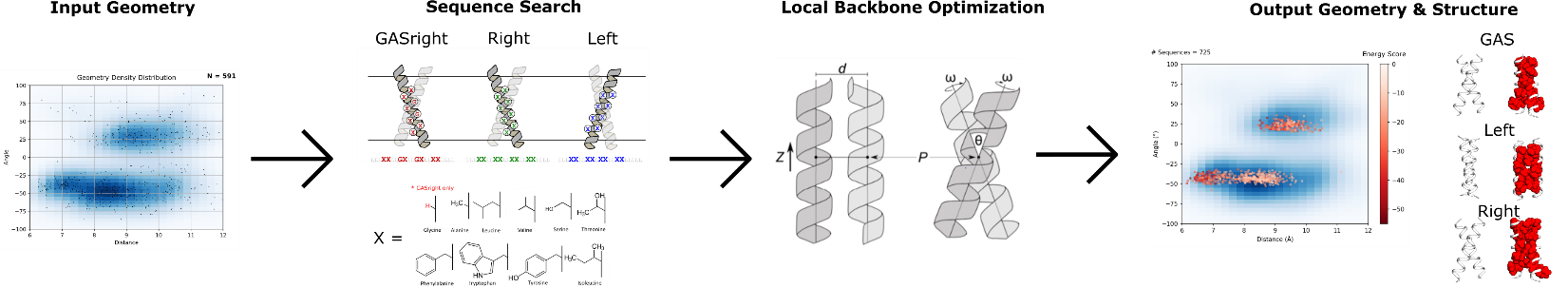
**Van der Waals packing facilitates membrane protein association**

**Abstract**

- TBD

**Graphical Abstract**



* Might need to find a way to combine 3 and 4 so I can add in fluorescence data as well

**Premise/Intro**

Membrane proteins comprise 25-30% of the proteins found within protein-coding genes of various organisms (Fagerberg et al., 2010). Proper membrane protein folding is critical for essential biological functions, including cell signaling, ion balance, and gene regulation. Misfolding of membrane proteins has been found to be involved in several human diseases such as Parkinson’s, cystic fibrosis, and cancer (Gregersen et al., 2006; Sanders & Myers, 2004). To understand how protein misfolding plays a role in disease states and progression, it is necessary to investigate how these proteins fold. However, studying membrane protein folding is inherently a difficult challenge because of their hydrophobic nature. Membrane proteins are difficult to express in yields high enough for biophysical experiments, and purification and solubilization of these proteins often lead to aggregation or unfolding (Carpenter et al., 2008). To combat these challenges, much of the research studying membrane protein folding is focused on understanding the biophysical forces that govern the folding process. Investigation of the biophysical forces that govern protein folding will allow us to determine why specific mutations result in diseases caused by misfolding. In addition, this knowledge can be applied to design new therapeutics that specifically target proteins in these misfolded states. Understanding the forces that enable membrane proteins to fold will contribute to our knowledge of how these fundamental forces are involved in human health and disease.

Proper membrane protein folding is regulated by a distribution of stabilizing hydrogen bonds, weak polar interactions, and van der Waals forces between the unfolded and folded states., but research is lacking on the contribution of van der Waals packing. This force is particularly important due to the nature of van der Waals interactions: Even if hydrogen bonding or polar interactions play a significant stabilizing role, because van der Waals occurs between any nonbonded atoms in close contact, it is a necessary force that is always present within the folded state. This means that van der Waals packing is essential for folding, but the extent at which packing can be a driving force for membrane protein folding is unclear. Without understanding the extent at which van der Waals packing contributes to folding, we cannot complete our understanding of how these other forces contribute to membrane protein association and folding.

