Chapter 4: Future Directions

Table of Contents

[4.1 Summary of Dissertation 3](#_Toc172901150)

[4.2 Studying the impact of sidechain packing with other forces 4](#_Toc172901151)

[4.2 Improving the protein design algorithm 7](#_Toc172901152)

[4.2.1 7](#_Toc172901153)

[4.2.2 Heterodimer design 8](#_Toc172901154)

[4.2.3 Shape Fitting 9](#_Toc172901155)

[4.3 Detecting protein concentration in high-throughput 9](#_Toc172901156)

4.1 Summary of Dissertation

In the previous chapters, I detailed the research that the Senes lab has focused on in terms of 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. In my research, we studied a subset of van der Waals packing known as sidechain packing. We found that sidechain packing is a weak driving force when it is the sole force involved in the association of small TMH systems.

In my 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 with only sidechain packing dimerized less than proteins designed with sidechain packing and hydrogen bonding (GASright). Additionally, our energetics correlate much better to the GASright region, suggesting that our tools are not well tuned to predict the association of weak MP dimers. In Chapter 3 I detail a majority of the computational methods that I developed to design dimers and the rationale for each decision 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. At the end of Chapter 3, I also detail additional experiments that were not included in my publication, where I studied the effects of mutating leftover hydrogen bonding residues that were used in the design process. These results show that we designed multiple dimers that associate by sidechain packing in the absence of any potential hydrogen bonding.

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 will expand on potential avenues for future research where this knowledge can be utilized. I first suggest future experiments to enhance our understanding of sidechain packing in the presence of other forces. I then detail my ideas for potential improvements to my design algorithm, including the potential for designing heterodimer sequences. Finally, I suggest improvements for our high-throughput sort-seq, with attempts to further characterize sequences by their expression level in cells.

4.2 Studying the impact of sidechain packing with other forces

To investigate the impact of sidechain packing on MP folding and association, I designed thousands of proteins using only a subset of AAs. However, this subset of AAs includes two AAs (Ser and Thr) with the potential to form hydrogen bonds (cite) and two AAs (Trp and Phe) with the potential for electrostatic interactions. Tyr, another AA included in our designs, has the potential to facilitate both interactions. Our current dataset has the potential to facilitate interactions outside of solely sidechain packing, but our data suggests that they associate primarily through van der Waals packing. Mutational testing on our sequences suggests that hydrogen bonding is unlikely to play a role in all our sequences, and the structures do not suggest that electrostatics are involved in association.

We included these AAs because of their relative abundance in MP sequences. Ser and Thr are typically found to hydrogen bond with the backbone carbonyl oxygens in monomeric helices (cite), so we expected the addition of these AAs not to have a large impact on folding through additional hydrogen bonding. Larger AAs (Trp, Phe, and Tyr) have an extensive van der Waals radius, which we did not want to compromise losing in our experiments. Excluding these AAs would restrict our design pool to four AAs (Leu, Ala, Val, and Ile). Because we decided to use a mutagenesis strategy implementing Ala and Ile, our designs would be even more restricted with the number of potential designable sequences as well as preventing them from being similar to MP sequences. However, a more purposeful study would be helpful to ascertain the overall impact of packing on folding and association.

To better study the impact of sidechain packing alongside other forces, we can rerun our design algorithm with a few different constraints. Instead of optimizing the sequences we choose solely by their calculated amount of van der Waals packing, we can instead choose sequences with a considerable amount of hydrogen bonding. We can see if an increase in the amount of predicted hydrogen bonding and number of hydrogen bonds results in higher dimerization, like GASright. One option is to include electrostatics in our energy calculation to determine if structures more amenable to these interactions are more stable. While the previous two ideas utilize the same AAs as the designs run in my paper, designing sequences with an alternative subset of AAs could also be an option. Charged interactions have been found to impact folding in a variety of MP systems. 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. Alternatively, designing with only a subset of AAs with the potential for hydrogen bonding (Ser and Thr) or electrostatic interactions (Phe or Trp) would allow us to better isolate sidechain packing with these forces. By designing sequences with different numbers of each of these AAs, we can tune the amount of hydrogen bonding and electrostatics even if our energetics are inaccurate.

Another idea is to design sequences for our most accurate energy model: GASright. GAS­right are involved in a variety of essential interaction including MP signaling cascades and …. We have previously studied GASright sequences from the human genome, demonstrating that a subset of GASright correlate well with our energetics (cite). However, designing GASright resulted in a slightly worse correlation (figure? Maybe CATM and my energies against the dimerization?). Instead of tuning our AA propensity to just GASright sequences, we used the propensity for all TMs found in our TMH extracted dataset. Recalculating the propensity for these previously studied GASright proteins and implementing them in design may result in better correlation with our energetics. We can additionally study the impact of hydrogen bonding, electrostatics, and charged interactions by using the previously suggested subsets of AAs for GASright design. By having a better understanding of GASright tuned association, we will learn how impactful other forces can be to dimerization outside of sidechain packing and interhelical hydrogen bonding.

