Title

Sequence-dependent scale for translocon-mediated insertion of interfacial helices in membranes.

Journal

Science Advances, 11(8)

Authors

Grau, Brayan
Kormos, Rian
Bañó-Polo, Manuel
et al.

Publication Date

2025-02-21

DOI

10.1126/sciadv.ads6804

Peer reviewed

BIOCHEMISTRY

Sequence-dependent scale for translocon-mediated insertion of interfacial helices in membranes

Brayan Grau[1]†, Rian Kormos[2]†, Manuel Bañó-Polo[1]‡, Kehan Chen[2]‡, María J. García-Murria[1], Fatlum Hajredini[3], Manuel M. Sánchez del Pino[1], Hyunil Jo[2], Luis Martínez-Gil[1], Gunnar von Heijne[4], William F. DeGrado[1]*, Ismael Mingarro[1,2]*

Biological membranes consist of a lipid bilayer studded with integral and peripheral membrane proteins. Most α-helical membrane proteins require protein-conducting insertases known as translocons to assist in their membrane insertion and folding. While the sequence-dependent propensities for a helix to either translocate through the translocon or insert into the membrane have been codified into numerical hydrophobicity scales, the corresponding propensity to partition into the membrane interface remains unrevealed. By engineering diagnostic glycosylation sites around test peptide sequences inserted into a host protein, we devised a system that can differentiate between water-soluble, surface-bound, and transmembrane (TM) states of the sequence based on its glycosylation pattern. Using this system, we determined the sequence-dependent propensities for transfer from the translocon to a TM, interfacial, or extramembrane space and compared these propensities with the corresponding probability distributions determined from the sequences and structures of experimentally determined proteins.

INTRODUCTION

While the sequence characteristics required for translocon-mediated insertion and stabilization of transmembrane (TM) helices in membrane proteins have been extensively studied for decades (*1*), there has been a paucity of corresponding data examining the insertion of sequence elements such as amphiphilic α helices into the membrane interface. The dynamic equilibrium between bulk water, the membrane surface, and TM states represents a delicate balance that is essential for the function of peptides such as antimicrobial peptides (AMPs) and lytic peptides. Moreover, proteins with helical fusogenic sequences similarly partition between water-soluble, membrane-associated, and TM fusion pore-forming states (*2*, *3*). Other surface-interacting peptides are known to stabilize membrane curvature (*4*). Thus, elucidating the features that dictate membrane surface versus TM associations is essential to a wide swath of natural proteins (*5*, *6*).

Biological membranes can be divided into two regions based on physicochemical properties: the highly hydrophobic core formed mainly by the lipid acyl chains and the interfaces on both sides of this central region that contain the polar head groups (*7*). The hydrophobic effect is the primary driving force of membrane partitioning to the former, but much less is known about the energetics of interface partitioning. The combined thickness of the interfaces is similar to the thickness of the hydrophobic core (~30 Å), and this region is able to accommodate unfolded and folded polypeptide chain (*8*). The interfacial region, occupied by the lipid headgroups and the associated hydration layer, is highly physically anisotropic. Because the interfaces are rich in groups with different chemical properties, a polypeptide chain in this region faces an environment enriched in possibilities to establish a variety of noncovalent interactions.

Translocons have evolved to recognize and insert hydrophobic TM segments into the bilayer in a manner that optimizes side-chain interactions with both the hydrophobic core and interfacial regions (*6*, *9*, *10*). Structural data have provided detailed insights concerning the insertion into the hydrophobic sector of the bilayer, showing that TM segments leave the translocon through a lateral gate in the channel wall that opens laterally toward the bilayer (*6*, *11*–*13*). Classically, it has been assumed that the translocation-dependent insertion and folding of TM proteins is a two-state process in which helices that are inserted in the translocon move directly into the membrane in a fully inserted state (*14*, *15*). The native state of a membrane protein then emerges from an ensemble of TM helices. However, recent work has suggested that large portions of the protein chain sample a denatured, surface-absorbed state during the folding process (*16*). Thus, it is important to understand the energetics of the surface-absorbed state relative to the aqueous and helical TM states in order to gain a more complete understanding of membrane protein folding.

To achieve a quantitative description of membrane protein insertion and folding, it becomes necessary to unravel the molecular processes by which a polypeptide segment exiting the translocon adopts a TM orientation (spanning the membrane) versus sliding toward the membrane surface in an interfacial disposition (*6*, *17*). Once a detailed description of TM segment recognition by the translocon has been established (*18*, *19*), a fundamental requirement for a quantitative characterization of protein sliding and folding in membrane interfaces is a suitable interfacial hydrophobicity scale. In this study, we present such a scale for the 20 natural amino acids derived from quantitative measurements of the translocon-mediated protein integration pathway into the endoplasmic reticulum (ER) membrane.

To develop this scale, we challenged the translocon both in vitro and in cellular membranes with a set of designed polypeptide sequences. Our designs are based on a peptide sequence derived from bacteriorhodopsin helix C (bRc) that out of the protein context does not insert into a lipid bilayer as a membrane-spanning helix

[1]Institute for Biotechnology and Biomedicine (BIOTECMED), Department of Biochemistry and Molecular Biology, University of Valencia, E-46100 Burjassot, Spain. [2]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA. [3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA. [4]Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, SE-10691 Stockholm, Sweden.
*Corresponding author. Email: Ismael.Mingarro@uv.es (I.M.); Bill.DeGrado@ucsf.edu (W.F.D.)
†These authors contributed equally to this work.
‡These authors contributed equally to this work.

at physiological pH but rather adopts a surface configuration with a high helical content in the presence of 1,2-dioleoyl-sn-glycero-3-phosphocholine (DOPC) liposomes (*20*). We then substituted different amino acids into this peptide sequence to modulate the energetics of partitioning into the TM versus surface state. Through the rational design of glycosylation sites, we devised a system that can differentiate among water-soluble, surface-bound, and TM states of the peptide sequence based on its glycosylation pattern. A quantitative analysis of the mole fraction of the protein in each state allows one to compute an apparent free energy of transfer (ΔG_{app}) from water to both the surface-absorbed and TM states for each of the 20 natural amino acids.

We compare the data from our ΔG_{app} scales with those obtained from (i) previous studies of translocon-mediated TM insertion (*18*, *19*); (ii) biophysical measurements (*21*, *22*) of partitioning into organic solvents and the surface of phospholipid bilayers; and (iii) previous (*23*) statistical analyses of the depth-dependent distribution of amino acids in membrane proteins of known structures. Our biologically and statistically derived scales show reasonable agreement with biophysical scales, with some interesting exceptions associated with the tendency of basic and aromatic residues to associate with the interfacial region of membrane bilayers. These sequence-specific propensities for the interface, even among residues that are charged or have expansive hydrophobic surface area, appear to enable more precise targeting of helices to the interface than could be achieved by merely selecting sequences with intermediate hydrophobicity. The derived scale is then shown to be useful for analysis of membrane-associated helical proteins.

RESULTS

Development of an assay to assess TM, interfacial, and water-soluble orientations

Previously, Hessa *et al.* (*18*, *19*) developed a "biological" hydrophobicity scale based on an in vitro assay for quantifying the efficiency of translocon-mediated membrane integration of TM helices into dog pancreas rough microsomes (RMs). In this method, a test sequence is engineered into the luminal P2 domain of the integral membrane protein leader peptidase (LepB), where it is flanked by two acceptor sites for N-linked glycosylation (Fig. 1A). The degree of membrane integration of the test sequence is quantified from SDS–polyacrylamide gel electrophoresis (SDS-PAGE) gels by measuring the fraction of singly versus doubly glycosylated LepB molecules. If the test sequence is inserted into the membrane adopting a TM orientation, then only one site (designated the G1 site) is glycosylated, while double glycosylation at G1 and G2 (or G2′) is observed when the test sequence fails to insert into the membrane and instead is translocated to the lumen. Here, we wished to additionally probe the sequence requirements for the test sequence to associate tightly with the luminal surface of the membrane. We hypothesized that a G2 site proximal to the test sequence would be protected from glycosylation if the test sequence associated tightly to the membrane (*24*, *25*). As a test sequence, we used a bRc-derived interfacial peptide, developed by Musial-Siwek *et al.* (*20*), which is helical in the presence of 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) liposomes (fig. S1) and, depending on conditions, capable of adopting a TM or interfacial conformation. bRc is considerably less amphiphilic than typical surface-seeking peptides such as melittin (bRc has a hydrophobic moment $m_H = 0.35$, melittin has $m_H = 0.53$, fig. S2), and is hence poised close to the threshold between TM and interfacial orientations. Thus, even minor changes in sequence should be able to push it one way or the other, making it an ideal test sequence for measuring the propensities of different amino acids for promoting interfacial versus TM orientations.

We thus began by testing the bRc sequence flanked by GGPG and GPGG tetrapeptides to insulate the interfacial segment from the surrounding sequence (*18*), placing the G2 site at two different positions. We first placed the G2 site only five residues after the test sequence such that it would not be accessible to the luminal oligosaccharyl transferase (OST) active site (*24*, *26*) if the test sequence partitioned into the interfacial region of the membrane. In this construct, with bRc as the test sequence, the G2 site was protected from glycosylation, while the G1 site remained efficiently glycosylated (Fig. 1B, lane 2). Proteinase K digestion revealed a luminal disposition of the test sequence, with the LepB P2 domain protected from degradation (Fig. 1B, lane 3). Moving the second glycosylation site further away from the tested sequence (G2′) yielded mainly doubly glycosylated molecules resistant to proteinase K treatment (Fig. 1B, lanes 5 and 6), confirming the luminal location of the P2 domain. These results were confirmed using a second peptide, melittin, known to form an interfacial helix (*27*, *28*) as the test sequence (fig. S3), which was also mostly protected from glycosylation at G2, accessible for glycosylation at G2′, and with the P2 loop protected from proteinase K degradation.

To further validate the assay, the "minimal glycosylation distance" for interfacial sequences was measured by placing the G2 site at different positions upstream and downstream the bRc sequence (fig. S4). We found that efficient glycosylation was observed when the Asn residue in the G2 site was placed at least 11 residues downstream (Fig. 1C and fig. S4) and at least 9 residues upstream of the bRc sequence (Fig. 1C and fig. S4).

bRc-derived sequence as a vehicle to study amino acid propensities

In order to develop a strategy to measure the effect of amino acid substitutions on membrane integration, we next modified the test sequence from the bRc peptide to favor either TM or fully exposed (luminal) conformations. The interfacial location of bRc is believed to result from three Asp residues, D6, D11, and D18 (Fig. 1A), and single substitutions within this peptide changed its disposition in previous studies with synthetic peptides (*20*) and in vitro transcription/translation experiments (*29*). In agreement with the earlier data, we found that the D11L and D18L variants, in which a single charged Asp is changed to an apolar Leu, becomes fully TM (i.e., the singly glycosylated band predominates and no proteinase K protection is observed, fig. S5B, lanes 2, 3, 11, and 12, in contrast to the WT sequence in Fig. 1B, lanes 2 and 3). Conversely, replacing an existing Leu with an Asp (variants L12D and L17D) shifts the glycosylation pattern to the doubly glycosylated, translocated state with the P2 domain protected from proteinase K degradation (fig. S5B, lanes 5, 6, 14, and 15, respectively). When both Asp and Leu residues in positions 11/12 or 17/18 were replaced by Trp residues (D11W/L12W or L17W/D18W), which strongly prefer the membrane interface (*21*), we observed singly glycosylated forms and PK treatment protection (fig. S5B, lanes 7 to 9 and 16 to 18, respectively), supporting a primarily interfacial disposition. Those results confirm that the bRc-derived sequence is able to provide a suitable vehicle to study the contribution of any single amino acid along its sequence, being

