# Investigating the effect of sidechain packing on membrane protein association

By

**Gilbert Jamilla Loiseau**

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

(Biochemistry)

at the

UNIVERSITY OF WISCONSIN-MADISON

2024

Date of final oral examination: August 29, 2024

The dissertation is approved by the following members of the Final Oral Committee:

Alessandro Senes, Associate Professor, Biochemistry
Helen Blackwell, Professor, Chemistry
Samuel Butcher, Professor, Biochemistry
Philip Romero, Assistant Professor, Biomedical Engineering, Duke
Baron Chanda, Associate Professor, Anesthesiology, Washington University

## Table of Contents

**Acknowledgements**

I've been extremely fortunate in my graduate career to have the types of support that have enhanced my ability to learn at the highest level. This PhD hasn't been easy. But even on the roughest days I've been able to find myself smiling because of the support that I've had around me.

First, I need to thank my advisor Alessandro Senes. Alessandro has taught me what it means to be a scientist, an expert in learning, a PhD. To remember to focus on the learning in every aspect, from teaching me how to present research to guiding me to figure out the questions I need to ask next. He's been encouraging, honest, and extremely understanding throughout my entire time in graduate school. I had very little coding experience when I joined his lab, but he was always patient with me, giving me ample time and opportunities to learn. Whenever data looked confusing or didn't follow what we expected, he emphasized that being able to discern what we learn matters the most. That the data will tell a story, and there's no issue if it's not the most exciting or groundbreaking knowledge because we ultimately learned something new. Thank you, Alessandro, I'll be forever grateful for your mentorship!

I've had the fortune to have 1-on-1 relationships with each of my graduate committee advisors. Helen Blackwell has repeatedly asked inciteful questions that helped me hone-in on weaknesses in explaining my research. Phil Romero has given advice and encouragement for many parts of my work, helping to keep me excited about the research in front of me. Baron Chanda has kept me grounded in membrane protein biology, having multiple exciting chats where we brainstorm ways to analyze data or think of potential experimental improvements and pitfalls. Sam Butcher has always been supportive and transparent, encouraging me whenever we've discussed the joys and difficulties of grad school and being a professor. Thank you all for giving me confidence in my research and willingly offering advice for after my PhD! Although I'm not sure what I'm doing next, I'm excited to be able to share it with you in the future.

The supportive community that Alessandro has fostered in the Senes lab is a major reason why I've been able to finish this PhD. From post-docs Beth Caselle and Kai Cai being willing to chat with me about so many things I knew nothing about in my first year (living in Madison, how NMR works, membrane protein folding basics), to all the graduate students that I've had the fortune to overlap with. Samuel Craven, my first bay mate and the one who most helped my critical reading skills. Samson Condon, who taught me the beauty of Vim and inspired me to learn computer skills after initially knowing next to nothing. Samantha Anderson, who was and still is always willing to chat back and forth about our troubles while having fun conversations about perspectives and life. Gladys Díaz-Vázquez showed me the beauty of learning at the highest level and inspired me to think harder about my research, becoming one of the main driving forces for me passing my 2$^{nd}$ prelim. Joshua Choi and I have worked closely on our projects, and he's always been willing to discuss experiments and code. Samridhi Garg and I have shared our countless scientific difficulties and issues, and we've continually encouraged each other to take time outside of the stresses of lab. Tamalika Kar's consistent positive attitude has invigorated me and helped me reflect on why I first came to grad school. Thank you all for teaching and encouraging me, and for always being willing to chat about the fun things (Magic, baking, food, culture, music, video games)!

I've been extremely fortunate to be a part of a few scientific communities in grad school. Thank you to Science and Medicine Graduate Research Scholars (SciMed) for funding and for creating an engaging environment for minorities in science. Meeting you all and vibing over culture and our shared experiences was always a joy! A huge shout out to the SciMed coordinators Sara Patterson, Beth Meyerand, Abbey Thompson, and Michelle Parmenter. Anytime I felt I lacked support, they were always willing to chat and talk through ideas to implement something in SciMed that would be helpful for us students. The Chemical Biology Interface (CBI) training program supported me both financially and scientifically throughout my tenure. Many thanks to Helen Blackwell, Cara Jenkins, and all the CBI trainees! The supportive environment helped me work on communicating my research, and I thoroughly enjoyed everyone's

willingness to chat and laugh about the ironic comedies of grad school. I was also a part of IPiB's Diversity, Equity, and Inclusion (DEI) committee, and I reveled in chatting about our own experiences as we discussed ways to make IPiB more welcoming! Thank you to Christina Hull, Mike Cox, Christine Hustmyer, Chase Freschlin, and Bianca Chavez for your camaraderie in the struggle!

A special shout out to the roommates and day ones who stuck around and supported me throughout. Peter Luong, Abel Ingle, Matthew Blackburn, Wojtek Delewski, Akshay Kholi, and Nithesh Chandrasekharan, from the movie nights and dinner parties to the board games and indoor hoops, I've experienced many core memories with you all. I felt revitalized just by being around you and bonding over our shared turmoil, yet still finding so many ways to have fun. Thanks for helping to keep the grad school journey light and upbeat. Excited to spend more time in the future!

A special thanks for the friends and family who helped me put this thesis together with their many helpful critiques and edits. Thank you to my dad, Brandon Phan, Cecilia Nguyen, Jerry Yu, Joshua Choi, Tram-Anh Nguyen, Samridhi Garg, and Wojtek Delewski for helping me to communicate my research thoughtfully and clearly while cleaning up my many grammatical errors!

My therapists have helped me find myself again during times of intense stress and struggle, giving me ways to abate my stress while encouraging me to think deeper about my emotions. I wouldn't have made it through grad school without those thoughtful conversations to keep me grounded. Thank you, therapy!

My family has been nothing but happy for me making the decision to go to graduate school. My dad inspired me to continue, despite the many times I wanted to leave. My mom supported me through my droughts without eating, sending copious amounts of food my way so that I wouldn't go hungry. My brother has consistently given me an outlet for my stress by being ready to play video games, meet up in Chicago, or chat about interesting business ideas for world improvement. Thanks for being there for me

and for being understanding throughout this entire journey! Excited to finally spend more intentional time :D!

And a final thanks to my cat, Jada, who's given me ample emotional support when I've needed it most.

## Abbreviations

**AA**: amino acid, **AMP**: ampicillin

**CAT**: chloramphenicol acetyl transferase, **CATM**: C-alpha transmembrane, **CTX**: cholera toxin promoter

**FACS**: Fluorescence-activated cell sorting, **FRET**: Förster resonance energy transfer

**GFP**: green fluorescent protein, **GpA**: glycophorin A

**KAN**: kanamycin

**MBP**: maltose binding protein, **MC**: Monte Carlo, **MD**: molecular dynamics, **MP**: membrane protein, **MSA**: multiple sequence alignment, **MSL**: molecular software library

**NGS**: next-generation sequencing

**OPM**: orientations of proteins in membranes

**PBS**: phosphate buffer saline, **PCR**: polymerase chain reaction, **PDB**: Protein Data Bank

**SASA**: solvent accessible surface area, **SDS-PAGE**: sodium dodecyl sulfate polyacrylamide gel electrophoresis, **SCMF**: self-consistent mean field, **SE-AUC**: sedimentation equilibrium analytical ultracentrifugation

**TM**: transmembrane, **TMH**: transmembrane helices, **tRFP**: TagRFP-t

**vdW**: van der Waals

**WT**: wild-type

# Chapter 1: Introduction

## 1.1 Introduction to membrane proteins

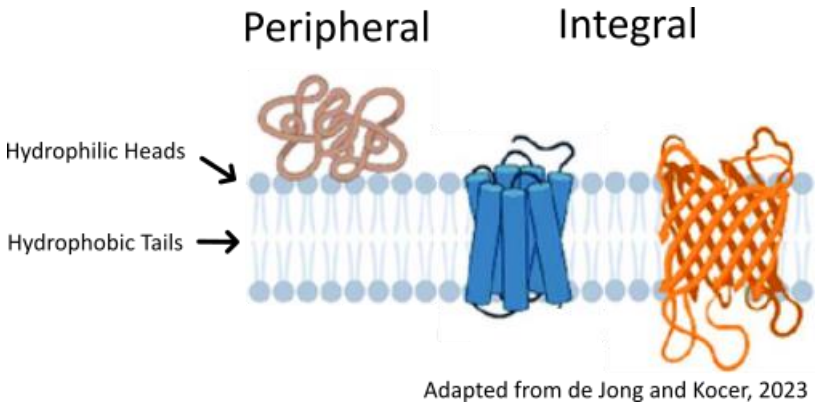



**Figure 1.1 Types of Membrane Proteins.** Membrane proteins associate with a bilayer composed of hydrophilic heads and hydrophobic tails. Peripheral proteins are found at the interface of the bilayer and solute, while integral proteins are embedded into the membrane.

The cell membrane is a bilayer that separates internal cellular components from the outside environment. The membrane bilayer is composed of phospholipids, amphipathic molecules made of two distinct components: hydrophilic (water-loving) heads and hydrophobic (water-fearing) tails. To form the bilayer, hydrophobic tails are sandwiched between hydrophilic heads exposed to the soluble cell cytoplasm and the outer environment. Despite this separation, communication outside of the cell is critical to sense external stimuli and maintain cell survival. This process is regulated by a class of proteins tethered to the membrane: membrane proteins (MPs).

MPs can be broken down into two groups: peripheral MPs and integral MPs (Fig. 1.1). Peripheral MPs are composed of both hydrophilic and hydrophobic components, allowing them to localize to the edges of the cell membrane while still exposed to the soluble environment. Unlike peripheral MPs, integral MPs are primarily hydrophobic; they are embedded within the membrane with minimal outside exposure. Integral MPs are made up of multiple structural subunits, such as β-sheets and α-helices. β-sheets often form open pores through the membrane, functioning as channels and transporters that allow ions and molecules to enter or exit the cell through the bilayer. Conversely, transmembrane (TM) α-helices are crammed into the membrane between lipids and assembled into complex multi-domain structures. Multiple TM α-helices

(TMH) can associate in response to environmental stimuli, signaling activation and deactivation of the appropriate genes.

The sequestering of hydrophobic tails into the center of the membrane yields a hydrophobic environment; the core of the bilayer is devoid of polar interactions, which are responsible for the hydrophobic effect that drives soluble protein folding (Tanford, 1980; Yang et al., 1992). For MPs to fold within this hydrophobic environment, they must strike a delicate balance of interactions while surrounded by lipids. Mutations within MPs can drastically affect these interactions, preventing them from folding properly. Misfolding of TMHs has been implicated in several human diseases such as Parkinson's, cystic fibrosis, and cancer (Gregersen et al., 2006; Sanders & Myers, 2004). To fully understand how to combat the progression of these diseases, it is necessary to understand the impact of the individual forces that govern the folding process; however, studying MPs is inherently a difficult challenge. MPs are difficult to express in high yields for biophysical experiments, and purification and solubilization of these proteins often lead to aggregation and unfolding (Carpenter et al., 2008). Alternatively, researchers have focused on using model systems of MPs to better understand folding.

My research focused on using a model TMH system to investigate the extent that van der Waals (vdW) packing can act as a driving force for MP folding and association. I developed an *in silico* protein design algorithm to study the association of single-pass TMH homodimers complemented with high-throughput experiments to validate my computational models. Before detailing my research and experiments, I review the contributions made to understanding driving forces in MP folding and association. I detail the forces and interactions involved in MP folding, while also highlighting the sequence and structural motif $GAS_{right}$, an important control for my research used to juxtapose differences between association by forces other than vdW packing. I then review the tools that have been used to study TMH association and folding before emphasizing the deficiency of research on the contribution of vdW packing.

## 1.2 The two-stage model of MP folding

Early MP research focused on identifying membrane embedded regions within proteins using hydrophobicity analysis: navigating through the protein amino acid (AA) sequence and scanning for stretches of hydrophobic AAs (Kyte & Doolittle, 1982). Hydrophobicity analysis was successful in predicting the helices in both bacterial photosynthetic reaction centers and bacteriorhodopsin (Engelman et al., 1982; Michel et al., 1986). This method was further developed to determine a charge bias known as the positive-inside rule, where charged AAs are likely to be found outside of the membrane (von Heijne, 1992). Multiple tools are now available, allowing researchers to easily identify TM regions from protein sequences (Wilkins et al., 1999).

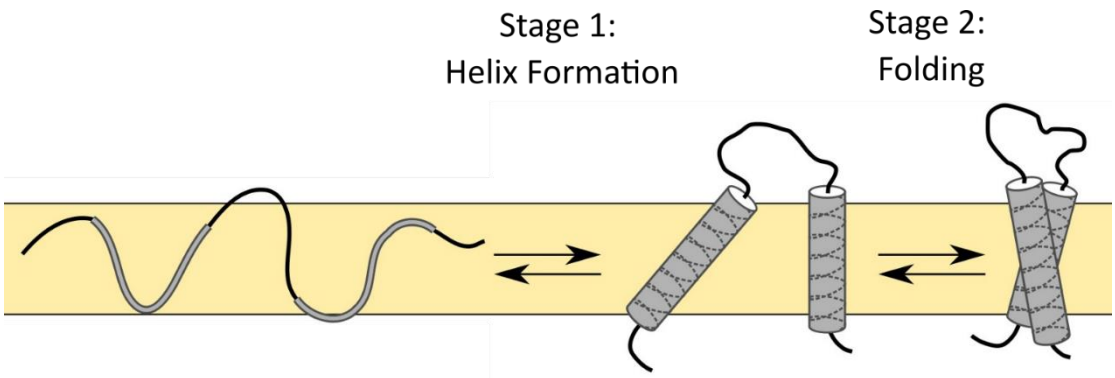



**Figure 1.2 The two-stage model.** In stage 1, TM helices begin to form while the protein is inserted into the membrane. In stage 2, helices oligomerize and assemble into a fully folded protein as a result of thermodynamic interactions which include hydrogen bonding, electrostatics, and vdW packing.

In 1990, Popot and Engelman proposed the two-stage model for MP folding (Fig. 1.2): As the protein is threaded into the membrane, TMHs begin to form (stage 1) prior to stabilizing into a fully folded protein (stage 2). TMHs first reach a thermodynamic equilibrium with the lipid environment before undergoing stage 2, where individual TMHs oligomerize and assemble into the folded protein (Popot & Engelman, 1990, 2000). While stage 1 is driven by the hydrophobic effect to coordinate insertion of hydrophobic protein sequences into the membrane, stage 2 is governed by interactions between individual TM domains. Research on bacteriorhodopsin gives credence to stage 2: denaturing two separate fragments of the protein and resuspending them in lipid vesicles results in an active, folded protein (Popot et al., 1987);

two chemically synthesized TMHs of the protein were reconstituted in lipid vesicles with a larger fragment, resulting in the reformation of the bacteriorhodopsin shown by X-ray crystallography (Kahn & Engelman, 1992); lastly, extraction and reconstitution of individual helices of the protein were found to yield activity (Marti, 1998). Additional research on large protein complexes pushed the field forward, showing that mutating the hydrophobic core of four-helix-bundle protein Rop and five-helix-bundle protein phospholamban decreases the stability of both proteins (Arkin et al., 1994; Munson et al., 1996). With reassembly of MPs being an effective model for studying MP folding, other groups continued to build on this research by exploring model systems of TMHs to determine how minute changes in sequence and structure influence stability.

In the two-stage model, there are unique forces involved in each stage of folding. In the first stage, insertion of proteins into the membrane is driven by the hydrophobic effect, as MPs are more stable in the membrane than in the soluble environment. When MPs are being translated by the ribosome, a signal sequence on the protein directs translation to the translocon (Dalbey et al., 2011). Together, the translocon and ribosome individually thread hydrophobic segments of the protein into the membrane (Hessa et al., 2005; Rapoport, 2007). In the subsequent stage of folding, vdW packing, electrostatics, hydrogen bonding, and weak polar interactions between individual TM domains contribute to guiding the MP to the folded state.

Near the turn of the century, MP studies advanced our understanding of MP folding beyond the simplicity of the two-stage model. A third stage was considered, which accounted for the thermodynamic impact of ligand binding domains, folding of loops outside of the membrane, and inserting other domains into the bilayer (Engelman et al., 2003). Rather than focusing on how bulk changes in forces impacts folding, researchers began to characterize the impact of individual AAs and the respective forces that drive MP folding.

## 1.3 Methods to study transmembrane helix oligomerization

To investigate MP folding, researchers have developed tools to study the oligomerization of TMHs. These tools strive to identify changes in stability between the unfolded and folded states, allowing researchers to uncover MP folding thermodynamics within a variety of systems. The oligomerization process is essential for influencing cell gene expression, including epidermal growth factor receptors and proteins involved in tyrosine kinase signaling cascades (Kumari & Yadav, 2019). Furthermore, this thermodynamic information can be used to assess and validate computational models for designing and engineering novel proteins. In this section, I detail the tools and techniques that have been implemented to further understand the driving forces in MPs.

### 1.3.1 *In vitro* techniques

Early tools used to study MP folding monitored the reversible folding of MPs. *In vitro* techniques focus on expressing and solubilizing proteins into suitable membrane mimetics, such as detergents. An initial method studied the thermodynamics of TMH association by observing differences in mobility within SDS-PAGE gels. SDS-PAGE was used to tease the effect of point mutations in a variety of TM sequences, including GpA (Choma et al., 2000; Lemmon, Flanagan, Hunt, et al., 1992; Lemmon, Flanagan, Treutlein, et al., 1992; Zhou et al., 2000). Sedimentation equilibrium analytical ultracentrifugation (SE-AUC) is another technique that explores different folding states by varying the concentrations of detergents used to solubilize the protein. These samples are centrifuged at high speeds, resulting in a concentration gradient that is then analyzed to quantitatively ascertain the transition of the protein at different folding states. SE-AUC has been used to investigate mutations of GpA to better understand its thermodynamics of association (Doura & Fleming, 2004; Fleming et al., 1997; Fleming & Engelman, 2001). These techniques developed our understanding of the thermodynamics of TMH association. However, they were low throughput and limited to studying MPs solubilized in detergents.

Multiple *in vitro* techniques were developed to observe MP dynamics within membrane like environments. Disulfide cross-linking was used to measure TMH oligomerization in both micelles and lipid vesicles to investigate the interfaces of a variety of proteins (Cristian et al., 2003; Hastrup et al., 2001; Kovalenko et al., 2005; Lu et al., 2010). Pulse proteolysis quantitatively measures the thermodynamic stability of MPs by selectively denaturing and subsequently digesting the unfolded MP (Park & Marqusee, 2005). Using this technique to study bacteriorhodopsin folding uncovered that folding was dependent on changing concentrations of mixed micelles, which was not determined previously (Schlebach et al., 2012; Schlebach et al., 2011). Steric trapping utilizes the streptavidin-biotin binding system to measure the binding affinity of associating TMs in lipid bilayers, and it has been used to further determine the folding energy landscapes of GpA and mutants affecting its association (Blois et al., 2009; Hong & Bowie, 2011; Hong et al., 2013; Howarth et al., 2006; Huang et al., 2022). Compared to previous research, these methods allowed for studying MP folding thermodynamics in lipids and mixed micelles, closer to the environments of the cell membrane. While these techniques approach understanding proteins in native environments, other techniques were developed to further study MPs within cells at higher throughput.

### 1.3.2 *In vivo* assays

*In vivo* assays have been utilized to investigate the folding and association of MPs in their natural environment. Double mutant cycle analysis, a method that quantitatively measures interaction in protein structures, mutates two non-interacting residues within a protein to assess the impact of coupled residues on thermodynamic stability (Carter et al., 1984). Double mutant cycles in bacterial two-hybrid and protein complementation assays allow researchers to determine the strength of protein-protein interactions by changes in cell growth due to mutation, which can be monitored in high-throughput (Horovitz et al., 2019; Salinas & Ranganathan, 2018; Tarassov et al., 2008). Genetic reporter assays allow cells to express MPs of interest fused to a DNA binding domain that can either inhibit or promote transcription of a reporter gene. GALLEX is a two-hybrid system where TMs are fused to DNA binding domain LexA. Association of the TMs

inhibits the β-galactosidase gene (Schneider & Engelman, 2003). Other reporter assay systems have utilized a chimera of the MPs of interest fused to ToxR, a dimeric transcription factor, to promote gene expression (Gurezka & Langosch, 2001; Russ & Engelman, 1999).

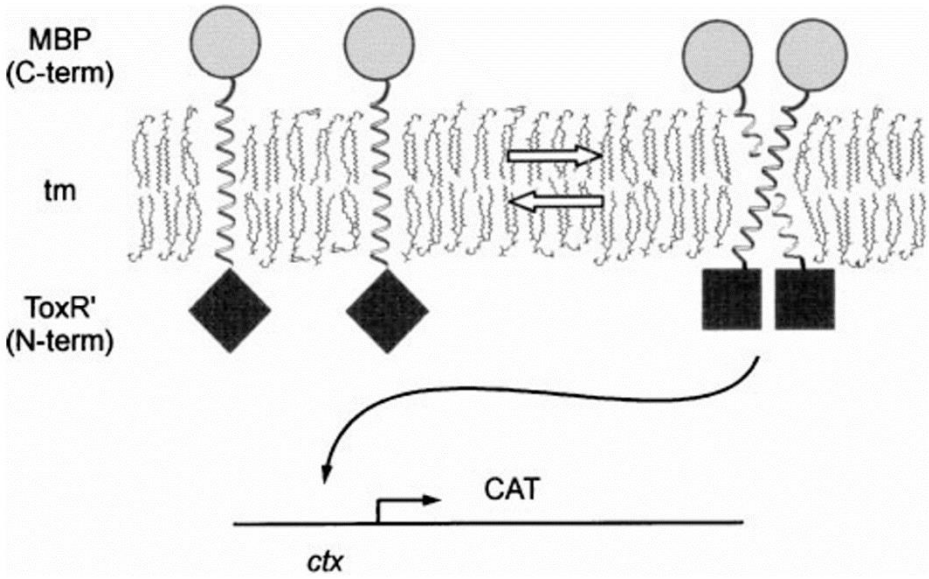



**Figure 1.3 TOXCAT.** TOXCAT is an experimental assay that has been used to study the self-association of helices. The TM of interest is expressed bound to maltose binding protein (MBP) and ToxR, a dimeric transcription factor originally found in *V. cholerae*. When the TM associates, ToxR does as well, resulting in the expression of a gene (CAT) that can be measured to determine the strength of the association.

TOXCAT has been used to study TM helix-helix interactions, where the TM of interest is fused to dimeric transcription factor ToxR (Fig. 1.3). When the TMs associate, ToxR dimerizes and promotes the expression of chloramphenicol acetyltransferase (CAT) which is measured to determine the strength of association. TOXCAT demonstrated that mutations of polar residues on GpA in the native membrane environment yield different results than the previous *in vitro* studies (Russ & Engelman, 1999; Zhou et al., 2000; Zhou et al., 2001). Johnson et al. expanded on these findings, suggesting that electrostatic interactions between charged and aromatic AAs facilitates oligomerization (Johnson et al., 2007). TMH association as measured by TOXCAT correlates to changes in the free energy of association of GpA and point mutations (Duong et al., 2007). Anderson et al. used TOXCAT to study the association of GpA and similar TMHs, suggesting that these proteins associate via a combination of hydrogen bonding and vdW

interactions (Anderson et al., 2017). TOXCAT is a well-studied system for probing TMH association, determining the impact of individual AAs and their respective forces on the thermodynamics of association.

Recently, TOXCAT has been adapted into the high-throughput assay TOXGREEN. The reporter gene CAT was replaced with green fluorescent protein (GFP), allowing fluorescent readings to be used to assess the association levels of the TMs of interest and their corresponding mutants (Armstrong & Senes, 2016). TOXGREEN can be used in high-throughput applications such as fluorescence activated cell sorting (FACS), where a library of TMs is expressed, sorted, and sequenced through next generation sequencing (NGS). The sequencing data can then be quantified to determine the relative association propensities for each protein present in the library (Anderson, 2019).

### 1.4 Computational methods to study MP structure

In conjunction with experimental methods to study MP folding, computational methods have been designed by evaluating previously solved MP structures. These methods look to further understand MP folding by establishing energetic terms that estimate the thermodynamics of MP folding. In this section, I review computational methods used to predict MP structures, highlighting unique features of each tool.

#### 1.4.1 Rosetta

Rosetta houses a variety of energy functions and prediction tools for soluble environments, including the ability to dock or design proteins *de novo* (Chaudhury et al., 2011; Koehler Leman et al., 2017; Leman et al., 2020; Weitzner et al., 2017; Yarov-Yarovoy, Schonbrun, et al., 2006). These energetics include but are not limited to terms parameterized by CHARMM (vdW and electrostatics), a hydrogen bond and disulfide function curated from polar contacts found in ~8000 high-resolution crystal structures, and a side chain conformation energy based on the probability of occurrence from the Dunbrack rotamer database (Alford et al., 2017; RICHARDSON et al.). These Rosetta energy functions have been adapted to accommodate predicting helical TMs within the membrane environment. The updated functions include an energy term that separates the membrane into layers designating atoms as water-exposed, polar, interface, or hydrophobic (White & Wimley, 1999; Yarov-Yarovoy, Schonbrun, et al., 2006). Successful predictions helped discover structural details in MPs associated with voltage sensing and gating mechanisms (Vargas et al., 2012; Yarov-Yarovoy, Baker, et al., 2006).

Recently, RosettaMP was devised to enhance the functionality of MP prediction. With RosettaMP, TM helices are modeled *de novo* from sequence, the membrane bilayer is represented, and MP assembly is simulated (Koehler Leman et al., 2017). These tools increase accessibility to MP structures, improving the ability to visualize and predict structures of MPs that have not yet been solved, as well as enhancing MP design (Duran & Meiler, 2018). Simultaneously, energetic predictions permit researchers to analyze structural mutations *in silico* prior to testing with *in vitro* or *in vivo* experiments (Thieker et al., 2022).

### 1.4.2 Molecular Software Library

Another computational tool for modeling protein structures is the Molecular Software Library, or MSL (Kulp et al., 2012). Similar to Rosetta, MSL contains a variety of tools to perform MP structure prediction. These tools include the ability to transform proteins in space, mutate specific residues, extract geometric information from known structures, and predict the structure and energetics of an input sequence. Additionally, MSL was used to develop the CATM structure prediction algorithm (Fig. 1.4).

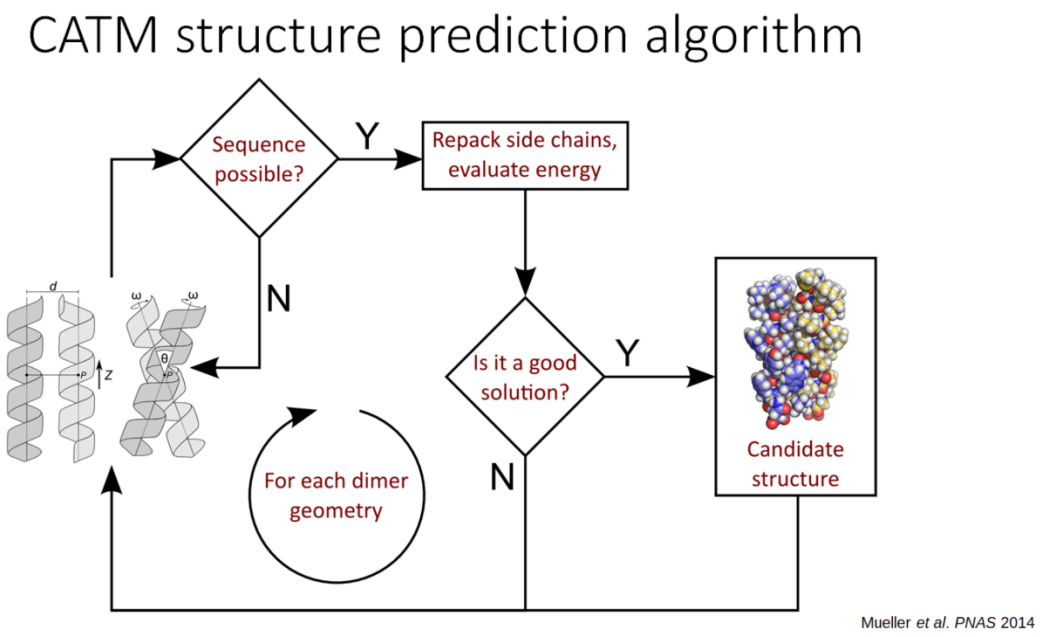



**Figure 1.4 CATM.** Using a refined set of dimer geometries for $GAS_{right}$ proteins, the CATM algorithm predicts the stability and structure of $GAS_{right}$ dimers from an input sequence.

Briefly, the CATM algorithm predicts structures of known and unknown TM dimers that associate by the $GAS_{right}$ motif, and experimental studies have shown that it accurately estimates the energetics of association (Anderson et al., 2017; Díaz Vázquez et al., 2023; Mueller et al., 2014). CATM uses the Energy-Based conformer library applied at the 95% level for side chain mobility (Subramaniam & Senes, 2012). Energetics of predicted proteins are determined in CATM using the CHARMM 22 vdW function (MacKerell et al., 1998), the IMM1 membrane implicit solvation model (Lazaridis, 2003), and the hydrogen bonding function SCWRL4 (Krivov et al., 2009). Each of these energy terms is applied to optimize the dimer

geometry by Monte Carlo (MC) backbone perturbation cycles where all parameters (xShift, zShift, axialRotation, and crossingAngle) are locally varied. The association energy is calculated as the energy of the dimer minus the energy of two monomers:

$$Eq.\ 1.1 \quad Dimer/Monomer\ Energy = CHARMM22 + IMM1 + SCWRL4$$

$$Eq.\ 1.2 \quad Energy\ Score = Dimer - (2 \times Monomer)$$

My research adapts the CATM algorithm to design structures with strong vdW packing in the absence of hydrogen bonding, allowing me to assess the extent at which packing can drive MP association. I further detail how I used MSL to design TM homodimers in chapter 3.

### 1.4.3 Topology prediction and docking Algorithms

| Tool | Source | Key Features |
|---|---|---|
| OCTOPUS | octopus.cbr.su.se | Models membrane dipping-regions and TM hairpins |
| HADDOCK | rascar.science.uu.nl/haddock2.4 | Allows input of interaction restraints to drive the docking process |
| PREDDIMER | model.nmr.ru/preddimer | Includes ranking and filtering of predicted structures and representative visuals of the interface |
| EVFOLD | evcouplings.org | Evolutionary constraints derived from multiple sequence alignments |
| TOPCONS | topcons.net | Separates signal peptides from TM regions |
| TMDOCK | membranome.org/tmdock | Calculates thermodynamic stabilities of insertion and association |

**Table 1.1** Docking tools, the websites they can be accessed at, and the key features for each tool.

Other methods used to determine interactions between MPs focus on predicting the topology or docking of TMHs (Table 1.1). OCTOPUS predicts TM topology using a combination of Markov models and neural networks (Viklund & Elofsson, 2008). HADDOCK can apply experimental knowledge of the interface region between proteins to refine docking (de Vries et al., 2010; Dominguez et al., 2003). PREDDIMER utilizes a novel surface-based modeling approach to predict and screen TM dimers for conformation heterogeneity (Polyansky et al., 2012). EVFold employs evolutionary-based, structural restraints to refine

their docked structures (Braun T et al., 2015). TOPCONS can identify signal peptides separate from TM regions, and displays homology to known structures as well as a predicted ΔG of insertion (Tsirigos et al., 2015). TMDOCK applies an all-atom model for helices, inserting them in the membrane and outputting a structure alongside a predicted ΔG of insertion and ΔG of association (Lomize & Pogozheva, 2017). Each of these methods is available online, where users can input the sequence and additional information to guide the process.

### 1.4.4 Molecular dynamics simulations

Molecular dynamics (MD) is a computationally intensive approach to predicting MP structures. This technique aims to simulate biological interactions using different representations of the bilayer and protein structures. Representations range in varying degrees of complexity, from simplified systems that reduce resolution by combining atoms into large single-body molecules (coarse-grained) to detailed including different types of lipids and other molecules present (atomistic). The chosen representation and force field can increase the time necessary for creating these nano- to microsecond timescale simulations of these interactions (Goossens & De Winter, 2018). Using force fields such as Charmm (Mackerell Jr. et al., 2004), Gromos (Soares et al., 2005), Amber (Wang et al., 2004), and MARTINI (Marrink et al., 2007), groups have used MD to investigate membrane interactions for many proteins including the potassium channel KcsA, rhodopsin, and GpA (Bond & Sansom, 2006; Bu et al., 2007; Deol et al., 2006; Grossfield et al., 2006; Mottamal et al., 2006).

### 1.4.5 AlphaFold and RoseTTaFold

At the 2020 Critical Assessment of Structure Prediction (CASP) conference, Google's DeepMind introduced the machine learning model AlphaFold. Unlike previously mentioned prediction algorithms, AlphaFold predicts structures without energetics. AlphaFold utilizes a combination of neural networks, training on multiple sequence alignments (MSAs) and solved protein structures to predict unknown structures to near atomic precision with a 95% confidence interval (Jumper et al., 2021). Shortly afterward,

David Baker's group introduced RoseTTAFold, improving on the Rosetta prediction by incorporating a similar architecture to AlphaFold, with the inclusion of a third track network that connects sequence, residue-residue distances, and atomic coordinates (Baek et al., 2021). Each of these methods drastically improved the ability to predict unknown protein structures using information from previously studied and solved proteins. With increasing interest in using these technologies, multiple free online tools have been established to enhance access to these advanced protein prediction algorithms (Mirdita et al., 2022; Roberts et al., 2024). However, these machine learning algorithms are limited by the amount of information available. AlphaFold struggles to predict proteins with <30 homologs in their MSAs, and accuracy decreases for multi-protein interactions, while RoseTTAFold has difficulty predicting higher-order oligomers (Agard et al., 2022). These limitations are amplified in MPs due to the lack of MP structures, making small or complex TM proteins difficult to predict. To better understand the dynamics of association and folding in MPs, it is necessary to advance our knowledge of the forces involved in folding.

## 1.5 Driving forces in MP folding

The elaborate nature of the lipid bilayer makes it difficult to directly study the forces involved in MP folding. As an initial approach, researchers aimed to solve the structures of MPs by identifying structural features necessary for the folded state. However, solving MP structures is an inherently difficult task due to the need to express and solubilize MPs for experiments (Carpenter et al., 2008). Alternative approaches to study folding utilized a combination of *in vitro* and *in vivo* experimental tools to determine the rules that govern TM folding. Folding of integral TM proteins involves a variety of energetic constraints resulting from the hydrophobic nature of the phospholipid bilayer. The translocon complex assists during translation, inserting TM domains into the membrane (White & von Heijne, 2004). TMs are composed of amide nitrogens and carbonyl oxygens within the protein backbone, atoms prone to forming hydrogen bonds. However, inserting hydrogen bonds into the hydrophobic core of the bilayer carries an energetic penalty (Marinko et al., 2019; Popot & Engelman, 1990, 2000). To satisfy the lack of hydrogen bonding within the membrane, TMs adopt standard α-helical and β-sheet structures where hydrogen bonds form along the protein backbone. Experimental tools have been developed to tease out folding interactions after insertion by using model MP systems. In this section, I will summarize advances in understanding driving forces in MP folding, with a particular focus on using single-pass TMHs.

### 1.5.1 Hydrogen bonding and polar interactions

Hydrogen bonding plays a key role in regulating MP structure and function, and many mutations on polar residues have been found to promote disease states (Choi et al., 2004; Partridge et al., 2002, 2004; Therien et al., 2001; Wehbi et al., 2008). Research characterizing the impact of polar residues on TMH association suggests that hydrogen bonding and polar interactions can drive TMH association. Using a wild-type like sequence of the GCN4 leucine zipper, a mutation from Asn to Val was found to decrease association on SDS-PAGE (Choma et al., 2000). Synthetic model poly-leucine peptides based on GCN4 were made with three different compositions of AAs at the interface, two being completely hydrophobic and

the other hydrophobic with a single Asn. When tested for their ability to associate on SDS-PAGE, only the sequence with Asn was found to have equal amounts of monomers and dimers (Zhou et al., 2000). These results suggest that Asn plays a role in driving TMH association.

Further research began to characterize the impact of other polar AAs in TMH association systems. Poly-leucine based peptides were made with single AA mutations to a variety of polar residues and tested using the *in vivo* experimental assay TOXCAT. Their results showed that larger polar residues (Asn, Asp, Gln, and Glu) capable of being both hydrogen bond donors and acceptors drive association more than poly-leucine alone (Zhou et al., 2001). A similar study showed that replacing hydrophobic AAs with large polar AAs on the GCN4 peptide resulted in association with higher stabilities (Gratkowski et al., 2001). These studies suggest that large polar AAs drive association. However, large polar AAs are not often found in MP sequences, whereas small polar AAs Thr and Ser are more common due to their ability to more readily form hydrogen bonds with backbone carbonyls on the same helix (Gray & Matthews, 1984; Liu et al., 2002).

The peptides used in the previous studies were made of bulky hydrophobic AAs, possibly preventing Ser and Thr from playing the roles in association that they do in naturally occurring sequences. Using TOXCAT, a library of TM sequences that mutated the interface of known dimer glycophorin A (GpA) were screened for their ability to associate. A majority of the proteins found to associate were composed of Thr and Ser at the interface, suggesting that these AAs are important for association (Russ & Engelman, 2000). Additional investigation into the structure of GpA suggests that Thr 87 forms interhelical hydrogen bonds at the interface, supporting previous research that mutations at this residue disrupt dimerization (Lemmon, Flanagan, Hunt, et al., 1992; Lemmon, Flanagan, Treutlein, et al., 1992; MacKenzie et al., 1997; Smith et al., 2002). Alongside earlier research, this data suggests that hydrogen bonding is a driving force that strongly stabilizes TMH association.

By observing MP structures, researchers have been able to identify and characterize hydrogen bonds between TM helices in multiple solved structures (Adamian & Liang, 2002; Freiberg et al., 2012; MacKenzie et al., 1997; White, 2005). Using double mutant cycle analysis, MPs with multiple TMHs were mutated to determine the contribution of interhelical hydrogen bonding to MP stability. The average contribution for hydrogen bonding in multiple proteins was found to be 0.5kcal/mol +/- 0.7 (Bowie, 2011). Using an SDS unfolding assay, the average contribution of eight hydrogen bonds was found to be 0.6 kcal/mol (Joh et al., 2008). Despite the relatively small contribution in larger protein, hydrogen bond energies estimated in vacuum have been calculated to be around ten times higher (Ben-Tal et al., 1997; Mitchell & Price, 1990; Rose & Wolfenden, 1993; Tsemekhman et al., 2007). Single mutants on a variety of MPs were also tested, determining on average that hydrogen bonding contributes similar stability to water soluble proteins (Bowie, 2011).

Hydrogen bonding and polar interactions are stabilizing forces in MP folding and association. In larger MP complexes, hydrogen bonding contributes similar stability as found in soluble proteins (Bowie, 2011). However, the hydrophobic nature of the membrane suggests that hydrogen bonding contributes more to stability. Most polar AAs are able to form hydrogen bonds with the backbone of the TM, and this bond must be broken prior to forming a stabilizing interhelical hydrogen bond (Chamberlain & Bowie, 2004). Additionally, mutating polar AAs to nonpolar AAs disrupts association, suggesting that hydrogen bonding drives association of TMHs (Lemmon, Flanagan, Treutlein, et al., 1992; Smith et al., 2002; Zhou et al., 2001).

### 1.5.2 Electrostatics and weak hydrogen bonding

Electrostatics interactions in the membrane can be broken down into two groups: π-π or cation-π. π-π interactions typically occur by burying surface area between aromatic rings, combining vdW and hydrophobic interactions. Cation-π interactions occur from attractive forces between charged AAs (Lys and Arg) and the electron clouds of aromatic AAs (Phe, Tyr, His, and Trp) (Johnson et al., 2007). These

interactions are found in a multitude of channels and G protein-coupled receptors, and are equally important for ligand binding of neurotransmitters, metal ions, and toxins (Infield et al., 2021). Charged AAs are not often found in MPs, but molecular dynamics simulations and potential of mean force calculations supports the thermodynamic stability of Arg in TMs (Ulmschneider et al., 2017). Electrostatic interactions have been studied between a variety of TMH interactions. Johnson et al. mutated a hydrophobic protein with a pair of charged and aromatic AAs. Using TOXCAT, they found that Lys coupled with Tyr, Trp, and Phe is able to drive these proteins to associate (Johnson et al., 2007). Another study looked at the role of aromatic AAs in the β-barrel outer MP OmpA, and using double mutant cycle analysis, found that each side chain contributes more than 1kcal/mol to stability (Hong et al., 2007). Additional SDS-PAGE analysis on helical hairpins demonstrated that TM-TM electrostatic interactions alongside helical turns promote folding (Bañó-Polo et al., 2013).

Similar to hydrogen bonding, electrostatics plays a strong stabilizing role in MP folding and association, able to drive association of TMHs with charged and aromatic interactions. Another force that has been shown to strongly influence association of TMHs are interhelical hydrogen bonds. These hydrogen bonds help to facilitate association and folding by the $GAS_{right}$ motif, one of the most prevalent sequence and structural motifs found in TM proteins (Walters & DeGrado, 2006). GAS is an acronym for the three AAs typically found in the sequence: Gly, Ala, and Ser. These small residues define the interface of the motif (G/A/S)xxx(G/A/S), resulting in a short interhelical distance between two TM helices. The "right" subscript in $GAS_{right}$ comes from an important structural feature in which TM helices associate at a right-handed crossing angle. $GAS_{right}$ proteins are frequently found to be involved in a variety of diseases: syndecan-2 overexpression has been found in colorectal cancer cell lines, neuropilin-1 has been shown to intensify symptoms of SARS-CoV-2, and Glycophorin A (GpA) misregulation is involved in sickle cell disease (Benedicto et al., 2021; Marshall et al., 2024; Vicente et al., 2013). Due to the prevalence of $GAS_{right}$

proteins in medical applications as well as its well-defined sequence and structural features, many groups have studied these proteins to further understand the forces governing TM association.

GpA is a well-studied protein that associates via the $GAS_{right}$ motif. Multiple *in vitro* studies worked to define the interface of GpA, making point mutations along the protein and visualizing the changes in dimerization using SDS-PAGE (Lemmon, Flanagan, Hunt, et al., 1992; Lemmon, Flanagan, Treutlein, et al., 1992). Using dimerization as a model system, researchers aimed to further characterize the thermodynamics of dimerization by monitoring changes in stability between the monomer and dimer state. Additional studies using *in vitro* techniques – sedimentation equilibrium analytical ultracentrifugation (SE-AUC) and Förster resonance energy transfer (FRET) – were able to ascertain differences in stability, effectively quantifying the thermodynamics of association for GpA (Fisher et al., 1999; Fleming et al., 1997).

After the structure of GpA was solved by solution nuclear magnetic resonance (NMR), groups analyzed the structure to further characterize their thermodynamic data (MacKenzie et al., 1997). Analysis of GpA in molecular dynamics simulations coupled with mutations on the NMR structure showed that the alternative association resulted from changes in vdW interactions (Fleming et al., 1997; MacKenzie & Engelman, 1998; MacKenzie et al., 1997; Petrache et al., 2000). However, further investigation into the unique sequence and defined structure of the $GAS_{right}$ motif has been shown to permit an uncommon structural feature. Small AAs at the interface allow TM backbones to associate with a short interhelical distance, resulting in the formation of a network of weak interhelical hydrogen bonds, where donors are Cα carbons and acceptors are carbonyl oxygens on the opposite helix (Cα–H···O=C, or Cα–H bonds) (Senes et al., 2001). Carbon atoms are not commonly considered as hydrogen bond donors because they are less electronegative than typical nitrogen and oxygen donors. However, these carbons are found near electronegative withdrawing groups on the peptide backbone, increasing their electronegativity. Estimates from quantum mechanics calculations suggest that the stabilizing energy of an Cα–H bond may contribute

between one third and half of that of an N—H donor in vacuum (Scheiner et al., 2001; Vargas et al., 2000). Measurements of the stretching frequency of these bonds in GpA suggests that it could contribute 0.9 kcal/mol of stability to the dimer (Arbely & Arkin, 2004).

Further research on $GAS_{right}$ TMs helped define the geometric structure for the network of hydrogen bonds. This research resulted in CATM, an algorithm that successfully predicted the structures of five known homodimer structures (Mueller et al., 2014). The CATM algorithm was used in conjunction with TOXCAT to determine the influence of this network of Cα–H bonds. By predicting the structures of $GAS_{right}$ TMs found in natural sequences and testing their stability using *in vivo* TOXCAT, they showed that structures predicted to have more Cα–H bonds are more thermodynamically stable (Anderson et al., 2017). Additionally, the free energy of association of $GAS_{right}$ structures was measured using *in vitro* FRET, concluding that the thermodynamic stability of $GAS_{right}$ proteins is well correlated with *in vivo* experiments (Díaz Vázquez et al., 2023). These studies suggest that $GAS_{right}$ proteins associate primarily through two forces: weak hydrogen bonding and vdW packing. Using a refined version of the CATM algorithm, I designed sequences to associate solely by vdW packing, leveraging $GAS_{right}$ sequences as controls. By evaluating sequences designed that associate through vdW packing, I was able to differentiate the impact of packing (designed sequences) versus both hydrogen bonding and packing ($GAS_{right}$) on association.

### 1.6 Understanding van der Waals as a driving force

As individual TMHs are threaded into the membrane, an interplay of biophysical forces produces helix-helix association. This process is regulated by an intricate distribution of hydrogen bonding, electrostatic interactions, and vdW forces that govern the stabilities of the unfolded and folded states. Each of these interactions are driven by the types of AAs present within the TM. Removing hydrogen bonds within TMs decreases stability (Duong et al., 2007; Gratkowski et al., 2001; He & Hristova, 2008; Li et al., 2006; Stanley & Fleming, 2007). Hydrogen bonding not only regulates the secondary structure of TMs, but also drives TM helix association when polar AAs Ser and Thr form interhelical hydrogen bonds between opposing helices (Johnson et al., 2007; Zhou et al., 2001). Additionally, electrostatic interactions between positively charged Lys and electronegative aromatic AAs Tyr, Trp, and Phe promote association between helices (Johnson et al., 2007). Other charged and aromatic interactions have been shown to contribute similar stability as in water soluble proteins (Bañó-Polo et al., 2013; Burley & Petsko, 1985; Hong et al., 2009; Hong et al., 2007; Ulmschneider et al., 2017). However, hydrogen bonding and electrostatic interactions only account for a subset of AAs typically present in MPs. The three AAs most frequently found in MPs (Leu, Ile, and Ala) are uncharged and lack the ability to form hydrogen bonds (Liu et al., 2002), thus TMs constituted of these AAs can only be stabilized by vdW forces.

VdW forces occur between atoms within close contact, including the interactions between MPs and the hydrophobic tails within the membrane. MP association motifs have been structurally characterized and studied, determining that tight vdW packing plays an important role in TMH association (Gurezka et al., 1999; Kim et al., 2004; MacKenzie et al., 1997; North et al., 2006; Russ & Engelman, 1999; Wu et al., 2005). Mutational studies on well packed residues in the core of MPs suggest that changes in packing can destabilize protein structure (Ash et al., 2004; Faham et al., 2004; Joh et al., 2009; Mravic et al., 2019; Yano

et al., 2002). However, accounting for vdW between MPs and phospholipids is complex, and not many studies have successfully investigated the influence of vdW forces on MP stability.

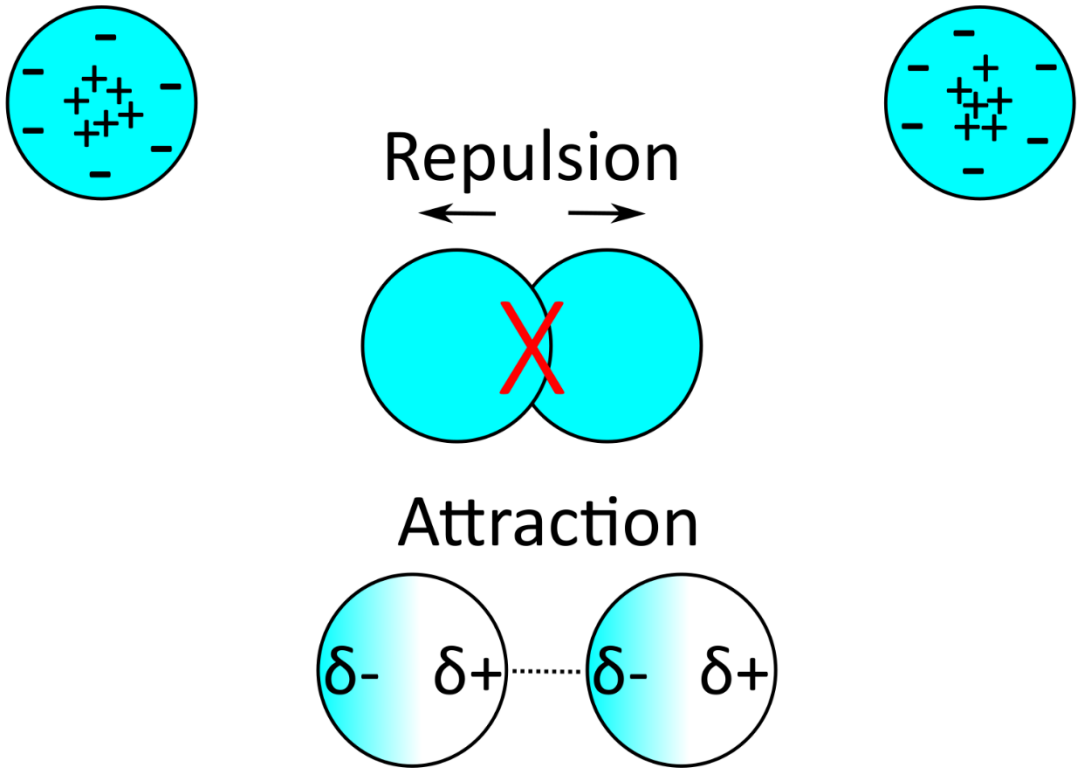



**Figure 1.5 VdW force.** VdW forces between two atoms are non-existent at long range distances. When too close, vdW is a repulsive force. However, vdW becomes an attractive force as atoms approach a minimum distance, or the vdW radius.

The physical properties of atoms are the foundation of intermolecular interactions. Atoms are composed of a nucleus of protons surrounded by an outer electron shell. The electron shell expands a finite distance away from the nucleus, constructing the space occupied by the atom, or the vdW radius (Batsanov, 2001). When atoms are found at a distance smaller than their combined vdW radii, the opposing electron shells repulse, pushing the atoms away. However, atoms and molecules undergo natural dipole moments where electrons are distributed unevenly, resulting in a slight positive and negative charge (Fig. 1.5). These dipoles result in a weak attraction between protons in the nucleus of one atom and the electrons of another (Holstein, 2001; "VdW Forces," 2013). This attraction is the core principle behind the vdW force: It is a favorable intermolecular interaction occurring between atoms in proximity.

The vdW force between two atoms can be calculated using the Lennard-Jones (LJ) Potential:

$$\textit{Eq. 1.3} \qquad V(r) = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^{6} \right]$$

The LJ potential calculates the intermolecular potential ($V_{LJ}$) between two atoms at a specified distance ($r$), using the strength of attraction between the atoms ($\epsilon$) and the distance where the potential is 0 ($\sigma$). This function expresses the repulsive force as $(\sigma/r)^{12}$ while the attractive force is represented as $(\sigma/r)^{6}$ (Smit, 1992). As the atoms approach the minimum distance, there is a distance $r$ that corresponds to the tightest attraction between the atoms $\epsilon$.

