Chapter 4: Future Directions

Table of Contents

[4.1 Summary of Dissertation 2](#_Toc174278203)

[4.2 Hydrogen bond mutations 4](#_Toc174278204)

[4.3 Studying the impact of sidechain packing with other forces 6](#_Toc174278205)

[4.4 Improving the protein design algorithm 9](#_Toc174278206)

[4.4.1 Heterodimer design 9](#_Toc174278207)

[4.4.2 Turning sequence entropy into a pairwise term 13](#_Toc174278208)

[4.4.3 Training energy terms 15](#_Toc174278209)

[4.5 Detecting protein concentration in high-throughput 16](#_Toc174278210)

[4.6 Supplementary Details 18](#_Toc174278211)

[4.7 References 19](#_Toc174278212)

4.1 Summary of Dissertation

In the previous chapters, I detailed the research in the Senes lab that has focused on understanding driving forces in MP folding. My research focused on improving our understanding of the impact that van der Waals packing has on facilitating MP folding and association. We studied a subset of van der Waals packing known as sidechain packing and found that it is a weak driving force when it is the sole force used to design the association between TMHs.

In my recently submitted 2024 paper (Chapter 2), I studied the effect of sidechain packing on thousands of MP dimers. Through computational design and mutagenesis of key interacting residues, we found that proteins designed to associate solely with sidechain packing dimerized less than proteins designed with both sidechain packing and interhelical hydrogen bonding (GASright). Additionally, our energetics correlated much better to the GASright region, suggesting that our computational model is not well-tuned to predict the association of weak MP dimers. In Chapter 3, I detailed many of the computational methods that I developed to design dimers and the rationale for decisions made during the design procedure. Explaining these methods in detail allows for students in our lab and others to utilize and/or co-opt my methods for future research.

My research has shown that sidechain packing is an essential force for MP folding and association, despite being a weak driving force. In this chapter, I expand on potential avenues for future research where this knowledge can be utilized. I first discuss additional experiments that were not included in my publication, where I studied the effects of mutating out all potential hydrogen bonding residues used in the design process on a subset of designs. These results confirm that we designed multiple dimers that associate by sidechain packing in the absence of any potential hydrogen bonding. I then suggest future experiments to enhance our understanding of sidechain packing in the presence of other forces. I detail my ideas for potential improvements to my protein design algorithm, including the potential for designing heterodimer sequences, converting sequence entropy into a pairwise term, and improving our energetics using machine learning. Finally, I suggest improvements for our high-throughput sort-seq method, with attempts to further characterize sequences by their expression level in cells.

4.2 Hydrogen bond mutations

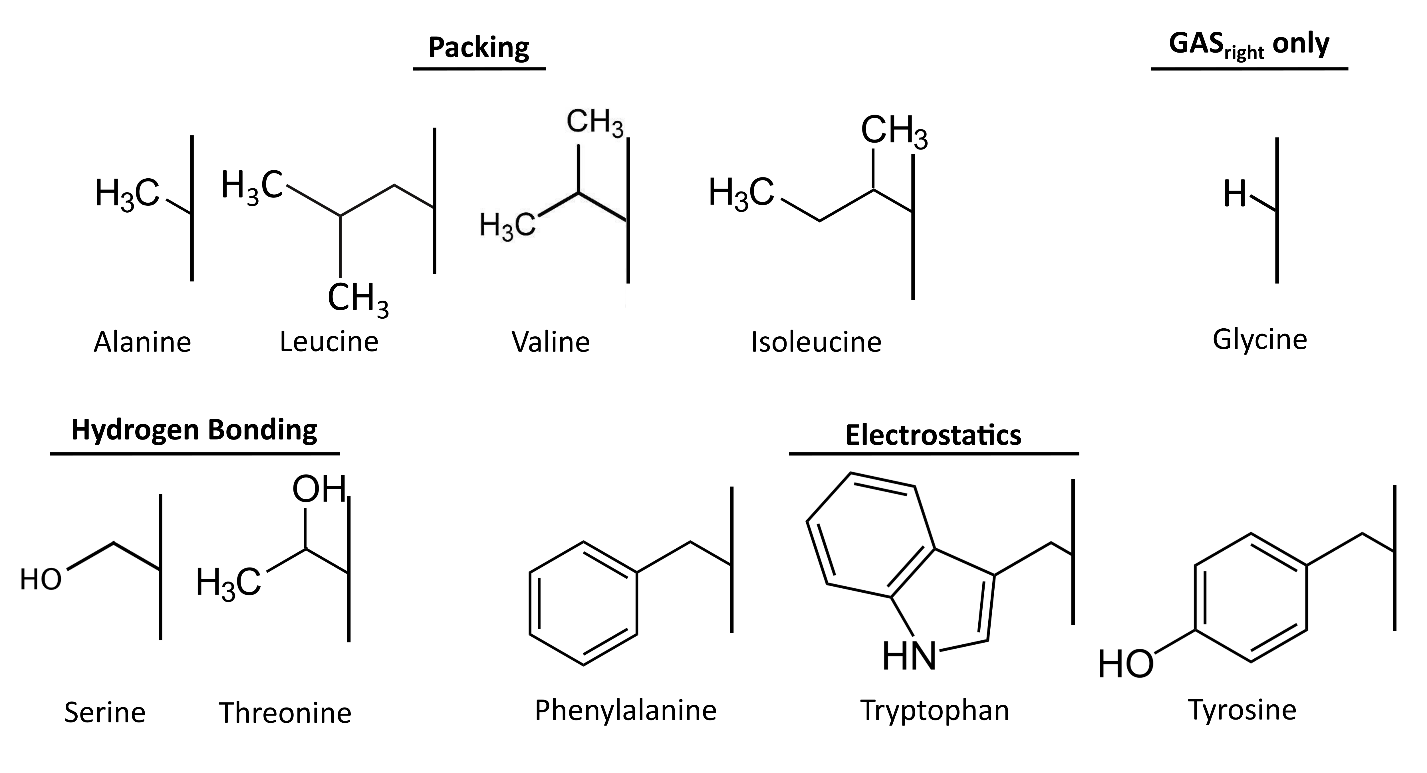
In our design procedure, we included AAs that had the potential to hydrogen bond due to how frequently they are found in membrane proteins (Figure S3.4). Our design energies predicted low levels of hydrogen bonding, suggesting that these AAs would not result in hydrogen bonding impacting association. However, because our energetics show little correlation with experimental dimerization propensity in the Left and Right design regions, it is difficult to confidently state that our sequences associate solely by van der Waals packing without the influence of other forces like hydrogen bonding.

