Coarse-grained simulation approaches for protein integration and translocation via the Sec translocon

Thesis by
Michiel Jacobus Maria Niesen

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2018
Defended March 23, 2018

Michiel Jacobus Maria Niesen
ORCID: 0000-0002-9255-6203

ACKNOWLEDGEMENTS

The work described in this thesis would not have been possible without the support of many people. I would like to thank everyone who has contributed in any way, both during my time at Caltech and on my journey towards that point.

At Caltech, I have had the great pleasure of working with Professor Thomas Miller. Tom has helped me to grow as a scientist by being a great example, and by giving me many opportunities to challenge and improve myself. I want to thank Tom and my thesis committee, Professor William Clemons, Professor Shu-ou Shan, and Professor William Goddard, for mentoring me during my PhD. I also want to thank my past mentors that have prepared me for my PhD: Professor Nagarajan Vaidehi, Professor John Shively, Professor Peter Hilbers, and Professor Natal van Riel.

Much of the research described in this thesis was done in close collaboration with other scientists. I want to especially thank Connie Wang and Reid Van Lehn for their help in developing the simulation method described in Chapter 2, and William Clemons and Stephen Marshall for their help in performing the experiments described in Chapters 4 and 5. I would also like to thank Matthew Zimmer, Shyam Saladi, and Bin Zhang for the excellent research meetings and discussions.

I have enjoyed living in California due to my friends here. I want to thank all members of the Miller group, past and present, who have made my time in the group enjoyable: Priscilla, Feizhi, Matt W, Ioan, Philip, Leanne, Sebastian, Jorge, Dan, Matt Z, Xuecheng, Jeongmin, Sherry, Brooke, Ralph, Mike, Connie, Brett, Reid, Josh, Taylor, Jason, Artur, Fran, Pengfei, Kuba, Nick, Lila, Joonho, Mark, Eric, and Bin. I would also like to thank my friends outside of the lab, too many to list, who have made California my new home.

Finally, I am truly grateful to my family for all their support and patience. Thank you Noriko, Xander, Sophie, and Choboroku for everything. Thank you Theo, Ans, Kuniharu, Tomiko, and the rest of my extended family and friends abroad for the visits and vacations during my time at Caltech.

ABSTRACT

This thesis describes coarse-grained approaches for simulating the co-translational integration and translocation of proteins via the Sec translocon, which is a key step in the biogenesis of membrane and secretory proteins. We present a coarse-grained simulation approach that is capable of simulating minute-timescale dynamics while retaining sufficient chemical and structural detail to capture sequence-specific interactions. The model is validated through comparison to existing experimental data and applied to characterize the forces that act on nascent proteins and drive successful integration and translocation. We also apply coarse-grained simulations of the integration of multi-spanning membrane proteins to understand the effect of sequence modifications on expression levels. We uncover the link between integration efficiency and observed expression levels for membrane proteins, and utilize coarse-grained simulations to predict sequence modifications that improve heterologous overexpression.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] Stephen S. Marshall, Michiel J. M. Niesen, A. Müller, K. Tiemann, S. M. Saladi, R. P. Galimidi, B. Zhang, W. Clemons, and Miller. T. F. A link between integral membrane protein expression and simulated integration efficiency. *Cell Reports*, 16(8):2169–2177, 2016. doi: http://dx.doi.org/10.1016/j.celrep.2016.07.042.
M.J.M.N designed research, performed simulations, contributed new reagents/analytic tools, analyzed data, and wrote the manuscript.

[2] Michiel J. M. Niesen, Stephen S. Marshall, Thomas F. Miller, and William M. Clemons. Improving membrane protein expression by optimizing integration efficiency. *Journal of Biological Chemistry*, 292(47):19537–19545, 2017. doi: 10.1074/jbc.M117.813469. URL `http://www.jbc.org/content/292/47/19537.abstract`.
M. J. M. N. conceived and designed the study, analyzed the data, executed and analyzed the CG simulations, and wrote the manuscript.

[3] Michiel J. M. Niesen, Connie Y. Wang, Reid C. Van Lehn, and Thomas F. Miller, III. Structurally detailed coarse-grained model for sec-facilitated co-translational protein translocation and membrane integration. *PLOS Computational Biology*, 13(3):1–26, 03 2017. doi: 10.1371/journal.pcbi.1005427. URL `https://doi.org/10.1371/journal.pcbi.1005427`.
M.J.M.N designed research, performed research, developed new simulation methods/analytic tools, analyzed data, and wrote the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

NOMENCLATURE

$\Delta G_{\mathbf{app}}$**.** Apparent free energy of membrane integration for an amino-acid sequence.

$f_{\mathbf{FL}}$**.** Fraction of full-length protein, used to determine the degree of stall-release in arrest peptide experiments.

2D-CG. Coarse-grained model for the simulation of Sec-facilitated protein integration and translocation. Described in the method section of Chapter 4 of this thesis.

3D-CG. Structurally detailed coarse-grained model for the simulation of Sec-facilitated protein integration and translocation. Described in Chapter 2 of this thesis..

***Aa*-tail.** The C-terminal loop of *Aa*TatC, defined as in Figure 4.3E.

AP. Arrest Peptide; peptide that will cause ribosomal translation to stall, the stall is released with a force-dependent rate.

AUC. Area Under the Curve.

C-tail. C-terminal loop of a TatC homolog, sequences defined as in Figure 4.3E.

CI. Confidence Interval.

Co-translational. Occurring during the process of ribosomal translation.

CV. Collective Variable; a variable that is defined to describe a collective motion in the molecular dynamic simulations described in this thesis.

GFP. Green Fluorescent Protein.

IMP. Integral Membrane Protein.

Integration. insertion of a polypeptide domain into the lipid membrane.

LG. Lateral Gate; region of the Sec translocon that can open to create an interface between the channel interior and the lipid membrane.

MARTINI force-field. Residue-based coarse-grained forcefield used for the simulation of biological systems.

PMF. Potential of Mean Force; free-energy surface along a chosen reaction coordinate (CV).

ROC. Receiver Operating Characteristic.

SecY/Sec61. Central component of the Sec translocon; a membrane spanning channel that can open laterally to facilitate protein integration and translocation.

SEM. Standard error of the mean.

TatC. Twin-Arginine transporter component C; a transporter of hydrophilic domains across the cell-membrane in bacteria.

Translocation. Passage across the cell membrane.

Chapter 1

INTRODUCTION

Co-translational integration and translocation via the Sec translocon

Most integral membrane proteins (IMPs) are co-translationally inserted into the membrane during biosynthesis via the Sec translocon, a multiprotein complex [19, 29, 98, 116]. In this process, a ribosome docks to the cytosolic opening of the Sec translocon and feeds a nascent polypeptide chain (NC) into the translocon channel (Figure 1.1). Secretory proteins and the soluble domains of IMPs translocate across the lipid membrane by passing through the translocon channel [29, 98]. Alternatively, the transmembrane domains (TMDs) of IMPs integrate directly into the lipid membrane via the translocon lateral gate (LG). Integration is facilitated by a conformational change in the channel that separates the two LG helices to create an opening between the channel interior and the hydrophobic core of the membrane [9, 32, 55]. The likelihood of integration or translocation of polypeptide segments depends on residue-specific chemical features of the nascent polypeptide chain, such as its hydrophobicity and charge [37, 68, 91, 93, 131], but is also governed by the dynamics of protein synthesis on the minute timescale [42, 70].

Experimental studies have elucidated many aspects of the structure and function of the Sec translocon, although their ability to directly probe the non-equilibrium co-translational integration process is limited. Structural characterization has revealed many of the components of the translocon complex in both eukaryotes [7, 11, 43, 94, 133] and prokaryotes [9, 32, 36, 65, 75, 122], while biophysical assays have investigated the functional effects of NC hydrophobicity [68, 91], charges flanking TMDs [37, 93, 131], soluble loop length [42, 70], and the forces exerted on a NC during translation [19, 26, 62]. Despite these findings, mechanistic details of the co-translational integration process remain in question [19] because most experiments are limited to probing final protein distributions – such as the fraction of protein in a specific topology [42] or the fraction of protein integrated in the membrane [57] – and do not typically resolve NC dynamics.

Computational models of co-translational integration and translocation

Atomistic-scale molecular dynamics simulations can be used to probe detailed aspects of co-translational integration, with recent simulations providing insight

into the energetics of TMD integration [49, 51], the dynamics of water inside the translocon [16], the effect of NC properties on LG opening [145], the dynamics of a NC during the initial stages of translation [48, 126, 147], and the dynamics of IMP integration in simplified system representations [101, 146]. However, the separation of timescales relevant to co-translational integration poses a significant challenge to conventional simulation methods: notably, ribosomal translation requires seconds to minutes to complete the biosynthesis of typical polypeptides [12, 60, 86, 143], while conformational fluctuations of the NC occur on the nanosecond timescale. Currently available simulation approaches either fail to reach the biological timescales of ribosomal translation [16, 48, 147] or lack sufficient detail to describe detailed features of the NC-translocon interactions and NC conformational dynamics [101, 102, 146]. The model presented here overcomes these limitations, allowing direct comparison with a broad range of available experiments.

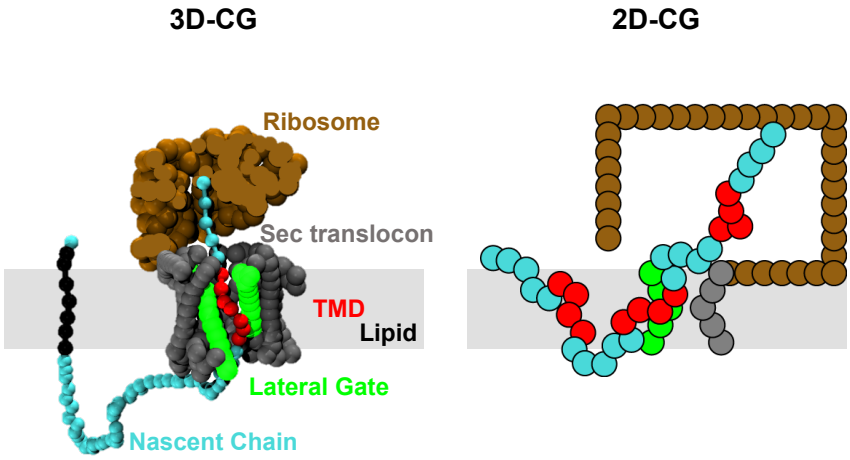

Figure 1.1: The Sec translocon in the coarse-grained representations used in this thesis. On the left, an example configuration in the 3D-CG model. On the right, an example configuration in the 2D-CG model. The same color labels are used in both panels; the ribosome (brown) sits at the cytosolic side of the Sec translocon (grey), the nascent chain (cyan) is translated and inserts into the translocon, TMDs (red) partition into the lipid bilayer (shaded) via the LG (green).

In previous work, a highly coarse-grained (CG) model of Sec-facilitated IMP integration was developed in which all system coordinates are projected onto a two-dimensional plane passing through the translocon LG [146]. This 2D-CG model (Figure 1.1), which is applied and described in Chapter 4 and 5 of this thesis, includes an explicit representation of NC translation, translocon LG conformational gating, and a sufficiently simple system description to enable minute-timescale unbi-

ased trajectories. Previous work has demonstrated that the 2D-CG model correctly predicts the distribution of topologies obtained by TMDs as a function of C-terminal soluble loop length [146], the probability of membrane integration as a function of TMD hydrophobicity [146], the effect of charge mutations on the topology of the dual-topology protein EmrE [129], and the effect of sequence modifications on the integration efficiency of the multispanning protein TatC [84]. The 2D-CG model was also used to demonstrate a link between IMP integration efficiency and expression levels for TatC [84], enabling the computational prediction of amino-acid sequence modifications that improve IMP expression. These successes illustrate the potential for using CG methods to capture the essential physics of the co-translational protein translocation and membrane integration processes. However, several shortcomings of the 2D-CG model have been identified. In particular, the ribosome and translocon are modeled without detailed structural features, sequence-specific ribosome and translocon chemical features are not mapped directly to the CG representation, and interactions between the NC and the translocon are independent of NC sequence. These shortcomings limit the ability of the 2D-CG model to investigate phenomena arising from sequence-specific structural and chemical features, such as variations among homologs of the Sec translocon [9, 32] or interactions between the NC and translocon [56, 61].

Chapter 2 of this thesis presents a new method for the simulation of co-translational integration and translocation via the Sec translocon. The new method (3D-CG, Figure 1.1) retains structural and chemical detail, and is demonstrated to more accurately capture the effect of amino-acid level changes on the integration and translocation process. Chapter 3 discusses the application of the 3D-CG model to characterize the forces that act on translating NC. Using 3D-CG simulations in combination with arrest-peptide (AP) experiments, performed by our collaborators in the group of Professor Gunnar von Heijne, we disentangle force contributions caused by NC-translocon and NC-ribosome interactions, membrane partioning, and coupling to the transmembrane electrostatic potential.

The membrane protein expression problem

IMPs play crucial roles in the transport of molecules, energy, and information across the membrane and are an important focus of structural and biophysical studies. However, the production of sufficient levels of IMPs is a limiting factor in their characterization [74]. Even among homologous IMP sequences, expression levels can vary widely [45, 71, 74, 77, 78, 84], and the mechanistic basis for this variability

is often unclear. Extensive efforts have been committed to identify IMP sequences, expression conditions, and host modifications that yield IMP expression at sufficient levels for further study [100, 109, 113, 134]. Despite these efforts, general guidelines for successful overexpression for IMPs are lacking.

To reach a stable folded structure, IMPs must integrate into the membrane with the correct topology (i.e., orientation of each TMD with respect to the membrane), which depends sensitively on the properties of both the NC and the translocon[116, 132]. Even single mutations to an IMP amino-acid sequence can disrupt integration and induce disease phenotypes [108] or decrease protein expression [33, 84, 142]; similarly, mutations to the translocon channel can inhibit IMP folding[25, 67, 68, 120, 127]. The important role for IMPs in cellular functions motivates the understanding of the effect of NC and translocon properties on the efficiency of co-translational integration. However, a detailed understanding of this process presents challenges for both theory and experiment due to the long range of timescales (from nanoseconds to minutes) that are involved.

Chapter 4 and 5 of this thesis describe how 2D-CG simulations of the integration of a multispanning membrane protein, TatC, can be used to predict sequence modifications that enable heterologous overexpression in *E. coli*. Chapter 4 demonstrates that there is a link between the efficiency with which a protein integrates in the 2D-CG simulations and the level at which it expresses in *E. coli*, as quantified by our experimental collaborators Professor William Clemons and Stephen Marshall. In Chapter 5 we show how this link between simulated integration efficiency and expression can be utilized to predict sequence modifications that improve expression for a large set of diverse sequence modifications.

Chapter 2

STRUCTURALLY DETAILED COARSE-GRAINED MODEL FOR SEC-FACILITATED CO-TRANSLATIONAL PROTEIN INTEGRATION AND TRANSLOCATION

Adapted from:

Niesen, M. J. M.*, Wang, C. Y.*, Van Lehn, R. C., Miller, T. F. M. (2017). "Structurally detailed coarse-grained model for Sec-facilitated co-translational protein translocation and membrane integration". In: *PLoS Comp. Biol.* 13(3): e1005427. DOI: 10.1371/journal.pcbi.1005427. (*) Equal contribution.

This chapter describes the development and validation of a coarse-grained simulation model that is capable of simulating the minute-timescale dynamics of protein translocation and membrane integration via the Sec translocon, while retaining sufficient chemical and structural detail to capture many of the sequence-specific interactions that drive these processes. The model includes accurate geometric representations of the ribosome and Sec translocon, obtained directly from experimental structures, and interactions parameterized from nearly 200 μs of residue-based coarse-grained molecular dynamics simulations. A protocol for mapping amino-acid sequences to coarse-grained beads enables the direct simulation of trajectories for the co-translational insertion of arbitrary polypeptide sequences into the Sec translocon. The model reproduces experimentally observed features of membrane protein integration, including the efficiency with which polypeptide domains integrate into the membrane, the variation in integration efficiency upon single amino-acid mutations, and the orientation of transmembrane domains. The central advantage of the model is that it connects sequence-level protein features to biological observables and timescales, enabling direct simulation for the mechanistic analysis of co-translational integration and for the engineering of membrane proteins with enhanced membrane integration efficiency.

2.1 Introduction

In this chapter, we describe a refined CG model that enables simulation of the long time- and length-scales that are relevant to co-translational protein integration, while preserving sequence-specific properties of the NC and translocon and capturing the structure of the ribosome-translocon complex. The new 3D-CG model extends the previous 2D-CG model[146] by providing a realistic three-dimensional representation of the ribosome-translocon complex mapped directly from high-resolution structural data [9, 133]. Additionally, the model is parameterized via a bottom-up approach to reproduce sequence-specific NC-translocon interactions, and it includes a protocol for directly mapping any input amino-acid sequence to a simulation representation, enabling simulation of any polypeptide using only the amino-acid sequence as input. The improved 3D-CG model is validated by reproducing experimental measurements of TMD integration efficiency [56] and signal peptide topogenesis [42]. The model further reproduces the "biological hydrophobicity" scale derived by von Heijne and co-workers [56], capturing the effects of single-residue mutations on stop-transfer efficiency. The strong agreement between simulation and experiment indicates that the 3D-CG model produces simulation predictions that can be confirmed by direct experimental analogues. The new model provides a framework for performing mutagenesis studies of the NC and ribosome-translocon complex to obtain a detailed mechanistic understanding of interactions that impact TMD integration and topogenesis, potentially enabling the prediction of IMP sequence modifications with enhanced membrane integration efficiency and stability.

2.2 Description of the coarse-grained simulation approach

We now present the details of the 3D-CG model of Sec-facilitated co-translational protein synthesis. The 3D-CG model preserves several features of the prior 2D-CG model [146], including (*i*) representation of the NC as a non-overlapping freely-jointed chain, (*ii*) 3:1 mapping of amino-acid residues to CG beads, (*iii*) implicit representation of the lipid membrane, (*iv*) stochastic opening and closing of the translocon LG, (*v*) explicit modeling of NC translation during the simulation trajectories, and (*vi*) sufficient computational efficiency to reach long second-minute timescales, achieved using a high level of coarse graining and the use of a partially tabulated potential energy function.

Significant improvements of the 3D-CG model described below include a three-dimensional representation of the ribosome/translocon/NC geometry (shown in

Fig 2.1 and residue-specific interactions between the NC and the translocon. The resulting 3D-CG model allows any input amino-acid sequence to be directly converted to a CG simulation representation. The 3D-CG model then simulates the dynamics of the nascent protein, including elongation of the polypeptide during ribosomal translation, the integration of protein segments into the membrane bilayer, and the retention or translocation of protein segments flanking transmembrane domains (shown in Fig 2.1D).

3D-CG model geometry

Fig 2.1A presents the components of the 3D-CG model compared to an image of the ribosome-translocon complex obtained from a cryo-EM structure [133]. The SecYEG translocon (grey/green), ribosome (brown), and the NC (cyan/red) are represented with explicit CG beads, while the implicit membrane is drawn as a shaded region. As in the 2D-CG model [146], each CG bead has a diameter of σ = 0.8 nm, the Kuhn length of a polypeptide chain [101, 146], and represents three amino-acid residues; σ sets the length scale for the 3D-CG model. The coordinate system is defined such that the origin is placed at the geometric center of the translocon channel Cα atoms, the implicit membrane spans the x-y plane with its midplane located at $z = 0\sigma$, and the axis of the translocon is aligned with the z-axis (Fig 2.1C).

The geometry of the Sec translocon is obtained by mapping all amino-acid residues of the translocon onto CG beads in a ratio of three amino acids to one CG bead, where the CG bead is positioned at the center of mass of the Cα atoms for each consecutive triplet of amino-acid residues in the translocon primary sequence. Triplets of amino acids with a net positive charge are assigned a +1 charge, and triplets of amino acids with a net negative charge are assigned a -1 charge. To determine the net charge of a triplet of amino acids the charges of the amino acids are summed, with arginine and lysine counted as +1, and aspartate and glutamate counted as -1 (see the section *Translocon CG bead charges* and Figure 2.12 for further discussion). The translocon is modeled in two distinct conformations, with the LG either closed or open (Figure 2.1B). CG bead coordinates for both conformations are obtained from residue-based coarse-grained simulations of the *Methanocaldococcus jannaschii* SecYEG translocon (PDB ID: 1RHZ) [9] (see the section *Translocon CG bead coordinates*). The 3D-CG model of the translocon is oriented such that the y-axis of the simulation coordinate system passes between the helices of the LG when the translocon is in the open conformation (Fig 2.1C).

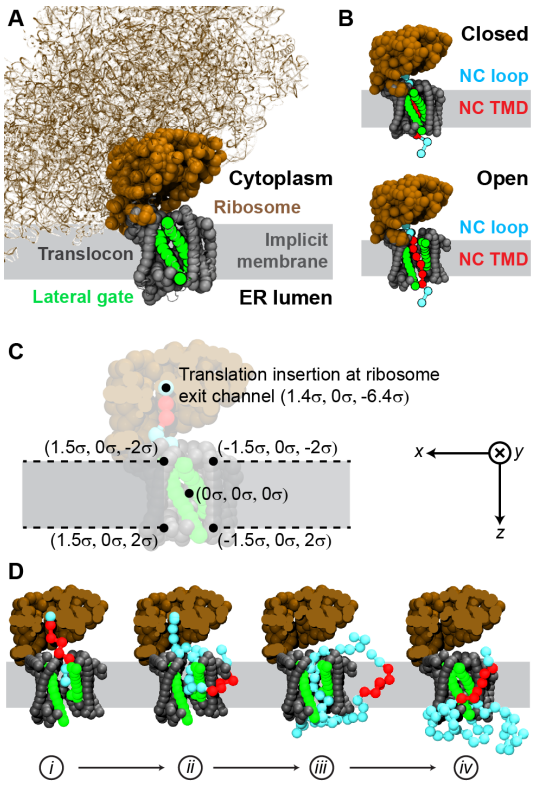

Figure 2.1: 3D-CG model geometry. (**A**) Components of the 3D-CG model overlaid on a high-resolution cryo-EM structure of the ribosome-translocon complex [133]. 3D-CG model beads are represented by opaque spheres and are labeled according to their color. The region representing the implicit membrane is drawn as a grey background. (**B**) 3D-CG model snapshots of the two possible translocon conformations, with a closed lateral gate (top) and with an open lateral gate (bottom). In each case, a NC is shown emerging from the ribosome exit channel and interacting with the translocon. (**C**) Coordinate system for the 3D-CG model. Coordinates for the translation insertion point at the ribosome exit channel, the origin, and four points illustrating the bounds of the implicit membrane are indicated. (**D**) Simulation snapshots showing representative states during a simulation trajectory, including: (*i*) the start of translation, (*ii*) topological inversion of a TMD during integration, (*iii*) release of the C-terminus at the end of translation, and (*iv*) the end of a simulation in which the TMD has integrated into the membrane, the lateral gate is closed, and all polypeptide segments have exited the channel.

The geometry of the ribosome is obtained by mapping the ribosome-translocon complex from a recent high-resolution cryo-EM structure (PDB ID: 3J7Q) onto CG beads [133]. Amino-acid residues are mapped onto CG beads in a 3:1 ratio following the same procedure used for the translocon. Each RNA nucleotide in the ribosome is mapped onto two CG beads; one bead represents the sugar-phosphate

backbone, while the other bead represents the nucleobase. This mapping is used to capture the excluded volume and the rigidity of the RNA scaffold and is consistent with previous work on coarse-grained DNA/RNA simulations [30, 81, 104]. Each CG bead representing a RNA sugar-phosphate backbone in the ribosome is assigned a -1 charge and each CG bead representing a nucleobase is neutral. Only the portion of the ribosome near the translocon channel is explicitly represented as CG beads in the final simulation system (Figure 2.1A; additional details are in the section *Ribosome CG bead coordinates* and Figure 2.11). Ribosome CG bead positions are identical for both translocon conformations.

To characterize whether the ith NC bead, with position $\mathbf{x}_i = (x_i, y_i, z_i)$, is located in the implicit membrane region, we define the characteristic function

$$S_{\text{mem}}(\mathbf{x}_i) = [1 - S(x_i, y_i)]S(z_i), \tag{2.1}$$

which assumes a value of 1 in the membrane and 0 elsewhere. $S(x, y)$ and $S(z)$ are smooth switching functions,

$$S(x, y) = \frac{1}{4}\left[1 + \tanh\left(\frac{\sqrt{x^2 + y^2} + 1.5\sigma}{0.25\sigma}\right)\right]\left[1 - \tanh\left(\frac{\sqrt{x^2 + y^2} - 1.5\sigma}{0.25\sigma}\right)\right], \tag{2.2}$$

and

$$S(z) = \frac{1}{4}\left[1 + \tanh\left(\frac{z + 2\sigma}{0.25\sigma}\right)\right]\left[1 - \tanh\left(\frac{z - 2\sigma}{0.25\sigma}\right)\right], \tag{2.3}$$

where $\sqrt{x^2 + y^2}$ is the radial distance from the coordinate system origin in the x-y plane. $S(x, y)$ is approximately 1 for the range -1.5σ < $\sqrt{x^2 + y^2}$ < 1.5σ and 0 elsewhere, while $S(z)$ is approximately 1 for the range -2σ < z < 2σ and 0 elsewhere (Fig 2.1C). Eq 2.1-2.3 are used in Eq 2.8 to define the solvation of a NC bead.

3D-CG model potential energy function

The potential energy function for the 3D-CG model is expressed

$$\begin{aligned} U(\mathbf{x}_{\text{n}}, \mathbf{x}_{\text{c}}; q, g) = U_{\text{bond}}(\mathbf{x}_{\text{n}}) + U_{\text{excl}}(\mathbf{x}_{\text{n}}) + U_{\text{elec}}(\mathbf{x}_{\text{n}}, \mathbf{x}_{\text{c}}; q) + U_{\text{solv}}(\mathbf{x}_{\text{n}}; g) \\ + U_{\text{chan}}(\mathbf{x}_{\text{n}}, \mathbf{x}_{\text{c}}; g) + U_{\text{ribo}}(\mathbf{x}_{\text{n}}), \end{aligned} \tag{2.4}$$

where $\mathbf{x}_{\text{n}}$ indicates the set of NC bead positions, $\mathbf{x}_{\text{c}}$ indicates the set of channel and ribosome bead positions, q is the set of all bead charges, and g is the set of all NC bead transfer free energies. All interactions in the 3D-CG model are defined using

an energy scale given by $\epsilon = k_B T$, where the temperature, T, is fixed at 310 K to represent physiological conditions.

Bonded interactions between consecutive NC beads are described using the finite extension nonlinear elastic (FENE) potential,

$$U_{\text{bond}}(\mathbf{x}_{\text{n}}) = -\frac{1}{2} K_0 R_0^2 \sum_{b \in \text{Bonds}} \ln\left(1 - \frac{r_b^2}{R_0^2}\right), \tag{2.5}$$

where the sum runs over all bonds in the NC, r_b is the distance between the NC beads that share bond b, $K_0 = 5.833\ \epsilon/\sigma^2$, and $R_0 = 2\sigma$. Short-ranged excluded volume interactions between pairs of NC beads are modeled using a purely repulsive Lennard-Jones (LJ) potential [138],

$$U_{\text{excl}}(\mathbf{x}_{\text{n}}) = \sum_{i,j \in \text{NC}} \left\{ \begin{array}{cc} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \epsilon_{ij} & , \quad r_{ij} < 2^{1/6}\sigma_{ij} \\ 0 & , \quad r_{ij} \geq 2^{1/6}\sigma_{ij} \end{array} \right\}, \tag{2.6}$$

where the sum runs over all pairs of NC beads, r_{ij} is the distance between NC beads i and j, and $\epsilon_{ij} = \epsilon$, and $\sigma_{ij} = \sigma$.

Electrostatic interactions are described using the Debye-Hückel potential,

$$U_{\text{elec}}(\mathbf{x}_{\text{n}}, \mathbf{x}_{\text{c}}; q) = \sum_{i,j \in \text{All}} \frac{l_B q_i q_j \epsilon}{r_{ij}} \exp\left(-\frac{r_{ij}}{\kappa}\right), \tag{2.7}$$

where the sum runs over all pairs of charged beads, l_B is the Bjerrum length, q_i is the charge of CG bead i in the NC, translocon, or ribosome, and κ is the Debye length. Assuming that electrostatic interactions are screened by physiological salt concentrations [5, 118], the electrostatic length scales are approximated by $\kappa = l_B = \sigma$.

NC bead interactions with the implicit solvent are described using a position-dependent potential,

$$U_{\text{solv}}(\mathbf{x}_{\text{n}}; g) = \sum_{i \in \text{NC}} g_i S_{\text{mem}}(\mathbf{x}_i), \tag{2.8}$$

where $\mathbf{x}_i$ is the position of NC bead i, and g_i is the transfer free energy for partitioning NC bead i from water to the membrane.

Residue-specific interactions between NC beads and translocon beads are given by

$$U_{\text{chan}}(\mathbf{x}_{\text{n}}, \mathbf{x}_{\text{c}}; g) = \sum_{i \in \text{NC}} [1 - S_{\text{mem}}(\mathbf{x}_i)] U_{\text{chan}}^{\text{aq}}(\mathbf{x}_i, \mathbf{x}_{\text{c}}; g_i) + [S_{\text{mem}}(\mathbf{x}_i)] U_{\text{chan}}^{\text{mem}}(\mathbf{x}_i, \mathbf{x}_{\text{c}}; g_i). \tag{2.9}$$

Eq 2.9 smoothly interpolates between NC bead-translocon interactions for which NC bead i is positioned in aqueous solution inside the channel ($U^{\text{aq}}_{\text{chan}}(\mathbf{x}_i, \mathbf{x}_c; g_i)$) or positioned in the membrane near the channel exterior ($U^{\text{mem}}_{\text{chan}}(\mathbf{x}_i, \mathbf{x}_c; g_i)$). The exact functional forms of $U^{\text{aq}}_{\text{chan}}(\mathbf{x}_i, \mathbf{x}_c; g_i)$ and $U^{\text{mem}}_{\text{chan}}(\mathbf{x}_i, \mathbf{x}_c; g_i)$ are described in the section *Parameterization of NC-translocon interactions*.

Interactions between NC beads and ribosome beads are included in the $U_{\text{chan}}(\mathbf{x}_n, \mathbf{x}_c; g)$ potential energy term (Eq 2.9). Contrary to interactions between NC beads and translocon beads, interactions between NC beads and ribosome beads are not bead-type specific; they are described by a repulsive soft-core LJ potential (Eq 2.17), with $\epsilon_{ij} = \epsilon$ and $\sigma_j = 1.2\sigma$. To prevent the NC from moving into the part of the ribosome that is not explicitly included in the simulations (see *3D-CG Model Geometry*), a repulsive sphere is centered at (-10σ, -0.5σ, 1.0σ) (Fig 2.1C). Repulsive interactions with this sphere are described using

$$U_{\text{ribo}}(\mathbf{x}_n) = \sum_{i \in \text{NC}} \left\{ \begin{array}{cc} 4\epsilon \left[\left(\frac{\sigma}{r_{ir} - 2\sigma} \right)^{12} - \left(\frac{\sigma}{r_{ir} - 2\sigma} \right)^{6} \right] + \epsilon & , \quad r_{ir} - 2\sigma < 2^{1/6}\sigma \\ 0 & , \quad r_{ir} - 2\sigma \geq 2^{1/6}\sigma \end{array} \right\}, \tag{2.10}$$

where r_{ir} is the distance of the NC bead i from the center of the sphere.

3D-CG Model Dynamics

The time evolution of the NC beads is modeled using overdamped Langevin dynamics with a first-order Euler integrator [2],

$$\mathbf{x}_n(t + \Delta t) = \mathbf{x}_n(t) - \beta D \nabla_{\mathbf{x}_n} U(\mathbf{x}_n(t), \mathbf{x}_c(t); q, g)\Delta t + \sqrt{2D\Delta t}\mathbf{R}(t), \tag{2.11}$$

where $\mathbf{x}_n(t)$ are the positions of the NC beads at time t, $U(\mathbf{x}_n(t), \mathbf{x}_c(t); q, g)$ is the 3D-CG model potential energy function (Eq 2.4), $\beta = 1/k_B T$, D is an isotropic diffusion coefficient, and $\mathbf{R}(t)$ is a random number vector drawn from a Gaussian distribution with zero mean and unit variance. The timestep, Δt = 300 ns, permits stable integration of the equations of motion with a diffusion coefficient of D = 253.0 nm^2/s (see section *Robustness to simulation timestep* for discussion and Table 2.4 for robustness with respect to timestep). Ribosome CG bead coordinates are fixed throughout the simulations. Translocon CG beads undergo stochastic transitions between fixed configurations associated with the open versus closed lateral gate.

NC-dependent conformational gating of the translocon is attempted at every simulation timestep. The probability that the translocon transitions from a closed ($\mathbf{x}_c^{\text{closed}}$)

to open ($\mathbf{x}_c^{open}$) conformation, $p_{open}(\mathbf{x}_n; q, g)$, is

$$p_{open}(\mathbf{x}_n; q, g) = \frac{1}{\tau_{LG}} \frac{\exp\left[-\beta \Delta G_{open}(\mathbf{x}_n; q, g)\right]}{1 + \exp\left[-\beta \Delta G_{open}(\mathbf{x}_n; q, g)\right]} \Delta t, \tag{2.12}$$

and the probability that the translocon transitions from an open to closed conformation, $p_{close}(\mathbf{x}_n; q, g)$, is

$$p_{close}(\mathbf{x}_n; q, g) = \frac{1}{\tau_{LG}} \frac{1}{1 + \exp\left[-\beta \Delta G_{open}(\mathbf{x}_n; q, g)\right]} \Delta t. \tag{2.13}$$

The timescale for attempting translocon conformational changes, τ_{LG} = 500 ns, is obtained from prior molecular dynamics simulations [145, 146]. The total free energy change for switching the translocon from the closed to open conformation, $\Delta G_{open}(\mathbf{x}_n; q, g)$, is given by

$$\Delta G_{open}(\mathbf{x}_n; q, g) = \Delta G_{empty} + U(\mathbf{x}_n, \mathbf{x}_c^{open}; q, g) - U(\mathbf{x}_n, \mathbf{x}_c^{closed}; q, g), \tag{2.14}$$

where $\Delta G_{empty} = 3\epsilon$ is the free energy penalty for opening a closed channel in the absence of a substrate [3], $U(\mathbf{x}_n, \mathbf{x}_c^{open}; q, g)$ is the 3D-CG model potential energy function (Eq 2.4) with the channel in the open configuration, and $U(\mathbf{x}_n, \mathbf{x}_c^{closed}; q, g)$ is the 3D-CG model potential energy function (Eq 2.4) with the channel in the closed configuration. Previous simulations have found the translocon to exhibit both closed and open lateral-gate conformations [145], and the timescale needed to perform this conformational switch is relatively small (500 ns) in comparison to the other timescales modeled in the 3D-CG model [147]. Therefore, as in the 2D-CG model [146], the lateral-gate conformational changes in the 3D-CG model are described in terms of instantaneous switches between the closed and open conformations. If an attempted conformational change is accepted, all bead positions in the translocon are immediately switched to the positions corresponding to the new channel conformation. The equations of motion described by Eq 2.11-2.14 rigorously obey detailed balance.

Translation of the NC is modeled by adding CG beads to the C-terminus of the NC during a simulation trajectory. At the initiation of the trajectory, the C-terminal NC bead is fixed at the translation insertion point (Fig 2.1C). For each simulation timestep in which translation is performed, the C-terminal bead is moved in the $+z$ direction by a distance equal to $\sigma \Delta t / t_{trans}$, where t_{trans} is the timescale for translating a single CG bead. t_{trans} is set to 0.6 seconds to reproduce a translation rate of 5 residues/second [12, 60, 86, 143] unless otherwise specified. The C-terminal NC

bead is otherwise held fixed, although all interactions between the C-terminal NC bead and other NC beads are included in Eq 2.4. The translation of the C-terminal bead is completed after a period of t_{trans} and its dynamics are described using Eq 2.11 for the remainder of the simulation trajectory. The next CG bead in the NC sequence is then positioned at the translation insertion point and the process is repeated until all NC beads have been translated.

For the combined dynamics of the ribosome-translocon-NC system, a series of five steps is iterated at each trajectory timestep: (*i*) forces acting on each NC bead are calculated, (*ii*) NC bead positions are time-evolved using Eq 2.11, (*iii*) conformational gating of the translocon is attempted (Eq 2.12 and 2.13), (*iv*) ribosomal translation is performed if all NC beads have not yet been translated, and (*v*) the simulation is terminated if user-defined conditions are met. Specific protocols for initializing and terminating simulation trajectories are provided for each workflow described in the *Results*.

3D-CG model parameterization

While the system geometry, 3D-CG model dynamics, and most terms in the 3D-CG model potential energy function (Eq 2.4) are fully described in the *Methods*, the functional forms of the NC-translocon interaction potentials, $U^{\text{aq}}_{\text{chan}}(\mathbf{x}_i, \mathbf{x}_{\text{c}}; g_i)$ and $U^{\text{mem}}_{\text{chan}}(\mathbf{x}_i, \mathbf{x}_{\text{c}}; g_i)$ in Eq 2.9, have yet to be specified. Here, we describe the protocol for obtaining these potentials, which determine sequence-specific NC bead-translocon interactions.

First, we define a protocol for assigning an effective water-membrane transfer free energy, g_i, and charge, q_i, to a NC bead, based on available experimental data. Second, potentials of mean force (PMFs) for translocating model tripeptide substrates across the translocon channel are calculated using the MARTINI residue-based coarse-grained force field. Finally, sequence-specific NC bead-translocon interactions in the 3D-CG model are parameterized by reproducing the MARTINI PMFs using the 3D-CG potential energy function.

Determination of Substrate Water-Membrane Transfer Free Energies and Charge

The water-membrane transfer free energy, or hydrophobicity, of a NC bead, g_i, is calculated by summing the transfer free energies of the associated trio of amino-acid residues. Residue-specific transfer free energies are obtained from the Wimley-

White octanol-water hydrophobicity scale, which measures the partitioning of pentapeptides between octanol and water in a well-defined experimental assay [139]. The Wimley-White hydrophobicity scale has been shown to correlate well with other biophysical hydrophobicity scales [56, 79, 85]. Hydrophobic residues have negative transfer free energies while hydrophilic and charged residues have positive transfer free energies; the full hydrophobicity scale is reproduced in units of ϵ in Table 2.3. The Wimley-White hydrophobicity scale assumes that each residue's peptide backbone participates in intramolecular hydrogen bonds typical of residues forming secondary structure elements. Peptide bonds that do not form intramolecular hydrogen bonds have an additional free energy cost for partitioning into the membrane [8, 92, 139]. Hence, the transfer free energy of a residue is increased by 1.78ϵ, the approximate cost for partitioning a peptide bond that lacks hydrogen bonds, if it is assumed to not form a secondary structure element as discussed in the section *Mapping amino-acid sequence properties to CG beads*.

The charge of a NC bead, q_i, is equal to the sum of the charges of the three associated amino-acid residues. It is assumed that arginine and lysine residues bear a +1 charge, glutamate and aspartate residues bear a -1 charge, and all other residues are neutral. The N- and C-terminal CG beads are assigned an additional +1 and -1 charge, respectively, and have 6ϵ added to their transfer free energies to account for the additional charge [80].