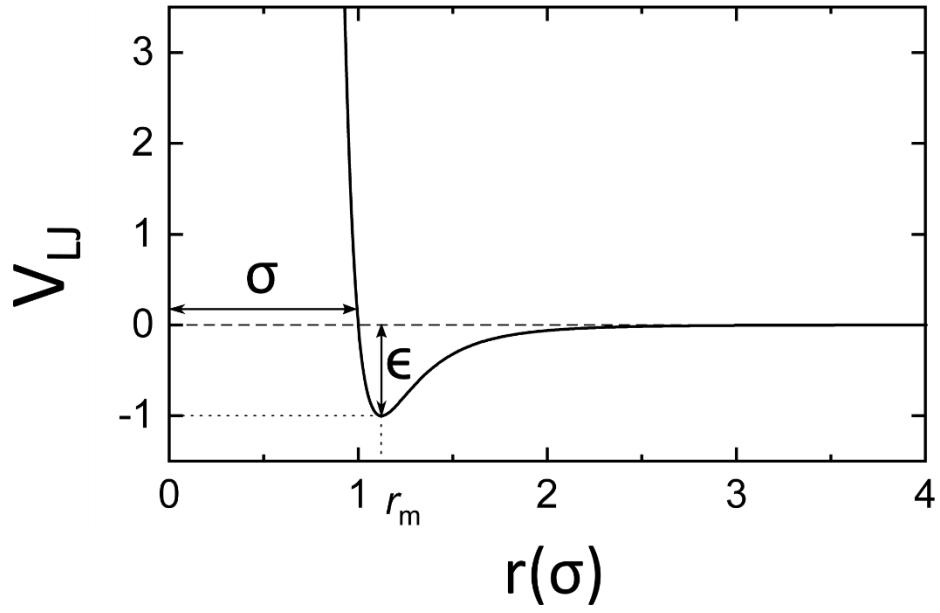



**Figure 1.6 Lennard-Jones Potential.** The intermolecular potential $V_{LJ}$ as a function of distance r between a pair of atoms. The minimum distance ($r_m$) for the most stable interaction energy (ϵ) and the distance where the potential is 0 (σ) are represented on the graph. Adapted from ("Lennard-Jones potential," 2024).

Soluble proteins are driven to fold by the hydrophobic effect, where nonpolar AAs are forced to the core of the folded state. These nonbonded atoms at the core of these proteins are found in tight contact with one another, compounding into a multitude of weak vdW interactions known as vdW packing (Lins & Brasseur, 1995; Pace, 1992). Although vdW packing is not a driving force for soluble protein folding, it is a necessary force that is always present in the folded state. For MPs situated in the core of the hydrophobic

membrane, the hydrophobic effect does not drive MP folding. This means that MPs must rely on other forces to reach the folded state. Although hydrogen bonding and polar interactions have been found to drive MP folding, the extent at which packing contributes to the folded state is unclear. Like soluble protein folding, MP folding eventuates in vdW packing. However, because MPs are engulfed within the crowded lipid bilayer, it is difficult to tease out the influence that vdW packing has on MP folding.

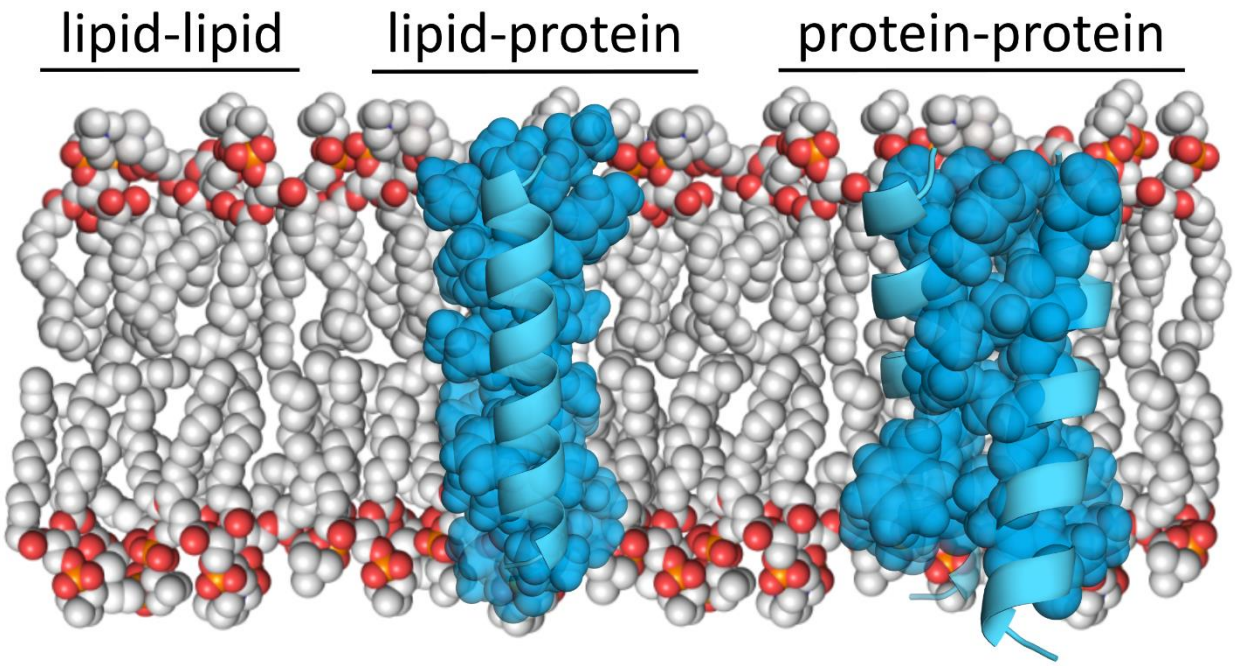



**Figure 1.7 Different types of vdW packing.** VdW packing can be separated into three interactions: lipid-lipid, lipid-protein, and protein-protein. Understanding the impact of each of these forces on folding is crucial to fully understand the impact that vdW has on MP folding and association.

The contribution of vdW packing to MP folding can be broken down into three distinct interactions: lipid-lipid packing, lipid-protein packing, and protein-protein (or sidechain) packing. Lipid-lipid packing involves individual lipid molecules nudged tightly against each other to keep the bilayer assembled. Lipid-protein packing occurs between these lipid molecules and the lipid exposed protein shell (Fattal & Ben-Shaul, 1995). Sidechain packing focuses on the stability gained between fragments of proteins in close contact (Bromberg & Dill, 1994). Each of these interactions plays a role in stabilizing an MP in the bilayer. When an individual protein subunit is inserted into the membrane, it must destabilize the lipid-lipid packing with more favorable lipid-protein packing interactions. For protein-protein packing to occur, these

newly formed lipid-protein interactions must be destabilized for a more favorable combination of protein-protein packing and lipid-lipid packing. This assortment of packing interactions takes place to keep the lipid bilayer intact while the MP reaches its folded state. But simultaneously accounting for all these interactions within the thermodynamics of MP folding is impractical using current technologies. Sidechain packing is a technically feasible starting point because of the ability to manipulate protein sequence and structure within a controlled environment.

Previous research has demonstrated that disruption of sidechain packing within the core of bacteriorhodopsin destabilizes protein structure (Faham et al., 2004; Joh et al., 2009). In addition, a recent study using MP design showed that optimized sidechain packing can stabilize the folded state of the 5-helix bundle protein phospholamban (Mravic et al., 2019). Although these studies suggest that sidechain packing plays a role in stabilizing MP structure, there has not been much investigation on the thermodynamic contribution of packing outside of individual MP systems. My research aims to characterize and quantify the extent to which sidechain packing is a driving force for MP folding for the general population of MP structures.

## 1.7 Thesis overview

My graduate research focused on using computational protein design in combination with high throughput assays to determine the extent at which vdW packing contributes to MP association and folding. Prior research on the impact of packing to the folded state of MPs honed-in on singular systems, and I aimed to expand this knowledge to a larger variety of MP structures.

**In Chapter 2**, I present the majority of my graduate schoolwork, which will be published in the near future. In this paper, I found that sidechain packing is a weak driving force for MP homodimer association. I data mined the Protein Databank (PDB) for all solved MP structures to determine the best TMH structures for computational design, developed an algorithm to design protein homodimers, and assessed the ability for these designs to association using high-throughput sort-seq. I found that proteins designed using sidechain packing associate mildly when compared to $GAS_{right}$ designs that rely on a combination of hydrogen bonding and vdW packing.

**In Chapter 3**, I discuss my computational methods. With improving experimental technologies, many studies at the forefront of research utilize a combination of high-throughput experiments and computational analysis. My research coupled high-throughput experiments with computational design to explore the impact of biophysical forces on protein association in a large range of structures. I discuss rationale for decisions made during the development of my design algorithm and data analysis, including the development of energy terms, choosing different interfaces, and converting sort-seq reconstructed fluorescence to TOXGREEN. I detail programs used in my research, aiming to convey my methods so that they can be implemented in future research.

**In Chapter 4**, I describe a variety of future directions for studying forces involved in MP association and folding using protein design. I include unpublished experiments where I mutated residues that can hydrogen bond on several designs to support that many of my designs associate solely through packing. I

then discuss how to expand my protein design algorithm, including how to design heterodimers, adapting our sequence entropy into a pairwise term, and using machine learning tools for design. I also include a section aimed to address a weakness in sort-seq, discussing how to detect our protein concentrations in high-throughput.

**In Chapter 5**, I share a collaboration with the SciFun program at UW-Madison, detailing my PhD journey through a chapter written for the public. I describe the premise of my research with simplified terminology while simultaneously reflecting on key moments during my graduate school journey, giving transparent thoughts on how my research affected my emotional and mental well-being. Through a combination of letter writing and music, I share how science and research has helped me grow during my time in graduate school.

## 1.8 References

Adamian, L., & Liang, J. (2002). Interhelical hydrogen bonds and spatial motifs in membrane proteins: polar clamps and serine zippers. *Proteins*, *47*(2), 209-218. https://doi.org/10.1002/prot.10071

Agard, D. A., Bowman, G. R., DeGrado, W., Dokholyan, N. V., & Zhou, H. X. (2022). Solution of the protein structure prediction problem at last: crucial innovations and next frontiers. *Fac Rev*, *11*, 38. https://doi.org/10.12703/r-01-0000020

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H.,…Gray, J. J. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, *13*(6), 3031-3048. https://doi.org/10.1021/acs.jctc.7b00125

Anderson, S. M. (2019). *Understanding the GASright Motif: Sequence, Structure, and Stability* (Publication Number 27548821) [Ph.D., The University of Wisconsin - Madison]. Dissertations & Theses @ Big Ten Academic Alliance; Dissertations & Theses @ University of Wisconsin at Madison; ProQuest Dissertations & Theses Global. United States -- Wisconsin. https://ezproxy.library.wisc.edu/login?url=https://www.proquest.com/dissertations-theses/understanding-gas-sub-right-motif-sequence/docview/2331244818/se-2?accountid=465https://resolver.library.wisconsin.edu/uwmad??url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations&sid=ProQ:ProQuest+Dissertations+%26+Theses+Global&atitle=&title=Understanding+the+GASright+Motif%3A+Sequence%2C+Structure%2C+and+Stability&issn=&date=2019-01-01&volume=&issue=&spage=&au=Anderson%2C+Samantha+Marie&isbn=9781392603215&jtitle=&btitle=&rft_id=info:eric/&rft_id=info:doi/

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

Arbely, E., & Arkin, I. T. (2004). Experimental measurement of the strength of a C alpha-H...O bond in a lipid bilayer. *J Am Chem Soc*, *126*(17), 5362-5363. https://doi.org/10.1021/ja049826h

Arkin, I. T., Adams, P. D., MacKenzie, K. R., Lemmon, M. A., Brünger, A. T., & Engelman, D. M. (1994). Structural organization of the pentameric transmembrane alpha-helices of phospholamban, a cardiac ion channel. *EMBO J*, *13*(20), 4757-4764. https://doi.org/10.1002/j.1460-2075.1994.tb06801.x

Armstrong, C. R., & Senes, A. (2016). Screening for transmembrane association in divisome proteins using TOXGREEN, a high-throughput variant of the TOXCAT assay. *Biochim Biophys Acta*, *1858*(11), 2573-2583. https://doi.org/10.1016/j.bbamem.2016.07.008

Ash, W. L., Stockner, T., MacCallum, J. L., & Tieleman, D. P. (2004). Computer modeling of polyleucine-based coiled coil dimers in a realistic membrane environment: insight into helix-helix interactions in membrane proteins. *Biochemistry*, *43*(28), 9050-9060. https://doi.org/10.1021/bi0494572

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R.,…Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871-876. https://doi.org/10.1126/science.abj8754

Batsanov, S. S. (2001). VdW Radii of Elements. *37*(9), 871-885.

Bañó-Polo, M., Martínez-Gil, L., Wallner, B., Nieva, J. L., Elofsson, A., & Mingarro, I. (2013). Charge pair interactions in transmembrane helices and turn propensity of the connecting sequence promote helical hairpin insertion. *J Mol Biol*, *425*(4), 830-840. https://doi.org/10.1016/j.jmb.2012.12.001

Ben-Tal, N., Sitkoff, D., Topol, I. A., Yang, A.-S., Burt, S. K., & Honig, B. (1997). Free energy of amide hydrogen bond formation in vacuum, in water, and in liquid alkane solution. *The Journal of Physical Chemistry B*, *101*(3), 450-457.

Benedicto, A., García-Kamiruaga, I., & Arteta, B. (2021). Neuropilin-1: A feasible link between liver pathologies and COVID-19. *World J Gastroenterol*, *27*(24), 3516-3529. https://doi.org/10.3748/wjg.v27.i24.3516

Blois, T. M., Hong, H., Kim, T. H., & Bowie, J. U. (2009). Protein unfolding with a steric trap. *J Am Chem Soc*, *131*(39), 13914-13915. https://doi.org/10.1021/ja905725n

Bond, P. J., & Sansom, M. S. (2006). Insertion and assembly of membrane proteins via simulation. *Journal of the American Chemical Society*, *128*(8), 2697-2704.

Bowie, J. U. (2011). Membrane protein folding: how important are hydrogen bonds? *Curr Opin Struct Biol*, *21*(1), 42-49. https://doi.org/10.1016/j.sbi.2010.10.003

Braun T, Koehler Leman J, & OF, L. (2015). **Combining Evolutionary Information and an Iterative Sampling Strategy for Accurate Protein Structure Prediction**. In: PLOS Computational Biology.

Bromberg, S., & Dill, K. A. (1994). Side-chain entropy and packing in proteins. *protein Science*, *3*(7), 997-1009.

Bu, L., Im, W., & Brooks, C. L. (2007). Membrane assembly of simple helix homo-oligomers studied via molecular dynamics simulations. *Biophysical journal*, *92*(3), 854-863.

Burley, S. K., & Petsko, G. A. (1985). Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science*, *229*(4708), 23-28. https://doi.org/10.1126/science.3892686

Carpenter, E. P., Beis, K., Cameron, A. D., & Iwata, S. (2008). Overcoming the challenges of membrane protein crystallography. *Curr Opin Struct Biol*, *18*(5), 581-586. https://doi.org/10.1016/j.sbi.2008.07.001

Carter, P. J., Winter, G., Wilkinson, A. J., & Fersht, A. R. (1984). The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus). *Cell*, *38*(3), 835-840. https://doi.org/10.1016/0092-8674(84)90278-2

Chamberlain, A. K., & Bowie, J. U. (2004). Analysis of side-chain rotamers in transmembrane proteins. *Biophys J*, *87*(5), 3460-3469. https://doi.org/10.1529/biophysj.104.044024

Chaudhury, S., Berrondo, M., Weitzner, B. D., Muthu, P., Bergman, H., & Gray, J. J. (2011). Benchmarking and analysis of protein docking performance in Rosetta v3. 2. *PloS one*, *6*(8), e22477.

Choi, M. Y., Cardarelli, L., Therien, A. G., & Deber, C. M. (2004). Non-native interhelical hydrogen bonds in the cystic fibrosis transmembrane conductance regulator domain modulated by polar mutations. *Biochemistry*, *43*(25), 8077-8083. https://doi.org/10.1021/bi0494525

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

Cristian, L., Lear, J. D., & DeGrado, W. F. (2003). Use of thiol-disulfide equilibria to measure the energetics of assembly of transmembrane helices in phospholipid bilayers. *Proc Natl Acad Sci U S A*, *100*(25), 14772-14777. https://doi.org/10.1073/pnas.2536751100

Dalbey, R. E., Wang, P., & Kuhn, A. (2011). Assembly of bacterial inner membrane proteins. *Annu Rev Biochem*, *80*, 161-187. https://doi.org/10.1146/annurev-biochem-060409-092524

de Vries, S. J., van Dijk, M., & Bonvin, A. M. (2010). The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc*, *5*(5), 883-897. https://doi.org/10.1038/nprot.2010.32

Deol, S. S., Domene, C., Bond, P. J., & Sansom, M. S. (2006). Anionic phospholipid interactions with the potassium channel KcsA: simulation studies. *Biophysical journal*, *90*(3), 822-830.

Dominguez, C., Boelens, R., & Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, *125*(7), 1731-1737. https://doi.org/10.1021/ja026939x

Doura, A. K., & Fleming, K. G. (2004). Complex interactions at the helix-helix interface stabilize the glycophorin A transmembrane dimer. *J Mol Biol*, *343*(5), 1487-1497. https://doi.org/10.1016/j.jmb.2004.09.011

Duong, M. T., Jaszewski, T. M., Fleming, K. G., & MacKenzie, K. R. (2007). Changes in apparent free energy of helix-helix dimerization in a biological membrane due to point mutations. *J Mol Biol*, *371*(2), 422-434. https://doi.org/10.1016/j.jmb.2007.05.026

Duran, A. M., & Meiler, J. (2018). Computational design of membrane proteins using RosettaMembrane. *Protein Science*, *27*(1), 341-355. https://doi.org/https://doi.org/10.1002/pro.3335

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

Engelman, D. M., Chen, Y., Chin, C. N., Curran, A. R., Dixon, A. M., Dupuy, A. D.,…Popot, J. L. (2003). Membrane protein folding: beyond the two stage model. *FEBS Lett*, *555*(1), 122-125. https://doi.org/10.1016/s0014-5793(03)01106-2

Engelman, D. M., Goldman, A., & Steitz, T. A. (1982). [11] The identification of helical segments in the polypeptide chain of bacteriorhodopsin. In *Methods in Enzymology* (Vol. 88, pp. 81-88). Academic Press. https://doi.org/https://doi.org/10.1016/0076-6879(82)88014-2

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

Fattal, D. R., & Ben-Shaul, A. (1995). Lipid chain packing and lipid-protein interaction in membranes. *Physica A: Statistical Mechanics and its Applications*, *220*(1-2), 192-216.

Fisher, L. E., Engelman, D. M., & Sturgis, J. N. (1999). Detergents modulate dimerization, but not helicity, of the glycophorin A transmembrane domain. *J Mol Biol*, *293*(3), 639-651. https://doi.org/10.1006/jmbi.1999.3126

Fleming, K. G., Ackerman, A. L., & Engelman, D. M. (1997). The effect of point mutations on the free energy of transmembrane alpha-helix dimerization. *J Mol Biol*, *272*(2), 266-275. https://doi.org/10.1006/jmbi.1997.1236

Fleming, K. G., & Engelman, D. M. (2001). Specificity in transmembrane helix-helix interactions can define a hierarchy of stability for sequence variants. *Proc Natl Acad Sci U S A*, *98*(25), 14340-14344. https://doi.org/10.1073/pnas.251367498

Freiberg, A., Kangur, L., Olsen, J. D., & Hunter, C. N. (2012). Structural implications of hydrogen-bond energetics in membrane proteins revealed by high-pressure spectroscopy. *Biophys J*, *103*(11), 2352-2360. https://doi.org/10.1016/j.bpj.2012.10.030

Goossens, K., & De Winter, H. (2018). Molecular Dynamics Simulations of Membrane Proteins: An Overview. *Journal of Chemical Information and Modeling*, *58*(11), 2193-2202. https://doi.org/10.1021/acs.jcim.8b00639

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

Gregersen, N., Bross, P., Vang, S., & Christensen, J. H. (2006). Protein misfolding and human disease. *Annu Rev Genomics Hum Genet*, *7*, 103-124. https://doi.org/10.1146/annurev.genom.7.080505.115737

Grossfield, A., Feller, S. E., & Pitman, M. C. (2006). A role for direct interactions in the modulation of rhodopsin by ω-3 polyunsaturated lipids. *Proceedings of the National Academy of Sciences*, *103*(13), 4888-4893.

Gurezka, R., Laage, R., Brosig, B., & Langosch, D. (1999). A heptad motif of leucine residues found in membrane proteins can drive self-assembly of artificial transmembrane segments. *J Biol Chem*, *274*(14), 9265-9270. https://doi.org/10.1074/jbc.274.14.9265

Gurezka, R., & Langosch, D. (2001). In vitro selection of membrane-spanning leucine zipper protein-protein interaction motifs using POSSYCCAT. *J Biol Chem*, *276*(49), 45580-45587. https://doi.org/10.1074/jbc.M105362200

Hastrup, H., Karlin, A., & Javitch, J. A. (2001). Symmetrical dimer of the human dopamine transporter revealed by cross-linking Cys-306 at the extracellular end of the sixth transmembrane segment. *Proc Natl Acad Sci U S A*, *98*(18), 10055-10060. https://doi.org/10.1073/pnas.181344298

He, L., & Hristova, K. (2008). Pathogenic activation of receptor tyrosine kinases in mammalian membranes. *J Mol Biol*, *384*(5), 1130-1142. https://doi.org/10.1016/j.jmb.2008.10.036

Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H.,…von Heijne, G. (2005). Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, *433*(7024), 377-381. https://doi.org/10.1038/nature03216

Holstein, B. R. (2001). The vdW interaction. *American Journal of Physics*, *69*(4), 441-449.

Hong, H., & Bowie, J. U. (2011). Dramatic destabilization of transmembrane helix interactions by features of natural membrane environments. *J Am Chem Soc*, *133*(29), 11389-11398. https://doi.org/10.1021/ja204524c

Hong, H., Chang, Y.-C., & Bowie, J. U. (2013). Measuring Transmembrane Helix Interaction Strengths in Lipid Bilayers Using Steric Trapping. In *Membrane Proteins: Folding, Association, and Design* (pp. 37-56). Humana Press. https://doi.org/10.1007/978-1-62703-583-5_3

Hong, H., Joh, N. H., Bowie, J. U., & Tamm, L. K. (2009). Methods for measuring the thermodynamic stability of membrane proteins. *Methods Enzymol*, *455*, 213-236. https://doi.org/10.1016/S0076-6879(08)04208-0

Hong, H., Park, S., Jiménez, R. H., Rinehart, D., & Tamm, L. K. (2007). Role of aromatic side chains in the folding and thermodynamic stability of integral membrane proteins. *J Am Chem Soc*, *129*(26), 8320-8327. https://doi.org/10.1021/ja068849o

Horovitz, A., Fleisher, R. C., & Mondal, T. (2019). Double-mutant cycles: new directions and applications. *Curr Opin Struct Biol*, *58*, 10-17. https://doi.org/10.1016/j.sbi.2019.03.025

Howarth, M., Chinnapen, D. J., Gerrow, K., Dorrestein, P. C., Grandy, M. R., Kelleher, N. L.,…Ting, A. Y. (2006). A monovalent streptavidin with a single femtomolar biotin binding site. *Nat Methods*, *3*(4), 267-273. https://doi.org/10.1038/nmeth861

Huang, B., Xu, Y., Hu, X., Liu, Y., Liao, S., Zhang, J.,…Liu, H. (2022). A backbone-centred energy function of neural networks for protein design. *Nature*, *602*(7897), 523-528. https://doi.org/10.1038/s41586-021-04383-5

Infield, D. T., Rasouli, A., Galles, G. D., Chipot, C., Tajkhorshid, E., & Ahern, C. A. (2021). Cation-π Interactions and their Functional Roles in Membrane Proteins. *J Mol Biol*, *433*(17), 167035. https://doi.org/10.1016/j.jmb.2021.167035

Joh, N. H., Min, A., Faham, S., Whitelegge, J. P., Yang, D., Woods, V. L., & Bowie, J. U. (2008). Modest stabilization by most hydrogen-bonded side-chain interactions in membrane proteins. *Nature*, *453*(7199), 1266-1270. https://doi.org/10.1038/nature06977

Joh, N. H., Oberai, A., Yang, D., Whitelegge, J. P., & Bowie, J. U. (2009). Similar energetic contributions of packing in the core of membrane and water-soluble proteins. *J Am Chem Soc*, *131*(31), 10846-10847. https://doi.org/10.1021/ja904711k

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

Kahn, T. W., & Engelman, D. M. (1992). Bacteriorhodopsin can be refolded from two independently stable transmembrane helices and the complementary five-helix fragment. *Biochemistry*, *31*(26), 6144-6151. https://doi.org/10.1021/bi00141a027

Kim, S., Chamberlain, A. K., & Bowie, J. U. (2004). Membrane channel structure of Helicobacter pylori vacuolating toxin: role of multiple GXXXG motifs in cylindrical channels. *Proc Natl Acad Sci U S A*, *101*(16), 5988-5991. https://doi.org/10.1073/pnas.0308694101

Koehler Leman, J., Mueller, B. K., & Gray, J. J. (2017). Expanding the toolkit for membrane protein modeling in Rosetta. *Bioinformatics*, *33*(5), 754-756. https://doi.org/10.1093/bioinformatics/btw716

Kovalenko, O. V., Metcalf, D. G., DeGrado, W. F., & Hemler, M. E. (2005). Structural organization and interactions of transmembrane domains in tetraspanin proteins. *BMC Struct Biol*, *5*, 11. https://doi.org/10.1186/1472-6807-5-11

Krivov, G. G., Shapovalov, M. V., & Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, *77*(4), 778-795. https://doi.org/10.1002/prot.22488

Kulp, D. W., Subramaniam, S., Donald, J. E., Hannigan, B. T., Mueller, B. K., Grigoryan, G., & Senes, A. (2012). Structural informatics, modeling, and design with an open-source Molecular Software Library (MSL). *J Comput Chem*, *33*(20), 1645-1661. https://doi.org/10.1002/jcc.22968

Kumari, N., & Yadav, S. (2019). Modulation of protein oligomerization: An overview. *Prog Biophys Mol Biol*, *149*, 99-113. https://doi.org/10.1016/j.pbiomolbio.2019.03.003

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, *157*(1), 105-132. https://doi.org/10.1016/0022-2836(82)90515-0

Lazaridis, T. (2003). Effective energy function for proteins in lipid membranes. *Proteins*, *52*(2), 176-192. https://doi.org/10.1002/prot.10410

Leman, J. K., Weitzner, B. D., Lewis, S. M., Adolf-Bryfogle, J., Alam, N., Alford, R. F.,…Barth, P. (2020). Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature methods*, *17*(7), 665-680.

Lemmon, M. A., Flanagan, J. M., Hunt, J. F., Adair, B. D., Bormann, B. J., Dempsey, C. E., & Engelman, D. M. (1992). Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices. *J Biol Chem*, *267*(11), 7683-7689.

Lemmon, M. A., Flanagan, J. M., Treutlein, H. R., Zhang, J., & Engelman, D. M. (1992). Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry*, *31*(51), 12719-12725. https://doi.org/10.1021/bi00166a002

Lennard-Jones potential. (2024). *Wikipedia*.

Li, E., You, M., & Hristova, K. (2006). FGFR3 dimer stabilization due to a single amino acid pathogenic mutation. *J Mol Biol*, *356*(3), 600-612. https://doi.org/10.1016/j.jmb.2005.11.077

Lins, L., & Brasseur, R. (1995). The hydrophobic effect in protein folding. *The FASEB journal*, *9*(7), 535-540.

Liu, Y., Engelman, D. M., & Gerstein, M. (2002). Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biol*, *3*(10), research0054. https://doi.org/10.1186/gb-2002-3-10-research0054

Lomize, A. L., & Pogozheva, I. D. (2017). TMDOCK: An Energy-Based Method for Modeling α-Helical Dimers in Membranes. *J Mol Biol*, *429*(3), 390-398. https://doi.org/10.1016/j.jmb.2016.09.005

Lu, C., Mi, L. Z., Grey, M. J., Zhu, J., Graef, E., Yokoyama, S., & Springer, T. A. (2010). Structural evidence for loose linkage between ligand binding and kinase activation in the epidermal growth factor receptor. *Mol Cell Biol*, *30*(22), 5432-5443. https://doi.org/10.1128/MCB.00742-10

MacKenzie, K. R., & Engelman, D. M. (1998). Structure-based prediction of the stability of transmembrane helix-helix interactions: the sequence dependence of glycophorin A dimerization. *Proc Natl Acad Sci U S A*, *95*(7), 3583-3590. https://doi.org/10.1073/pnas.95.7.3583

MacKenzie, K. R., Prestegard, J. H., & Engelman, D. M. (1997). A transmembrane helix dimer: structure and implications. *Science*, *276*(5309), 131-133. https://doi.org/10.1126/science.276.5309.131

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J.,…Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, *102*(18), 3586-3616. https://doi.org/10.1021/jp973084f

Mackerell Jr., A. D., Feig, M., & Brooks III, C. L. (2004). Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry*, *25*(11), 1400-1415. https://doi.org/https://doi.org/10.1002/jcc.20065

Marinko, J. T., Huang, H., Penn, W. D., Capra, J. A., Schlebach, J. P., & Sanders, C. R. (2019). Folding and Misfolding of Human Membrane Proteins in Health and Disease: From Single Molecules to Cellular Proteostasis. *Chem Rev*, *119*(9), 5537-5606. https://doi.org/10.1021/acs.chemrev.8b00532

Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., & De Vries, A. H. (2007). The MARTINI force field: coarse grained model for biomolecular simulations. *The journal of physical chemistry B*, *111*(27), 7812-7824.

Marshall, J. N., Klein, M. N., Karki, P., Promnares, K., Setua, S., Fan, X.,…Fontaine, M. J. (2024). Aberrant GPA expression and regulatory function of red blood cells in sickle cell disease. *Blood Adv*, *8*(7), 1687-1697. https://doi.org/10.1182/bloodadvances.2023011611

Marti, T. (1998). Refolding of bacteriorhodopsin from expressed polypeptide fragments. *J Biol Chem*, *273*(15), 9312-9322. https://doi.org/10.1074/jbc.273.15.9312

Michel, H., Epp, O., & Deisenhofer, J. (1986). Pigment-protein interactions in the photosynthetic reaction centre from Rhodopseudomonas viridis. *EMBO J*, *5*(10), 2445-2451. https://doi.org/10.1002/j.1460-2075.1986.tb04520.x

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nat Methods*, *19*(6), 679-682. https://doi.org/10.1038/s41592-022-01488-1

Mitchell, J. B., & Price, S. L. (1990). The nature of the N□ H… O□ C hydrogen bond: An intermolecular perturbation theory study of the formamide/formaldehyde complex. *Journal of computational chemistry*, *11*(10), 1217-1233.

Mottamal, M., Zhang, J., & Lazaridis, T. (2006). Energetics of the native and non-native states of the glycophorin transmembrane helix dimer. *PROTEINS: Structure, Function, and Bioinformatics*, *62*(4), 996-1009.

Mravic, M., Thomaston, J. L., Tucker, M., Solomon, P. E., Liu, L., & DeGrado, W. F. (2019). Packing of apolar side chains enables accurate design of highly stable membrane proteins. *Science*, *363*(6434), 1418-1423. https://doi.org/10.1126/science.aav7541

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

Munson, M., Balasubramanian, S., Fleming, K. G., Nagi, A. D., O'Brien, R., Sturtevant, J. M., & Regan, L. (1996). What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties. *Protein Sci*, *5*(8), 1584-1593. https://doi.org/10.1002/pro.5560050813

North, B., Cristian, L., Fu Stowell, X., Lear, J. D., Saven, J. G., & DeGrado, W. F. (2006). Characterization of a Membrane Protein Folding Motif, the Ser Zipper, Using Designed Peptides. *Journal of Molecular Biology*, *359*(4), 930-939. https://doi.org/https://doi.org/10.1016/j.jmb.2006.04.001

Pace, C. N. (1992). Contribution of the hydrophobic effect to globular protein stability. *Journal of molecular biology*, *226*(1), 29-35.

Park, C., & Marqusee, S. (2005). Pulse proteolysis: a simple method for quantitative determination of protein stability and ligand binding. *Nat Methods*, *2*(3), 207-212. https://doi.org/10.1038/nmeth740

Partridge, A. W., Therien, A. G., & Deber, C. M. (2002). Polar mutations in membrane proteins as a biophysical basis for disease. *Biopolymers*, *66*(5), 350-358. https://doi.org/10.1002/bip.10313

Partridge, A. W., Therien, A. G., & Deber, C. M. (2004). Missense mutations in transmembrane domains of proteins: phenotypic propensity of polar residues for human disease. *Proteins*, *54*(4), 648-656. https://doi.org/10.1002/prot.10611

Petrache, H. I., Grossfield, A., MacKenzie, K. R., Engelman, D. M., & Woolf, T. B. (2000). Modulation of glycophorin A transmembrane helix interactions by lipid bilayers: molecular dynamics calculations. *J Mol Biol*, *302*(3), 727-746. https://doi.org/10.1006/jmbi.2000.4072

Polyansky, A. A., Volynsky, P. E., & Efremov, R. G. (2012). Multistate organization of transmembrane helical protein dimers governed by the host membrane. *J Am Chem Soc*, *134*(35), 14390-14400. https://doi.org/10.1021/ja303483k

Popot, J. L., & Engelman, D. M. (1990). Membrane protein folding and oligomerization: the two-stage model. *Biochemistry*, *29*(17), 4031-4037. https://doi.org/10.1021/bi00469a001

Popot, J. L., & Engelman, D. M. (2000). Helical membrane protein folding, stability, and evolution. *Annu Rev Biochem*, *69*, 881-922. https://doi.org/10.1146/annurev.biochem.69.1.881

Popot, J. L., Gerchman, S. E., & Engelman, D. M. (1987). Refolding of bacteriorhodopsin in lipid bilayers. A thermodynamically controlled two-stage process. *J Mol Biol*, *198*(4), 655-676. https://doi.org/10.1016/0022-2836(87)90208-7

Rapoport, T. A. (2007). Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature*, *450*(7170), 663-669. https://doi.org/10.1038/nature06384

RICHARDSON, J. S., KEEDY, D. A., & RICHARDSON, D. C. “THE PLOT” THICKENS: MORE DATA, MORE DIMENSIONS, MORE USES. In *Biomolecular Forms and Functions* (pp. 46-61). https://doi.org/10.1142/9789814449144_0004

Roberts, J. B., Nava, A. A., Pearson, A. N., Incha, M. R., Valencia, L. E., Ma, M.,…Keasling, J. D. (2024). Foldy: An open-source web application for interactive protein structure analysis. *PLoS Comput Biol*, *20*(2), e1011171. https://doi.org/10.1371/journal.pcbi.1011171

Rose, G. D., & Wolfenden, R. (1993). Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annual review of biophysics and biomolecular structure*, *22*(1), 381-415.

Russ, W. P., & Engelman, D. M. (1999). TOXCAT: a measure of transmembrane helix association in a biological membrane. *Proc Natl Acad Sci U S A*, *96*(3), 863-868. https://doi.org/10.1073/pnas.96.3.863

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

Salinas, V. H., & Ranganathan, R. (2018). Coevolution-based inference of amino acid interactions underlying protein function. *Elife*, *7*. https://doi.org/10.7554/eLife.34300

Sanders, C. R., & Myers, J. K. (2004). Disease-related misassembly of membrane proteins. *Annu Rev Biophys Biomol Struct*, *33*, 25-51. https://doi.org/10.1146/annurev.biophys.33.110502.140348

Scheiner, S., Kar, T., & Gu, Y. (2001). Strength of the Calpha H..O hydrogen bond of amino acid residues. *J Biol Chem*, *276*(13), 9832-9837. https://doi.org/10.1074/jbc.M010770200

Schlebach, J. P., Cao, Z., Bowie, J. U., & Park, C. (2012). Revisiting the folding kinetics of bacteriorhodopsin. *Protein Sci*, *21*(1), 97-106. https://doi.org/10.1002/pro.766

Schlebach, J. P., Kim, M. S., Joh, N. H., Bowie, J. U., & Park, C. (2011). Probing membrane protein unfolding with pulse proteolysis. *J Mol Biol*, *406*(4), 545-551. https://doi.org/10.1016/j.jmb.2010.12.018

Schneider, D., & Engelman, D. M. (2003). GALLEX, a measurement of heterologous association of transmembrane helices in a biological membrane. *J Biol Chem*, *278*(5), 3105-3111. https://doi.org/10.1074/jbc.M206287200

Senes, A., Ubarretxena-Belandia, I., & Engelman, D. M. (2001). The Cα—H···O hydrogen bond: A determinant of stability and specificity in transmembrane helix interactions. *Proceedings of the National Academy of Sciences*, *98*(16), 9056-9061. https://doi.org/doi:10.1073/pnas.161280798

Smit, B. (1992). Phase diagrams of Lennard-Jones fluids. *The Journal of chemical physics*, *96*(11), 8639-8640.

Smith, S. O., Eilers, M., Song, D., Crocker, E., Ying, W., Groesbeek, M.,…Aimoto, S. (2002). Implications of threonine hydrogen bonding in the glycophorin A transmembrane helix dimer. *Biophys J*, *82*(5), 2476-2486. https://doi.org/10.1016/S0006-3495(02)75590-2

Soares, T. A., Hünenberger, P. H., Kastenholz, M. A., Kräutler, V., Lenz, T., Lins, R. D.,…van Gunsteren, W. F. (2005). An improved nucleic acid parameter set for the GROMOS force field. *Journal of Computational Chemistry*, *26*(7), 725-737. https://doi.org/https://doi.org/10.1002/jcc.20193

Stanley, A. M., & Fleming, K. G. (2007). The role of a hydrogen bonding network in the transmembrane beta-barrel OMPLA. *J Mol Biol*, *370*(5), 912-924. https://doi.org/10.1016/j.jmb.2007.05.009

Subramaniam, S., & Senes, A. (2012). An energy-based conformer library for side chain optimization: improved prediction and adjustable sampling. *Proteins*, *80*(9), 2218-2234. https://doi.org/10.1002/prot.24111

Tanford, C. (1980). *The hydrophobic effect: formation of micelles and biological membranes 2d ed*. J. Wiley.

Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Molina, M. M. S., Shames, I.,…Michnick, S. W. (2008). An in vivo map of the yeast protein interactome. *Science*, *320*(5882), 1465-1470.

Therien, A. G., Grant, F. E., & Deber, C. M. (2001). Interhelical hydrogen bonds in the CFTR membrane domain. *Nat Struct Biol*, *8*(7), 597-601. https://doi.org/10.1038/89631

Thieker, D. F., Maguire, J. B., Kudlacek, S. T., Leaver-Fay, A., Lyskov, S., & Kuhlman, B. (2022). Stabilizing proteins, simplified: A Rosetta-based webtool for predicting favorable mutations. *Protein Science*, *31*(10), e4428. https://doi.org/https://doi.org/10.1002/pro.4428

Tsemekhman, K., Goldschmidt, L., Eisenberg, D., & Baker, D. (2007). Cooperative hydrogen bonding in amyloid formation. *Protein science*, *16*(4), 761-764.

Tsirigos, K. D., Peters, C., Shu, N., Käll, L., & Elofsson, A. (2015). The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Research*, *43*(W1), W401-W407. https://doi.org/10.1093/nar/gkv485

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

VdW Forces. (2013). *Chemistry LibreTexts*.

Vargas, E., Yarov-Yarovoy, V., Khalili-Araghi, F., Catterall, W. A., Klein, M. L., Tarek, M.,…Roux, B. (2012). An emerging consensus on voltage-dependent gating from computational modeling and molecular dynamics simulations. *J Gen Physiol*, *140*(6), 587-594. https://doi.org/10.1085/jgp.201210873

Vargas, R., Garza, J., Dixon, a. D. A., & Hay, B. P. (2000). How Strong Is the Cα−H···OC Hydrogen Bond? *Journal of the American Chemical Society*, *122*, 4750-4755.

Vicente, C. M., Ricci, R., Nader, H. B., & Toma, L. (2013). Syndecan-2 is upregulated in colorectal cancer cells through interactions with extracellular matrix produced by stromal fibroblasts. *BMC Cell Biol*, *14*, 25. https://doi.org/10.1186/1471-2121-14-25

Viklund, H., & Elofsson, A. (2008). OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics*, *24*(15), 1662-1668. https://doi.org/10.1093/bioinformatics/btn221

von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol*, *225*(2), 487-494. https://doi.org/10.1016/0022-2836(92)90934-c

Walters, R. F., & DeGrado, W. F. (2006). Helix-packing motifs in membrane proteins. *Proc Natl Acad Sci U S A*, *103*(37), 13658-13663. https://doi.org/10.1073/pnas.0605878103

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., & Case, D. A. (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry*, *25*(9), 1157-1174. https://doi.org/https://doi.org/10.1002/jcc.20035

Wehbi, H., Gasmi-Seabrook, G., Choi, M. Y., & Deber, C. M. (2008). Positional dependence of non-native polar mutations on folding of CFTR helical hairpins. *Biochim Biophys Acta*, *1778*(1), 79-87. https://doi.org/10.1016/j.bbamem.2007.08.036

Weitzner, B. D., Jeliazkov, J. R., Lyskov, S., Marze, N., Kuroda, D., Frick, R.,…Gray, J. J. (2017). Modeling and docking of antibody structures with Rosetta. *Nature protocols*, *12*(2), 401-416.

White, S. H. (2005). How hydrogen bonds shape membrane protein structure. *Adv Protein Chem*, *72*, 157-172. https://doi.org/10.1016/S0065-3233(05)72006-4

White, S. H., & von Heijne, G. (2004). The machinery of membrane protein assembly. *Curr Opin Struct Biol*, *14*(4), 397-404. https://doi.org/10.1016/j.sbi.2004.07.003

White, S. H., & Wimley, W. C. (1999). Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct*, *28*, 319-365. https://doi.org/10.1146/annurev.biophys.28.1.319

Wilkins, M. R., Gasteiger, E., Bairoch, A., Sanchez, J. C., Williams, K. L., Appel, R. D., & Hochstrasser, D. F. (1999). Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol*, *112*, 531-552. https://doi.org/10.1385/1-59259-584-7:531

Wu, T., Malinverni, J., Ruiz, N., Kim, S., Silhavy, T. J., & Kahne, D. (2005). Identification of a multicomponent complex required for outer membrane biogenesis in Escherichia coli. *Cell*, *121*(2), 235-245. https://doi.org/10.1016/j.cell.2005.02.015

Yang, A.-S., Sharp, K. A., & Honig, B. (1992). Analysis of the heat capacity dependence of protein folding. *Journal of Molecular Biology*, *227*(3), 889-900. https://doi.org/https://doi.org/10.1016/0022-2836(92)90229-D

Yano, Y., Takemoto, T., Kobayashi, S., Yasui, H., Sakurai, H., Ohashi, W.,…Matsuzaki, K. (2002). Topological stability and self-association of a completely hydrophobic model transmembrane helix in lipid bilayers. *Biochemistry*, *41*(9), 3073-3080. https://doi.org/10.1021/bi011161y

Yarov-Yarovoy, V., Baker, D., & Catterall, W. A. (2006). Voltage sensor conformations in the open and closed states in ROSETTA structural models of K(+) channels. *Proc Natl Acad Sci U S A*, *103*(19), 7292-7297. https://doi.org/10.1073/pnas.0602350103

Yarov-Yarovoy, V., Schonbrun, J., & Baker, D. (2006). Multipass membrane protein structure prediction using Rosetta. *Proteins*, *62*(4), 1010-1025. https://doi.org/10.1002/prot.20817

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

# Chapter 2: Van der Waals forces alone are a weak design principle for transmembrane helix interaction stability

This chapter was prepared for publication as:

Gilbert J. Loiseau and Alessandro Senes “Van der Waals forces alone are a weak design principle for transmembrane helix interaction stability” In preparation.

## 2.1 Abstract

Membrane protein folding occurs as a result of an equilibrium of biophysical forces. Polar and charged interactions can drive folding within the hydrophobic environment of the membrane. The stabilizing effects of sidechain packing on membrane protein folding were deduced in a variety of foundational bacteriorhodopsin studies, and sidechain packing as a driving force has been demonstrated through modern protein re-design of phospholamban. However, the extent at which packing can act as a driving force in general in membrane protein systems is not well understood. We investigate the impact of sidechain packing on membrane protein association by using computational design. By sampling dimer conformations from the PDB, we designed hundreds of sequences that self-associate. Structural characterization through mutations on these proteins suggests that packing is a weak driving force in a variety of membrane protein systems.

## 2.2 Introduction

Proper membrane protein folding is regulated by a distribution of stabilizing hydrogen bonds, weak polar interactions, and vdW (vdW) forces between the unfolded and folded states. Hydrogen bonding and polar interactions have been found to modestly stabilize both large protein complexes (Ben-Tal et al., 1997; Bowie, 2011; Mitchell & Price, 1990; Rose & Wolfenden, 1993; Tsemekhman et al., 2007) and association of smaller peptides such as glycophorin A (GpA) (MacKenzie et al., 1997; Smith et al., 2002). Mutations to key hydrogen bonding residues plays a role in promoting disease states such as cystic fibrosis (Partridge et al., 2004; Therien et al., 2001; Wehbi et al., 2008). Misfolding of membrane proteins has been found to be involved in several genetic diseases such as Parkinson's and cancer (Gregersen et al., 2006; Sanders & Myers, 2004). Understanding how forces drive membrane protein folding is a critical task to further prevent disease. However, research is lacking on the contribution of vdW packing.

The contribution of vdW packing to membrane protein folding can be broken down into three distinct interactions: lipid-lipid packing, lipid-protein packing, and protein-protein packing. Protein-protein (or sidechain) packing is a technically feasible starting point because of the ability to manipulate sequences and determine changes in stability due to mutation. Previous research has demonstrated that disruption of packing within the core of bacteriorhodopsin destabilizes protein structure, suggesting that it is a necessary force for proteins to reach optimal stability (Faham et al., 2004; Joh et al., 2009). Additionally, a recent study using membrane protein design has shown that optimized sidechain packing can stabilize the folded state of phospholamban, demonstrating that packing can drive protein stability (Mravic et al., 2019). Although these studies suggest that sidechain packing plays a role in stabilizing membrane proteins, there has not been much investigation of the thermodynamic contribution of packing outside of individual structures.

Research using homodimerization as a model system has measured the contributions of both hydrogen bonding and weak polar interactions in the membrane, determining that these forces drive membrane protein folding (Johnson et al., 2007; Yano et al., 2002; Zhou et al., 2001). In this study, we investigate the effect of sidechain packing within a multitude of dimer structures. Using large-scale computational design on the most common dimeric backbone geometries found within the PDB, we designed thousands of homodimer sequences and characterized their propensity to associate with a complementary high-throughput method, sort-seq (Anderson, 2019). Using this simple and tractable model system, successfully designed dimers that associate, suggesting that sidechain packing is a weak driving force involved in TM association.

## 2.3 Results and Discussion

### 2.3.1 Design strategy

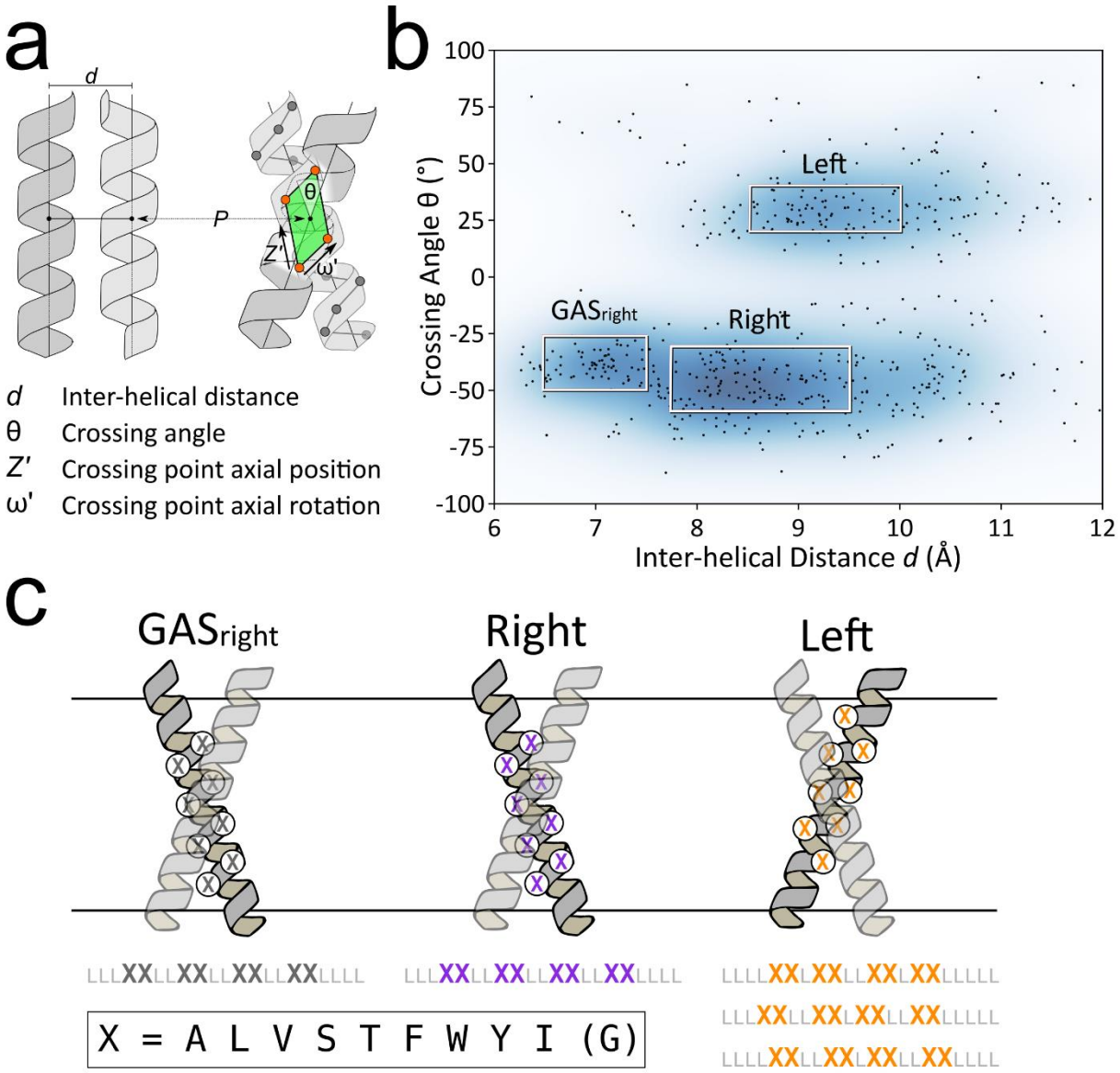



**Figure 2.1 Membrane protein dimer design. a)** Homodimer interfacial sequences are designed on straight helical structures placed at 4 geometric parameters: distance (d), angle (θ), axial rotation (ω), and Z-shift (Z). **b)** Helices within close contact were extracted from the Orientations of Proteins in membranes (OPM) in September 2019. The geometric terms were extracted and overlaid over the kernel density estimation of the membrane protein contact space. **c)** Interfaces for designs on poly-leucine backbones, where x is an interfacial position that can be designed using a library of the most prevalent amino acids in membrane proteins. Glycine is only used for $GAS_{right}$ designs.