To ascertain if our proteins associate solely by packing, I conducted an experiment where I mutated our designed proteins to remove the potential for hydrogen bonding. To identify proteins with the potential for hydrogen bonding, I wrote a Python script that searches through my protein structures and identifies any oxygen atoms within 3Å, a generous threshold for potential hydrogen bond formation. I identified 17 proteins that associated from mildly weak to strong dimers (>40% GpA) with the potential for at least 1 hydrogen bond. I mutated all hydrogen bonding AAs in these sequences to hydrophobic AAs with similar steric bulk: Thr→Val, Tyr→Phe, and Ser→Ala. I ordered the original design sequences, their respective clash mutants, and the hydrophobic mutants as gblocks from Twist Bioscience and successfully cloned 13/17 proteins into the TOXGREEN plasmid for experiments. 6 biological replicates were run in TOXGREEN and the data assessed for consistency with our sort-seq results, aiming to determine if our sequences associate according to the designed interface and if hydrophobic mutants associate similarly to the wild type design.

**A close-up of a graph

Description automatically generated**8/13 designs were found to properly associate according to our designed interface, where either the clash (C) or hydrophobic (HP) mutations resulted in a significant decrease in association when compared to the wild type design (Figure 4.1). 4/8 of the HP mutations displayed similar or increased association as compared to the design. Although we are unable to predict association using our energy terms, this result indicates that we have successfully designed proteins that associate without hydrogen bonding. This data suggests that although a weak force, van der Waals packing drives the association of a variety of our designed membrane proteins.

**Figure 4.1 TOXGREEN on hydrogen bonding AAs.** TOXGREEN fluorescence of 6 biological replicates for each design (D), clash (C), and hydrophobic (HP) mutations. **A)** Table of sequences for each design with the interface highlighted in bold. Clash mutant highlighted in red. Hydrophobic mutates all T, Y, and S on design sequence to V, F, and A, respectively. T-tests were conducted between sequences and mutants, and at least one of the C or HP mutants significantly decreases fluorescence. **B)** Bar graph of each design (blue) next to their respective clash (red) and HP (yyellow) mutants.

4.3 Studying the impact of sidechain packing with other forces

**Figure 4.2 Chemical Structures of Design AAs.** Design AAs separated by structures and potential energetic contributions to MP folding.

To investigate the impact of sidechain packing on MP folding and association, I designed thousands of proteins using only a subset of AAs (Figure 4.2). However, this subset of AAs includes two AAs (Ser and Thr) that can form hydrogen bonds (Russ & Engelman, 2000) and three aromatic AAs (Trp, Tyr, and Phe) that can facilitate electrostatic π-π stacking interactions (Johnson et al., 2007). Our current dataset has the potential to facilitate these interactions outside of solely sidechain packing, however, our energetics suggest that they associate primarily through packing. Mutational testing on our sequences suggests that hydrogen bonding is unlikely to play a role in all our sequences (section 4.2), and visual inspection of structures does not suggest that electrostatics interactions are involved in association.

