these two effects, with the elongated C-terminal domain

allowing more time for the protein conformation to interconvert between states *d* and *c* prior to release from the ribosome (Figure C.23) and with a decreased rate of back-sliding from state *d* with longer CTL (Figure C.25). Finally, part B_4 reveals that decreased hydrophobicity of the C-terminal domain residues leads to increased stop-transfer efficiency. Without hydrophobic patches, the C-terminal domain residues in the translocon channel do little to stabilize opening of the LG; therefore, once the system reaches state *c* along the pathway to membrane integration, it is less likely that the H-domain will return to the channel interior and then undergo secretion (Figure C.17).

Figure 4.5D provides a more complete view of the connection between CTL, ribosomal translation rate, and stop-transfer efficiency. At relatively long CTL ($\geq$ 75 res.), stop-transfer efficiency decreases for longer proteins and for slower ribosomal translation, as was previously discussed in connection with Figures 4.5B_1 and 4.5B_3. However, at short CTL ($\leq$ 50 res.), stop-transfer efficiency increases for longer proteins and exhibits no dependence on the ribosomal translation rate. In the short-CTL regime, slowing ribosomal translation affords little additional time for the protein conformation to interconvert between states *d* and *c* prior to release from the ribosome (Figure C.23); there is thus no enhancement of the nonequilibrium population for state *d* and no corresponding change in stop-transfer efficiency. Previous experimental studies of stop-transfer efficiency involving relatively short CTL find no dependence of stop-transfer efficiency on translation rate [71], as is consistent with the results in Figure 4.5D; experimental results for longer CTL would be of significant interest.

4.4 Discussion

We have introduced a CG model for the direct simulation of cotranslational protein translocation and membrane integration on biological timescales. The model, which is based on MD simulations and limited experimental data, captures a striking array of experimentally observed features of integral membrane protein topogenesis and stop-transfer efficiency. The success of the model suggests that regulation of Sec-facilitated protein translocation and membrane integration arises from simple features of the translocon machinery, including the confined geometry of the ribosome and translocon channel, conformational flexibility the translocon LG, and electrostatic and hydrophobic driving forces. Analysis of over $40,000$ minute-timescale CG trajectories provides detailed insight into the mechanistic origin of the observed trends in protein targeting. In simulations of integral membrane protein topogenesis, the ensemble of CG trajectories suggests that the experimentally observed dependence of signal orientation on the ribosomal translation rate [7] arises from the slow reorientation (i.e., flipping) of the SP in the confined environment of the translocon channel. In simulations of TM partitioning, the ensemble of CG trajectories suggests that the experimentally observed sigmoidal relationship between stop-transfer efficiency and the H-domain hydrophobicity [5] arises from rapid local equilibration of the H-domain across the translocon LG. Finally, we utilize the CG model to predict the dependence of cotranslational protein stop-transfer efficiency on the ribosomal translation rate, protein nascent chain sequence, and protein CTL. The theoretical framework put forward in this chapter provides a basis for testing and refining the mechanistic understanding of Sec-facilitated protein targeting.

4.5 Methods

Here, we present the CG model for direct simulation of cotranslational protein translocation and membrane integration. The model introduces necessary simplifications to reach the long timescales associated with these biological processes. It is parameterized using the results of MD simulations and transferable experimental data. Numerical testing, reported in the Results Section, and in the appendix C, indicates that the CG model is consistent with independent experimental measurements of protein translocation and membrane integration and that reported conclusions are robust with respect to the details of the model parameterization.

The most aggressive simplification employed in the CG model is projection of the nascent protein dynamics onto the plane that passes along the translocon channel axis and between the helices of the LG (see Figure 4.1, as well the more detailed description below). The model includes explicit opening and closing of the translocon LG, which corresponds to the LG helices passing into and out-of the plane of the nascent protein dynamics, but the nascent protein is itself confined to the planar subspace. This dimensionality reduction is necessary to make tractable the minute-timescale trajectories for protein translocation and membrane integration. Similar approaches are well established for the study of biomolecule transport and translocation systems. Planar models have been utilized for the theoretical analysis [87–89] and computer simulation [90–94] of protein and DNA translocation through nanometer-lengthscale pores, and they have been used to investigate both thermodynamic and kinetic features of protein folding pathways [95–97]. Even more simplified one-dimensional models of protein translocation have proven use-

ful [55, 98–100]. The success of such models follows from the pseudo-one-dimensional nature of pore-transport phenomena; kinetic bottlenecks are largely governed by progress transverse to the narrow pore, enabling dramatic simplification of other degrees of freedom. Although the CG model presented here is novel in that it explicitly describes translocon LG motions and ribosomal translation, it is based on the foundation of these earlier physical models.

Parameterization of the CG model utilizes MD simulations and transferable experimental data. Free energy calculations and direct MD simulations determine the energetics and timescales of LG opening, including the dependence of the LG energetics on the nascent-protein amino acid sequence; microsecond-timescale all-atom simulations and experimental measurements determine the diffusive timescale for the CG representation of the nascent protein; and experimental amino acid water/membrane transfer free energies determine the solvation energetics of the CG nascent protein residues.

Following initial parameterization, the CG model is left unchanged throughout the remainder of the study. Numerical tests indicate that the reported conclusions are robust with respect to geometric features of the translocon (Figure C.2) and the ribosome (Figure C.3) , the timescales for translocon LG motion and nascent protein diffusion (Figures C.4-C.9), features of the nascent protein sequence (Figures C.13, C.15, C.18 and C.17), and the effects of lumenal biasing factors, such as BiP (Figures C.11 and C.12). These validation studies, as well as comparison of the simulations with experimental results, (Figures 4.3, 4.5, C.13-C.19, C.21 and C.22), suggest that the model captures the essential features of translocon-guided protein translocation and membrane integration.

Nonetheless, limitations of the CG model are emphasized from the outset. In addition to enforcing planar constraints on the motion of the nascent protein, the model provides a coarsened representation for nascent-protein, translocon, and membrane bilayer that includes only simple aspects of electrostatic and hydrophobic driving forces; potentially important details of residue-specific interactions are thus neglected [70]. Backbone interactions along the nascent protein chain are also neglected, such that effects due to the onset of nascent protein secondary structure are ignored, and effects due to translocon conformational changes other than LG motion are not explicitly included. Moreover, the possible roles of membrane-bound chaperones or oligomerization of the translocon channel [101] are not considered here. In principle, the CG model can be modified to incorporate greater accuracy and detail, as well as additional complexity and computational expense. In its current form, which is described in detail below, the model provides a minimalist description of Sec-facilitated protein translocation and membrane integration.

4.5.1 The System

The model employs CG particles, or beads, to describe the Sec translocon, protein nascent chain, hydrophobic membrane interior, and confinement effects due to the translating ribosome. The beads are constrained to the plane that lies normal to the lipid bilayer membrane and that bisects the translocon channel interior and the LG helices (Figure 4.1). CG beads corresponding to the residues of the translating nascent chain (Figure 4.1 inset) evolve subject to overdamped Brownian dynamics, whereas beads representing the Sec translocon (light and dark green) and the docked ribosome (brown) are fixed with respect to the

membrane bilayer. To explicitly incorporate the conformational gating of the translocon LG helices, beads representing the LG helices (dark green) undergo stochastic transitions between closed-state interactions, which occlude the passage of the nascent chain from the Sec channel to the membrane interior, and open-state interactions, for which the steric barrier to membrane integration is removed. Structural features of the channel and ribosomal confinement are obtained from crystallographic and electron microscopy studies [10, 79]. The positions for the translocon and ribosome beads are reported in Table C.2.

4.5.2 Interactions

We employ a CG bead diameter of $\sigma = 8$ Å, which is typical of the Kuhn length for polypeptide chains [102, 103]; the protein nascent protein chain is thus modeled as a freely jointed chain with each CG bead corresponding to approximately three amino acid residues. Bonding interactions between neighboring beads in the nascent chain are described using the finite extension nonlinear elastic (FENE) potential [104], $U(r) = -\frac{1}{2}kR_0^2\ln(1-r^2/R_0^2)$, where $k = 7\varepsilon/\sigma^2$, $R_0 = 2\sigma$, and $\varepsilon = 0.833k_{\mathrm{B}}T$; all simulations are performed using $T =$ 300 K. The bonding interactions are sufficiently strong to avoid self-crossing of the protein nascent chain.

For the description of nonbonded interactions, the CG beads are categorized into various types. For the protein nascent chain, the CG bead types correspond to positively charged (R), negatively charged (E), neutral-hydrophobic (L), neutral-hydrophilic (Q), mildly hydrophobic (V), amphiphilic (P), and variable-hydrophobic ($\mathscr{L}$) groups of amino acid residues. Additional CG bead types correspond to residues of the ribosome, residues for

the translocon LG in the closed state (LG_c), residues for the translocon LG in the open state (LG_o), and residues for the translocon that are not part of the LG (LG_n).

Short-ranged nonbonding interactions are modeled using the Lennard-Jones (LJ) potential energy function,

$$U_{\mathrm{LJ}}(r) = \begin{cases} 4\varepsilon_{\mathrm{lj}}\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] + \varepsilon_{\mathrm{cr}} & , \; r_{\mathrm{cl}} < r \leq r_{\mathrm{cr}} \\ 0 & , \text{otherwise} \end{cases} \tag{4.6}$$

where the constant $\varepsilon_{\mathrm{cr}}$ ensures that the pairwise interaction vanishes at r_{cr}. For each pair of CG bead types, the corresponding LJ parameters are reported in Table C.3. For the nonbonding interactions among the beads of the protein nascent chain and between beads of the nascent chain and the ribosome, the LJ parameters correspond to soft-walled, excluded volume interactions [105]. Weak attractive interactions account for the affinity of the protein nascent chain for the LG helices of the translocon, as has been observed in cross-linking experiments [32]. For the open state of the LG, repulsions between the LG and protein nascent chain beads are truncated to allow the peptide to laterally exit the translocon channel.

Pairwise Coulombic interactions are modeled using the Debye-Hückel potential, $U_{\mathrm{DH}}(r) = \sigma q_1 q_2 (\beta r)^{-1} \exp[-r/\kappa]$, where q_1 and q_2 are the charges for the various CG beads (Table C.1). We employ a Debye length of $\kappa = 1.4\sigma$ that is typical for electrostatic screening under physiological conditions. Two additional charges are included to model charge distribution among the residues of the translocon; a charge of $q = -2$ and $q = 2$ are included on first and fourth beads of the LG, where the LG beads are ordered with respect to their

distance from the cytosol. The justification for the LG beads charges is discussed in appendix C section: *Model Parameterization and Validation*. When the LG is in the open state, the electrostatic potential between the beads on the LG and on the protein nascent chain is capped from below to avoid the singularity in the Debye-Hückel potential, such that

$$U(r) = \begin{cases} U_{\mathrm{DH}}(r) & , \; r > \sigma \\ U_{\mathrm{DH}}(\sigma) & , \text{otherwise.} \end{cases} \tag{4.7}$$

Solvation energetics for each CG bead is described using the position-dependent potential energy function

$$U_{\mathrm{solv}}(x,y) = g\, S(x;\phi_x,\psi_x)\,[1 - S(y;\phi_y,\psi_y)], \tag{4.8}$$

where x and y are the Cartesian coordinates for the CG bead (Figure C.1), and g is the corresponding water-membrane transfer FE (Table C.1). Smooth transitions for the bead solvation energy upon moving from aqueous to membrane environments are achieved using the switching function

$$S(x;\phi,\psi) = \frac{1}{4}\left(1 + \tanh\frac{x-\phi}{b}\right)\left(1 - \tanh\frac{x-\psi}{b}\right), \tag{4.9}$$

where the switching lengthscale is $b = 0.25\sigma$. The parameters that describe the switching between the aqueous and membrane regions of the system are $\phi_x = -2.0\sigma$, $\psi_x = 2.0\sigma$, $\phi_y = -1.5\sigma$, $\psi_y = 1.5\sigma$.

4.5.3 Dynamics

The time-evolution of the system is modeled using a combination of Brownian dynamics for the nascent protein chain and stochastic opening and closing of the translocon LG. The off-lattice nascent chain dynamics is evolved using the first-order Euler integrator [106]

$$x_i(t+\Delta t) = x_i(t) - \beta D \frac{\partial V(\mathbf{x}(t))}{\partial x_i}\Delta t + \sqrt{2D\Delta t}\,\eta_i, \tag{4.10}$$

where $x(t)$ is a Cartesian degree of freedom for the nascent chain at time t, $V(\mathbf{x}(t))$ is the potential energy function for the full system, D is the isotropic diffusion constant for the CG beads, $\beta = (k_{\mathrm{B}}T)^{-1}$, and η is a random number drawn from the Gaussian distribution with zero mean and unit variance. As is described in appendix C section: *Model Parameterization and Validation* , a CG bead diffusion constant of $D = 758.7\,\mathrm{nm}^2/\mathrm{s}$ reproduces experimentally observed timescales for nascent chain diffusion through the translocon channel [100, 107] and is consistent with microsecond all-atom MD simulations. With this diffusion constant and the previously described interaction parameters, equation (4.10) can be stably integrated with a timestep of $\Delta t = 100\,\mathrm{ns}$.

At every simulation timestep, the probability of LG opening/closing is $p_{\text{open/close}} = k_{\text{open/close}}\Delta t$, where

$$k_{\text{open}} = \frac{1}{\tau_{\text{LG}}}\frac{\exp(-\beta\Delta G_{\text{tot}})}{1+\exp(-\beta\Delta G_{\text{tot}})}, \tag{4.11}$$

$$k_{\text{close}} = \frac{1}{\tau_{\text{LG}}}\frac{1}{1+\exp(-\beta\Delta G_{\text{tot}})}. \tag{4.12}$$

Here, τ_{LG} corresponds to the timescale for attempting LG opening or closing events, and

ΔG_{tot} is the FE cost associated with LG opening. As is described in appendix C section: *Model Parameterization and Validation*, The calculation of ΔG_{tot}, as well as the dependence of this FE cost on the nascent chain contents of the translocon channel, is based on MD simulations of the channel/peptide-substrate/membrane system [40]. The timescale $\tau_{\text{LG}} = 500$ ns is likewise determined from MD simulations [40]. Equations (4.10)-(4.12) satisfy detailed balance, ensuring that the CG dynamics is consistent with equilibrium Boltzmann statistics.

4.5.4 Modeling Translation

Ribosomal translation is directly modeled in the CG simulations via growth of the nascent chain at the ribosome exit channel (Figure 4.1 inset, red). The C-terminus of the protein nascent chain is held fixed at the exit channel throughout translation, and beads are sequentially added at the C-terminal tail, elongating the protein nascent chain. Upon completion of translation, the nascent chain is released from the exit channel, and the small subunit of the ribosome dissociates from the cytosolic mouth of translocon [108, 109]; we model ribosomal dissociation by eliminating interactions associated with the ribosome CG beads. Ribosomal translation proceeds at a pace of approximately 10-20 amino acid residues per second (res/s) [110, 111], although this rate can be reduced approximately fourfold upon addition of cycloheximide [7, 112]; we thus consider ribosomal translation rates in the range of 6-24 res/s (2-8 beads/s) in the current study.

The binding immunoglobulin protein (BiP) is an essential component of the eukaryotic Sec translocon machinery [113]. In appendix C section: *Explicit Modeling of Lumenal*

BiP, we consider the explicit inclusion of BiP binding within the CG model and show that it gives rise to only modest effects in the calculated results for protein translation and membrane integration. Unless otherwise stated, explicit BiP binding is not included in the reported simulation results.

Chapter 5

Conclusions and future work

5.1 Conclusions

Decades of studies on the Sec-facilitated protein translocation and integral membrane protein topogenesis have cumulated a large amount of valuable information. However, due to the different experimental approaches and setups used, establishing connections between these studies to propose a unified framework for the general mechanistic understanding of Sec-facilitated protein targeting has proven to be challenging. In this thesis, I presented our effort toward that goal using first principle computer simulation. In the following, I summarize some of our contributions.

- **Conformational landscape of the Sec translocon.** Using rigorous FE calculations, we have demonstrated that the archaeal crystal structure [10] is the only metastable conformation for the bare translocon. We further showed that the translocon conformational landscape can be regulated with the presence of peptide substrate, and inclusion of a hydrophobic peptide substrate stabilizes an open conformation of the LG that facilitates membrane integration. Our study here provides mechanistic in-

sight on the dependence of protein integration on peptide hydrophobicity and lays down the foundation for the further development of CG model.

- **Molecular features of early-stage protein translocation.** Using a novel nonequilibrium insertion protocol, we directly simulated the early-stage nascent-protein insertion into the Sec translocon. This study provides molecular pictures for a series of events suggested from biochemical studies, including signal-peptide docking at the translocon LG, large-lengthscale conformational rearrangement of the translocon LG helices, and partial membrane integration of hydrophobic nascent-protein sequences. It also helps to elucidate the role of nascent protein sequence, physicochemical properties of the translocon, and lipid composition in the regulation of integral membrane protein topogenesis.

- **Molecular mechanism of protein integration and topogenesis.** Using a CG modeling approach that enables direct simulation of co-translational protein translocation on experimental timescale, we studied the regulation of integral membrane protein topogenesis and stop-transfer efficiency. We uncovered multiple kinetic pathways for protein integration, and a Type II flipping pathway in which the nascent protein undergoes slow-timescale reorientation, or flipping, in the confined environment of the translocon channel is responsible for the experimentally observed kinetic dependence of protein topology on ribosomal translation rate and protein length. We further demonstrated that sigmoidal dependence of stop-transfer efficiency on TM hydrophobicity arises from local equilibration of the TM across the translocon LG, and final commitment of the TM to integration product is subject to long-timescale

and long-ranged kinetic regulation.

5.2 Future Work

The CG model we introduced in chapter 4 provides a powerful simulation framework that enables direct interrogation of the dynamics of protein translocation and integral membrane protein topogenesis at the single molecule level. As summarized above, our approach provides unprecedented insight regarding the regulation of membrane partition for a single TM segment, and its orientation with respect to the membrane. Future work to extend the study to the biogenesis of multispanning integral membrane protein will be of great interest. The significant complication going from single- to multispanning membrane protein hinders intuitive interpretation of experimental results, and we expect simulation studies to be critical in unveiling the underlying molecular mechanism.

A multispanning integral membrane protein by definition has multiple TM segments, and a natural question is how the different TM segments are integrated into the membrane bilayer. There is unfortunately no straightforward answer to this simple question, and two contradicting models exist, both supported with some experimental evidence. The earliest and the most intuitive one is the *sequential* model proposed by Blobel [114]. As shown in Figure 5.1A, the TMs transfer from the translocon to the lipid interior independently with corresponding orientations as in the final product. The orientation of the entire integral protein is fixed at the end of translation and is determined solely by the first TM segment. An excellent example of the sequential model is aquaporin AQP4, whose topology was shown to be established one helix at a time via cross-linking experiments with truncated

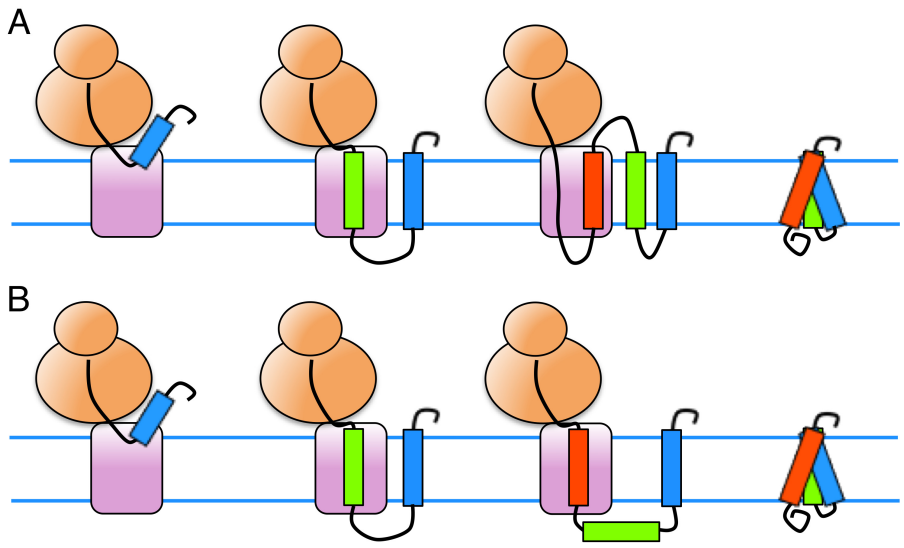


Figure 5.1: Co-translational integration for multispanning integral membrane proteins. **(A)** Sequential model. **(B)** Flipping model. See text for detailed discussion.

integration intermediates [115]. In an alternative *flipping* model, however, the TMs can integrate with incorrect orientation or even translocate to the ER lumen temporarily during the early stage of translation (Figure 5.1B). Not until a later stage of translation, or even after termination of translation will these TMs convert back to the correct topology through large-scale reorientation. This model has been used to explain topological dependence of front TM segments on distant residues that are not even translated at the time when these TMs partition into the membrane [116–118]. Though the macroscopic pictures of the two pathways are clear, many of the molecular details remain to be revealed.

Using the CG model we developed, we will directly simulate the co-translational integration of multispanning membrane protein. We will address the question what properties of a given protein determine which pathway it will choose for integration, and how will this propensity be regulated under the change of ribosomal translation rate, physicochemical properties of the translocon, and lipid composition. More generally, we seek to

reconcile positive-inside rule in the context of these pathways, and to relate the established topological rules for single-spanning membrane proteins to the multispanning counterparts. These questions are hard to address experimentally, as they require direct characterization of transient dynamical intermediates along the co-translational process. We will also strive to make better connection with experiments by proposing unique signatures of different pathways that are experimentally measurable.

A crucial step of the flipping model is the reorientation of the TM segments with incorrect topology, and under what environment this occurs is still under debate. Though the lipid composition induced reorganization of lactose permease protein topology strongly argues for a lipidic environment [117], general acknowledgement of this notion is hindered due to the conceivably substantial barrier for translocating charged and/or hydrophilic peptide groups across membrane bilayer. Determination of the precise barrier for flipping a TM segment across membrane has proven to be challenging experimentally, in part due to uncertainty on the role of other TMs in facilitating this transition. Computer simulation with atomistic force field, when assisted with rare-event sampling techniques, can in principle provide accurate estimation of the energetic barrier without any prior mechanistic assumption of the underlying reaction. For example, with string method and Markovian milestoning, one can calculate the free energy barrier, and mean first passage time of a reaction from first principle. Future work of applying these methods to investigate the regulation of the reorientation rate of a TM helix across the lipid bilayer by proton motive force, lipid composition, and neighboring TMs will be of great interest.

Another important subject of membrane protein study is the topology prediction from

amino acid sequence, and we expect incorporating the mechanistic insight from direct simulation of co-translational membrane protein integration will help to improve the accuracy of these prediction algorithms. Currently, even the most sophisticated algorithm available can only predict 80% of the TM segments with correct topology [119]. A major drawback of these methods is a lack of physical foundation for the ad hoc choice of various parameters. One typical example is the hydrophobicity scale of amino acid residues that is used to identify the TM segments. There are quite a few distinct data sets available and which one to choose is unclear. Improved understanding of the co-translational integration mechanism will provide further guidance on this choice. For example, if the sequential model is the dominant pathway for integration, and the TMs partition directly from translocon to the membrane interior, then the "biological hydrophobic scale" determined from the von Heijne lab shall be used [5]. On the other hand, if the protein is subject to large-scale reorganization of its TM segments in the lipid bilayer post-translationally as in the flipping model, then the Wimley-White scale [120] that directly probe the water to lipid hydrophobic core transfer FE is more appropriate in determining which TM eventually stay in the membrane.

In summary, we expect the CG model we developed will prove useful in extracting the general mechanistic principle for the Sec-facilitated integration of multispanning integral membrane protein.

Appendix A

Supporting Information for Chapter 2

A.1 Atomistic Simulations

All atomistic simulations were implemented within the TCL scripting protocol of the NAMD package [121].

All simulations were performed on the Sec channel from the archaeal *Methanococcus jannaschii* species, for which a high-resolution crystal structure of has been reported [10]. The MD protocol for that we employ follows closely that of Gumbart and Schulten [27]. The channel was simulated in an explicit membrane composed of 254 POPC lipid molecules and an explicit solvent of 24296 rigid water molecules. Interactions were described using the CHARMM27 force field [19], including the TIP3P model for the water molecules. Counterions (Na^+ and Cl^-) were included to achieve electroneutrality and a salt concentration of approximately 50 mM (see Figure A.1). The protonation state of the histidine residues was chosen to be neutral. The initial system contains 115402 atoms in a simulation cell of size 110 Å $\times$ 110 Å $\times$ 100 Å. The system was described using orthorhombic periodic boundary conditions. Long-range electrostatics were calculated using the PME technique [20]. From the initialized configuration, a 5 ns MD simulation was

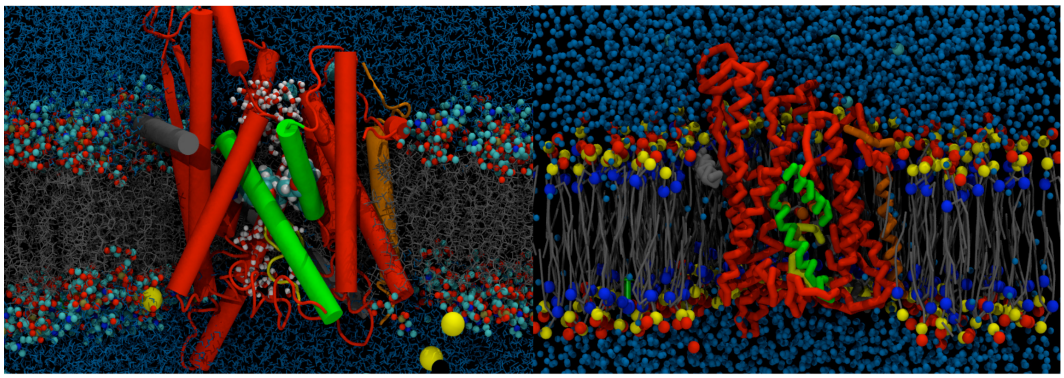


Figure A.1: Snapshots of the all-atom (**Left**) and CG (**Right**) simulation systems. The translocon (SecY in red, SecE in grey and SecG in orange) is shown in cartoon representation, with the two LG helices colored in green. Water molecules are drawn as blue beads and lipid molecules are drawn as black lines.

performed to relax the system for production runs. This simulation was composed of a 2000-step minimization, followed by a 0.5 ns NVT simulation with harmonic restraints (k = 2.0 kcal/mol/Å^2) applied to all atoms except the lipid tails, followed by a 1 ns NPT simulation with harmonic restraints applied only to the protein backbone, followed by a 3.5 ns NPT simulation with no restraints.

Production runs were performed in the NPT ensemble, with Langevin dynamics (damping coefficient 5 ps^{-1}) to keep the system at 300 K and with Nose-Hoover Langevin barostat (damping period 200 fs, damping time 100 fs) [122, 123] to maintain the pressure at 1 atm. The dynamics were integrated using a multiple-time-stepping approach [124], in which a 1 fs timestep was used for bonded interactions, a 2 fs timestep was used for short-range nonbonded interactions, and a 4 fs timestep was used for long-range electrostatic interactions. Short-range interactions were truncated at a distance 12 Å, using a smoothing function in the range of distances from 10 to 12 Å.

A.2 Coarse-grained Simulations

All CG simulations were implemented within the TCL scripting protocol of the NAMD package [121].

We employ a CG representation that combines the MARTINI coarse-graining algorithm for the lipid molecules, water molecules, and ions [22] and the residue-based coarse-graining scheme of Shih *et al.* for the amino acid residues [21]. The CG system was initialized by positioning the CG particles for the amino acid residues and the lipid molecules at the center of mass of the corresponding moieties in atomistic model. The system was then solvated with CG particles representing water molecules and the counterions. The ionic charge and atom types were set to be the same as in the corresponding atomistic system. The final CG system contained a total of 9882 particles. All CG simulations were run in the NPT ensemble, with Langevin dynamics (damping coefficient 5 ps^{-1}) to keep the system at 323 K and Langevin piston (damping period 5 ps, damping time 2.5 ps) to maintain the pressure to 1 atm. A temperature of 323 K, rather than 300 K, was used because CG simulations at the higher temperature were found to better reproduce all-atom simulations of the lipid structure at 300 K [21, 22].

From the initialized configuration, the system was relaxed in preparation for production runs. A 5000-step minimization was first performed with the protein backbone and lipid heads harmonically restrained (k = 1.0 kcal/mol/$Å^2$). Then, a 5 ns CG MD simulation run was performed with a 5 fs timestep and with the protein backbone and the lipid heads harmonically restrained (k = 1.0 kcal/mol/$Å^2$ and 0.2 kcal/mol/$Å^2$, respectively). Finally, the whole system was released to relax along a 10 ns long trajectory with a 10 fs timestep.

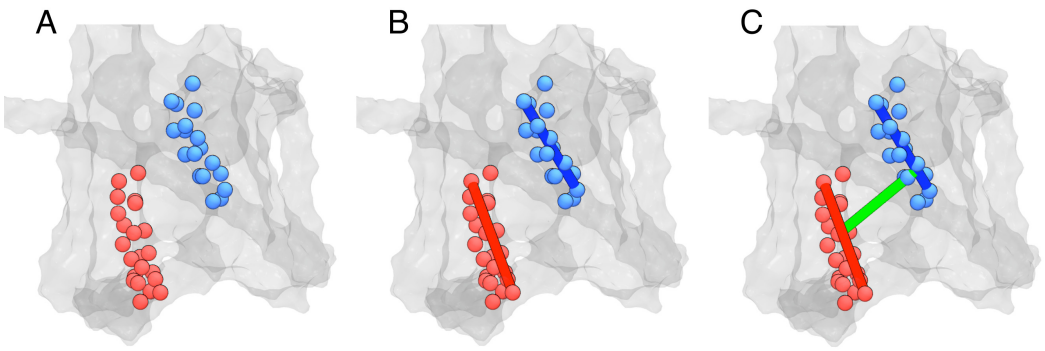


Figure A.2: The LG distance collective variable. (A) Residues forming the two LG helices: TM2b (blue), and TM7 (red). (B) Least-squares-fit lines for each helix. (C) The minimum distance (green) between the two fit lines.

All other CG simulations were also performed using a 10 fs timestep.

A.3 Collective Variables

Detailed definitions and illustrations for all collective variables employed in chapter 2 are presented here. In general, if an α-carbon is used to define a collective variable in the atomistic representation, then the corresponding backbone CG particle is used in the CG representation.

A.3.1 Lateral Gate Distance

The LG distance, d_{LG}, is defined as the distance of minimum approach between the line of least-squares fitting for the α-carbons of residues in the TM2b helix (residues Ile^{75}, Gly^{76}, Val^{79}, Thr^{80}, Ile^{84}, Leu^{87}, Ser^{91}, Gly^{92} in the α-subunit) and the corresponding fitting line for residues in the TM7 helix (residues Ile^{257}, Pro^{258}, Ile^{260}, Leu^{261}, Ala^{264}, Leu^{265}, Asn^{268}, Leu^{271}, Trp^{272}, Ala^{275}, Leu^{276}, Arg^{278} in the α-subunit).

An illustrated explanation of d_{LG} is provided in Figure A.2. In Figure A.2A, the

residues shown in blue correspond to TM2b helix, and residues shown in red correspond to the TM7 helix. In Figure A.2B, the blue line corresponds to h_1, the least-squares-fit line through the α-carbons of the residues used to define the TM2b helix; the red line corresponds to h_2, the least-squares-fit line through the α-carbons of the residues used to define the TM7 helix. In Figure A.2C, the green segment corresponds to d_{LG}, the distance of minimum approach between lines h_1 and h_2; this distance is calculated using

$$d_{\mathrm{LG}} = |r_{12} + \mu e_2 - \lambda e_1|, \qquad (r_{12} = r_2 - r_1), \tag{A.1}$$

where

$$\begin{aligned} \lambda &= [r_{12} \cdot e_1 - (r_{12} \cdot e_2)(e_1 \cdot e_2)] \,/\, [1 - (e_1 \cdot e_2)^2], \text{ and} \\ \mu &= -[r_{12} \cdot e_2 - (r_{12} \cdot e_1)(e_1 \cdot e_2)] \,/\, [1 - (e_1 \cdot e_2)^2]. \end{aligned} \tag{A.2}$$

Here, r_1 is an arbitrary point on h_1 and e_1 is the unit vector that is parallel to h_1; r_2 and e_2 are similarly defined.