Our strategy for investigating whether pure packing can be a strong force for stabilizing helix-helix interaction in membrane proteins was to choose a wide variety of commonly found inter-helical geometries and design their interfaces with hydrophobic side chains, producing different levels of optimal packing. The designed interfaces were limited to the hydrophobic amino acids most commonly found in membrane proteins (A, V, F, W, L, I, S, T and Y) (Liu et al., 2002). For variety, we chose to include Ser, Thr, and Tyr in the set of amino acids since they are common in membrane proteins (> 4%, Fig. S2.2). These amino acids contain a hydroxyl group and could potentially for inter-helical hydrogen bonds. However, these Ser and Thr have tendency to satisfy their potential by forming intra-helical hydrogen bonds with carbonyl groups at i-3 and i-4 (Bower et al., 1997). Tyr is an amino acid that has a tendency to be enriched in the membrane head-group region and it has a tendency to "snorkel", and thus expose its η-hydroxyl group to water and polar groups in the head-group region (Liang & Tamm, 2016). These amino acids are thus compatible with a helix-helix interface mediated primarily by non-polar interactions. We chose not to include Gly because it is the central amino acid in the interface of the strong $GAS_{right}$ motifs, to prevent the accidental occurrence of $GAS_{right}$ capable dimers (Fig. 2.1b). All non-interfacial positions were standardized to Leu.

To produce the backbones for the designs, we selected the most common regions of helix-helix interaction geometry, under the assumption that these would be the most favorable regions for protein design. To identify these regions, we analyzed the geometry of all pairs of interacting helices found in known membrane protein structures, using a database of 1541 membrane protein structures from Orientations of Proteins in Membranes, filtered by sequence similarity from the PDB (Lomize et al., 2006; Steinegger & Söding, 2017). Any two helices in close contact in the structures were considered as an individual helical pair and the conformational parameters of each pair were computed (i.e., the crossing angle θ, the inter-helical distance *d*, the axial rotation of each helix ω, and the displacement in along the helical axis or *Z*-shift, see Fig. 2.1a and Methods). Although similar analyses were performed previously

(Senes et al., 2001; Walters & DeGrado, 2006), we restricted the analysis to consider only the subset of helical pairs that interact through segments that have regular α-helix conformation, since the inclusion of segments that contain kinks, curvature, $3_{10}$ helix and other deviations from standard α-helix, introduces errors in the estimation of the helical parameters (i.e. θ, *d*, ω, *Z*) of the pairs.

Fig. 2.1b illustrates the resulting helix-helix interaction landscape for parallel helical pairs (i.e. the pairs whose N-termini are on the same side of the membrane), plotted as a scatterplot of the crossing angle θ and interhelical distance *d*. Two major high-density regions that are suitable for vdW-based protein design were identified. In the left-handed region (positive crossing angle θ), the helical pairs are found to interact most frequently in the range between approximately 8.5 to 10 Å of inter-helical distance and a 20 to 40° crossing angle (a region that we call "Left", highlighted as a box in Fig. 2.1b). The right-handed region presents a broader range of distances from 6.5 to over 10 Å of distance and crossing angles between -30 and -60°. We selected the maximum density region with inter-helical distance between 7.75 and 9.5 Å and a -30 to -60° crossing angle ("Right" box) as the second source of backbones.

The section of the right-handed region (θ = -25 to -55°) with the closest inter-helical distance (*d* = 6.5 to 7.5 Å) corresponds to the $GAS_{right}$ motif, a well-know helix-helix dimerization and interaction motif (Mueller et al., 2014; Russ & Engelman, 2000). It has a sequence signature of small amino acids (Gly, Ala, Ser) at the dimer interface, forming its characteristic GxxxG and similar sequence motifs (GxxxA, SxxxG, etc.). These small amino acids allow for the short inter-helical distance that brings the helical backbones in contact, which results in the formation of characteristic networks of inter-helical weak hydrogen bonds between between activated Cα–H carbon donors and backbone carbonyl acceptor groups on the opposing helix (Cα–H···O=C hydrogen bonds) (Senes et al., 2001). We previously showed that these networks of Cα–H are major contributor to the stability of $GAS_{right}$ along with vdW interactions (Anderson et al., 2017). Since $GAS_{right}$ can form strong dimers, we used designs in this region as a positive control in our procedure

along with the vdW only design based on the “Left” and “Right” regions. The set of amino acids used for the design of the interfaces of $GAS_{right}$ designs included Gly in addition to the other 9 amino acids used for the Left and Right region designs.

### 2.3.2 Selection of backbones and Computational Design Strategy

Within the boundaries of the three regions of Fig. 2.1b, we selected a total of 10000 starting backbones with crossing angle and inter-helical distance randomly assigned within their boundaries. Since no specific dependency was identified for the axial rotation ω, and the *Z* displacement on the crossing angle and inter-helical distance, these two parameters were also assigned randomly (Fig. S2.7, S2.8, and 2.9). In creating these constructs we followed an approach used previously in a TOXGREEN study (Anderson et al., 2017). The designs we based on a 21 amino acid hydrophobic segment, with the crossing point between the two monomers placed near the middle of the membrane. For each dimer, only 8 positions at the helix-helix interface were set as variable for the design. All non-interfacial positions of the backbone were standardized to Leucine. This standardization reduces variability in hydrophobicity, which helps obtaining consistent expression and insertion into the membrane of the TOXGREEN constructs. The poly-Leu backbone was used because this sequence has low propensity for association in TOXCAT based assays, reducing the risk of non-specific association outside of our designed interface (Zhou et al., 2000; Zhou et al., 2001; Ruan et al., 2003).

The pattern of variable and fixed position used for the constructs is illustrated in Fig. 2.1c. For the $GAS_{right}$ region, we adopted the typical pattern that spaces two interfacial positions (x) with two fixed positions (L), resulting in a LLL**xx**LL**xx**LL**xx**LL**xx**LLLL pattern (Anderson et al., 2017; Mueller et al., 2014; Russ & Engelman, 2000). The same interface was applied to the Right region designs, which has geometry similar to $GAS_{right}$ outside of a larger inter-helical distance. In the Left region, the interfacial positions follows the typical L**xx**LL**x**L heptad repeat that is typical in leucine zippers, knobs-into-holes, and coiled

coils (Ash et al., 2004; Bornberg-Bauer et al., 1998; Walshaw & Woolfson, 2003). However, the specific registry of this pattern is dependent on the Z-shift and axial rotation coordinate of the dimer (Fig. S2.10, Methods). For this reason three different interfacial patterns, all consisting of 8 variable positions, were adopted for dimeric designs in this region (Fig. 2.1c).

The protein design was based on a Monte Carlo procedure that included an initial fixed backbone design, followed by an iterative local backbone refinement using the resulting design sequence (Huang et al., 2022; Kuhlman et al., 2003; Kulp et al., 2012; Nash et al., 2015; Senes, 2011) (see Methods). As in our previous $GAS_{right}$ prediction and design studies (Anderson et al., 2017; Mueller et al., 2014), each protein was evaluated using a computational energy score composed of a vdW function (MacKerell et al., 1998), hydrogen bonding (Krivov et al., 2009), and the IMM1 implicit membrane solvation (Lazaridis, 2003).

To control in part for the possibility that some of the constructs would associate through a different interface than their design, we designed mutations at positions expected to disrupt association of the given structure. For each of these designs, an interfacial position was mutated either small-to-large (isoleucine) or large-to-small (alanine) (Fig. 2.2a). The isoleucine mutants were designed to create significant clashes within the structure (“Clash mutants”), whereas the alanine mutations (”Large→Ala”) were designed to produce a significant reduction of packing at the helix-helix interface. Both Clash and Large→Ala mutants were expected to decrease association, with Clash mutants expected to result in drastic decreases as in previous studies (Khadria et al., 2014; LaPointe et al., 2013; Lawrie et al., 2010; Wei et al., 2011, 2013).

### 2.3.3 Experimental determination of dimerization propensities

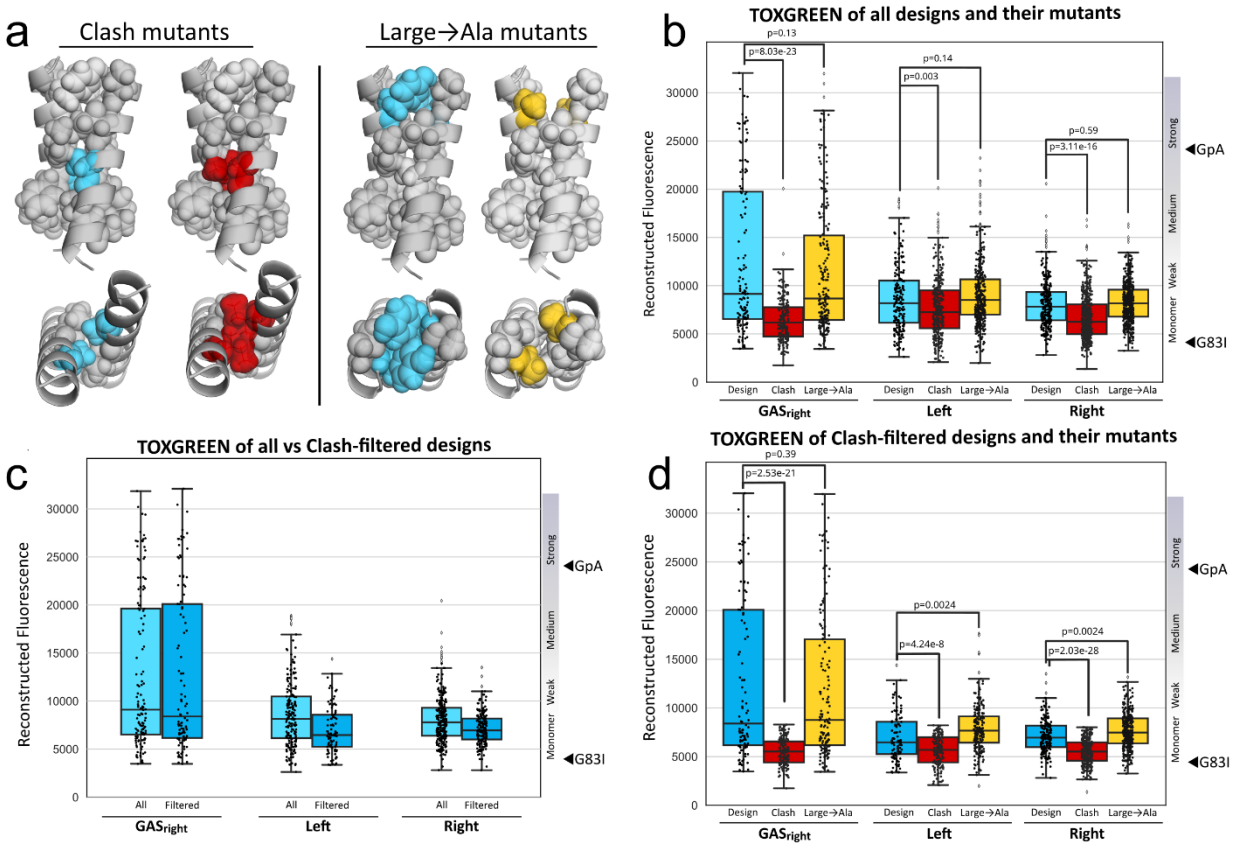



**Figure 2.2 Using mutants to validate designed sequences. a)** Mutations made to each WT design. Clash mutants result in atoms from one helix overlapping with atoms from the other helix, Void mutants result in pockets at the interface lacking atoms. **b-d)** Box and whisker plots separate the data by designating the lowest 25% and highest 25% of the data as whiskers outside of the box. Outliers are represented as diamonds outside of the range of the box and whisker plot. p-values determining the significant difference between design fluorescence and mutant fluorescence in b and d are shown over the corresponding boxplot. The WT (dark green) fluorescence distribution is compared to the fluorescence distributions of their corresponding clash (red) and void (orange) mutants in all data (**b**) and clash filtered (**d**). Panel **c** compares the clash filtered dataset of designs (light green) to all designs found in the sort-seq data (dark green).

We selected 1045 design sequences with a range of both structure and stability. The propensity for dimerization of these sequences and their mutants was assessed in the *Escherichia coli* membrane using the high throughput TOXGREEN sort-seq assay (Anderson, 2019). TOXGREEN (Armstrong & Senes, 2016) is the GFP-based version of TOXCAT, a widely used *in vivo* reporter assay that is sensitive to the relative association of TM dimers in a biological membrane. It is based on a chimeric protein in which the TM domain of interest is fused to the ToxR transcriptional activator. Dimerization of these constructs in the

inner membrane promotes allows for ToxR binding to a specific promoter, resulting in the expression of chloramphenicol acetyltransferase (CAT, in TOXCAT) of GFP (in TOXGREEN). Quantification of these reporter proteins provides an indication of a TM domain's propensity for oligomerization. We recently demonstrated that the TOXCAT signal of a series of $GAS_{right}$ dimers based on a similar standardized poly-Leu backbone correlates well with their thermodynamic stability measured *in vitro* (Díaz Vázquez et al., 2023).

The recent development of a high-throughput TOXGREEN assay based on sort-seq (Anderson, 2019) allows for the measurement of thousands of constructs simultaneously. The assay utilizes a fluorescent activated cell sorting (FACS) to separate cells based on their GFP expression in bins. The relative frequency of a construct in the bins is then measured using next generation sequencing (NGS) and this profile is used to reconstruct the GFP fluorescence of each individual construct (Anderson, 2019).

### 2.3.4 The vdW-based designs are weak in comparison to $GAS_{right}$

The NGS analysis of the FACS fractions recovered 91% of the designed sequences (Fig. S2.3 and S2.4). After removing 241 constructs that do not grow in maltose media (suggesting that these constructs may have poor membrane insertion), we obtained a reconstructed TOXGREEN fluorescence signal for a total of 613 designs and their respective mutants (138 $GAS_{right}$, 198 Left and 277 Right). Fig. 2.2b shows the reconstructed GFP fluorescence distribution of the designs in the three regions (dark green) and their Clash (red) and Large→Ala mutants (orange). As a reference, the fluorescence of the two classic controls is indicated, the strong dimer glycophorin A (GpA) and its monomeric G83I variants.

It is notable that the fluorescence of the designs within the $GAS_{right}$ region spans a larger range than both the Left and Right designs, including constructs with fluorescence comparable or greater than the strong GpA dimer, confirming the propensity of $GAS_{right}$ for strong self-association. In contrast, the majority of the designs in the Left and Right regions appear monomeric or have a weak propensity to

oligomerize. For example, approximately 40% of the $GAS_{right}$ designs have a reconstructed fluorescence above 60% of the GpA standard (14,500 fluorescent units in the figure), compared to only 8% and 4% of the Left and Right designs. These data suggests that $GAS_{right}$ is a motif that is highly designable for stability, whereas the sequences in the Right and Left region are much weaker. In turn, this suggests that, at least in the conditions tested, vdW packing is not a strong driver for dimerization, and thus that the network of weak hydrogen bonds that characterizes the interaction interface of $GAS_{right}$ plays an important role for the stability of this common motif. It is however notable that even in the Right and Left regions, a small number of sequences are present whose TOXGREEN signals rise to the level corresponding to moderate dimerization propensity.

Fig. 2.2b shows the reconstructed TOXGREEN fluorescence of the Clash (red) and Large→Ala (orange) controls side by side with the original design (green). For the $GAS_{right}$ designs, the introduction of clashing mutations result in a dramatic decrease in the distribution of fluorescence. Since these mutations were designed to be incompatible with their structural model, the results are consistent with the assumption that a majority of these constructs associate according to their designed interface. The reconstructed fluorescence distribution observed for the Clash mutants of the Left and Right design regions is less dramatic (as expected because these constructs already contain a large fraction of constructs that are monomeric or weakly dimeric) but statistically significant. This suggest the positions involved participate in the predicted interface also in a significant fractions of the Left and Right constructs.

Surprisingly, the set of Large→Ala mutants do not show a general decrease in fluorescence for any of the design regions. A small but not statistically significant decrease is observed for the designs in the $GAS_{right}$ region, whereas the mutants in the Right and Left region show a small increase (also not significant). These mutants consist of individual Ala substitutions of larger (generally F, L, Y, and I) that were designed to reduce the amount of packing by a small amount (Fig. S2.5 and S2.6), although they do

not generally create cavities, given the nature of the long and narrow nature of the helix-helix interface. Large-to-small mutations are often well tolerated in mutagenesis of single pass TM dimers but they can reduce dimerization (Doura & Fleming, 2004; Fleming & Engelman, 2001; Howitt et al., 1996; Metcalf et al., 2007; Mingarro et al., 1996). For this reason we hypothesized that we would observe a shift of the fluorescence distribution. It is possible that the design procedure does not have sufficient precision for producing effect from small alterations of packing that are measurable with TOXGREEN, if the changes in dimerization propensity are small.

### 2.3.5 A mutation-validated subset confirms that vdW-based designs are weak

We used the Clash mutations to create a set of validated constructs that behave consistently with the predicted interface. To retain a design in this filtered dataset, we imposed that all of its Clash mutants are in the monomeric range using a stringent threshold as the criterion (defined as below 35% GpA TOXGREEN signal). The selection resulted in a dataset that included 384 of the original 1045 designs (37%; ST2.1, ST2.2, and ST2.3).

Fig. 2.2c shows the distribution of reconstructed TOXGREEN signal of the Clash-filtered dataset, compared to all designs. The distribution does not change for the $GAS_{right}$ designs, which retains a wide range of dimerization propensities, including constructs with fluorescent signals comparable to the strong GpA dimer. Conversely, we observe a significant reduction of fluorescence for the designs based on vdW of the Right and Left regions. This data may indicate that a substantial number of designs in these region associate through a different interface than the one designed. Fig. 2.2d compares the dimerization propensity of the mutants of the Clash-filtered designs. The Clash mutants of all designs are below 35% GpA (corresponding to 8350 reconstructed fluorescence), otherwise they would have been discarded. For the $GAS_{right}$ set, the distribution of fluorescence of the Large→Ala mutants remains similar to the distribution of the original designs, as it was for the unfiltered set. In the Left and Right regions, we

observed a small but statistically significant overall apparent stabilization of the Large→Ala mutants relative to the original designs. We speculate that Ala residues may slightly favor random weak association over larger amino acids.

### 2.3.6 $GAS_{right}$ designs follow a previous energetic model of association

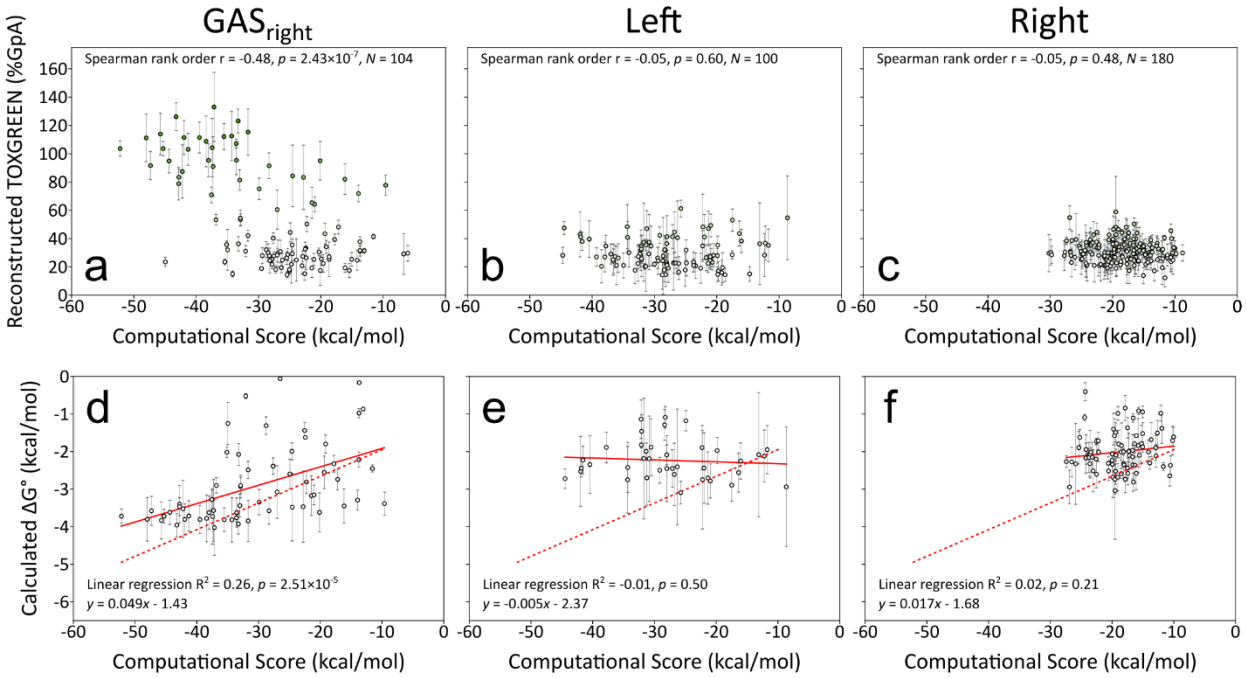



**Figure 2.3 Computational score vs dimerization propensity. a)** Design energy score plotted vs dimerization propensity in %GpA (a-c) and calculated ΔG (d-f). Spearman rank order correlation shows significance with computational score in $GAS_{right}$, but not other design regions. Dimerization propensity was converted to biophysical ΔG using methods from an *in vitro* FRET studying GASright association (Díaz Vázquez et al., 2023).

After experimentally determining the propensities of the three groups of Clash-filtered designs, we compared them to their computational energies (Tables ST1-3). Fig. 2.3a-c show scatterplots of the fluorescence of the designs against our energy score for each region. In a previous structure-based analysis of computationally predicted $GAS_{right}$ dimers, we showed they followed a simple model of association in which a combination of Cα-H hydrogen bonding and vdW packing modulates their stability (Anderson et al., 2017). We found a similar proportionality between TOXGREEN dimerization propensity and computational score for this motif (Fig. 2.3a $p$=2.43×10$^{-7}$, using Spearman Rank Order statistics because the TOXGREEN signal is not expected to be linearly dependent on dimerization free energy).

Conversely, there is no notable correlation found for the Left (Fig. 2.3b) and Right (Fig. 2.3c) regions. Since the dimerization propensity of these designs populates only the monomeric and weakly dimeric ranges of TOXGREEN signal, it is not possible to assess whether the design would follow a similar trend against the computational energetics (which, in their case, would be based primarily on vdW and a solvation term, since these constructs are design not to form inter-helical hydrogen bonds).

### 2.3.7 Conversion of TOXGREEN signals to theoretical association free energies

In the final step of our analysis, we applied a conversion factor that enabled us to estimate the free energy of association of constructs from their TOXGREEN signal. The conversion is based on a recent study that found good agreement between the TOXGREEN dimerization propensities and the ΔG° of association of a series of poly-Leu $GAS_{right}$ dimers measured in a decyl-β-maltopyranoside detergent environment using FRET (Díaz Vázquez et al., 2023). Once again, we found that there is a statistically significant linear correlation between the computational energy score of the $GAS_{right}$ designs and their calculated Δ*G*° of association (Fig. 2.3d). The linear regression fit obtained here ($y = 0.049x - 1.43$, solid line, where $y$ is the calculated Δ*G*° and $x$ is the computational score) is similar to the original model of Diaz Vazquez ($y = 0.076x - 1.33$, dashed line), indicating that the large set of $GAS_{right}$ designs of the present analysis behaves consistently with the previous set. As before, no correlation was found for the two sets of vdW-based designs (Fig. 2.3e and f). It should be noted that the fact that these designs only populate a small window in the low stability range prevent us to assess whether the energetic model would apply to these constructs.

## 2.4 Conclusion

We have performed a high-throughput analysis of transmembrane helix association to attempt to address the question of what extent vdW interactions can, as a design element, provide stability in membrane protein oligomerization and folding. The approach we took consists of a large-scale exploration of hundreds of dimers whose interfaces were designed to be mediated only by packing of non-polar side chains. These dimers were created to cover the entire geometric ranges of two common regions for helix-helix interaction, one characterized by left-handed crossing angles and the other in the right-handed region. As a comparison and control, we also designed hundreds of $GAS_{right}$ dimers, a motif whose stability is also dependent on hydrogen bonding contributions.

The design is based on a simple energy score that consists of a vdW term, hydrogen bonding and the Lazaridis IMM1 implicit solvation. This energy score was successful at predicting the structure and energetics of a smaller subset of 26 $GAS_{right}$ dimer before (Anderson et al., 2017; Díaz Vázquez et al., 2023). Those studies indicated that $GAS_{right}$ dimerization is promoted by the combination of vdW interactions and the network of backbone-to-backbone Cα-H hydrogen bonds that characterizes the motif. The vdW component represented the largest contributor of the energetic model (by about a factor of 2 over hydrogen bonding, even when the desolvation cost was considered) (Anderson et al., 2017). This suggested that in the absence of inter-helical hydrogen bonding, vdW forces would be sufficient to produce rather stable transmembrane dimers. Our experimental data suggest that this is not the case since the majority of the vdW-based designs are monomeric or weakly oligomeric at best.

A possibility is that the procedure for designing the helical dimer is somehow more effective when applied to $GAS_{right}$ configurations. In all three cases, we selected the most common regions of backbone space (and thus presumably the most conductive to stability) and applied the same design approach based on a series cycles of side chain optimization and local backbone geometry optimization, with the same

energetic score. Alternatively, a more likely possibility is that the contributions of the vdW interactions in $GAS_{right}$ association are over-estimated by our model. Finally, it may be possible that vdW forces and hydrogen bonding work cooperatively in $GAS_{right}$ dimers, possibly at the expense of some of the entropic costs of dimer association, which are not considered in our rigid body model.

Overall, the data confirms that $GAS_{right}$, with its combination of packing and weak hydrogen bonding is indeed a special configuration for designing stable transmembrane helical pairs. It also suggests that vdW packing alone is not a strong enough force for designing strong dimers. It should be noted that among the designs in the Left and Right regions there are instances of designs that appear to rise to the level of moderately stable dimers (over 60% of GpA TOXGREEN signal), although most of them were removed by our stringent Clash-filtering procedure. Assessing those dimers individually (measuring their individual TOXGREEN signal, controlling for their expression and validating their conformation by extensive mutagenesis) would be laborious and it is outside of the scope of this work. However, future studies could address the possibility that, when fully optimized, packing of non-polar side chains leads to biologically significant stability.

## 2.5 Methods

### 2.5.1 Membrane protein helical pair extraction

We developed a program in MSL that extracts helical pairs from PDBs from OPM. To ensure that we don't extract redundant helical pairs, the MP structures from OPM were trimmed by sequence similarity. Only unique structures with less than 30% sequence similarity (Steinegger & Söding, 2017) were analyzed. We first identify the top and bottom z-axes of the membrane in the OPM structure. The segments of the protein within the membrane are then assessed for their helical nature. Cartesian points for quadruplets of Cα carbons are then assessed for their helical nature. The height (1.25-1.75Å), twist (90-110°), and radius (2.12-2.42Å) are measured, with loose restrictions against the ideal values (1.5Å, 100°, 2.27Å) for each parameter. Helical segments composed of at least 13 AAs in length are extracted as individual helices, and the distance is measured between Cα carbons on each unique helical pair. Any two helices with at least 3 Cα carbons within 9Å of each other are extracted as an individual helical pair. The crossing angles and distances between these pairs are plotted on Fig. 2.1b.

### 2.5.2 Computational Sequence Design

The algorithm designs an interfacial sequence along a given poly-Leu backbone geometry. Using a Monte Carlo sequence optimization, an interfacial position is switched to a random amino acid most prevalently found in membrane protein sequences. The side chain mobility was modeled using the energy based conformer library, with interfacial side chains given higher mobility than non-interfacial (Subramaniam & Senes, 2012). To ensure that these designed sequences are membrane sequences, we developed a sequence entropy term. The sequence entropy term evaluates the current interface as it's likelihood to occur in membrane protein sequences, and converts this likelihood to an energy applied during the sequence search. Energetics for the dimer were calculated using the CHARMM22 vdW function, the IMM1 membrane implicit solvation term, and the hydrogen bonding function SCWRL4. To estimate

the free energy of the sequences during the search, we typically calculate the energy of a monomeric helix with the same sequence and subtract it from the dimer energy made of only these energetic terms.

$$\textit{Eq. 2.1} \quad Dimer = CHARMM\ VDW + IMM\ 1 + SCWRL\ 4\ HBOND\,(for\ 2\ helices)$$

$$\textit{Eq. 2.2} \quad Monomer = CHARMM\ VDW + IMM\ 1 + SCWRL\ 4\ HBOND\,(for\ 1\ helix)$$

$$\textit{Eq. 2.3} \quad Total\ Energy = Dimer - (2 \times Monomer)$$

We found that computing the monomer energy created a bottleneck during our sequence search. To reduce computational time, we developed an energy term that estimates the monomer energy for each designed sequence. This baseline monomer energy was determined by measuring the CHARMM_VDW, IMM1, and SCWRL4_HBOND energies for individual amino acids on a monomeric helix. The self (single AA) and pair (two AAs) energies were computed for 10000 random sequences and averaged. We found that there was a strong correlation between the computed monomer energy and this baseline monomer energy and utilized this energy during our sequence search. The baseline monomer energy and the sequence entropy are subtracted from the dimer energy to find the sequence with the most stable energy during the sequence search.

$$\textit{Eq. 2.4}$$
$$Search\ Energy = Dimer - (2 \times BASELINE\ MONOMER) - (SEQUENCE\ ENTROPY \times Weight)$$

After the sequence is designed for the geometry, the structure undergoes Monte Carlo backbone optimization cycles where all parameters (distance, z-shift, axial rotation, and crossing angle) are varied. The final energy from backbone optimization is used to assess our designed constructs against their dimerization propensity. The total energy (Eq. 2.3) for the refined structure is used as the acceptance criteria. These energies are used to evaluate our sequences against their dimerization propensity determined in sort-seq.

### 2.5.3 Sequence Entropy

We developed a sequence entropy term that outputs an energy based on how similar a sequence is to the composition of natural membrane protein sequences. Previous studies have determined the amino acid composition of membrane proteins (Liu et al., 2002), but we chose to evaluate the composition of amino acids found in our helical pair TMs (Fig. S2.2). We first removed all amino acids that were represented less than 2% (C, P, H). We then trimmed the amino acid pool for amino acids likely to form disulfide bonds (M), hydrogen bonds (N, D, Q, E) except for a few that were well represented (S, T, Y), and charged interactions (K, R), leaving us with a pool of 10 amino acids for design (Fig. S2.2). To convert the frequency of AAs in a membrane sequence to an energy term, we utilized the following equation based on the Boltzmann entropy formula:

$$\textit{Eq. 2.5} \qquad SEQUENCE\ ENTROPY = -\log(probability) \times RT$$

where R is the gas constant and T is temperature defaulted to 298K (RT = 0.592). To compute the sequence entropy, we calculated the probability that the sequence occurs in membrane. First, the number of each AA (AA1, AA2, etc.) is counted within the sequence. Using these values, we calculated the number of possible permutations for the sequence. This is determined using the following equation:

$$\textit{Eq. 2.6} \qquad permutations = \frac{n!}{(AA1! \times AA2! \times \ldots)}$$

where n is the number of positions, which is divided by the factorial for #AA in the sequence multiplied, or the total number of combinations possible. The probability is computed using the frequency of each AA in membrane protein sequences ($freq_{AA}$) to the power of the number of each AA in the sequence multiplied by the permutations:

$$\textit{Eq. 2.7} \qquad probability = (freq_{AA}1^{AA1} \times freq_{AA}2^{AA2} \times \ldots) \times permutations$$

This probability is inserted into the sequence entropy equation, returning a value that can be applied as an energy term during the sequence search.

### 2.5.4 Left-Handed Interfaces

During visual inspection of our randomized backbone geometries, we found that left-handed dimers were able to accommodate multiple interfaces. Interfaces differ based on the input axial rotation and z-shift (Fig. S2.10). LLL**xx**LL**xx**L**xx**LL**xx**LLILI: 0-40° axial rotation and 1-6Å z-shift (left stripe). LLLL**xx**L**xx**LL**xx**L**xx**LLILI: 30-90° axial rotation and 0-6Å z-shift (middle stripe). LLLL**xx**LL**xx**L**xx**LL**xx**LILI: 80-100° axial rotation and 0-2Å z-Shift (right stripe).

### 2.5.5 Cloning for bacterial cell expression

Sequences were ordered as an oligo pool from Twist Biosciences. Individual segments were amplified using qPCR (Roche KAPA SYBR Fast). Segments were digested using restriction enzymes NheI-HF and DpnII. TOXGREEN vector was digested in preparation for ligation using restriction enzymes NheI-HF, BamHI-HF, and CIP. Plasmids were assembled by incubating segment DNA with TOXGREEN vector in a 1:10 backbone:insert ratio with ElectroLigase for 2 hours at room temperature and 10 minutes at 65°C to inactivate the enzyme. 5uL of the ligation mixture were added to 50uL of NEB DH10B *E. coli* cells and cloned through electroporation (BIO-RAD MicroPulser Electroporator), outgrown for 1 hour in 950uL NEB SOM media in 37°C shaker, and diluted 1mL in 4mL LB Amp 100. 50 uL of a 1:10 and 1:100 dilution were plated on LB Amp 100 plates. Plates were grown for 14-16 hours at 37°C and colony forming units (CFUs) calculated for each segment (# Colonies x dilution/# Sequences in segment). A CFU > 5 was accepted and overnight cultures corresponding to plates were spun down and miniprepped (Qiagen). Segment plasmids were cloned through electroporation into 50uL in house mm39 cells for fluorescence readout, outgrown for 1 hour in 950uL SOC media in 37°C shaker, and 100uL of 1:10 dilution plated on LB Amp 100 plates with the rest being grown overnight in 3mL LB Amp 100. CFUs were again counted and any plates > 5 were

accepted. Overnights for each segment were stored in 25% glycerol stocks. 50uL samples and controls from glycerol stocks were grown in 3mL LB Amp100 for 2-4 hours to reach ~0.1 OD600 and the corresponding number of sequences per sample used to calculate how much to add to a full library glycerol stock. All enzymes acquired from NEB.

### 2.5.6 Sort-Seq and NGS Preparation Protocol

50uL of library glycerol stock were grown in 3mL LB Amp 100 overnight for 14-16 hours in 37C shaker. Samples were diluted in PBS buffer to appropriate concentration for recording 10000 events per second in Sony MA900 fluorescence activated cell sorter (UWCCC Flow Cytometry Laboratory). Controls GpA, G83I, and NoTM were flowed to calibrate the instrument and determine proper gating to remove dead cells. Individual sample libraries were first flowed through the instrument and the fluorescence profile separated into 4 bins. 100000 events were sorted into bins 1-3, and 50000 events sorted into bin 4. Each library sample was sorted in triplicate from biological replicates of overnights. Sorted populations were grown for 1-5 hours to reach ~0.3 OD600 and minipreppred. Samples were prepared for NGS using PCR and amplified with individual primers. The samples were then sent to the DNA Sequencing center at UW Biotech Center for library preparation and next generation sequencing. Library Preparation Services: Index PCR (TruSeq). Sequencing Services: Illumina (NovaSeq) Sequencing [2x150 Shared (10M read increments)].

### 2.5.7 Immunoblotting

Protein expression was confirmed for an assortment of designs and mutants using immunoblotting (Fig. S2.1). Cell lysates were normalized by ThermoScientific Pierce Protein BCA Assay Kit and loaded onto NuPage 4-12% bis-tris SDS-PAGE gels (ThermoFisher) and then transferred to PVDF membranes (VWR) for 1h at 100 millivolts. Blots were blocked using 5% milk (Nestle Instant Nonfat Dry Milk) in TBS-Tween buffer (50 mM Tris, 150 mM NaCl, 0.05% Tween 20) for 2 h at 4°C and incubated overnight with peroxidase-

conjugated monoclonal anti-maltose binding protein antibodies (Sigma-Aldrich). Blots were developed with Pierce ECL Western Blotting Substrate Kit; 1mL of ECL solution was added to the blot and incubated for 90s. Chemiluminescence was measured using an ImageQuant LAS 4000 (GE Healthsciences).

## 2.6 Supplementary Information

### Table ST2.1 Validated set of $GAS_{right}$ designs

| Design ID | Sequence | Computational Score (kcal/mol) | GpA (%) | ΔG (kcal/mol) |
|---|---|---|---|---|
| G_001 | LLL**TA**LL**VG**LL**GG**LL**FL**LILI | -24.58 | 0.29 ± 0.05 | NA |
| G_002 | LLL**LF**LL**AT**LL**GY**LL**SI**LILI | -31.69 | 0.42 ± 0.04 | -2.49 ± 0.23 |
| G_003 | LLL**LF**LL**GT**LL**GA**LL**AY**LILI | -31.69 | 1.15 ± 0.16 | -3.85 ± 0.54 |
| G_004 | LLL**VT**LL**FS**LL**GA**LI**SL**LILI | -17.19 | 0.48 ± 0.05 | -2.74 ± 0.28 |
| G_005 | LLL**YI**LL**GL**LL**GT**LI**AS**LILI | -32.08 | 0.31 ± 0.04 | -0.52 ± 0.07 |
| G_006 | LLL**LV**LL**TG**LL**FG**LL**YA**LILI | -26.51 | 0.31 ± 0.08 | -0.06 ± 0.01 |
| G_007 | LLL**FA**LL**AS**LL**GI**LL**TL**LILI | -26.35 | 0.19 ± 0.04 | NA |
| G_008 | LLL**TF**LL**AV**LL**GL**LL**GI**LILI | -21.45 | 0.65 ± 0.11 | -3.18 ± 0.53 |
| G_009 | LLL**SF**LL**VL**LL**GT**LI**AY**LILI | -21.46 | 0.19 ± 0.07 | NA |
| G_010 | LLL**YL**LL**GA**LL**GT**LI**SF**LILI | -32.9 | 0.53 ± 0.04 | -2.9 ± 0.22 |
| G_011 | LLL**TV**LL**GL**LL**GL**LL**GF**LILI | -32.92 | 0.54 ± 0.06 | -2.93 ± 0.3 |
| G_012 | LLL**YF**LL**GT**LL**GA**LL**GL**LILI | -33.02 | 0.81 ± 0.07 | -3.44 ± 0.3 |
| G_013 | LLL**AV**LL**TS**LL**GG**LL**FL**LILI | -21.13 | 0.25 ± 0.04 | NA |
| G_014 | LLL**VT**LL**LG**LL**FA**LI**ST**LILI | -26.16 | 0.25 ± 0.05 | NA |
| G_015 | LLL**TF**LL**GL**LL**GY**LI**AS**LILI | -33.27 | 1.23 ± 0.08 | -3.93 ± 0.27 |
| G_016 | LLL**YV**LL**GA**LL**GI**LI**TL**LILI | -33.55 | 0.95 ± 0.1 | -3.62 ± 0.39 |
| G_017 | LLL**YF**LL**GV**LL**GT**LI**GL**LILI | -33.58 | 1.07 ± 0.03 | -3.76 ± 0.11 |
| G_018 | LLL**YA**LL**AL**LL**GT**LL**GF**LILI | -16.12 | 0.19 ± 0.03 | NA |
| G_019 | LLL**YL**LL**AS**LL**GI**LI**AF**LILI | -34.17 | 0.15 ± 0.02 | NA |
| G_020 | LLL**YL**LL**GA**LL**GA**LI**TF**LILI | -34.32 | 1.12 ± 0.17 | -3.82 ± 0.59 |
| G_021 | LLL**TV**LL**LG**LL**FG**LL**FA**LILI | -22.26 | 0.5 ± 0.05 | -2.81 ± 0.3 |
| G_022 | LLL**TV**LL**AL**LL**GF**LL**GS**LILI | -9.61 | 0.78 ± 0.07 | -3.39 ± 0.3 |
| G_023 | LLL**YV**LL**TL**LL**GA**LL**GF**LILI | -22.36 | 0.26 ± 0.08 | NA |
| G_024 | LLL**YV**LL**GL**LL**GT**LI**AF**LILI | -34.97 | 0.32 ± 0.14 | -1.25 ± 0.56 |
| G_025 | LLL**YF**LL**AT**LL**GL**LL**GV**LILI | -35.12 | 0.36 ± 0.02 | -2.02 ± 0.14 |
| G_026 | LLL**VG**LL**FA**LL**GI**LI**TL**LILI | -33.23 | 0.36 ± 0.05 | -2.08 ± 0.29 |
| G_027 | LLL**TF**LL**AI**LL**GL**LL**GV**LILI | -21.01 | 0.64 ± 0.05 | -3.15 ± 0.26 |
| G_028 | LLL**AA**LL**AL**LL**GV**LL**GT**LILI | -17.8 | 0.39 ± 0.04 | -2.33 ± 0.22 |
| G_029 | LLL**AA**LL**FS**LL**GL**LL**GT**LILI | -26.92 | 0.3 ± 0.09 | NA |
| G_030 | LLL**YA**LL**TA**LL**GG**LL**FL**LILI | -19.63 | 0.22 ± 0.06 | NA |
| G_031 | LLL**LT**LL**VG**LL**FA**LL**YG**LILI | -27.85 | 0.25 ± 0.02 | NA |
| G_032 | LLL**YI**LL**GL**LL**GT**LI**AF**LILI | -28.2 | 0.25 ± 0.09 | NA |
| G_033 | LLL**LV**LL**AT**LL**GG**LI**SI**LILI | -19.93 | 0.26 ± 0.05 | NA |
| G_034 | LLL**YG**LL**FA**LL**GI**LI**TL**LILI | -28.27 | 0.28 ± 0.01 | NA |
| G_035 | LLL**ST**LL**GA**LL**GL**LI**AV**LILI | -28.31 | 0.91 ± 0.09 | -3.58 ± 0.36 |
| G_036 | LLL**YV**LL**FT**LL**GA**LL**GL**LILI | -27.82 | 0.22 ± 0.03 | NA |
| G_037 | LLL**YV**LL**AT**LL**GL**LL**GF**LILI | -28.48 | 0.3 ± 0.06 | NA |
| G_038 | LLL**YL**LL**GT**LL**GI**LI**AF**LILI | -19.35 | 0.43 ± 0.07 | -2.55 ± 0.43 |
| G_039 | LLL**YG**LL**FL**LL**GV**LI**AT**LILI | -27.64 | 0.4 ± 0.04 | -2.39 ± 0.22 |
| G_040 | LLL**YF**LL**AT**LL**GV**LL**GL**LILI | -19.14 | 0.34 ± 0.05 | -1.8 ± 0.24 |
| G_041 | LLL**YL**LL**TT**LL**AG**LL**FS**LILI | -20.09 | 0.17 ± 0.1 | NA |
| G_042 | LLL**TF**LL**AL**LL**GY**LL**GS**LILI | -20.12 | 0.95 ± 0.14 | -3.62 ± 0.52 |
| G_043 | LLL**VT**LL**LT**LL**GA**LI**SA**LILI | -27.56 | 0.28 ± 0.03 | NA |
| G_044 | LLL**LT**LL**VA**LL**AG**LL**FS**LILI | -20.31 | 0.31 ± 0.03 | NA |
| G_045 | LLL**YG**LL**FL**LL**GV**LI**AS**LILI | -28.77 | 0.32 ± 0.06 | -1.31 ± 0.23 |
| G_046 | LLL**YV**LL**AL**LL**GT**LL**GF**LILI | -27.25 | 0.18 ± 0.03 | NA |
| G_047 | LLL**TA**LL**VG**LL**GL**LL**GF**LILI | -18.68 | 0.27 ± 0.02 | NA |
| G_048 | LLL**YL**LL**GA**LL**GV**LI**ST**LILI | -24.54 | 0.84 ± 0.22 | -3.48 ± 0.9 |
| G_049 | LLL**LA**LL**FA**LL**GI**LI**ST**LILI | -18.68 | 0.25 ± 0.12 | NA |
| G_050 | LLL**YF**LL**AV**LL**GL**LL**GT**LILI | -29.4 | 0.28 ± 0.06 | NA |
| G_051 | LLL**YG**LL**FA**LL**GT**LI**SL**LILI | -29.53 | 0.19 ± 0.01 | NA |
| G_052 | LLL**VA**LL**AL**LL**GT**LI**SF**LILI | -27.01 | 0.28 ± 0.05 | NA |
| G_053 | LLL**YL**LL**GA**LL**GI**LI**ST**LILI | -29.9 | 0.75 ± 0.08 | -3.35 ± 0.34 |
| G_054 | LLL**TF**LL**AV**LL**GL**LI**GS**LILI | -27.01 | 0.6 ± 0.14 | -3.08 ± 0.7 |
| G_055 | LLL**YA**LL**FL**LL**GT**LI**SI**LILI | -15.45 | 0.17 ± 0.05 | NA |
| G_056 | LLL**YF**LL**AL**LL**GT**LL**GS**LILI | -35.41 | 0.23 ± 0.02 | NA |
| G_057 | LLL**TA**LL**FL**LL**GV**LI**AS**LILI | -16.15 | 0.82 ± 0.11 | -3.45 ± 0.45 |
| G_058 | LLL**AG**LL**FL**LL**GV**LI**AT**LILI | -22.42 | 0.33 ± 0.01 | -1.62 ± 0.03 |
| G_059 | LLL**YV**LL**AT**LL**GG**LL**FL**LILI | -23.03 | 0.21 ± 0.06 | NA |
| G_060 | LLL**TA**LL**LA**LL**FG**LL**FS**LILI | -13.75 | 0.38 ± 0.03 | -2.21 ± 0.2 |
| G_061 | LLL**YV**LL**TL**LL**GG**LL**FA**LILI | -13.73 | 0.32 ± 0.04 | -0.98 ± 0.12 |
| G_062 | LLL**AY**LL**GT**LL**GL**LL**GF**LILI | -45.74 | 1.14 ± 0.15 | -3.83 ± 0.49 |
| G_063 | LLL**VL**LL**AT**LL**GS**LL**GT**LILI | -13.7 | 0.31 ± 0.07 | -0.17 ± 0.04 |
| G_064 | LLL**AL**LL**GL**LL**GT**LI**SF**LILI | -45.31 | 1.04 ± 0.05 | -3.72 ± 0.18 |
| G_065 | LLL**LV**LL**TT**LL**GG**LL**FA**LILI | -25.32 | 0.29 ± 0.04 | NA |
| G_066 | LLL**AY**LL**GL**LL**GF**LL**GT**LILI | -39.49 | 1.11 ± 0.11 | -3.81 ± 0.37 |
| G_067 | LLL**YT**LL**AI**LL**GL**LL**GF**LILI | -44.96 | 0.24 ± 0.03 | NA |
| G_068 | LLL**TS**LL**VG**LL**AY**LL**GL**LILI | -11.56 | 0.41 ± 0.02 | -2.45 ± 0.09 |
| G_069 | LLL**AY**LL**GV**LL**GL**LI**AF**LILI | -35.55 | 1.12 ± 0.09 | -3.81 ± 0.31 |
| G_070 | LLL**YF**LL**AV**LL**GL**LL**GS**LILI | -23.94 | 0.25 ± 0.15 | NA |

| | | | | |
|---|---|---|---|---|
| G_071 | LLLSYLLGVLLGLLLGTLILI | -44.36 | 0.95 ± 0.08 | -3.62 ± 0.32 |
| G_072 | LLLALLLGTLLGFLLSILILI | -43.23 | 1.26 ± 0.1 | -3.96 ± 0.31 |
| G_073 | LLLTALLAGLLFLLLSVLILI | -13.06 | 0.31 ± 0.02 | -0.87 ± 0.05 |
| G_074 | LLLSYLLGVLLGLLLATLILI | -41.31 | 1.03 ± 0.11 | -3.71 ± 0.41 |
| G_075 | LLLSYLLGVLLGTLLGFLILI | -42.81 | 0.79 ± 0.12 | -3.4 ± 0.5 |
| G_076 | LLLSYLLGLLLGTLLGFLILI | -42.79 | 0.83 ± 0.06 | -3.47 ± 0.24 |
| G_077 | LLLYLLLGALLAFLLSTLILI | -25.05 | 0.18 ± 0.01 | NA |
| G_078 | LLLTYLLGVLLGLLLGFLILI | -42.23 | 0.87 ± 0.19 | -3.52 ± 0.76 |
| G_079 | LLLLVLLGALLGTLLSFLILI | -24.93 | 0.44 ± 0.07 | -2.6 ± 0.39 |
| G_080 | LLLAYLLGVLLGLLLGSLILI | -41.95 | 1.11 ± 0.12 | -3.81 ± 0.41 |
| G_081 | LLLALLLGVLLGTLLGFLILI | -38.02 | 0.95 ± 0.12 | -3.62 ± 0.44 |
| G_082 | LLLSYLLGLLLGFLLSTLILI | -38.42 | 1.09 ± 0.18 | -3.78 ± 0.62 |
| G_083 | LLLYLLLGALLGFLLSTLILI | -48.02 | 1.11 ± 0.17 | -3.8 ± 0.58 |
| G_084 | LLLYSLLFALLGALLTLLILI | -24.61 | 0.36 ± 0.04 | -1.99 ± 0.22 |
| G_085 | LLLFVLLTILLGALLSLLILI | -15.06 | 0.25 ± 0.05 | NA |
| G_086 | LLLYTLLVALLALLLGLLILI | -22.53 | 0.33 ± 0.05 | -1.44 ± 0.22 |
| G_087 | LLLYSLLALLLGTLLGFLILI | -22.59 | 0.15 ± 0.05 | NA |
| G_088 | LLLALLLTVLLGGLLFSLILI | -25.5 | 0.26 ± 0.04 | NA |
| G_089 | LLLTSLLLGLLFALLAYLILI | -24.73 | 0.22 ± 0.13 | NA |
| G_090 | LLLYLLLGILLGTLLAFLILI | -36.83 | 0.53 ± 0.04 | -2.9 ± 0.2 |
| G_091 | LLLVTLLLGLLFGLLFALILI | -22.79 | 0.83 ± 0.23 | -3.47 ± 0.95 |
| G_092 | LLLYILLTLLLGALLAFLILI | -6.71 | 0.29 ± 0.14 | NA |
| G_093 | LLLFYLLGVLLGLLLASLILI | -47.35 | 0.92 ± 0.1 | -3.58 ± 0.39 |
| G_094 | LLLYVLLFLLLGALLGTLILI | -25.05 | 0.17 ± 0.02 | NA |
| G_095 | LLLTLLLGLLLGVLLGSLILI | -37.29 | 0.91 ± 0.21 | -3.57 ± 0.83 |
| G_096 | LLLALLLGLLLGVLLGSLILI | -37.44 | 1.04 ± 0.21 | -3.73 ± 0.74 |
| G_097 | LLLTVLLGILLGLLLAFLILI | -37.57 | 0.71 ± 0.06 | -3.28 ± 0.25 |
| G_098 | LLLSSLLFLLLGVLLGTLILI | -6.06 | 0.3 ± 0.06 | NA |
| G_099 | LLLVLLLTGLLFALLAYLILI | -14.18 | 0.25 ± 0.07 | NA |
| G_100 | LLLAVLLFSLLGLLLGTLILI | -22.9 | 0.27 ± 0.04 | NA |
| G_101 | LLLTALLFLLLGVLLGILILI | -13.97 | 0.72 ± 0.06 | -3.29 ± 0.27 |
| G_102 | LLLAGLLFLLLGTLLSYLILI | -25.42 | 0.14 ± 0.02 | NA |
| G_103 | LLLYVLLGLLLGFLLSTLILI | -52.21 | 1.04 ± 0.05 | -3.72 ± 0.19 |
| G_104 | LLLAFLLGTLLGVLLGLLILI | -37.15 | 1.33 ± 0.24 | -4.02 ± 0.74 |

