Design of Novel Repeat Proteins Containing Beta Strands

Dmitri D. Zorine

A dissertation

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

David Baker, Chair

Frank Dimaio

Neil King

Hao Yuan Kueh

Program Authorized to Offer Degree:

Bioengineering

University of Washington

Abstract

Design of Novel Repeat Proteins Containing Beta Strands

Dmitri D Zorine

Chair of the Supervisory Committee:

David Baker

Biochemistry

Repeat protein design is a natural direction for protein bioengineering thanks to inherent adaptability, geometric properties, and facile extensibility. Natural systems use repeat proteins as materials scaffolds, sensors, polymer binders, and functional superstructures. Protein engineers have both co-opted and adapted natural systems as well as created novel systems inspired by nature to serve new purposes. New deep learning approaches in both scaffold and sequence design have enabled significantly less constrained design approaches with greater success rates in designing soluble, stable biomolecules. This work presents a greatly expanded methodology for repeat protein design architectures. As particular examples, closed and unclosed repeat proteins with both alpha-helical and beta-sheet character are described, with morphology differing from both known natural and engineered proteins. A bioinspired terminal capping strategy is also presented, which greatly increases solubility and yields of repeat proteins containing beta strands.

Chapter 1 - **Introduction**

Repeat proteins serve a variety of functions in nature, where the repetitive nature of the protein intrinsically satisfies some biomolecular niche. Repeat proteins in nature function as cellular sensors, structural subunits, and even anti-freeze ice-binding proteins [1–5]. Scientists and protein engineers have sought to harness the advantages of repeat geometry to solve various technological challenges, including catalysis, ligand binding, and DNA editing. These efforts have led to a variety of incredible successes and advancements to the field, including de novo designed TIM barrels, TALE proteins with novel DNA binding targets, and repeat proteins with the potential to replace antibodies as biosensor molecules in some specific applications[6–9].

Alongside innovation in mutation, selection, and some proofs of concept, investigators have also established a variety of new and broadly useful frameworks for design of new repeat proteins as well as scaffold pools with inter-combinable component sequences. Notably, protein designers have established mutable and constant regions of a large variety of natural backbone geometries (LRRs or Leucine Rich Repeats, ankyrins, DARPins or Designed Akyrin Repeats, and TALEs or TAL Effectors)[10–15]. Researchers working on a tangential trajectory have further broadened this pool of designable repeat proteins by establishing repeat scaffolds fully novel in terms of structure and sequence, or *de novo* repeat scaffolds: new repeat proteins completely outside the scope of known natural proteins. These de novo proteins not only greatly expand the scope of potential geometries, but offer promise as scaffolds useful for construction of structured nanoparticles and biomolecular fibers[16,17].

Some fundamental design limitations continue to exist in repeat protein research. Repeat protein design has largely been restricted to two broad schools: strictly constrained design of natural-like protein architectures and largely unconstrained but all-helical *de novo* backbone design. These repeat protein “scaffolds” form the backbone of studies to which a repetitive or extensible geometry is uniquely suited. Scientific challenges with new geometric constraints often return investigators to the proverbial drawing board; if an existing scaffold set cannot be rapidly sequence designed to fit a particular binding or scaffolding challenge, an all-new repeat protein backbone must first be discovered.

A special case of repeat proteins, closed repeat structures in nature, and in particular, pseudosymmetric proteins (where a monomer adopts conformational repeat symmetry similar to a symmetric oligomer), accomplish a variety of functions tied specifically to their symmetry. Notable examples of such proteins, TIM-barrels and the beta-barrel fluorescent protein family, all use a combination of alpha-helices and beta-strands to robustly pre-structure active sites of some form [18,19]. These structures feature central pores, active sites, and otherwise structured regions, thanks to the inherent steric properties of pseudosymmetric proteins. No symmetric structural elements can lie directly on the axis: the repetition of units symmetrically precludes this geometry sterically. TIM-barrels take advantage of the central pocket for ligand specificity, while fluorescent proteins like Green Fluorescent Protein (GFP) sequester an auto-catalytic active site which generates the fluorophore for which the protein is known in the central region[20–22].

One key challenge to all novel protein design, but acutely felt in repeat protein design is the challenge of limiting intermolecular self-aggregation. The inherently repetitive nature of the core regions of an exact sequence repeat provides not only a repeating intra-molecular core, but also exactly identical geometry and residue identity displayed at the termini. As a consequence, the very design principles which create a successful hydrophobic folding nucleus which drives folding also template an ideal inter-molecular aggregation modality. A solution to approach this challenge can be found by looking to nature: the design of capping features. Natural proteins often have conserved end-cap features which serve to nucleate folding even with a large number of mutations to the repeat regions. Protein designers have used these features extensively, sometimes even largely unchanged from natural analogs [7,12,13].

Chapter 2 - **Repeat Protein Design Review**

Simply defined: a repeat protein is one whose monomeric unit has some repetitive region where either structure, sequence, or often both are heavily conserved. In nature, repeat proteins have emerged to fill niches where a combination of relatively static global geometry and highly variable local geometry is desirable. Two notable functional groups from nature are cell-surface receptors and DNA binding proteins. In both cases, a large family of these proteins are simultaneously structurally similar enough to function in the same cellular location with broadly similar collections of binding partners, while maintaining variable regions which allow for specific affinity to targets of interest.

Such broadly similar but minutely distinct features are valuable for protein designers because of their joint flexibility and designability. Often, extensive surface regions of repeat proteins can be mutated heavily, without changing a substantial percentage of the original sequence identity [8,11,13]. In this manner, natural evolutionary pressures have produced many very similar proteins derivative of a single (presumably ancient) repetitive motif, originally generated through DNA sequence duplication [18], but serving a variety of distinct but broadly related functions. High mutation rates of such proteins can yield still folded, but diversely functional biomolecules, as conserved regions of extended repeat proteins tend to be distributed throughout the sequence (localized in space by repeat symmetry, but distributed evenly in sequence space).

Other naturally occurring contexts for repeat proteins are niches where their repetitive nature allows for large scale fiber assembly from repetitive sub-components. Keratin and collagen use repetitive structure to aggregate into ordered fibrillar structures, yielding complex materials [2,3,23]. While the direct mechanism is difficult to establish, ice-binding antifreeze proteins exploit highly local and fine repetitive features to disrupt regular crystal formation [4,5,19]. Protein designers take inspiration from these natural features of repeat proteins for design of novel and function: including DNA binding, ligand binding, and other applications. Functions of other families of natural repeat proteins are enabled in other ways.

Leucine Rich Repeat (LRR) proteins in nature, across different species, display striking capability as binders of a variety of large molecule and peptide hormone substrates with great specificity and affinity [24,25]. The inherent geometric properties of LRR proteins contribute greatly to these properties: LRRs are inherently modular, with similar, but distinct repeat units. They form a concave, relatively flat surface, which offers a substantially increased molecular surface contact area over helical proteins with approximately similar dimensions. They are somewhat flexible, thanks to their repetitive, extended nature, but less so than all-helical structures (which can easily form a variety of soluble degenerate helical folds). Their usefulness is characterized by an omnipresence in a variety of species of plant, animal, and even prokaryote, though the exact evolutionary origin is unknown, and perhaps the result of independent processes converging on a useful fold [24]. For protein designers, the only downside of these modules is a scarcity in geometric diversity.

LRR-derivative proteins have been used to create small protein binding pockets [11]. In these designs, surface residues in the sheet region can be mutated extensively, altering the charge and steric contours of the molecular surface, while having nearly no effect on overall folding and stability of the original protein. A common feature of the overall topology of these proteins, like many beta-strand biomolecules, strand residues are stabilized by the hydrophobic, inward-facing residues of the sheet, pre-stabilizing a concave binding surface for potential targets [9,11,26]. Furthermore, protein designers have built LRR proteins with modular components derived from natural protein backbones and capping features [12,26]. These repeat proteins are excellent scaffolds for the particular problems which they have been used for, limited only by the imitation of a relatively small pool of natural scaffolds.

TAL effector proteins fused to nucleases (TALENs) are a perfect example of such a functionalization: with just two mutations per repeat unit, a designer can create a specific protein binder to just about any short DNA sequence, while maintaining over 90% sequence identity to the original protein, even accounting for compensatory, optimization mutations [7,8,13]. While TALENs have largely been overshadowed by CRISPR/Cas9 as the engineered nuclease of choice, the application of modular protein DNA binders remains a relevant direction in several bioengineering disciplines [27,28].

This limitation of restricted scaffold pool size, has been addressed a number of times through development of a variety of modular and semi-modular repeat scaffolds for functional applications. Notably, the Designed Helical Repeats (DHRs) from the Baker group [29,30] span geometric space previously unsampled by the whole set of repurposed natural proteins. These proteins are readily fused to functional motifs [31], adapted to structured oligomeric states [32], and functionalized to form fibrous superstructures [17]. Their all helical architecture is amenable to a variety of straightforward protein fusions on secondary structure, and a broad diversity of repeat geometry enables a wide variety of superstructures.

While these modular structures represent one of the biggest steps toward universal protein design building blocks, some problems remain unsolved. All-helical proteins have relatively compact volume per sequence length, but therefore cannot span large distances as efficiently as structures containing beta strands. Helical backbones are compatible with a variety of similar sequences, but can therefore have degenerate folded states, leading to apparent “flexibility” of structural fusions. All-helical structures with small superhelical radii cannot have twists as tight as beta-containing structures without also having large rise, because the helical secondary structure itself has a larger radius than does a beta strand, and so such architecture sterically occludes these tightly wound backbone configurations. Another gap in the designed repeat protein repertoire is the simple inherent absence of the beta strand secondary structure. While fusion to helical structures is a matter of straightforward sequence fusion, many proteins of interest have beta-strands near the termini which can be better targets for fusion in certain cases. The current best approach to surmount this obstacle is sampling and characterization of helical fusions to the beta-strand of interest, in order to create a “fusion handle” for the repeat protein sequence to extend off of.

A straightforward approach to design of a new protein with desired backbone architecture is fragment assembly of secondary structures pieces having the desired properties.

Fragment assembly has been applied successfully to symmetric and asymmetric design challenges [12,33–36]. The key advantage of fragment assembly comes from usage of backbone patterns repeated in known structures. These small, structured features represent plausible conformers for peptide backbones, and thus, increase success of design trajectories over usage of purely idealized fragments. These protocols are made more efficient by two broad geometric evaluations during backbone design. Fragments must be placed such that hydrophobic core residues can be plausibly assigned to positions between fragments and hydrogen bonding backbone residues are either solvent accessible or fully satisfied. These constraints tend to reduce sampling of beta-stranded backbones to known architectures, specifically with respect to satisfying the hydrogen bonding residues of extended strand regions. During the fragment sampling process, it must be known exactly which backbone position will pair with another to form a beta-strand pair. There have been several successful expansions of natural folds by this methodology [12,36,37]. Repeat proteins with beta strands are particularly amenable to design by this sort of constrained pair sampling because a small repeat unit can easily have beta-strand pairs mapped with *a priori* knowledge, and still be propagated to form a much larger protein.

Another, perhaps more significant, challenge is the facile and rapid generation of new repeating, modular protein scaffolds to solve emerging challenges. Existing methods are sufficiently straightforward to adapt for a new design objective, but produce soluble, folded sequences in timescales on the order of tens to hundreds of CPU hours per useful outcome. Several rounds of iteration are typically required to establish the correct collection of constraints and sequence design parameters. Modular fusion approaches [12,15,29] (where existing backbones are fused together to alter structural parameters) require a large amount of iterative testing to meet specific scientific objectives, even after protein modules are characterized and published. These approaches offer access to a variety of intriguing structures and geometries, but are ultimately too complex and involved for translation to other applications with the established methods.

In addition to extended repeat proteins, cyclic repeat proteins occupy an important niche in natural systems which protein engineers have worked to adapt into designable systems. TIM-barrel proteins have been widely investigated as targets for novel enzyme generation through recombination of repeat features [6,22]. Fully *de novo* TIM-barrel-like folds have been designed and characterized as starting points for future work. All-helical toroidal structures have been generated *de novo* and functionalized with binding domains of interest [38]. Natural functions, such as ion specific nano-pores, circular architectures, and others are promising targets for future protein designs, if a robust framework for generating mixed-fold cyclic proteins can be established.

Existing methods for repeat scaffold generation are often limited by bespoke constraint selection, the necessity for repeated design iteration, massive sampling, and careful tuning of computational metric filters. New high accuracy prediction methods, like AlphaFold2 and RosettaFold offer promise with respect to reducing these design constraints. The combination of simple random walk sequence iteration with geometric or structural constraints and machine learning derived structural prediction greatly reduces the burden of these challenges. Higher rates of designable backbones and greater rates of designed, folded proteins in vitro than

previous fragment assembly based methods show promise to rapidly provide a much more complex space of backbone geometries to apply to emerging challenges.

Chapter 3 - **ABR Design**

Design Objectives

Given the expansive nature of existing repeat protein design and the efficacy of existing methods, the primary design objective was chosen as exploration of novel geometry, both at the local level of the repeat unit and at the higher level of repeat protein symmetry.

In order to approach this objective, we chose a model fold architecture: strand-loop-helix-loop, repeating. The initial objective was an investigation of the viability of this fold as a pure repeat (at the backbone level), with the possibility of breaking symmetry at the terminal repeats using sequence alone. The first concrete design strategy was the evaluation of existing fragment assembly methods for applicability to this challenge. Both more native-like "high constraint" schemes and "exploratory" low constraint schemes were envisioned for the project. Repeat proteins would be evaluated by their general solution behavior, their stability when fused to proteins of interest, and the presence of apparent beta-character.

The design objectives evolved further to evaluate the relative efficacy of Deep Learning (DL) hallucination based methods using AF2 for backbone design and the Protein Message Passing Neural Network (MPNN) for sequence design as compared to Rosetta based methods previously successful in designing repeat proteins.

Fragment Assembly and Rosetta Fast Design

In order to construct new repeat protein architectures, we first established backbone design trajectories using Rosetta Remodel and FastDesign protocols, similar to previously published works [33–35,39,40]. These methods were based on existing methodology: both those which focused on generation of alpha-helical repeat proteins, as well as that of a more general design strategy for rational design of beta-strand containing proteins.

We modeled initial design blueprints (design models ABR 1-14) coarsely after TLR and LRR proteins with respect to fragment dimensions and repeat unit length [24]. Figure 3.1 graphically illustrates the Remodel with constraints pipeline. Subsequent design rounds sampled larger variations in helix and strand size, but restricted loop size to 2 or 3 residues. The rationale for this was multifold: final backbone scaffolds, after design and FastRelax, often saw secondary structure units unfold into adjacent loop regions slightly. Shorter initial loops resulted in a greater effective sampling pool. Additionally, the design rules from Koga et al[35] demonstrate the greater likelihood for Remodel to produce a "designable" backbone structure with loop lengths of 2 or 3 when that loop is between a helix and a strand. While early sampling trajectories did produce promising designs and one soluble sequence with interesting properties, the proportion of unsuccessful trajectories to viable design models was unacceptably high for our design objective.

Repeat unit length was sampled between 20 and 40 residues. Repeat units were sampled as poly-alanine or poly-leucine using the centroid based score function, which models

residue sidechains as single pseudo-atoms. Repeat symmetry was strictly enforced for the backbone generation process. Fragment lengths for strands were sampled between 3 and 7, for helices between 14 and 24. Distance constraints were applied between alpha-carbon atoms, modeled after distances observed in native proteins. Loops were sampled broadly in length. Successful backbone candidate models were selected based on coarse grained clash filters and a basic fragment quality check in rosetta ("worst9mer") as well as the presence of beta strands identified by protein DSSP [41] with the version incorporated into Rosetta.

Subsequent design rounds featured variations on backbone constraints as well as more tightly controlled fragment sampling, in line with previous work for asymmetric beta-stranded protein design [35]. One type of backbone constraint chosen were constraints between both alpha and beta carbons, in order to reduce twist between repeat units, and increase the proportion of trajectories resulting in strand-paired repeats. Another type of constraints applied were loose distance constraints between helical residues, in order to enforce a more compact fold. The first round of designs screened (ABR 1 - ABR 14) were chosen from a pool of over 50,000 initial backbones. Each subsequence design round was of approximately similar size, though unused backbone models were only stored temporarily during the design process and not rigorously quantified throughout the process.

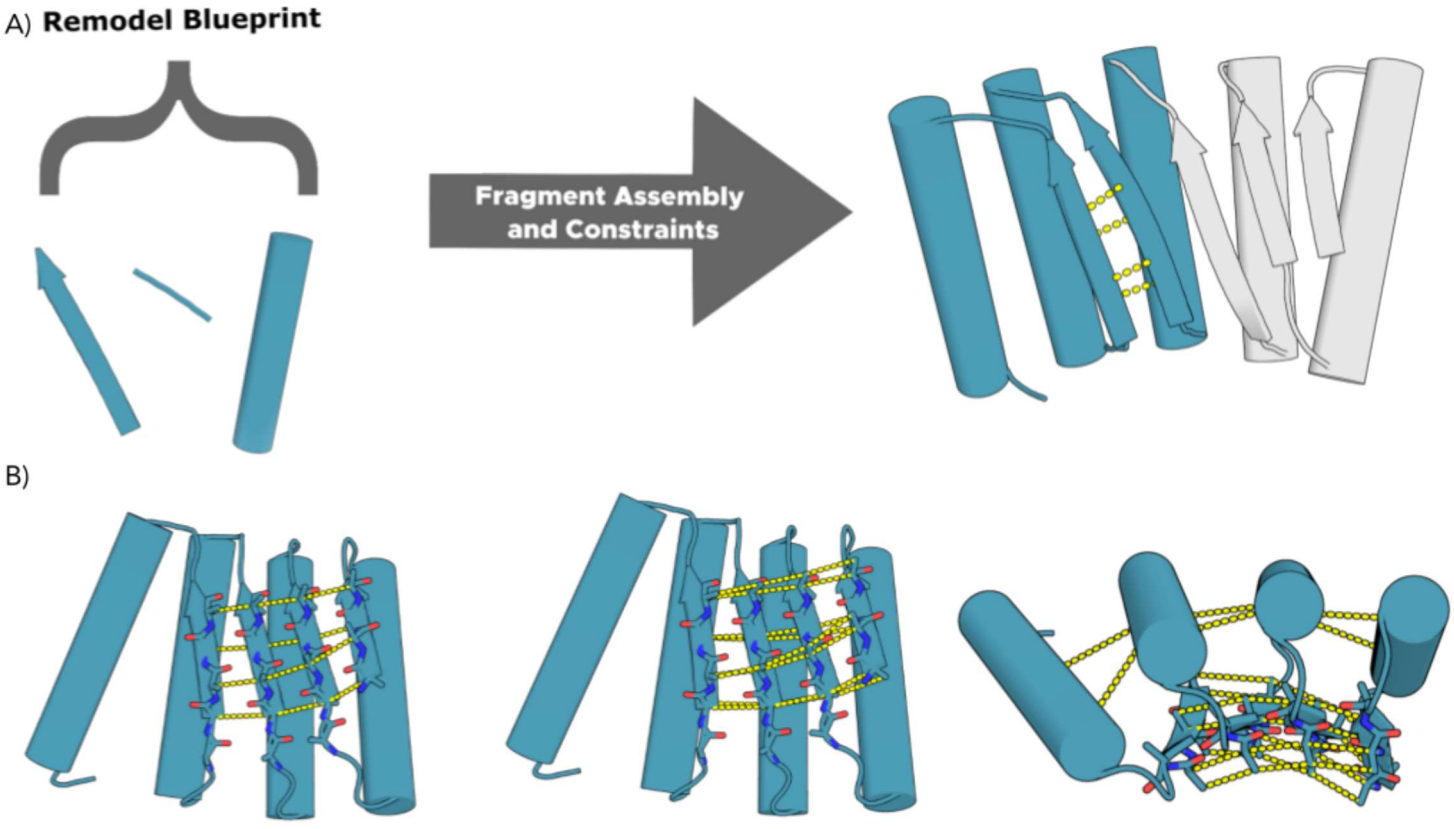


Figure 3.1 - Schematic Diagram of the AB repeat Fragment Assembly Procedure
A) Rosetta Remodel Blueprint file specifies the fragment sizes, order, and even residue types. A Fragment assembly protocol, with scoring and repeat-constraints is used to generate the backbone positions, represented as a cartoon with dashed lines between two repeat units, shown in blue. **B)** A representation of distance constraint types. On the left, only backbone alpha-carbons are constrained. In the center, alpha-carbons and beta-carbons are constrained. This constrains the twist of the sheet more strongly, since distance between beta carbons is sensitive to strand twisting. The model is rotated on the right to show the intense constraint of helix-helix distances and inter-strand constraints. Constraining inter-helix distance, even slightly, greatly constrains the search space.

Of these initial backbone models, only approximately 5% were found to have beta strands formed after fragment assembly, a rate consistent with previously published "rules-based" fragment assembly protocols [35]. Later design rounds with less variability allowed with respect to the distance constraint between both the alpha- and beta-carbons of strands meant to pair, as well as shorter allowed loop lengths, produced modestly more "designable" backbones per remodel trajectory. These design rounds had higher rates of strand pairing found after fragment assembly, but substantially less diversity in twist and rise among final models, an expected downside of increased constraint on the search space (Figure A8.2).

Deep Learning Based Backbone Hallucination

For DL design of protein backbones we conducted similar tandem repeat hallucination Markov Chain Monte Carlo (MCMC) sequence search enhanced by AlphaFold2 (AF2) prediction as previously described [42,43], but with constraint on predicted secondary structure content rather than superhelical parameters or global architecture (Fig 3.2). In brief, the trajectories consisted of a sequence space Markov Chain Monte Carlo (MCMC) optimization protocol. We specified, within each trajectory, a length L and a number N of repeating units. Trajectories were seeded with a sequence predicted (by AF2) to fold (at any confidence) to a structure with alpha-helices and beta sheets, then tandemly repeated N times. We then used a scaling mutation rate from 5 to 1 (based on how many steps had elapsed). We derived a final quality score by taking the worst value among AF2 model confidence (predicted LDDT), predicted TMScore (pTM) and adherence to a mixed ratio of alpha helices and beta sheets (dssp_loss). We finally accepted or rejected each collection of substitutions according to the standard Metropolis criterion in every step.