The contribution of van der Waals packing to membrane protein folding can be broken down into three distinct interactions: lipid-lipid packing, lipid-protein packing, and protein-protein packing. Protein-protein (or sidechain) packing, is a technically feasible starting point because of the ability to manipulate sequences and determine changes in stability due to mutation. Previous research has demonstrated that disruption of packing within the core of bacteriorhodopsin destabilizes protein structure, suggesting that it is a necessary force for proteins to reach optimal stability (Faham et al., 2004; Joh et al., 2009). Additionally, a recent study using membrane protein design has shown that optimized sidechain packing can stabilize the folded state of phospholamban, demonstrating that packing can drive protein stability (Mravic et al., 2019). Although these studies show that sidechain packing plays a crucial role in stabilizing membrane protein structure in individual systems, the energetic contribution of sidechain packing to the folded state of membrane proteins more generally has not yet been determined (Hong, 2014). In this study, we use homodimerization as a simple and tractable model system to characterize the extent at which sidechain packing is a driving force for membrane protein association (Popot & Engelman, 1990). Previous research using this model system has measured the contributions of both hydrogen bonding and weak polar interactions in the membrane and determined that these forces can drive membrane protein folding (Johnson et al., 2007; Yano et al., 2002; Zhou et al., 2001). Using a large-scale computational design on common dimeric backbone geometries found within the PDB paired with a complementary high throughput method to measure transmembrane (TM) association, we have determined that sidechain packing is a weak driving force involved in TM association.

**Results**

A diagram of a graph

Description automatically generated with medium confidence**Figure MP\_design**

Figure MP\_design: Membrane protein dimer design. (a) Helices within close contact were extracted from the Orientations of Proteins in membranes (OPM) in September 2019. The geometric terms were extracted and overlaid over the kernel density estimation of the membrane protein contact space. (b) Design algorithm schematic used to design structures for this study. (c) Geometries and energy score of the designed sequences overlaid on the membrane protein contact space.

**Extracting membrane protein geometries from the Protein Databank (PDB)**

We began by analyzing membrane protein structures found in the PDB. To do so, we extracted all membrane protein structures solved through x-ray crystallography and cryo-EM found in the Orientations of Proteins in Membranes (OPM) database in September 2019 and any two helices in close contact were extracted as a helical pair. The geometric parameters for each of these pairs were extracted and plotted as a scatterplot that was used to determine the density distribution in geometric space between two parameters: distance and crossing angle (**Figure MP\_design A**). We were able to identify three high-density regions: between 8.5 and 10.5 angstroms in distance with a left-handed crossing angle from 20 to 40, between 8.5 and 10.5 angstroms with a right-handed crossing angle between -30 and -60, and between 6.5 and 7.5 angstroms in distance with a right-handed crossing angle between -25 and -55. This third region is known as GASright, which is known to be stabilized by a combination of sidechain packing and hydrogen bonding (cite). To investigate the impact of sidechain packing in the absence of other forces, we decided to use GASright as a control region where we can compare the impact of the additional stability included from hydrogen bonding. For the rest of this study these regions will be referred to as GASright, Left, and Right. Each of these geometric regions were used as inputs for computational design, with the expectation that these common regions would be most amenable for protein design.

**Computational Design Strategy**

In order to reduce heterogeneity of our sequences, we have chosen to design sequences using a poly-Leu backbone sequence (Zhou et al., 2001). This strategy ensures that all designs have the same TM domain length, reduces variability in hydrophobicity, and aims to normalize expression for each of these designs. For each of the geometric regions, we identified 8 interfacial positions at the helix-helix interface for design. This has been implemented successfully in the past to determine the dimerization propensity of GASright proteins (Anderson et al., 2017). When mutating poly-leucine for GASrights, the interfacial positions are found in a LLLxxLLxxLLxxLLxxLILI pattern, where x is an interfacial position that will be allowed to alternate for design. This set of interfacial positions ensures that the designed interface is centered in the middle of the membrane. For the right-handed design region, we were able to use the same set of interfacial positions as GASrights for sequence design. However, for the left-handed design region, we found that alternating the crossing angle resulted in different interfacial positions. We chose three different patterns for the left-handed interface, accommodating for interfaces typically found in knobs-into-holes packing and leucine zippers found in left-handed coiled coils (cite). To ensure that we designed sequences that would insert properly into the membrane, we used a library of the most prevalent amino acids found in membrane proteins (Liu et al., 2002), with Glycine only being used in the GASright sequences to prevent from designing sequences with interhelical hydrogen bonding in the other regions.