## Table ST2.2 Validated set of Left designs

| Design ID | Sequence | Computational Score (kcal/mol) | GpA (%) | ΔG (kcal/mol) |
|---|---|---|---|---|
| L_001 | LLLLLLLTALLIFLVSLLILI | -24.86 | 0.23 ± 0.09 | NA |
| L_002 | LLLLYVLIALLTLLFSLLILI | -19.91 | 0.28 ± 0.05 | NA |
| L_003 | LLLVALLLILTFLLASLLILI | -19.85 | 0.35 ± 0.06 | -1.97 ± 0.35 |
| L_004 | LLLLFVLTILLALLYSLLILI | -27.78 | 0.19 ± 0.1 | NA |
| L_005 | LLLLAYLLALLFLLTSLLILI | -24.92 | 0.32 ± 0.07 | -1.18 ± 0.27 |
| L_006 | LLLYALLFVLTALLLSLLILI | -20.96 | 0.49 ± 0.08 | -2.78 ± 0.45 |
| L_007 | LLLLYVLLATLLALLSFLILI | -20.52 | 0.28 ± 0.08 | NA |
| L_008 | LLLLAILLTLLAVLFSLLILI | -25.76 | 0.61 ± 0.06 | -3.09 ± 0.3 |
| L_009 | LLLVALLLLLFLLLTSLLILI | -22.22 | 0.48 ± 0.23 | -2.75 ± 1.31 |
| L_010 | LLLYFLLAILVTLLLSLLILI | -22.59 | 0.19 ± 0.11 | NA |
| L_011 | LLLLFVLLALLTILLSLLILI | -19.85 | 0.15 ± 0.05 | NA |
| L_012 | LLLLAVLLFLLATLISLLILI | -26.1 | 0.22 ± 0.05 | NA |
| L_013 | LLLLLLLLFILAVLLSTLILI | -26.11 | 0.16 ± 0.05 | NA |
| L_014 | LLLLYVLLATLLFLLSLLILI | -21.98 | 0.41 ± 0.16 | -2.44 ± 0.94 |
| L_015 | LLLLTLLVALLLFLISLLILI | -26.31 | 0.41 ± 0.26 | -2.4 ± 1.54 |
| L_016 | LLLLYFLLALLAVLTSLLILI | -21.69 | 0.27 ± 0.09 | NA |
| L_017 | LLLYLLLTVLALLLFSLLILI | -21.39 | 0.23 ± 0.06 | NA |
| L_018 | LLLLTLLLALIVLLFSLLILI | -21.35 | 0.47 ± 0.12 | -2.7 ± 0.69 |
| L_019 | LLLLVLLLTILFLLLSALILI | -26.56 | 0.23 ± 0.04 | NA |
| L_020 | LLLLYLLLFILAVLLSTLILI | -23.39 | 0.21 ± 0.09 | NA |
| L_021 | LLLLLFLLATLLILLSVLILI | -21.08 | 0.25 ± 0.05 | NA |
| L_022 | LLLLYLLLAVLTALLFSLILI | -21.06 | 0.27 ± 0.09 | NA |
| L_023 | LLLLYVLLALLFTLISLLILI | -25.06 | 0.18 ± 0.12 | NA |
| L_024 | LLLVTLLLALLLLLFSLLILI | -20.94 | 0.27 ± 0.06 | NA |
| L_025 | LLLLYALLFLLTALLFSLILI | -26.79 | 0.45 ± 0.14 | -2.61 ± 0.83 |
| L_026 | LLLLLLLLFVLATLLSILILI | -26.85 | 0.23 ± 0.07 | NA |
| L_027 | LLLFSLLVLLTALLYLLLILI | -20.71 | 0.23 ± 0.05 | NA |
| L_028 | LLLLTILLLVLALLLSFLILI | -26.94 | 0.23 ± 0.19 | NA |
| L_029 | LLLLFLLLAILLVLLSTLILI | -23.7 | 0.21 ± 0.05 | NA |
| L_030 | LLLLYVLLALLFTLLSILILI | -23.75 | 0.23 ± 0.07 | NA |
| L_031 | LLLLAVLLFLLALLTSLLILI | -27.1 | 0.41 ± 0.08 | -2.45 ± 0.5 |
| L_032 | LLLVALLTILLALLFSLLILI | -27.98 | 0.41 ± 0.15 | -2.45 ± 0.88 |
| L_033 | LLLVSLLILLTALLFLLLILI | -22.28 | 0.35 ± 0.11 | -1.89 ± 0.59 |
| L_034 | LLLLFALLTILVALLSLLILI | -28.01 | 0.36 ± 0.05 | -2.09 ± 0.26 |
| L_035 | LLLLAVLTALLLLLFSLLILI | -36.08 | 0.21 ± 0.13 | NA |
| L_036 | LLLLYVLTALLFLLISLLILI | -35.8 | 0.26 ± 0.1 | NA |
| L_037 | LLLLVLLLTILLFLLSALILI | -34.96 | 0.21 ± 0.09 | NA |
| L_038 | LLLLWYLVSLLILLTFLLILI | -34.36 | 0.48 ± 0.16 | -2.75 ± 0.89 |
| L_039 | LLLLVSLLILLTALLFLLILI | -34.32 | 0.41 ± 0.11 | -2.42 ± 0.67 |
| L_040 | LLLLAYLTVLLSILFLLILI | -16.04 | 0.38 ± 0.08 | -2.25 ± 0.48 |
| L_041 | LLLLAILLALLFVLTSLLILI | -34.1 | 0.3 ± 0.1 | NA |
| L_042 | LLLLFVLLLLLATLISLLILI | -36.31 | 0.27 ± 0.08 | NA |
| L_043 | LLLYTLLAVLIALLLSLLILI | -16.38 | 0.44 ± 0.09 | -2.56 ± 0.52 |
| L_044 | LLLLFILLAVLLTLLLSLILI | -32.64 | 0.22 ± 0.13 | NA |
| L_045 | LLLVALLLFLLSLLYTLLILI | -32.59 | 0.27 ± 0.08 | NA |
| L_046 | LLLLVLLLAILFTLISLLILI | -32.33 | 0.29 ± 0.07 | NA |
| L_047 | LLLVALLFLLTSLLYLLLILI | -27.98 | 0.21 ± 0.07 | NA |
| L_048 | LLLLFLLLAVLYLLLSTLILI | -32.17 | 0.32 ± 0.14 | -1.14 ± 0.51 |
| L_049 | LLLLSYLLAVLFTLLLILILI | -32.13 | 0.34 ± 0.08 | -1.82 ± 0.4 |
| L_050 | LLLLFLLLLALTVLLSILILI | -32.09 | 0.33 ± 0.1 | -1.46 ± 0.44 |
| L_051 | LLLYFLLAVLITLLLSLLILI | -33.12 | 0.18 ± 0.09 | NA |
| L_052 | LLLLTYLLAILFVLLLSLILI | -32.09 | 0.25 ± 0.12 | NA |
| L_053 | LLLVLLLAILFLLLTSLLILI | -36.54 | 0.24 ± 0.09 | NA |
| L_054 | LLLYTLLAVLLLLLASLLILI | -14.73 | 0.15 ± 0.05 | NA |
| L_055 | LLLYALLLTLIVLLFSLLILI | -8.65 | 0.55 ± 0.3 | -2.94 ± 1.59 |
| L_056 | LLLLLLLLAVLTALLFILILI | -44.72 | 0.28 ± 0.06 | NA |
| L_057 | LLLLVALLLLTSLLFLLILI | -44.48 | 0.48 ± 0.05 | -2.72 ± 0.26 |
| L_058 | LLLYALLLVLITLLFSLLILI | -11.73 | 0.35 ± 0.12 | -1.95 ± 0.64 |
| L_059 | LLLFALLLILTALLYVLLILI | -12.21 | 0.37 ± 0.12 | -2.12 ± 0.69 |
| L_060 | LLLLFYLTALLILLVSLLILI | -12.34 | 0.28 ± 0.15 | NA |
| L_061 | LLLLVALLLLLTALLFILILI | -41.95 | 0.43 ± 0.16 | -2.53 ± 0.93 |
| L_062 | LLLLAILLAVLFLLLSTLILI | -36.63 | 0.25 ± 0.11 | NA |
| L_063 | LLLLLALLFVLTSLLYLLILI | -41.84 | 0.41 ± 0.04 | -2.44 ± 0.24 |
| L_064 | LLLLYVLIALLLLLTSLLILI | -13.11 | 0.36 ± 0.29 | -2.08 ± 1.65 |
| L_065 | LLLLYWLLTILAFLLSLLILI | -40.46 | 0.4 ± 0.13 | -2.34 ± 0.77 |
| L_066 | LLLLYSLTALLFVLLLLILI | -39.3 | 0.27 ± 0.06 | NA |
| L_067 | LLLLTVLLALLFLLISLLILI | -38.49 | 0.2 ± 0.12 | NA |
| L_068 | LLLLVALLLLLTSLIFLLILI | -37.95 | 0.27 ± 0.11 | NA |
| L_069 | LLLLFVLLAILYLLLSTLILI | -37.79 | 0.35 ± 0.08 | -1.89 ± 0.41 |
| L_070 | LLLFALLLLLVTLLYSLLILI | -36.99 | 0.29 ± 0.07 | NA |

| | | | | |
|---|---|---|---|---|
| L_071 | LLLLFWLLAVLTALLYLLILI | -41.58 | 0.38 ± 0.06 | -2.23 ± 0.37 |
| L_072 | LLLLAVLLLTLALLSFLILI | -32.06 | 0.26 ± 0.03 | NA |
| L_073 | LLLLVALLLILFLLLSTLILI | -32.21 | 0.28 ± 0.08 | NA |
| L_074 | LLLLSYLVALLLTLFLLLILI | -31.72 | 0.37 ± 0.28 | -2.18 ± 1.61 |
| L_075 | LLLLLVLLAILLFLLSTLILI | -29.41 | 0.24 ± 0.1 | NA |
| L_076 | LLLLYWLLTLLVALLSFLILI | -29.15 | 0.42 ± 0.06 | -2.5 ± 0.34 |
| L_077 | LLLYLLLAVLTSLLFLLLILI | -28.74 | 0.14 ± 0.1 | NA |
| L_078 | LLLLALLLLFLTVLLSILILI | -28.66 | 0.24 ± 0.07 | NA |
| L_079 | LLLLLFLTALLYVLISLLILI | -28.61 | 0.28 ± 0.27 | NA |
| L_080 | LLLLYVLLLALTILLFSLILI | -19.09 | 0.19 ± 0.11 | NA |
| L_081 | LLLLYLLLAVLLSLLFTLILI | -29.49 | 0.26 ± 0.08 | NA |
| L_082 | LLLYFLLAVLLTLLFSLLILI | -19.2 | 0.16 ± 0.09 | NA |
| L_083 | LLLLLALLLILVTLLSFLILI | -28.35 | 0.32 ± 0.1 | -1.29 ± 0.41 |
| L_084 | LLLLWTLIALLLVLFSLLILI | -28.25 | 0.32 ± 0.08 | -1.09 ± 0.29 |
| L_085 | LLLYALLFLLTSLLFSLLILI | -28.21 | 0.14 ± 0.08 | NA |
| L_086 | LLLLLLLLAVLYTLLSFLILI | -28.07 | 0.23 ± 0.04 | NA |
| L_087 | LLLYALLLLLVTLLFSLLILI | -28.03 | 0.22 ± 0.08 | NA |
| L_088 | LLLYALLFVLLLLTSLLILI | -19.75 | 0.17 ± 0.07 | NA |
| L_089 | LLLLAILLLLLTVLFSLLILI | -28.43 | 0.18 ± 0.08 | NA |
| L_090 | LLLLYILLALLFVLLSTLILI | -30 | 0.22 ± 0.02 | NA |
| L_091 | LLLYALLFVLLILLLFSLLILI | -18.68 | 0.14 ± 0.05 | NA |
| L_092 | LLLLALLFLLTSLLYVLILI | -31.34 | 0.36 ± 0.04 | -1.99 ± 0.22 |
| L_093 | LLLYWLLTVLFLLLASLLILI | -17.52 | 0.28 ± 0.03 | NA |
| L_094 | LLLLYFLLATLVALLLSLILI | -31.23 | 0.47 ± 0.13 | -2.7 ± 0.73 |
| L_095 | LLLLILLLALLFVLTSLLILI | -30.54 | 0.35 ± 0.14 | -1.89 ± 0.76 |
| L_096 | LLLAALLLLLVTLLFSLLILI | -17.48 | 0.53 ± 0.08 | -2.89 ± 0.43 |
| L_097 | LLLLAVLLLILTLLLSFLILI | -30.86 | 0.38 ± 0.2 | -2.19 ± 1.14 |
| L_098 | LLLLTVLLAILFLLLSLLILI | -30.91 | 0.29 ± 0.09 | NA |
| L_099 | LLLYSLLLLLTFLLAVLLILI | -31.39 | 0.2 ± 0.18 | NA |
| L_100 | LLLLAVLLALLLILLSTLILI | -31.53 | 0.2 ± 0.02 | NA |

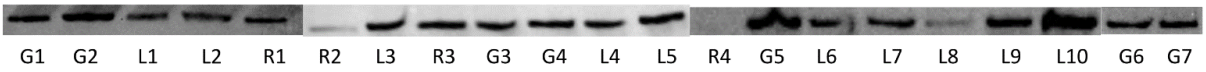
## Table ST2.3 Validated set of Right designs

| Design ID | Sequence | Computational Score (kcal/mol) | GpA (%) | ΔG (kcal/mol) |
|---|---|---|---|---|
| R_001 | LLL**YT**LL**VL**LL**AF**LL**SI**LILI | -26.75 | 0.3 ± 0.07 | NA |
| R_002 | LLL**YS**LL**LT**LL**FV**LL**AL**LILI | -30.22 | 0.3 ± 0.13 | NA |
| R_003 | LLL**YT**LL**IA**LL**FV**LL**SL**LILI | -26.69 | 0.28 ± 0.04 | NA |
| R_004 | LLL**YS**LL**VA**LL**FT**LL**FL**LILI | -26.6 | 0.23 ± 0.1 | NA |
| R_005 | LLL**LS**LL**FA**LL**IA**LL**VT**LILI | -24.71 | 0.29 ± 0.03 | NA |
| R_006 | LLL**YL**LL**FT**LL**AV**LL**SI**LILI | -8.79 | 0.3 ± 0.06 | NA |
| R_007 | LLL**YA**LL**VA**LL**FT**LL**SL**LILI | -24.88 | 0.26 ± 0.05 | NA |
| R_008 | LLL**TY**LL**FV**LL**AI**LL**SL**LILI | -24.89 | 0.29 ± 0.1 | NA |
| R_009 | LLL**YT**LL**IS**LL**VA**LL**FL**LILI | -25.05 | 0.25 ± 0.1 | NA |
| R_010 | LLL**YT**LL**VS**LL**AL**LL**AF**LILI | -25.05 | 0.31 ± 0.07 | 0.43 ± 0.1 |
| R_011 | LLL**LY**LL**VA**LL**IS**LL**FT**LILI | -25.89 | 0.39 ± 0.18 | -2.33 ± 1.07 |
| R_012 | LLL**YS**LL**LA**LL**FV**LL**TI**LILI | -27.63 | 0.3 ± 0.06 | NA |
| R_013 | LLL**LV**LL**IA**LL**FT**LL**SY**LILI | -25.24 | 0.33 ± 0.04 | -1.62 ± 0.21 |
| R_014 | LLL**SY**LL**VL**LL**AF**LL**TI**LILI | -25.39 | 0.35 ± 0.08 | -1.9 ± 0.43 |
| R_015 | LLL**YT**LL**IS**LL**FV**LL**AL**LILI | -26.81 | 0.31 ± 0.08 | 0.25 ± 0.07 |
| R_016 | LLL**YF**LL**VA**LL**TL**LL**SI**LILI | -26.06 | 0.27 ± 0.09 | NA |
| R_017 | LLL**SY**LL**FV**LL**AL**LL**ST**LILI | -26.54 | 0.39 ± 0.07 | -2.28 ± 0.4 |
| R_018 | LLL**SY**LL**TL**LL**AF**LL**SI**LILI | -26.9 | 0.55 ± 0.08 | -2.94 ± 0.45 |
| R_019 | LLL**WT**LL**VF**LL**AI**LL**SL**LILI | -29.79 | 0.29 ± 0.06 | NA |
| R_020 | LLL**SY**LL**FT**LL**AI**LL**TL**LILI | -29.93 | 0.3 ± 0.09 | NA |
| R_021 | LLL**YS**LL**IA**LL**FV**LL**TL**LILI | -27.75 | 0.28 ± 0.1 | NA |
| R_022 | LLL**SY**LL**FT**LL**AL**LL**AV**LILI | -27.41 | 0.38 ± 0.03 | -2.26 ± 0.15 |
| R_023 | LLL**YA**LL**LT**LL**VA**LL**SI**LILI | -10.03 | 0.33 ± 0.05 | -1.61 ± 0.26 |
| R_024 | LLL**YT**LL**IA**LL**VL**LL**FS**LILI | -24.33 | 0.31 ± 0.18 | -0.41 ± 0.24 |
| R_025 | LLL**FY**LL**VI**LL**AL**LL**ST**LILI | -16.37 | 0.29 ± 0.14 | NA |
| R_026 | LLL**FY**LL**VA**LL**IS**LL**TL**LILI | -16.35 | 0.34 ± 0.05 | -1.79 ± 0.24 |
| R_027 | LLL**YI**LL**AT**LL**AF**LL**SL**LILI | -16.29 | 0.27 ± 0.12 | NA |
| R_028 | LLL**FY**LL**VA**LL**LT**LL**SL**LILI | -16.26 | 0.4 ± 0.08 | -2.36 ± 0.48 |
| R_029 | LLL**YA**LL**IA**LL**FT**LL**VL**LILI | -16.15 | 0.3 ± 0.03 | NA |
| R_030 | LLL**YA**LL**TT**LL**SV**LL**FL**LILI | -16.09 | 0.44 ± 0.07 | -2.58 ± 0.43 |
| R_031 | LLL**TA**LL**IA**LL**FV**LL**SL**LILI | -16.08 | 0.33 ± 0.01 | -1.57 ± 0.04 |
| R_032 | LLL**YS**LL**LT**LL**VA**LL**IF**LILI | -15.9 | 0.24 ± 0.11 | NA |
| R_033 | LLL**YS**LL**LA**LL**AT**LL**VF**LILI | -15.86 | 0.24 ± 0.07 | NA |
| R_034 | LLL**YV**LL**TI**LL**AL**LL**SF**LILI | -15.76 | 0.34 ± 0.13 | -1.84 ± 0.72 |
| R_035 | LLL**YL**LL**AT**LL**FV**LL**SI**LILI | -15.7 | 0.31 ± 0.03 | -0.91 ± 0.1 |
| R_036 | LLL**YT**LL**IA**LL**AV**LL**SL**LILI | -15.58 | 0.37 ± 0.05 | -2.1 ± 0.31 |
| R_037 | LLL**YA**LL**VA**LL**IT**LL**FL**LILI | -15.3 | 0.3 ± 0.07 | NA |
| R_038 | LLL**FY**LL**IT**LL**VA**LL**SL**LILI | -15.25 | 0.34 ± 0.09 | -1.79 ± 0.49 |
| R_039 | LLL**YL**LL**AT**LL**AI**LL**FS**LILI | -15.07 | 0.14 ± 0.06 | NA |
| R_040 | LLL**YL**LL**AT**LL**AV**LL**SF**LILI | -16.45 | 0.25 ± 0.07 | NA |
| R_041 | LLL**YS**LL**LT**LL**AT**LL**AF**LILI | -16.53 | 0.21 ± 0.08 | NA |
| R_042 | LLL**VS**LL**LT**LL**FA**LL**AY**LILI | -16.54 | 0.29 ± 0.03 | NA |
| R_043 | LLL**TA**LL**VA**LL**AY**LL**SF**LILI | -16.55 | 0.5 ± 0.04 | -2.81 ± 0.24 |
| R_044 | LLL**TY**LL**TI**LL**AF**LL**SL**LILI | -17.5 | 0.41 ± 0.05 | -2.44 ± 0.27 |
| R_045 | LLL**LV**LL**IT**LL**FA**LL**SY**LILI | -17.49 | 0.32 ± 0.06 | -1.37 ± 0.27 |
| R_046 | LLL**YL**LL**TT**LL**VA**LL**SF**LILI | -17.43 | 0.44 ± 0.05 | -2.58 ± 0.31 |
| R_047 | LLL**YL**LL**VT**LL**AI**LL**SF**LILI | -17.35 | 0.17 ± 0.09 | NA |
| R_048 | LLL**YS**LL**VA**LL**AT**LL**FL**LILI | -17.33 | 0.26 ± 0.07 | NA |
| R_049 | LLL**AI**LL**VT**LL**FT**LL**SL**LILI | -17.29 | 0.36 ± 0.09 | -1.99 ± 0.48 |
| R_050 | LLL**YA**LL**TA**LL**SV**LL**FL**LILI | -17.19 | 0.37 ± 0.03 | -2.1 ± 0.18 |
| R_051 | LLL**YI**LL**TA**LL**VA**LL**SL**LILI | -15.06 | 0.31 ± 0.05 | -0.94 ± 0.15 |
| R_052 | LLL**FS**LL**LT**LL**VA**LL**SY**LILI | -17.12 | 0.26 ± 0.08 | NA |
| R_053 | LLL**VY**LL**TA**LL**IA**LL**SL**LILI | -16.69 | 0.37 ± 0.05 | -2.14 ± 0.31 |
| R_054 | LLL**YS**LL**LT**LL**VA**LL**SF**LILI | -16.68 | 0.29 ± 0.1 | NA |
| R_055 | LLL**YA**LL**VA**LL**TA**LL**SL**LILI | -16.67 | 0.35 ± 0.08 | -1.96 ± 0.42 |
| R_056 | LLL**TA**LL**LA**LL**AV**LL**SF**LILI | -16.65 | 0.4 ± 0.06 | -2.35 ± 0.38 |
| R_057 | LLL**AV**LL**LT**LL**FA**LL**SY**LILI | -16.62 | 0.26 ± 0.06 | NA |
| R_058 | LLL**YI**LL**LT**LL**AV**LL**SF**LILI | -16.59 | 0.24 ± 0.04 | NA |
| R_059 | LLL**YF**LL**VA**LL**LA**LL**ST**LILI | -16.58 | 0.32 ± 0.03 | -1.1 ± 0.11 |
| R_060 | LLL**YI**LL**FA**LL**TV**LL**SL**LILI | -17.06 | 0.26 ± 0.06 | NA |
| R_061 | LLL**YV**LL**IA**LL**AT**LL**SL**LILI | -15.04 | 0.33 ± 0.06 | -1.52 ± 0.29 |
| R_062 | LLL**YA**LL**VA**LL**SL**LL**FT**LILI | -15.04 | 0.35 ± 0.03 | -1.87 ± 0.14 |
| R_063 | LLL**YS**LL**TA**LL**VA**LL**FL**LILI | -14.78 | 0.29 ± 0.08 | NA |
| R_064 | LLL**TY**LL**LV**LL**AI**LL**SF**LILI | -12.04 | 0.19 ± 0.12 | NA |
| R_065 | LLL**YV**LL**IA**LL**FT**LL**SL**LILI | -11.97 | 0.26 ± 0.07 | NA |
| R_066 | LLL**YV**LL**TI**LL**AL**LL**AF**LILI | -11.94 | 0.32 ± 0.09 | -1.38 ± 0.4 |
| R_067 | LLL**YL**LL**ST**LL**AF**LL**AI**LILI | -11.93 | 0.28 ± 0.04 | NA |
| R_068 | LLL**TY**LL**IV**LL**AL**LL**SF**LILI | -11.69 | 0.41 ± 0.02 | -2.4 ± 0.1 |
| R_069 | LLL**YL**LL**FT**LL**AA**LL**TS**LILI | -11.61 | 0.12 ± 0.06 | NA |
| R_070 | LLL**YA**LL**FA**LL**LS**LL**VT**LILI | -11.25 | 0.27 ± 0.02 | NA |

| | | | | |
|---|---|---|---|---|
| R_071 | LLLTYLLAVLLAILLFLLILI | -12.11 | 0.32 ± 0.07 | -0.98 ± 0.21 |
| R_072 | LLLYVLLTTLLFALLSLLILI | -11.22 | 0.28 ± 0.08 | NA |
| R_073 | LLLAALLLSLLFTLLVFLILI | -10.66 | 0.46 ± 0.05 | -2.65 ± 0.27 |
| R_074 | LLLYLLLATLLAVLLSILILI | -10.64 | 0.24 ± 0.07 | NA |
| R_075 | LLLYALLLSLLTALLVFLILI | -10.61 | 0.28 ± 0.09 | NA |
| R_076 | LLLTALLIALLFLLLSVLILI | -10.55 | 0.4 ± 0.04 | -2.36 ± 0.25 |
| R_077 | LLLYVLLFALLTVLLSLLILI | -10.4 | 0.3 ± 0.02 | NA |
| R_078 | LLLYSLLFVLLAILLTLLILI | -10.33 | 0.25 ± 0.1 | NA |
| R_079 | LLLAYLLVLLLATLLSFLILI | -10.19 | 0.34 ± 0.07 | -1.71 ± 0.38 |
| R_080 | LLLYVLLAILLTLLLSFLILI | -10.68 | 0.27 ± 0.1 | NA |
| R_081 | LLLYTLLATLLVLLLSFLILI | -17.69 | 0.25 ± 0.07 | NA |
| R_082 | LLLYVLLATLLALLLSFLILI | -12.44 | 0.26 ± 0.03 | NA |
| R_083 | LLLYALLLTLLVALLSFLILI | -12.82 | 0.25 ± 0.08 | NA |
| R_084 | LLLYALLLTLLSILLVFLILI | -14.69 | 0.22 ± 0.04 | NA |
| R_085 | LLLYALLLALLFSLLVTLILI | -14.68 | 0.3 ± 0.1 | NA |
| R_086 | LLLLYLLTALLIALLFVLILI | -14.55 | 0.27 ± 0.06 | NA |
| R_087 | LLLLILLSTLLATLLVFLILI | -14.36 | 0.23 ± 0.05 | NA |
| R_088 | LLLYILLLTLLVALLSFLILI | -14.33 | 0.28 ± 0.01 | NA |
| R_089 | LLLLALLFALLVSLLTWLILI | -14.2 | 0.3 ± 0.0 | NA |
| R_090 | LLLTVLLLALLFALLSILILI | -14.11 | 0.36 ± 0.06 | -2.01 ± 0.32 |
| R_091 | LLLVYLLTALLLALLSFLILI | -12.66 | 0.33 ± 0.07 | -1.51 ± 0.34 |
| R_092 | LLLYVLLFTLLAALLSLLILI | -13.95 | 0.28 ± 0.06 | NA |
| R_093 | LLLVYLLLALLTSLLFILILI | -13.89 | 0.39 ± 0.03 | -2.3 ± 0.16 |
| R_094 | LLLYSLLLILLAVLLTFLILI | -13.86 | 0.24 ± 0.07 | NA |
| R_095 | LLLTYLLVFLLALLLSILILI | -13.66 | 0.34 ± 0.09 | -1.68 ± 0.45 |
| R_096 | LLLYALLLTLLSVLLFILILI | -13.65 | 0.3 ± 0.07 | NA |
| R_097 | LLLALLLVTLLAILLSFLILI | -13.32 | 0.17 ± 0.07 | NA |
| R_098 | LLLTVLLALLLAFLLSILILI | -12.94 | 0.35 ± 0.06 | -1.9 ± 0.33 |
| R_099 | LLLTYLLVALLIALLFLLILI | -12.88 | 0.37 ± 0.05 | -2.12 ± 0.3 |
| R_100 | LLLYVLLLTLLFALLSLLILI | -13.89 | 0.3 ± 0.11 | NA |
| R_101 | LLLYWLLVALLISLLTLLILI | -24.44 | 0.32 ± 0.03 | -1.09 ± 0.09 |
| R_102 | LLLYLLLVALLATLLSFLILI | -17.7 | 0.22 ± 0.08 | NA |
| R_103 | LLLFYLLLTLLVALLSILILI | -17.85 | 0.48 ± 0.06 | -2.74 ± 0.36 |
| R_104 | LLLTYLLFTLLAVLLSLLILI | -22.18 | 0.36 ± 0.04 | -2.01 ± 0.25 |
| R_105 | LLLYWLLTALLLSLLFVLILI | -22.11 | 0.21 ± 0.05 | NA |
| R_106 | LLLYVLLLTLLATLLSFLILI | -21.91 | 0.22 ± 0.12 | NA |
| R_107 | LLLTYLLVALLAILLSLLILI | -21.41 | 0.28 ± 0.02 | NA |
| R_108 | LLLYSLLLTLLAVLLFSLILI | -21.39 | 0.21 ± 0.1 | NA |
| R_109 | LLLLYLLITLLFSLLVALILI | -21.29 | 0.3 ± 0.05 | NA |
| R_110 | LLLYSLLLALLFALLVTLILI | -21.29 | 0.27 ± 0.05 | NA |
| R_111 | LLLYSLLVALLFTLLILLILI | -21.29 | 0.28 ± 0.12 | NA |
| R_112 | LLLFSLLLTLLAALLSYLILI | -21.16 | 0.21 ± 0.07 | NA |
| R_113 | LLLYILLSTLLALLLAFLILI | -21.11 | 0.21 ± 0.04 | NA |
| R_114 | LLLYSLLITLLFVLLALLILI | -20.93 | 0.25 ± 0.09 | NA |
| R_115 | LLLYTLLATLLFLLLSVLILI | -20.85 | 0.27 ± 0.11 | NA |
| R_116 | LLLYVLLFILLALLLSTLILI | -20.83 | 0.22 ± 0.1 | NA |
| R_117 | LLLYILLSVLLALLLTFLILI | -20.82 | 0.19 ± 0.08 | NA |
| R_118 | LLLYTLLVALLFALLSLLILI | -20.72 | 0.3 ± 0.07 | NA |
| R_119 | LLLVYLLIALLFSLLTLLILI | -22.22 | 0.34 ± 0.04 | -1.71 ± 0.23 |
| R_120 | LLLYALLTILLTFLLSLLILI | -22.25 | 0.28 ± 0.15 | NA |
| R_121 | LLLYALLTSLLVALLFLLILI | -22.5 | 0.34 ± 0.02 | -1.73 ± 0.08 |
| R_122 | LLLYSLLLALLFALLTILILI | -22.52 | 0.45 ± 0.06 | -2.6 ± 0.36 |
| R_123 | LLLYTLLVSLLFALLILLILI | -24.31 | 0.23 ± 0.06 | NA |
| R_124 | LLLVYLLTALLISLLFLLILI | -24.25 | 0.37 ± 0.07 | -2.17 ± 0.4 |
| R_125 | LLLWYLLLVLLATLLSFLILI | -24.21 | 0.35 ± 0.07 | -1.95 ± 0.39 |
| R_126 | LLLYSLLLTLLATLLVFLILI | -24.19 | 0.21 ± 0.08 | NA |
| R_127 | LLLYILLATLLTLLLSFLILI | -24.1 | 0.21 ± 0.11 | NA |
| R_128 | LLLFVLLATLLAYLLSLLILI | -23.75 | 0.36 ± 0.03 | -2.07 ± 0.14 |
| R_129 | LLLYSLLVALLFALLTLLILI | -23.75 | 0.37 ± 0.07 | -2.16 ± 0.4 |
| R_130 | LLLYLLLAVLLAFLLTSLILI | -20.67 | 0.24 ± 0.09 | NA |
| R_131 | LLLSYLLVFLLATLLSLLILI | -23.6 | 0.39 ± 0.06 | -2.32 ± 0.34 |
| R_132 | LLLYSLLLTLLALLLAFLILI | -23.54 | 0.3 ± 0.08 | NA |
| R_133 | LLLYSLLLTLLFALLVILILI | -23.43 | 0.36 ± 0.02 | -2.01 ± 0.12 |
| R_134 | LLLYVLLATLLAFLLSLLILI | -23.31 | 0.29 ± 0.06 | NA |
| R_135 | LLLLYLLVTLLFALLSILILI | -23.2 | 0.37 ± 0.07 | -2.15 ± 0.4 |
| R_136 | LLLATLLISLLFTLLVLLILI | -23.06 | 0.43 ± 0.09 | -2.51 ± 0.56 |
| R_137 | LLLYLLLATLLFVLLSLLILI | -22.94 | 0.21 ± 0.12 | NA |
| R_138 | LLLYALLLALLVSLLFTLILI | -22.59 | 0.38 ± 0.06 | -2.2 ± 0.38 |
| R_139 | LLLVYLLLALLTALLSILILI | -23.55 | 0.36 ± 0.11 | -2.07 ± 0.63 |
| R_140 | LLLFYLLVALLTALLSLLILI | -20.62 | 0.39 ± 0.03 | -2.31 ± 0.17 |

| | | | | |
|---|---|---|---|---|
| R_141 | LLL**VY**LL**IT**LL**FA**LL**SL**LILI | -20.57 | 0.43 ± 0.04 | -2.51 ± 0.25 |
| R_142 | LLL**YI**LL**TV**LL**AL**LL**SF**LILI | -20.38 | 0.26 ± 0.05 | NA |
| R_143 | LLL**YT**LL**VA**LL**ST**LL**FL**LILI | -19.39 | 0.23 ± 0.09 | NA |
| R_144 | LLL**LT**LL**AT**LL**AV**LL**SF**LILI | -19.38 | 0.37 ± 0.03 | -2.09 ± 0.16 |
| R_145 | LLL**VY**LL**LA**LL**FT**LL**SL**LILI | -19.07 | 0.47 ± 0.06 | -2.68 ± 0.36 |
| R_146 | LLL**YV**LL**TA**LL**FA**LL**SL**LILI | -19.06 | 0.32 ± 0.06 | -0.98 ± 0.18 |
| R_147 | LLL**LT**LL**IA**LL**VA**LL**SF**LILI | -18.94 | 0.34 ± 0.07 | -1.8 ± 0.37 |
| R_148 | LLL**FV**LL**TT**LL**AI**LL**SL**LILI | -18.76 | 0.22 ± 0.02 | NA |
| R_149 | LLL**YF**LL**AI**LL**AV**LL**TL**LILI | -18.7 | 0.24 ± 0.06 | NA |
| R_150 | LLL**LT**LL**IA**LL**FV**LL**YS**LILI | -19.44 | 0.43 ± 0.06 | -2.52 ± 0.34 |
| R_151 | LLL**VA**LL**LA**LL**FA**LL**ST**LILI | -18.56 | 0.36 ± 0.04 | -2.01 ± 0.23 |
| R_152 | LLL**VA**LL**LT**LL**AI**LL**SF**LILI | -18.48 | 0.28 ± 0.05 | NA |
| R_153 | LLL**YI**LL**VA**LL**AT**LL**TL**LILI | -18.48 | 0.34 ± 0.11 | -1.84 ± 0.61 |
| R_154 | LLL**YL**LL**TV**LL**AF**LL**SI**LILI | -18.43 | 0.27 ± 0.05 | NA |
| R_155 | LLL**TY**LL**VA**LL**IA**LL**SL**LILI | -18.14 | 0.36 ± 0.02 | -2.0 ± 0.09 |
| R_156 | LLL**LA**LL**VA**LL**ST**LL**IF**LILI | -18.12 | 0.39 ± 0.06 | -2.31 ± 0.37 |
| R_157 | LLL**TY**LL**IA**LL**LA**LL**FV**LILI | -18.08 | 0.4 ± 0.01 | -2.35 ± 0.05 |
| R_158 | LLL**YS**LL**VA**LL**TL**LL**FI**LILI | -17.88 | 0.31 ± 0.13 | -0.84 ± 0.34 |
| R_159 | LLL**LA**LL**FA**LL**TV**LL**SI**LILI | -18.52 | 0.44 ± 0.02 | -2.59 ± 0.13 |
| R_160 | LLL**YS**LL**LA**LL**AV**LL**TI**LILI | -17.72 | 0.31 ± 0.22 | 0.32 ± 0.23 |
| R_161 | LLL**AI**LL**LT**LL**FV**LL**SL**LILI | -19.44 | 0.33 ± 0.06 | -1.49 ± 0.26 |
| R_162 | LLL**TY**LL**VA**LL**TF**LL**SL**LILI | -19.5 | 0.59 ± 0.25 | -3.04 ± 1.29 |
| R_163 | LLL**AV**LL**LT**LL**FT**LL**SY**LILI | -20.37 | 0.21 ± 0.11 | NA |
| R_164 | LLL**YS**LL**VA**LL**FL**LL**AT**LILI | -20.33 | 0.28 ± 0.1 | NA |
| R_165 | LLL**YS**LL**LT**LL**AA**LL**VF**LILI | -20.13 | 0.23 ± 0.08 | NA |
| R_166 | LLL**LY**LL**IA**LL**VA**LL**ST**LILI | -20.11 | 0.3 ± 0.05 | NA |
| R_167 | LLL**YT**LL**VA**LL**FL**LL**SL**LILI | -20.05 | 0.27 ± 0.1 | NA |
| R_168 | LLL**FY**LL**TA**LL**IA**LL**SL**LILI | -20 | 0.38 ± 0.03 | -2.23 ± 0.15 |
| R_169 | LLL**YW**LL**VA**LL**LA**LL**ST**LILI | -19.98 | 0.36 ± 0.01 | -2.06 ± 0.04 |
| R_170 | LLL**YI**LL**FT**LL**VA**LL**SL**LILI | -19.46 | 0.19 ± 0.06 | NA |
| R_171 | LLL**FA**LL**TI**LL**AY**LL**SL**LILI | -19.95 | 0.4 ± 0.06 | -2.34 ± 0.34 |
| R_172 | LLL**YA**LL**VA**LL**LT**LL**SW**LILI | -19.84 | 0.27 ± 0.04 | NA |
| R_173 | LLL**YS**LL**FT**LL**AV**LL**AL**LILI | -19.77 | 0.32 ± 0.1 | -1.4 ± 0.42 |
| R_174 | LLL**YW**LL**LA**LL**VA**LL**ST**LILI | -19.77 | 0.3 ± 0.04 | NA |
| R_175 | LLL**LA**LL**IA**LL**FV**LL**ST**LILI | -19.62 | 0.27 ± 0.05 | NA |
| R_176 | LLL**YT**LL**VS**LL**AI**LL**TL**LILI | -19.61 | 0.33 ± 0.09 | -1.58 ± 0.45 |
| R_177 | LLL**FY**LL**TA**LL**IT**LL**VL**LILI | -19.54 | 0.31 ± 0.09 | NA |
| R_178 | LLL**SY**LL**VF**LL**AL**LL**AT**LILI | -19.53 | 0.48 ± 0.07 | -2.74 ± 0.39 |
| R_179 | LLL**TY**LL**LA**LL**VS**LL**FI**LILI | -19.87 | 0.47 ± 0.09 | -2.68 ± 0.54 |
| R_180 | LLL**YV**LL**AT**LL**TF**LL**SL**LILI | -9.81 | 0.29 ± 0.02 | NA |

Data for validated sets of designs where clashing mutations < 35% GpA
Interfacial designed positions of each sequence in bold
Designs where a positive ΔG was calculated are designated NA




| Western ID | Sequence | Design ID |
|---|---|---|
| G1 | LLL**AV**LL**LT**LL**GG**LL**FA**LILI | |
| G2 | LLL**YV**LL**GA**LL**GI**LL**TL**LILI | G_016 |
| L1 | LLLLL**FL**LL**AT**LL**IL**LL**SV**LILI | L_021 |
| L2 | LLL**FS**LL**LL**LV**VA**LL**LT**LLILI | |
| R1 | LLL**YI**LL**TA**LL**VA**LL**SL**LILI | R_051 |
| R2 | LLL**FY**LL**VA**LL**TA**LL**SL**LILI | R_140 |
| L3 | LLLL**AV**LL**FL**LL**AL**LL**TS**LLILI | L_031 |
| R3 | LLL**TV**LL**LA**LL**FA**LL**SI**LILI | R_090 |
| G3 | LLL**TA**LL**LA**LL**FG**LL**FS**LILI | G_060 |
| G4 | LLL**TA**LL**IG**LL**FG**LL**VL**LILI | |
| L4 | LLLL**YL**LL**AV**LL**TA**LL**FS**LILI | L_022 |
| L5 | LLLL**LA**LL**LI**LL**FV**LL**ST**LILI | |
| R4 | LLL**TS**LL**LA**LL**FV**LL**IL**LILI | |
| G5 | LLL**VA**LL**AL**LL**GT**LL**SF**LILI | G_052 |
| L6 | LLLL**FA**LL**LA**LL**YV**LL**TS**LLILI | |
| L7 | LLLL**AI**LL**AV**LL**FT**LL**SL**LILI | |
| L8 | LLL**YA**LL**FV**LL**TA**LL**LS**LLILI | L_006 |
| L9 | LLLL**AV**LL**LF**LL**AL**LL**TS**LLILI | L_031 |
| L10 | LLL**FA**LL**LV**LL**IT**LL**YS**LLILI | |
| G6 | LLL**TV**LL**AL**LL**GF**LL**GS**LILI | G_022 |
| G7 | LLL**VA**LL**AL**LL**GT**LL**SF**LILI | G_052 |

**Figure S2.1 Immunoblotting of designed constructs.** A subset of designed constructs were verified for their expression levels with western blots. While $GAS_{right}$ and Left show similar levels of expression, Right designs often have fainter bands on western blots. Sequences without a Design ID are present in our sort-seq data but not part of our validated set (clashing mutants < 35% GpA)

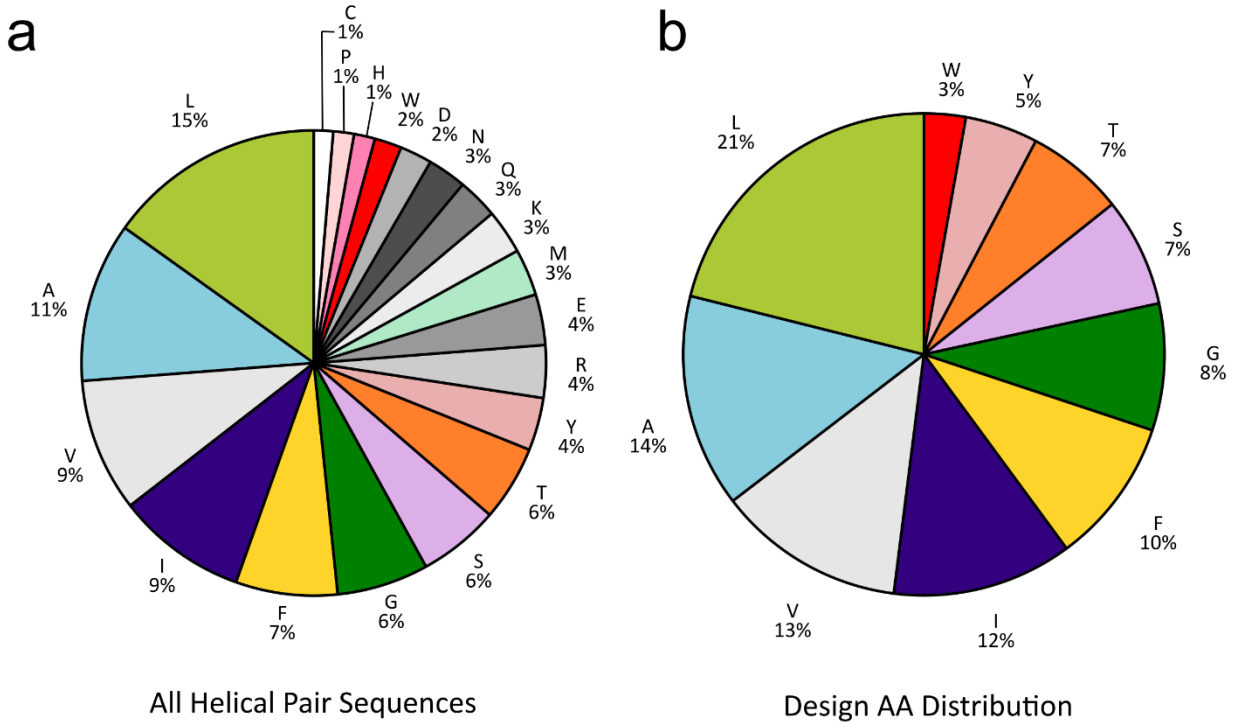



**Figure S2.2 Amino acid composition. a)** Frequency of amino acids from all transmembrane helical sequences extracted from OPM. **b)** The frequency of the design amino acids, adjusted to add up to 100% after removing the non-design amino acids. Sequences were designed with interfaces aiming to match the frequency of the design amino acid distribution (SEQUENCE_ENTROPY).

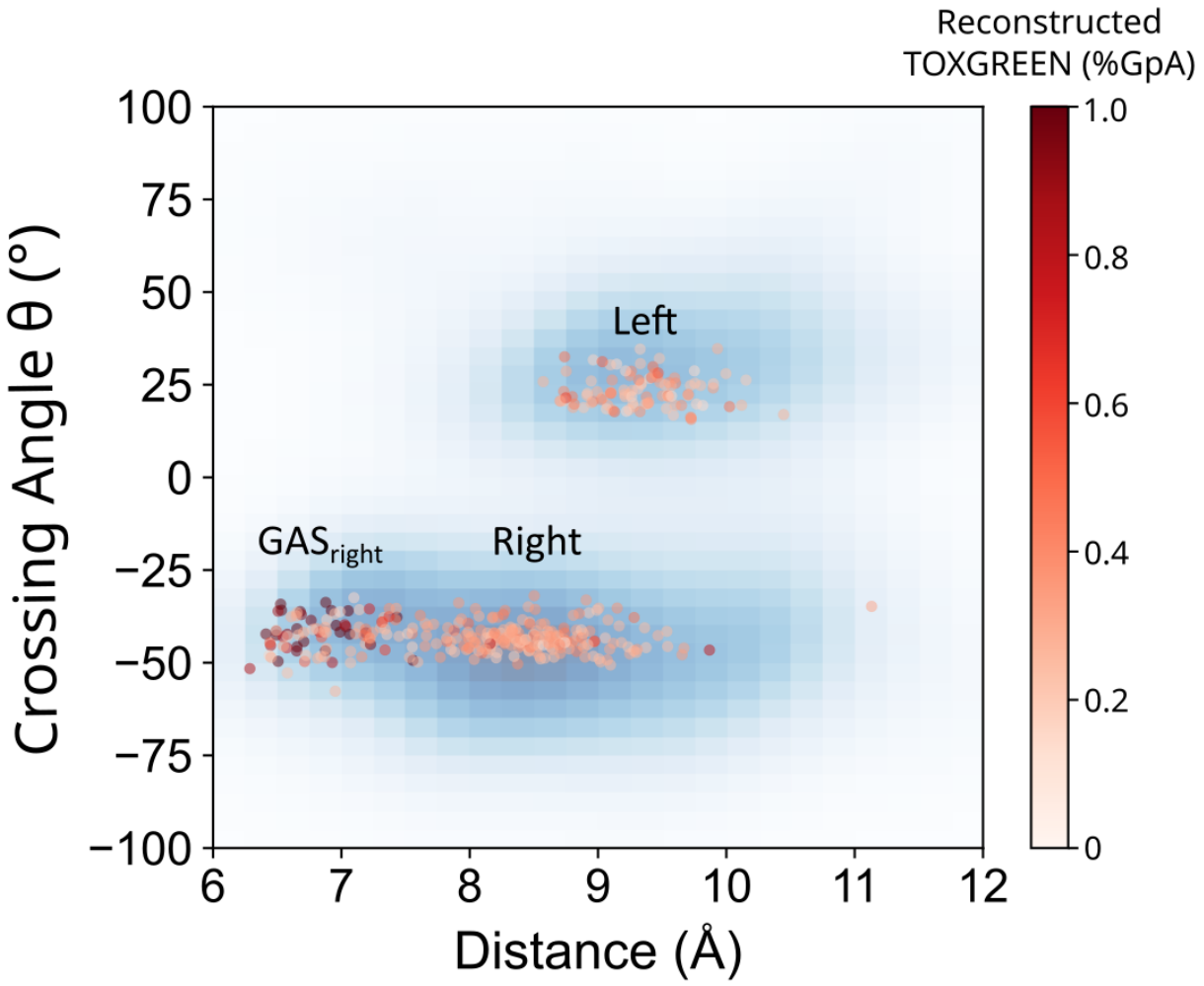



**Figure S2.3 Output Geometry vs Dimerization Propensity.** Distance and crossing angle of validated designed sequences plotted against %GpA in red. Sequences with highest dimerization propensity are dark red, most often found in GAS$_{right}$.

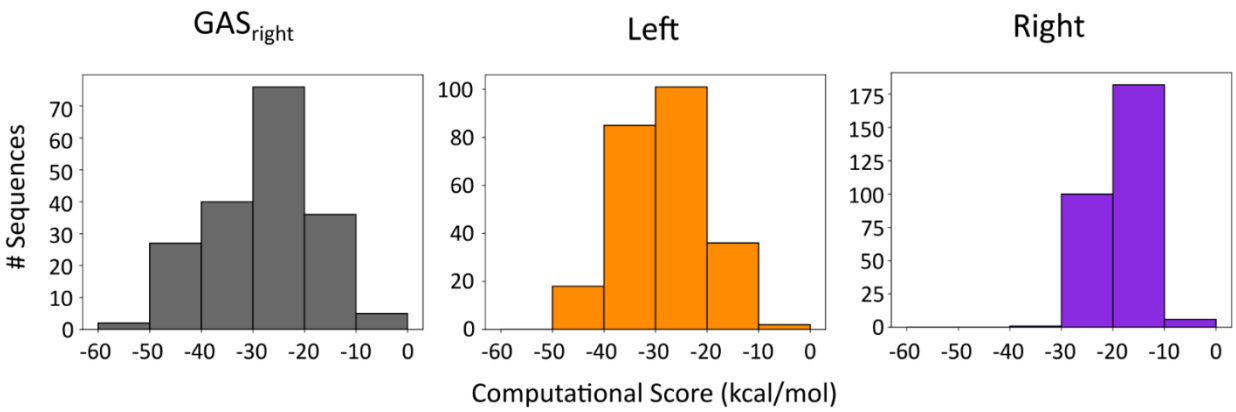



**Figure S2.4 Frequency of sequences by computational score.** Computational energy score of sequences present in sort-seq and the frequency for each design region.

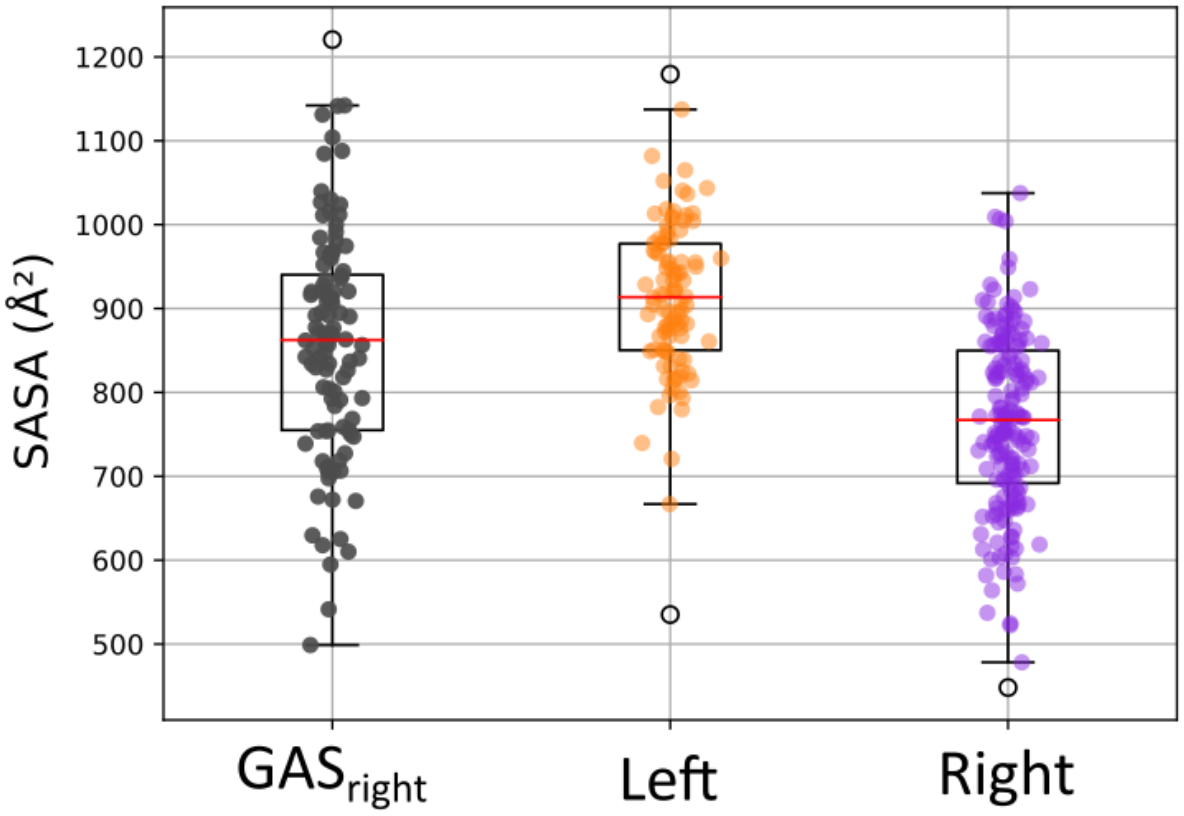



**Figure S2.5 Interface SASA.** Boxplots of calculated SASA at the interface of design structures of validated sequences. $GAS_{right}$ and Left design interfaces are larger than Right, which may contribute to Right designs having a lower computational energy score.