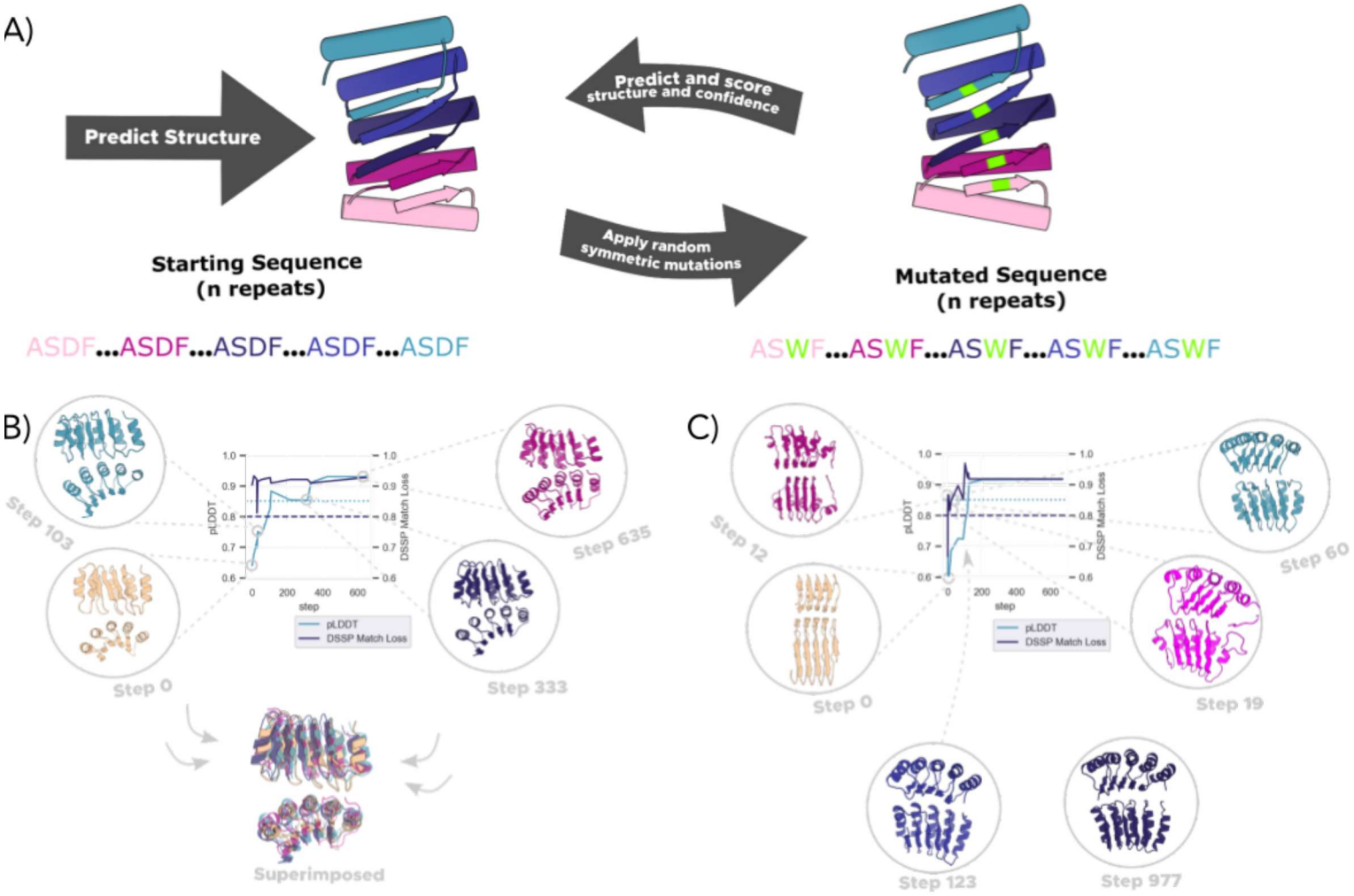


Figure 3.2 - Schematic diagram of Hallucination-based Design

A) MCMC sequence based hallucination is conducted on a tandemly repeating starting sequence. Repeat symmetric mutation sites are represented in green, and shown in corresponding regions of sequence. Structures are scored by confidence and adherence to mixed alpha and beta character. **B)** Scores of models accepted by the MCMC search on pLDDT and desired alpha/beta ratio are shown as a trace. Steps of interest are rendered as cartoons. A superimposed overlay shows changes to the structure over many steps, even while the fold remains the same. Dashed lines represent quality cutoffs for each metric, where a backbone will be chosen for design. **C)** Some trajectories started with models not matching the desired fold, but converged rapidly. Step 977 (not shown on the trace) resembles step 123 very strongly.

Sequence Design

We conducted sequence design with RosettaFast Design and ProteinMPNN. For Rosetta FastDesign, we used strict sequence symmetry for core residue positions and unconstrained sequence design for solvent accessible regions. We conducted MPNN design by iterating between MPNN sequence design, Rosetta FastRelax, and AF2 prediction. We evaluated MPNN designed sequences and determined promising targets to characterize through prediction by AF2 and pLDDT, pTM, further narrowing the pool by selecting the models with the lowest SAP score, lowest mismatch probability to psipred predicted structure, and best rosetta sidechain shape complementarity scores, as well as other metrics detailed in Appendix 6. We evaluated models designed exclusively with Rosetta by these metrics, as well as the best fragment quality (worst9mer) and lowest number of buried polar residues in the design models.

We retrospectively analyzed AF2 scores between Rosetta-only designs and DL-design assisted models (AF2 was not initially available for earlier design trajectories) and found that higher degrees of constraint tended to improve RMSD of the design model to prediction model as well as improving pLDDT. Sequence design methods and rationale are further described in detail in Appendix 3. MPNN sequence quality was primarily evaluated by best AF2 prediction pLDDT for each sequence, and sub-selected by lowest SAP score.

Capping Feature Design

We reasoned we could improve our success rates through designed helical capping features. Native beta-strand-containing proteins almost always have such caps, and rapidly folding helix-containing segments could potentially block slower forming intermolecular strand-strand interactions. One of our design models (ABR10) was found to have a reversible, concentration dependent soluble aggregation mode and displayed more obviously beta character signal on CD at high concentrations (Appendix 9). We hypothesized that if the transition from a folded mixed fold to a lower energy solenoid or amyloid aggregate was driving aggregation of some of our designs, helical, low contact order, capping features may work as a kinetic barrier to prevent this transition.

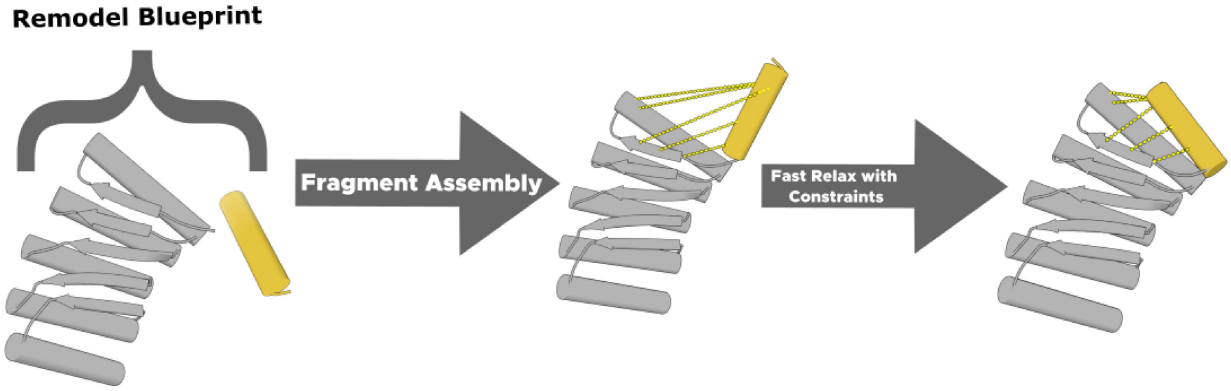


Figure 3.3 - Cap Design Scheme
A variety of cap sizes and loop lengths were sampled through rosetta remodel, and then FastRelaxed under distance constraints to a rigid body model of the protein.

Helical capping features were designed with Rosetta remodel, sequence design was conducted with distance constraints anchoring the alpha-helical capping features in place at the termini of the repetitive region. Constraints were enforced between the cap terminus and the nearest residue in the proximal repeat unit after each cycle of Rosetta FastRelax. Quality of capping was determined by changes in the Surface Aggregation Propensity (SAP) score measured with and without the cap in the final models. After AF2 prediction of capped proteins sequence designed at the cap interface with RosettaFastRelax, many models showed great deviation from the original design. Serendipitously, the development of the Protein Message Passing Neural Network (ProteinMPNN, or MPNN) [44] allowed for an alternative strategy for sequence design, where only backbone sampling was required to test many sequences. Iteration of MPNN, FastRelax, and AlphaFold2 for 3 cycles was used to design final sequences.

Fusion to Designed Heterodimers

Following the successful design, expression and characterization of the base repeat proteins, we looked to dock and fuse them along the edge strands to functional beta strand containing proteins. We chose the Large Heterodimers (LHDs) previously developed in our group as a target with immediate relevancy and relative modularity. These proteins have been fused to DHRs previously, though this process is limited by location of helices on the LHD subunit, and results in fusions which twist dramatically away from the plane along which the molecules dimerize (Fig 4.7). This was an opportunity to demonstrate the advantage of using different repeat architectures to enhance the robustness of the repeat protein “building block” space. We docked our new models along terminal strands with a simple strand template alignment script and designed backbone connections with Protein Inpainting [45]. We then designed these models as described above with MPNN, AF2, and FastRelax: conserving the original heterodimer interface, but allowing mutations to accommodate the new fusion.

Chapter 4 - **ABR Findings**

Expression, Purification and Characterization

We conducted several iterations of design and characterization: a total of 85 proteins were screened from the initial Rosetta-based protocol and an additional 47 from the DL-Design protocol. Proteins were named ABR1-85 and 86-132 respectively. All DL-designed designed models featured both an N- and C-terminal helical cap.

Genes were obtained, cloned, transformed into E. coli, and proteins were expressed with autoinduction. Expression cultures were lysed by sonication and purified through immobilized metal affinity chromatography (IMAC) and size exclusion chromatography (SEC) via fast liquid protein chromatography (FPLC). Circular Dichroism (CD) spectra were collected and thermal melt scans conducted for models with promising solubility characteristics and SEC peak character.

Further details of expression, characterization and additional data collected are presented in Appendix 4.

Rosetta-Designed Models

The vast majority of Rosetta designed models were either not present in the soluble fraction of lysate or did not remain soluble overnight after IMAC but before SEC. Rosetta-based designs which displayed promising behavior and their respective SEC traces are presented in figure 4.1. Models ABR 10, ABR 34, and ABR 35 were also screened for crystallography. While crystals were obtained for ABR 10 and ABR 35, diffraction data were not obtainable from the experiment. SEC traces as well as discussion of the scores, design models, and corresponding AF2 predictions for these design models are discussed in greater detail in Appendix 5.

Of particular interest is ABR 10, which displayed a reversible concentration dependent aggregation mode characterized by SEC. This protein also showed a concentration dependent shift in its CD spectrum. ABR 10 was easily concentrated to well above 50 mg/ml, and remained soluble at room temperature for well over a year, though at a lower concentration. These data are presented in Appendix 9.

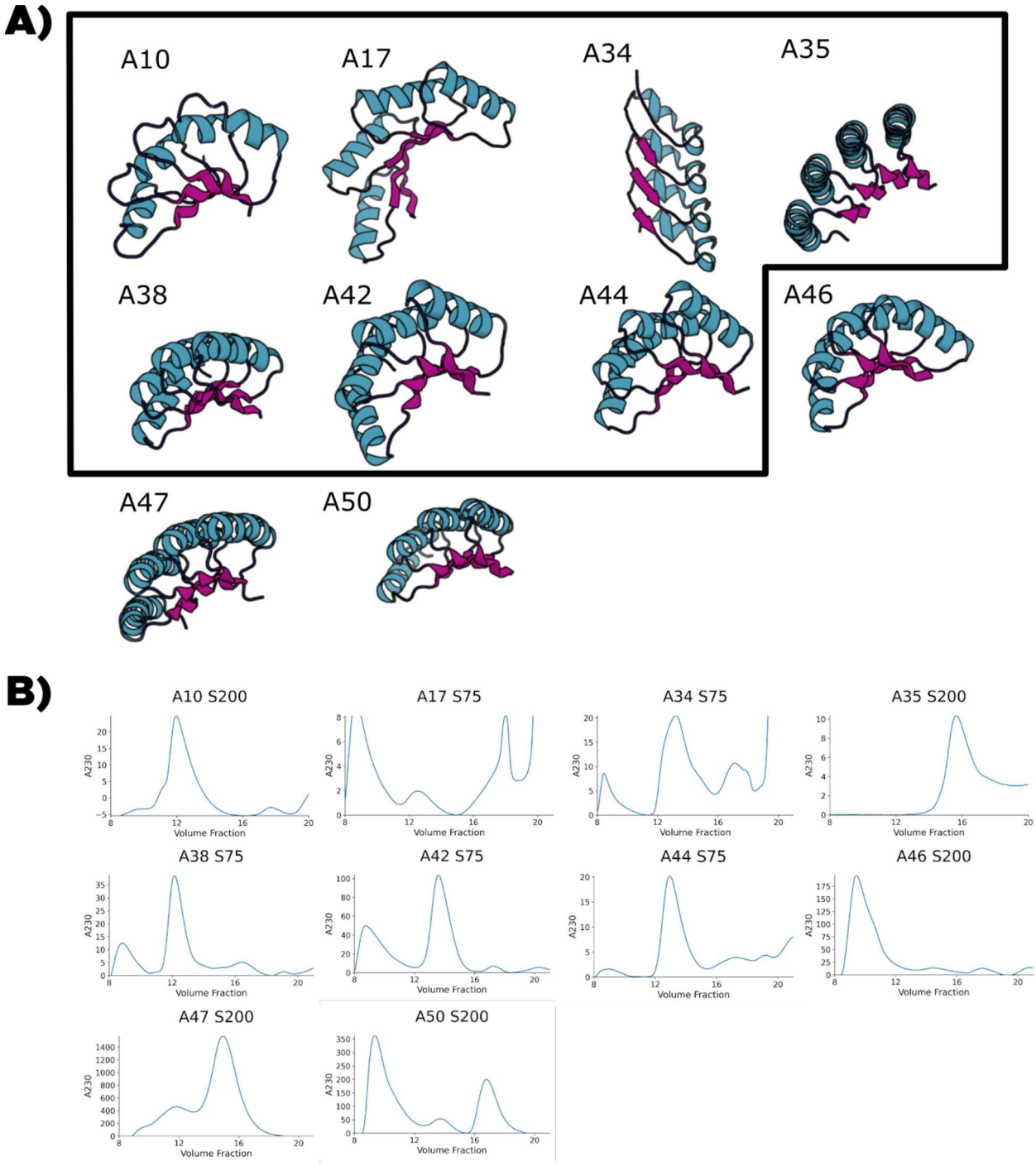


Figure 4.1 Rosetta Designed Models and SEC traces

A) Design models. Ca-carbon constraint-only models shown outlined in black, others shown outside border. **B)** SEC traces for all models shown. Note the scale on the y-axis is adjusted so the maximum is thresholded to the highest peak on the interval between 9 and 19 ml retention volume; where protein of interest and degradation products are expected. Yields outside soluble aggregate range were poor except for A47 and 50.

DL-Designed Models

A substantially larger proportion of DL-designed models were found to be soluble and stable enough for SEC than models designed exclusively with Rosetta. Clearest and best representative models are shown in Figure 4.2. Two key differences separate the Rosetta design pipeline from the DL design pipeline. While both collections of models had surface positions designed with Rosetta FastDesign, the remainder of residue positions in the DL designed models were designed with MPNN. Additionally, explicitly designed capping appears to have played a part in greatly improving overall properties of these proteins.

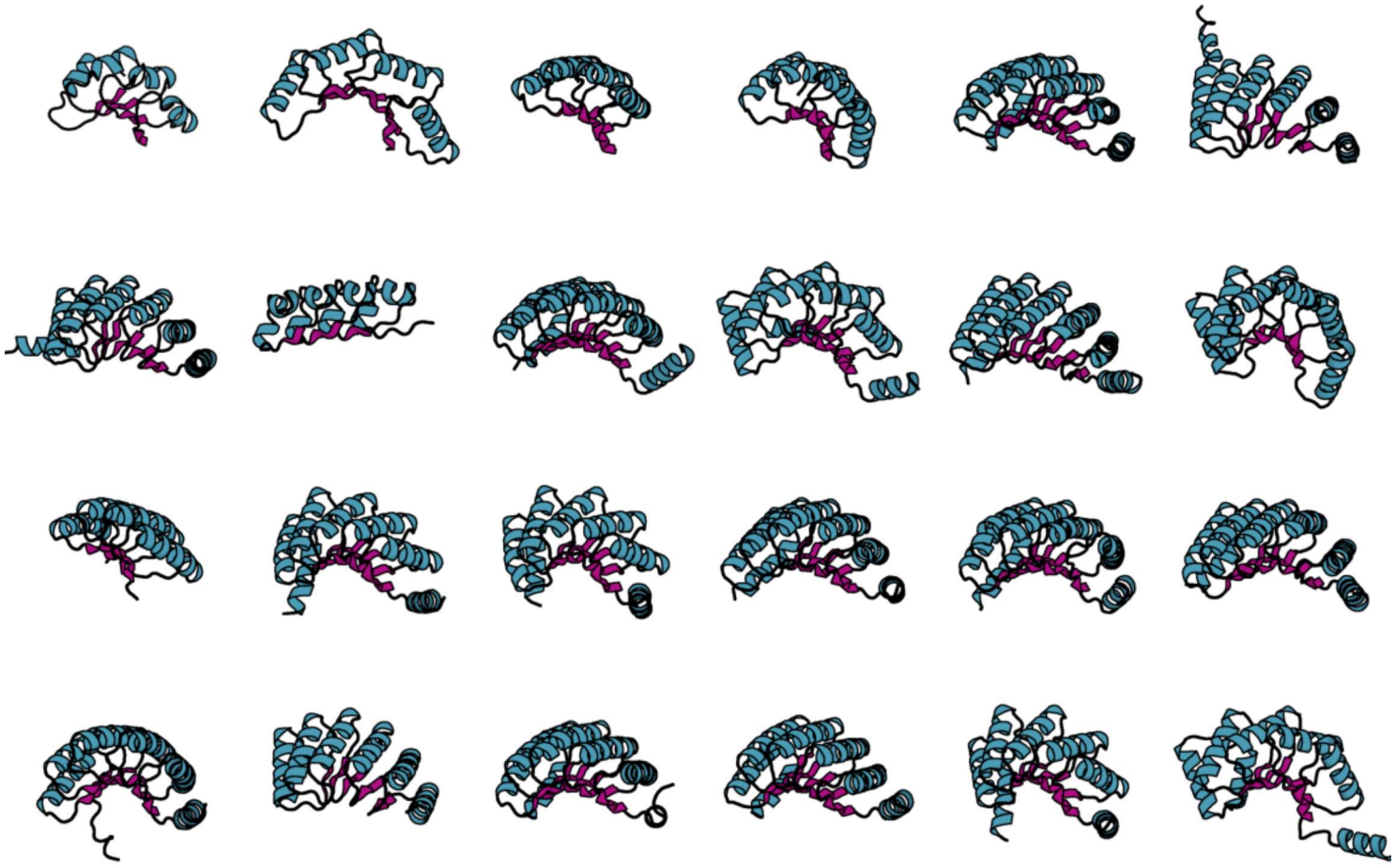


Figure 4.2 - Diversity of soluble ABRs

Selected ABR models are roughly aligned to their superhelical repeat axis to demonstrate their diversity in rise, twist and radius. loop, sheet, and helix regions are colored in dark blue, magenta, and teal, respectively.

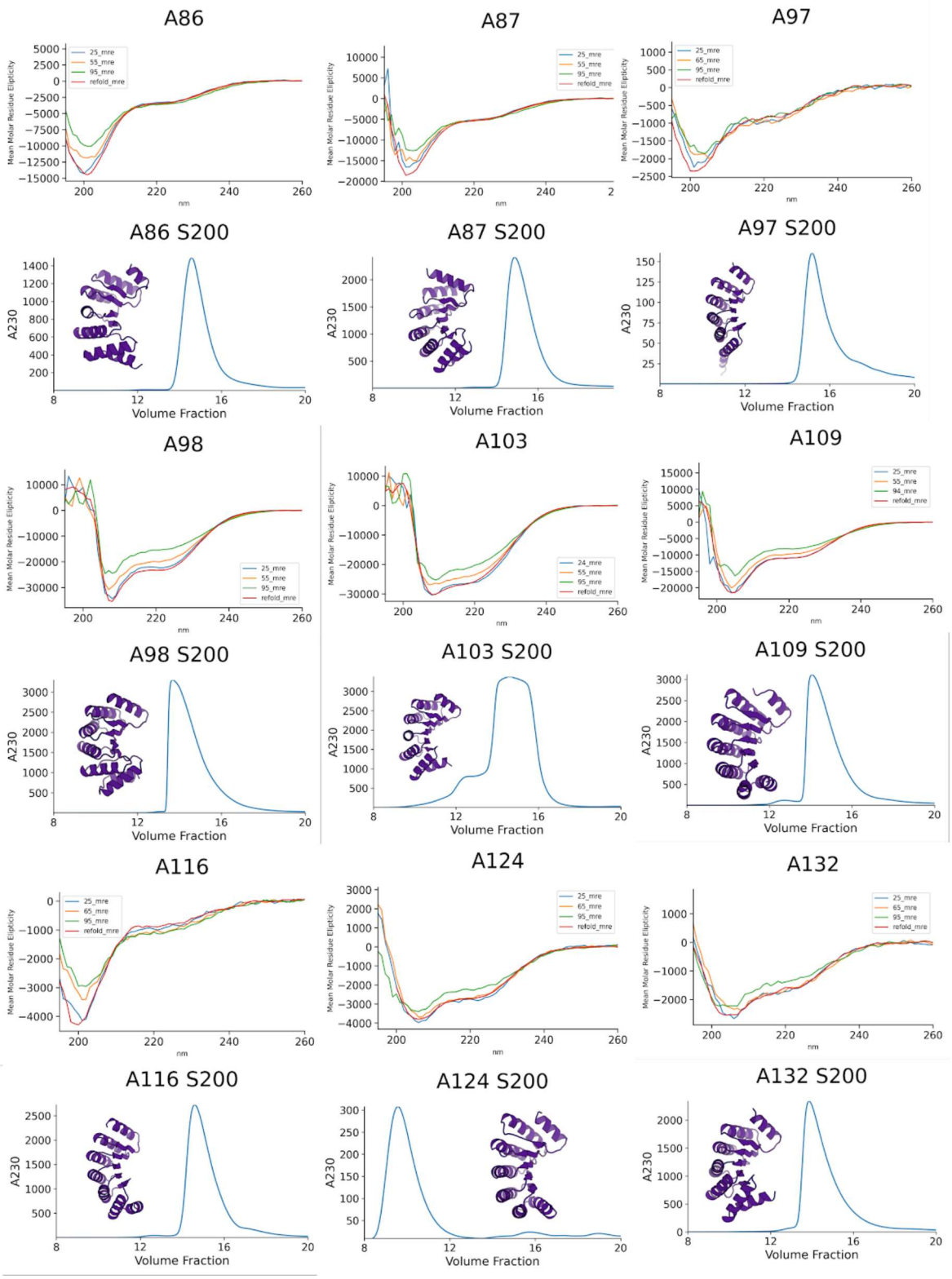


Figure 4.3 SEC traces and CD melts of DL designed ABRs

SEC traces were obtained by collecting fractions from the largest peak in an initial purification S75 SEC run, and re-purifying again with an S200 column for increased resolution. CD melts include a scan, in red, at 25C after cooling. Traces in blue, orange, and green represent 25C, 55C, and 95C scans respectively, except where 65C was the intermediate temperature recorded, as noted. CD units are mean residue elipticity

Advantages of DL Hallucination

Introduction of deep learning hallucination to the design of ABRs overcame a number of challenges which produced more high quality models, introduced more diversity to the models produced, and resulted in a greater proportion of proteins passing characterization screens. Initial approaches to fragment assembly for backbone design produced a number of proteins present in soluble fraction after lysis and even a few with significant peaks at approximately the expected retention volume on SEC. Only a small proportion of the models tested displayed these favorable characteristics, summarized in Table 1. Many proteins displayed slow aggregation or were absent in soluble fraction of culture lysate entirely.