To computationally design our dimers of interest, we utilized a combination of fixed backbone design with iterative backbone refinement (Huang et al., 2022; Kuhlman et al., 2003; Kulp et al., 2012; Nash et al., 2015; Senes, 2011). We randomly generated 10000 geometries within the high-density regions (MP\_design A) as inputs for our design algorithm (Methods). A random geometry was used as the fixed backbone input for each design, and positions on the defined interface were allowed to mutate between all amino acids in the given library. Each sequence was evaluated by a simple set of energetic terms that make up our design energy score: van der Waals packing, hydrogen bonding, and implicit solvation(Krivov et al., 2009; Lazaridis, 2003; MacKerell et al., 1998). The sequence with the lowest energy score then underwent a local backbone minimization to find the most stable structure for the designed sequence. Our protein design algorithm resulted in a set of 1075 wild type (WT) design sequences with a range of both energy and stability (Figure MP\_design C; change this figure, I think it’s the one for post filtering from ngs). However, due to the large scale of this study it would be difficult to determine the success of our design algorithm using these methods. As an alternative, we added two types of mutants to our libraries: clashing and void mutants. For each of these mutations, each interfacial amino acid is individually mutated to either isoleucine (clash) or alanine (void) (**Figure clash dataset A**). The clashing mutations result in a significant overlap between atoms in our structures, while void mutants result in a large hole at the helix-helix interface. Each of these mutations is expected to result in a significant decrease in the dimerization propensity, allowing us to filter for sequences with structures that associate as expected by our design principles.

**A screenshot of a computer

Description automatically generatedFigure Experimental\_methodology**

Figure experimental\_methodology: Experimental methods from the cellular level to the interpreted fluorescent data. Plasmids coding for the transmembrane sequence of interest are expressed in *E. coli* cells. These are then grown overnight and can be evaluated for fluorescence using both a high and low throughput method. Using high throughput fluorescence activated cell sorting in combination with next generation sequencing, we are able to get fluorescence data for 1000s of sequences at a time. A subset of those sequences is evaluated using low throughput TOXGREEN run in a fluorescence plate reader. These sequences are normalized to our positive control GpA, allowing us to transform and normalize the data by our positive control.

**Determining the dataset with experiment and filtering**

To assess the dimerization propensity of our designs, we used sort-seq (**cite sort-seq**). This high throughput assay utilizes a combination of fluorescent activated cell sorting (FACS) and next generation sequencing (NGS) to determine the fluorescence profile of a library of plasmids coding for our designs cloned into *E. coli* (**Figure exp\_method high throughput**). To evaluate the dimerization propensity of our designs, we use a linear transformation using a set of controls and individual clones from our libraries to the values found in TOXGREEN (**Figure exp\_method low throughput**). This transformation normalizes the association strength for each of our libraries, allowing us to compare their association strength.

After running each of the libraries through sort-seq in triplicate, sequences are filtered for proper insertion in the membrane by their ability to survive in maltose media (**cite sort-seq**). The maltose test recovered x sequences **…maltose detail here after results come back…**

We first looked at the fluorescence distribution of the WT designs found in our libraries and compared them to each group of their respective mutants (**Figure clash dataset B**). We separated the data for each design region by their classification as mutant or WT design, with the WT in dark green, void mutants in light green, and clash mutants in white. It is notable that the fluorescence of the GASright regions spans a larger fluorescence range than both the left- and right-handed regions, demonstrating GASrights ability to self-associate with the addition of hydrogen bonding. However, many proteins in both the left- and right-handed designs are above the fluorescence of our monomeric control (25% GpA), meaning that we have successfully designed sequences that associate in those regions. We then used a t-test to calculate the significant difference between the fluorescence distributions of the WT designs against the mutant groups to determine the impact of each mutation on association. Clashing mutations significantly decrease the association of WTs, with a p-value < 0.005 in each region. This supports our initial hypothesis of using clashing mutations as a method for supporting our designed interfaces. However, void mutations were not found to have a significant effect on association in any of the regions. Although structurally we expected voids to decrease association, both left- and right-handed designs show an increase in their void mutants’ respective fluorescence, while void mutations on GASright proteins show an overall decrease in fluorescence that is not significant. We speculate that creating these voids in dimers may not necessarily be destabilizing because of the potential for these sequences to undergo dimerization with an alternative structure that fills the void, resulting in structures that are stabilized more favorably than their WTs for sequences primarily stabilized by tight sidechain packing.