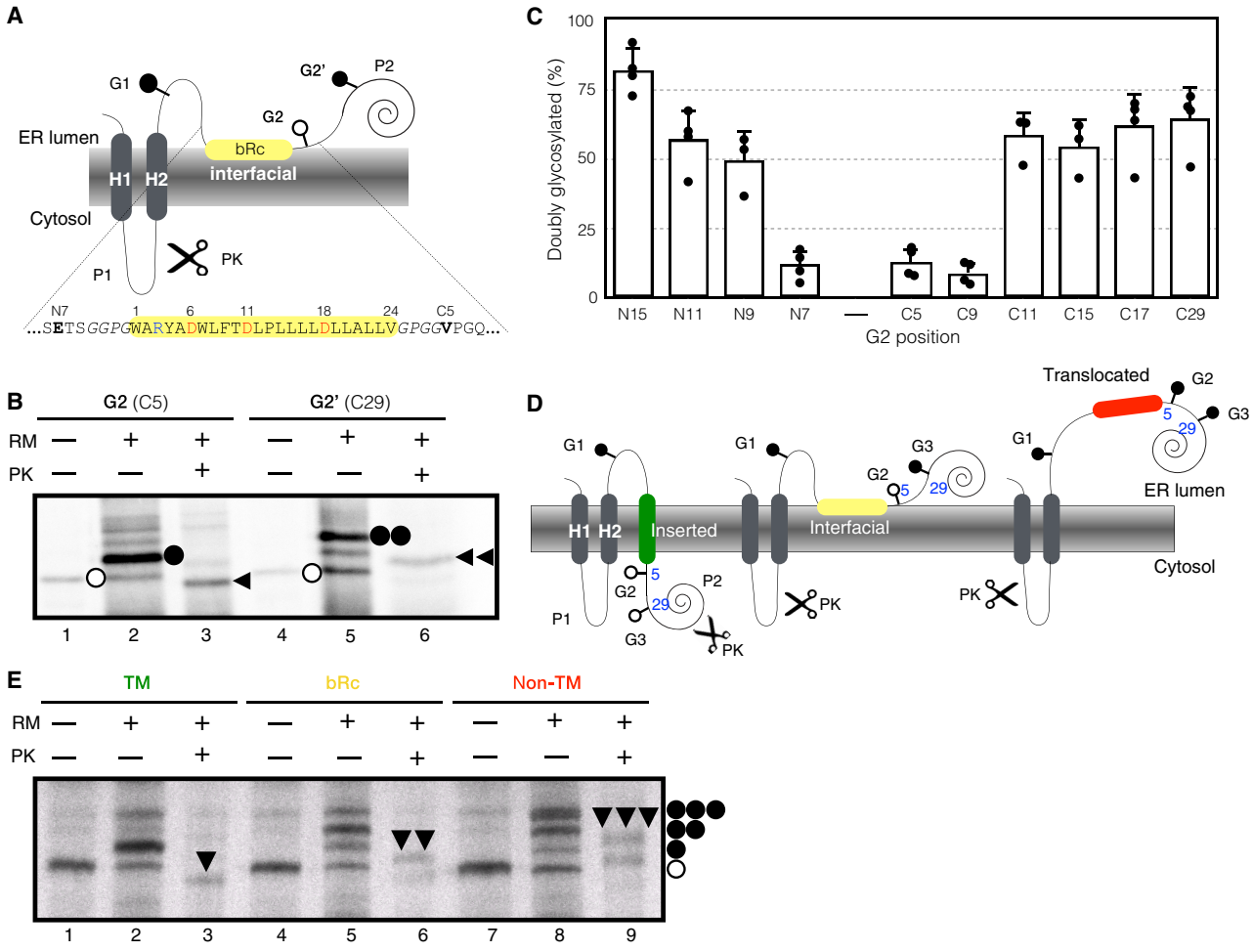

Fig. 1. Interfacial segment disposition in the microsomal membrane via a glycosylation-based assay. (**A**) Schematic of the modified glycosylation-based assay introduced by Hessa *et al.* (*18*, *19*) for distinguishing TM and aqueous peptide dispositions. Wild-type Lep contains two N-terminal TM segments (H1 and H2) and a luminal domain (P2) where the interfacial segment is inserted (yellow). Glycosylation sites, placed N-terminal (G1) and C-terminal (G2 and G2s′), indicate the membrane insertion topology. G1 is fixed at positions 96 to 98, while G2 is near the insulating GPGG sequence or 29 residues downstream, reflecting the final disposition of tested sequences. (**B**) Plasmids encoding Lep/bRc constructs were transcribed and translated in vitro with (+) or without (–) rough microsomes (RM) and treated with proteinase K (PK). The tested bRc sequence is highlighted in yellow. Nonglycosylated proteins are marked by a white dot; singly and doubly glycosylated proteins are marked by one or two black dots. Arrowheads identify PK-protected fragments: one for singly and two for doubly glycosylated fragments. (**C**) Minimal glycosylation distance for interfacial sequences was determined by moving G2 across the N-terminal (N15, N11, N9, and N7) and C-terminal (C5, C9, C11, C15, C17, and C29) position of bRc. Absence of G2 is marked as "–". Error bars represent S from ≥3 experiments. Black and white rectangles indicate when doubly glycosylated molecules dominate or do not, respectively. (**D**) Schematic of the LepG3 construct, where glycosylation sites (G1, G2, and G3) reflect the final peptide disposition (inserted, surface, and translocated). (**E**) In vitro translation of LepG3 constructs with TM (turnip crinkle virus) (*31*), bRc, and translocated pseudo-randomized sequence (*51*) was performed with (+) or without (–) RM. Non-, single-, double-, and triple-glycosylated proteins are marked by white or black dots (one to three). PK assay fragments are protected by glycosylation are identified by arrowheads (two or three for double or triple-glycosylated proteins).

its final membrane disposition successfully switched with punctual mutations. Ultimately, we chose the more centered 11 and 12 positions over the more peripheral 17 and 18 positions, because the latter bias the initiation of helix formation (*30*).

Developing the LepG3 assay for a multiple orientation assessment in a single experiment

To facilitate experimental assessment of TM, interfacial, and water-soluble orientations, we introduced a third glycosylation site G3 (Fig. 1D) in all subsequent experiments, taking into account the minimal glycosylation distance from the previously assayed interfacial sequences (Fig. 1C and fig. S4). While the G2 site was kept 5 residues away from the test sequence, the extra G3 site was placed 29 residues away from the test sequence (the farthest tested position in the minimal glycosylation distance screening) to be able to determine whether the C terminus of the construct was luminal (as in a surface-absorbed state) or cytoplasmic (as in the TM state). With these three sites (G1, G2, and G3) simultaneously present (LepG3), the glycosylation state can be used to determine whether the test sequence in the translation product is TM, interfacial, or soluble (Fig. 1D). In the TM state, the G2 and G3 sites locate to the cytosol, and hence, only the G1 site is glycosylated, leading to a singly glycosylated band on an SDS-PAGE

electrophoresis gel. If the test sequence instead associates tightly with the luminal membrane surface, a doubly glycosylated product is observed due to modifications at G1 and G3. Last, a triply glycosylated product is the predominant product when the test sequence is fully translocated into the luminal space. The expected single-glycosylation pattern was observed when the test sequence was substituted with a known TM sequence (*31*) (Fig. 1E, TM, lane 2), and the triply glycosylated pattern predominates when a polar C-terminal peptide sequence from LepB served as the test sequence (Fig. 1E, non-TM, lane 8). Moreover, as expected, the bRc sequence gave a mixture of singly, doubly, and triply glycosylated products, with doubly glycosylated molecules being the most abundant (Fig. 1E, bRc, lane 5). Protection of the P2 domain from proteinase K digestion confirmed this membrane disposition (Fig. 1E, lane 6).

Once the LepG3 design was established, we also challenged our assay with a set of 19-residue test sequences composed entirely of Ala and Leu (*18*), with compositions ranging from two to five Leu residues. This family of sequences showed a clear transition from triple to single glycosylation as the number of Leu residues was increased (fig. S6), showing that the assay using model Ala and Leu hydrophobic sequences can resolve the state preferences of sequences that partition into a mixture of soluble and TM states, but not the interfacial state.

We further established the suitability in our LepG3 assay of bRc-derived sequence as a backbone to study the contribution of single amino acids. We substituted Asp at position 11 with Leu and Ala (fig. S7, constructs #1 and #2), which increased the fraction in the TM-inserted state. We next increased the number of Asp residues in the peptide by substituting the hydrophobic residues at positions 12 to 17 with increasing numbers of Asp residues (fig. S7, constructs #4 to #7). The consecutive aspartate substitutions shift the banding pattern toward the fully exposed, triply glycosylated state, reaching a maximum after only two Asp substitutions are introduced, which creates a triplet of Asp residues considering the presence of the original D11. This result again shows the sensitivity of the system to small changes.

We next compared the effect of replacing the original residues D11 and L12 at the center of the bRc-derived sequence with pairs of identical amino acids or residues D11, L12, and P13 with triplets of identical amino acids (fig. S8). Triplet substitutions were chosen to maximize the effects, displaying a clearer glycosylation pattern than pair substitutions (fig. S8). For the hydrophobic leucine triplets, the singly glycosylated form was predominant (fig. S8C, lane 2), indicative of a TM disposition, whereas the tryptophan triplet mainly produced doubly glycosylated molecules (fig. S8C, lane 5), indicating a surface location. Last, the construct harboring three Asp residues yielded triply glycosylated bands (fig. S8C, lane 8) as expected for the fully exposed, luminal orientation of this sequence. We therefore chose the LepG3 system for a systematic study of the effects of substitutions on membrane association and insertion, and focused on triplet substitutions because they display a clearer glycosylation pattern than pair substitutions. As seen in fig. S2, in the triplet substitutions, one of the three substituted amino acids is located on the hydrophobic face of the bRc helix.

While different amino acids have different helix propensities in water-soluble proteins, only Pro often distorts the conformation of TM helices (*32*, *33*). The DDD-substituted bRc helix showed helical conformation in the presence of POPC liposomes (fig. S1), despite the fact that Asp has one of the lowest helical propensities (only Gly and Pro have lower) (*34*). Thus, all of the triplet-substituted bRc constructs tested below (except possibly the PPP substitution) likely adopt a helical conformation in both the TM and interfacial states.

Biological interfacial and TM scales

We used triplet substitutions in the bRc-derived sequence to quantify the mole fractions of the three glycosylation states, from which we computed the apparent free energies of transfer of each of the 20 naturally occurring amino acid side chains from water to the membrane interface ($\Delta G_{app}^{Wat\rightarrow Int}$) or to the hydrophobic core of the membrane ($\Delta G_{app}^{Wat\rightarrow TM}$). First, we investigated the contribution of each position within the triplet, and results obtained when one to three Gly residues are introduced in an Ala triplet showed an overall linear relationship between the number of Gly residues and ΔG_{app}, a simple outcome consistent with energy additivity (fig. S9). As described in Materials and Methods, we evaluated a total of 112 measurements on 27 variants using a rabbit reticulocyte-derived translation system to obtain the biologically derived in vitro scale. As expected, the $\Delta G_{app}^{Wat\rightarrow TM}$ values obtained from the three-state system (Fig. 2A and table S1) correlate well with the depth-dependent insertion scale of Hessa *et al.* (*18*, *19*). The $\Delta G_{app}^{Wat\rightarrow TM}$ scale correlates best with the Hessa values for transfer of an amino acid to locations near the center of the bilayer ($R^2 = 0.78$ to 0.79 within the range spanning ±4.5 Å of the center, Fig. 2, B and D). By contrast, the $\Delta G_{app}^{Wat\rightarrow Int}$ values correlate best with the corresponding values for transfer of a side chain from water to a depth more consistent with the boundary between the bilayer core and the interface regions ($R^2 = 0.79$ to 0.81 within the pair of ranges spanning 7.5 ± 1.5 Å and −7.5 ± 1.5 Å from the center, Fig. 2, C and D). These findings are in good agreement with our expectation that $\Delta G_{app}^{Wat\rightarrow TM}$ values should correlate with the center of the bilayer, while $\Delta G_{app}^{Wat\rightarrow Int}$ values should reflect transfer to a more interfacial location in the bilayer.