A.3.2 Pore-Plug Distance

The PP distance, d_{PP}, is defined as the distance between the center of mass of the α-carbons for the residues that comprise the isoleucine ring of the channel (Ile^{75}, Val^{79}, Ile^{170}, Ile^{174}, Ile^{260}, and Leu^{406} in the α-subunit) and the center of mass of the α-carbons for the residues of the plug domain (Ile^{55}-Ser^{65} in the α-subunit).

An illustrated explanation of d_{PP} is provided in Figure A.3. In Figure A.3A, the residues shown in blue correspond to the pore of the channel, and the residues shown in red correspond to the plug moiety. In Figure A.3B, the blue bead, p_1, corresponds to the center of

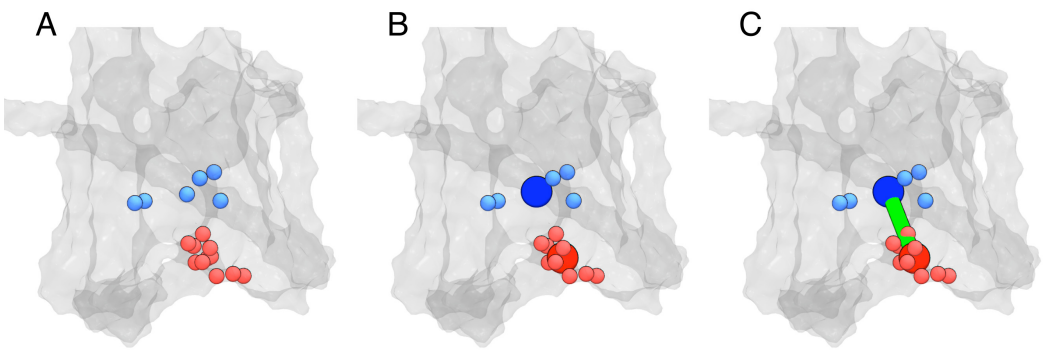


Figure A.3: The PP distance collective variable. (A) Residues forming the pore (blue) and plug (red). (B) The center of mass of the pore residues (dark blue bead), and the center of mass of the plug residues (dark red bead). (C) The center of mass distance between the pore and the plug (green line).

mass of the α-carbons that define the channel pore; the red bead, p_2, corresponds to the center of mass of the α-carbons that define the channel plug. In Figure A.3C, the green segment corresponds to d_{PP}, the distance between points p_1 and p_2.

A.3.3 Plug-Peptide Orientation Parameter

The plug-peptide orientation parameter, θ, is defined to be the angle between a vector v_1 that points from the peptide substrate to the plug moiety and a vector v_2 that points outward from the the opening of the LG. If $\cos(\theta) > 0$, then the plug is between peptide and the LG, as is shown for the snapshot of the hydrophilic peptide in Figure 2.6A of chapter 2. For $\cos(\theta) < 0$ corresponds to the reversed orientation in which the peptide is between the plug and the LG.

An illustrated explanation of θ is provided in Figure A.4. Figure A.4A shows the two LG helices (TM2b and TM7) in green and the rest of the channel in gray. The TM2b helix is defined in terms of the residues Ile^{75}-Gly^{92} in the α-subunit of the translocon, and the TM7 helix is defined in terms of the residues Ile^{257}-Arg^{278} in the α-subunit of the

translocon. Figure A.4B shows the lines h_1 and h_2, which are the least-squares fits through the α-carbons of the residues that compose the helix TM2b and TM7, respectively. Figure A.4C introduces the vector $n_1 = h_1 \times h_2$ (red), and Figure A.4D shows n_2 (blue) which is aligned with the z-axis of the simulation cell (and is perpendicular to the plane of the lipid bilayer). Together, the vectors n_1 and n_2 define the plane of the LG that separates the channel interior and the membrane exterior. Finally, Figure A.4E shows $v_2 = n_1 \times n_2$ (green), which is the vector that points outward from the opening of the LG.

In Figure A.4F, the residues shown in orange, referred to as the “lower residues" of the peptide substrate, are determined as follows. For any configuration of the system, we consider the Cartesian coordinates for the α-carbons of the peptide substrate; the lower residues are defined to be those 15 substrate residues with the lowest values of the Cartesian coordinate along the z-axis. The residues shown in red correspond to the translocon plug moiety (Ile^{55}-Ser^{65} in the α-subunit). In Figure A.4G, the orange bead, p_2, is the center of mass for the 15 α-carbons from the lower residues; the red bead, p_3, is the center of mass of the α-carbons for the plug moiety. In Figure A.4H, the arrow v_1 (yellow) connects p_2 and p_3, pointing from the peptide substrate to the plug moiety. We then obtain

$$\cos(\theta) = \frac{v_1 \cdot v_2}{|v_1||v_2|}. \tag{A.3}$$

A.3.4 Lateral Gate Surface Area

The LG surface area is illustrated in Figure A.5 and is calculated as follows. The z-axis is first uniformly discretized at a resolution of Δz between the bottom, z_0, and top, z_N, of

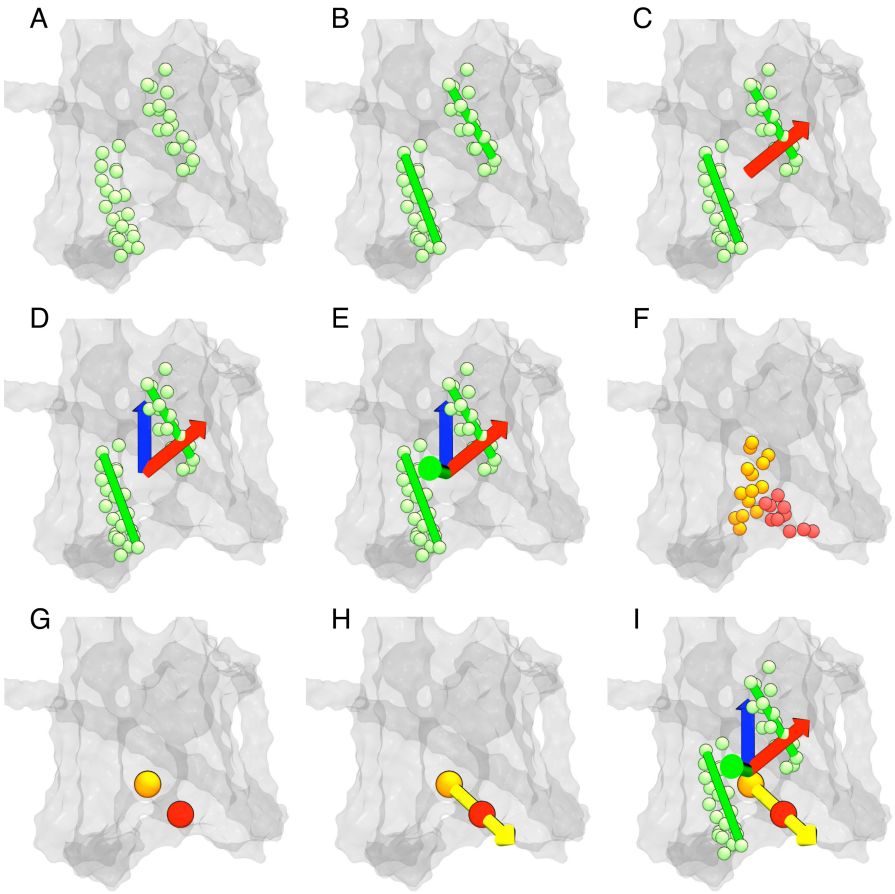


Figure A.4: The plug orientation order parameter. (A) Residues forming the two LG helices. (B) The least-square fit lines for the TM2b and TM7 helices, h_1 and h_2, respectively. (C) The vector $n_1 = h_1 \times h_2$ (red). (D) The vector n_2 that is aligned with the z-axis of the simulation cell (blue). (E) Vector $v_2 = n_1 \times n_2$ pointing outward from the opening of the LG (green). (F) Residues forming the lower half the peptide substrate (orange), and residues forming the plug moiety (red). (G) Centers of mass for the lower peptide residues (orange bead) and the plug residues (red bead). (H) The vector v_1 pointing from the peptide substrate to the plug moiety (yellow). (I) Combined figure showing the relative direction of the v_1 and v_2 vectors, which define the plug-peptide orientation parameter (equation (A.3)).

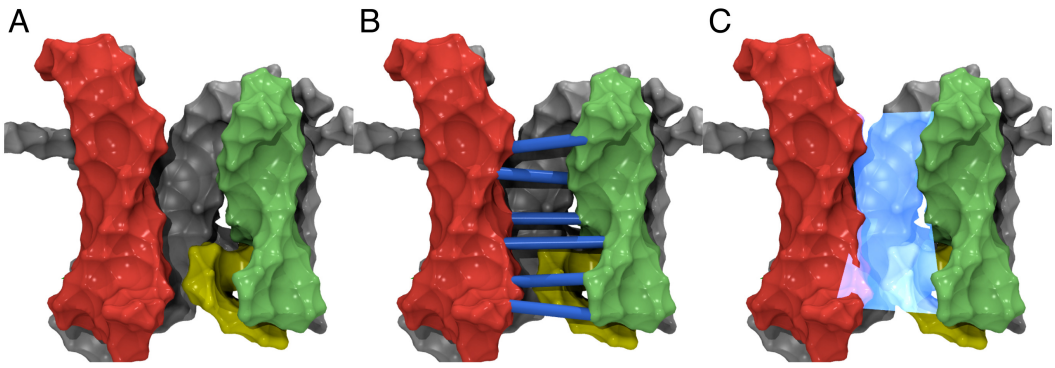


Figure A.5: Illustration of the LG surface area. (A) TM7-9 (Lys^{250}-Gly^{400} in the α-subunit) are shown in red, TM2b-4 (Met^{70} to Ile^{160} in the α-subunit) are shown in green, and the plug moiety (Leu^{40} to Met^{70} in the α-subunit) is shown in yellow. (B) The width of the LG opening at various points along the z-axis. (C) The surface area is obtained by integrating the width of the LG opening over the z-axis, as is described in the text.

the lipid bilayer; z_0 and z_N are defined in terms of the centers of mass for the lipid head groups of each leaf of the bilayer. For each discretized value along the z-axis, z_j, there is a corresponding slab that is parallel to the x-y plane, that is of thickness Δz, and that is centered around z_j.

The width of the LG opening for each slab is determined by considering the backbone CG particles within the z_j slab of the simulation cell and within two particular subsets of the translocon residues. The first subset, shown in red in Figure A.5A, includes residues in TM7-9 (Lys^{250}-Gly^{400} in the α-subunit). If $\cos(\theta) < 0$, the second subset, shown in green in Figure A.5A, includes residues Met^{70} to Ile^{160} in the α-subunit. If $\cos(\theta) > 0$, the second subset, shown in both green and yellow in Figure A.5A, includes residues Leu^{40} to Ile^{160} in the α-subunit. The width of the LG opening width for a given slab, w_j, is defined as the minimum distance from any CG particle in the first subset to any particle in the second subset; this definition accounts for the effect of the plug-substrate orientation on the LG surface area.

The LG surface area is finally obtained from the sum $\sum_{j=1}^{N} w_j \Delta z$, where $N = 20$. This is illustrated in Figure A.5B and C.

A.4 Initializing the Peptide Substrate

The hydrophobic (Leu_{30}) and hydrophilic (Gln_{30}) peptides were initialized as idealized α-helices with all-atom resolution using PyMOL [125]. The idealized α-helices for both peptides were built with Ramachandran angles of ($\phi = -60°$, $\psi = -45°$). Simulations including the peptide substrate were initialized by inserting the idealized α-helix into a configuration for the channel with the plug displaced (i.e., a configuration of the translocon with $d_{PP} = 20$ Å and $d_{LG} = 6$ Å that was drawn from the restrained MD simulations used to calculate Figure 2.3 in chapter 2). The helix was positioned in the channel by aligning it with the z-axis of the simulations cell (perpendicular to the lipid bilayer) and placing the center of mass of the helix at the same position as the center of mass of the channel pore residues (Ile^{75}, Val^{79}, Ile^{170}, Ile^{174}, Ile^{260}, and Leu^{406} in the α-subunit). Having initialized the channel-substrate system with full atomistic resolution, the system was mapped onto the CG representation as described in "Coarse-grained simulations" and then equilibrated for 700-800 ns. As is discussed in the text, the centers of mass for the backbone CG particles of the substrate and the channel pore residues were tethered to each other with a weak harmonic restraint of 0.5 kcal/mol/Å^2 to allow for arbitrary long simulations without the possibility of peptide diffusion out of the channel.

The long (over 700 ns) equilibration timescale for these simulations allows for extensive sampling of the peptide and translocon configuration space, such that the calculated

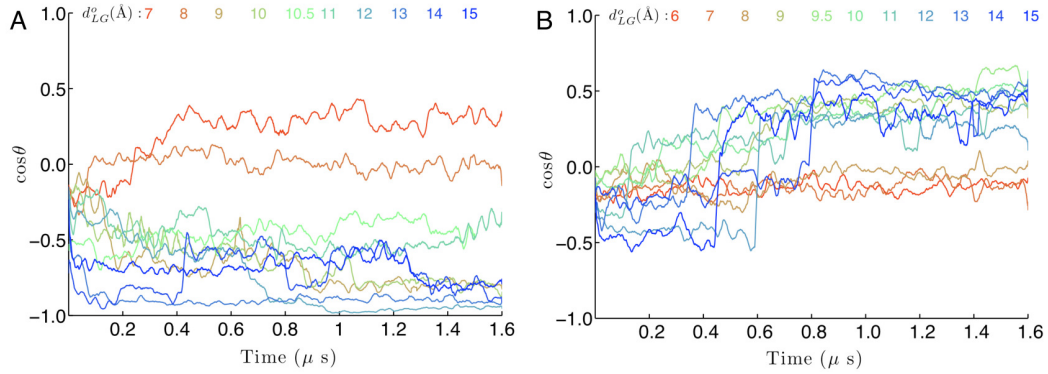


Figure A.6: The CG MD sampling trajectories for the translocon in the presence of (A) the hydrophobic substrate and (B) the hydrophilic substrate, plotted as a function of the plug-peptide orientation parameter. Each trajectory is restrained to a different value for the LG distance, indicated by color. All curves correspond to 10 ns rolling averages of the simulation data. See text for details.

FE profiles are not dependent on the details of the initialization protocol described above. The slowest relaxation timescale that was found during equilibration corresponds to the relative orientation of the plug residue and the peptide substrate. This point is illustrated in Figure A.6, in which the plug-peptide orientation parameter is plotted as a function of the simulation time for the sampling trajectories. In part A, it is seen that the trajectories with the hydrophobic substrate relax relatively quickly (within about 400 ns) with respect to the initial orientation of the substrate and the plug. However, for many of the trajectories with the hydrophilic substrate (part B), the initial configuration appears to be a metastable conformation that eventually relaxes on a longer timescale. For the trajectories for which the LG distance is restrained to $d^\circ_{\mathrm{LG}}/\text{Å} = 9$, 9.5, 10, 11, 13, and 15, it was found that that the plug moiety and peptide substrate undergo an abrupt, kinetically frustrated reorientation on the timescale of hundreds of nanoseconds. Recognizing this clear tendency for reorientation, the $d^\circ_{\mathrm{LG}}/\text{Å} = 12$ trajectory was reinitialized from the $d^\circ_{\mathrm{LG}}/\text{Å} = 11$ trajectory at 600 ns, and the $d^\circ_{\mathrm{LG}}/\text{Å} = 14$ trajectory was reinitialized from the $d^\circ_{\mathrm{LG}}/\text{Å} = 13$ trajectory at 800

ns. This initially frustrated reorientation seems to be thermodynamically favorable for the hydrophilic substrate, since reorientation in the reverse direction was never observed.

Finally, we emphasize that the definition of the FE profiles reported in Figures 2.3, 2.4 and 2.7 in chapter 2 make no assumptions about the configuration of the peptide within the channel. The error bars in these plots, obtained from 160 ns block averages of the production data, and the trajectories in Figure A.6 suggest that the thermal distribution of configurations has been thoroughly sampled. However, as in any molecular simulation of a complex system, it is possible that 1.5+ μs trajectories are not adequate to discover all thermodynamically dominant configuration of the system; at the very least, the calculations reported in Figures 2.4 and 2.7 offer a meaningful characterization of long-lived basins of stability.

A.5 Side chain Transfer Free Energies for the CG Residues

A measure of the accuracy of the residue-based CG models is obtained by considering the hydrocarbon/water transfer free energies for the amino acid residues. Recent simulation studies have demonstrated that the MARTINI [126] and the Sansom [127] residue-based coarse-graining methods, which are closely related to the CG potential employed here [21], give rise to side chain transfer free energies that exhibit strong correlation and reasonable absolute agreement with experimental results. For the CG potential employed in the current study [21], we have also calculated the transfer free energies for all of the amino acid side chains.

The transfer free energies for CG amino acid side chains is obtained from the difference

of the side chain solvation free energies (ΔG) in water and in nonpolar solvent. Each solvation FE is calculated using the FE perturbation formula [128].

$$\Delta G^{\text{solv}} = -k_{\text{B}}T \sum_{i=1}^{N} \ln \left\langle e^{-\beta[\mathscr{H}(\mathbf{x},\mathbf{p};\lambda_{i+1})-\mathscr{H}(\mathbf{x},\mathbf{p};\lambda_i)]} \right\rangle_i , \tag{A.4}$$

where

$$\mathscr{H}(\mathbf{x},\mathbf{p};\lambda) = \lambda \mathscr{H}_{\text{solv}}(\mathbf{x},\mathbf{p}) + (1-\lambda)\mathscr{H}_{\text{vac}}(\mathbf{x},\mathbf{p}), \tag{A.5}$$

and where the angle brackets correspond to the ensemble average for the system with Hamiltonian $\mathscr{H}(\mathbf{x},\mathbf{p};\lambda_i)$. The classical Hamiltonians for the system with and without interactions between the solvent and the side chain are given by $\mathscr{H}_{\text{solv}}$ and $\mathscr{H}_{\text{vac}}$, respectively. We employed the NAMD implementation of this method.

Each solvation FE was obtained from $N = 20$ ensemble averages of a system containing 1878 CG solvent molecules and one CG side chain particle, with $\lambda_i = \{0, 0.00001, 0.0001, 0.001, 0.05, 0.10, 0.15, \ldots, 0.85, 0.90, 0.95, 0.99, 0.999, 0.9999, 0.99999\}$; the additional values of λ_i in the limits approaching $\lambda \to 0$ or 1 were included to avoid numerical instabilities. Each ensemble average was calculated from a 2 ns MD trajectory at constant temperature (300 K) and constant pressure (1 atm) with a timestep 40 fs. The first 40 ps of the trajectory was discarded as equilibration. The potential energy parameters for the nonpolar solvent particles are the same as those for the CG particles in saturated lipid tails (i.e., the hydrophobic-apolar CG particle type). It was confirmed that the calculated free energies are converged with respect to equilibration time, the MD timestep, and the number of discretizations in λ.

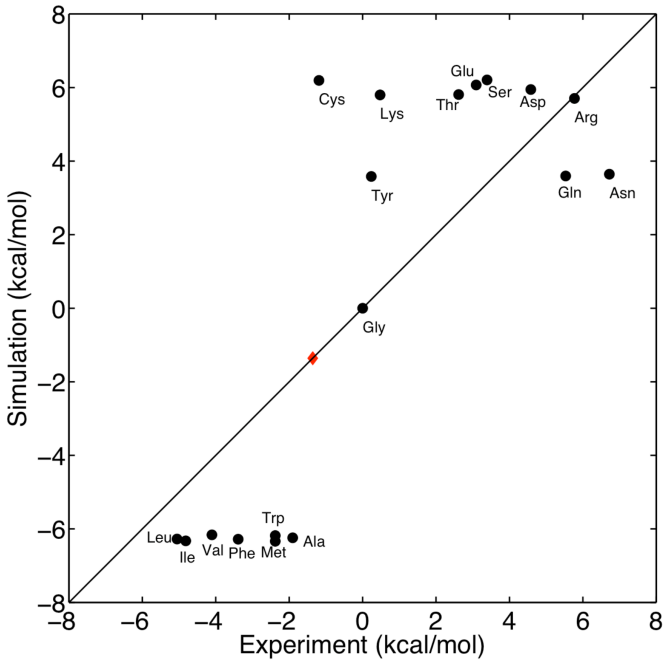


Figure A.7: Correlation plot between the oil/water transfer free energies for the amino acid side chains obtained experimentally and from the CG simulations. The red diamond indicates the transfer FE for an additional CG side chain particle of intermediate hydrophobicity.

The results for the transfer free energies of the amino acid side chains in the CG model are presented in Figure A.7. The statistical error for the simulations was typically of the size of the plotted symbols. The experimental results correspond to cyclohexane/water side chain transfer free energies [129]. Comparison of these results reveals reasonable correlation between the CG model and the experimental results. With regard to the amino acid residues that form the peptide substrates in our simulations, the Leu transfer free energies are within 1 kcal/mol of the experimental results, whereas the CG model underestimates the hydrophilicity of the Gln residues by approximately 2 kcal/mol. Although there is much room to improve the CG model, the accuracy suggested in Figure A.7 is not inconsistent with the level of accuracy observed in fully atomistic models. In particular, precise atomistic simulations have demonstrated that the choice of all-atom water potential leads to

deviations of up to 1.5 kcal/mol in amino acid side chain solvation free energies [130], and the choice of molecular mechanics force field for the amino acid changes the calculated solvation FE by over 1 kcal/mol in many cases [131]. The calculations presented in Figure A.7, along with the comparison between the atomistic and CG FE profiles in Figures 2.3 and A.8, suggest that the CG potentials employed in this study form a reasonable basis for the qualitative interpretation of the simulation results.

For other tests reported in this supporting information, it is useful to have a CG side chain particle that is of intermediate hydrophobicity with respect to the Leu and Gln residues. We developed such a side chain particle with the same potential energy functional form as the other CG side chain particles [21]. The interaction parameters for this “Int" CG side chain particle are presented in Table A.1, and the corresponding transfer FE is reported as the red diamond in Figure A.7.

A.6 Scaffolding Contribution to the Free Energy Profile

Figure 2.3 in chapter 2 presents the FE profile for the translocon as a function of d_{LG} and d_{PP}. Here, in Figure A.8A, these results are replotted with error estimates. The red/yellow-shaded surface corresponds to the atomistic FE profile, and the blue/orange-shaded surface below it corresponds to the CG FE profile. The error bars correspond to the standard deviation of the mean FE profile obtained from the five block averages of the simulation data.

To investigate the impact of the scaffolding interactions on the calculated CG FE profile, it was recalculated for the CG model without scaffolding. Following the same protocol as

Table A.1: Potential energy parameters for the interaction of the particle of intermediate hydrophobicity with the other CG particle types.

CG Particle Type	CG Particle Type [a]	Lennard-Jones Energy-scale, ε (kcal/mol)	Lennard-Jones Lengthscale, [b] $R_{\min}$ (Å)
Int	C	-0.621	5.300
Int	Nx	-0.812	5.300
Int	No	-0.813	5.300
Int	Nd	-0.812	5.300
Int	Na	-0.812	5.300
Int	Qx	-0.812	5.300
Int	Qo	-0.621	5.300
Int	Qd	-0.717	5.300
Int	Qa	-0.717	5.300
Int	P	-0.717	5.300
Int	Nxx	-0.812	5.300
Int	Nxg	-0.812	5.300
Int	Ca	-0.621	5.300
Int	Qdr	-0.717	5.300
Int	Nxn	-0.812	5.300
Int	Qad	-0.717	5.300
Int	Pc	-0.717	5.300
Int	Nxq	-1.195	5.300
Int	Qae	-0.717	5.300
Int	Ph	-0.717	5.300
Int	Qdh	-0.717	5.300
Int	Ci	-0.621	5.300
Int	Int	-1.195	5.300
Int	Cl	-0.621	5.300
Int	Qdk	-0.717	5.300
Int	Cm	-0.621	5.300
Int	Cf	-0.621	5.300
Int	Cp	-0.621	5.300
Int	Ps	-0.717	5.300
Int	Pt	-0.717	5.300
Int	Cw	-0.621	5.300
Int	Nxy	-0.812	5.300
Int	Cv	-0.621	5.300
Int	Nap	-0.812	5.300
Int	CDB	-0.621	5.300

[a] CG particle-type names are consistent with Ref.21.

[b] $R_{\min}$ is related to the usual Lennard-Jones lengthscale via $R_{\min} = \sigma 2^{1/6}$.

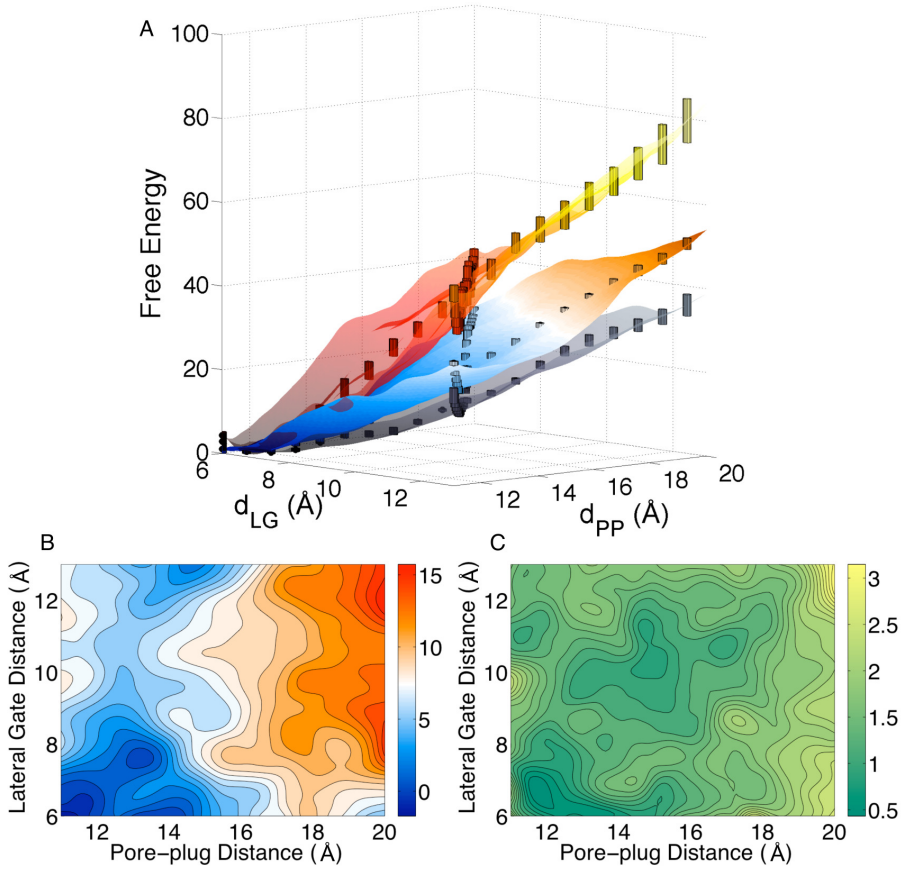


Figure A.8: Free energy profiles for the translocon as a function of the LG and PP distances, calculated using the atomistic potential (A, red/yellow-shaded) the CG potential with scaffolding (A, blue/orange-shaded), and the CG potential without scaffolding (A, grey). (B) The difference between the FE profiles obtained using the CG potential with and without scaffolding. (C) The statistical uncertainty in this difference. All energies in kcal/mol.

was used to obtain Figure 2.3B, an additional set of 80 CG MD sampling trajectories was performed without scaffolding interactions, each of which was of length 20 ns and was harmonically restrained for the collective variables $d_{\mathrm{LG}}/\text{Å} \in [6,13]$ and $d_{\mathrm{PP}}/\text{Å} \in [11,21]$. Using the WHAM algorithm, the FE profile without scaffolding was constructed and is plotted as the grey surface in Figure A.8A. All three profiles are vertically shifted to have a minimum at 0 kcal/mol. Comparison of the CG FE profiles with and without scaffolding in Figure A.8A indicates that the features of the CG profile are not dramatically altered by the inclusion of the scaffolding interactions.

For a more detailed comparison, the difference between the FE profiles for the CG model with and without scaffolding is plotted in Figure A.8B, and the statistical uncertainty of this difference is plotted in A.8C. The difference between the CG FE profiles appears to be more sensitive to changes in d_{PP} than d_{LG}, and for most of the domain, the difference between the CG surfaces does not exceed the statistical uncertainty by more that 5 kcal/mol. These results indicate that the scaffolding interactions do give rise to some changes in the calculated CG FE profile, although the differences are relatively small in comparison to the other features on the surface.

A.7 Free-Energy Surface Cross Sections

Figure A.9 presents FE profiles as a function of the LG coordinate for fixed values of the PP distance. The red curve in part (B) corresponds to the cross section of the FE profile at 12 Å that is indicated by the red band in part (A). The blue curve in part (B) is similarly obtained from the cross section at 19 Å. The curves in part (B) are vertically shifted to be 0 kcal/mol at their minimum.

A.8 Additional Trajectories

A.8.1 Trajectories without Scaffolding

To confirm that the closing of the Sec translocon with the hydrophilic peptide inside is not an artifact of the scaffolding interactions, we performed additional CG MD trajectories without scaffolding for the translocon initialized from open configurations ($d_{\mathrm{LG}} = 14$ Å) of

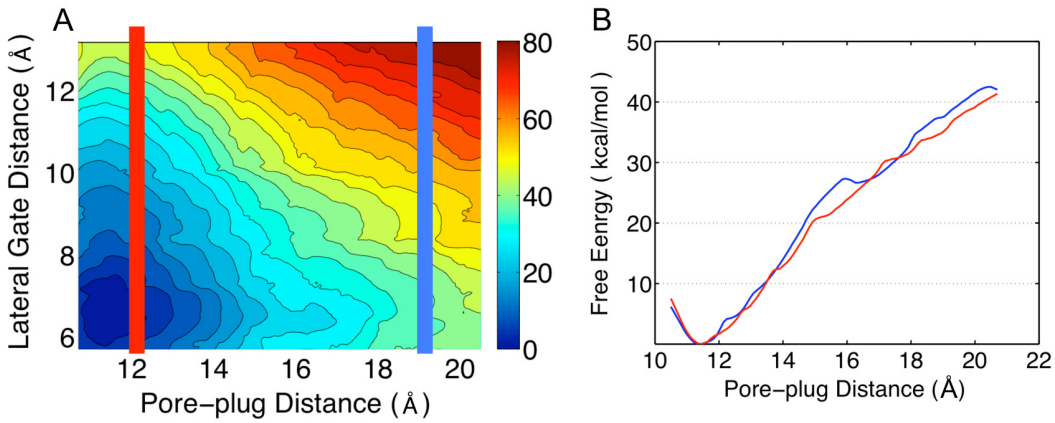


Figure A.9: Free-energy profiles as a function of the LG coordinate for fixed values of the PP distance. The red curve in part (B) corresponds to the cross section of the FE profile at 12 Å that is indicated by the red band in part (A). The blue curve in part (B) is similarly obtained from the cross section at 19 Å. The curves in part (B) are vertically shifted to be 0 kcal/mol at their minimum.

the LG. For both the hydrophobic substrate and the hydrophilic substrate, three independent trajectories of length 500 ns are presented in Figure A.10; as in Figure 2.5, the initial configurations for the trajectories were drawn from the substrate-containing trajectories with scaffolding that were restrained with respect to d_{LG}.

The results in Figure A.10 are broadly consistent with simulations that include scaffolding in Figure 2.5. For the trajectories with the hydrophilic substrate, the initially open LG distance closes during the simulation (Figure A.10A) to the same extent that was seen in Figure 2.5. Furthermore, the LG surface area for the hydrophilic trajectories trend downwards over the simulated timescale, although to varying degrees, and the substrate is found to remain within the translocon channel.

