Chapter 3: Computational Methodology

Table of Contents

[3.1 Abstract 3](#_Toc174306308)

[3.2 Introduction 4](#_Toc174306309)

[3.3 Protein Design Algorithm 6](#_Toc174306310)

[3.3.1 Analysis of membrane protein PDBs 7](#_Toc174306311)

[3.3.2 Choosing amino acids for membrane protein design 10](#_Toc174306312)

[3.3.3 Defining the Interface 11](#_Toc174306313)

[3.3.4 Developing the energy terms 12](#_Toc174306314)

[3.3.5 Sequence Search 15](#_Toc174306315)

[3.3.6 Backbone Refinement 16](#_Toc174306316)

[3.4 Analysis 18](#_Toc174306317)

[3.4.1 Software 18](#_Toc174306318)

[3.4.2 Design Analysis and Choosing Interface Mutations 19](#_Toc174306319)

[3.4.3 Fluorescence Reconstruction 20](#_Toc174306320)

[3.4.4 TOXGREEN Conversion 21](#_Toc174306321)

[3.4.5 Determining proper membrane insertion 22](#_Toc174306322)

[3.4.6 Identifying proteins associating by designed interface 23](#_Toc174306323)

[3.4.7 Comparison to energetics 24](#_Toc174306324)

[3.5 Summary 25](#_Toc174306325)

[3.6 Supplementary Figures 26](#_Toc174306326)

[3.7 Supplementary Tables 32](#_Toc174306327)

[3.7.1 MSL Scripts 32](#_Toc174306328)

[3.7.2 Python Scripts 34](#_Toc174306329)

[3.9 References 36](#_Toc174306330)

3.1 Abstract

Recent advances in experimentation allow researchers to collect data in high-throughput. Computational tools and software are invented to complement, designing experiments to collect and analyze large datasets. My research implements a protein design algorithm paired with high-throughput sort-seq to characterize these designs. 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 mentioned previously, many of the techniques used take years to ascertain conditions that 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 van der Waals 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 Zn2+ 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.

A diagram of a dna sequence

Description automatically generated3.3 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 van der Waals packing on membrane protein association, I opted for a high-throughput design approach. 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.

A diagram of angles and distance

Description automatically generated3.3.1 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 Membrane Proteins (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 (Figure 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 then 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 then extracted two parameters: the distance (x-shift, **d**) and the angle (crossing angle, **θ**) (Figure 3.2A), which I plotted as a scatterplot and analyzed using kernel density estimation (Figure 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 Figure 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 (Figure 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 GASright, which is composed of dimers with a short interhelical distance (6.5-7.5 Å) and crossing angles of -25° to -55°. The GASright ­is well characterized and known to be stabilized by a combination of van der Waals 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 A black and white grid

Description automatically generatedstabilized solely by van der Waals packing and GASright.

**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**) (Figure 3.2A). To determine these features for the corresponding angles and distances, I created a grid of template geometries for each design region (Figure 3.3). I used MSL to place poly-Leu sequences at each geometry with Ala (Left and Right) or Gly (GASright) at the interface (described in section 3.3.3). TMHs were assessed for clashing at the interface by measuring the van der Waals energy. If the structure is clashing with the small AA Ala (Left and Right) or Gly (GASright) 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 (Figures S2.7, S2.8, and S2.9). Finally, I generated 1000s of geometries for each design region, where the angle and distance are chosen from the membrane protein 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.

A colorful circle with numbers and letters

Description automatically generated3.3.2 Choosing amino acids for membrane protein 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 membrane protein sequences (Figure 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 (Figure 3.4B). Each of these AAs was allowed to be designed along the interface during the sequence search described in section 3.3.5.

A screenshot of a cell phone

Description automatically generated3.3.3 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 GASright 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 membrane protein analysis (Figure 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 (Figure S3.1).

A diagram of a dna strand

Description automatically generatedFor subsequent design runs, I standardized the interface for each region (Figure 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 LxxLLxL 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 (Figure 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.

**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.

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 van der Waals 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:

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).