We included these AAs because of their relative abundance in MP sequences (Figure S3.4). Ser and Thr typically form hydrogen bonds with the backbone carbonyl oxygens in monomeric helices (Gray & Matthews, 1984), so we expected the addition of these AAs not to minimal impact on folding through additional hydrogen bonding. Larger AAs (Trp, Phe, and Tyr) have an extensive van der Waals radius due to their size and steric bulk, therefore we wanted to include them in our experiments. Excluding these hydrogen bonding and aromatic AAs would restrict our design pool to four AAs (Ala, Leu, Val, and Ile), restricting us to a small number of potential designable sequences. Additionally, sequences designed with a small pool of AAs would not be representative of MP-like sequences. With our current methods and pool of AAs, a follow-up study might consider addressing the impact of other forces alongside sidechain packing.

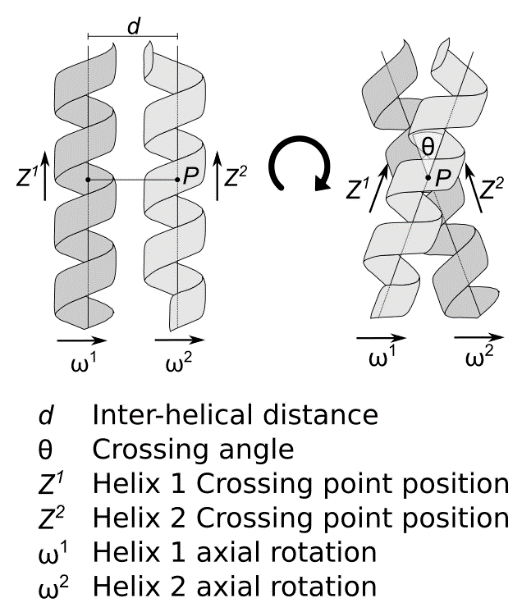
To study the impact of sidechain packing alongside other forces, we can rerun our design algorithm and change how we pick sequences for experiments. I chose sequences based on the calculated van der Waals energy, but a follow up study could instead choose sequences with a considerable amount of hydrogen bonding. We can see if increasing amounts of predicted hydrogen bonding results in higher dimerization, like GASright. Another option is to include electrostatics in our energy calculation (Patel et al., 2004; Zhu et al., 2012), which may result in designing structures amenable to forming electrostatic interactions. If the energetics are found to match dimerization propensity for structures with more hydrogen bonding or electrostatics, then it would suggest that these other forces play a more prominent role in helix-helix association than packing. This would support the data found in my initial study suggesting that packing is a weak driving force. By choosing sequences with a range of packing energies for each of these other energetic variables, we may identify a trend between packing and its impact on association alongside other forces.

In addition to studying forces other than packing in MP association, we can use an alternative set of AAs in protein design. We can continue to include the AAs that can only pack (Ala, Leu, Val, and Ile) alongside other subsets of AAs. Designing sequences with hydrogen bonding (Ser and Thr) or electrostatics (Phe, Tyr, and Trp) AAs would allow us to better isolate sidechain packing with these forces. Alternatively, charged interactions have been found to impact folding in a variety of MP systems (Ulmschneider et al., 2017). Designing sequences using a subset including the two charged AAs (Arg and Lys) to facilitate association would allow us to determine the impact of packing with different numbers of charged AAs. By designing sequences with similar geometries with these subsets of AAs, we can potentially tune the amount of hydrogen bonding, electrostatics, or charged interactions for each sequence. Similar studies have mutated single residues on helices to determine the effect of different AAs (Choma et al., 2000; Zhou et al., 2000; Zhou et al., 2001). If we find that mutations to specific residues increase or decrease association, then we can determine how different forces influence association alongside changes in packing.

We can further assess the impact of sidechain packing on association by altering the backbone sequence. In our previous research, we have primarily used a poly-Leucine backbone to assess dimerization of interface sequences. Another approach is to assess the dimerization for sequences with the same interface on poly-Leu, poly-Ala, and poly-Val backbones. Changing Leu→Val results in the loss of a single CH2, and changing Val→Ala results in a loss of 2 CH3, both reducing the size of the sidechain. Each of these AAs undergoes sidechain packing, however, the loss of steric bulk from Leu→Val→Ala suggests that poly-Val and poly-Ala backbones would exhibit less van der Waals packing. By investigating how these small changes impact association of our design interfaces, we can learn how minute changes in packing result in differences in association. Additionally, these experiments may elucidate the van der Waals packing interactions with the membrane environment. This includes a more detailed understanding of the entropic cost of these AAs on their role in MP stability. Overall, these experiments may allow us to correlate the changes in association with change in AA size, which could contribute to tuning our van der Waals energy term to differences in steric bulk. These experiments may give us insight into how to better predict the van der Waals packing between the membrane and the protein backbone.

4.4 Improving the protein design algorithm

In Chapter 3, I detailed how I developed my protein design algorithm. Although minor changes have been made throughout my research, most of the program was built in 2021. Additionally, a variety of machine learning algorithms have been developed that predict protein structures to near atomic precision. In particular, AlphaFold2 revolutionized the way that we look at the puzzle of solving protein structures. My data suggests that I have successfully designed sequences that associate along the designed interface; however, our energetic terms are unable to predict the ability for these proteins to associate. Inspired by recent literature, in this section I detail ways to develop and build off my algorithm for future research.