The current study uses an experimental scale to evaluate the effects of substitutions on an interfacially oriented versus a TM-inserted helical sequence under identical experimental conditions. Thus, it is of interest to compare the interfacial and TM scales (table S1). A plot of $\Delta G_{app}^{Wat\rightarrow Int}$ versus $\Delta G_{app}^{Wat\rightarrow TM}$ (Fig. 2E) shows that the two scales are well correlated ($R^2 = 0.93$), as anticipated from the fact that both are measures of the transfer of a side chain from water to a more apolar environment. The trendline has a slope near unity (1.27), indicating that $\Delta G_{app}^{Wat\rightarrow Int}$ is similar in magnitude to $\Delta G_{app}^{Wat\rightarrow TM}$.

Despite the high overall correlation between $\Delta G_{app}^{Wat\rightarrow Int}$ and $\Delta G_{app}^{Wat\rightarrow TM}$, closer comparison of the two scales reveals important information about the particular amino acids (Fig. 2E, bold annotations) that most strongly favor an interfacial versus TM disposition. The positively charged residues Arg and Lys showed the largest deviation from the regression trendline, favoring the interfacial region over a TM orientation. This finding is in good agreement with their positive charge, which is complementary to the anionic membrane lipids and compatible with the well-documented phenomenon of "snorkeling," whereby the cations at the end of the long side chains of these amino acids extend out of the interface into water (*35*–*37*). Consistent with this observation, the anionic residues Asp and Glu have less favorable values of $\Delta G_{app}^{Wat\rightarrow Int}$ compared to Arg and Lys (Fig. 2A), though only Glu shows a substantial deviation from the trendline (Fig. 2E). Tyr and Trp also show significant deviations favoring the surface orientation, as they have an amphiphilic structure with a polar OH or indole NH, respectively, connected to an otherwise apolar aromatic core side chain. Amphiphilicity would appear

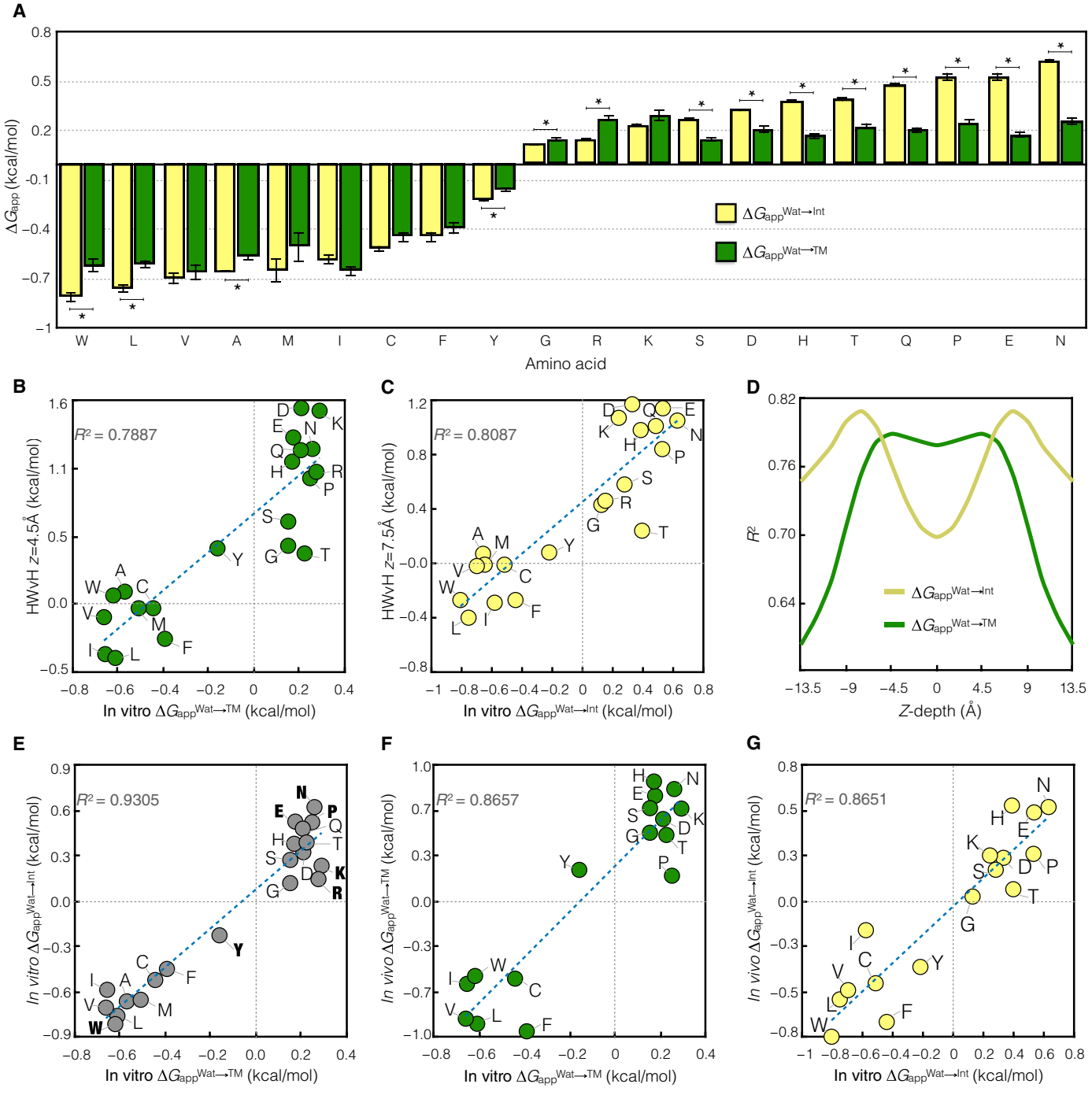

Fig. 2. Biological and biophysical ΔG_{app} scales. (**A**) ΔG_{app} of transferring for each amino acid type from the aqueous translocon channel to the membrane interface ($\Delta G_{app}^{Wat \rightarrow Int}$, yellow bars) and to the hydrophobic core of the membrane ($\Delta G_{app}^{Wat \rightarrow TM}$, green bars). Amino acids for which the difference between $\Delta G_{app}^{Wat \rightarrow Int}$ and $\Delta G_{app}^{Wat \rightarrow TM}$ was statistically significant are marked with asterisks. (**B**) Correlation between $\Delta G_{app}^{Wat \rightarrow TM}$ and the position-dependent scale of Hessa *et al.* (*18*, *19*) at a position ±3 residues, or equivalently 4.5 Å, from the center of the bilayer. (**C**) Correlation between $\Delta G_{app}^{Wat \rightarrow Int}$ and the position-dependent scale of Hessa *et al.* (*18*, *19*) at a position ±5 residues, or equivalently 7.5 Å, from the center of the bilayer. (**D**) Coefficient of determination (R^2) values for linear least squares trendlines between all position-dependent scales of Hessa *et al.* (*19*) and $\Delta G_{app}^{Wat \rightarrow Int}$ (yellow line) or $\Delta G_{app}^{Wat \rightarrow TM}$ (green line). (**E**) Correlation between $\Delta G_{app}^{Wat \rightarrow TM}$ and $\Delta G_{app}^{Wat \rightarrow Int}$. Amino acids with major dispersion of the trendline are marked in bold. (**F**) Correlation between in vivo $\Delta G_{app}^{Wat \rightarrow Int}$ scale and in vitro $\Delta G_{app}^{Wat \rightarrow Int}$ scale. (**G**) Correlation between in vivo $\Delta G_{app}^{Wat \rightarrow TM}$ scale and in vitro $\Delta G_{app}^{Wat \rightarrow TM}$ scale.

to be more important than aromatic character, as Phe does not deviate strongly from the trendline. Last, helix-breaking residues such as Pro and Asn are destabilizing to a surface association. This finding might be at least partially a result of the coil-to-helix orientation that accompanies the binding of peptides to the membrane interface (*20*, *38*, *39*), which is expected to be reflected in the coupled energetics associated with such binding. By contrast, a helix is the default orientation for sequences as they are transmitted from the translocon to a TM orientation, and even Pro substitutions are easily accommodated in the TM helices of membrane proteins (*25*, *32*).

Development of biological scales for $\Delta G_{app}^{Wat \rightarrow TM}$ and $\Delta G_{app}^{Wat \rightarrow Int}$ based on expression in mammalian cells

While the above studies were conducted using an in vitro assay, it was important to ascertain that the results extend to a cellular milieu. To address this question, we tagged appropriated control sequences plus 17 representative variants of bRc and expressed them in human embryonic kidney (HEK) 293T cells to measure an in vivo scale. As shown in fig. S10A, an unambiguous glycosylation pattern arises, with the construct harboring the TM sequence being singly glycosylated, the one encoding the bRc variant sequence being doubly glycosylated, and the construct harboring a non-TM sequence being triply glycosylated, denoting an inserted, interfacial, and translocated location, respectively. As before, triplet substitutions were chosen to maximize the effects and because they displayed a clearer glycosylation pattern than pair substitutions in mammalian cells (fig. S10, B and C). The in vitro and in vivo scales are highly correlated ($R^2 = 0.87$ for both, Wat➔Int and Wat➔TM, Fig. 2, F and G, respectively), and the deviations from linearity were largely within the experimental error seen in the individual in vitro measurements (Fig. 2A). These deviations can also arise from the presence in the cellular system of targeting factors and/or insertase components not included in the in vitro microsomal membranes. These findings indicate that the energetics measured in vitro are quite predictive of those observed in a cellular milieu.

Comparison of biological interfacial and TM scales with scales derived from structural informatics

Biologically derived hydrophobicity scales (*19*) have previously been shown to correlate well with knowledge-based scales derived from the frequency of occurrence of the 20–amino acid side chains in membrane proteins, although the absolute range of the free energies obtained are reduced for knowledge-based versus biologically derived scales. It has been suggested that this attenuation arises from a lack of consideration of interior versus exposed positions during derivation of statistical potentials (*40*). The increase in the number of structures of membrane proteins in recent years has now allowed us to calculate the propensities of each amino acid as a function of both lipid exposure and depth in the lipid bilayer. A total of 2229 membrane proteins (1,159,085 residues) were analyzed, and their positions within the bilayer (designated z positions) were taken from the assignments in the Orientations of Proteins in Membranes (OPM) database (*41*). The propensities for each residue type varied considerably based on both their membrane depth and surface accessibility. We used these propensity scales to calculate the apparent free energy required to move an amino acid from an exposed water-soluble state to varying depths using the reverse Boltzmann approximation as described previously (*40*, *42*). The center of the bilayer was designated as $z = 0$, and we computed separate statistics for residues that were exposed versus buried in the protein interior (Fig. 3). The frequency of occurrence in the exposed outermost bins, between $z = \pm 34$ and 40 Å, was used as an aqueous standard state. The z-dependent profiles for the exposed positions provide an approximation of the free energy of transfer from water to various regions of the membrane (ΔG_{PDB}) (*40*, *42*).

