Chapter 4: Future Directions

Table of Contents

[4.1 Summary of Dissertation 3](#_Toc173176992)

[4.2 Studying the impact of sidechain packing with other forces 5](#_Toc173176993)

[4.3 Improving the protein design algorithm 8](#_Toc173176994)

[4.3.1 Heterodimer design 8](#_Toc173176995)

[4.3.2 Turning sequence entropy into a pairwise term 11](#_Toc173176996)

[4.3.3 Machine Learning Ideas 12](#_Toc173176997)

[4.4 Detecting protein concentration in high-throughput 12](#_Toc173176998)

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 sidechain packing is a weak driving force when it is the sole force involved in the association of small TMH systems.

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 with only sidechain packing dimerized less than proteins designed with both sidechain packing and hydrogen bonding (GASright). Additionally, our energetics correlate 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 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 discussed 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 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 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 and improving our energetics using machine learning. 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 (**cite**). Tyr, another AA included in our designs, has the potential to facilitate both interactions (**cite**). Our current dataset has the potential to facilitate 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, 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?**). 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 wanted to include in our experiments. Excluding these hydrogen bonding and ring 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. Additionally, sequences designed with a small pool of AAs would not be representative of MP-like sequences. However, studying the impacts of forces alongside packing could be addressed using our current methods.

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 decided on sequences by the calculated van der Waals energy, but another study could instead choose sequences with a considerable amount of hydrogen bonding. We can then see if an increase in the amount of predicted hydrogen bonding and number of hydrogen bonds results in higher dimerization, like GASright. Another option is to include electrostatics in our energy calculation. Including this term may result in designing structures amenable to electrostatic interactions such as pi-pi stacking (**cite**). If the energetics are found to match dimerization propensity for structures with more hydrogen bonding or electrostatics, 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 also be able to identify a trend between packing and its impact on association alongside other forces.

While the previous ideas utilize the same AAs as the designs run in my paper, designing sequences with an alternative set of AAs could also be an option. We can continue to include the AAs that can only pack (Leu, Ala, Val, and Ile) alongside other subsets of AAs. Designing with hydrogen bonding (Ser and Thr) or electrostatic (Phe 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 (**cite**). Including a subset of 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 been done mutating single residues on poly-Leucine helices to determine the affect of different AAs (cite). If we find that mutations to particular residues increase or decrease association, we can determine how different forces influence association alongside changes in packing.

Additionally, we can further assess the impact of sidechain packing on association by altering the backbone sequence. By assessing the association differences for sequences with the same interface on 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 better tune our van der Waals energy term to differences in steric bulk.

4.3 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, the major 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 the recent literature and thinking of ways to continue the research, in this section I detail ways to develop and build off my algorithm for future research.

4.3.1 Heterodimer design

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 placed at the same geometric value for each term. 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 for design, namely each helix 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 defines the crossing point of the interface respective to each helix (**figure**). With this increased geometric search space, the parts of the algorithm that rely on symmetry are no longer applicable and must be developed for individual helices.

As an initial approach we aim to design heterodimeric GASright 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 from known homodimer geometries is a reasonable initial approach. As starting geometries to design, we can use data from known GASright proteins. 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 designs in my study and other research in the lab, we now have a database of hundreds of GASright homodimers. We can further pinpoint local geometries for different levels of protein association. By aggregating these geometries, we can identify input geometries to start designing heterodimeric GASright proteins. (**figure**/**I should do this for my stuff and CATM geometries as kind of data here**)

To simplify heterodimer design, we can design sequences against a single helix (figure). From data in our previous work, we have characterized multiple homodimer sequences that do not associate alone. We can use each of these sequences as templates for design, aiming to find sequences that associate as heterodimers. 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. Any sequences found to associate in CATM can then be removed from our pool of sequences that we plan for experiments. By designing sequences against a single helix, this simultaneously makes experimental design easier. Other students in my lab are developing a two-plasmid TOXGREEN system, where one plasmid codes for the template helix and the other for the designed helix. We can clone the template helix into a bacterial strain with resistance to an antibiotic, then clone a library of our designed heterodimer sequences into that strain with resistance to another antibiotic.

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)
2. **Backbone optimization changes** 
   * **Look into the code**

After designing many sequences to associate to a sequence of interest, either TOXGREEN or 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 of our designed heterodimers will yield a range of dimerization.

One limitation of using homodimer backbones 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/SxxxG/A/S at the interface. To account for this, we can also study the effect of making point mutations along one helix. 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, 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, where van der Waals packing contributes … to stability more than hydrogen bonding (**reference or adapt figure here; could this also be true with my stuff?**).

To further expand on heterodimer geometries, we can explore the stability helices at combinations of axial rotations and z-shifts. 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 (**figure**?). 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 and figure**). Rather than brute forcing this for a multitude of crossing angles and x-shifts, we can explore favorable heterodimer geometries for the local geometries we’ve found from previous homodimeric GASright proteins.

4.3.2 Turning sequence entropy into a pairwise term

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 with similar to 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 (cites). Our term designs sequences similar to MPs, but does not consider the reliance of AAs on each other. To further investigate the impact of packing on these changes, we can calculate 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 previous (figure). Similar to the original sequence entropy term, I determined the distribution of pairs of AAs being found within the same sequence to account for the frequency AAs are paired.

**I also analyzed the distribution for sequences that successfully associate, pass maltose test, and … for designs**

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 …

4.3.3 Training energy terms

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. Another graduate student in the lab is currently using regression training to fit each of our energetic terms against the dimerization propensity (cite). This 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.

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>