Residue-based coarse-grained simulations

Residue-based coarse-grained simulations are performed using the MARTINI force field, version v2.2P, with the MARTINI polarizable water model [22, 144]. In the MARTINI model, each amino-acid residue is represented by a backbone particle and one or more side-chain particles. MARTINI simulations include the translocon embedded within a lipid membrane containing 368 palmitoyloleoylphosphatidylcholine (POPC) lipids and solvated by an electroneutral 50 mM NaCl salt solution containing 6,225 CG polarizable water molecules (Fig 2.2A). The ribosome is not included due to its large size, and the plug region (Ala^{48}-Leu^{70}) was excluded from the MARTINI representation of the continuous translocon sequence to avoid slow-timescale sampling issues [145]. The translocon is restrained during these simulations to either the closed or open conformation by applying a biasing potential; the minimum distance between any pair of backbone particles in separate LG helices is restrained to be 0.88σ in the closed conformation and 1.75σ in the open

conformation based on previous molecular dynamics results [145]. The described simulation system is used to determine bead positions for the 3D-CG model channel geometry (*3D-CG Model Geometry*) and for PMF calculations. Complete details on the MARTINI simulations, collective variable definitions, and PMF calculation are provided in the section *Additional methods & discussion*.

PMFs for translocating homogeneous tripeptide substrates through the translocon are calculated from umbrella-sampling simulations. The collective variable, d_z, is defined as the distance along the z-axis (i.e., the channel axis) between the center-of-mass of the tripeptide and the center-of-mass of the six hydrophobic pore residues in the translocon [I75, V79, I170, I174, I260, L406] (Figure 2.2A and Figure 2.7). In each umbrella-sampling trajectory, the substrate is kept near a specific value of d_z using a harmonic restraint, confined within a cylinder of radius 1.5σ, and sampled for 400 ns. At least 50 umbrella-sampling trajectories, spanning a range of d_z values between -5.0σ and 4.5σ, are performed for each substrate. Additional simulation trajectories are generated for a restricted range of d_z to improve convergence as needed (summarized in Table 2.2). Each translocation PMF is obtained from the set of corresponding umbrella-sampling trajectories using the Weighted Histogram Analysis Method [73]. Additional details on the umbrella-sampling simulations are provided in the section *MARTINI simulations for translocation PMF profiles*.

Translocation PMFs are calculated for homogeneous leucine (LLL), glutamine (QQQ), and aspartate (DDD) tripeptides. These substrates are selected because their water-membrane transfer free energies span a range from very hydrophobic (LLL) to very hydrophilic (DDD). In the MARTINI force field, each residue is represented by a backbone particle and one or more side chain particles, with the backbone particle type assigned based on the secondary structure of the residue. The LLL substrate is assigned the more hydrophobic "helix" backbone type, the DDD substrate is assigned the more hydrophilic "coil" backbone type, and the QQQ substrate, of intermediate hydrophobicity, is simulated twice, once with the helix backbone type ($\mathrm{QQQ_{helix}}$) and once with the coil backbone type ($\mathrm{QQQ_{coil}}$). The difference in backbone particle type affects only the non-bonded interactions between the backbone particle and other particles; given the short length of the tripeptides, the change in the backbone type does not affect tripeptide structure.

Fig 2.2B shows PMFs calculated from the MARTINI simulations for the translocation of all four substrates and both channel conformations. Previous work has shown that amino-acid water-lipid transfer free energies calculated using MARTINI

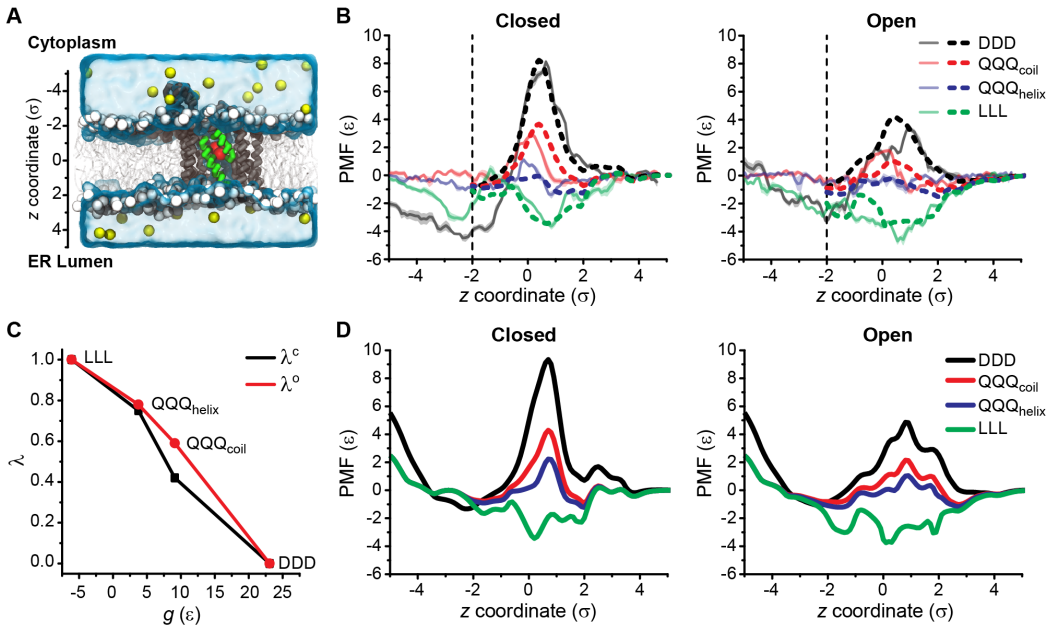

Figure 2.2: Bottom-up parameterization of NC bead-translocon interactions. (**A**) Simulation snapshot of the residue-based coarse-grained simulation system using the MARTINI force field. The translocon is in its closed conformation, a tripeptide substrate is shown in red, lipids are shown with head groups in white and tail groups in grey, water is represented as a transparent surface, and ions are shown as yellow spheres. (**B**) PMFs for translocating homogeneous tripeptides across the closed (left) and open (right) channel conformations. PMFs calculated using MARTINI for all four tripeptides are plotted as transparent lines, with shaded regions indicating the estimated error. The MARTINI PMFs are scaled by a factor of 0.25 and are vertically shifted such that the average value for $4.0\sigma \leq z \leq 4.5\sigma$ is 0. Best-fit PMFs calculated using the 3D-CG model are plotted as opaque dashed lines, and are fit in the range $z \geq -2\sigma$ (dashed vertical line). All PMFs are presented as a function of z, rather than d_z, since these values differ only by an offset of 0.1σ. (**C**) Piecewise linear interpolation relating values of λ^{c} and λ^{o} to the substrate hydrophobicity g. The endpoints of the piecewise linear interpolation correspond to the four substrates in (B). (**D**) PMFs calculated using the 3D-CG model and the best-fit parameters, for the same four peptides as in B, but with the ribosome and translocon plug domain included.

correlate well with the Wimley-White transfer free energy scale, but the correlation has a slope of 3.69 [79]; to treat NC-lipid interactions and NC-translocon interactions in the 3D-CG model on an equal footing, the MARTINI PMFs are rescaled by a factor of 0.25 and the rescaled PMFs are presented in Fig 2.2B. The hydrophobic LLL substrate (green in Fig 2.2) and hydrophilic DDD substrate (black in Fig 2.2) demonstrate opposing behavior in both channel conformations; LLL is attracted to

the center of the channel, which is lined with hydrophobic residues [9, 68], while DDD is repelled. These results qualitatively agree with the atomistic simulations of similar substrates performed by Gumbart et al. [49, 50]. The more hydrophobic QQQ_{helix} substrate is more attracted to the center of the channel than the QQQ_{coil} substrate, while PMFs for both QQQ substrates lie in between the LLL and DDD PMFs. These results show that NC bead-channel interaction ranges from attractive to repulsive as the substrate becomes more hydrophilic.

Parameterization of NC-translocon interactions

Residue-specific NC bead-translocon interactions (Eq 2.9) are obtained by parameterizing the 3D-CG model to fit the MARTINI PMFs shown in Fig 2.2B. Based on the MARTINI results, we assume that: (*i*) NC bead-translocon interactions are a function of substrate hydrophobicity, (*ii*) interactions with the LLL and DDD tripeptides represent the most attractive and most repulsive possible channel interactions, respectively, and (*iii*) all other NC bead-translocon interactions vary between these extremes. Further, we assume that $U^{\text{mem}}_{\text{chan}}(\mathbf{x}_i, \mathbf{x}_c; g_i)$ is independent of NC bead properties. Therefore, the $U^{\text{aq}}_{\text{chan}}(\mathbf{x}_i, \mathbf{x}_c; g_i)$ term in Eq 2.9, which describes pairwise interactions between NC bead i and channel bead j, is decomposed into four separate interactions, given by

$$U^{\text{aq}}_{\text{chan}}(\mathbf{x}_i, \mathbf{x}_c; g_i) = \sum_{j \notin \text{NC}} \left\{ \begin{array}{lll} \lambda^{\text{o}}(g_i) U^{\text{open}}_{\text{LLL}}(r_{ij}) + [1 - \lambda^{\text{o}}(g_i)]\, U^{\text{open}}_{\text{DDD}}(r_{ij}) & , & \text{open channel} \\ \lambda^{\text{c}}(g_i) U^{\text{closed}}_{\text{LLL}}(r_{ij}) + [1 - \lambda^{\text{c}}(g_i)]\, U^{\text{closed}}_{\text{DDD}}(r_{ij}) & , & \text{closed channel} \end{array} \right\}, \tag{2.15}$$

and the $U^{\text{mem}}_{\text{chan}}(\mathbf{x}_i, \mathbf{x}_c; g_i)$ term in Eq 2.9 contains a single term that is not bead-type dependent

$$U^{\text{mem}}_{\text{chan}}(\mathbf{x}_i, \mathbf{x}_c; g_i) = \sum_{j \notin \text{NC}} U_{\text{out}}(r_{ij}), \tag{2.16}$$

where r_{ij} is the distance between NC bead i and translocon channel bead j, $U^{\text{open}}_{\text{LLL}}(r_{ij})$ and $U^{\text{closed}}_{\text{LLL}}(r_{ij})$ are the interactions in the 3D-CG model between a NC bead representing a LLL tripeptide and the open or closed channel, respectively, and $U^{\text{open}}_{\text{DDD}}(r_{ij})$ and $U^{\text{closed}}_{\text{DDD}}(r_{ij})$ are the interactions in the 3D-CG model between a NC bead representing a DDD tripeptide and the open or closed channel, respectively. $\lambda^{\text{o}}(g_i)$ and $\lambda^{\text{c}}(g_i)$ are NC bead-specific parameters that interpolate the channel interactions for NC bead i between the most attractive interaction ($U^{\text{open}}_{\text{LLL}}(r_{ij})$ for $\lambda^{\text{o}}(g_i) = 1$ or $U^{\text{closed}}_{\text{LLL}}(r_{ij})$ for $\lambda^{\text{c}}(g_i) = 1$) to the most repulsive interaction ($U^{\text{open}}_{\text{DDD}}(r_{ij})$ for $\lambda^{\text{o}}(g_i) = 0$ or $U^{\text{closed}}_{\text{DDD}}(r_{ij})$ for $\lambda^{\text{c}}(g_i) = 0$), depending on the bead hydrophobicity, g_i.

The functional form for $U_{\text{out}}(r_{ij})$, $U_{\text{LLL}}^{\text{open}}(r_{ij})$, $U_{\text{LLL}}^{\text{closed}}(r_{ij})$, $U_{\text{DDD}}^{\text{open}}(r_{ij})$, and $U_{\text{DDD}}^{\text{closed}}(r_{ij})$ is a soft-core LJ potential with three free parameters per unique channel bead [10],

$$U(r_{ij}) = \left\{ \begin{array}{cc} 4\epsilon_j^{\text{int}} \left(\frac{1}{\left[\alpha_j + (r_{ij}/\sigma_j)^6\right]^2} - \frac{1}{\left[\alpha_j + (r_{ij}/\sigma_j)^6\right]} \right) - \epsilon_j^{\text{cr}} & , \quad r_{ij} < r_{ij}^{\text{cr}} \\ 0 & , \quad r_{ij} \geq r_{ij}^{\text{cr}} \end{array} \right\}, \tag{2.17}$$

where ϵ_j^{int} is the interaction energy, r_{ij}^{cr} is the right cut-off radius, and σ_j is the diameter of channel bead j. The term ϵ_j^{cr} is the value of the potential at the right cut-off radius, and $\alpha_j = 0.02(\epsilon_j^{\text{int}}/\epsilon)\left[\sqrt{1 + 100(\epsilon/\epsilon_j^{\text{int}})} - 1\right]$ is chosen to cap the maximum value of the potential to prevent infinite energies during the stochastic gating of the translocon conformation, as described in *3D-CG Model Dynamics* (Eq 2.14). For the beadtype independent interactions with the channel exterior, $U_{\text{out}}(r_{ij})$, we assign the free parameters ϵ_j^{int}, r_{ij}^{cr}, and σ_j to represent interactions between NC beads and the hydrophobic channel exterior in a lipid environment (Table 2.1). For the beadtype dependent interactions with the channel interior, $U_{\text{chan}}^{\text{aq}}(\mathbf{x}_i, \mathbf{x}_{\text{c}}; g_i)$, the free parameters, ϵ_j^{int}, r_{ij}^{cr}, and σ_j are fit for each of the four potential energy terms in Eq 2.15, as described below.

In order to find parameters for the 3D-CG model that best reproduce the MARTINI PMF data, corresponding PMFs must be calculated using the 3D-CG model. The PMF for translocating a single CG bead, i, across the channel in the 3D-CG model can be calculated using numerical integration if all interactions for that NC bead with the channel and solvent are defined. As all potential terms other than $U_{\text{chan}}^{\text{aq}}(\mathbf{x}_i, \mathbf{x}_{\text{c}}; g_i)$ (Eq 2.15) are now defined, the MARTINI PMF data is used to define the remaining potential terms. First, parameters for $U_{\text{LLL}}^{\text{open}}(r_{ij})$, $U_{\text{LLL}}^{\text{closed}}(r_{ij})$, $U_{\text{DDD}}^{\text{open}}(r_{ij})$, and $U_{\text{DDD}}^{\text{closed}}(r_{ij})$, are determined independently by fixing the channel in a single conformation, either open or closed, and setting the value of $\lambda(g_i)$ to either 1 or 0 such that only one of the potential terms contributes to the interactions with CG bead i. Specifically, for the open channel configuration, a PMF calculated with $\lambda^{\text{o}}(g_{\text{LLL}})$=1, where g_{LLL}=-6.1ϵ is the water-lipid transfer free energy for a LLL substrate, is fit to the MARTINI PMF for LLL in the open channel to determine parameters for $U_{\text{LLL}}^{\text{open}}(r_{ij})$. A PMF calculated with $\lambda^{\text{o}}(g_{\text{DDD}})$=0, where g_{DDD}=23.1ϵ is the water-lipid transfer free energy for a DDD substrate, is fit to the MARTINI PMF for DDD in the open channel to determine parameters for $U_{\text{DDD}}^{\text{open}}(r_{ij})$. Similarly, for the closed channel configuration, a PMF calculated with $\lambda^{\text{c}}(g_{\text{LLL}})$=1 is fit to the MARTINI PMF for LLL in the closed channel to determine parameters for $U_{\text{LLL}}^{\text{closed}}(r_{ij})$, and a

PMF calculated with $\lambda^{c}(g_{DDD})$=0 is fit to the MARTINI PMF for DDD in the closed channel to determine parameters for $U^{closed}_{DDD}(r_{ij})$. We find that fitting the MARTINI PMFs requires at least two bead types for the translocon channel; one "normal" bead type, and one "confined" bead type, that have distinct parameter values. The values for all resulting parameters are summarized in Table 2.1. Details for the fitting process and the assignment of channel bead types are included in the section *Channel CG bead type assignment and fitting MARTINI PMFs*. Fig 2.2B shows the best-fit PMFs calculated using numerical integration for the 3D-CG model potential energy function with the parameters listed in Table 2.1 (opaque dashed lines) overlaid on the corresponding MARTINI PMFs (transparent solid lines).

Potential	$\epsilon_j^{int}[\epsilon]$	$\epsilon_j^{cr}[\epsilon]$	$r_j^{cr}[\sigma]$	$\alpha_j[\sigma]$	$\sigma_j[\sigma]$
U^{open}_{LLL}	0.46	-0.008	2.500	0.127	1.0
U^{closed}_{LLL}	0.30	-0.005	2.500	0.104	1.0
U^{open}_{DDD}	0.30	-0.005	2.500	0.104	1.0
U^{closed}_{DDD}	0.30	-0.005	2.500	0.104	1.0
$U^{open\text{-}confined}_{LLL}$	1.38	-0.023	2.500	0.209	1.0
$U^{closed\text{-}confined}_{LLL}$	1.41	-0.023	2.500	0.211	1.0
$U^{open\text{-}confined}_{DDD}$	9.85	-9.85	1.075	0.461	1.2
$U^{closed\text{-}confined}_{DDD}$	0.51	-0.51	1.110	0.133	1.2
U^{mem}_{chan}	0.50	-0.008	2.500	0.132	1.0

Table 2.1: Parameters defining NC-translocon interactions

Having obtained parameters for $U_{out}(r_{ij})$, $U^{open}_{LLL}(r_{ij})$, $U^{closed}_{LLL}(r_{ij})$, $U^{open}_{DDD}(r_{ij})$, and $U^{closed}_{DDD}(r_{ij})$, we define a mapping between the transfer free energy (g_i) of any NC bead and its corresponding channel interactions ($\lambda^{o}(g_i)$ and $\lambda^{c}(g_i)$) to fully specify Eq 2.15. These mappings for the LLL, DDD, QQQ_{helix}, and QQQ_{coil} substrates are determined by fitting the MARTINI PMFs. For a CG bead with an arbitrary value of g_i, the corresponding value of $\lambda^{o}(g_i)$ and $\lambda^{c}(g_i)$ is determined by linear interpolation between these four points. As described previously, the values of $\lambda^{o}(g_i)$ and $\lambda^{c}(g_i)$ for the LLL substrate are set to 1, the values of $\lambda^{o}(g_i)$ and $\lambda^{c}(g_i)$ for the DDD substrate are set to 0. For QQQ_{helix} and QQQ_{coil} the values of $\lambda^{o}(g_i)$ and $\lambda^{c}(g_i)$ are determined as follows. First, the channel is fixed in the open conformation and the PMF for translocating a QQQ_{helix} substrate across the open channel in the 3D-CG model is calculated using numerical integration. The QQQ_{helix} 3D-CG model PMF is then fit to the MARTINI PMF for translocating the QQQ_{helix} substrate across the open channel, with $\lambda^{o}(g_{QQQ})$ as a fitting parameter, where g_{QQQ}=3.8ϵ is the

water-lipid TFE of a QQQ helix bead. This procedure is repeated for translocating a QQQ_{helix} substrate across the closed channel to obtain a best-fit value of $\lambda^c(g_{QQQ})$ for the QQQ_{helix} substrate.

Next, the transfer free energy for the QQQ_{coil} CG bead in the 3D-CG model is assigned by increasing the transfer free energy for the QQQ_{helix} CG bead by 5.3ϵ, which is the cost for partitioning three peptide bonds that lack hydrogen bonds between water and alkane (see *Determination of Substrate Water-Membrane Transfer Free Energies*) [8, 92, 139]. PMFs for translocating the QQQ_{coil} substrate across both the open and closed channels for the 3D-CG model are calculated using numerical integration and fit to the corresponding MARTINI PMFs to obtain best-fit values of $\lambda^o(g_{QQQc})$ and $\lambda^c(g_{QQQc})$, where g_{QQQc}=9.1ϵ is the water-lipid transfer free energy of a QQQ_{coil} bead. Best-fit values of the translocation PMFs for the QQQ_{helix} and QQQ_{coil} substrates are shown in Fig 2.2B.

Having obtained $\lambda^o(g_i)$, and $\lambda^c(g_i)$ values for LLL, DDD, QQQ_{coil}, and QQQ_{helix} by direct fitting to the MARTINI PMF profiles, a piecewise linear interpolation between these four sets of g_i, $\lambda^o(g_i)$, and $\lambda^c(g_i)$ values is then performed to define values of $\lambda^o(g_i)$, and $\lambda^c(g_i)$ for a CG bead with an arbitrary value of g_i, as shown in Fig 2.2C. In principle, this mapping between CG bead hydrophobicity and channel interactions could be further refined by simulating translocation PMFs with the MARTINI force field for all possible tripeptide substrates, including heterogeneous tripeptides, and then fitting independent channel interactions in the 3D-CG model for each tripeptide; however, due to the significant computational expense of the MARTINI calculations, we use the piecewise linear interpolation scheme specified above, which yields good agreement with experiments (see *Results*). Future work may further refine the relationship between substrate properties and channel interactions.

The bottom-up parameterization process completely specifies all terms in the 3D-CG potential energy function that define interactions between a CG bead with hydrophobicity g_i and the translocon channel. One caveat is that all translocation PMFs used in the fitting procedure are calculated in the absence of the ribosome and plug domain, which are present in the full 3D-CG model. Fig 2.2D shows PMFs calculated using numerical integration for the same four tripeptide substrates using the 3D-CG model with best-fit values, and including the ribosome and plug domain. Comparing Fig 2.2B and 2.2D shows that the plug domain does not have a large effect on the PMF. The only minor effect associated with including the plug domain appears to be a small shift in the position of the barrier for QQQ_{helix} with the

translocon in the closed configuration; inclusion of the ribosome has no observable effect on the PMFs. The final PMFs, presented in Fig 2.2D, are thus representative of the interactions of CG beads with the translocon during 3D-CG model simulations.

Mapping amino-acid residues to CG beads

The interactions between a general NC bead and the rest of the system is defined by four parameters: g_i, q_i, $\lambda^{\mathrm{o}}(g_i)$, and $\lambda_i^{\mathrm{c}}(g_i)$. These parameters are determined as described in detail in section *3D-CG Model Parameterization*. Specifically, the NC bead transfer free energy, g_i, is equal to the sum of the transfer free energies of the three amino-acid residues associated with the bead according to the Wimley-White hydrophobicity scale (Table 2.3). For each residue that does not form secondary structure, g_i is increased by 1.78ϵ, the cost for partitioning a peptide bond that lacks hydrogen bonds. The CG bead charge, q_i, is equal to the sum of the charges of the three associated amino-acid residues. The N- and C-terminal CG beads are assigned an additional +1 and -1 charge, respectively, and have 6ϵ added to their transfer free energies to account for the additional charge [80]. The scaling parameters for NC-channel interactions, $\lambda^{\mathrm{o}}(g_i)$ and $\lambda^{\mathrm{c}}(g_i)$, are determined from g_i using the piecewise-linear interpolation scheme shown in Fig 2.2C. Fig 2.3 demonstrates the mapping procedure for an example amino-acid sequence.

To start a 3D-CG simulation, both an input amino acid-sequence and a secondary structure assignment for this sequence must be provided. For the membrane integration simulations, the secondary structure of the experimental sequence is reported in the UniProt database and is assigned in the model directly from the available information [17]. For simulations of TMD topology, the secondary structure is not available through the UniProt database and is instead assigned using the PSIPRED secondary structure prediction server [66].

2.3 Validation of the coarse-grained simulation approach

Having fully specified the features and parameters of the 3D-CG model, we now validate the model by simulating three biophysical assays and comparing the simulation results to previously published experimental data. The CG model is used to calculate (*i*) the probability of membrane integration as a function of NC segment hydrophobicity [56], (*ii*) the residue-specific change in the probability of membrane integration (i.e., the “biological hydrophobicity scale") for all twenty amino-acid residues [56], and (*iii*) the distribution of final topologies of a hydrophobic TMD as a function of C-terminal soluble loop length and translation rate [42]. Together,

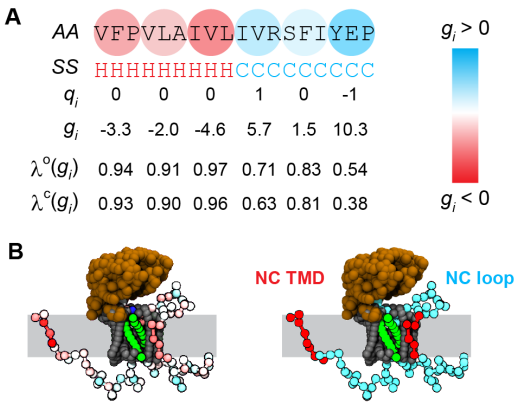

Figure 2.3: Example sequence mapped to 3D-CG model representation (**A**) An input amino-acid sequence (*AA*) and secondary structure assignments (*SS*; H for helix and C for coil) are mapped to 3D-CG beads and assigned values of q_i, g_i, $\lambda^c(g_i)$, and $\lambda^o(g_i)$ based on the properties of sequential amino-acid triplets. (**B**) Visualization of heterogeneous NC properties and correspondence with structural elements. Left, a snapshot of a NC with each CG bead colored by g_i; red beads are hydrophobic, while cyan beads are hydrophilic. Right, the same snapshot colored by assigning each NC bead to a domain.

these tests demonstrate the ability of the 3D-CG model to correctly predict the integration and orientation of TMDs with minimal input, as well as the effect of sequence mutations.

Probability of membrane integration for NC segments of varying hydrophobicity

TMDs typically contain a large number of hydrophobic residues to improve stability within the lipid membrane [140]. von Heijne and co-workers measured the probability with which a designed segment (H-segment) of the leader peptidase (Lep) protein integrates into the membrane, demonstrating that the translocon is more likely to integrate hydrophobic NC segments [56]. It was found that increasing the hydrophobicity of a poly-alanine H-segment, through mutation of alanine residues to leucine residues, monotonically increased the probability of H-segment membrane integration. Previous simulations using model sequences and the 2D-CG simulation model revealed that this trend is caused by local equilibration of the H-segment across the translocon lateral gate [146]. Reproducing the same assay using the 3D-CG model, with full structural detail and an direct mapping of the NC amino acid sequence, provides a first means to quantitatively validate model predictions.

To simulate the H-segment membrane integration assay with the 3D-CG model, the

Lep protein sequence is mapped to CG beads following the procedure described in *Mapping amino-acid sequence properties to CG beads*. Three helical secondary structure elements, including the H-segment are identified via the UniProt database (ID:P00803). Eight 19-residue H-segments are studied. Each H-segment contains between 0 to 7 leucine residues and the remaining H-segment residues are alanine [56]. All trajectories are initialized from configurations in which the two N-terminal TMDs are already translated. To reduce computational cost, simulations are initiated with the second TMD pre-inserted in the lipid membrane (Figure 2.4A). The simulated sequences are limited to 90 CG beads in length, corresponding to a continuous stretch of amino acids starting from the second TMD (All simulated sequences are included as an SI datafile with Ref. [88], and can be found online[1]) σ away from the translocon and span the membrane (integration, Fig 2.4A, Movie S1 in Ref. [88]) or when all CG beads have translocated to the lumenal side of the membrane (translocation, Fig 2.4A, Movie S2 in Ref. [88]). The probability of membrane integration is defined as the fraction of simulation trajectories that terminate by H-segment integration.

Fig 2.4B shows the comparison of the experimental versus the simulated probability of H-segment membrane integration as a function of the number of leucine residues in the H-segment. The results of the experimental assay [56] are plotted in black squares and the shaded region indicates outcomes within 1 kcal/mol of the experimental measurement as determined by a best fit of the apparent free energy of integration via a sigmoidal curve [56]. The calculated results from the 3D-CG model simulations are plotted in red circles. In agreement with the experiments, the 3D-CG model shows that H-segment integration increases with the number of leucines. Although slightly shifted to the right of the experimental curve, the simulation results recover the same sigmoidal dependence of integration on leucine content and are within 1 kcal/mol accuracy of the experiment [56]. These results indicate that the 3D-CG model correctly predicts trends in NC membrane integration using only information about the protein sequence as input.

Fig 2.4C and 2.4D investigate the issue of mapping from trios of amino-acid residues to a single CG bead. There are three possible CG representations (frameshifts) of the NC sequence that arise from the 3:1 mapping of amino-acid residues to CG beads as shown in Figure 2.4C. Since there is no basis for choosing any one frameshift over the other two, each of the possible frameshifts is simulated, and the calculated membrane integration probabilities shown in Fig 2.4B are the averaged value over

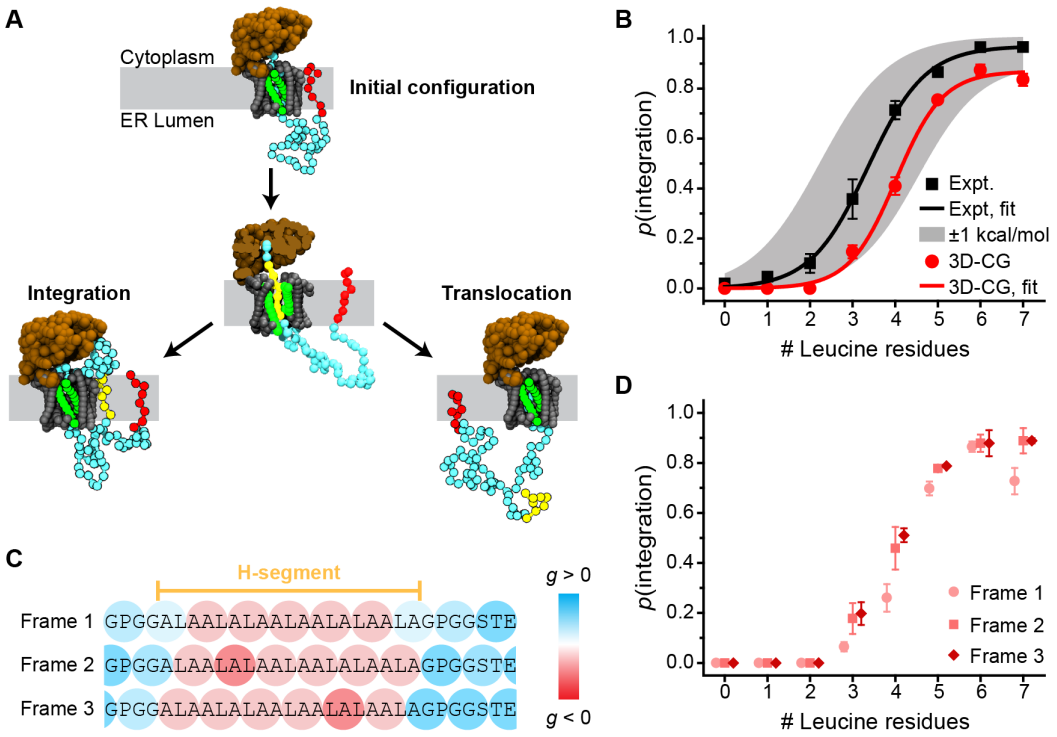

Figure 2.4: 3D-CG model predictions of membrane integration versus secretion. (**A**) Snapshots of the initial system configuration, an intermediate state in which the H-segment (yellow) enters the channel, and two possible simulation products. Simulations are initialized with the TMD upstream of the H-segment (red) integrated into the membrane. (**B**) Probability of membrane integration (p(integration)) as a function of the number of leucine residues in the H-segment. Experimental results from Hessa et al. [56] are reproduced in black, while results from the 3D-CG model are shown in red. Each point for the 3D-CG model is the average of all three frameshifts. The solid lines are sigmoidal fits to each data set. (**C**) Schematic representation of three possible 3D-CG representations of the same sequence (i.e., frameshifts). The example sequence is the Lep construct with a 7 leucine H-segment (identified in yellow region). Each triplet is colored according to its value of g. (**D**) Probability of membrane integration as a function of the number of leucine residues in the H-segment for each individual frameshift.

all three frameshifts. For each frameshift and for each of the eight H-segment sequences, 100 trajectories are calculated (ranging from 20-3000 s in time) leading to 2,400 total simulations which required a total of 15,520 CPU hours on 2.6-2.7 GHz Intel Xeon processors. All CG bead sequences used in the simulations are provided in Ref. [88] and online[1]. Fig 2.4D shows the membrane integration probability for the H-segment sequences for each individual frameshift. Results based on individual frameshifts are comparable, with a notable discrepancy for the

7 leucine H-segment in Frame 1 where the particular grouping of amino acids into triplets resulted in an H-segment for which the integration probability is relatively low. This sensitivity to the choice of triplets is addressed by simply averaging the results over all three frameshifts, which is done for the results in Fig 2.4B.

Effect of single-residue mutations on the probability of membrane integration
As shown in Fig 2.4B, experiments and the 3D-CG model simulations both show that increasing the hydrophobicity of a H-segment by mutating alanine residues to leucine residues increases the probability of H-segment membrane integration. Von Heijne and co-workers have extended this analysis by determining the effect of all twenty amino acids on the probability of H-segment membrane integration in the context of the Lep construct [56]. Assuming that there is an effective two-state equilibrium between the integration and translocation outcomes, the probability of integration can be converted into an apparent free energy of integration, ΔG_{app}, defined by [56]

$$\Delta G_{\text{app}} = -kT \ln \left[p(\text{integration}) / p(\text{secretion}) \right] . \tag{2.18}$$

By mutating the central residue of the H-segment in the same Lep construct used in the section *Probability of membrane integration for NC segments of varying hydrophobicity*, von Heijne and coworkers measured $\Delta G^{\text{aa}}_{\text{app}}$, or the single-residue apparent free energy of integration, for all twenty naturally occurring amino-acid residues, thus deriving a "biological hydrophobicity scale" in analogy to other hydrophobicity scales [79]. Calculating the probability of membrane integration of the same set of H-segments with the 3D-CG model provides a means to validate the predicted effect of single amino-acid residue mutations.

The simulation procedure for calculating the biological hydrophobicity scale is the same as illustrated in Fig 2.4A). To determine $\Delta G^{\text{aa}}_{\text{app}}$ for all 20 amino acids, 22 experimentally studied constructs of the mutated Lep sequence are mapped to a CG representation. Results are averaged over all three frameshifts for each of the 22 constructs, requiring a total of 66 CG bead sequences. All CG bead sequences modeled are provided in the supplemental information of Ref. [88] and online[1]. The probability of H-segment membrane integration is calculated from an ensemble of 200 trajectories (ranging from 20-2000 s in time) per sequence, leading to a total of 13,200 simulations which required a total of 77,003 CPU hours on 2.6-2.7 GHz Intel Xeon processors.

The probability of H-segment membrane integration is converted to a ΔG^{aa}_{app} following the procedure of von Heijne and coworkers described below [56]. The ΔG^{aa}_{app} for alanine and leucine are determined first from a linear fit of ΔG_{app} for H-segments with 3 to 7 Leucine residues from the simulated membrane integration probability curves (Fig 2.4B) using

$$\Delta G_{app} = n_{Leu}\left(\Delta G^{Leu}_{app} - \Delta G^{Ala}_{app}\right) + 19\Delta G^{Leu}_{app}. \tag{2.19}$$

ΔG^{aa}_{app} for alanine and leucine are found to be 0.13 kcal/mol and -0.43 kcal/mol respectively. Experimentally determined values for alanine and leucine are 0.1 kcal/mol and -0.6 kcal/mol respectively. The difference in ΔG^{aa}_{app} between simulation and experiment for leucine gives rise to the slight rightward shift of the simulated membrane integration probability curve compared to the experiment in Fig 2.4B.

To obtain ΔG^{aa}_{app} for the remaining amino acids, we employ [56]

$$\Delta G^{aa}_{app} = \Delta G^{x[aa]x}_{app} - \Delta G^{x[ref]x}_{app} + \Delta G^{ref}_{app}. \tag{2.20}$$

$\Delta G^{x[aa]x}$ is the apparent free energy of integration for an H-segment construct with the probed amino acid (aa) at the midpoint of the H-segment $\Delta G^{x[ref]x}$ is the apparent free energy of integration for the same H-segment where the probed amino acid is replaced by a reference amino acid with a known apparent free energy of integration, ΔG^{ref}_{app}. The reference amino acids employed match those used in Ref. [56] and are specified in the supplemental information of Ref. [88].

The H-segment constructs were chosen to have a leucine content such that the probability of membrane insertion for the sequence is nearly 50% to yield maximum sensitivity in the experimental assay [56]. For cysteine and methionine, we added two additional leucines to the simulated H-segment constructs compared to the experimental constructs to yield additional sensitivity in the computation.

Fig 2.5 compares the values of ΔG^{aa}_{app} determined experimentally to the values of ΔG^{aa}_{app} calculated using the 3D-CG model. Each point represents a single amino acid. Points are colored by grouping amino-acid residues as charged (black), polar (red), aromatic (blue), or non-polar (green). The solid line is a linear fit to the data, while the dashed line illustrates a perfect correlation as a guide to the eye. Each ΔG^{aa}_{app} value is calculated from the average of three frameshifts (defined as in Figure 2.4). The average standard deviation between the frameshift results is 0.2 kcal/mol, the error bars indicate the standard error of the mean. Individual frameshift values are reported in Table 2.5. The experimental and 3D-CG simulation scales are highly

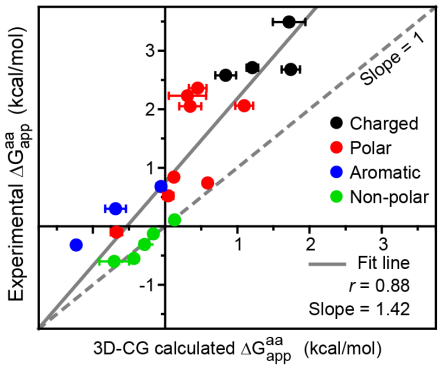

Figure 2.5: Experimental versus simulated predictions of the single-residue apparent free energy of integration. Each point corresponds to a different amino acid, with the character of the amino acid indicated by its plotted color. Each 3D-CG calculated ΔG^{aa}_{app} value the average of three frameshifts, the error bars indicate the standard error of the mean.

correlated ($r = 0.88$), confirming that the 3D-CG model reproduces trends in ΔG^{aa}_{app} with high fidelity. The data points largely lie above the dashed line, indicating that the 3D-CG simulations slightly overestimate the experimentally observed degree of integration. These results thus indicate that the 3D-CG is capable of reproducing the effect of single-residue mutations in good agreement with available biophysical measurements, although the quantitative agreement with experiments may still be improved via further model refinements.

Kinetic regulation of TMD topology

In addition to determining whether NC segments integrate into the membrane as TMD domains, the translocon regulates the orientation with which TMD segments integrate (Fig 2.6A) [26, 41, 42]. In particular, Spiess and co-workers found that an engineered TMD signal anchor (H1Δ22) integrates in either the N_{ER}/C_{cyt} (i.e., Type 1) or the N_{cyt}/C_{ER} (i.e. Type 2) topology; it was also found that decreasing the rate of ribosomal translation by adding cycloheximide increases the preference for the Type 2 topology [42]. Furthermore, increasing the length of the soluble loop flanking the C-terminus of the TMD segment also increases the probability that the TMD segment obtains the Type 2 topology until the probability eventually plateaus for a sufficiently long loop length. Previous work using the 2D-CG model qualitatively captured both these trends and revealed that the mechanistic basis for the kinetic effect is flipping of the NC from the Type 1 topology to the Type 2 topology as a function of time [146]. However, due to the lack of residue-specific interactions in

the 2D-CG model, this work employed model sequences. Additionally, due to the simplified geometric representation of the 2D-CG model, it predicted that p(Type 2) plateaus at shorter C-terminal lengths than is observed in the experiments. While the 2D-CG model can provide mechanistic insights [146], quantitative agreement with the experiments is poor compared to the 3D-CG model when directly mapping the amino-acid sequence (Figure 2.14 and corresponding discussion). Here, we test the 3D-CG model for predicting TMD topogenesis and the effect of translation kinetics on topology.

