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

Gilbert Loiseau

Advisor: Alessandro Senes

Committee Meeting

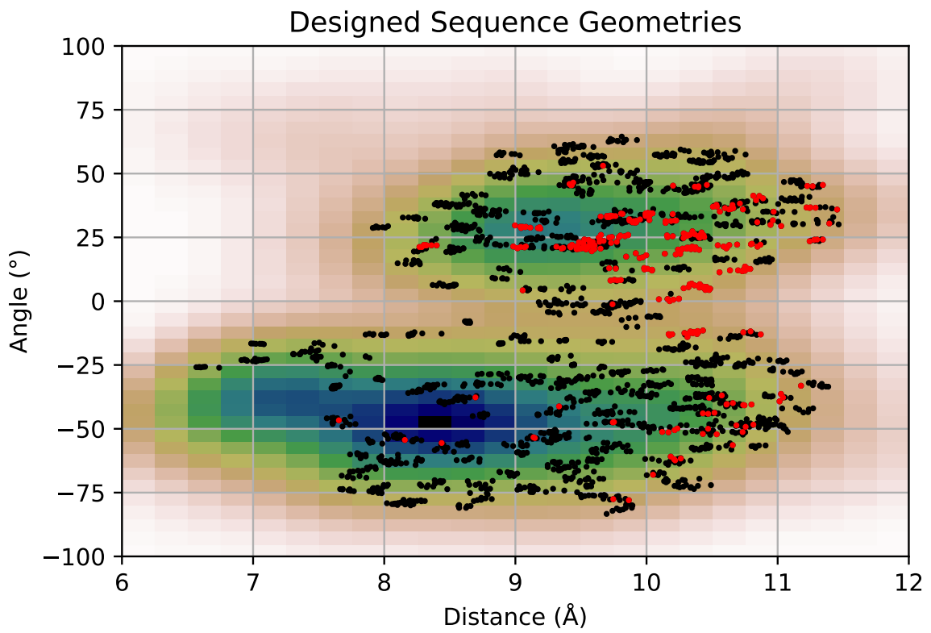
June 3rd, 2021

1:00 PM

**Introduction**

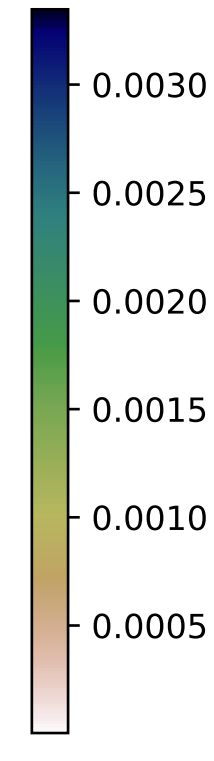
Proper membrane protein folding is necessary for essential biological functions such as cell signaling and gene regulation. Misfolding of membrane proteins often leads to disease phenotypes including growth defects and cancer. A variety of forces contribute to proper membrane protein folding including 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: Computational design of homodimeric sequences**



|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Sequence** | LLLLL**AA**L**WS**LL**F**LL**IV**LLLL | LLLL**FS**LL**A**LL**IA**LL**S**LLLLL |
| **vdW** | -30.9 | -13.8 |
| **Hbond** | 0.6 | 0.8 |
| **Solvation** | 8.2 | 4.3 |
| **Energy Score** | -22.1 | -8.7 |

**Fig. 1: Preliminary Sequence Design Data. A) Kernel Density Estimator overlayed with designed sequence geometries.** Kernel Density Estimation was used to identify dense areas of the geometric space by smoothing the data to give a more accurate representation of the density of the geometric space. Overlayed over the geometric density data are points from sequence design runs to explore which parts of the space are amenable to homodimer design. In **black** are all geometries from all designed sequences and in **red** are geometries from sequences that were identified stable as dimers. Areas around the points in red will be further explored to design more stable sequences for experiments. **B) Comparison of well-packed vs poorly packed designs.** Well-packed design on the left vs poorly-packed design on the right. Some designs are found to have poor packing like the design on the right, despite the negative design energy score.