A low success rate in solubility and aggregation screens does not independently present an insurmountable challenge for design of such proteins, but, combined with the quantity of sampling required to generate the models, represented a significant obstacle to design. We sampled tens of thousands of backbones per backbone design trajectory. Between all design rounds and constraint schemes, this totalled an estimated low millions of backbone geometries sampled to produce under one hundred final models. Furthermore, sequence design expanded the search space many-fold, producing dozens of fully designed models per input backbone, with RMSD deviation from the original model ranging to values above 2.

The primary failure modes of these fully designed models were detectable only at the sequence design step. Many backbone models were simply incompatible with any sequence we sampled, primarily because of buried unsaturated polar heavy atoms. While some backbones had positions which were incompatible both with hydrophobic residues and polar ones, others had backbone positions with buried amide bonds and no means of electrostatically satisfying the dipole on the polar nitrogen atom through sequence design, even while employing backbone-backbone geometric evaluation metrics, such as MotifHash (a specialized coarse-grained backbone quality metric in Rosetta). Because the designation of surface, core, and boundary are determined by sidechain density and atomic depth, these problematic backbone positions can only be identified after a full round of sequence design. Differences in solubility according to Rosetta metrics are further evaluated in Appendix 6.

Strikingly, while insoluble Rosetta design models predicted with AF2 showed more unsaturated polar atoms, this difference was not replicated in DL designed models. DL designed models, even ones with excellent solubility behavior feature atoms marked as buried and unsaturated in Rosetta. We propose that this may be a function of training data featuring implicit presence of solvent and small molecules in crystal structures, while Rosetta models can only evaluate a very broad implicit solvent term. Further investigation of this complex question is beyond the scope of these findings.

Constraints on the backbone design protocol and optimization to the sequence design protocol were able to produce larger pools of models with better overall quality metrics. These improvements restricted the design space to the specific constrained geometries dictated. This tradeoff produces more viable models within the geometric space explored, but reduces the search radius overall compared to an unconstrained strategy.

AF2 prediction of Rosetta Designed models demonstrates a clear trend. Few of the Rosetta based design models reach the minimum pLDDT cutoff required for the hallucination models (Figure 4.4). The number of models passing initial screening increases dramatically for DL designed models. Models with greater constraints appear to have lower RMSD to their AF2 predicted backbones, and, combined with higher pLDDT and higher likelihood to pass screens, support that computational constraints during the design process may correspond to a more well-folded geometry in vitro.

The overall runtime of the Rosetta based protocol varies highly based on scaffold properties and speed of MCMC convergence. A start to finish trajectory of backbone hallucination and MPNN sequence design corresponds roughly to the runtime of a Rosetta trajectory with the same output. The main advantage of the DL design trajectories are their incredible diversity and much higher quality, which reduces the number of sampling trajectories required to find a viable model geometry. Under 500 backbone design trajectories were run to generate all the DL design ABRs, with approximately 1500 backbones chosen for further sequence design, compared to the millions of backbones generated via Rosetta Remodel.

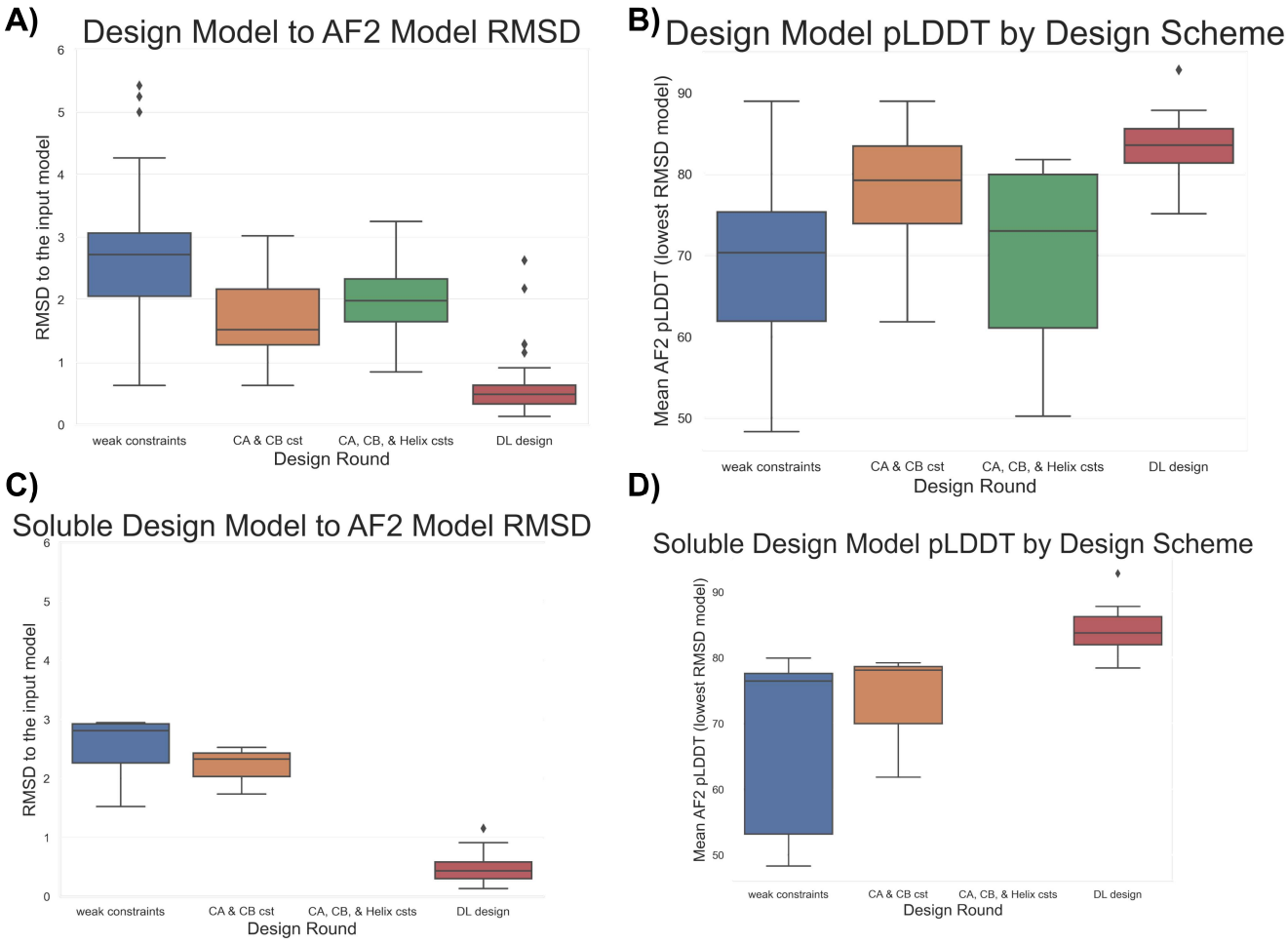


Figure 4.4 AF2 Predictions of different design Streams

A) box plots of the RMSD from design model to the AF2 prediction of that sequence. One outlier at 12 RMSD is omitted from display for clarity, but included in calculations. The mean RMSD of Ca & Cb cst model pool was found to be significantly different than that of the Ca-only cst pool at the 0.001 level. **B)** Box plots pLDDT of the lowest RMSD prediction for each design scheme. Ca & Cb mean was found to be different from the weak constraints mean at the 0.001 level. **C)** Plots of only the soluble models from **A).** Means were not found to be significantly different between constraint groups. No models from the three-constraint group were found to be soluble. **D)** Plots of only soluble models from **B).** Means were not found to be significantly different between constraint groups for pLDDT. Boxes show mean and interquartile range, whiskers show population maximums and minimums except outliers. Outliers are plotted as points except for the cropped outlier in **A).**

Contrasting Solubility properties of different design approaches

Though the design process was made more efficient with respect to the number of CPU-hours spent to generate a promising model pool, constraints were not universally successful in producing a greater number of successful models. As summarized above, well-chosen constraints tended to bias the design procedure to higher quality models (albeit with lower diversity) by pLDDT and Rosetta metrics, but did not result in a greater proportion of these top models showing desirable solubility behavior when investigated in vitro. In contrast, the capped DL-designed approach generated a substantially higher proportion of highly soluble constructs.

Design Type	Count Highly soluble	Fraction Highly Soluble	Total Tested
Weak Constraints	7	0.16	45
CA & CB CST	3	0.14	21
CA,CB & Helix CST	0	0	19
DL Design	25	0.425	47

Table 4.1 - Successful Design Hits by Backbone Generation Protocol
The fraction of models characterized as highly soluble as assayed by distinct peak with no shoulder on SEC, at a retention volume outside the column void volume. DL design produced a dramatically higher success rate among investigated models.

Geometric properties of ABRs

We evaluated the superhelical parameters of the ABR design models found to be soluble compared to the parameters found for published DHR models. ABR repeat parameters were found to occupy wholly different space. This is not entirely unexpected. Helices have a larger steric occlusion radius than strands, and consist of more densely packed sequences of residues. These geometric properties produce ABRs which occupy similar twist and radius ranges as DHRs, but with substantially smaller rise. The unique geometric niches of ABRs vs DHRs are summarized graphically in Figure 4.5, with additional data presented in Fig A8.1. ABRs have, across the board, shorter repeat units than DHRs, meaning that subject to design constraints in which a compact, short, fusion is needed, ABRs fill the role with which DHRs are incompatible.

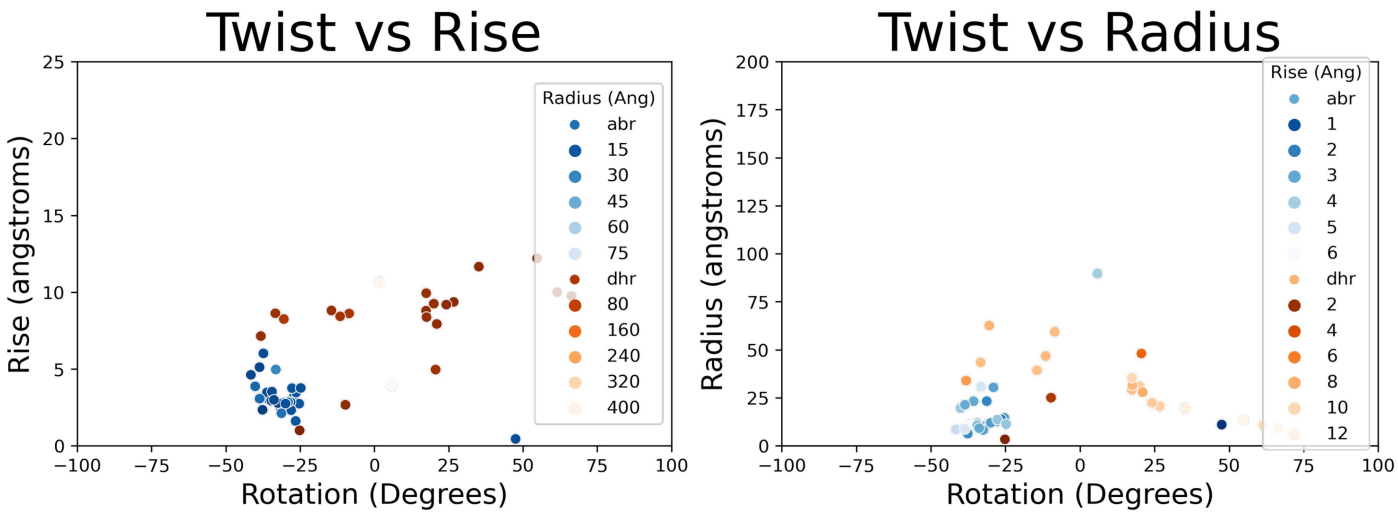


Figure 4.5 Super-helical Geometry of ABRs vs DHRs

The unique geometry of soluble ABR designs contrasted against well-characterized DHR models. WHile DHR models cover an impressive degree of super-helical geometric space, ABR proteins offer the ability to make “tight turns” (low rise, low radius, high twist).

Capping Effects

In order to identify the joint influence of helical capping and MPNN sequence design on model quality, we performed deletions of the capping regions, as well as compensatory mutations. As a proxy for model quality, we used protein yield of the primary peak on SEC for each model. Of the six base designs tested, 5 performed best with at least one helical cap (Figure 4.6).

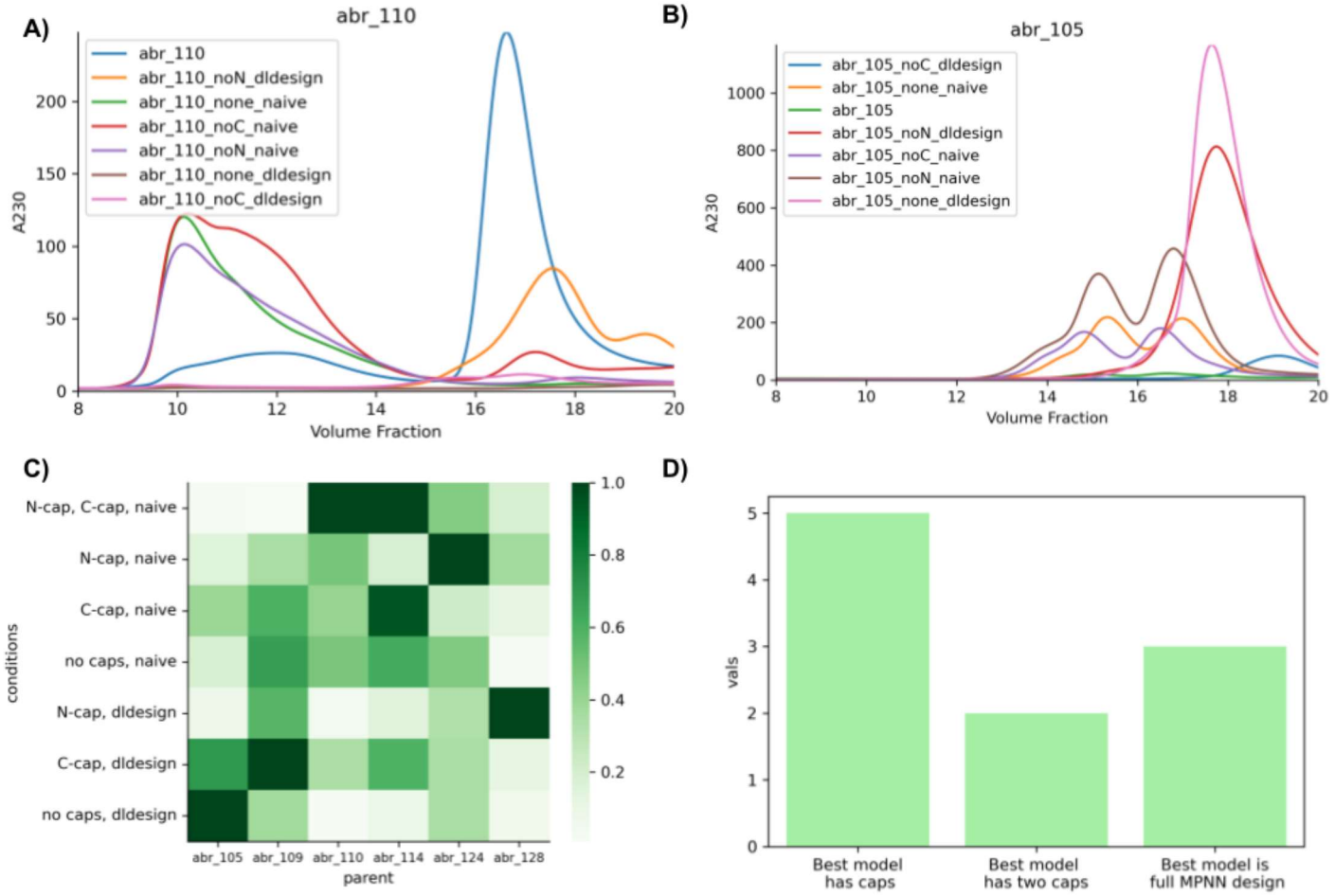


Fig 4.6 - Inspection of cap deletions on SEC behavior.
Best peak on SEC was chosen by eliminating broad peaks, peaks with shoulders, and favoring higher expression over lower expression. Peaks eluting far from the expected retention volume were also eliminated. Example SEC traces after IMAC purification for A110 and A105 respectively are shown in A) and B). Aggregated data are show in in C), where the maximum peak height is represented as a ratio to the strongest peak in that group. Naive represents the original sequence, designed with an early version of MPNN, as well as Rosetta redesign of surface residues. In all cases but one, the best model contained at least one terminal helix cap, and in many cases the N- and C-capped version was the best model. Of the five designs where capped versions performed best, three were fully sequence designed with MPNN, while two were the original, doubly capped hybrid design scheme. The only design where the best model had no caps was A105, though a singly capped design was shown to perform almost as well. Interestingly, ABR 109 was the lowest yield sequence originally, but deleting one or both caps dramatically improved solution behavior, both with and without DLdesign, implying that some repetitive model geometries are fully designable without explicit end caps.

The model which performed best with caps deleted was fully MPNN designed after deletion. Capless models were generally not among the top yielding models, and generally had lower yields and larger soluble aggregate peaks than the corresponding capped models.

The impact of simple helical fusion caps offers an additional benefit for future design. Existing DHRs can only be structurally fused directly to helical regions of proteins of interest.

Helically capped ABRs offer the possibility to fuse to either strand or helix directly, as helical caps are designed structurally in place. This background investigation also shows the sufficiency of a single capping feature; these repeat architectures can often be sufficient to reduce aggregations and improve yields.

LHD fusion constructs

As a preliminary application of the ABRs to a fusion challenge, we conducted trial fusion of ABRs to previously established LHD proteins. Two of the designed pairs were notably functional on an SEC shift screen, with depletion of the independent monomer peaks. The retention volume observed corresponds to an appropriate heterodimer retention volume, but not to a larger order species. This implies that fusion has not disrupted the original binding interface, though further characterization is necessary to concretely establish the structure of the complex.

Previous fusions to LHDs were restricted to pairs with available helix-loop-helix regions at a terminus, distant in structure from the hetero-dimer interface region. These constructs show proof of concept of fusion of an extensible repeat unit along a strand region: connected with a hallucinated piece of structure and locally sequence optimized. While such facile and hyper-modular approaches to protein architecture were proposed as the natural outcome of de novo design, these models are straightforward demonstration of the simplicity and efficacy which DL hallucination provides previously laborious and sampling intensive design approaches.

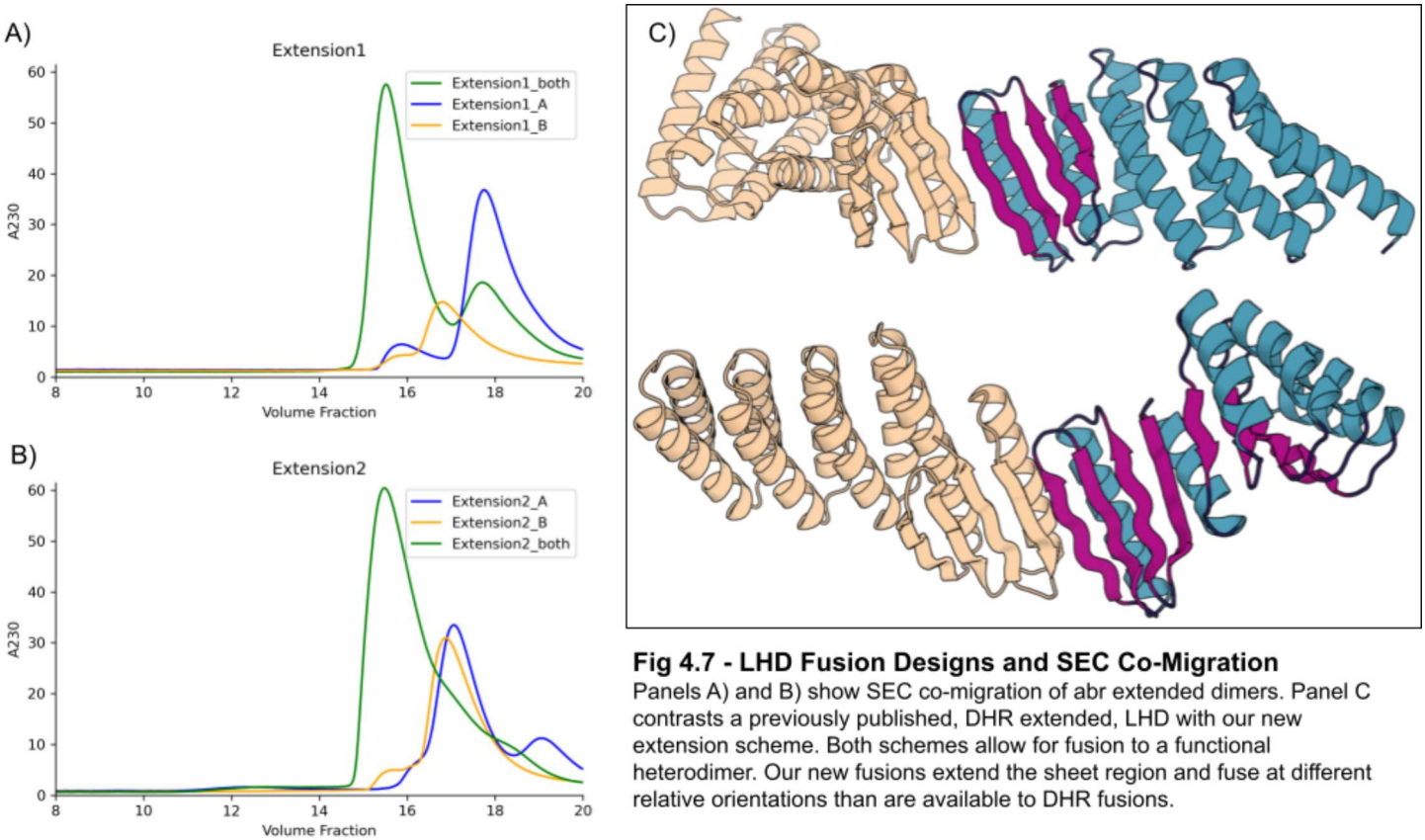


Fig 4.7 - LHD Fusion Designs and SEC Co-Migration
Panels A) and B) show SEC co-migration of abr extended dimers. Panel C contrasts a previously published, DHR extended, LHD with our new extension scheme. Both schemes allow for fusion to a functional heterodimer. Our new fusions extend the sheet region and fuse at different relative orientations than are available to DHR fusions.

Chapter 5 - **Pseudocycle Design**

Introduction

While open form repeat proteins are supremely useful as “building blocks” for connecting superstructures or binding to large targets, closed cyclic repeats can serve other functions. Geometrically, such proteins are essentially a special case of super helical or screw axis repeat parameters, where the superhelical rise of the repeat structure is zero, and the rotation about the superhelical axis, or “twist” is equal to the unit fraction represented by 360 degrees over N (where N is the number of repeat units). The radius can be left unconstrained. These structures feature an inherent geometric advantage over nominally similar globular proteins in some respects. We call these monomeric pseudosymmetric designs “pseudocycles”.