The trajectories with the hydrophobic substrate show a greater range of behavior. In one case (royal blue), the initially open LG remains fully open over the course of the simulation, as was seen in Figure 2.5. In a second case (dark blue), the hydrophobic peptide exits the

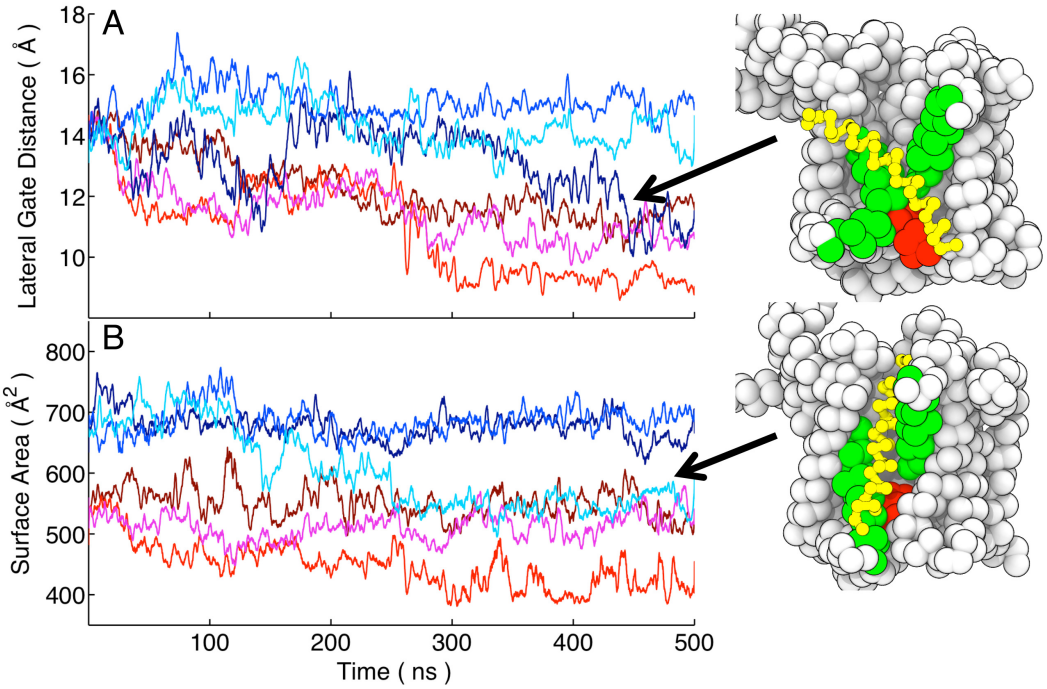


Figure A.10: CG MD trajectories without scaffolding for the translocon containing either the hydrophobic (blue-shaded) or hydrophilic (red-shaded) substate are initialized from open configurations of the LG. Three independent trajectories for each substrate are performed. (A) The LG distance d_{LG} for the trajectories is plotted as a function of simulation time. (B) The LG surface area for the trajectories is plotted as a function of simulation time. The lines indicate 1 ns rolling averages. Also shown are snapshots from the two trajectories in which the hydrophobic substrate partially exits from the channel; the substrate is indicated in yellow, the LG helices are indicated in green, and the plug moiety is indicated in red.

channel, and the LG distance closes behind it in such a way that the LG surface area remains relatively large. In a third case (light blue), the peptide partially exits the channel, and the LG surface area closes in such a way that the LG distance remains relatively large. In interpreting these results for the hydrophobic substrate, it must be remembered that the CG model without scaffolding does not fully preserve structural features of the translocon on these long timescales (Figure 2.2C); it is possible that unphysical distortions in the channel are facilitating the exit of the substrate.

A.8.2 Trajectories with Substrate of Intermediate Hydrophobicity

The trajectories presented in Figure 2.5 in chapter 2 illustrate the metastability of the translocon in the presence of strongly hydrophobic and strongly hydrophilic substrates. Here, we explore the metastability of the translocon with a peptide substrate of intermediate hydrophobicity. Six CG MD trajectories of length 500 ns were performed with the substrate Int_{30}, a linear peptide composed of 30 CG amino acids with side chains that exhibit a transfer FE between that of the Leu and Gln residues (see section: "Side chain Transfer Free Energies for the CG Residues," Figure A.7, and Table A.1). As in Figure 2.5, scaffolding interactions for the translocon were employed.

Figure A.11 presents the CG MD trajectories performed with the Int_{30} substrate. These additional trajectories were initialized from the same configurations as the six trajectories reported in Figure 2.5 (which include closed, partially open, and fully open configurations for the LG). The color scheme in Figure A.11 identifies which of the additional trajectories shares the same initial configuration as a given trajectory in Figure 2.5.

Figure A.11 indicates that the Int_{30} substrate supports both the metastable open and closed configurations of the translocon LG. For the two trajectories that are initialized with $d_{\text{LG}} = 6$ Å, this distance remains relatively unchanged over the course of the simulations and the LG surface remains closed in the range of 400-450 Å^2. Similarly, for the two trajectories that are initialized with $d_{\text{LG}} = 15$ Å, this distance remains relatively unchanged over the course of the simulations, but the LG surface areas relaxes into the range of 600-650 Å^2 that is sufficiently open to allow for large substrate exposure to the membrane (Figure 2.5C).

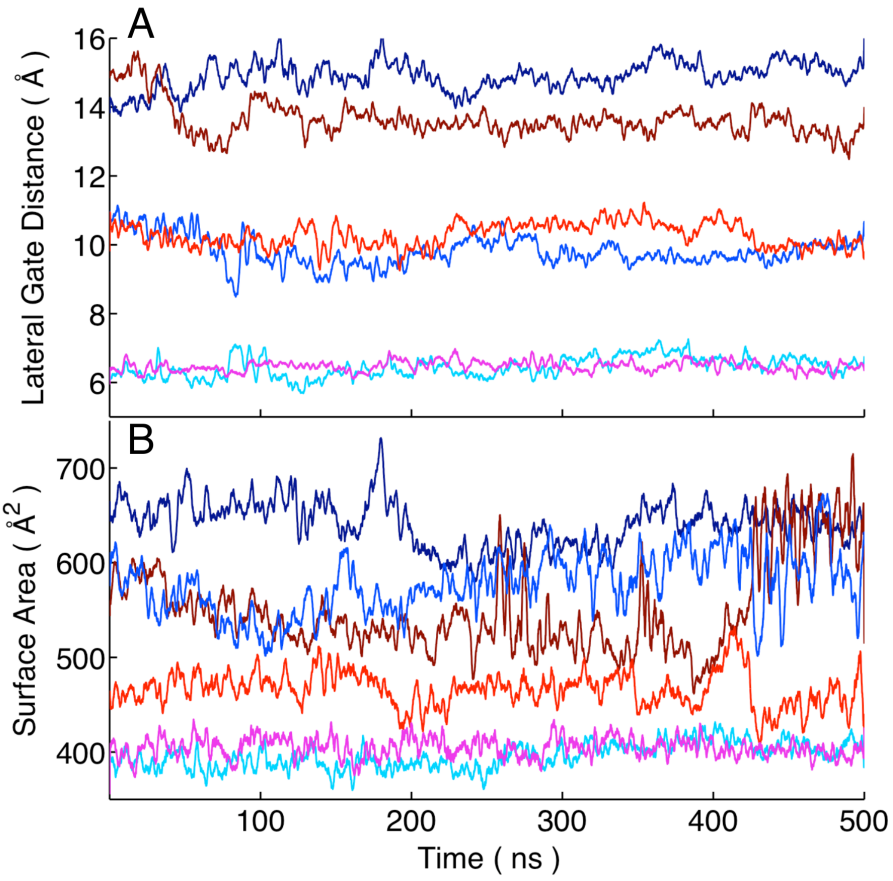


Figure A.11: CG MD trajectories for the translocon with a substrate of intermediate hydrophobicity, plotted as a function of (A) LG distance and (B) LG surface area. These trajectories are initialized from the same configurations as the six trajectories reported in Figure 2.5 (chapter 2), and the color scheme identifies which of the trajectories shares the same initial configuration as a given trajectory in Figure 2.5. The lines indicate 1 ns rolling averages.

Of the two trajectories that were initialized from the partially open LG ($d_{\mathrm{LG}} = 11$ Å), one (plotted in blue) exhibits LG surface area values in the open range (600-650 Å^2), whereas the other (plotted in red) exhibits LG surface area values in the closed range (400-450 Å^2). After 500 ns of simulation time, all six of the trajectories in Figure A.11 exhibit LG surface areas either in the range of 400-450 Å^2 or in the range of 600-650 Å^2, although the LG distance seems to relax on a slower timescale. Consistent with Figure 2.5, the trajectories in Figure A.11 suggest that even in the presence of a substrate of intermediate

hydrophobicity, the LG surface area exhibits two-state behavior with respect to opening and closing of the LG.

A.9 Mutations in the Translocon Pore Residues

Here, we explore how mutations in the translocon pore residues alter the FE cost for opening the LG. These pore residues have been demonstrated to have significant impact on the functioning of the Sec translocon [10, 17]. In a first set of calculations, we replace the six amino acid residues that comprise the hydrophobic translcon pore moiety (Ile^{75}, Val^{79}, Ile^{170}, Ile^{174}, Ile^{260}, and Leu^{406} in the α-subunit) with either six hydrophilic (Gln) residues or six intermediate (Int) residues, and we calculate the corresponding changes in the FE profiles in Figure 2.7. (For a discussion of the Int residues, see section: "Side chain Transfer Free Energies for the CG Residues," Figure A.7, and Table A.1.) The mutated FE profiles are calculated from the simulation data used to construct Figure 2.7, using

$$F_{\mathrm{mut}}(A_{\mathrm{LG}}) = -k_{\mathrm{B}}T \ln P_{\mathrm{mut}}(A_{\mathrm{LG}}), \tag{A.6}$$

where

$$P_{\mathrm{mut}}(A_{\mathrm{LG}}) \propto \langle \delta(A_{\mathrm{LG}} - A_{\mathrm{LG}}(\mathbf{x})) e^{-(U_{\mathrm{mut}}(\mathbf{x}) - U_{\mathrm{wt}}(\mathbf{x}))/(k_{\mathrm{B}}T)} \rangle_{U_{\mathrm{wt}}}, \tag{A.7}$$

$A_{\mathrm{LG}}(\mathbf{x})$ is the LG surface area as a function of the positions x of the CG particles, $U_{\mathrm{wt}}(\mathbf{x})$ is the potential energy surface for the wild-type system, and $U_{\mathrm{mut}}(\mathbf{x})$ is the potential energy surface for the mutated system [132]. The angle brackets indicate thermal averaging on the wild-type potential energy surface. The FE surfaces for the wild-type and mutant translo-

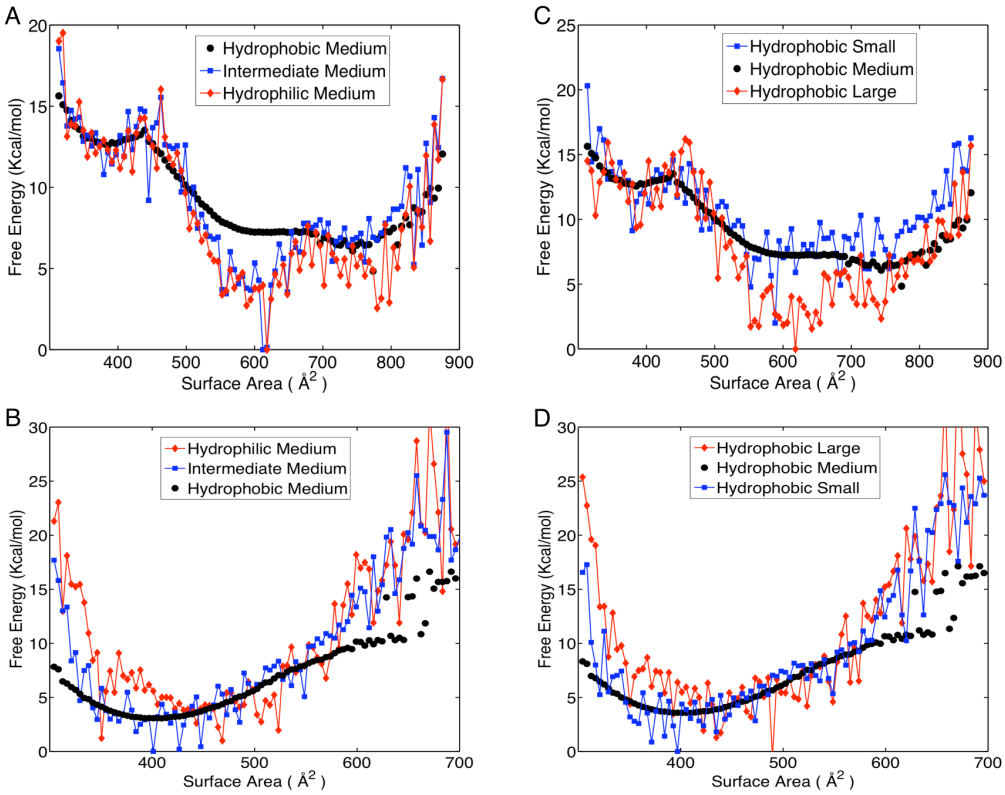


Figure A.12: Mutations in the translocon pore residues alter the FE profiles in Figure 2.7 of chapter 2. (A) and (B) The dependence of the FE profiles on the pore residue hydrophobicity for the translocon with (A) hydrophobic and (B) hydrophilic substrate. (C) and (D) The dependence of the FE profiles on the pore residue bulkiness for the translocon with (C) hydrophobic and (D) hydrophilic substrate. See text for details.

cons are plotted in Figure A.12. Although the variance of the exponential term in equation (A.7) leads to diminished statistical certainty in the FE profiles for the mutants, clear trends seem to emerge.

Figures A.12A and B illustrate the effect of hydrophobicity of the pore residue side chains. Part A shows the results for the translocon with the hydrophobic substrate, and part B shows the results for the hydrophilic substrate. For the case of the hydrophobic substrate, making the pore more hydrophilic favors the open LG. This trend can be easily understood in terms of the preference of the hydrophobic substrate for the membrane environment over the increasingly hydrophilic environment of the mutated translocon channel. However,

part B shows that for the case of the hydrophilic substrate, the changes are less dramatic. Indeed, over the range of 400-600 $\AA^2$ for the surface area, which we find to be the range of LG opening in our MD trajectories, there is very little difference in the FE profiles for the hydrophilic substrate.

We can use a similar analysis to study the effect of the pore-residue bulkiness. We consider two translocon mutants in which all six hydrophobic pore residues are replaced with CG particles for which the Lennard-Jones radius for the CG side chain particle is either increased by 5% or decreased by 10%, and the FE profiles are recalculated using equations (A.6) and (A.7). Figure A.12C shows the resulting profiles for the translocon with the hydrophobic substrate, and Figure A.12D shows the resulting profiles for the hydrophilic substrate. Figure A.12C suggests that increasing the bulkiness of the pore residues leads to the relative stabilization of the open LG. Physically, this is reasonable. With the bulkier pore residues, less room is available inside the channel, which favors the open configurations of the LG in which the hydrophobic substrate partially extends into the membrane (see Figure 2.6B in chapter 2). Figure A.12C suggests that the bulkiness of the pore residue side chains does not significantly change the FE profile in the range of 400-600 $\AA^2$ for the surface area. This can be rationalized by observing that since the hydrophilic substrate is tucked behind the plug residue (Figure 2.6A), the opening of the LG does not relieve the confined environment of the substrate.

The simulations presented in Figure A.12 indicate that the employed CG model is sensitive to the molecular details of the translocon system.

Appendix B

Supporting Information for Chapter 3

B.1 Materials and Methods

B.1.1 Simulation Protocols

Initialization. The crystal structure from Ref. 12 is used as the starting point for simulations of the SecA-SecYEG complex. To meet the size-constraints of the Anton hardware [37, 38], only SecA residues within 15 Å of the translocon in the crystal structure are included in the simulations. Specifically, if any atom within a residue of SecA is less than 15 Å from any atom within a residue of the translocon (SecYEG), then all atoms of that SecA residue are included in the simulations; otherwise, all atoms of that SecA residue are deleted from the simulations. The partial atomic charges of the SecA residues are left unchanged following truncation; however, since the force fields employed in this study exhibit integer values for the net charge of each amino acid residue, the simulation protocol does not introduce any net fractional charges into the simulation cell. In all simulations, the nonhydrogen atoms of SecA are harmonically restrained to their positions in the reported crystal structure with a force constant of $k = 2.0$ kcal/mol/Å^2.

Residues 42-61 of the SecY protein, which are unresolved in the crystal structure, are constructed in a random loop configuration and then refined against a pseudoenergy function using the MODELLER protocol [133] that consists of conjugate gradient minimization and MD with simulated annealing [134, 135]; the employed pseudoenergy function includes interactions from the CHARMM22 force field [19] and restraints based on the statistical distributions of known protein structures [136]. The small number of crystallographically unresolved residues at the C-terminal end of the SecY protein (Res. 424-431), which are not expected to play a significant role in the nascent protein insertion process or in establishing the structural integrity of the translocon channel, are neglected [12].

The SecA-SecYEG complex is embedded in a membrane composed of 222 POPC lipid molecules and surrounded by 25767 explicit water molecules; Na^+ and Cl^- counterions are included to achieve electroneutrality in the simulation cell at a salt concentration of approximately 50 mM. In orienting the SecA-SecYEG complex relative to the membrane, we obtain coordinates from the Orientations of Proteins in Membranes (OPM) database [137], in which the position of the protein relative to the lipid bilayer minimizes the transfer FE from water to the membrane hydrophobic core [138]. The system is described using orthorhombic periodic boundary conditions. The initial system contains 121107 atoms in a simulation cell of size 105 Å$\times$105 Å$\times$120 Å; the membrane bilayer lies parallel to the *x*-*y* plane of the simulation cell.

All equilibration and nascent-protein growth simulations (described immediately below) are performed using the GROMACS molecular package, version 4.5.3 [139]. Interactions are described using the CHARMM36 force field with the TIP3P water model

[19, 140]. Long-range electrostatics are calculated using the PME technique [20], with the real space contribution to this potential cut off at 12 Å. Short-range van der Waals interactions are smoothly switched off over the distances from 10 to 12 Å. The GROMACS neighbor list for all short-ranged interactions is cut off at 12 Å and updated every 20 fs. All bond distances are constrained using the P-LINCS algorithm [6], and a time step of 2 fs is employed. Simulations are performed either at constant temperature and constant volume (i.e., the NVT ensemble) or at constant temperature and constant pressure (i.e., the NPT ensemble). Constant temperature simulations are fixed at 300 K using the Nose-Hoover thermostat [141, 142]; three separate thermostats, each with a coupling constant of 0.4 ps, are applied to the protein, lipids and water molecules, respectively. Constant pressure simulations are fixed at 1 bar using the Parrinello-Rahman barostat [143] with a coupling constant of 4 ps. Pressure coupling is applied semi-isotropically, such that the x and y dimensions of the simulation cell remain equal to each other and deform independently of the z dimension of the simulation cell.

Equilibration. The initial system is equilibrated using the following four-step process. First, the energy of the system is minimized using the steepest-descent method to eliminate steric clashes that lead to forces in excess of 11.9 kcal/mol/Å; during this minimization, harmonic restraints ($k = 2.0$ kcal/mol/Å^2) are applied to the lipid head groups and the heavy atoms of SecYEG. Second, the system is relaxed using a 1 ns simulation in the NVT ensemble with harmonic restraints ($k = 2.0$ kcal/mol/Å^2) applied to the lipid head groups and the heavy atoms of SecYEG. Third, the system is relaxed using a 1 ns simulation in the NPT ensemble with harmonic restraints ($k = 2.0$ kcal/mol/Å^2) applied to the heavy atoms

Table B.1: Cartesian positions for the α-carbon atoms of SecA residues 780-785.

Residue ID	Position (Å)		
	x	y	z
780	-2.073	-9.178	22.870
781	-3.665	-11.383	20.201
782	-6.463	-10.542	17.788
783	-7.417	-6.878	17.970
784	-4.841	-4.080	18.208
785	-2.543	-3.605	21.212

of SecYEG. Finally, the system is relaxed using a 50 ns simulation in the NPT ensemble without restraints on the lipid or SecYEG.

Additional discussion and testing of the equilibration process is provided in section: *Robustness of the insertion trajectory initialization*.

Nascent-protein growth. Following equilibration of the SecA-SecYEG complex, we introduce a nascent protein composed of $n+4$ amino acid residues; the four N-terminal residues are initially positioned in a β-strand configuration and aligned with the axis of the translocon channel, and the center of mass positions of the remaining n residues are initially placed at an “insertion point" at the cytosolic mouth of the translocon channel. As is described in the chapter 3, we consider two insertion points. The insertion points are defined relative to the positions of the atoms of SecA, which are restrained in absolute space to the geometry of SecA-SecYEG crystal structure, as described under *Simulation Protocols*. Coordinates in absolute space associated with the geometry of the SecA-SecYEG crystal structure are given in Table B.1. In terms of this unique coordinate system, the position of insertion point IP2 is $\{x-a, y-b, z-c\}$, where $\{x,y,z\}$ is the Cartesian center of mass for residues 780-785 of the two-helix-finger domain of SecA, and where $a = -6$ Å, $b = 3$ Å, and $c = -1$ Å are chosen to avoid steric clashes; the coordinates for IP1 are simply shifted

+10 Å from IP2 along the z-axis.

Each of the n residues on the C-terminal end of the nascent protein exists in either an off-state, in which nonbonding interactions between each residue and the rest of the system are excluded, or an on-state, in which all interactions are included. Residues in the off-state are tethered to the insertion point via harmonic restraints with $k = 23.8$ kcal/mol/Å^2. Upon sequentially switching each residue from the off-state to the on-state, the simulation cell is subjected to a partial minimization to avoid large steric clashes associated with the newly introduced nascent protein residue; this minimization is only performed with respect to forces that exceed a magnitude of 23.8 kcal/mol/Å, such that only atoms in the immediate vicinity of the newly introduce amino-acid residue are primarily affected. After switching each residue from the off-state to the on-state, it is pulled from the insertion point toward the center of mass of the translocon pore residues for a period of 5 ns to create space for the next amino-acid residue in the nascent-protein sequence; this is achieved by harmonically tethering ($k = 2.38$ kcal/mol/Å^2) the center of mass of the nascent-protein residue to a virtual bead that moves with constant velocity $v = 1$ Å/ns. Since these simulations are performed in the NPT ensemble, the simulation cell volume appropriately relaxes upon inclusion of the additional nascent-protein residues. Since the nascent-protein growth involves only the introduction of uncharged amino-acid residues, the total simulation cell remains neutral.

Nascent-protein evolution. At nascent-protein lengths corresponding to 15, 30 and 45 amino acid residues, the simulation cell is ported to the Anton computing system for microsecond-timescale relaxation. In these simulations, short-ranged Van der Waals and

electrostatic interactions are cut off at 9.48 Å, and long-ranged electrostatic contributions are included using the k-space Gaussian Split Ewald method [144] with a cubic ($64 \times 64 \times 64$) k-point grid, an electrostatic screening parameter of $\sigma = 2.44$ Å, and a Gaussian charge-spreading width of $\sigma_s = 1.72$ Å. Bond lengths involving hydrogen atoms are constrained using the M-SHAKE algorithm [145]. We employ the RESPA numerical integration scheme [146] with a timestep of 2 fs; short-ranged interactions are updated every timestep, and long-range electrostatic interactions are updated every 6 fs. The Berendsen coupling scheme [147] is applied to keep the simulation at a temperature of 300 K and a pressure of 1 bar; the thermostat employs a coupling timescale of τ = 1.0 ps, and the barostat employs a semi-isotropic coupling timescale of τ = 2.0 ps.

B.1.2 Channel Axis Definition

To analyze configuration changes during nascent-protein insertion, we define a one dimensional coordinate associated with the translocon channel axis. For the initial configuration of each insertion trajectory, the channel axis coordinate is defined as the z-component of the Cartesian coordinate system for the simulation cell with origin positioned at the center of mass of the SecA-SecYEG complex. To avoid artifacts due to fluctuations or drift in the translocon position during the long insertion trajectories, atomic positions in each subsequent configuration of the trajectories are aligned to those of the initial configuration, using a protocol that minimizes mean-square displacement in the α-carbons of SecYEG [25].

B.1.3 Translocon Lateral Gate Width Profile

The LG width profile illustrated in Figure 3.3A in the chapter 3 is calculated as follows. The channel-axis coordinate is uniformly discretized at a resolution of $\Delta\eta = 2$Å, and the simulation cell is thus divided into corresponding parallel slabs, $\{\eta_j\}$. For each slab, the width of the LG opening is determined by considering the translocon α-carbon atoms that lie within the slab and that correspond to one of two particular subsets of the translocon atoms. The first subset, which is shown in green in Figure 3.3A in chapter 3 and which corresponds to one half of the translocon LG, includes residues in transmembrane (TM) helices TM7-8 (V270-I335 in the SecY protein). The second subset, which is shown in yellow in Figure 3.3A in chapter 3 and which corresponds to the other half of the LG, includes residues M80-S142 in the SecY protein. The width of the LG opening for each slab, w_j, is defined as the minimum distance from any α-carbon in the first subset to any in the second subset. We note that a similar definition is used to calculate the LG surface area in chapter 2.

For each insertion trajectory, we obtain the LG width profile as a function of both simulation time and the channel axis coordinate. In making Figures 3.3B and 3.3C in chapter 3 and Figure B.8, we smooth the LG width profile in both dimensions, with raw input data corresponding to the channel width profile evaluated at time intervals of 240 ps. The smoothed output, obtained using a modified ridge estimator with dimensionless smoothing parameter $s = 1$, is provided on a 50×50 grid, with spacing $\Delta t = 0.076$ μs ($\Delta t = 0.090$ μs for Figure B.8) and $\Delta\eta = 0.7$ Å. As shown in Figure B.1, this smoothing does not affect any of the trends discussed in connection with Figure 3.3.

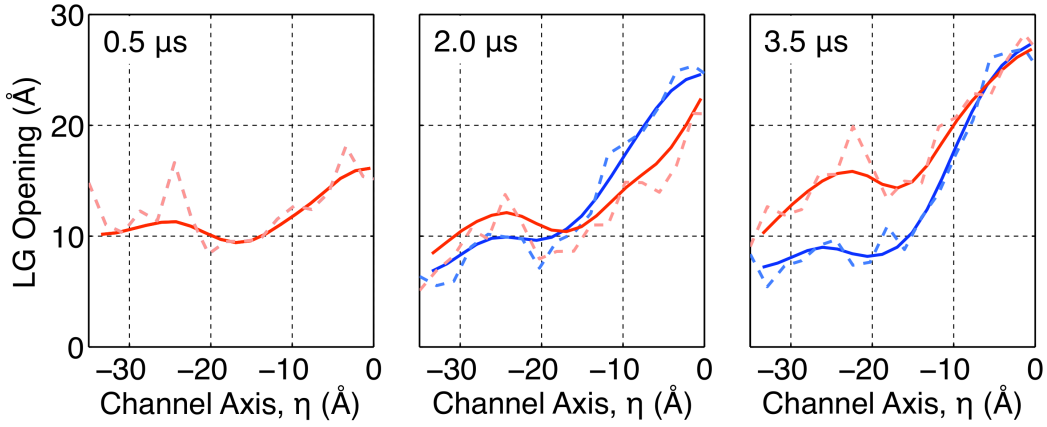


Figure B.1: Comparison of the raw (dashed) and smoothed (solid) LG width profiles. See the *LG width profile* section for details. The data presented here correspond to the results for trajectory T_1 that appear in Figure 3.3B of chapter 3.

B.1.4 Hydrophobic Contact Area

To quantify the extent of the hydrophobic contact between the nascent-protein SP and the lipid molecules in Figures 3.5C and 3.5D in chapter 3 and Figures B.10C and B.10D, we determine the interfacial contact area using the INTERVOR tool [148] in the Visual Molecular Dynamics (VMD) program [149]. The area is calculated via Delaunay triangulation of the Voronoi surface that separates two groups of atoms in the system. Atoms in the nascent-protein SP are treated as one group, and atoms in the lipid molecules that are within a cutoff distance $d_0 = 10$ Å of the translocon LG helices (SecY residues 80-100 and 270-290) are treated as the other group. Increasing the cutoff distance to $d_0 = 20$ Å results in no significant differences from the data presented in either Figure 3.5 or B.10.

B.1.5 Homology Modeling

To enable comparison between the residues of the *Thermotoga maritima* SecY protein studied in this work and the residues of the *S. cerevisiae* Sec61p on which previous experimental mutagenesis studies have been performed [63], we determine the structure for Sec61p using homology modeling. All homology modeling is performed using the MODELLER program [133].

We first construct 50 initial structural models for Sec61p from the alignment of its sequence with the *Methanococcus jannaschii* SecY protein (PDB: 1RHZ)[10]; the Cartesian coordinates for the atoms in these initial models are locally randomized with respect to a common reference structure. The structure for each model is then optimized with respect to (*i*) homology-derived restraints from the alignment, (*ii*) molecular mechanics force field from CHARMM22 [19], and (*iii*) statistically derived potentials from a representative set of known protein structures; the final homology-determined structure then corresponds to the optimized model with the lowest discrete optimized protein energy (DOPE) score [150]. A comparison between the homology-determined structure for Sec61p and the *Thermotoga maritima* SecY crystal structure is shown in Figure B.2A. The *Thermotoga maritima* SecY residues that are homologous to residues E382, E106, and E460 in the *S. cerevisiae* Sec61p are then identified as the closest negatively charged residues in the aligned structures. We note that by alternatively performing this analysis with the *Thermotoga maritima* SecY protein (PDB: 3DIN) as the structural template for Sec61p, the determined homologous residues are unchanged, although the spatial distance between residue E330 in SecY and residue E382 in Sec61p increases.

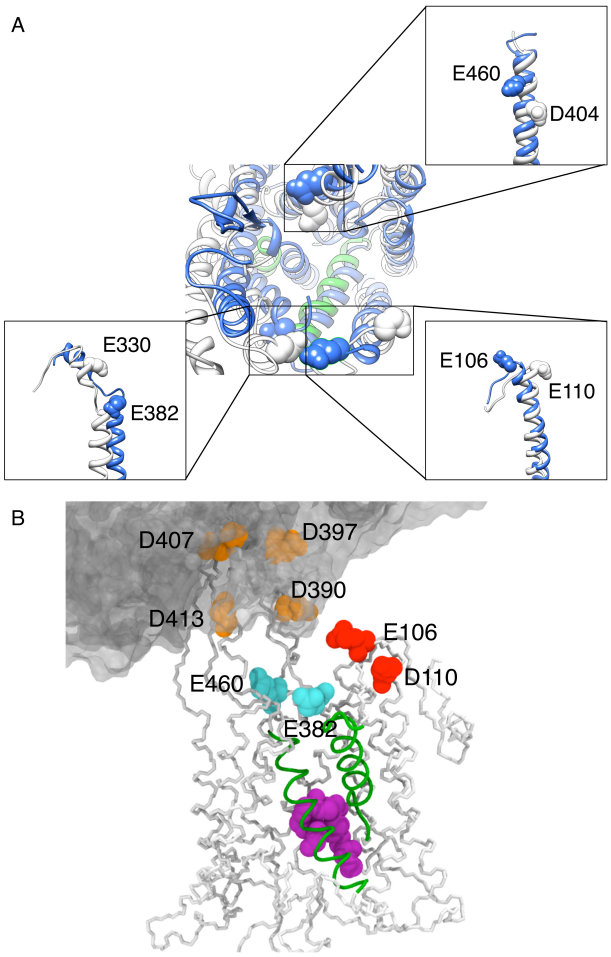


Figure B.2: Homology-determined structures for the *S. cerevisiae* Sec61p protein, which has been characterized in experimental mutagenesis studies [63]. The procedure for constructing these structures is described in the *Homology modeling* section. (A) Comparison of the *Thermotoga maritima* SecY crystal structure (white and green) and the homology-determined *S. cerevisiae* Sec61p structure (blue). The two structures are aligned via minimization of the root mean squared displacement of the α-carbon atoms between residues 80 to 140 of the two molecules. The negatively charged residues of *Thermotoga maritima* SecY that we consider in this paper (shown in white) are in close proximity to the negatively charged residues that are studied via mutagenesis in Ref. 63 (blue). (B) The homology-determined structure for Sec61p in complex with the ribosome (see text for details). The ribosome is rendered in a gray transparent surface, and the translocon is drawn in white with the LG helices in green and the plug moiety in red. Charged residues on the translocon are labeled and shown in space-filling representation. From the results of Ref. 63, it follows that mutation of residues in orange decreases the integration of protein with either Type II or Type III orientation; mutation of residues in purple enhances the integration of protein with Type II orientation and suppresses the integration with Type III orientation; and mutation of residues in cyan suppresses the integration of protein with Type II orientation and enhances the integration with Type III orientation. This figure, along with the presented insertion simulations, emphasizes that the position of charged residues impacts the regulation of integral membrane protein topogenesis.