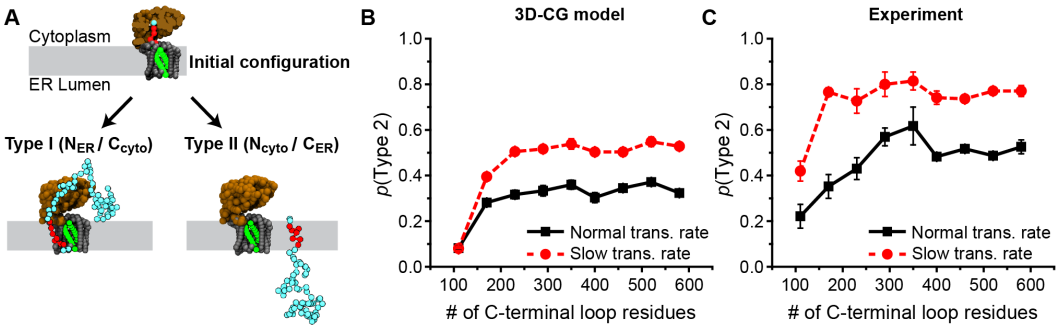

Figure 2.6: 3D-CG model predictions for TMD topology. (**A**) Snapshots of the initial system configuration and the two possible TMD topologies. (**B**) 3D-CG model simulation results showing the fraction of trajectories that reach the Type 2 topology as a function of the number of C-terminal loop residues, plotted for a normal translational rate (solid black) and a slowed translation rate (dashed red). (**C**) Experimental results from Göder et al [42], with a normal translation rate (solid black) and with the addition of cyclohexamide, a translation rate inhibitor (dashed red).

The simulation approach for modeling TMD topogenesis is summarized in Fig 2.6A (see Movie S3 and Movie S4 in Ref. [88] for example trajectories). The H1Δ22 sequence is mapped to CG beads, and the results are averaged over all three frameshifts. Nine different lengths of the C-terminal soluble loop are mapped directly from the experimental constructs used in [42]. The first 99 residues of the sequence are assumed to be part of helical domains based on secondary structure predictions from the PSIPRED server [15, 66]. Simulations are initialized from configurations in which four CG beads are translated and have not yet entered the translocon. Simulations are terminated when CG beads of the TMD have all integrated into the lipid bilayer in either an Type 1 or Type 2 topology and diffuse 10σ away from the translocon. The final TMD topology is determined from the position of the C-terminal CG bead relative to the membrane upon simulation termination (Fig 2.6B). All simulations are performed with either the default translation rate of 5 residues/second

(fast translation) or with a reduced translation rate of 1.25 residues/second (slow translation) to model the effect of adding cyclohexamide in the experimental assay. 200 trajectories (ranging from 25-1200 s in time) are simulated for each of the three frameshifts and for each of the nine loop lengths and at both translation rates, leading to a total of 10,800 trajectories which required a total of 149,009 CPU hours on 2.6-2.7GHz Intel Xeon processors.

Fig 2.6B compares the simulated and experimental results for the probability with which the TMD obtains the Type 2 topology as a function of the length of the C-terminal soluble loop. The results of the experimental assay are plotted on the right for reference. Results for the normal translation rate are in solid black lines, while results for the reduced translation rate are in dashed red lines. The simulation results correctly reproduce the trends observed in the experiments, including the increased probability of the Type 2 topology for longer C-terminal loop lengths and the eventual plateau in the probability of the Type 2 topology at long C-terminal loop lengths. Furthermore, like the experimental results, the CG model predicts a significant shift to greater Type 2 integration upon reducing the rate of ribosomal translation.

2.4 Methods

MARTINI simulation initialization and equilibration

In this section, we describe the initialization of simulations from the section *Residue-based coarse-grained simulations*. Residue-based coarse-grained (RBCG) simulations of the translocon are set up using GROMACS 4.5 [96]. The initial system is prepared by converting the crystal structure of the α, β, and γ-subunits of the Archaeal Sec-translocon (PDB ID: 1RHZ) to a MARTINI RBCG representation using the martinize.py script [22]. Scaffolding interactions are introduced to correctly preserve protein tertiary structure [145]. Scaffolding interactions are included for a pair of CG particles if both are contained in one of the following subsets of the translocon: (*i*) residues Lys^{2}-Val^{45} and Ile^{71}-Pro^{205} in the α-subunit, and the entire β-subunit; (*ii*) residues Trp^{29}-Lys^{66} in the γ-subunit; and (*iii*) residues Pro^{205}-Leu^{433} in the α-subunit. Scaffolding interactions are also included between particles in subsets *i* and *ii*, and between particles in subsets *ii* and *iii*. Scaffolding interactions are only included between CG beads that are separated by 5-9 Å in the original mapping from the crystal structure, and that do not already share a bonded interaction. Scaffolding interactions between pairs of CG particles are weak harmonic distance restraints with an equilibrium distance equal to the distance in the original crystal

structure mapping and a force constant equal to 100 kJ mol^{-1} nm^{-2}.

The RBCG translocon is oriented with respect to a pre-equilibrated lipid bilayer consisting of 400 POPC lipid molecules using the Lambada package [111]. The lipids are then packed around the translocon using the inflategro2 package [111], and lipids that clash with the translocon are removed from the simulation. CG water molecules are added using the *genbox* command in GROMACS, and ions are added using the *genion* command in GROMACS to reach charge neutrality and a physiological salt concentration (~50 mM). The final system contains the translocon, 368 POPC molecules, 6209 CG water molecules, 6 sodium ions, and 17 chloride ions.

The entire system is equilibrated in the MARTINIv2.2 force field using the following protocol: (*i*) 50 steps of steepest descent energy minimization, (*ii*) a 20 ps NPT simulation at 310 K and 1 bar with 2 fs timesteps, and (*iii*) a 100 ns NPT simulation at 310 K and 1 bar with 20 fs timesteps. During steps *ii* and *iii* protein backbone CG beads are position restrained during the equilibration using harmonic constraints with a force constant of 1000 kJ mol^{-1} nm^{-2}. Both NPT simulations use the leap-frog integrator, Berendsen temperature coupling using a temperature coupling constant, τ_T, of 0.5 ps, and semi-isotropic pressure coupling using the Berendsen barostat with a pressure coupling constant, τ_p, of 1.2 ps and an isothermal compressibility of 4.5 x 10^{-5} bar^{-1}.

For simulations with tripeptide substrates, MARTINI representations of NC substrates are added to the system and overlapping water molecules were removed. The new system with the substrate is then equilibrated again using the three step equilibration cycle as described above.

Collective variables used in MARTINI simulations

Here, we describe the collective variables used in section *Residue-based coarse-grained simulations*. Three collective variables (CVs) are used in the MARTINI simulations. This section lists the CVs and provides details on any biasing force applied to these CVs.

The first CV, $d_{LG}(\mathbf{r})$, describes the opening of the translocon lateral gate (LG). It is defined as the minimum distance between the CG backbone beads in TM2b (Ile^{75}-Gly^{92}) and TM7 (Ile^{257}-Arg^{278}) in the translocon α-subunit (Figure 2.7A).

Specifically, $d_{\mathrm{LG}}(\mathbf{r})$ is expressed as

$$d_{\mathrm{LG}}(\mathbf{r}) = \frac{\alpha}{\ln\left[\sum_{ij} \exp(\alpha/|r_{ij}|)\right]}, \tag{2.21}$$

where the sum is over all pairs i, j for which i is a backbone CG bead in TM2b and j is a backbone CG bead in TM7, $|r_{ij}|$ is the distance between CG bead i and CG bead j and α is a large number, the value for α is chosen depending on the distance at which the $d_{\mathrm{LG}}(\mathbf{r})$ is constrained, to avoid precision errors in evaluating the exponential. For simulations of the translocon in the closed LG conformation, a harmonic constraint is placed on $d_{\mathrm{LG}}(\mathbf{r})$ with equilibrium distance 0.7 nm, force constant 2000 kJ mol^{-1} nm^{-2}, and $\alpha = 250$. For simulations of the translocon in the open LG conformation, a harmonic restraint is placed on $d_{\mathrm{LG}}(\mathbf{r})$ with equilibrium distance 1.4 nm, force constant 2000 kJ mol^{-1} nm^{-2}, and $\alpha = 500$.

The second CV, $d_z(\mathbf{r})$, describes the position of the NC substrate along the translocon channel axis, perpendicular to the lipid bilayer. It is the dot product of a distance vector, v_{s}, between the geometric center of the NC substrate and the geometric center of CG beads describing the translocon pore residues (Ile^{75}, Val^{79}, Ile^{170}, Ile^{174}, Ile^{260}, and Leu^{406}) (red vector in Figure 2.7B), and a normal vector, v_{c}, in the positive z-direction originating from the geometric center of CG beads describing the translocon pore residues (blue vector in Figure 2.7B). For the umbrella-sampling trajectories used to construct the PMFs for NC substrate translocation along the channel axis, a harmonic restraint is placed on $d_z(\mathbf{r})$, the equilibrium distance, $d_{z,0}$, and force constant, κ_z, used is listed in the description of the relevant simulations (Table 2.2).

The third CV, $d_{xy}(\mathbf{r})$, describes the distance between the NC substrate and the translocon channel axis in the plane parallel to the lipid bilayer. A soft-wall potential is used to ensure that the NC substrate does not diffuse far from the translocon (Figure 2.7C). The potential is expressed as

$$U_{\mathrm{wall}}(\mathbf{r}) = \left\{ \begin{array}{cc} \kappa_{\mathrm{w}}(d_{xy}(\mathbf{r}) - d_{\mathrm{w}})^2 & , \quad d_{xy}(\mathbf{r}) > d_{\mathrm{w}} \\ 0 & , \quad d_{xy}(\mathbf{r}) \leq d_{\mathrm{w}} \end{array} \right\}, \tag{2.22}$$

where the force constant, κ_{w}, is set to 2000 kJ mol^{-1} nm^{-2}, and the wall distance, d_{w}, is set to 1.2 nm.

All collective variables used in the MARTINI simulations were implemented using PLUMED version 2 [125].

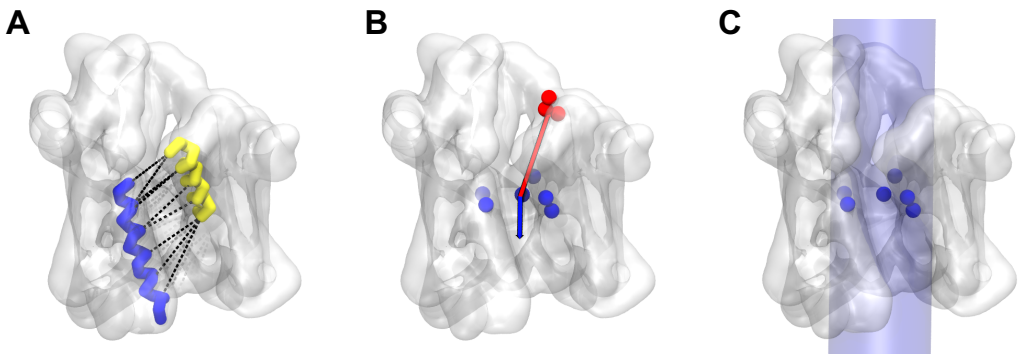

Figure 2.7: Visual representation of the collective variables used in the MARTINI simulations. (**A**) The minimum distance between the backbone CG beads in TM2b (yellow) and TM7 (blue) defines the conformational state of the translocon lateral gate. (**B**) The position of the NC substrate along the channel axis is calculated as the dot product between the NC-pore vector (red) and the channel-axis vector (blue). (**C**) The radial distance of the NC substrate to the channel axis is constrained to be inside a 1.2 nm cylinder (blue region). In all panels the translocon is shown as a white transparent surface.

Details on MARTINI simulations for translocation PMF profiles

In this section, we describe the calculation of the translocation PMF profiles used in the section *Residue-based coarse-grained simulations*. MARTINI simulations are performed using a Langevin dynamics integrator with a 20 fs timestep. Lennard-Jones interactions are shifted from 0.9 to 1.2 nm. Electrostatic interactions are calculated using the smooth Particle Mesh Ewald (PME) method with a grid spacing of 0.12 nm and a short-range cutoff of 1.2 nm. The dielectric constant is set to 2.5 as recommended for the MARTINI polarizable water model. The simulation temperature is maintained at 310 K using a Langevin dynamics integrator. The simulation pressure is maintained at 1 bar via a semi-isotropic Parrinello-Rahman barostat with a coupling time constant of 12 ps and an isothermal compressibility of 3 x 10^{-4} bar^{-1}. These parameters follow recent recommendations for optimal MARTINI simulations using the polarizable water molecule and PME electrostatics [23].

To fully sample the PMF along the channel axis, 51-64 umbrella-sampling trajectories are performed for each tri-peptide substrate in which $d_{z,0}$ is restrained (simulations summarized in Table 2.2). Collective variables were restrained as described in the section *Collective variables used in MARTINI simulations*, and

simulations were carried out in both the open and closed channel conformation. For each simulation reported, at least 100 ns of equilibration is performed followed by 400 ns which is sampled for the calculation of the PMF. Translocation PMFs are obtained from the umbrella-sampling trajectories using the Weighted Histogram Analysis Method [73].

Convergence of MARTINI simulations

We assess the convergence of the MARTINI simulations described in the section *Residue-based coarse-grained simulations* by plotting the PMF as a function of increasing sampling time Figure 2.8 and observe that all PMFs have converged with respect to simulation time. We also plot the overlap in umbrella sampling windows in Figure 2.9 and observe sufficient overlap between all windows.

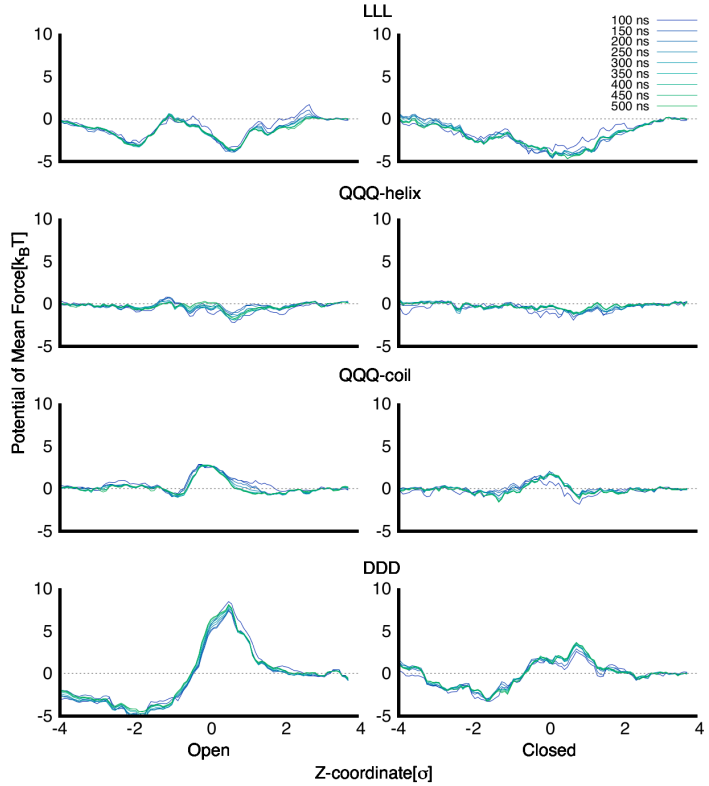

Figure 2.8: Convergence of MARTINI PMFs as the sampling time increases. The PMF is calculated after 10 ns, 50 ns, 100 ns, 150 ns, 200 ns, 250 ns, 300 ns, 350 ns, and 400 ns of sampling time and plotted as a gradient of blue to green lines. The similarity in calculated PMFs as the simulation time increases is used to assess convergence.

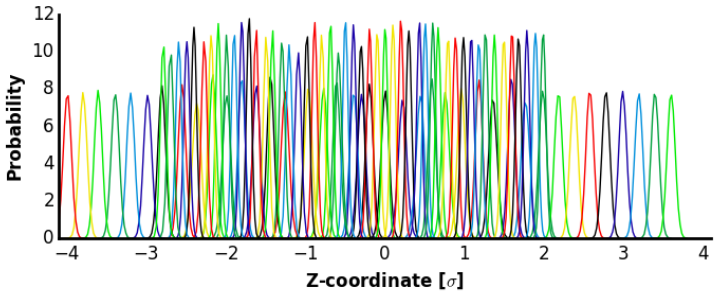

Figure 2.9: Overlap in umbrella sampling windows for the LLL substrate in the open channel. Additional windows with stiffer springs are added to improve overlap between z = -3σ and z = 2σ

Translocon CG bead coordinates

In this section, we describe the determination of translocon channel CG bead coordinates used in the section *3D-CG Model Geometry*. Channel coordinates are obtained from equilibrium MARTINI simulations of the Sec-translocon in explicit solvent. Simulations are set up as described in the section *MARTINI simulation initialization and equilibration*, with a alpha-helical poly-leucine (L_{30}) substrate inside the channel. The distance between the geometric center of the substrate and the geometric center of the translocon pore residues is restrained about zero using a harmonic potential with force constant 200 kJ mol^{-1} nm^{-2}.

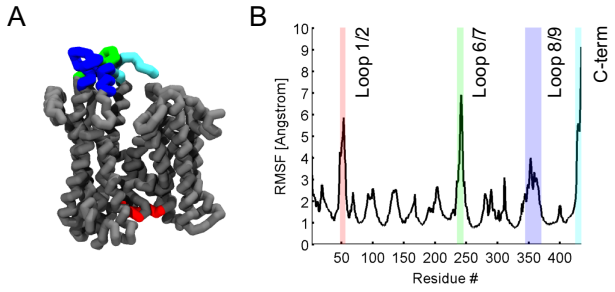

Figure 2.10: Flexible loops identified by equilibrium MARTINI simulations. (**A**) Representative snapshot from a MARTINI simulation of the translocon in the open conformation. Flexible loops are colored as follows; Thr^{47}-Phe^{56} red, Pro^{234}-Gly^{245} green, Lys^{343}-Arg^{369} blue, and Val^{423}-Lys^{433} cyan. Non-protein atoms are not shown for clarity. (**B**) Root-mean squared fluctuation (RMSF) per residue backbone bead calculated from 500 ns of equilibrium MARTINI simulation on the Sec-translocon in the absence of a substrate. Flexible loop regions are highlighted, with colors as in part (A).

To obtain coordinates for the channel in the closed conformation, the system is

Substrate	LG state	$d_{z,0}$ [nm] range	Spacing	Equil. time [ns]	κ_z [kJ mol^{-1} nm^{-2}]
LLL	closed	-4.0 — -1.8	0.2	100	1000
LLL	closed	-1.6 — 1.4	0.1	100	2000
LLL	closed	1.6 — 3.6	0.2	100	1000
LLL	open	-4.0 — -3.0	0.2	600	1000
LLL	open	-2.8 — 2.0	0.1	600	2000
LLL	open	2.2 — 3.6	0.2	600	1000
QQQ_{helix}	closed	-4.0 — -3.0	0.2	100	1000
QQQ_{helix}	closed	-2.8 — -1.4	0.1	100	2000
QQQ_{helix}	closed	-1.3 — 0.8	0.1	600	2000
QQQ_{helix}	closed	0.9 — 1.4	0.1	600	2000
QQQ_{helix}	closed	1.6 — 2.0	0.2	600	1000
QQQ_{helix}	closed	2.2 — 3.6	0.2	100	1000
QQQ_{helix}	open	-4.0 — -3.0	0.2	600	1000
QQQ_{helix}	open	-2.8 — 2.2	0.1	600	2000
QQQ_{helix}	open	2.4 — 3.6	0.2	600	1000
QQQ_{coil}	closed	-4.0 — -3.0	0.2	100	1000
QQQ_{coil}	closed	-2.8 — -1.4	0.1	100	2000
QQQ_{coil}	closed	-1.3 — 1.0	0.1	600	2000
QQQ_{coil}	closed	1.1 — 2.0	0.1	100	2000
QQQ_{coil}	closed	2.2 — 3.6	0.2	100	1000
QQQ_{coil}	open	-4.0 — -2.4	0.2	600	1000
QQQ_{coil}	open	-2.2 — 2.2	0.1	600	2000
QQQ_{coil}	open	2.4 — 3.6	0.2	600	1000
DDD	closed	-4.0 — -1.8	0.2	100	1000
DDD	closed	-1.6 — 1.0	0.1	100	2000
DDD	closed	1.2 — 3.6	0.2	100	1000
DDD	open	-4.0 — -2.4	0.2	600	1000
DDD	open	-2.2 — 1.0	0.1	600	2000
DDD	open	1.2 — 3.6	0.2	600	1000

Table 2.2: Summary of MARTINI simulations used for translocation PMF construction.

simulated for 500 ns with $d_{\mathrm{LG}}(\mathbf{r})$ = 0.7 nm Similarly, to obtain coordinates for the channel in the open conformation the system is simulated for 500 ns with $d_{\mathrm{LG}}(\mathbf{r})$ = 1.4 nm. The average Sec-translocon protein backbone coordinates over the simulation are mapped into the 3D-CG model using the mapping procedure as described in the Main Text section *3D-CG Model Geometry*. Flexible regions of the channel, defined as loop regions with a high RMSF measured from equilibrium MARTINI simulations (Figure 2.10), are excluded from the mapping. An additional bead is included to represent the stable salt bridge between K26 and E421. Final channel

coordinates are included as a supplemental dataset for Ref. [88], and are available online.[2]

Ribosome CG bead coordinates

In this section, we describe the determination of ribosome CG bead coordinates from the section *3D-CG Model Geometry*. The geometry of the ribosome is obtained by mapping the ribosome-translocon complex from a recent high-resolution cryo-EM structure [133] (PDB ID: 3J7Q) onto CG beads. Because CG bead coordinates for the translocon are obtained from the equilibrated average coordinates of residue-based coarse-grained simulations in the absence of the ribosome, the CG model ribosome must be aligned with the CG model translocon by the procedure described below (shown in Figure 2.11). The ribosome-translocon cryo-EM structure [133], which contains a translocon in a partially cracked conformation similar to the closed conformation, is aligned with the CG model of the translocon in the closed conformation by minimizing the root mean squared deviation between the LG helices in the CG model translocon and the LG helices mapped from the cryo-EM structure. After alignment, only CG beads of the ribosome that are within 9σ of origin and with 5σ of the channel axis are explicitly retained as CG beads in the final simulation system.

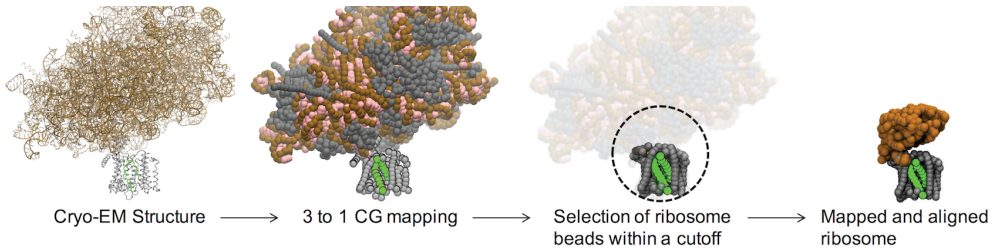

Figure 2.11: Mapping and alignment of ribosome in 3D-CG model. The starting coordinates of the mammalian ribosome-translocon complex from the Voorhees *et al.* cryo-EM structure are shown in the stick representation. These coordinates are mapped in place onto CG beads with CG beads from protein residues shown in gray, CG beads associated with the RNA backbone shown in brown and CG beads associated with RNA nucleobases shown in pink. The lateral gate CG beads, shown in green, are aligned with the equilibrated translocon CG beads from MARTINI simulations and ribosome beads within 9σ of the origin are retained for the final model.

In total, the ribosome consists of 357 CG beads. Complete coordinate files for the CG representation of the ribosome-translocon complex with the translocon in both conformations are included in S1 Dataset.

Translocon CG bead charges

To avoid the assignment of large net charges to the Sec translocon CG beads via the 3:1 mapping protocol, channel beads were assigned a +1 charge if the net charge of the underlying amino acids was positive and a -1 charge if the net charge of the underlying amino acids was negative. In contrast, for nascent chain beads the charge of the CG bead is the sum of the charge of the underlying amino-acid residues. This was done because the CG representation of the translocon was not simulated with all possible frameshifts, as was done for the nascent chain. Only four translocon beads are affected by this issue, and all of these beads are located on the channel exterior. As a result, the PMF for the translocation of charged residues in the 3D-CG model is not sensitive to this subtlety related to determining the charge of the translocon CG beads (Figure 2.12).

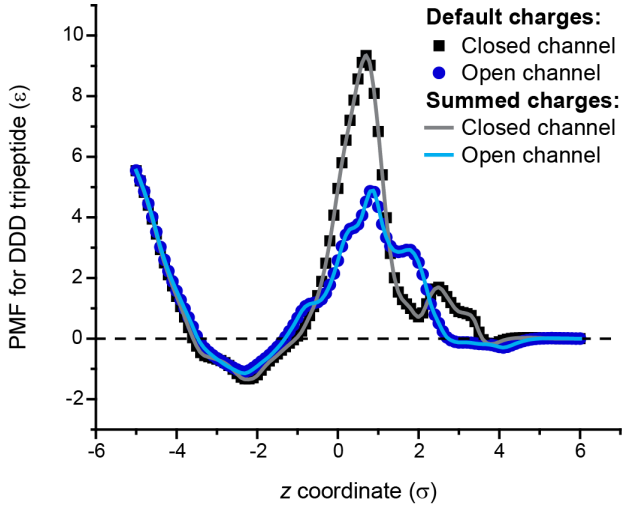

Figure 2.12: Comparison of the PMF obtained with translocon bead charges as described in the section *3D-CG model geometry* (symbols), or with translocon bead charges equal to the sum of the charge of the underlying amino-acid residues (lines). No discernible difference is observed.

Tabulation of the NC-translocon potential energy surface for computational efficiency

To increase the computational efficiency of the 3D-CG model and to exploit the static nature of the translocon and ribosome CG beads (i.e., non-NC CG beads), interactions between NC and non-NC CG beads are tabulated. At the start of a 3D-CG model trajectory, potential energy tables are generated for these interactions. The tables store interaction energies on a rectangular grid surrounding the translocon/ribosome beads. The grid size is chosen such that there is a distance

of at least the non-bonded interaction cut-off length between the grid edge and any translocon/ribosome bead. With a grid-spacing of 0.1σ this yields 121x122x115-element tables. Separate tables are generated for Lennard-Jones and electrostatic interactions. All simulations and analytical potentials of mean force for the 3D-CG model described in this work utilize these potential energy tables.

From the potential energy tables, the interaction energy for a NC bead with the translocon/ribosome beads at any point in Cartesian space is obtained via tri-linear interpolation of the potential energy at the nearest eight grid-points. Specifically,

$$\begin{aligned} U(\mathbf{r}) = {} & \bar{x}\bar{y}\bar{z}U_{111} + \bar{x}\bar{y}(1-\bar{z})U_{110} + \bar{x}(1-\bar{y})\bar{z}U_{101} \\ & + \bar{x}(1-\bar{y})(1-\bar{z})U_{100} + (1-\bar{x})\bar{y}\bar{z}U_{011} + (1-\bar{x})\bar{y}(1-\bar{z})U_{010} \\ & + (1-\bar{x})(1-\bar{y})\bar{z}U_{001} + (1-\bar{x})(1-\bar{y})(1-\bar{z})U_{000}, \end{aligned} \tag{2.23}$$

where the local coordinate, $\bar{x} = (x - x_0)/(x_1 - x_0)$; with x the x-coordinate of the NC bead, x_1 the x-coordinate of the nearest grid-points with $x_1 > x$, and x_0 the x-coordinate of the nearest grid-points with $x_0 < x$. Potential energy values at the grid-points are labeled using the local coordinates at the grid-point, $U_{\bar{x}\bar{y}\bar{z}}$.

Forces on a NC bead at any point in Cartesian space are similarly obtained using interpolation. For example, for the force in the x-direction is given by,

$$\begin{aligned} F_x(\mathbf{r}) = {} & [U_{011} - U_{111}]\bar{y}\bar{z} \\ & + [U_{010} - U_{110}]\bar{y}(1-\bar{z}) \\ & + [U_{001} - U_{101}](1-\bar{y})\bar{z} \\ & + [U_{000} - U_{100}](1-\bar{y})(1-\bar{z}). \end{aligned} \tag{2.24}$$

Amino Acid	WW TFE (kcal/mol)	WW TFE (epsilon)
A	0.5	0.81
C	-0.02	-0.03
D	3.64	5.91
E	3.63	5.89
F	-1.71	-2.78
G	1.15	1.87
H	2.33	3.78
I	-1.12	-1.82
K	2.8	4.55
L	-1.25	-2.03
M	-0.67	-1.09
N	0.85	1.38
P	0.14	0.23
Q	0.77	1.25
R	1.81	2.94
S	0.46	0.75
T	0.25	0.41
V	-0.46	-0.75
W	-2.09	-3.39
Y	-0.71	-1.15

Table 2.3: Wimley-White water-octanol transfer free energy values for all twenty naturally occurring amino acids

Reduced units in the 3D-CG model

Simulations with the 3D-CG model use Lennard-Jones reduced units, which are described in this section. The length-scale is set by the Kuhn-length of a polypeptide chain, σ = 0.8 nm, and the energy unit is $\epsilon = 1k_BT$. The diffusion coefficient used in the 3D-CG simulations is based on previous work [146, 147]. It is reduced by 3-fold compared to the value used in 2D-CG simulations to account for the faster translocation times in three-dimensions, $\tau_{tr} \propto \frac{R_g^2}{D}$. The reduced mass unit, m^* = 300 Da, is the approximate mass of a CG bead containing three amino acid residues. From this, one can derive the reduced time-unit in the 3D-CG simulations,

$$t^* = \sigma\sqrt{\frac{m^*}{\epsilon}}s, \tag{2.25}$$

which yields $t^* \approx 0.01$ ns, the same value as used in the previously published 2D-CG model [146]. This reduced time unit is used to compare simulation results to experimental data. Although caution should be taken when relating CG timescales

with real time, the agreement between 3D-CG simulations and experiments probing kinetic effects on TMD integration (Figure 2.6) are encouraging.

Robustness to simulation timestep

Table 2.4 presents a test of the robustness of the reported results with respect to the timestep employed in the 3D-CG simulations. In addition to the results obtained using the 300 ns timestep employed throughout this chapter, results were recomputed using a timestep of 150 ns. The probability for membrane integration, reported in Figure 2.4B, is recomputed for H-segments containing 4 and 5 Leucine residues. The probability of Type 2 topology for a H1Δ22 signal anchor, reported in Figure 2.6B, is recomputed for sequences with a C-terminal length, L, of 400 residues. In all cases, the results obtained with a 150 ns timestep are within error of the results presented in Figure 2.6.

Experiment	Δt=300 ns	Δt=150 ns
Membrane integration, 4 Leu	26% ± 5%	27 % ± 7%
Membrane integration, 5 Leu	70% ± 3%	70% ± 2%
Type 2 topology, L=400 residues, 5 res/sec	40% ± 2%	40 % ± 5%
Type 2 topology, L=400 residues, 1.25 res/sec	66% ± 2%	64% ± 2%

Table 2.4: Comparison of results with different simulation time steps

Channel CG bead type assignment and fitting MARTINI PMFs

Here, we describe the fitting of the Martini PMFs described in the section *Parameterization of NC-Translocon interactions*. The MARTINI PMFs are fit by defining two bead types for the translocon channel; one “normal” bead type, and one “confined” bead type, that have distinct parameter values. When fitting the PMF for translocating a LLL substrate across the closed and open channel, the “confined” bead type are assigned as beads for which $0.3 < z < 1.0$ and $x^2 + y^2 < 2.25$. All other beads are assigned the “normal” bead type. These bead type assignments are shown in Fig 2.13 with normal bead types shown in gray and confined bead types shown in red. When fitting the PMF for translocating a DDD substrate across the closed and open channel the “confined” bead type are assigned as those beads for which $-0.1 < z < 1.1$ and $x^2 + y^2 < 4.8$. These bead type assignments are shown in Fig 2.13 with normal bead types shown in gray and “confined” bead types as the set of both red and pink beads.

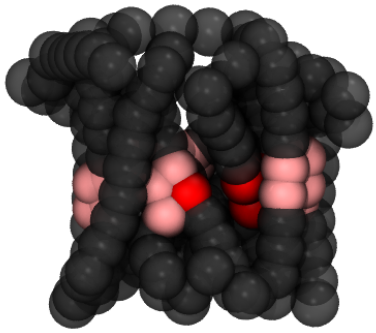

Figure 2.13: Channel bead assignment. Bead assignments in the closed channel conformation with the “default" bead types shown in gray, the repulsive confined beads are shown in red spheres and the attractive confined beads are both the red and pink spheres.

We obtain the values of ϵ_j^{int} using the lmfit Python module to perform a weighted least squares fit between $-2 < z < 5$ where the absolute error is exponentially weighted to prioritize fitting the peaks and valleys. The region $z < -2$ is not used for the fit because this region contains contributions from flexible loop regions of the translocon in the MARTINI PMF simulations; these loop regions have been

removed in the 3D-CG model and therefore features in this region are not expected to be captured in the 3D-CG model. In order to match the MARTINI residue-based coarse-grained simulations for which the plug domain has been removed and the ribosome is not present, the CG beads corresponding to the ribosome and the plug domain are removed during the fitting procedure. For the 3D-CG calculations the plug domain beads are assumed to have normal bead type interactions.

Effect of single-residue mutations on membrane integration by frameshift

Residue	Frame 1 ΔG_{app}	Frame 2 ΔG_{app}	Frame 3 ΔG_{app}	$\langle \Delta G_{\text{app}} \rangle$	SE
I	-0.99	-0.82	-0.30	-0.70	0.21
F	-1.28	-1.26	-1.16	-1.23	0.04
V	-0.41	-0.26	-0.18	-0.28	0.07
C	-0.23	-0.16	-0.10	-0.16	0.04
M	-0.71	-0.79	-0.52	-0.67	0.08
W	-0.85	-0.40	-0.80	-0.68	0.14
T	-0.05	0.18	0.00	0.04	0.07
Y	0.05	-0.21	-0.01	-0.06	0.06
G	0.56	0.54	0.67	0.59	0.04
S	0.14	0.08	0.14	0.12	0.02
N	0.13	0.64	0.27	0.35	0.15
H	1.29	1.03	0.97	1.10	0.12
P	0.02	0.36	0.56	0.31	0.26
Q	0.61	0.39	0.36	0.45	0.12
K	1.46	1.03	1.13	1.21	0.08
R	0.73	1.00	0.79	0.84	0.14
E	1.88	1.70	1.65	1.74	0.13
D	2.00	1.71	1.44	1.72	0.22

Table 2.5: ΔG_{app} values for each residue, separated by frameshift. The result presented in Figure 2.5 is the average over the three possible frameshifts, $\langle \Delta G_{\text{app}} \rangle$. All values are presented in kcal/mol, SE indicates the standard error.

Comparison with the 2D-CG model

In Figure 2.14, we present a comparison of the probability of a H1Δ22 signal anchor obtaining the Type 2 topology calculated using the 3D-CG model and the 2D-CG model. As in Figure 2.6, the probability of Type 2 topology, p(Type 2), is plotted as a function of the C-terminal length, L, for two different translation rates. In the 2D-CG model, the normal translation rate corresponds to 24 residues/second and the slow translation rate corresponds to 6 residues/second.

Simulations performed using the 2D-CG model utilize a direct 3:1 mapping procedure to assign the hydrophobicity and charge of each CG bead, replicating the procedure used in the 3D-CG model. For each of the three frameshifts, 100 trajectories are simulated. This differs from our previous study of topogenesis using the 2D-CG model [146], in which the amino-acid sequence of the experimental construct was not directly mapped to a CG representation; instead, in the earlier work, each construct was represented as a "model" sequence in which CG beads in the C-terminal loop were assigned uniform hydrophilic transfer free energies, CG beads in the transmembrane domain were assigned a uniform hydrophobic transfer free energy, and charges were added to CG beads to approximate the charge distribution of the experimental construct. This previous work demonstrated that the 2D-CG model qualitatively captures the effect of C-terminal length and translation rate on the probability of Type 2 topology for the model sequences. Here, we compare the 2D-CG model versus the 3D-CG model for sequences that are directly mapped from the underlying amino-acid sequence (i.e., not model sequences).

As seen in Figure 2.14, when the CG beads are directly mapped from the amino-acid sequence, the 3D-CG model both qualitatively and quantitatively outperforms the 2D-CG model by more accurately capturing the probability of Type 2 topology and correctly demonstrating the experimentally observed trends. Specifically, the probability of Type 2 insertion plateaus for long C-terminal domain lengths and the probability of Type 2 insertion increases with decreasing translation rate (Figure 2.14). In comparison, the 2D-CG model fails to capture either trend and significantly underestimates the probability of Type 2 insertion. We emphasize that this failure is not inherent to the 2D-CG model itself, since application with model sequences yields these qualitative trends [146]. Nonetheless, these results indicate that the 3D-CG model improves significantly over the 2D-CG model in applications involving the direct mapping of the amino-acid sequence.

2.5 Discussion

This chapter describes a refined CG model for co-translational membrane protein integration via the Sec translocon that captures the detailed three-dimensional geometry of the ribosome-translocon complex from high-resolution structural data [9, 133] and that describes residue-specific interactions between the NC and translocon based on detailed MD simulations. The bottom-up parameterization approach utilized here employs extensive residue-based coarse-grained simulations to inform the model parameters without the need for additional experimental inputs. The re-

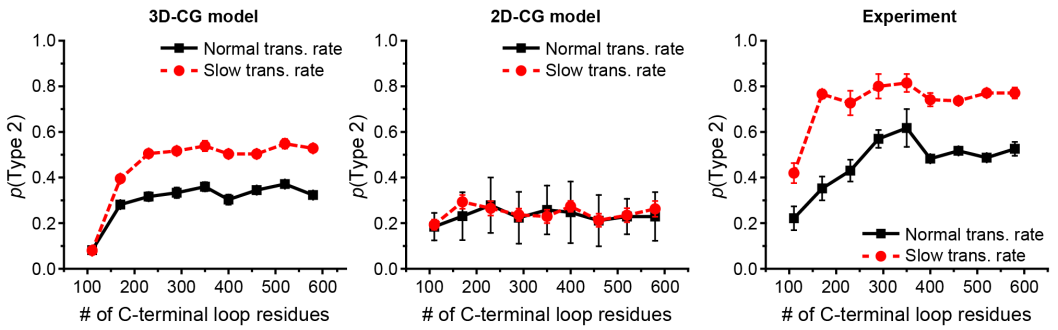

Figure 2.14: Comparison of 3D-CG model simulation results and 2D-CG simulation results showing the fraction of trajectories that reach the Type 2 topology as a function of the number of C-terminal loop residues, plotted for a normal translational rate (solid black) and a slowed translation rate (dashed red). Data for the 3D-CG model and the experiment are reproduced from Figure 2.6

sulting 3D-CG model is applied to calculate the membrane integration efficiency and topology of TMDs, where the only required input is the amino-acid sequence and NC secondary structure. The 3D-CG model captures the experimentally observed [56] sigmoidal dependence of the probability of TMD integration on substrate hydrophobicity. We extend this analysis to study the effect of all twenty amino-acids on the membrane integration probability yielding values of residue-specific TMD membrane integration probabilities in good agreement with the experimentally observed "biological hydrophobicity" scale [56]. These results demonstrate that the 3D-CG model successfully combines factors that are known from previous work to affect TMD integration at the translocon, such as interactions of the nascent chain and the translocon channel interior [16, 51, 147], the non-equilibrium nature of peptide elongation [51, 146], and the sequence context of the TMD [54]. This suggests that the 3D-CG model is well suited for future applications to investigate phenomena such as the experimentally observed position dependence of the biological hydrophobicity scale [57] and the dependence of the observed hydrophobicity values on the amino-acid residues flanking the TMD [54]. The specific interactions between the NC and the translocon, determined as part of this study, already suggest a mechanism by which flanking residues can affect TMD integration; the high barrier for the translocation of charged residues limits translocation, resulting in more integration. Finally, the 3D-CG model accurately describes the experimentally observed effect of translation rate and C-terminal loop length on TMD topogenesis [42]. The 3D representation of the model ensures the correct ribosome-translocon geometry and volume scaling behavior necessary to capture the C-terminal length

dependence of TMD topology, an effect not captured in a previous 2D model [146].