High Density

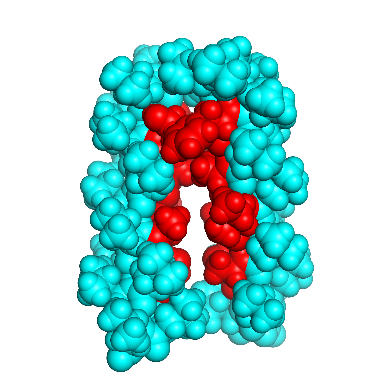
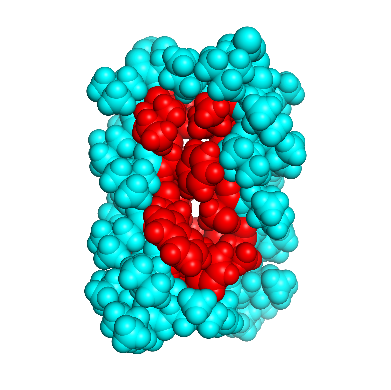
Low Density

**A**

**B**

**Well-Packed**

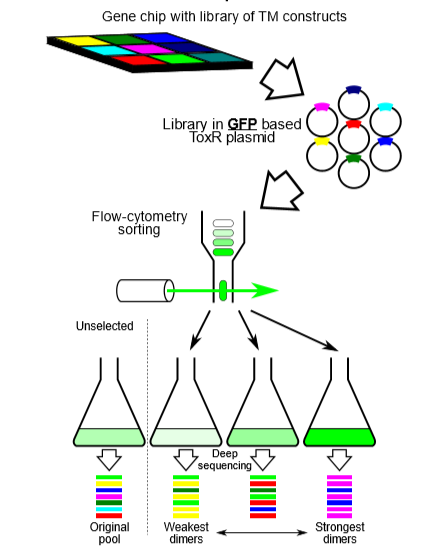
**Poorly-Packed**



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 (Figure 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 1A). The overall workflow of my computational design algorithm is found in the supplement (Figure S2). In particular, there is a high density of points where I was able to design sequences with negative energy scores with a crossing angle around 25 degrees and a distance of 9-10 angstroms. Using this subset of geometries, I am currently designing more sequences, aiming to distinguish hundreds of sequences that rely on stabilization by solely packing.

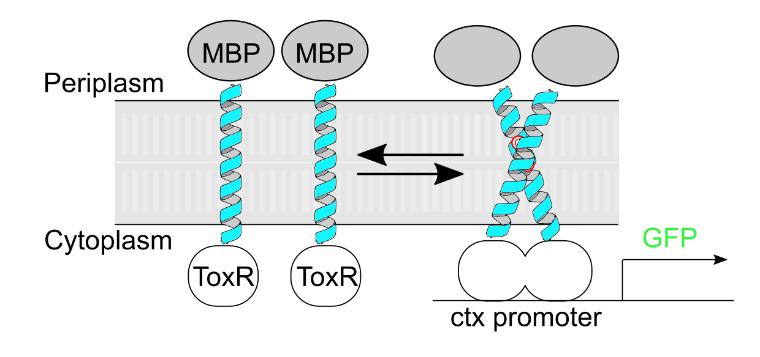
Although I have already identified a subset of well-packed structures, many structures that my design algorithm has determined as a successful design are not as well-packed as expected (Figure 1B). I am planning to implement a way to determine how well a design is packed so that I do not have to manually inspect each structure. One option to determine if a structure is well-packed is to calculate the occluded surface area of the structure: if a water molecule can fit between two atoms, then the area between the atoms is not buried, and likely not well-packed (Fleming and Richards, 2000). Another option is to determine the SASA difference between monomer and dimer, where a well-packed structure would bury much of the SASA of residues at the interface. I should then be able to identify a trend for the SASA of well-packed structures vs poorly packed structures, allowing me to continue forward into experiments with structures with well-packed interfaces.