To further illustrate the spatial distribution of the Sec61p charged residues relative to the bound ribosome, we build another homology-determined structure from the cryo-electron-microscopy-derived structure of SecY in association with the ribosome (PDB: 3KC4, 3KCR) [151]. The same method explained above is employed, and the resulting structure is shown in Figure B.2B.

B.2 Robustness of the Insertion Trajectory Initialization

Here, we examine the robustness of the equilibration and insertion protocol employed for the insertion trajectories in chapter 3. We consider potential biases due to the relatively low resolution (4.5 Å) of the experimental crystal structure that is used for a starting structure in these simulations, as well as potential biases due to the use of harmonic restraints to stabilize the truncated SecA protein in the reported insertion trajectories. Several important points suggest that this issue does not impact the reliability of the results in chapter 3.

First, we note that although the experimental resolution of the electron density for the SecA-SecYEG complex is relatively low (4.5 Å) [12], the crystal structure was solved with the assistance of higher resolution crystal structures for both the translocon (Ref. 10, 3.2 Å resolution) and SecA (Ref. 51, 2.2 Å resolution). The electron density map for the SecA-SecYEG complex provides sufficient quality for the unique assignment of all of the SecYEG TM helices and each of the SecA domains [12]. Incorporation of conserved interactions from the higher-resolution structures, which exhibit strong sequence homology with the components of the SecA-SecYEG complex [10, 51], improves the credibility of the resulting structure for the complex.

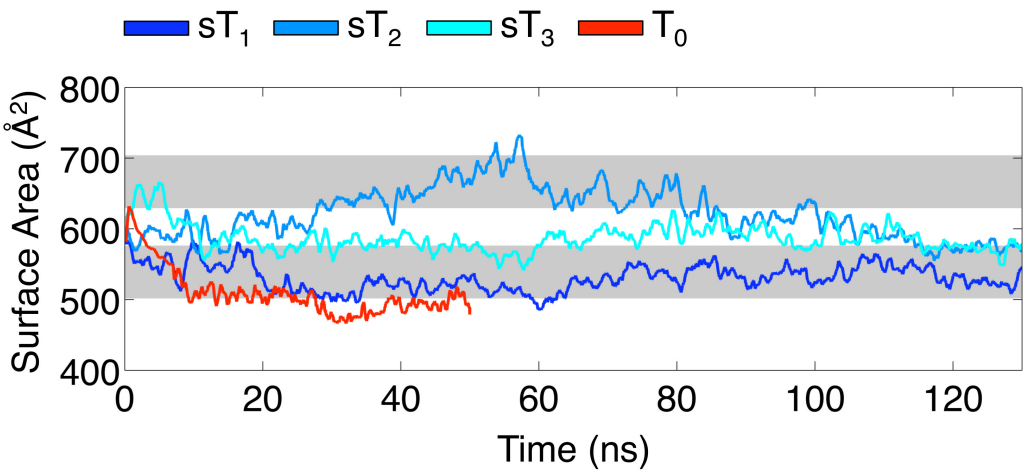


Figure B.3: Comparison of additional equilibration trajectories from the SecA-SecYEG crystal structure (sT_1-sT_3) with equilibration trajectory that is employed in chapter 3 (T_0). The surface area of the translocon LG is plotted as a function of simulation time. In all cases, the trajectories relax to partially closed conformations for the translocon LG. The shaded gray areas indicate values for the surface area that are consistent with the open and partially closed conformations for the translocon LG. See text for details.

Second, equilibration of the system in the MD simulations reduces bias associated with the choice of the initial structure. As described in section: *Simulation Protocols*, the initial SecA-SecYEG crystal structure was equilibrated to performing the μs-timescale nascent protein insertion trajectories. During this equilibration, the LG relaxed from a relatively open conformation associated with the SecA-SecYEG crystal structure to a more closed conformation that is more like the archael crystal structure of the translocon [10]. To illustrate this conformational change, the surface area associated with the opening of the translocon LG for the initialization trajectory employed in chapter 3 (T_0) is plotted as a function of simulation time in Figure B.3. The LG surface area is obtained via quadrature of the LG width profile along the LG channel axis over the range [-38 Å, 2 Å]; the LG width profile and channel axis coordinate are defined in section: *Materials and Methods*.

To confirm the robustness of the equilibrated starting configurations for the nascent

protein insertion trajectories, additional equilibration trajectories are reported in Figure B.3. Each of the additional three trajectories is initialized from the SecA-SecYEG crystal structure configuration but with different initial velocities drawn from Maxwell-Boltzmann distribution at a temperature of T = 300 K. Unlike the equilibration trajectory employed in chapter 3 (trajectory T_0), the additional trajectories were obtained using the OPLS force field and without truncation of SecA (See section: *Simulation Protocols* for details); however, the results are the same. In each case, the equilibration trajectories eventually relax to the more closed conformation of the translocon LG.

It is clear that for trajectory sT_2 in Figure B.3, the timescale for the closing of the LG is slower than that observed in trajectories T_0 and sT_1. The origin of this slow timescale is the intercalation of lipid molecules between the LG helices, which hinders LG closing. To overcome this slow timescale, we removed the three lipid molecules that were in closest proximity to the LG helices in trajectories sT_1, sT_2, and sT_3 at the simulation time of 80 ns; as is seen in Figure B.3, elimination of these lipid molecules enables the “hung" trajectory sT_2 to then relax to the more closed LG configuration. In no case was it found that lipid molecules intercalate between the LG helices after the closed LG conformation was reached. The results in Figure B.3 thus indicate that equilibration of the system relaxes the initial bias of the crystal structure and consistently leads to partial closing of the translocon LG.

Third, we find that restraints that are applied to SecA in chapter 3, which are necessary for the μs-timescale stability of the truncated SecA structure, do not appear to bias the conformational distribution of the translocon. In Figure B.4A, trajectories sT_4-sT_6 are

performed exactly as trajectories sT_1-sT_3, except that trajectories sT_4-sT_6 do not include harmonic restraints on the heavy atoms of SecA, whereas sT_1-sT_3 do include these restraints (as in chapter 3). As before, to avoid slow relaxation timescales associated with the intercalation of molecules in the LG, we delete the three nearest lipid molecules to the LG helices after 80 ns of simulation time. The trajectories performed without SecA restraints are qualitatively unchanged from those that employ restraints; in all cases, the trajectories relax to the partially closed configurations that are consistent with the equilibration trajectory employed in chapter 3. In Figure B.4B, we show the root mean squared displacement (RMSD) of the SecY protein in these six trajectories; again, no major effect associated with the SecA harmonic restraints is found.

Fourth, the translocon LG conformational changes that are emphasized in chapter 3 are large and qualitatively distinct from biases associated with the SecA harmonic restraints. Figure B.4C shows the LG surface area as a function of time in the nascent protein insertion trajectory T_1. This trajectory shows a pronounced opening of the LG that accompanies insertion of the nascent protein into the translocon channel. This trend is in marked contrast to the LG closing and conformational relaxation that is observed in the trajectories that are initialized from the SecA-SecYEG crystal structure (Figure B.4A), and it is distinct from any effects associated with the transient intercalation of lipid molecules between the LG helices in the initialized system. Furthermore, the conformational opening of the translocon LG that is observed in Figure B.4C is far more pronounced than any observed bias in Figure B.4A or Figure B.4B that is associated with the use of harmonic restraints on the SecA residues.

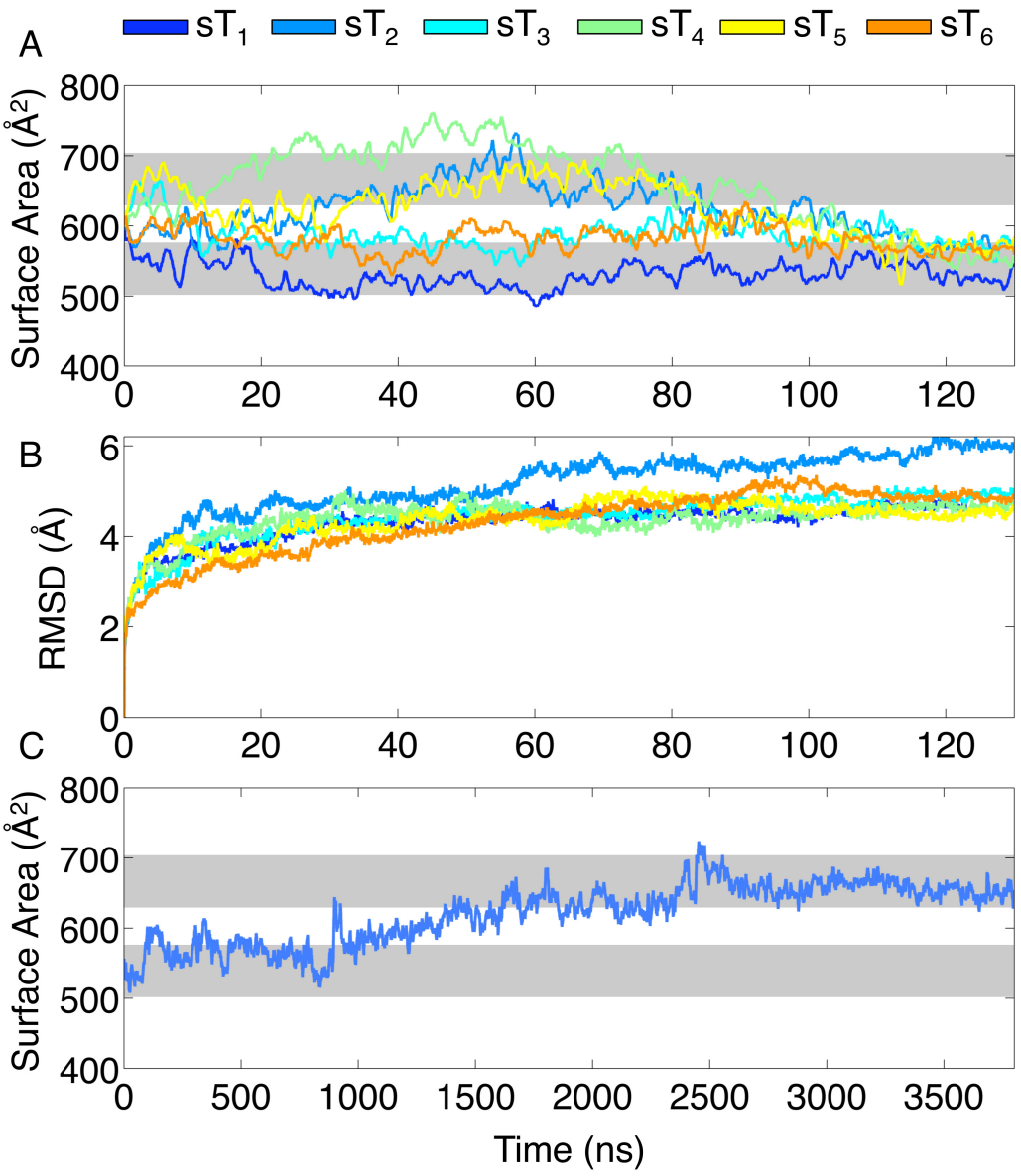


Figure B.4: Comparison of additional equilibration trajectories from the SecA-SecYEG crystal structure that are performed with (sT_1-sT_3) and without (sT_4-sT_6) harmonic restraints applied to the SecA residues. (A) The translocon LG surface area is plotted as a function of time. (B) The RMSD of the SecY protein with respect to the crystal structure geometry [12] is plotted as a function of time. (C) The translocon LG surface area plotted during of insertion trajectory T_1 from chapter 3. Comparison of the results in this figure with those of part A illustrate that conformational changes associated with relaxation from the initial structure are qualitatively different than the trends associated with nascent protein insertion that are emphasized in chapter 3. Furthermore, any biases due to the use of harmonic restraints on SecA are apparently very small in comparison to the qualitative conformational changes that accompany nascent protein insertion.

In summary, concerns over the reliability of the starting configuration of the simulations are at least partially mitigated by *(i)* the fact that the solution of the 4.5 Å resolution SecA-SecYEG crystal structure was obtained with input from much higher-resolution structures of the component proteins, and *(ii)* the observation in Figure B.3 that equilibration of the SecA-SecYEG complex relaxes the initial conformation of the crystal structure and leads to more closed conformations of the translocon LG. Furthermore, concerns about the use of harmonic restraints to stabilize the SecA structure are minimized due to *(iii)* the observation in Figures B.4A and B.4B that harmonic restraints have little effect on the conformation of the translocon LG, and *(iv)* the fact that results and conclusions that are emphasized in chapter 3 involve trends and conformational changes (such as those in Figure B.4C) that are far more pronounced than any observed biases due to the initial crystal structure or the use of harmonic restraints on the SecA residues.

B.3 Insertion Trajectories with Different Periods of Growth Evolution

For comparison with trajectories T_1-T_4 that are discussed in chapter 3, we performed two additional insertion simulations in which each nascent-protein growth period involves the addition of only two additional nascent-protein residues at a pace of one residue per five nanoseconds and in which the intervening nascent-protein evolution periods span 100 ns. The details for these two additional insertion trajectories (T_5 and T_6) are presented in Figure B.2. We note that the protocol employed for these additional trajectories allows equilibra-

Table B.2: Summary of the additional insertion trajectories.

Trajectory	IP	Mature Domain	Growth Cycles	Total Length
T_5	1	L_{30}	15	2.5 μs
T_6	1	Q_{30}	15	2.5 μs

tion after smaller periods of growth than the trajectories discussed in chapter 3; however, since the evolution periods in trajectories T_5 and T_6 span only 100 ns, the net pace of nascent-protein insertion in trajectories T_5 and T_6 is faster than the pace of insertion for trajectories T_1-T_4. As is seen in Figure B.5, this faster pace of insertion in the additional trajectories leads to simulation results that do not exhibit the anticipated features of protein translocation or membrane integration. The figure compares trajectories T_5 and T_6 with trajectories T_1 and T_2; all four trajectories share the same insertion point, trajectories T_1 and T_5 model nascent proteins with the same sequence, and trajectories T_2 and T_6 model nascent proteins with the same sequence. As is extensively discussed in chapter 3, the slower insertion protocol employed in trajectories T_1 and T_2 leads to docking of the nascent-protein SP at the translocon LG and associated conformational changes in the LG. However, the more rapidly inserted trajectories T_5 and T_6 do not exhibit these mechanistic features; instead, the nascent protein becomes jammed at the cytosolic mouth of the translocon and exhibits little conformational sampling on the timescale of the simulations.

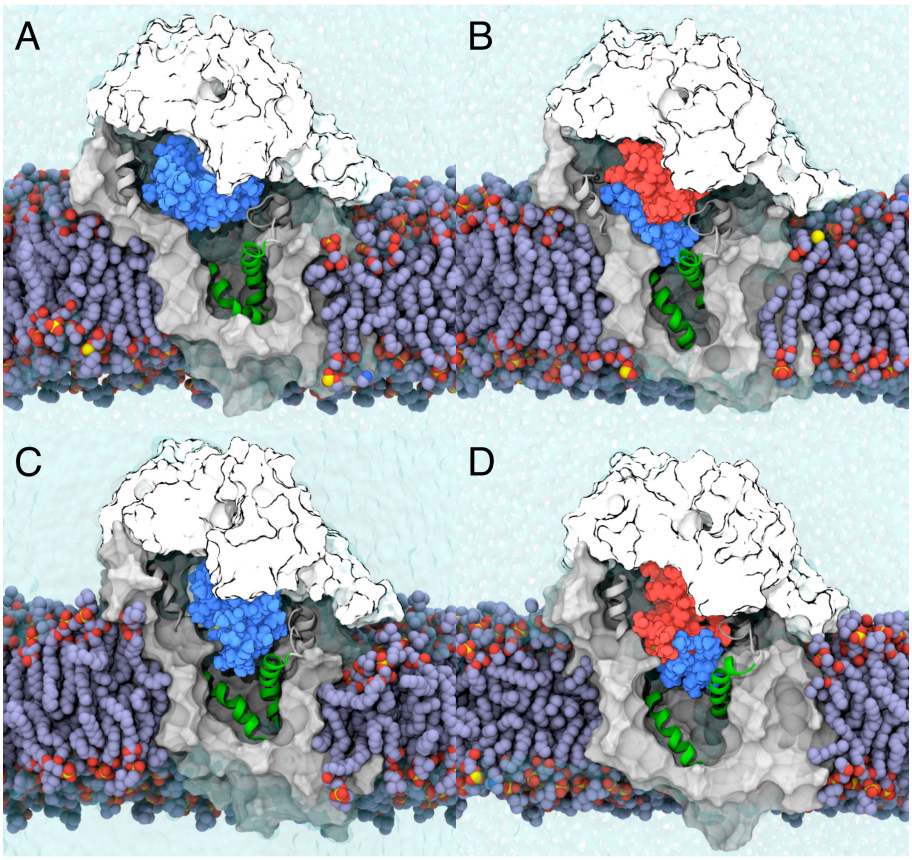


Figure B.5: The final configuration from the insertion simulations associated with (A) trajectory T_5, (B) trajectory T_6, (C) trajectory T_1, and (D) trajectory T_2. The slower insertion protocol employed in trajectories T_1 and T_2 leads to docking of the nascent-protein SP at the translocon LG and associated conformational changes in the LG. However, the more rapidly inserted trajectories T_5 and T_6 do not exhibit these mechanistic features; instead, the nascent protein becomes jammed at the cytosolic mouth of the translocon.

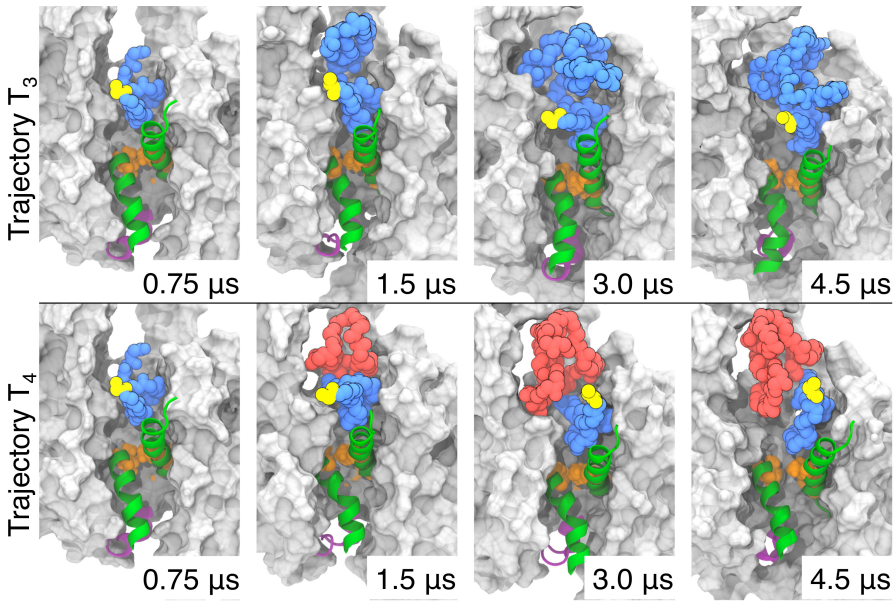


Figure B.6: Structural features of the nascent protein and translocon at various times along the insertion trajectories T_3 and T_4. This figure is comparable to Figure 3.2 in chapter 3 and employs the same representation. As in Figure 3.2, the trajectories exhibit localization of the SP residues in the LG region. Both trajectories shown here exhibit loop configurations of the nascent protein SP, with the N-terminus exposed to lipid head-groups and with some hydrophobic segments of the SP buried within the translocon channel. The trajectories in Figure 3.2 exhibit a different nascent-protein conformation, with the SP N-terminus buried in the translocon interior and with the hydrophobic segments of the SP exposed to the membrane.

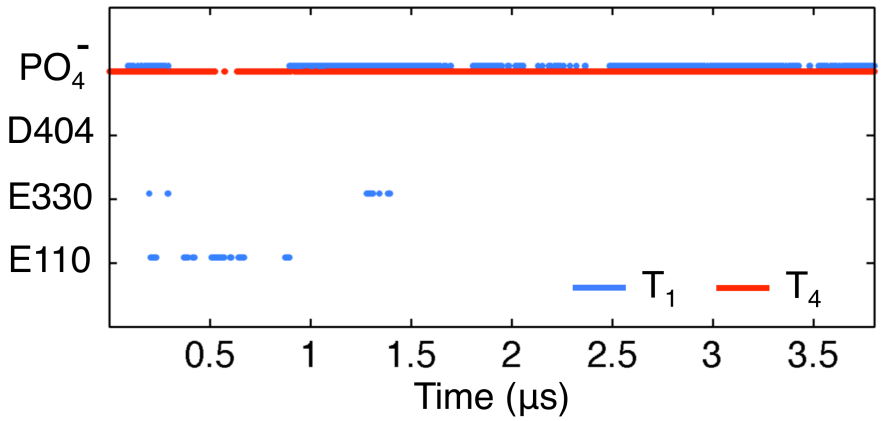


Figure B.7: The time evolution of the salt bridges along trajectories T_1 and T_4. This figure is directly comparable to the results for trajectories T_2 and T_3 in Figure 3.6 of chapter 3. As seen in Figure 3.6, the current figure indicates that salt-bridge contacts involving the N-terminus of the nascent protein form almost immediately and persist throughout the duration of the insertion simulations.

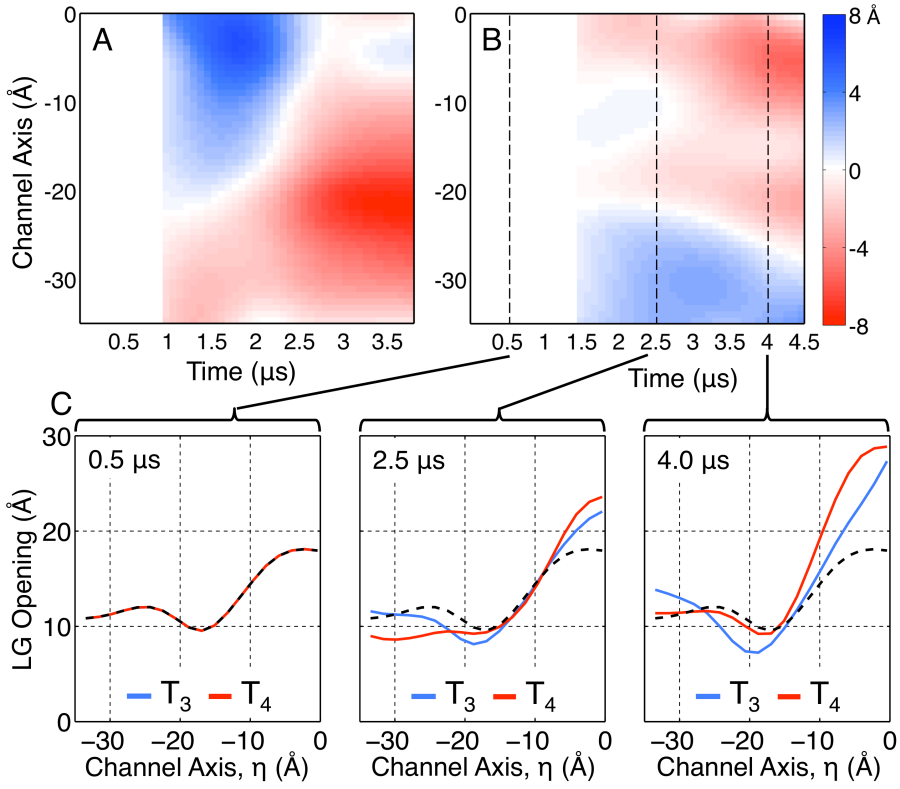


Figure B.8: The LG width profiles for trajectories T_3 and T_4. (A) For reference, the difference of the LG width profiles between trajectories T_1 and T_2, which is reproduced from Figure 3.3C in chapter 3. (B) The difference of the LG width profiles between trajectories T_3 and T_4. (C) The LG width profiles for trajectories T_3 (blue) and T_4 (red) at various times. The data at time 0.5 μs is repeated in the dashed black curve. The translocon LG undergoes similar conformational changes in both trajectories T_3 and T_4; this is consistent with the observation from Figure B.6 that the nascent protein adopts a loop configuration in both trajectories, such that both trajectories exhibit similar interactions between the nascent protein and the translocon LG.

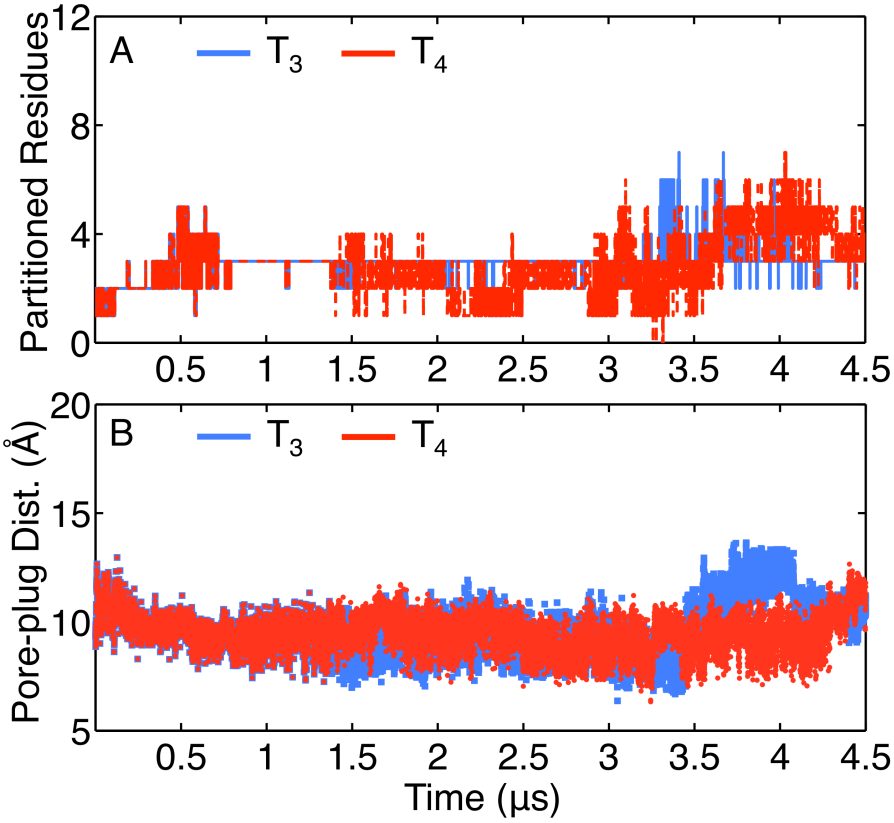


Figure B.9: Conformational dynamics of the nascent protein and the translocon plug moiety during insertion trajectories T_3 and T_4. This figure is comparable to the results for trajectories T_1 and T_2 in Figure 3.4 of chapter 3. (A) The time evolution of the number of membrane-integrated residues, $\mathcal{N}$, for the insertion trajectories T_3 and T_4. Neither of the two trajectories presented here exhibit the extensive degree of membrane integration that is observed for trajectory T_1 in Figure 3.4. The observed values of $\mathcal{N} \approx 3$ in the current figure arise from the close contact of the SP N-terminal residues with the phosphate lipid head groups. This result can again be understood as a consequence of the loop configuration that is assumed by the nascent-protein SP in both trajectories T_3 and T_4; since this configuration buries the protein mature domain within the tranlsocon channel interior, little membrane integration is possible. (B) The time evolution of the pore-plug distance for trajectories T_3 and T_4. Little displacement of the plug moiety is observed during the course of these simulations.

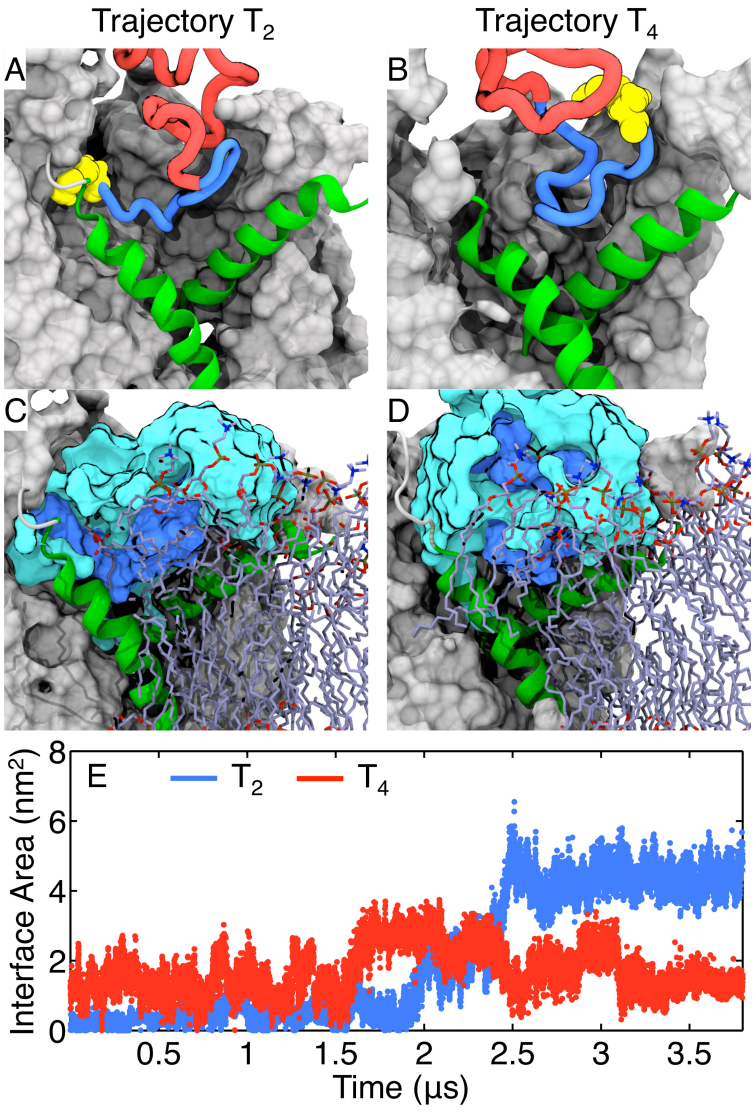


Figure B.10: Conformation of the nascent-protein SP and its solvation environment during insertion trajectories T_2 and T_4. This figure is directly comparable to the results for trajectories T_1 and T_3 in Figure 3.5 of chapter 3; the same representations and color schemes are employed. (A, B) The configuration of the nascent-protein SP is shown after 3.5 μs of simulation time in the two trajectories. (C, D) The corresponding solvation environment for the nascent-protein SP in these two snapshots of the system. In parts A and C, it is seen that the SP adopts a configuration in which the N-terminus is buried inside the translocon, whereas in parts B and D, the SP adopts a loop configuration with its N-terminus exposed to the phosphate lipid head groups. (E) The hydrophobic contact area between the SP and the surrounding lipid molecules is plotted as a function of time for trajectories T_2 (blue) and T_4 (red). Similar to the results presented in Figure 3.5, the SP configuration with the N-terminus buried in the channel interior (A, C) experiences more extensive hydrophobic contact with the lipid bilayer than the SP configuration with the N-terminus exposed to the lipid head groups (B, D).

Appendix C

Supporting Information for Chapter 4

C.1 Model Parameterization and Validation

Here, we describe the parameterization of the CG model from MD simulation results and transferable experimental data.

C.1.1 CG Bead Transfer Free Energies and Charges

Transfer FE values for bead-types R, E, L, Q, V, P used in this study (Table C.1) are comparable to experimental water-octanol transfer free energies for single Arg, Glu, Leu, Gln, Val, and Pro residues [120], respectively. Bead-types R and E, which are employed only in the topogenesis simulations in *Signal Orientation and Protein Topogenesis* in chapter 4, include charges of +2 and -2 to model the charged residues that flank the signal peptide (SP) in the engineered H1ΔLeu22 protein considered in previous experimental work [7]; the two positive charges correspond to the N-terminal Met residue and a neighboring Arg residue, and the two negative charges correspond to the two Glu residues at the opposite end of the SP.