4.4.1 Heterodimer design

**Figure 4.3 Heterodimer Geometry.** Heterodimeric sequences do not need to have a symmetric geometry, where the z-shift and axial rotations of each helix can be different from its partner.

My research focused on assessing the stability of homodimer proteins, partially due to their simplicity to design. When designing homodimer sequences, we take advantage of symmetry: Helices are made up of the same sequence and each helix is computationally placed at the same value for each geometric term. However, important biological interactions such as regulating gene expression are carried out by heterodimeric receptor tyrosine kinases (Del Piccolo et al., 2017). Heterodimer design adds multiple variables for design, namely each helix is composed of a different sequence and can be placed at non-symmetric geometries. The two most important geometric terms for heterodimer design are the axial rotation and the z-shift. Each of these terms helps to define where the crossing point of the interface is found, with respect to the individual helix (Figure 4.3). This exponentially widens the geometric space for heterodimer association: Instead of a 1-to-1 ratio for each axial rotation versus z-shift, all rotations on one A screenshot of a graph

Description automatically generatedhelix must be assessed against all rotations on the other helix, and the same must be applied for z-shifts.

**Figure 4.3 Data separated by groups of %GpA. A)** Data from Anderson et al. 2017, showing that CATM predicts sequences with a higher dimerization propensity to have distinct geometries and energies. **B)** Data from CATM runs on my GASright designs. Similar results are found, with CATM predicting a better energy score, van der Waals, and hydrogen bonding on average for more stable designs, as well as a narrower crossing angle and interhelical distance.

As an initial approach, we aim to design heterodimeric GASright sequences. Although heterodimers are not necessarily symmetrical like homodimers, it is possible for heterodimers to have symmetric geometries. Therefore, attempting to design heterodimer sequences from known homodimer geometries is a reasonable initial approach. In a previous study in our lab, we found that distinct geometries in GASright result in different levels of association according to our CATM prediction algorithm (Anderson et al., 2017; Mueller et al., 2014). I ran CATM on my designed GASright sequences and created a table of data similar to as found in Anderson et al. 2017 (Figure 4.3). As demonstrated previously, CATM predicts the association GASright sequences and shows that sequences with different levels of association (S4.1) have a geometric dependency. The interhelical distance and crossing angles narrow as dimerization propensity increases. There does not seem to be a large dependency on axial rotation and z-shift, but these values give us A diagram of a dna structure

Description automatically generatedstarting points that can be applied in heterodimer design.

**Figure 4.4 Heterodimer design strategy.** To simplify heterodimer design, we can take known non-homodimerizing sequences (templates) from previous research and design a sequence to associate with it (design).

To simplify heterodimer design, we can design sequences against a single template sequence (Figure 4.4). I expand on how I would implement this into my design algorithm in Supplementary Details 4.6.1 to 4.6.3. From data in our previous work, we have characterized many sequences that do not homodimerize. These non-homodimerizing sequences are ideal templates to use for heterodimer design. After designing sequences against these templates, we need to ensure that the designed helices do not homodimerize. We can determine this by predicting their homodimerization in CATM (Anderson et al., 2017; Mueller et al., 2014). Any sequences found to associate with a stable energy in CATM can be A diagram of a diagram of a diagram

Description automatically generatedremoved from our pool of sequences that we plan for experiments.

**Figure 4.5 Heterodimer experimental strategy.** By expressing two plasmids that code for different antibiotic resistance (AMP and KAN) and each expressing a different TM sequence (Template and Design), we can investigate the dimerization propensity of heterodimers.

After designing sequences to associate to a template non-homodimerizing sequence, either TOXGREEN or sort-seq can be used to evaluate the dimerization propensity. Other students in the Senes lab are developing a two-plasmid TOXGREEN system, where one plasmid codes for the template sequence and the other for the design. We can clone the template sequence into a bacterial strain with resistance to an antibiotic, then clone a library of our designed sequences into the bacterial strain that expresses the template sequence with resistance to an additional antibiotic (Figure 4.5). Because we are designing GASright sequences which show better correlation to our energetics, we expect that exploring a range of energies of our designed heterodimers will yield a range of dimerization.

One limitation of using homodimer geometries for heterodimer design is the potential lack of sequence diversity at the interface. The symmetrical nature of these backbones may often result in similar (G/A/S)xxx(G/A/S) at the interface. To account for this, we can also study the effect of making point mutations along the design sequence. By mutating interfacial positions to another AA often found in GASright, we can compare dimerization propensity based on small changes in packing (G→A/S) or addition of interhelical hydrogen bonding (A/S→G). If our energetics determine that the main differences between two sequences with one AA difference is mostly due to a loss of van der Waals packing, then the data would suggest that packing is a stronger driving force when applied alongside interhelical hydrogen bonding. The energetic data from previous research suggests that this is a possibility (Anderson et al., 2017; Díaz Vázquez et al., 2023), where van der Waals packing contributes to stability more than hydrogen bonding.