**Future Work: *In vivo* characterization of designed sequences using TOXGREEN and sort-seq**



**Fig. 2: 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.



A

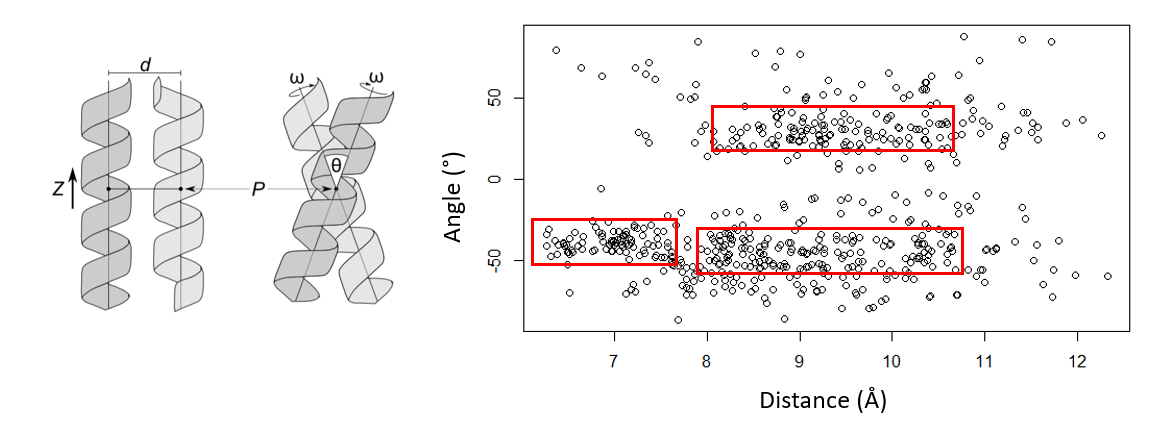
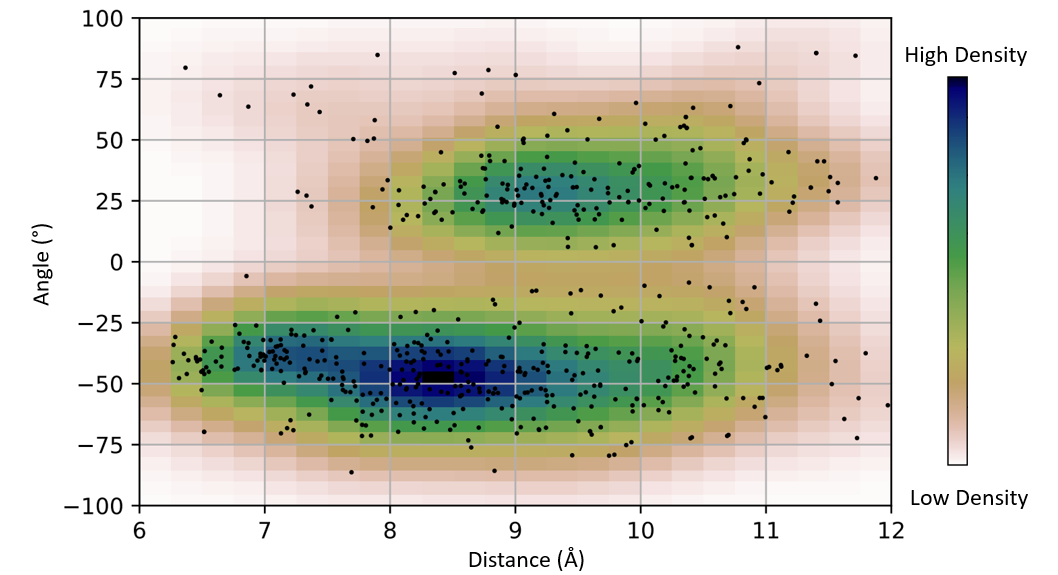
B