The pseudocycle model design strategy avoids the problem of capping terminal repeats: terminal repeats self-contact from N-terminus to C-terminus intramolecularly inside the cyclic monomer. Terminal fusion design of such structures is often less straightforward (requiring extensive redesign of the protein of interest), though other design advantages exist specific to the topology.

True cyclic proteins have a fold which repeats about a central axis. By virtue of steric exclusion, no structural elements of a cyclic protein lie exactly on the axis: such a feature would clash with corresponding units from other symmetric repeats located, by definition, equidistant from this axis. Our pseudocyclic proteins have some deviations from the precise symmetry of purely cyclic proteins (deviations which may be a design advantage with respect to installation of binding cavities), but they all feature a central cavity of some sort. These vary in dimension, size, and solvent accessibility based on a variety of factors. Some cavities are composed of single-angstrom “voids” in the modeled structure, and others, such as those within beta barrels, comprising substantially larger cavities for solvent accessibility or ligand docking.

Design Objectives

One design objective of the pseudocycle framework was to establish the efficacy and diversity of space explored through repeat sequence sampling hallucination. Another primary objective was the design of compact monomeric proteins with structured pocket regions. Secondary objectives were the evaluation of the general efficacy of the MPNN sequence prediction with backbones not observed in the training set (newly generated backbones with little or no structural homology to known natural proteins) as well as paired sequence and structure quality of hallucinated backbones of this type.

Backbone Design

In order to design closed repeat proteins of this form, we applied our MCMC sequence based hallucination protocol with an additional “cyclicity” score. This score represented a mean between the deviation from ideal rise and twist of the repeat units (considered as rigid bodies). This score was computed alongside pLDDT and pTM, and used as part of the overall loss.

Because of non-ideality in the predicted models, the superhelical parameters were computed from an averaged rigid body transform between subunits. Each transform between corresponding residue pairs was reduced to a rotation part and a translation part. The rotation parts were converted to normalized quaternions and approximately averaged, while a simple mean was taken for the translation part. This approximate average transform between subunits was then decomposed to a cross product matrix from which the appropriate Rodrigues rotation vector was derived. This allowed for the computation of the relative rotation about, translation along, and radius of the screw axis representation of the transform. These parameters correspond to twist, rise, and superhelical radius of a rigid body representation of each subunit of a repeat protein.

A graphical summary of this design process is outlined in Figure 5.1.

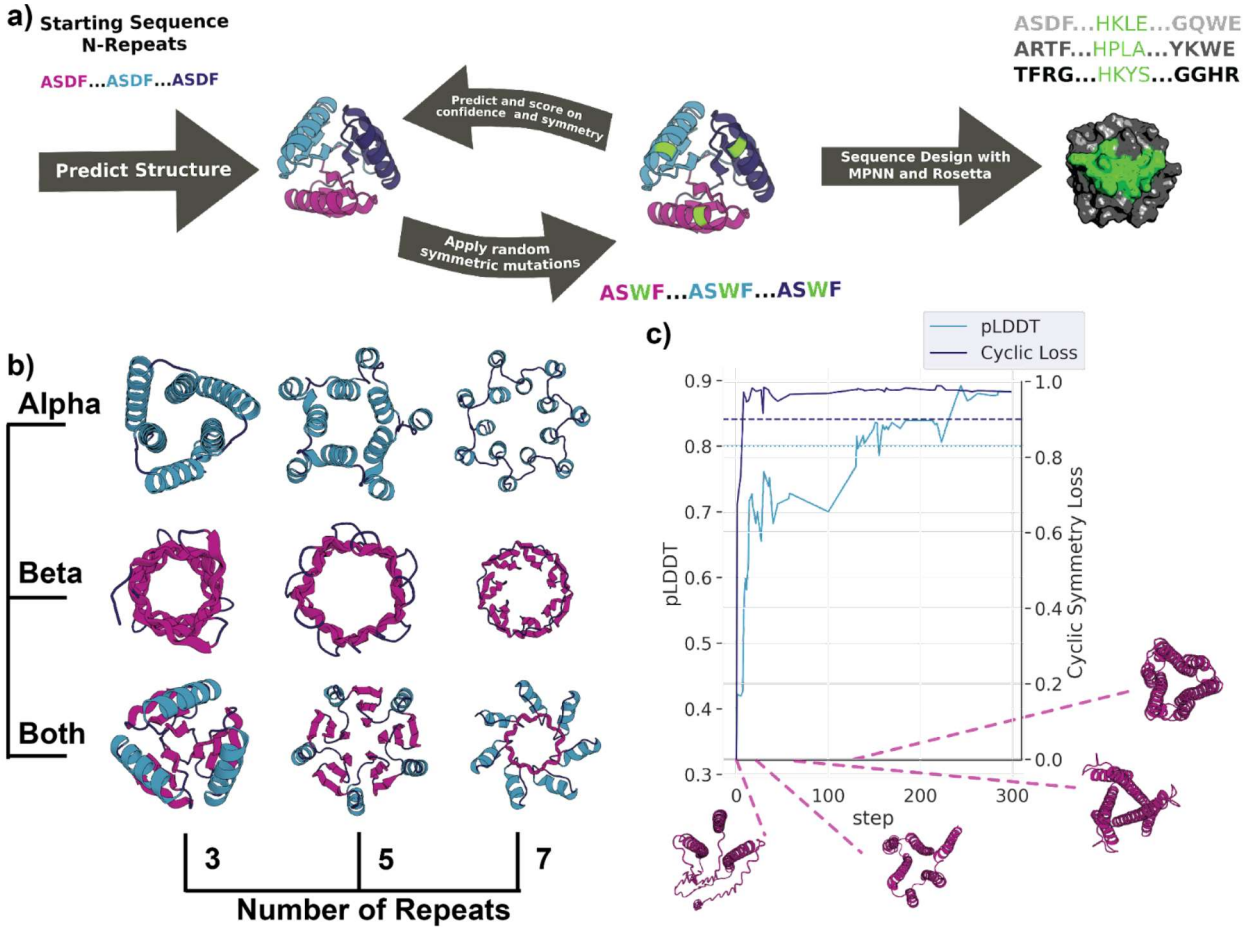


Figure 5.1 - Pseudocyclic protein design

a) Schematic representation of the scaffold hallucination and design pipeline. **b)** Selected output proteins featuring 3, 5 or 7 repeats, and all-α (Alpha), all-β (Beta) or mixed α/β (Both) topologies. **c)** Representative design trajectory showing the optimization of pLDDT (teal) and cyclic loss (dark blue) over 300 steps, with dashed lines indicating our selected score cutoffs. Protein structure cartoons are snapshots at indicated steps in the trajectory; loop, sheet, and helix regions are colored in dark blue, magenta, and teal, respectively.

Introducing a direct loss on helical parameters effectively reduced the median time to convergence while greatly increasing the proportion of trajectories successful in finding sequences predicted to form cycles with pLDDT and pTM within a preliminary cutoff range.

To reduce redundancy of the roughly 20,000 backbones obtained through this search, we computed an all by all matrix of TMalign scores between the models. We were able to reduce the population to approximately 9800 clusters, subject to the constraint of attempting to minimize the number of singleton clusters without increasing the intra-cluster TMscores.

In order to investigate whether we had reached diminishing returns in our search, we evaluated the model redundancy of several random subsamples chosen from representative models for each cluster. We found that as we increased the number of models drawn, the number of models which were totally unrelated to every other structure dropped dramatically. These findings are presented in Appendix 2. This leads us to conclude that we have sampled a large proportion of the space available with this specific methodology.

Sequence Design of Hallucinated Closed Repeat Proteins

We found that the sequences and structures produced by the hallucination had a large proportion of solvent accessible surface hydrophobic residues. In order to improve the sequence and structural properties of the design models, we applied protein MPNN. In order to enhance sampling and validate model quality, we employed a design protocol which cycled MPNN, FastRelax, and AF2, as with the open form repeat proteins. We conducted design trajectories for every representative model in the clusters we generated from hallucination outputs. Models which deviated by more than 2 angstroms in backbone RMSD from the input were discarded.

Of the final design models, we selected 96 for further characterization. A subset of these models were chosen based on a percentile rank cutoff for SAP score, sidechain shape complementarity, further subdivided by other metrics used for the ABR designs, and others were chosen based on novelty of the fold and potential impact. All models chosen were predicted by AF2 with pLDDT of at least 0.85 and a pTMscore of at least 0.75.

Chapter 6 - **Pseudocycles Findings**

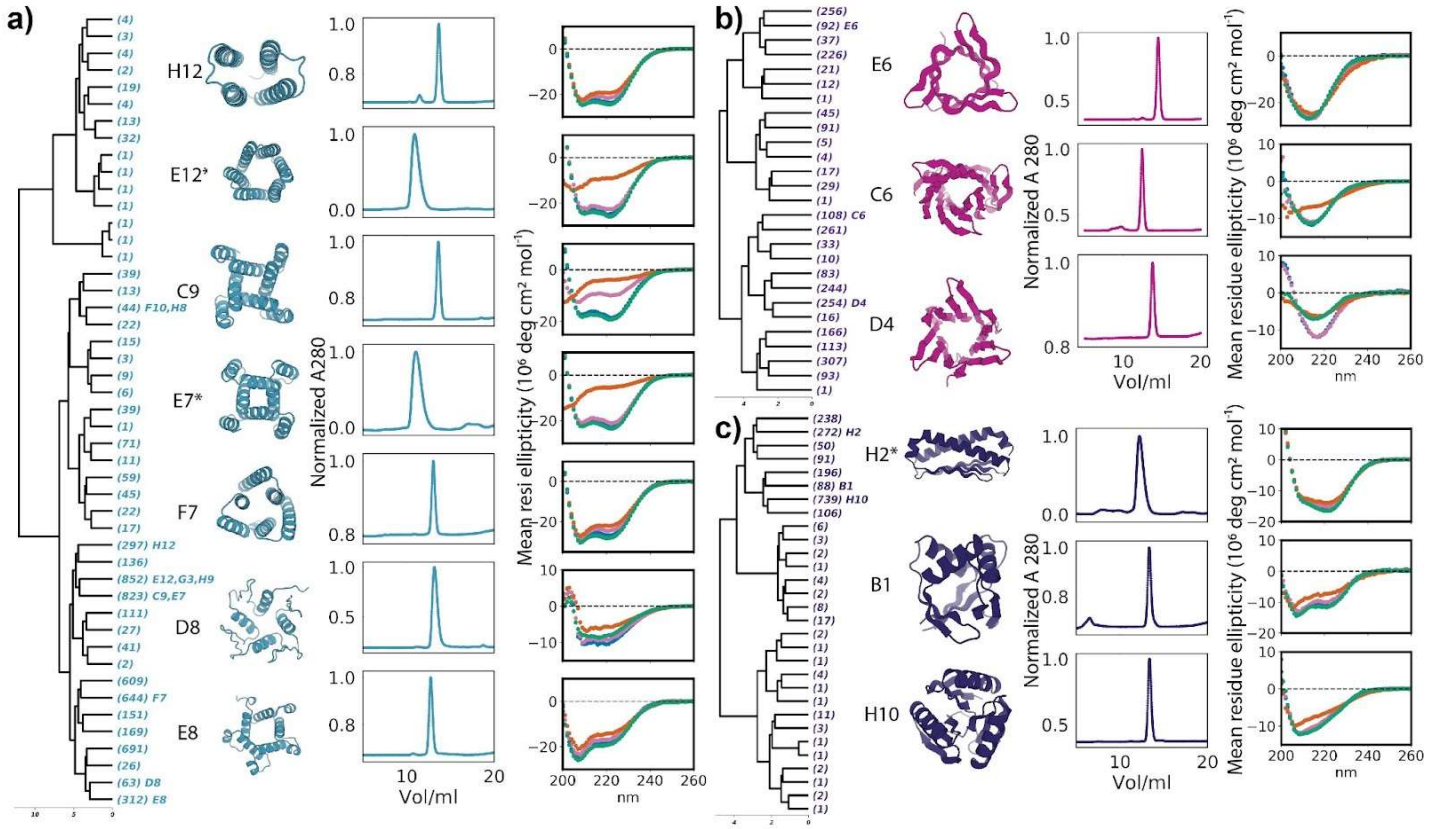


Figure 6.1 - **Biophysical characterization.**

Left panel in a, b, c: hierarchical clustering of designed pseudocycles. The number of sub-branches are indicated in brackets. 2nd panel: cartoon diagrams of designs selected for experimental characterization; identifiers indicate position in dendrograms. Third panel: SEC trace. Protocols are described in the supplementary methods, expression and purification of selected proteins. Proteins prepared following protocol1 are marked with star(*). Fourth panel: CD spectra at different temperatures (25 ℃ in blue, 55 ℃ in orange, 95 ℃ in pink, followed by refolding at 25 ℃ in green). a) α helical topologies (colored teal), b) β sheet topologies (colored magenta), c) mixed α/β topologies (colored dark blue).

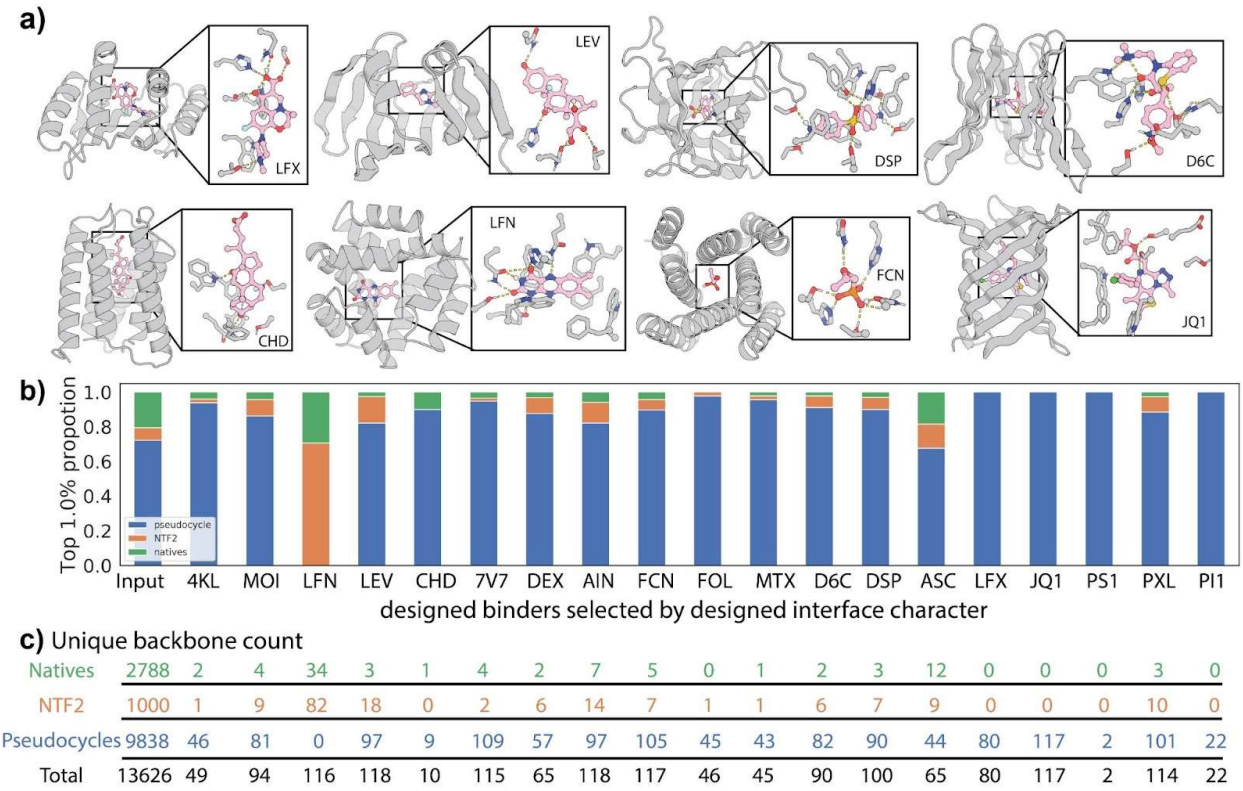
Expression, Purification and Characterization of Closed Repeat Proteins

The designs we selected included a wide spread of repeat number, radius, unit length, secondary structure properties, and dimensions all having little identity to folds observed in the Protein Data Bank (PDB) (Figure A2.5). We found these proteins to have divergent structures from those found in the PDB, with a median BLAST e-value of 0.018 and TMscores observed between 0.33 - 0.87 (Figure A2.6). We found that following expression in Escherecia coli 81 of the 96 designs were identifiable by SDS-PAGE in the soluble fraction of culture lysate. Of these 81 soluble designs, we found that 31 designs had sufficiently high purification yields for further investigation. We obtained Circular Dichroism (CD) spectra for these proteins, and we concluded that they were folded with secondary structures corresponding to those of the predicted design model. 17 of these 38 designs were found by size exclusion chromatography to elute at the expected retention volume for a monomer and found to be monodisperse, while and additional 15 were found to be polydisperse but with the most prominent SEC peak present at the expected retention volume for a monomer.

Ligand Docking and Computational Pocket Characterization

As a means of establishing the pseudocycle's potential for scaffolding ligand binding pockets, we selected 19 small molecule ligand candidates of interest with a variety of sizes, chemical properties, and molecular morphologies. Using representative models from each of our structural clusters of sequence designed pseudocycles, we carried out RIF docking[46] and pocket design calculations (Figure 6.2)[34]. As a computational reference, we also carried out our docking and scoring protocol for 1000 previously published de novo designed proteins with NTF2-like folds [37] and 2787 single-chain natural small molecule-binding proteins chosen from the PDB (via PDBBind[47]), using those same 19 candidate ligands. Structures of all candidate ligands are detailed in Figure A2.10.

We conducted an identical, Rosetta-based ligand aware design protocol on the residues proximal to the ligand binding site for each of these ligand docks: native and designed protein. Scaffolds and ligands were paired based on shape complementarity, computationally predicted binding energy, and other previously established computational metrics[34]. For each collection of docks considered to be "designable" by our computational docking experiment, we recorded the number of unique scaffolds compatible with the ligand for each category of scaffold. For most ligands, pseudocycles were found to have the largest number of docks relative to the total size of the scaffold pool (Figure 6.2 and Appendix 1).



c) Unique backbone count

Natives	2788	2	4	34	3	1	4	2	7	5	0	1	2	3	12	0	0	0	3	0
NTF2	1000	1	9	82	18	0	2	6	14	7	1	1	6	7	9	0	0	0	10	0
Pseudocycles	9838	46	81	0	97	9	109	57	97	105	45	43	82	90	44	80	117	2	101	22
Total	13626	49	94	116	118	10	115	65	118	117	46	45	90	100	65	80	117	2	114	22

Figure 6.2 - Computational Docking of Ligands to Central Pockets of Pseudocycles, Natural Proteins, and NTF2s

a) Examples of computationally designed binding interactions for diverse ligands in pockets of diverse pseudocyclic scaffolds. Designed proteins are shown as gray cartoons/sticks and the ligands are shown in pink sticks. Oxygen, nitrogen, phosphorus, and chlorine elements are colored in red, blue, orange, and green, respectively. **b)** Barplots showing the composition of input scaffolds (pseudocyclic designs, native structures from the PDBBind database, and designed NTF2's from Basanta et al[37]) and subsequent composition of the best ranked small molecule binders (each scaffold can contribute up to 30 binders) for diverse ligands. **c)** The numbers of unique backbone scaffolds selected based on the top 1% designed interface character from each type of scaffold are listed. 3-letter ligand codes are shown with molecular names and structures in Fig A2.10.

Solved crystal structures for pseudocycle designs

Of several crystal candidates, we obtained diffraction and were able to phase and refine structures for three designs. All three structures show very close resemblance to the AF2 predicted design model: closed repeat proteins (Figure 6.3). We found that each of the three structures solved contained central cavities along the pseudo-symmetric axis, consistent with the majority of our design models.

We found the structure of the first model to be a 5-fold pseudo-symmetric helical bundle with a marked central cavity (Figure 6.3.a). This fold is stabilized by a repeating cyclic hydrophobic interface between subunits formed between the smaller helical stub on the exterior of the bundle and the larger interior helices. The most similar structure observed in the PDB at this time is chain B of 3WFB, with a TMScore of 0.392. The deviation from the design model was found to be minimal with respect to alpha-carbon root-mean-squared-deviation (Ca-RMSD): 0.6 Å.

The second structure, a pseudo-C2 symmetric double helix hairpin, is a simple 4-helix bundle (Figure. 6.3.b); the closest structure in the PDB (5OXF) has a TMscore of 0.65, which is an accepted threshold for a protein having a similar fold. The design model accords strongly with the crystal structure at 0.8 Å Ca-RMSD.

The final structure, which contains a 3-stranded beta sheet packed against a helix, and forms a more complex pseudo-C3 symmetric fold (Figure. 6.3.c). The fold appears to be stabilized by the interdigitation of hydrophobic residues between repeat units. These seven hydrophobic residues fill a groove between the space where each helix-strand motif meets the opposite face of the equivalent sheet in the following repeat unit (Figure. 6.3.d). Strikingly, we observe an ordered water containing hydrogen bond network along the central axis of the pseudocyclic symmetry (Figure. 6.3.e). The symmetry axis lies where three strands, one from each unit, meet within range of short range electrostatic contact energies. This structure was found to be within 0.5 Å Ca-RMSD of the AF2 design model prediction, and dissimilar from all folds found in the PDB (closest TMscore of 0.38 to 1U7Z).

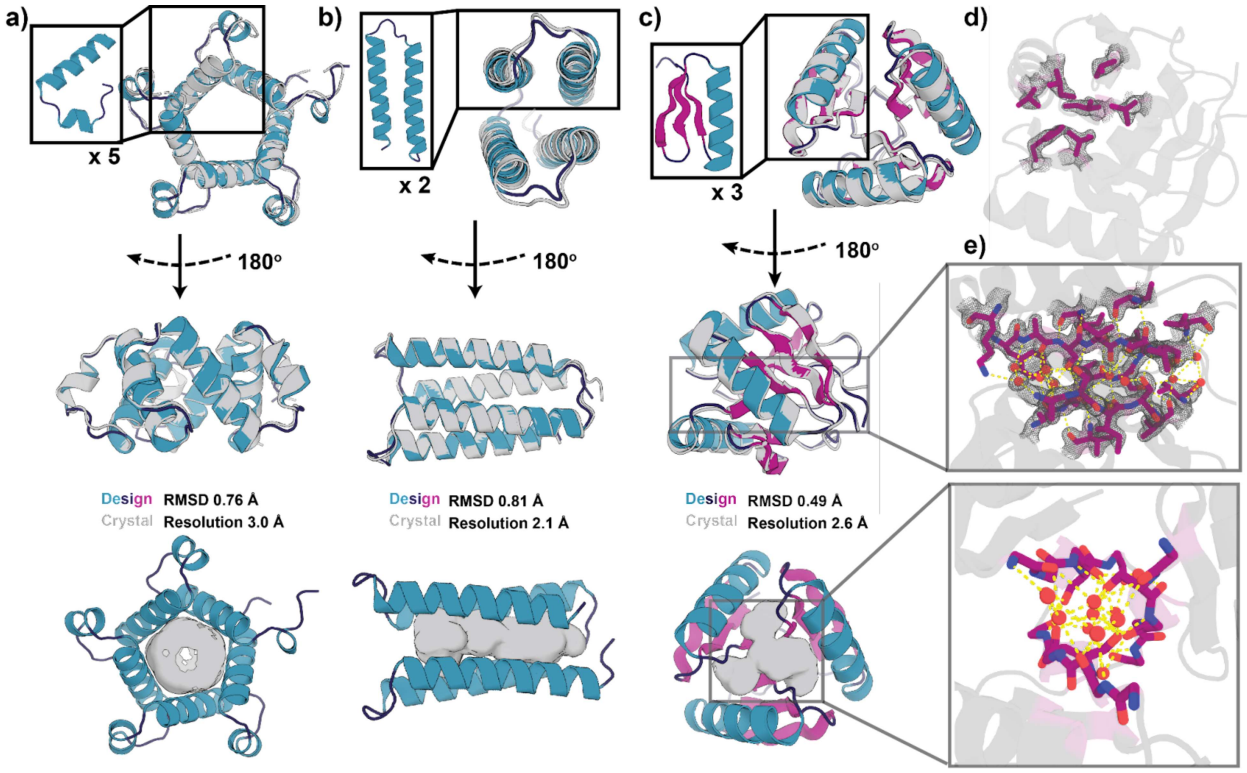


Figure 6.3 - X-ray Crystal Structures Match Computational Models

Crystal structures of the five-repeat design E8 **(a)**, two-repeat design H12 **(b)**, and the three-repeat design H10 **(c-e)** are shown in gray cartoons, the loop, sheets, and helix of the design are shown in dark blue, magenta, and teal, respectively. Central pockets in the designs are shown in gray sphere **(a, b, c)**. The secondary structure interface **(d)** and the center water mediated hydrogen bond network **(e)** of the refined crystal structure of design H10 are shown using sticks. The electron density map of the interface and the center hydrogen bond network and water are shown in gray mesh. The oxygen, nitrogen, and carbon are colored in red, blue, magenta in d and e; hydrogen bond networks are shown in yellow dashed lines, and water molecules in red spheres.

