Assessing van der Waals packing as a driving force in membrane protein association and folding

Gilbert Loiseau
Advisor: Alessandro Senes

Committee Meeting
June 10th, 2022
1:00 PM

Introduction

Proper membrane protein folding is necessary for essential biological functions such as cell signaling and gene regulation. Misfolding of membrane proteins often leads to disease phenotypes including growth defects and cancer. A variety of forces contribute to proper membrane protein folding including hydrogen bonding, weak polar interactions, and van der Waals packing. In order to fully understand how membrane proteins fold for proper function, it is necessary to elucidate the energetic contribution of each of these interactions to the folded state. The ability to drive the transition from the unfolded to the folded state has been characterized and quantified for hydrogen bonding and weak polar interactions, but research is lacking on the contribution of van der Waals packing. Previous research has demonstrated that disruption of packing within the core of bacteriorhodopsin destabilizes the protein structure (Faham et al., 2004; Joh et al., 2009) while membrane protein design has shown that optimized packing stabilizes a redesigned phospholamban structure (Mravic et al., 2019). However, outside of individual systems, the contribution of van der Waals packing to the folded state of membrane proteins has not yet been determined. With my research, I aim to characterize and quantify the extent at which van der Waals packing contributes to membrane protein association.

Results: Sort-Seq and fluorescence reconstruction of designed homodimers

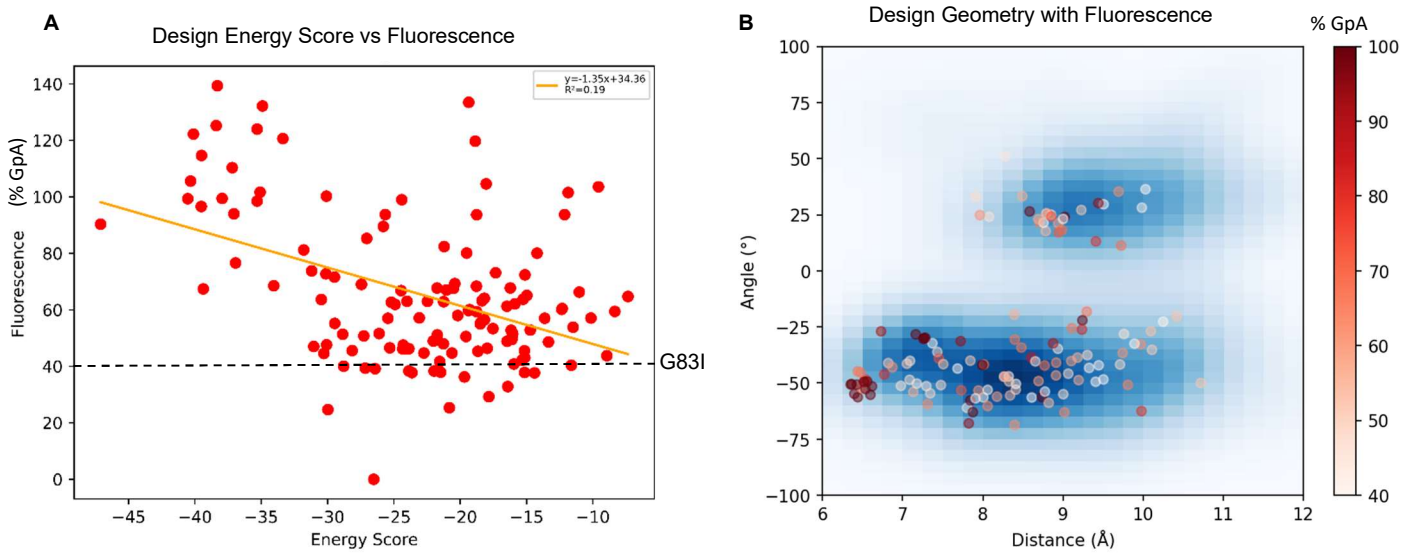


Figure 1: Fluorescence of designed sequences. A) Energy Score plotted against fluorescence for designed sequences. I plotted the expected energy score and the fluorescence output from sort-seq to evaluate the overall correlation between design and our *in vivo* experiments (standard deviation not shown, supplement figure S2). Fluorescence is output as a function of percentage of a known strong dimer, glycophorin A (GpA), which has been shown in the past to dimerize on SDS-Page gels. The dashed black line represents the cutoff for dimerization, represented by the fluorescence of G83I, a monomerizing mutant of GpA. As shown on this graph, there is little correlation between my designed energy scores and fluorescence. **B) Kernel density overlayed with design geometries and their fluorescence.** Kernel density estimation is used to identify the dense areas of geometric space obtained from any membrane proteins with helices in close contact. Overlayed on top of this are the geometries that I ended up using for this run of design. Each is colored a deeper red depending on the fluorescence output from sort-seq. The bar on the right is the fluorescence in terms of percent GpA. There is a deep red area in the GASright region of the space (~6.5 Distance and -50 Angle) as well a scatter of other deep red points. This data demonstrates that I am able to design sequences that fluoresce well *in vivo*, however, the lack of correlation between our designed energy score and the fluorescence output is an issue that needs to be addressed.