As expected, we find that the z-dependent values of ΔG_{PDB} depend quite markedly on the burial of residues in a protein, with the exposed residues having a much greater tendency than buried residues to match their physical properties to that of the environment. Thus, the penalty for bringing a polar residue such as Asn, Asp, Gln, Arg, and Lys to the center of the bilayer is 1 to 2 kcal/mol greater when their side chains are lipid exposed versus buried in the protein interior, where they can engage in favorable electrostatic and hydrogen-bonded interactions. Similarly, apolar residues have the strongest tendency to be exposed to the membrane lipids near the center of the bilayer. Last, Trp and, to a lesser extent, Tyr have a pronounced tendency to accumulate at surface sites near the headgroup region, which display strong minima in ΔG_{PDB} at the headgroup region (approximately ±12 to 15 Å relative to the bilayer center) for exposed but not buried positions. These findings help rationalize the stabilizing influence of Tyr and Trp on $\Delta G_{app}^{Wat \rightarrow Int}$ versus $\Delta G_{app}^{Wat \rightarrow TM}$ in our biological scales. Moreover, the aromatic residue Phe behaves more like simple apolar side chains such as Ile and Leu than the interface-seeking Tyr and Trp residues in both the biological and PDB-derived scales.

To examine similarities among the amino acids, we clustered the 20 residue types based on their position-dependent values of ΔG_{PDB} (Fig. 4A). The ranking of the amino acids shows many expected features. For example, polar residues cluster together, with Lys + Arg and Glu + Asp in separate subclusters. Trp and Tyr cluster together, far from Phe, again showing that amphiphilicity plays an important role in defining locational preferences for these amino acids. Small polar (G, S, and T) and hydrophobic residues (L, I, V, M, F, and A) cluster together. The observed clustering is quite similar to the rankings seen in the values of $\Delta G_{app}^{Wat \rightarrow Int}$ and $\Delta G_{app}^{Wat \rightarrow TM}$.

We next compared our biological scales with the ΔG_{PDB} values computed at varying depths in the bilayer with the expectation that $\Delta G_{app}^{Wat \rightarrow TM}$ would correlate best with ΔG_{PDB} values computed near the bilayer center, while $\Delta G_{app}^{Wat \rightarrow Int}$ would correlate better with values computed near the headgroup region. $\Delta G_{app}^{Wat \rightarrow TM}$ correlates best with ΔG_{PDB} at $z = -7$ Å (Fig. 4B). This location is consistent with the position of the variable residues in the bRc peptide, which were slightly displaced from the TM region. By contrast, $\Delta G_{app}^{Wat \rightarrow Int}$ correlates best with ΔG_{PDB} at $z = -11$ Å (Fig. 4C). This value represents a minimum in ΔG_{PDB} for interfacially disposed residues such as Tyr and Trp. It is also a region where the value of ΔG_{PDB} is highly sensitive to the z position for both polar and apolar residues. Thus, even small adjustments in side-chain conformation or rigid-body shifts of an interfacial helix at this depth of insertion would cause a large change to the free energy of association. This heightened sensitivity manifests itself as a sharper peak in R^2 for $\Delta G_{app}^{Wat \rightarrow Int}$ at $z = -11$ Å than the corresponding peak $\Delta G_{app}^{Wat \rightarrow TM}$ at $z = -7$ Å (Fig. 4D). Hence, these correlations are consistent with the behavior expected for the assumed equilibrium between TM and interfacial locations.

We also considered the possibility that differences in the experimental systems contribute to a lack of perfect correlation between the biologically and statistically derived scales. In particular, the

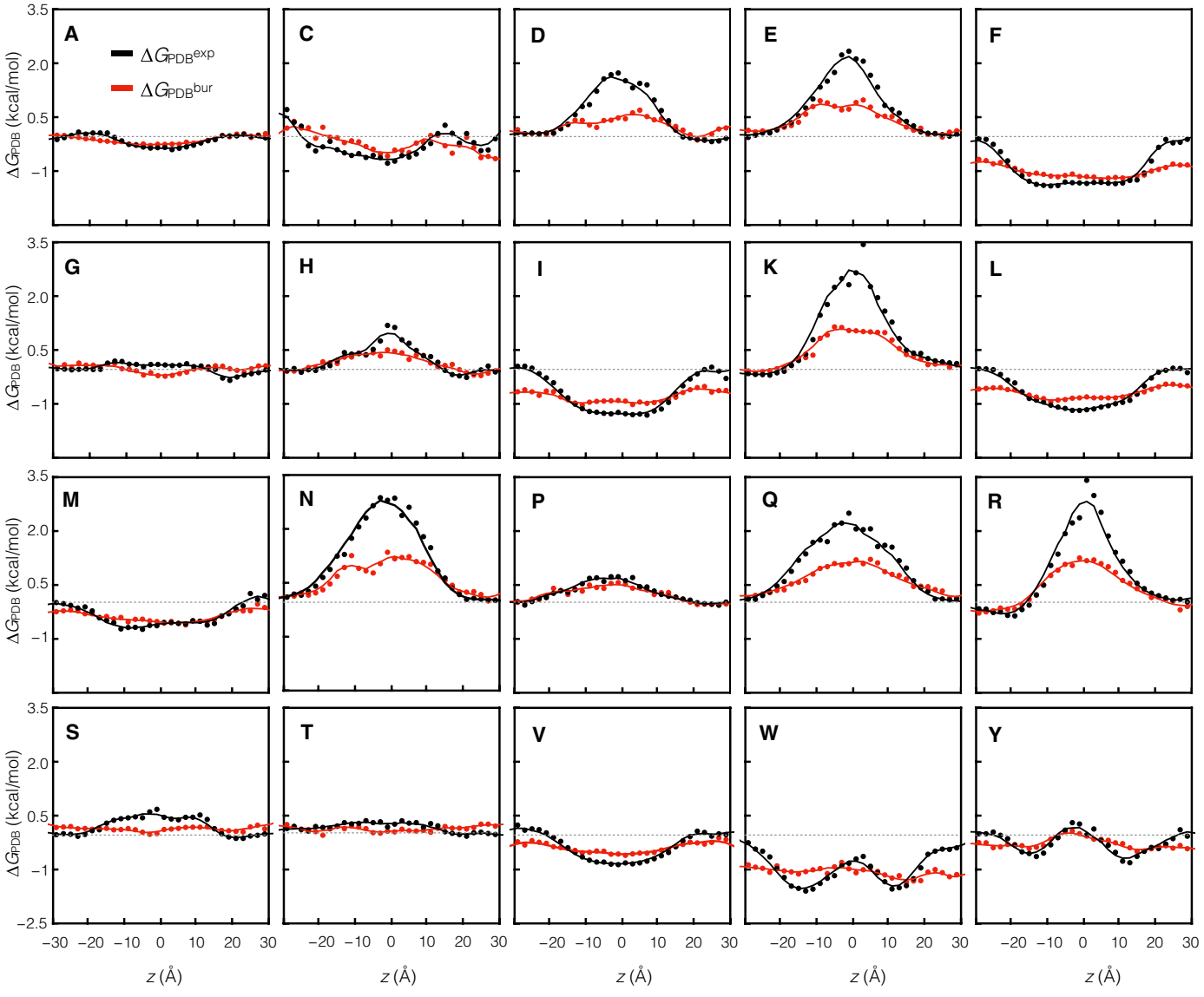

Fig. 3. Free energy of transfer from water to various regions of the membrane (ΔG_{PDB}) calculated from propensity differences. Each plot shows the free energy of transfer for moving an amino acid (indicated as a one-letter amino acid code in the top left corner of each chart) from an exposed position in water (averaged from propensities at $z = \pm 34$ to 40 Å) to a position at various z heights, either exposed at the protein exterior (ΔG_{PDB}^{exp}) or buried inside the protein (ΔG_{PDB}^{bur}). ΔG_{PDB} values are fitted with sigmoids, a Gaussian, or a combination of the two, as seen most appropriate.

interfacial helical state is not well represented in our PDB files, which are dominated by multispan TM helices. We therefore sought reduced dimensionality representation of sequence space rather than structural space for water-soluble and membrane helices to allow comparison of our bRc peptide sequences to natural proteins.

UMAP analysis of helical peptide segments

Uniform manifold approximation and projection (UMAP) has recently emerged as a useful means of visualizing high-dimensional data in a low-dimensional representation (*43*). To better understand how our experimentally characterized sequences relate to the broader natural distribution of helical protein segments, we generated a UMAP from a dataset of 3672 sequences of helices from soluble proteins and 2941 sequences of TM helices from the UniProt database (*44*). Each sequence was converted into a 24-feature vector consisting of the fractional amino acid composition, the average of the z-dependent ΔG_{PDB} values computed for each amino acid by assuming that the helix spans the membrane (Materials and Methods), the hydrophobic moment computed using the same ΔG_{PDB} values, the net charge of the peptide segment at pH 7, and the length of the sequence.

The resulting UMAP shows a two-lobed structure (Fig. 5A) with sequences from soluble helices in one large cluster (left lobe) and sequences from TM helices in the other (right lobe), with minimal overlap. When the bRc-derived sequences studied here are embedded into the same UMAP, their locations within the two lobes were