To evaluate successfully designed sequences, I plan to use TOXGREEN, an *in vivo* dimerization assay that quantifies dimerization propensity through the output of sfGFP (Fig. 2A). Prior to testing a large subset of sequences, I am going to test a subset of designed sequences using TOXGREEN, an in vivo dimerization assay that assesses dimerization propensity. This subset of sequences will consist of sequences that my design algorithm considers to be the most likely to dimerize, and another groups of sequences that are not likely to dimerize. I will clone these sequences into TOXGREEN plasmids and then based on the expression of sfGFP, I will determine the amount of fluorescence for each of my sequences. By testing this subset of sequences, I will determine the fluorescence threshold of my best sequences, allowing me to better estimate the spread of the fluorescence distribution from the designs. This will allow me to compare the potential fluorescence distribution for my data to other dimers that have been shown to dimerize with TOXGREEN. This first trial of sequences will evaluate how well my algorithm designs sequences and assess the extent at which solely van der Waals packing can stabilize membrane protein dimers.

To further determine the extent at which van der Waals packing stabilizes membrane protein dimers, the designed population of sequences will be evaluated for their experimental stability using sort-seq, a high-throughput method which combines fluorescence activated cell sorting (FACS) with next-generation sequencing (NGS) to evaluate dimerization propensity of TM domains (Fig. 2B). An oligo pool library consisting of my designed TM sequences will be cloned into TOXGREEN plasmids and the reporter gene sfGFP will quantify dimerization (Armstrong and Senes, 2016). Based on the expression of sfGFP, cells are sorted into different bins with different fluorescence thresholds. These plasmids can then be 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, will be reconstructed for each dimer. Saturating point mutations of each sequence to various hydrophobic amino acids will also be characterized on sort-seq, and a sensitivity to each mutation will be generated based on change in dimerization propensity. This mutagenesis and sensitivity profile will be used to confirm that the designed sequences associate via the predicted structure (Anderson (Thesis), 2019). Using this technique, I will be able to determine if there is a correlation between strength of sidechain packing and membrane protein association in hundreds of designed sequences at a time.

Finally, it is possible that our results show little correlation between our designed energies and dimerization propensity. We expect that small changes in sequence should have little effect on the monomer energy of our TMs and that these changes would average out with a large number of sequences for analysis, but this would suggest that this is not always the case. This could be the result of changes in expression level due to differences in sequence composition, a possible variable that cannot be measured using sort-seq but that can be checked on Western blots. One approach to address this issue is to maintain the sequence composition of a group of designed sequences, aiming to control sequence expression. I can shuffle amino acids at the interface of a designed sequence and run these through a helix-helix docking algorithm. This docking algorithm will give me a subset of sequences from my shuffles that can be tested for correlation between dimerization and designed energy.

**Supplementary Figures**



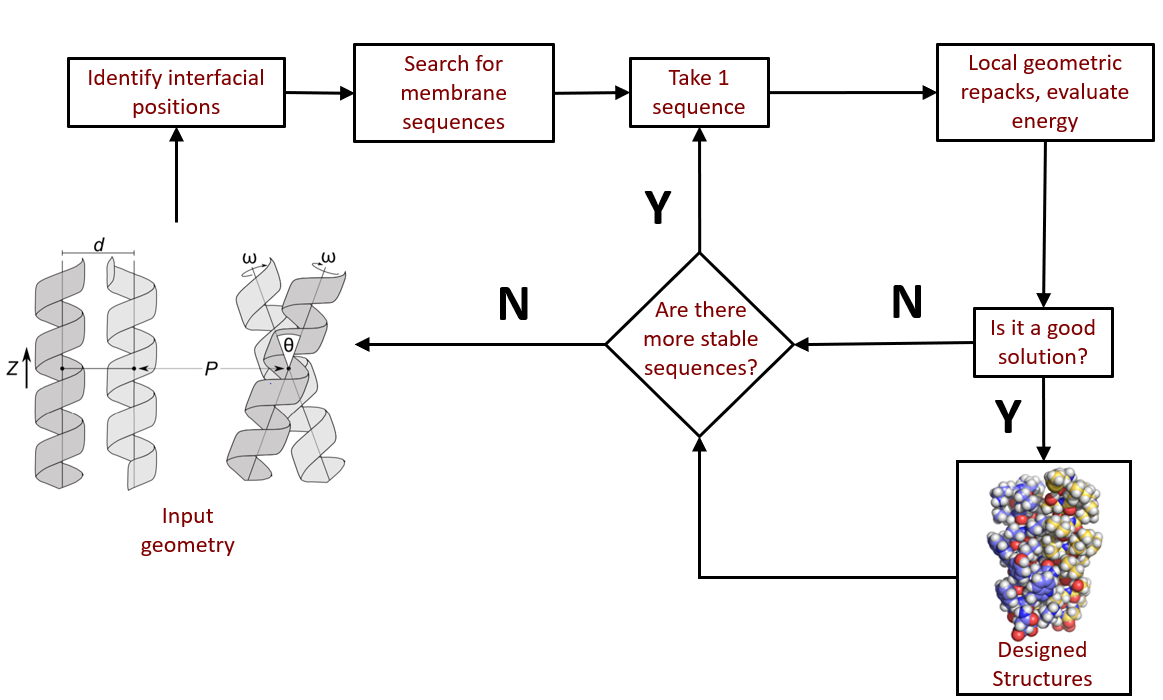
**Fig. S1: Analysis of nonredundant membrane proteins determines common geometries of helix-helix association.**

