Chapter 4: Future Directions

Table of Contents

[4.1 Summary of Dissertation 2](#_Toc173449330)

[4.2 Studying the impact of sidechain packing with other forces 4](#_Toc173449331)

[4.3 Improving the protein design algorithm 7](#_Toc173449332)

[4.3.1 Heterodimer design 7](#_Toc173449333)

[4.3.2 Turning sequence entropy into a pairwise term 10](#_Toc173449334)

[4.3.3 Training energy terms 11](#_Toc173449335)

[4.4 Detecting protein concentration in high-throughput 12](#_Toc173449336)

[4.5 Supplementary Details 14](#_Toc173449337)

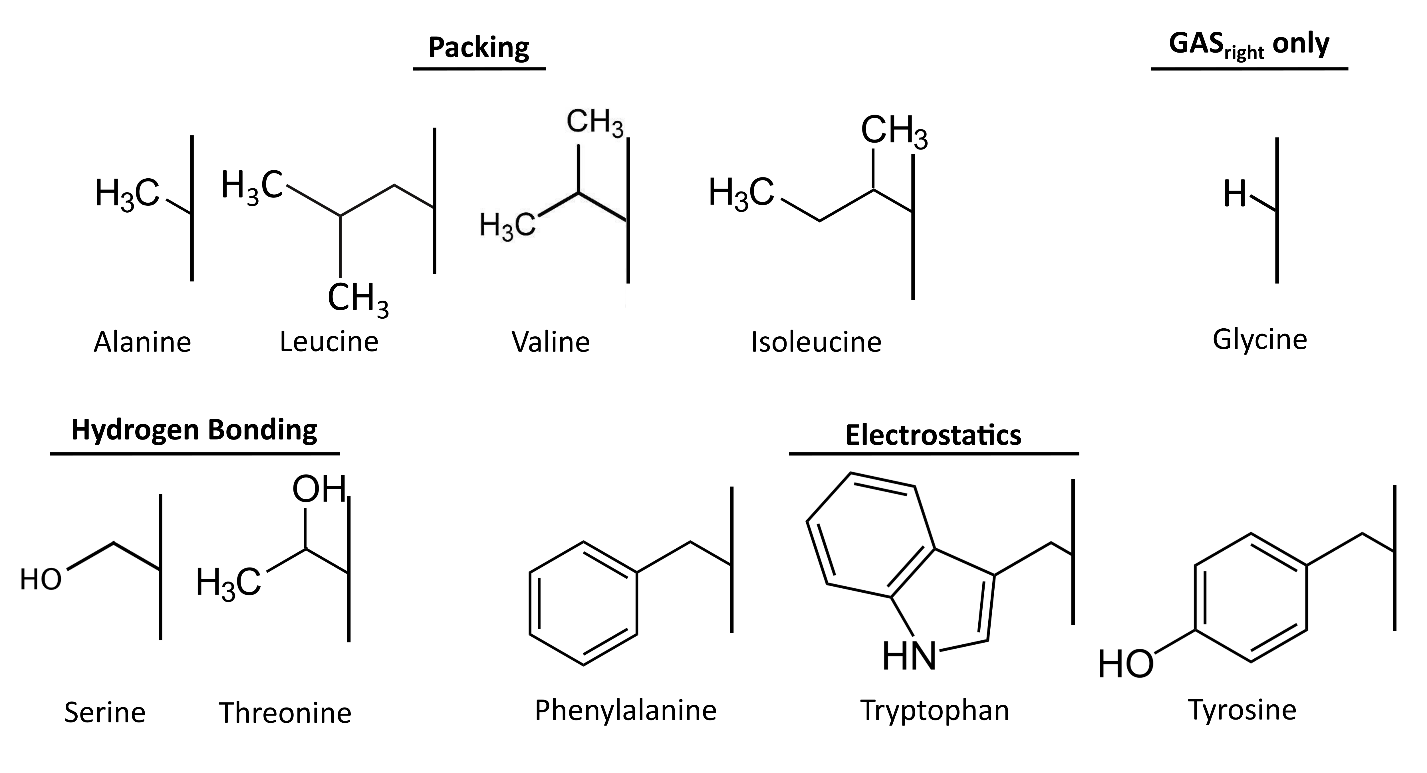
[4.6 References 15](#_Toc173449338)

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 solely sidechain packing dimerized less than proteins designed with both sidechain packing and interhelical 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

**Figure 4.1 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 4.1**). However, this subset of AAs includes two AAs (Ser and Thr) with that can form hydrogen bonds (Russ & Engelman, 2000) and three 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, 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 (**ref fig Chapter 3?**). Ser and Thr are typically found to hydrogen bond with the backbone carbonyl oxygens in monomeric helices (Gray & Matthews, 1984), 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 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). Because we decided to use a mutagenesis strategy implementing Ala and Ile, our designs would be 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.