For my first sort-seq run, I tested 160 designed sequences. These sequences have a range of stabilities as assessed by an energy score composed of van der Waals packing, hydrogen bonding, and implicit membrane solvation. These designs are primarily driven to associate by an increase in calculated van der Waals packing, allowing me to assess the range at which van der Waals packing affects dimerization. In addition to my designed sequences, I made mutations at the interfacial residues of my sequences and predicted their energy scores. I expected these mutations to recapitulate the trend expected by our energy scores, demonstrating that the structure stabilized by van der Waals packing is either disrupted or stabilized by mutagenesis. This data will give support for the structures of our designed sequences in a high-throughput manner.

After cloning my sequences into the TOXGREEN plasmid, I ran three replicates of sort-seq (methods, Fig. S1A-B) to appropriately quantify the amount of association for my designed sequences and their mutants. These sequences were then sent for next generation sequencing to determine the counts present within the

appropriate bins. Using the counts from NGS (methods), I reconstructed the fluorescence of each of these sequences (Kosuri et al., 2013). In order to simplify our understanding of this fluorescence as dimerization, we compare the fluorescence of each of our sequences to that of a known strong dimer glycophorin A (Walters and DeGrado, 2006). In addition, we can assess whether our sequences are associating by comparing to the monomerizing mutant of GpA, known as G83I which has a glycine to an isoleucine mutation (Anderson thesis, 2019). In figure 3A, I have plotted the correlation of the energy scores of my designed sequences against their reconstructed fluorescence in terms of percent GpA, where 100 is the association strength of GpA. Unfortunately, this first run does not have very strong correlation between our predicted energy scores and the fluorescence. However, there does seem to be a clear trend above a predicted energy score at -30, where our energy score is better predicting the association strength of these dimers.

To confirm the interface of my designed sequences, I analyzed individual designs and their mutants by making graphs of energy score against fluorescence as a function of percent GpA (Fluorescence of design/Fluorescence of GpA x 100%). In my initial design pool, I included sequences containing a sequence signature known as GASright, where small amino acids glycine, alanine, and serine are typically found at the interface, allowing the sequence to be stabilized by interhelical hydrogen bonding (Anderson et al., 2017). Many of these sequences were stabilized by increased hydrogen bonding and were found within the geometric region that these GASright sequences are often found. Within these sequences, I was able to determine good correlation between the energy and *in vivo* fluorescence of our mutant sequences and our WT designs (Fig. S4). In contrast, many of the sequences without this GASright sequence signature show a negative correlation, where mutations that are predicted to be stabilizing have low fluorescence and mutations that are predicted to break dimerization have high fluorescence (data not shown). Sequences without the GASright sequence signature are also found to be quite less stable in energy score. I am still in the process of analyzing this data, aiming to determine what led to negative correlation between designed energy and *in vivo* fluorescence. For my next round of design and sort-seq, I will aim to design sequences that have similar stability as these GASright sequences that were able to recapitulate our expected trend. In doing so, I am likely to observe a similar trend and will be able to confidently identify the interface of sequences stabilized by van der Waals packing.

