**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

**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.

* Using the molecular software library, we have been able to predict the structures of a variety of membrane protein dimers (cite Ben and Samantha)
  + Some of these proteins have shown agreement with our energetics, suggesting that we can explore thermodynamics for membrane protein dimers using our current method (Gladys paper)
  + 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.
  + Explore the geometric conformational space of membrane protein dimers by extracting protein-protein interactions from membrane protein structures solved in the PDB (non-NMR structures, cite OPM) (figure A)
    - X proteins were extracted from orientations in membrane proteins
    - Any two helices in close contact (define) were extracted from each membrane protein structure (total number), and the helical geometric information was extracted using MSL
    - These helical geometries were used to determine a density plot of the potential designable space for membrane protein dimer geometries
* Membrane protein design using the membrane protein geometric landscape from the pdb (figure B and C):
  + Protein design using MSL (minor detail, add rest to methods)
  + Show and define the reason for using our library of Aas for each region
  + Designed x proteins with a range of expected stabilities
    - Energy score made of vdW, hbonding, implicit solvent (cite).
  + Talk about the design regions (GAS, right, left) (figure C)
    - GAS corresponds to the proteins that we’ve studied previously with our methods, stabilized by a combination of vdW and noncanonical hbonding
    - Right are proteins with similar geometries to GAS, but with decreased potential for noncanonical hbonding without glycine in the library and with a larger interhelical distance
      * Should give us a better understanding of how essential packing is to stabilizing proteins with these geometries and have a direct comparison to proteins with hydrogen bonding in the GAS
    - Left handed proteins have been found to be stabilized by knobs-into-holes packing in coiled coils and leucine zipper motifs (cite)
      * Exploring how packing interactions similar to these results in association

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**

* To further understand the impact of packing on association, we decided to take a high throughput approach:
  + With our thousands of proteins, we made mutations for each corresponding to the following:
    - Clash: mutate a well packed position to isoleucine. Take the protein structure, replace each amino acid with isoleucine at the interface, then allow the structure to optimize for the sequence. Chose two sequences for every structure with the highest vdW energy (which would be the least stable)
      * These mutants are expected to abolish/significantly decrease association if our structures are correct, supporting whether our designed structure is correct
    - Void: mutate a well packed position to alanine. Similar method to above but alanine. Chose two sequences that had the largest SASA difference between mutant and WT, suggesting the largest loss of packing (should explain why not just use the vdW energy here?
      * These mutants are expected to decrease association, however unlike previous literature (Bowie and others), these dimers may compensate for the loss of packing by finding an alternative structure. Mutants that we can use to understand the extent at which a loss of packing may affect association.
  + Complementary high througput method:
    - Sort-Seq
      * Cite, but also might be good to have a quick figure for how it works
        + Fluorescent signal is given off from a cloned library and sorted in populations. These populations can then be determined by NGS and reconstructed for each sequence.

Figure notes:

* I’m picturing a figure that has a kind of flow chart for my mutants
  + Show a few structures, pick one out, split to void and clash mutations, show the distribution of all data for each (those boxplots)
    - I think since this would be partially a figure summarizing the method and what we expect from the data, it could be a good idea to show the overall results before breaking it down into the clashing data
* Above the above, I can have a quick breakdown of how sort seq works:
  + cell with dimer → large population of cells with other dimers → sorted into a couple bins → fluorescence reconstructed -> conversion to TOXGREEN values method