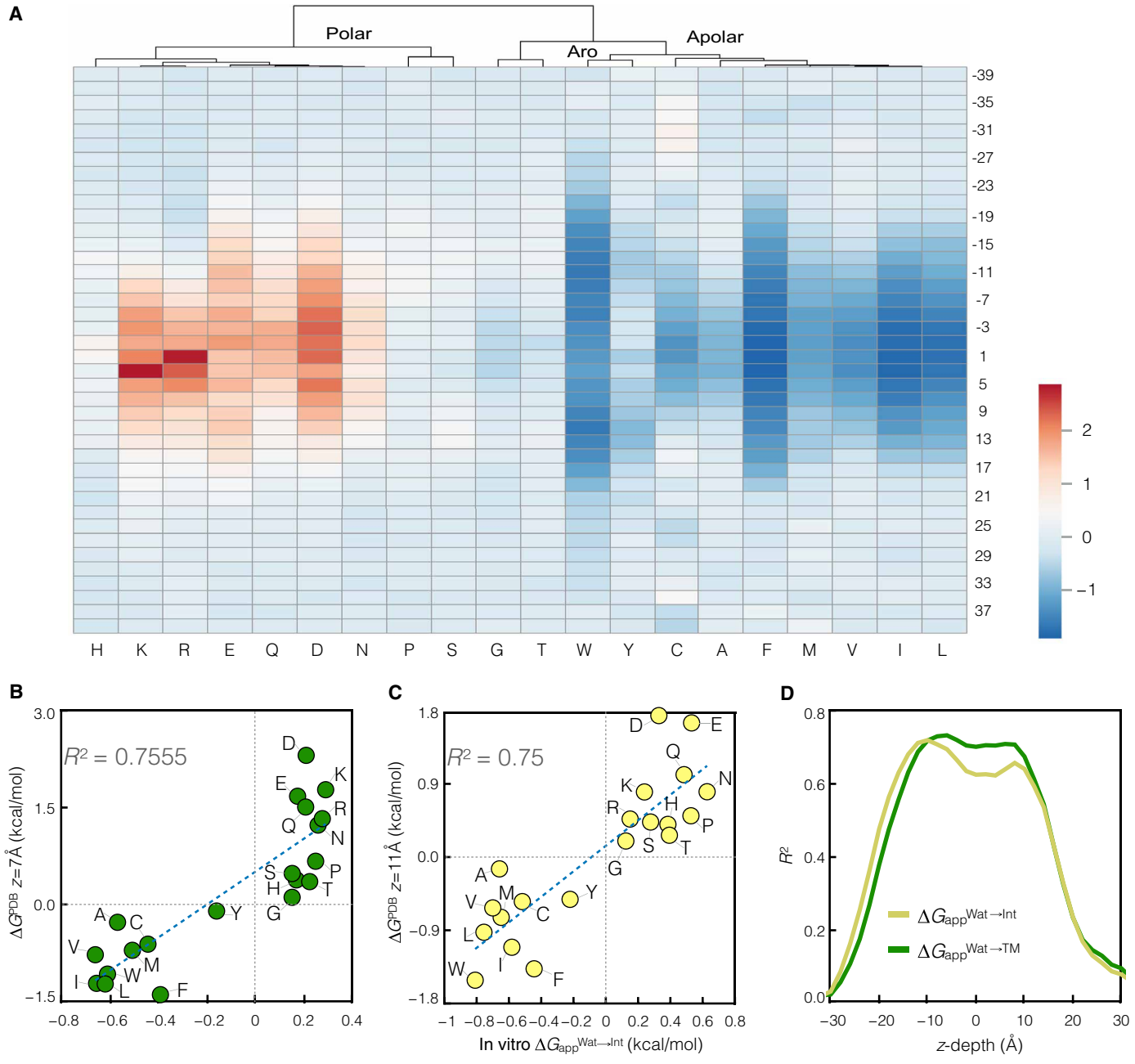

Fig. 4. Comparison between experimentally derived scales and Δ*G* from propensity calculations. Amino acids are marked beside their corresponding data points by single-letter names. (**A**) Heatmap-dendrogram of ΔG_{PDB} values at different *z* heights. Data are ΔG_{PDB} values of moving amino acids from a generic position in water to various positions in the membrane and exposed at the protein exterior. Data are clustered using the ClustVis web tool (*52*). Columns are clustered using correlation distance and complete linkage. (**B**) Experimentally derived water-to-TM scale ($\Delta G_{app}^{Wat\rightarrow TM}$) in the in vitro case compared to ΔG_{PDB} from an exposed position in water into a position exposed to the lipid bilayer at $z = -7$ Å. (**C**) Experimentally derived water-to-interface scale ($\Delta G_{app}^{Wat\rightarrow Int}$) in the in vitro case compared to ΔG_{PDB} from a generic position in water into a position exposed to the lipid bilayer at $z = -11$ Å. (**D**) Coefficient of determination (R^2) values for linear least squares trendlines between all *z*-dependent values of ΔG_{PDB} and $\Delta G_{app}^{Wat\rightarrow TM}$ (green line) and $\Delta G_{app}^{Wat\rightarrow Int}$ (yellow line) scale values.

reflective of the propensities of the helices for the aqueous, interfacial, and TM states. The individual points are aggregated into eight clusters (table S2), to allow us to examine how a position in the UMAP relates to the f_{Wat}, f_{Int}, and f_{TM}. Satisfyingly, there is qualitative agreement between the positions in the UMAP versus the fractions of these states, with the TM and interfacial states being increasingly populated as one moves from left to right. To place this qualitative observation on more quantitative footing, we derived moments describing the predilection of the helices to form the three species based on their position in the UMAP (Fig. 5A). The three moments represent a direction along which the embeddings are most highly correlated with the respective probabilities of the three states. The

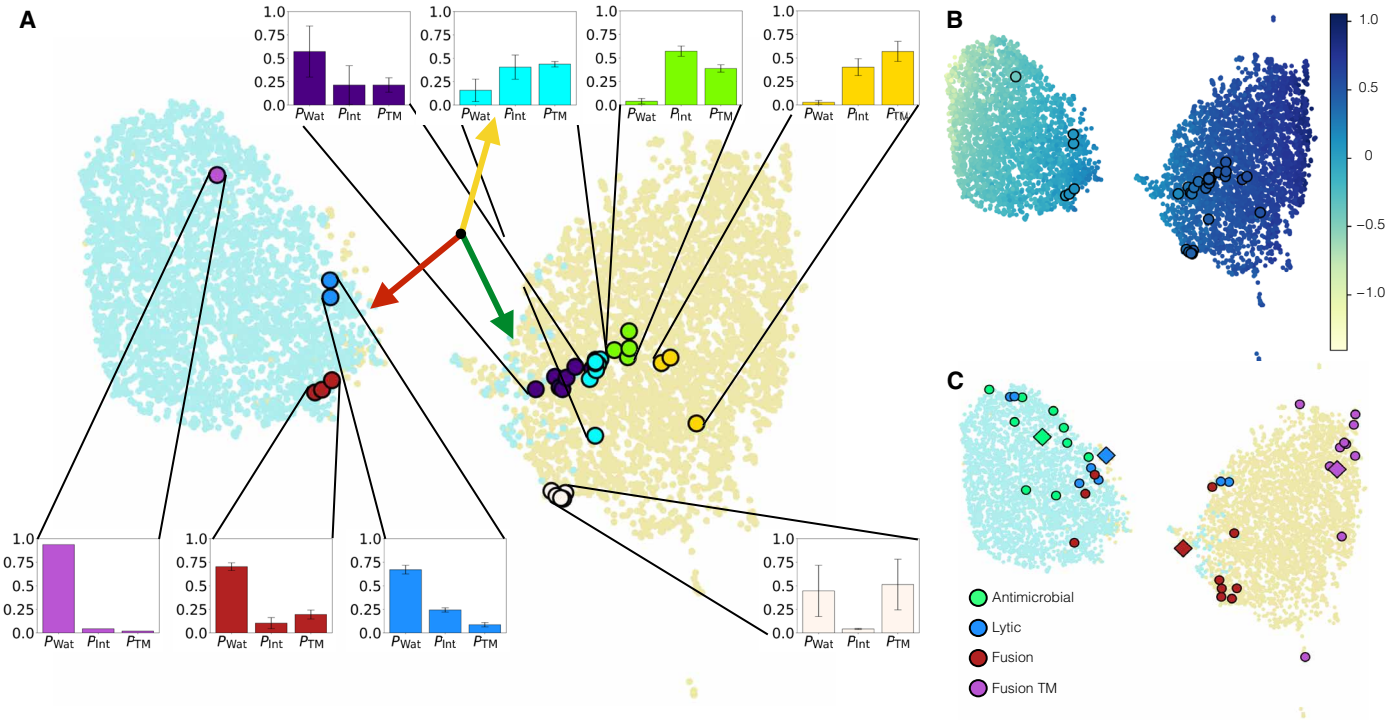

Fig. 5. UMAP embeddings of peptide sequences from UniProt exhibiting known helical disposition within soluble or TM proteins. The experimentally tested sequences are embedded as large points with black outlines. (**A**) UMAP embeddings with UniProt-derived sequences colored light blue if found in soluble domains and beige if found in TM domains. The UMAP embeddings cluster into two large lobes, delineating helices within soluble proteins from TM helices. All sequences tested with the LepB glycosylation assay (table S2) were visually clustered based on their embeddings and the predicted statistics over the three dispositions [aqueous (P_{Wat}), interfacial (P_{Int}), and TM (P_{TM})] are shown as histograms. Each of the three helical dispositions rapidly becomes more dominant along a particular direction in the UMAP space, shown by the three vectors using the same color scheme for the three dispositions as in Fig. 1D (green, yellow, and red for inserted, interfacial, and translocated location, respectively). (**B**) The same UMAP embeddings, colored according to average hydrophobicity computed using the ΔG_{PDB} scale, show a clear trend toward higher hydrophobicity (i.e., lower average ΔG_{PDB}) for helices from soluble proteins from left to right. (**C**) UMAP embeddings of helical peptides with known function (*23*) appear in distinct regions of the UMAP space. AMPs, which tend to be soluble and negatively charged, cluster in the upper half of the left lobe. Lytic peptides and the C-terminal segments of viral fusion proteins occupy the boundary region where helices from soluble proteins and TM proteins can be found. TM viral fusion domains, which must insert into a TM state to function, can be found at the rightmost edge of the right lobe, implying a strong preference for the TM disposition. Points labeled with circular markers represent individual sequences, while diamonds represent means in the UMAP space over a functional family.

results show good agreement with the qualitative observation that the horizontal axis roughly determines partitioning between water and TM states. The vertical axis reflects additional information relating to increased interfacial propensity.

Last, we analyzed the capacity of the UMAP projection to meaningfully cluster the sequences based on physical properties and functional families. Coloring the embeddings by physical properties, we found that the horizontal axis of the UMAP reflects hydrophobicity as assessed by our lipid-exposed ΔG_{PDB} scale (Fig. 5B), while the vertical axis reflects net charge at pH 7 within the left lobe (fig. S11A). Sequences with high helical hydrophobic moments, also calculated using our lipid-exposed ΔG_{PDB} scale, congregated in the upper left section of the left lobe (fig. S11B). To examine whether the UMAP could meaningfully cluster peptide sequences with different functions, we embedded the membrane-associating sequences from four distinct functional classes of previously characterized peptides and proteins whose functions are dictated by their ability to bind to membrane surfaces into the UMAP (Fig. 5C).

We first examined helical AMPs and cytotoxic peptide sequences. Both bind to membrane surfaces, attracted to microbial bilayers that are richer in acidic phospholipids than mammalian membranes, while cytotoxic peptides bind more indiscriminately to eukaryotic and bacterial membranes. Both disrupt cell membranes by mechanisms that involve surface-absorbed as well as membrane-spanning states. Our analysis shows that AMPs embed into the basic and water-soluble helical regions of the UMAP, while cytotoxic peptides tend to have embeddings directly between the TM and water-soluble states. Thus, on average, cytotoxic peptides would be predicted to have higher affinities for membranes and be more stable in TM states. This finding is in keeping with their more aggressive and nonselective behavior toward bilayers relative to AMPs. [For references on this subject, see the review (*39*) and other reviews referenced within.]

Helical fusogenic sequences present another interesting class of proteins, which are released from membrane-embedded fusogenic proteins to bind to the surface of target membranes and ultimately help stabilize a dynamic membrane-spanning fusion pore. The dichotomy between the embeddings for the TM regions of fusogenic peptides and their fusion pores is functionally interesting in this regard. The TM regions lie in the far right of the plot, consistent with their function as membrane anchors, while the fusogenic peptides lie close to the interface between water-soluble

and membrane-spanning helices in keeping with their dynamic requirements for membrane binding. These considerations extend our analysis of model peptides to sequences that, like the bRc peptides, are poised for membrane insertion and surface association.

DISCUSSION

A major goal of this study was to understand the balance between the surface-absorbed and TM states, particularly given the important role that surface-absorbed states play during the folding process of membrane proteins. Given that they must interconvert during folding, it would seem that similar forces stabilize both states, making it difficult to assess the essential differential features experimentally. To address this challenge, we take advantage of a bRc-derived peptide that can equilibrate between surface-absorbed and TM states. The placement of a residue in a helical surface-absorbed peptide is influenced by its position along the amphiphilic helix. Residues on the polar side of an amphiphilic helix can be replaced by other polar residues without consequence, but the same substitution would greatly change the amphiphilicity of the helix if the polar residue were placed on the apolar face of the helix. By replacing three consecutive residues in bRc, we ensure that substitutions are made on both its polar and apolar faces, which enables the contribution of each amino acid to the amphiphilicity of the structure to be unambiguously determined.

As a result, this work provides a large-scale, systematic examination of the role of side chains in mediating the propensity of peptides in a helical conformation to associate with the membrane interface during translocon-mediated protein insertion into the endoplasmic reticulum. In pioneering biophysical studies (*21*, *22*), Wimley, White, and coworkers examined short (nonhelical) host-guest pentapeptides to see how variations in sequence led to changes in the energetics of binding to the surface of artificial liposomes to provide the first experimental scale for surface association. Their interfacial scale showed a significant correlation with hydrophobic scales, such as one derived from octanol-water partition coefficients ($R^2 = 0.86$) (fig. S12). Amphiphilic aromatic residues, Trp and Tyr, showed enhanced affinity for the membrane interface, although charged residues such as Arg and Lys, which clearly contribute to interfacial recognition in our own and previous studies (*19*), did not appear to enhance affinity for the neutral lipid bilayers used in these studies. We also observe good correlation between $\Delta G_{app}^{Wat \rightarrow Int}$ and other hydrophobicity scales (fig. S13), most notably the *z*-dependent scale for inserted helices elucidated by Hessa *et al.* (Fig. 2, B and C).