The main advantage of the 3D-CG model presented here, compared to previous work, is that it requires few assumptions. NC properties are directly mapped from the underlying amino acid sequence, the ribosome/translocon geometry is mapped from available structural data, and there is no projection onto a two-dimensional subspace. Provided only with an amino acid sequence and a secondary structure assignment, the 3D-CG model obtains striking agreement with experiment, validating the ability of the 3D-CG model to predict key aspects of Sec-fascilitated protein translocation and membrane integration.

We additionally emphasize that the 3D-CG model provides a refinable framework for simulating IMP co-translational membrane integration via the Sec translocon. Currently, the bottom-up parameterization strategy uses MARTINI PMFs for four distinct tripeptide substrates as input information. The 3D-CG model parameterization could be refined, either by calculating the PMF of other substrates using the MARTINI force field, by considering the role of changes in substrate protonation state in the channel interior, or by calculating PMFs using an atomistic force field. Furthermore, improved methods for parameterization and uncertainty quantification can be employed to determine parameter sets consistent with the available data [6]. All of these refinements can be made within the current 3D-CG model framework, and they will enable incorporation of additional information and improved quantitative prediction. This framework can also be naturally extended to include additional complexity, such as NC secondary and tertiary structure, other proteins that are part of the Sec translocon complex, and a heterogeneous translation rate. Future studies aimed at the prediction of multispanning IMP topology will guide model development.

The 3D-CG model presented here broadens the capability of computer simulation approaches for future studies of the TMD membrane insertion process. In particular, by providing residue-specific NC-translocon interactions the current model enables direct comparison to biophysical measurements of forces between the NC and the translocon due to hydrophobic and electrostatic forces [61, 62]. Furthermore, the realistic representation of the structure and interactions enables future mutational studies and comparison of species-specific features of the ribosome-translocon complex to obtain a detailed understanding of key residues that impact TMD integration and topogenesis. The encouraging agreement between 3D-CG model simulation outcome and experiments for single-spanning TMDs displays the capabilities of

the 3D-CG framework. It enables the calculation of minute-timescale trajectories in three dimensions, facilitating computational studies that are not possible using existing models with less detail, or existing models that are unable to reach the biologically relevant timescales. The 3D-CG model, with initial model parameters obtained here using a bottom-up strategy, provides a systematically improvable framework for the simulation of co-translational membrane protein integration via the Sec translocon.

Chapter 3

FORCES ON NASCENT POLYPEPTIDES DURING MEMBRANE INSERTION AND TRANSLOCATION VIA THE SEC TRANSLOCON

During ribosomal translation, nascent polypeptide chains (NCs) undergo a variety of physical processes that determine their fate in the cell. We resolve the mechanisms and interactions that govern co-translational integration and translocation of NCs via the Sec translocon, using a combination of translation arrest peptide (AP) experiments and coarse-grained molecular dynamics (CGMD) simulations. For a variety of NC sequences, AP experiments measure the external pulling forces that are exerted on the NC at different points during translation, and CGMD is used to simulate the full dynamics of co-translational integration and translocation. Direct comparison between experimentally measured pulling forces and those obtained using CGMD provide validation for the computational model. The CGMD simulations additionally provide a connection between the pulling forces and the underlying molecular interactions that generate those forces, disentangling the contributions from NC-translocon and NC-ribosome interactions, membrane partioning, and coupling to the transmembrane electrostatic potential. The complementary simulation and experiment approach in this work provides a detailed view of the physical processes that determine the fate of proteins in the cell during biosynthesis.

3.1 Introduction

Co-translational protein biogenesis is tightly regulated to ensure that newly synthesized proteins are correctly targeted and folded within the cellular environment. Throughout this process, a nascent polypeptide chain (NC) is exposed to a complex range of forces and interactions, the study of which is complicated by the crowded, stochastic nature of the cell. The work described in this chapter combines 3D-CG simulations and arrest peptide (AP) experiments to connect the forces experienced by a NC to the underlying molecular processes associated with membrane integration and translocation via the Sec translocon.

AP experiments provide a means of probing the co-translational forces that act on the NC, providing a signature of the underlying interactions between the NC and the translocon during co-translational membrane integration. Once an AP is synthesized by the ribosome, it stalls further NC translation; [63] the stall is released with a rate that is dependent on the pulling forces that are experience by the NC. [44] APs are used in nature to control NC translation [63] and have recently been applied to gain insight into physical processes such as integration into the cell membrane, [61] co-translational folding, [18, 89] and electrostatic interactions. [62] In this study, we use AP experiments to measure the forces exerted on model integrating and translocating NCs. To complement the AP experiments, 3D-CG simulations are performed [88] allowing for the direct computation of the NC dynamics, interactions, and resulting pulling forces. The combination of simulation and experiment reveals pulling forces acting on the NC at specific points during translation. Depending on the nature of the NC residues, we observe pulling forces due to NC-translocon and NC-ribosome interactions, membrane integration, and coupling to the membrane potential.

3.2 Results

We investigate a series of integrating and translocating NC substrates to validate the combined simulation and experimental approach, and to provide new insight into the molecular interactions that govern co-translational NC integration and translocation. All NC substrates described in this work utilize a well-established model system, with an engineered domain (H segment) inserted in leader peptidase (Lep) protein (Figure 3.1A, bottom). [56, 61, 62] We study the forces exerted during the integration of a model transmembrane domain, translocation and integration of non-spanning hydrophobic domains, and the translocation of model hydrophilic and charged domains. CGMD calculated forces are compared to previous [61, 62] and

new experimental data, providing validation for the CGMD trajectories. Detailed investigation of the CGMD trajectories, combined with new AP experiments and CGMD with modified interactions are then used to provided molecular level insight into the interactions that govern co-translational NC integration and translocation via the Sec translocon.

Forces on integrating hydrophobic domains

We begin by investigating the co-translational integration of a transmembrane helix, focusing on an engineered transmembrane domain (H segment) in the leader peptidase (Lep) protein (Figure 3.1A, bottom). Previously published AP experiments [61] reveal the points during translation at which forces are exerted on the NC (Figure 3.1B). In the experiment, an AP was inserted downstream of the H segment, and the number of residues between the AP and the H segment, L, was varied (Figure 3.1A, bottom). The H segment consists of 19 residues that are either leucine or alanine, various H segment compositions were tested. In the experiment, the degree of stall-release quantified using the fraction of full-length protein, f_{FL}, is measured as a proxy for the external forces acting on the AP, with greater forces leading to increased stall-release (see *Methods*). Two peaks in f_{FL} were observed around $L = 28$ and $L = 39$ (Figure 3.1B), indicating that force is exerted on the NC as the H segment reaches the translocon and the lipid membrane.

To validate CGMD, we apply it here to calculate the forces exerted on the NC for the same sequences as those tested in the experiment. Experimental sequences are mapped into a coarse-grain representation (Figure 3.1A, top) and simulated to compare calculated f_{FL} values to the experimental data (see Methods). The model successfully captures peaks in f_{FL} at the correct values of L (Figure 3.1C, dashed vertical lines). Consistent with the experiment, the peaks in f_{FL} are dependent on the number of leucine-residues, $nLeu$, in the H segment (Figure 3.1C). We will next demonstrate that the model exhibits two separable peaks, and use the available trajectory data to elucidate the cause of the forces exerted on the NC at $L = 27$ and $L = 39$.

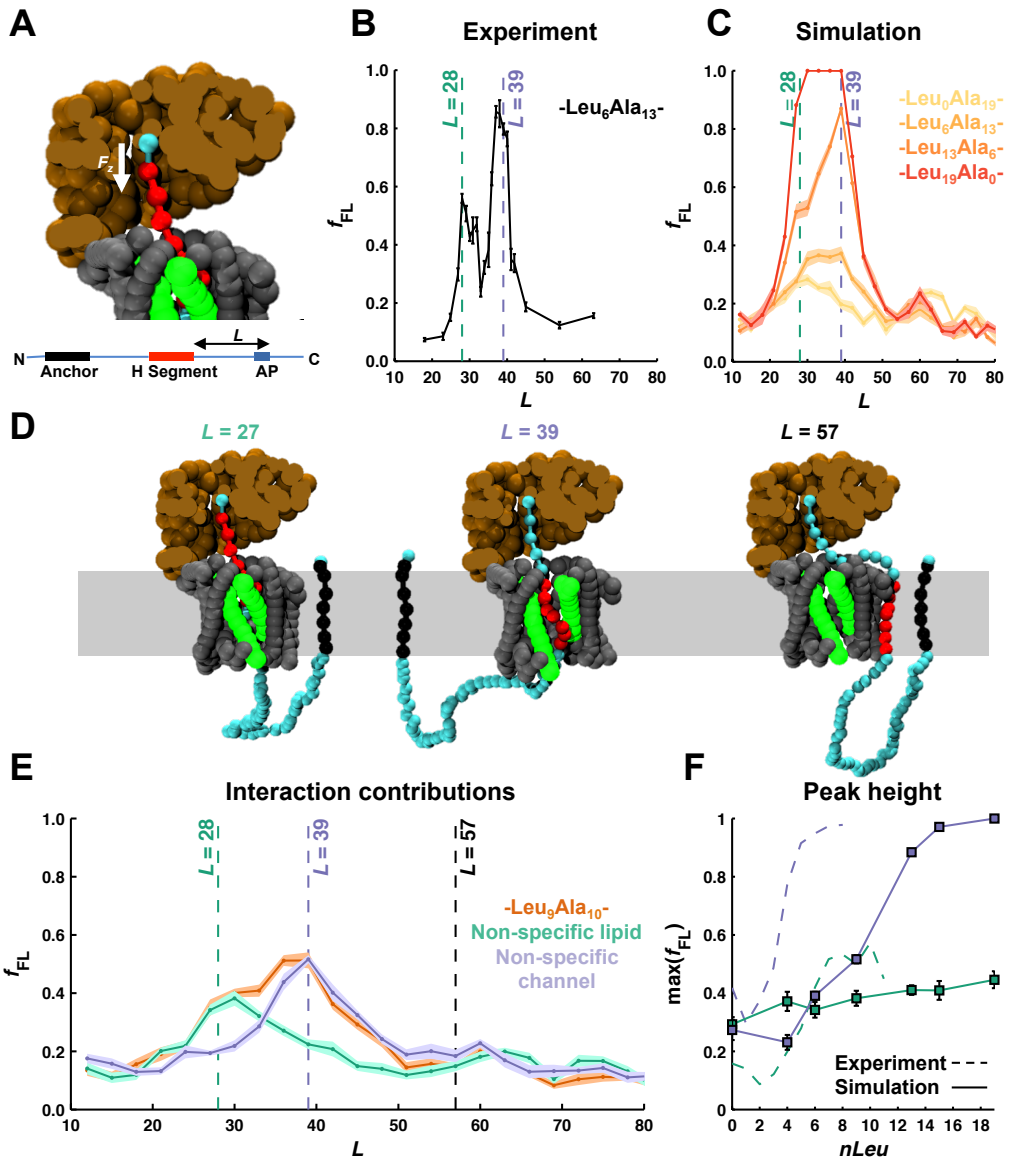

Figure 3.1: Characterization of the physical processes that drive H segment integration. (**A**) Simulation setup used to directly measure the force on the Lep protein with an engineered H segment (red), during co-translational integration. Shown is a simulation snapshot at $L = 27$, the C-terminal bead is held in place and forces exerted by the nascent protein on that bead are collected. (**B**) Experimental data reproduced from ref [61]. Two peaks in f_{FL} are observed. (**C**) Simulation data for H segments of varying Leucine content. Peaks in f_{FL} are identified at similar values of L as compared to the experiment (dashed vertical lines). (**D**) Simulation snapshots showing representative conformations at $L = 27$, $L = 39$, and $L = 57$. (**E**) f_{FL} for an H segment with nine leucine residues with; (red) default interactions, (green) non-specific lipid interactions, and (blue) non-specific channel interactions. (**F**) The maximum value of f_{FL}, $\max(f_{\mathrm{FL}})$, extracted from simulations in which the peaks were isolated as shown in (e). Experimental data [61] is plotted using dashed lines for comparison.

Mechanism of the biphasic pulling force

To identify the physical processes that underlie the observed peaks in force, we analyze the CGMD trajectories. Trajectories where translation was stalled at $L = 27$ exhibit states where the H segment has just reached the cytosolic opening of the translocon (Figure 3.1D, left), indicating that the first peak in force could be due to interaction of the N-terminal part of the H segment with the translocon, in agreement with experimental data on the effects of point mutations in the H segment. [61] Trajectories where translation was stalled at $L = 39$ exhibit states where the H segment is inside the translocon and is partitioning into the lipid membrane via the opened lateral gate (Figure 3.1D, middle), indicating that the second peak in force could be due to integration of the H segment into the lipid membrane, again in agreement with available mutagenesis data. [61] At greater values of L there are no forces observed (Figure 3.1B and Figure 3.1C), at these lengths the trajectories exhibit states where the H segment has completed integration into the lipid membrane (Figure 3.1D, right) and the NC is no longer under tension.

The mechanistic basis of the observed peaks in pulling force suggested by the unbiased simulation trajectories is further confirmed using CGMD with modified interactions. Figure 3.1E shows f_{FL} calculated from simulations either without residue-specific water-lipid transfer free energies (green), or without residue-specific interactions between the NC and the translocon (blue). Simulations without residue-specific interactions between the NC and the translocon do not display a peak in f_{FL} at $L = 27$ (blue line in Figure 3.1E); this result confirms that specific interactions with the translocon give rise to the peak at $L = 27$. Simulations without residue-specific water-lipid transfer free energies do not display a peak in f_{FL} at $L = 39$ (green line in Figure 3.1E); this result confirms that interactions with the lipid membrane give rise to the peak at $L = 39$ Similar results are obtained for all tested H segments, the f_{FL} profiles can be separated into two underlying peaks (Figure 3.2). This demonstrates that in the CGMD simulations two peaks in force are observed, consistent with the experimental data. Additionally, the peaks can be unambiguously assigned to specific physical processes; the peak at $L = 27$ is due to interactions between the NC and the cytosolic opening of the translocon, and the peak at $L = 39$ is due to interaction between the NC and the lipid membrane.

To investigate the experimentally observed sensitivity of force magnitude on H segment hydrophobicity [61] we calculated the maximum in f_{FL} observed for NC sequences with varying H segment hydrophobicity. Figure 3.1F shows the peak

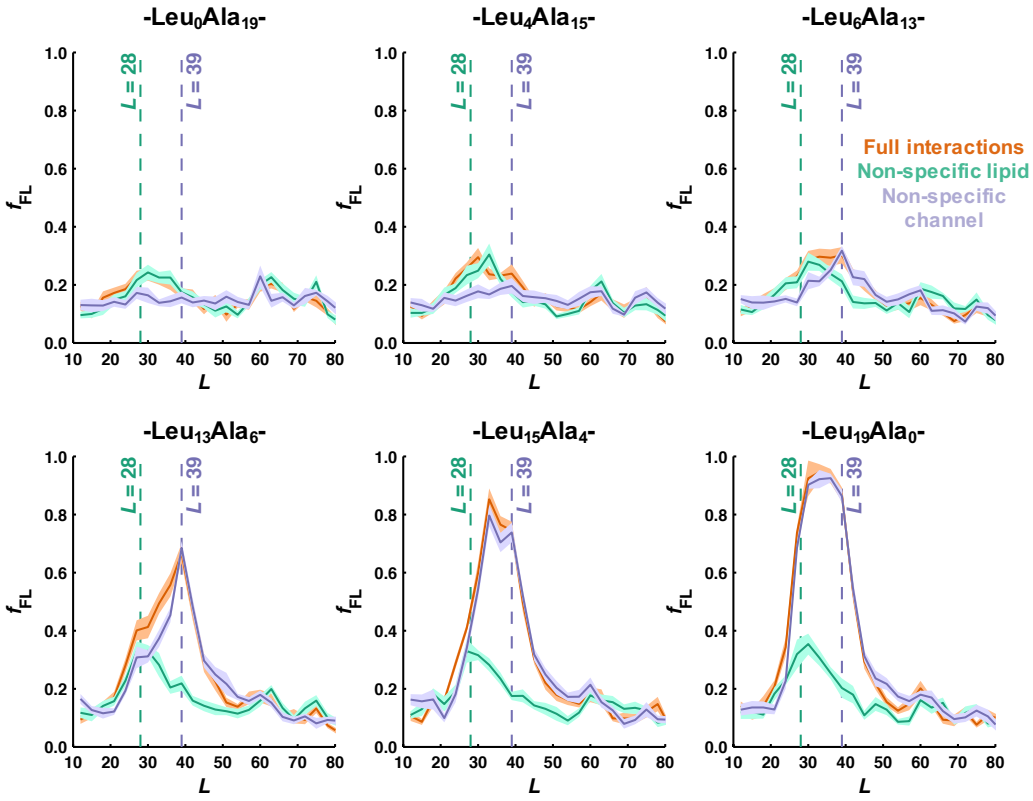

Figure 3.2: f_{FL} for H segments with varying leucine content with; (red) default interactions, (green) non-specific lipid interactions, and (blue) non-specific channel interactions. This data shows that the two separable peaks observed in Figure 3.1E can be observed for all tested H segments. The maximum values in these data were used to construct Figure 3.1F.

values in f_{FL} as a function of the number of leucine residues in the H segment, calculated from trajectories without specific interactions between the NC and the translocon (blue) or without specific interactions between the NC and the lipid membrane (green). The result shows that forces arising from interactions between the NC and the translocon plateau at lower f_{FL} and are much less sensitive to the number of leucine residues in the H segment than the forces due to membrane integration. Experimentally determined f_{FL} values are shown as dashed lines for comparison ($L = 28$ in green and $L = 39$ in blue). Although the model underestimates the magnitude of f_{FL} compared to the experiment, the hydrophobicity dependence and the reduced sensitivity of the peak at $L = 28$ are qualitatively reproduced.

Forces on hydrophobic domains of variable length

To determine if integration is required for lipid interactions to exert a force on the NC, we investigate hydrophobic poly-leucine H segments of varying length using a combination of CGMD simulations and new AP experiments. In contrast to the H segments studied previously, some of the poly-leucine H segments are too short to span the cell membrane (i.e. Figure 3.3A shows a simulation snapshot for a NC with a 8-leucine block stalled at $L + n = 46$). This may affect the forces exerted on the NC due to lipid interactions, because the short H segments will not integrate fully into the cell membrane. For easier comparison of the results for H segments of varying length, n, we define $L + n$ as the number of residues from the start of the H segment to the end of the AP (Figure 3.3A, bottom).

To compare the forces acting on the poly-leucine H segments of varying length to those observed for model transmembrane domains, the forces exerted on the NC are determined using both CGMD and new AP experiments. Both simulation and experiment show a single peak in force (Figure 3.3B and C), compared to the two peaks observed for model transmembrane domains (Figure 3.1B and C). Figure 3.3B shows calculated values of f_{FL} versus NC length, $L + n$, for poly-leucine substrates with varying length. Comparing these results with those obtained for NCs with a 19-residue H segment (Figure 3.1C), the lipid interaction peak, expected at the blue vertical dashed line, is no longer observed. As more leucine-residues are added the peak widens and increases in magnitude. Experiments on NCs with blocks of leucine residues (Figure 3.3C) mirror these trends; as the number of leucines increases the peak shifts up and to greater values of $L + n$.

To confirm whether the lipid-interaction peak is absent for poly-leucine substrates that do not span the membrane CGMD simulations with modified interactions were performed. Surprisingly, in simulations with non-specific channel interactions (Figure 3.3D, blue) a large peak in f_{FL} was observed, and this peak was greatly reduced in simulations with non-specific lipid interactions (Figure 3.3D, green). This indicates that the lipid-interaction peak is not absent for poly-leucine substrates that do not span the membrane; it is not observed because it is shifted to smaller values of $L + n$ than expected and overlaps with the channel-interaction peak. Consistent results are obtained for leucine blocks of various sizes (Figure 3.4). Figure 3.3E demonstrates that lipid-interactions are the dominant source of forces exerted on the NC (Figure 3.3E, blue), with forces due to channel interactions relatively insensitive to the number of leucine residues in the H-segment (Figure

3.3E, green). This is in agreement with previous work, [19] which suggests that hydrophobic segments in the NC slide along the lateral gate of the translocon and are in contact with the lipid membrane.

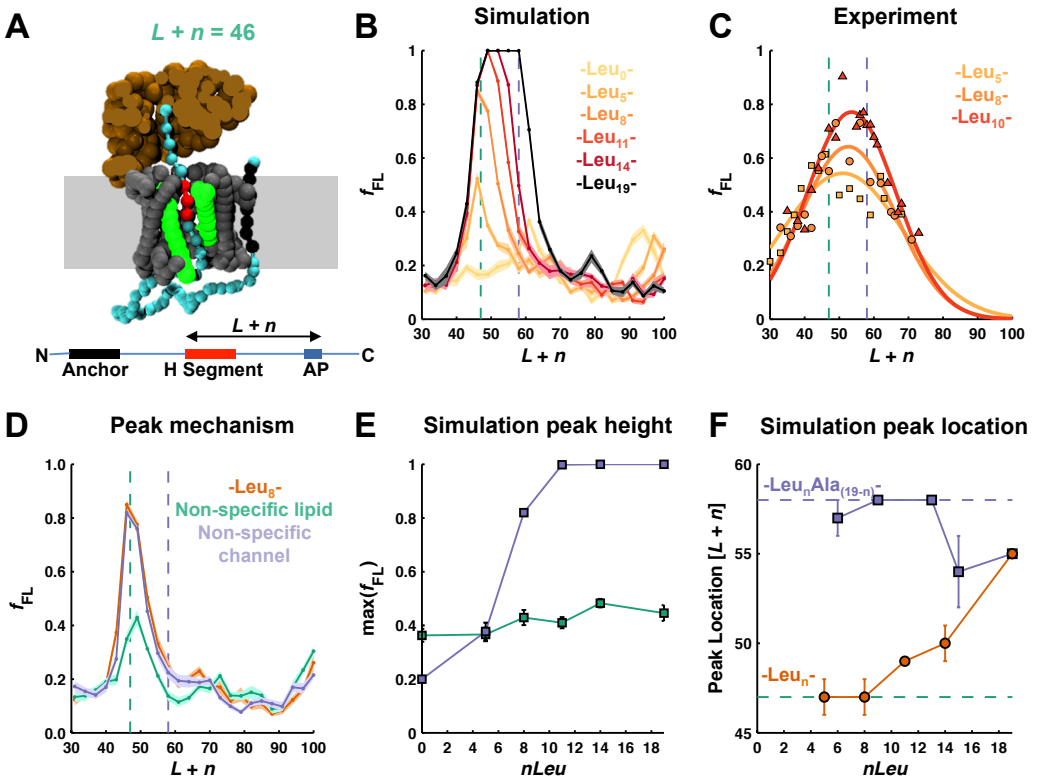

Figure 3.3: Forces exerted on poly-leucine H segments of varying length. (**A**) Simulation snapshot for a H segment with eight leucine residues (red), stalled at $L + n = 46$. The f_{FL} determined from simulation (**B**) and from experiment (**C**) for poly-leucine H segments with increasing numbers of leucine residues. Scatter points reflect experimental data-points, the solid lines are a single-gaussian fit of the experimental data-points. (**D**) f_{FL} for an H segment with eight leucine residues with; (red) default interactions, (green) non-specific lipid interactions, and (blue) non-specific channel interactions. (**E**) The maximum value of f_{FL} extracted from simulations in which the peaks were isolated as shown (d). (**F**) Location of the integration peak in f_{FL} as a function of $nLeu$. For poly-leucine H segments (red) and for 19-residue H segments consisting of alanine and leucine (blue). Error bars indicate the standard error of the mean for simulations with three different frameshifts. The dashed lines correspond to the $L + n$ value at which the channel interaction peak (green) and the lipid interaction peak (blue) are expected, based on the peak locations observed in experiment (Figure 3.1B).

The greatest difference observed between model transmembrane domains and non-spanning hydrophobic segments is the location of the integration peak. We next

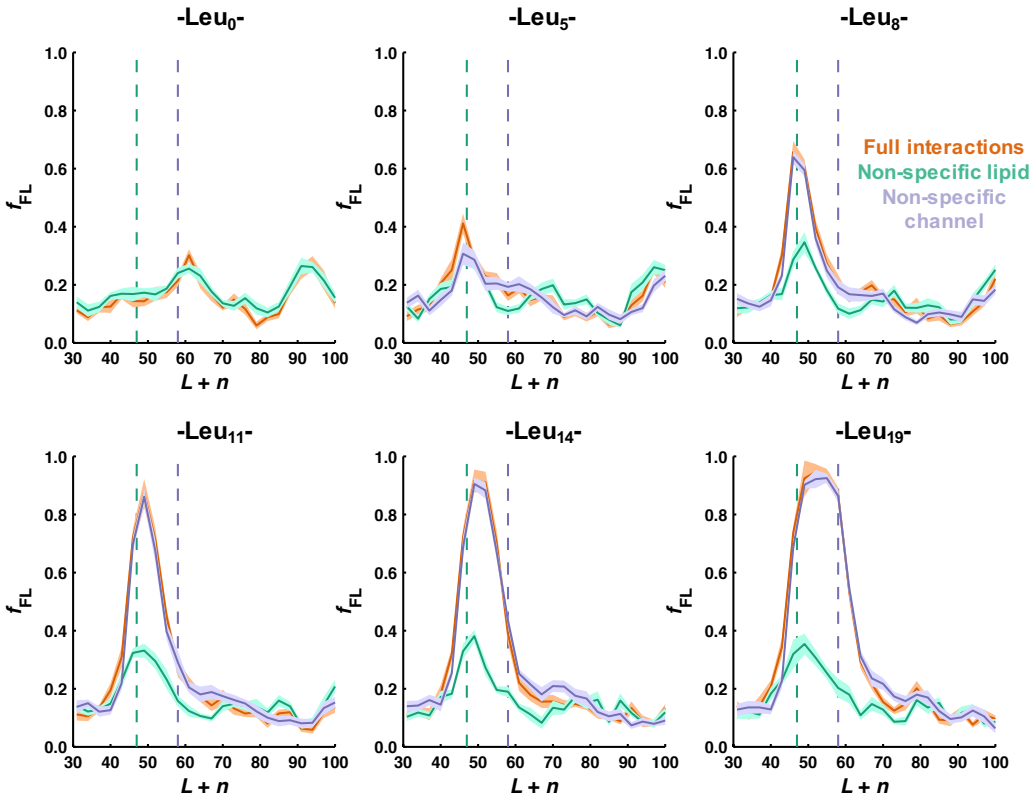

Figure 3.4: f_{FL} for poly-leucine H segments with varying numbers of residues with; (red) default interactions, (green) non-specific lipid interactions, and (blue) non-specific channel interactions. This data shows that the two separable peaks observed in Figure 3.3D can be observed for all tested H segments. The maximum values in these data were used to construct Figure 3.3E, and the value of L at which the maximum in f_{FL} occurs is reported in Figure 3.3F.

compare the location of the integration peak between spanning H segments of varying leucine content and poly-leucine H segments of varying length. Figure 3.3F shows the location of the lipid-interaction peak, determined using CGMD, as a function of the number of leucine residues in the H segment. For H segments with a consecutive block of leucine residues (red) the peak is left-shifted compared to 19-residue H segments (blue), reflected by a peak location at lower values of $L + n$. The left-shift in lipid-interaction peak location for poly-leucine H segments causes the peak to overlap with the channel-interaction peak, leading to difficulties in identifying the two separate physical processes that exert force on the NC in the experiment. As the number of leucine residues increases, the peak shifts to greater values of $L + n$, causing an apparent right-shift in the combined peak (Figure 3.3B and Figure 3.3C). These results indicate that the translocon lateral gate opens as soon

as the short poly-leucine H segments reach the translocon, allowing the hydrophobic domain to interact with the hydrophobic lipid membrane. In agreement, we find that the lateral gate is more likely in the open confirmation at $L + n = 46$ in simulations with the poly-leucine H segments than in simulations with model transmembrane domains (Figure 3.5).

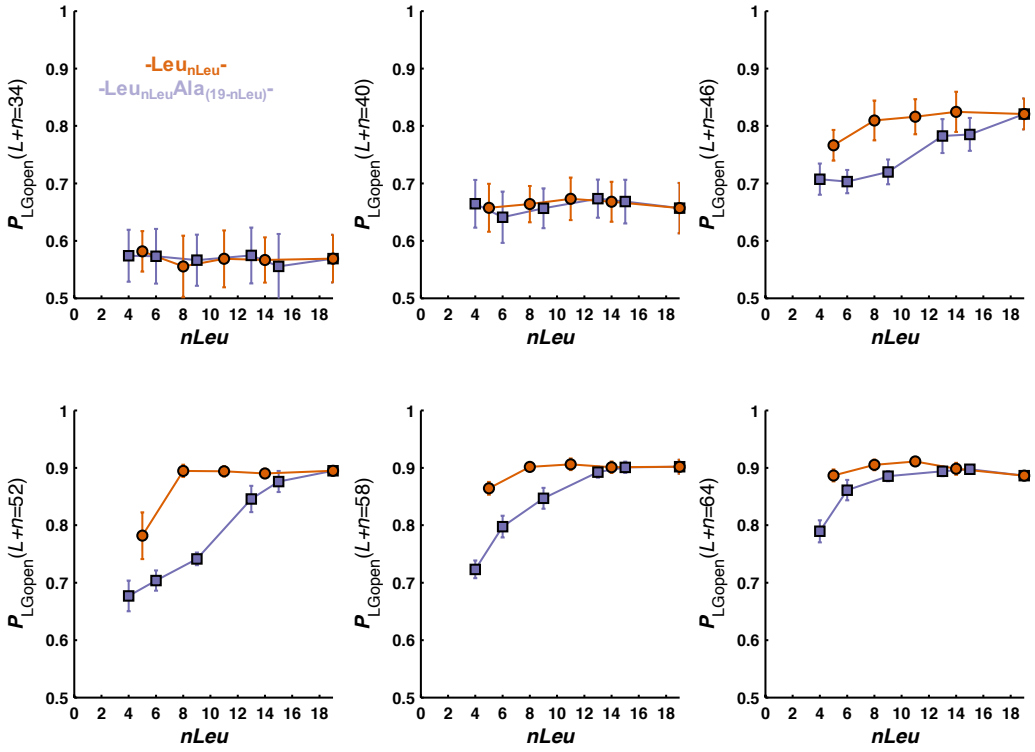

Figure 3.5: The probability that the lateral gate is in the open conformation, $P_{\mathrm{LGopen}}(L + n)$, plotted for various values of $L + n$. Data is shown for both poly-leucine H segments (red) and model transmembrane domains consisting of leucine and alanine residues (blue). At values of $L + n$ corresponding to NC lengths at which the H segment has just reached the translocon ($L + n$=46) we observe that, in substrates with low leucine content, the lateral gate is more likely in the open conformation for the hydrophobic poly-leucine H segments (red). This results in the lipid interaction peak occurring at lower $L + n$ values for these substrates (Figure 3.3E).

Forces on translocating substrates

In addition to the integration of hydrophobic transmembrane domains, the translocon facilitates the translocation of hydrophilic substrates across the cell membrane. Previously published experimental work using APs demonstrated that there are significant forces exerted on the NC due to the translocation segments of negatively

charged residues. [62] The main contributor to these forces was suggested to be the membrane potential, since systematic reduction of the membrane potential by addition of indole led to monotonic reduction of the observed f_{FL}. [62] Here we will apply CGMD simulations to provide a direct connection between observed increases in force on the NC during translation and the underlying physical processes. We show that the dominant feature in the CGMD simulations agrees with AP experiments, and we identify two additional features from the CGMD simulations and determine their underlying physical processes.

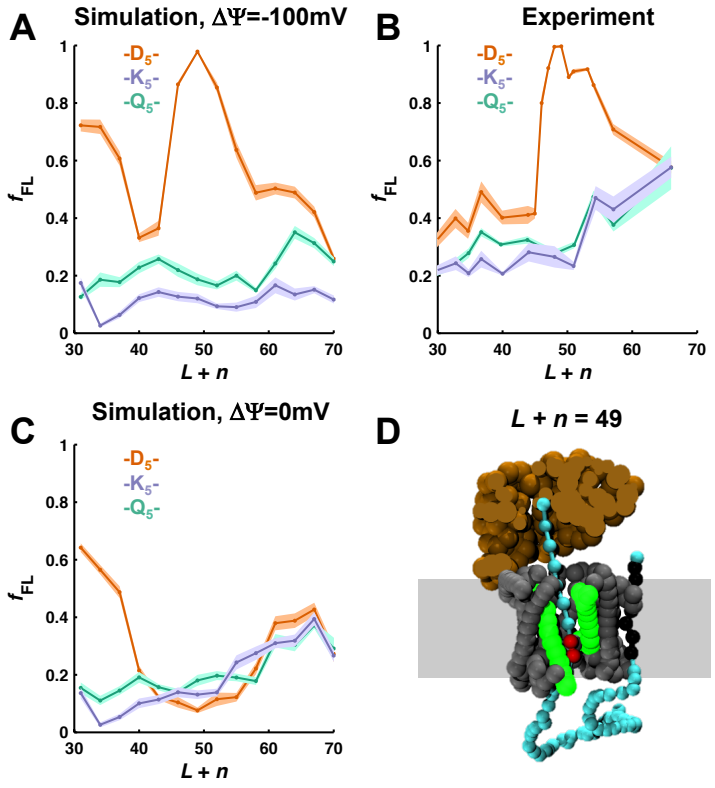

Figure 3.6: Forces exerted on translocating H segments. The f_{FL} determined from simulation (**A**) and from experiment (**B**) for negatively charged (D_5, red), positively charged (K_5, blue), and neutral (Q_5, green) 5-residue H segments. (**C**) As in (A), but for simulations without a membrane potential. (**D**) Simulation snapshot for a D_5 H segment (red), stalled at $L + n = 49$.

To validate CGMD, we apply it here to calculate the forces exerted on the NC for the sequences used in previously published AP experiments using hydrophilic H segments. [62] Figure 3.6A shows f_{FL} as a function of $L + n$ calculated using CGMD. Consistent with experiment [62] (Figure 3.6B), a dominant peak around $L+n = 50$ is observed for a negatively charged H segment (D_5, red). This dominant

peak is not observed for a positive (K_5, blue) or a neutral (Q_5, green) hydrophilic H segment. Additionally, f_{FL} for the negatively charged H segment is higher at all values of $L + n$. Previous work [62] proposed that the dominant peak in f_{FL}, observed for a negatively charged H segment, is due to the membrane potential. To test this proposed mechanism, CGMD simulations were performed where the membrane potential is removed. Comparing the result for simulations without a membrane potential (Figure 3.6C) with the result for simulations with a membrane potential (Figure 3.6A) clearly shows that the dominant feature observed for the D_5 H segment is due to the membrane potential, validating the previously proposed mechanism. [62] Analysis of the CGMD trajectories reveals that at $L+n = 50$ the H segment is in the process of translocating across the cell membrane (Figure 3.6D), suggesting that forces due to the membrane potential could aid in the translocation of negatively charged residues.

In addition to the dominant feature at $L+n = 50$, two additional features are observed in Figure 3.6C; the NC with a negatively charged H segment experiences a force at $L + n < 40$, and all hydrophilic substrates show an increase in f_{FL} for $L + n > 60$. CGMD trajectories for $L + n < 40$ exhibit states where the H segment has not yet reached the translocon, and is instead near the ribosomal exit (Figure 3.7A). The surface of the ribosome is strongly negatively charged, and electrostatic repulsion between the ribosomal surface and the D_5 H segment could generate force on the NC. Consistent with this hypothesis, in simulations without ribosomal charges (Figure 3.7B, black) the feature at $L + n < 40$ is no longer observed. CGMD trajectories for $L + n > 60$ exhibit states where the H segment has translocated across the cell membrane (Figure 3.7C). Interactions between a hydrophilic H segment and the translocon are mostly repulsive, [88] when the H segment translocates it is displaced by residues that potentially have more favorable interactions with the translocon interior. This could generate a driving force for the translocation of a hydrophilic H segment, regardless of the sign of the charge on that H segment. Consistent with this hypothesis, in simulations with non-specific channel interactions (Figure 3.7D, blue) the feature at $L + n > 60$ is no longer observed. CGMD simulations on a mutated NC, where hydrophobic residues C-terminal of the H segment were mutated to alanine, also show a reduction in the feature at $L + n > 60$ (Figure 3.8).

3.3 Discussion

The reported results help to clarify the molecular interactions and conformational changes that govern the co-translational membrane integration and translocation

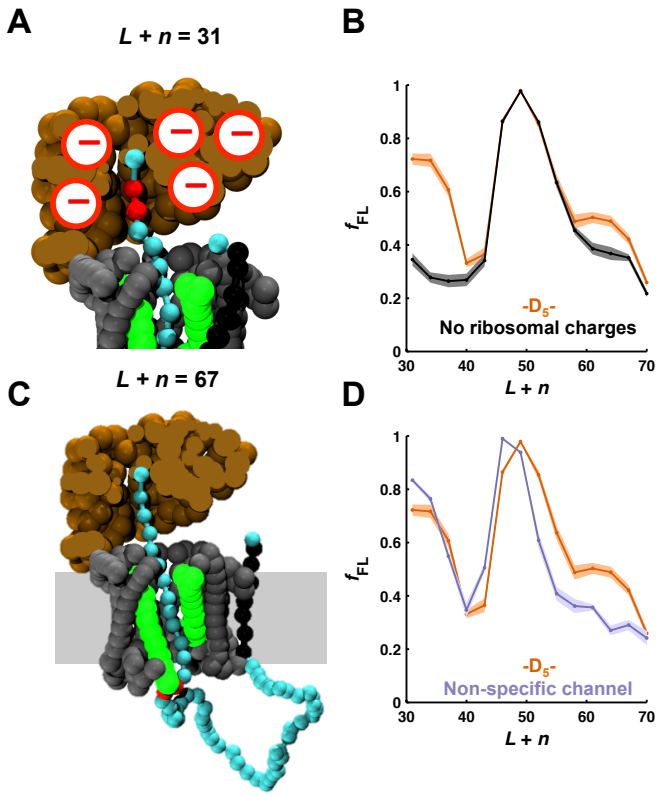

Figure 3.7: Characterization of the physical processes that drive H segment translocation. (**A**) Simulation snapshot for a D_5 H segment (red), stalled at $L + n = 31$. (**B**) The f_{FL} for a D_5 H segment with (red) and without (black) ribosomal charges. (**C**) Simulation snapshot for a D_5 H segment, stalled at $L + n = 67$. (**D**) The f_{FL} for a D_5 H segment with specific channel interactions (red) and with non-specific channel interactions (blue).

of nascent polypeptides. Previously, assigning experimentally observed forces to molecular interactions requires assumptions [31, 44] and further experimental effort. In the current study we combine new AP experiments with minute-timescale CGMD simulations to bridge the gap between experimentally measured forces and the underlying molecular interactions. CGMD has been demonstrated in previous work to capture the effect of amino-acid mutations on the co-translational integration process [84, 87, 88], making it ideally suited to study the molecular interactions that underly the experimentally observed forces.