To further expand on geometries for heterodimer design, we can explore the stability of helices at various combinations of axial rotations and z-shifts. For homodimers, I assessed the energetics of poly-Leu sequences with Ala or Gly at all interfacial positions for randomized symmetric axial rotations and z-shifts (Section 3.3.3). We can use this method to identify asymmetric axial rotations and z-shifts. However, the heterodimer space is quite large, and can expand an even larger range depending on how finely we grid the space. For example, if we wanted to explore energetic trends for 100° of axial rotation in 1° increments, we would have to calculate the energetics for 1002 conformations. Simultaneously, we’d also need to consider non-symmetric changes in z-shift for each pair of axial rotations, increasing the extensiveness of the geometric space. Rather than trying to brute forcibly compute possible energetic trends for all heterodimer axial rotations, we can first explore favorable parameters from previously studied homodimeric GASright proteins. Combining the GASright designs in my study with other research in the lab, we have a database of hundreds of GASright homodimers. Using combinatorial testing of the most common axial rotations and z-shifts (Figure 4.3), we may be able to identify regions that are energetically favorable for heterodimer design.

A graph of different types of numbers

Description automatically generated with medium confidence4.4.2 Turning sequence entropy into a pairwise term

**Figure 4.6 Amino acid frequency from TMs extracted from OPM. A)** Each AA was counted as found in TMs identified from OPM and divided by the total count of all AAs to determine the frequency. Data separated by year (2021, light blue; 2024, white). **B)** Frequency of AA pairs found in at least 10 designed sequences compared with the frequency of same AA pairs in all TMs. The largest differences in AA pair frequencies are shown here, with negative differences (red) corresponding to AA pairs more often expressed TMs than in designed sequences (gray). Pairs found in all regions marked in red, and pairs found in Left and Right marked in blue.

To design sequences similar to natural MP proteins, we created a SEQUENCE\_ENTROPY term detailed in section 3.3.4. Briefly, this term uses the natural distribution of AAs in MPs to design a sequence like MPs. It is currently implemented as a similarity score, with higher values being determined as more similar. We expected this term to help normalize experimental expression and insertion; however, we may not be maintaining packing interactions found in natural MP structures. Previous research has shown that protein activity and folding are affected by small changes in AA sequence (Faham et al., 2004; Gratkowski et al., 2001; Johnson et al., 2007; Russ & Engelman, 2000; Zhou et al., 2000; Zhou et al., 2001). To further investigate how these changes impact sidechain packing, we can determine how individual AAs might affect those around them.

I recalculated the composition of AAs for all non-redundant TMHs in OPM as of May 14, 2024. The composition is quite similar to the composition of AAs for all non-redundant TMHs determined in 2021 (Figure 4.6A). I then chose to analyze the pairs of AAs found within the interface of my successful designs (as defined by SASA) and those within the TMHs (Figure 4.6B). Although we successfully designed interfacial sequences that reflect the frequency of AAs in TMs (Figure S3.1), there are multiple pairs of AAs that deviate between natural TMs and our designs. For example, Figure 4.6B shows that the AA pair of GV is expressed >20% more frequently in TMs than in our designed sequences. In a future design run that considers pair frequencies of AAs, we can remove these biases to more accurately design sequences like natural TMs. However, this comparison is between all AAs within TMH sequences and not the interfaces of TMs. It may be more informative to first identify interfaces between TMs from solved structures and use this AA frequency for future designs. Additionally, identifying relationships between AA frequency by position (i.e. AA3 = Ala, Ala separated by 3 bases) could have the added benefit of designing sequences that maintain atomic interactions found within natural MPs. I reference where changes to our sequence entropy term can be made in Supplementary Details 4.6.4.

4.4.3 Machine learning ideas

One final approach to improving the design procedure is to better optimize our energetic algorithm using machine learning. Another student in the Senes lab is using machine learning to optimize the weights for CATM, aiming to gain a better understanding of which forces contribute more to stability. By applying similar regression training to fit each of our energy terms ("A Review on Linear Regression Comprehensive in Machine Learning," 2020), we may be able to identify the reason why our energetics do not correlate well outside of GASright. By taking each of our energy terms and re-weighting, we may find that there is a relationship between the two terms most crucial for our designs non-hydrogen bonding designs: packing and implicit solvation. For example, the implicit solvation term may be weighted too low and packing too high. This would allow us to readjust the energy terms for each of our designs and to better understand the extent at which packing can drive association in these regions.