Table C.1: CG bead charges (q) and water/membrane transfer free energies (g).

	R	E	L	Q	V	P	$\mathscr{L}$
q	2.0	-2.0	0.0	0.0	0.0	0.0	0.0
g/ε	4.0	4.0	-4.0	2.0	-2.0	0.0	variable

C.1.2 Translocon Geometry and Charges

The positions of the CG beads that model the Sec translocon (Table C.2, Figure C.1) reflect the hour-glass-shaped profile of the translocon from atomic-resolution crystal structures [10, 12, 13]. At its cytosolic and lumenal mouths, the channel diameter widens to approximately 24 Å, and it narrows to approximately 8 Å in the membrane interior [10].

To investigate the sensitivity of the simulation results to the translocon channel dimensions, we explore the degree to which integral membrane protein topogenesis is altered by reducing the width of the translocon channel from 24 to 16 Å at its cytosolic and lumenal openings (Figures C.2A and C.2B). Although narrowing the channel does alter the fraction of Type II membrane integration (Figure C.2C), the dependence of Type II membrane integration fraction with the mature domain length (MDL) remains qualitatively unchanged. Furthermore, the mechanism followed by the CG trajectories, including the competition between Type III integration and the flipping mechanism for Type II integration (Figure C.2D), remains qualitatively unchanged upon narrowing the channel.

The translocon exhibits charged residues that play an important role in establishing integral membrane protein topogenesis [63]. The CG model thus includes charges of $q = -2$ and $q = 2$ on first and fourth beads of the lateral gate (LG), where the LG beads are ordered with respect to their distance from the cytosol. The negatively charged CG bead models the electrostatic effect of residues Glu^{106} and Glu^{382} near the LG helices, and accessible lipid

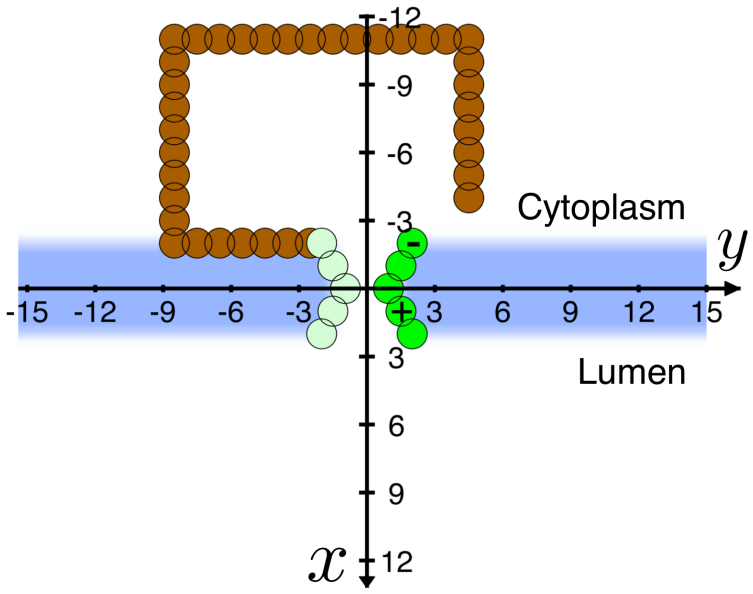


Figure C.1: Coordinate system for the CG model. CG beads for the ribosome and translocon are shown in brown and green, respectively; the membrane is shown in blue. The coordinates are reported in distance units of $\sigma = 8$ Å.

heads at the cytosolic cup; the positively charged CG bead models the effect of residue Arg^{67}, Arg^{74} and Lys^{313} at the lumenal mouth of the channel. The residue orders are based on the Sec61p molecule in *S. cerevisiae* [152].

C.1.3 Ribosome Geometry

Confinement effects due to the ribosome are explicitly included in the CG model (Table C.2, Figure C.1). Electron microscopy (EM) structures of the ribosome in complex with the translocon reveal a large lateral opening above the cytosolic cup of the translocon, which is about 20 Å wide [14, 79, 153]. The CG model likewise includes a ribosomal enclosure that is of comparable size with respect to the volume occupied by nascent chain residues in the CG representation. Near the translocon LG, the ribosomal enclosure is partially open to the cytosol, as is seen in the EM structures [14, 79, 153]; this opening prevents steric hindrance

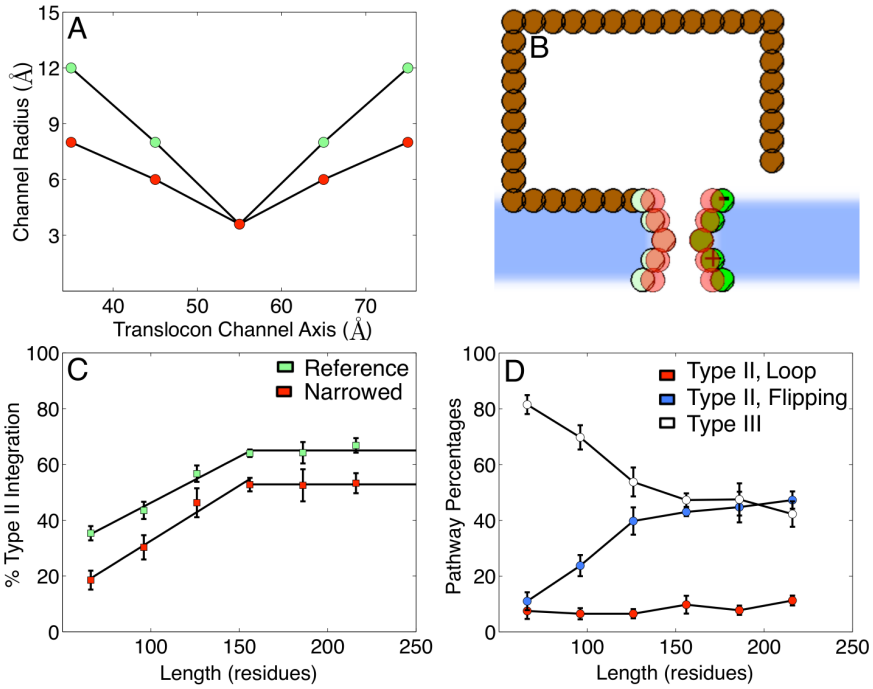


Figure C.2: Testing the sensitivity of the CG model to the coordinates of the translocon. **(A)** The width of the space within the translocon channel, plotted along the channel axis. The profile is reported for the translocon CG bead coordinates employed in chapter 4 (green) and for coordinates that correspond to narrowing the cytosolic and lumenal openings of the channel (red). **(B)** Comparison between the translocon CG bead coordinates used in chapter 4 (green) and those that correspond to the narrowed channel (red). **(C)** Changes in protein topogenesis upon narrowing the translocon channel. The protein nascent chain SP sequence of RL_6E is employed. The green data set demonstrates the MDL dependence of the Type II integration fraction for the coordinates of the translocon employed in chapter 4, and the red data set shows the corresponding results for the narrowed translocon channel. The green data set reported here is identical to that reported for the RL_6E sequence in Figure 4.3B of chapter 4. It is clear from part C that although narrowing the channel does slightly shift the simulation results, the qualitative trend in the MDL dependence is unchanged. **(D)** For the narrowed CG channel, the MDL-dependence of the fraction of CG trajectories that follow the Type II loop pathway (red), the Type II flipping pathway (blue), and the Type III integration pathway (white). As in Figure 4.3F of chapter 4, the CG simulations are performed using a protein nascent chain with SP sequence RL_6E. Comparison of part D with Figure 4.3F reveals that qualitative features of the integration mechanism are unchanged upon narrowing the translocon channel. The most significant effect is that the narrowed channel exhibits a smaller fraction of trajectories that follow the loop mechanism.

Table C.2: CG bead positions for the ribosome and the translocon (units in σ). The CG bead for the ribosomal exit tunnel is located at $[-10, -5]$. An illustration of the coordinate system is provided in Figure C.1.

Ribosome					Translocon	
x	y	x	y		x	y
-2	-2.5	-11	-5.5		-2.0	-2.0
-2	-3.5	-11	-4.5		-1.0	-1.5
-2	-4.5	-11	-3.5		0.0	-0.95
-2	-5.5	-11	-2.5		1.0	-1.5
-2	-6.5	-11	-1.5		2.0	-2.0
-2	-7.5	-11	-0.5		-2.0	2.0
-2	-8.5	-11	0.5		-1.0	1.5
-3	-8.5	-11	1.5		0.0	0.95
-4	-8.5	-11	2.5		1.0	1.5
-5	-8.5	-11	3.5		2.0	2.0
-6	-8.5	-11	4.5			
-7	-8.5	-10	4.5			
-8	-8.5	-9	4.5			
-9	-8.5	-8	4.5			
-10	-8.5	-7	4.5			
-11	-8.5	-6	4.5			
-11	-7.5	-5	4.5			
-11	-6.5	-4	4.5			

of membrane integration in the CG model and enables access of the protein nascent chain to the cytosolic exterior of the membrane [154].

To investigate the sensitivity of the simulation results to the effects of ribosomal confinement, we explore the degree to which integral membrane protein topogenesis is altered by reducing the volume of the ribosomal enclosure in the CG model. Figure C.3 shows that reducing the ribosomal enclosure by 17% leads to no statistically significant change in the fraction of Type II integration as a function of MDL. Further reduction of the ribosome enclosure by 33% leads to only a modest effect on the results, suggesting that the CG calculations are robust with respect to the details of the ribosome enclosure size.

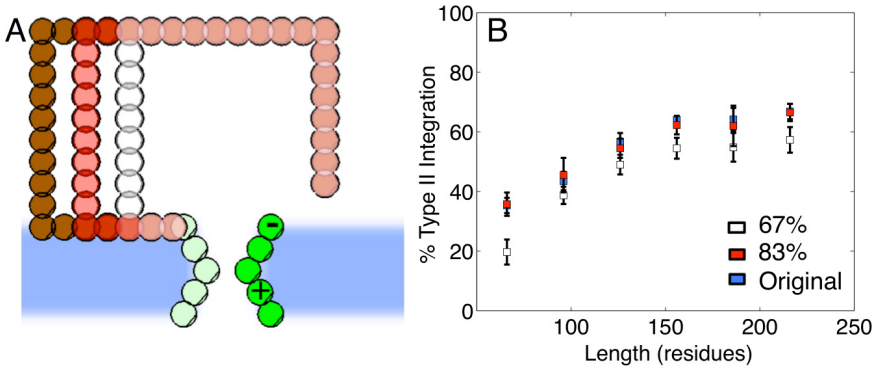


Figure C.3: Testing the sensitivity of the CG model to the volume enclosed by the ribosome CG beads. **(A)** We consider CG simulations in which the volume enclosed by the ribosome CG beads (brown) is reduced to 83% (red) and 67% (white) of its original size. In the first two cases, the position of the ribosome exit channel is unchanged (Table C.2); for the smallest volume, the ribosome exit channel is moved to position $[-10, -3]$ to avoid overlapping with the ribosome CG beads. **(B)** The MDL-dependence of the fraction of CG trajectories that undergo Type II integration, obtained using the various sizes of the volume enclosed within the ribosome beads. The protein nascent chain SP sequence of RL_6E is employed. The blue data set reported here is identical to that reported for the RL_6E sequence in Figure 4.3B of chapter 4. These results suggest that significant reductions in the size of the ribosome enclosure have little effect on the calculated trends in protein topogenesis.

C.1.4 Timescale for LG Opening

The opening and closing of the translocon LG is modeled stochastically with rates defined in equations (4.11) and (4.12) in chapter 4. In these expressions, the parameter τ_{LG} corresponds to the timescale for attempting LG opening or closing events. As in classical rate theory [155, 156], this attempt timescale is related to the timescale required for the system to transiently pass between the open and closed configurations for the LG, which we have observed in previous MD simulations of translocon/peptide-substrate/membrane systems. In our previous work [40], it was shown that spontaneous translocon LG closing in the presence of a peptide substrate occurs on the timescale of approximately 300-500 ns; specifically, this was shown in Figure 5 of the stated reference. To explore the robustness of the CG model to this parameter, we calculate the dependence of the Type II integration as a

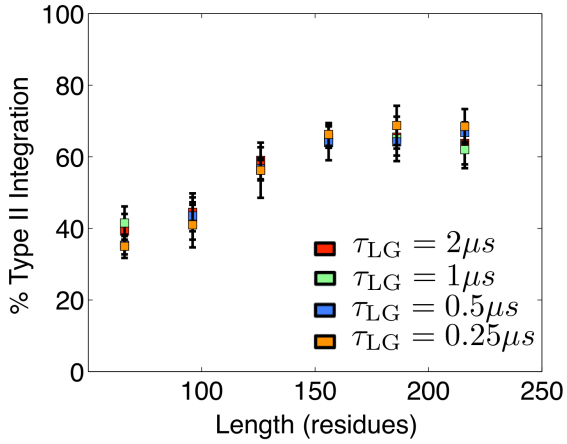


Figure C.4: Testing the sensitivity of the CG model to the timescale τ_{LG} for stochastic opening/closing of the translocon LG. The MDL-dependence of the fraction of CG trajectories that undergo Type II integration is presented, obtained using various values for τ_{LG} in the CG model dynamics. The protein nascent chain SP sequence of RL_6E is employed. The blue data set reported here is identical to that reported for the RL_6E sequence in Figure 4.3B of chapter 4. These results suggest that significant changes in τ_{LG} have little effect on the calculated trends in protein topogenesis.

function of MDL for the RL_6E SP sequence (blue curve in Figure 4.3B of chapter 4), using $\tau_{LG} = 250$, 500, 1000, and 2000 ns; these results, which are reported in Figure C.4, show no significant differences among the four data sets and suggest that the CG calculations are very robust with respect to τ_{LG}. A value of $\tau_{LG} = 500\,\text{ns}$ is employed throughout chapter 4.

C.1.5 FE for LG Opening

In previous work [40], we performed extensive atomistic and residue-based CG simulations to explore the energetics of conformational changes in the translocon LG. These simulations found that the FE for LG opening depends approximately linearly on the hydrophobicity of peptide substrates in the channel, such that $\Delta G_{tot} \approx \Delta\sigma A$; here, A is the surface area created between the membrane and channel interiors upon LG opening, and

$\Delta\sigma$ is the difference in surface tensions between the substrate-membrane interface and a substrate-aqueous interface associated with the channel interior. This energy expression reflects changes in the large-lengthscale hydrophobic interactions of the peptide substrate upon LG opening [157].

In the CG model that is developed in the current paper, we utilize this simple relationship between LG energetics and substrate hydrophobicity. Specifically, for cases in which peptide substrate occupies the full length of the channel, the CG model employs

$$\Delta G_{\mathrm{tot}} = \Delta\sigma_{\mathrm{occ}} A_{\mathrm{occ}} + \Delta E, \tag{C.1}$$

where $\Delta\sigma_{\mathrm{occ}}$ is the difference in surface tensions between the substrate-membrane interface and a substrate-aqueous interface, and A_{occ} is the corresponding surface area; these quantities depend only on the peptide substrate residues that occupy the translocon channel. The second term in this equation is $\Delta E = E^{\mathrm{o}}_{\text{p-LG}} - E^{\mathrm{c}}_{\text{p-LG}}$, where $E^{\mathrm{o}}_{\text{p-LG}}$ is the sum of the pairwise interactions between the CG beads of the protein nascent chain and the CG beads of the LG in the open state, and $E^{\mathrm{c}}_{\text{p-LG}}$ is the corresponding sum of interactions for the LG in the closed state. The ΔE term prevents the sterically forbidden closing of the LG when the protein nascent chain is in transit between the channel interior and the membrane; it makes no contribution to ΔG_{tot} unless the beads of the protein nascent chain overlap with the translocon LG.

To calculate $\Delta\sigma_{\mathrm{occ}} A_{\mathrm{occ}}$, we assume that each nascent chain CG bead in the channel makes an equal contribution to the interfacial surface area, such that $A_{\mathrm{occ}} = aM$, where M is the number of nascent chain CG beads that occupy the channel in a particular configuration

of the system and a is the surface area per bead. Similarly, $\Delta\sigma_{\text{occ}} = \sum_{i=1}^{M} \Delta\sigma_i / M$, where $\Delta\sigma_i$ is the surface tension difference associated with a particular CG bead that occupies the channel. It follows that $\Delta\sigma_{\text{occ}} A_{\text{occ}} = \sum_{i=1}^{M} \Delta\sigma_i a$. For these calculations, a nascent chain CG bead is defined as occupying the channel if its position falls within $\mathscr{V}$, the volume enclosed by the polygon whose vertices correspond to the centers of the translocon CG beads (Table C.2). Finally, relating the hydrophobicity of each CG bead to its marginal contribution to the interfacial FE (i.e., $g_i = \Delta\sigma_i a$), we obtain $\Delta\sigma_{\text{occ}} A_{\text{occ}} = \sum_{i=1}^{M} g_i$, where g_i is the water-membrane transfer FE (Table C.1) for each CG bead that occupies the translocon channel in a given configuration.

To account for cases in which the translocon channel is only partially occupied by the protein nascent chain, equation (C.1) is generalized such that

$$\Delta G_{\text{tot}} = \Delta\sigma_{\text{occ}} A_{\text{occ}} + \Delta E + \Delta G_{\text{empty}} \chi_{\text{empty}}. \tag{C.2}$$

Here, the first two terms on the right-hand side are unchanged from equation (C.1), and the third term accounts for the FE of opening portions of the translocon LG that are not occupied by protein nascent chain residues. Specifically, ΔG_{empty} corresponds to the FE cost for opening the translocon LG in the absence of peptide substrate, and χ_{empty} corresponds to the fraction of the translocon channel that is not occupied by nascent chain beads. When the protein nascent chain occupies the full length of the translocon channel, equation (C.2) reduces to equation (C.1); when the channel is completely empty, equation (C.2) reduces to $\Delta G_{\text{tot}} = \Delta G_{\text{empty}}$. Following our previously reported simulations of translocon LG opening [40], we employ $\Delta G_{\text{empty}} = 16\varepsilon \approx 8$ kcal/mol. To calculate χ_{empty} for a given

configuration of the protein nascent chain, the volume $\mathscr{V}$ is equally partitioned into four subvolumes along the channel axis; it follows that $\chi_{\text{empty}} = \mathscr{M}/4$, where $\mathscr{M}$ is the number of subvolumes that are empty of CG beads of the protein nascent chain.

From numerical tests of the robustness of these parameters, the dependence of Type II integration on the rate of ribosomal translation is found to be sensitive to the magnitude of the parameter ΔG_{empty}. Larger values of this parameter, which would better match the earlier calculation of $\Delta G_{\text{empty}} \approx$16-20 kcal/mol [40], lead to smaller changes in the Type II integration fraction upon slowing ribosomal translation. However, we note that these earlier calculations did not account for the effect of ribosomal binding, which is expected to reduce the FE barrier associated with LG opening [12, 153].

C.1.6 Alternative Approaches to Modeling the FE for LG Opening

Here, we consider alternative schemes for describing the opening and closing of the translocon LG. In the first case, we assume that FE for the open LG is far lower than that of the closed LG, such that the LG is always open. In the second case, we assume that the FE for LG opening remains fixed at a value that favors the closed LG conformation, regardless of the nascent protein conformation. Unlike the approach employed in chapter 4 (and described in the previous section), numerical tests reveal that both of these alternative descriptions of the LG energetics lead to qualitatively incorrect results.

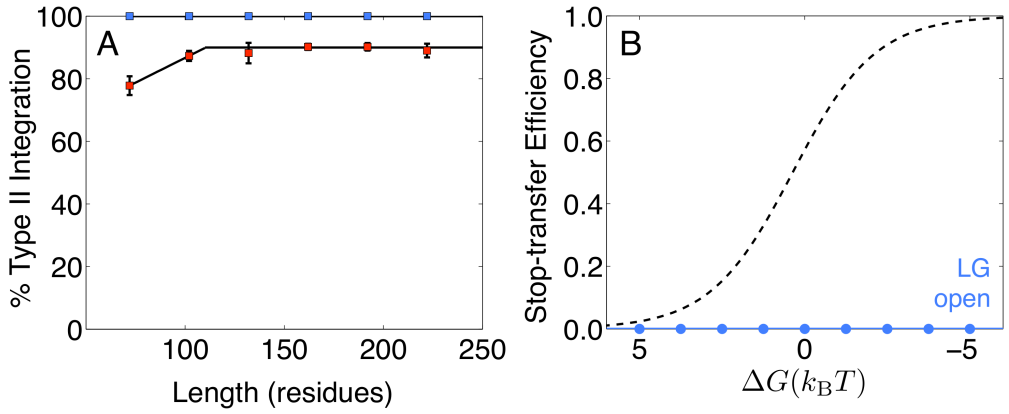


Figure C.5: Testing the effect of LG motions on protein topogenesis and stop-transfer efficiency. We consider the alternative description of the translocon LG in which the LG is always left open. **(A)** The figure plots Type II integration fraction as a function of MDL. The red data set corresponds to the protein topogenesis results presented for the RL_4E SP sequence in Figure 4.3A of chapter 4. The blue data set is obtained using the same protein sequences and employs the alternative description in which the LG is open at all times in the simulations. **(B)** The figure plots stop-transfer efficiency as a function of H-domain hydrophobicity. The black dashed line is the sigmoidal fit to the data presented in Figure 4.5A of chapter 4. The blue data set is obtained using the same protein sequences and employs the assumption that the LG is open at all times in the simulations.

C.1.6.1 Assumption that the LG is Always Open

Figure C.5A shows that neglecting LG opening/closing significantly impacts the calculated results for nascent protein topogenesis. The red data set corresponds to the protein topogenesis results presented for the RL_4E SP sequence in Figure 4.3A of chapter 4. The blue data set is obtained using the same protein sequences and assuming that the LG is open at all times in the simulations. The neglect of LG opening/closing leads to the complete loss of Type III membrane integration in this case. In the absence of the slow timescale for LG opening, the SP readily adopts the thermodynamically favorable Type II orientation.

Figure C.5B shows similarly discouraging results for stop-transfer efficiency. The dashed line in the figure corresponds to the stop-transfer efficiency results reported in Figure 4.5A of chapter 4. The blue data set is obtained using the same protein sequences and

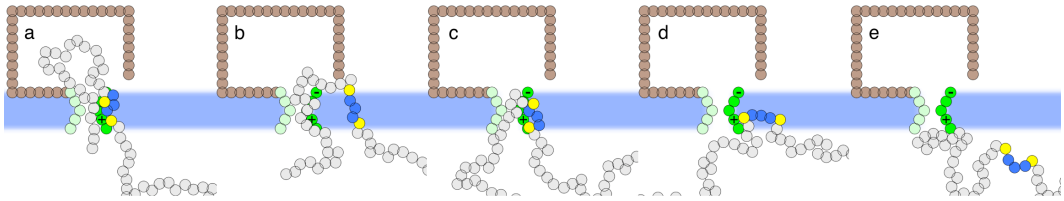


Figure C.6: Translocation mechanism observed in simulations for which the LG is kept open. Although the open LG enables the H-domain to partition from the channel interior to the membrane interior (a, b), diffusion of the H-domain away from the translocon to yield the membrane integration product is hindered by the attraction of the hydrophilic C-terminal domain for the channel interior (c). Without allowing for LG closing, the C-terminus effectively tethers the H-domain to the ribosome until secretion of the C-terminal tail occurs (d), exclusively leading to the secretion product (e). For very hydrophobic H-domain, the final transition from d to e does not occur on the timescale of the simulation performed here, yet secretion of the nascent-protein mature domain is complete.

employs the assumption that the LG is open at all times in the simulations. The alternative treatment of the LG leads to complete loss of stop-transfer efficiency; all trajectories lead to the translocation of the C-terminal domain. The mechanistic basis for this result is illustrated in Figure C.6. Although the open LG enables the H-domain to partition from the channel interior to the membrane interior, diffusion of the H-domain away from the translocon to yield the membrane integration product is hindered by the attraction of the hydrophilic C-terminal domain for the channel interior. Without allowing for LG closing, the C-terminus effectively tethers the H-domain to the ribosome until secretion of the C-terminal tail occurs, exclusively leading to the secretion product.

C.1.6.2 Assumption that the FE for LG Opening is Unaffected by the Nascent Protein

We now consider the alternative description in which the FE for LG opening is independent of the nascent protein contents of the channel. Specifically, we perform simulations in which the relative rates for LG opening and closing (equations (4.11) and (4.12)) corre-

spond to $\Delta G_{\text{tot}} = \Delta G_{\text{empty}}$, regardless of the nascent protein configuration; we employ the same value for the opening/closing attempt timescale, $\tau = 500$ ns, as is used in chapter 4.

Figure C.7A shows the effect of the alternative description on nascent protein topogenesis. The red data set corresponds to the protein topogenesis results presented for the RL_4E SP sequence in Figure 4.3A of chapter 4. The blue data set is obtained using the same protein sequences and using the alternative description for the LG energetics. This leads to almost complete loss of Type II integration for all MDL. Without the role of hydrophobic SP residues in stabilizing open LG configurations, closed LG configurations dominate. Both the direct and flipping pathways for Type II integration are thus eliminated. Type III integration survives by having the SP enter directly into the membrane interior from the ribosome enclosure, without passing through the translocon channel interior. These results are clearly inconsistent with the experimental observation of Type II membrane integration.

Figure C.7B illustrates the effect of the alternative description on the stop-transfer simulations. The dashed line in the figure corresponds to the stop-transfer results reported in Figure 4.5A of chapter 4; the blue data set employs the alternative description, which neglects the effect of the nascent protein on the FE for LG opening. The alternative description leads to a significant shift toward reduced membrane integration, although sigmoidal behavior of the stop-transfer efficiency as a function of H-domain hydrophobicity is still observed. As for the topogenesis simulations, closed LG conformations dominate in the alternative description, due to the neglect of the role of H-domain residues in stabilizing the open configurations of the LG. The rapid equilibration of the CG trajectories between states b and c^* (Figure 4.4) is thus shifted toward state b, which favors the subsequent for-

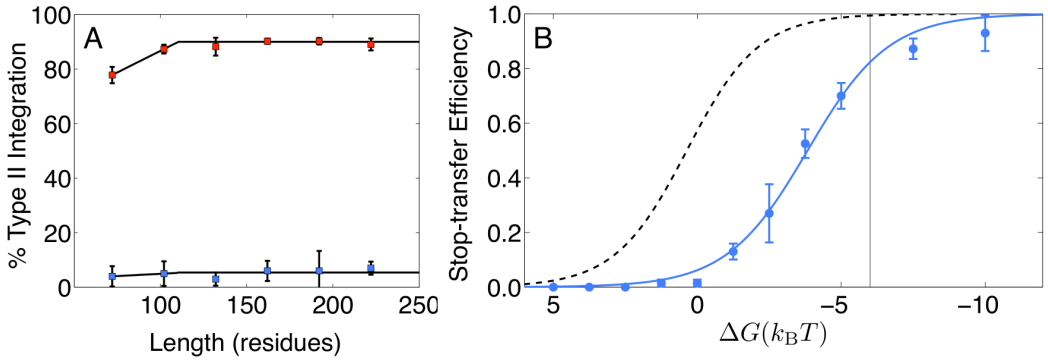


Figure C.7: Testing the effect of LG motions on protein topogenesis and stop-transfer efficiency. We consider the alternative description of the translocon LG in which the FE for opening the LG is assumed to be unaffected by the nascent-protein chain. **(A)** The figure plots Type II integration fraction as a function of MDL. The red data set corresponds to the protein topogenesis results presented for the RL_4E SP sequence in Figure 4.3A of chapter 4. The blue data set is obtained using the same protein sequences and employs the alternative description, which neglects the effect of the nascent protein on the FE for LG opening. **(B)** The figure plots stop-transfer efficiency as a function of H-domain hydrophobicity. The black dashed line is the sigmoidal fit to the data presented in Figure 4.5A of chapter 4. The blue data set is obtained using the same protein sequences and employs the alternative description, which neglects the effect of the nascent protein on the FE for LG opening.

mation of the secretion product (as is described in connection with Figure 4.5 in chapter 4 or in *Analytical Model for TM Partitioning*).

C.1.7 CG Bead Diffusion Coefficient

The diffusion coefficient D for the CG beads of the protein nascent chain (equation (4.10)) is parameterized to reproduce the experimentally observed timescale for protein diffusion across the Sec translocon. Specifically, we consider the measurements by Rapoport and co-workers of post-translational translocation times for the 165-residue pre-pro-α factor (ppαF) [107]. In these experiments, the protein substrate is initially bound to the Sec translocon in proteoliposomes; translocation is initiated via addition of adenosine triphosphate (ATP) and binding immunoglobulin protein (BiP), and the fraction of translocated

protein is monitored as a function of time (Figure C.8, red).

To model this experiment, simulations using the CG model are performed on a weakly hydrophilic 165-residue peptide substrate (i.e., 55 type-Q CG beads). Following the experimental study, CG trajectories are initialized with the N-terminal bead of the substrate positioned near the center of the translocon channel (i.e., $(x/\sigma, y/\sigma) = (0, 0.5)$); the remainder of the CG beads for the substrate are positioned in a linear configuration along the channel axis in the cytosolic direction with an inter-bead separation distance of σ. Given the post-translational experimental protocol, the ribosome is excluded in these CG simulations; furthermore, to mimic the effect of the BiP protein in preventing complete backsliding of the nascent chain to the cytosolic side of the membrane, the following soft-wall interaction is applied to the N-terminal bead of the protein nascent chain,

$$U(x,y) = \begin{cases} \frac{1}{2}\kappa(x - 0.5\sigma)^2 & x < 0.5\sigma \\ 0 & x \geq 0.5\sigma, \end{cases} \tag{C.3}$$

where $\kappa = 7k_{\mathrm{B}}T/\sigma^2$ and the coordinate system for the CG beads is provided in Figure C.1. Following equilibration of the tail of the protein nascent chain (i.e., 10^6 CG timesteps with the N-terminal bead held fixed) each CG trajectory is evolved until the C-terminal tail of the protein nascent chain reaches the lumenal side of the membrane. Repeating this process for 800 trajectories, we determine the fraction of translocated peptides as a function of time, and the CG bead diffusion coefficient is tuned to match the experimental timescale.

Using $D = 758.7\,\mathrm{nm}^2/\mathrm{s}$, excellent agreement between the simulated and experimental protein translocation times is obtained (Figure C.8). We note that this value for D is

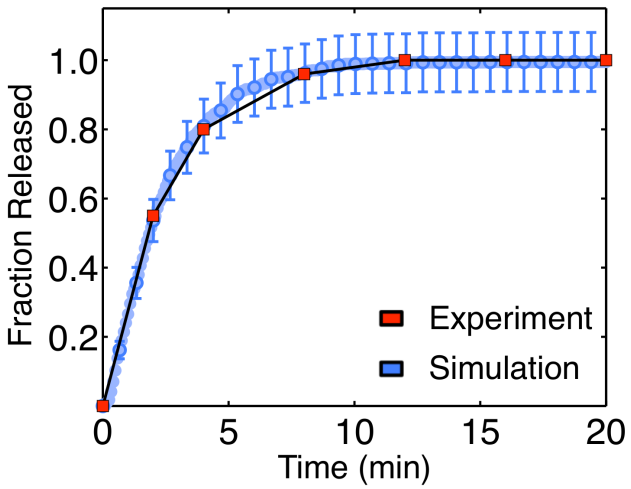


Figure C.8: Comparison of experimental and simulated timescales for secretion of the 165-residue pre-pro-α factor via the Sec translocon. The red data set is taken from Figure 1C of an earlier paper from Rapoport and co-workers [107]. The blue data set is obtained with the CG model using a bead diffusion coefficient of $D = 758.7\ \mathrm{nm}^2/\mathrm{s}$.

significantly smaller than protein diffusion coefficients for small proteins in aqueous and membrane environments (approximately $10^8\,\mathrm{nm}^2/\mathrm{s}$ and $10^6\,\mathrm{nm}^2/\mathrm{s}$, respectively [158]), which reflects the highly confined environment of channel interior [98]. We utilize $D = 758.7\,\mathrm{nm}^2/\mathrm{s}$ for all reported CG simulations.