A screenshot of a computer screen

Description automatically generated**Figure clash dataset**

Figure clash dataset. A) Mutations made to each WT design. Clash mutants result in overlaps between atoms from each helix, Void mutants result in large pockets lacking atoms. B) The WT (dark green) fluorescence distribution is compared to the fluorescence distributions of their corresponding mutants. C) Separated graphs for normalized fluorescence against the design energy score in A.

**Structural and energetic validation by clashing mutations**

We next decided to evaluate the structures with interfaces supported by our clashing mutations. We limited this dataset to WT sequences with at least one clashing mutation with a mild monomeric cutoff (< 35% GpA) or with fluorescence 50% less than the WT. We then plotted the fluorescence against our design energy score, allowing us to evaluate the accuracy of our design method (**Figure clash dataset C**). There were two things that were apparent from this data: GASrights are much more stable and appear to follow our design principles better than the other regions. GASright sequences expressed the most fluorescence of all three design regions, suggesting that interhelical hydrogen bonding is necessary for high level association in membrane proteins. Additionally, our design energy score expected many of the left-handed structures to be more stable than right-handed, which can be seen in Figure clash dataset B. However, it doesn’t appear that our design energy score is able to capture differences between many structures in the lower threshold of association (0.25-0.4). This suggests that our energetic parameters are not well tuned to evaluate low levels of homodimer association.

* If wanted more info here, could also have a dataset where we get rid of the clashing data (below 25%) and analyze that

**top 5**

A screenshot of a computer

Description automatically generated

Figure top 5: Frontal and side view of predicted structures for the top 5 fluorescent sequences for each region.

As a representation of our structures, we’ve extracted the top 5 structures from each region (figure top 5). Each of these sequences are quite well packed, with the left- and right-handed sequences having small voids at the interface compared to GASrights. It is possible that GASright sequence stability is influenced by increased packing compared to the other regions, accompanied by the increased stability with interhelical hydrogen bonding. To verify the structures of these sequences, we ran these through TMDOCK (Lomize & Pogozheva, 2017), an online membrane protein modeling tool. 10 of the 15 sequences have a Cα RMSD below 1.5 angstroms, showing good agreement between our designed structure and the TMDOCK predictions.

* Could make graphs for any of the data in the table

**Delta G data**

A screenshot of a graph

Description automatically generated

Figure DeltaG. Top:DeltaG vs the predicted total energy of all of our structures. Bottom: DeltaG vs the predicted van der Waals energy.

Finally, we converted the fluorescence to biophysical ΔG using a method from an *in vitro* study on GASright proteins (Díaz Vázquez et al., 2023). We then compared these ΔG values to the energetics from our design algorithm (**figure deltaG data**). Although our energetics do not seem to capture the overall trend of stability, it allows us to visualize the distribution of the calculate biophysical stability of these sequences. This data shows that both left- and right-handed structures, stabilized to a ΔG around -3 to -3.5 while many of the most stable GASrights are calculated to have a ΔG -4 to -5. This data suggests that interactions that stabilize the left- and right-handed structures result in around 60% of the stability found in GASrights. **… I think we need some way here to say that these proteins are likely to be stabilized just by vdW and nothing else…** Suggesting that packing has a major impact on stability, but other contributions from hydrogen bonding to polar interactions are necessary for many interactions that result in highly stable structures.

* Maybe just bin this data instead of showing it against energetics?