Machine learning could also be applied to improving upon design without reliance on energy terms. Alphafold2 uses multiple sequence alignments (MSA) as a foundation for determining structures from a given sequence (Jumper et al., 2021). Protein design is a similar problem, where we aim to find a sequence that can accommodate a given structure. Proteins have been successfully designed using recurrent neural networks including anticancer peptides (Grisoni et al., 2018), and another group recently developed the ProteinSolver web server that uses deep graph neural networks to design sequences using distance matrices between AAs (Strokach et al., 2020). Applying the previous methods or other algorithms could be a relevant option for membrane protein design. In ProteinSolver, we are able to supply a reference sequence or structure to design against, and we can set positions on the design to specific AAs. This algorithm then generates designed sequences that are expected to associate to the given sequence. In architecture, it is quite like what we aim to do with heterodimer design, except using machine learning. Although not specifically built to design membrane proteins, using ProteinSolver is an alternative to our current design procedure. Additionally, we have a database of 3413 unique TMH pairs extracted from OPM as of May, 2024. Similar to ProteinSolver, we can create our own distance matrices between AA pairs in these structures and use these as restraints to train a deep graph neural network for membrane protein design.

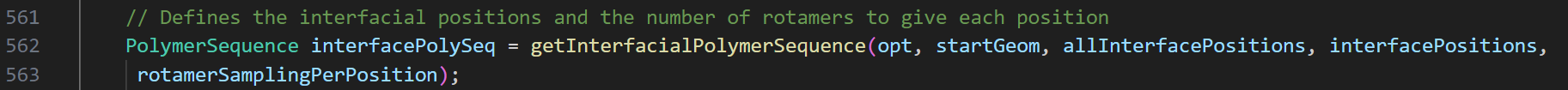
4.5 Detecting protein concentration in high-throughput

One of the weaknesses of our TOXGREEN assay is the inability to accurately determine the expression levels of each of our proteins in the membrane. Currently, much of this research holds the assumption that our proteins express at the same level. We designed experiments to control for this variable previously by studying the dimerization of interfaces on poly-Leu backbones and in my study by maintaining the sequence composition as found in natural MP sequences. Subsets of sequences are then extracted and analyzed for their ability to express using western blots. Although we found that sequences designed in each region had similar expression, the Right-handed designs displayed noticeably less expression compared to Left-handed and GASright (Figure S2.1). Additionally, the western blot only assesses total protein concentration, not considering just the proteins that are inserted into the membrane. Therefore, it is possible that our Right-handed designs express and insert the same in the membrane as other designed sequences. However, we currently do not have the tools to assess this. To improve our accuracy in measuring MP association, we would ideally develop a technique to determine protein concentration in high-throughput. We could then leverage the protein concentration to normalize the fluorescence yield of each sequence to its protein concentration.

One way to measure the expression of our MPs is to use fluorescence. By labeling our proteins with fluorescent proteins that are viable in bacteria, we can measure the amount of expressed protein through fluorescence. Recently, fluorescent proteins have been used to track movements of membrane proteins within bacterial cells (Lyu et al., 2022; Navarro et al., 2022). Since we can measure the fluorescence of expressed superfolder GFP (sfGFP) in TOXGREEN, we need a molecule that emits a different wavelength of light such as TagRFP-t (tRFP) (Yang et al., 2021). tRFP can be expressed in our TOXGREEN cell line fused to MBP on our proteins. We can detect the amount of tRFP signal through FACS. Even if the total fluorescence output from tRFP yields a wide range and our sequences don’t express similarly, we can use FACS to identify cells that express our proteins similarly. By setting the FACS to only count cells found within an acceptable window of fluorescence, we can perform multiple sorting runs on populations of cells with similar expression. Finally, we can normalize dimerization propensity by the amount of tRFP-t, allowing us to assess association of our sequences independent of protein expression.

4.6 Supplementary Details

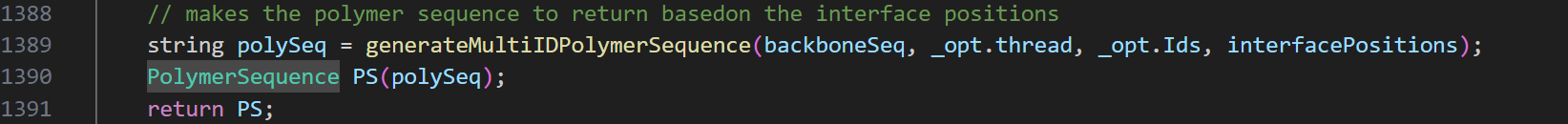
**4.6.1 Separate helices: one as template and one to designed**

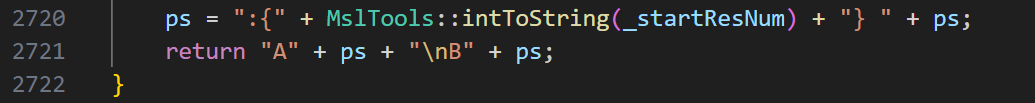
This function creates a sequence with alternate identities at the given interfacial positions.