A wide variety of Sec substrates are analyzed, including model transmembrane domains (Figure 3.1), non-spanning hydrophobic domains (Figure 3.3), and translocating hydrophilic domains (Figure 3.6 and 3.7). For each NC substrate CGMD is validated by the agreement between the calculated forces and the experimentally

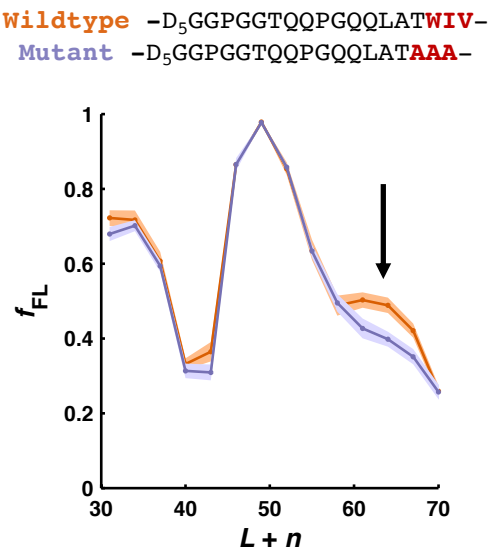

Figure 3.8: The f_{FL} for a D_5 H segment with the wildtype C-terminal flanking loop (red) and with a mutated C-terminal flanking loop (blue). The mutation removes a hydrophobic patch C-terminal of the H segment and replaces it with three alanine residues. The relevant portion of the C-terminal flanking loop is shown at the top of the figure, with the mutation site highlighted in red.

measured stall-release probability; either comparing to previously published experimental data [61, 62], or to new AP experiments. In-depth analysis of the CGMD trajectories, combined with additional simulations using modified interactions, allows us to unambiguously ascribe the forces acting on the NC to specific molecular interactions.

Forces acting on model transmembrane domains are found to be due to attractive interactions between the transmembrane domain and the Sec translocon, and due to interactions between the transmembrane domain and the lipid membrane (Figure 3.1). Non-spanning hydrophobic domains are found to experience forces due to the same molecular interactions as model transmembrane domains (Figure 3.3). However, the forces due to interaction with the lipid membrane occur at an earlier point during translation for these substrates (Figure 3.3F), indicating that the hydrophobic domain samples the lipid membrane as soon as it reaches the Sec translocon. Further analysis confirms that, for the simulated hydrophobic domains, the lateral gate opens as soon as the hydrophobic domain reaches the translocon (Figure 3.5). This finding is consistent with previous simulation [146, 147] and experimental studies [19] that show that hydrophobic domains are able to sample the lipid membrane due to opening of the lateral gate. Forces acting on translocating substrates are due to

repulsive interactions with the ribosome, repulsive interactions with the translocon, and the membrane potential (Figure 3.6 and Figure 3.7). Specifically, hydrophilic domains with a negative charge are repelled from the ribosome (Figure 3.7A and B) and experience a force due to the transmembrane potential (Figure 3.6A and C). For all tested hydrophilic domains, we observe a small but noticeable driving-force for hydrophilic domains to translocate that is dependent on the presence of a C-terminal hydrophobic domain (Figure 3.7C,D, and Figure 3.8). This is an interesting effect of sequence context on co-translational translocation, similar to the effect of sequence context on transmembrane domain integration. [54, 146]

The current study combines AP experiments and CGMD simulations to bridge the gap between minute-timescale biophysical experiments and the underlying nanosecond-timescale protein dynamics for the co-translational integration and translocation of NCs. AP experiments are a powerful experimental tool for the study of co-translational processes, such as co-translational protein folding, and coarse-grain simulations have shown great promise as a tool for simulating the dynamics of these co-translational processes. [126] This suggests additional target applications where a combined AP experiment and CGMD approach could provide molecular insights into systems that are otherwise intractable.

3.4 Methods

In this study we apply a recently developed coarse-grain simulation approach [88] (3D-CG) to measure forces acting on a translating nascent chain (NC) as it is co-translationally inserted into the lipid membrane, or translocated across the lipid membrane. The method is implemented as described in detail in Chapter 2 of this thesis, [88] with the addition of a transmembrane potential as described below.

Simulation setup

Simulations were initiated as described in previous work [88] for the Leader peptidase (Lep) protein with an engineered H segment and the SecM(Ms) arrest peptide. To simulate stalling, translation is halted after a variable number of residues have been translated. The point at which translation is halted is defined by the variable L (Figure 3.1A), defined as the number of residues counted from the end of the H segment to the end of the arrest peptide (AP). After translation is halted the simulation continues for 15s, during this time the force component pointing away from the ribosome and toward the membrane (Figure 3.1A, top), F_z, exerted by the NC on the stalled bead is calculated and stored at 300ns intervals. For each H segment and each

value of L, forces are calculated for 100 independent trajectories. Each H segment is simulated for all three possible frameshifts [88] and the reported results are the average over the results for individual frameshifts, with the error bar representing the standard error of the mean. In total, the presented data is determined from 4500s of simulation time per data point.

Simulations with non-specific interactions were set up and analyzed in exactly the same manner, with as only difference a modification to the inter-particle interactions as described. Instead of having interaction parameters as based on the underlying amino-acid sequence, [88] specific interaction parameters were set to a constant value irrespective of the amino-acid sequence. Specifically; for simulations with non-specific channel interactions all NC beads interact with the translocon using the parameters for a QQQ tri-peptide ($\lambda_c = 0.75$ and $\lambda_o = 0.78$), and for simulations with non-specific lipid interactions all NC beads have a constant water-lipid TFE set to 5ϵ.

Addition of a membrane potential

To investigate forces exerted on H segments containing charged residues (Figure 3.6), a membrane potential, U_{mp} was added to the 3D-CG model based on previous work. [62] The functional form of this membrane potential acting on a bead, i, is described in Eq. 3.1,

$$U_{mp}(z_i; q_i, \Delta\Psi) = q_i \frac{\Delta\Psi}{1 + e^{\kappa z_i}}, \tag{3.1}$$

where z_i is the position of bead i along the channel axis, q_i is the charge of bead i, $\kappa = 1.6\sigma^{-1}$ (0.2Å^{-1}) the lengthscale of the membrane potential drop, and $\Delta\Psi$ is the maximum membrane potential difference either -3.74ϵ ($-100mV$, used in Figure 3.6A and Figure 3.7) or 0ϵ (used elsewhere).

Calculation of fraction full-length protein

To compare to available experimental data the calculated force, F_z, has to be converted to a predicted fraction of full-length protein, f_{FL}. A stalled ribosome continues translation with a force dependent rate, k_{FL}, which is calculated here assuming Bell's model [44, 62] (Eq. 3.2).

$$k_{FL} = k_0 \left\langle e^{\beta \Delta x^{\ddagger} F_z} \right\rangle, \tag{3.2}$$

where k_0 is the rate without an applied force, $\beta = 1/k_B T$, $\Delta x^{\ddagger}$ is an arrest peptide dependent characteristic distance, F_z is the instantaneous applied force, and $\langle .. \rangle$ indicates a time-average over the available trajectory data. The value used for

$\Delta x^{\ddagger}$=0.5 nm is based on previous work, [44, 62] no value for k_0 has to be defined in the final calculation of f_{FL} as it is multiplied by an unknown observation time, t (Eq. 3.4).

The force dependent rate for breaking the translation arrest, k_{FL}, is used to calculate a fraction of full-length protein, f_{FL}, that can be compared to the experimental results using Eq. 3.3. This equation assumes a first-order kinetic scheme in which stalled ribosomes irreversibly resume translation, with force dependent rate k_{FL}, yielding full-length protein. The fraction full-length protein is then (Eq. 3.3-3.4);

$$f_{\mathrm{FL}} = 1 - \exp[k_{\mathrm{FL}} t], \tag{3.3}$$

re-arranging and using Eq. 3.2 for k_{FL} we obtain

$$f_{\mathrm{FL}} = 1 - \exp\left[k_0 t \left\langle e^{\beta \Delta x^{\ddagger} F_z} \right\rangle\right]. \tag{3.4}$$

The only undetermined parameter in this equation is $(k_0 t)$, this parameter is dependent on the arrest peptide, the experimental system and the observation time. The value for $(k_0 t)$ is determined in this work by fitting to the baseline f_{FL} observed in the experiment. $(k_0 t)$ is fit only once, using the data in Figure 3.1B for $L >= 51$, and is held constant in all results shown in this work.

An alternate kinetic scheme assuming the presence of a competing process acting on stalled ribosome, is investigated in the SI and leads to qualitatively similar results as those presented in the main text.

Experimental methods

All plasmids were designed as in [61], i.e., H segments of different amino acid composition and flanked by “insulating" GPGG. . . .GGPG segments were inserted into the periplasmic P2 domain of the *E. coli* inner membrane protein LepB. The 17-residue long AP from the *E. coli* SecM protein (ref) was inserted at varying distances downstream of the C-terminal end of the H segment, leaving a 23-reside C-terminal tail after the AP to ensure that arrested and full-length protein products were of sufficiently different molecular weight to allow separation by SDS-PAGE. Constructs with poly-leucine H segment of composition 5L, 8L, and 10L were expressed in *E. coli*, and analyzed by pulse-labeling, immunoprecipitation, and SDS-PAGE as described in ref [61]. The fraction full-length protein, f_{FL}, was calculated as $f_{\mathrm{FL}} = I_{\mathrm{FL}}/(I_{\mathrm{FL}} + I_{\mathrm{A}})$, where I_{FL} and I_{A} are the intensities of the bands corresponding to, respectively, the full-length and arrested forms of the protein on the SDS-PAGE gel.

Chapter 4

A LINK BETWEEN INTEGRAL MEMBRANE PROTEIN EXPRESSION AND SIMULATED INTEGRATION EFFICIENCY

Adapted from:

Marshall, S. S*., Niesen, M. J. M*., et al. (2016). "A link between integral membrane protein expression and simulated integration efficiency". In: *Cell Reports* 16(8): 2169-2177. DOI: 10.1016/j.celrep.2016.07.042. (*) Equal contribution.

Integral membrane proteins (IMP) control the flow of information and nutrients across cell membranes, yet IMP mechanistic studies are hindered by difficulties in expression. We investigate this issue by addressing the connection between IMP sequence and observed expression levels. For homologs of the IMP TatC, observed expression levels widely vary and are affected by small changes in protein sequence. The effect of sequence changes on experimentally observed expression levels strongly correlates with the simulated integration efficiency obtained from coarse-grained modeling, which is directly confirmed using an *in vivo* assay. Furthermore, mutations that improve the simulated integration efficiency likewise increase the experimentally observed expression levels. Demonstration of these trends in both *Escherichia coli* and *Mycobacterium smegmatis* suggests that the results are general to other expression systems. This work suggests that IMP integration is a determinant for successful expression, raising the possibility of controlling IMP expression via rational design.

4.1 Introduction

Closely related IMP homologs can vary dramatically in the amount of protein available after expression [74], which raises a fundamental question: What differentiates the expression of IMP homologs? The hypothesis raised here is that the efficiency with which an IMP is integrated into the membrane is a key determinant in the degree of observed IMP expression.

A fundamental step in the biosynthesis of most IMPs involves their targeting to and integration into the membrane via the Sec protein translocation channel [98]. Integration of IMP transmembrane domains (TMDs) into the membrane is facilitated primarily through interaction between the nascent chain and SecY, which forms the core of the protein translocation complex, or translocon. Following the co-translational or post-translational insertion of nascent-protein sequences into the translocon channel, hydrophobic segments pass through the lateral gate of SecY into the membrane to form TMDs. Factors such as TMD hydrophobicity [53, 56] and loop charge [42, 130] have been shown to affect the efficiency of TMD integration and topogenesis. For example, TMD hydrophobicity is directly related to the probability with which TMDs partition into the lipid bilayer, while positively charged residues in the loop alter TMD orientation by preferentially occupying the cytosol [42, 56, 130].

In this chapter, we investigate the connection between observed IMP expression levels and Sec-facilitated IMP integration efficiency (i.e., the probability of membrane integration with the correct multi-spanning topology). Systematic investigation of chimeras within an IMP family leads to the identification of sequence elements that modulate expression levels. *in silico* modeling of IMP integration at the Sec translocation channel finds that the sequence modifications that increase the calculated IMP integration efficiency correlates with *in vivo* overexpression improvements, suggesting that IMP integration efficiency is a determinant for successful expression. The result is found to be general across distinct expression systems (*E. coli* and *M. smegmatis*). Furthermore, an *in vivo* assay based on antibiotic resistance in *E. coli* experimentally confirmed the model that the integration efficiency of an individual TMD correlates with the observed IMP expression levels. The strong link between the effect of sequence modifications on simulated integration efficiency and experimentally measured expression levels offers future promise for the rational design of IMP systems with increased expression levels.

4.2 Results

As a detailed case study, the TatC IMP family is employed for all experimental and computational results reported here. A component of the bacterial twin-arginine translocation pathway, TatC plays a key role in the transport of folded proteins across the cytoplasmic membrane [13]. The employment of TatC is well-suited for the current study as it is reasonably sized (only six TMDs (Figure 4.1A)), non-essential, and found broadly throughout bacteria; furthermore, TatC homologs have previously been observed to exhibit widely varying expression levels in *E. coli* [97], suggesting the importance of sequence-level details in the expression of this IMP.

Variance in wild-type and chimeric protein expression levels in *E. coli*

It is first demonstrated that homologs of the IMP TatC exhibit large variance in observed expression levels in *E. coli*. For a quantitative measure of IMP expression, we employ a C-terminal fusion-tag of a green fluorescent protein (GFP) variant [135] (Figure 4.1A) and measure whole-cell fluorescence by flow cytometry. Whole-cell fluorescence intensity of this fusion-tag has been validated in numerous previous studies to correlate strongly with the amount of folded IMP, rather than the total level of IMP translated [27, 34, 38, 47, 136]; we further validate the expression levels measured from whole-cell fluorescence (Figure 4.1B) using in-gel fluorescence (Figure 4.1C and Figure 4.2, Pearson correlation coefficient, $r = 0.914$) and western blot analysis (Figure 4.2A). With this approach, expression levels in *E. coli* are experimentally measured for TatC homologs from a variety of bacteria, including *Aquifex aeolicus* (*Aa*), *Bordetella parapertussis* (*Bp*), *Campylobacter jejuni* (*Cj*), *Deinococcus radiodurans* (*Dr*), *Escherichia coli* (*Ec*), *Hydrogenivirga* species 128-5-R1 (*Hy*), *Mycobacterium tuberculosis* (*Mt*), *Staphylococcus aureus* (*Sa*), *Vibrio cholera* (*Vc*), and *Wolinella succinogenes* (*Ws*) (sequences available online[3]).

Figure 4.1B demonstrates the wide range of expression levels that are exhibited by the TatC homologs in *E. coli*. Previous expression trials of TatC homologs identified that *Aa*TatC is readily produced at high levels in *E. coli*, which enabled the solution of its structure [97, 99]. In contrast, low expression is found for both the *Mycobacterium tuberculosis* TatC (hereafter referred to as *Mt*TatC(Wt-tail) and a modified sequence truncating the un-conserved 38-residue sequence of the C-terminal loop (hereafter referred to as *Mt*TatC) [97].

To examine the parts of the protein sequence that affect expression, “swap chimeras" were generated by exchanging entire loops and TMDs between *Aa*TatC and *Mt*TatC

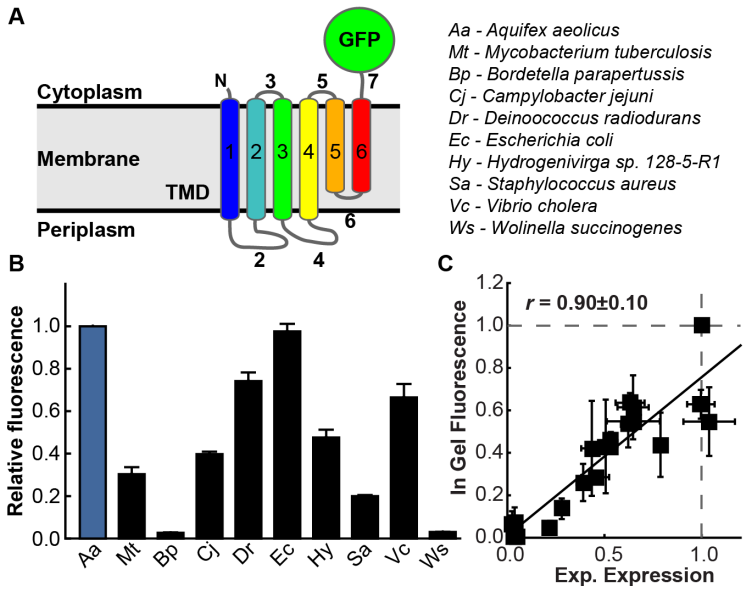

Figure 4.1: Variation in the expression of TatC homologs in *E. coli*. (**A**) A topology representation of TatC with a GFP C-terminal tag, as used in the expression studies. TMDs and loops are indicated in colors and gray, respectively, and are numbered. (**B**) Expression levels of various TatC homologs in *E. coli*, measured by TatC-GFP fluorescence, with expression levels normalized to *Aa*TatC (blue). Error bars indicate the standard error of mean. (**C**) Correlation of the in-gel fluorescence quantified for each band versus the experimental expression measured by flow cytometry. Both metrics are highly correlated across multiple trials (Pearson correlation coefficient $r = -0.9 \pm 0.1$) with in-gel fluorescence showing the same trends in expression yield as seen by flow-cytometry. Error bars indicate the SEM. See also Figure 4.2.

(sequences available online[3]). The TMDs and loops were defined by comparing sequence alignments and membrane topology predictions (Figure 4.3B) [117, 128]. The swap chimeras exhibited a wide range of expression results (Figure 4.3A). The C-terminal loop sequence, referred to as the C-tail and labeled as loop 7 in Figure 4.1A, was found to have a significant effect on expression levels (shaded bars in Figure 4.3A). Removal of the *Mt*TatC C-tail improves expression. Removal of the C-tail from the *Aa*TatC sequence leads to a corresponding decrease in expression. Strikingly, swapping the *Aa*TatC C-tail (*Aa*-tail) into the *Mt*TatC sequence leads to a significant improvement in expression.

The positive effect of the *Aa*-tail on *Mt*TatC expression raises the question of whether expression can be similarly improved in other TatC homologs by substituting the corresponding C-tail sequence (Figure 4.3E) with that of *Aa*TatC. Swapping the C-tail of the various TatC homologs with the *Aa*-tail improved expression in seven

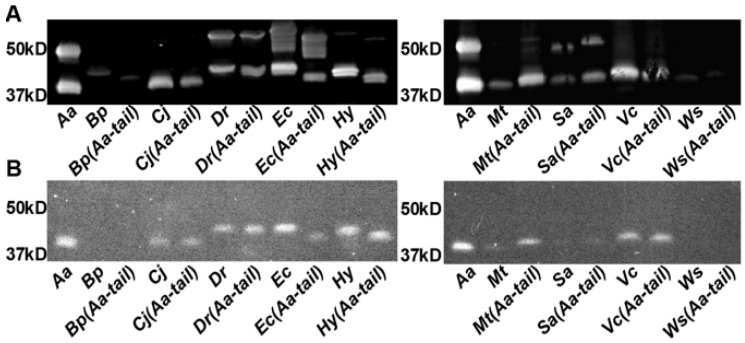

Figure 4.2: (**A**) Anti-GFP western blot results for TatC homologs and the corresponding *Aa*-tail swap chimeras. Two bands were observed for all lanes where TatC-GFP was at high relative concentrations with the lower bands active by in-gel fluorescence and therefore determined to be folded protein[135]. (**B**) In-gel fluorescence of SDS-PAGE for TatC homologs and the corresponding *Aa*-tail swap chimeras. Bands that exhibit fluorescence represent folded protein. The results exhibit the same trends in expression yield as seen by flow-cytometry. This data was used in Figure 4.1C.

out of nine cases (Figure 4.3D). Taken together, the results in Figure 4.3 indicate that the C-tail is a significant factor in determining TatC expression across homologs.

Simulation suggests integration efficiency as a cause for variation in expression levels

To investigate the mechanistic basis for the experimentally observed effect of the C-tail on expression, we employ a recently developed *in silico* coarse-grained approach that models co-translational translocation on unbiased biological timescales [146]. The coarse-grained model, which is derived from over 16 μs of molecular dynamics simulations of the Sec translocation channel, the membrane bilayer, and protein substrates [145, 147], has been validated for the description of Sec-facilitated membrane integration, including experimentally observed effects of amino-acid sequence on the membrane topology of single-spanning IMPs [146] and multi-spanning dual-topology proteins [129]. IMP sequences are mapped onto a Brownian dynamics model of the ribosome/translocation-channel/nascent-protein system, and the Sec translocon-facilitated integration of the IMP into the lipid bilayer is directly simulated in 1,200 independent minute-timescale trajectories for each TatC (Figure 4.4A). The current implementation of the coarse-grained model does not distinguish between expression systems.

Using the results of the coarse-grained model, Figure 4.4B presents the simulated

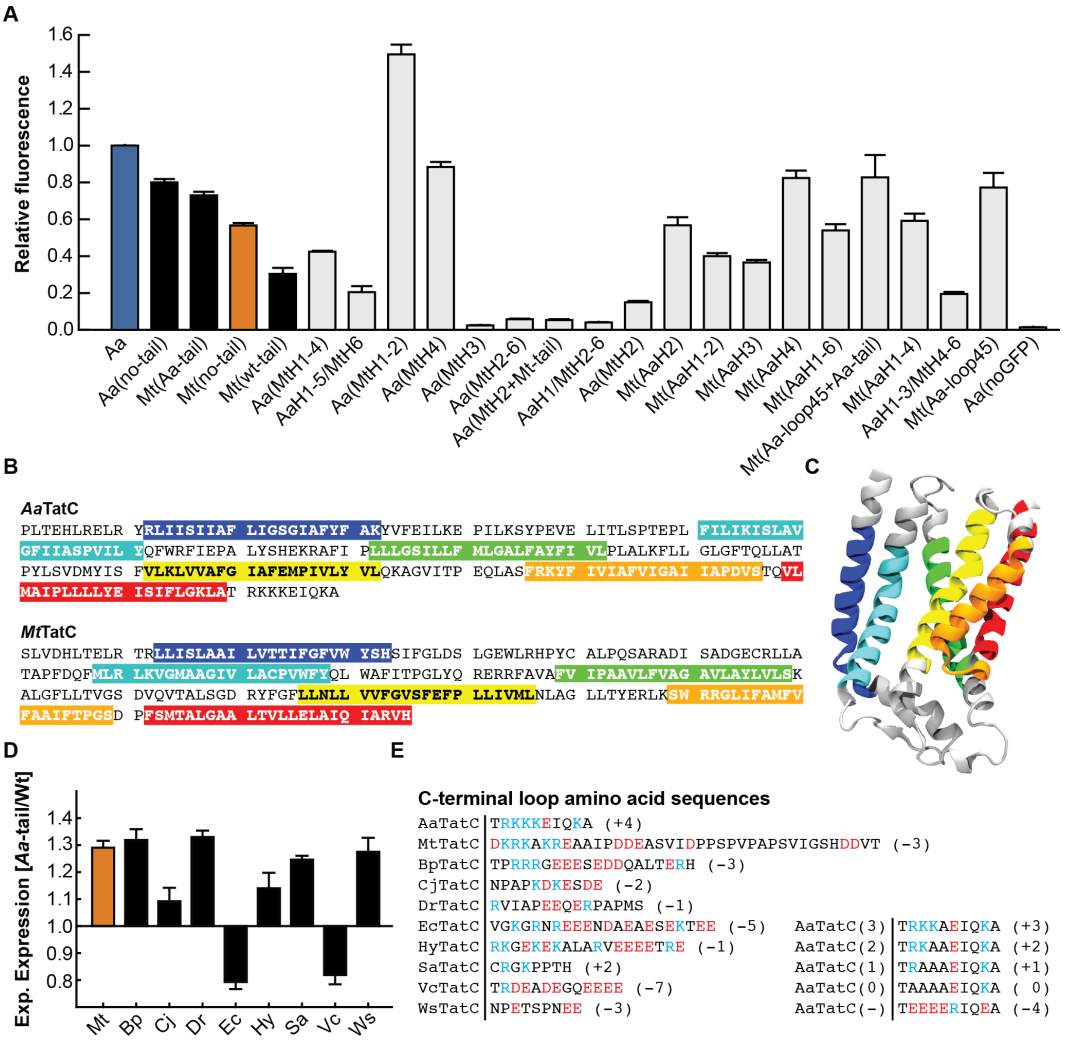

Figure 4.3: Effect of the C-tail on TatC expression in *E. coli*. (**A**) Measured expression levels of the *Aa*TatC and *Mt*TatC chimera proteins, normalized to *Aa*TatC. Shaded bars represent wild-type TatC homologs and mutants with C-tail modifications. (**B**) Domain definitions used in generating the swap chimeras, with TMDs highlighted. (**C**) A ribbons diagram of the structure of *Aa*TatC (RCSB PDB: 4HTS). TMDs are colored according to the highlights used in part (B). (**D**) For each homolog, the ratio of the measured expression level for the *Aa*-tail chimera to that of the corresponding wild-type sequence. (**E**) TatC wild-type and charge mutant C-tail sequences. Positive residues are in blue and negative residues are in red. The net charge is shown to the right of each sequence. Error bars indicate the SEM.

integration efficiency (i.e., the simulated integration efficiency is defined to be the fraction of trajectories that lead to the correct membrane topology) for several TatC sequences. Unless otherwise specified, we define membrane topology in terms of the final orientation of the C-tail; Figure 4.5 confirms that analyzing the

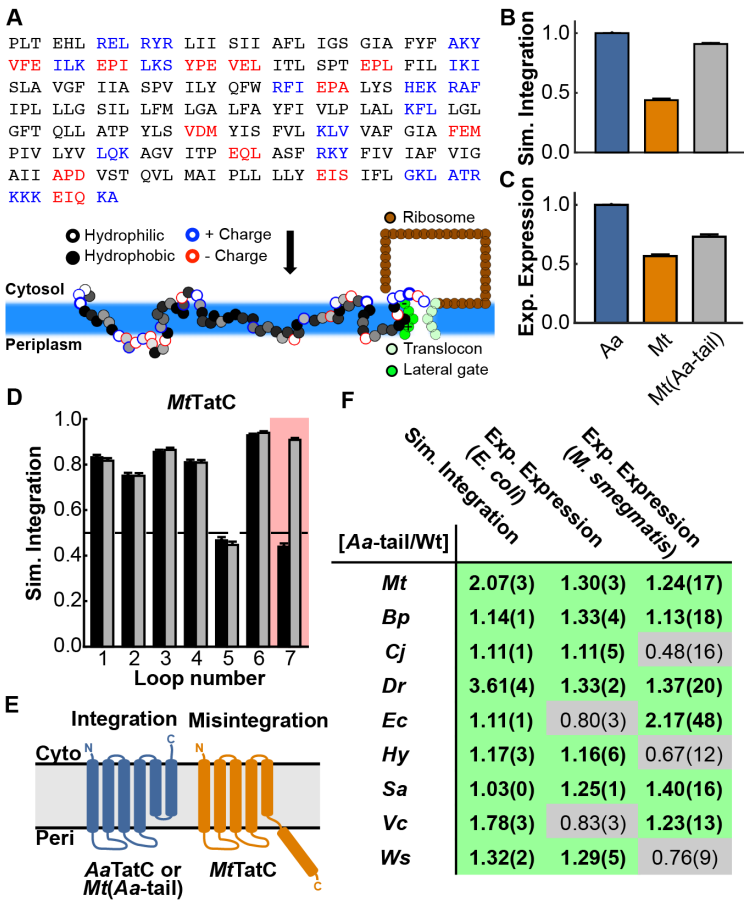

Figure 4.4: Calculation of TatC integration efficiencies. (**A**) Schematic illustration of the CG simulation model that is used to model co-translational IMP membrane integration. The amino-acid sequence of the IMP is mapped onto CG beads, with each consecutive trio of amino-acid residues in the nascent protein sequence mapped to an associated CG bead; the underlying properties of the amino-acid residues determine the interactions of the CG beads, as described in the text. (**B**) Simulated integration efficiency of the *Aa*TatC, *Mt*TatC, and Mt(*Aa*-tail) sequences. (**C**) Experimental expression of the *Aa*TatC, *Mt*TatC, and Mt(*Aa*-tail) sequences. (**D**) The simulated integration efficiency for individual loops of both the wild-type *Mt*TatC sequence (black bars) and the *Aa*-tail swap chimera (grey bars), with loop 7 highlighted. (**E**) Schematic of the correct and incorrect TatC topologies observed in the simulations; misintegration of loop 7 and translocation of TMD 6 leads to an incorrect final topology for *Mt*TatC. Error bars indicate the standard error of mean. (**F**) For each homolog, comparison between the experimental expression levels in *E. coli* and *M. smegmatis* and the simulated integration efficiencies, reporting the ratio of the *Aa*-tail chimera result to that of the corresponding wild-type sequence. Ratios exceeding unity are highlighted in green, indicating enhancement due to the *Aa*-tail. Values in parentheses indicate the SEM.

trajectories in terms of this single-loop definition for membrane topology correlates with defining topology in terms of all loops, while reducing the statistical noise. The *Aa*TatC homolog exhibits significantly higher simulated integration efficiency than the *Mt*TatC homolog, which is consistent with the relative experimental expression levels for the two homologs in Figure 4.4C. Figure 4.4B further shows that the Mt(*Aa*-tail) chimera recovers the high levels of simulated integration efficiency seen for the *Aa*TatC homolog, further mirroring the experimental trends in IMP expression (Figure 4.4C). Figure 4.4D presents an analysis of the orientation of each loop, indicating that only loop 7 is significantly affected swapping the C-tail in the simulations. As is shown schematically in Figure 4.4E, the simulations find that *Mt*TatC exhibits a large fraction of trajectories in which the C-tail resides in the periplasm, such that the C-terminal TMD (TMD 6) fails to correctly integrate into the membrane.

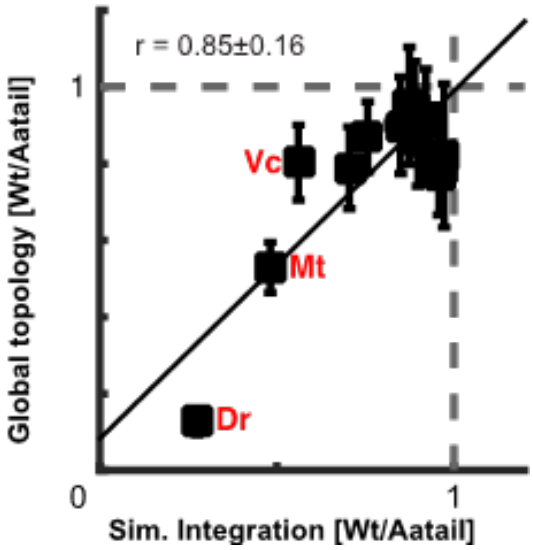

Figure 4.5: Correlation of the simulated integration efficiency calculated using only the loop that was modified (x-axis) versus the simulated integration efficiency calculated using the full multispanning topology (y-axis). Both metrics are highly correlated (Pearson correlation coefficient $r = 0.85$), but use of only the modified loop avoids statistical error due to fluctuations in the topology of the remaining loops. This figure includes data of all the chimeras that were computationally studied, error bars indicate the SEM.

Additional simulations were performed for the full set of the experimentally characterized TatC homologs (Figure 4.6), allowing comparison of the computationally predicted shifts in IMP integration with those observed experimentally for IMP expression. For each homolog, Figure 4.4F compares the effect of swapping the

wild-type C-tail with the *Aa*-tail on both the experimental expression level and the simulated integration efficiency. With the exception of *Vc*TatC and *Ec*TatC, Figure 4.4F shows consistent agreement between the computational and experimental results in *E. coli* upon introducing the *Aa*-tail.

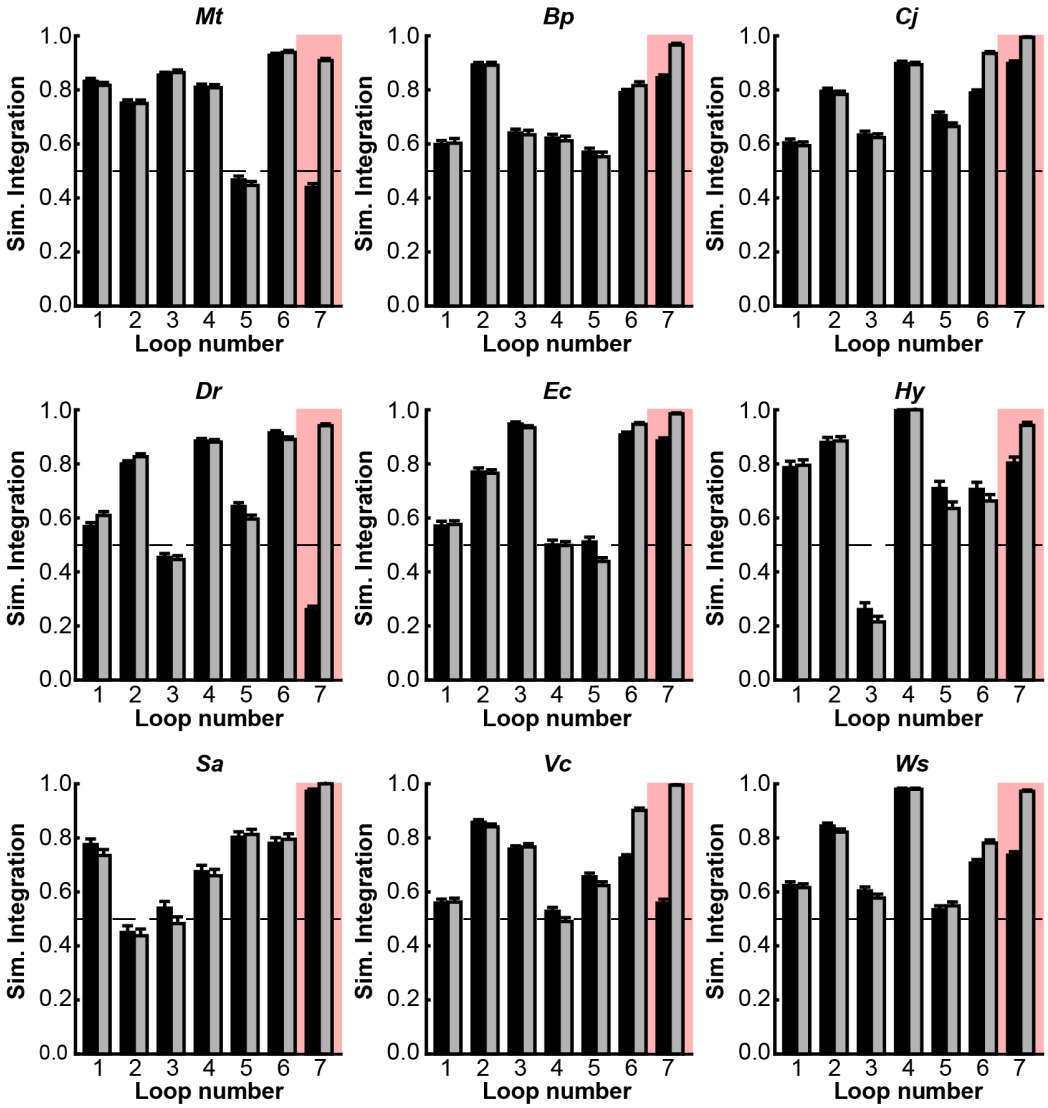

Figure 4.6: For each considered TatC homolog, the simulated integration efficiency for the individual loops for both the wild-type sequence (black bars) and the *Aa*-tail chimeras (grey bars). It is seen that the *Aa*-tail generally leads to a significant effect on the integration efficiency of loop 7 (highlighted), with smaller effects on the other loops. Error bars indicate the SEM.

Experimental confirmation of the simulated integration efficiency changes

The comparison between simulation and experiment in the previous sections suggests a mechanism in which translocation of the C-tail of TatC into the periplasm leads to a reduction in the observed expression level. To validate this, an experimental *in vivo* assay based on antibiotic resistance in *E. coli* is employed. The C-terminal GFP tag was replaced by β-lactamase, such that an incorrectly oriented C-tail would confer increased resistance to β-lactam antibiotics (Figure 4.7A); an inverse correlation between antibiotic resistance and GFP fluorescence is thus expected. *Aa*TatC, Mt, and Mt(*Aa*-tail) constructs containing the β-lactamase tag were expressed using the same protocol as before. Following expression, the cells were diluted to an OD_{600} of 0.1 in fresh media without inducing agent and then grown to an OD_{600} of approximately 0.5 at which point ampicillin was added. 1.5 hours after ampicillin treatment, equal amounts of the media were plated on LB agar plates without ampicillin (Figure 4.7B). The number of observed colonies is used to quantify the relative cell survival (Figure 4.7C, bottom). The survival rate of Mt(*Aa*-tail), Mt, and *Aa*TatC inversely correlates with the simulated integration efficiency of the C-tail (Figure 4.7C), validating the proposed mechanism.

Tail-charge as an expression determinant: Experimental tests of computational predictions

To further establish the connection between the simulated integration efficiencies and the experimentally observed expression levels, we examine the effect of C-tail mutations. We focus on modifications of the C-tail amino-acid sequences that involve the introduction or removal of charged residues, which are known to affect IMP topology and stop-transfer efficiency [42, 114, 146].

We begin by investigating the generic effect of the C-tail charge magnitude on TatC simulated integration efficiency. Figure 4.8A presents the results of coarse-grained simulations in which the magnitude of the charges on the C-tail of the Mt(*Aa*-tail) sequence were scaled by a multiplicative factor, , keeping all other aspects of the protein sequence unchanged. The simulations reveal that reducing the charge magnitude on the C-tail leads to lower simulated integration efficiency.

To examine the corresponding effect of C-tail charge magnitude on expression levels, Figure 4.8B plots the ratio of experimentally observed expression for each wild-type homolog relative to its corresponding *Aa*-tail swap chimera versus the total charge magnitude on the wild-type C-tail. Without exception in these data, the expression of

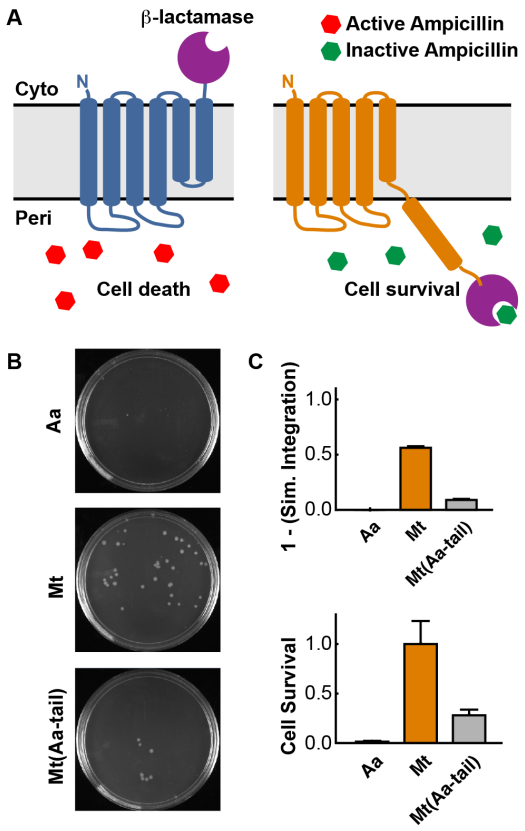

Figure 4.7: Correlation of antibiotic resistance to membrane topology. (**A**) Schematic of the cytoplasmic and periplasmic topologies of the TatC C-tail with the fused β-lactamase enzyme. Misintegration of loop 7 leads to periplasmic localization of the β-lactamase, resulting in enhanced antibiotic resistance and cell survival. (**B**) Representative plates from the ampicillin survival test. (**C**) Comparison of the simulated integration efficiency (top) and relative ampicillin survival rate (bottom) for *Aa*TatC, *Mt*TatC, and Mt(*Aa*-tail). The reported cell survival corresponds to the ratio of counted cells post-treatment versus prior to treatment with ampicillin; all values are reported relative to *Mt*TatC. Error bars indicate the SEM.

wild-type homologs with weakly charged C-tails (relative to the *Aa*-tail) is improved upon swapping with the *Aa*-tail, whereas the expression of homologs with strongly charged C-tails is reduced upon swapping with the *Aa*-tail (i.e., all data points in Figure 4.8B fall into the unshaded quadrants).