Conclusion

There is little correlation between the energy scores output from design and the fluorescence recovered from sort-seq (Fig. 1A). However, I was able to successfully design 112 of the 160 sequences, as their reconstructed fluorescence is higher than G83I, a sequence that does not associate. In addition, many of the mutants fluoresce as well, but for sequences that are purely stabilized by van der Waals packing, we often see a negative correlation between predicted energy score and fluorescence. One reason for this may be that we did not run any helix to helix docking between our sequences and our mutants. Through docking, we may be able to identify unexpected interfaces and identify energy scores for our sequences that are more stable, as seen in the fluorescence measurement. In the following months, I plan to run docking on my sequences and continue analyzing the rest of this data before running another design run with saturating mutations and with sequences of increased stability, as detailed in my future work section.

Future Work

To obtain designed sequences with increased stability, I will make slight modifications to my design algorithm. One major change I aim to implement is a method to iteratively make local backbone adjustments during the design procedure, inspired by a recent backbone design paper and from the Baker lab (Huang et al., 2022; Kuhlman et al., 2003). Rather than find a sequence for a fixed backbone structure, this will allow me to design sequences that fit a novel backbone without needing an additional step for backbone optimization. In addition, I will be able to follow the sequence and structural changes, resulting in more than a single sequence for a design run. This should increase the speed at which I am able to design sequences with a range of designed energies.

To more thoroughly ensure that we are designing sequences with the correct interface, I will perform saturating mutagenesis along the entire sequence of every sequence in my design pool rather than just the

interfacial positions. I will then be able to recreate the expected interface for each of our sequences from their reconstructed fluorescence profile in sort-seq. For all sequences that do recapitulate our designed interface, I will analyze the sequences that recapitulate our designed interface, aiming to determine the correlation between our designs and energy scores. If we see sequences that do not recapitulate our expected interface in sort-seq, I will be able to analyze their reconstructed fluorescence to determine if there is a mutational pattern that we did not expect. This will allow me to determine if my design algorithm identified an alternate, less stable interface compared to the one seen in our experiment. Data from these alternate interfaces can be used to further improve our design algorithm for future experiments, such as heterodimer design.

In my initial investigation into the impact of packing in membrane protein association, I analyzed the geometric space of tight packing between pairs of helices found within solved membrane proteins structures. Within this analysis, I also recovered the sequences of these pairs. In order to validate our mutagenesis strategy with known structures outside of my designs, I will identify the homodimer sequences found within these pairs and extract their interfacial sequences. These sequences will then be placed onto standardized poly-Leucine backbones and their stability will be predicted using our energy functions. Any of these sequences that are found to be primarily stabilized by van der Waals packing will be added to the sequence pool to be run on sort-seq. We expect our mutagenesis strategy to accurately convey the interface of these dimers of known structure.

Methods

I aim to explore the effect of sidechain packing on homodimerization. I have identified common geometries from the PDB as structural templates for computational design of sequences of membrane protein sequences. I was able to identify areas of high density in the geometric space obtained from the PDB, allowing me to choose from a variety of geometries to explore for protein design (Fig. S1). Each of these geometries is standardized with a poly-leucine backbone to control for expression and insertion of our sequences (Zhou et al., 2001; Anderson et al., 2017). To specifically vary the sidechain packing contributing to association, positions at the dimer interface are identified for each geometry using solvent accessible surface area (SASA). To calculate the SASA, non-mutated helices are placed at the chosen crossing angle with a tight axial distance that allows for simulation of sequences with tight packing. The SASA for each residue is calculated and any positions with less than the average of the total SASA is considered interfacial and allowed to mutate to an array of amino acids. I then used well-known computational algorithms to filter and search sequence space for amino acid combinations that pack at the dimerization interface (Koehl and Delarue, 1994; Hansmann and Okamoto, 1999). Using a minimalistic set of energy functions that measure van der Waals packing, hydrogen bonding, and membrane implicit solvation (IMM1) (MacKerell et al., 1998; Lazaridis, 2003; Krivov et al., 2009), I have measured the stability of sequences at a variety of geometries, determining geometries where design of well-packed sequences is possible (Figure 1B). The overall workflow of my computational design algorithm is found in the supplement (Figure S2). Using this subset of geometries, I designed a test subset of 160 sequences that have been evaluated using a complementary high-throughput assay.