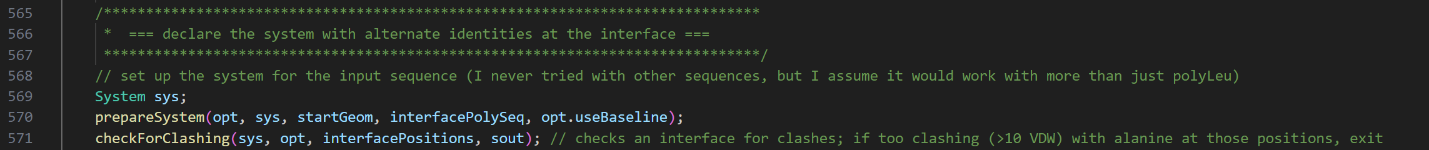
Within this function are two functions of importance.

A computer screen shot of a computer code

Description automatically generatedThe function below creates a backboneSeq (backbone sequence) to design on depending on the given options backboneAA, backboneLength, and useAlaAtTermini (i.e. if inputs are L, 21, and False, it will design on a 21 AA poly-Leu helix with no Ala at the termini positions).

The program uses that backboneSeq to generate a polymer sequence, which is read into a System to build helices with multiple identities at the interface positions.

For ease of transitioning to heterodimers, remaking the above function as a heterodimer function that takes an additional input (i.e. \_opt.templateSequence) should result in a heterodimer being built into the System. To do this, make sure you add templateSequence as an option in the Options structure (line 64) and the parseOptions function (line 3620).

The end of the generateMultiIDPolymerSequence function uses a copy of the ps (polymerSequence) for symmetry. It will need to be edited so that it instead makes a polymer sequence for the \_opt.templateSequence and the backboneSeq of interest. Example: return “A” + ps1 + “\nB” + ps2;

If helices have been successfully separated, outputting the System as a PDB using the following function after the above lines should show that Chain A and Chain B are different sequences.

A computer code on a black background

Description automatically generated

A screen shot of a computer screen

Description automatically generated**4.6.2 Creating a non-symmetric version of the switchSequence function**

This function uses the setActiveSequence function to change the System to a given sequence.

A computer screen shot of a program code

Description automatically generatedIt is currently set to change all chains in the System to a given sequence. An alternate version of this function could instead set specific chains to a given sequence.

**Example**

Remake the setActiveSequence function for specific chains:

setActiveSequenceChainA and setActiveSequenceChainB

Then use setActiveSequenceChainA for design sequence and setActiveSequenceChainB for the template.

If this works properly, you should be able to change all instances of switchSequence to the new version, and change all instances of setActiveSequence to your new version. This should result in only your design sequence being changed throughout the design process.

A black background with colorful text

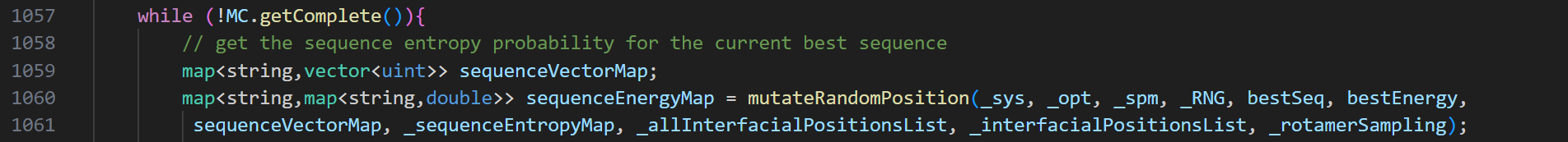
Description automatically generated**4.6.3 Outputting the designed heterodimer sequence**

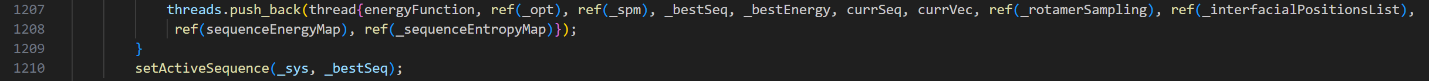
To output the current sequence, only Chain A is used throughout most if not all of the program. If the sequence to design is set to Chain A and the template to design against set as Chain B, this should continue to work properly when outputting the designed sequence.

**4.6.4 Adjusting sequence entropy**

At the beginning of the sequence search MC, a random position on the sequence is mutated to all AAs and the best AA is assessed against the current sequence.

A black screen with many lines

Description automatically generated with medium confidenceWithin this function is a function that calculates the energy terms for each iteration of the designed sequence.

If you are aiming to change how the sequence entropy is calculated, you will need to alter the math in the calculateInterfaceSequenceEntropy function (detailed in section 3.3.4).

A screen shot of a computer code

Description automatically generated

4.7 References

A Review on Linear Regression Comprehensive in Machine Learning. (2020). *Journal of Applied Science and Technology Trends*, *1*(2), 140-147.