**Discussion and Conclusion**

**…**

**Other things that can be included:**

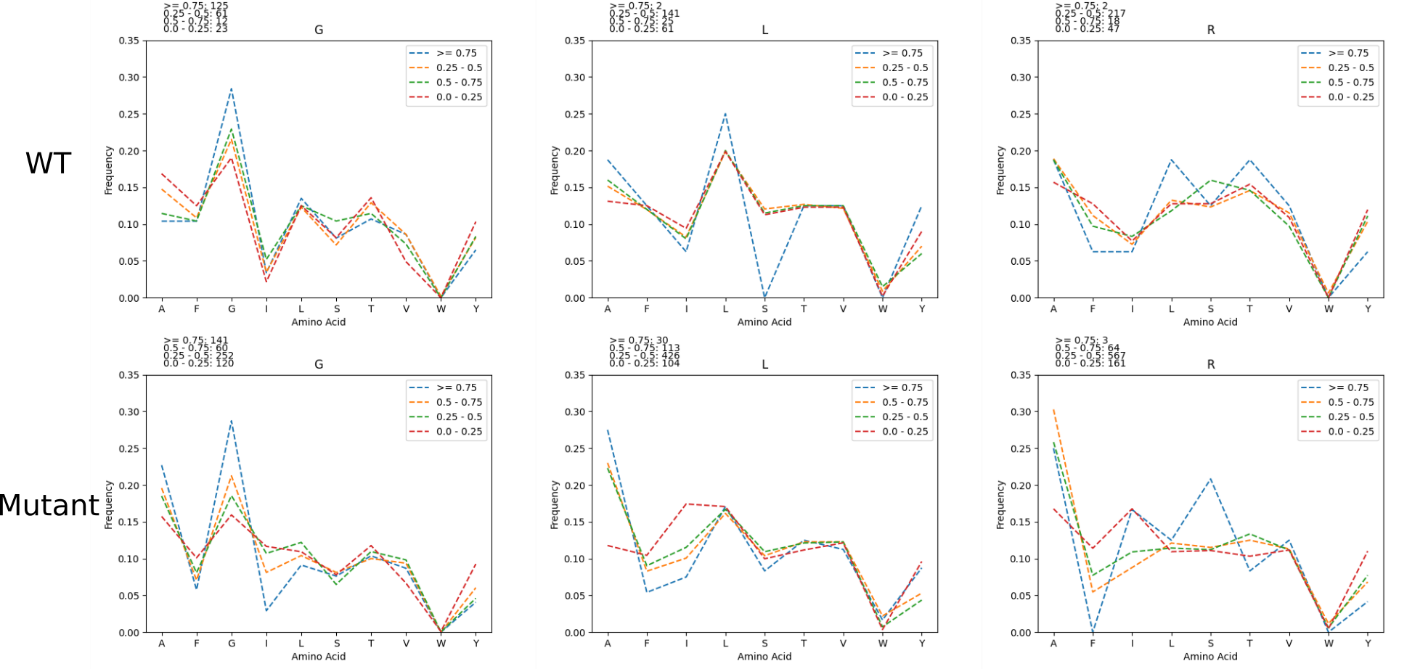
**Amino acid distribution per Fluorescence**

Figure: The amino acid frequency was calculated for sequences based on their fluorescence (>= 0.75, 0.5 – 0.75, 0.25-0.5, and < 0.25). This was done for designs and all mutant sequences.

* In all regions, isoleucine is more prevalently found in sequences with lower fluorescence
* Phenylalanine is also more prevalently found in mutant sequences with lower fluorescence
* Tyrosine is kind of variable, but more prevalently found in mutant sequences with lower fluorescence
  + Suggest that these AAs disrupt association
* Alanine is more prevalent in sequences with higher fluorescence
  + Suggests that it helps facilitate association
* Distributions of hbond AAs (Ser, Thr) are similar throughout, suggesting that potential hbonding by these AAs is not the main driving force for association

**Figure A diagram of a dna sequence

Description automatically generatedA diagram of a dna sequence

Description automatically generatedexp\_design**

Figure exp\_design. Controlling for expression and insertion variability. (a) [Design algorithm name] was run on a 21 amino acid poly-leucine backbone, allowing for 8 variable amino acid positions. GASright and right handed positions are identical, while left handed positions are based on knobs into holes packing motifs found within coiled coils. (b) TOXGREEN fuses the designed protein to maltose binding protein (MBP) in the periplasm and ToxR, a dimeric transcription factor, in the cytoplasm. Dimerization of designs results in the dimerization of ToxR, and subsequently the transcription and translation of GFP as a readout of association strength.