As an alternative measure of the timescale for nascent chain diffusion, we perform a single, 2.3 microsecond MD trajectory of the translocon/nascent-chain/membrane system with all-atom resolution. This simulation, a snapshot of which is shown in Figure C.9A, considers the diffusion of a Leu_{30} peptide sequence inside the channel of the archaeal Sec translocon [10]. The trajectory is performed on the specialized architecture Anton [38], and it is initialized from a previously equilibrated configuration for the system [40]. Details of the atomistic MD simulation protocol are as follows. The system contains 115651 atoms in a simulation cell of initial dimensions $104\,\text{Å} \times 104\,\text{Å} \times 103\,\text{Å}$. The system is described using orthorhombic periodic boundary conditions, and the atomistic interactions

are described using the CHARMM27 force field for protein and CHARMM36 for the lipid with the TIP3P water model [19, 140]. Nonbonded van der Waals interactions employ a distance cut off of 9.48 Å. Short-ranged electrostatic interactions are also cutoff at 9.48 Å, and long-ranged electrostatic contributions are computed using the k-space Gaussian Split Ewald method [144] with a cubic $64 \times 64 \times 64$ grid, a electrostatic splitting parameter $\sigma = 2.44$Å, and a Gaussian width $\sigma_s = 1.72$Å. Bond lengths involving hydrogen atoms are constrained using the M-SHAKE algorithm [145]. We employ the RESPA numerical integration scheme [146] with a timestep of 2 fs for the atom positions and short-ranged interactions; long-range electrostatic interactions are updated every 6 fs. The Berendsen coupling scheme [147] is applied to keep the simulation at a temperature of 300 K and a pressure of 1 bar; the thermostat employs a coupling timescale of $\tau = 1.0$ ps, and the barostat uses a semi-isotropic coupling timescale of $\tau = 2.0$ ps.

In this microsecond-timescale MD trajectory, z-component of the displacement between the centers of mass for the protein nascent chain and the translocon, $z(t)$, is plotted as a function of simulation time (Figure C.9B). By time averaging over this trajectory, we then obtain the mean squared displacement (MSD) of the relative z-component displacement as a function of time, $\langle z(t) - z(0) \rangle^2$. The MSD shows a short timescale for motion (≈ 10 ns) that is associated with local rearrangements of the nascent chain, and a slower timescale that is associated with diffusion of the chain through the channel. Fitting this slower timescale with $\langle z(t) - z(0) \rangle^2 = 2NDt$, where $N = 10$ is the number of CG beads associate with a 30-residue peptide chain [159], we obtain a diffusion coefficient for a single CG bead of $D = 410\,\mathrm{nm}^2/\mathrm{s}$. Conclusions drawn from this single MD trajectory must be regarded with

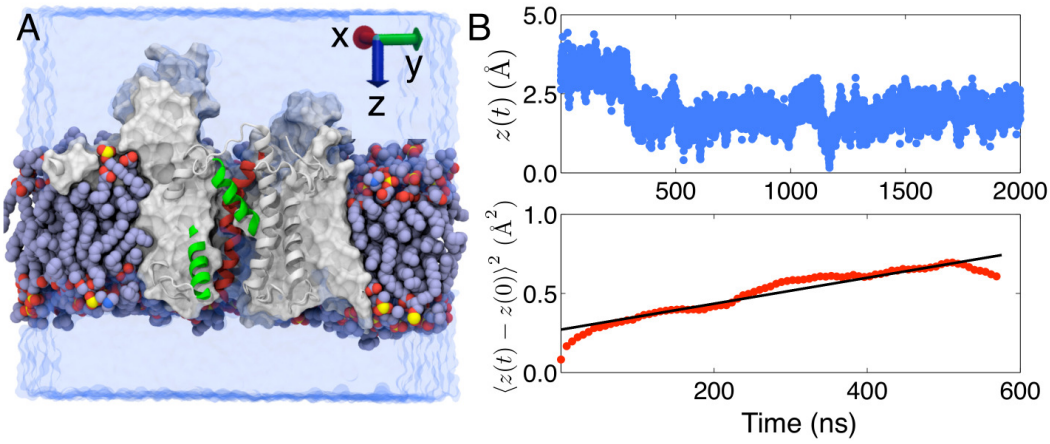


Figure C.9: Characterizing the CG bead diffusion coefficient using microsecond-timescale all-atom MD simulations. **(A)** Snapshot of the full periodic simulation cell for the all-atom system, including the Sec translocon (grey, with the LG helices in green) and the protein nascent chain (red). The explicit lipid bilayer is shown in space-filling representation, and the explicit aqueous solution is show in surface representation. **(B)** The top panel shows the z-component of the displacement between the center of mass of the protein nascent chain and the translocon as a function of time in the 2.3 microsecond MD simulation. The bottom panel shows the MSD for this motion along the channel axis (red). The black line is a linear fit to the MSD in the range of 10 - 570 ns, the slope of which is used to estimate the CG bead displacement.

caution, since more extensive simulations are needed to fully converge the MSD curve and to assess the range of timescales associated with diffusion of a peptide chain through the translocon; nonetheless, we note that the bead diffusion coefficient of $D = 410\,\mathrm{nm}^2/\mathrm{s}$ obtained here is in reasonable agreement with the value of $D = 758.7\,\mathrm{nm}^2/\mathrm{s}$ obtained above by fitting the CG model to experimental data.

C.2 Simulation Protocols

C.2.1 Trajectory Initialization and Termination (Protein Topogenesis Simulations)

Ribosomal translation is directly modeled in the CG simulations via growth of the nascent chain at the ribosome exit channel (shown in red, Figure 4.2). CG trajectories are initialized from equilibrated configurations for the peptide of nine beads long. Different initial random number seeds are used for each independent simulation. During translation, CG beads are introduced sequentially at the C-terminus, such that the nascent chain elongates; during this elongation process, the bead at the C-terminal tail is held fixed at the ribosome exit channel, and all other protein and translocon degrees of freedom are simulated as described in the *Materials and Methods* in chapter 4.

Upon completion of protein translation, the C-terminus of the inserted protein detaches from the ribosome exit channel, and the small subunit of the ribosome releases from the cytosolic mouth of translocon [108]. Experimentally observed leakage of small molecules across the translocon following this ribosomal release suggests that the ribosome no longer seals the cytosolic mouth of the translocon [109]; we thus model ribosomal release by eliminating interactions associated with the ribosome CG beads.

Membrane integration trajectories are terminated after full translation of the protein mature domain, either when the SP integrates into the membrane in the Type III orientation and diffuses to a distance of 16 nm from the translocon (state *d*, Figure 4.2) or when the SP integrates into the membrane in the Type II orientation. To meet the distance criterion,

the y-coordinate for each bead in the nascent protein SP must be greater than 16 nm, using the coordinate system illustrated in Figure C.1. For proteins in the Type II orientation, rather than running the CG trajectories until the SP diffuses a distance of 16 nm from the translocon (state g, Figure 4.2), trajectories are terminated when the trajectories reach state f, for which the SP is integrated into the membrane and the translocon LG is closed. As is demonstrated in Figure C.11, termination of the Type II integration trajectories at state f accounts for the effect of BiP binding to the lumenally exposed portions of the protein nascent chain.

C.2.2 Trajectory Initialization and Termination (Stop-Transfer Simulations)

As in the topogenesis simulations, ribosomal translation is modeled via addition of peptide residues to the nascent chain at the ribosomal exit channel (shown in red, Figure 4.4). The stop-transfer trajectory is initialized from the ensemble of equilibrated configurations for the protein with only 15 residues of the C-terminal domain translated, with the H-domain residing in the ribosome translocon junction (Figure 4.4a).

Unbinding of the ribosome at the end of translation is modeled as in the topogenesis simulations. Upon completion of translation, the constraint on the C-terminus of the protein nascent chain is removed and interactions between the CG beads of the ribosome and protein nascent chain are eliminated.

Each CG trajectory is terminated after full translation of the protein C-terminal domain, either when the H-domain integrates into the membrane and diffuses a distance of 16 nm

from the translocon (state f, Figure 4.4) or when both the H-domain and the C-terminal domain fully translocate into the lumenal region (state e, Figure 4.4). The N-terminal signal anchor of the protein is fixed at a distance of 20 nm from the translocon; the simulations thus assume that the H-domain membrane integration mechanism does not involve direct helix-helix contacts with the protein anchor domain [85].

C.2.3 Definition of State *c* in Figure 4.2 (Protein Topogenesis Simulations)

State c includes protein nascent chain configurations for which *(i)* the SP adopts the N_{exo}/C_{cyt} orientation, *(ii)* all the hydrophobic beads in the SP occupy the membrane interior (Figure C.10, Region C), and *(iii)* the translocon LG is closed.

C.2.4 Definition of States in Figure 4.4 (Stop-Transfer Simulations)

For the purposes of quantitatively defining the states in Figure 4.4, the configuration space for each CG bead is divided into four regions (Figure C.10). These regions include the cytosolic region (Figure C.10A), the translocon region (Figure C.10B), the membrane region (Figure C.10C), and the lumenal region (Figure C.10D). State a (Figure 4.4) is then defined to include configurations of the protein nascent chain for which all CG beads of the H-domain occupy the cytosolic region and for which no beads of the protein nascent chain (except those in the anchor domain) occupy the membrane region. State d includes configurations for which all CG beads of the H-domain occupy the lumenal region and for which no CG beads of the protein nascent chain (except those in the anchor domain) occupy the

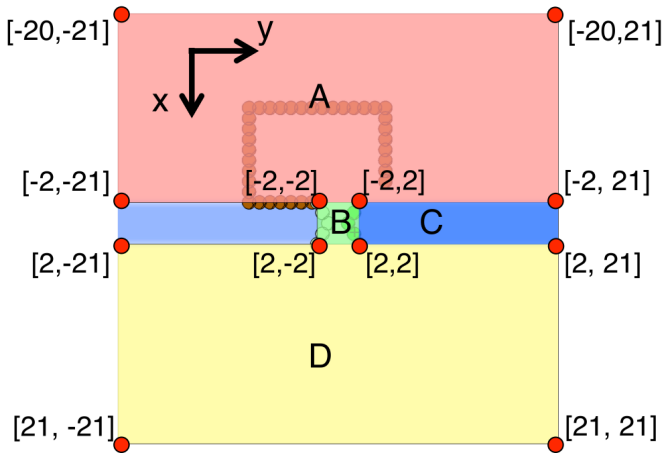


Figure C.10: Regions of the CG model used in defining intermediates states for protein translocation and membrane integration. Region **A** (red) encloses the cytosolic region; region **B** (green) includes the translocon channel; region **C** (blue) consists of the hydrophobic interior of the membrane; region **D** (yellow) includes the lumenal region. Each region is defined as a rectangle with the indicated vertex position; all coordinates are reported in coordinate system described in Figure C.1 and in Table C.2.

membrane region. State c includes configurations for which all three of the $\mathscr{L}$-type CG beads of the H-domain occupy the membrane region and for which the translocon LG is in the closed state. State b includes configurations for which the center of mass of the H-domain occupies the translocon region, while none of the three $\mathscr{L}$-type beads occupies the membrane region. State c^* includes configurations for which all three of the $\mathscr{L}$-type beads occupies the membrane region, at least one of the other CG beads in the H-domain occupies the translocon region, and the translocon LG is the open state.

C.2.5 Equilibrium Rate Calculations

The thermal rate constants reported in Figures 4.5C and C.25 are computed from long, equilibrium CG trajectories. Specifically, we utilize 100 independent CG trajectories, each of length $T = 40$ seconds. The trajectories are performed with a fixed MDL for the protein nascent chain, and its C-terminal bead is held fixed at the ribosome exit channel. The

ribosome remains in complex with the translocon throughout these equilibrium simulations. The equilibrium transition rates are obtained from the frequency of interstate transitions in a long trajectory [160, 161], using $k_{ij} = \frac{N_{ij}}{T_i}$. Here, a transition from state i to state j is defined as an event in which the trajectory leaves state i and reaches state j before visiting any other state. The term N_{ij} corresponds to the total number of transitions from state i to state j in a trajectory of length T. The term T_i corresponds to the amount of time that the systems occupies state i during the trajectory. This estimate for k_{ij} is obtained by averaging over estimates from the independent trajectories.

This protocol is repeated for the different values of ΔG reported in Figure 4.5C and the different values for the protein C-terminal tail length (CTL) in Figure C.25.

C.3 Explicit Modeling of Lumenal BiP

BiP is an essential component of the Sec translocon machinery that resides in the ER lumen and binds to translocated regions of the protein nascent chain [107, 113, 162]. To keep the CG model as simple as possible, the explicit role of BiP on the dynamics and mechanism of protein translocation is generally not included in the simulations reported in chapter 4. To validate this simplification of the CG model, the current section investigates the sensitivity of our simulation results to the explicit inclusion of BiP. In no case do we find that explicit inclusion of BiP alters the conclusions reported in chapter 4.

C.3.1 Algorithm

Molecular motors such as BiP interact with the portion of the nascent chain exposed to the ER lumen, thus biasing or rectifying the motion of protein domains that undergo translocation across the membrane [55]. The following algorithm is employed to explicitly model this “Brownian ratcheting" effect of BiP on the mechanism and kinetics of protein translocation; unless otherwise indicated, the CG simulations presented in this study do not utilize this algorithm to explicitly include the effects of lumenal BiP.

For both the protein topogenesis simulations and the stop-transfer simulations, the entire mature domain of the protein nascent chain is grouped into nonoverlapping BiP binding sites, each of which consists of four consecutive CG beads (12 amino-acid residues) [99]. Upon reaching the lumenal region of the system (defined as configurations for which all four of the corresponding CG beads occupy the lumenal region), each binding site adopts either an on-state, in which it is in complex with BiP, or an off-state, in which it is available for BiP binding. Stochastic transitions between the on- and off-states are attempted at every CG timestep with rates of $k_{\text{on}} = 60\ \text{min}^{-1}$ and $k_{\text{off}} = 1\ \text{min}^{-1}$ [99, 100]. For CG beads that comprise an occupied BiP binding site, a biasing force of 2.0 pN is applied when the bead position approaches within a distance of 2σ from the lumenal mouth of the transocon (at position $(x,y)/\sigma = (2,0)$); the biasing force is aligned with the x-axis in the CG coordinate system.

C.3.2 Numerical Tests of Explicit BiP Binding

Figures C.11 and C.12 demonstrates the effect of explicit BiP binding on CG simulations of protein topogenesis and stop-transfer efficiency. In Figure C.11, for CG simulations of protein topogenesis, the fraction of Type II integration is plotted as a function of MDL for several sets of simulations, and we compare the effect of including explicit BiP binding and of terminating the CG trajectories at state f. Comparison of the two data sets that utilize BiP binding (blue, filled and unfilled) indicates no effect upon terminating trajectories at state f rather than state g. Comparison of the two sets of trajectories that are terminated at state f (unfilled blue and red) indicates no effect upon inclusion of BiP binding. The physical basis for these results is that the SP reverse-flipping ($f \rightarrow c$) transition occurs on a timescale that is slow in comparison to BiP binding, such that BiP locks the SP into the Type II orientation upon arriving at state f; the results presented in Figure C.11 would not be expected to fully hold if SP reverse-flipping occurred more rapidly than BiP binding. In summary, Figure C.11 indicates that CG simulations performed with explicit BiP and that terminate Type II integration at state g yield no significant differences from CG simulations that are performed without explicit BiP binding and that terminate Type II integration at state f. The latter protocol requires shorter CG trajectories and is thus employed in chapter 4.

Figures C.12A and C.12B illustrate the effect of BiP binding on stop-transfer efficiency. In each case, explicit BiP binding reduces stop-transfer efficiency, since backsliding along the secretion pathway is inhibited. However, qualitative features of the CG simulation results are unaffected by inclusion of explicit BiP binding. In particular, Figure C.12A shows

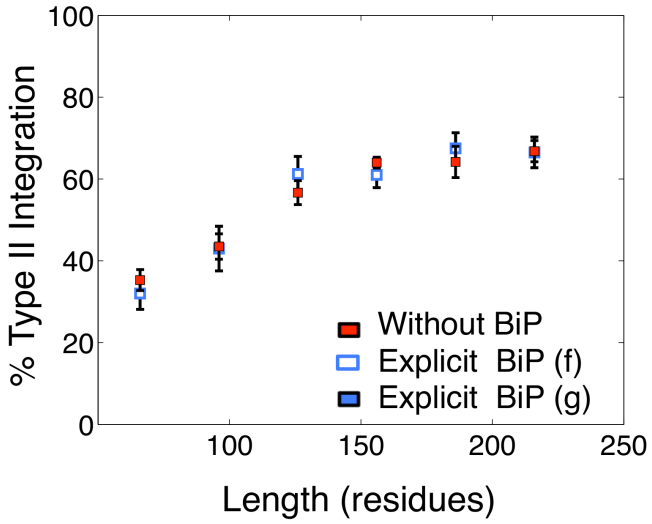


Figure C.11: The effect of explicit BiP binding on CG simulations of protein topogenesis. In CG simulations of protein topogenesis, the fraction of Type II integration is plotted as a function of MDL for several sets of simulations. The first set (blue, filled) employs explicit BiP binding, and the CG trajectories undergoing Type II integration are terminated only upon reaching state *g* (Figure 4.2). The second set (blue, open) differs only in that CG trajectories undergoing Type II integration are terminated upon reaching state *f* (Figure 4.2). The third set (red) does not include explicit BiP binding, and the CG trajectories undergoing Type II integration are terminated upon reaching state *f*. All data sets are obtained using the same insertion rate and SP sequence; the red data set is identical to that reported for the RL_6E sequence in Figure 4.3B of chapter 4. These results indicate that explicit BiP binding leads to no significant change in the CG simulations of SP orientation.

that the sigmoidal dependence of stop-transfer efficiency on the H-domain hydrophobicity is preserved in the presence of BiP. Figure C.12B illustrates that regardless of explicit BiP binding, stop-transfer efficiency with CTL for short C-terminal domains, and it decreases with CTL for long C-terminal domains. It is also seen that regardless of explicit BiP binding, stop-transfer efficiency decreases with ribosomal translation rate for proteins with long CTL, whereas no such effect is observed at short CTL. Even though explicit BiP binding affects the quantitative value for stop-transfer efficiency, it does not influence the underlying mechanism for TM partitioning or any of the conclusions presented in chapter 4.

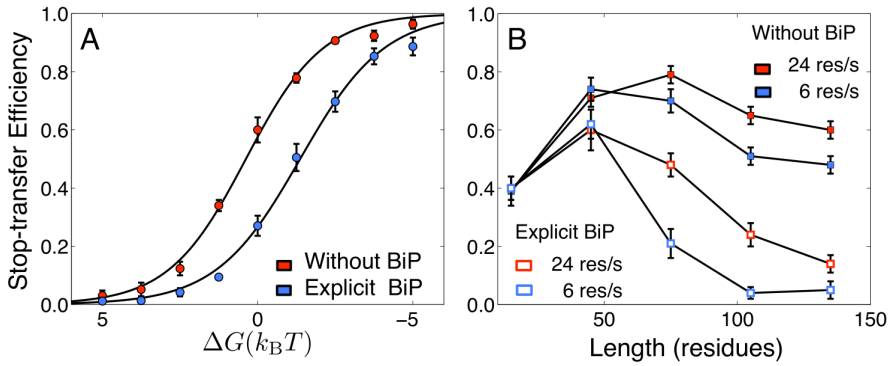


Figure C.12: The effect of explicit BiP binding on CG simulations of stop-transfer efficiency. **(A)** Stop-transfer efficiency as a function of H-domain hydrophobicity, for CG simulations that either include (blue) or do not include (red) explicit BiP binding. All data sets are reported using the same insertion rate and CTL; the red data set is identical to that reported in Figure 4.5A of chapter 4, and the blue data set is identical to that reported in Figure 4.5B$_2$ of chapter 4. **(B)** The dependence of stop-transfer efficiency on CTL, insertion rate, and explicit BiP binding. The data sets that are reported with filled symbols do not include explicit BiP binding; these two data sets are identical to the results presented in Figure 4.5D of chapter 4. The data sets that are presented with open symbols correspond to simulations that differ only in that they include explicit BiP binding. For the CG simulations of stop-transfer efficiency, including explicit BiP binding leads to reduced stop-transfer efficiency, since backsliding along the secretion pathway is inhibited. However, qualitative features of the CG simulation results are unaffected by inclusion of explicit BiP binding.

C.4 Additional Validation and Predictions for Protein Topogenesis

C.4.1 Hydrophobic Patches in the Mature Domain

Figure C.13 demonstrates significant dependence of the CG simulations of protein topogenesis on both the hydrophobicity (Figure C.13A) and the location (Figure C.13B) of hydrophobic patches in the mature domain of the protein nascent chain. The results can be understood from the effect of the hydrophobic patches in the Type II flipping pathway for membrane integration, which involves reorientation of the SP from the N_{exo}/C_{cyt} to the opposite topology. The flipping transition is facilitated by the transient opening of the LG,

the energetics of which depend on the hydrophobicity of the protein nascent chain beads that occupy the channel interior. The probability of undergoing the flipping transition thus decreases as the hydrophobic patch plays a smaller role in stabilizing the transient opening of the translocon LG, either because the patch is less hydrophobic or because it occupies a more distant region of the mature domain.

We note that evidence for both SP and mature-domain effects in protein topogenesis have been observed experimentally. For example, Hegde and co-workers have found that the functionality of the SP in gating the translocon is evolutionarily matched with the mature domain to facilitate efficient translocation [163]. Of course, given the simplicity of the CG model developed here, it is important to avoid overinterpreting mechanistic details of the simulations. We simply emphasize that the SP flipping transition gives rise to a slow timescale in Type II membrane integration that leads to characteristic trends in protein topogenesis, and the hydrophobic patches in the mature domain play a significant role in our CG model of facilitating this flipping transition.

C.4.2 Charged-Residue Mutations on the Translocon

Figure C.14 illustrates that charged residue mutations on the translocon lead to significant changes in integral membrane protein topology. The red data set corresponds to the protein topogenesis results presented for the RL_4E SP sequence in Figure 4.3A of chapter 4. The blue data set is obtained using the same protein sequences and removing the positive charge on the lumenal side of the translocon LG (see Figure 4.2 of chapter 4); the negatively charged CG bead on the cytosolic side of the translocon LG is left unchanged. As is seen in

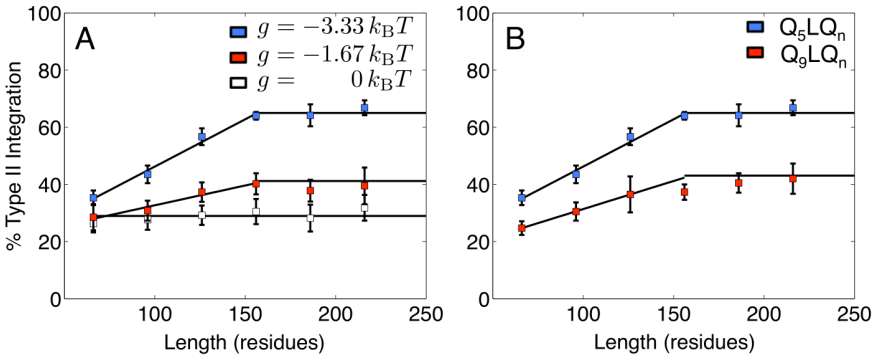


Figure C.13: Testing the sensitivity of SP orientation to hydrophobic patches in the mature domain of protein nascent chain. **(A)** The MDL-dependence of the fraction of CG trajectories that undergo Type II integration, obtained using various values for the water-membrane transfer FE of the L-type CG bead in the protein mature domain (i.e., the CG bead representing the hydrophobic patch). The protein nascent chain SP sequence of RL_6E is employed; the blue data set reported here is identical to that reported for the RL_6E sequence in Figure 4.3B of chapter 4. These results indicate that the CG model exhibits significant dependence of the SP orientation to hydrophobic patches in the mature domain. **(B)** The MDL-dependence of the fraction of CG trajectories that undergo Type II integration, obtained with different spacing between the SP and the hydrophobic patch in the mature domain. The blue data set was obtained using the mature domain sequence Q_5LQ_n, as in chapter 4; this data set is identical to the blue data set in part (A). The red data set was obtained by changing the mature domain sequence to Q_9LQ_n. The effect of the mature-domain hydrophobic patch in the CG model diminishes with its separation from the SP.

the figure, the charge mutation leads to reduction of Type II integration. General features of the nascent-protein length dependence remain unchanged. The plateau value for the Type II integration at long MDL is reduced by approximately 10%. These results illustrate the role of charged translocon residues in establishing the “positive-inside rule" for integral membrane protein topogenesis. Similar charge-mutation trends have been experimentally observed [63, 84]; these studies considered Arg-to-Glu mutations of residues R67 and R74 on the translocon plug domain, which lead to 10-20% reduction of Type II integration. Despite the good agreement between simulation and experiment found here, we note that the CG model does not explicitly represent the conformation of the translocon plug moiety, which has also been suggested to impact topogenesis [84].

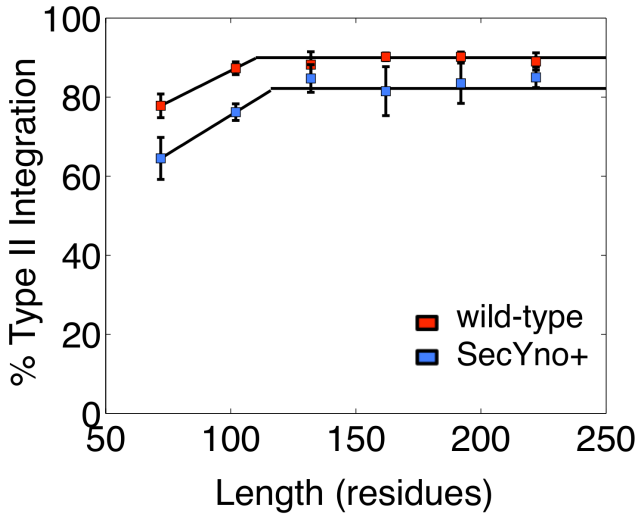


Figure C.14: Testing the effect of charged-residue mutations on the translocon. The figure plots Type II integration fraction as a function of MDL. The red data set corresponds to the protein topogenesis results in Figure 4.3A for the RL_4E SP sequence. The blue data set is obtained using the same protein sequences and removing the positive charge on the lumenal side of the translocon LG (see Figure 4.2 of chapter 4); the negatively charged CG bead on the cytosolic side of the translocon LG is left unchanged.

C.4.3 Charged-Residue Mutations on the Nascent-Protein Mature Domain: A Multispanning Protein Example

One of the most remarkable recent experimental results on protein topogenesis is that distant C-terminal residues can control the overall topology of a multispanning integral membrane protein [164]. In Figure C.15, we illustrate that this effect is also captured in the CG model presented here. The figure presents results from the direct simulation of membrane integration for a multispanning integral membrane protein. Specifically, we consider two different nascent protein sequences, each of which exhibits three hydrophobic TM domains (Figure C.15A). The distribution of flanking charges for the first two TM domains is identical for the two protein sequences. For Protein 1, the third TM domain includes a single positively charged bead at its N-terminal end, whereas for Protein 2, the third TM domain

includes three positively charged beads at its C-terminal end. In complete detail, the sequence of CG beads for Protein 1 is $RL_4EQ_3L_4RQ_3RL_4Q_{28}$, and the sequence for Protein 2 is $RL_4EQ_3L_4RQ_4L_4R_3Q_{25}$.

Membrane integration of Proteins 1 and 2 is directly simulated using the same membrane topogenesis protocol as in chapter 4. The CG trajectories are terminated when all of the following criteria are met: *(i)* ribosomal translation is completed, *(ii)* all three TM domains span the membrane (Figure C.15B), and *(iii)* the first two TM domains at the N-terminal end of the protein have diffused to a distance of 16 nm from the translocon (which is sufficient to ensure that the third TM has also released from the channel).

Figure C.15C presents the calculated fraction of trajectories that lead to the N_{cyt}/C_{exo} orientation for the two protein sequences. Both protein sequences exhibit a final product that is consistent with the positive-inside rule, despite the fact that this rule is dictated by the third TM domain. Consistent with the earlier experimental study [164], these simulations suggest that overall integral membrane topology can remain undetermined until the final stages of ribosomal translation.

C.4.4 Positive vs. Negative N-terminal Charges on the Nascent Protein

The results in Figure 4.3A of chapter 4 emphasize that the model captures the essential features of the positive-inside rule for protein topogenesis. Specifically, comparison of nascent proteins for which the SP has a positively charged N-terminus (RL_4E) with those for which the SP has a neutral N-terminus (QL_4E) indicates that the positive charge leads

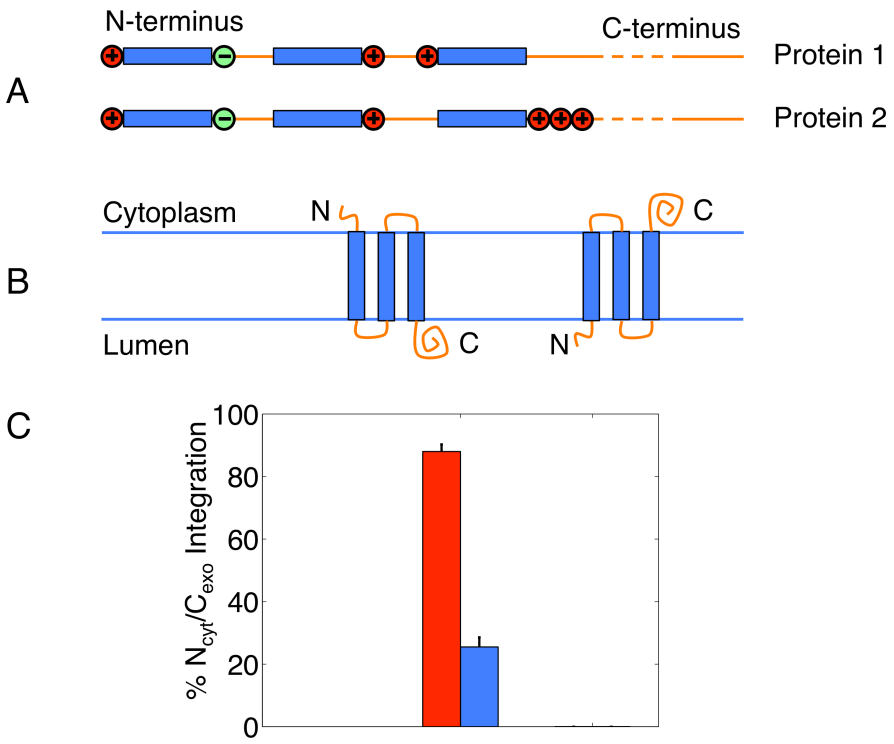


Figure C.15: Testing the effect of distant charged-residue mutations on the nascent-protein mature domain. (**A**) Schematic representation of the CG bead sequences for Proteins 1 and 2, which have three TM domains and which differ only with respect to the charge distribution in the third. (**B**) Illustration of the possible overall topologies for the two multispanning proteins. (**C**) The fraction of CG insertion trajectories that lead to the N_{cyt}/C_{exo} topology for Protein 1 (red) and Protein 2 (blue).

to a greater fraction of Type II integration. This effect is well established experimentally [165].

A natural question, then, is whether the CG model also predicts a “negative-outside" bias, for which a negatively charged SP N-terminus leads to a greater degree of Type III integration. This effect is less clearly established experimentally, with studies both observing [166, 167] and not observing [168] the negative-outside bias on protein topology.

