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

**Abstract**

- TBD

- Start calling figures by a descriptive name instead of by numbers so that it’s easier to move them around later

- Start bulletting the story

**Graphical Abstract d**

- Have a nice figure of how my design algorithm works, maybe tie it into the figure 1 (extracted geometries), finally output the protein and show fluorescence

- I also think visually for my presentation, having some sort of pymol driven video would be nice

**Premise/Intro**

- Membrane protein folding and association is important (look at Mravic et al. and other papers to understand how to write this)

* Structure is dictated by the interactions occurring at protein-protein interfaces
* Previous research has shown that hydrogen bonding and polar interactions act as forces that facilitate association
* Van der Waals packing is an necessary force, but the extent at which packing facilitates association is not well understood (why? Limitations of previous research like Mravic and Bowie?)
* This research uses a high throughput approach to study the impact of packing on membrane protein association
* Could talk about the fact that packing is a required feature,
  + not necessarily a driving force (bowie: cost of void is similar to voids in soluble; trading protein-protein interaction for lipid could mean that it’s necessary vs a driving force)
  + mention protein-protein packing vs protein-lipid packing
  + taking a thermodynamic approach
  + may also need to talk a bit about the balance with entropy

Transition to results:

Recently, the Degrado lab successfully re-designed a 5-helix bundle protein using solely tight sidechain packing. However, the extent at which sidechain packing is involved in stabilizing and driving membrane protein folding in a variety of systems is still an open question for debate. Our research uses a high throughput approach to assess the extent to which sidechain packing is a driving force in membrane protein association. Pairing computational homodimer design with a high throughput dimerization assay, our research suggests that sidechain packing plays a crucial role, although likely a smaller one than hydrogen bonding in the association and folding of membrane proteins.

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 density 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.

Previous research on GASright proteins suggests that using a simple set of energetic functions, we are able to capture the biophysical trend of association (Anderson et al., 2017). Additional studies have shown that this association *in vivo* can be correlated to thermodynamic stability *in vitro* (Díaz Vázquez et al., 2023; Fleming et al., 1997). By pairing computational design with our high throughput *in vivo* dimerization assay, we aim to to determine the contribution of sidechain packing to thermodynamic stability.

To 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). To assess packing in natural structures, we first extracted fixed backbones from membrane protein structures to be used in our design algorithm. We first searched through x-ray crystallography and cryo-EM solved structures of proteins found in the Orientations of Proteins in Membranes (OPM) database in September 2019. Each of these structures was input into the molecular software library (MSL) and any two helices in close contact were extracted as a helical pair (**define?**) (Kulp et al., 2012). The geometric parameters for each of these helical pairs was extracted and then against the overall density of these pairs in geometric space between two parameters: distance and crossing angle (**Figure MP\_design A**). This resulted in three high density regions in dark blue (**define**) that we expected to be most amenable for protein design.

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.

Previous research has shown that mutating positions along a polyleucine transmembrane helix influences its ability to dimerize (Anderson et al., 2017; Zhou et al., 2001). This has been implemented successfully in the past to determine the dimerization propensity of GASright proteins (Anderson et al., 2017), but not for the left and right design regions. 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. For the right handed design region, we used the same set of interfacial positions as GASrights for sequence design. For the left handed design region, we chose three different patterns for the interface to accommodate for potential knobs-into-holes packing and leucine zippers typically found in left handed coiled coils (cite). In each of these design regions, we randomly generated … geometries as inputs for our design algorithm (Methods). 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 GAS sequences in the other regions. Each designed sequence is evaluated using a combination of van der Waals packing, hydrogen bonding, and implicit solvation(Krivov et al., 2009; Lazaridis, 2003; MacKerell et al., 1998)*.* These designed sequences then undergo a local backbone minimization to find the most stable structure using our energy terms. Our protein design algorithm resulted in a set of x sequences with a range of both energy and stability.

**A screenshot of a computer

Description automatically generatedFigure Experimental\_methodology**

In order to assess the dimerization propensity of our designs, we used high throughput sort-seq (**cite sort-seq**). This assay utilizes a combination of fluorescent activated cell sorting (FACS) and next generation sequencing (NGS) to determine the fluorescence profile of a library of 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**). As an additional way to determine the success of our designed sequences, 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**). 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)**.

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**). 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. However, many proteins in both the left- and right-handed designs are above the fluorescence of our monomeric control, meaning that we can successfully design sequences that associate in those regions as well. The fluorescence distribution of WTs is compared to the fluorescence of both mutants, with significance level above the respective boxplot. Clashing mutations significantly decrease the association of their respective WTs, as shown by their respective p-values < 0.005. 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. 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.

A screenshot of a computer screen

Description automatically generated**Figure clash dataset**

Figure 3. 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 of 35% GpA or 50% less than the WT % GpA. 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. Additionally, our design energy score expected many of the left-handed structures to be more stable than right, and many sequences with successful clashing mutations are below our threshold for monomer suggesting that these are monomeric as well.