Sequence logo info:

* Have these images somewhere for sequences from the PDB

We then took the WT sequences found in these dense regions and created sequence logos, aiming to determine if there are any common motifs to be used in design. These sequence logos show that the region in the bottom left of Figure MP\_design A is the GASright, a well characterized dimerization motif with a known sequence signature where small amino acids glycine, alanine, and serine are typically found at the interface (Figure MP\_design logos). This sequence signature of small amino acids allows GASright helices to come into close contact, resulting in stabilization by a combination of tight van der Waals packing and interhelical hydrogen bonding. There were no common motifs found in the other regions, but the GASright sequence motif will be used to design sequences in the GASright region.

Mutants:

We chose two of the least stable mutants by energy (clash) or largest differences in solvent accessible surface area (void) for each sequence **(supplement methods)**.

Baseline term:

To reduce computational expense, we developed a baseline energetic term that functioned as an estimate for the energy of the monomeric state, allowing us to accept sequences with the best stability between the dimer-monomer state.

**Delta Fluorescence Figure**

A graph with lines and numbers

Description automatically generated

Fluorescence of wild type designs subtracted by fluorescence of their corresponding mutants is plotted above.

* Plot this as an x vs y plot, might likely be supplement of just a subset of the previous figure
  + However, there are a variety of sequences above the 40% GpA in both the right and left handed regions (conservative limit for association)
    - OR take the data for some of those sequences from TOXGREEN and compare them in an x vs y plot or something? If done here, then can also talk about the expression checks through westerns (Supplement western)

**A screenshot of a graph

Description automatically generatedPercent SASA figure**

Anderson, S. M., Mueller, B. K., Lange, E. J., & Senes, A. (2017). Combination of Cα-H Hydrogen Bonds and van der Waals Packing Modulates the Stability of GxxxG-Mediated Dimers in Membranes. *J Am Chem Soc*, *139*(44), 15774-15783. <https://doi.org/10.1021/jacs.7b07505>

Carpenter, E. P., Beis, K., Cameron, A. D., & Iwata, S. (2008). Overcoming the challenges of membrane protein crystallography. *Curr Opin Struct Biol*, *18*(5), 581-586. <https://doi.org/10.1016/j.sbi.2008.07.001>

Díaz Vázquez, G., Cui, Q., & Senes, A. (2023). Thermodynamic analysis of the GAS. *Biophys J*, *122*(1), 143-155. <https://doi.org/10.1016/j.bpj.2022.11.018>

Fagerberg, L., Jonasson, K., von Heijne, G., Uhlén, M., & Berglund, L. (2010). Prediction of the human membrane proteome. *Proteomics*, *10*(6), 1141-1149. <https://doi.org/10.1002/pmic.200900258>

Faham, S., Yang, D., Bare, E., Yohannan, S., Whitelegge, J. P., & Bowie, J. U. (2004). Side-chain contributions to membrane protein structure and stability. *J Mol Biol*, *335*(1), 297-305. <https://doi.org/10.1016/j.jmb.2003.10.041>

Fleming, K. G., Ackerman, A. L., & Engelman, D. M. (1997). The effect of point mutations on the free energy of transmembrane alpha-helix dimerization. *J Mol Biol*, *272*(2), 266-275. <https://doi.org/10.1006/jmbi.1997.1236>

Gregersen, N., Bross, P., Vang, S., & Christensen, J. H. (2006). Protein misfolding and human disease. *Annu Rev Genomics Hum Genet*, *7*, 103-124. <https://doi.org/10.1146/annurev.genom.7.080505.115737>

Hong, H. (2014). Toward understanding driving forces in membrane protein folding. *Arch Biochem Biophys*, *564*, 297-313. <https://doi.org/10.1016/j.abb.2014.07.031>