Sequence-based bioinformatic analysis via UMAP also provided complementary insights into the experimental assay. The clustering of sequences in the UMAP space based on aqueous or TM disposition, as well as net charge, hydrophobicity, and known function, showed that UMAP could embed short peptide sequences in semantically meaningful ways. Each experimentally tested sequence had numerous close neighbors from endogenous proteins, representing a broad sample of the known sequence space. Moreover, they clustered near other experimentally tested sequences with similar propensities for the three states, with the horizontal coordinate in the UMAP space roughly determining the partitioning between the aqueous and TM states and, for more hydrophobic sequences, the vertical coordinate roughly determining the partitioning between the aqueous and TM states. We hypothesize that endogenous sequences near the experimentally tested sequences in the UMAP space will have comparable preferences for the three states, presenting a potential direction for future study.

Last, both the experimental $\Delta G_{app}^{Wat \rightarrow Int}$ and the informatic scale, ΔG_{PDB}, provide important information for the de novo design of membrane proteins. The large difference between ΔG_{PDB} for buried versus exposed sites has not previously been appreciated. The scale was derived in a manner that allows easy incorporation into algorithms for sequence design given a backbone structure. Thus, our studies should enable understanding of natural proteins as well as de novo design of peptides and proteins in the highly heterogenic milieu of the membrane, spanning a wide range of topologies and functions. We posit, on the basis of our findings, that the vital process by which cells determine the relation of their myriad proteins to the plasma membrane can be reduced to a collection of straightforward sequence-based rules.

MATERIALS AND METHODS

Enzymes and chemicals

All enzymes as well as plasmid pGEM1 and the TnT coupled transcription/translation system were from Promega (Madison, WI). SP6 RNA polymerase and ER RMs from dog pancreas were from tRNA Probes (College Station, TX). EasyTagTM EXPRESS35S Protein Labeling Mix, [^{35}S]-L-methionine and [^{35}S]-L-cysteine, for in vitro labeling was purchased from Perkin Elmer (Waltham, MA, USA). Restriction enzymes and Endoglycosidase H were from Roche Molecular Biochemicals (Basel, Switzerland). The DNA plasmid, RNA cleanup, and PCR purification kits were from Qiagen (Hilden, Germany). The QuikChange PCR mutagenesis kit was from Stratagene (La Jolla, CA). All the oligonucleotides were from Macrogen Inc. (South Korea).

DNA manipulation

Oligonucleotides encoding the different bRc variants and melittin flanked by GGPG…GPGG tetrapeptides intended to "insulate" the central peptide from the surrounding LepB sequence were introduced into the pGEM1Lep plasmid (*18*, *45*) between the Spe I and Kpn I sites by using four double-stranded oligonucleotides (38 to 58 nucleotides long) with overlapping overhangs at the ends and phosphorylated at 5′ ends. Pairs of complementary oligonucleotides were first annealed at 85°C for 10 min followed by slow cooling to 30°C. After that, the pair-annealed double-stranded oligos were mixed, incubated at 65°C for 5 min, cooled slowly to room temperature, and ligated into the vector. All bRc inserts were confirmed by sequencing of plasmid DNA (Macrogen).

The bRc site-directed mutagenesis was performed using the QuikChange mutagenesis kit (Stratagene) following the manufacturer's protocol. The DNA encoding LepG3 proteins was PCR amplified to incorporate a c-Myc tag at the N terminus and subcloned into the mammalian pCAGGS vector for in vivo assays using the In-Fusion HD technology (Takara), according to the manufacturer's instructions. All DNA manipulations were confirmed by sequencing of plasmid DNAs. Site-directed mutagenesis was also used to introduce acceptor sites for N-linked glycosylation at appropriate positions.

In vitro transcription and translation

LepB-bRc constructs were transcribed and translated in the TnT Quick system (Promega). One microgram of DNA template, 1 μl (5 μCi) of ^{35}S-Met/Cys (PerkinElmer), and 1 μl of microsomes (tRNA

Probes) were added at the start of the reaction, and samples were incubated for 90 min at 30°C. After polypeptide synthesis membranes were collected by ultracentrifugation and analyzed by SDS-PAGE, gels were lastly visualized on a Fuji FLA3000 phosphorimager using the Image Gauge software.

For the proteinase K protection assay, the translation mixture was supplemented with 1 μl of 50 mM $CaCl_2$ and 1 μl of proteinase K (4 mg/ml) and then digested for 20 min on ice. Adding 1 mM phenylmethylsulfonyl fluoride before SDS-PAGE analysis stopped the reaction.

After polypeptide synthesis membranes were collected by ultracentrifugation and analyzed by SDS-PAGE, gels were lastly visualized on a Fuji FLA3000 PhosphorImager using the Image Gauge software.

Mammalian cells assay

HEK293T cells (ATTC, CRL-3216) were grown in Dulbecco's modified Eagle's medium (DMEM) (Gibco) supplemented with 10% fetal bovine serum (FBS) and penicillin-streptomycin (P/S) (100 U/ml) at 37°C, 5% CO_2. Cells were plated in 24-well plates (2×10^6 cells per plate) and transfected after 24 hours. For transfection procedure, 500 ng per well of plasmids encoding c-Myc tagged LepG3 were added to a mixture of 2 μl of polyethylenimine (PEI) MW 25,000 (1 mg/ml) (Alfa Aesar) diluted in 100 μl of Opti-MEM reduced serum medium (Gibco). Transfection mixture was incubated for 20 min at room temperature and then added dropwise to 24-hour cultured cells in 500 μl of DMEM containing FBS and P/S.

At 48 hours posttransfection, cells were collected in lysis buffer [TBS (20 mM tris-HCl, pH 7.5, and 150 mM NaCl) and 1% SDS] and put through three freeze-thaw cycles. The suspensions were clarified by centrifugation (13,000*g*). Supernatants were mixed with SDS-PAGE sample buffer, heated for 5 min at 95°C, and loaded on 12% SDS-PAGE. Next, proteins were transferred onto PVDF membranes. Immuno-identification of the LepG3 system proteins was done using a-c-Myc antibody (Merck) followed by a secondary horseradish peroxidase–conjugated a-rabbit antibody (Merck). Chemiluminescence was visualized by an ImageQuant LAS 4000 (GE Healthcare). Bands were quantified using ImageJ (NIH).

Peptide synthesis

The bRc peptides were synthesized using standard Fmoc chemistry by an automated microwave-assisted solid- phase peptide synthesizer (Biotage Initiator+Alstra) on a 0.1 mmol scale. Each cycle included (i) Fmoc deprotection [20% 4-methyl piperidine with HOBt (0.1 M) in *N*,*N*′-dimethylformamide (DMF), 4.5 ml, 5 min, 75°C]; (ii) coupling with N-α-Fmoc-amino acid (5 eq, 0.5 M in DMF), O-(6-Chlorobenzotriazol-1-yl)-N,N,N′,N′-tetramethyluronium hexafluorophosphate (HCTU, 4.95 eq, 0.5 M in DMF), and N,N-Diisopropylethylamine (DIPEA) (10 eq, 0.5 M in DMF) (5 min, 75°C). N-terminal acetylation was done by treatment or resin with Ac_2O (10 eq) and DIPEA (20 eq) in DMF and the final cleavage was performed using trifluoroacetic acid:triisopropylsilane:water (TFA:TIPS:H2O, 95:2.5:2.5). The crude peptide was obtained by cold ether precipitation and purified by Reverse Phase High Pressure Liquid Chromatography (RP-HPLC). Their chemical entity and purity were confirmed by matrix-assisted laser desorption/ionization and analytical HPLC.

CD spectrometry

Small unilamellar vesicles of POPC were prepared by combining ethanolic solution of POPC and the peptide ([peptide]/[lipid] = 1/100), drying to a film and lyophilized overnight. The thin film was resuspended in phosphate buffer (20 mM, pH 8.0) to a phosholipid concentration of 25 mM and tip-sonicated (Fisher Sonic Dismembrator Model 500, 20% power, 5 min, 2 s on, 2 s off). The resulting mixture was further diluted to 5 μM (peptide concentration) in 3 ml of phosphate buffer (20 mM, pH 8.0) in a 1-cm cuvette. Circular dichroism data were collected on a JASCo J-810 (8 s average, 2 nm bandwidth, triplicate measurement).

Analysis of the observed free energy of transfer from the experimental data

The fraction of the peptide sequences in the water-exposed, interfacial, or TM locational states were approximated from their relative mole fractions. To obtain these mole fractions from the experimentally observed radiometric counts, a statistical model was developed to account for the effects of incomplete glycosylation of the nascent chain. Beginning with the counts C_i, with $i \in \{1, 2, 3\}$ denoting the number of glycosylations, the probabilities $P(i)$ of each observed number of glycosylations were calculated as

$$P(i) = \frac{C_i}{C_1 + C_2 + C_3} \tag{1}$$

The glycosylation state probabilities of Eq. 1 were then expressed in terms of the conditional probabilities of observing i glycosylations given that the peptide sequence is in a known locational state, as follows

$$P(i) = P(i\,|\,\mathrm{TM})p(\mathrm{TM}) + P(i\,|\,\mathrm{Int})p(\mathrm{Int}) + P(i\,|\,\mathrm{Wat})p(\mathrm{Wat}) \tag{2}$$

The three instances of Eq. 2, one for each i, may be collected into a matrix-vector product

$$\begin{bmatrix} P(1) \\ P(2) \\ P(3) \end{bmatrix} = \begin{bmatrix} P(1\,|\,\mathrm{TM}) & P(1\,|\,\mathrm{Int}) & P(1\,|\,\mathrm{Wat}) \\ P(2\,|\,\mathrm{TM}) & P(2\,|\,\mathrm{Int}) & P(2\,|\,\mathrm{Wat}) \\ P(3\,|\,\mathrm{TM}) & P(3\,|\,\mathrm{Int}) & P(3\,|\,\mathrm{Wat}) \end{bmatrix} \begin{bmatrix} p(\mathrm{TM}) \\ p(\mathrm{Int}) \\ p(\mathrm{Wat}) \end{bmatrix} \tag{3}$$

The columns of the matrix may be interpreted as the probabilities of observing each of the three possible glycosylation states given that the protein is in a known locational state. While the TM state is assumed to never be doubly or triply glycosylated and the interfacial state is assumed to never be triply glycosylated, incomplete glycosylation by the OST is expected to lead to a mixture of single and double glycosylation for the interfacial state and single, double, and triple glycosylation for the aqueous state. For $P(2\,|\,\mathrm{Int})$, we retained to three significant digits the baseline double glycosylation probability of 86.5% from the original study by Hessa *et al.* (*18*), while the values of $P(i\,|\,\mathrm{Wat})$ were determined from our experimental glycosylation data for the soluble sequence GDKQEGEWPTGLRLSIGGI (corresponding to residues 304 to 322 in the translocated P2 domain of LepB). We computed $P(i\,|\,\mathrm{Wat})$ for $i = 1, 2, 3$ by averaging the glycosylation state probability $P(i)$, determined via Eq. 1, across the four experimental replicates for the P2 domain. This analysis generated the following values

$$\begin{bmatrix} P(1\,|\,\mathrm{TM}) & P(1\,|\,\mathrm{Int}) & P(1\,|\,\mathrm{Wat}) \\ P(2\,|\,\mathrm{TM}) & P(2\,|\,\mathrm{Int}) & P(2\,|\,\mathrm{Wat}) \\ P(3\,|\,\mathrm{TM}) & P(3\,|\,\mathrm{Int}) & P(3\,|\,\mathrm{Wat}) \end{bmatrix} = \begin{bmatrix} 1 & 0.1350 & 0.0957 \\ 0 & 0.8650 & 0.0977 \\ 0 & 0 & 0.8066 \end{bmatrix} \tag{4}$$

As the matrix given in Eq. 4 is invertible, the locational state probabilities were determined from multiplying both sides of Eq. 3 by the inverse matrix. As a result

$$\begin{bmatrix} p(\mathrm{TM}) \\ p(\mathrm{Int}) \\ p(\mathrm{Wat}) \end{bmatrix} = \begin{bmatrix} 1 & -1.15607 & -0.09974 \\ 0 & 1.15607 & -0.14003 \\ 0 & 0 & 1.23977 \end{bmatrix} \begin{bmatrix} P(1) \\ P(2) \\ P(3) \end{bmatrix} \quad (5)$$