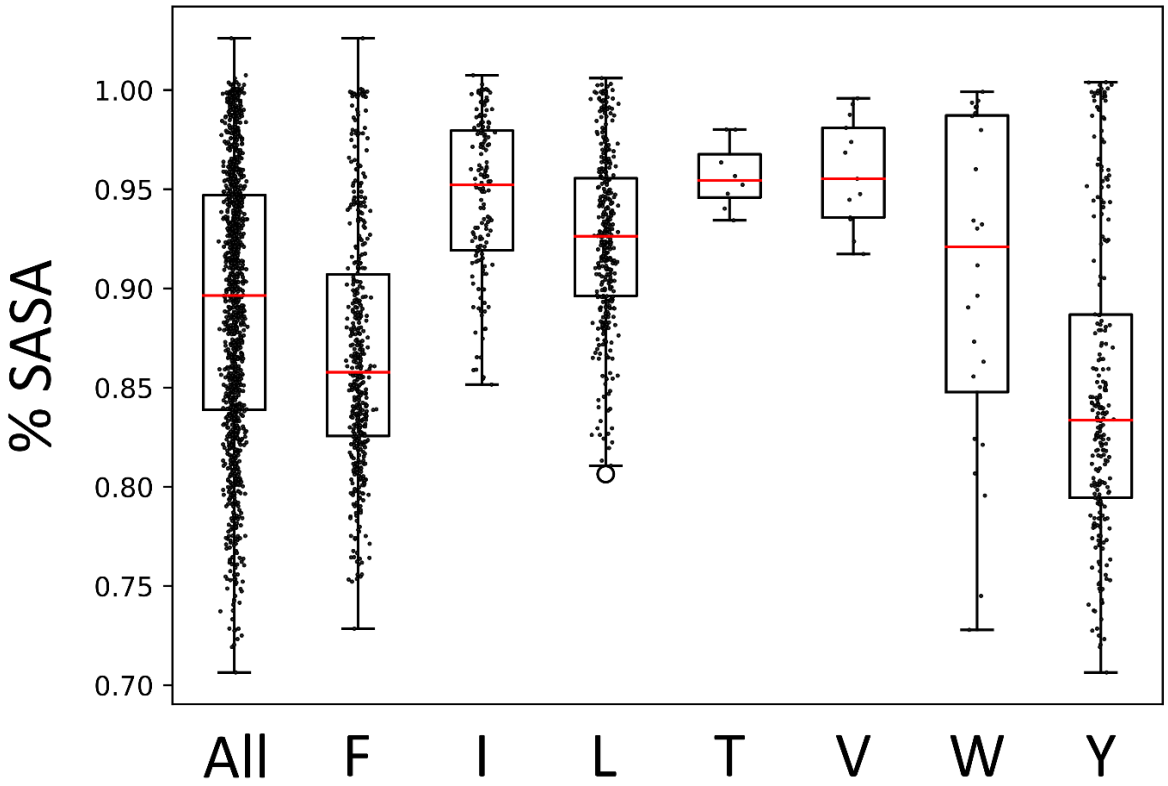



**Figure S2.6 % SASA of Large→Ala.** Single alanine point mutations were made on all interfacial positions designed structures. Mutants where the interface decreased the most, or the smallest % SASA (Design Interface SASA/Mutant Interface SASA **x** 100%), were selected.

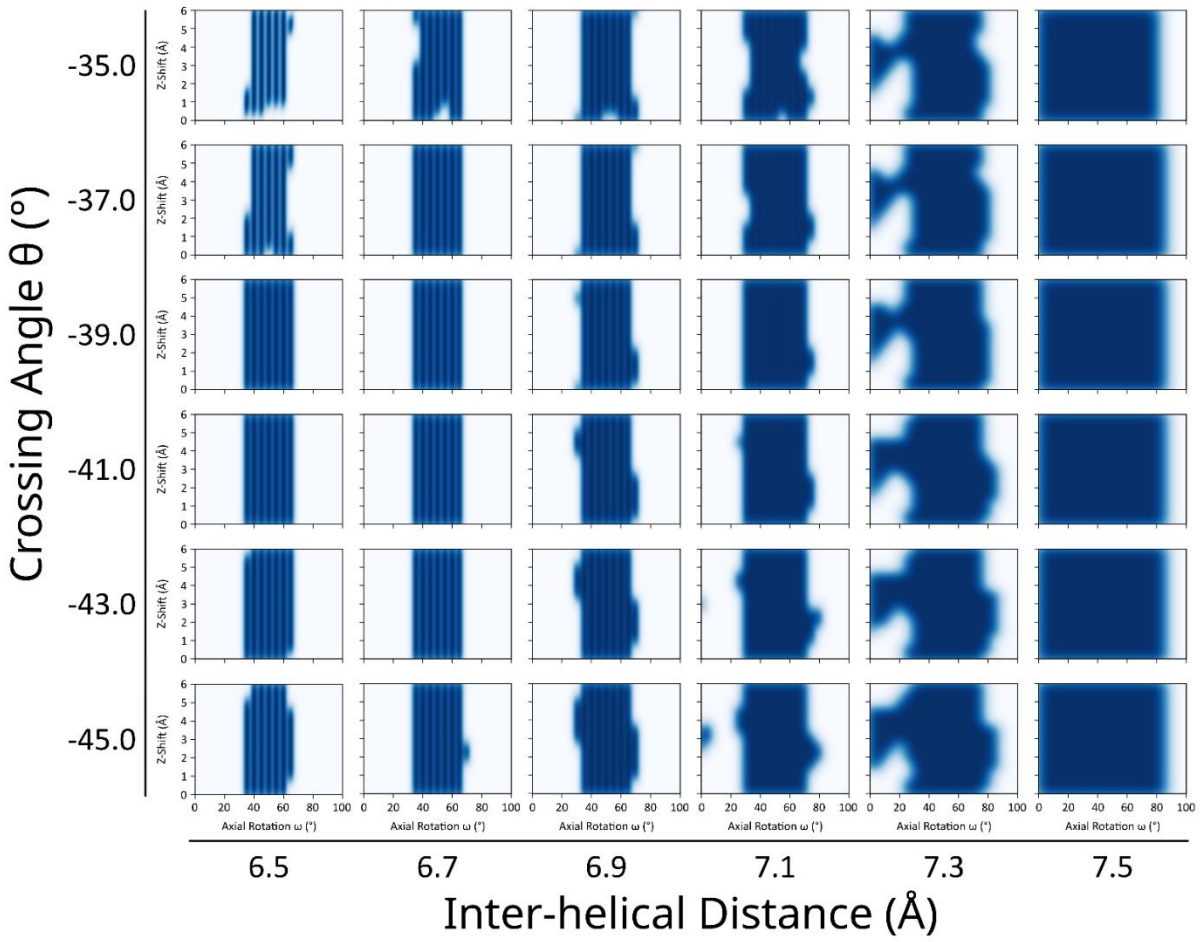



**Figure S2.7 $GAS_{right}$ axial rotations and z-shifts.** A 4-dimensional grid of geometries were randomly generated to determine axial rotations and z-shifts for design. Poly-leucine helices with Ala at the design interface positions were placed at 3000 geometries and computational energy score measured. Geometries where energy score < 10 kcal/mol were plotted on a KDE plot of Z-Shift vs Axial Rotation. Values with KDE > 0.8 were randomly selected for design.

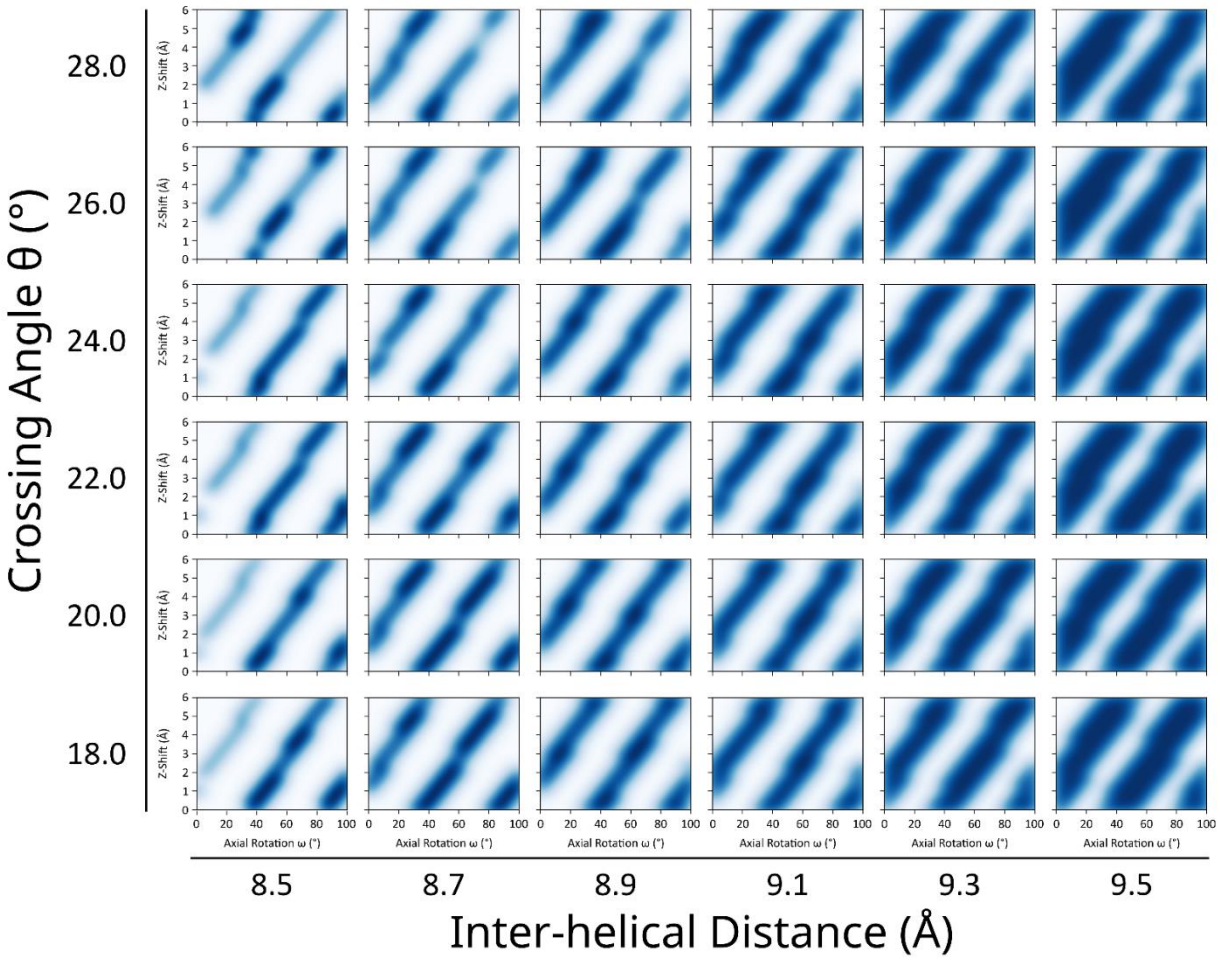



**Figure S2.8 Left axial rotations and z-shifts.** A 4-dimensional grid of geometries were randomly generated to determine axial rotations and z-shifts for design. Poly-leucine helices with Ala at the design interface positions were placed at 3000 geometries and computational energy score measured. Geometries where energy score < 10 kcal/mol were plotted on a KDE plot of Z-Shift vs Axial Rotation. Values with KDE > 0.8 were randomly selected for design.

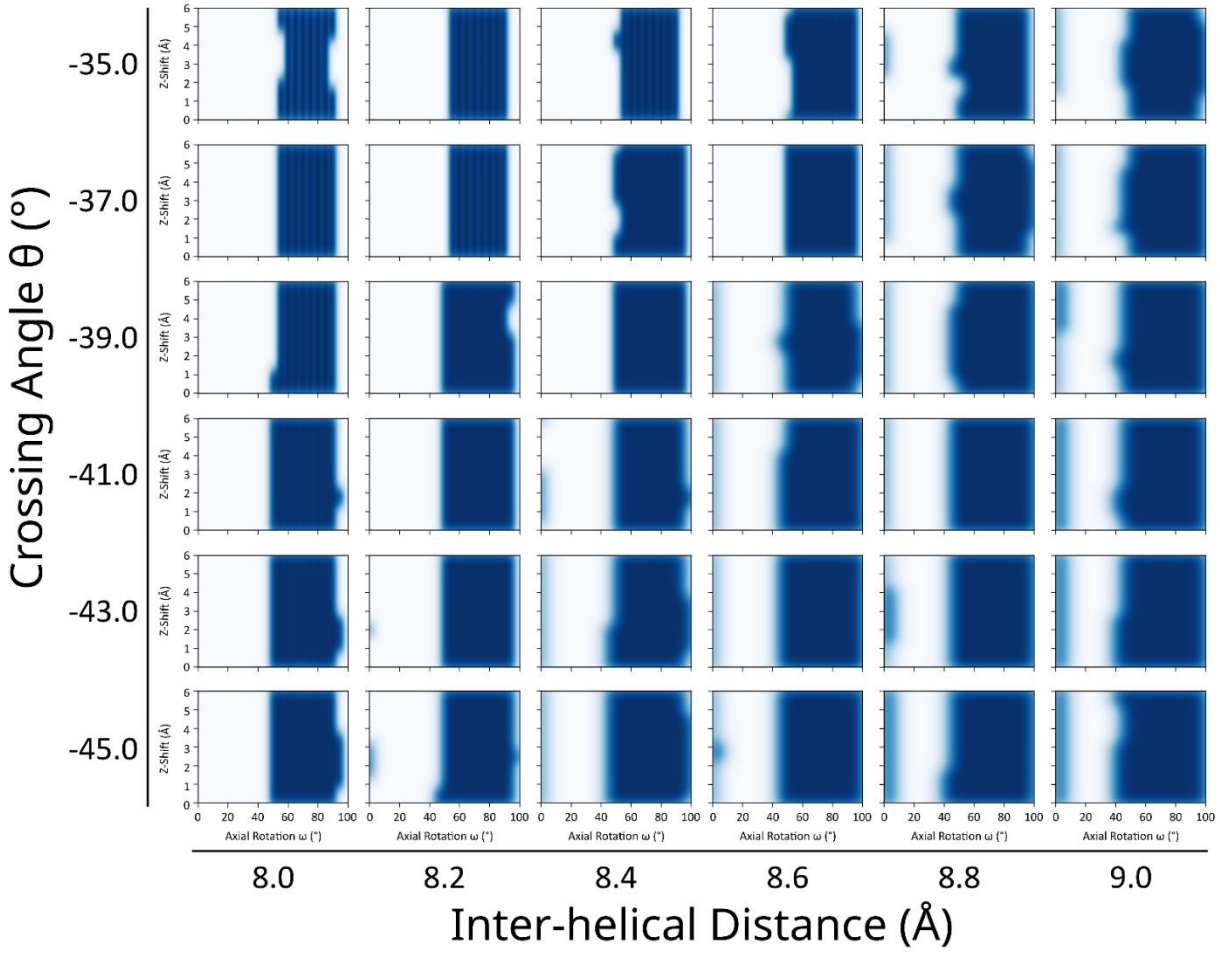



**Figure S2.9 Right axial rotations and z-shifts.** A 4-dimensional grid of geometries were randomly generated to determine axial rotations and z-shifts for design. Poly-leucine helices with Ala at the design interface positions were placed at 3000 geometries and computational energy score measured. Geometries where energy score < 10 kcal/mol were plotted on a KDE plot of Z-Shift vs Axial Rotation. Values with KDE > 0.8 were randomly selected for design.

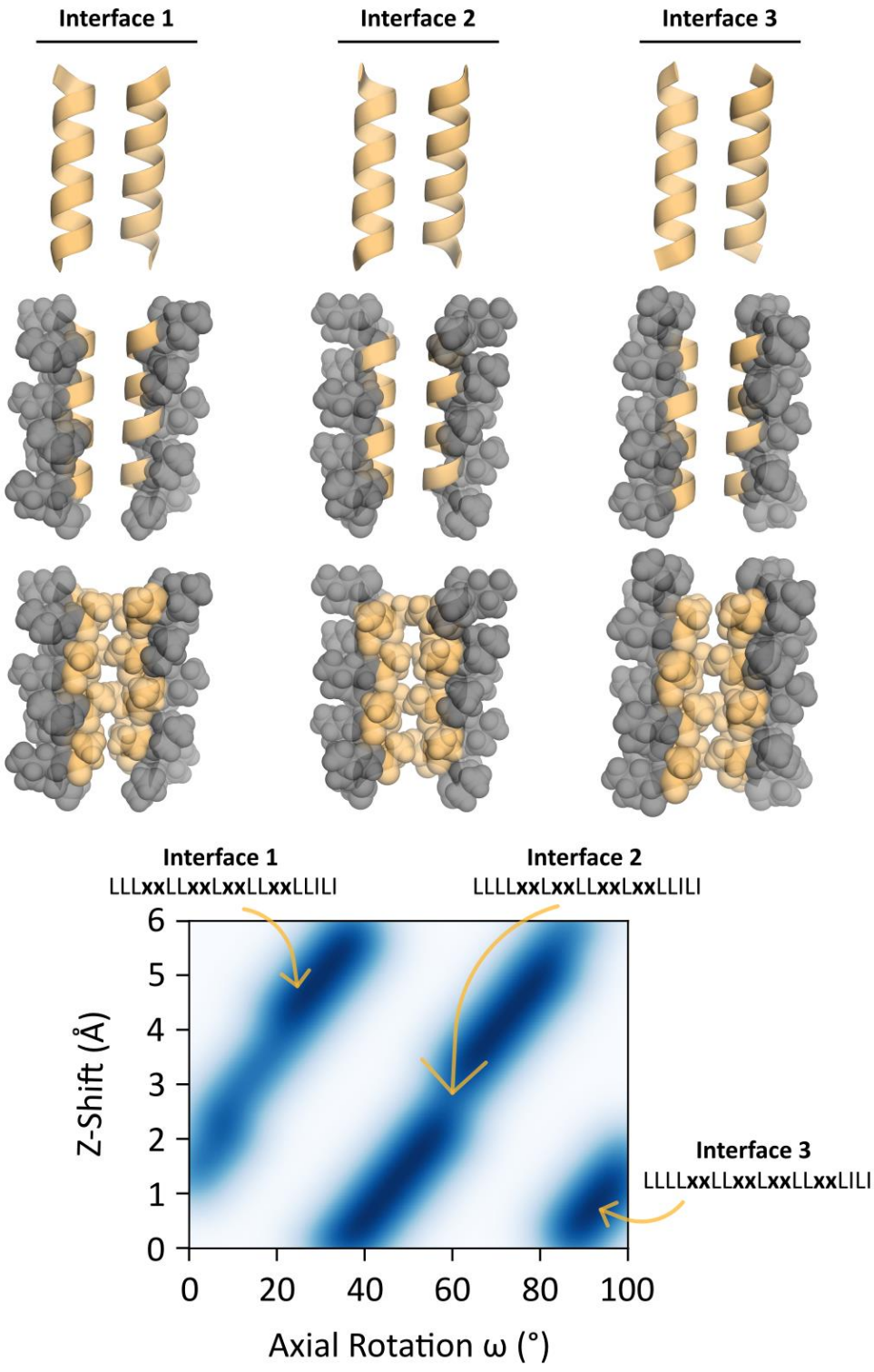



**Figure S2.10 Left-handed design interfaces.** Left-handed areas of high density were visually inspected to determine interfacial positions. Each design interface corresponds to a striped region of the axial rotation and z-shift density. Structures with Ala at corresponding interfacial positions are as visual representations.

## 2.7 References

Anderson, S. M. (2019). *Understanding the GASright Motif: Sequence, Structure, and Stability* [Ph.D., The University of Wisconsin - Madison]. https://www.proquest.com/docview/2331244818/abstract/2C4D47E16DE047ADPQ/1

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

Armstrong, C. R., & Senes, A. (2016). Screening for transmembrane association in divisome proteins using TOXGREEN, a high-throughput variant of the TOXCAT assay. *Biochim Biophys Acta*, *1858*(11), 2573–2583. https://doi.org/10.1016/j.bbamem.2016.07.008

Ash, W. L., Stockner, T., MacCallum, J. L., & Tieleman, D. P. (2004). Computer modeling of polyleucine-based coiled coil dimers in a realistic membrane environment: Insight into helix-helix interactions in membrane proteins. *Biochemistry*, *43*(28), 9050–9060. https://doi.org/10.1021/bi0494572

Ben-Tal, N., Sitkoff, D., Topol, I. A., Yang, A.-S., Burt, S. K., & Honig, B. (1997). Free energy of amide hydrogen bond formation in vacuum, in water, and in liquid alkane solution. *The Journal of Physical Chemistry B*, *101*(3), 450–457.

Bornberg-Bauer, E., Rivals, E., & Vingron, M. (1998). Computational approaches to identify leucine zippers. *Nucleic Acids Research*, *26*(11), 2740–2746. https://doi.org/10.1093/nar/26.11.2740

Bower, M. J., Cohen, F. E., & Dunbrack, R. L. (1997). Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool1. *Journal of Molecular Biology*, *267*(5), 1268–1282. https://doi.org/10.1006/jmbi.1997.0926

Bowie, J. U. (2011). Membrane protein folding: How important are hydrogen bonds? *Curr Opin Struct Biol*, *21*(1), 42–49. https://doi.org/10.1016/j.sbi.2010.10.003

Díaz Vázquez, G., Cui, Q., & Senes, A. (2023). Thermodynamic analysis of the GASright transmembrane motif supports energetic model of dimerization. *Biophysical Journal*, *122*(1), 143–155. https://doi.org/10.1016/j.bpj.2022.11.018

Doura, A. K., & Fleming, K. G. (2004). Complex Interactions at the Helix–Helix Interface Stabilize the Glycophorin A Transmembrane Dimer. *Journal of Molecular Biology*, *343*(5), 1487–1497. https://doi.org/10.1016/j.jmb.2004.09.011

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

Fleming, K. G., & Engelman, D. M. (2001). Computation and mutagenesis suggest a right-handed structure for the synaptobrevin transmembrane dimer. *Proteins: Structure, Function, and Bioinformatics*, *45*(4), 313–317. https://doi.org/10.1002/prot.1151

Gregersen, N., Bross, P., Vang, S., & Christensen, J. H. (2006). Protein misfolding and human disease. *Annu Rev Genomics Hum Genet*, *7*, 103–124. https://doi.org/10.1146/annurev.genom.7.080505.115737

Howitt, S. M., Rodgers, A. J. W., Jeffrey, P. D., & Cox, G. B. (1996). A Mutation in Which Alanine 128 Is Replaced by Aspartic Acid Abolishes Dimerization of the b-Subunit of the F0F1-ATPase from Escherichia coli(∗). *Journal of Biological Chemistry*, *271*(12), 7038–7042. https://doi.org/10.1074/jbc.271.12.7038

Huang, B., Xu, Y., Hu, X., Liu, Y., Liao, S., Zhang, J., Huang, C., Hong, J., Chen, Q., & Liu, H. (2022). A backbone-centred energy function of neural networks for protein design. *Nature*, *602*(7897), Article 7897. https://doi.org/10.1038/s41586-021-04383-5

Joh, N. H., Oberai, A., Yang, D., Whitelegge, J. P., & Bowie, J. U. (2009). Similar energetic contributions of packing in the core of membrane and water-soluble proteins. *J Am Chem Soc*, *131*(31), 10846–10847. https://doi.org/10.1021/ja904711k

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

Khadria, A. S., Mueller, B. K., Stefely, J. A., Tan, C. H., Pagliarini, D. J., & Senes, A. (2014). A Gly-Zipper Motif Mediates Homodimerization of the Transmembrane Domain of the Mitochondrial Kinase ADCK3. *Journal of the American Chemical Society*, *136*(40), 14068–14077. https://doi.org/10.1021/ja505017f

Krivov, G. G., Shapovalov, M. V., & Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, *77*(4), 778–795. https://doi.org/10.1002/prot.22488

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., & Baker, D. (2003). Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science*, *302*(5649), 1364–1368. https://doi.org/10.1126/science.1089427

Kulp, D. W., Subramaniam, S., Donald, J. E., Hannigan, B. T., Mueller, B. K., Grigoryan, G., & Senes, A. (2012). Structural informatics, modeling, and design with an open-source Molecular Software Library (MSL). *J Comput Chem*, *33*(20), 1645–1661. https://doi.org/10.1002/jcc.22968

LaPointe, L. M., Taylor, K. C., Subramaniam, S., Khadria, A., Rayment, I., & Senes, A. (2013). Structural Organization of FtsB, a Transmembrane Protein of the Bacterial Divisome. *Biochemistry*, *52*(15), 2574–2585. https://doi.org/10.1021/bi400222r

Lawrie, C. M., Sulistijo, E. S., & MacKenzie, K. R. (2010). Intermonomer Hydrogen Bonds Enhance GxxxG-Driven Dimerization of the BNIP3 Transmembrane Domain: Roles for Sequence Context in Helix–Helix Association in Membranes. *Journal of Molecular Biology*, *396*(4), 924–936. https://doi.org/10.1016/j.jmb.2009.12.023

Lazaridis, T. (2003). Effective energy function for proteins in lipid membranes. *Proteins*, *52*(2), 176–192. https://doi.org/10.1002/prot.10410

Liang, B., & Tamm, L. K. (2016). NMR as a tool to investigate the structure, dynamics and function of membrane proteins. *Nat Struct Mol Biol*, *23*(6), 468–474. https://doi.org/10.1038/nsmb.3226

Liu, Y., Engelman, D. M., & Gerstein, M. (2002). Genomic analysis of membrane protein families: Abundance and conserved motifs. *Genome Biol*, *3*(10), research0054. https://doi.org/10.1186/gb-2002-3-10-research0054

Lomize, M. A., Lomize, A. L., Pogozheva, I. D., & Mosberg, H. I. (2006). OPM: orientations of proteins in membranes database. *Bioinformatics*, *22*(5), 623–625.

MacKenzie, K. R., Prestegard, J. H., & Engelman, D. M. (1997). A transmembrane helix dimer: Structure and implications. *Science*, *276*(5309), 131–133. https://doi.org/10.1126/science.276.5309.131

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., … Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, *102*(18), 3586–3616. https://doi.org/10.1021/jp973084f

Metcalf, D. G., Law, P. B., & DeGrado, W. F. (2007). Mutagenesis data in the automated prediction of transmembrane helix dimers. *Proteins: Structure, Function, and Bioinformatics*, *67*(2), 375–384. https://doi.org/10.1002/prot.21265

Mingarro, I., Whitley, P., Heijne, G. V., & Lemmon, M. A. (1996). Ala-insertion scanning mutagenesis of the glycophorin a transmembrane helix: A rapid way to map helix-helix interactions in integral membrane proteins. *Protein Science*, *5*(7), 1339–1341. https://doi.org/10.1002/pro.5560050712

Mitchell, J. B., & Price, S. L. (1990). The nature of the NH… O=C hydrogen bond: An intermolecular perturbation theory study of the formamide/formaldehyde complex. *Journal of Computational Chemistry*, *11*(10), 1217–1233.

Mravic, M., Thomaston, J. L., Tucker, M., Solomon, P. E., Liu, L., & DeGrado, W. F. (2019). Packing of apolar side chains enables accurate design of highly stable membrane proteins. *Science*, *363*(6434), 1418–1423. https://doi.org/10.1126/science.aav7541

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), E888-95. https://doi.org/10.1073/pnas.1319944111

Nash, A., Notman, R., & Dixon, A. M. (2015). De novo design of transmembrane helix-helix interactions and measurement of stability in a biological membrane. *Biochim Biophys Acta*, *1848*(5), 1248–1257. https://doi.org/10.1016/j.bbamem.2015.02.020

Partridge, A. W., Therien, A. G., & Deber, C. M. (2004). Missense mutations in transmembrane domains of proteins: Phenotypic propensity of polar residues for human disease. *Proteins*, *54*(4), 648–656. https://doi.org/10.1002/prot.10611

Rose, G. D., & Wolfenden, R. (1993). Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annual Review of Biophysics and Biomolecular Structure*, *22*(1), 381–415.

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

Sanders, C. R., & Myers, J. K. (2004). Disease-related misassembly of membrane proteins. *Annu Rev Biophys Biomol Struct*, *33*, 25–51. https://doi.org/10.1146/annurev.biophys.33.110502.140348

Senes, A. (2011). Computational design of membrane proteins. *Curr Opin Struct Biol*, *21*(4), 460–466. https://doi.org/10.1016/j.sbi.2011.06.004

Senes, A., Ubarretxena-Belandia, I., & Engelman, D. M. (2001). The Cα—H···O hydrogen bond: A determinant of stability and specificity in transmembrane helix interactions. *Proceedings of the National Academy of Sciences*, *98*(16), 9056–9061. https://doi.org/10.1073/pnas.161280798

Smith, S. O., Eilers, M., Song, D., Crocker, E., Ying, W., Groesbeek, M., Metz, G., Ziliox, M., & Aimoto, S. (2002). Implications of threonine hydrogen bonding in the glycophorin A transmembrane helix dimer. *Biophys J*, *82*(5), 2476–2486. https://doi.org/10.1016/S0006-3495(02)75590-2

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*(11), 1026–1028.

Subramaniam, S., & Senes, A. (2012). An energy-based conformer library for side chain optimization: Improved prediction and adjustable sampling. *Proteins*, *80*(9), 2218–2234. https://doi.org/10.1002/prot.24111

Therien, A. G., Grant, F. E., & Deber, C. M. (2001). Interhelical hydrogen bonds in the CFTR membrane domain. *Nat Struct Biol*, *8*(7), 597–601. https://doi.org/10.1038/89631

Tsemekhman, K., Goldschmidt, L., Eisenberg, D., & Baker, D. (2007). Cooperative hydrogen bonding in amyloid formation. *Protein Science*, *16*(4), 761–764.

Walshaw, J., & Woolfson, D. N. (2003). Extended knobs-into-holes packing in classical and complex coiled-coil assemblies. *Journal of Structural Biology*, *144*(3), 349–361.

Walters, R. F. S., & DeGrado, W. F. (2006). Helix-packing motifs in membrane proteins. *Proceedings of the National Academy of Sciences*, *103*(37), 13658–13663. https://doi.org/10.1073/pnas.0605878103

Wehbi, H., Gasmi-Seabrook, G., Choi, M. Y., & Deber, C. M. (2008). Positional dependence of non-native polar mutations on folding of CFTR helical hairpins. *Biochim Biophys Acta*, *1778*(1), 79–87. https://doi.org/10.1016/j.bbamem.2007.08.036

Wei, P., Liu, X., Hu, M.-H., Zuo, L.-M., Kai, M., Wang, R., & Luo, S.-Z. (2011). The dimerization interface of the glycoprotein Ibβ transmembrane domain corresponds to polar residues within a leucine zipper motif. *Protein Science*, *20*(11), 1814–1823. https://doi.org/10.1002/pro.713

Wei, P., Zheng, B.-K., Guo, P.-R., Kawakami, T., & Luo, S.-Z. (2013). The Association of Polar Residues in the DAP12 homodimer: TOXCAT and Molecular Dynamics Simulation Studies. *Biophysical Journal*, *104*(7), 1435–1444. https://doi.org/10.1016/j.bpj.2013.01.054

Yano, Y., Takemoto, T., Kobayashi, S., Yasui, H., Sakurai, H., Ohashi, W., Niwa, M., Futaki, S., Sugiura, Y., & Matsuzaki, K. (2002). Topological stability and self-association of a completely hydrophobic model transmembrane helix in lipid bilayers. *Biochemistry*, *41*(9), 3073–3080. https://doi.org/10.1021/bi011161y

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

# Chapter 3: Computational Methodology

## 3.1 Abstract

Recent advances in experimentation allow researchers to collect data in high-throughput. Computational tools and software are invented in complement, designing experiments to collect and analyze large datasets. My research implements a protein design algorithm paired with high-throughput sort-seq to characterize sequences designed to associate by sidechain packing. A portion of the methods are described in my publication in Chapter 2, but much of the rationale for specific details of my algorithm and the analysis are not covered. Being able to effectively share these tools and algorithms is necessary for conveying science and ensuring reproducibility (Greener et al., 2022; Mougeot et al., 2022; Na, 2020; van Iterson et al., 2012). By understanding minute details in previous research, future studies can improve and build upon the former results. This chapter focuses on explaining the details and rationale of my design algorithm, alongside the analysis utilized for my research.

## 3.2 Introduction

To study MP folding, researchers aimed to identify common structural patterns found among MP systems. The Protein Data Bank (PDB) was established to share discovered protein structures globally. This tool lets researchers deposit solved protein structures for others to access and evaluate (Berman et al., 2000). Initially, protein structures were studied primarily using x-ray crystallography, which has contributed to solving ~80% of MP structures (Kermani, 2021). MP structures have also been solved by nuclear magnetic resonance (NMR). Solid-state NMR bypassed the need for detergents in crystallography, obtaining structures of MPs with less than 50 residues within lipid bilayers or nanodiscs (Liang & Tamm, 2016). More recently, cryo-EM has been utilized to solve MP structures. Cryo-EM enables MP structures to be studied in a large variety of different environments, allowing researchers to study alternative structures of MPs by changing solubilization conditions. In addition to bilayers and nanodiscs, it is possible to solubilize and obtain the structures of MPs within detergents, saposin-lipoprotein nanoparticles, amphipols, and peptidiscs (Januliene & Moeller, 2021).

Despite advancements in MP structural characterization, many techniques take years to ascertain conditions to successfully solve structures in high resolution. MPs make up ~30% of known protein coding genes and integral MPs make up 60% of all drug targets (Arinaminpathy et al., 2009; Overington et al., 2006); however, only 4.6% of structures deposited in the PDB are MPs (April 2024; PDB). Solving the structures of MPs is difficult due to the need to reproduce interactions found between the lipid bilayer and protein. Additionally, MPs are difficult to express and purify in quantities necessary for structural experiments. Instead of focusing on structural determination, some groups utilize information from known structures to advance MP research. Using previously solved protein structures as datasets, researchers have developed computational algorithms and tools that identify common motifs and patterns among MP structures. These tools leverage our current understanding of structures to deduce the impact of forces such as vdW packing or hydrogen bonding.

Computational tools have been developed to help assess our understanding of the forces that drive MP association. By deriving the contributions of these forces to protein stability, we can predict structures from protein sequence and/or design sequences for given structures. Molecular dynamics simulations permit researchers to use established statistical and energetic potentials to simulate MP folding over time (Karplus & Petsko, 1990; MacKerell et al., 1998). Structure prediction tools use known information from previously solved structures to estimate the structure of MP folded states (Elofsson & von Heijne, 2007). Protein design strategies incorporate these structure prediction tools into simple model systems that can be used to assess the current understanding of MP folding (Ghirlanda, 2009). MP design to study TMH systems has been successful: peptides were engineered to associate with the TMH of integrins and a cytokine receptor EpoR (Mravic et al., 2024; Shandler et al., 2011; Yin et al., 2007), a non-natural integral MP was engineered to transfer electrons across the lipid bilayer (Korendovych et al., 2010), a 4-helix bundle was designed to transport $Zn^{2+}$ across the bilayer (Joh et al., 2014), and phospholamban was redesigned using packing interactions (Mravic et al., 2019).

My research expands on previous prediction and design studies. I surveyed possible TMH dimer conformations by extracting backbone helix-helix conformations from MPs found in the PDB. I then sampled different AA combinations, designing the interface along a standardized backbone sequence. Predictions of how these designed proteins associate were made using established energetic functions. Additionally, the stability of these designed proteins was assessed using a complimentary high-throughput assay (Anderson, 2019). This combination of techniques allowed me to develop an algorithm to design thousands of TMHs to study in high-throughput. In this chapter, I detail the development of my computational algorithm and tools used to analyze my high-throughput data.

### 3.3 Protein design algorithm

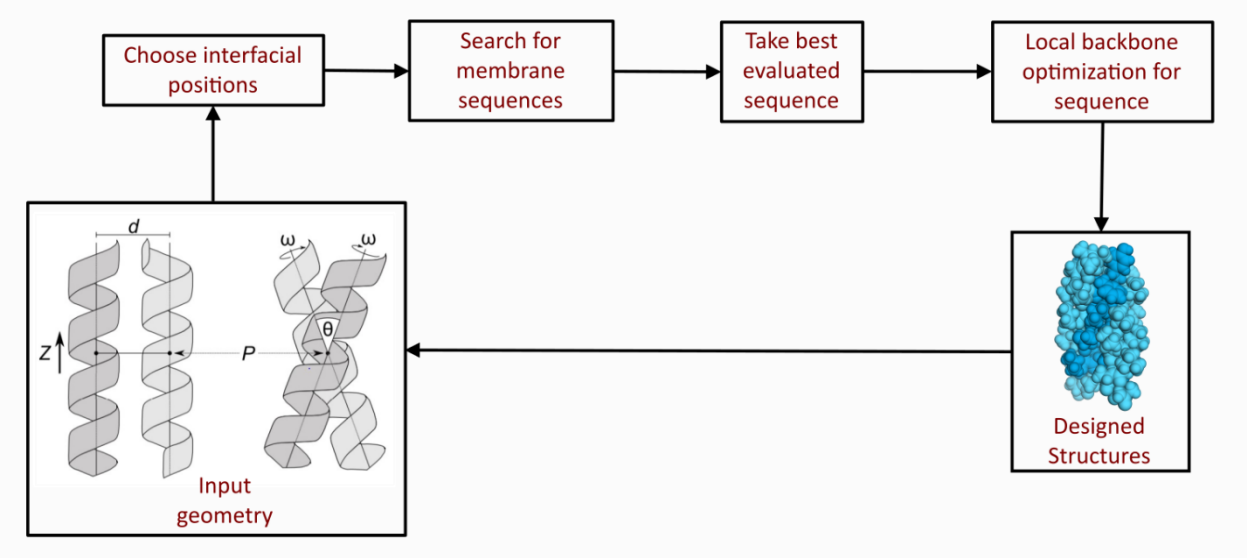



**Figure 3.1 Protein design algorithm workflow.** An input geometry is fed into the algorithm, and the interfacial positions are mutated searching for membrane protein like sequences. The most energetically stable sequence then undergoes a backbone optimization, where the geometry is optimized for the designed sequence. Designed structures are output and that designed geometry can be reinput into the algorithm to design other sequences with similar geometry and different energy.

To investigate the impact of vdW packing on MP association, I opted for a high-throughput design approach (Fig. 3.1). I created a sequence search algorithm that can design thousands of homodimer MP structures using MSL v. 1.1, an open source C++ library that is freely available at http://msl-libraries.org (Kulp et al., 2012). I coupled this algorithm with a structural backbone refinement program also built in MSL. Below, I detail the algorithm alongside experiments and tests that aided in its development.

### 3.3.1 Analysis of membrane protein PDBs

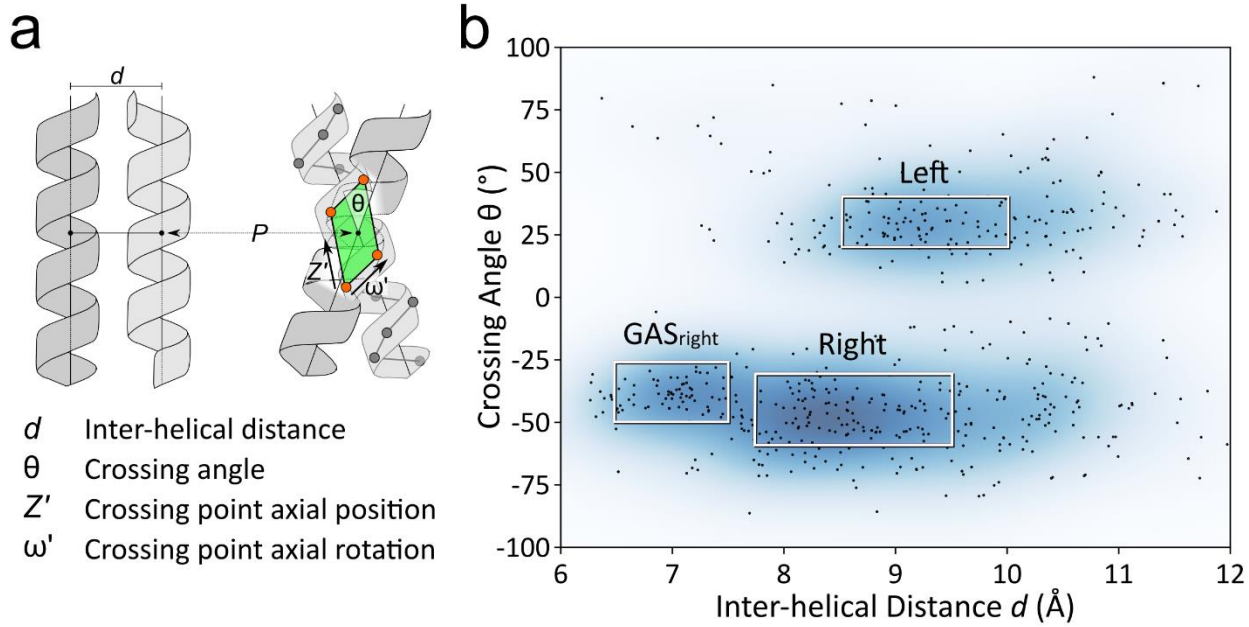



**Figure 3.2 MP helix-helix density distribution. a)** The geometric terms for homodimer proteins as referenced at the crossing point (P) between helices: interhelical distance (x-shift, d), crossing angle (θ), axial rotation (ω), vertical shift (z-shift, Z). **b)** Helix-helix interactions extracted from the PDB in February 2020. Plotted against the angle and distance of each interaction, and density map defined using kernel density estimation.

To computationally design homodimers, I first extracted backbone geometries from all unique MPs found in the Orientations of Proteins in Membranes (OPM) database (Lomize et al., 2006). To ensure redundant helical pairs were not extracted, the MP structures from OPM were trimmed by sequence similarity. Only unique structures with less than 30% sequence similarity were analyzed (Steinegger & Söding, 2017). We then developed a program in MSL that reads in a structure and identifies helical segments (Table ST3.1). We first identify the top and bottom z-axes of the membrane in the OPM structure. The segments of the protein within the membrane are then assessed for their helical nature. Cartesian points for quadruplets of Cα carbons are then assessed for their helical nature. The height (1.25-1.75Å), twist (90-110°), and radius (2.12-2.42Å) are measured, with loose restrictions against the ideal values (1.5Å, 100°, 2.27Å) for each parameter. To ensure that helices were of sufficient length for a dimeric interface, any helical segments composed of at least 13 AAs are extracted as individual helices. The distance is measured between Cα carbons on each unique helical pair. Any two helices with at least 3 Cα

carbons within 9Å of each other are extracted as an individual helical pair. I extracted two parameters: the distance (x-shift, **d**) and the angle (crossing angle, **θ**) (Fig. 3.2A), which I plotted as a scatterplot and analyzed using kernel density estimation (Fig. 3.2B).

The density plot identifies the most common interaction motifs for dimeric proteins. We expected regions of high density to correlate with designability: By applying my design algorithm to the most common geometries found in nature, they are more likely to successfully interact. First, it is important to define how these helical geometries are commonly referred to in scientific literature. Dimers can be characterized as right-handed or left-handed, depending on which dimer appears closer to us. If we look at the dimer to the right in Fig. 3.2A, the helix in front is pointing up and to the right. We refer to these dimers with a negative crossing angle as right-handed. The opposite is true for positive crossing angles, where the helix in front is pointing up and to the left. We refer to these dimers as left-handed.

There are three high density regions present in the MP helix-helix dataset (Fig. 3.2B). The first design region is present in the left-handed region. Helical pairs interact frequently in the range between 8.5 to 10 Å of interhelical distance and 20° to 40° of crossing angle. Because there is only a single patch of high density, we refer to this region as the left-handed design region (Left). The other two high density regions are found in the right-handed region. Helical pairs are found at a much broader range of distances from 6.5 to over 10 Å of interhelical distance. The region with the most density is found between 7.75 to 9.5 Å interhelical distance and -30° to -60° crossing angle. We refer to this as the right-handed design region (Right). Finally, the third region corresponds to a known dimerization motif called $GAS_{right}$, which is composed of dimers with a short interhelical distance (6.5-7.5 Å) and crossing angles of -25° to -55°. The $GAS_{right}$ is well characterized and known to be stabilized by a combination of vdW packing and the formation of interhelical weak hydrogen bonds between helices (Anderson et al., 2017; Mueller et al.,

2014). I decided to design this region as a control, allowing me to compare the stability between proteins stabilized solely by vdW packing and $GAS_{right}$.

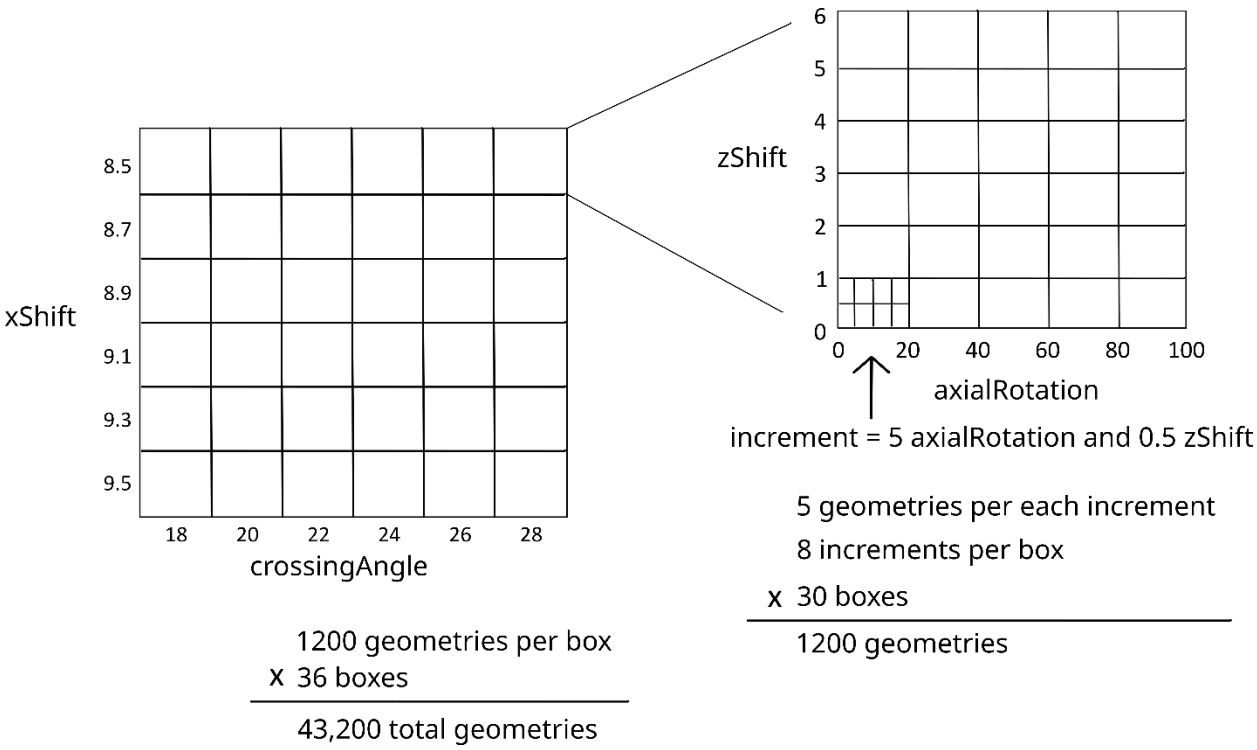



**Figure 3.3 Template Geometry Grid.** To determine the axial rotations (ω) and z-shifts (Z) that would be favorable for protein design, I created a grid of randomized geometries that were assessed for possible clashes at the dimer interface.

Two additional geometric features need to be considered when designing homodimer proteins: the rotation of the helix (axial rotation, **ω**) and the vertical shift in the membrane (z-shift, **Z**) (Fig. 3.2A). To determine these features for the corresponding angles and distances, I created a grid of template geometries for each design region (Fig. 3.3). I used MSL to place poly-Leu sequences at each geometry with Ala (Left and Right) or Gly ($GAS_{right}$) at the interface (described in section 3.3.3). TMHs were assessed for clashing at the interface by measuring the vdW energy. If the structure is clashing with the small AA Ala (Left and Right) or Gly ($GAS_{right}$) at the interface, then the structure is less likely to be designable as it would not be able to accommodate the rest of the larger design AAs. Any structures that corresponded to an energy of less than 10 kcal/mol were saved, allowing some leeway for potential clashes in each design region that could be mitigated with backbone refinement. I plotted the saved geometries on density maps

and extracted the ranges of axial rotations and z-shifts (Fig. S2.7, S2.8, and 2.9). Finally, I generated 1000s of geometries for each design region, where the angle and distance are chosen from the MP density map, and the rotation and z-shift are chosen from the identified ranges where clashing did not occur. These geometries were used as input backbone templates for protein design.

### 3.3.2 Choosing amino acids for MP design

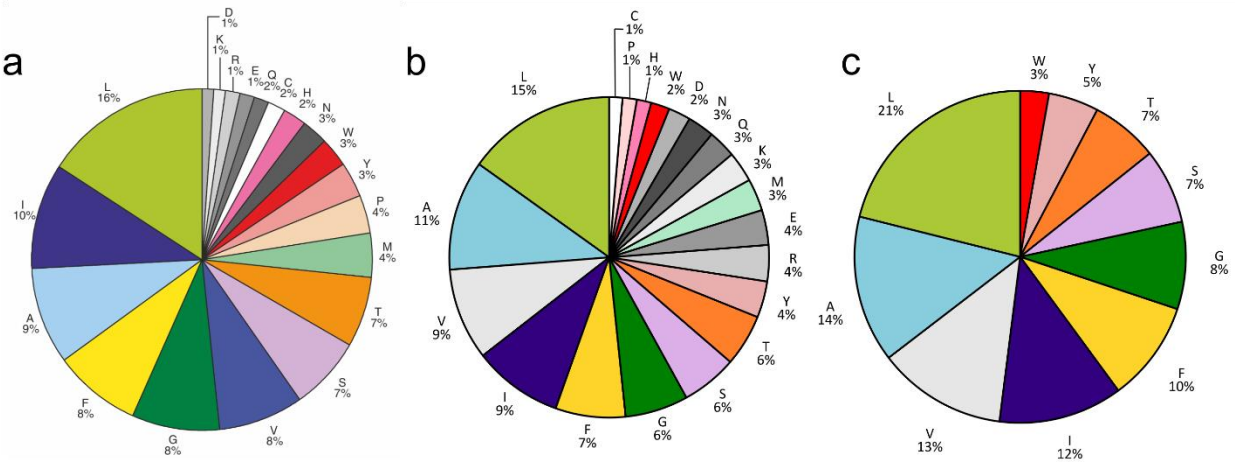



**Figure 3.4 Membrane Protein AA Percent Composition. a)** Percent composition of TM-helical regions for all sequences and consensus sequences (Liu et al., 2002). **b)** Percent composition of all helical pairs extracted from OPM. **c)** Percent composition of AAs chosen for design after removing non-design AAs and adjusting to add up to 100%.