Figure 4.8C further illustrates the effect of charge magnitude on expression by presenting the experimentally observed expression levels for *Aa*-tail(-) swap chimeras, in which the introduced C-tail sequence preserves the charge magnitude of the *Aa*-tail sequence while reversing the net charge (see Figure 4.3E for the C-tail sequences). Despite the complete reversal of the C-tail charge, the observed correlation between

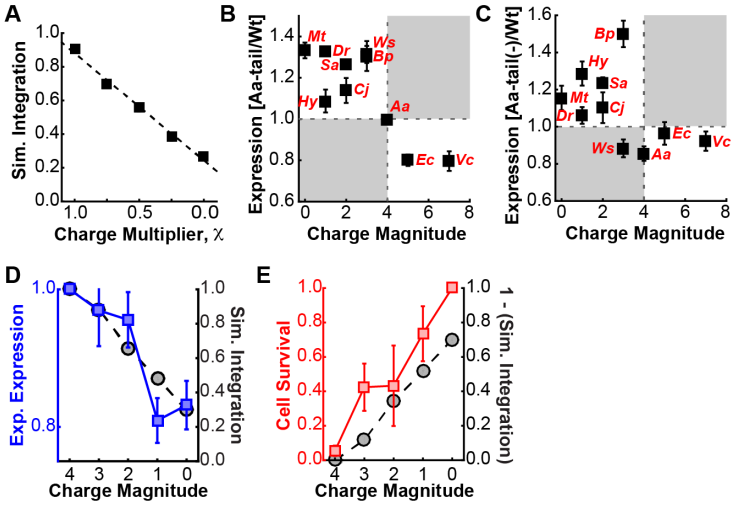

Figure 4.8: Mechanistic basis associated with charged C-tail residues. (**A**) Simulated integration efficiency of the Mt(*Aa*-tail) chimera, as a function of scaling the charges of the C-tail residues. (**B**) Correlation of the ratio of the measured expression for the *Aa*-tail swap chimeras to that of the corresponding wild-type sequence versus the charge magnitude of the wild-type C-tail (data from Figure 4.3B and Figure 4.3E). Linear regression yields a fit of $r = -0.8 \pm 0.2$. (**C**) Correlation of the ratio of the measured expression for the *Aa*-tail(-) swap chimeras to that of the corresponding wild-type sequence versus the charge magnitude of the wild-type C-tail, where the *Aa*-tail(-) swap chimeras include a variant of the *Aa*-tail with net negative charge and the same overall charge magnitude. (**D**) Experimental expression levels in *E. coli* (blue, left axis) and simulated integration efficiency (black, right axis) for a series of mutants of the Mt(*Aa*-tail) sequence, in which positively charged residues in the *Aa*-tail are mutated to alanine residues. Reported values are normalized to Mt(*Aa*-tail). (**E**) Relative ampicillin survival rate in *E. coli* (red, left axis) and simulated integration efficiency (black, right axis) for a series of mutants of the Mt(*Aa*-tail) sequence, in which positively charged residues in the *Aa*-tail are mutated to alanine residues. Simulation results are normalized as in part (D), while ampicillin survival is normalized to the highest survival rate (i.e., with zero charge magnitude). Error bars indicate the SEM.

expression and C-tail charge magnitude for these two sets of chimeras is strikingly similar (compare Figures 4.8B and C).

Finally, we considered a series of mutants of the Mt(*Aa*-tail) chimera, in which the charge magnitude of the *Aa*-tail is reduced by mutating positively charged residues to alanine residues (see Figure 4.3E for the C-tail sequences). For this series of mutants, Figure 4.8D (black) shows that the simulated integration efficiency decreases with the charge of the C-tail, which predicts a corresponding decrease in the experimental expression levels; indeed, the subsequent experimental measure-

ments confirm the predicted trend (Figure 4.8D, blue). Again using the antibiotic resistance assay to validate the connection between simulated integration efficiency and observed expression, Figure 4.8E confirms that the simulation results correlate with the relative survival of the Mt(*Aa*-tail) alanine mutants with a β-lactamase tag (Figure 4.8E, red). In addition to providing evidence for the connection between simulated integration efficiency and observed expression levels, the results in Figure 4.8 suggest that this link can be used to control IMP expression.

Transferability to expression systems other than *E. coli*

Beyond the *E. coli* overexpression host, we now examine the transferability of the relation between simulated integration efficiency and experimental expression levels. We employ *Mycobacterium smegmatis*, a genetically tractable model organism that is phylogenetically distinct from *E. coli*. All coding sequences were transferred into an inducible *M. smegmatis* vector, including the linker and C-terminal GFP, and expressed; expression levels were then measured by flow cytometry and validated by western blot.

Figure 4.9A demonstrates that, as in *E. coli*, the experimentally observed expression levels vary widely among the wild-type TatC homologs in *M. smegmatis*. However, comparison of Figure 4.9A with Figure 4.1B reveals that the total expression levels for the homologs in *M. smegmatis* are different from those seen in *E. coli*, although for both systems, the *Aa*TatC homolog expresses strongly and *Mt*TatC expresses poorly (which is perhaps surprising, given the close evolutionary link between *M. smegmatis* and M. tuberculosis). Figure 4.4F also shows that replacing the wild-type C-tail with the *Aa*-tail in *M. smegmatis* generally increases the experimentally observed expression levels, in general agreement (six out of nine homologs) with the previously discussed simulated integration efficiency results.

Figure 4.4F further shows that the subset of homologs for which the *Aa*-tail swap chimeras led to increased levels of expression in *M. smegmatis* is overlapping but different from the subset associated with the *E. coli* results. This emphasizes that although the computed levels of simulated integration efficiency agree with the observed changes in expression levels in both expression systems, the observed expression levels depend upon the expression system, while the simulated integration efficiencies calculated using the current implementation of the coarse-grained model are independent of the expression system. In short, simulated integration efficiency is a predictor of the expression levels in both systems, but it is not the only factor

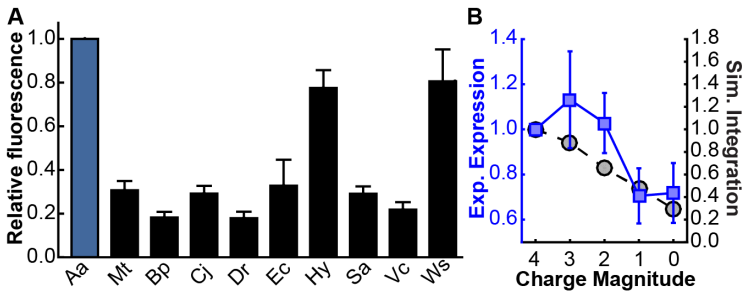

Figure 4.9: *M. smegmatis* expression tests. (**A**) Expression levels of various TatC homologs in *M. smegmatis*, measured by TatC-GFP fluorescence, with expression levels normalized to *Aa*TatC (blue). (**B**) Simulated integration efficiency (blue, left axis) and measured expression levels in *M. smegmatis* (black, right axis) for a series of mutants of the Mt(*Aa*-tail) sequence, in which positively charged residues in the *Aa*-tail are mutated to alanine residues. Error bars indicate the SEM.

contributing to the observed expression levels.

Continuing with the *M. smegmatis* expression system, Figure 4.9B repeats the comparison between the simulated integration efficiency and the observed expression levels for the series of mutants of the Mt(*Aa*-tail) chimera, in which the positive charge of the *Aa*-tail is reduced by mutating positively charged residues to alanine residues. The simulated integration efficiencies, identical to those in Figure 4.8D, are predicted to decrease as charges are removed. The experimental expression levels for *M. smegmatis* in Figure 4.9B likewise show a decrease. Taken together, the results obtained for the *M. smegmatis* expression system suggest that the connection between simulated integration efficiency and observed expression levels may be generalizable beyond *E. coli*.

Predicting the effect of mutations other than in the C-tail

As seen in Figure 4.4D, the coarse-grained simulations predict poor integration efficiency for loop 5, suggesting an additional location (beyond the C-tail, loop 7) in the *Mt*TatC sequence that can be optimized for expression. Figure 4.10A presents the simulated integration efficiency for loop 5 in each of the TatC homologs, revealing a significant range of efficiencies. Selecting the four homologs with the highest predicted simulated integration efficiency for loop 5 (Sa, Hy, Cj, and Vc), chimera proteins were derived from the *Mt*TatC sequence by swapping loop 5 of *Mt*TatC with the corresponding loop 5 sequence from each of these homologs (Figure 4.10B). Figure 4.10C compares the simulated integration efficiency and experimentally observed expression level for each chimera, revealing agreement for

three out of four cases. Comparing the simulation results in Figure 4.10, note that the degree of improvement for the simulated integration efficiency obtained from the coarse-grained simulations of the chimeras (Figure 4.10C) is different from that anticipated by naïve comparison of the individual loops in the wild-type sequences (Figure 4.10A); this emphasizes that the simulated integration efficiency is sensitive to elements of the IMP sequence beyond the local segment that is being swapped. The results in Figure 4.10 suggest the simulated integration efficiency can be used to identify regions beyond the TatC C-tail for modification to improve experimental expression; more generally, it suggests the potential for identifying local segments of an IMP amino-acid sequence that may be modified to yield increased experimental expression.

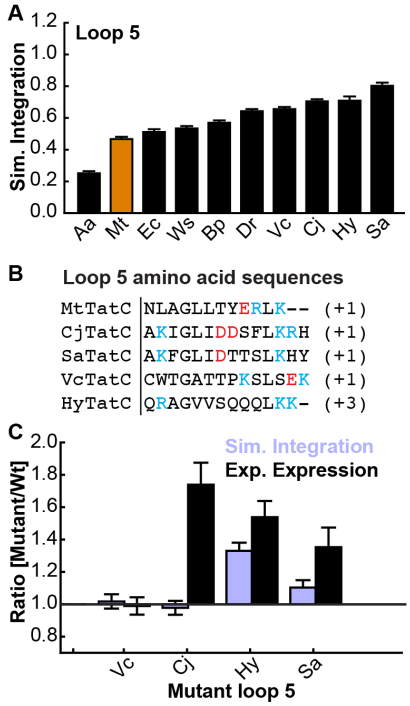

Figure 4.10: Loop 5 analysis for *Mt*TatC. (**A**) Simulated integration efficiency of loop 5 for the TatC homologs. (**B**) Loop 5 amino-acid sequence for various TatC homologs. (**C**) Experimental expression (black) and simulated integration efficiency (purple) for the loop 5 swap chimeras of *Mt*TatC in which the entire loop 5 sequence of wild-type *Mt*TatC is replaced with the corresponding sequence of other homologs. Error bars indicate the SEM.

4.3 Discussion

The mechanistic picture that emerges from the experimental and theoretical analysis of the TatC IMP family is that the efficiency of Sec-facilitated membrane integra-

tion, which is impacted by the IMP amino-acid sequence, is a key determinant in the degree of observed protein expression. We observed that TatC homologs had varying levels of expression (Figures 4.1B and 4.9A). Swap chimeras between *Aa*TatC and *Mt*TatC revealed a significant effect of the C-tail in determining expression yields (Figure 4.3A), with the *Aa*-tail having a largely positive effect that was transferrable to other homologs (Figure 4.4F). Coarse-grained modeling predicted a large, sequence-dependent variation of the simulated integration efficiency for the C-tail (Figure 4.4), suggesting the underlying mechanism by which the *Aa*-tail enhances the expression of other TatC homologs. Validation of this mechanism was experimentally demonstrated using an antibiotic-resistance assay (Figure 4.7). Additional point-charge mutations in the C-tail were shown to change the simulated integration efficiency, which in turn predicted changes in the IMP expression levels according to the proposed mechanism; these predictions were experimentally confirmed in both *E. coli* (Figure 4.8) and *M. smegmatis* (Figure 4.9). Finally, the link between simulated integration efficiency and experimental expression was exploited to design *Mt*TatC chimeras with improved expression based on the loop 5 simulated integration efficiency (Figure 4.10).

The observed correlation between IMP integration efficiency and observed expression levels presented here is consistent with earlier observations that expression can be modulated by mutations of the sequence [46, 106, 137], as well as recent work in which mis-integrated dual-topology IMPs are shown to be degraded by FtsH [142]. However, these earlier studies did not provide a clear mechanistic basis for the relation between IMP sequence modifications and observed expression levels. In the current work, we demonstrate the relation between integration efficiency and observed expression levels, and we demonstrate a tractable coarse-grained approach for computing the simulated integration efficiency and its changes upon sequence modifications. This work also raises the possibility of using simulated integration efficiencies to optimize experimental expression levels, which has been demonstrated here via the computational prediction and subsequent experimental validation of individual charge mutations in the C-tail and of loop-5 swap chimeras.

A few comments are worthwhile with regard to the scope of the conclusions drawn here. Firstly, the current work focused on comparing protein expression levels among IMP sequences that involve relatively localized changes, such as single mutations or loop swap chimeras, as opposed to predicting relative expression levels among dramatically different IMP sequences. Secondly, the current work

examines experimental conditions for the overexpression of IMPs using the same plasmids, which may be expected to isolate the role of membrane integration in determining the relative expression levels of closely related IMP sequences. The prediction of expression levels among IMPs that involve more dramatic differences in sequence may well require the consideration of other factors, beyond just the simulated integration efficiency. Moving forward, we expect that a useful strategy will be to systematically combine the simulated IMP integration efficiency with other sequence-based properties to predict IMP expression levels [21].

The experimental and computational tools used here are readily applicable to many systems, potentially aiding the understanding and enhancement of IMP expression in many other systems, as well as providing fundamental tools for the investigation of co-translational IMP folding. By demonstrating inexpensive *in silico* methods for predicting protein expression, we note the potential for computationally guided protein expression strategies to significantly impact the isolation and characterization of many IMPs.

4.4 Methods

Overview of the experimental procedures

Briefly, for *E. coli* all expression plasmids were derived from pET28(a+)-GFP-ccdB, with the final expressed sequences containing a Met-Gly N-terminus followed by the IMP sequence, a tobacco etch virus (TEV) protease site, a GFP variant [135], and an eight His tag. For β-lactamase constructs, the GFP sequence was replaced by a β-lactamase sequence. For *M. smegmatis* expression plasmids, the entire coding region of the TatC homologs were sub-cloned and transferred into pMyNT vector [90]. *E. coli* constructs were grown in BL21 Gold (DE3) cells (Agilent Technologies) at 16 °C and induced with 1 *m*M IPTG at an OD_{600} of 0.3 then analyzed after 16 hours. *M. smegmatis* constructs were grown in mc^2 155 cells (ATCC) at 37 °C and induced at an OD_{600} of 0.5 with 0.2% acetamide then analyzed after six hours. A 200 μL sample of each expression culture was pelleted and resuspended in 1 *m*L of PBS. Whole-cell GFP fluorescence was measured using a MACSQuant10 Analyzer (Miltenyi). For the ampicillin survival assay, the cells were diluted to an OD_{600} of 0.1 in fresh media following expression without inducing agent and then grown to an OD_{600} of approximately 0.5 at which point ampicillin was added. 1.5 hours after ampicillin treatment, equal amounts of the media were plated on LB agar plates without ampicillin. The number of observed colonies is used to quantify the relative cell survival. The following sections provide additional details on the experimental

assays used in this chapter.

Designing and cloning of TatC chimeras

The parent plasmid used for cloning, pET28(a+)-GFP-ccdB, was derived from an IMP-GFP vector used in Ref. [28]. TatC homologs and chimeras were prepared from genomic DNA, with the exception of wild-type M. tuberculosis and A. aeolicus TatC genes which were synthesized by primer extension as applied in DNAWorks (NIH) [58]. In most cases, the Gibson assembly cloning protocol was used for cloning [39]. Expression of a vector containing *Aa*TatC with an N-terminal ten-His tag and without the GFP fusion-tag was used as a negative control for in-gel fluorescence, western blot analysis, and flow cytometry. For constructs containing the β-lactamase tag, the GFP sequence was removed and replaced with a β-lactamase sequence using Gibson cloning. For generation of *M. smegmatis* compatible plasmids, the entire coding region of the TatC homologs including the entire GFP sequence and the poly-His tag were PCR amplified out of their respective pET28(a+)-GFP-ccdB vector using primers with compatible regions for placement into the pMyNT vector using Gibson assembly[90]. For β-lactamase constructs, the GFP sequence was replaced by a β-lactamase sequence using Gibson assembly.

Expression in *E. coli*

Plasmids were transformed into BL21 Gold (DE3) cells and transferred onto LB agar plates containing 50 μg/mL kanamycin plates after one-hour incubation. After overnight incubation at 37 °C, colonies were scraped off the plates into 5 mL of LB, resuspended, and the OD_{600} was determined. These samples were then diluted into 50 mL 2xYT containing 50 μg/mL kanamycin in 125 mL baffled flasks to a starting OD_{600} of approximately 0.01. Cultures were grown in an orbital shaker at 37 °C until they reached an OD_{600} of 0.15. The temperature of the orbital shaker was then reduced to 16 °C. Upon reaching an OD_{600} of 0.3, IPTG was added to final concentration of 1mM to induce expression. Cultures were grown for a further 16 hours prior to analysis.

β-lactamase survival test

Plasmids containing the β-lactamase tag were expressed overnight at 16 °C as previously described. Cells from each overnight culture were washed with phosphate buffered saline (PBS) to remove IPTG and then diluted into fresh 50 mL 2xYT

media containing 50 μg/mL kanamycin to a starting OD_{600} of 0.1 in 125 mL baffled flasks. Cultures were grown at 37 °C to an OD_{600} of approximately 0.5 where a control sample from each culture was taken, diluted 10,000 times in PBS, and 50 μL was plated onto LB agar plates containing 50 μg/mL kanamycin. To each culture, 50 μg/mL ampicillin was added and shaken at 37 °C for a further 90 minutes. A sample from each culture was taken, diluted 200 times in PBS, and 50 μL was plated onto LB agar plates containing 50 μg/mL kanamycin. Plates were grown overnight ($\approx$16 hours) and the number of colonies on each plate was counted. Colony counts from the second plating were normalized by the colony counts from the first plating to account for variation in the OD_{600} at which ampicillin was added to determine relative survival. The procedure was performed in triplicate and standard errors of normalized values were calculated. For each plot of relative survival, the values are normalized to the highest survival rate of the samples in the figure.

Expression in *M. smegmatis*

For *M. smegmatis* overexpression, constructs were transformed into mc^2 155 cells using electroporation and transferred onto Middlebrook 7H11 plates (10.25 g Middlebrook 7H11 Agar Base, 1 vial ADC growth supplement, 2.5 g glycerol, 1 mM CaCl2, 50 μg/mL carbenicillin, 10 μg/mL cyclohexamide, 50 μg/mL hygromycin, and water to 500 mL) after a three hour incubation in 1 mL Middlebrook 7H9 culture media (2.35 g Middlebrook 7H9 Broth Base, 1 vial ADC growth supplement, 0.5 g Tween-80, 1 mM CaCl2, 50 μg/mL carbenicillin, 10 μg/mL cyclohexamide, and water to 500 mL). Plates were grown for three to four days until colonies formed. Single colonies were picked into 5 mL Middlebrook 7H9 culture media containing 50 μg/mL hygromycin. The following day, 50 mL cultures of Middlebrook 7H9 expression media (2.35 g Middlebrook 7H9 Broth Base, 0.25 g Tween-80, 1 g glycerol, 1 g glucose, 1 mM CaCl2, 50 μg/mL carbenicillin, 10 μg/mL cyclohexamide, 50 μg/mL hygromycin, and water to 500 mL) were inoculated at a starting OD_{600} of 0.005. Cultures were grown at 37 °C and expression was induced with 0.2% acetamide at an OD_{600} of 0.5. Cultures were grown for six hours after induction prior to analysis.

Flow cytometry

A 200 μL sample of each expression culture was centrifuged at 4000g for 3 minutes to pellet the cells and then the supernatant was removed. Cells were resuspended in

1 mL of PBS and 200 μL of each were dispensed into 96-well plates and kept on ice for analysis. Whole-cell GFP fluorescence was determined using a MACSQuant10 Analyzer. Forward scattering, side scattering, and total fluorescence at 488 nm were considered during analysis. Measured events were gated based on the negative control sample to contain the lowest 90% of both forward and side scattering values to remove anomalous particles, such as dead or clumped cells. Mean cell fluorescence was calculated for the gated population as a measure of folded TatC. At least four independent expression trials were performed for each sequence tested to ascertain expression variance. Flow cytometry data analysis was performed with FlowJo Software. Flow cytometry data is normalized to a standard for each day data was collected. For example, for "*Aa*-tail/wild-type" data points, the mean fluorescence of the *Aa*-tail swap chimeras were normalized by the mean fluorescence of their respective homologs containing the wild-type tail for that day's trial. Similarly, for relative fluorescence data points in which wild-type *Aa*TatC was the standard, the mean fluorescence of each sample was normalized by the mean fluorescence of the *Aa*TatC sample for that day's trial. In both cases, final calculated values are averages of the normalized values over at least four trials with error bars representing standard errors of the mean for those normalized values.

In-gel fluorescence and western blot analyses

In-gel fluorescence and western blot analyses were used as an alternative measure of total expressed proteins. 5 mL of expression samples were centrifuged and supernatant discarded. Samples were resuspended to an OD_{600} of 3.0 in PBS. 1 mL of each sample was collected and 250 μL lysis buffer (375 mM Tris-HCl pH 6.8, 6% SDS, 48% glycerol, 9% 2 Mercaptoethanol, 0.03% bromophenol blue) was added. Samples were lysed via freeze fracturing by three rounds of freezing using liquid nitrogen and thawing using room temperature water. 20 μL of each lysed sample was subjected to SDS-PAGE. SDS-PAGE gels were imaged for fluorescence using a UV gel imager with a filter for GFP fluorescence to determine in-gel fluorescence. For western blot analysis, the samples were transferred from the gel onto a nitrocellulose membrane using the Trans-Blot Turbo System. The membranes were washed three times with 15 mL TTBS (50 mM Tris pH 7.6, 150 mM NaCl, 0.05% Tween-20), incubated one hour with 15 mL 5% milk powder in TTBS, washed three times with 15 mL of TTBS, and then incubated with 1:5000 anti-GFP Mouse primary antibody (EMD Millipore, Lot # 2483215) in 15 mL 5% milk powder in TTBS overnight. Membranes were washed three times with 15 mL TTBS, incubated with 1:15000

IRDye 800CW Donkey anti-Mouse secondary antibody (LI-COR, Lot # C31024-04) in 15 mL 5% milk powder in TTBS for one hour, washed three times with 15 mL TTBS, and then visualized using a Licor IR western blot scanner. ImageJ was used to process the images.

Overview of the CG Model

Modeling of IMP integration in the current study was performed using a previously developed coarse-grained (CG) method for the direct simulation of co-translational protein translocation and membrane integration [146]. Ribosomal translation and membrane integration of nascent proteins are thus simulated on the minute timescale, enabling direct comparison between theory and experiment. The CG model was previously parameterized using extensive molecular dynamics simulations of the translocon and nascent protein in explicit lipid and water environment [145, 147]. The CG model has been validated against available experimental data and shown to correctly capture effects related to nascent protein charge, hydrophobicity, length, and translation rate in both IMP integration and protein translocation studies [129, 146].

The CG model is employed with only minor modifications from [146], all of which are specified below. Key features of the CG model and its implementation are provided here; for a full discussion of the CG model, the reader is referred to [146].

As described in [146], the CG model explicitly describes the configurational dynamics of the nascent-protein chain, conformational gating in the Sec translocon, and the slow dynamics of ribosomal translation. The nascent chain is represented as a freely jointed chain of beads, where each bead represents three amino acids and has a diameter of 8Å, the typical Kuhn length for polypeptide chains [52, 119]. Bonding interactions between neighboring beads are described using the finite extension nonlinear elastic (FENE) potential [72], short-ranged nonbonding interactions are modeled using the Lennard-Jones potential, electrostatic interactions are modeled using the Debye-Hückel potential, periplasmic binding is included as described in (Zhang and Miller, 2012) for BiP, and solvent interactions are described using a position-dependent potential based on the water-membrane transfer free energy for each CG bead; all parameters are the same as used previously [146], unless otherwise stated. The time evolution of the nascent protein is modeled using overdamped Langevin dynamics, with the CG beads confined to a two-dimensional subspace that runs along the axis of the translocon channel and between the two helices of the

lateral gate (LG). Conformational gating of the translocon LG corresponds to the LG helices moving out of the plane of confinement for the CG beads, allowing the nascent chain to pass into the membrane bilayer. The rate of stochastic LG opening and closing is dependent on the sequence of the nascent protein CG beads that occupy the translocon channel. Ribosomal translation is directly simulated via growth of the nascent protein at the ribosome exit channel; throughout translation, the C-terminus of the nascent protein is held fixed, and new beads are sequentially added at a rate of 24 residues per second. Upon completion of translation, the C-terminus is released from the ribosome. It has been confirmed that the results presented in the current study are robust with respect to changes in the rate of ribosomal translation (Pearson correlation coefficient between Wt/*Aa*-tail ratios obtained using a rate of translation of 24 residues/sec and 6 residues/sec, $r = 0.99 \pm 0.06$).

The CG simulation model: Implementation details

Two changes to the protocol for the CG simulation model were introduced in the current study, with respect to the protocol used in [146]. These modifications were included to remove unphysical artifacts in the simulations, although it is emphasized that conclusions in the main text are qualitatively unchanged by these modifications (Pearson correlation coefficient between Wt/*Aa*-tail ratios obtained with and without the modifications to the simulation protocol, $r = 0.97 \pm 0.09$).

The first change in the CG model is that the ribosome is assumed to remain associated with the translocon following translation of the nascent protein. In the previously implementation of the model, the ribosome was assumed to dissociate from the translocon immediately following stop-translation, which was found in the current study to lead to artifacts for nascent proteins with extremely short C-terminal domains. Furthermore, this modification is consistent with experimental evidence that indicates that the timescale for ribosomal dissociation is slower than the trajectories simulated here [95, 107].

The second change in the CG model relates to the potential energy cost of flipping hydrophilic nascent-protein loops across the lipid membrane at significant distances from the translocon. The Wimley-White water-octanol transfer free energy scale [141] that was used to parameterize the interactions of the CG beads with the membrane is appropriate for describing the transfer of amino acids between an aqueous region and either the phospholipid interface or the region of the membrane interior that is close to the translocon lateral gate [79]. However, the flipping

of hydrophilic nascent-protein loops across the membrane at significant distances from the translocon involves moving CG beads through the hydrophilic core of the membrane interior, which will incur a large potential energy barrier [79]. To account for this effect, and to avoid unphysical flipping of short hydrophilic loops across the membrane, an additional potential energy term was included in potential energy function that describes the interactions between the CG beads and the membrane,

$$U_{core}(x, y) = gS(x; \phi_x, \psi_x)[1 - S(y; \phi_y, \psi_y)], \tag{4.1}$$

$$S(x; \phi_x, \psi_x) = \frac{1}{4}\left(1 + \tanh\frac{x - \phi_x}{b}\right)\left(1 + \tanh\frac{x - \psi_x}{b}\right), \tag{4.2}$$

where $\phi_x = -1\sigma$, $\psi_x = 1\sigma$, $\phi_y = -2.5\sigma$, $\psi_y = 2.5\sigma$, and $b = 0.25\sigma$. The parameters σ and g are respectively the diameter of the CG beads and the water-octanol transfer free energy for the CG beads, both of which appear in the original model. We emphasize that this new term has no noticeable effect on the potential energy function for the CG beads at distances within 8Åto the translocon channel; it simply affects unphysical flipping of the TMD domains across the membrane at larger distances from the channel. This artifact was not observed in the earlier study using the CG model, since only processes involving the translocation or membrane integration of a single TMD domain were considered.

Mapping IMP amino-acid sequences to the CG model

In the current study, amino-acid sequences for the TatC homologs are mapped onto sequences of CG beads as follows. Each consecutive trio of amino acid residues in the nascent protein sequence is mapped to an associated CG bead. The water-membrane transfer free energy for each CG bead is taken to be the sum of the contributions from the individual amino acids; these values are taken from the experimental water-octanol transfer free energies for single residues [141]. The charge for each CG bead is taken to be the sum of the contribution from the individual amino acids. As in [146], positively charged residues (arginine and lysine) were modeled with a +2 charge to capture significant effects on topology due to changes in the nascent protein sequence. Histidine residues were modeled with a +1 charge to account for the partial protonation of these residues, and negatively charged residues (glutamate and aspartate) were modeled with a charge of -1. The mapping procedure for *Aa*TatC is depicted in Figure 4.4A as an example.

In the *Mt*TatC chimeras where loop 5 was replaced (Figure 4.10), the mapping protocol was modified to avoid a frame-shift in the three-to-one mapping of amino

acids. Specifically, prior to mapping amino acids to beads as described previously, 0,1 or 2 dummy amino acids were added to the sequence immediately following loop 5. The number of dummy amino acids was chosen such that the amino acid to bead mapping was identical to that of *Mt*TatC wildtype for TMD 6 onwards, avoiding a frame-shift. Dummy amino acids have zero charge and zero water-membrane transfer free energy.

CG simulation model: Calculation details

For the results presented in this chapter other than in Figure 4.10, the co-translational membrane integration for each TatC sequence is simulated using 1200 independent CG trajectories; for the results in Figure 4.10, each sequence is simulated using over 400 independent trajectories. As in [146], each CG trajectory is performed with a timestep of 100 ns. All trajectories were terminated 30 seconds after the end of translation for the protein sequence.

Analysis of CG simulation results

To determine whether a given trajectory leads to integration in the correct multispanning topology, the topology of a nascent protein configuration can be characterized by the location of the soluble loops that connect the TMD. We thus specify a collective variable λ_i associated with each loop, with $i = 1$ corresponding to the loop that leads TMD 1 in the sequence (i.e. the N-terminal sequence) and $i = 7$ corresponding to the loop that follows TMD 6 (i.e. the C-tail). If loop i is in the cytosol, then $\lambda_i = 1$; if loop i is in the periplasm, then $\lambda_i = -1$; otherwise, $\lambda_i = 0$. Whether a given loop is in the cytosol, in the membrane, or in the periplasm is determined by the tracking position of a representative bead in that loop (Table 4.1). Representative beads were chosen based on having the lowest probability of being inside the lipid region compared to other beads in that loop. A given trajectory is determined to have reached correct IMP integration ($\lambda_i = -1$ for periplasmic loops and, $\lambda_i = 1$ for cytosolic loops) if a configuration with the loops in the correct orientation is sampled during a time window of 6 seconds taken 25 seconds after the end of translation; the time window of 25 seconds was found sufficient to allow the nascent protein to finish the integration/translocation of TMD 6.

Figure 4.6 shows the fraction of trajectories that exhibit the correct topology for each individual loop for all TatC homologs and chimeras considered in this study. It is clear from Figure 4.6 that the changes to the amino-acid sequence considered

	Loop 1	Loop 2	Loop 3	Loop 4	Loop 5	Loop 6	Loop 7
*Aa*TatC	7-9	43-45	88-90	145-147	181-183	202-204	238-239
*Mt*TatC	7-9	61-63	112-114	151-153	193-195	220-222	244-246
Mt(*Aa*-tail)	7-9	61-63	112-114	151-153	193-195	220-222	244-246
*Bp*TatC	25-27	64-66	112-114	160-162	196-198	220-222	253-255
Bp(*Aa*-tail)	25-27	64-66	112-114	160-162	196-198	220-222	250-252
*Cj*TatC	13-15	55-57	100-102	139-141	187-189	208-210	238-240
Cj(*Aa*-tail)	13-15	55-57	100-102	139-141	187-189	208-210	238-240
*Dr*TatC	28-30	73-75	118-120	166-168	202-204	229-231	262-264
Dr(*Aa*-tail)	28-30	73-75	118-120	166-168	202-204	229-231	247-249
*Ec*TatC	10-12	55-57	103-105	142-144	190-192	211-213	244-246
Ec(*Aa*-tail)	10-12	55-57	103-105	142-144	190-192	211-213	244-246
*Hy*TatC	7-9	40-42	94-96	139-141	184-186	205-207	232-234
Hy(*Aa*-tail)	7-9	40-42	94-96	139-141	184-186	205-207	232-234
*Sa*TatC	7-9	43-45	91-93	142-144	178-180	199-201	229-231
Sa(*Aa*-tail)	7-9	43-45	91-93	142-144	178-180	199-201	229-231
*Vc*TatC	16-18	52-54	103-105	145-147	190-192	211-213	247-249
Vc(*Aa*-tail)	16-18	52-54	103-105	145-147	190-192	211-213	241-243
*Ws*TatC	10-12	61-63	97-99	148-150	181-183	205-207	241
Ws(*Aa*-tail)	10-12	61-63	97-99	148-150	181-183	205-207	235-237

Table 4.1: Loop definitions used in simulation trajectory analysis. Each loop is specified in terms of the amino-acid residue sequence numbers (end-points inclusive) associated with the wild-type sequence.

in this study largely only impact the topology of the domain where the changes to the amino acid sequence were introduced; the topology of the rest of the protein is not predicted by the CG simulation model to be significantly affected by the sequence changes. The calculated results are robust with respect to the details of the definition of simulated integration efficiency (Pearson correlation coefficient between Wt/Mutant ratios obtained analyzing only the loop that was modified and those obtained analyzing all loops, $r = 0.85 \pm 0.16$) (Figure 4.5); to minimize statistical error, for all simulation results presented in this chapter, the topology of the IMP is thus characterized in terms of only the loop of interest.

Chapter 5

IMPROVING MEMBRANE PROTEIN EXPRESSION BY OPTIMIZING INTEGRATION EFFICIENCY

Adapted from:

Niesen, M. J. M*., Marshall, S. S*., Miller, T. F. M, Clemons, W. M. C. (2016). "Improving membrane protein expression by optimizing integration efficiency". In: *J. Biol. Chem.* 292(47): 19537-19545. DOI: 10.1074/jbc.M117.813469. (*) Equal contribution.

The heterologous overexpression of integral membrane proteins in *Escherichia coli* often yields insufficient quantities of purifiable protein for applications of interest. The current study leverages a recently demonstrated link between co-translational membrane integration efficiency and protein expression levels to predict protein sequence modifications that improve expression. Membrane integration efficiencies, obtained using a coarse-grained simulation approach, robustly predicted effects on expression of the integral membrane protein TatC for a set of 140 sequence modifications, including loop-swap chimeras and single-residue mutations distributed throughout the protein sequence. Mutations that improve simulated integration efficiency were four-fold enriched with respect to improved experimentally observed expression levels. Furthermore, the effect of double mutations, on both simulated integration efficiency and experimentally observed expression levels were cumulative and largely independent, suggesting that multiple mutations can be introduced to yield higher levels of purifiable protein. This work provides a foundation for a general method for the rational overexpression of integral membrane proteins based on computationally simulated membrane integration efficiencies.

5.1 Introduction

Integral membrane proteins (IMPs) play crucial roles in the transport of molecules, energy, and information across the membrane and are an important focus of structural and biophysical studies. However, the production of sufficient levels of IMPs is a limiting factor in their characterization [74]. Even among homologous IMP sequences, expression levels can vary widely [45, 71, 74, 77, 78, 84], and the mechanistic basis for this variability is often unclear. Extensive efforts have been committed to identify IMP sequences, expression conditions, and host modifications that yield IMP expression at sufficient levels for further study [100, 109, 113, 134]. Despite these efforts, general guidelines for successful overexpression for IMPs are lacking.

Biogenesis of IMPs in *E. coli* involves multiple steps that are potential bottlenecks for overexpression, including correct targeting to the inner membrane [1, 105], membrane integration [19, 29, 84, 98, 116, 146], and folding [35, 76, 129, 142]. For a given sequence, understanding how each of these steps affects observed expression levels may lead to improved strategies for IMP overexpression.

As demonstrated in *Chapter 4*, the Sec-facilitated membrane integration step of biogenesis is a limiting factor in the overexpression of the TatC IMP [84]. Sequence changes in the C-tail that alter the efficiency of membrane integration efficiency – determined either from coarse-grained (CG) simulations or experimentally – were shown to correlate with experimentally observed IMP expression levels. Further work is necessary to explore the generality of this link and its potential for enabling the rational enhancement of IMP expression.

The results in this chapter demonstrate the predictive capacity of simulated integration efficiency for experimental expression by examining a wide range of sequence modifications to TatC homologs across the protein sequence. The studied sequence modifications include point mutations, loop-swap chimeras, and double-loop-swap chimeras, and it is shown that the simulated integration efficiency – as predicted by CG simulations – broadly correlates with IMP expression. An ampicillin resistance assay is employed to directly validate the simulated integration efficiencies and to confirm the mechanistic interpretation. We further demonstrate cumulative and largely independent effect of multiple mutations on both the simulated integration efficiency and the experimentally observed expression levels. Finally, we provide a methodology that can be used to generally identify sequence regions in other IMPs that may exhibit correlations like those elucidated here for TatC, yielding a

broadly applicable tool for the computational prediction of sequence modifications that improve IMP overexpression.

5.2 Results

TatC expression levels are sensitive to loop swaps

TatC is an IMP with six transmembrane domains (TMD) and a cytoplasmic N- and C-terminus (Figure 5.1A) that is a component of the bacterial twin-arginine translocation pathway [13]. A representative pool of 111 loop-swap chimeras was generated by replacing a single loop in one of ten wild-type TatC homologs (*Aquifex aeolicus* (*Aa*), *Bordetella parapertussis* (*Bp*), *Campylobacter jejuni* (*Cj*), *Deinococcus radiodurans* (*Dr*), *Escherichia coli* (*Ec*), *Hydrogenivirga* species 128-5-R1 (*Hy*), *Mycobacterium tuberculosis* (*Mt*), *Staphylococcus aureus* (*Sa*), *Vibrio cholera* (*Vc*), and *Wolinella succinogenes* (*Ws*)) with the corresponding loop from one of the other nine homologs (Figure 5.1A). Loop domains were identified by sequence alignment and membrane topology predictions [128] (sequences available online[4]). Both mutant and wild-type expression levels were determined using a C-terminal GFP tag [27] (see Methods), and the relative effect of each mutation on expression was quantified in terms of the ratio (Eq. 5.1)

$$\text{Exp. Expression} = \frac{expression(\text{mutant})}{expression(\text{wild-type})}. \tag{5.1}$$

Values greater than unity (> 1.0) indicate improvement in expression due to the sequence modification.