Locational state probabilities were estimated from the glycosylation state probabilities via Eq. 5 for all peptide sequences that were experimentally tested. To amplify the apparent differences in mole fraction between the singly and doubly glycosylated states, three substitutions were simultaneously made, replacing a DLP sequence (residues 11, 12, and 13 in bRc-derived sequence) with a new sequence $X_1X_2X_3$. Free energy differences between the water-exposed state and the interfacial and TM states, respectively, were calculated from probability ratios as follows

$$\Delta G_{\mathrm{Wat}\to\mathrm{Int}}(\mathrm{X_1X_2X_3}) = -RT\ln\left[\frac{p(\mathrm{Int})}{p(\mathrm{Wat})}\right] \quad (6)$$

$$\Delta G_{\mathrm{Wat}\to\mathrm{TM}}(\mathrm{X_1X_2X_3}) = -RT\ln\left[\frac{p(\mathrm{TM})}{p(\mathrm{Wat})}\right] \quad (7)$$

In Eqs. 6 and 7, the value of 0.602 or 0.612 kcal mol^{-1} was used for the ideal gas constant and temperature, RT, for in vitro or in vivo data, respectively. The dependence of these free energy differences upon the sequence $X_1X_2X_3$ is modeled as a sum over $\Delta\Delta G$ values associated with the substitution $Z_i \to X_i$, where $Z_1Z_2Z_3$ is a hypothetical sequence for which $\Delta G_{\mathrm{Wat}\to\mathrm{Int}}(\mathrm{Z_1Z_2Z_3}) = 0$ and $\Delta G_{\mathrm{Wat}\to\mathrm{TM}}(\mathrm{Z_1Z_2Z_3}) = 0$. Formally

$$\begin{aligned}\Delta G_{\mathrm{Wat}\to\mathrm{State}}(\mathrm{X_1X_2X_3}) = {} & \Delta\Delta G_{\mathrm{State}}(\mathrm{Z_1}\to\mathrm{X_1}) + \\ & \Delta\Delta G_{\mathrm{State}}(\mathrm{Z_2}\to\mathrm{X_2}) + \Delta\Delta G_{\mathrm{State}}(\mathrm{Z_3}\to\mathrm{X_3})\end{aligned} \quad (8)$$

While the form of Eq. 8 implies a positional dependence of ΔG upon the three substitutions, we saw that permutations of the sequences AAG and AGG showed much smaller differences in $\Delta G_{\mathrm{Wat}\to\mathrm{TM}}$ or $\Delta G_{\mathrm{Wat}\to\mathrm{Int}}$ between sequences of the same composition versus between sequences of different composition (table S3). Thus, we assume ΔG is independent of order and thus $\Delta\Delta G_{\mathrm{State}}(\mathrm{Z_i}\to\mathrm{X_i}) \stackrel{\mathrm{def}}{=} \Delta\Delta G_{\mathrm{State}}(\mathrm{X_i})$. Therefore, ΔG can be expressed as a sum over all 20 amino acids X of the corresponding $\Delta\Delta G_{\mathrm{State}}(\mathrm{X})$ values, weighted by the number of occurrences n_X of the amino acid in the sequence

$$\begin{aligned}\Delta G_{\mathrm{Wat}\to\mathrm{State}} = {} & n_{\mathrm{A}}\Delta\Delta G_{\mathrm{State}}(\mathrm{A}) + \\ & n_{\mathrm{C}}\Delta\Delta G_{\mathrm{State}}(\mathrm{C}) + \cdots + n_{\mathrm{Y}}\Delta\Delta G_{\mathrm{State}}(\mathrm{Y})\end{aligned} \quad (9)$$

A total of 113 replicates were tested experimentally across 27 sequences $X_1X_2X_3$, including DLP and the 20 possible sequences XXX for each of the amino acids. The replicates for each sequence were analyzed for outliers that deviated from the sample mean by more than twice the sample SD, resulting in the identification of one outlier for the sequence AAA, which was discarded. The 112 remaining instances of Eq. 9 for each of the experiments were concatenated into matrix form

$$\begin{bmatrix} \Delta G_{\mathrm{Wat}\to\mathrm{State}}(\mathrm{AAA}) \\ \vdots \\ \Delta G_{\mathrm{Wat}\to\mathrm{State}}(\mathrm{CCC}) \\ \vdots \end{bmatrix} = \begin{bmatrix} 3 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \Delta\Delta G_{\mathrm{State}}(\mathrm{A}) \\ \vdots \\ \Delta\Delta G_{\mathrm{State}}(\mathrm{C}) \\ \vdots \end{bmatrix} \quad (10)$$

The two vectors of 20 values for $\Delta\Delta G_{\mathrm{Int}}(\mathrm{X})$ and $\Delta\Delta G_{\mathrm{TM}}(\mathrm{X})$ were estimated by weighted linear least-squares regression using the NumPy Python package (*46*). The weights were set to be the sample SDs of the $\Delta G_{\mathrm{Wat}\to\mathrm{State}}$ values across the experimental replicates of each three-residue sequence. The errors are given by the standard errors of each regression variable.

Database for propensity calculation

Distinct structures of bitopic and α-helical polytopic proteins from the OPM database (*41*) as of February 2022 were used for propensity calculation. Proteins in this database are in their native biological assemblies and have an average TM helix tilt of 0°, and the average centroid of the TM region is at the bilayer center. Therefore, their structures were used without further alignment. Only structures, or chains in multimeric structures, with resolution better than 3.5 Å and a maximum sequence identity of 70% were selected for further analysis. The final list of PDB accession codes and chain IDs for propensity calculation can be found in the data S1 Excel file. The list contains a total of 2193 structures with 7057 unique chains. Note that biological assemblies were used for the correct identification of interior(buried)/exterior(exposed) residue positions, while the asymmetric units of the structures from this nonredundant list were used for all other statistics and calculations.

Propensity calculation

Propensities were calculated in a similar fashion to previous results (*23*, *41*, *47*). Structures from the OPM database were aligned so that the membrane normal was parallel to the z axis, with $z = 0$ at the bilayer center and negative z toward the cytoplasm. Residue positions were defined by the coordinates of C_β (C_α for Gly). The structural data were divided into bins of 2 Å along the z axis and the occurrence of each residue at each binned z value was counted. Propensities were calculated using the equation

$$P_{x,z} = \frac{n_{x,z}/n_z}{n_x/N} \quad (11)$$

where $P_{x,z}$ is the propensity of a given residue x in a given z value bin z, $n_{x,z}$ is the total count of residue x in bin z, n_z is the total count of all residue types in this bin, n_x is the total count of residue x in all bins, and N is the total count of all residue types in all bins.

Residues were identified as "exterior" (exposed) or "interior" (buried) using a convex hull algorithm (*48*, *49*). The coordinates for C_α and C_β atoms were used to define two surfaces. If a C_β atom fell outside the surface of the C_β hull, that residue was identified as exterior; conversely, if a C_β atom was found within the surface of the C_α hull, that residue was identified as interior (C_β atoms were appended to the C_α atoms of Gly for this calculation by converting the residue to Ala). The advantage of using C_α and C_β atoms is that this treatment can be used in future studies for protein design, in which only the backbone, but not yet the sequence, has been generated (in which case the values of $\Delta\Delta G_{\mathrm{PDB}}$ can be used directly in a design energy function or to bias sampling). Moreover, this approach eliminates

problems with missing side chains or limited resolution for surface residues in experimental structures. The radius of the alpha-sphere used to define the surfaces was 8 Å. Propensities for exterior (exposed) and interior (buried) residues were calculated the same way as described above, but for exterior and interior residues, respectively.

ΔG_{PDB} calculation

For the "reaction" of moving a residue from one position (e.g., in the aqueous environment) to another position (e.g., in the interior of a protein at $z = 0$ Å), a pseudo free energy ΔG_{PDB} was defined by Eq. 12

$$\Delta G_{PDB} = -RT\ln\left(\frac{P_{pos2}}{P_{pos1}}\right) \tag{12}$$

where P_{pos1} and P_{pos2} are the propensities of this residue at positions 1 and 2, respectively, R is the gas constant, and T is temperature in kelvin. As in Eqs. 6 and 7, the value of 0.593 kcal mol^{-1} was used for the ideal gas constant and temperature, RT, in Eq. 12. Considering that a residue is sufficiently distant from the lipid membrane when $z < -35$ Å or $z > +35$ Å, the propensity for a residue to occupy the aqueous environment was defined by the average of its propensities at the two z value ranges of −39 to −35 Å and +35 to +39 Å.

UMAP analysis

Protein sequences of known α helices were retrieved from the UniProt database (*44*). The sequences corresponding to TM helices were derived from integral membrane proteins with no more than four membrane-spanning segments. This was done to ensure that all sampled helices are at least partially exposed to the lipid environment and not buried within the cores of large TM protein domains. In addition, given that TM helices must span the bilayer and thus are seldom shorter than 20 amino acids in length, a minimum length of 15 amino acids was imposed upon the helices from soluble proteins. This resulted in a dataset consisting of 3672 TM helices and 2941 helices from soluble proteins. Each α helix was featured as a 24-dimensional vector, with entries given by the fractional composition of the sequence by each of the 20 amino acids, the average z-dependent ΔG_{PDB} across the sequence, the hydrophobic moment computed using the calculated ΔG_{PDB} values for each residue, the net charge Q_{net} at pH 7, and the length L of the peptide segment. The z-dependent ΔG_{PDB} values were calculated according to Eq. 12 under the assumption that the TM reference state consists of the apolar portion of a membrane helix inserted at an angle across the membrane with its centroid at $z = 0$ and the angle determined such that the N and C termini are at $z = 15$ and $z = -15$, respectively. Helices with $L \leq 21$ are assumed to be oriented perpendicular to the bilayer normal. Formally, for residue indices $i \in [\![1, L]\!]$

$$z(i) = \begin{cases} 1.5 \cdot \dfrac{2i-1-L}{2} & \text{if } L \leq 21 \\ 1.5 \cdot \dfrac{2i-1-L}{2} \cdot \dfrac{20}{L-1} & \text{else} \end{cases} \tag{13}$$

For $L > 21$, Eq. 13 linearly interpolates the interval $[-15, 15]$ to determine the z coordinates of each residue, which is equivalent to averaging the z coordinate of each residue over all possible rotations of the helix about its axis (fig. S11). Hydrophobic moments were computed using the method of Eisenberg *et al.* (*50*). Moreover, because the net charge at pH 7 and helix length L were initially much larger in magnitude than the other 22 features, they were normalized to lie within the interval $[0, 1]$ as follows

$$Q_{net}(\text{seq}) = \frac{Q_{net}(\text{seq}) - \min(Q_{net})}{\max(Q_{net}) - \min(Q_{net})} \tag{14}$$

$$L(\text{seq}) = \frac{L(\text{seq}) - \min(L)}{\max(L) - \min(L)} \tag{15}$$

The 24-dimensional feature vectors from the UniProt-derived database were then passed as input to the UMAP fitter in the UMAP Python package, using default parameters. Feature vectors for the experimentally tested sequences and the sequences from the four functional peptide classes were preprocessed using the minimum and maximum values for Q_{net} and L from the UniProt-derived database and embedded in the previously fit UMAP space for comparative analysis.