Chapter 7 - **Conclusions**

Beta Strands in Repeat Protein Design Enhances Structural Sampling

Previous attempts at repeat protein design were limited largely by two forces: the capacity to well-specify a search space, and the ability to search that space efficiently. Mimicry of natural protein backbones pre-specifies the search space and allows sequence oversampling and backbone splicing to achieve this objective. All-helical repeat protein design reduces the difficulty of well-specifying the search space: helical backbone positions are substantially less sensitive to small movements during both backbone and sequence design. This is both because backbone positions have fully satisfied hydrogen bonding residues and because local sequence patterns can lead to cooperative helical folding, less influenced by global forces on the biomolecule. Sampling with DL hallucination is able to overcome these challenges by exploiting secondary structure propensity of certain residue identities with evolutionary information. With this new toolset, repeat protein design can be conducted reliably without prior constraint of secondary structure types.

DL Design Fundamentally Changes Structure Sampling

The sampling space is reduced to plausible local backbone geometry for real biomolecules, even if intermediate steps predict physically implausible sequence/structure pairings. This can be observed through a major shift in a key Rosetta Energy metric: long range backbone hydrogen bonding energy (score term differences are illustrated in Appendix 6). Additionally: there is no need to specify distance constraints between sheet residues: the unproductive models sampled through fragment assembly with extended, unpaired, unstructured loop regions (composed of originally beta fragments) are dramatically less likely to be sampled in the hallucination space for any sequence, regardless of mismatch between secondary structure and sequence in the fragment identity.

Another benefit of hallucination based design is the ability to expand upon known geometries. The intimate relationship between structure and sequence means that finding the appropriate mutations to create a slightly different structure from a starting point can be explored with a simple random walk. The rapid convergence of trajectories implies that AF2, in particular, has the propensity, given a non-folding sequence, has a propensity to predict a structure similar to that of a *related* folding sequence. The final, high confidence structures may not yield productive proteins, but they are extremely plausible backbones, and can be designed.

While not all protein folds are expected to lie readily within a useful number of random trajectories, we have identified that the broad range of repeating alpha-beta LRR-adjacent geometries appear to be well suited to investigation in this manner. These repeats cover a geometric space previously inaccessible to designable repeat proteins, both purely helical geometry and LRR derivative proteins. The potential impact is therefore a great expansion in the sorts of super-structures which can be assembled, as well as further expansion of the way in which existing building blocks can be recombined. Additionally, we show that the space of

pseudocyclic, compact proteins can be searched for a broad variety of intriguing, previously not observed folds, and then fully designed as soluble, crystallizable biomolecules.

Rudimentary Structured Caps Effectively Improve Yields

We generated capping regions for the ABR designs, with the expectation that rapidly folding helical regions would prevent slower intermolecular aggregation of exposed beta-strands. With this approach we observed that presence of an additional helix at one or both termini prevented aggregation and greatly improved yields when compared to similar models.

The primary motivation for using such a straightforward, design-agnostic capping strategy was its generalizability. We found that very minimal helical caps, generated in a simple sampling protocol, produced a robust effect across a variety of models. While more construct-specific caps may prove to be better in specific contexts, this protocol offers, essentially, a robust solubilization domain for an entire class of protein folds composed of a single, 10-20 residue region. This approach is facilitated by MPNN, because of its robust ability to predict a consensus sequence for a structure, and the relatively small size of the cap relative to the whole protein. The effectiveness of this design strategy in non-repetitive proteins remains to be investigated.

Small helical caps have great potential, not only in repeat protein design, but in any design challenge with solvent exposed beta strands. More than half of the composition of a beta strand is generally hydrophobic, simply because a hydrophilic structure is penalized entropically for forming a structured intramolecular ordered feature. As a result, any designs with this secondary structure at the edges of the model volume stand to benefit from shielding via simple addition of an alpha-helix. While this may not be a viable strategy for all possible designs, we provide some evidence that a terminal sequence with high helical propensity may protect some designs from intermolecular aggregation.

DL Design Methods Improve Strand Fusion

Fusion along strands was previously a difficult challenge, solved in a number of specific cases as a proof of concept [48,49]. Previous approaches required careful matching of structure pairs to combine compatible strand twists, as well as extensions of strand length to accommodate the fusion. Most rigid fusions not along helices have required some form of fragment based loop closure to connect distinct structures. DL based inpainting allowed fusion of one exposed beta strand to another with only a coarse grained docking procedure of the desired strands to an arbitrary idealized strand pair. A combination of MPNN based sequence design and AF2 prediction ensured we could select models which retained the desired strand pairing after fusion. By using starting models predicted by AF2, and a backbone-only sequence design method, we are better able to design consensus sequences for combined domains. Previous approaches were highly sensitive to small changes in rigid body orientations, because they could produce incompatible loop endpoint positions or backbone positions incompatible with any residue identity appropriate to that molecular context. Prediction based approaches only sample plausible backbone geometries, greatly reducing this sensitivity, especially with

domains preselected for coarse grained compatibility.

The combination of modular beta-strand containing repeat proteins and molecular space filling tools signals a shift in viable design strategies. Previous efforts to recombine repeat proteins with desired rigid fusions required installation of a robust terminal helix region. Using ABRs and new DL design technology opens the opportunity to design repeat protein fusions along strands at the outside region of the molecule. This offers the potential to stably extend functional sheets, branch a wider variety of proteins, and generally reduces design times of multi-component fusion design procedures.

Afterword

The space of repeat protein design remains, of course, largely unexplored. Even a coarse-grained map, like the three-dimensional plot of superhelical repeat parameters, remains largely empty in spaces where steric clashes do not forbid proteins to lie. New innovations like the ABRs and pseudocycles help shore up some empty places, but the protein universe, even the much smaller repeat protein galaxy, remains largely unexplored by designers. New frontiers have been opened wide by accurate DL structural prediction, and greatly enhanced the toolbox for functional protein engineering.

The ABR models provide a handy collection of repeat proteins ready to be fused with precisely spaced protein components for macroassemblies, superstructures, or cellular activators. The continuous strand surface of these designs expands the volume of potential protein binder backbones with new, concave native-like high-surface area regions. While preliminary applications show the extension of a model protein, many more applications have yet to be realized.

Acknowledgements

This thesis work and all work I had the privilege of conducting at the Institute for Protein Design was enabled by a great number of people's generous contributions, enthusiasm, goodwill, and shared technical expertise. It is impossible to overstate the impact of this highly collaborative and open environment on my experience, so I will instead understate it. Reflecting back on all the different people who have made this journey possible is moving beyond words. This experience of trial and triumph, of creativity and collaboration, and of rigor and reimagining has changed me as a person, and changed my perspective on what it means to participate in modern science. Despite the obstacles, challenges, and frustrations this experience was replete with, I could not feel more grateful for having been accepted to experience it.

Individuals acknowledged in each section are listed alphabetically by last name.

Defense Committee

I would first and foremost like to thank my supervisory committee for volunteering their time, expertise, and efforts towards helping me navigate the process of completing my doctoral degree. Their insights and guidance have been integral to developing my scientific communication skills and completing an effective PhD level body of work.

David Baker
Patrick Boyle
Philip Bradley
Frank Dimaio
Neil King
Hao Yuan Kueh
Georg Seelig

Department

I would like to thank all the administrative and academic staff at the Departments of Bioengineering and Biochemistry at the UW for enabling a multitude of resources and providing guidance for the Bioengineering Doctorate of Philosophy. In particular, I would like to thank the following Graduate Program staff who personally assisted me in ensuring requirements and deadlines were fulfilled throughout the process.

Kalei Combs
Erin Kirschner

Institute for Protein Design

I would like to thank senior IPD staff Lauren Carter, Luki Goldschmidt, and Lance Stewart for their continuous efforts and leadership with respect to keeping the institute functioning and effective.

Direct Collaborators

I would like to thank all of my co-authors and people who directly contributed material, intellectual, or editorial support to this specific work.

Linna An
Asim K. Bera
Lauren Carter
Alexis Courbet
Justas Dauparas
Stacey Gerben
Derrick R. Hicks
Alex Kang
Lukas Milles
Hannah Nguyen
Basile Wicky

Direct Software Support

The following individuals directly contributed unpublished software or directly assisted in software debugging which was used in this work.

Ivan Anishchanka
Minkyung Baek
Brian Coventry
Adam Moyer
Will Sheffler

Mentors at the IPD

These people guided my journey for years as a young scientist and a graduate student, helping with challenges large and small, and showing me not only how to think for myself, but how to avoid thinking too much of myself, and too much in general.

Florian Praetorius
Nicholas Woodall
Danny Sahtoe

Other Support at the IPD

I would like to thank staff at the IPD who have helped ensure critical tasks get done behind the scenes to ensure all of our work is possible at all.

Ian Haydon

Xinting Li
Zari Magness
Hernan Nunez-Ortega
Austin Smith
Kandise VanWormer

Close Discussions

I will remember scientific and non-scientific discussions over the years at the IPD with great fondness. In particular, I would like to acknowledge the moral and intellectual support contributed to my thesis work by these individuals

Adam Broerman
Shane Caldwell
Jung Ho Chun
Fatima Davila
Natasha Edman
Hannah Han
Yang Hsia
Phil Leung
Sidney Lisanza
Ryan Kibler
Yakov Kipnis
Sanaa Mansoor
Jacob O'Connor
Arvind Pillai
Harley Pyles
Christian Richardson
Amijai Saragovi
Thomas Schlichthärle
Meerit Said
Jerimiah Sims
Will White

Family

The support of my family has been invaluable, but in particular the journey would have been much harder if not for the efforts, energy, and enthusiasm of my wife Ceren Savasan. Thanks, Buggy.

I would also like to thank my sister, Sophia Berezin, for her moral support, and whose own journey into the natural sciences inspired me to keep going in gloomy times.

Khare Lab

I would like to thank Sagar Khare, and all members of the Khare lab at my time there between 2016 and 2018. This was my first exposure to the Rosetta community, and my first training with computational protein design, and I would not be here without you.

Funding

This work was supported by a grant from the Department of Defense (DOD-0001039633), a grant from the National Institute on Aging (5U19AG065156.), a gift from the Washington Research Foundation, and the Audacious Project at the Institute for Protein Design.

This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

1. Akira, S. & Takeda, K. Toll-like receptor signalling. *Nat. Rev. Immunol.* **4**, 499–511 (2004).
2. Collagen family of proteins - Van Der Rest - 1991 - The FASEB Journal - Wiley Online Library. https://faseb.onlinelibrary.wiley.com/doi/abs/10.1096/fasebj.5.13.1916105.
3. Patino, M. G., Neiders, M. E., Andreana, S., Noble, B. & Cohen, R. E. Collagen: An Overview. *Implant Dent.* **11**, 280–285 (2002).
4. Bredow, M. & Walker, V. K. Ice-Binding Proteins in Plants. *Front. Plant Sci.* **8**, (2017).
5. Davies, P. L. Ice-binding proteins: a remarkable diversity of structures for stopping and starting ice growth. *Trends Biochem. Sci.* **39**, 548–555 (2014).
6. Huang, P.-S. *et al.* De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016).
7. Becker, S. & Boch, J. TALE and TALEN genome editing technologies. *Gene Genome Ed.* **2**, 100007 (2021).
8. Praetorius, F. & Dietz, H. Self-assembly of genetically encoded DNA-protein hybrid nanoscale shapes. *Science* **355**, eaam5488 (2017).
9. Uribe, K. B. *et al.* Engineered Repeat Protein Hybrids: The New Horizon for Biologic Medicines and Diagnostic Tools. *Acc. Chem. Res.* **54**, 4166–4177 (2021).
10. Jost, C. & Plückthun, A. Engineered proteins with desired specificity: DARPins, other alternative scaffolds and bispecific IgGs. *Curr. Opin. Struct. Biol.* **27**, 102–112 (2014).
11. Choi, Y. *et al.* Computer-guided binding mode identification and affinity improvement of an LRR protein binder without structure determination. *PLOS Comput. Biol.* **16**, e1008150 (2020).
12. Park, K. *et al.* Control of repeat-protein curvature by computational protein design. *Nat. Struct. Mol. Biol.* **22**, 167–174 (2015).
13. Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.* **39**, e82–e82 (2011).
14. Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P. & Plückthun, A. Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins. *J. Mol. Biol.* **332**, 489–503 (2003).
15. Parmeggiani, F. & Huang, P.-S. Designing repeat proteins: a modular approach to protein design. *Curr. Opin. Struct. Biol.* **45**, 116–123 (2017).
16. Hsia, Y. *et al.* Design of multi-scale protein complexes by hierarchical building block fusion. *Nat. Commun.* **12**, 2294 (2021).

17. Shen, H. *et al.* De novo design of self-assembling helical protein filaments. *Science* **362**, 705–709 (2018).

18. Nagano, N., Orengo, C. A. & Thornton, J. M. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321**, 741–765 (2002).

19. Ibraheem, A. & Campbell, R. E. Designs and applications of fluorescent protein-based biosensors. *Curr. Opin. Chem. Biol.* **14**, 30–36 (2010).

20. Reid, B. G. & Flynn, G. C. Chromophore Formation in Green Fluorescent Protein. *Biochemistry* **36**, 6786–6791 (1997).

21. Romero-Rivera, A., Corbella, M., Parracino, A., Patrick, W. M. & Kamerlin, S. C. L. Complex Loop Dynamics Underpin Activity, Specificity, and Evolvability in the (βα)8 Barrel Enzymes of Histidine and Tryptophan Biosynthesis. *JACS Au* **2**, 943–960 (2022).

22. Kamondi, S., Szilágyi, A., Barna, L. & Závodszky, P. Engineering the thermostability of a TIM-barrel enzyme by rational family shuffling. *Biochem. Biophys. Res. Commun.* **374**, 725–730 (2008).

23. Bragulla, H. H. & Homberger, D. G. Structure and functions of keratin proteins in simple, stratified, keratinized and cornified epithelia. *J. Anat.* **214**, 516–559 (2009).

24. Kobe, B. & Deisenhofer, J. Proteins with leucine-rich repeats. *Curr. Opin. Struct. Biol.* **5**, 409–416 (1995).

25. Kobe, B. & Kajava, A. V. The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* **11**, 725–732 (2001).

26. Marcos, E. *et al.* Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017).

27. Gogolok, S., Garcia-Diaz, C. & Pollard, S. M. STAR: a simple TAL effector assembly reaction using isothermal assembly. *Sci. Rep.* **6**, 33209 (2016).

28. Lin, S. *et al.* An Effective and Inducible System of TAL Effector-Mediated Transcriptional Repression in Arabidopsis. *Mol. Plant* **9**, 1546–1549 (2016).

29. Brunette, T. *et al.* Modular repeat protein sculpting using rigid helical junctions. *Proc. Natl. Acad. Sci.* **117**, 8870–8875 (2020).

30. Brunette, T. J. *et al.* Exploring the repeat protein universe through computational protein design. *Nature* **528**, 580–584 (2015).

31. Reconfigurable asymmetric protein assemblies through implicit negative design. https://www.science.org/doi/full/10.1126/science.abj7662?casa_token=UXeBDLVOsBUAAAAA%3ApJCaCyQWba-wgS624ANnrZ0nU_p1GwI_pZZBSHmImsoHXAIBffYiB0Fu3LoR7H-tfKCTapwIheVrcg.

32. De novo design of protein homodimers containing tunable symmetric protein pockets. https://www.pnas.org/doi/10.1073/pnas.2113400119 doi:10.1073/pnas.2113400119.

33. Doyle, L. *et al.* Rational design of α-helical tandem repeat proteins with closed architectures. *Nature* **528**, 585–588 (2015).

34. Cao, L. *et al.* Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).

35. Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).

36. Marcos, E. *et al.* De novo design of a non-local β-sheet protein with high stability and accuracy. *Nat. Struct. Mol. Biol.* **25**, 1028–1034 (2018).

37. Basanta, B. *et al.* An enumerative algorithm for de novo design of proteins with diverse pocket structures. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 22135–22145 (2020).

38. Hallinan, J. P. *et al.* Design of functionalised circular tandem repeat proteins with longer repeat topologies and enhanced subunit contact surfaces. *Commun. Biol.* **4**, 1–14 (2021).

39. Loshbaugh, A. L. & Kortemme, T. Comparison of Rosetta flexible-backbone computational protein design methods on binding interactions. *Proteins Struct. Funct. Bioinforma.* **88**, 206–226 (2020).

40. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).

41. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

42. An, L. *et al.* Hallucination of closed repeat proteins containing central pockets. 2022.09.01.506251 Preprint at https://doi.org/10.1101/2022.09.01.506251 (2022).

43. Wicky, B. I. M. *et al.* Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).

44. Dauparas, J. *et al.* Robust deep learning based protein sequence design using ProteinMPNN. 2022.06.03.494563 Preprint at https://doi.org/10.1101/2022.06.03.494563 (2022).

45. Wang, J. *et al.* Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).

46. Dou, J. *et al.* De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485–491 (2018).

47. Su, M. *et al.* Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* **59**, 895–913 (2019).

48. Lin, Y.-R., Koga, N., Vorobiev, S. M. & Baker, D. Cyclic oligomer design with de novo αβ-proteins. *Protein Sci.* **26**, 2187–2194 (2017).

49. Stranges, P. B., Machius, M., Miley, M. J., Tripathy, A. & Kuhlman, B. Computational design of a symmetric homodimer using β-strand assembly. *Proc. Natl. Acad. Sci.* **108**, 20562–20567 (2011).

50. Michael J. McCarthy. *Introduction to theoretical kinematics.* (MIT press, 1990).

51. Voynov, V., Chennamsetty, N., Kayser, V., Helk, B. & Trout, B. L. Predictive tools for stabilization of therapeutic proteins. *mAbs* **1**, 580–582 (2009).

52. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).

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

54. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

55. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **72**, 171–179 (2016).

56. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminformatics* **3**, 33 (2011).

57. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38, 27–28 (1996).

58. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci. Publ. Protein Soc.* **30**, 70–82 (2021).

59. Park, H., Zhou, G., Baek, M., Baker, D. & DiMaio, F. Force Field Optimization Guided by Small Molecule Crystal Lattice Data Enables Consistent Sub-Angstrom Protein–Ligand Docking. *J. Chem. Theory Comput.* **17**, 2000–2010 (2021).

60. Sharp, P. M. & Li, W. H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).

61. Dang, B. *et al.* SNAC-tag for sequence-specific chemical protein cleavage. *Nat. Methods* **16**, 319–322 (2019).

62. Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).

63. Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. HKL-3000: the integration of data reduction and structure solution--from diffraction images to an initial model in minutes. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 859–866 (2006).

64. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).

65. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

66. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).

67. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

68. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci. Publ. Protein Soc.* **27**, 293–315 (2018).

Appendix 1- **Pseudocycle Detailed Methods**

Preface

This appendix contains material directly reproduced from a collaborative work[42] in which the author was a primary contributor.

Protein generation and sequence design pipeline

Sequence space was sampled via MCMC to generate AF2-based backbones through hallucination. Initial models were generated by AF2 prediction of a random, tandem repeated, sequence of length L (chosen between 20 and 60). At each step, random mutations were evaluated by propagating the mutation to each repeat position and generation of a structure through AF2 prediction. This structure was then scored on a linear combination of cyclic character, pLDDT, and pTMscore. Cyclic character was defined as a helical rise near zero and per-unit rotation near 360/N degrees, where N is the number of repeat units. The degree of cyclicity or "closure score" in each model was set as the mean between a "rise score" and a "rotation" score. Each of these was the raw difference (in angstroms and degrees respectively) from ideal values (0 and 360/N) rescaled logistically to a value between 0 and 1, where the midpoint of the logistic curve (a score of 0.5) was set as a rise of 2 and a rotation deviation of 4 degrees respectively, with smaller deviation trending to 0 and larger deviations asymptotically approaching 1.

Mutations were made at a rate of 1-3 per MCMC step, with more mutations at the beginning of the trajectory and fewer as the trajectory proceeded. Different trajectory parameters were tried, and average time to reach a minimum quality as well as average number of successful trajectories were largely unaffected by the ranges investigated, with the exception of a larger weight on cyclic score leading to roughly 5-fold faster convergence times than equal weighted scores. Larger weights tended to increase the convergence time. Mutation rate, MCMC temperature, or mutation type weights did not dramatically affect convergence times.

Closure score is a score from 0 to 1 (0 indicates perfect closure, values approaching 1 indicate significant deviation from ideal parameters), which is a linear combination of rescaled delta rise and rescaled delta rotation. Delta rise and delta rotation are computed by extrapolating a superhelical axis (screw axis) from a homogeneous relative transform derived between repeat units. Relative transforms between matching repeats where derived from protein backbone positions, using the Nitrogen, α-Carbon, and Carboxyl Carbon to define the local coordinate plane for each residue position, and then deriving the homogenous relative transform between two coordinate frames. Relative transforms were then converted to a smoothed transform estimating the approximate rigid body relationship between repeat units. This was accomplished by averaging the quaternions corresponding to the relative transforms, and directly computing the mean translation vector from relative transforms. A helical axis is derived from this transform, as well as rise along and rotation about that axis through geometric transformations using the Rodrigues rotation vector and Rodrigues rotation matrix and standard kinematic equations [50].

The logistic rescaling function used was:

$$\frac{1}{1 + e^{-s(x-m))}}$$

Where *s* is the slope factor, *x* is the delta value, and *m* is the logistic scale midpoint. *m* and *s* for rotation were 4 and 1.5 respectively. Those values for rise were 2 and 2, respectively. The mean between these rescaled values was used as a score for closure quality.