The set of loop swaps exhibit a wide range of values for this experimental expression ratio, as shown in Figure 5.1B. The effect of single loop swaps range from 0.02- to 40-fold changes, with 43% of the studied loop swaps yielding improved expression. Control studies were performed to confirm that the C-terminal GFP tag does not substantially alter the experimentally measured expression levels. A set of 11 single-loop-swap chimeras and their corresponding wild-type sequences were cloned into an alternative construct containing an N-terminal Strep tag (WSHPQFEK) with no C-terminal tag (see Methods). The experimental expression ratio in Eq. 5.1 was measured for each N-terminal Strep tag construct and compared against quantification via C-terminal GFP fluorescence. Figure 5.1C shows this comparison, revealing agreement for all studied cases between measured expression levels using either tag. This result, additionally supported by extensive studies in which IMP-GFP fluorescence is shown to be a robust quantifier of expression [27, 34], indicates

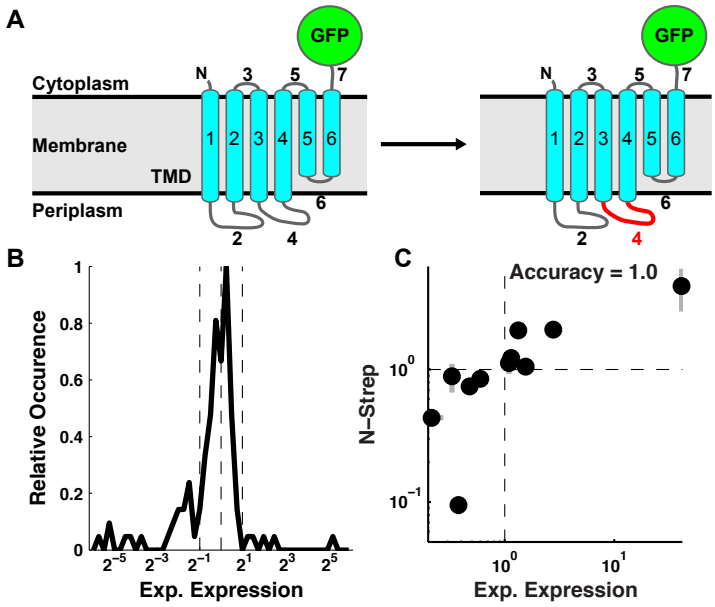

Figure 5.1: TatC loop-swap chimeras demonstrate a range of expression outcomes. (**A**) A schematic of a wild-type (left) and loop-swap chimera (right) sequence for the TatC IMP with a C-terminal GFP tag. Corresponding loop domains are swapped between TatC homologs to create loop-swap chimeras, as illustrated for loop 4. (**B**) The distribution of experimental expression values (mutant/wild-type) for the pool of 111 single-loop-swap TatC chimeras. Vertical dashed lines indicate two-fold change in experimental expression about the mean of the distribution. (**C**) Correlation between experimental expression levels quantified using a C-terminal GFP tag (Exp. Expression) versus using an N-terminal Strep tag (N-strep).

that the experimental expression outcomes are robust with respect to the means of quantifying the expression levels.

Simulated integration efficiency is predictive of TatC expression

Correlation between simulated integration efficiency and experimentally observed expression levels was previously identified in TatC based on a limited set of mutations [84]; here, we systematically test the predictive capacity of simulated integration efficiency for expression in a diverse set of 111 loop-swap chimeras. CG simulations were performed for each chimera and wild-type sequence (see Methods), and the effect of each mutation on simulated integration efficiency was quantified in terms of the ratio (Eq. 5.2)

$$\text{Sim. Integration} = \frac{P_{C_{in}}(\text{mutant})}{P_{C_{in}}(\text{wild-type})}, \tag{5.2}$$

where $P_{C_{in}}$ corresponds to the fraction of simulated trajectories for which the C-tail domain is correctly localized with respect to the cell membrane; in a later Results section, we investigate the use of sequence features other than the C-tail for quantifying integration efficiency. Receiver operator characteristic (ROC) curves (Figure 5.2A) [121] provide a statistical measure of the predictive capacity of simulated integration efficiency, with values in excess of 0.5 for the area under the ROC curve (AUC) indicating predictive capacity.

ROC curves in Figure 5.2A are shown for datasets corresponding to all 111 loop-swap chimeras (blue) and to the subset of 82 loop-swap chimeras that exclude C-tail swaps (green). This plot demonstrates the predictive capacity of simulated integration efficiency for experimental expression, with AUC values exceeding 0.5 beyond 95% statistical confidence. The similarity of the two curves indicates that the predictive capacity of the simulated integration efficiency is relatively insensitive to whether the loop-swap involves the C-tail domain.

Also, indicated in Figure 5.2A (blue and green dots) are the points along the ROC curve that correspond to the cut-off value (defining positive prediction) for the simulated integration efficiency ratio in Eq. 5.2 that offers the greatest predictive capacity for experimentally observed expression; for both datasets, this optimal value is found to be 1.0, indicating that increases or decreases in the simulated integration efficiency straightforwardly predict the corresponding changes in experimental expression levels.

Experimental confirmation of changes in integration efficiency predicted by simulation

To experimentally confirm that the *in vivo* integration efficiency is correctly described by the CG simulations, we apply a previously developed ampicillin resistance assay [84] (see Methods). Upon fusing a C-terminal β-lactamase tag to the TatC sequence, ampicillin resistance is imparted when the C-tail is mislocalized (i.e., oriented into the periplasm) during expression. Therefore, an increase in ampicillin resistance is a direct *in vivo* test of any decrease in correct C-tail localization predicted from the CG simulations.

The survival metric reported in Figure 5.2B is the ratio of colonies observed following ampicillin treatment between a loop-swap chimera and the corresponding wild-type TatC sequence. For a subset of 14 loop-swap chimeras, Figure 5.2B compares the relative survival to simulated integration efficiency; this subset was

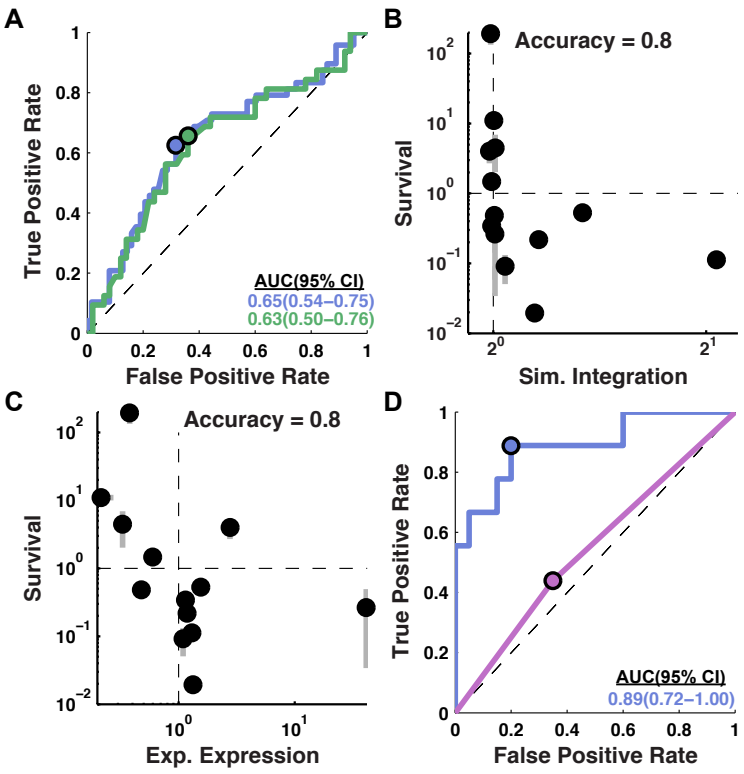

Figure 5.2: C-tail localization is predictive of experimental expression. (**A**) The predictive capacity of simulated integration efficiency for experimental expression is assessed using a ROC curve for all single-loop-swap chimeras (blue, 111 sequence modifications) and all single-loop-swap chimeras excluding those in which the C-tail was swapped (green, 82 sequence modifications). Significant predictive capacity is observed for both sets, as indicated by the area under the curve (AUC) values (bottom right, in colors matching the corresponding ROC curves). (**B**) Comparison of simulated integration efficiency and ampicillin resistance for TatC loop-swap chimeras. A negative correlation between survival and simulated integration efficiency indicates that the C-tail topology predicted by the CG simulations occurs *in vivo*. One sequence had a survival level below the plotted range. The reported measure of accuracy corresponds to the fraction of sequences for which the simulation predicts changes in topology that are consistent with the direction of changes in the experimental expression. (**C**) Comparison of experimental expression with relative ampicillin resistance for TatC loop-swap chimeras. A negative correlation between survival and experimental expression indicates that the C-tail mislocalizes in poorly expressing chimeras, consistent with the mechanism predicted by the CG simulations. One sequence had a survival level below the plotted range. (**D**) The predictive capacity of simulated integration efficiency for experimental expression assessed using a ROC curve for TatC point mutants (29 sequence modifications). Simulated integration efficiency from the CG model (blue) has greater predictive capacity for experimental expression than the positive inside rule (purple).

selected randomly from the full set of single-loop swap chimeras and including four C-tail-swap chimeras (sequences available online[4]). For 11 of these 14 cases, the corresponding data points in Figure 5.2B fall into the diagonal quadrants of the plot, indicating good agreement between the experimental and simulated measures of integration efficiency (Accuracy= 0.8 ± 0.2, 95% confidence interval).

Figure 5.2C plots the correlation between ampicillin resistance and experimental expression for the same set of loop-swap chimeras. As expected (given the positive correlation between simulated integration efficiency and experimental expression in Figure 5.2A, and the negative correlation between the simulated integration efficiency and the survival assay in Figure 5.2B), Figure 5.2C indicates strong negative correlation between ampicillin resistance and experimental expression, with 11 of the 14 data points falling in the diagonal quadrants (Accuracy= 0.8 ± 0.2, 95% confidence interval). Taken together, Figs. 5.2B and 5.2C demonstrate that simulated integration is a reliable predictor of the C-tail orientation, which is in turn a reliable predictor of experimental expression.

The effect of point mutations on simulated integration efficiency is predictive for expression

Rather than loop-swap mutations, we now consider the effect single-point mutations on both experimental expression and simulated integration efficiency. Point mutants introduce minimal changes to the wild-type sequence and are often used for protein-sequence design [82, 110, 115]. The blue curve in Figure 5.2D shows the ROC curve for a set of 29 point mutants; each exhibits a single mutation at a position in the wild-type sequence that is not universally conserved across homologs, with the mutation either increasing or decreasing the charge at that position (sequences available online[4]). The blue curve in Figure 5.2D indicates that the simulated integration efficiencies from the CG method have predictive capacity (AUC= 0.89) that is even higher than was found in Figure 5.2A for loop-swap mutations (AUC= 0.65).

For comparison, the purple curve in Figure 5.2D explores the predictive capacity of a simpler measure of integration efficiency based only on the positive inside rule, which observes that positively charged residues are more likely to be localized to the cytosolic side of the cell membrane [130] and that modification of the positively charged residues can change IMP topology [35, 114, 129, 142]. As employed here, the positive inside rule simply predicts that a mutation will have increased integration efficiency (and thus a positive effect on expression) if it increases the

net charge of the cytosolic loops minus the net charge of the periplasmic loops, and vice versa. It is clear from the Figure 5.2D that in contrast to the prediction of the CG model (blue), the positive inside rule has little predictive capacity for expression when employed in this way. These results emphasize that the molecular processes and interactions that govern IMP integration are more complex, and they are more completely described using the CG simulations than by simple analysis of charged residues.

Multiple sequence modifications have a combinatory effect on simulated integration efficiency and expression

To determine whether multiple sequence modifications have a combinatory effect on expression and simulated integration efficiency, a set of 12 double-loop-swap chimeras was generated (sequences available online[4]) and tested against the corresponding effect of the constituent single-loop-swap mutations. Figure 5.3 shows that for both simulated integration efficiency (part A) and experimental expression (part B) comparison of the fold-change (Eq. 5.1-5.2) observed for the double-loop-swap chimera is strongly correlated with the product of fold-changes for the corresponding single-loop-swap chimeras (Pearson's correlation coefficient, $r = 0.9$). Linear fits of the data are plotted as solid lines. The slope of the linear fits for both simulated integration efficiency (Figure 5.3A, slope= 0.8) and experimental expression (Figure 5.3B, slope= 0.7) deviate only slightly from unity, indicating that the effect of each mutation is largely independent. The results in Figure 5.3 suggest that the introduction of multiple mutations is a viable strategy for enhancing expression, and that simulated integration efficiency largely captures the effect of these multiple mutations.

TatC topology features, other than C-tail localization, are not predictive for expression

Using the fraction of CG trajectories for which the TatC C-tail reaches correct localization with the respect to the membrane as the measure of successful IMP integration, the results in Figure 5.2, along with previous work [84], support the conclusion that simulated integration efficiency reliably predicts experimental expression in TatC. However, other features of the TatC topology (such as the localization of other soluble loops) could have been employed to quantify IMP integration from the CG simulations. We now investigate the predictive capacity of the CG simulations for experimental expression, using alterative measures of IMP integration.

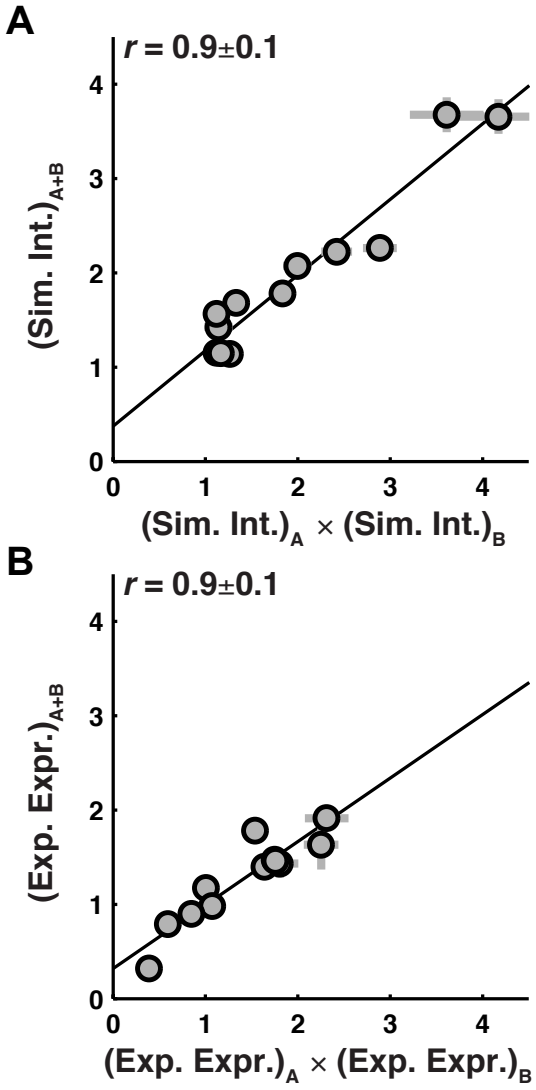

Figure 5.3: Effect of multiple sequence modifications on simulated integration efficiency and experimental expression is cumulative and nearly independent. (**A**) The simulated integration efficiency of double loop-swap chimeras (vertical axis) versus the product of the simulated integration efficiencies of the constituent single-loop-swap chimeras (horizontal axis). The guideline with slope of 0.8 indicates that the effect of loop-swap mutations on simulated integration efficiency is cumulative and largely independent. (**B**) The experimental expression of double loop-swap chimeras (vertical axis) versus the product of the experimental expression values of the constituent single-loop-swap chimeras (horizontal axis). The guideline with slope of 0.7 indicates that the effect of loop-swap mutations on experimental expression is also cumulative and largely independent.

The alternative measures of IMP integration that are considered include (1) $p^{(i)}$, the fraction of CG trajectories for which soluble loop i reaches correct localization with

the respect to the membrane, (2) $p^{(\mathrm{All})}$, the fraction of CG trajectories for which all soluble loops reach correct localization, and (3) $p^{(N)}$, the fraction of CG trajectories for which correct localization is achieved for the soluble loop that includes the mutation. In this notation, the previously discussed measure of IMP integration based on the C-tail is given by $p^{(7)}$.

Using each of these measures of IMP integration, we obtained ROC curves that compare the simulated integration efficiency with observed experimental expression, and the corresponding AUC values are presented in Figure 5.4A. In all cases, the ROC curves were determined using the dataset with all 140 TatC loop-swap and point mutations discussed in the preceding sections. The AUC for the C-tail measure ($p^{(7)}$ is 0.73, indicating the strong predictive capacity of this measure. However, it is clear that all other measures of integration efficiency fail to offer predictive capacity (yielding AUC values that are within 95% confidence of 0.5). Even when the measure of integration efficiency is based on the localization of the loop in which the mutation occurs (i.e., $p^{(N)}$), the predictive capacity is significant worse than using the C-tail (i.e., $p^{(7)}$).

The results in Figure 5.4A raise the question of the underlying mechanism for the predictive capacity of the C-tail localization for TatC. One hypothesis is that the C-tail acts as an "aggregator" of all preceding errors in the IMP integration, providing a cumulative report on the TatC topology. A second hypothesis is that the C-tail is akin to a "canary in the coal mine", particularly sensitive to mutations, regardless of where in the sequence the mutation occurs. Finally, a third hypothesis is that the unique features of the C-tail could make it more amenable to accurate description by the CG method than the other TatC loops.

We directly test the aggregator hypothesis by investigating the degree to which the C-tail measure of integration efficiency is predictive of the alternative measures. Figure 5.4B presents the resulting AUC values, obtained from ROC curves for $p^{(7)}$ versus the alternative measures, using the full dataset of 140 TatC loop-swap and point mutations. It is clear from the figure that there is no significant correlation between $p^{(7)}$ and the other measures, a finding that is inconsistent with the aggregator hypothesis. Both Figure 5.4A and 5.4B emphasize that the C-tail is a unique reporter of TatC integration efficiency, at least among the diverse set of measures considered here.

The second hypothesis reasons that the C-tail of TatC is particularly sensitive to sequence modification and is thus a useful reporter of integration efficiency, regardless

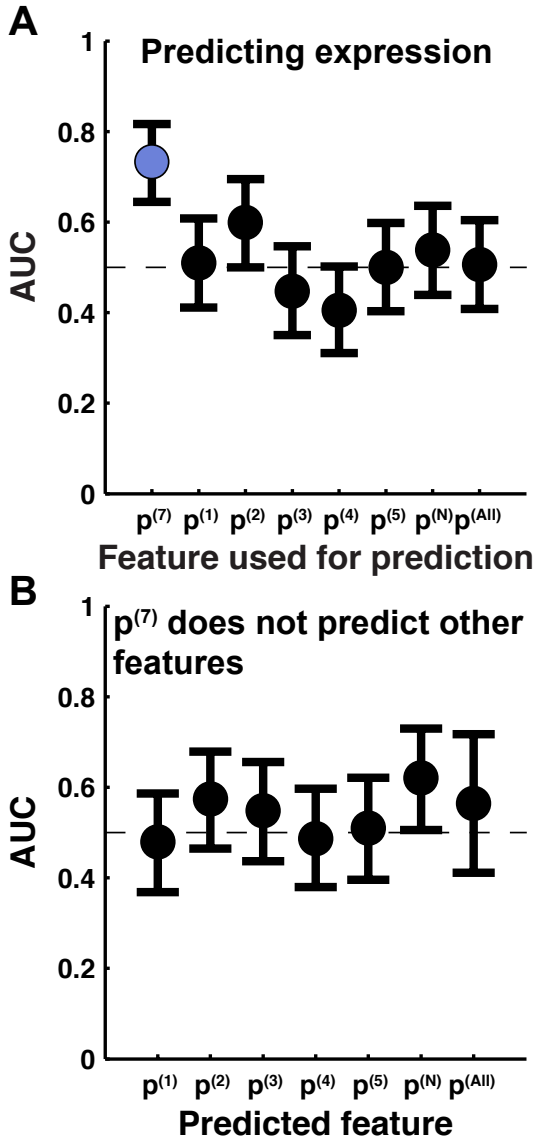

Figure 5.4: Simulated integration efficiency using the C-tail ($p^{(7)}$) measure of integration is predictive of experimental expression of TatC, while other measures are not. (**A**). AUC obtained by using various measures of integration efficiency ($p^{(1)}$, $p^{(2)}$, $p^{(3)}$, $p^{(4)}$, $p^{(5)}$, $p^{(7)}$, $p^{(N)}$, and $p^{(\mathrm{All})}$; defined in text) to predict experimental expression. $p^{(7)}$ (i.e., C-tail localization) is the only measure with statistically significant predictive capacity. (**B**) AUC obtained by using C-tail localization ($p^{(7)}$) to predict other measures of integration efficiency. Error bars indicate 95% confidence intervals.

of where in the sequence the mutation occurs. Although this hypothesis is difficult to directly test, it is consistent with the results from the ampicillin resistance assay, which found that C-tail localization was substantially impacted by mutations in other parts of the TatC sequence, even for mutations in other loops. Possibly contributing

to the conformational sensitivity of the C-tail is that the preceding TMD domains (TM5 and TM6) are relatively short and do not fully span the cell membrane in the *Aa*TatC structure [97, 99].

With regard to the third hypothesis, we note that the CG model does not explicitly describe sequence-specific interactions and packing effects among the TMD domains; the model is thus expected to be most reliable for describing the topology of TMD domains with weak tertiary interactions, such as the C-tail of TatC [97, 99]. This explanation leaves open the possibility that improvements to the CG model in terms of its description of tertiary IMP interactions could lead to more robust measures of simulated integration efficiency [88].

The analysis in this section is central to the question of how generally the CG simulations will be able to predict membrane protein expression for IMPs other than TatC. It is very possible that for other IMPs, the C-tail localization will not be the most useful measure of IMP integration for predicting expression levels [103]. In the next section, we thus describe a simple strategy for identifying a useful measure of IMP integration, on the basis of limited experimental expression data.

Predictors for expression can be identified from limited training data

Utilization of simulated integration efficiency to predict IMP expression in IMPs other than TatC requires a useful measure of IMP integration to compute from the CG simulations. The results in Figs. 5.2 and 5.3 use C-tail localization for this purpose, but as is illustrated in Figure 5.4, other reasonable measures of simulated integration efficiency are not predictive for expression. For the study of an arbitrary IMP, we are thus faced with determining, as efficiently as possible, a measure of simulated integration efficiency to compute from the CG method.

Here, we present a simple strategy for identifying a useful measure of IMP integration, based on comparison of the CG simulations with limited experimental expression data. For the case of TatC, Figure 5.5 presents the results of an analysis in which the predictive capacity of various candidate measures of IMP integration is evaluated using a limited number of comparisons between experimental expression measurements and CG simulations. We consider randomly selected subsets of the full dataset of 140 TatC loop-swap and point mutations, and for each subset, we employ the various measures of integration efficiency to evaluate the AUC that reflects the predictive capacity of simulated integration efficiency in comparison to experimental expression data. As a function of the subset size, the figure plots the

fraction ($M(i)$) of random subsets for which each measure of integration efficiency (indexed by i) yields the highest AUC value. These results show that with expression data for only a small training set, the most predictive measure of IMP integration can be identified. In the case of TatC, fewer than 20 sequences are needed to determine $p^{(7)}$ as most predictive.

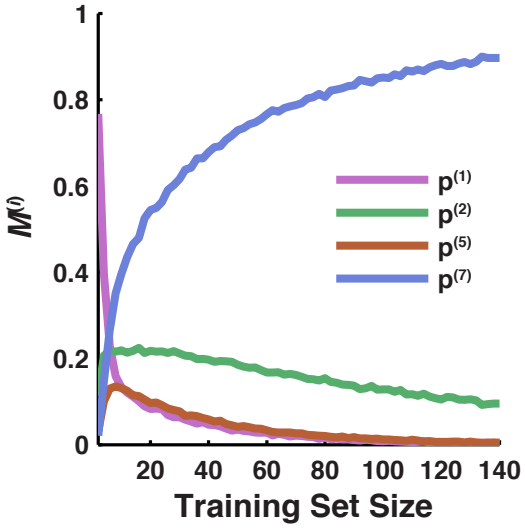

Figure 5.5: Determination of useful measures of integration efficiency based on limited data. The probability that each measure of integration is the most predictive for expression ($M^{(i)}$), described in text), based on training datasets of increasing size. The $p^{(7)}$ measure (based on C-tail localization) is identified as the most predictive based on datasets with fewer than 20 sequences. For clarity, only features with values of $M^{(i)}$ greater than 0.1 are shown in the plot; not shown but included in the analysis are $p^{(3)}$, $p^{(4)}$, $p^{(N)}$, and $p^{(\mathrm{All})}$.

The strategy in Figure 5.5 illustrates that for cases in which limited IMP expression data is available, a useful measure of IMP integration from the CG simulations can be identified without other prior knowledge, thus yielding a general strategy for enhancing IMP expression in systems other than TatC. However, there will be cases in which even limited IMP expression data is not available. For these cases, a reasonable strategy is to use a measure of IMP integration that involves a sequence domain that is expected to be prone to mislocalization with respect to the cell membrane. Analyses of sequence conservation [4] and residue co-evolution [69, 83, 124] provide reasonable strategies for identifying such sequence domains. For the case of TatC, this approach would again be consistent with the use of the C-tail for measuring of integration efficiency, since this sequence domain is

not conserved across homologs and was not resolved in the reported TatC crystal structures [97, 99].

5.3 Discussion

We address the problem of heterologous IMP expression in *E. coli* by utilizing the link between simulated integration efficiency and experimental expression outcomes [84] to predict sequence modifications that improve expression for the TatC. Simulated integration efficiency is determined using CG molecular dynamics of the co-translational integration of the IMP via the Sec-translocon [146] and is compared against experimental expression measurements for a set of 140 TatC sequence modifications. For both loop-swap modification (Figure 5.2A) and point mutations (Figure 5.2D) the simulated integration efficiency is shown to provide clear predictive capacity of experimental expression, and the effect of multiple sequence modifications (Figure 5.3) is shown to be cumulative and likewise captured by the simulated integration efficiency. For the combined set of 140 sequence modifications, the diagnostic odds ratio [40] obtained from comparison of simulated integration efficiency with experimental expression yields a value of 3.9 (1.9-9.1, 95% confidence interval), indicating that sequence modifications that improve simulated integration efficiency are four-fold enriched in terms of improved experimental expression.

Although successful strategies for improving IMP overexpression have been previously demonstrated [109, 113, 134], these approaches leave unclear the mechanism by which expression is improved, requiring a case-by-case implementation that can be costly in terms of both time and material resources. The strategy employed in the current work aims to optimize IMP expression on the basis of a particular step in IMP biogenesis – successful integration into the membrane and adoption of the correct multi-spanning topology. Additional work is needed to demonstrate the degree to which improving membrane integration efficiency will lead to improved expression levels in other IMPs, but the central role of membrane integration in IMP biogenesis suggests that the approach may prove successful for other IMPs.

Finally, we note that the current work is unique in that CG simulations form the basis for the prediction of enhanced IMP expression. Although molecular simulations have been successfully employed in the context of other biomolecular design problems – such as the de novo protein structure design [20, 24, 59] or enzyme design [14, 64, 123] – the current work suggests that rational enhancement of IMP

expression is a new application domain in which molecular simulations may prove useful.

5.4 Methods

Cloning

All TatC coding sequences were created using either primer extension or were synthesized by Twist Bioscience (San Francisco, CA). Loop-swap chimeras involved modification of loops 1-5 and 7, avoiding the short loop 6. The pool of 111 loop-swap chimera sequences were selected from all 540 possible combinations. Each wild-type homolog was used between 6 to 15 times as a parent, between 7 to 19 times as a source for the mutant loop, and each loop was mutated between 8 to 29 times. Point mutants were chosen to affect a change in charge through mutation of neutral residues to charged residues or through mutation of charged residues to the opposite charge. All sequences used are provided in the supplemental data. Each loop-swap chimera coding sequences was cloned into the pET28(a+)-GFP-ccdB vector [28, 84] using the Gibson cloning protocol [39], resulting in each IMP possessing a C-terminal GFP tag. For constructs containing the β-lactamase tag, the GFP sequence was replaced with a β-lactamase sequence using Gibson cloning. For constructs containing the N-terminal Strep tag, the GFP and poly-His sequence was removed during PCR and the Strep tag was added using primer extension; the final vectors were constructed using Gibson cloning.

Heterologous expression in *E. coli*

Heterologous expression of IMPS in *E. coli* was performed as previously described [84]. In short, IMPs were expressed in BL21 Gold (DE3) (Agilent Technologies, Santa Clara, CA) cells at 16 °C for approximately 16 hours prior to either flow cytometry, western blot, or ampicillin resistance analysis.

Flow cytometry

Flow cytometry was performed as previously described [84]. In short, cultures of cells expressing TatC IMPs with a C-terminal GFP tag were resuspended in PBS and subjected to flow cytometry. Whole cell fluorescence from the B1/FITC channel was measured using a MACSQuant10 Analyzer (Miltenyi Biotec, Bergisch Galbach, Germany). Mean fluorescence values are calculated using FlowJo (Ashland, OR).

Western blotting

All samples of cells expressing IMPs with an N-terminal Strep tag were subjected to the following protocol for western blot analysis. Samples were normalized to an OD_{600} of 3.0 in PBS and subjected to three freeze-thaw cycles using liquid nitrogen and applied to 10% SDS-PAGE followed by western blotting. Relative protein levels were determined by incubation of the western blot membrane with an anti-Strep tag primary rabbit antibody (NWSHPQFEK Antibody, GenScript, Piscataway, NJ) followed by incubation with an IRDye 800CW Donkey anti-rabbit secondary antibody (LI-COR, Lincoln, NE) and visualization using a LI-COR IR western blot scanner (LI-COR, Lincoln, NE). Relative band intensities were quantified using ImageJ [112].

Description of the CG simulations

We apply a previously developed CG approach [84, 146, 147] to simulate the minute-timescale dynamics of co-translational membrane integration via the Sec translocon. The CG model is applied and implemented as described in detail in [84] and in Chapter 4 of this thesis, with key features of the CG model summarized here.

The CG simulations explicitly describe the configurational dynamics of the IMP, conformational gating of the Sec translocon lateral gate, and ribosomal translation (at 24 residues/second). The IMP is represented as a freely jointed chain of CG beads, where each CG bead represents three amino acids and has a diameter of 8Å, equal to the Kuhn length of a polypeptide chain [52, 119]. To avoid a frameshift in the mapping of amino acids to CG beads upon a loop-swap sequence modification, dummy atoms were introduced, as described previously [84]. Bonding interactions between neighboring CG beads are described using the finite extension nonlinear elastic (FENE) potential [72], short-ranged non-bonding interactions are modeled using a Lennard-Jones potential, and electrostatic interactions are modeled using the Debye-Hückel potential. Factors that prevent backsliding of large translocated hydrophilic loops are included, as described in [146], for consistency with previous work but have only a modest effect in TatC. Solvent interactions are described using a position-dependent potential based on the water-membrane transfer free energy for each CG bead [84].

The configuration of the IMP is time evolved using overdamped Langevin dynamics, with the CG beads confined to a two-dimensional subspace that runs along the axis of the translocon channel and between the two helices of the LG. Conformational gating

of the LG corresponds to the LG helices moving out of the place of confinement for the IMP, allowing the IMP to pass into the membrane bilayer. The rate of stochastic LG opening and closing is dependent on the sequence of the CG beads that occupy the translocon channel [145, 146]. Ribosomal translation is directly simulated via growth of the IMP at the ribosomal exit channel; throughout translation, the C-terminus of the IMP is held fixed, and new beads are sequentially added at a rate of 24 residues per second. Upon completion of translation, the C-terminus is released from the ribosome.

Trajectories use a step-size of 100 ns for time integration and are terminated 31 s after the end of translation. For each protein sequence, at least 400 independent trajectories are calculated.

Determination of measures of integration from CG simulations

The simulated integration efficiency for a protein sequence is calculated from the CG model as previously described [84] in Chapter 4 of this thesis. The topology of a protein is analyzed over the last 6 s of the CG simulation trajectories, starting 25 s after the end of protein translation by the ribosome. For each loop, i, the location of the loop during this time-window is described by a variable λ_i, where $\lambda_i = 1$ if the loop is in the cytosol, $\lambda_i = -1$ if the loop is in the periplasm, and $\lambda_i = 0$ otherwise. For each trajectory, we assess whether a given measure of integration is visited during the analysis time window. The various measures of integration efficiency used in this work are described in the text.

Ampicillin resistance assay

The ampicillin resistance assay was performed as previously described [84]. In short, cells that had expressed IMPs with a C-terminal β-lactamase-tag overnight at 16 °C were resuspended to an OD_{600} of 0.1 and grown to an OD_{600} of 0.5, after which ampicillin was added; cells were then incubated for an additional 1.5 hours, followed by plating on kanamycin LB agar plates. The relative number of observed colonies between loop-swap chimera and wild-type was used to determine the change in C-tail translocation, with a ratio greater than one indicating an increase in translocation of the C-tail to the periplasm due to the sequence modification.

Statistical significance calculations

Reported experimental measurements, including values for experimental expression, survival, and protein levels quantified using western-blot, correspond to averages

over at least 3 independent trials, with error bars representing the standard error of the mean unless otherwise noted. Simulated integration efficiencies represent the average outcome of at least 400 independent CG simulations trajectories, with error bars indicating the standard error of the mean. Confidence intervals on AUC values were determined by bootstrapping. Specifically, $1,000,000$ samples of simulated integration and expression pairs, with size equal to the set of sequence modification, were drawn with replacement from the set of sequence modifications; the AUC was calculated for each sample, and the relevant percentile of the resulting AUC-value distribution determines the confidence intervals.

A similar procedure was used to generate the randomly selected subsets of the full dataset of 140 TatC loop-swap and point mutations used in Figure 5.5. For each subset size, $1,000,000$ independent samples of that size were chosen with replacement from the full dataset of 140 TatC loop-swap and point mutations.

BIBLIOGRAPHY

[1] D. Akopian, K. Shen, X. Zhang, and S. O. Shan. Signal recognition particle: an essential protein-targeting machine. *Annu Rev Biochem*, 82:693–721, 2013. ISSN 1545-4509 (Electronic) 0066-4154 (Linking). doi: 10.1146/annurev-biochem-072711-164732. URL https://www.ncbi.nlm.nih.gov/pubmed/23414305.

[2] Mike P Allen and Dominic J Tildesley. *Computer simulation of liquids*. Oxford University Press, 1989. ISBN 0198556454.

[3] William John Allen, Robin Adam Corey, Peter Oatley, Richard Barry Sessions, Sheena E Radford, Roman Tuma, and Ian Collinson. Two-way communication between SecY and SecA suggests a Brownian ratchet mechanism for protein translocation. *eLife*, 5:1–23, 2016. ISSN 2050-084X. doi: 10.7554/eLife.15598.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990. ISSN 0022-2836 (Print) 0022-2836 (Linking). doi: 10.1016/S0022-2836(05)80360-2. URL https://www.ncbi.nlm.nih.gov/pubmed/2231712.

[5] Tadashi Ando and Jeffrey Skolnick. Crowding and hydrodynamic interactions likely dominate in vivo macromolecular motion. *Proc. Natl. Acad. Sci. U S A*, 107(43):18457–18462, 2010. ISSN 1091-6490. doi: 10.1073/pnas.1011354107.

[6] Panagiotis Angelikopoulos, Costas Papadimitriou, and Petros Koumoutsakos. Bayesian uncertainty quantification and propagation in molecular dynamics simulations: A high performance computing framework. *J Chem Phys*, 137 (14):144103, 2012. doi: http://dx.doi.org/10.1063/1.4757266.

[7] Thomas Becker, Shashi Bhushan, Alexander Jarasch, Jean-Paul Armache, Soledad Funes, Fabrice Jossinet, James Gumbart, Thorsten Mielke, Otto Berninghausen, Klaus Schulten, Eric Westhof, Reid Gilmore, Elisabet C Mandon, and Roland Beckmann. Structure of monomeric yeast and mammalian Sec61 complexes interacting with the translating ribosome. *Science*, 326(5958):1369–1373, 2009. ISSN 0036-8075. doi: 10.1126/science.1178535.

[8] Nir Ben-Tal, Avinoam Ben-Shaul, Anthony Nicholls, and Barry Honig. Free-energy determinants of alpha-helix insertion into lipid bilayers. *Biophys. J.*, 70 (4):1803–1812, 1996. ISSN 00063495. doi: 10.1016/S0006-3495(96)79744-8.

[9] Bert Van Den Berg, William M Clemons Jr, Ian Collinson, Yorgo Modis, Enno Hartmann, Stephen C Harrison, Tom A Rapoport, Bert Van den Berg, and William M Clemons. X-ray structure of a protein-conducting channel. *Nature*, 427(6969):36–44, 2004. ISSN 1476-4687. doi: 10.1038/nature02218.

[10] Thomas C Beutler, Alan E Mark, Rem C Van Schaik, Paul R Gerber, and Wilfred F Van Gunsteren. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem Phys Lett*, 222:529–539, 1994.

[11] L Bischoff, S Wickles, O Berninghausen, E O van der Sluis, and R Beckmann. Visualization of a polytopic membrane protein during SecY-mediated membrane insertion. *Nat Commun*, 5:4103, 2014. ISSN 2041-1723. doi: 10.1038/ncomms5103.

[12] K W Boehlke and J D Friesen. 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 (2):429–433, 1975. ISSN 0021-9193.

[13] E. G. Bogsch, F. Sargent, N. R. Stanley, B. C. Berks, C. Robinson, and T. Palmer. An essential component of a novel bacterial protein export system with homologues in plastids and mitochondria. *J Biol Chem*, 273(29):18003–6, 1998. ISSN 0021-9258 (Print) 0021-9258 (Linking). URL `http://www.ncbi.nlm.nih.gov/pubmed/9660752`.

[14] D. N. Bolon and S. L. Mayo. Enzyme-like proteins by computational design. *Proc Natl Acad Sci U S A*, 98(25):14274–9, 2001. ISSN 0027-8424 (Print) 0027-8424 (Linking). doi: 10.1073/pnas.251555398. URL `https://www.ncbi.nlm.nih.gov/pubmed/11724958`.

[15] Daniel W A Buchan, Federico Minneci, Tim C O Nugent, Kevin Bryson, and David T. Jones. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.*, 41:349–357, 2013. ISSN 13624962. doi: 10.1093/nar/gkt381.

[16] Sara Capponi, Matthias Heyden, Ana-Nicoleta Bondar, Douglas J Tobias, and Stephen H White. Anomalous behavior of water inside the SecY translocon. *Proc. Natl. Acad. Sci. U S A*, 112(29):9016–9021, 2015. doi: 10.1073/pnas.1424483112.

[17] The UniProt Consortium. Uniprot: a hub for protein information. *Nucleic Acids Res.*, 43:D204–12, 2014. ISSN 0305-1048. doi: 10.1093/nar/gku989.

[18] Florian Cymer and Gunnar von Heijne. Cotranslational folding of membrane proteins probed by arrest-peptide-mediated force measurements. *Proc. Natl. Acad. Sci.*, 110(36):14640–5, 2013. ISSN 1091-6490. doi: 10.1073/pnas.1306787110.

[19] Florian Cymer, Gunnar Von Heijne, and Stephen H. White. Mechanisms of integral membrane protein insertion and folding. *J. Mol. Biol.*, 427(5): 999–1022, 2015. ISSN 10898638. doi: 10.1016/j.jmb.2014.09.014.

[20] B. I. Dahiyat and S. L. Mayo. De novo protein design: fully automated sequence selection. *Science*, 278(5335):82–7, 1997. ISSN 0036-8075 (Print) 0036-8075 (Linking). URL https://www.ncbi.nlm.nih.gov/pubmed/9311930.

[21] Daniel O. Daley, Mikaela Rapp, Erik Granseth, Karin Melén, David Drew, and Gunnar von Heijne. Global topology analysis of the escherichia coli inner membrane proteome. *Science*, 308(5726):1321–1323, 2005. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1109730. URL http://www.sciencemag.org/content/308/5726/1321.

[22] Djurre H. de Jong, Gurpreet Singh, W. F. Drew Bennett, Clement Arnarez, Tsjerk A. Wassenaar, Lars V. Schäfer, Xavier Periole, D. Peter Tieleman, and Siewert J. Marrink. Improved Parameters for the Martini Coarse-Grained Protein Force Field. *J. Chem. Theory Comput.*, 9:687–697, 2013.

[23] Djurre H de Jong, Svetlana Baoukina, Helgi I Ingólfsson, and Siewert J Marrink. Martini straight: Boosting performance using a shorter cutoff and gpus. *Comp. Phys. Comm.*, 199:1–7, 2016.