As is seen in Figure C.16, the CG model also finds mixed results with respect to negative-outside bias. Simulations presented in the figure employ the same protein to-

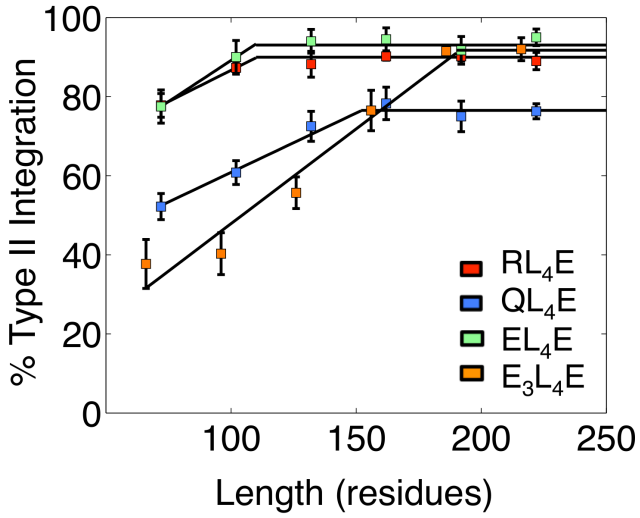


Figure C.16: Testing the effect of negatively charged N-terminal residues on SP orientation. The figure plots Type II integration fraction as a function of MDL. The red and blue data set corresponds to the protein topogenesis results in Figure 4.3A for the RL_4E and QL_4E SP sequences, respectively. Also shown are results for which the SP sequence includes either one (EL_4E) or three (E_3L_4E) negatively charged N-terminal CG beads.

pogensis simulation protocol as is used for Figure 4.3A in chapter 4. In addition to the results for the SP with an uncharged (QL_4E) and a positively charged N-terminus (RL_4E), we also include results for which the SP exhibits a single negatively charged N-terminal bead (EL_4E) or three negatively charged beads (E_3L_4E). Remarkably, inclusion of a single negatively charged bead at the SP N-terminus (EL_4E) is found to have essentially the same effect as a single positively charged bead (RL_4E); this result is inconsistent with a negative-outside bias. However, upon inclusion of additional negatively charged beads (E_3L_4E), the negative-outside bias is observed for relatively short MDL. The competition of factors associated with negative-outside bias are found to be more complex than those leading to the positive-inside rule, which may help to explain the variation in experimental findings. We further note that detailed molecular interactions of the charged residues with the lipid bilayer, which are greatly simplified in the CG model, may substantially impact

these findings [70, 169].

C.5 Additional Validation and Predictions for Stop-Transfer Efficiency

C.5.1 Hydrophobic Patches in the C-terminal Domain

Figure C.17A illustrates the dependence of the CG simulations of stop-transfer efficiency on hydrophobic patches in the C-terminal domain of the protein nascent chain. Removal of the hydrophobic patches leads to a shift in favor of increased membrane integration. Without hydrophobic patches, the C-terminal domain residues in the translocon channel do little to stabilize opening of the LG; therefore, once the system reaches state *c* (Figure 4.4) along the pathway to membrane integration, it is less likely that the H-domain will return to the channel interior and then undergo secretion. The result is an increase in membrane integration upon removal of the hydrophobic patches from the C-terminal domain. We note that this interpretation is consistent with the observed enhancement of the nonequilibrium state population for state *c*, P_c, upon removal of the hydrophobic patches from the C-terminal domain (Figure C.17B). Sensitivity of stop-transfer efficiency to C-terminal domain sequence has also been observed in experimental studies [170].

C.5.2 Charged-Residue Mutations Flanking the H-domain

Experimental studies have also found that charged residues flanking the nascent-protein H-domain affect stop-transfer efficiency [2, 171, 172]. Figure C.18 illustrates this effect in

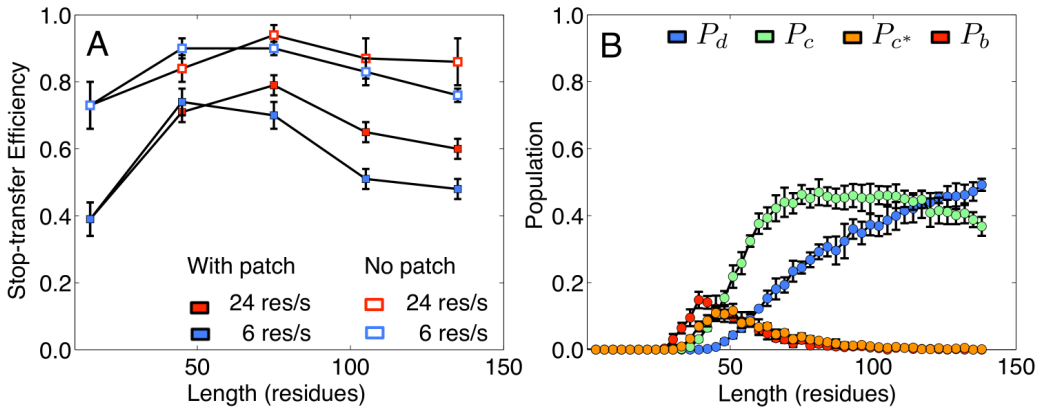


Figure C.17: Testing the effects of hydrophobic patches in the C-terminal domain on H-domain stop-transfer efficiency. **(A)** The dependence of stop-transfer efficiency on the peptide CTL, ribosomal translation rate, and mature domain sequence. The results reported with filled data points are identical to those reported in Figure 4.5D. The results reported with open data points correspond to the same calculations with the CG model, except that hydrophobic patches in the C-terminal domain of the protein nascent chain are removed. Specifically, the V-type beads in the C-terminal domain of the protein sequence used to construct Figure 4.5D are substituted with Q-type CG beads. **(B)** Nonequilibrium populations of the states in Figure 4.4 at the time of stop-translation for proteins of various CTL. The CG trajectories used to make this figure employed protein sequences without hydrophobic patches and a ribosomal translation rate of 24 res/s; the results correspond exactly to the open-red data points in part A. Comparison of the nonequilibrium populations in this figure with the results obtained for proteins that include hydrophobic patches (Figure C.23A) reveals enhancement of P_c.

the CG model presented here. The dashed line in the figure corresponds to the stop-transfer efficiency results reported in Figure 4.5A of chapter 4. The blue data set is obtained using the same protein sequences, except that the three CG beads in the C-terminal domain that directly flank the nascent protein H-domain are mutated from being hydrophilic and neutral (Q-type) to being hydrophilic and positively charged (R-type).

As is seen in Figure C.18, the charged-residue mutations lead to a substantial shift toward increased membrane integration of the nascent-protein H-domain. Analysis of the CG trajectories reveals the mechanistic basis for this trend. Whereas progress along the secretion pathway (state *b* to state *d* in Figure 4.4 of chapter 4) involves sacrificing the favorable electrostatic interaction between positively charged flanking beads on the nascent

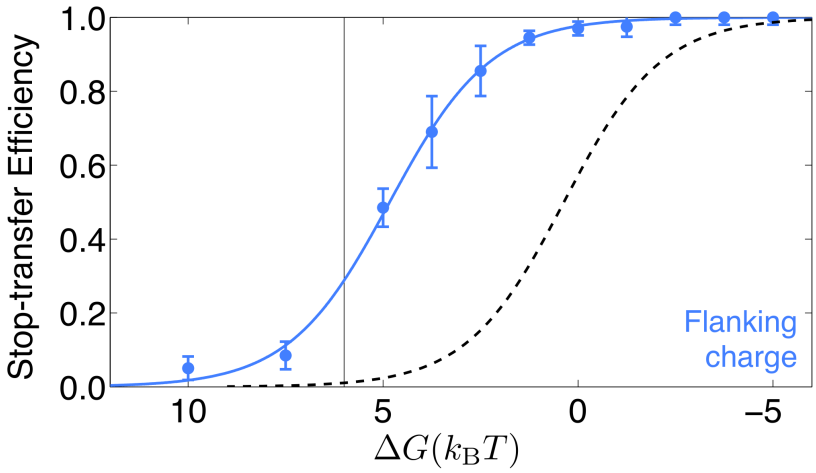


Figure C.18: Testing the effect of charged-residue mutations flanking the nascent-protein H-domain. The figure plots stop-transfer efficiency as a function of H-domain hydrophobicity. The black dashed line is the sigmoidal fit to the data presented in Figure 4.5A of chapter 4. The blue data set is obtained using the same protein sequences, except that the three CG beads in the C-terminal domain that directly flank the nascent protein H-domain are mutated from being hydrophilic and neutral (Q-type) to being hydrophilic and positively charged (R-type).

protein and the negatively changed bead on the translocon, progress along the integration pathway (state b to state c^* to state c) allows this electrostatic contact to be preserved. In effect, the charged residues lead to enhancement of the nonequilibrium population of state c in favor of state d, which leads to an enhancement of the membrane integration product. These simulations suggest that the C-terminal positive charges enhance the stop-transfer efficiency of marginally hydrophobic TM segments, which is consistent with experimental observation [2, 171, 172].

C.5.3 Dependence of Protein Translocation Time on Nascent Protein Hydrophobicity

Previous stop-transfer experiments have concluded that hydrophobic nascent-protein segments exhibit stalling, or pausing, in the translocon channel [2]. Protein translocation mod-

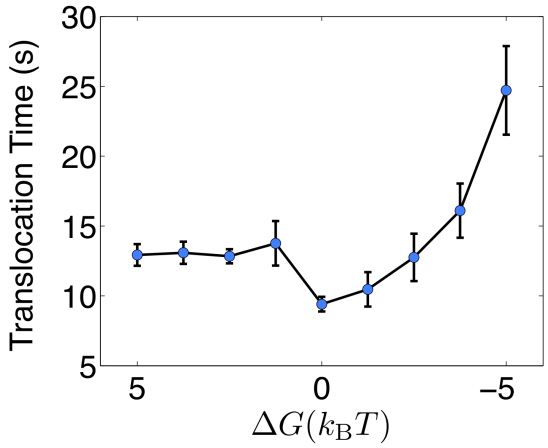


Figure C.19: Dependence of the average translocation time for secreted proteins on H-domain hydrophobicity. The protein sequences and stop-transfer simulation protocols employed here are the same as those used to construct Figure 4.5A in chapter 4.

eling has also led to the prediction that hydrophobic segments retard translocation due to lateral partitioning [99]. Figure C.19 investigates this effect using the current CG model. We calculate the average simulation time for trajectories to reach the secretion product; trajectories that lead to the membrane integration product are not included in the average. The protein sequences and stop-transfer simulation protocols used to construct Figure C.19 are the same as those used to construct Figure 4.5A in chapter 4.

For hydrophilic and amphiphilic H-domain sequences ($\Delta G > -2k_{\mathrm{B}}T$), the CG model predicts relatively weak dependence of the protein translocation time on the H-domain hydrophobicity (Figure C.19). However, for more strongly hydrophobic H-domain sequences, the translocation time is found to increase by a factor of 2-3. This increase in translocation time is in qualitative agreement with the experimental study in Ref.2. Furthermore, the results in Figure C.19 bear striking resemblance to the exponential increase in translocation time with H-domain hydrophobicity that is predicted in Ref. 99. We emphasize that any experimental or theoretical measurement of protein translation time must take care (as is

done here) to avoid contamination due to the increased formation of membrane integration product with strongly hydrophobic H-domain sequences.

C.6 Analytical Model for TM partitioning

We derive an analytical kinetic model based on Markovian transitions among the states illustrated in Figure 4.4 in chapter 4. The time evolution of the state populations is described using a master equation that includes

$$\frac{dP_c(t;\Delta G)}{dt} = P_{c^*}(t;\Delta G)k_{c^*c} - P_c(t;\Delta G)(k_{cc^*} + k_{cf}), \tag{C.4}$$

and

$$\frac{dP_d(t;\Delta G)}{dt} = P_b(t;\Delta G)k_{bd} - P_d(t;\Delta G)(k_{db} + k_{de}). \tag{C.5}$$

The analytical model invokes the primary assumptions that *(i)* partitioning of the H-domain across the LG (i.e., transitions between states b and c^*) occurs on a faster timescale than all other transitions in the system, such that these states are always in equilibrium, *(ii)* the populations of states b and c^* are slowly varying on the timescale of translocation and integration (i.e., these populations satisfy a steady-state approximation), and *(iii)* only the rate of transitions between states b and c^* depend on the H-domain hydrophobicity.

From the first two assumptions, we arrive at equation (4.2) in chapter 4:

$$\ln\left[\frac{P_b(t;\Delta G)}{P_{c^*}(t;\Delta G)}\right] = \ln\left[\frac{P_b(\Delta G)}{P_{c^*}(\Delta G)}\right] = -\beta\alpha\Delta G + C, \tag{C.6}$$

where the time dependence for $P_b(t;\Delta G)$ and $P_{c^*}(t;\Delta G)$ is removed as a consequence of the steady-state assumption; in the second equality, the relative FE of states b and c is assumed to decompose into additive contributions from a term that is proportional to the H-domain transfer FE (ΔG) and a remaining term that includes contributions from the relative entropy.

We now focus on deriving equations (4.3) and (4.4) in chapter 4 from equations (C.4) and (C.5). We begin by considering equation (C.4), which is a linear ordinary differential equation with solution [173]

$$u(t)P_c(t;\Delta G) = P_{c^*}(\Delta G)\int u(t)k_{c^*c}dt + \text{const.}, \tag{C.7}$$

where $u(t) = e^{\int(k_{cc^*}+k_{cf})dt}$. In general, the transition rates are time dependent, since they vary with the elongation of the protein nascent chain (Figure C.25). However, using that the local relaxation time for the protein nascent chain *within* each of the intermediates states (i.e., milliseconds) is fast in comparison to the timescale for elongation of the protein nascent chain (i.e., seconds), it follows that the transition rates are piecewise constant functions of time, such that $k_{c^*c}(t) = k^n_{c^*c}$, where $t \in [t_{n-1}, t_n]$, $n = \text{floor}(t/\Delta t)$ corresponds to the number of protein nascent chain beads that have been translated since the initialization of the CG trajectories, Δt is the time increment between ribosomal translation events, and $t_j = j\Delta t$ for integer values of j. Note that this step introduces no new assumptions about the relative timescales for transitions *between* intermediates states and the timescale for protein elongation.

Using that $k^n_{c^*c}$ is constant over the time increment associated with each value of n,

solution of equation (C.4) in the time interval $t \in [t_{n-1}, t_n]$ yields

$$\begin{aligned} P_c(t;\Delta G) &= P_{c^*}(\Delta G) h_n(t) \\ &+ P_c(t_{n-1};\Delta G) e^{-(k^n_{cc^*}+k^n_{cf})(t-t_{n-1})}, \end{aligned} \tag{C.8}$$

where

$$h_n(t) = \frac{k^n_{c^*c}}{k^n_{cc^*}+k^n_{cf}} \left(1 - e^{-(k^n_{cc^*}+k^n_{cf})(t-t_{n-1})}\right). \tag{C.9}$$

Finally, using assumption *(iii)*, it follows that $h_n(t)$ is independent of ΔG.

Using induction, we now argue that equation (C.8) implies that for each time interval $t \in [t_{n-1}, t_n]$,

$$P_c(t;\Delta G) = P_{c^*}(\Delta G) f_n(t), \tag{C.10}$$

where $f_n(t)$ is a function that is independent of ΔG.

Firstly, we note that equation (C.10) holds for the case of $n = 1$. At the start of the translation, state c is unoccupied, such that $P_c(0;\Delta G) = 0$. It then follows from equation (C.8) that for $t \in [t_0, t_1]$,

$$P_c(t;\Delta G) = P_{c^*}(\Delta G) h_1(t), \tag{C.11}$$

where $h_1(t)$ is defined in equation (C.9) and is independent of ΔG.

In performing the induction step, we argue that if equation (C.10) holds for the case of $n-1$, it must also hold for the case of n. Assume that in the time interval $t \in [t_{n-2}, t_{n-1}]$,

$$P_c(t;\Delta G) = P_{c^*}(\Delta G) f_{n-1}(t), \tag{C.12}$$

where $f_{n-1}(t)$ is independent of ΔG. At the end of this time interval, $P_c(t_{n-1};\Delta G) = P_{c^*}(\Delta G) f_{n-1}(t_{n-1})$. Inserting this result into equation (C.8), it follows that in the time interval $t \in [t_{n-1}, t_n]$,

$$P_c(t;\Delta G) = P_{c^*}(\Delta G) f_n(t), \tag{C.13}$$

where $f_n(t) = h_n(t) + f_{n-1}(t_{n-1}) e^{-(k^n_{cc^*}+k^n_{cf})(t-t_{n-1})}$ is independent of ΔG. This completes the induction step, as well as the demonstration that equation (C.10) holds for each time interval.

Applying the analogous series to equation (C.5) leads to the result that for each time interval $t \in [t_{n-1}, t_n]$,

$$P_d(t;\Delta G) = P_b(\Delta G) s_n(t), \tag{C.14}$$

where $s_n(t)$ is a function that is independent of ΔG.

Combining the results of equations (C.14) and (C.10), we obtain that for each time interval $t \in [t_{n-1}, t_n]$,

$$\begin{aligned} \ln\left[\frac{P_d(t;\Delta G)}{P_c(t;\Delta G)}\right] &= \ln\left[\frac{P_b(\Delta G)}{P_{c^*}(\Delta G)} \frac{f_n(t)}{s_n(t)}\right] \\ &= -\beta\alpha\Delta G + C + \ln\left[\frac{f_n(t)}{s_n(t)}\right] \\ &= -\beta\alpha\Delta G + \delta(t), \end{aligned} \tag{C.15}$$

where $\delta(t)$ is independent of ΔG. This is the result stated in equation (4.3) of chapter 4; the remaining derivation of equations (4.4) and (4.5) in chapter 4 is straightforward.

The analytical model for TM partitioning derived here assumes that trajectories for the stop-transfer simulations flow entirely through the two primary pathways shown in Figure 4.4 in chapter 4. The assumptions of this analysis weaken for cases in which minor pathways for integration and secretion (such as those illustrated in Figure C.24) become significant. This is indeed the cause of the slight deviation of the parameter α associated with the sigmoidal fit in Figure 4.5B$_3$ (i.e., $\alpha = -0.60$ for the sigmoidal fit of the data in Figure 4.5B$_3$, whereas $\alpha \approx -0.80$ for the other data sets in Figures 4.5A and 4.5B). A minor pathway for membrane integration (P_{I}, Figure C.24A) plays an increasing role in proteins with elongated CTL (Figure C.24B).

Table C.3: Parameters for the nonbonded interactions (equation (4.6)).

		R	E	L	Q	V	P	$\mathscr{L}$	Ribosome	LG_n	LG_o	LG_c
$\varepsilon_{lj}/\varepsilon$	R	1	1	1	1	1	1	1	1	1.5	1.5	1.5
	E		1	1	1	1	1	1	1	1.5	1.5	1.5
	L			1	1	1	1	1	1	1.5	1.5	1.5
	Q				1	1	1	1	1	1.5	1.5	1.5
	V					1	1	1	1	1.5	1.5	1.5
	P						1	1	1	1.5	1.5	1.5
	$\mathscr{L}$							1	1	1.5	1.5	1.5
r_{cl}/σ	R	0	0	0	0	0	0	0	0	0	1.0	0
	E		0	0	0	0	0	0	0	0	1.0	0
	L			0	0	0	0	0	0	0	1.0	0
	Q				0	0	0	0	0	0	1.0	0
	V					0	0	0	0	0	1.0	0
	P						0	0	0	0	1.0	0
	$\mathscr{L}$							0	0	0	1.0	0
r_{cr}/σ	R	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	2.5	2.5	2.5
	E	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	2.5	2.5	2.5
	L	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	2.5	2.5	2.5
	Q	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	2.5	2.5	2.5
	V	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	2.5	2.5	2.5
	P	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	2.5	2.5	2.5
	$\mathscr{L}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	$2^{\frac{1}{6}}$	2.5	2.5	2.5

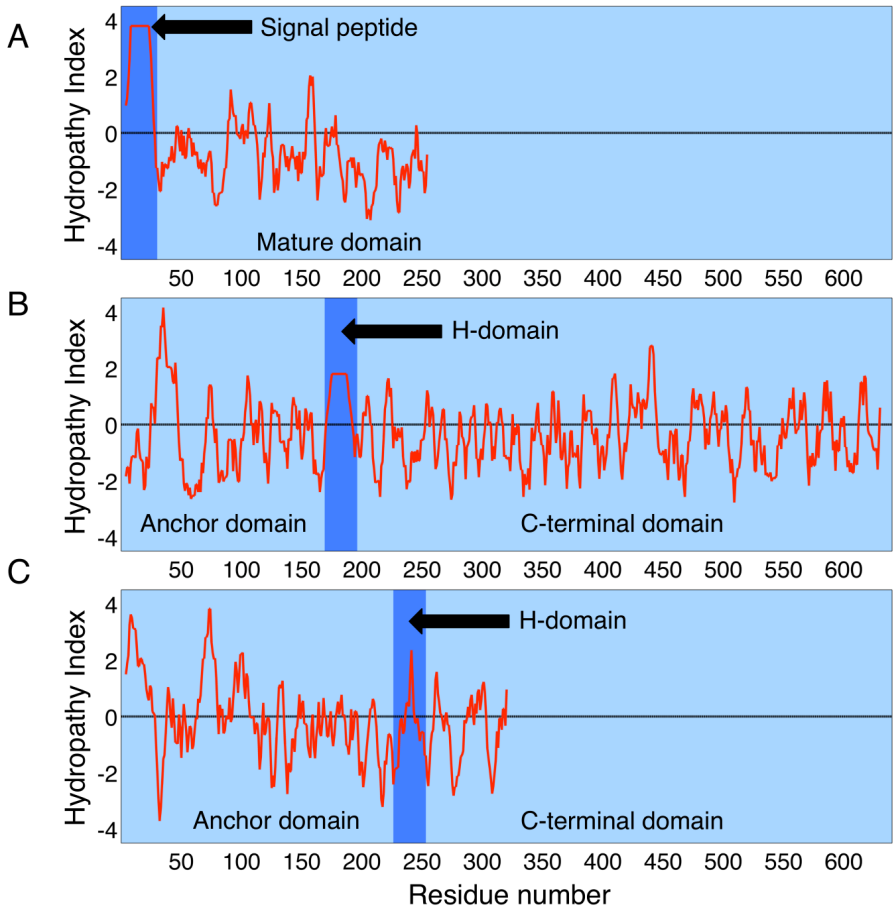


Figure C.20: The hydropathy profile for the protein sequences that are modeled in the current study. Each residue is scored according to the Kyte-Doolittle hydropathy measure [174], which is more positive for hydrophobic residues and more negative for hydrophilic residues; the profile is then plotted as a rolling average over seven-residue segments along the protein sequence. **(A)** The profile for the engineered protein H1ΔLeu22 [7], which serves as the model for the CG protein sequence employed in the *Signal Orientation and Protein Topogenesis* section of chapter 4. The SP and mature domain regions are indicated by the heavy and light blue regions, respectively. The mature domain sequence is generally hydrophilic, with periodic increases in local hydrophobicity (i.e., hydrophobic patches) about residue number 50, 100, etc. The sensitivity of the CG model to the strength and position of these hydrophobic patches are discussed in Figure C.13. **(B)** The profile for the modified dipeptidyl aminopeptidase B (DPAPB) protein sequence [58]. **(C)** The profile for the leader peptidase (Lep) protein sequence [5]. The protein sequences in (B) and (C) serve as the model for the CG protein nascent chain sequence employed in the *Regulation of Stop-Transfer Efficiency* section of chapter 4. The N-terminal anchor domain, H-domain, and C-terminal domain are indicated by the light, heavy, and light blue regions, respectively. In both cases, C-terminal domain sequence is generally hydrophilic, with closely spaced periodic patches. The predicted sensitivity of stop-transfer efficiency to these hydrophobic patches is discussed in Figure C.17.

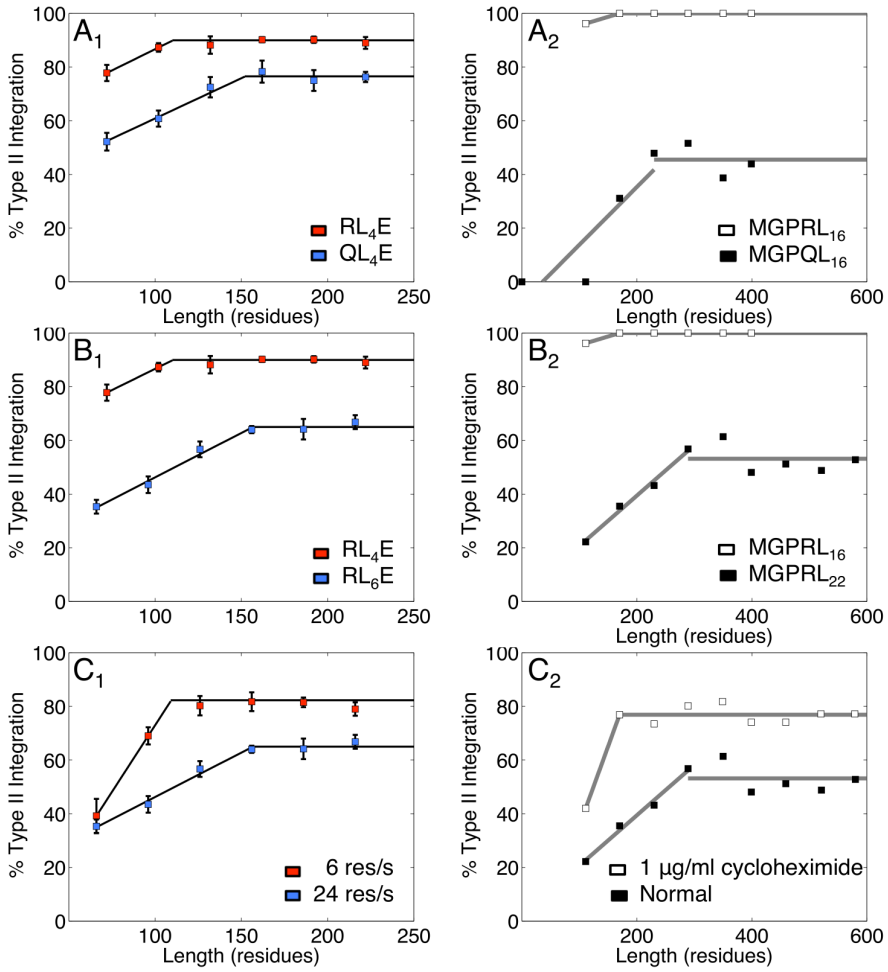


Figure C.21: Comparison of CG simulation results and experimental results for integral membrane protein topogenesis. The left column reproduces the simulation results from Figure 4.3 of chapter 4. The right column presents experimental results for the Type II integration fraction as a function of MDL, SP charge distribution, SP hydrophobicity, and ribosomal translation rate [7]. In all the three cases, we find the simulation predictions agree quantitatively with the experimental results. **(A)** Mutating off the positive charges at the N-terminus of the SP reduces the Type II integration. **(B)** Increasing the number of hydrophobic residues in SP reduces the Type II integration. **(C)** Increasing the ribosomal translation rate reduces the Type II integration.

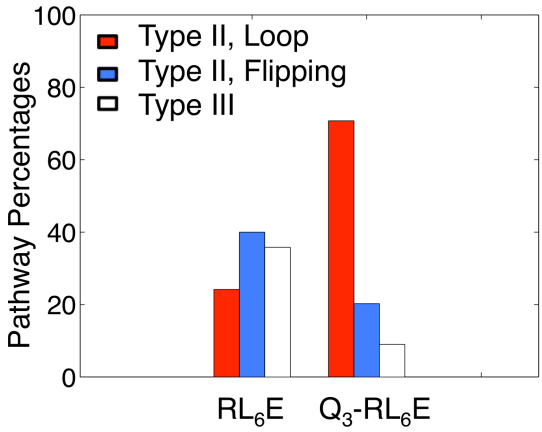


Figure C.22: Analysis of the membrane integration mechanism upon increasing the length of the nascent chain N-terminal domain length. For two data sets, the fraction of CG trajectories that pass through each of the kinetic pathways in Figure 4.2 is presented. The RL_6E data set is identical to that presented in Figure 4.3D. The Q_3-RL_6E data set is obtained in the same way, except that the protein nascent chain sequence is modified to include three additional Q-type CG beads at its N-terminus. Comparison of the two data sets indicates that increasing the N-terminal domain length leads to a substantial decrease in the relative fraction of trajectories that undergo Type II integration via the flipping mechanism.

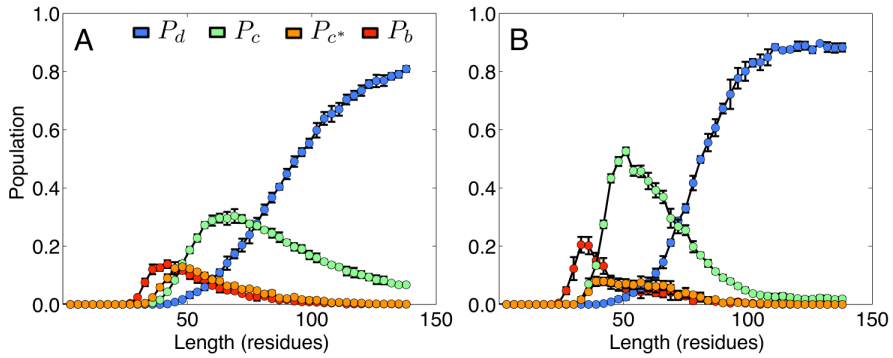


Figure C.23: Nonequilibrium populations of the states in Figure 4.4 at the time of stop-translation for proteins of various CTL, obtained from over 2000 CG trajectories for co-translational TM partitioning. These results are obtained using the same protein nascent chain sequences described in the *Direct simulation of co-translational TM partitioning* section of chapter 4, with H-domain hydrophobicity $\Delta G = -1.25 k_B T$. The CG trajectories employ a ribosomal translation rate of either **(I)** 24 res/s or **(II)** 6 res/s. Three observations from this figure pertain to the discussion in the *Kinetic and CTL effects in TM partitioning* section of chapter 4. Firstly, at longer CTL (≥ 75 residues), slowing ribosomal translation leads to an enhancement of P_d with respect to P_c. Secondly, at shorter CTL (≤ 75 residues), slowing translation does not lead to enhancement of P_d with respect to P_c. Thirdly, at both insertion rates, P_d increases monotonically, such that longer CTL always correspond to more population in state d at the time at which translation ends.

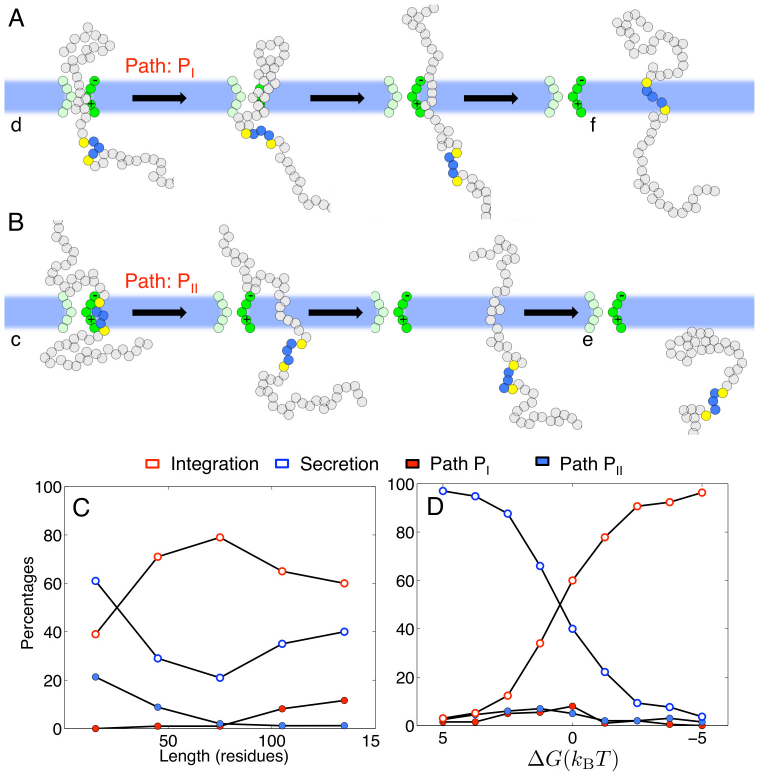


Figure C.24: Minor pathways observed for co-translational TM partitioning. **(A)** For pathway P_I, the protein nascent chain partitions into the membrane directly from state d (i.e., with the H-domain on the lumenal side of the membrane), and then the H-domain backslides into the membrane without passing through the translocon. States d and f are defined as in Figure 4.4. Trajectories are determined to have passed through the P_I pathway if neither state b nor state c^* is visited along the transition from d to f. **(B)** For pathway P_{II}, the H-domain transits to the lumenal side of the membrane from state c without reentering the translocon channel, and the C-terminal domain translocates across the membrane bilayer without passing through the translocon channel. States c and e are defined as in Figure 4.4. Trajectories are determined to have passed through the P_{II} pathway if neither state b nor state c^* is visited along the transition from c to e. **(C)** As a function of CTL, the percentage of CG trajectories that undergo membrane integration (red, open) versus secretion (blue, open), as well as the percentage that follow the P_I (red, closed) and P_{II} (blue, closed) pathways. We employ a protein nascent chain sequence for which the H-domain transfer FE is $\Delta G = -1.25 k_B T$; the ensemble of trajectories analyzed here corresponds to the red data set in Figure 4.5D of chapter 4. Membrane integration via the P_I pathway is most often observed at long CTL, since a longer CTL provides more opportunities for the C-terminal tail to partition through the translocon LG. H-domain secretion via the P_{II} pathway is most often observed at short CTL, since short C-terminal domains create a smaller energetic barrier to direct translocation through the membrane (i.e., with short C-terminus, the H-domain is less stably anchored in the membrane). Note that throughout the full range of CTL considered here, neither of these pathways is the dominant mechanism for protein translocation or membrane integration. **(D)** As a function of H-domain hydrophobicity, the percentage of CG trajectories that undergo membrane integration (red, open) versus secretion (blue, open), as well as the percentage that follow the P_I (red, closed) and P_{II} (blue, closed) pathways. We employ a protein nascent chain sequence for which the CTL is 75 residues; the ensemble of trajectories analyzed here corresponds to the data in Figure 4.5A of chapter 4.