After generation of initial pseudocycle scaffolds, we used ProteinMPNN[44] to design the sequences as asymmetric monomeric proteins and used AF2 and RF to generate structural models for the new sequences.

After initial generation and visual inspection of pseudocycle design models, a problematic density of apolar residues on solvent-accessible regions of the molecular surface became apparent. These patches were later quantified with the SAP score (see Figure. A2.2 and A2.3). To resolve this issue, we used ProteinMPNN[44] and the Rosetta sequence design suite[40] to perform a redesign of the pseudocycles of their surface residues. For each pseudocycle model, we generated 100 sequences using ProteinMPNN[44] and used these newly generated sequences to generate a scaffold-specific Position Specific Score Matrix (PSSM) file. Using Rosetta sequence design (FastDesign), surface residues of the pseudocycle models with high spatial aggregation propensity (SAP)[51] score were selected, designed with the non-hydrophobic amino acid preference provided by the PSSM file, and scored with Rosetta metrics. These sequences were then used to predict new pseudocycle structures with AF2 and RoseTTAFold (RF). All 5 available AF2 models were used for prediction, but the final prediction model used as the design structure was the highest among the 5 predictions for each sequence, which we designated “rank1 prediction”. AF2 metrics including pLDDT and pTM as well as RF metrics including pLDDT, CCE, KL divergence, and RMSD of the predicted structure to the original design model are presented in Figure. A2.4.

To generate the final pseudocycle backbone scaffold list for further design applications, we examined Rosetta metrics, AF2/RF metrics. We removed the models which were predicted to fold into scaffolds with Ca-RMSD over 2 Å to the original model by AF2 and generated a finalized pseudocycle list which consists of 21,021 designed proteins.

TMalign methods

We used pyrosetta[52] to calculate the average TMscore between two input proteins by averaging the TMscore obtained when each of the two pdbs was treated as the reference pdb for sequence length normalization.

TMalign to natives

We curated a set of sequence nonredundant (using mmseq[53]), high resolution (< 1.8 Å resolution), relaxed monomer structures from the PDB which yielded 6,111 structures. We then executed TMalign as described above to find designs with native structure equivalents.

mTMalign of 96 characterized designs against PDB

We used the mTMalign server (http://yanglab.nankai.edu.cn/mTM-align/) to compare the designs we characterized to the whole PDB, selecting the top database hit for TMalign comparison.

Protein clustering

The design objective of clustering was jointly to minimize redundant computation for downstream design and to productively categorize the backbone space searched. As such, within each search trajectory group (repeat number, DSSP type and ordering of final models) we hoped to find the smallest number of clusters where each cluster generally still represented extremely self-similar models. The goal was to construct clusters with TM score of above 0.85, and where each model within a cluster, when sampled for design/docking produces similar results to every other model. Loss of fine grained diversity or some similarity between groups was acceptable under the constraint that each final cluster should produce a scaffold capable of hosting unique pockets, and a single model from each cluster could reasonably serve as a proxy for the rest.

All scaffolds were first grouped based on their initial symmetry number (2-7), subsequently sub-grouped based on their repeat unit DSSP, followed by empirical clustering via AgglomerativeCluster (from scikitlearn[54]) through an all-by-all matrix of TMalign scores within each sub-group. Clustering parameters for AgglomerativeCluster were iteratively tuned to determine a useful consensus: final cluster number was chosen by minimizing the proportion of clusters with fewer than 2 members (singletons) with respect to the joint constraints of maintaining a high mean TMScore within each cluster and a low standard deviation of intra-cluster TMScores. The final clustering yielded a mean intra-group TMScore of 0.88 (not including singleton clusters) with only 9.42% singleton clusters.

Ligand docking to pseudocycles, NTF2, and Native proteins

The design objective of ligand docking small molecules to pseudocycles was to identify the likelihood of such scaffolds to have potential for hosting designable internal region along the central axis. The exact symmetry, volume, and dimensions of such pockets were not constrained. As an *in silico* control measure for our docking procedure, we evaluated the pocket quality of proteins known to have such features: the de novo designed NTF2-fold library[37] and native ligand-binding proteins. The design objective was to establish that we could detect and design ligand binding pockets at similar rates in this new scaffold set as with previous scaffold

set. Methods described were adapted from previous work to generate small molecule binding scaffolds from the NTF2 pool.

The pocket residues of pseudocycles were annotated using a python script. This procedure identified the largest internal cavity bounded by a convex hull wrapped around the protein, generates a poly-alanine backbone model and then identifies all side chain residues whose beta-carbons contact this internal cavity. The pockets of 100 randomly selected pseudocycles were manually inspected to verify annotation accuracy (approximate correspondence to a central cavity). A similar method of scaffold based annotation of pocket residues of NTF2s has been reported previously[37]. Previously verified native small molecule-binding proteins were taken from the PDBBind database[47]. Only single-chain native small molecule-binding proteins were selected to be comparable with pseudocycles. The binding pockets were selected based on the annotation provided by PDBBind database. The native proteins were relaxed with backbone and sidechain constraints using Rosetta to remove clashes before any further computational experiments.

All 19 ligands were docked to pseudocycles, NTF2, and native proteins following the same procedure. One to eight rotamers of each ligand were extracted from PDB or Cambridge Structure Database[55]. Hydrogens were added to ligand rotamers using OpenBabel[56] or VMD[57] with visual inspection. The conjugation and charge were edited/added with VMD or Chimera[58] with visual inspection. The parameter file of the ligand was generated using the python script from Rosetta application.

Rifgen/RIFdock suite[46] was used to perform ligand docking to all the proteins. Various amino acid rotamers (referred to as RIF, or Rotamer Interaction Field) which provide hypothetical polar, aromatic, and apolar interactions to the ligand rotamer were generated for each ligand using Rifgen function with the requirements of polar interactions to all heavy atoms from the ligand rotamer. The polar interaction requirement was relaxed if the generated RIF size was smaller than 1 MB, since such a small file size for RIFs generally corresponds to a degree of sparseness which does not yield meaningful docking data. With the RIF generated for each ligand, these RIFs, which encode geometry and energy information of potential interaction between amino acid rotamers and the ligand rotamers, were used to compute the docking quality of pseudocycles, NTF2s, and native proteins at their annotated pocket residues to spatially locked ligand molecules using RIFdock function. All remaining requirements to make polar interactions to ligand heavy atoms were kept during the docking procedure. A maximum of 30 docks were generated for each protein scaffold. Scaffolds which produced 0 docks were not designed or scored further.

The generated docks were designed using Rosetta sequence design suite to provide score terms to identify the protein scaffolds most suitable for holding each ligand. Each generated dock was designed using a fast version of fixed-backbone sequence design procedure adapted from protein-protein binder design to be ligand-aware[34,59]. Interface metrics including, “contact_molecular_surface”, “ddG” were used to select the top percentile of binders designed for each ligand[34]. The top 1% of the selected docks for each ligand type were identified with their scaffold type origin to compare scaffold group hit rates (Figure. 6.2).

Expression and purification of selected proteins

Designs were reverse translated into DNA using a custom python script that attempts to maximize host-specific codon adaptation index[60] and IDT synthesizability, which includes optimizing whole gene and local GC content as well as removing repetitive sequences, and ordered as Eblocks from IDT. Eblocks were cloned into a pET29b-derived vector with C-terminal SNAC cleavable His Tags using Golden Gate assembly and transformed into *E. coli* BL21 strain. The solubility of the proteins was first assessed using small-scale expression. 1 mL cultures were grown in a round-bottom 96 deep-well plate covered with breathable film and shaken at 200 x rpm overnight at room temperature, the cultures were harvested using centrifugation for 10 min at 4,000 x rpm and resuspended in bugbuster lysis buffer (1x bugbuster (Millipore), 25mM Tris, 100 mM NaCl, pH 8). The lysed cells were centrifuged and 10 uL of soluble fraction was investigated for protein of interest via SDS-PAGE and coomassie stain. Protein bands at expected molecular range were used as judgment for protein expression and solubility (an empty vector culture was used as background to assist in excluding native E. coli protein bands). Soluble designs were subsequently grown in 50 mL autoinduction media in 250 mL baffled erlenmeyer flasks for assay-scale production (6 hours at 37 ℃ followed by 24 hours at 18 ℃ shaking at 180 rpm). Cells for each design culture were harvested and resuspended in 30 mL of lysis buffer (25mM Tris 100 mM NaCl, pH 8, with protease inhibitor tablet) and sonicated to lyse (3 min sonication, 10 s pulse, 10 s pause, 60% amplitude). After centrifugation for 30 min at 14,000 x rpm, soluble fractions were bound to 1 mL Ni-NTA resin (Qiagen) in a Econo-Pac® gravity column (BIO-RAD) at 4 ℃ for 1 hour with rotation. The resin was washed with 20 CV (column volume) low salt buffer (50 mM tris, 100 mM NaCl, 50 mM Imidazole, pH 8) and with 20 CV high salt buffer (50 mM tris, 1000 mM NaCl, 50 mM Imidazole, pH 8).

For initial characterization using SEC and CD, proteins were eluted with 2 CV of elution buffer (25 mM tris, 100 mM NaCl, 500 mM Imidazole, pH 8) and purified on a superdex 75 increase 10/300 GL column connected to ÄKTA protein purification systems in TBS buffer (25 mM Tris, 100 mM NaCl, pH 8).

For crystallography, the samples were treated identically except 4 to 8 flasks of 50 mL cultures were pooled together before sonication and His tags were cleaved on bead following the SNAC cleavage protocol[61] before subsequent SEC purification.

For small-angle X-ray scattering (SAXS) studies, the samples were treated identically except for 4 flask of 50 mL culture were pooled together before sonication, and SNAC [61] tags upstream of His-tag were cleaved on column by overnight incubation in 40ml SNAC cleavage buffer before subsequent SEC purification. The buffer of the samples were exchanged via serial concentration in centrifugal 3k micropore concentrator units to 20 mM tris, 100 mM NaCl, 2% glycerol (v/v) for SAXS studies.

Circular dichroism characterization of selected proteins

Circular dichroism spectra were measured with a Jasco J-1500 CD spectrometer. Samples were diluted to concentrations of approximately 0.25 mg/mL (individual samples ranging 0.1 - 0.5 mg/ml) in 25 mM phosphate buffer titrated to pH 8 by combination of dibasic and monobasic sodium phosphate according to a standard table, and a 1-mm path length cuvette was used. The CD signal was converted to mean residue ellipticity by dividing the raw spectra by N × C × L × 10, where N is the number of residues, C is the concentration of protein, and L is the path length (0.1 cm).

Crystallographic analysis

Crystals were produced using the sitting drop vapor diffusion method. Drops with volumes of 200 nL in ratios of 1:1, 2:1, and 1:2 protein to crystallization were set up in 96 well plates at 20°C, using the Mosquito from SPT Labtech. Drops were monitored using the JANSi UVEX imaging system.

For E8, diffraction quality crystals appeared in a mixture of 0.2M DL-Glutamic acid monohydrate, 0.2M, DL-Alanine, 0.2M Glycine, 0.2M DL-Lysine, 1.0M imidazole, MES monohydrate (acid), 37.5 % v/v of (25% v/v MPD; 25% PEG 1000; 25% w/v PEG 3350).

For H10, diffraction quality crystals appeared in a mixture of 0.12 M D-glucose, 0.12 M D-mannose, 0.12 M D-galactose, 0.12 M L-fucose, 0.12 M D-xylose, 0.12 M N-acetyl-D-glucosamine, 0.0499 M HEPES, 0.0501 M MOPS (acid), 20% v/v PEG 500 MME, and 10% w/v PEG 20,000.

For H12, diffraction quality crystals appeared in a mixture of 0.09 M sodium fluoride, 0.09 M sodium bromide, 0.09 sodium iodide, 0.0499 M HEPES, 0.0501 M MOPS (acid), 12.5% v/v MPD, 12.5% PEG 1000, and 12.5% w/v PEG 3350.

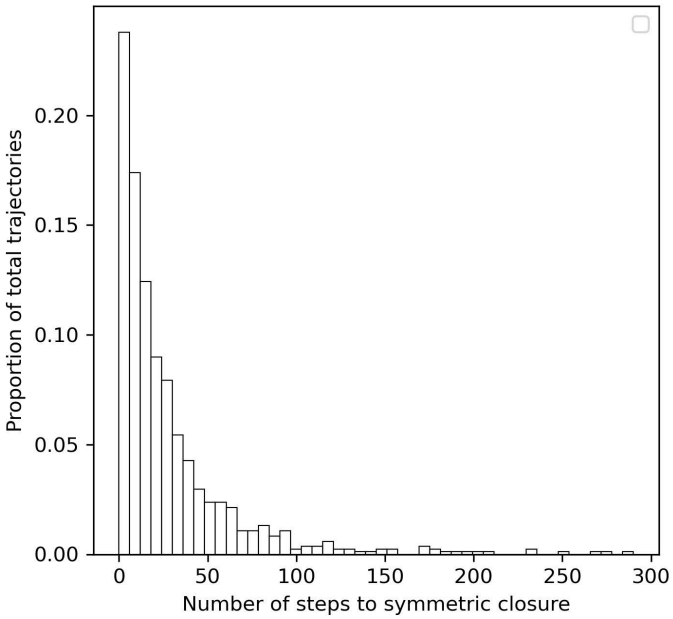
Crystals were cryoprotected prior to flash freezing in liquid nitrogen before shipping for data collection at synchrotron. Data collection was performed with synchrotron radiation at the Advanced Photon Source (APS) on beamline 24ID-C.

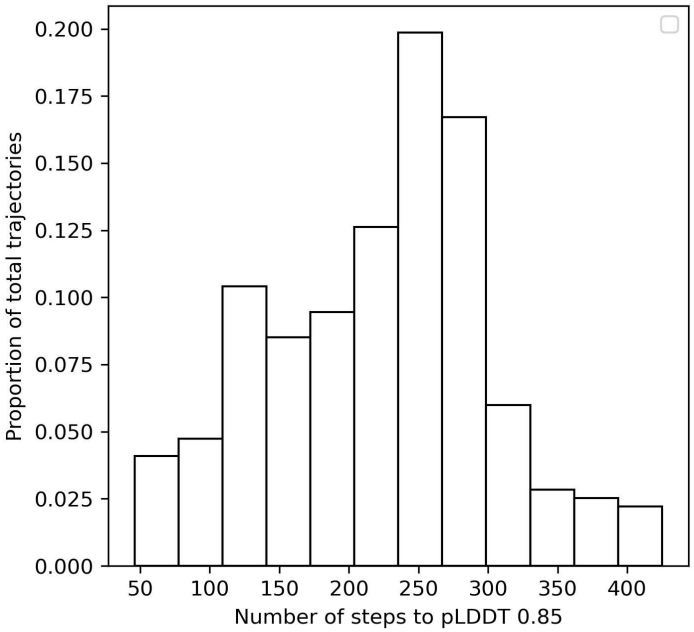
X-ray intensities and data reduction were evaluated and integrated using either XDS[62] or HKL3000[63] and merged/scaled using Pointless/Aimless in the CCP4 program suite[64]. Structure determination and refinement starting phases were obtained by molecular replacement using Phaser[65] using the design model for the structures. Following molecular replacement, the models were improved using phenix autobuild[66]; efforts were made to reduce model bias by setting rebuild-in-place to false, and using simulated annealing. Structures were refined in Phenix[66]. Model building was performed using COOT[67]. The final model was evaluated using MolProbity[68]. Data deposition, atomic coordinates, and structure factors were deposited in the PDB (PDB ID 8FJE for E8, 8FJF for H10 and 8FJG for H12).

Appendix 2 - **Pseudocycle Additional Figures**

Preface

This appendix contains material directly reproduced from a collaborative work[42] in which the author was a primary contributor.





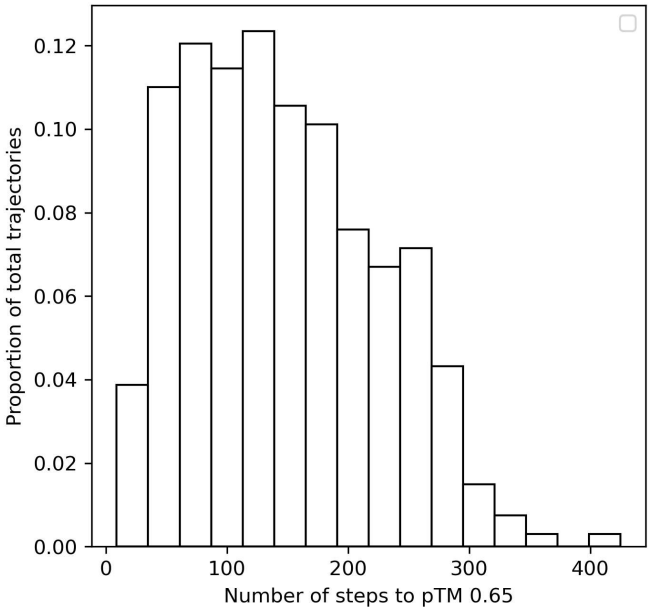


Figure A2.1. Histogram of MCMC Step Count to Convergence

The number of steps to closure for a representative sample of trajectories is shown here. Symmetric closure convergence is defined as a “closure score” of 0.1 or less. A clear trend is that AF2 readily predicts closed, cyclic structures from random repetitive sequences, but with initially low confidence.

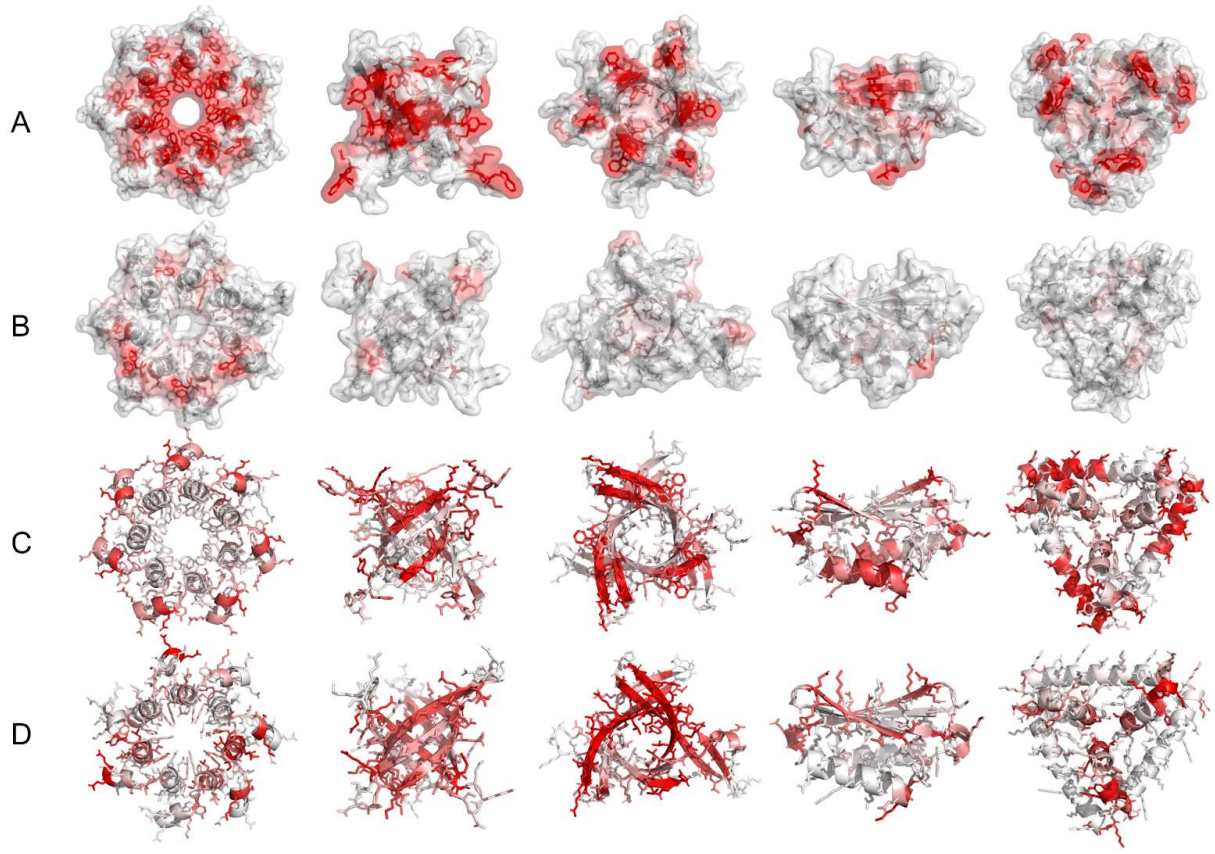


Figure A2.2 - Cartoons of designed pseudocycles with per residue SAP and psipred scores before and after ProteinMPNN & Rosetta redesign

a) 5 diverse representative proteins following the hallucination procedure colored by SAP score. Color scales from white (no aggregation propensity) to red (high aggregate propensity). b) The same 5 proteins after ProteinMPNN redesign and Rosetta surface optimization colored by SAP score. c) The same 5 hallucinated proteins colored by agreement of single sequence psipred prediction with the intended secondary structure. Color scales from white (perfect agreement) to red (no agreement). d) The redesigned proteins colored by agreement of single sequence psipred prediction with the intended secondary structure.

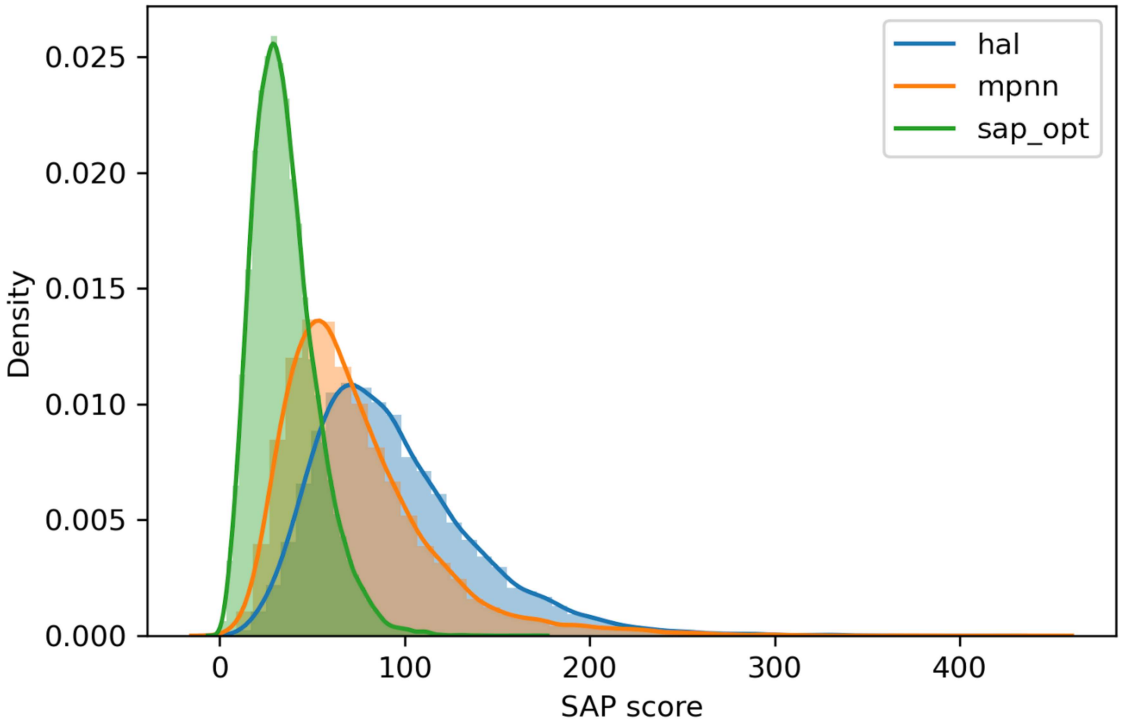


Figure A2.3 SAP Score Improvement During Design

Histogram of SAP score for original hallucinations (hal), after ProteinMPNN (mpnn) redesign, and after Rosetta surface optimization (sap_opt).