Huang, B., Xu, Y., Hu, X., Liu, Y., Liao, S., Zhang, J., . . . Liu, H. (2022). A backbone-centred energy function of neural networks for protein design. *Nature*, *602*(7897), 523-528. <https://doi.org/10.1038/s41586-021-04383-5>

Joh, N. H., Oberai, A., Yang, D., Whitelegge, J. P., & Bowie, J. U. (2009). Similar energetic contributions of packing in the core of membrane and water-soluble proteins. *J Am Chem Soc*, *131*(31), 10846-10847. <https://doi.org/10.1021/ja904711k>

Johnson, R. M., Hecht, K., & Deber, C. M. (2007). Aromatic and cation-pi interactions enhance helix-helix association in a membrane environment. *Biochemistry*, *46*(32), 9208-9214. <https://doi.org/10.1021/bi7008773>

Krivov, G. G., Shapovalov, M. V., & Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, *77*(4), 778-795. <https://doi.org/10.1002/prot.22488>

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, *302*(5649), 1364-1368. <https://doi.org/10.1126/science.1089427>

Kulp, D. W., Subramaniam, S., Donald, J. E., Hannigan, B. T., Mueller, B. K., Grigoryan, G., & Senes, A. (2012). Structural informatics, modeling, and design with an open-source Molecular Software Library (MSL). *J Comput Chem*, *33*(20), 1645-1661. <https://doi.org/10.1002/jcc.22968>

Lazaridis, T. (2003). Effective energy function for proteins in lipid membranes. *Proteins*, *52*(2), 176-192. <https://doi.org/10.1002/prot.10410>

Liu, Y., Engelman, D. M., & Gerstein, M. (2002). Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biol*, *3*(10), research0054. <https://doi.org/10.1186/gb-2002-3-10-research0054>

Lomize, A. L., & Pogozheva, I. D. (2017). TMDOCK: An Energy-Based Method for Modeling α-Helical Dimers in Membranes. *J Mol Biol*, *429*(3), 390-398. <https://doi.org/10.1016/j.jmb.2016.09.005>

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., . . . Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, *102*(18), 3586-3616. <https://doi.org/10.1021/jp973084f>

Mravic, M., Thomaston, J. L., Tucker, M., Solomon, P. E., Liu, L., & DeGrado, W. F. (2019). Packing of apolar side chains enables accurate design of highly stable membrane proteins. *Science*, *363*(6434), 1418-1423. <https://doi.org/10.1126/science.aav7541>

Nash, A., Notman, R., & Dixon, A. M. (2015). De novo design of transmembrane helix-helix interactions and measurement of stability in a biological membrane. *Biochim Biophys Acta*, *1848*(5), 1248-1257. <https://doi.org/10.1016/j.bbamem.2015.02.020>

Popot, J. L., & Engelman, D. M. (1990). Membrane protein folding and oligomerization: the two-stage model. *Biochemistry*, *29*(17), 4031-4037. <https://doi.org/10.1021/bi00469a001>

Sanders, C. R., & Myers, J. K. (2004). Disease-related misassembly of membrane proteins. *Annu Rev Biophys Biomol Struct*, *33*, 25-51. <https://doi.org/10.1146/annurev.biophys.33.110502.140348>

Senes, A. (2011). Computational design of membrane proteins. *Curr Opin Struct Biol*, *21*(4), 460-466. <https://doi.org/10.1016/j.sbi.2011.06.004>

Yano, Y., Takemoto, T., Kobayashi, S., Yasui, H., Sakurai, H., Ohashi, W., . . . Matsuzaki, K. (2002). Topological stability and self-association of a completely hydrophobic model transmembrane helix in lipid bilayers. *Biochemistry*, *41*(9), 3073-3080. <https://doi.org/10.1021/bi011161y>

Zhou, F. X., Merianos, H. J., Brunger, A. T., & Engelman, D. M. (2001). Polar residues drive association of polyleucine transmembrane helices. *Proc Natl Acad Sci U S A*, *98*(5), 2250-2255. <https://doi.org/10.1073/pnas.041593698>