**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**

- 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, we have begun to tease out the contribution of packing in membrane protein stability.

Results

**Figure MP\_design**

A diagram of a graph

Description automatically generated with medium confidence

Figure MP\_design: Membrane protein dimer design. (a) Helices within close contact (define how) were extracted from all membrane protein dimers submitted to the PDB in (date; 2019). Orientations of Proteins in membranes (OPM, cite) was used to only choose the proteins found in membranes. These geometric terms are plotted and overlaid over the density of these in space. (b) Design algorithm schematic used to design structures for this study. (c) Geometries and energy score of the designed sequences.

**Background paragraph on GASright protein research? Might just be in the intro:**

Previous research on GASright proteins suggests that using a simple set of energetic functions (van der Waals, hydrogen bonding, and implicit solvation), we are able to capture the trend of association. Additional studies have shown that this association *in vivo* can be correlated to thermodynamic stability *in vitro* (cite Gladys), suggesting that utilizing a high throughput method will allow us to assess the potential thermodynamics stability of a large number of membrane protein dimers.

To design our dimers of interest, we utilized a combination of fixed backbone design with iterative backbone refinement (**cite MSL and other fixed backbone design papers, and Rosetta backbone refinement**). In order to have fixed backbones for our design algorithm, we first searched through all membrane proteins found in the protein databank (**not NMR**) and searched for any two helices in close contact (**define?**). Using the molecular software library, we extracted the geometric parameters for each of these helical pairs and plotted them against their density (Figure MP\_design A). We found three high density regions (**define, what is the density in those regions**) that we expected to be amenable for protein design: (**add in the box limits that we used for each region here**). The GASright region is a well characterized dimerization motif with a known sequence signature where small amino acids glycine, alanine, and serine are typically found at the interface. This allows GASright proteins to be stabilized by a combination of tight van der Waals packing and interhelical hydrogen bonding. By designing sequences in this region, we effectively have a way to compare the extent at which packing influences association in the absence of hydrogen bonding in both the left and right regions.

Previous research has shown that mutating positions along a polyleucine transmembrane helix influences its ability to dimerize (cite first polyleu paper). This has been implemented successfully in the past to determine the dimerization propensity of GASright proteins (cite SMA), but not for the left and right design regions. For right handed, we used the same set of interfacial positions 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). In each of these design regions, we randomly generated x 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, with Glycine only being used in the GASright sequences to prevent from designing GAS sequences in the other regions **(cite MP aas paper from SMA)**. *Each designed sequence is evaluated using a combination of van der Waals packing, hydrogen bonding, and implicit solvation (****cite energetics****).* 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.

* + We aimed to design protein structures with a range of stabilities influenced specifically by van der Waals packing, allowing us to determine the extent at which packing influences association in a variety of membrane protein systems

Figure 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

**Figure 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 exp\_method). The mutated structure undergoes local backbone minimization we chose two of the least stable (clash) or largest differences in solvent accessible surface area (void) mutants 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).

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. 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 right and left handed designs are above the fluorescence of our monomeric control, meaning that we are able to successfully design sequences that associate in those regions as well. The fluorescence distribution of WTs from all three design regions is significantly different from their respective clashing mutants, suggesting that the clashing mutations significantly decrease the association of their respective WTs. This allows us to view 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.