To evaluate successfully designed sequences, I used TOXGREEN, an *in vivo* dimerization assay that quantifies dimerization propensity through the output of sfGFP (Fig. S1A). This dimerization assay has been optimized for a high-throughput approach known as sort-seq, which combines fluorescence activated cell sorting (FACS) with next-generation sequencing (NGS) to evaluate dimerization propensity of TM domains (Fig. S1B). I expressed an oligo pool library consisting of my designed sequences and cloned them into TOXGREEN plasmids, allowing me to simultaneously study the association of hundreds to thousands of designed constructs. Based on the expression of sfGFP, cells were sorted into different bins with different fluorescence thresholds. These plasmids were then purified out of the cells in each bin and enumerated via NGS. Based on the counts present within each bin, the fluorescence profile, and thus dimerization propensity, was reconstructed for each dimer (Fig. S1C).

Supplementary Figures

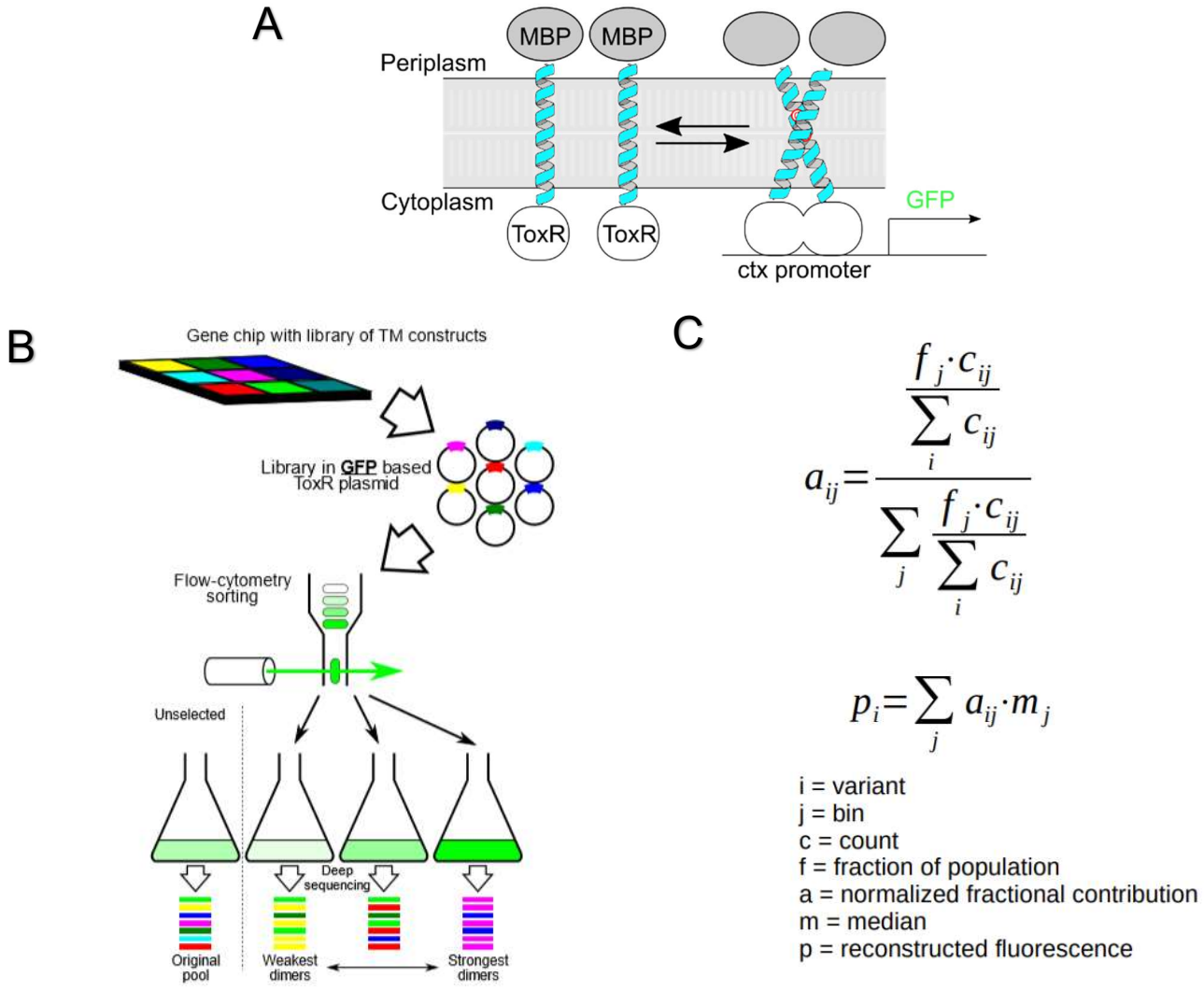


Fig. S1: TOXGREEN sort-seq schematic. A) TOXGREEN. TOXGREEN is an *in vivo* assay that reports TM helix-helix association through the expression of reporter gene sfGFP.
B) Sort-seq. Individual TM domains are synthesized on a chip using oligo pool technology and cloned into TOXGREEN plasmids. In the GFP-based assay, cells will fluoresce based on GFP expression correlating to the dimerization propensity of the expressed TM domain chimera. This GFP expression will be sorted by fluorescence-activated cell sorting (FACS) followed by next-generation sequencing of the bins.
C) Fluorescence Reconstruction. Fluorescence is reconstructed according to the above equation. Sequence counts (c) for each sequence (i) in each bin (j) is multiplied by the fraction of the population of the bin (f). This gives the total percent of a sequence within a particular bin. This is then divided by the total percent of sequence withing all bins, which results in the normalized fractional contribution of a sequence within a bin (a). After getting all of the contributions for a sequence within all bins, and contribution for a sequence per bin (a) is multiplied by the median of its respective bin (m). These are then summed together, resulting in the reconstructed fluorescence of a sequence across all bins (p).