Liu et al. 2002 identified the percent composition of AAs as found in all TM-helical regions in MP sequences (Fig. 3.4A). Inspired by the previous literature, I determined the composition of AAs found in my nonredundant extracted helical pair dataset. Since I aimed to study the effect of sidechain packing on association, I chose to design with a subset of AAs to decrease the potential for association by alternate forces at the interface. To prevent the formation of disulfide bridges, the two sulfur containing AAs (Cys and Met) were removed (Lim et al., 2019; SRINIVASAN et al., 1990). AAs with the potential to form charged interactions (Lys and Arg) were also excluded (Li et al., 2013). Histidine, often forming cation-π interactions, was excluded. AAs that often form hydrogen bonds (Asp, Glu, Asn, Gln) were removed, with the exclusion of Ser (7%), Thr (7%), and Tyr (5%) due to how frequently they were found in our TM helical pairs. Finally, Pro, which is known to form kinks in helices, was excluded from the design pool. The remaining 10 AAs

(Leu, Ala, Val, Ile, Phe, Gly, Ser, Thr, Tyr, Trp) were used for protein design (Fig. 3.4B). Each of these AAs was allowed to be designed along the interface during the sequence search described in section 3.3.5.

### 3.3.3 Defining the interface

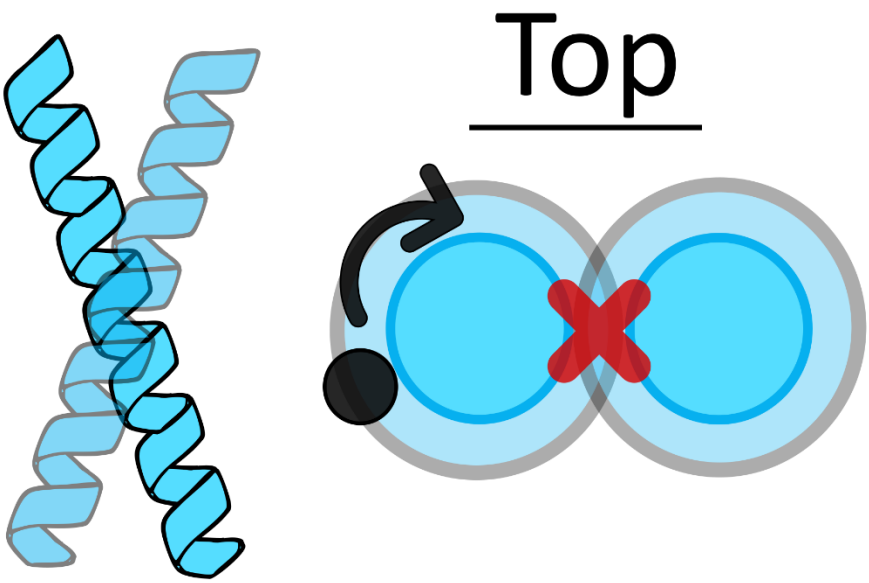



**Figure 3.5 Solvent Accessible Surface Area (SASA).** Side view and top-down view of helical dimers. SASA is determined by the amount of area on both helices that is not in contact. Consider rolling a ball along the surface of the protein structure: The ball can only access parts of the protein that are not in contact, or outside of the X (the most buried region of the dimer interface).

To reduce heterogeneity in protein expression, I designed the interface of a standardized TM helix of 21 AAs consisting of a poly-Leu backbone, a strategy previously applied to study the association of $GAS_{right}$ proteins (Anderson et al., 2017). In my first protein design run, I used Solvent Accessible Surface Area (SASA) to identify the interfacial positions of dimers placed at a specified geometry from my MP analysis (Fig. 3.5). SASA was calculated for each position on the dimer and the interface defined as the 8 positions with the least amount of access to the solvent, or the most buried positions. Although the designed sequences were able to associate, our energy score showed little correlation to association. However, I found that sequences with similar interfaces had better correlation with the predicted energy score (Fig. S3.1).

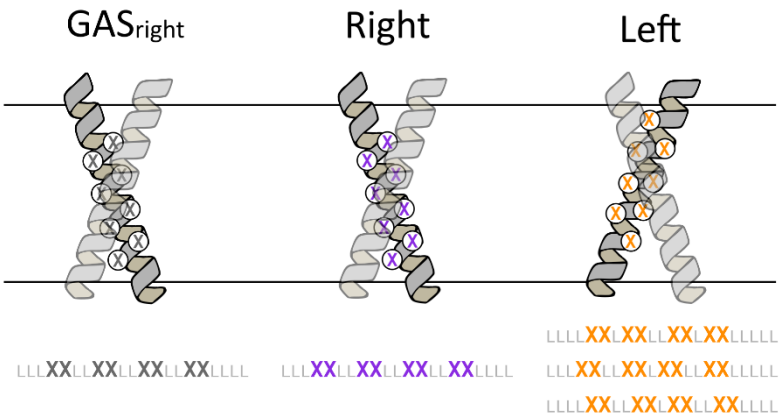



**Figure 3.6 Protein design interfaces.** Sequence design was performed along a poly-Leucine backbone sequence with defined interfaces for each design region. Interface designated by positions with X.

For subsequent design runs, I standardized the interface for each region (Fig. 3.6). For GASright, I used the typical pattern that spaces two interfacial positions (x) with two fixed positions (L), resulting in a LLL**xx**LL**xx**LL**xx**LL**xx**LILI pattern (Anderson et al., 2017; Mueller et al., 2014; Russ & Engelman, 2000). This same interface was applied to Right designs, which has similar geometry outside of the larger interhelical distance. In the Left region, I used an interface that applies the typical L**xx**LL**x**L heptad repeat common in leucine zippers, knobs-into-holes, and coiled coils (LLLL**xx**L**xx**LL**xx**L**xx**LLILI) (Ash et al., 2004; Bornberg-Bauer et al., 1998; Walshaw & Woolfson, 2003). Through visual inspection of the poly-Leu structures assessed for clashing in section 3.3.1, I found that the Left region could accommodate three potential interfaces, each of which are used for Left design (Fig. S2.10). Standardizing the interfaces allowed me to come up with a consistent mutational strategy to assess my proteins for their association at the given interface that is described in section 3.3.6.

### 3.3.4 Developing the energy terms

To predict the stability of my designed proteins, I applied the same set of energy terms used previously by CATM (Anderson et al., 2017; Mueller et al., 2014): CHARMM_VDW for vdW packing (MacKerell et al., 1998), SCWRL4_HBOND for hydrogen bonding (Krivov et al., 2009), and CHARMM_IMM1 to estimate the interactions found in the membrane environment (Lazaridis, 2003). These terms are calculated for each

protein during the sequence search to find the best interfacial sequence for the input geometric structure. To determine the stability of my designed dimers, I calculated the interaction energies of the dimer structure and two monomers, and then subtracted the monomer energy from the dimer energy:

$$\textit{Eq. 3.1} \qquad \text{Total Energy} = \text{Dimer} - (2 \times \text{Monomer})$$

However, calculating the monomer energy for each sequence during the sequence search is time consuming, resulting in a bottleneck in the algorithm and limiting the number of sequences I could design. To account for this, I developed an energy term that estimates the monomer energy of each sequence.

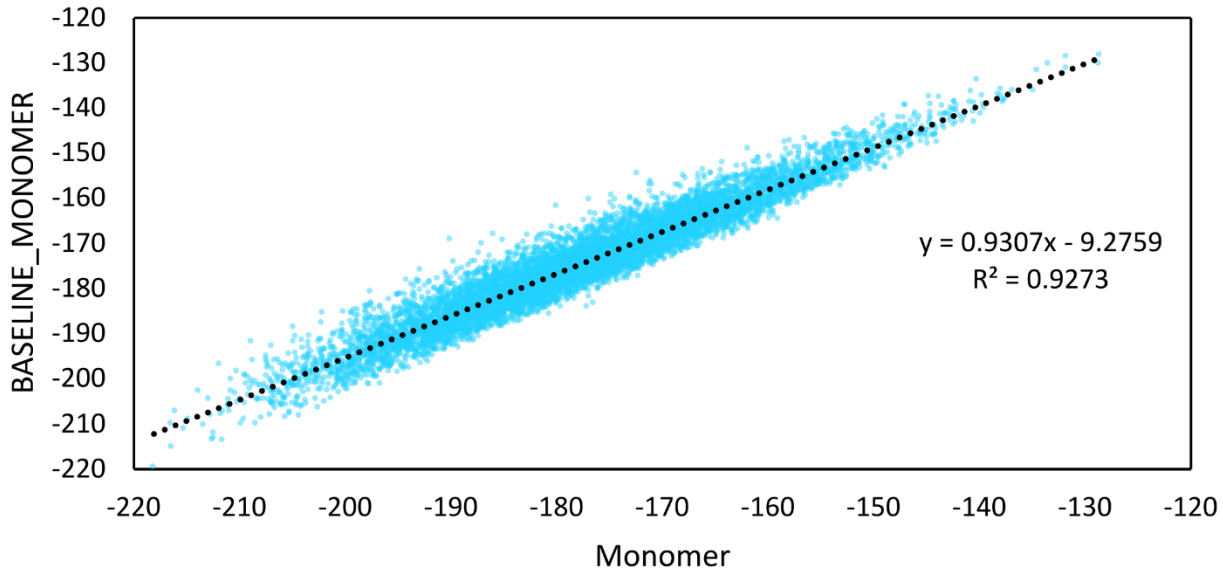



**Figure 3.7 Developing the BASELINE_MONOMER term.** A baseline energy term was developed to increase computational speed, estimating the stability of the sequence as a monomer. A strong correlation was found between the calculated Monomer energy (x-axis) and the BASELINE_MONOMER (y-axis).

The BASELINE_MONOMER term was created by measuring the energy of the previously mentioned terms for each individual amino acid on a monomeric helix. I calculated the energetics of 10000 random sequences and measured the self and pair energies for each amino acid. Self-energy represents the energy contribution for an individual amino acid to the protein stability alone, while the pair energies represent the energy contribution between any two interacting amino acids (Desmet et al., 1992). I measured the pair energies for all AA pairs on the sequence and found that pair interactions between amino acids more

than 10 bases away from each other returned an energy of 0 kcal/mol. Therefore, only pair energies for amino acids up to 10 bases away were calculated. From each iteration, I calculated the average of all self and pair energies and saw a strong correlation between the measured monomer energy and the BASELINE_MONOMER term (Fig. 3.7). This BASELINE_MONOMER term was made only for the subset of amino acids used in design (Fig. 3.4B) and would need to be recalculated to establish the term for any additional amino acids. This term was helpful in decreasing computational time, enabling me to design 1000s of sequences within a week.

Another issue I encountered was that many of my initial designs were often composed of only 2-3 different AAs. This result could impact our protein expression and insertion, as natural MP sequences are often made of a diverse set of AAs. To account for this sequence diversity, I developed a SEQUENCE_ENTROPY term that outputs an energy based on how similar an AA sequence is to the composition of a natural MP sequence (Fig. 3.4). To convert the composition of AAs in a membrane sequence to an energy term, I utilized the following equation based on the Boltzmann entropy formula:

*Eq. 3.2* $$SEQUENCE\_ENTROPY = -\log(probability) \times RT$$

where R is the gas constant and T is temperature defaulted to 298K (RT = 0.592). To compute the sequence entropy, I calculated the probability that the sequence is expressed as an MP. First, the number of each AA (#AA) is counted within the sequence. Using these values, I then calculated the number of possible permutations for the sequence. This is determined using the following equation:

*Eq. 3.3* $$permutations = {}^{n!}/_{(\#AA1! \times \#AA2! \times \dots)}$$

where n is the number of positions, divided by the total number of combinations possible for a sequence of AAs, which is the product of the factorials for each #AA. The probability is computed using the frequency

of each AA in MP sequences (freq_AA) to the power of the number of each AA in the sequence multiplied by the permutations:

*Eq. 3.4* $$probability = (freq\_AA1^{\#AA1} \times freq\_AA2^{\#AA2} \times \ldots) \times permutations$$

This probability is inserted into the sequence entropy equation, returning a value that can be applied as an energy term for each sequence. My algorithm utilizes this term as an additional energy, predicting the stability of my designed homodimer sequences and adjusting the energy by the likelihood that they are found in natural MP sequences.

### 3.3.5 Sequence search

For each input geometry, the algorithm first defines the best sequence using the Self Consistent Mean Field (SCMF) theory as implemented in MSL. This method estimates the conformational entropy of each design AA as a probability that it is present within the dimer interface (Koehl & Delarue, 1994). The sequence from the SCMF is then run through a Monte Carlo (MC) sequence optimization: A random position on the interface is switched to a random AA, and the energy is calculated using the combination of the CATM energy terms and the developed energy terms defined in section 3.4.3.

Each energy term can be multiplied by an optional weight variable, meaning the total energy can be changed proportionately by the given weights. Previous research in our lab defaulted the weights of the CATM energy terms to 1, which was repeated in this study. To determine if SEQUENCE_ENTROPY performed better at different weights, I ran a test using weights of 1, 5, 10, 50, and 100 and calculated the AA composition in these designs. Weights greater than 10 were found to be optimal for mimicking the AA composition found in TM helical pairs. I chose to use 10 as it resulted in the SEQUENCE_ENTROPY term

affecting the total energy at the same order of magnitude as the other terms. The equation for the search energy is as follows:

*Eq. 3.5*

$$\text{Search Energy} = \text{Dimer} - (2 \times \text{BASELINE\_MONOMER}) - (\text{SEQUENCE\_ENTROPY} \times \text{Weight})$$

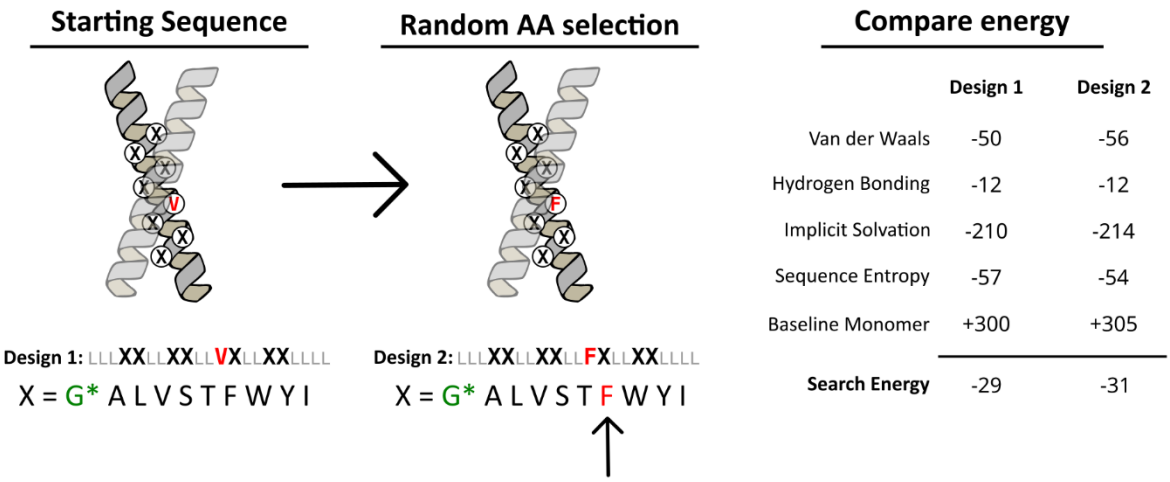



**Figure 3.8 Sequence Search Example.** A random position on the input sequence (red) is selected and switched to another AA. The energy is then calculated for the new sequence and compared to the energy of the previous sequence. Sequences with more stable energies (more negative) are always accepted.

The search energy is used for the acceptance criteria during the sequence search, and the MC searches for a multitude of sequences before reaching an energetic minimum. The sequences accepted during the search are saved into an output trajectory file alongside the energy (Table ST3.2). The sequence with the best total energy is saved, and that single design undergoes backbone refinement (Fig. 3.8).

### 3.3.6 Backbone refinement

After initially starting with a specific backbone template, the newly designed sequence undergoes an MC based structural refinement procedure. The structure undergoes MC backbone perturbations, where one of the four inter-helical parameters (Fig. 3.1A: d, θ, ω, Z) is chosen and shifted during each cycle. The total energy for the refined structure is used as the acceptance criteria, with the BASELINE_MONOMER energy being replaced by the computed monomer energy and the SEQUENCE_ENTROPY term no longer applied:

*Eq. 3.1* $$\text{Total Energy} = \text{Dimer} - \text{Monomer}$$

The backbone refined geometry can be input into the sequence search to find other designable sequences. The sequence, energetics, geometries, and their corresponding structures are output to a folder for analysis (Table ST 3.2). After initially only refining my backbone in my design script, I developed another program in MSL that runs a more thorough refinement protocol. I found that this program improved the structure and energetics described in section 3.3.6. This updated algorithm was used for determining energies for all sequences and is detailed in Fig. S3.2 and Table ST3.3.

## 3.4 Analysis

### 3.4.1 Software

The following calculations, analyses, and graphing were implemented and performed using Python v. 2.7.

Relevant packages include:

Pandas: (McKinney, 2011)

DNAChisel: (Zulkower & Rosser, 2020)

Numpy: (McKinney, 2012)

Matplotlib: (Tosi, 2009)

Seaborn: (Waskom, 2021)

Scipy: (Virtanen et al., 2020)

Sklearn: (Pedregosa et al., 2011)

Logomaker: (Tareen & Kinney, 2019)

Pymol: (DeLano, 2002)

All programs and code can be found on **Github**.

### 3.4.2 Design analysis and mutational strategy

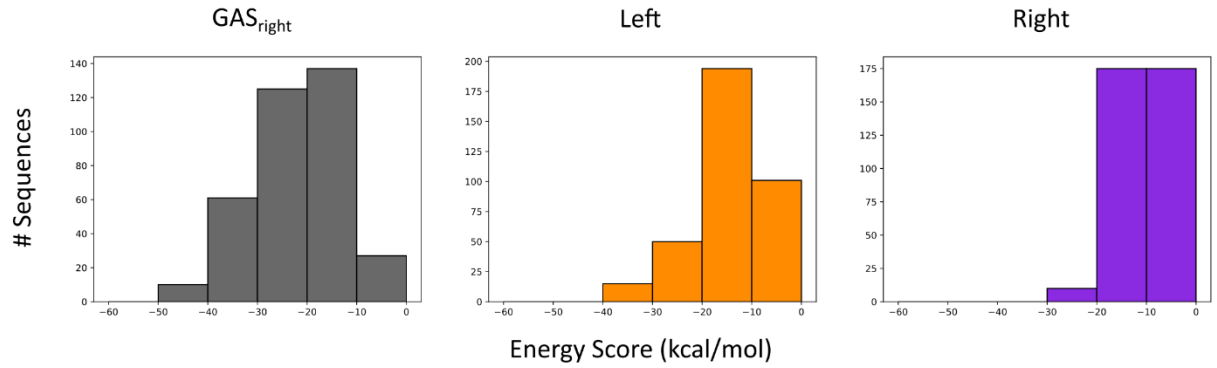



**Figure 3.8 Energy Score Spread.** Frequency of energy scores for sequences chosen for experiments.

To analyze my designs, I wrote a script that compiles all design directories and outputs a variety of plots (Fig. S3.3 and Table ST3.6). To ensure that we have a spread of energy scores for our design pool, a subset of 1045 designed sequences were selected for experiments (Fig. 3.8 and S3.4). Because it is not feasible to solve the structures of all designed sequences, I decided on a mutational approach to confirm that my designs dimerize at the designed interface. Mutations expected to decrease association were chosen by two additional programs developed in MSL.

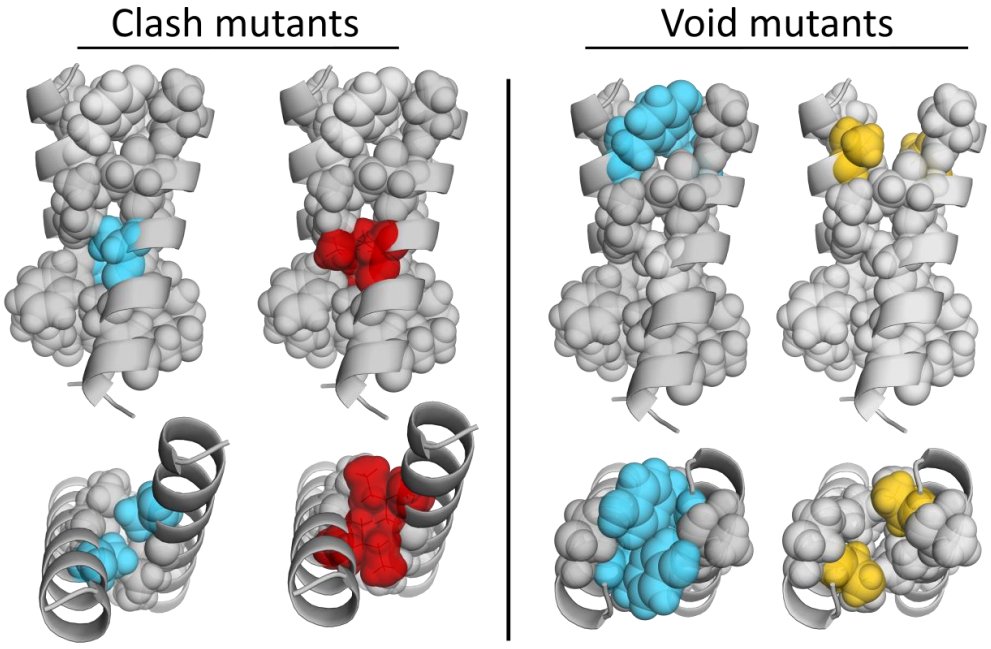



**Figure 3.9 Clash and Void mutations.** Mutations to Ile (Left) results in interfacial positions on one helix overlapping with atoms on the opposite helix. Mutations to Ala (Right) results in holes at the interface. Each of these mutations was expected to decrease association.

The first program identifies clash mutants, where an interfacial position was mutated to an Ile which can protrude into the opposing helix, often disrupting the ability for a protein to associate (Table ST3.4). The second program identifies void mutants, where an interfacial position was mutated to a smaller Ala, aiming to decrease the amount of packing at the interface (Table ST3.5). The designed protein structure is read into each program as an input, and the interface mutated one at a time to either Ile (clash) or Ala (void). Each program outputs either an energy score (clash) or SASA (void) for the mutant sequence. The two clashing mutations with the highest energy (least stable) and the two void mutants with the largest increase in SASA (less packed) were chosen for experiments. We expected these mutants to enable us to determine if our proteins associate by the designed interface (Fig. 3.9). The designed sequences, their respective mutants, and a variety of control sequences were ordered in an oligo pool library from Twist Bioscience and cloned into plasmids for TOXGREEN sort-seq.

### 3.4.3 Fluorescence reconstruction

As detailed previously in Anderson, 2019 thesis, a library of genes coding for designed TMs is cloned into the ToxR plasmid, allowing each design to be expressed in *E. coli*. These plasmids are used to assess dimerization by TOXGREEN, which is detailed in section 1.3.2. Each cell outputs fluorescence corresponding to the dimerization propensity of the expressed sequence. A population of *E. coli* containing the library of sequences is sorted into separate bins through fluorescence activated cell sorting (Fig. S3.5). Plasmids obtained from the sorted populations of *E. coli* are sent for Next Generation Sequencing (NGS). The sequencing returns counts for sequences found in each bin, which are used to reconstruct the fluorescence profile for each sequence. This reconstructed fluorescence is used to assess the dimerization propensity of all sequences in the population. Reconstructed fluorescence levels were calculated as a weighted average (Kosuri et al., 2013). This method normalizes the reads per protein per bin with the fraction of the population found in that bin. The normalized fractional contribution of each bin (j) for each protein (i), $a_{ij}$ is calculated as:

*Eq. 3.6*

$$a_{ij} = \frac{\frac{f_j \cdot c_{ij}}{\sum_i c_{ij}}}{\sum_j \frac{f_j \cdot c_{ij}}{\sum_i c_{ij}}} \qquad p_i = \sum_j a_{ij} \cdot m_j$$

i = variant
j = bin
c = count
f = fraction of population
a = normalized fractional contribution
m = median
p = reconstructed fluorescence

where the normalized fractional contribution is then multiplied by the median fluorescence of that bin ($m_j$). Finally, the contributions for each sorted bin are summed to determine the reconstructed fluorescence. The sequencing data is run through a Python pipeline that computes the above reconstructed fluorescence for all sequences found in the NGS data (Table ST3.7). The reconstruction resulted in 949 of our 1045 designs (91%) present within each replicate of the NGS data.

### 3.4.4 TOXGREEN conversion

Studies using TOXCAT systems typically normalize the dimerization propensity by comparing dimerization propensity to the well-studied strong dimer Glycophorin A (GpA), which is included as a control in each of our libraries. This normalization is calculated as a percentage of GpA fluorescence (%GpA). The reconstructed fluorescence is converted to %GpA using the following equation:

*Eq. 3.7*

$$\%GpA = {}^{Design\ Fluorescence}/_{GpA\ Fluorescence} \times 100\%$$

To calibrate our fluorescence properly to %GpA, we also include a variety of control sequences that we have previously studied using TOXGREEN. Upon initial inspection, the controls present in our experiment were reconstructed to a lower %GpA value than we've previously found in TOXGREEN.

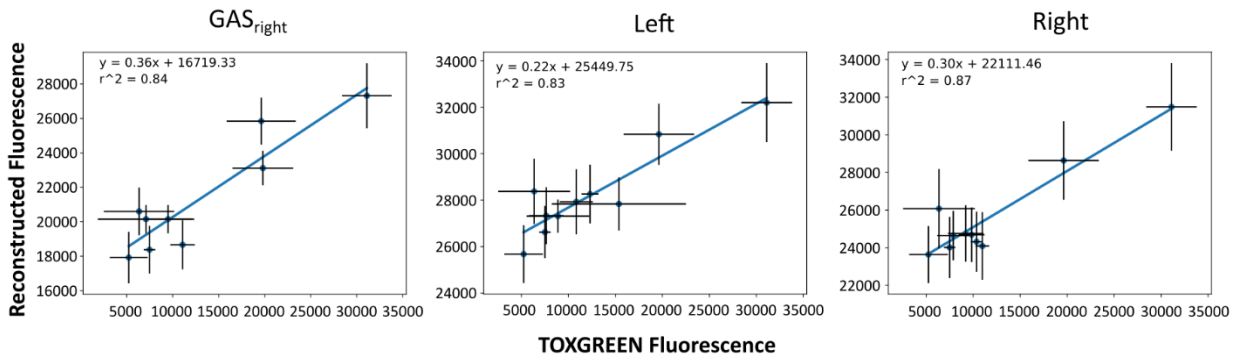



**Figure 3.10 TOXGREEN Converted Fluorescence.** Fluorescence from reconstruction is converted to TOXGREEN fluorescence using correlation plots between a set of controls and subset of designs previously tested in TOXGREEN**.**

I conducted a separate low-throughput TOXGREEN experiment on the control sequences and a subset of my designed sequences. When I compared the TOXGREEN to the reconstructed data, the reconstructed values were noticeably smaller for most sequences. However, when we plotted TOXGREEN versus the reconstruction, we found a clear correlation between the two datasets (Fig. 3.10). To calibrate our reconstruction to TOXGREEN, we converted the values from the reconstruction data to TOXGREEN fluorescence (Table ST3.8). We applied the equation found by the correlation between reconstructed fluorescence and TOXGREEN fluorescence. The reconstructed fluorescence is multiplied by the slope and then subtracted by the y-intercept of the correlation. The values for each sequence are averaged with their corresponding replicates and normalized to the GpA sorted in each design population. This conversion allows us to differentiate between different levels of dimerization propensity as seen previously in TOXGREEN experiments: monomers (0-35%), weak dimers (35-60%), and strong dimers (>60%).

### 3.4.5 Determining proper membrane insertion

A liquid maltose growth assay was run in triplicate as in Anderson, 2019 thesis. Briefly, cultures composed of *E. coli* containing plasmids for the designs, mutants, and controls are grown overnight. These are normalized by OD600 in the morning and the normalized population added to flasks of liquid maltose media and grown for 36 hours (36H), with timepoints taken every 6 hours. The populations for each timepoint are spun down, plasmids extracted through miniprep, and prepared for NGS. Within each

population are control sequences that are known not to insert in the membrane, as shown by failure to grow on maltose plates. To assess whether the sequences properly insert into the membrane, we compare the relative abundance of our designs to these controls. The relative abundance from the overnight growth (0H) and the growth in liquid maltose at 30H to determine the ability to insert:

$$Eq.\ 3.8 \qquad Relative\ Abundance = {}^{NGS\ Count\ 0H}/_{NGS\ Count\ 30H}$$

Sequences that are more abundant than these controls are considered properly inserted. 708 of the 949 designs present (75%) pass our insertion test.

### 3.4.6 Identifying proteins associating by designed interface

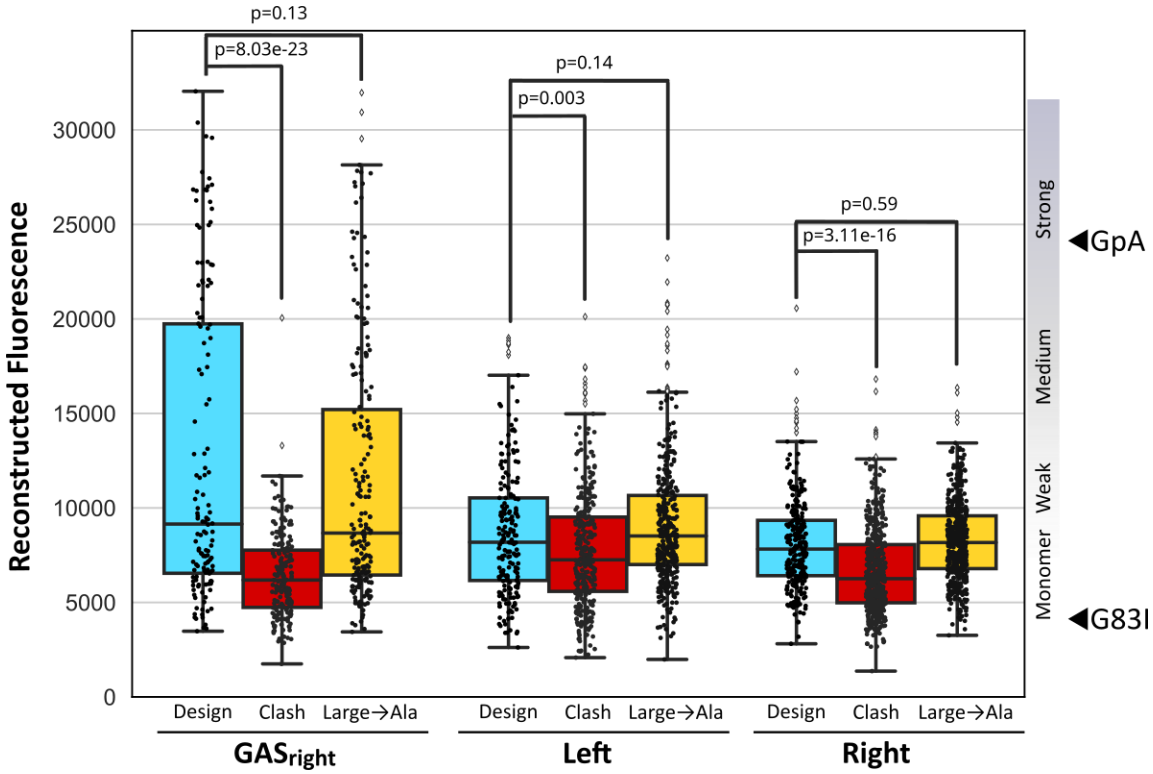



**Figure 3.11 Reconstructed Fluorescence of Designs versus Mutants.** Fluorescence of Designs (light blue), clash (red), and void/Large->Ala (yellow) mutants. Significance for designs versus each mutant group is calculated and displayed above each group.

To determine if sequences dimerize along the designed interface, we sought to identify sequences where the mutant results in a significant decrease in association. We analyzed the fluorescence for the clash and void mutants against the fluorescence of our designed sequences (Fig. 3.11 and Table ST3.9).

When comparing the clash mutations to the fluorescence of the design, we saw a significant decrease in their association ($p<0.05$). However, the void mutants did not show this same decrease, often resulting in similar fluorescence as the WT designs. This data suggests that mutating larger amino acids to the smaller Ala to reduce packing does not significantly impact association. It is possible that these mutants dimerize by an alternate interface than our designed structures. We decided to move forward by trimming our data using the clashing mutants, which appear to disrupt association by our designed interface. We trimmed our data for any designs where the clashing mutation was monomeric (< 35% GpA). This resulted in 379 out of the 708 designs that pass the maltose test (54%) associate by our designed interface.

### 3.4.7 Comparison to energetics

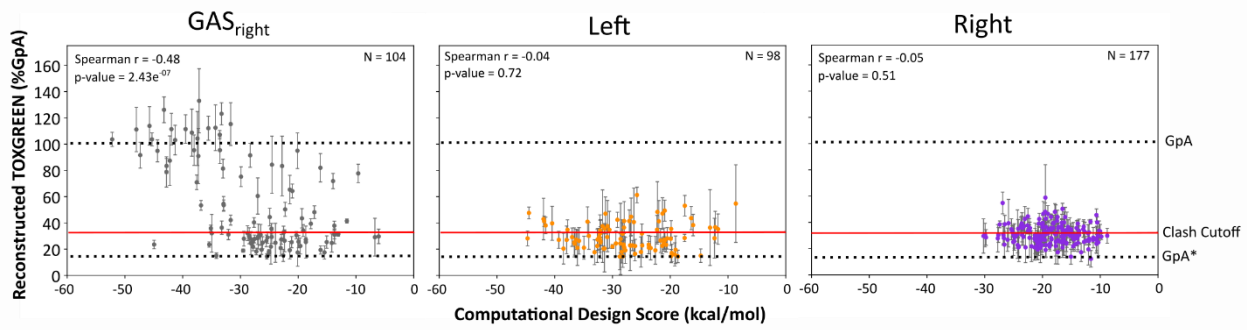


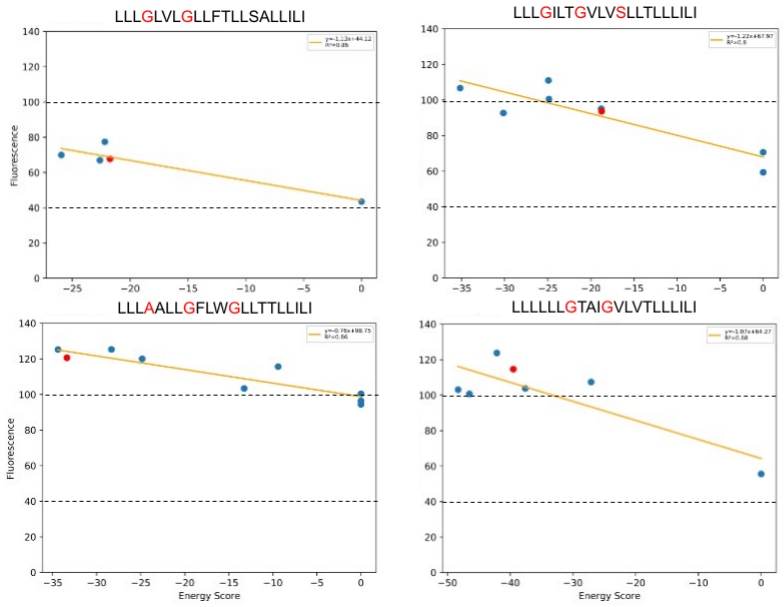

**Figure 3.12 Energy vs Reconstructed TOXGREEN.** Dimerization propensity in terms of %GpA against the predicted energy score. Control sequence GpA and it's monomerizing mutant (GpA*) are represented as dashed lines at 100% and 18% GpA, respectively. The cutoff for clash mutants is represented as a solid red line at 35% GpA.

We plotted the energy score against the dimerization propensity in terms of %GpA for each protein and separated the data by design region (Table ST3.10). Spearman ranked correlations between the energy score and the dimerization propensity were calculated. The energy score does not correlate well to proteins outside of the GAS$_{right}$ region (Fig. 3.12). This data suggests that although we were able to design sequences that associate (>35% GpA), we are unable to predict the dimerization propensity of proteins associated solely by vdW packing using our energetics. However, many of our designs outside of the GAS$_{right}$ associate as weak dimers (35-60% GpA). This suggests that our energetics may not be well tuned to predict weakly dimerizing proteins.

### 3.5 Summary

In this section, I detailed the computational methods I used in my research. Because I did not produce a web-based design script, describing these methods is necessary for reproducibility and understanding by future researchers. The methods described include the following: protein design algorithm, defining the interface for geometries, development of energy terms, backbone repack scripts, and mutation programs built using MSL. I also describe analysis scripts built in python which include the following: fluorescence reconstruction from NGS, conversion to TOXGREEN fluorescence, membrane insertion of maltose NGS, and identifying proteins by clash mutations. The outputs for each of these programs are referenced in the supplementary figures and tables, and each of these programs can be found on **Github**. Finally, I reference the results from my paper that can be improved upon in future experiments. Ways to further improve my design algorithm and to study the impact of other forces on MP folding and association is further described in Chapter 4: Future Directions.

## 3.6 Supplementary Figures




| Sequence | xShift | crossingAngle |
|---|---|---|
| LLLLLLGTAIGVLVTLLLILI | 6.35232 | -50.6541 |
| LLLAALLGFLWGLLTTLLILI | 7.26946 | -29.9582 |
| LLLGILTGVLVSLLTLLLILI | 7.72338 | -31.3153 |
| LLLGLVLGLLFTLLSALLILI | 8.39897 | -30.639 |

**Figure S3.1 Correlation for similar interfaces.** Energy scores (x-axis) plotted against the dimerization propensity in terms of %GpA. Correlations for interfaces with $R^2 > 0.6$ from my first sort-seq design run.

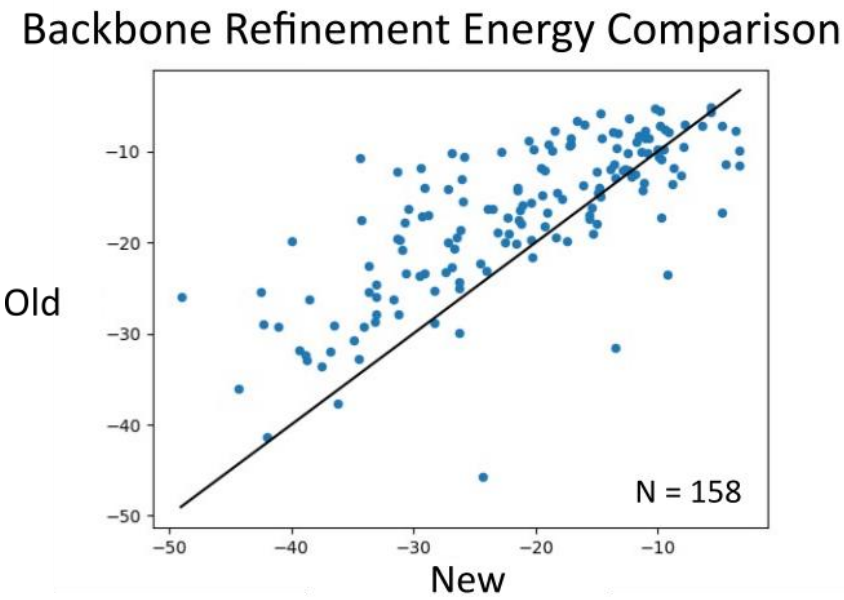



**Figure S3.2 Updated backbone refinement comparison.** Energies from previous backbone refinement (y-axis) versus the improved version (x-axis). Line to delineate x=y. Most of the points are found to the left of the line, showing that the improved version of the refinement results in more stable energies and better packed structures (data not shown).

The program reads an input structure and sequence. To ensure that the mutated sequence can find alternate dimer interfaces, the structure undergoes a more detailed MC backbone refinement. In the original backbone refinement, each geometric shift randomly alters the structure by choosing a value between 0 and an input upper limit. For example: When the geometric x-shift is chosen, a random value from 0-0.5 Å is applied to the structure. In this version, the geometric term procedurally decreases to a lower limit, such as 0.1 Å. Each cycle, the chosen geometric term is decreased by multiplying it by the metropolis criteria until it reaches the lower limit. Once the lower limit is reached, this value is always used when this term is shifted again. After initially testing this process on my designs, I found that the new backbone refinement resulted in more stable energies for my designed proteins. The energetics from this refinement is utilized to evaluate my structures and mutants against their reconstructed fluorescence in sort-seq.

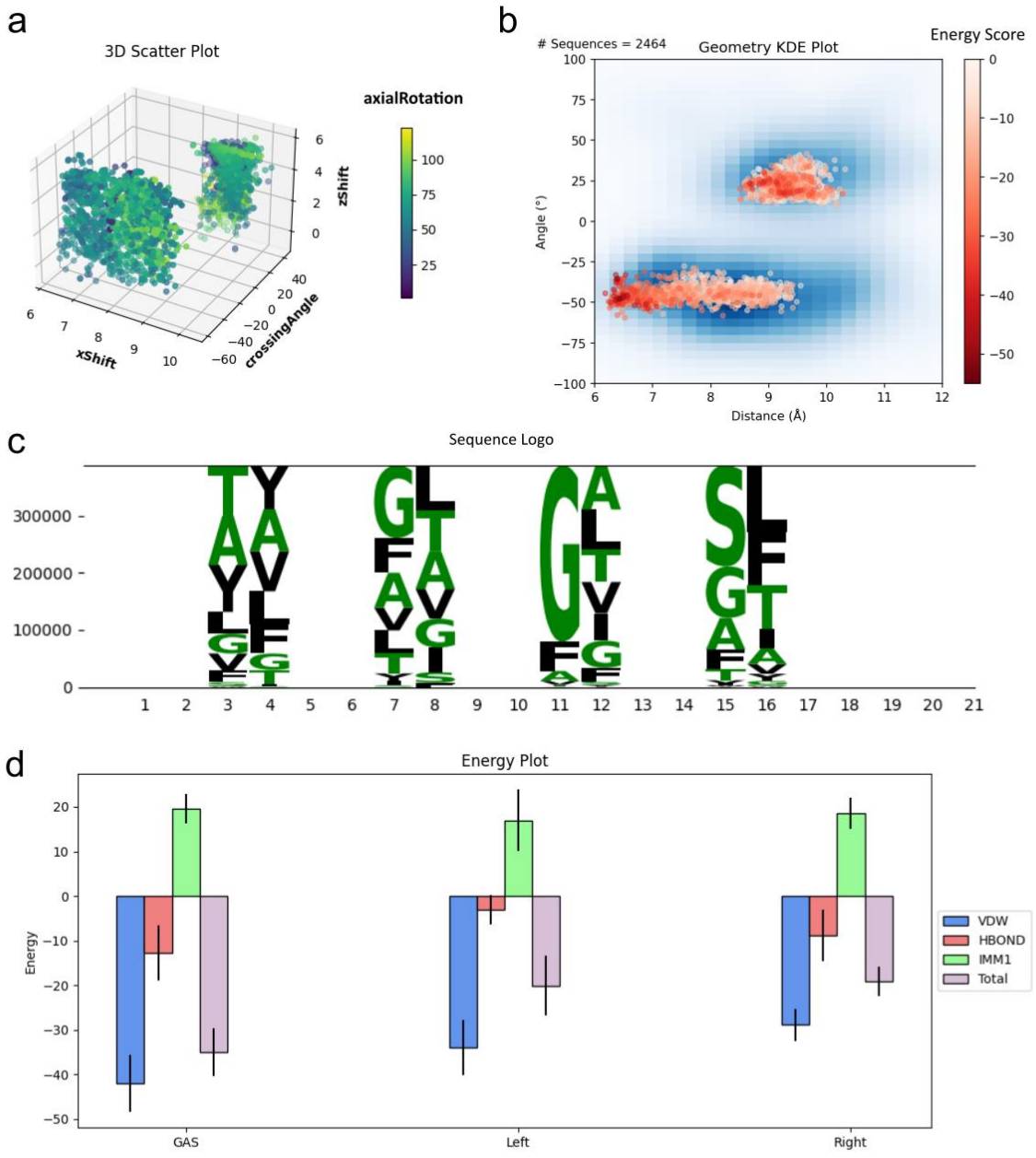



**Figure S3.3 Design Analysis Outputs. a)** Scatterplot of output of all geometric parameters. **b)** Scatterplot of xShift and crossingAngle with energy score color bar overlaid on the density map. **c)** Sequence logos for the interface of each design region are output. **d)** A bar plot of the average energy score for each energy term.

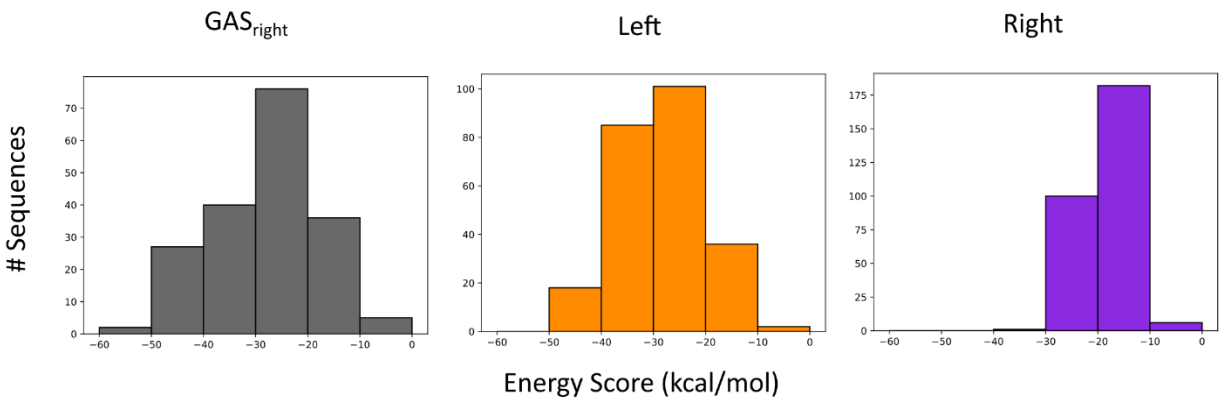



**Figure S3.4 Finalized Data Energy Score Ranges.** Range of energy scores for sequences found to associate by our designed interface. Energies were recalculated using backbone repack described in S3.2.

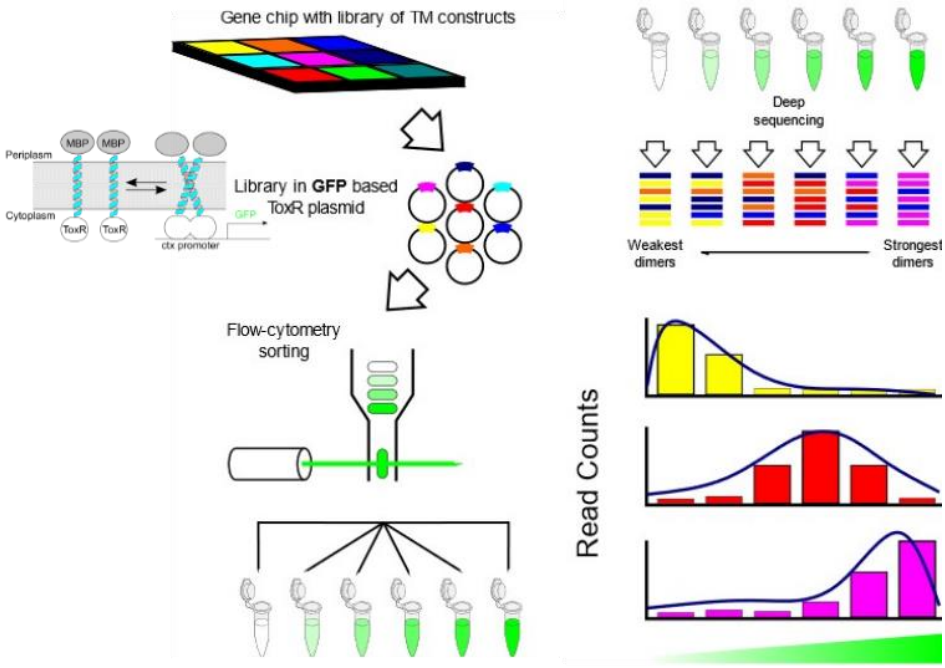



**Figure S3.5 Sort-Seq.** A library of genes coding for my designed sequences is cloned into the ToxR plasmid and cloned into *E. coli* cells. The fluorescence from each cell is assessed through fluorescence activated cell sorting. Sorted populations are then sent for deep sequencing, where we get counts for each of the sequences found in each bin. These counts are used to reconstruct the fluorescence profile for each sequence, allowing us to determine the dimerization propensity for every sequence in the population.

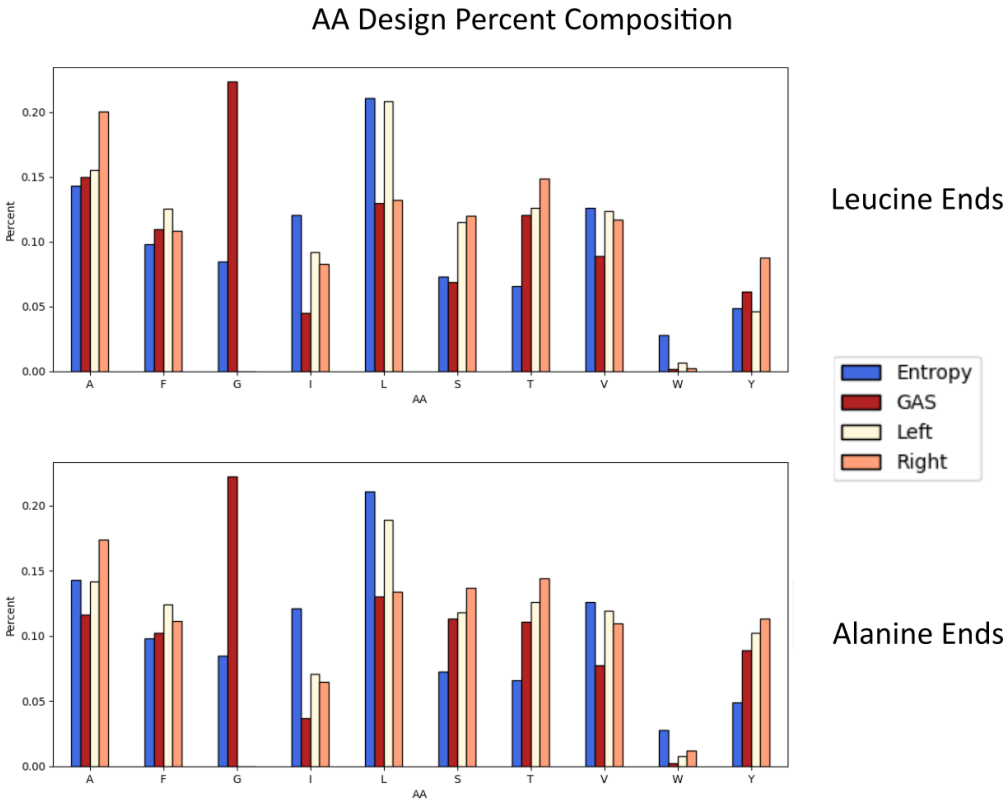



**Figure S3.6 Leucine and Alanine End Interface Composition.** Amino acid composition for sequences designed with Leucine vs Alanine Termini.