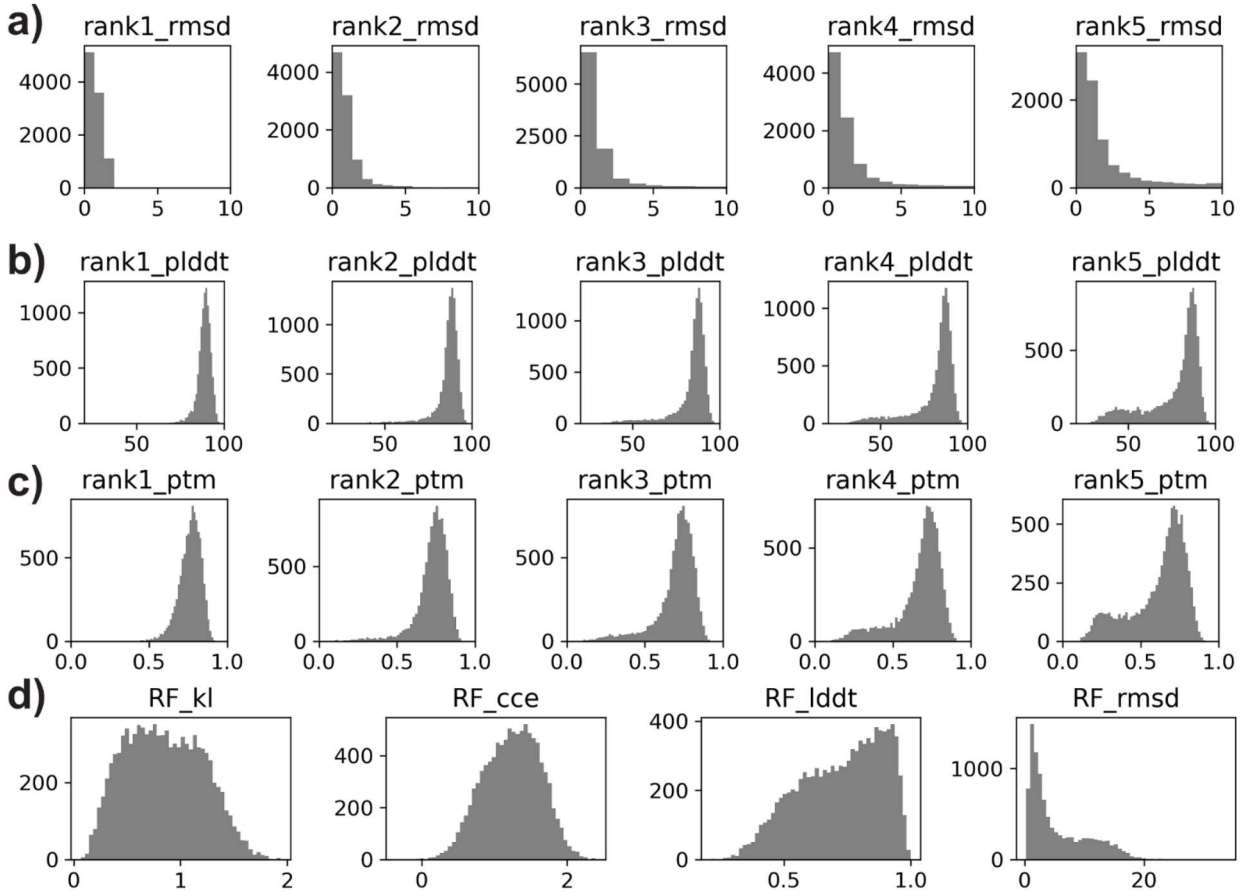


Figure A2.4 - AF2 and RF metric histograms for 9838 pseudocycle cluster representatives

a). AF2 Ca-RMSD to design models for 5 AF2 models by AF2 rank. b). AF2 plDDT for predictions. c). AF2 ptm for predictions. d). RosettaFold (RF) Ca-RMSD to design model, RF lddt , RF KL divergence, and RF CCE for predictions.

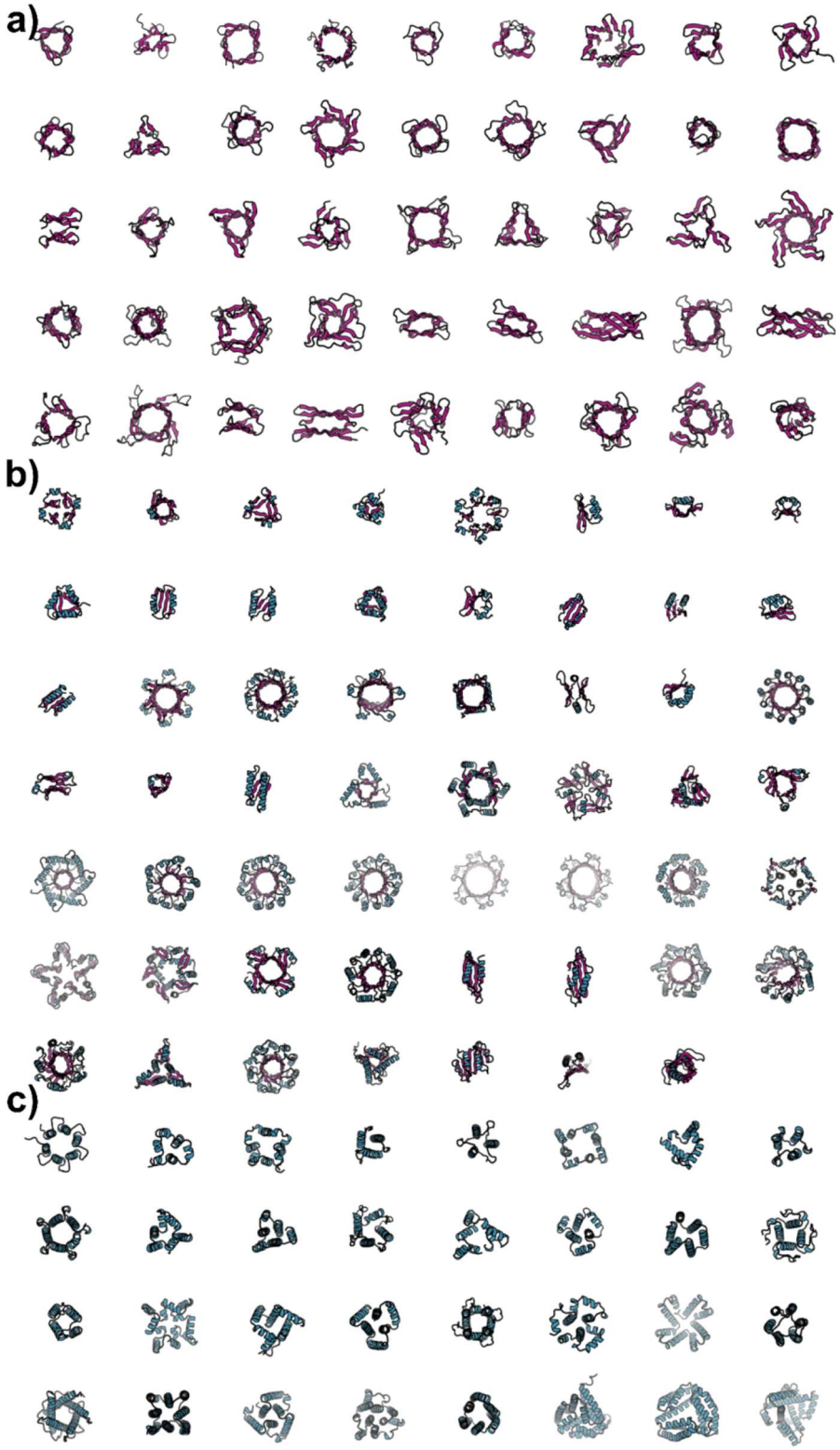


Figure A2.5 - Backbone Diversity of Pseudocycle Structures

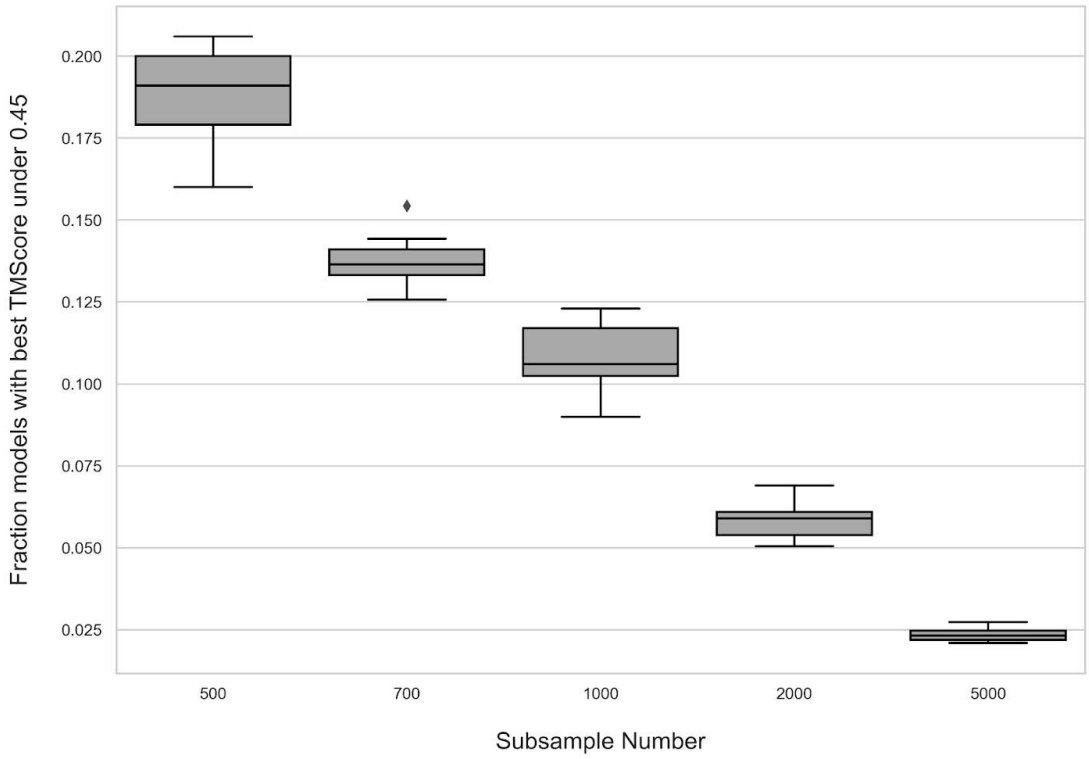


Figure A2.6 - Pseudocycle Structural Space Sampling

We randomly constructed 10 subsamples from our pool of design scaffolds (after removing structurally redundant models as described in Methods Protein Clustering) and recorded the number of models we found with TMScore of 0.45 or less to every other model in the pool. This represents models which are significantly different from every other model in the sample. Increased sample size shows that a smaller fraction of the models are structurally unique, this further implies that we have sampled the majority of the space available with this method. Central line shows median, box shows interquartile range, and whiskers show range, except for one outlier for the 700 sample group, shown as a diamond.

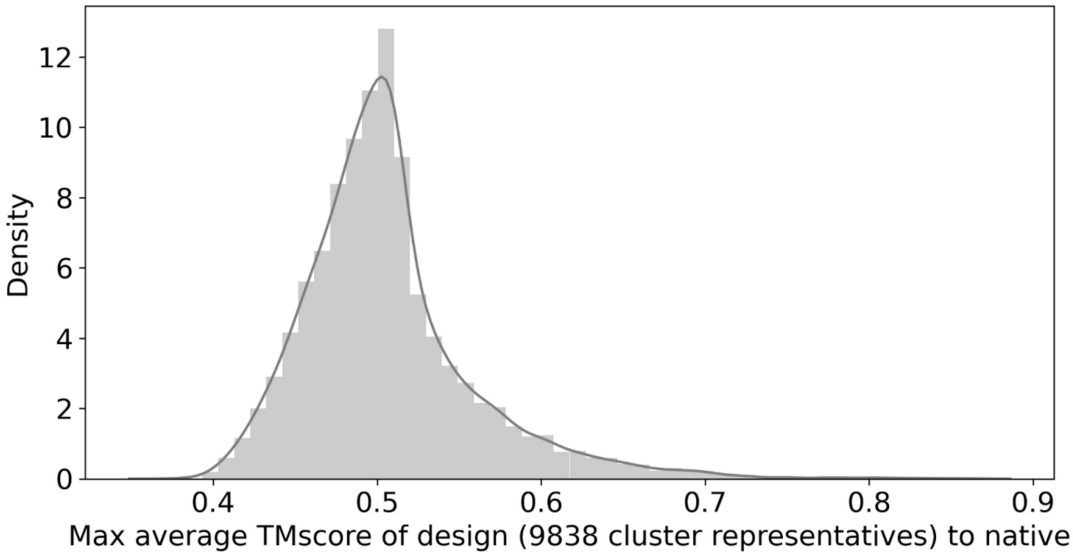


Figure A2.6 - TMscore of Designs to Natives

For each of the 9838 design cluster representatives, the max average TMscore to native structures plotted as a histogram.

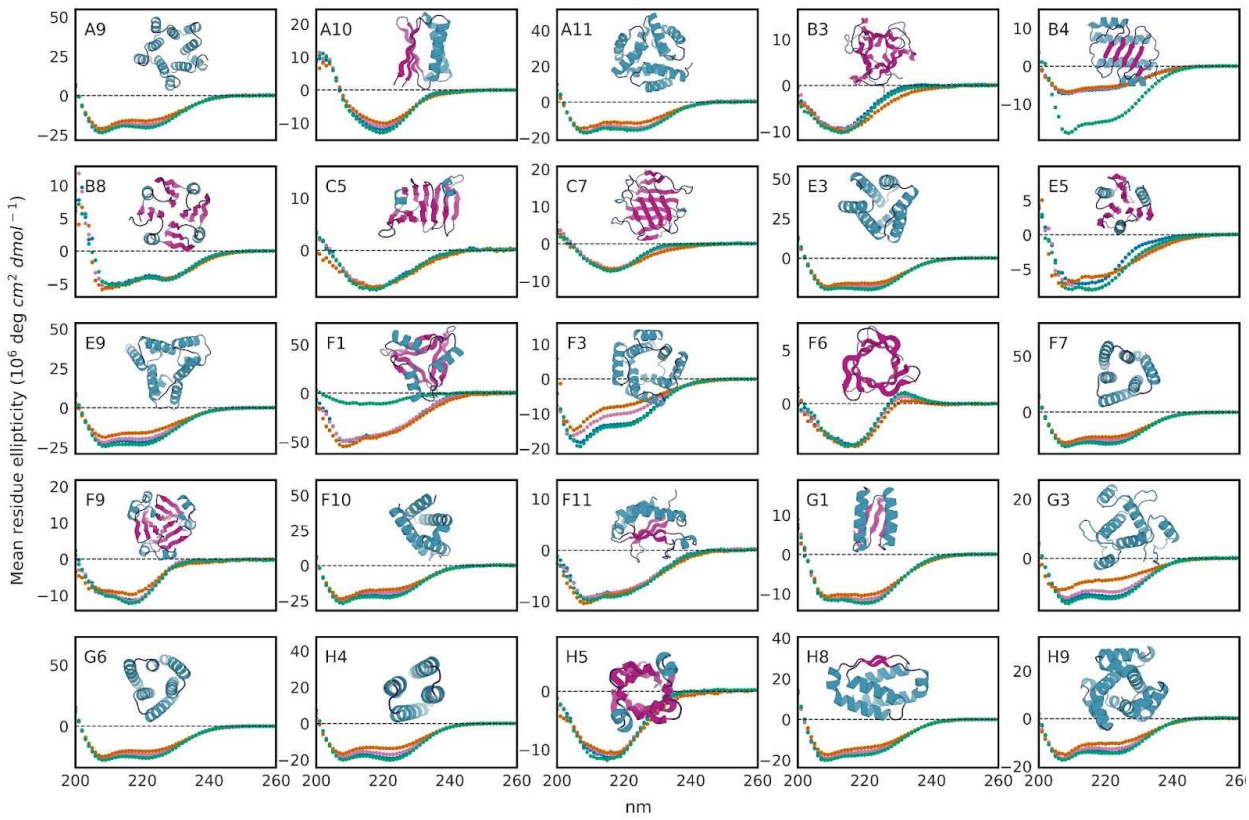


Figure A2.7 - CD data for 25 designs not shown in figure 4.3

Different temperatures of the CD scan spectra are plotted as follows: 25 ℃ in blue, 55 ℃ in orange, 95 ℃ in pink, refolding at 25 ℃ in green. The cartoon of the corresponding designed pseudocycle is shown with each CD spectra. The sheet, helix, loop substructures are colored in magenta, teal, and dark blue, respectively.

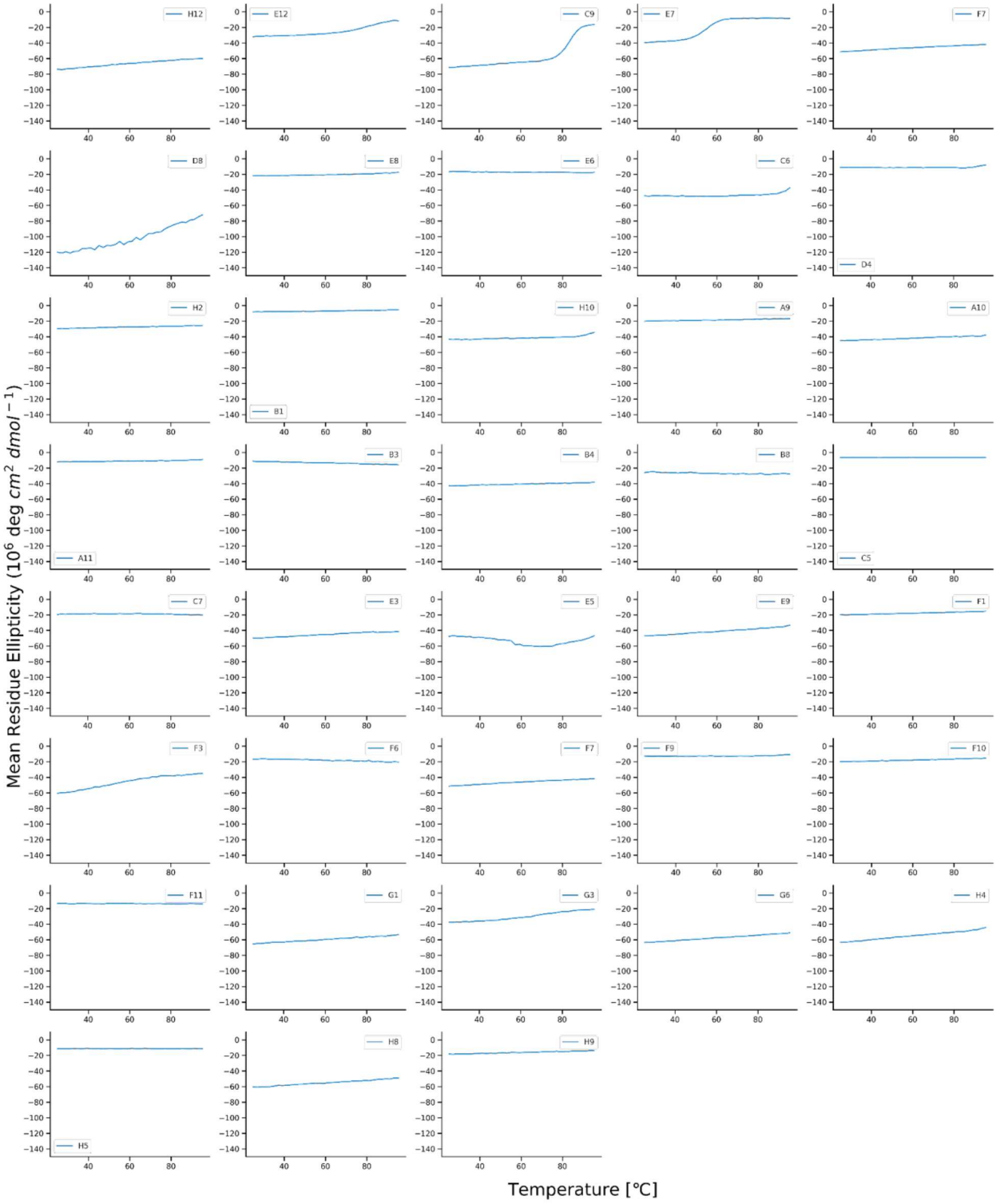


Figure A2.8 - Thermal Melt Profiles at 200 nm

Traces of mean residue ellipticity as a function of temperature for all designed pseudocycles with monomer fractions (as judged by SEC traces).

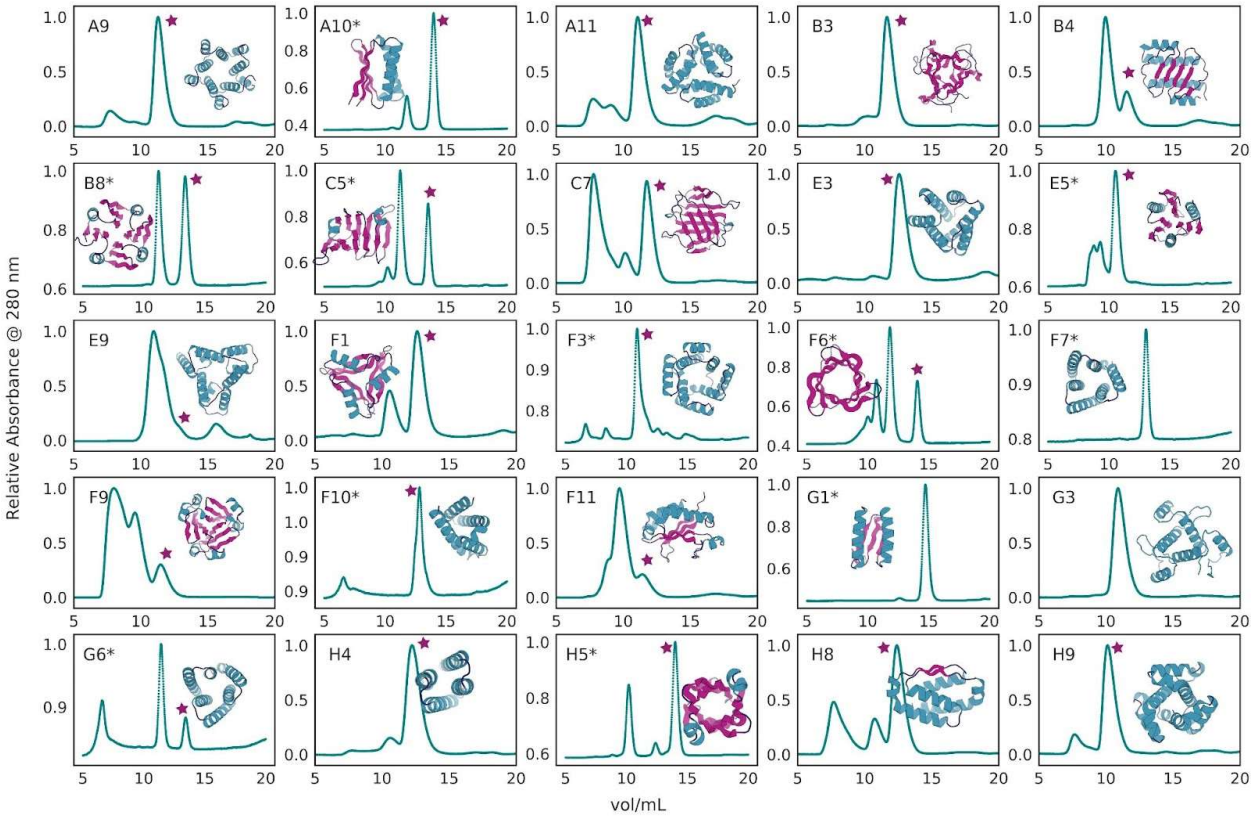


Figure A2.9 - SEC data for 25 designs not shown in figure 4.3

Monomeric fraction was marked out using a magenta star. The cartoon of the corresponding designed pseudocycle is shown with each subplot. The sheet, helix, loop substructures are colored in magenta, teal, and dark blue, respectively.