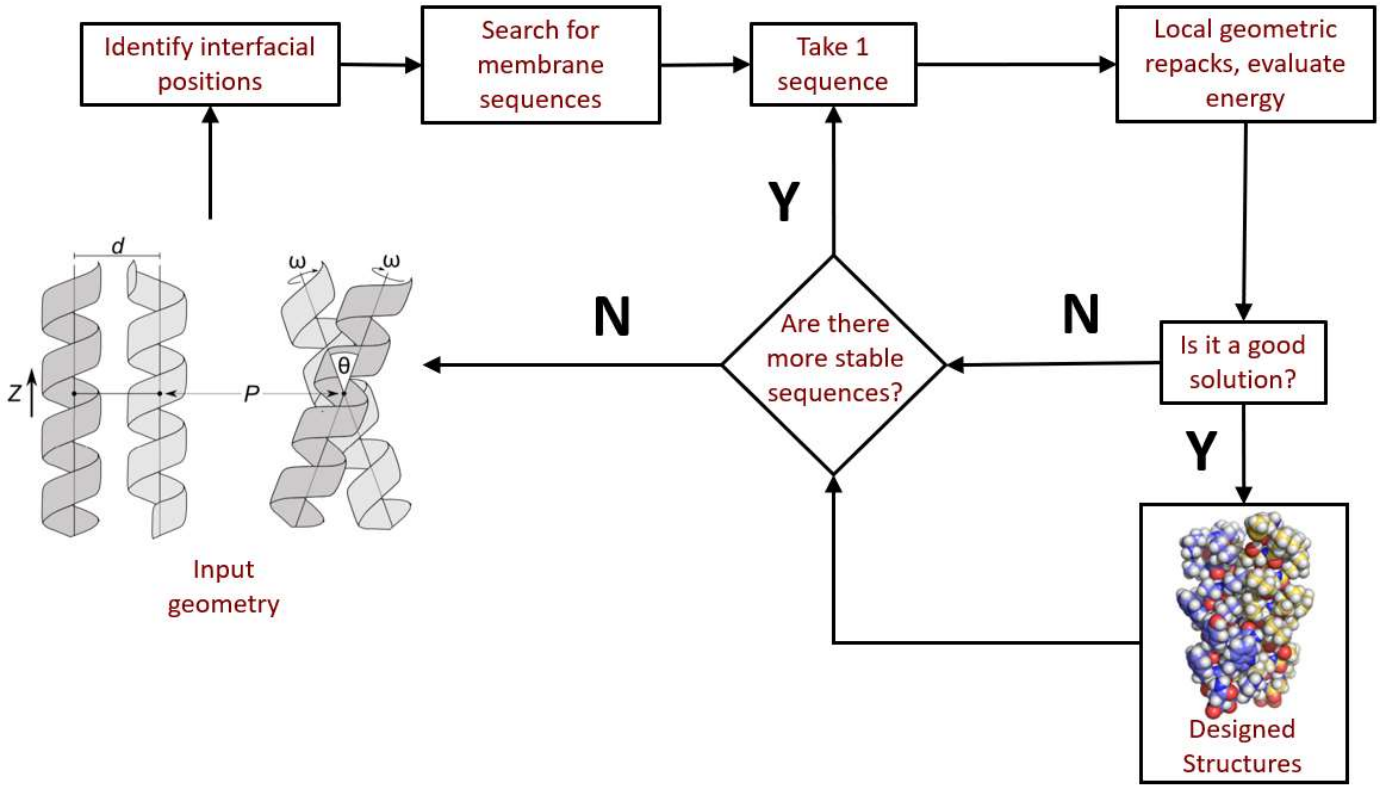


Fig. S2: Flowchart of the computational design algorithm. A point from the density estimate obtained from geometries from the PDB is input as a structural template. Positions found at the interface of this geometry are identified and common computational methods search sequence space for well-packed homodimers. One of those sequences undergoes uses local geometric repacks to search for a geometry with the most stable energy. This energy is compared to the monomeric state and the design is added to a pool of successfully design sequences if it is more stable. The algorithm then repeats this process until there are no more sequences with good packing at the interface. This process can be repeated will other common geometries from the PDB, allowing me to design hundreds of sequences with an array of expected dimerization based on sidechain packing energies.