During my design run, I encountered an issue upon visual inspection of some of my poly-Leucine standardized sequence PDBs: interfaces often included voids to accommodate Leucine at the termini, preventing clashing interactions. Although these designs had a considerable amount of vdW packing according to our energetics, we wanted to ensure that the interface was driving association. I chose to repeat my design process with the smaller AA Alanine at the termini. These structures were found to include less voids and a well packed interface. Although the termini are unchanged in our experiment, we included these Alanine termini designs in our dataset with the assumption that helices in the experiment would be more flexible to accommodate these interfaces than our rigid helices. The sequences designed with Alanine termini were converted to Leucine ends in the backbone repack script, and all energetics were assessed for sequences with Leucine termini.

## 3.7 Supplementary Tables

### 3.7.1 MSL Scripts

**Table ST3.1: interhelicalCoordinates.cpp**

| Output File | Description |
|---|---|
| fit_#-#.pdb | Helices in a straight helix representation made by MSL |
| helix_#.pdb | Structure of each identified helix of at least 13 AAs |
| pair_#-#.pdb | Structure of each pair of helices, included with centroids of each helix |
| pairGeometryReport.csv | Geometric information extracted from each pair, including sequence, helical stretches, and points of closest approach |
| pairReport.csv | Geometric information extracted from each pair |
| proteinReport.csv | Information extracted from each protein structure, including identifying phi and psi angles, potential for hydrogen bonding, and helicity of each position |
| rerun_conf.txt | The configuration file that can be used to rerun |
| segmentReport.csv | Information about identified helices with start and end positions and length of each helix |

**Table ST3.2: seqDesign.cpp**

| Output File | Description |
|---|---|
| #_#.pdb | Structure of the designed interface |
| bbRepack_#.out | Information about the repack for the structure |
| bbRepack_trajectory_#.pdb | Structural trajectory for the backbone repack |
| energyFile.csv | File to be analyzed; contains the energies, geometries, sequence, rotamers, interface, and SASA for each design |
| errors.out | Error output file |
| rerun.config | The configuration file that can be used to rerun |
| seqSearch_#.pdb | Pdb for each design pre-backbone repack |
| sequence_search_trajectory_#.pdb | Pdb for the sequences accepted during the search trajectory for each design interface |
| sequenceSearchEnergyLandscape_#.out | Energy landscape trajectory for each design |
| summary.out | Summary output file for the run, includes the elapsed time of the program through each step of design |
| x#_cross#_ax#_z#_vdW#.pdb | Pdb for the input design geometry, with the vdW energy output from the clashing check |

**Table ST3.3: bbRepack.cpp**

| Output File | Description |
|---|---|
| #.pdb | Repacked structure |
| #_repack.out | Output file for each structure repack, with the time elapsed and the before and after repack energies |
| summary.out | Summary file including the elapsed time for each step of the repack |
| energyFile.csv | File to be analyzed; contains the energies, geometries, sequence, rotamers, interface, and SASA for each design |
| errors.out | Error output file |
| initialPdb.pdb | Pdb for the input sequence set at the geometry of the input pdb |
| monomer.pdb | Monomer pdb structure |
| rerun.config | The configuration file that can be used to rerun |
| startPdb.pdb | Copied structure of the input pdb |

**Table ST3.4: getClashMutants.cpp (formerly calcMutantEnergy.cpp)**

| Output File | Description |
|---|---|
| LLLxxLLxxLLxxLLxxLILI.pdb | All sequences with Ile (or designated AA) at mutant positions are output as separate pdbs |
| energyFile.txt | Text file with the energy outputs for each sequence |

**Table ST3.5: getVoidMutants.cpp (formerly findPdbSASAVoids.cpp)**

| Output File | Description |
|---|---|
| LLLxxLLxxLLxxLLxxLILI.pdb | All sequences with Ala at mutant positions are output as separate pdbs |
| sasaMap.txt | Text file with SASA values for all of the mutated pdbs |

### 3.7.2 Python Scripts

**Table ST3.6: Design Analysis**

| Script | Description |
|---|---|
| main.py | Driver script that runs the other scripts by reading in a config file |
| compileEnergyFiles.py | Compiles the energy files from output from the sequence design script |
| analyzeDesignData.py | Main analysis script, outputting plots and csv files |
| createPymolSessionFiles.py | Makes pymol session files for the designed sequences with the most stable energies |
| createBackboneRepackFile.py | Creates a csv file for inputting into bbRepack script |

**Table ST3.7: NGS Reconstruction**

| Script | Description |
|---|---|
| main.py | Driver script that runs the other scripts by reading in a config file; contains helpful functions for the other scripts |
| fastqToTxt.pl | Converts the fastq NGS data to a txt file that can be analyzed |
| ngsAnalysis.py | Reconstructs the fluorescence from the converted NGS data |

**Table ST3.8: Convert to TOXGREEN Fluorescence**

| Script | Description |
|---|---|
| toxgreenConversion.py | Driver script that runs the other scripts by reading in a config file |
| adjustFluorByControlFlow_percentGpA_stdFix_fluor.py | Script that converts the fluorescence |
| filterWithComputation_percentGpA_stdFix.py | Outputs the design computational data with their sequences and filters the data standard deviation; removes sequences with fluorescence < 0 or where fluorescence – stddev < 0; outputs are used in pdbOptimizationAnalysis |

**Table ST3.9: Analyze the designs vs mutation data in boxplots**

| Script | Description |
|---|---|
| sequenceAnalysis.py | Driver script that runs the other scripts by reading in a config file |
| addNecessaryColumns.py | Using the mutant and sequence files from the clash filtered data in pdbOptimizationAnalysis, adds columns for analysis (WT and mutant AA, position of mutation, Type of sequence WT, Clash, Void) |
| plotBoxplotsPerAAPosition.py | Plots boxplots for differences between AA positions |
| plotBoxplotsCombined.py | Plots boxplots for differences between each design region |
| graphDeltaFluorescence.py | Graphs plots for the change in fluorescence between WT and Mutant sequence |

**Table ST3.10: Assess association by updated repack energy and mutations**

| Script | Description |
|---|---|
| pdbOptimizationAnalysis.py | Driver script that runs the other scripts by reading in a config file |
| stripSequenceEnds.py | Removes the first 3 letters and last 4 letters of all sequences (also removed in the later parts and reinserted later; to match up ala and leu designs) |
| keepMaltoseData.py | Filters the data for sequences that pass maltose test |
| compileFilesFromDirectories.py | Compiles the energyFile.csv from bbRepack to use in this analysis |
| addPercentGpaToDf.py | Appends the fluorescence and percentGpA data to the energy data |
| keepBestClashing.py | Filters data using the given clashing checks |
| combineFilesAndPlot.py | Combines the clash filtered files and the energy data from the maltose passing data, then plots using analyzeData.py |
| makeKdePlots.py | Outputs the kde plots for each dataset |
| convertToDeltaG.py | Converts the fluorescence data to deltaG |
| graphDeltaG.py | Graphs the deltaG data |
| analyzeData.py | Outputs plots of the energy terms against the fluorescence and %GpA |

## 3.8 References

Anderson, S. M. (2019). *Understanding the GASright Motif: Sequence, Structure, and Stability* (Publication Number 27548821) [Ph.D., The University of Wisconsin - Madison]. Dissertations & Theses @ Big Ten Academic Alliance; Dissertations & Theses @ University of Wisconsin at Madison; ProQuest Dissertations & Theses Global. United States -- Wisconsin. https://ezproxy.library.wisc.edu/login?url=https://www.proquest.com/dissertations-theses/understanding-gas-sub-right-motif-sequence/docview/2331244818/se-2?accountid=465https://resolver.library.wisconsin.edu/uwmad??url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations&sid=ProQ:ProQuest+Dissertations+%26+Theses+Global&atitle=&title=Understanding+the+GASright+Motif%3A+Sequence%2C+Structure%2C+and+Stability&issn=&date=2019-01-01&volume=&issue=&spage=&au=Anderson%2C+Samantha+Marie&isbn=9781392603215&jtitle=&btitle=&rft_id=info:eric/&rft_id=info:doi/

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

Arinaminpathy, Y., Khurana, E., Engelman, D. M., & Gerstein, M. B. (2009). Computational analysis of membrane proteins: the largest class of drug targets. *Drug Discovery Today*, *14*(23), 1130-1135. https://doi.org/https://doi.org/10.1016/j.drudis.2009.08.006

Ash, W. L., Stockner, T., MacCallum, J. L., & Tieleman, D. P. (2004). Computer modeling of polyleucine-based coiled coil dimers in a realistic membrane environment: insight into helix-helix interactions in membrane proteins. *Biochemistry*, *43*(28), 9050-9060. https://doi.org/10.1021/bi0494572

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H.,…Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, *28*(1), 235-242.

Bornberg-Bauer, E., Rivals, E., & Vingron, M. (1998). Computational approaches to identify leucine zippers. *Nucleic Acids Research*, *26*(11), 2740-2746. https://doi.org/10.1093/nar/26.11.2740

DeLano, W. L. (2002). Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr*, *40*(1), 82-92.

Desmet, J., Maeyer, M. D., Hazes, B., & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, *356*(6369), 539-542. https://doi.org/10.1038/356539a0

Elofsson, A., & von Heijne, G. (2007). Membrane protein structure: prediction versus reality. *Annu Rev Biochem*, *76*, 125-140. https://doi.org/10.1146/annurev.biochem.76.052705.163539

Ghirlanda, G. (2009). Design of membrane proteins: toward functional systems. *Current Opinion in Chemical Biology*, *13*(5), 643-651. https://doi.org/https://doi.org/10.1016/j.cbpa.2009.09.017

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nat Rev Mol Cell Biol*, *23*(1), 40-55. https://doi.org/10.1038/s41580-021-00407-0

Januliene, D., & Moeller, A. (2021). Single-Particle Cryo-EM of Membrane Proteins. *Methods Mol Biol*, *2302*, 153-178. https://doi.org/10.1007/978-1-0716-1394-8_9

Joh, N. H., Wang, T., Bhate, M. P., Acharya, R., Wu, Y., Grabe, M.,…DeGrado, W. F. (2014). De novo design of a transmembrane $Zn^{2+}$-transporting four-helix bundle. *Science*, *346*(6216), 1520-1524. https://doi.org/10.1126/science.1261172

Karplus, M., & Petsko, G. A. (1990). Molecular dynamics simulations in biology. *Nature*, *347*(6294), 631-639. https://doi.org/10.1038/347631a0

Kermani, A. A. (2021). A guide to membrane protein X-ray crystallography. *FEBS J*, *288*(20), 5788-5804. https://doi.org/10.1111/febs.15676

Koehl, P., & Delarue, M. (1994). Application of a Self-consistent Mean Field Theory to Predict Protein Side-chains Conformation and Estimate Their Conformational Entropy. *239*(2), 249-275.

Korendovych, I. V., Senes, A., Kim, Y. H., Lear, J. D., Fry, H. C., Therien, M. J.,…Degrado, W. F. (2010). De novo design and molecular assembly of a transmembrane diporphyrin-binding protein complex. *J Am Chem Soc*, *132*(44), 15516-15518. https://doi.org/10.1021/ja107487b

Kosuri, S., Goodman, D. B., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P.,…Church, G. M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc Natl Acad Sci U S A*, *110*(34), 14024-14029. https://doi.org/10.1073/pnas.1301301110

Krivov, G. G., Shapovalov, M. V., & Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, *77*(4), 778-795. https://doi.org/10.1002/prot.22488

Kulp, D. W., Subramaniam, S., Donald, J. E., Hannigan, B. T., Mueller, B. K., Grigoryan, G., & Senes, A. (2012). Structural informatics, modeling, and design with an open-source Molecular Software Library (MSL). *J Comput Chem*, *33*(20), 1645-1661. https://doi.org/10.1002/jcc.22968

Lazaridis, T. (2003). Effective energy function for proteins in lipid membranes. *Proteins*, *52*(2), 176-192. https://doi.org/10.1002/prot.10410

Li, L., Vorobyov, I., & Allen, T. W. (2013). The Different Interactions of Lysine and Arginine Side Chains with Lipid Membranes. *The Journal of Physical Chemistry B*, *117*(40), 11906-11920. https://doi.org/10.1021/jp405418y

Liang, B., & Tamm, L. K. (2016). NMR as a tool to investigate the structure, dynamics and function of membrane proteins. *Nat Struct Mol Biol*, *23*(6), 468-474. https://doi.org/10.1038/nsmb.3226

Lim, J. M., Kim, G., & Levine, R. L. (2019). Methionine in Proteins: It's Not Just for Protein Initiation Anymore. *Neurochem Res*, *44*(1), 247-257. https://doi.org/10.1007/s11064-017-2460-0

Liu, Y., Engelman, D. M., & Gerstein, M. (2002). Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biol*, *3*(10), research0054. https://doi.org/10.1186/gb-2002-3-10-research0054

Lomize, M. A., Lomize, A. L., Pogozheva, I. D., & Mosberg, H. I. (2006). OPM: orientations of proteins in membranes database. *Bioinformatics*, *22*(5), 623-625.

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J.,…Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, *102*(18), 3586-3616. https://doi.org/10.1021/jp973084f

McKinney, W. (2011). pandas: a Foundational Python Library for Data Analysis and Statistics.

McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.".

Mougeot, G., Dubos, T., Chausse, F., Péry, E., Graumann, K., Tatout, C.,…Desset, S. (2022). Deep learning -- promises for 3D nuclear imaging: a guide for biologists. *J Cell Sci*, *135*(7). https://doi.org/10.1242/jcs.258986

Mravic, M., He, L., Kratochvil, H. T., Hu, H., Nick, S. E., Bai, W.,…DeGrado, W. F. (2024). De novo-designed transmembrane proteins bind and regulate a cytokine receptor. *Nat Chem Biol*. https://doi.org/10.1038/s41589-024-01562-z

Mravic, M., Thomaston, J. L., Tucker, M., Solomon, P. E., Liu, L., & DeGrado, W. F. (2019). Packing of apolar side chains enables accurate design of highly stable membrane proteins. *Science*, *363*(6434), 1418-1423. https://doi.org/10.1126/science.aav7541

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

Na, D. (2020). User guides for biologists to learn computational methods. *J Microbiol*, *58*(3), 173-175. https://doi.org/10.1007/s12275-020-9723-1

Overington, J. P., Al-Lazikani, B., & Hopkins, A. L. (2006). How many drug targets are there? *Nature Reviews Drug Discovery*, *5*(12), 993-996. https://doi.org/10.1038/nrd2199

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,…Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

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

Shandler, S. J., Korendovych, I. V., Moore, D. T., Smith-Dupont, K. B., Streu, C. N., Litvinov, R. I.,…DeGrado, W. F. (2011). Computational design of a β-peptide that targets transmembrane helices. *J Am Chem Soc*, *133*(32), 12378-12381. https://doi.org/10.1021/ja204215f

SRINIVASAN, N., SOWDHAMINI, R., RAMAKRISHNAN, C., & BALARAM, P. (1990). Conformations of disulfide bridges in proteins. *International Journal of Peptide and Protein Research*, *36*(2), 147-155. https://doi.org/https://doi.org/10.1111/j.1399-3011.1990.tb00958.x

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, *35*(11), 1026-1028.

Tareen, A., & Kinney, J. B. (2019). Logomaker: beautiful sequence logos in Python. *Bioinformatics*, *36*(7), 2272-2274. https://doi.org/10.1093/bioinformatics/btz921

Tosi, S. (2009). *Matplotlib for Python developers*. Packt Publishing Ltd.

van Iterson, M., van Haagen, H. H., & Goeman, J. J. (2012). Resolving confusion of tongues in statistics and machine learning: a primer for biologists and bioinformaticians. *Proteomics*, *12*(4-5), 543-549. https://doi.org/10.1002/pmic.201100395

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D.,…Contributors, S. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*, *17*(3), 261-272. https://doi.org/10.1038/s41592-019-0686-2

Walshaw, J., & Woolfson, D. N. (2003). Extended knobs-into-holes packing in classical and complex coiled-coil assemblies. *144*(3), 349-361.

Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Yin, H., Slusky, J. S., Berger, B. W., Walters, R. S., Vilaire, G., Litvinov, R. I.,…DeGrado, W. F. (2007). Computational design of peptides that target transmembrane helices. *Science*, *315*(5820), 1817-1822. https://doi.org/10.1126/science.1136782

Zulkower, V., & Rosser, S. (2020). DNA Chisel, a versatile sequence optimizer. *Bioinformatics*, *36*(16), 4508-4509.

# Chapter 4: Future Directions

## 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 vdW packing has on facilitating MP folding and association. We studied a subset of vdW 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 ($GAS_{right}$). Additionally, our energetics correlated much better to the $GAS_{right}$ 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 evaluate my algorithm versus CATM predictions of designed $GAS_{right}$ sequences, and I further discuss ways to use my design algorithm to design heterodimers. I discuss ideas for potential improvements to my protein

design algorithm by converting sequence entropy into a pairwise term and improving our ability to design by using machine learning. Finally, I suggest improvements for our high-throughput sort-seq method, attempting to measure expression of sequences within cells to more accurately assess and compare dimerization propensity.

### 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 MPs (Fig. 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 vdW 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

| | Sequence | C p-value | HP p-value |
|---|---|---|---|
| D1 | LLLLAILLTLLAVLFSLLILI | $9.508e^{-5}$ | 0.008 |
| D2 | LLLVALLTILLALLFSLLILI | 0.009 | 0.041 |
| D3 | LLLLTLLVALLSTLLIFLILI | 0.369 | 0.002 |
| D4 | LLLTALLVALLFALLSLLILI | 0.0002 | 0.206 |
| D5 | LLLYALLTTLLSVLLFLLILI | 0.001 | 0.0002 |
| D6 | LLLYALLTVLLALLLSFLILI | 0.009 | 0.34 |
| D7 | LLLALLLVTLLAYLLSFLILI | 0.005 | 0.337 |
| D8 | LLLFYLLLTLLVALLSILILI | 0.06 | 0.004 |

b

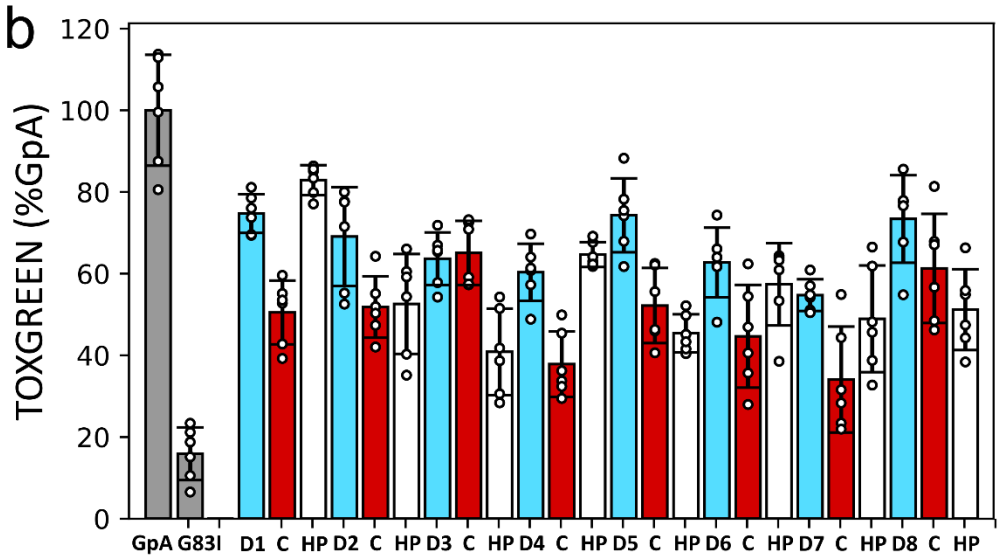



**Figure 4.1 TOXGREEN on hydrogen bonding AAs.** TOXGREEN fluorescence of 6 biological replicates for each design (D), clash (C) mutant, and hydrophobic (HP) mutant. **A)** Table of sequences for each design with the interface highlighted in bold. Clash mutant highlighted in red. Hydrophobic mutates 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 (white) mutants.

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 (Fig. 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, vdW packing drives the association of a variety of our designed MPs.

## 4.3 Studying the impact of sidechain packing with other forces

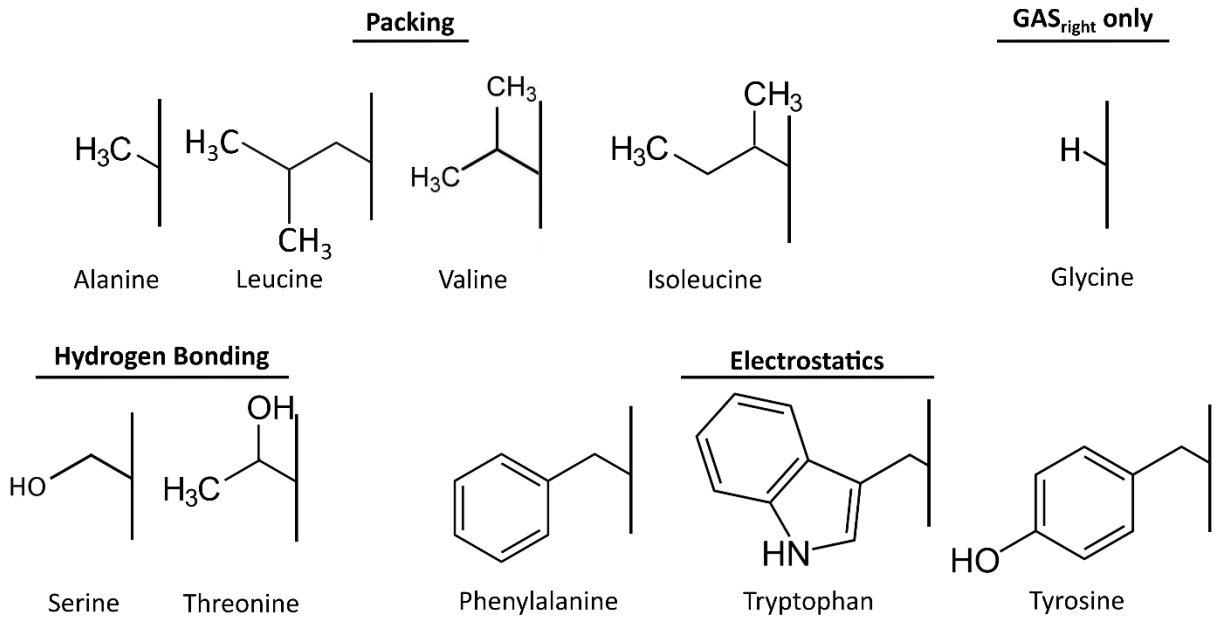



**Figure 4.2 Chemical Structures of Design AAs.** Design AAs separated by structures and potential energetic contributions to MP folding.

To investigate the impact of sidechain packing on MP folding and association, I designed thousands of proteins using only a subset of AAs (Fig. 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 (Fig. 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 to have minimal impact on folding through additional hydrogen bonding. Larger aromatic AAs (Trp, Phe, and Tyr) have an extensive vdW 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 reduce 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 vdW energy, but another 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 $GAS_{right}$. 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 either hydrogen bonding (Ser and Thr) or aromatic (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 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 $CH_2$, and changing Val→Ala results in a loss of 2 $CH_3$, 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 vdW 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 vdW packing interactions with the membrane environment. Overall, these experiments may allow us to correlate the changes in association with change in AA size, which could contribute to tuning our vdW energy term to differences in steric bulk. Additionally, it may be possible to gain insight into the entropic cost of exposing these sidechains to the membrane rather than buried within a protein interface. Experiments with designed interfaces along different backbones may give us insight into how to better predict the vdW packing between the membrane and the protein backbone.

## 4.4 Heterodimer design

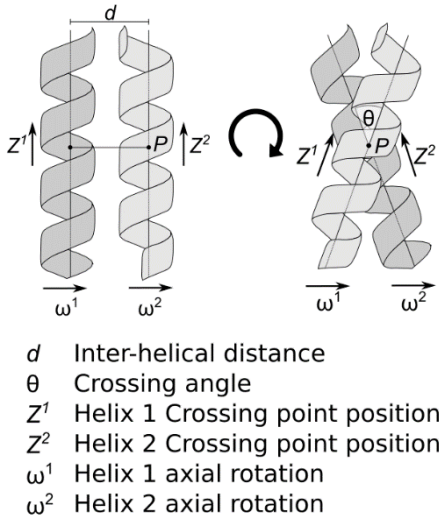



**Figure 4.3 Heterodimer Geometry.** Heterodimeric sequences do not need to have a symmetric geometry, where the z-shift and axial rotations of each helix can be different from its partner.

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 (Fig. 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 helix must be assessed against all rotations on the other helix, and the same must be applied for z-shifts.

### 4.4.1 Predicting designed $GAS_{right}$ sequences with CATM

a

| TOXCAT range (% GpA) | 25–50% | 50–75% | 75–100% | 100–125% | 125+% |
|---|---|---|---|---|---|
| number of constructs | 7 | 7 | 4 | 4 | 3 |
| average TOXCAT (% GpA) | 37 ± 1 | 58 ± 2 | 88 ± 4 | 118 ± 2 | 141 ± 7 |
| CATM energy score (kcal/mol)[b] | −14.7 ± 2.5 | −27.1 ± 2.5 | −30.0 ± 2.0 | −39.1 ± 2.8 | −41.7 ± 7.5 |
| van der Waals (kcal/mol) | −26.2 ± 5.3 | −33.7 ± 4.5 | −33.6 ± 2.1 | −39.3 ± 2.4 | −39.0 ± 11.1 |
| Cα-H hydrogen bonding (kcal/mol) | −5.2 ± 1.1 | −8.0 ± 1.9 | −9.7 ± 0.5 | −12.0 ± 2.3 | −13.0 ± 0.8 |
| solvation (kcal/mol) | 16.7 ± 1.9 | 14.2 ± 1.9 | 13.3 ± 2.0 | 11.7 ± 2.4 | 10.6 ± 2.7 |
| crossing angle (deg) | −51 ± 4 | −47 ± 6 | −49 ± 2 | −41 ± 7 | −39 ± 7 |
| number of Cα-H hydrogen bonds | 4.6 ± 1.0 | 5.1 ± 1.1 | 6.0 ± 0.0 | 7.5 ± 1.0 | 8.0 ± 0.0 |
| interface surface area (Å$^2$) | 4810 ± 490 | 4660 ± 500 | 4630 ± 190 | 4770 ± 540 | 4510 ± 280 |
| interhelical distance (Å) | 7.1 ± 0.2 | 6.7 ± 0.3 | 6.5 ± 0.1 | 6.4 ± 0.1 | 6.5 ± 0.0 |

from Anderson et al. 2017

b

| TOXGREEN range (%GpA) | 25-50% | 50-75% | 75-100% | 100+% |
|---|---|---|---|---|
| number of constructs | 50 | 9 | 20 | 17 |
| average TOXGREEN (%GpA) | 33 ± 7 | 60 ± 8 | 87 ± 6 | 113 ± 8 |
| ΔG | -2.01 ± 0.65 | -3.06 ± 0.18 | -3.51 ± 0.09 | -3.83 ± 0.09 |
| TOXGREEN Fluorescence | 8045.3 ± 1585.81 | 14572.8 ± 1936.22 | 20936.35 ± 1564.81 | 27315.87 ± 2012.4 |
| CATM Energy score (kcal/mol) | -29.14 ± 7.78 | -38.79 ± 10.67 | -46.81 ± 9.78 | -49.36 ± 9.73 |
| CHARMM_VDW (kcal/mol) | -37.92 ± 5.29 | -41.36 ± 6.09 | -47.45 ± 5.47 | -47.73 ± 7.56 |
| SCWRL4_HBOND (kcal/mol) | -4.93 ± 4.28 | -8.41 ± 4.33 | -10.0 ± 4.39 | -12.22 ± 3.32 |
| CHARMM_IMM1 (kcal/mol) | -10.75 ± 3.72 | -9.44 ± 3.87 | -7.73 ± 3.68 | -8.1 ± 3.69 |
| CHARMM_IMM1REF (kcal/mol) | 24.45 ± 5.86 | 20.41 ± 6.37 | 18.36 ± 5.08 | 18.7 ± 4.98 |
| crossing angle (°) | -46.52 ± 6.24 | -42.49 ± 6.96 | -38.87 ± 6.95 | -39.01 ± 6.43 |
| interhelical distance (Å) | 6.96 ± 0.29 | 6.78 ± 0.38 | 6.6 ± 0.29 | 6.51 ± 0.19 |
| z-shift (Å) | 3.31 ± 1.63 | 4.08 ± 1.72 | 4.47 ± 1.81 | 4.28 ± 1.88 |
| axial rotation (°) | -62.64 ± 12.76 | -73.72 ± 13.31 | -74.16 ± 14.46 | -74.22 ± 14.34 |
| numModels | 42 ± 29 | 85 ± 42 | 116 ± 40 | 134 ± 32 |

**Figure 4.4 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 $GAS_{right}$ designs. Similar results are found, with CATM predicting a better energy score, vdW, and hydrogen bonding on average for more stable designs, as well as a narrower crossing angle and interhelical distance.

As an initial approach, we can design heterodimeric $GAS_{right}$ 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 $GAS_{right}$ 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 $GAS_{right}$ sequences and created a table similar to as found in Anderson et al. 2017 (Fig. 4.4). 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 starting points that can be applied in heterodimer design.

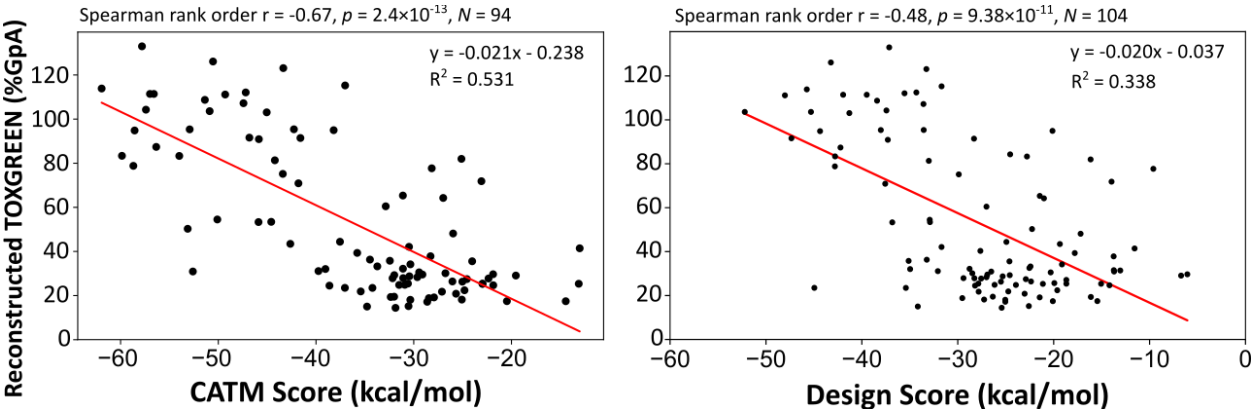



**Figure 4.5 CATM and Design score comparison.** Energy scores output from CATM and the design algorithm are plotted against dimerization propensity (standard deviation not shown) for designed $GAS_{right}$. CATM scores have improved linear and spearman correlations than design scores.

I further analyzed CATM against my design algorithm, aiming to see if correlation improves using CATM energetics for $GAS_{right}$ proteins. Since my design algorithm uses the same energetics as CATM, I wanted to see if there are large differences between the energetics for these my designed sequences. I found that there is a better correlation between CATM energies and dimerization propensity in both $R^2$ linear and Spearman rank order correlations (Fig. 4.5). CATM more extensively searches $GAS_{right}$ geometric space than my design algorithm, effectively exploring multiple potential interfaces for a single sequence (Mueller et al., 2014). Because my design algorithm only places interfaces along the center of a dimer structure, it is likely that CATM predicts proteins with different interfaces and structures. I further probed this possibility by assessing CATM structures against designs. I calculated the Cα RMSD between each sequence's CATM and design structure and separated them into three groups: RMSD<1, 1<RMSD<2, RMSD>2. I then plotted the energy score of CATM and the design against the dimerization propensity for each group (Fig. S4.1). Interestingly, proteins with RMSD<1 did not have a strong linear correlation between energy score and dimerization propensity. A majority of these proteins are found in the weak dimerization range (0.2-0.4), supporting that our energy terms are unable to accurately predict weak dimers. As a final assessment, I plotted the CATM score against the design score for each separate RMSD group (Fig. S4.2). The CATM

energy scores correlate well with the design scores, suggesting that the energetics between these two programs is consistent despite differences in their ability to predict dimerization propensity.

### 4.4.2 Heterodimer design strategy

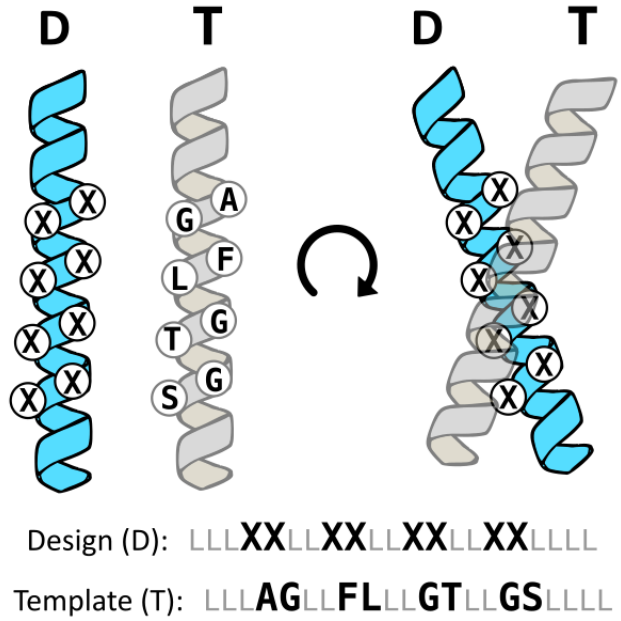



**Figure 4.6 Designing heterodimers against a template sequence.** To simplify heterodimer design, we can take known non-homodimerizing sequences (templates) from previous research and design a sequence to associate with it (design).

To simplify heterodimer design, we can design sequences against a single template sequence (Fig. 4.6). I expand on how I would implement this into my design algorithm in Supplementary Details 4.3 to 4.5. 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 removed from our pool of sequences that we plan for experiments.

### 4.4.3 Heterodimer experimental strategy

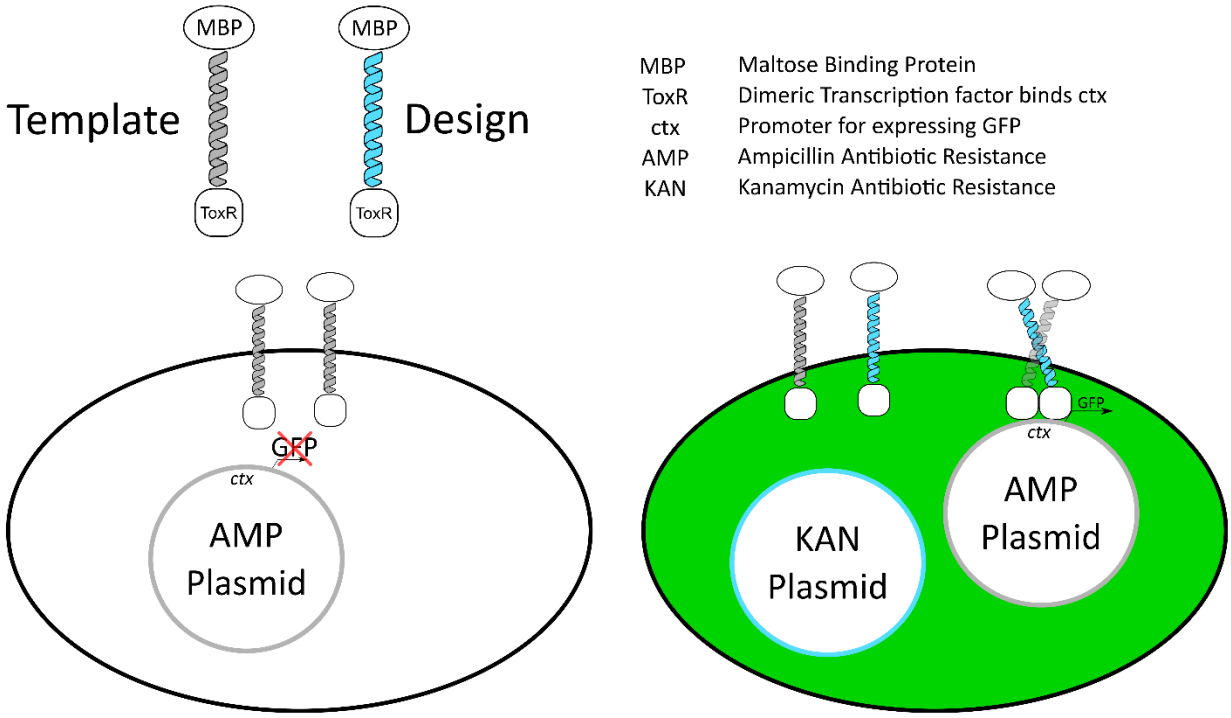



**Figure 4.7 Heterodimer experiments.** By expressing two plasmids that code for different antibiotic resistance (AMP and KAN) and each expressing a different TM sequence (Template and Design), we can investigate the dimerization propensity of heterodimers.

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 (ampicillin/AMP), then clone a library of plasmids containing our designed sequences into that same strain. To ensure that both of our plasmids are in the bacteria, the library can be cloned into a plasmid expressing another antibiotic (kanamycin/KAN). Any cells that grow in media containing both AMP and KAN should express both the template and design sequences. This will result in the expression of GFP from cells containing any heterodimer pairs (Fig. 4.7). Because we plan to design $GAS_{right}$ sequences which show better correlation to our energetics (Fig. 2.3), we expect that exploring a range of energies of our designed heterodimers will yield a range of heterodimerization.

### 4.4.4 Simplifying sequence and geometric space for heterodimers

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 $GAS_{right}$, 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 vdW 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 vdW 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 $100^2$ 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 $GAS_{right}$ proteins (section 4.4.1). Combining the $GAS_{right}$ designs in my study with other research in the lab, we have a database of hundreds of $GAS_{right}$ homodimers. Using combinatorial testing

of the most common axial rotations and z-shifts (Fig. 4.4), we may be able to identify regions that are energetically favorable for heterodimer design.

### 4.5 Turning sequence entropy into a pairwise term

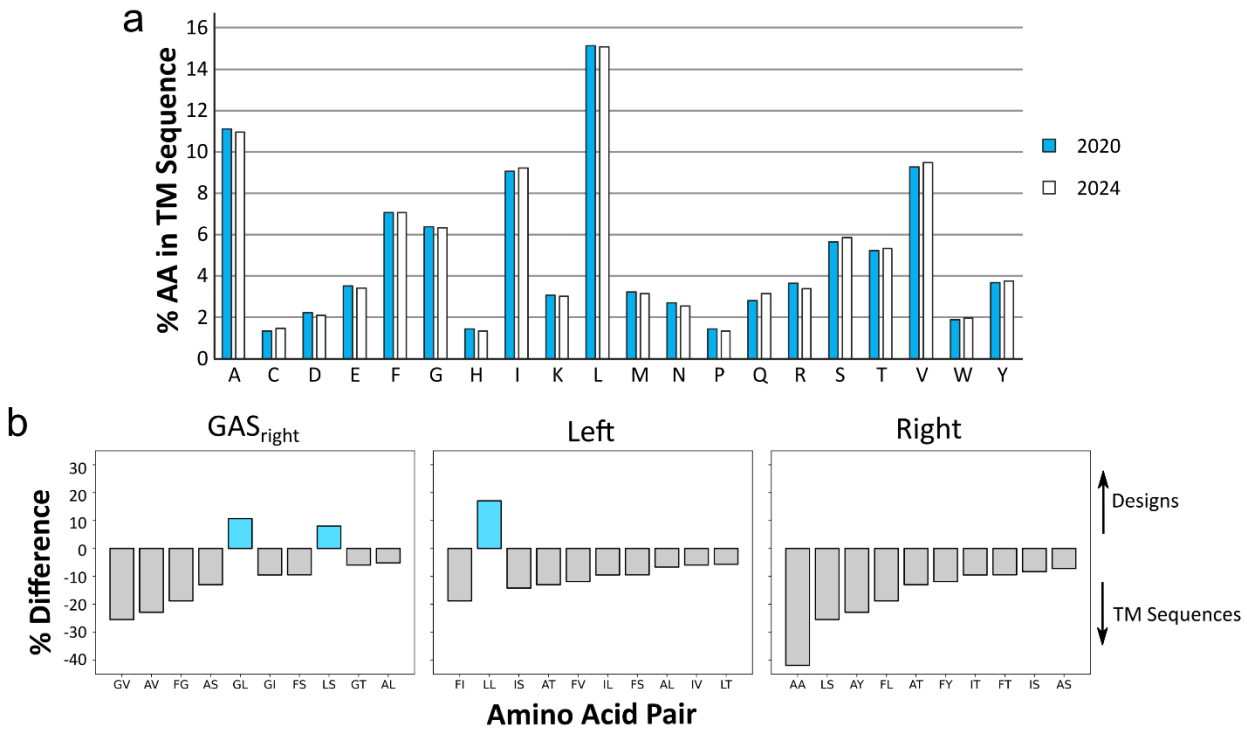



**Figure 4.8 Amino acid frequency from TMs extracted from OPM.** A) Each AA was counted as found in TMs identified from OPM and divided by the total count of all AAs to determine the frequency. Data separated by year (2021, blue; 2024, white). B) Frequency of AA pairs found in at least 10 designed sequences compared with the frequency of same AA pairs in all TMs. The largest differences in AA pair frequencies are shown here, with negative differences (gray) corresponding to AA pairs more often expressed in TMs, and positive difference are pairs more often found in designed sequences (light 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 2020 (Fig. 4.8A). 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 (Fig. 4.8B). Although we successfully designed interfacial sequences that reflect the frequency of AAs in TMs (Fig. S3.6), there are multiple pairs of AAs that deviate between natural TMs and our designs. For example, Fig. 4.6B shows one such bias where the Ala-Ala pair is expressed >40% more frequently in TMs than in our Right design sequences. Additionally, a few pairs were found to occur more frequently in TMs in more than one design region (Ala-Ser, Ala-Leu, Ile-Ser, Ala-Thr). This data suggests that our design algorithm cannot replicate some pair interactions found within MP sequences. In a future design run that considers pair frequencies of AAs, we can work to remove these biases to preserve pair interactions found in natural TM sequences. 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 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.

### 4.6 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 $GAS_{right}$. 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 MP 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 similar to what we aim to do with heterodimer design except using machine learning. Although not specifically built to design MPs, 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 MP design.

### 4.7 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 $GAS_{right}$ (Fig. 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 MPs 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.

## 4.8 Supplementary Details

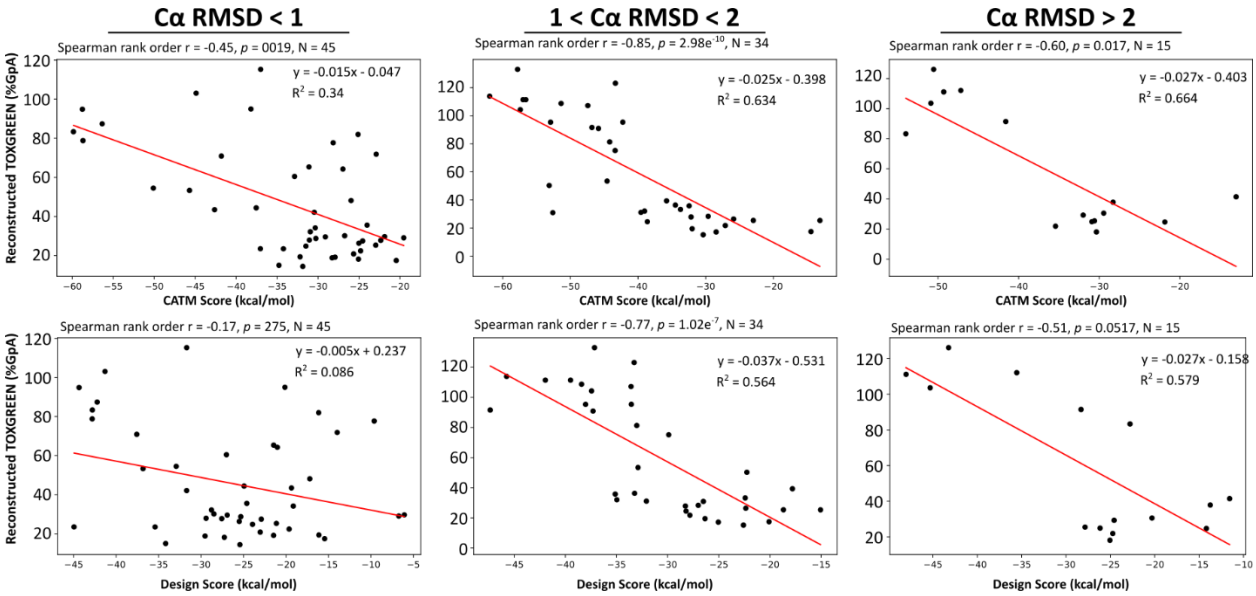



**Figure S4.1 Dimerization propensity vs CATM/Design score.** Cα RMSD was calculated between designed $GAS_{right}$ structures and CATM predicted structures of each sequence. Energy scores were plotted against the dimerization propensity (standard deviation not shown) for each group.

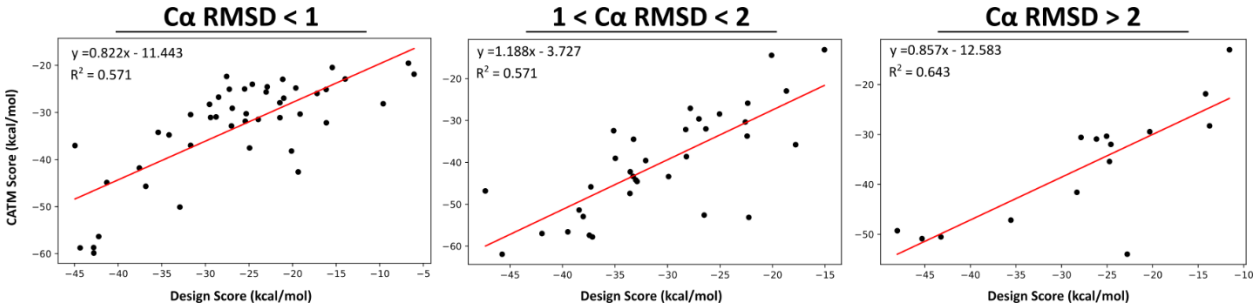



**Figure S4.2 CATM vs Design score.** CATM energy scores from predicted designed $GAS_{right}$ sequences plotted against design energy scores.

### S4.3 Separate helices: one as template and one to designed

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

```
561     // Defines the interfacial positions and the number of rotamers to give each position
562     PolymerSequence interfacePolySeq = getInterfacialPolymerSequence(opt, startGeom, allInterfacePositions, interfacePositions,
563      rotamerSamplingPerPosition);
```

Within this function are two functions of importance.

The function below creates a backbone sequence (backboneSeq) 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 the typical LLL and ILI termini, rather than AAA at the termini).

```
1331    // if sequence is not empty, use polyLeu
1332    string backboneSeq;
1333    if (_opt.sequence == ""){
1334        backboneSeq = generateBackboneSequence(_opt.backboneAA, _opt.backboneLength, _opt.useAlaAtTermini);
1335    } else {
1336        backboneSeq = _opt.sequence;
1337    }
```

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.

```
1388    // makes the polymer sequence to return basedon the interface positions
1389    string polySeq = generateMultiIDPolymerSequence(backboneSeq, _opt.thread, _opt.Ids, interfacePositions);
1390    PolymerSequence PS(polySeq);
1391    return PS;
```

```
2720        ps = ":{" + MslTools::intToString(_startResNum) + "} " + ps;
2721        return "A" + ps + "\nB" + ps;
2722    }
```

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;

```
565     /*****************************************************************************
566      *  === declare the system with alternate identities at the interface ===
567      *****************************************************************************/
568     // set up the system for the input sequence (I never tried with other sequences, but I assume it would work with more than just polyLeu)
569     System sys;
570     prepareSystem(opt, sys, startGeom, interfacePolySeq, opt.useBaseline);
571     checkForClashing(sys, opt, interfacePositions, sout); // checks an interface for clashes; if too clashing (>10 VDW) with alanine at those positions, exit
```

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.

```
2574    // function for writing a pdb
2575    void writePdb(System &_sys, string _outputDir, string _pdbName){
2576        PDBWriter writer;
2577        writer.open(_outputDir + "/" + _pdbName + ".pdb");
2578        writer.write(_sys.getAtomPointers(), true, false, false);
2579        writer.close();
2580    }
```

### S4.4 Creating a non-symmetric version of the switchSequence function

```
764   // switches to the starting sequence (if given, otherwise set to polyleu)
765   void switchSequence(System &_sys, Options &_opt, string _sequence){
766       if (_sequence == ""){
767           _sequence = generateBackboneSequence(_opt.backboneAA, _opt.backboneLength, _opt.useAlaAtTermini);
768       }
769       setActiveSequence(_sys, _sequence);
770   }
```

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

```
2553   // set the active identity for each position to the identity in the given sequence
2554   void setActiveSequence(System &_sys, string _sequence){
2555       // loop through the sequence
2556       for (uint i=0; i<_sequence.size(); i++){
2557           // get the ith residue in the sequence
2558           string res = _sequence.substr(i,1);
2559           // loop through all of the chains in the system
2560           for (uint j=0; j<_sys.chainSize(); j++){
2561               Chain &chain = _sys.getChain(j);
2562               // get the ith position in the system
2563               Position &pos = chain.getPosition(i);
2564               // get the position id for the ith position
2565               string posId = pos.getPositionId();
2566               // convert the residue id to three letter code
2567               string aa = MslTools::getThreeLetterCode(res);
2568               // set active identity
2569               _sys.setActiveIdentity(posId, aa);
2570           }
2571       }
2572   }
```

It 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 the new version. This should result in only your design sequence (chain A; S4.5) being changed throughout the design process.

### S4.5 Outputting the designed heterodimer sequence

```
610     // get the starting sequence
611     Chain &chainA = sys.getChain("A");
612     string seq = convertPolymerSeqToOneLetterSeq(chainA);
613     cout << "Sequence before stateMC: " << seq << endl;
```

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.