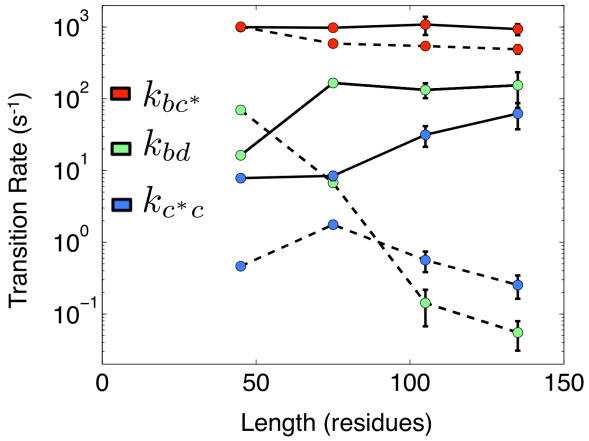


Figure C.25: Equilibrium transition rates between the states in Figure 4.4 as a function of CTL for the protein nascent chain. For each color, the forward rate is indicated with the solid line, and the reverse rate is indicated with dashed line. As is described in connection with Figure 4.5C of chapter 4, the transition rates are calculated from long, equilibrium CG trajectories for which the protein C-terminus is fixed at the ribosome exit channel. Firstly, note that the forward and reverse transition rates between states b and c^* are fast in comparison to the other transitions and relatively independent of CTL. Secondly, note that the forward transition rate k_{bd} increases with CTL, whereas the reverse transition rate k_{db} (green, dashed) dramatically decreases with increasing CTL. This decreased backsliding of the H-domain from state d into state b is of relevance to the discussion in the *Kinetic and CTL effects in TM partitioning* section of chapter 4. State b is destabilized relative to state d at long CTL because of crowding in the ribosome-translocon junction. A similar trend is seen in the forward and reverse rates between states c and c^*.

References

[1] Heinrich, SU, Mothes, W, Brunner, J, Rapoport, TA (2000) The Sec61p complex mediates the integration of a membrane protein by allowing lipid partitioning of the transmembrane domain. *Cell* 102:233–244.

[2] Sääf., A, Wallin., E, von Heijne, G (1998) Stop-transfer function of pseudo-random amino acid segments during translocation across prokaryotic and eukaryotic membranes. *Eur. J. Biochem.* 251:821–829.

[3] Duong, F, Wickner, W (1998) Sec-dependent membrane protein biogenesis: SecYEG, preprotein hydrophobicity and translocation kinetics control the stop-transfer function. *EMBO J.* 17:696–705.

[4] White, SH, von Heijne, G (2008) How translocons select transmembrane helices. *Annu. Rev. Biophys.* 37:23–42.

[5] Hessa, T et al. (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* 433:377–381.

[6] Hessa, T et al. (2007) Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* 450:1026–1030.

[7] Goder, V, Spiess, M (2003) Molecular mechanism of signal sequence orientation in the endoplasmic reticulum. *EMBO J.* 22:3645–3653.

[8] Higy, M, Gander, S, Spiess, M (2005) Probing the environment of signal-anchor sequences during topogenesis in the endoplasmic reticulum. *Biochemistry* 44:2039–2047.

[9] Rapoport, TA, Jungnickel, B, Kutay, U (1996) Protein transport across the eukaryotic endoplasmic reticulum and bacterial inner membranes. *Annu. Rev. Biochem.* 65:271–303.

[10] Berg, Bvd et al. (2004) X-ray structure of a protein-conducting channel. *Nature* 427:36–44.

[11] Li, W et al. (2007) The plug domain of the SecY protein stabilizes the closed state of the translocation channel and maintains a membrane seal. *Mol. Cell* 26:511–521.

[12] Zimmer, J, Nam, Y, Rapoport, TA (2008) Structure of a complex of the ATPase SecA and the protein-translocation channel. *Nature* 455:936–943.

[13] Tsukazaki, T et al. (2008) Conformational transition of Sec machinery inferred from bacterial SecYE structures. *Nature* 455:988–991.

[14] Becker, T et al. (2009) Structure of monomeric yeast and mammalian Sec61 complexes interacting with the translating ribosome. *Science* 326:1369–1373.

[15] Tam, PC, Maillard, AP, Chan, KK, Duong, F (2005) Investigating the SecY plug movement at the SecYEG translocation channel. *EMBO J.* 24:3380–3388.

[16] du Plessis, DJ, Berrelkamp, G, Nouwen, N, Driessen, AJ (2009) The lateral gate of SecYEG opens during protein translocation. *J. Biol. Chem.* 284:15805–15814.

[17] Smith, MA, Clemons, WM, DeMars, CJ, Flower, AM (2005) Modeling the effects of prl mutations on the *Escherichia coli* SecY complex. *J. Bacteriol.* 187:6454–6465.

[18] Rapoport, TA, Goder, V, Heinrich, SU, Matlack, KE (2004) Membrane-protein integration and the role of the translocation channel. *Trends Cell Biol.* 14:568–575.

[19] MacKerell, AD et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* 102:3586–3616.

[20] Darden, T, York, D, Pedersen, L (1993) Particle mesh Ewald: An N.logN method for Ewald sums in large systems. *J. Chem. Phys.* 98:10089–10092.

[21] Shih, AY, Arkhipov, A, Freddolino, PL, Schulten, K (2006) Coarse-grained protein-lipid model with application to lipoprotein particles. *J. Phys. Chem. B* 110:3674–3684.

[22] Marrink, SJ, de Vries, AH, Mark, AE (2004) Coarse-grained model for semiquantitative lipid simulations. *J. Phys. Chem. B* 108:750–760.

[23] Bond, PJ, Sansom, MS (2007) Bilayer deformation by the Kv channel voltage sensor domain revealed by self-assembly simulations. *Proc. Natl. Acad. Sci. U.S.A.* 104:2631–2636.

[24] Arkhipov, A, Yin, Y, Schulten, K (2008) Four-scale description of membrane sculpting by BAR domains. *Biophys. J.* 95:2806–2821.

[25] Kabsch, W (1978) Discussion of solution for best rotation to relate 2 sets of vectors. *Acta Crystallogr. Sect. A: Found. Crystallogr.* 34:827–828.

[26] Kumar, S, Rosenberg, JM, Bouzida, D, Swendsen, RH, Kollman, PA (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* 13:1011–1021.

[27] Gumbart, J, Schulten, K (2006) Molecular dynamics studies of the archaeal translocon. *Biophys. J.* 90:2356–2367.

[28] Tian, P, Andricioaei, I (2006) Size, motion, and function of the SecY translocon revealed by molecular dynamics simulations with virtual probes. *Biophys. J.* 90:2718–2730.

[29] Haider, S, Hall, BA, Sansom, MS (2006) Simulations of a protein translocation pore: SecY. *Biochemistry* 45:13018–13024.

[30] Gumbart, J, Schulten, K (2007) Structural determinants of lateral gate opening in the protein translocon. *Biochemistry* 46:11147–11157.

[31] Gumbart, J, Schulten, K (2008) The roles of pore ring and plug in the SecY protein-conducting channel. *J. Gen. Physiol.* 132:709–719.

[32] Plath, K, Mothes, W, Wilkinson, BM, Stirling, CJ, Rapoport, TA (1998) Signal sequence recognition in posttranslational protein transport across the yeast ER membrane. *Cell* 94:795–807.

[33] White, SH, Wimley, WC (1999) Membrane protein folding and stability: Physical principles. *Annu. Rev. Biophys. Biomol. Struct.* 28:319–365.

[34] Becker, OM, Jr, ADM, Roux, B, Watanabe, M (2001) *Computational Biochemistry and Biophysics* (CRC Press).

[35] Simon, SM, Blobel, G (1991) A protein-conducting channel in the endoplasmic reticulum. *Cell* 65:371–380.

[36] Crowley, KS, Liao, S, Worrell, VE, Reinhart, GD, Johnson, AE (1994) Secretory proteins move through the endoplasmic reticulum membrane via an aqueous, gated pore. *Cell* 78:461–471.

[37] Shaw, DE et al. (2007) Anton, a special-purpose machine for molecular dynamics simulation. *SIGARCH Comput. Archit. News* 35:1–12.

[38] Shaw, DE et al. (2010) Atomic-level characterization of the structural dynamics of proteins. *Science* 330:341–346.

[39] Bondar, AN, del Val, C, Freites, JA, Tobias, DJ, White, SH (2010) Dynamics of SecY translocons with translocation-defective mutations. *Structure* 18:847–857.

[40] Zhang, B, Miller, TF (2010) Hydrophobically stabilized open state for the lateral gate of the Sec translocon. *Proc. Natl. Acad. Sci. U.S.A.* 107:5399–5404.

[41] van der Wolk, JP, de Wit, JG, Driessen, AJ (1997) The catalytic cycle of the escherichia coli SecA ATPase comprises two distinct preprotein translocation events. *EMBO J.* 16:7297–7304.

[42] Erlandson, KJ et al. (2008) A role for the two-helix finger of the SecA ATPase in protein translocation. *Nature* 455:984–987.

[43] Liang, FC, Bageshwar, UK, Musser, SM (2009) Bacterial Sec protein transport is rate-limited by precursor length: A single turnover study. *Mol. Biol. Cell* 20:4256–4266.

[44] Zimmer, J, Rapoport, TA (2009) Conformational flexibility and peptide interaction of the translocation ATPase SecA. *J. Mol. Biol.* 394:606–612.

[45] Kusters, I, Driessen, AJ (2011) SecA, a remarkable nanomachine. *Cell. Mol. Life Sci.* 68:2053–2066.

[46] Lycklama, ANJA, Driessen, AJ (2012) The bacterial Sec-translocase: Structure and mechanism. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367:1016–1028.

[47] Or, E, Boyd, D, Gon, S, Beckwith, J, Rapoport, T (2005) The bacterial ATPase SecA functions as a monomer in protein translocation. *J. Biol. Chem.* 280:9097–9105.

[48] Jilaveanu, LB, Oliver, D (2006) SecA dimer cross-linked at its subunit interface is functional for protein translocation. *J. Bacteriol.* 188:335–338.

[49] Kusters, I et al. (2011) Quaternary structure of SecA in solution and bound to SecYEG probed at the single molecule level. *Structure* 19:430–439.

[50] Economou, A, Wickner, W (1994) SecA promotes preprotein translocation by un-

dergoing ATP-driven cycles of membrane insertion and deinsertion. *Cell* 78:835–843.

[51] Osborne, AR, Clemons, W. M., J, Rapoport, TA (2004) A large conformational change of the translocation ATPase SecA. *Proc. Natl. Acad. Sci. U.S.A.* 101:10937–10942.

[52] Erlandson, KJ, Or, E, Osborne, AR, Rapoport, TA (2008) Analysis of polypeptide movement in the SecY channel during SecA-mediated protein translocation. *J. Biol. Chem.* 283:15709–15715.

[53] Liang, FC, Bageshwar, UK, Musser, SM (2012) Position-dependent effects of polylysine on Sec protein transport. *J. Biol. Chem.* 287:12703–12714.

[54] Connolly, T, Gilmore, R (1986) Formation of a functional ribosome-membrane junction during translocation requires the participation of a GTP-binding protein. *J. Cell Biol.* 103:2253–2261.

[55] Simon, SM, Peskin, CS, Oster, GF (1992) What drives the translocation of proteins? *Proc. Natl. Acad. Sci. U.S.A.* 89:3770–3774.

[56] Rapoport, TA (2007) Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature* 450:663–669.

[57] Lycklama, ANJA, Wu, ZC, Driessen, AJ (2011) Conformational dynamics of the plug domain of the SecYEG protein-conducting channel. *J. Biol. Chem.* 286:43881–43890.

[58] Junne, T, Kocik, L, Spiess, M (2010) The hydrophobic core of the Sec61 translocon defines the hydrophobicity threshold for membrane integration. *Mol. Biol. Cell* 21:1662–1670.

[59] Hikita, C, Mizushima, S (1992) Effects of total hydrophobicity and length of the hydrophobic domain of a signal peptide on in vitro translocation efficiency. *J. Biol. Chem.* 267:4882–4888.

[60] Wahlberg, JM, Spiess, M (1997) Multiple determinants direct the orientation of signal-anchor proteins: The topogenic role of the hydrophobic signal domain. *J. Cell Biol.* 137:555–562.

[61] Harley, CA, Holt, JA, Turner, R, Tipper, DJ (1998) Transmembrane protein insertion orientation in yeast depends on the charge difference across transmembrane segments, their total hydrophobicity, and its distribution. *J. Biol. Chem.* 273:24963–24971.

[62] Shaw, AS, Rottier, PJ, Rose, JK (1988) Evidence for the loop model of signal-sequence insertion into the endoplasmic reticulum. *Proc. Natl. Acad. Sci. U.S.A.* 85:7592–7596.

[63] Goder, V, Junne, T, Spiess, M (2004) Sec61p contributes to signal sequence orientation according to the positive-inside rule. *Mol. Biol. Cell* 15:1470–1478.

[64] Dorairaj, S, Allen, TW (2007) On the thermodynamic stability of a charged arginine side chain in a transmembrane helix. *Proc. Natl. Acad. Sci. U.S.A.* 104:4943–4948.

[65] Johansson, AC, Lindahl, E (2009) Protein contents in biological membranes can explain abnormal solvation of charged and polar residues. *Proc. Natl. Acad. Sci. U.S.A.* 106:15684–15689.

[66] Rychkova, A, Vicatos, S, Warshel, A (2010) On the energetics of translocon-assisted insertion of charged transmembrane helices into membranes. *Proc. Natl. Acad. Sci. U.S.A.* 107:17598–17603.

[67] Gumbart, J, Chipot, C, Schulten, K (2011) Free-energy cost for translocon-assisted insertion of membrane proteins. *Proc. Natl. Acad. Sci. U.S.A.* 108:3596–3601.

[68] Schow, EV et al. (2010) Arginine in membranes: The connection between molecular dynamics simulations and translocon-mediated insertion experiments. *J. Membr. Biol.* 239:35–48.

[69] Ojemalm, K et al. (2011) Apolar surface area determines the efficiency of translocon-mediated membrane-protein integration into the endoplasmic reticulum. *Proc. Natl. Acad. Sci. U.S.A.* 108:E359–364.

[70] Dowhan, W, Bogdanov, M (2009) Lipid-dependent membrane protein topogenesis. *Annu. Rev. Biochem.* 78:515–540.

[71] Hessa, T, Monne, M, von Heijne, G (2003) Stop-transfer efficiency of marginally hydrophobic segments depends on the length of the carboxy-terminal tail. *EMBO Rep.* 4:178–183.

[72] Goder, V, Spiess, M (2001) Topogenesis of membrane proteins: Determinants and dynamics. *FEBS Lett.* 504:87–93.

[73] Beltzer, JP et al. (1991) Charged residues are major determinants of the transmembrane orientation of a signal-anchor sequence. *J. Biol. Chem.* 266:973–978.

[74] Parks, GD, Lamb, RA (1991) Topology of eukaryotic type II membrane proteins: importance of N-terminal positively charged residues flanking the hydrophobic domain. *Cell* 64:777–787.

[75] Heijne, GV (1986) The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J.* 5:3021–3027.

[76] Garrison, JL, Kunkel, EJ, Hegde, RS, Taunton, J (2005) A substrate-specific inhibitor of protein translocation into the endoplasmic reticulum. *Nature* 436:285–289.

[77] Maifeld, SV et al. (2011) Secretory protein profiling reveals TNF-α inactivation by selective and promiscuous Sec61 modulators. *Chem. Biol.* 18:1082–1088.

[78] Devaraneni, PK et al. (2011) Stepwise insertion and inversion of a type II signal anchor sequence in the ribosome-Sec61 translocon complex. *Cell* 146:134–147.

[79] Beckmann, R et al. (2001) Architecture of the protein-conducting channel associated with the translating 80S ribosome. *Cell* 107:361–372.

[80] Crowley, KS, Reinhart, GD, Johnson, AE (1993) The signal sequence moves through a ribosomal tunnel into a noncytoplasmic aqueous environment at the ER membrane early in translocation. *Cell* 73:1101–1115.

[81] Jungnickel, B, Rapoport, TA (1995) A posttargeting signal sequence recognition event in the endoplasmic reticulum membrane. *Cell* 82:261–270.

[82] Rutkowski, DT, Lingappa, VR, Hegde, RS (2001) Substrate-specific regulation of the ribosome-translocon junction by N-terminal signal sequences. *Proc. Natl. Acad. Sci. U.S.A.* 98:7823–7828.

[83] Denzer, AJ, Nabholz, CE, Spiess, M (1995) Transmembrane orientation of signal-anchor proteins is affected by the folding state but not the size of the N-terminal domain. *EMBO J.* 14:6311–6317.

[84] Junne, T, Schwede, T, Goder, V, Spiess, M (2007) Mutations in the Sec61p channel affecting signal sequence recognition and membrane protein topology. *J. Biol. Chem.* 282:33201–33209.

[85] Meindl-Beinker, NM, Lundin, C, Nilsson, I, White, SH, von Heijne, G (2006) Asn- and Asp-mediated interactions between transmembrane helices during translocon-mediated membrane protein assembly. *EMBO Rep.* 7:1111–1116.

[86] Sokal, RR, Rohlf, FJ (1994) *Biometry: The Principles and Practices of Statistics in Biological Research* (W. H. Freeman), 3rd edition.

[87] Sung, W, Park, PJ (1996) Polymer translocation through a pore in a membrane. *Phys. Rev. Lett.* 77:783–786.

[88] Muthukumar, M (1999) Polymer translocation through a hole. *J. Chem. Phys.* 111:10371–10374.

[89] Panja, D, Barkema, GT, Ball, RC (2007) Anomalous dynamics of unbiased polymer translocation through a narrow pore. *J. Phys.: Condens. Matter* 19:432202–432202.

[90] Chuang, J, Kantor, Y, Kardar, M (2002) Anomalous dynamics of translocation. *Phys. Rev. E* 65:011802.

[91] Huopaniemi, I, Luo, K, Ala-Nissila, T, Ying, SC (2006) Langevin dynamics simulations of polymer translocation through nanopores. *J. Chem. Phys.* 125:124901.

[92] Wei, D, Yang, W, Jin, X, Liao, Q (2007) Unforced translocation of a polymer chain through a nanopore: The solvent effect. *J. Chem. Phys.* 126:204901.

[93] Luo, K, Ala-Nissila, T, Ying, SC, Bhattacharya, A (2007) Influence of polymer-pore interactions on translocation. *Phys. Rev. Lett.* 99:148102.

[94] Luo, K, Ala-Nissila, T, Ying, SC, Bhattacharya, A (2008) Sequence dependence of DNA translocation through a nanopore. *Phys. Rev. Lett.* 100:058101.

[95] Go, N, Taketomi, H (1978) Respective roles of short- and long-range interactions in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* 75:559–563.

[96] Li, MS, Cieplak, M (1999) Folding in two-dimensional off-lattice models of proteins. *Phys. Rev. E* 59:970–976.

[97] Dill, KA et al. (1995) Principles of protein folding–a perspective from simple exact models. *Protein Sci.* 4:561–602.

[98] Chauwin, JF, Oster, G, Glick, BS (1998) Strong precursor-pore interactions constrain models for mitochondrial protein import. *Biophys. J.* 74:1732–1743.

[99] Liebermeister, W, Rapoport, TA, Heinrich, R (2001) Ratcheting in post-translational protein translocation: a mathematical model. *J. Mol. Biol.* 305:643–656.

[100] Elston, TC (2000) Models of post-translational protein translocation. *Biophys. J.* 79:2235–2251.

[101] Hizlan, D et al. (2012) Structure of the SecY complex unlocked by a preprotein mimic. *Cell Rep.* 1:21–28.

[102] Staple, DB, Payne, SH, Reddin, ALC, Kreuzer, HJ (2008) Model for stretching and unfolding the giant multidomain muscle protein using single-molecule force spectroscopy. *Phys. Rev. Lett.* 101:248301.

[103] Hanke, F, Serr, A, Kreuzer, HJ, Netz, RR (2010) Stretching single polypeptides: The effect of rotational constraints in the backbone. *EPL* 92.

[104] Kremer, K, Grest, GS (1990) Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. *J. Chem. Phys.* 92:5057–5086.

[105] Weeks, JD, Chandler, D, Andersen, HC (1971) Role of repulsive forces in determining the equilibrium structure of simple liquids. *J. Chem. Phys.* 54:5237–5247.

[106] Stoer, J, Bulirsch, R (2002) *Introduction to Numerical Analysis* (Springer, New York).

[107] Matlack, KE, Misselwitz, B, Plath, K, Rapoport, TA (1999) BiP acts as a molecular ratchet during posttranslational transport of prepro-α factor across the ER membrane. *Cell* 97:553–564.

[108] Seiser, RM, Nicchitta, CV (2000) The fate of membrane-bound ribosomes following the termination of protein synthesis. *J. Biol. Chem.* 275:33820–33827.

[109] Heritage, D, Wonderlin, WF (2001) Translocon pores in the endoplasmic reticulum are permeable to a neutral, polar molecule. *J. Biol. Chem.* 276:22655–22662.

[110] Boehlke, KW, Friesen, JD (1975) Cellular content of ribonucleic acid and protein in *Saccharomyces cerevisiae* as a function of exponential growth rate: Calculation of the apparent peptide chain elongation rate. *J. Bacteriol.* 121:429–433.

[111] Bilgin, N, Claesens, F, Pahverk, H, Ehrenberg, M (1992) Kinetic properties of *Escherichia coli* ribosomes with altered forms of S12. *J. Mol. Biol.* 224:1011–1027.

[112] Abou Elela, S, Nazar, RN (1997) Role of the 5.8S rRNA in ribosome translocation. *Nucleic Acids Res.* 25:1788–1794.

[113] Brodsky, JL, Goeckeler, J, Schekman, R (1995) BiP and Sec63p are required for both co- and posttranslational protein translocation into the yeast endoplasmic reticulum. *Proc. Natl. Acad. Sci. U.S.A.* 92:9643–9646.

[114] Blobel, G (1980) Intracellular protein topogenesis. *Proc. Natl. Acad. Sci. U.S.A.* 77:1496–1500.

[115] Sadlish, H, Pitonzo, D, Johnson, AE, Skach, WR (2005) Sequential triage of transmembrane segments by Sec61α during biogenesis of a native multispanning membrane protein. *Nat. Struct. Mol. Biol.* 12:870–878.

[116] Nilsson, I, Witt, S, Kiefer, H, Mingarro, I, von Heijne, G (2000) Distant downstream sequence determinants can control N-tail translocation during protein insertion into the endoplasmic reticulum membrane. *J. Biol. Chem.* 275:6207–6213.

[117] Bogdanov, M, Heacock, PN, Dowhan, W (2002) A polytopic membrane protein displays a reversible topology dependent on membrane lipid composition. *EMBO J.* 21:2107–2116.

[118] Skach, WR (2009) Cellular mechanisms of membrane protein folding. *Nat. Struct. Mol. Biol.* 16:606–612.

[119] Elofsson, A, von Heijne, G (2007) Membrane protein structure: Prediction versus reality. *Annu. Rev. Biochem.* 76:125–140.

[120] Wimley, WC, Creamer, TP, White, SH (1996) Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry* 35:5109–5124.

[121] Phillips, JC et al. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26:1781–1802.

[122] Martyna, GJ, Tobias, DJ, Klein, ML (1994) Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* 101:4177–4189.

[123] Feller, SE, Zhang, Y, Pastor, RW, Brooks, BR (1995) Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.* 103:4613–4621.

[124] Tuckerman, ME, Berne, BJ, Martyna, GJ (1991) Molecular dynamics algorithm for multiple time scales: Systems with long range forces. *J. Chem. Phys.* 94:6811–6815.

[125] Schrödinger, LLC (2010) The PyMOL Molecular Graphics System, Version 1.3r1.

[126] Marrink, SJ, Risselada, HJ, Yefimov, S, Tieleman, DP, de Vries, AH (2007) The MARTINI force field: Coarse-grained model for biomolecular simulations. *J. Phys. Chem. B* 111:7812–7824.

[127] Bond, PJ, Wee, CL, Sansom, MS (2008) Coarse-grained molecular dynamics simulations of the energetics of helix insertion into a lipid bilayer. *Biochemistry* 47:11321–11331.

[128] Chipot, C, Pohorille, A (2007) *Free Energy Calculations: Theory and Applications in Chemistry and Biology* (Springer).

[129] Radzicka, A, Wolfenden, R (1988) Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* 27:1664–1670.

[130] Shirts, MR, Pande, VS (2005) Solvation free energies of amino acid side

chain analogs for common molecular mechanics water models. *J. Chem. Phys.* 122:134508.

[131] Shirts, MR, Pitera, JW, Swope, WC, Pande, VS (2003) Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J. Chem. Phys.* 119:5740–5761.

[132] Frenkel, D, Smit, B (2001) *Understanding Molecular Simulation, Second Edition: From Algorithms to Applications* (Academic Press), 2nd edition.

[133] Eswar, N et al. (2006) Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* Chapter 5:Unit 5.6.

[134] Sali, A, Blundell, TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234:779–815.

[135] Fiser, A, Sali, A (2003) ModLoop: Automated modeling of loops in protein structures. *Bioinformatics* 19:2500–2501.

[136] Fiser, A, Do, RK, Sali, A (2000) Modeling of loops in protein structures. *Protein Sci.* 9:1753–1773.

[137] Lomize, MA, Lomize, AL, Pogozheva, ID, Mosberg, HI (2006) OPM: Orientations of proteins in membranes database. *Bioinformatics* 22:623–625.

[138] Lomize, AL, Pogozheva, ID, Lomize, MA, Mosberg, HI (2006) Positioning of proteins in membranes: A computational approach. *Protein Sci.* 15:1318–1333.

[139] Hess, B, Kutzner, C, van der Spoel, D, Lindahl, E (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* 4:435–447.

[140] Klauda, JB et al. (2010) Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J. Phys. Chem. B* 114:7830–7843.

[141] Nose, S (1984) A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* 81:511–519.

[142] Hoover, WG (1985) Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* 31:1695.

[143] Parrinello, M, Rahman, A (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* 52:7182–7190.

[144] Shan, Y, Klepeis, JL, Eastwood, MP, Dror, RO, Shaw, DE (2005) Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. *J. Chem. Phys.* 122:54101–54113.

[145] Krautler, V, Van Gunsteren, WF, Hunenberger, PH (2001) A fast SHAKE: Algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* 22:501–508.

[146] Tuckerman, M, Berne, BJ, Martyna, GJ (1992) Reversible multiple time scale molecular dynamics. *J. Chem. Phys.* 97:1990–2001.

[147] Berendsen, HJC, Postma, JPM, Vangunsteren, WF, Dinola, A, Haak, JR (1984) Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.

[148] Loriot, S, Cazals, F (2010) Modeling macro-molecular interfaces with Intervor. *Bioinformatics* 26:964–965.

[149] Humphrey, W, Dalke, A, Schulten, K (1996) VMD: Visual molecular dynamics. *J Mol Graph* 14:33–38, 27–38.

[150] Shen, MY, Sali, A (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 15:2507–2524.

[151] Gumbart, J, Trabuco, LG, Schreiner, E, Villa, E, Schulten, K (2009) Regulation of the protein-conducting channel by a bound ribosome. *Structure* 17:1453–1464.

[152] Sanders, SL, Whitfield, KM, Vogel, JP, Rose, MD, Schekman, RW (1992) Sec61p and BiP directly facilitate polypeptide translocation into the ER. *Cell* 69:353–365.

[153] Frauenfeld, J et al. (2011) Cryo-EM structure of the ribosome-SecYE complex in the membrane environment. *Nat. Struct. Mol. Biol.* 18:614–621.

[154] Cheng, Z, Gilmore, R (2006) Slow translocon gating causes cytosolic exposure of transmembrane and lumenal domains during membrane protein integration. *Nat. Struct. Mol. Biol.* 13:930–936.

[155] Chandler, D (1986) Roles of classical dynamics and quantum dynamics on activated processes occurring in liquids. *J. Stat. Phys.* 42:49–67.

[156] Hummer, G (2004) From transition paths to transition states and rate coefficients. *J. Chem. Phys.* 120:516–523.

[157] Chandler, D (2005) Interfaces and the driving force of hydrophobic assembly. *Nature* 437:640–647.

[158] Ramadurai, S et al. (2009) Lateral diffusion of membrane proteins. *J. Am. Chem. Soc.* 131:12650–12656.

[159] Doi, M, Edwards, SF (1988) *The Theory of Polymer Dynamics* (Oxford University Press, USA).

[160] Sriraman, S, Kevrekidis, LG, Hummer, G (2005) Coarse master equation from Bayesian analysis of replica molecular dynamics simulations. *J. Phys. Chem. B* 109:6479–6484.

[161] Buchete, NV, Hummer, G (2008) Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B* 112:6057–6069.

[162] Tyedmers, J, Lerner, M, Wiedmann, M, Volkmer, J, Zimmermann, R (2003) Polypeptide-binding proteins mediate completion of co-translational protein translocation into the mammalian endoplasmic reticulum. *EMBO Rep.* 4:505–510.

[163] Kim, SJ, Mitra, D, Salerno, JR, Hegde, RS (2002) Signal sequences control gating of the protein translocation channel in a substrate-specific manner. *Dev. Cell* 2:207–217.

[164] Seppala, S, Slusky, JS, Lloris-Garcera, P, Rapp, M, von Heijne, G (2010) Control of membrane protein topology by a single C-terminal residue. *Science* 328:1698–1700.

[165] von Heijne, G (1989) Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature* 341:456–458.

[166] Andersson, H, Bakker, E, von Heijne, G (1992) Different positively charged amino acids have similar effects on the topology of a polytopic transmembrane protein in *Escherichia coli*. *J. Biol. Chem.* 267:1491–1495.

[167] Kida, Y, Morimoto, F, Mihara, K, Sakaguchi, M (2006) Function of positive charges following signal-anchor sequences during translocation of the N-terminal domain. *J. Biol. Chem.* 281:1152–1158.

[168] Nilsson, I, von Heijne, G (1990) Fine-tuning the topology of a polytopic membrane protein: Role of positively and negatively charged amino acids. *Cell* 62:1135–1141.

[169] Bogdanov, M, Xie, J, Heacock, P, Dowhan, W (2008) To flip or not to flip: lipid-protein charge interactions are a determinant of final membrane protein topology. *J. Cell Biol.* 182:925–935.

[170] Hedin, LE et al. (2010) Membrane insertion of marginally hydrophobic transmembrane helices depends on sequence context. *J. Mol. Biol.* 396:221–229.

[171] Lerch-Bader, M, Lundin, C, Kim, H, Nilsson, I, von Heijne, G (2008) Contribution of positively charged flanking residues to the insertion of transmembrane helices into the endoplasmic reticulum. *Proc. Natl. Acad. Sci. U.S.A.* 105:4127–4132.

[172] Fujita, H, Kida, Y, Hagiwara, M, Morimoto, F, Sakaguchi, M (2010) Positive charges of translocating polypeptide chain retrieve an upstream marginal hydrophobic segment from the endoplasmic reticulum lumen to the translocon. *Mol. Biol. Cell* 21:2045–2056.

[173] Polyanin, AD, Zaitsev, VF (1995) *Handbook of Exact Solutions for Ordinary Differential Equations* (CRC-Press), 1st edition.

[174] Kyte, J, Doolittle, RF (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105–132.