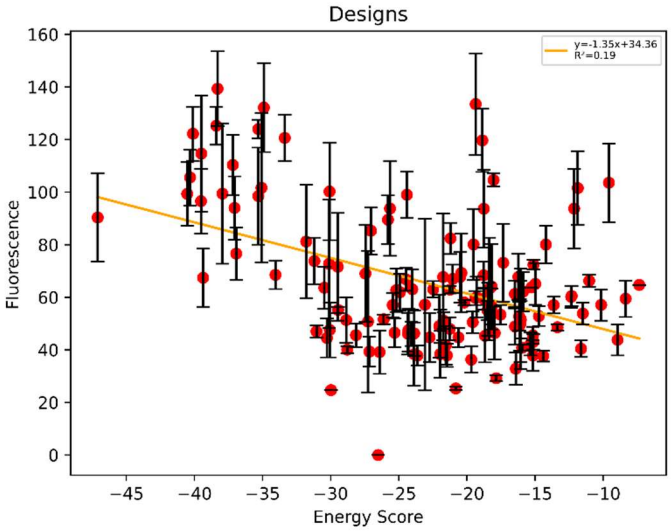
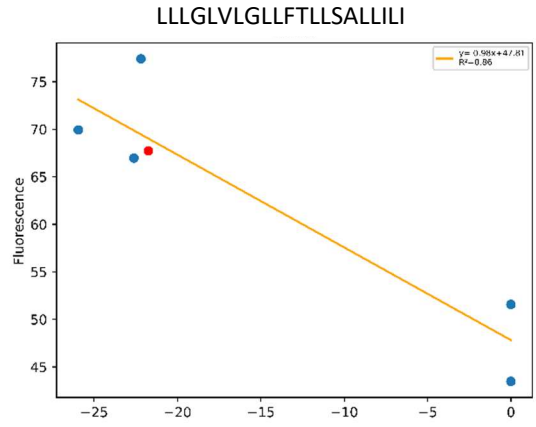
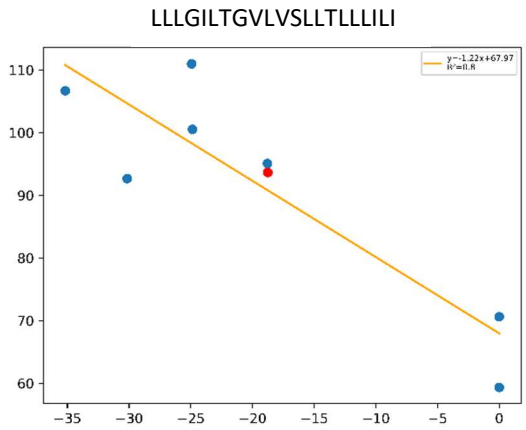
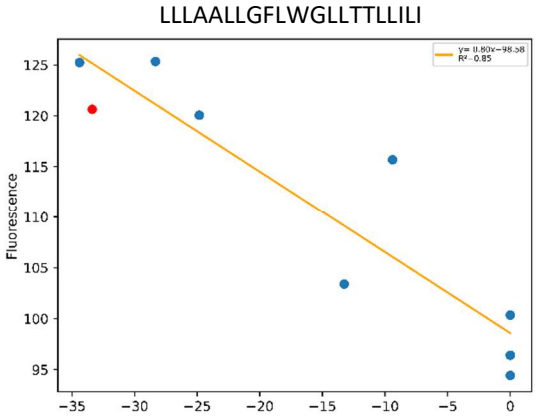
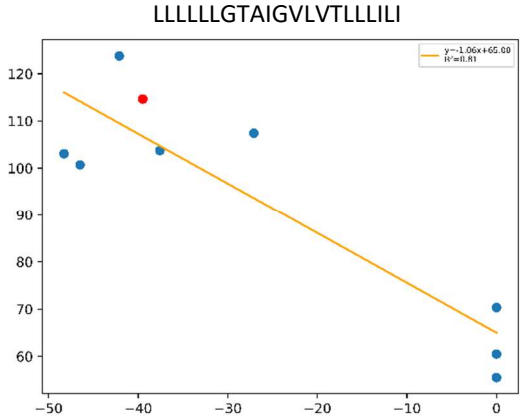


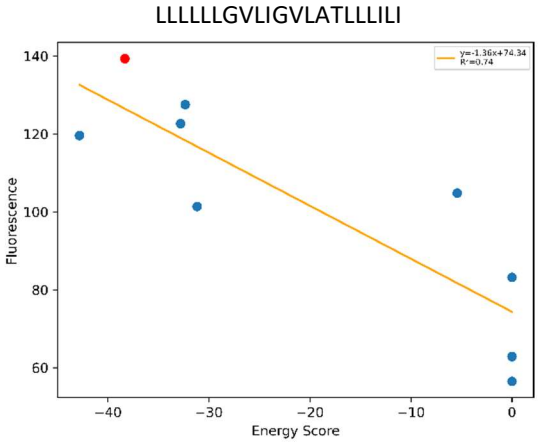
Figure S3: Designed energy scores graphed for correlation against *in vivo* fluorescence. Version of Figure 1A with the