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>

Choma, C., Gratkowski, H., Lear, J. D., & DeGrado, W. F. (2000). Asparagine-mediated self-association of a model transmembrane helix. *Nat Struct Biol*, *7*(2), 161-166. <https://doi.org/10.1038/72440>

Del Piccolo, N., Sarabipour, S., & Hristova, K. (2017). A New Method to Study Heterodimerization of Membrane Proteins and Its Application to Fibroblast Growth Factor Receptors. *Journal of Biological Chemistry*, *292*(4), 1288-1301.

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>

Faham, S., Yang, D., Bare, E., Yohannan, S., Whitelegge, J. P., & Bowie, J. U. (2004). Side-chain contributions to membrane protein structure and stability. *J Mol Biol*, *335*(1), 297-305. <https://doi.org/10.1016/j.jmb.2003.10.041>

Gratkowski, H., Lear, J. D., & DeGrado, W. F. (2001). Polar side chains drive the association of model transmembrane peptides. *Proc Natl Acad Sci U S A*, *98*(3), 880-885. <https://doi.org/10.1073/pnas.98.3.880>

Gray, T. M., & Matthews, B. W. (1984). Intrahelical hydrogen bonding of serine, threonine and cysteine residues within alpha-helices and its relevance to membrane-bound proteins. *J Mol Biol*, *175*(1), 75-81. <https://doi.org/10.1016/0022-2836(84)90446-7>

Grisoni, F., Neuhaus, C. S., Gabernet, G., Müller, A. T., Hiss, J. A., & Schneider, G. (2018). Designing Anticancer Peptides by Constructive Machine Learning. *ChemMedChem*, *13*(13), 1300-1302. <https://doi.org/https://doi.org/10.1002/cmdc.201800204>

Johnson, R. M., Hecht, K., & Deber, C. M. (2007). Aromatic and cation-pi interactions enhance helix-helix association in a membrane environment. *Biochemistry*, *46*(32), 9208-9214. <https://doi.org/10.1021/bi7008773>

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O.,…Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589. <https://doi.org/10.1038/s41586-021-03819-2>

Lyu, Z., Yahashiri, A., Yang, X., McCausland, J. W., Kaus, G. M., McQuillen, R.,…Xiao, J. (2022). FtsN maintains active septal cell wall synthesis by forming a processive complex with the septum-specific peptidoglycan synthases in E. coli. *Nature Communications*, *13*(1), 5751. <https://doi.org/10.1038/s41467-022-33404-8>

Mueller, B. K., Subramaniam, S., & Senes, A. (2014). A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Cα-H hydrogen bonds. *Proc Natl Acad Sci U S A*, *111*(10), E888-895. <https://doi.org/10.1073/pnas.1319944111>

Navarro, P. P., Vettiger, A., Ananda, V. Y., Llopis, P. M., Allolio, C., Bernhardt, T. G., & Chao, L. H. (2022). Cell wall synthesis and remodelling dynamics determine division site architecture and cell shape in Escherichia coli. *Nature Microbiology*, *7*(10), 1621-1634. <https://doi.org/10.1038/s41564-022-01210-z>

Patel, S., Mackerell Jr., A. D., & Brooks III, C. L. (2004). CHARMM fluctuating charge force field for proteins: II Protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *Journal of Computational Chemistry*, *25*(12), 1504-1514. <https://doi.org/https://doi.org/10.1002/jcc.20077>

Russ, W. P., & Engelman, D. M. (2000). The GxxxG motif: a framework for transmembrane helix-helix association. *J Mol Biol*, *296*(3), 911-919. <https://doi.org/10.1006/jmbi.1999.3489>

Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A., & Kim, P. M. (2020). Fast and Flexible Protein Design Using Deep Graph Neural Networks. *Cell Systems*, *11*(4), 402-411.e404. <https://doi.org/10.1016/j.cels.2020.08.016>

Ulmschneider, M. B., Ulmschneider, J. P., Freites, J. A., von Heijne, G., Tobias, D. J., & White, S. H. (2017). Transmembrane helices containing a charged arginine are thermodynamically stable. *Eur Biophys J*, *46*(7), 627-637. <https://doi.org/10.1007/s00249-017-1206-x>

Yang, X., McQuillen, R., Lyu, Z., Phillips-Mason, P., De La Cruz, A., McCausland, J. W.,…Xiao, J. (2021). A two-track model for the spatiotemporal coordination of bacterial septal cell wall synthesis revealed by single-molecule imaging of FtsW. *Nature Microbiology*, *6*(5), 584-593. <https://doi.org/10.1038/s41564-020-00853-0>

Zhou, F. X., Cocco, M. J., Russ, W. P., Brunger, A. T., & Engelman, D. M. (2000). Interhelical hydrogen bonding drives strong interactions in membrane proteins. *Nat Struct Biol*, *7*(2), 154-160. <https://doi.org/10.1038/72430>

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>

Zhu, X., Lopes, P. E. M., & MacKerell Jr, A. D. (2012). Recent developments and applications of the CHARMM force fields. *WIREs Computational Molecular Science*, *2*(1), 167-185. <https://doi.org/https://doi.org/10.1002/wcms.74>