Finally, we converted the fluorescence to deltaG using a method from an in vitro study on GASright proteins (Díaz Vázquez et al., 2023). We then compared these deltaG values to the energetics from our design algorithm (**figure deltaG data**). Although our energetics does not seem to capture the overall trend of stability, it allows us to visualize the distribution stability of these sequences against a couple of different terms. This data shows that both left- and right-handed structures, which are specifically designed using solely packing, are stabilized to a deltaG of around -3.5 to -4 while many of the most stable GASrights are around -4 to -5. This suggests that packing contributes to at least 60% of the stability of these structures.

**Delta G data**

A screenshot of a graph

Description automatically generated

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

**A screenshot of a computer

Description automatically generatedtop 5**

Figure: 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. Additionally, since many void mutations results in increased stability, it is possible that those sequences also have increased packing compared to their respective WT sequences.

**Discussion and Conclusion**

**Other things that can be included:**

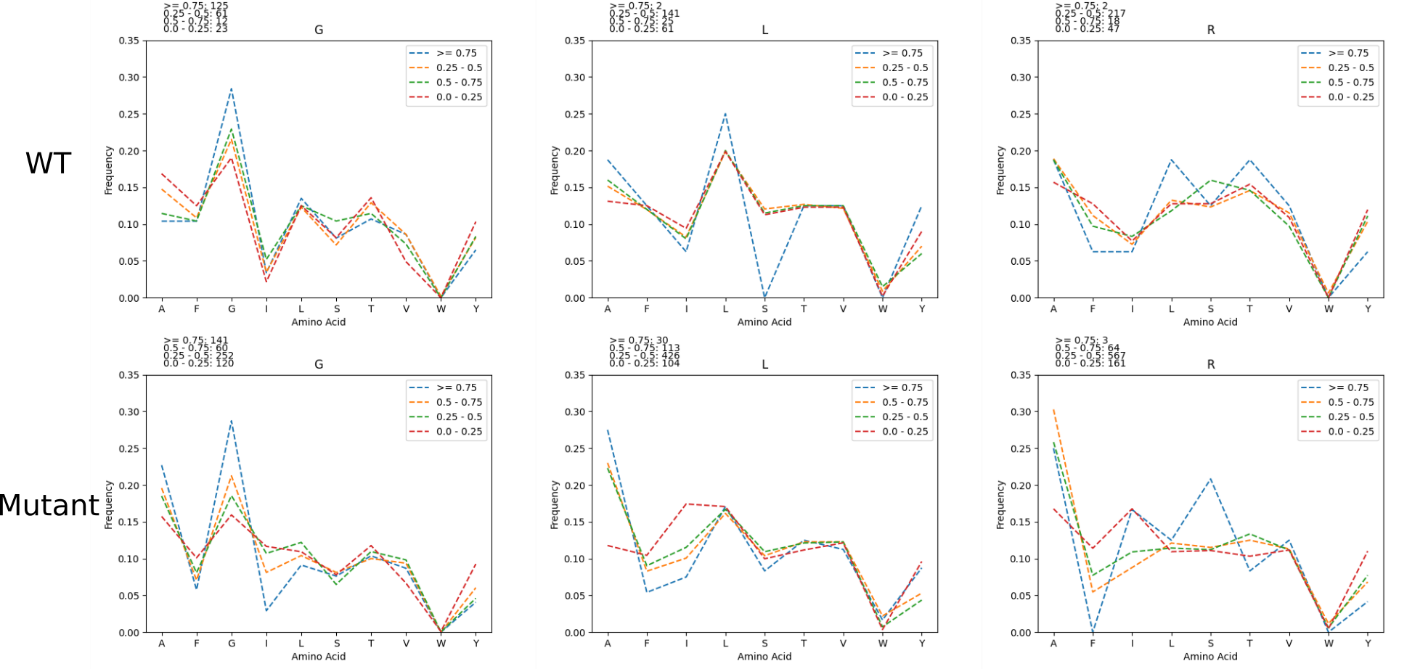
**Amino acid distribution per Fluorescence**

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.

Previous research has shown that mutating 8 designated positions along a polyleucine transmembrane helix influences it’s ability to self associate (cite SMA and first polyleu paper). For both GASright and right handed, we used the same 8 positions as the interface to design sequences. For the left handed design region, we chose three different patterns for the interface to accommodate for potential knobs-into-holes packing and leucine zippers typically found in left handed coiled coils (cite).

**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**

Figure MP\_design notes:

* Maybe add an oval around each of the design sections corresponding to the colors that are later in the figures
* See if my sequences correspond to sequences found in nature (outside of the polyleu backbone)
* I think I can actually add additional detail to this design process? How can I make it understandable and concise

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>

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>

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>

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>

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>

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>

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>

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>

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>