fluorescence standard deviations included for each sequence.

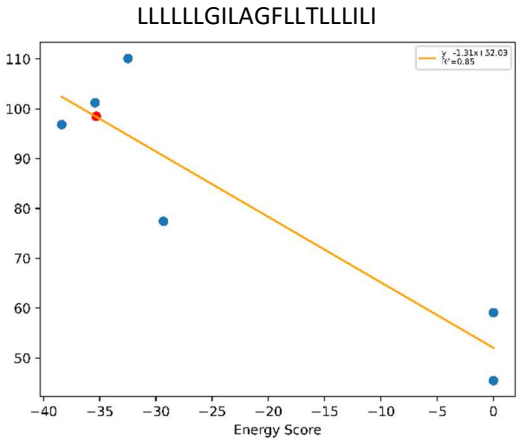












Sequence	xShift	crossingAngle
LLLLLLGILAGFLLTLLLILI	6.53676	-52.6586
LLLLLLGVLIGVLATLLLILI	6.55858	-49.3007
LLLLLLGTAIGVLVTLLLILI	6.35232	-50.6541
LLLAALLGFLWGLLTTLLILI	7.26946	-29.9582
LLLGILTGVLVSLLTLLLILI	7.72338	-31.3153
LLLGLVLGLLFTLLSALLILI	8.39897	-30.639