**Figure clash dataset**

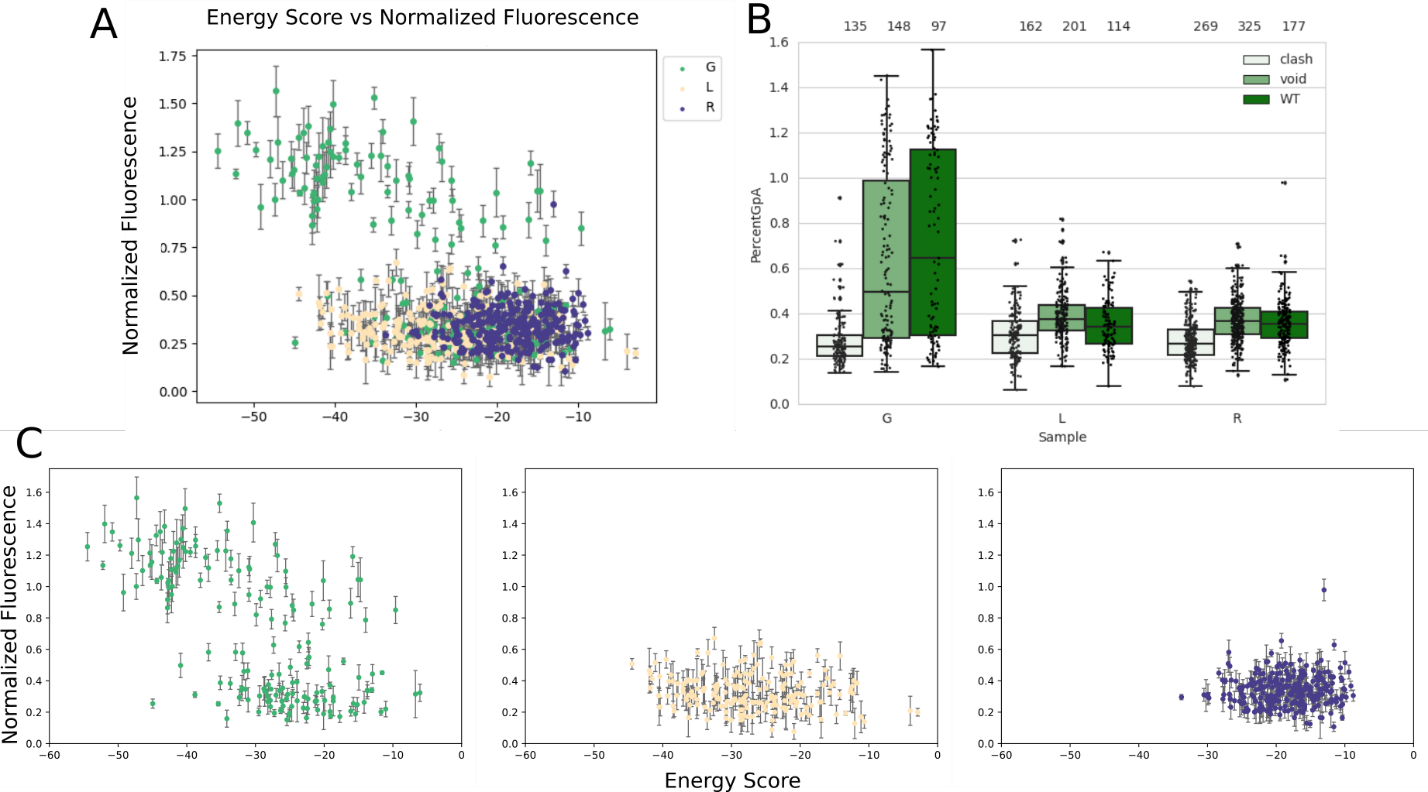


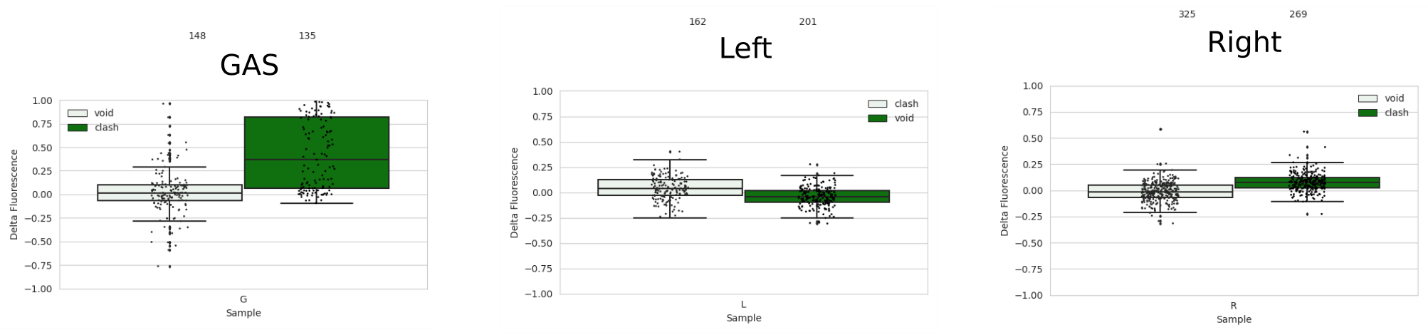
Figure 3. A) The normalized fluorescence plotted against the design energy score for all of the designed sequences in the dataset. The plot is separated by the geometric region of design with GASright in green, Left in yellow, and right in purple. B) The WT (dark green) fluorescence distribution is compared 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. 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. To evaluate these sequences further, we decided to look at the top 5 structures from each region (new data here: monomer SASA/dimer SASA, things greater than 35%, sequence analysis, comparison to TMDOCK, etc.)