Additionally, we can further assess the impact of sidechain packing on association by altering the backbone sequence. By assessing the association differences for the same interface between poly-Leucine, poly-Alanine, and poly-Valine backbones, we can gain insight into the impact of how minute changes in a protein result in differences in association. Although these backbone AAs are not often found at the interface, the slight loss in steric bulk (Leu->Val->Ala) may result in changes in association. Therefore, there may be a way to correlate the changes in association with change in AA size, allowing us to tune our van der Waals energy term by assessing these differences in steric bulk.

4.3 Improving the protein design algorithm

In Chapter 3, I detailed how I developed my protein design algorithm. I utilized known literature and improved my algorithm over time with the knowledge at hand. However, most of my algorithm was finished prior to the development of AlphaFold, a machine learning based algorithm that predicts both known and unknown protein structures to near atomic precision. According to the data in my paper in Chapter 2, many of my structures associate according to the design interface. However, our energetic terms are unable to predict the ability for these proteins to associate. By studying AlphaFold and other machine learning based algorithms, we can improve the ability of the algorithm to better design structures and our energetics for predicting association.

One of the advantages of AlphaFold is the ability to predict structures without energetics. AlphaFold2 (describe AlphaFold in detail here; maybe take the image?). However, many groups aim to design proteins based on structures found in nature rather than sequence. I am unable to predict the structures of my designed proteins due to the sequence dependency of AlphaFold2. However, the way that AlphaFold2 connects sequence information to structure is relevant for improving design.

4.3.1

By extracting sequence information from known MP structures, we can develop a prediction algorithm that considers the frequency of amino acids and their nearby partners. The current iteration of the design algorithm uses **(describe types of terms here)** energetics to decide on amino acids at the interface. But utilizing natural sequence interfaces for design could improve the algorithm, as well as give insight into how to improve the energy terms. In the current algorithm, we have developed a term like this idea: **…explain membrane sequence similarity term here**. Using this similarity score, our sequences were designed to resemble MP sequences based on overall frequency. **Say caveats for the previous term here**

By tuning a new sequence predictor term **(maybe explain the math a bit here? With a theoretical formula?)** and designing sequences with different values from 0-100%, we may see that tuning closer to natural sequences at the interface results in increased association. However, with this new term it is possible that we see that the AAs near each other in weaker sequences are not as frequently found in MP structures as in stronger sequences. This gives credence to why our energetics might be better able to predict GASright sequences. The GASright sequence signature is maintained in our structures, and the frequency at the interface likely …

4.3.2 Heterodimer design

My research focused on assessing the stability of homodimer proteins, partially due to their simplicity to design. However, important biological interactions like … are composed of heterodimers (cites). I am currently working with another graduate student in the lab to develop our protein design algorithm for heterodimer sequences. Heterodimer design adds multiple variables of complexity for design: When designing homodimer sequences, we can simplify our algorithm by setting individual helices at symmetric geometric terms. However, with heterodimers, two of the geometric terms are no longer symmetric: z-shift and axial rotation. Each of these terms is a representative of the alignment between interfacial amino acids and their placement within the membrane respective to each helix. The z-shift defines the placement of the crossing point within the membrane while the axial rotation defines the part of the helix at the interface (**figure**). With this increased geometric search space, much of the algorithm that relies on symmetry is no longer applicable and must be developed for individual helices.

Since our energy terms work best with GASright structures, as an initial approach we aim to design heterodimeric GASright sequences. As geometries to start with, we can use GASrights found within the MP PDB space. For homodimers, I assessed the energetics of poly-Alanine sequences at crossing angles and x-shifts for a randomized set of symmetric axial rotations and z-shifts. Geometries that had an energy below 10kcal/mol were accepted, as we expect backbone optimization to relax structures to a stable energy. We can use this method to identify asymmetric axial rotations and z-shifts. However, the search space for applying this method is exponentially larger than previous: for every rotation, every other rotation must be applied and the same must be done for z-shifts (**equation here**).

Rather than brute forcing this for multiple crossing angles and x-shifts, the first approach is to utilize our current homodimer geometries to design heterodimer sequences. Although heterodimers have an increased geometric space, the geometries for homodimers are a portion of that space (**say this better/figure**). Therefore, attempting to design heterodimer sequences is a reasonable initial approach. To simplify this process, we can design sequences against a single helix. This simultaneously makes experimental design easier (**figure**).

Below, I detail the changes I would make to the design code in order of priority (maybe make this a figure; could also directly reference pieces of my code?):

1. Separate helices: one as stationary, or undesigned, and the other as designed
   * Add an additional option for the stationary helix so that an input sequence can be given on command line or in the configuration file
   * Set both helices to the input sequence, then only add AAs for design to the design helix
   * Adjust the (add function here) to only change the AA on the design helix
   * Check that the BASELINE\_MONOMER still works for heterodimers (if not, adjust it)

After designing hundreds of sequences to associate to a sequence of interest, TOXGREEN sort-seq can be used to evaluate the dimerization propensity. Because we are designing GASright sequences which show better correlation to our energetics, we expect that exploring a range of energies will result in a range of dimerization.