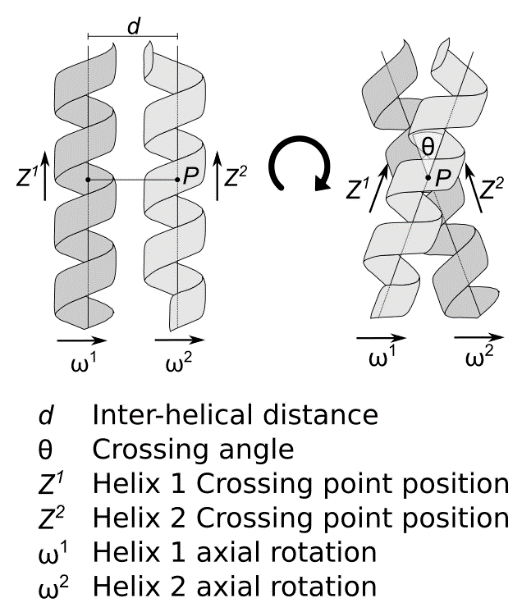
To study the impact of sidechain packing alongside other forces, we can rerun our design algorithm and change how we pick sequences for experiments. I decided on sequences by the calculated van der Waals energy, but a follow up study could instead choose sequences with a considerable amount of hydrogen 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 (Patel et al., 2004; Zhu et al., 2012). Including this term 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, 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 (Ala, Leu, Val, and Ile) alongside other subsets of AAs. Designing sequences with hydrogen bonding (Ser and Thr) or electrostatic (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 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 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. We have primarily used a poly-Leucine backbone in our previous research to assess dimerization of the interface. However, another approach is to assess the dimerization for sequences with the same interface on poly-Leucine, poly-Alanine, and poly-Valine 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 gain insight into how minute changes in packing result in differences in association. Additionally, these experiments may give better insight into studying 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 additionally give us insight into how to better predict the van der Waals packing between the membrane and the protein backbone.

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, a majority 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 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

**Figure 4.2 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 placed at the same geometric value for each term. However, important biological interactions like regulating gene expression are carried about by heterodimeric receptor tyrosine kinases (Del Piccolo et al., 2017). 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 helps to define where the crossing point of the interface is found, with respect to the individual helix (Figure 4.2). 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 are not necessarily symmetric 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. 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 result in different levels of association (Anderson et al., 2017). However, this was only for 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 from our previous research, we can identify input geometries to start designing heterodimeric GASright proteins (**plan to make figure/table for this here**).

A diagram of a dna structure

Description automatically generated

**Figure 4.3 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 4.3). I expand on how I would implement this into my design algorithm in Supp. 4.6.1. 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. 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 (Anderson et al., 2017; Mueller et al., 2014). Any sequences found to associate with a stable energy in CATM can be removed from our pool of sequences that we plan for experiments.

A diagram of a diagram of a diagram

Description automatically generated

**Figure 4.4 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 my 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 4.4). 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/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, 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 heterodimer geometries, we can explore the stability helices at combinations of axial rotations and z-shifts. For homodimers, I assessed the energetics of poly-Leucine sequences with Ala at all interfacial positions 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: Instead of a 1-to-1 search space for axial rotation versus z-shift, all rotations on one helix must be assessed against all rotations on the other helix, and the same must be done for z-shifts. Using the gridding strategy explained in Chapter 3, Figure 3.2, I explored 1200 geometries for each symmetric z-shift and axial rotation. We would now have to explore 12001200. 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. Starting by testing different combinations of the homodimer axial rotations and z-shifts discussed earlier from our homodimers, we can see if small shifts result in favorable energies. **(going to rewrite this part after I do analysis mentioned on page 8)**

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**). To further investigate the impact of packing on these changes, we can calculate how individual AAs might affect those A graph of different types of lines

Description automatically generated with medium confidencearound them.

**Figure 4.5 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. Separated by year that data was extracted. B) Frequency of the pairs found in at least 10 designed sequences compared with the largest differences between designs and all TMs. Most pairs are found more often in TMs (more negative). Pairs found to be all regions marked in red, and pairs found in Left and Right marked in blue.

I recalculated the composition of AAs for all non-redundant TMHs in OPM as of May 14, 2024. The composition is quite similar to when determined in 2021 (Figure 4.5A). 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 4.5B). Although we were able to successfully design interfacial sequences that reflect the frequency of AAs in TMs (Supp. 3….), these are pairs of AAs that are most different from the frequency in all TMHs. 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, beginning to identify relationships between AA frequency by position (AA3 = alanine, alanine separated by 3 bases) could also benefit designing sequences that maintain atomic interactions found within natural MPs.

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.

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 (**reference westerns from paper**). 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.

4.5 Supplementary Details

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 template and the other as designed
   * Add an additional option for the template 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**

4.6 References

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>

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>

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>

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>

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>