**A)** The geometric landscape of homodimerization is defined by four parameters: interhelical distance (d), crossing angle (*θ*), rotation of the helix around its axis (*ω*), and the vertical position (*Z*) relative to the crossing point between the helices (P). **B)** Each circle in this scatterplot represents one helical pair of the 835 pairs extracted from nonredundant membrane proteins found within the PDB. These pairs were plotted by their angle and distance, allowing me to determine areas with common geometries that may be amenable for sequence design. Kernel Density Estimation was used to identify dense areas of the geometric space by smoothing the data to give a more accurate representation of the density of the geometric space.

**A**

**B**

**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.



**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 GASright motif: sequence, structure, and stability. Thesis. University of Wisconsin, Madison, WI.

In addition, if I decide that confirming the structure of some of my constructs is important to determine the accuracy of my designs, I will have a large population of proteins that I could test for expression, purification, and crystallization ability, which may make it possible to solve the structure of some of my designs. Another caveat is that I am assuming that if a geometry is more prevalent in the PDB analysis, then it may be more amenable to design (Zhou and Grigoryan, 2015).

I may have to use more sensitive methods to confirm differences between the energy of the monomeric and dimeric states of my designs such as molecular dynamics simulations, giving me a better understanding of why particular sequences do not follow a trend. These sequences can then be removed from the analysis, allowing me to determine if a trend between sidechain packing and dimerization exists with sequences where the changes in the monomer energy have little effect on dimerization.

Talk about needing to calculate occluded surface area or SASA for structures and determining if they are less than the monomer (compare to our data on GxxxG.

Elaborate on the below: and talk first about how I will use a gblock with TXOGREEN (maybe pull up the TOXGREEN figure too)

* Add in some specific mutations (if I have a sequence already for example, I can point out positions that might be mutable because of results in my design runs (sequences that are positive: figure out how to say this within my writing about my sequence design process, and then explained here (maybe an image of the actual process?)

I don’t think I need to harp on the caveat here too much, but say the things I’m going to be doing to test: making mutants, how I can test some good sequences, then a block of poor sequences and see how poorly they associate to controls (G83I: figure out how much fluorescence)

Protein design papers typically result in the structural determination of each of their designs, confirming the expected orientation of the design. However, my research aims to design hundreds of structures, making it difficult to confirm the structure of each design. To address this lack of structural data, I will use sort-seq to test point mutations at each position on my homodimers. If changes in dimerization are particularly sensitive at the interface, then I will have supporting information for the structures of my designs. To determine if this is the case, I will also design sequences for geometries that are not as prevalent within the analysis. If these designed sequences are overall less energetically stable than our other designs and have lower dimerization propensity, then it will support our assumption that common geometries are more favorable for design.