Figure S4: Sequences with GASright sequence signature and structure. Above: Graphs of correlation between the computed energy score and in vivo fluorescence in terms of percent GpA (fluorescence of sequence/fluorescence of GpA x 100%). The WT design sequence is highlighted in red and the mutant sequences are in blue. For many of these, there is a range of energy and fluorescence, giving us some confidence in our ability to design these sequences with the correct interface. Left: The table contains the geometry information, showing that these sequences are found within the expected regions that GASright forms tight interhelical hydrogen bonding (6-7.5 angstrom distance and -45+/- 10 degrees crossingAngle).

References

1. Faham, S., Yang, D., Bare, E., Yohannan, S., Whitelegge, J.P., and Bowie, J.U. (2004). Side-chain Contributions to Membrane Protein Structure and Stability. Journal of Molecular Biology *335*, 297–305.
2. Joh, N.H., Oberai, A., Yang, D., Whitelegge, J.P., and Bowie, J.U. (2009). Similar Energetic Contributions of Packing in the Core of Membrane and Water-Soluble Proteins. J. Am. Chem. Soc. *131*, 10846–10847.
3. Mravic, M., Thomaston, J.L., Tucker, M., Solomon, P.E., Liu, L., and DeGrado, W.F. (2019). Packing of apolar side chains enables accurate design of highly stable membrane proteins. Science 363, 1418–1423.
4. Zhou, F.X., Merianos, H.J., Brunger, A.T., and Engelman, D.M. (2001). Polar residues drive association of polyleucine transmembrane helices. Proc Natl Acad Sci U S A 98, 2250–2255.
5. Anderson, S.M., Mueller, B.K., Lange, E.J., and 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, 15774–15783.
6. Koehl, P., and Delarue, M. (1994). Application of a Self-consistent Mean Field Theory to Predict Protein Side-chains Conformation and Estimate Their Conformational Entropy. Journal of Molecular Biology 239, 249–275.
7. Hansmann, U.H.E., and Okamoto, Y. (1999). New Monte Carlo algorithms for protein folding. Current Opinion in Structural Biology 9, 177–183.
8. MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102, 3586–3616.
9. Lazaridis, T. (2003). Effective energy function for proteins in lipid membranes. Proteins 52, 176–192.
10. Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77, 778–795.
11. Fleming, P.J.and F.M. Richards (2000) Protein Packing:Dependence on Protein Size, Secondary Structure and Amino Acid Composition. J.Mol.Biol. 299, 487-498.
12. Armstrong, C.R., and Senes, A. (2016). Screening for transmembrane association in divisome proteins using TOXGREEN, a high-throughput variant of the TOXCAT assay. Biochim Biophys Acta 1858, 2573–2583.
13. Anderson, S. M. (2019). Understanding the GAS_{right} motif: sequence, structure, and stability. Thesis. University of Wisconsin, Madison, WI.
14. Kosuri, S., Goodman, D.B., Cambray, G., Mutalik, V.K., Gao, Y., Arkin, A.P., Endy, D., and Church, G.M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. 110, 14024–14029.
15. Huang, B., Xu, Y., Hu, X., Liu, Y., Liao, S., Zhang, J., Huang, C., Hong, J., Chen, Q., Liu, H. (2022). A backbone-centred energy function of neural networks for protein design. Nature 602, 523–528.
16. 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, 1364–1368.
17. Walters, R.F.S., DeGrado, W.F., 2006. Helix-packing motifs in membrane proteins. Proc Natl Acad Sci U S A 103, 13658–13663.