**Figure clash dataset**

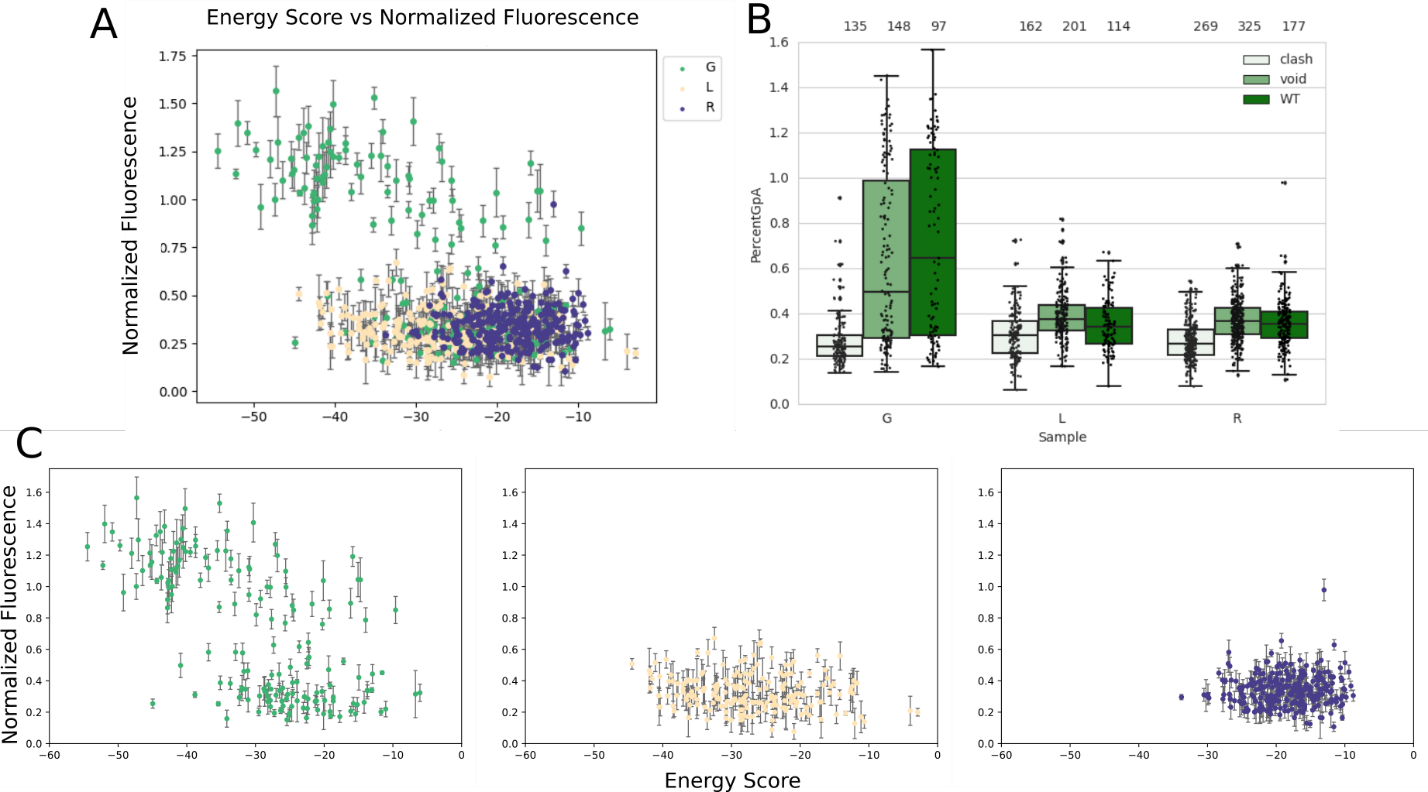
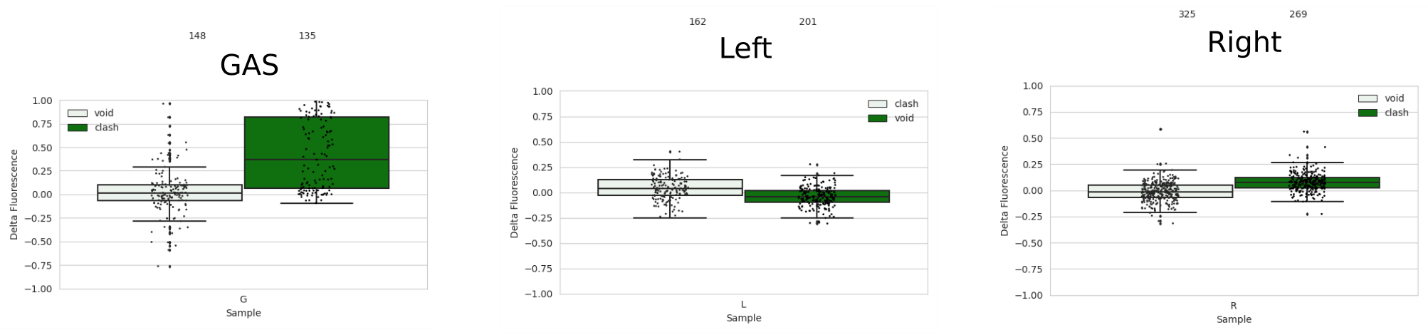


Figure 3. A) The normalized fluorescence plotted against the design energy score for all of the designed sequences in the dataset. The plot is separated by the geometric region of design with GASright in green, Left in yellow, and right in purple. B) The WT (dark green) fluorescence distribution is compared to the fluorescence distributions of their corresponding mutants. C) Separated graphs for normalized fluorescence against the design energy score in A.