## S4.6 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.

```
1057        while (!MC.getComplete()){
1058            // get the sequence entropy probability for the current best sequence
1059            map<string,vector<uint>> sequenceVectorMap;
1060            map<string,map<string,double>> sequenceEnergyMap = mutateRandomPosition(_sys, _opt, _spm, _RNG, bestSeq, bestEnergy,
1061             sequenceVectorMap, _sequenceEntropyMap, _allInterfacialPositionsList, _interfacialPositionsList, _rotamerSampling);
```

Within this function is a function that calculates the energy terms for each iteration of the designed sequence.

```
1207                threads.push_back(thread{energyFunction, ref(_opt), ref(_spm), _bestSeq, _bestEnergy, currSeq, currVec, ref(_rotamerSampling), ref(_interfacialPositionsList),
1208                 ref(sequenceEnergyMap), ref(_sequenceEntropyMap)});
1209            }
1210            setActiveSequence(_sys, _bestSeq);
```

```
1219    // threaded function that gets energies for a sequence and appends to the sequence energy map
1220    void energyFunction(Options &_opt, SelfPairManager &_spm, string _prevSeq, double _prevEnergy, string _currSeq, vector<uint> _currVec,
1221     vector<uint> &_rotamerSampling, vector<uint> &_interfacePositions, map<string,map<string,double>> &_seqEnergyMap,
1222     map<string,double> &_sequenceEntropyMap){
1223        // variable setup
1224        map<string,double> energyMap;
1225
1226        // Compute dimer energy
1227        outputEnergiesByTerm(_spm, _currVec, energyMap, _opt.energyTermList, "Dimer", true);
1228        double currEnergy = _spm.getStateEnergy(_currVec);
1229
1230        // initialize variables for this thread
1231        double sequenceProbability = 1;
1232        double sequenceEntropy = calculateInterfaceSequenceEntropy(_currSeq, _sequenceEntropyMap, _interfacePositions, sequenceProbability, _opt.weight_seqEntropy); // calculate the sequence entropy for the interface
1233
1234        // output info
1235        double baseline = _spm.getStateEnergy(_currVec, "BASELINE")+_spm.getStateEnergy(_currVec, "BASELINE_PAIR");
1236        double energyAndEntropyTotal = currEnergy+sequenceEntropy;
1237        energyMap["Dimerw/Baseline"] = currEnergy;
1238        energyMap["Dimer"] = currEnergy-baseline;
1239        energyMap["Baseline"] = baseline;
1240        energyMap["SequenceProbability"] = sequenceProbability;
1241        energyMap["SequenceEntropy"] = sequenceEntropy;
1242        energyMap["Totalw/Entropy"] = energyAndEntropyTotal;
1243        _seqEnergyMap[_currSeq] = energyMap;
1244    }
```

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

```
668   double calculateInterfaceSequenceEntropy(string _sequence, map<string, double> _sequenceEntropyMap, vector<uint> _interfacePositions,
669    double &_sequenceProbability, double _seqEntropyWeight){
670       //Get residue name for each interfacial identity
671       vector<string> seqVector;
672       int numInterfacials = _interfacePositions.size();
673       for (uint i=0; i<numInterfacials; i++){
674           stringstream tmp;
675           tmp << _sequence[_interfacePositions[i]];
676           string aa = tmp.str();
677           string resName = MslTools::getThreeLetterCode(aa);
678           seqVector.push_back(resName);
679           //cout << _interfacialPositionsList[i] << " : " << resName << endl;
680       }
681       map<string,int> seqCountMap = getAACountMap(seqVector); // get the count of each amino acid in the sequence
682       double numberOfPermutations = calcNumberOfPermutations(seqCountMap, numInterfacials); // calculate the number of permutations for the sequence
683       _sequenceProbability = calculateSequenceProbability(seqCountMap, _sequenceEntropyMap, numberOfPermutations); // calculate the sequence probability
684       double seqEntropy = -log(_sequenceProbability)*0.592*_seqEntropyWeight; // multiply by the weight and -1
685       //double seqEntropy = _seqEntropyWeight/_sequenceProbability; // divide the weight by the probability to get the entropy
686       return seqEntropy;
687   }
```

## 4.9 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

# Chapter 5: Chapter for the Public

## Why I wrote this chapter

Early on in my PhD career, my advisor inspired me to train my brain to learn at the highest level with a simple credo: know what you don't know.

To think about what you know so deliberately that you're able to figure out what there is left to learn.

And it's worked: I've finished my research and am working to publish my work as an infinitesimally small stamp in history. The PhD finish line is full of triumphs: a published paper, a final defense to showcase and explain your research, and recognition as an academic expert.

But PhDs typically aren't straight forward. For me it's been a long, intense personal journey: traversing the valleys and mountains of knowledge about cell membranes and biophysical forces. Sometimes, trekking and running energetically through a field of fluorescent green flowers. But more often, barely learning anything and feeling completely stuck, as if trudging through multiple feet of snow in the dark.

So when I saw the opportunity to include a chapter in my thesis about "the parts of the story of science that don't get told in scientific publication", I felt moved to write.

**Sonder**: the realization that everyone has a life as real and full as your own. It's one of my favorite words, expressing how connected we are as humans, going through our own emotions and personal turmoil, figuring out our lives as we go. Research is typically presented without mentioning the rigorous mental fortitude, the exhaustive emotional toll, the strains on life that it takes to succeed. But with this opportunity, I wanted to flip that narrative and share some of the mental and emotional swings of my PhD journey.

10 years from now, I'm not sure how I'll feel about graduate school. 7 whole years. A project that brought me deeper into the niches of science than I ever thought I could go. What is van der Waals packing anyways? This miniscule attractive force that relies on the periphery of atoms in space. I've spent years investigating subatomic interactions within both theoretical and physical experimentation and making sense of the results. And here, finally at the end, realizing that my thesis is a translation of my findings that this superficially nanoscopic (it's actually smaller!) force has on membrane protein folding and association.

But discovery and novelty are extremely difficult to quantify, and even more so to describe at a level that makes sense to everyone. So I've melded two approaches that I felt comfortable with: The creation of a playlist that embodies the experiences I've had paired with personable letter writing. All aiming to answer the question:

*What did it take for me to become a PhD?*

Thank you to SciFun, Wisconsin Initiative for Science Literacy (WISL), and WISL staff for allowing me this opportunity to share transparent reflections on my PhD. Thank you to Professor Bassam Shakhashiri, Cayce Osborne, and Elizabeth Reynolds, for helping me to develop and analogize the bits of science included in here!

And additional thanks to my dear friends Diego Lanao and Tram-Anh Nguyen for critiquing and giving feedback on my drafts. I wouldn't have been able to write something even semi-coherent without their help.

Thanks for reading, and best of luck on whatever journey you're on. Sending love and good vibes your way :D.

## Glossary

### Science

- **Protein** – molecules necessary for many important biological functions: supporting cells, building immunity, sensing changes in environment
- **Membrane protein** – proteins found in the cell membrane (the biological membrane that separates the inside of the cell from the outside environment); important for helping cells adapt and react to change
- **Van der Waals packing** – "static" like attraction between proteins in close contact
- **Algorithm** – sequence of computational instructions I made to build models of proteins with different amounts of "static"
- **Computational model** – predicted, visual representations of what a protein looks like
- **Associate/Association** – when two proteins stick together, like partners coming together in a choreographed dance
- **GFP** – green fluorescent protein, can be used like a marker to measure biological experiments
- **Cell sorter** – machine that can measure the amount of GFP produced from individual cells

### Other

- **Frisson** – aesthetic chills, psychogenic shivers; commonly tingling of the skin when listening to music
- **Leitmotif** – short, recurring musical theme accompanying a person, place, or idea
- **Gjetost** – Scandinavian cheese that tastes like caramel
- **Tsundoku** – the art of buying books and never reading them
- **Preliminary exam** – candidacy exam to verify that you have the potential to succeed in grad school
- **Suicidal ideation** – having ideas about committing suicide
- **Interrobang** – the combination of a question mark and exclamation point (‽ or ?!), for questions asked with a feeling of surprise or disbelief
- **Petrichor** – the pleasant smell of fresh rain
- **Mental spiral** – a gradually escalating, overwhelming cycle of negative thoughts
- **Imposter Syndrome** – internalized feeling of doubt in one's skill, talent, or intelligence; feeling like you don't deserve success and that much of it is attributed to things out of your control
- **Burnout** – state of emotional, mental, and physical exhaustion brought on by prolonged/repeated stress
- **Halcyon** – calm, peaceful
- **Geosmin** – the molecule responsible for petrichor
- **Phosphenes** – the light-like swirls, colors, shapes, etc. that you see when you close your eyes
- **Kaleidoscope** – toy made of internal mirrors that reflect in a seemingly infinite variety of patterns

**From *The Dictionary of Obscure Sorrows* by John Koenig**

- **Sonder** – the realization that everyone has a life as real and full as your own
- **Trumspringa** – the longing to wander off your career track in pursuit of a simple life
- **Etterath** – the feeling of emptiness after a long and arduous process is complete

**And some words that don't make it in but are fun anyways because I have the space:**

- **Zarf** – a coffee sleeve
- **Aglet** – the plastic end of a shoelace
- **Occlupanid** – the little plastic tags that are typically used to close bread bags
- **Griffonage** – illegible handwriting
- **Wamble** – feeling nauseous or queasy
- **Crapulence** – the sick feeling you get after eating or drinking too much
- **Vocables** – syllables with no referential meaning, like 'na na na' and 'la la la'

**Letters for my PhD**



**letters for my phd**

**A letter always seemed to me like immortality because it is the mind alone without corporeal friend.**

— *Emily Dickinson*

**I do**

*Flipside-postlude by Kid Quill*

Dear Reader, April 2017

*What do you want to be when you grow up?*

I've answered variations of this question over the years: What are you majoring in? What's next after college?

Do you want to go to medical school?

I'm a first-generation Haitian-Filipino American, and no one in my family is a scientist. But going into medicine and becoming a doctor was preached as the ideal life by the adults in the immigrant community I grew up in.

And I've always loved science. An elementary school field trip to the botanical gardens sticks with me: they gave each of us magnifying glasses and I was the kid getting left behind, needing to be reminded to keep up with the group. Getting lost in the observation of flower petals, mesmerized by this new perspective on nature.

When I got to college, I majored in biology to assuage my curiosity, to learn more about how life works.

I studied how nature puzzle pieces molecules together; I've seen the beauty in how cellular systems work. Even got the chance to do some independent research, learning a bunch about mice hormones and neurons.

But I ended up taking most of the classes necessary for med school. I knew that my path could result in me becoming a doctor. Internally, it felt like I was afraid to turn away from the expectations that adults had for me.

So when my biology professor told me that a PhD might be a good fit for me, I got emotional. Befuddled. Elated.

I never thought that I could complete an advanced degree. I was rejected from several labs during college, and I didn't even know that PhD programs existed until my junior year. But I think he saw my passion for learning: this deep-seeded interest in diving into subjects and a willingness to bang my head against a wall full of knowledge.

I applied, and I've recently been accepted to a biochemistry program at the University of Wisconsin-Madison!

I'm not particularly gifted, I end up being average at everything I do, and although I wanted to help people from a young age, I couldn't see myself going to med school at this stage in my life. But I'm fascinated by science. This route to becoming a PhD "doctor" feels semi-validating and soothes those immigrant expectations within me.

College didn't exactly prepare me for adulthood, but I'm excited to use the skills I've learned on this PhD journey! To begin this path towards becoming a professor who instills confidence in students just as mine have for me!

Gilbert Jamilla Loiseau

*P.S.* Did you grow up with expectations that you felt defined by?

*P.P.S.* **Frisson**: The aesthetic chills from the layering of instruments, the tingling up your spine from hypnotic harmony, engulfing you in tantalizing bliss. When I close my eyes and listen to music, I feel sound pouring into my being, my body literally resonating. Some of these letters might be convoluted, but I hope you'll understand my feelings by listening to the music: a **leitmotif** for each of these letters!

*P.P.P.S.* This song is about wanting to be heard. About having bold ideas and passions, about wanting to share. I'm hoping that a PhD will allow me to cultivate ideas, determine my strengths in learning, and allow me to confidently share something that I'm passionate about with the world!

**It's so hard to swim against the tide**

*Swim Against the Tide by The Japanese House*

Dear Reader, October 2017

Graduate school classes have a conversation-like feel. Instead of lectures where professors talk about a subject for an hour, classes focus on maximizing time to discuss what we students are most curious about.

Recently, one professor posed a personal question: What are your goals with your graduate education?

"Gilbert, what about you?"

"I'm not sure." I've never been good with public speaking. I never raise my hand in class, and being called on is *extremely* uncomfortable.


"That's okay, just say anything that feels right."

With that little bit of encouragement, I used my avoidant gaze, stared at the ceiling, and thought. *What could I do?*

"I want to find a way to replace PowerPoint."

"Okay! Why is that?"

"I feel like there are weaknesses in how it communicates concepts and that something better could be made."

In biochemistry, there's a technique called polymerase chain reaction, or PCR. Using our knowledge of how DNA replicates in cells, we're able to effectively replicate DNA with PCR!

But how did we come up with PCR, this fundamental tool used for forensic screening and diagnosing diseases?

Kary Mullis, the inventor of PCR said it best: "I was looking for something else…PCR was the possible outcome of a solution to a hypothetical problem that didn't really exist."

The best science comes from harnessing creativity, trying to see things that haven't been imagined, and discovering questions that haven't yet been answered.

*How does van der Waals packing impact membrane protein association?*

Research doesn't look for a specific answer. It focuses on understanding why things are as they are. My mind is racing, questioning what I know about my project, striving to delve deeper into membrane protein research.

What sticks with me is that this professor gave me the opportunity to share an idea of an idea.

Asking "why" helped me feel comfortable and supported my growth. And although it currently feels impossible, this journey into the unknown reaches of science is beginning to feel a little more comfortable.

GJL

*P.S.* How often do you go outside of your comfort zone?

*P.P.S.* Grad school is showing me that personal growth is enhanced when outside of your comfort zone. Whether it's getting called out in a class or moving to a state where the temperatures reach an unimaginable -40 degrees (the same in both Fahrenheit and Celsius!), there are a variety of ways to get outside of your comfort zone to grow! Why does **Gjetost** exist, why does it taste like caramel, and how the heck does it taste so delicious? By harnessing creativity and learning when to ask why, I'm hoping to find my stride on this journey for discovering knowledge.

**The world sayin' what you are because you're young and black**

*Outside by Childish Gambino*

Dear Reader, September 2018

I was chatting with someone recently about TV shows. They were surprised that I had never watched *How I Met Your Mother* and *Arrested Development*.

"Why not?"

"Honestly, there are too many white people in them."

A couple of days later, they recounted the conversation: "I looked up the demographics in those shows, and they don't skew too far from the US population." They didn't understand that those shows aren't likely to portray experiences I can relate to.

As a person of mixed descent, I've found it difficult to figure out how I belong. How do I fit in while being myself? What's the correct answer for surveys asking for my ethnicity? I love sharing culture, ideas, and listening to a wide variety of perspectives. But why don't I know who I am?

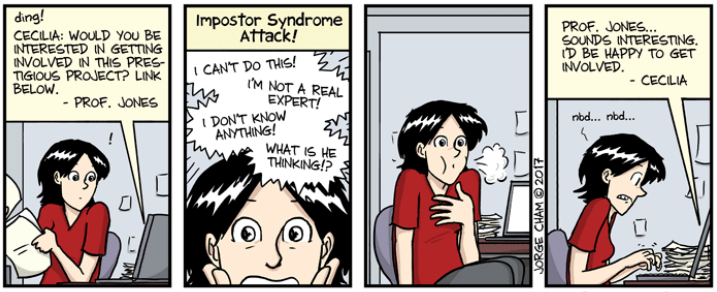
Why do I feel like an outsider?

In science I've often felt the same. And now after majoring in biology, my transition into a biochemistry PhD has been more arduous than expected. I'm finding it difficult to communicate how I understand science with the appropriate wording and depth. As if the words I'm saying don't mean what I think they mean.

*What knowledge am I missing? Do I have the ability to navigate the field of biochemistry? Do I even belong here?*

**Imposter syndrome**: a persistent, unjustified feeling that one's success is fraudulent.

I know how the world perceives me by the color of my skin, so I can quickly spiral into a mix of negative thoughts. *I'm a token minority for my program to look good, people don't understand me because I'm not smart enough to be understood, I wouldn't have been accepted if I wasn't a minority, I don't deserve to be here.*




In both my personal and professional life, I feel like an outsider. Isolated from people who are like me.

*Does that feeling ever go away?*

GJL

*P.S.* When was the last time you felt like you didn't belong?

*P.P.S.* I thought coming here would allow me to feel more comfortable because I'd find people like me: People who want to understand more about science, to better understand different cultures and perspectives, to learn by challenging the norm and gaining insights into the unique experiences of others. So why does it often feel like I have to conform to what other people expect of me rather than being given the chance to share my own identity?

**There will be mountains you won't move**

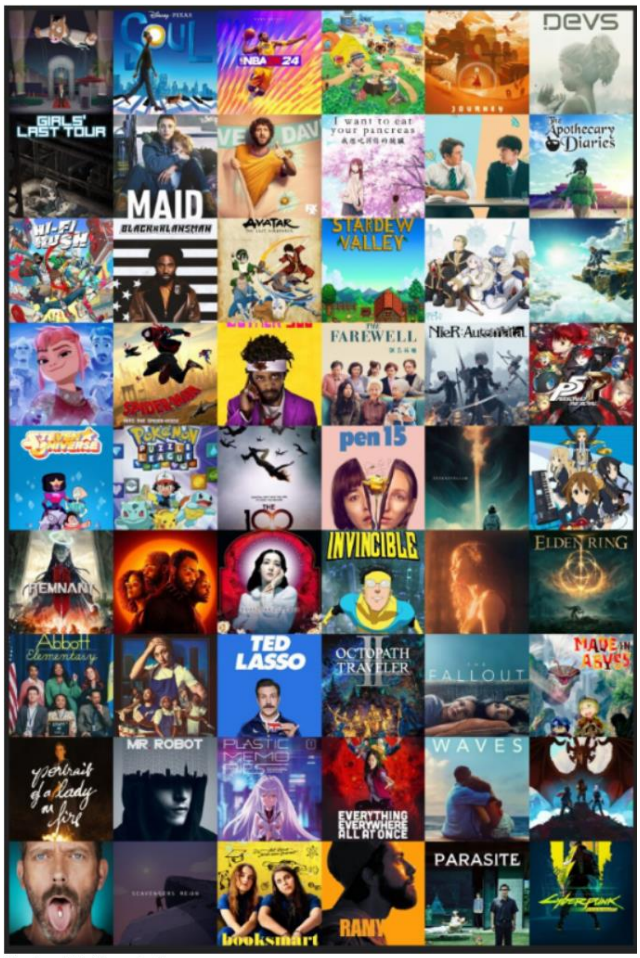
*Godspeed by Frank Ocean*

Salutations,

In the previous letter, I reflected on how a conversation about TV shows made me feel like an outsider and amplified my feelings of imposter syndrome.

I can't change how people see or treat me, but I'm still trying to treat others the way that I want to be treated. We all deserve to have opportunities to share our thoughts and feelings, likes and dislikes. So before getting deep into the mental anguishes of grad school, I wanted to share some things that helped break up my grad school journey.

Below is a mosaic highlighting some of the TV shows, video games, and movies that kept me going while helping me reflect on my grad school journey.




Made with Mosaically

Bojack Horseman | Soul | NBA 2K24 | Animal Crossing | Journey | Devs

Girls' Last Tour | Maid | Dave | I want to eat your pancreas | Heartstopper | The Apothecary Diaries

Hi-Fi Rush | BlacKkKlansman | Avatar | Stardew Valley | Sousou no Frieren | Zelda: Tears of the Kingdom

Nimona | Spiderman: Into the Spiderverse | Sorry to Bother You | The Farewell | Nier:Automata | Persona 5: Royale

Steven Universe | Pokemon Puzzle League | The 100 | Pen15 | Interstellar | K-On

Remnant II | Atlanta | Sympathy for Lady Vengeance | Invincible | Euphoria | Elden Ring

Abbott Elementary | The Bear | Ted Lasso | Octopath Traveler II | The Fallout | Made in Abyss

Portrait of a Lady on Fire | Mr. Robot | Plastic Memories | Everything Everywhere All At Once | Waves | The Dragon Prince

House | Scavengers Reign | Booksmart | Ramy | Parasite | Cyberpunk Edgerunners

**Tsundoku**: the art of buying books and never reading them. I didn't get to read much other than scientific journal articles…but I'll get to books eventually!

*Hope you're well and taking time to take care of yourself!*

GJL

*P.S.* If a friend asked you to give them something so they could know you better, what would you give and why?

**I don't belong here**

*atlas by Keshi*

Reader, May 2019

Today I had my **preliminary exam**, or prelim for short. It's the most unique exam I've taken: After a year of conducting independent research and reading copious academic papers, I prepared a presentation detailing how I'm going to successfully complete my research.

I was put into a room with my advisors – 5 professors who I've asked to supervise my progress during my PhD. I stood tall and explained a carefully thought-out research proposal to experts who each have published **many** scientific publications.

"What's the definition of van der Waals? What will you do if your experiments don't work as you expect? If a tree falls in a forest and no one is around to hear it, does it make a sound?"

After an hour of answering questions on my project, I left the room to allow my advisors to discuss how I did.

Deflated, exasperated, mind afloat, I remembered to breathe.

"Did you just finish your prelim?" I nodded to the student passing by in the hall. "Congrats, the worst part is over."

After sitting on the floor for 27 minutes, I'm called back into the room.

"Gilbert, we'd like to thank you for the presentation, but we can't give you a pass. There are some weaknesses…"

But I tuned the rest out. They didn't say the word, but I knew what it meant.

*i failed*

"…however, you're making progress, and we look forward to seeing you have another opportunity next year."

I cried for an hour in that windowless, dimly lit room. I didn't want to exist.

**For many prelim failures, the journey to the PhD ends here. You're given a master's degree and asked to leave.**

I get a second chance. But do I have what it takes to move on from this gut-wrenching result to try again next year?

I'm one of the few black people in my research program. I'm insecure about my identity and my ability to communicate. But I want to leave Wisconsin on my own terms. To do that, I have to figure out what I'm missing in my learning. What can I do to pass next year? Is something about my identity making me not good enough?

But for now, I'm in the comfort of my bed. Is it still considered crying if your eyes no longer have tears to shed?

GJL

*P.S.* Have you discovered any limitations about yourself recently? If so, how have you pushed through them?

*P.P.S.* I think this song fits how I'm feeling: strained by feeling constantly overwhelmed with a hint of wanting to be better. These lyrics have been pulsing through my head amidst my existential dread. *i don't want to fail, i don't want to feel like i have a second chance just because i'm a minority, i don't want to feel like a token student in this white state. i want to do science and learn, to discover knowledge and support other minorities, do i have to live with these feelings of failure forever? am i okay with that when will i be okay with that should i be okay with that? i hate myself, will i always feel like a failure no matter what i do in my life, am i worth anything other than work? why am i here?*

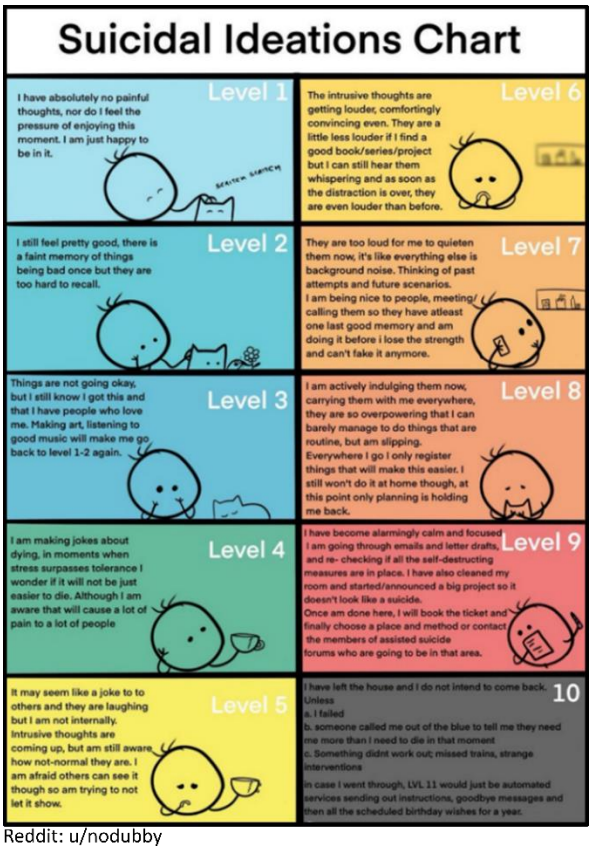
***Trigger Warning: mental health, suicidal ideation.*** *I sincerely hope that no burden in your life is too heavy to bear.*

**If you can't survive, just try**

*I Always Wanna Die (Sometimes) by The 1975*

My Dear Reader, October 2019

A few months ago, I had a dark notion. As I was waiting for the bus, my head felt heavy and everything went black. My eyes were open, yet things looked motionless, greyscale. Intrusive thoughts flitted in and out of my head.



Reddit: u/nodubby

*You're tired. What if you just stopped trying?*

But wouldn't my family and friends be disappointed in me?

*Nah, no one would care. Stop now, give up, what's the point.*

I want to end this on my own terms.

*You still can. Just walk forward. Don't think. Move.*

Into the middle of the street?

*It won't take longer than a second.*

Thankfully the bus arrived. I've had bouts with depression in the past, but never to the extent that my inner suicidal ideations were so evidently clear.

Then one day as I was preparing to start my day with a walk through the farmers market, I found myself unable to move. My mind was awake, eyes were open, and my inner voice saying "Move!", "Get up!", "Go over there".

But my body wouldn't listen.

Physically isolated from family and feeling like a burden to friends, my mental health continued to fade. I struggled to move, I stopped eating, and knew I needed help.

My therapist helped me state my feelings, rationalize, and reflect on my thoughts. I realized that I live most days at levels **6-8** on the above **suicidal ideations** chart. She reminded me to eat, gave me actionable suggestions to help me. My journey back to myself began with her advice: "Take risks".

I find it ironically comical that risking my mental health is the motivation to pass my prelim because that feels riskier than just leaving. But I'm still here. And it's time to put in the work to learn some science.

GJL

*P.S.* What's the riskiest thing you did in the past year?

*P.P.S.* My body stopped listening to my inner voice. Although my mind contemplated feelings of wanting to disappear, something within me felt that it wasn't right and resulted in outright rejections of simple thoughts of movement. Literally petrifying.

*P.P.P.S.* My limits are being tested here, mentally, physically, emotionally. But I'm still here. I've never acted on my intrusive thoughts, but the risk is there. The strings fill me with hopeful melancholy, accentuated by the strain in the singer's voice. As if things will get better with time, even though it doesn't feel that way right now.

**Don't worry 'bout tomorrow**

*WELCOME TO SOUL, PRESENT by Q*

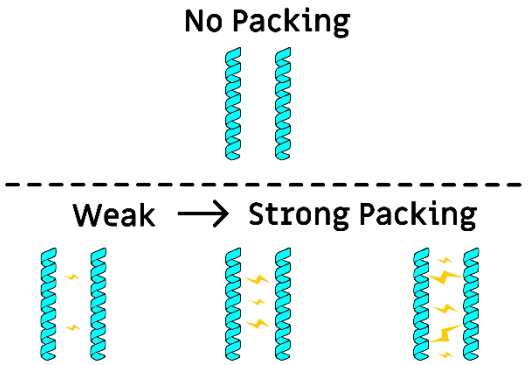
My Dear Reader, February 2020

I have a few more months before my second chance to pass my prelim. I'm anxious. I think not understanding how to effectively communicate my project was a glaring weakness during my first prelim. I need to learn how to speak confidently about my science if I'm going to be successful in this pursuit of knowledge.

*How does van der Waals packing impact membrane protein association?*

The focus of my research is to further humanity's understanding of how proteins found within cell membranes, or membrane proteins, interact.

**Membrane proteins** are complex structures that help our cells adapt to stimuli. From helping us recover from cuts and bruises to signaling to our brain that something feels hot or cold, these proteins are responsible for a multitude of our body's natural responses to the environment.




The proteins I work with are tiny, helix-like structures. These proteins naturally like to stick together, or associate. I'm trying to discover if different amounts of a force called van der Waals packing changes how well proteins associate in the membrane.

**Van der Waals packing** is like the "static" that comes from rubbing a balloon against hair: it's a weak attraction between things in close contact, sticking them together. We have tools to predict the "static" strength in between proteins, but no one knows how strongly it sticks membrane proteins together. Is packing more like the lid of a jar before or after asking someone to open it: impossible to twist or so loose that it's open in 3 seconds?

My current goal is to make a computational **algorithm** to build, or design, proteins with different amounts of packing. Using software produced by my lab, I can create **computational models** of proteins. We know that protein shape affects the amount of "static", so I'm currently developing a way to design proteins with different shapes.

My second goal is to test these designed proteins with experiments to see if my predictions agree with the data. But that comes after my prelim. If my computation and experiments are similar, then I'll be able to design membrane protein targeting drugs to prevent numerous viral infections and diseases.

Wouldn't that be a story? Researcher at UW-Madison finds the cure for Alzheimer's, cystic fibrosis, or cancer.

But that's MUCH farther away, tens of graduate students of research after me. For now, my research will have the potential to be a "possible outcome of a solution" that impacts humanity.

I'm feeling a smidge more confident in communicating my research. But this burden to pass to prove myself as a scientist still feels heavy. *if i don't pass, then did i ever even have potential in science? why does it vaguely feel like my failure would reflect poorly on not just myself…but other minorities in science?*

GJL

*P.S.* When was the last time that you felt proud of yourself?

*P.P.S.* Have you ever felt indebted to those who gave you an opportunity? I've been given this chance, but I feel like if I fail, I'll reflect poorly on my advisors who trusted me to succeed. I'll be an example of why there aren't many minorities in upper-level degree programs. So I'm focused on learning today instead of worrying about the result.

**Time has come, take it all in**

*5 Year Plan by Chance the Rapper*

To My Humble Acquaintance, June 2020

*Have you ever thought of how you could impact humanity?*

With the world currently entrenched in a global pandemic, with minorities being abused and killed in a time of global strife, why is understanding how proteins interact important?

*If we better understand how membrane proteins interact, could we engineer proteins to disrupt these interactions?*

It's bizarre to know that I'm living through this piece of history. For anyone studying virology, they get to see the applicability of their research in real time. But for a biochemist trying to understand the forces involved in membrane protein folding, could my research ever help anyone?

*But how do you engineer proteins? We'd need to understand how van der Waals packing sticks proteins together.*

I've learned to question whatever I don't know about a subject. I've developed critical reading skills, marking any paragraph, sentence, word that I can't fully comprehend in my brain. I take that information and go deeper, peeling away layer after layer until I reach the core bit of knowledge that I need to understand.

*If we can computationally predict how viral proteins interact with human cell membrane proteins and then support that with an experiment, we could engineer drugs to combat viral infections!*

Know what you don't know. I joined my advisors on a Zoom call and pitched my research to them.

After 40 minutes of defending and answering questions, my prelim exam was over.

***I passed***

*But did I really do that much better? Was I that much more prepared?*

Yes and yes. But all of this deep learning has left me searching for more questions: *What don't I know?*

With the state of police brutality and empowerment of the Black Lives Matter movement happening right now, *did I pass because it would look bad if one of the few black people in the program got kicked out for failure*?

GJL

*P.S.* What are 5 things you're excited to do in the next 5 years?

*P.P.S.* A PhD is usually 5-6 years, but who can plan for failing a prelim AND for a global pandemic. The chords in this song remind me of a sunshower: a sprinkle of refracted sunlight, dancing on your skin amidst the soothing **petrichor**. It's like a prescription of hope for the future. Even if you can't predict how things will go, you can appreciate what happened because you're alive. It makes me feel optimistic for a future in which I'll feel nostalgic about the present.

*P.P.P.S.* My feelings of my identity driving my success still weigh heavily on me. There are still a lot of things that I don't understand about my project, but I'm trusted to take it to completion. My research may not be used for anything impactful. But the fact that it'll be a small bubble on the expansion of human knowledge is enough for me.

Tumblr: Cannonbreed

**Into the woods**

*Into the Woods by Mree*

Rest Here Weary Traveler,

When I need a break from thinking, from emotions, from stress, I close my eyes and find respite in music.

This song channels both the fear of traversing the unknown, and the excitement of the discovery waiting ahead.

You've made it to the middle stages of my journey. The pandemic slowed everything down for me while I was in grad school. Those few years blurred together. Then as if no time passed, stores began to open, mask mandates ended, and people started to gather again.

Have you ever felt so much pressure that you felt like you couldn't carry on? Like there's a huge weight on your shoulders and you need to finish something for it to go away?

I've learned how to learn, but science isn't kind. It's purely honest. And if your hypothesis or experiment isn't good enough, it's difficult to find data to make conclusions worthy to be shared with the scientific community.

And it's equally difficult to know your timeline. To have to convey that you're a year away from being a year away from graduation to family and friends, and then repeating the same thing each year after that.

But I'm nearing what it means to find knowledge, to ascertain truth, to discover.

Thanks for reading, sending much love and support your way! I hope you enjoy the rest of these letters.

GJL

*P.S.* If you're still not convinced that research like mine could be helpful to humanity, another group doing similar stuff was able **to use protein design to combat Coronavirus**.

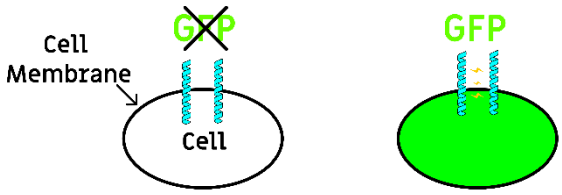
**Why don't you leave if you wanna leave**

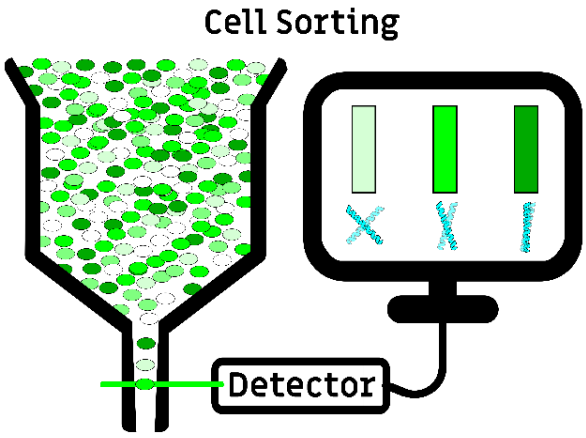
*Leave If You Wanna by Overcoats*

Dear Friend, July 2023

I've fine-tuned my algorithm over the last few years, tested my designed proteins in small experiments, and it's time to do a larger test. My algorithm can do three things: 1) build 1000s of proteins with unique shapes, 2) predict how well the proteins stick together (**associate**), and 3) estimate the amount of packing ("static").

A few years ago, my lab developed an experiment that I can use to test my proteins. We created a way for bacterial cells to produce our proteins alongside green fluorescent protein, or **GFP**. When the proteins associate stronger in the cell, more GFP is produced. We can't count exactly how much GFP is made, BUT the GFP makes cells light up green!




We can measure the cells' green light using a machine called a **cell sorter**. The machine passes bacterial cells through a detector that records the green light for each cell. This allows us to estimate how much GFP is produced per cell, effectively telling me how well our proteins stick together! I sorted billions of cells producing different proteins, and calculated how well each protein sticks together from the data. And for once, it looks promising! My designed proteins associate!

In my first 3 years of graduate school, no experiments worked, and I grew accustomed to failure. But I've finally reached the point in my graduate career where instead of taking an hour to work up the mental bandwidth to conduct an experiment, I can plan 2 or 3 experiments around one another. My experiments work on the first or second try. I'm a well-trained, diligent grad student.

But the laboriousness of research is draining me. I know this is the best data I've discovered, so why do I feel so detached from my science? Why is my excitement so fleeting? Why do I miss my family and friends so much?

*Why is the only thing on my mind this desire to leave?*

GJL

*P.S.* What was the last family/friend gathering that you regret missing?

*P.P.S.* After years of failure, I've finally found my stride in grad school. My project is taking shape, and the data I've discovered will ever so slightly push the boundary of membrane protein knowledge forward. I've learned to think on my feet and to dissect experimental data. I've become a good scientist. But after experiencing some success, I now feel this sudden urge to leave. I'm realizing how much I've neglected my life outside of science. Missing events with family, not seeing close friends for years. Am I losing my passion for research? Is science moving on from me?

*P.P.P.S. Why am I still here?*

**'Cause is it really love if it don't tear you apart?**

*When You're Breaking My Heart by Gatlin*

Confidante, September 2023

*When was the last time you had your heart broken?*

My time in graduate school is finally coming to an end.

I was given the acknowledgement of my advisors that I'm close to finishing up: "We can see the story forming, and we think you'll be able to graduate next summer."

Have you ever experienced a **mental spiral**? Where your thoughts begin in one place and start looping into another. And then another. Another. An infinite abyss of issues, problems, and conundrums to work out in your head.

*I need to rest but I should organize the data for my thesis, I guess I'll just do that this weekend, but then when do I get to rest and just not think? I also need to go grocery shopping, but is eating even worth it this week, do I even have enough money for food? after I graduate how will I make money to live, I have no marketable skills, I still don't know what I really want to do, is academia worth it? should I join industry, potentially do something insufferable like optimizing protocols? would I be able to listen to music every day? would I feel like more of an imposter wearing professional clothing after years of hoodies and sweatpants? am I even good enough to do anything?*

Pensive, contemplative, unrelenting, my stream of consciousness spirals out:

*What is my relationship with my PhD?*

It's an intense relationship, where my research can do no wrong and I'm always at fault. Experiment after experiment, failure after failure, I attribute to myself.

All this failure has resulted in a fear of rejection.

I've been selfish. I'm no longer reliable. I haven't given friends support, rarely give intentional time to my family. Asking for help feels like I'll be rejected just as I've rejected my relationships.

At times my PhD has felt like a black hole, sucking everything in, leaving me with nothing.

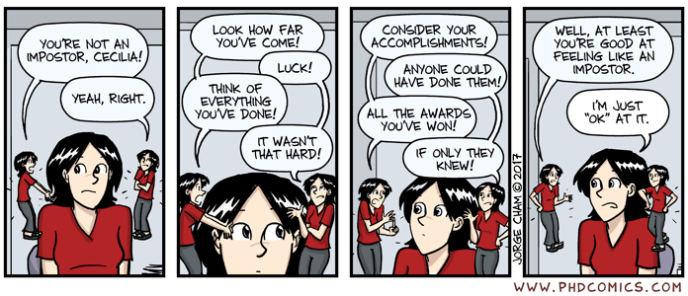
I'm not smart, I don't pick up on things quickly, but I'm doing my best to learn how to learn at the highest level.

And now that I'm getting close to the end, and it feels like I'm finally getting something back, is it fair that I can't even smile when I think about graduating? Why does it feel like I'm being pushed away?

GJL

*P.P.S.* In this song, Gatlin realizes that she loved the chase of a relationship more than the person.

*P.P.P.S.* For a while now, I've stopped thinking about my future. I don't know if I want to be a professor anymore, to support and mentor students, or even remain in science. I've been focused on chasing this PhD. *What do I love? Is there anything that I would be happy doing for the rest of my life? The pursuit of finding knowledge is what kept me here, but do I even enjoy it?*



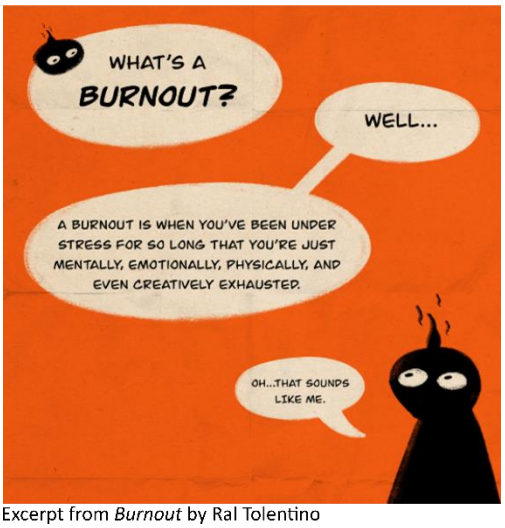

*P.P.P.P.S.* Those questions from years ago ring in my head: *Do I belong here?*

**You can't stay in bed forever**

*GOOD MORNING SUNSHINE by The Narcissist Cookbook*

Hello Supportive Friend, November 2023

I didn't realize the end of this journey would lead to such paralyzing and everlasting **burnout**.



Excerpt from *Burnout* by Ral Tolentino

My fierce zeal for science, to understand the auspicious beauty hidden within intricate systems that naturally form life, feels like it's nearing an end. That once unwavering fire inside me is fading as I run out of kindling: patience, time, and willpower.

It feels like I've already given a lifetime's worth of energy to this endeavor called PhD. My body is perpetually tired, my mind ailing with doubt, anguish, and distress. Another sleepless night. Hours pass and I continue to feel useless, thinking of all the things that I have to accomplish to graduate, yet being unable to do any of it.

Everyday feels like I'm searching for something: *What's the one thing that will pull me out of bed?*

Too many goals, temporary objectives, efforts to maintain content and find solace in the chase.

*Today I'm going to re-analyze my experimental data. I'm going to write the methods portion of my thesis.*

Frustration begins to set in. The goals shrink as the hours pass, as does my willingness to do anything.

*I'll prepare for my presentation next week, maybe take out the trash, go grocery shopping, do laundry for the first time this month, clean the kitchen, wash dishes from the fried rice I made 2 weeks ago, feed my cat, eat SOMETHING, shower for the first time this week, brush my teeth.*

Overwhelmed and powerless, I'm forced to abandon that teeny glimmer of hope that I would make it to work today. Only one goal is left.

*I just want to get up and leave the comfort of my bed.*

It's 3 pm and I'm ready to start my day. I'm alive.

*But am I well?*

GJL

*P.S.* When was the last time you felt burnt out? Were you able to take care of yourself? Have you learned what types of support you need/want?

*P.P.S.* At a time when I just needed ANYTHING to feel good about, I discovered this song that immediately resonated with me. Most days, it's hard to get out of bed. It's difficult to motivate myself to do even the most inconsequential things. I find ways to work through it as best as I can. Even with all these anxieties bogging down my mind, I find a way to tell myself that I've done enough each day. Telling myself that gets me through, keeps me productive, and gets me closer to the end of grad school.

**You don't cross my mind, you live in it**

*Off Day by Lyn Lapid*

Sup Bud, February 2024

*What does it mean to live?*

As I write this, I should be sleeping, prepping my body and mind for the next day. I close my eyes, but thoughts continue ringing in my head, drowning out my tinnitus. I peek at my phone: 12:19 AM. Readjust, pull the sheets closer. I don't enjoy sleeping tight. It feels restrictive and reminds me of being unable to force my body to move.

2:13 AM. I stare at my eyelids, no rest or reprieve.

During the months leading up to my 2nd prelim, I found myself dreaming of science. My subconscious was hard at work while I slept, imagining how to design proteins: conceptualizing electrons, envisioning the vibration of atoms in space, building amino acids, trying to make sense of why proteins associate.

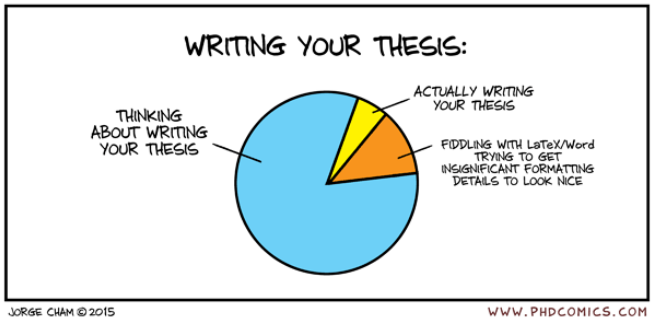
I thought it was a superpower.

But now, with a deeper understanding of my research, the main thing left is to relay my findings. No more anxiety about redoing experiments or thinking of new ways to interpret data. I need to write, to put my thoughts into coherent words on paper. But how?

Instead of dreaming, my mind is anxious *to work.*

I can't sleep because if I do, I'm not working. But if I can't sleep, I can't work well. A delightful feedback loop at the end of my PhD.

But during my 3AM trips to the lab, I'm one of the few beings filling the **halcyon** silence. I'm finding a deeper appreciation within this stillness.




Cherishing the freedom and fluidity of time in graduate school. Taking advantage of my sleeplessness. Remembering to breathe. Allowing myself to be mesmerized by blinking traffic lights. Catching myself smiling in the early glow of the sun as the brisk air fills with birdsong.

*Time doesn't feel like it's moved much for me during my PhD.*

And I'm finally starting to appreciate it.

GJL

*P.S.* When was the last time you closed your eyes and just took a moment to pause, breathe, and take in life?

*P.P.S.* I'm going to miss this environment that pushes me to discovery. Given me time to allow my brain to acclimate to the idea of working at the boundary of human knowledge. Allowing me to excise bias in pursuit of truth. It reminds me of being a kid again – staring at the sky, letting thoughts freely flow in and out of my mind. I feel like graduate school harnesses this latent ability, allowing for deeper exploration of whatever I find interesting.

*P.P.P.S.* This song feels like it bundles you up in a freshly washed blanket on a snowy day. A blend of tranquil and reassuring warmth amidst the frigid winter. I've finally made it to a place of what feels like mutual respect. Pushed to grow by finding comfort outside of my comfort zone. Pulled out of the ebbs of 36-hour water diets, to the flows of stress binge eating donuts and ice cream. Birther of my restless mind, bringer of appreciating peaceful subtleties of the day. My PhD has been that warm blanket for me, and I never expected it to be so comfortable.

**Once I saw fire…Did I let it go?**

*Before the Line by dodie*

To You Who I Hold In High Regard, May 2024

*If you could do it all over again, would you still go to graduate school?*

I've been caught within the greyscale of science for so long. Crafting my contribution to the scientific community, developing my research to ascertain a nuanced glance into membrane protein association.

But these reflections have reminded me how much I've learned about myself in graduate school.

From stumbling early on in my quest for discovery, to now being ever so close to reaching the mountain top. I've been able to reflect on my shortcomings, explore my love for science in depth, while discovering and mending cracks in my mental health.

I've willingly put myself through this grind because of how big this opportunity feels to me: I'll be the first person in my family to receive a PhD, in a field as prestigious as science. I recognize how important it is to become another minority in a field still growing in diversity.

And I've been fortunate during graduate school. My family is healthy, my friends are understanding, and my advisors and lab mates have been exactly the type of academic support I've needed.

It hasn't been easy. But I've learned to take care of myself through intense mental strangulation, and now I'm re-learning to cherish my life.

No longer just doing things as a distraction from stress.

Walking, breathing, smelling the morning **geosmin**. Running, playing basketball, investing in video games, music, and anime. Baking, composing playlists, writing letters.

Playing with my cat Jada, who's saved me from anxious bursts and emotional downfall countless times.

And I'm realizing now how little time I've taken for myself.

Jada atop a pile of notebooks, jelly beans, and miscellaneous items

For most of my life I've continuously reached for the next academic accomplishment: Diplomas, Bachelors, PhD.

But for once, I feel an odd bit of freedom: *I don't know what comes next*.

**Trumspringa**: the longing to wander off your career track in pursuit of a simple life.

The feeling that an escape from what you're currently doing will be enough to bring you back to the grind. To remember how to be excited about what's next. To feel like you can take time out of your day without being stressed about all the things you have to do. To appreciate the world around you.

*I can't shake this sense of journey, of needing to go somewhere new. To remind myself to…experience.*

GJL

*P.S.* If you could do it all again – as a bright eyed, younger version of your current self – would you make that big decision that led you to where you are today?

**I'm trying to start my life again**

*Wallflower by mxmtoon*

My Dear Friend, August 2024

After my interview for grad school, students took me on a night walk around Madison. It was eerily quiet. No insects or animals, little noise of a bustling city, as if the sound was dampened by the snow mounds littering the streets.

We made our way to the terrace, tiptoeing around icy, slippery sidewalks. We went ice skating on a small lake earlier, so I understood that Madison could get quite cold. But I was astounded when I saw Lake Monona, 13km/5mi of water across, completely frozen over.

We ignored the orange tape and traffic cones and walked onto the lake. I stared into darkness, wind gusting chill against my corneas. I closed my eyes, seeing the same darkness with one difference: the **phosphenes** on the backs of my eyelids, clearer than ever before.

Photo by pauliefred on flickr

That initial memory foreshadowed my PhD: Serenity in the quiet, astonished by the unanticipated display of nature, as biting cold winds made me question why I'm here.

It's been 7 long years.

I've fallen in love with the rich varieties of cheese and ice cream, got acclimated to the many farmer's markets, became accustomed to thanking bus drivers at stops, and appreciated people biking on even the chilliest days.

Madison is a charming city, a welcoming environment despite my many ruminations of feeling like an outsider.

A lot can happen in a PhD. I failed my prelim but got a second chance. But my story isn't the only one, and things can hinder you professionally and personally. Someone could embarrass you at a conference, or you might be forced to switch labs. You might need to prioritize a family member over science, or your partner could break up with you.

For all of us privileged to learn at the highest level, it's still life. Troubles with relationships, money, work-life balance. We juggle these things, trying not to break down while thinking critically and finding comfort in the unknown.

As I'm writing this, I haven't yet defended, the thesis isn't finished, and I'm not even sure if my committee will award me my PhD. But my research is submitted to be published. To occupy the same library as the discoveries of DNA and protein structure, albeit in a much less frequented bookcase.

The "static" packing that I designed between proteins isn't as strong as other forces. But hundreds of my proteins resulted in green, fluorescent light within millions of cells. I designed proteins from scratch and showed that "static" packing makes membrane proteins stick together and dance.

But my model isn't perfect.

I designed these proteins and even though they associate, they don't always match the prediction. There's still more work to do to mathematically understand how strong "static" packing can be. More to learn before we can use this information to engineer effective membrane protein targeting drugs.

I hope someone can take what I've done to inform future models. To use what I learned about packing to dissect its impact on other forces. Does "static" packing affect the impact of other forces? Is packing always only a weak force? Or is it stronger in the presence of other forces?

If my research can nudge just one future scientist in the right direction, that's more than enough for me.

*How does it feel to discover something, to study at the boundary of knowledge?*

Do you remember the first time you held a **kaleidoscope**? For those first few seconds you see a glistening, prismatic repeat of color and shapes so overwhelming that it seems unable to fit into the small toy in your hand. It's infinite.

That fleeting moment in admiration of unexpected beauty is probably something I've been chasing my entire life.

Despite my struggles, I've loved my PhD. I've thoroughly enjoyed searching in the unknown, trekking through journal articles. Embracing that childlike, innocent curiosity while searching for deeper understanding. Becoming enthralled with that feeling of comprehending something new and building it into a lifelong passion.

In the end, my love for learning won out.

I can see myself doing the same thing forever. Searching for that same joy in discovery, that kaleidoscopic spark.

But my head is throbbing from banging against walls full of knowledge. From forgetting what it's like to take time not thinking about my research.

From prioritizing learning over my physical, mental, and emotional well-being. Over friends and family.

I'm tired of constantly putting pressure on myself to succeed. For now, I've used up all my passion and determination.

**Etterath**: the feeling of emptiness after a long and arduous process is complete.

Every time I wanted to leave, I felt exhausted from the grind. But now it's more like watching the last episode of a TV show or beating the final boss in a video game. Bittersweet. Not exactly ready…but knowing it's time to let go.

I'm not yet sure what I'll be doing next. But first I want to regain my ability to experience. Travel to new places, use my coding skills to analyze basketball stats, do a darkness retreat. Just give myself *time* to explore.

No one can take away your education. A break from the grind won't lead me to forget what I've learned.

Thank you, PhD, for teaching me how to learn at the highest level. I'm not the smartest student, and this journey wasn't anywhere near a perfect PhD. But it helped me grow into the person that I am: a tryhard, a thinker, and someone who knows I can make a difference somewhere if I just put my mind to it to learn.

GJL

*P.S.* To my mom, dad, and brother, thank you for all your support. From sending food or just making time to distract me from personal turmoil, I appreciate all the love you've sent my way.

*P.P.S.* To my advisors, friends, roommates, lab mates, therapists, communities (shoutout SciMed, CBI, and IPiB DEI!), thank you for every bit of support you gave me along the way. Whether it was thoughtfully chatting about science to discover my weaknesses, finding ways to develop a more supportive environment for students, or just discussing life and food and the mundane things – I truly appreciate every conversation, every moment of time that you've given me. This PhD wouldn't be complete without it!

*P.P.P.S.* And to you reader, thank you for sharing this journey with me. Academia isn't always the most vulnerable, so I wanted to create a transparent way to personally highlight one grad school experience. I hope that I succeeded! Wherever you are on your own journey, best of luck on finding solace and happiness.

**letters for my phd-the extras**