A blue line with black dots

Description automatically generatedThe 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 (Figure 3.7). This BASELINE\_MONOMER term was made only for the subset of amino acids used in design (Figure 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 membrane protein 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 membrane protein sequence (Figure 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:

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 a membrane protein. 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:

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 membrane protein sequences (freq\_AA) to the power of the number of each AA in the sequence multiplied by the 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 membrane protein 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 A diagram of a dna helix

Description automatically generatedenergy is as follows:

**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 (Figure ST 3.2). The sequence with the best total energy is saved, and that single design undergoes backbone refinement (Figure 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 (Figure 3.1A: d, θ, ω, Ζ) 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:

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 (Figure 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 Figures S3.2and 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)

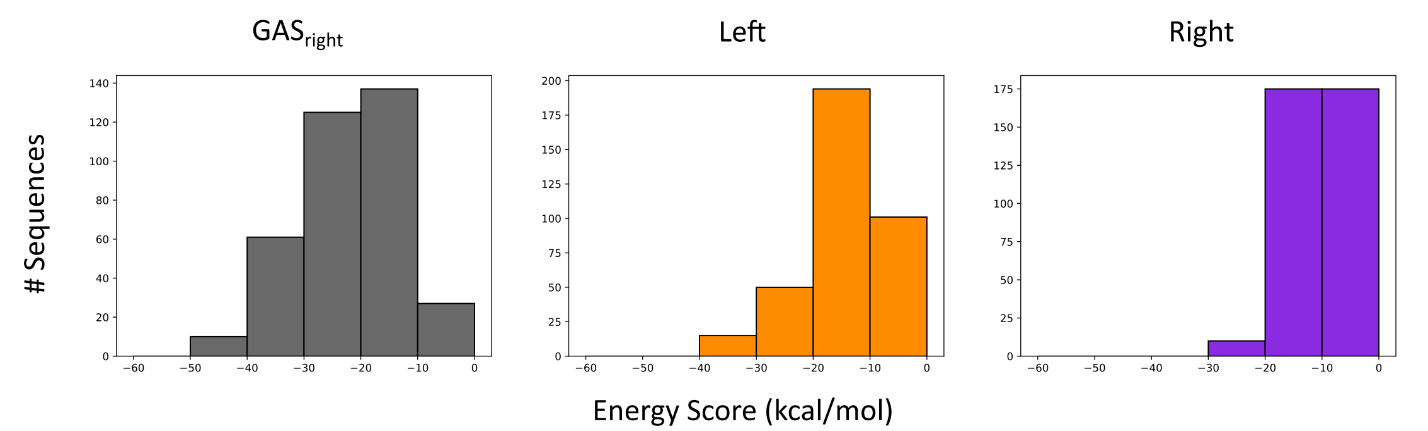
Sklearn: (Pedregosa et al., 2011)

Logomaker: (Tareen & Kinney, 2019)

Pymol: (DeLano, 2002)

All programs and code can be found on [Github](https://github.com/gjowl/Sequence-Design).

3.4.2 Design Analysis and Choosing Interface Mutations



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

A screenshot of a video game

Description automatically generatedTo analyze my designs, I wrote a script that compiles all design directories and outputs a variety of plots (Figures S3.3 and 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 (Figure 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 (top) results in interfacial positions on one helix overlapping with atoms on the opposite helix. Mutations to Ala (bottom) 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 (Figure 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 (Figure 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 (Figure 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 (Figure 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), aij is calculated as:

A mathematical equation with black letters

Description automatically generated with medium confidenceA diagram of a function

Description automatically generated

where the normalized fractional contribution is then multiplied by the median fluorescence of that bin (mj­). 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 (Figure 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:

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.

A graph with lines and numbers

Description automatically generatedI 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 (Figure 3.10). To calibrate our reconstruction to TOXGREEN, we converted the values from the reconstruction data to TOXGREEN fluorescence (Figure 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%).

**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**.**

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:

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

A graph of different colored squares

Description automatically generated3.4.6 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 (Figures 3.11 and 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.

A graph with orange and black lines

Description automatically generated3.4.7 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 (Figure 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 GASright region (Figure 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 van der Waals packing using our energetics. However, many of our designs outside of the GASright 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](https://github.com/gjowl/Sequence-Design). Finally, I reference the results from my paper that can be improved upon in future experiments. Ways to further improve my design algorithm, the correlation to energy terms, and to study the impact of other forces on membrane protein folding is further described in Chapter 4: Future Directions.