* From the full dataset, we then filtered our design sequences by the clashing mutations that support our designed interfaces:
  + Keep any wt sequence with at least 1 of the clashing mutants below 35% GpA OR 50% lower than the wt fluorescence
    - Reasoning: some sequences likely don’t have both clashing mutants, and this allows us to maximize the amount of data
  + Separated these into three graphs that compare the fluorescence to energy score.
  + Overall, GASrights fluoresce more highly than the other regions, likely due to their potential for stability through interhelical hydrogen bonding.
  + 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**

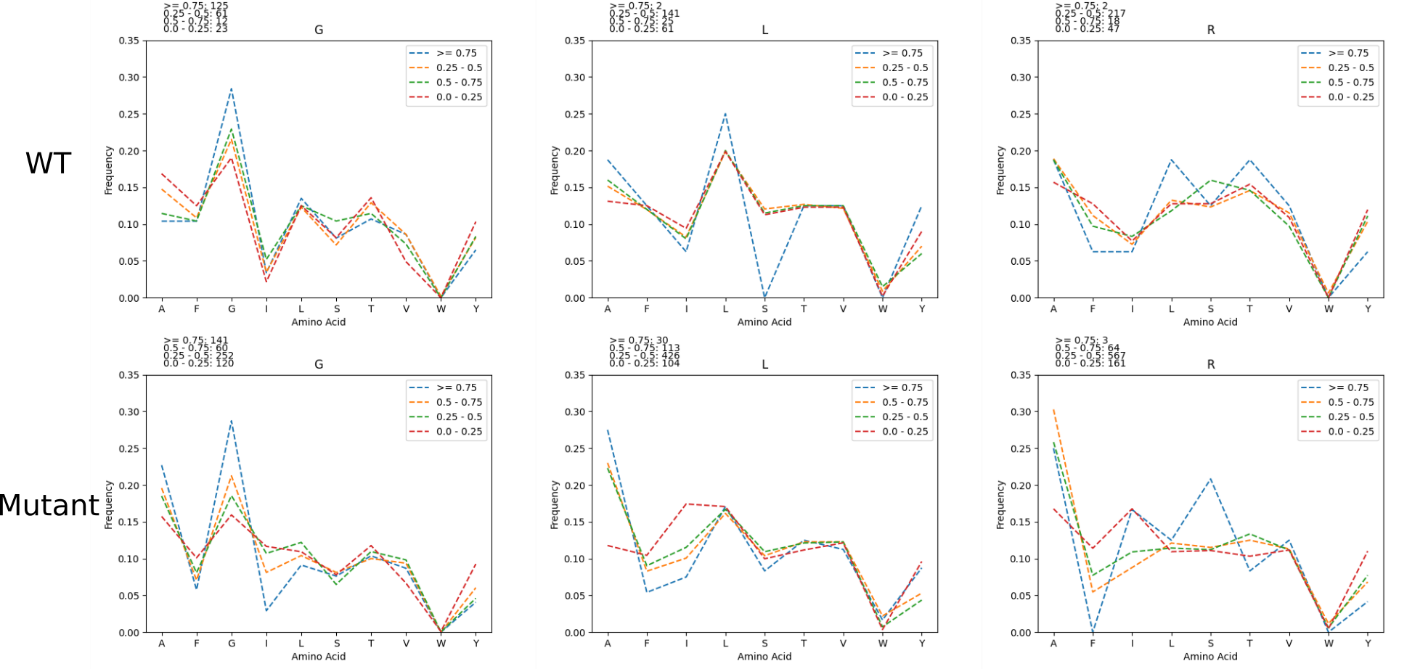


Figure: The amino acid frequency was calculated for sequences based on their fluorescence (>= 0.75, 0.5 – 0.75, 0.25-0.5, and < 0.25). This was done for designs and all mutant sequences.

* 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.