Supplementary Materials

The PDF file includes:

Figs. S1 to S13
Tables S1 to S3
Legend for data S1

Other Supplementary Material for this manuscript includes the following:

Data S1

REFERENCES AND NOTES

1. A. E. Johnson, M. A. van Waes, The translocon: A dynamic gateway at the ER membrane. *Annu. Rev. Cell Dev. Biol.* **15**, 799–842 (1999).
2. J. E. Donald, Y. Zhang, G. Fiorin, V. Carnevale, D. R. Slochower, F. Gai, M. L. Klein, W. F. DeGrado, Transmembrane orientation and possible role of the fusogenic peptide from parainfluenza virus 5 (PIV5) in promoting fusion. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 3958–3963 (2011).
3. R. M. Epand, R. F. Epand, Modulation of membrane curvature by peptides. *Biopolymers* **55**, 358–363 (2000).
4. N. W. Schmidt, A. Mishra, J. Wang, W. F. Degrado, G. C. L. Wong, Influenza virus A M2 protein generates negative Gaussian membrane curvature necessary for budding and scission. *J. Am. Chem. Soc.* **135**, 13710–13719 (2013).
5. I. D. Pogozheva, H. I. Mosberg, A. L. Lomize, Life at the border: Adaptation of proteins to anisotropic membrane environment. *Protein Sci.* **23**, 1165–1196 (2014).
6. F. Cymer, G. von Heijne, S. H. White, Mechanisms of integral membrane protein insertion and folding. *J. Mol. Biol.* **427**, 999–1022 (2015).
7. M. C. Wiener, S. H. White, Structure of a fluid dioleoylphosphatidylcholine bilayer determined by joint refinement of x-ray and neutron diffraction data. III. Complete structure. *Biophys. J.* **61**, 434–447 (1992).
8. S. H. White, W. C. Wimley, Membrane protein folding and stability: Physical principles. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 319–365 (1999).
9. R. S. Hegde, R. J. Keenan, A unifying model for membrane protein biogenesis. *Nat. Struct. Mol. Biol.* **31**, 1009–1017 (2024).
10. K. Corin, J. U. Bowie, How physical forces drive the process of helical membrane protein folding. *EMBO Rep.* **23**, e53025 (2022).
11. P. F. Egea, R. 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**, 17182–17187 (2010).
12. M. Gogala, T. Becker, B. Beatrix, J.-P. Armache, C. Barrio-Garcia, O. Berninghausen, R. Beckmann, Structures of the Sec61 complex engaged in nascent peptide translocation or membrane insertion. *Nature* **506**, 107–110 (2014).
13. R. M. Voorhees, R. S. Hegde, Structure of the Sec61 channel opened by a signal sequence. *Science* **351**, 88–91 (2016).
14. J. L. Popot, D. M. Engelman, Membrane protein folding and oligomerization: The two-stage model. *Biochemistry* **29**, 4031–4037 (1990).

15. M. Bañó-Polo, C. Baeza-Delgado, S. Tamborero, A. Hazel, B. Grau, I. Nilsson, P. Whitley, J. C. Gumbart, G. von Heijne, I. Mingarro, Transmembrane but not soluble helices fold inside the ribosome tunnel. *Nat. Commun.* **9**, 5246 (2018).
16. K. A. Gaffney, R. Guo, M. D. Bridges, S. Muhammednazaar, D. Chen, M. Kim, Z. Yang, A. L. Schilmiller, N. F. Faruk, X. Peng, A. Daniel Jones, K. H. Kim, L. Sun, W. L. Hubbell, T. R. Sosnick, H. Hong, Lipid bilayer induces contraction of the denatured state ensemble of a helical-bundle membrane protein. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2109169119 (2022).
17. L. Kater, B. Frieg, O. Berninghausen, H. Gohlke, R. Beckmann, A. Kedrov, Partially inserted nascent chain unzips the lateral gate of the Sec translocon. *EMBO Rep.* **20**, e48191 (2019).
18. T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S. H. White, G. von Heijne, Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433**, 377–381 (2005).
19. T. Hessa, N. M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, I. Nilsson, S. H. White, G. von Heijne, Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* **450**, 1026–1030 (2007).
20. M. Musial-Siwek, A. Karabadzhak, O. A. Andreev, Y. K. Reshetnyak, D. M. Engelman, Tuning the insertion properties of pHLIP. *Biochim. Biophys. Acta* **1798**, 1041–1046 (2010).
21. W. C. Wimley, T. P. Creamer, S. H. White, Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry* **35**, 5109–5124 (1996).
22. W. C. Wimley, S. H. White, Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* **3**, 842–848 (1996).
23. C. A. Schramm, B. T. Hannigan, J. E. Donald, C. Keasar, J. G. Saven, W. F. Degrado, I. Samish, Knowledge-based potential for positioning membrane-associated structures and assessing residue-specific energetic contributions. *Structure* **20**, 924–935 (2012).
24. I. M. Nilsson, G. von Heijne, Determination of the distance between the oligosaccharyltransferase active site and the endoplasmic reticulum membrane. *J. Biol. Chem.* **268**, 5798–5801 (1993).
25. M. Orzáez, J. Salgado, A. Giménez-Giner, E. Pérez-Payá, I. Mingarro, Influence of proline residues in transmembrane helix packing. *J. Mol. Biol.* **335**, 631–640 (2004).
26. K. Braunger, S. Pfeffer, S. Shrimal, R. Gilmore, O. Berninghausen, E. C. Mandon, T. Becker, F. Förster, R. Beckmann, Structural basis for coupling protein transport and N-glycosylation at the mammalian endoplasmic reticulum. *Science* **360**, 215–219 (2018).
27. K. Hristova, C. E. Dempsey, S. H. White, Structure, location, and lipid perturbations of melittin at the membrane interface. *Biophys. J.* **80**, 801–811 (2001).
28. I. Mingarro, E. Pérez-Payá, C. Pinilla, J. R. Appel, R. A. Houghten, S. E. Blondelle, Activation of bee venom phospholipase A2 through a peptide-enzyme complex. *FEBS Lett.* **372**, 131–134 (1995).
29. M. Bañó-Polo, L. Martínez-Gil, F. N. Barrera, I. Mingarro, Insertion of bacteriorhodopsin helix C variants into biological membranes. *ACS Omega* **5**, 556–560 (2020).
30. H. L. Scott, J. M. Westerfield, F. N. Barrera, Determination of the membrane translocation pK of the pH-low insertion peptide. *Biophys. J.* **113**, 869–879 (2017).
31. L. Martínez-Gil, A. E. Johnson, I. Mingarro, Membrane insertion and biogenesis of the *Turnip Crinkle Virus* p9 movement protein. *J. Virol.* **84**, 5520–5527 (2010).
32. S. Yohannan, D. Yang, S. Faham, G. Boulting, J. Whitelegge, J. U. Bowie, Proline substitutions are not easily accommodated in a membrane protein. *J. Mol. Biol.* **341**, 1–6 (2004).
33. A. Senes, D. E. Engel, W. F. Degrado, Folding of helical membrane proteins: The role of polar, GxxxG-like and proline motifs. *Curr. Opin. Struct. Biol.* **14**, 465–479 (2004).
34. C. Sen, V. Logashree, R. D. Makde, B. Ghosh, Amino acid propensities for secondary structures and its variation across protein structures using exhaustive PDB data. *Comput. Biol. Chem.* **110**, 108083 (2024).
35. J. A. Killian, G. von Heijne, How proteins adapt to a membrane-water interface. *Trends Biochem. Sci.* **25**, 429–434 (2000).
36. M. Bañó-Polo, C. Baeza-Delgado, M. Orzáez, M. A. Marti-Renom, C. Abad, I. Mingarro, Polar/ionizable residues in transmembrane segments: Effects on helix-helix packing. *PLOS ONE* **7**, e44263 (2012).
37. K. Ojemalm, T. Higuchi, P. Lara, E. Lindahl, H. Suga, G. von Heijne, Energetics of side-chain snorkeling in transmembrane helices probed by nonproteinogenic amino acids. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 10559–10564 (2016).
38. H. T. Kratochvil, L. C. Watkins, M. Mravic, J. L. Thomaston, J. M. Nicoludis, N. H. Somberg, L. Liu, M. Hong, G. A. Voth, W. F. DeGrado, Transient water wires mediate selective proton transport in designed channel proteins. *Nat. Chem.* **15**, 1012–1021 (2023).
39. H. T. Kratochvil, R. W. Newberry, B. Mensa, M. Mravic, W. F. Degrado, Spiers memorial lecture: Analysis and de novo design of membrane-interactive peptides. *Faraday Discuss.* **232**, 9–48 (2021).
40. J. Koehler, N. Woetzel, R. Staritzbichler, C. R. Sanders, J. Meiler, A unified hydrophobicity scale for multispan membrane proteins. *Proteins* **76**, 13–29 (2009).
41. M. A. Lomize, I. D. Pogozheva, H. Joo, H. I. Mosberg, A. L. Lomize, OPM database and PPM web server: Resources for positioning of proteins in membranes. *Nucleic Acids Res.* **40**, D370–D376 (2012).
42. A. Senes, D. C. Chadi, P. B. Law, R. F. S. Walters, V. Nanda, W. F. DeGrado, E_z, a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: Derivation and applications to determining the orientation of transmembrane and interfacial helices. *J. Mol. Biol.* **366**, 436–448 (2007).
43. L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction. *J. Open. Source Softw.* **3**, 861 (2018).
44. The UniProt Consortium, UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
45. L. Martínez-Gil, J. Pérez-Gil, I. Mingarro, The surfactant peptide KL_4 sequence Is inserted with a transmembrane orientation into the endoplasmic reticulum membrane. *Biophys. J.* **95**, L36–L38 (2008).
46. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy. *Nature* **585**, 357–362 (2020).
47. S.-Q. Zhang, D. W. Kulp, C. A. Schramm, M. Mravic, I. Samish, W. F. Degrado, The membrane- and soluble-protein helix-helix interactome: Similar geometry via different interactions. *Structure* **23**, 527–541 (2015).
48. J. Liang, K. A. Dill, Are proteins well-packed? *Biophys. J.* **81**, 751–766 (2001).
49. N. F. Polizzi, W. F. DeGrado, A defined structural unit enables de novo design of small-molecule–binding proteins. *Science* **369**, 1227–1233 (2020).
50. D. Eisenberg, R. M. Weiss, T. C. Terwilliger, The helical hydrophobic moment: A measure of the amphiphilicity of a helix. 299, 371–374 (1982).
51. A. Sääf, E. Wallin, G. von Heijne, Stop-transfer function of pseudo-random amino acid segments during translocation across prokaryotic and eukaryotic membranes. *Eur. J. Biochem.* **251**, 821–829 (1998).
52. T. Metsalu, J. Vilo, ClustVis: A web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Res.* **43**, W566–W570 (2015).

Acknowledgments

Funding: This work was supported by grants PID2020-119111GB-I00, PID2023-152568NB-I00, and PRX21/00348 from the Spanish Ministry of Science, Innovation and Universities (MCIN/AEI/10.13039/501100011033) and CIPROM/2022/062 from the Generalitat Valenciana (to I.M.). R.K. was supported by the U.S. Department of Defense (DoD) through the National Defense Science & Engineering Graduate (NDSEG) Fellowship Program. **Author contributions:** Conceptualization: I.M., B.G., M.B.-P., W.F.D., M.M.S.d.P., and R.K. Methodology: B.G., R.K., M.B.-P., K.C., M.J.G.-M., F.H., L.M.-G., W.F.D., G.v.H., and I.M. Investigation: B.G., R.K., M.B.-P., K.C., M.J.G.-M., H.J., L.M.-G., and F.H. Visualization: B.G., R.K., K.C., W.F.D., and I.M. Supervision: L.M.-G., W.F.D., and I.M. Writing—original draft: B.G., R.K., K.C., G.v.H., W.F.D., and I.M. Writing—review and editing: B.G., R.K., M.J.G.-M., M.M.S.d.P., F.H., L.M.-G., G.v.H., W.F.D., and I.M. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 23 August 2024
Accepted 15 January 2025
Published 19 February 2025
10.1126/sciadv.ads6804