A screenshot of a computer screen

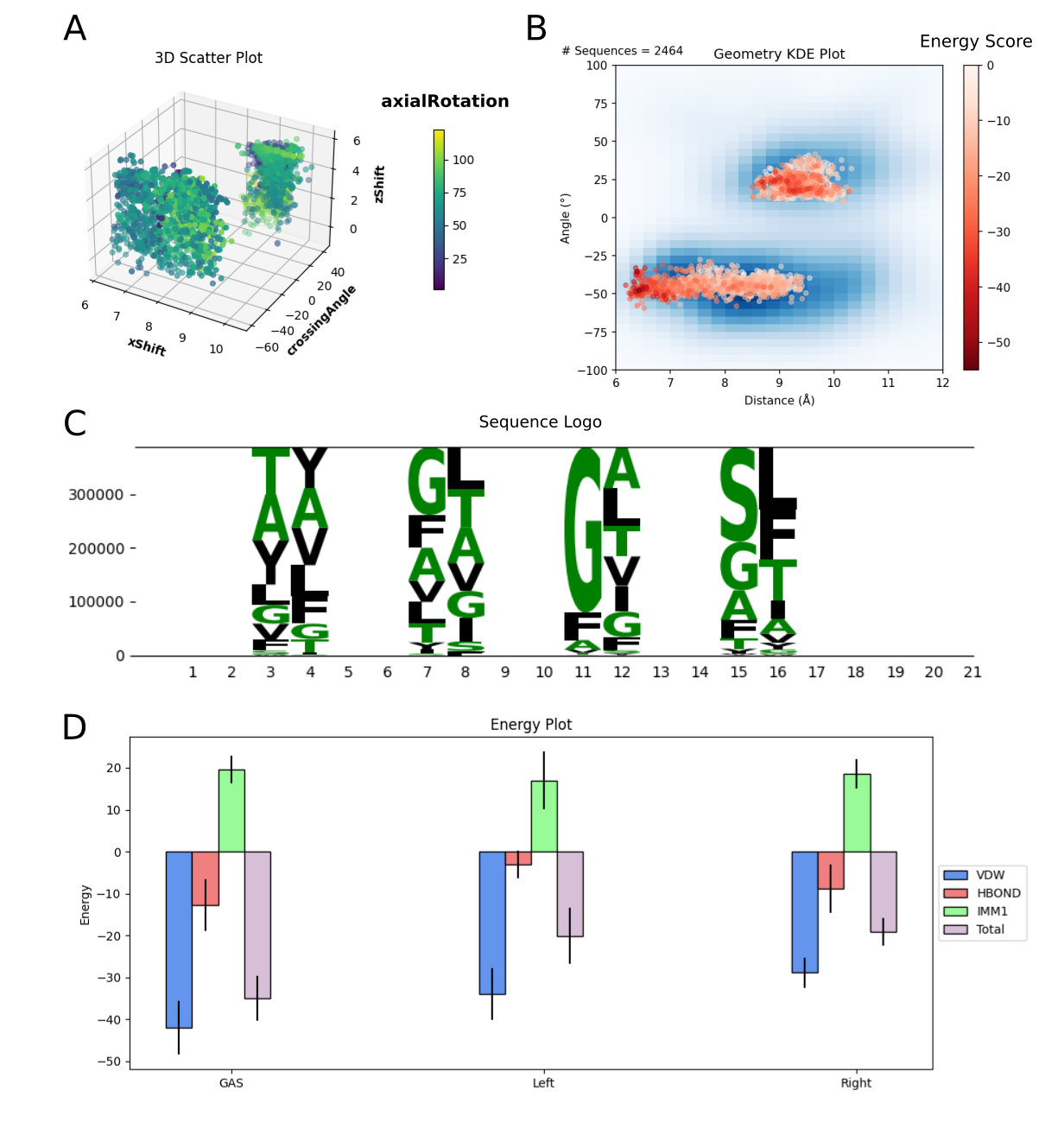
Description automatically generated3.6 Supplementary Figures

**Figure S3.1 Correlation for similar interfaces.** Energy scores (x-axis) plotted against the dimerization propensity in terms of %GpA. Correlations for interfaces with R2 > 0.6 from my first sort-seq design run.

A graph showing the difference between a refinement and a new

Description automatically generatedThe 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 S3.2 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. Structures are also shown to be more well-packed than previous (**add here**)

****

**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.

A graph of energy score

Description automatically generated

**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.

A diagram of a gene editing process

Description automatically generated with medium confidence

**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.

A screenshot of a graph

Description automatically generated

**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 van der Waals 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 Zn²⁺-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.