* + Separated these into three graphs that compare the fluorescence to energy score.
  + However, there are a variety of sequences above the 40% GpA in both the right and left handed regions (conservative limit for association)
    - Data figure for these: What could I do to summarize that data for those sequences? Maybe just boxplots of those sequences?
    - 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)
  + This data is not yet maltose filtered, but when I get the maltose filtering back I can remake these figures
    - Could also do some sort of clustering analysis of these points to see if their geometries are similar, and then if their sequences are similar?

**Delta Fluorescence Figure**



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

Structures for the top 5 designs for each region

A screenshot of a computer

Description automatically generated

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

* Overall they look well packed, with some slight voids in the left and right regions
* Add in the energies for each of these, and the SASA (compare that to the VDW per square angstrom (vdw/SASA))

**Amino acid distribution per Fluorescence**

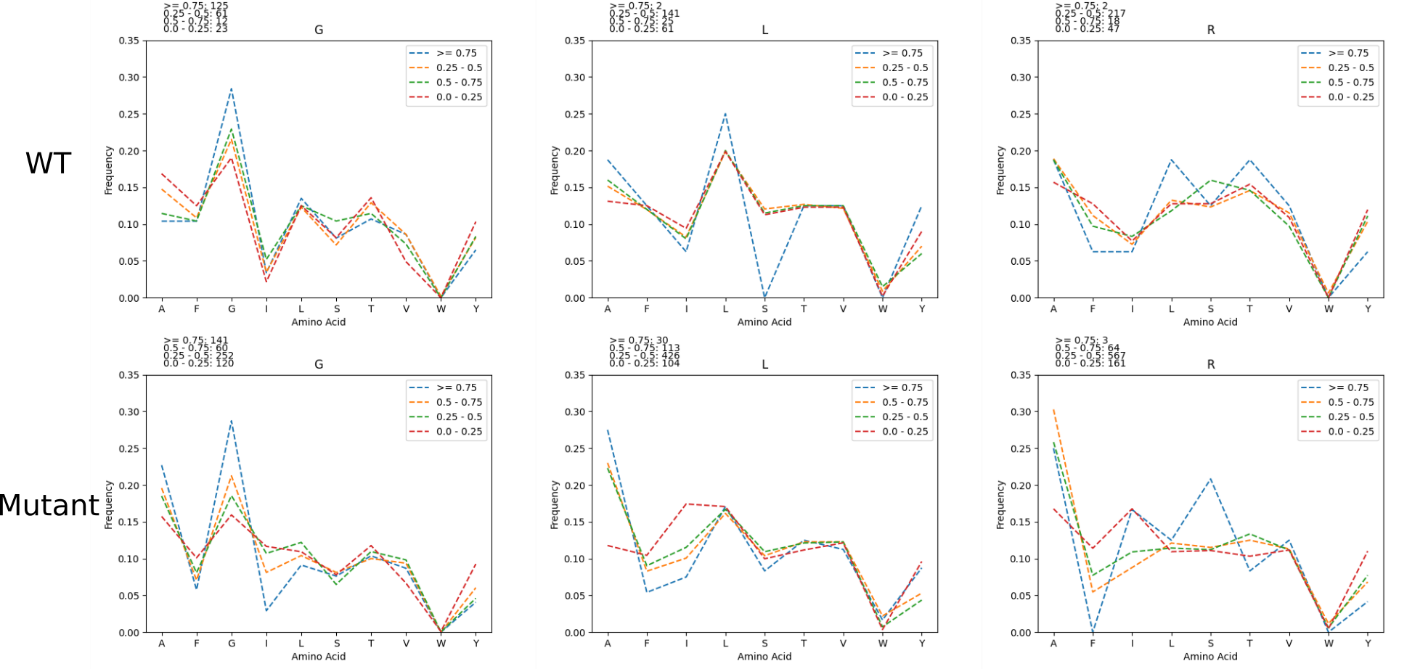


Figure: The amino acid frequency was calculated for sequences based on their fluorescence (>= 0.75, 0.5 – 0.75, 0.25-0.5, and < 0.25). This was done for designs and all mutant sequences.

* 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

**Delta G figure**

A graph of a graph with numbers and dots

Description automatically generated with medium confidence

Figure 4. Converting fluorescence to free energy.

* Detail here for how we convert from fluorescence to delta G
  + Cite Gladys paper
* Talk about how we trimmed to this dataset:
  + This image is for:
    - At least 1 clashing mutant < 35% OR the difference between the WT and mutant is > 50%
  + Give more information about clashing mutations here (probably also earlier in the design section)
* Talk about the impact of van der Waals packing in left and right versus GASright

SI Figures

**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).