While preparing the design code for heterodimer design, we can also continue to think about ways to explore heterodimer geometries. In a previous study in our lab, we found that distinct geometries in GASright results in different levels of association. However, this was only a small study of 26 proteins. Combining the GASright from recent research in the lab, we now have a database of hundreds of GASright proteins. We can further pinpoint local geometries for different levels of protein association. By taking the aggregate geometries for these, we can identify input geometries to start with. **(I should do this for my stuff and CATM geometries as kind of data here)** These can be used as starter geometries for heterodimer design before we identify heterodimeric structures.

One limitation of using homodimer backbones is the potential lack of sequence diversity at the interface. Although GASright are typically made of G/A/SxxxG/A/S the symmetrical nature of these backbones may result in the small Gly being the most favorable AA. To account for this, we can adjust the backbone optimization to shift only the geometric parameters for a single helix. We can then mutate A and S at these positions to determine the stability for these sequences with non-symmetric backbones. By comparing the dimerization propensity of similar designed sequences against the stationary helix, we can get a better understanding of the importance of packing vs interhelical hydrogen bonding. For example, 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, the data would suggest that packing is able to be a stronger driving force when applied alongside interhelical hydrogen bonding. The energetic data from Anderson et al. 2017 suggests that this is a possibility (**reference or adapt figure here**).

Below I detail the changes that we can make for backbone optimization:

* **Need to look into code to figure out how to implement this**

4.3.3 Machine Learning Ideas

One final approach to improving the design procedure is to better optimize our energetic algorithm through the use of machine learning. Another student in the lab is using machine learning to optimize the weights for CATM, giving us a better understanding of what forces contribute more to stability. By taking his approach for sequences outside of the GASright region, we may also be able to identify the reason why our energetics do not correlate well outside of GASright. Using linear regression models to fit each of our energetic terms against the dimerization propensity in a 70% train/30% split (cite), will provide us the ability to better predict the energetics of sequences designed outside of GASright.

An alternative use of machine learning is to …

4.4 Detecting protein concentration in high-throughput

One of the weaknesses of our TOXGREEN assay is the inability to accurately assess the expression levels of each our proteins in the membrane. Currently, much of the 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 (cites) 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 both left and GASright (figure). 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 assessing MP association, developing a way to assess the concentration of our proteins in high throughput would be beneficial, allowing us to normalize the fluorescence yield of each sequence to the 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 measure the fluorescence of expressed superfolder GFP (sfGFP), a molecule expressing a different wavelength of light such as TagRFP-t (tRFP) (Yang et al., 2021) can be fused to the MBP on our proteins. We can detect the amount of tRFP signal alongside our sfGFP signal on a FACs. However, the total fluorescence output from tRFP would relay the same information as western blots. Instead of just reading the tRFP information for each counted cell, we can set the FACs to identify an acceptable window of fluorescence; Cells that express beyond the upper and lower bounds of tRFP fluorescence would not be counted. Some sequences may inherently express more than others, so we can run multiple sorting runs on populations with different bounds of tRFP. This would allow us to measure the dimerization propensity of proteins with similar levels of MP expression.

We could then normalize by the amount of TagRFP-t and assess dimerization propensity against expression.

* I think there was a paper I read for chem bio seminar that was pretty relevant

So I just realized that there’s truly a problem with not knowing the protein concentration. In Samantha’s study, she normalized to %GpA by TOXCAT, done by lots of people previously. But in Gladys’ study, she saw that the proteins that she chose from Samantha’s study were all weaker than GpA in vitro, WHEN NORMALIZED BY CONCENTRATION. This means that GpA inherently expresses lower than those other constructs, which is a big part of the reason why our proteins look like they’re associating better. And this is likely the same with G83I. GpA has a deltaG of -6 while G83I is -2.5. Hilariously enough, most of my proteins are right around G83I, a known monomer. So this also means that since it’s expressing lower, we’re likely seeing better results in TOXGREEN and it only gets enhanced when doing it in high-throughput sort-seq. This is a big reason why sort-seq needs to be normalized for protein concentration in the future. It also might mean that a lot of my designed proteins that I think are associating aren’t actually associating. Well…actually I just rethought it a bit. My designs show pretty good association in TOXGREEN, suggesting that they’re fine. Sort-seq actually might be kind of cool though: it’s simultaneously supposed to be more accurate AND more sensitive. But I think we get a bit of noise along the way: converting counts to reconstructed fluorescence to TOXGREEN to TOXCAT to Kd to deltaG. So I actually think that although it’s not 100% good, it’s a fair way to assess association.

Address the above somehow

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>

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>

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>