[24] W. F. DeGrado, C. M. Summa, V. Pavone, F. Nastri, and A. Lombardi. De novo design and structural characterization of proteins and metalloproteins. *Annu Rev Biochem*, 68:779–819, 1999. ISSN 0066-4154 (Print) 0066-4154 (Linking). doi: 10.1146/annurev.biochem.68.1.779. URL https://www.ncbi.nlm.nih.gov/pubmed/10872466.

[25] Erhan Demirci, Tina Junne, Sefer Baday, Simon Bernèche, and Martin Spiess. Functional asymmetry within the Sec61p translocon. *Proc. Natl. Acad. Sci. U. S. A.*, 110(47):18856–61, 2013. ISSN 1091-6490. doi: 10.1073/pnas.1318432110.

[26] Prasanna K. Devaraneni, Brian Conti, Yoshihiro Matsumura, Zhongying Yang, Arthur E. Johnson, and William R. Skach. Stepwise insertion and inversion of a type II signal anchor sequence in the ribosome-Sec61 translocon complex. *Cell*, 146(1):134–147, 2011. ISSN 00928674. doi: 10.1016/j.cell.2011.06.004.

[27] D. Drew, D. J. Slotboom, G. Friso, T. Reda, P. Genevaux, M. Rapp, N. M. Meindl-Beinker, W. Lambert, M. Lerch, D. O. Daley, K. J. Van Wijk, J. Hirst, E. Kunji, and J. W. De Gier. A scalable, gfp-based pipeline for membrane protein overexpression screening and purification. *Protein Sci*, 14(8):2011–7, 2005. ISSN 0961-8368 (Print) 0961-8368 (Linking). doi: 10.1110/ps.051466205. URL http://www.ncbi.nlm.nih.gov/pubmed/15987891.

[28] D. E. Drew, G. von Heijne, P. Nordlund, and J. W. de Gier. Green fluorescent protein as an indicator to monitor membrane protein overexpression in escherichia coli. *FEBS Lett*, 507(2):220–4, 2001. ISSN 0014-5793 (Print) 0014-5793 (Linking). URL http://www.ncbi.nlm.nih.gov/pubmed/11684102.

[29] Arnold J M Driessen and Nico Nouwen. Protein translocation across the bacterial cytoplasmic membrane. *Annu. Rev. Biochem.*, 77:643–667, 2008. ISSN 0066-4154. doi: 10.1146/annurev.biochem.77.061606.160747.

[30] Karen Drukker and George C Schatz. A Model for Simulating Dynamics of DNA Denaturation. *J. Phys. Chem. B*, 104(26):6108–6111, 2000. ISSN 1520-6106. doi: 10.1021/jp000550j.

[31] Olga K. Dudko, Gerhard Hummer, and Attila Szabo. Theory, analysis, and interpretation of single-molecule force spectroscopy experiments. *Proceedings of the National Academy of Sciences*, 105(41):15755–15760, 2008. doi: 10.1073/pnas.0806085105. URL http://www.pnas.org/content/105/41/15755.abstract.

[32] Pascal F Egea and Robert M Stroud. Lateral opening of a translocon upon entry of protein suggests the mechanism of insertion into membranes. *Proc Natl Acad Sci U S A*, 107(40):17182–17187, 2010. doi: 10.1073/pnas.1012556107.

[33] A. Elazar, J. Weinstein, I. Biran, Y. Fridman, E. Bibi, and S. J. Fleishman. Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane. *eLife*, 5:e12125, 2016. ISSN 01962892. doi: 10.1109/TGRS.2004.834800.

[34] N. Fluman, S. Navon, E. Bibi, and Y. Pilpel. mrna-programmed translation pauses in the targeting of e. coli membrane proteins. *Elife*, 3, 2014. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.03440. URL http://www.ncbi.nlm.nih.gov/pubmed/25135940.

[35] N. Fluman, V. Tobiasson, and G. von Heijne. Stable membrane orientations of small dual-topology membrane proteins. *Proc Natl Acad Sci U S A*, 114 (30):7987–7992, 2017. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1706905114. URL https://www.ncbi.nlm.nih.gov/pubmed/28698365.

[36] Jens Frauenfeld, James Gumbart, Eli O van der Sluis, Soledad Funes, Marco Gartmann, Birgitta Beatrix, Thorsten Mielke, Otto Berninghausen, Thomas Becker, Klaus Schulten, and Roland Beckmann. Cryo-EM structure of the ribosome-SecYE complex in the membrane environment. *Nat. Struct. Mol. Biol.*, 18(5):614–621, 2011. ISSN 1545-9993. doi: 10.1038/nsmb.2026.

[37] Hidenobu Fujita, Marifu Yamagishi, Yuichiro Kida, and Masao Sakaguchi. Positive charges on the translocating polypeptide chain arrest movement through the translocon. *J. Cell Sci.*, 124(24):4184–93, 2011. ISSN 1477-9137. doi: 10.1242/jcs.086850.

[38] E. R. Geertsma, M. Groeneveld, D. J. Slotboom, and B. Poolman. Quality control of overexpressed membrane proteins. *Proc Natl Acad Sci U S A*, 105(15):5722–7, 2008. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.0802190105. URL http://www.ncbi.nlm.nih.gov/pubmed/18391190.

[39] D. G. Gibson, L. Young, R. Y. Chuang, J. C. Venter, 3rd Hutchison, C. A., and H. O. Smith. Enzymatic assembly of dna molecules up to several hundred kilobases. *Nat Methods*, 6(5):343–5, 2009. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi: 10.1038/nmeth.1318. URL http://www.ncbi.nlm.nih.gov/pubmed/19363495.

[40] A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. Bossuyt. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol*, 56(11):1129–35, 2003. ISSN 0895-4356 (Print) 0895-4356 (Linking). URL https://www.ncbi.nlm.nih.gov/pubmed/14615004.

[41] Veit Goder and Martin Spiess. Topogenesis of membrane proteins: Determinants and dynamics. *FEBS Lett.*, 504(3):87–93, 2001. ISSN 00145793. doi: 10.1016/S0014-5793(01)02712-0.

[42] Veit Goder and Martin Spiess. Molecular mechanism of signal sequence orientation in the endoplasmic reticulum. *EMBO J.*, 22(14):3645–3653, 2003. ISSN 02614189. doi: 10.1093/emboj/cdg361.

[43] Marko Gogala, Thomas Becker, Birgitta Beatrix, Jean-Paul Armache, Clara Barrio-Garcia, Otto Berninghausen, and Roland Beckmann. Structures of the Sec61 complex engaged in nascent peptide translocation or membrane insertion. *Nature*, 506(7486):107–10, 2014. ISSN 1476-4687. doi: 10.1038/nature12950.

[44] Daniel H. Goldman, Christian M. Kaiser, Anthony Milin, Maurizio Righini, Ignacio Tinoco, and Carlos Bustamante. Mechanical force releases nascent chain–mediated ribosome arrest in vitro and in vivo. *Science*, 348(6233):457–460, 2015. ISSN 0036-8075. doi: 10.1126/science.1261909. URL http://science.sciencemag.org/content/348/6233/457.

[45] E. Gordon, R. Horsefield, H. G. Swarts, J. J. de Pont, R. Neutze, and A. Snijder. Effective high-throughput overproduction of membrane proteins in escherichia coli. *Protein Expr Purif*, 62(1):1–8, 2008. ISSN 1096-0279 (Electronic) 1046-5928 (Linking). doi: 10.1016/j.pep.2008.07.005. URL https://www.ncbi.nlm.nih.gov/pubmed/18692139.

[46] R. Grisshammer, R. Duckworth, and R. Henderson. Expression of a rat neurotensin receptor in escherichia coli. *Biochem J*, 295 (Pt 2):571–6, 1993. ISSN 0264-6021 (Print) 0264-6021 (Linking). URL http://www.ncbi.nlm.nih.gov/pubmed/8240259.

[47] L. Guglielmi, V. Denis, N. Vezzio-Vie, N. Bec, P. Dariavach, C. Larroque, and P. Martineau. Selection for intrabody solubility in mammalian cells using gfp fusions. *Protein Eng Des Sel*, 24(12):873–81, 2011. ISSN 1741-0134 (Electronic) 1741-0126 (Linking). doi: 10.1093/protein/gzr049. URL http://www.ncbi.nlm.nih.gov/pubmed/21997307.

[48] James Gumbart and Klaus Schulten. Molecular Dynamics Studies of the Archaeal Translocon. *Biophys. J.*, 90(7):2356–2367, 2006. ISSN 0006-3495. doi: 10.1529/biophysj.105.075291.

[49] James Gumbart, Christophe Chipot, and Klaus Schulten. Free-energy cost for translocon-assisted insertion of membrane proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 108(9):3596–3601, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1012758108.

[50] James Gumbart, Christophe Chipot, and Klaus Schulten. Free energy of nascent-chain folding in the translocon. *J. Am. Chem. Soc.*, 133(19):7602–7607, 2011. ISSN 00027863. doi: 10.1021/ja2019299.

[51] James C. Gumbart and Christophe Chipot. Decrypting protein insertion through the translocon with free-energy calculations. *Biochimica et Biophysica Acta (BBA) -Biomembranes*, 1858(7, Part B):1663 – 1671, 2016. ISSN 0005-2736. doi: http://dx.doi.org/10.1016/j.bbamem.2016.02.017. New approaches for bridging computation and experiment on membrane proteins.

[52] F. Hanke, a. Serr, H. J. Kreuzer, and R. R. Netz. Stretching single polypeptides: The effect of rotational constraints in the backbone. *EPL (Europhysics Lett.*, 92(5):53001, 2010. ISSN 0295-5075. doi: 10.1209/0295-5075/92/53001.

[53] C. A. Harley, J. A. Holt, R. Turner, and D. J. Tipper. Transmembrane protein insertion orientation in yeast depends on the charge difference across transmembrane segments, their total hydrophobicity, and its distribution. *J Biol Chem*, 273(38):24963–71, 1998. ISSN 0021-9258 (Print) 0021-9258 (Linking). URL http://www.ncbi.nlm.nih.gov/pubmed/9733804.

[54] Linnea E. Hedin, Karin Öjemalm, Andreas Bernsel, Aron Hennerdal, Kristoffer Illergård, Karl Enquist, Anni Kauko, Susana Cristobal, Gunnar von Heijne, Mirjam Lerch-Bader, IngMarie Nilsson, and Arne Elofsson. Membrane insertion of marginally hydrophobic transmembrane helices depends on sequence context. *Journal of Molecular Biology*, 396(1):221 – 229, 2010. ISSN 0022-2836. doi: http://dx.doi.org/10.1016/j.jmb.2009.11.036.

[55] Sven U Heinrich, Walther Mothes, Josef Brunner, and Tom A Rapoport. The Sec61p complex mediates the integration of a membrane protein by allowing lipid partitioning of the transmembrane domain. *Cell*, 102(2):233–44, 2000. ISSN 0092-8674. doi: 10.1016/S0092-8674(00)00028-3.

[56] Tara Hessa, Hyun Kim, Karl Bihlmaier, Carolina Lundin, Jorrit Boekel, Helena Andersson, Ingmarie Nilsson, Stephen H White, and Gunnar von Heijne. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, 433(7024):377–381, 2005. ISSN 1476-4687. doi: 10.1038/nature03216.

[57] Tara Hessa, Nadja M Meindl-Beinker, Andreas Bernsel, Hyun Kim, Yoko Sato, Mirjam Lerch-Bader, IngMarie Nilsson, Stephen H White, and Gunnar von Heijne. Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature*, 450(7172):1026–1030, 2007. ISSN 0028-0836. doi: 10.1038/nature06387.

[58] D. M. Hoover and J. Lubkowski. Dnaworks: an automated method for designing oligonucleotides for pcr-based gene synthesis. *Nucleic Acids Res.*, 30(10):e43, 2002. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=12000848.

[59] P. S. Huang, S. E. Boyken, and D. Baker. The coming of age of de novo protein design. *Nature*, 537(7620):320–7, 2016. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature19946. URL https://www.ncbi.nlm.nih.gov/pubmed/27629638.

[60] Nicholas T. Ingolia, Liana F. Lareau, and Jonathan S. Weissman. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802, 2011. ISSN 00928674. doi: 10.1016/j.cell.2011.10.002.

[61] Nurzian Ismail, Rickard Hedman, Nina Schiller, and Gunnar von Heijne. A biphasic pulling force acts on transmembrane helices during translocon-mediated membrane integration. *Nat. Struct. Mol. Biol.*, 19(10):1018–1022, 2012. ISSN 1545-9993. doi: 10.1038/nsmb.2376.

[62] Nurzian Ismail, Rickard Hedman, Martin Lindén, and Gunnar von Heijne. Charge-driven dynamics of nascent-chain movement through the SecYEG translocon. *Nat. Struct. Mol. Biol.*, 22(2):145–149, 2015. ISSN 1545-9993. doi: 10.1038/nsmb.2940.

[63] Koreaki Ito, Shinobu Chiba, and Kit Pogliano. Divergent stalling sequences sense and control cellular physiology. *Biochemical and Biophysical Research Communications*, 393(1):1 – 5, 2010. ISSN 0006-291X. doi: https://doi.org/10.1016/j.bbrc.2010.01.073. URL http://www.sciencedirect.com/science/article/pii/S0006291X10001221.

[64] L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Rothlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, 3rd Barbas, C. F., D. Hilvert, K. N. Houk, B. L. Stoddard, and D. Baker. De novo computational design of retro-aldol enzymes. *Science*, 319(5868):1387–91, 2008. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1152692. URL https://www.ncbi.nlm.nih.gov/pubmed/18323453.

[65] Ahmad Jomaa, Daniel Boehringer, Marc Leibundgut, and Nenad Ban. Structures of the E. coli translating ribosome with SRP and its receptor and with the translocon. *Nat. Commun.*, 7:10471, 2016. ISSN 2041-1723. doi: 10.1038/ncomms10471.

[66] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195–202, 1999. ISSN 0022-2836. doi: 10.1006/jmbi.1999.3091.

[67] Tina Junne, Torsten Schwede, Veit Goder, and Martin Spiess. Mutations in the Sec61p channel affecting signal sequence recognition and membrane protein topology. *J. Biol Chem*, 282(45):33201–33209, 2007. ISSN 00219258. doi: 10.1074/jbc.M707219200.

[68] Tina Junne, Lucyna Kocik, and Martin Spiess. The Hydrophobic Core of the Sec61 Translocon Defines the Hydrophobicity Threshhold for Membrane Integration. *Mol. Biol. Cell*, 21(24):1662–1670, 2010. ISSN 1939-4586. doi: 10.1091/mbc.E10.

[69] H. Kamisetty, S. Ovchinnikov, and D. Baker. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A*, 110(39):15674–9, 2013. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1314045110. URL https://www.ncbi.nlm.nih.gov/pubmed/24009338.

[70] Lucyna Kocik, Tina Junne, and Martin Spiess. Orientation of internal signal-anchor sequences at the sec61 translocon. *J. Mol. Biol.*, 424(5):368–378, 2012. ISSN 00222836. doi: 10.1016/j.jmb.2012.10.010.

[71] A. Korepanova, F. P. Gao, Y. Hua, H. Qin, R. K. Nakamoto, and T. A. Cross. Cloning and expression of multiple integral membrane proteins from mycobacterium tuberculosis in escherichia coli. *Protein Sci*, 14(1):148–58, 2005. ISSN 0961-8368 (Print) 0961-8368 (Linking). doi: 10.1110/ps.041022305. URL https://www.ncbi.nlm.nih.gov/pubmed/15608119.

[72] K KREMER and G S GREST. Molecular-Dynamics (Md) Simulations for Polymers. *J. Phys. Condens. Matter*, 2:SA295—-SA298, 1990. ISSN 0953-8984.

[73] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comp. Chem.*, 13(8): 1011–1021, 1992. ISSN 1096-987X. doi: 10.1002/jcc.540130812.

[74] O. Lewinson, A. T. Lee, and D. C. Rees. The funnel approach to the precrystallization production of membrane proteins. *J Mol Biol*, 377(1):62–73, 2008. ISSN 1089-8638 (Electronic) 0022-2836 (Linking). doi: 10.1016/j.jmb.2007.12.059. URL http://www.ncbi.nlm.nih.gov/pubmed/18241890.

[75] Long Li, Eunyong Park, JingJing Ling, Jessica Ingram, Hidde Ploegh, and Tom A. Rapoport. Crystal structure of a substrate-engaged SecY protein-translocation channel. *Nature*, 531(7594):395–399, 2016. ISSN 0028-0836. doi: 10.1038/nature17163.

[76] Yun Lu, Isaiah R. Turnbull, Alvina Bragin, Kristin Carveth, A. S. Verkman, and William R. Skach. Reorientation of aquaporin-1 topology during maturation in the endoplasmic reticulum. *Molecular Biology of the Cell*, 11(9):2973–2985, 2000. ISSN 1059-1524, 1939-4586. URL http://www.molbiolcell.org/content/11/9/2973.

[77] K. Lundstrom. Structural genomics for membrane proteins. *Cell Mol Life Sci*, 63(22):2597–607, 2006. ISSN 1420-682X (Print) 1420-682X (Linking). doi: 10.1007/s00018-006-6252-y. URL https://www.ncbi.nlm.nih.gov/pubmed/17013556.

[78] P. Ma, F. Varela, M. Magoch, A. R. Silva, A. L. Rosario, J. Brito, T. F. Oliveira, P. Nogly, M. Pessanha, M. Stelter, A. Kletzin, P. J. Henderson, and M. Archer. An efficient strategy for small-scale screening and production of archaeal membrane transport proteins in escherichia coli. *PLoS One*, 8 (10):e76913, 2013. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). doi: 10.1371/journal.pone.0076913. URL https://www.ncbi.nlm.nih.gov/pubmed/24282478.

[79] Justin L. MacCallum and D. Peter Tieleman. Hydrophobicity scales: A thermodynamic looking glass into lipid-protein interactions. *Trends Biochem. Sci.*, 36(12):653–662, 2011. ISSN 09680004. doi: 10.1016/j.tibs.2011.08.003.

[80] Justin L. MacCallum, W. F. Drew Bennett, and D. Peter Tieleman. Distribution of amino acids in a lipid bilayer from computer simulations. *Biophys. J.*, 94(9):3393–3404, 2008. ISSN 1542-0086. doi: 10.1529/biophysj.107.112805.

[81] Christopher Maffeo, Thuy T M Ngo, Taekjip Ha, and Aleksei Aksimentiev. A coarse-grained model of unstretched single-stranded DNA derived from atomistic simulation and single-molecule experiment. *J. Chem. Theory Comput.*, 10:2891–2896, 2014. ISSN 1549-9618. doi: 10.1021/ct500193u.

[82] F. Magnani, Y. Shibata, M. J. Serrano-Vega, and C. G. Tate. Co-evolving stability and conformational homogeneity of the human adenosine a2a receptor. *Proc Natl Acad Sci U S A*, 105(31):10744–9, 2008. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.0804396105. URL https://www.ncbi.nlm.nih.gov/pubmed/18664584.

[83] D. S. Marks, T. A. Hopf, and C. Sander. Protein structure prediction from sequence variation. *Nat Biotechnol*, 30(11):1072–80, 2012. ISSN 1546-1696 (Electronic) 1087-0156 (Linking). doi: 10.1038/nbt.2419. URL https://www.ncbi.nlm.nih.gov/pubmed/23138306.

[84] Stephen S. Marshall, Michiel J. M. Niesen, A. Müller, K. Tiemann, S. M. Saladi, R. P. Galimidi, B. Zhang, W. Clemons, and Miller. T. F. A link between integral membrane protein expression and simulated integration efficiency. *Cell Reports*, 16(8):2169–2177, 2016. doi: http://dx.doi.org/10.1016/j.celrep.2016.07.042.

[85] C. Preston Moon and Karen G. Fleming. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proc Nat Acad Sci U S A*, 108(25):10174–10177, 2011. doi: 10.1073/pnas.1103979108.

[86] Tatsuya Morisaki, Kenneth Lyon, Keith F. DeLuca, Jennifer G. DeLuca, Brian P. English, Zhengjian Zhang, Luke D. Lavis, Jonathan B. Grimm, Sarada Viswanathan, Loren L. Looger, Timothee Lionnet, and Timothy J. Stasevich. Real-time quantification of single rna translation dynamics in living cells. *Science*, 352(6292):1425–1429, 2016. ISSN 0036-8075. doi: 10.1126/science.aaf0899.

[87] Michiel J. M. Niesen, Stephen S. Marshall, Thomas F. Miller, and William M. Clemons. Improving membrane protein expression by optimizing integration efficiency. *Journal of Biological Chemistry*, 292(47):19537–19545, 2017. doi: 10.1074/jbc.M117.813469. URL http://www.jbc.org/content/292/47/19537.abstract.

[88] Michiel J. M. Niesen, Connie Y. Wang, Reid C. Van Lehn, and Thomas F. Miller, III. Structurally detailed coarse-grained model for sec-facilitated cotranslational protein translocation and membrane integration. *PLOS Computational Biology*, 13(3):1–26, 03 2017. doi: 10.1371/journal.pcbi.1005427. URL https://doi.org/10.1371/journal.pcbi.1005427.

[89] Ola B Nilsson, Adrian A Nickson, Jeffrey J Hollins, Stephan Wickles, Annette Steward, Roland Beckmann, Gunnar von Heijne, and Jane Clarke. Cotranslational folding of spectrin domains via partially structured states. *Nat Struct Mol Biol*, 24(3):221–225, 03 2017. URL http://dx.doi.org/10.1038/nsmb.3355.

[90] E. E. Noens, C. Williams, M. Anandhakrishnan, C. Poulsen, M. T. Ehebauer, and M. Wilmanns. Improved mycobacterial protein production using a

mycobacterium smegmatis groel1deltac expression strain. *BMC Biotechnol*, 11:27, 2011. ISSN 1472-6750 (Electronic) 1472-6750 (Linking). doi: 10.1186/1472-6750-11-27. URL http://www.ncbi.nlm.nih.gov/pubmed/21439037.

[91] Karin Öjemalm, Takashi Higuchi, Yang Jiang, Ülo Langel, IngMarie Nilsson, Stephen H White, Hiroaki Suga, and Gunnar von Heijne. Apolar surface area determines the efficiency of translocon-mediated membrane-protein integration into the endoplasmic reticulum. *Proc Natl Acad Sci U S A*, 108(31): E359–E364, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1100120108.

[92] Karin Öjemalm, Takashi Higuchi, Patricia Lara, Erik Lindahl, Hiroaki Suga, and Gunnar von Heijne. Energetics of side-chain snorkeling in transmembrane helices probed by nonproteinogenic amino acids. *Proc Nat Acad Sci U S A*, 113(38):10559–10564, 2016. doi: 10.1073/pnas.1606776113.

[93] Griffith D. Parks and Robert A. Lamb. Topology of eukaryotic type II membrane proteins: Importance of N-terminal positively charged residues flanking the hydrophobic domain. *Cell*, 64(4):777–787, 1991. ISSN 00928674. doi: 10.1016/0092-8674(91)90507-U.

[94] Stefan Pfeffer, Laura Burbaum, Pia Unverdorben, Markus Pech, Yuxiang Chen, Richard Zimmermann, Roland Beckmann, and Friedrich Förster. Structure of the native Sec61 protein-conducting channel. *Nat. Commun.*, 6:8403, 2015. ISSN 2041-1723. doi: 10.1038/ncomms9403.

[95] M. D. Potter and C. V. Nicchitta. Endoplasmic reticulum-bound ribosomes reside in stable association with the translocon following termination of protein synthesis. *J Biol Chem*, 277(26):23314–20, 2002. ISSN 0021-9258 (Print) 0021-9258 (Linking). doi: 10.1074/jbc.M202559200. URL http://www.ncbi.nlm.nih.gov/pubmed/11964406.

[96] Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R Shirts, Jeremy C Smith, Peter M Kasson, David van der Spoel, Berk Hess, and Erik Lindahl. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29: 845–854, 2013.

[97] S. Ramasamy, R. Abrol, C. J. Suloway, and Jr. Clemons, W. M. The glove-like structure of the conserved membrane protein tatc provides insight into signal sequence recognition in twin-arginine translocation. *Structure*, 21(5):777–88, 2013. ISSN 1878-4186 (Electronic) 0969-2126 (Linking). doi: 10.1016/j.str.2013.03.004. URL http://www.ncbi.nlm.nih.gov/pubmed/23583035.

[98] Tom A Rapoport. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature*, 450(7170):663–669, 2007. ISSN 1476-4687. doi: 10.1038/nature06384.

[99] S. E. Rollauer, M. J. Tarry, J. E. Graham, M. Jaaskelainen, F. Jager, S. Johnson, M. Krehenbrink, S. M. Liu, M. J. Lukey, J. Marcoux, M. A. McDowell, F. Rodriguez, P. Roversi, P. J. Stansfeld, C. V. Robinson, M. S. Sansom, T. Palmer, M. Hogbom, B. C. Berks, and S. M. Lea. Structure of the tatc core of the twin-arginine protein transport system. *Nature*, 492(7428):210–4, 2012. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature11683. URL https://www.ncbi.nlm.nih.gov/pubmed/23201679.

[100] P. A. Romero and F. H. Arnold. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol*, 10(12):866–76, 2009. ISSN 1471-0080 (Electronic) 1471-0072 (Linking). doi: 10.1038/nrm2805. URL https://www.ncbi.nlm.nih.gov/pubmed/19935669.

[101] Anna Rychkova and Arieh Warshel. Exploring the nature of the translocon-assisted protein insertion. *Proc. Natl. Acad. Sci.*, 110(2):495–500, 2013. ISSN 1091-6490. doi: 10.1073/pnas.1220361110.

[102] Anna Rychkova, Spyridon Vicatos, and Arieh Warshel. On the energetics of translocon-assisted insertion of charged transmembrane helices into membranes. *Proc. Natl. Acad. Sci. U. S. A.*, 107(41):17598–603, 2010. ISSN 1091-6490. doi: 10.1073/pnas.1012207107.

[103] Shyam M Saladi, Nauman Javed, Axel Müller, and William M Clemons. A statistical model for improved membrane protein expression using sequence-derived features. *Journal of Biological Chemistry*, 2018. doi: 10.1074/jbc.RA117.001052. URL http://www.jbc.org/content/early/2018/01/29/jbc.RA117.001052.abstract.

[104] M. Sales-Pardo, R. Guimerà, A. A. Moreira, J. Widom, and L. A. N. Amaral. Mesoscopic modeling for nucleic acid chain dynamics. *Phys. Rev. E*, 71(5):1–13, 2005. ISSN 15393755. doi: 10.1103/PhysRevE.71.051902.

[105] I. Saraogi and S. O. Shan. Co-translational protein targeting to the bacterial membrane. *Biochim Biophys Acta*, 1843(8):1433–41, 2014. ISSN 0006-3002 (Print) 0006-3002 (Linking). doi: 10.1016/j.bbamcr.2013.10.013. URL https://www.ncbi.nlm.nih.gov/pubmed/24513458.

[106] C. A. Sarkar, I. Dodevski, M. Kenig, S. Dudli, A. Mohr, E. Hermans, and A. Pluckthun. Directed evolution of a g protein-coupled receptor for expression, stability, and binding selectivity. *Proc Natl Acad Sci U S A*, 105(39):14808–13, 2008. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.0803103105. URL http://www.ncbi.nlm.nih.gov/pubmed/18812512.

[107] J. Schaletzky and T. A. Rapoport. Ribosome binding to and dissociation from translocation sites of the endoplasmic reticulum membrane. *Mol Biol Cell*, 17(9):3860–9, 2006. ISSN 1059-1524 (Print) 1059-1524 (Linking).

doi: 10.1091/mbc.E06-05-0439. URL http://www.ncbi.nlm.nih.gov/pubmed/16822833.

[108] Jonathan P. Schlebach and Charles R. Sanders. Influence of Pathogenic Mutations on the Energetics of Translocon-Mediated Bilayer Integration of Transmembrane Helices. *J. Membr. Biol.*, 248:371–381, 2014. ISSN 0022-2631. doi: 10.1007/s00232-014-9726-0.

[109] S. Schlegel, M. Klepsch, D. Gialama, D. Wickstrom, D. J. Slotboom, and J. W. de Gier. Revolutionizing membrane protein overexpression in bacteria. *Microb Biotechnol*, 3(4):403–11, 2010. ISSN 1751-7915 (Electronic) 1751-7915 (Linking). doi: 10.1111/j.1751-7915.2009.00148.x. URL http://www.ncbi.nlm.nih.gov/pubmed/21255339.

[110] K. M. Schlinkmann, A. Honegger, E. Tureci, K. E. Robison, D. Lipovsek, and A. Pluckthun. Critical features for biosynthesis, stability, and functionality of a g protein-coupled receptor uncovered by all-versus-all mutations. *Proc Natl Acad Sci U S A*, 109(25):9810–5, 2012. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1202107109. URL https://www.ncbi.nlm.nih.gov/pubmed/22665811.

[111] Thomas H Schmidt and Christian Kandt. Lambada and inflategro2: efficient membrane alignment and insertion of membrane proteins for molecular dynamics simulations. *J. Chem. Inf. Model.*, 52(10):2657–2669, 2012.

[112] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri. Nih image to imagej: 25 years of image analysis. *Nat Methods*, 9(7):671–5, 2012. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). URL https://www.ncbi.nlm.nih.gov/pubmed/22930834.

[113] D. J. Scott, L. Kummer, D. Tremmel, and A. Pluckthun. Stabilizing membrane proteins through protein engineering. *Curr Opin Chem Biol*, 17(3):427–35, 2013. ISSN 1879-0402 (Electronic) 1367-5931 (Linking). doi: 10.1016/j.cbpa.2013.04.002. URL http://www.ncbi.nlm.nih.gov/pubmed/23639904.

[114] Susanna Seppälä, Joanna S Slusky, Pilar Lloris-Garcerá, Mikaela Rapp, and Gunnar von Heijne. Control of membrane protein topology by a single c-terminal residue. *Science*, 2010. doi: 10.1126/science.1188950. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=20508091.

[115] M. J. Serrano-Vega, F. Magnani, Y. Shibata, and C. G. Tate. Conformational thermostabilization of the beta1-adrenergic receptor in a detergent-resistant form. *Proc Natl Acad Sci U S A*, 105(3):877–82, 2008. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.0711253105. URL https://www.ncbi.nlm.nih.gov/pubmed/18192400.

[116] Sichen Shao and Ramanujan S Hegde. Membrane protein insertion at the endoplasmic reticulum. *Annu. Rev. Cell Dev. Biol.*, 27:25–56, 2011. ISSN 1530-8995. doi: 10.1146/annurev-cellbio-092910-154125.

[117] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol Syst Biol*, 7:539, 2011. ISSN 1744-4292 (Electronic) 1744-4292 (Linking). doi: 10.1038/msb.2011.75. URL http://www.ncbi.nlm.nih.gov/pubmed/21988835.

[118] Jan Spitzer and Bert Poolman. The role of biomacromolecular crowding, ionic strength, and physicochemical gradients in the complexities of life's emergence. *Microbiol. Mol. Biol. Rev.*, 73(2):371–88, 2009. ISSN 1098-5557. doi: 10.1128/MMBR.00010-09.

[119] Douglas B. Staple, Stephen H. Payne, Andrew L C Reddin, and Hans Jürgen Kreuzer. Model for stretching and unfolding the giant multidomain muscle protein using single-molecule force spectroscopy. *Phys. Rev. Lett.*, 101(24): 1–4, 2008. ISSN 00319007. doi: 10.1103/PhysRevLett.101.248301.

[120] C J Stirling, J Rothblatt, M Hosobuchi, R Deshaies, and R Schekman. Protein translocation mutants defective in the insertion of integral membrane proteins into the endoplasmic reticulum. *Mol. Biol. Cell*, 3(2):129–142, 1992. doi: 10.1091/mbc.3.2.129.

[121] J. A. Swets, R. M. Dawes, and J. Monahan. Better decisions through science. *Sci Am*, 283(4):82–7, 2000. ISSN 0036-8733 (Print) 0036-8733 (Linking). URL https://www.ncbi.nlm.nih.gov/pubmed/11011389.

[122] Yoshiki Tanaka, Yasunori Sugano, Mizuki Takemoto, Takaharu Mori, Arata Furukawa, Tsukasa Kusakizako, Kaoru Kumazaki, Ayako Kashima, Ryuichiro Ishitani, Yuji Sugita, Osamu Nureki, and Tomoya Tsukazaki. Crystal Structures of SecYEG in Lipidic Cubic Phase Elucidate a Precise Resting and a Peptide-Bound State. *Cell Rep.*, 13(8):1561–1568, 2015. ISSN 22111247. doi: 10.1016/j.celrep.2015.10.025.

[123] D. J. Tantillo, J. Chen, and K. N. Houk. Theozymes and compuzymes: theoretical models for biological catalysis. *Curr Opin Chem Biol*, 2(6):743–50, 1998. ISSN 1367-5931 (Print) 1367-5931 (Linking). URL https://www.ncbi.nlm.nih.gov/pubmed/9914196.

[124] M. L. Tress and A. Valencia. Predicted residue-residue contacts can help the scoring of 3d models. *Proteins*, 78(8):1980–91, 2010. ISSN 1097-0134 (Electronic) 0887-3585 (Linking). doi: 10.1002/prot.22714. URL https://www.ncbi.nlm.nih.gov/pubmed/20408174.

[125] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.*, 185(2):604–613, 2014. ISSN 0010-4655. doi: http://dx.doi.org/10.1016/j.cpc.2013.09.018.

[126] Fabio Trovato and Edward P. O'Brien. Insights into cotranslational nascent protein behavior from computer simulations. *Annu. Rev. of Biophys*, 45 (1):345–369, 2016. doi: 10.1146/annurev-biophys-070915-094153. PMID: 27297399.

[127] Steven F. Trueman, Elisabet C. Mandon, and Reid Gilmore. A gating motif in the translocation channel sets the hydrophobicity threshold for signal sequence function. *J. Cell Biol*, 199(6):907–918, 2012. doi: 10.1083/jcb.201207163.

[128] K. D. Tsirigos, C. Peters, N. Shu, L. Kall, and A. Elofsson. The topcons web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res*, 43(W1):W401–7, 2015. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkv485. URL http://www.ncbi.nlm.nih.gov/pubmed/25969446.

[129] Reid C. Van Lehn, Bin Zhang, and Thomas F. Miller. Regulation of multi-spanning membrane protein topology via post-translational annealing. *eLife*, 4:1–23, 2015. ISSN 2050084X. doi: 10.7554/eLife.08697.

[130] G. von Heijne. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J*, 5(11):3021–7, 1986. ISSN 0261-4189 (Print) 0261-4189 (Linking). URL http://www.ncbi.nlm.nih.gov/pubmed/16453726.

[131] Gunnar von Heijne. Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature*, 341:456–458, 1989. ISSN 0028-0836. doi: 10.1038/341456a0.

[132] Gunnar von Heijne. Membrane-protein topology. *Nat. Rev. Mol. Cell Biol.*, 7(12):909–918, 2006. ISSN 1471-0072. doi: 10.1038/nrm2063.

[133] Rebecca M. Voorhees, Israel S. Fernández, Sjors H. W. Scheres, and Ramanujan S. Hegde. Structure of the mammalian ribosome-Sec61 complex to 3.4 A resolution. *Cell*, 157(7):1632–1643, 2014. ISSN 10974172. doi: 10.1016/j.cell.2014.05.024.

[134] S. Wagner, M. L. Bader, D. Drew, and J. W. de Gier. Rationalizing membrane protein overexpression. *Trends Biotechnol*, 24(8):364–71, 2006. ISSN 0167-7799 (Print) 0167-7799 (Linking). doi: 10.1016/j.tibtech.2006.06.008. URL http://www.ncbi.nlm.nih.gov/pubmed/16820235.

[135] G. S. Waldo, B. M. Standish, J. Berendzen, and T. C. Terwilliger. Rapid protein-folding assay using green fluorescent protein. *Nat Biotechnol*, 17(7): 691–5, 1999. ISSN 1087-0156 (Print) 1087-0156 (Linking). doi: 10.1038/10904. URL http://www.ncbi.nlm.nih.gov/pubmed/10404163.

[136] Z. Wang, Q. Xiang, G. Wang, H. Wang, and Y. Zhang. Optimizing expression and purification of an atp-binding gene gsia from escherichia coli k-12 by using gfp fusion. *Genet Mol Biol*, 34(4):661–8, 2011. ISSN 1678-4685 (Electronic) 1415-4757 (Linking). doi: 10.1590/S1415-47572011005000043. URL http://www.ncbi.nlm.nih.gov/pubmed/22215971.

[137] T. Warne, M. J. Serrano-Vega, J. G. Baker, R. Moukhametzianov, P. C. Edwards, R. Henderson, A. G. W. Leslie, C. G. Tate, and G. F. X. Schertler. Structure of a beta(1)-adrenergic g-protein-coupled receptor. *Nature*, 454 (7203):486–U2, 2008. ISSN 0028-0836. doi: 10.1038/nature07101. URL <GotoISI>://WOS:000257860300044.

[138] John D. Weeks, David Chandler, and Hans C. Andersen. Role of repulsive forces in determining the equilibrium structure of simple liquids. *J. Chem Phys*, 54(12):5237–5247, 1971. doi: 10.1063/1.1674820.

[139] S H White and W C Wimley. Membrane protein folding and stability: Physical principles. *Annu. Rev. Biophys. Biomol. Str.*, 28:319–365, 1999.

[140] Stephen H White and Gunnar von Heijne. How translocons select transmembrane helices. *Annu. Rev. Biophys.*, 37:23–42, 2008. ISSN 1936-122X. doi: 10.1146/annurev.biophys.37.032807.125904.

[141] William C Wimley and Stephen H White. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature*, 3(10):842–848, 1996. ISSN 1072-8368. doi: 10.1038/nsb1096-842.

[142] N. B. Woodall, Y. Yin, and J. U. Bowie. Dual-topology insertion of a dual-topology membrane protein. *Nat Commun*, 6:8099, 2015. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/ncomms9099. URL http://www.ncbi.nlm.nih.gov/pubmed/26306475.

[143] Bin Wu, Carolina Eliscovich, Young J. Yoon, and Robert H. Singer. Translation dynamics of single mrnas in live cells and neurons. *Science*, 352(6292): 1430–1435, 2016. ISSN 0036-8075. doi: 10.1126/science.aaf1084.

[144] Semen O. Yesylevskyy, Lars V. Schäfer, Durba Sengupta, and Siewert J. Marrink. Polarizable water model for the coarse-grained MARTINI force field. *PLoS Comput. Biol.*, 6(6):1–17, 2010. ISSN 1553734X. doi: 10.1371/journal.pcbi.1000810.

[145] B. Zhang and T. F. Miller. Hydrophobically stabilized open state for the lateral gate of the Sec translocon. *Proc. Natl. Acad. Sci.*, 107(12):5399–5404, 2010. ISSN 0027-8424. doi: 10.1073/pnas.0914752107.

[146] Bin Zhang and Thomas F. Miller. Long-Timescale Dynamics and Regulation of Sec-Facilitated Protein Translocation. *Cell Rep.*, 2(4):927–937, 2012. ISSN 22111247. doi: 10.1016/j.celrep.2012.08.039.

[147] Bin Zhang and Thomas F. Miller. Direct simulation of early-stage sec-facilitated protein translocation. *J. Am. Chem. Soc.*, 134(33):13700–13707, 2012. ISSN 00027863. doi: 10.1021/ja3034526.

Online Data

[1]Simulated sequences: https://doi.org/10.1371/journal.pcbi.1005427.s009

[2]3D-CG model channel coordinates: https://doi.org/10.1371/journal.pcbi.1005427.s008

[3]TatC chimera sequences: https://ars.els-cdn.com/content/image/1-s2.0-S2211124716309603-mmc2.xlsx

[4]TatC sequences discussed in Chapter 5: http://www.jbc.org/content/suppl/2017/09/16/M117.813469.DC1/SI_Data.xlsx