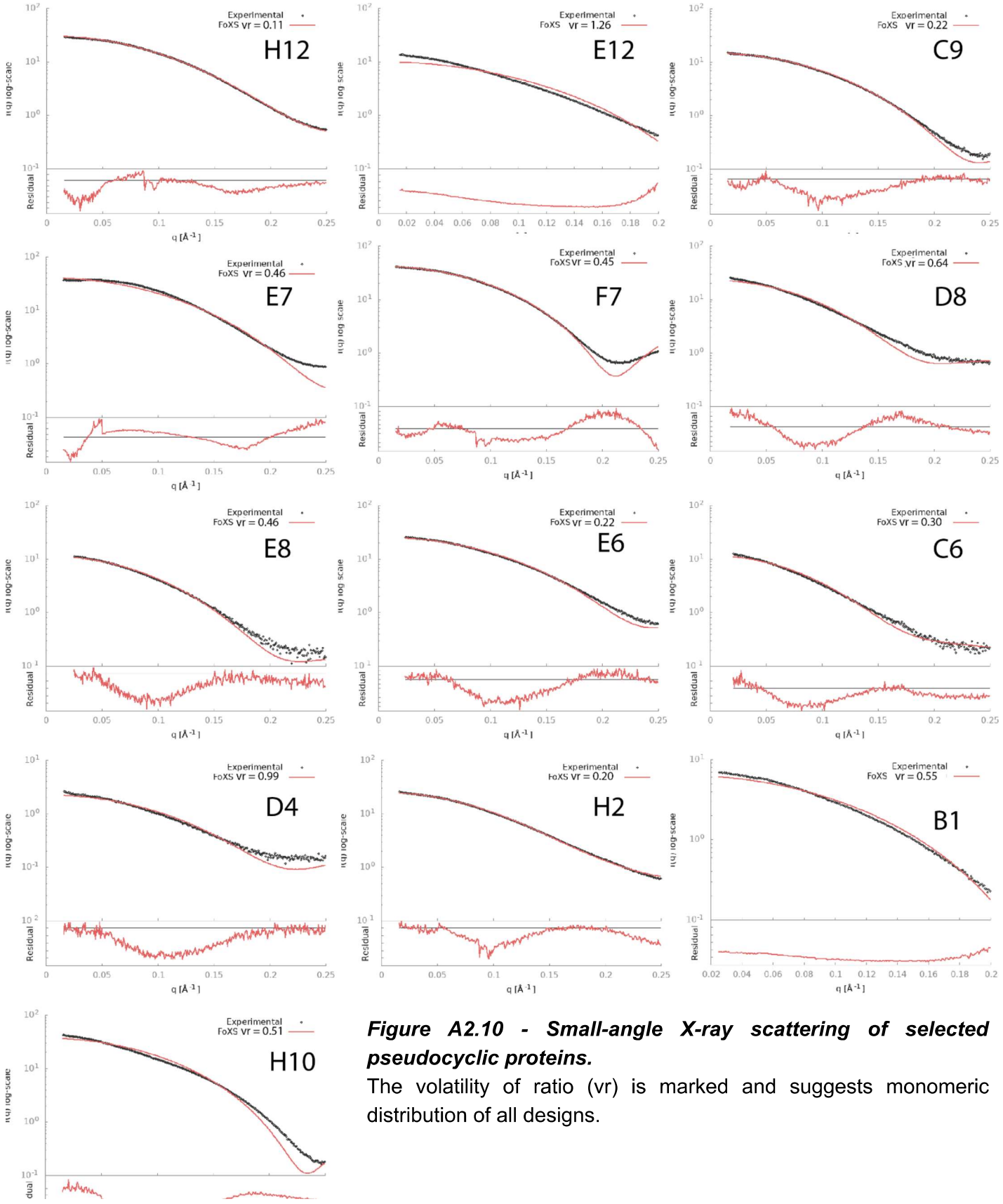


Figure A2.10 - Small-angle X-ray scattering of selected pseudocyclic proteins.

The volatility of ratio (vr) is marked and suggests monomeric distribution of all designs.

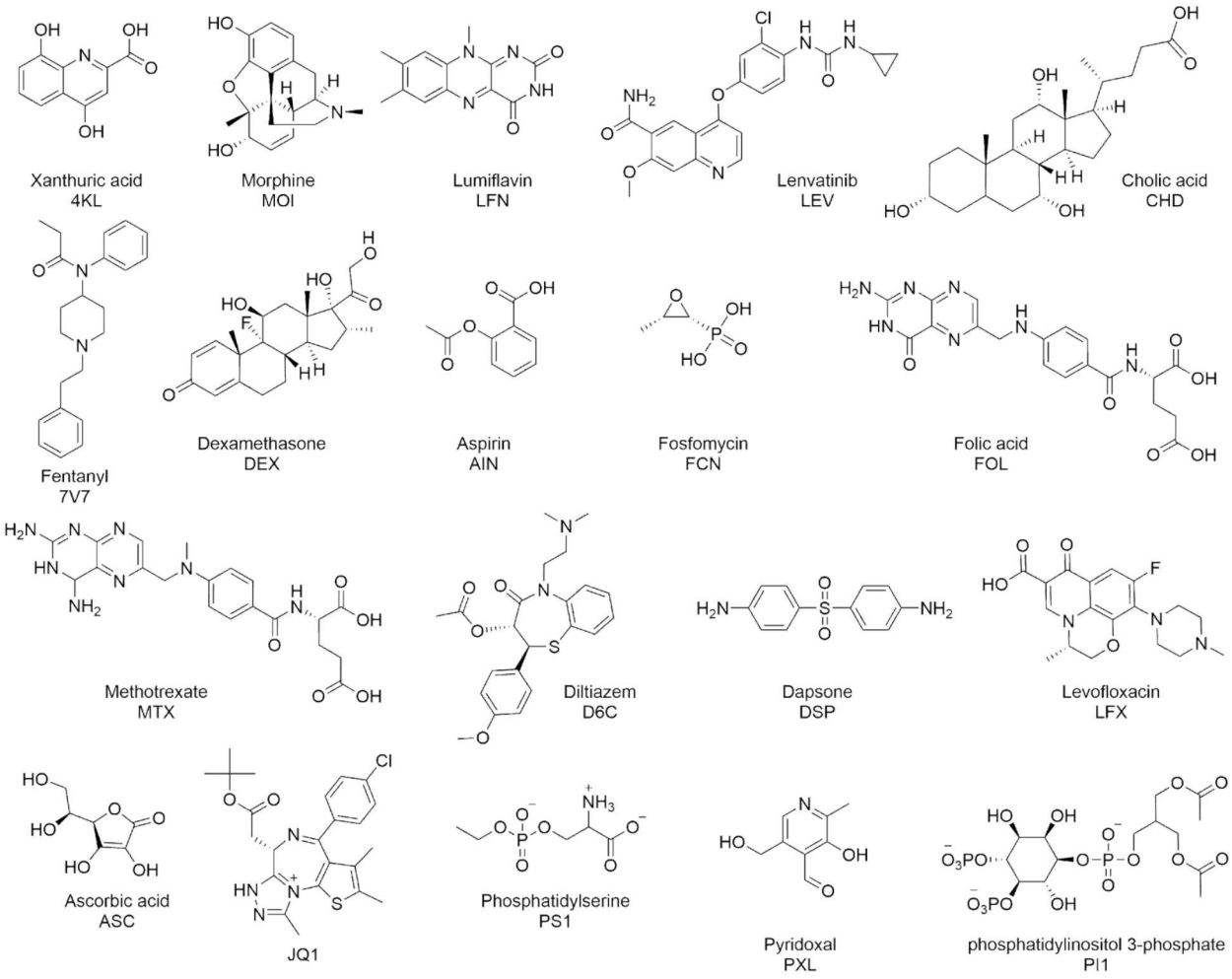


Figure A2.11 - Ligand Structures Used For Docking

The structure of ligands used for docking and design in this study. The three-letter names used in Fig 6.2 are indicated alongside a full chemical name.

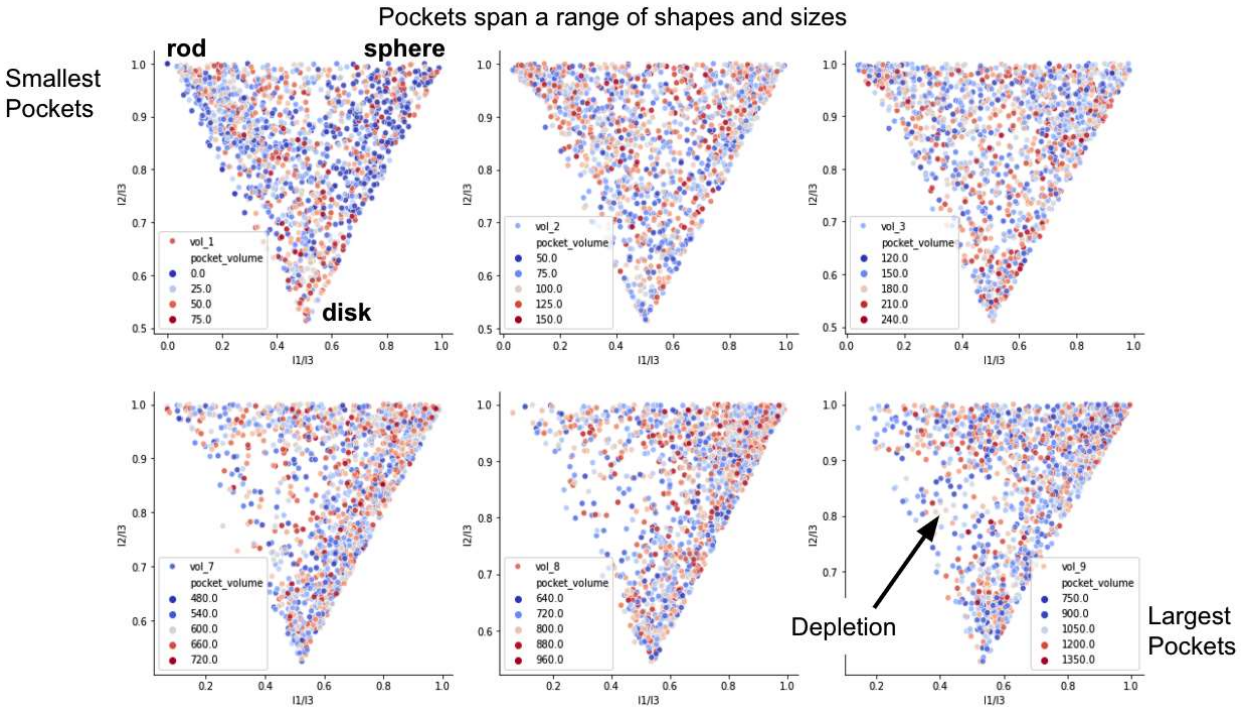


Figure A2.12 - Pocket Shapes

Pockets were detected for 21,021 designs with sequences converted to poly-alanine to show max possible pocket size. Plots show the I2/I3 vs I1/I3 ratio which dictate the pocket shape. We show plots for pockets binned from small (75 cubic Å) to large (1400 cubic Å).

Appendix 3 - **ABR Sequence Design and Rosetta Metrics Detailed Methods**

FastDesign and Sequence design procedures

Rosetta sequence design was conducted through a multistep process with a number of small modifications to the default parameters. The FastDesign procedure can be broken down into three subsequent design protocols. Each design protocol consists of MCMC sequence sampling and backbone angle minimization steps. Mutation permutations are rapidly sampled through “packing” steps, where coarse grained atom coordinate trees map relative self-compatibility of various sequences, and minimization is conducted via explicit gradient descent of Rosetta Energy via changes to backbone torsion angles and residue chi angles. Detailed explanation of the three packing/minimization protocols follow.

“Sandpig” QuickPack

A fast packing and minimization round was conducted with only the set of residues with one-letter codes SANDPVG. This procedure helps to relax poly-leucine and poly-alanine backbones into geometries more compatible with FastDesign sampling, while keeping backbone hydrogen bonds satisfied. After sandpig packing, glycine and proline positions installed at this step are fixed within the structure for the remainder of design. This procedure is borrowed from[34].

Symmetric Layer Design

Symmetric design was conducted using the “LayerDesign" framework in Rosetta. Layers were chosen based on the number of neighbors of each residue. Neighbors are residues whose neighbor atoms (generally the beta-carbon for canonical residues and Ca for glycine) are within a threshold distance. High neighbor count residues are considered “core”, low count are considered “surface”, and values between the two are designated as “boundary”. In the LayerDesign protocol, certain residue identities are forbidden for use in packing. Core positions forbid use of polar residues, while surface positions forbid the use of hydrophobic residues (except proline). Boundary positions use a combination of the two, but are forbidden from using bulky residues, such as tryptophan and isoleucine. As well as long chain polars such as lysine, arginine, and glutamate. In addition, our layer design scheme featured discrimination between secondary structure features, and forbade use of “helix breaking” residues in helices (Asparagine, Aspartate, proline) and bulky residues in loop regions. Additionally, a custom relaxscript file with adjusted reference weights was used to reduce heavy bias within Rosetta to assign histidine to every position in parallel beta strands. Standard reference weights were used during final FastRelax and scoring.

Symmetric layer design was conducted for 4 cycles of packing and minimization. The “approx_buried_unsat” term was used during packing to penalize buried, unsaturated polar residue head groups and backbone polar heavy atoms.

This LayerDesign protocol was conducted with the repeat symmetric constraints in Rosetta, where residue identity and geometry between repeat units is exactly fixed and accounted for in scoring.

Surface Design

Surface positions were designed with an identical procedure to that described for symmetric layer design. The key difference was the absence of symmetric constraints and the residues marked as designable. In this protocol, only surface residues and boundary residues on strands were marked as designable. This allowed for the sequence modification of terminal repeats to break symmetry, as well as improving residue diversity on the strand surfaces. Repeat parallel strands have closely spaced beta-carbons, and so symmetric constraints in Rosetta FastDesign at time of writing result in sheets consisting entirely of a single residue identity (often Histidine) in order to avoid clashes.

Scoring and Filtering

A variety of score terms in rosetta were applied to discriminate promising models for investigation. Differences in these scores between DL designs and Rosetta designs are summarized in Figure A6.1. Rosetta designs were chosen by selecting the models with the lowest number of buried, unsaturated heavy atoms and the best scores among “worst9mer” (a fragment quality metric), sidechain shape complementarity, and rosetta score per residue. For each of the 7 characterization rounds, approximately 50,000-100,000 backbone models were generated, of which only approximately 500-2,500 were suitable for design. Between 5 and 25 sequence design trajectories were run on each backbone, as computational resources allowed.

A target number of designs per design round was set at 10-20, as this allowed for best throughput in terms of various lab apparatus availability. In one case, this allowed for the investigation of ABR34, which has a designed register shift in its strands, ostensibly discouraged by the Ca constraints during design, because of a smaller characterization round.

Score ranges and differences between Rosetta designs, AF2 predicted models of the Rosetta designs, and DL design are presented in Appendix 6.

Appendix 4 - **Detailed Methods for ABR Proteins**

Expression and Purification

We obtained cloned gene plasmids and Golden Gate compatible gene sequences encoding the structures of our design models. For the gene sequences, we conducted golden gate cloning in-house. Genes of interest were tagged with hex-histidine tags for purification. Golden Gate genes were transformed into a vector containing a cleavable tag (SNAC tag) with downstream hexahistidine tag.

We transformed these plasmids into BL-21 DE3 E. coli and expressed protein via auto-induction. Autoinduction was conducted in 50 ml cultures with 6 hour outgrowth at 37C from either a 50% glycerol stock prepared from an overnight culture or from an agar plate colony, followed by a 24 hour expression before culture harvest at 18C. TerrificBroth (TB) with 5052 and Kanamycin was used as auto-induction media. Similar yields were observed between different expressions from glycerol stocks, single colony inoculations, and small variations in expression time at 18C. Cultures were harvested by centrifugation in 50ml conical tubes, and cell pellets were frozen until purification. Lysis was conducted by sonication on ice in 20 mM Tris, 100mM NaCl, and 50 mM Imidazole buffer at pH 8 with protease cocktail tablets used according to manufacturer instructions. Sonicated lysate was chilled before centrifugation. After centrifugation at 20,000g for 40 minutes, lysate supernatant was purified by immobilized metal affinity chromatography (IMAC) with nickel NTA resin by batch binding at 4C followed by elution with 20mM Tris, 100 mM NaCl and 500 mM imidazole. Eluates were stored at room temperature overnight before fast protein liquid chromatography (FPLC). On the subsequent day, eluates were centrifuged at minimum 20,000g for 12 minutes, and supernatant was passed through a 0.2 micron filter before FPLC. For initial characterization using SEC and CD, proteins were eluted with 2 CV of elution buffer (20 mM tris, 100 mM NaCl, 500 mM Imidazole, pH 8) and purified on a superdex 75 increase 10/300 GL column or superdex 200 increase 10/300 GL connected to ÄKTA protein purification systems in TBS buffer (25 mM Tris, 100 mM NaCl, pH 8).

Circular dichroism characterization of selected proteins

Circular dichroism spectra were measured with a Jasco J-1500 CD spectrometer. Samples were diluted to concentrations of approximately 0.25 mg/mL (individual samples ranging 0.1 - 0.5 mg/ml) in 25 mM phosphate buffer titrated to pH 8 by combination of dibasic and monobasic sodium phosphate according to a standard table, and a 1-mm path length cuvette was used. The CD signal was converted to mean residue ellipticity by dividing the raw spectra by N × C × L × 10, where N is the number of residues, C is the concentration of protein, and L is the path length (0.1 cm).

We conducted a thermal temperature melt from 25C to 95C slowly ramping over 3 hours with fixed temperature at 25, 55 or 65, and 95 for spectral scans. We observed spectra consistent with a structure having both alpha-helical and beta-strand character.

Appendix 5 - **ABR Additional SEC and CD Data**

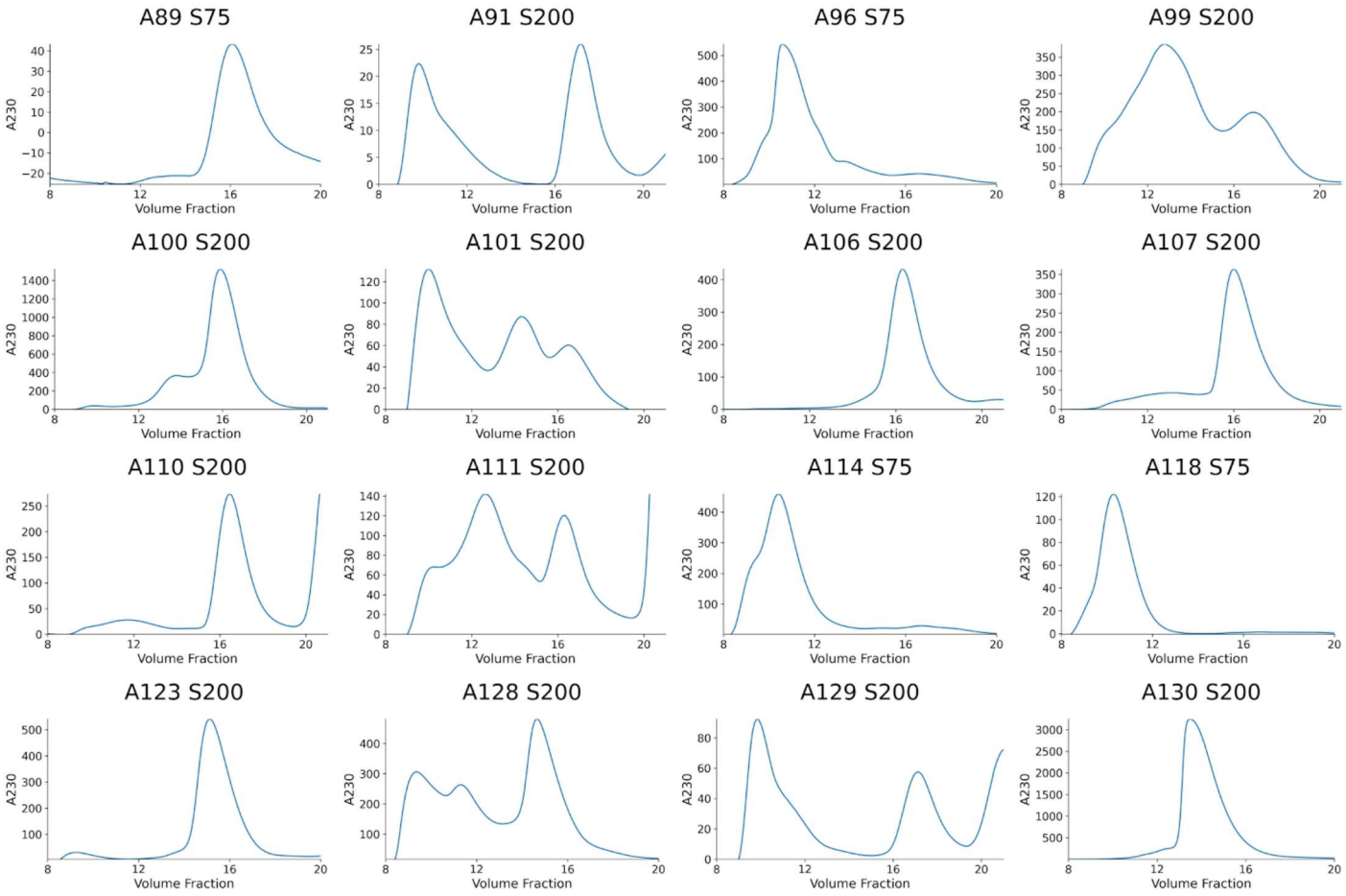


Figure A5.1 - Selected Additional ABR SEC traces

Select purification and analytical traces from the column indicated for each sample. Each protein was selected either because the peak appears at the expected retention volume, or a high-concentration peak is present at a retention volume greater than the column void volume.

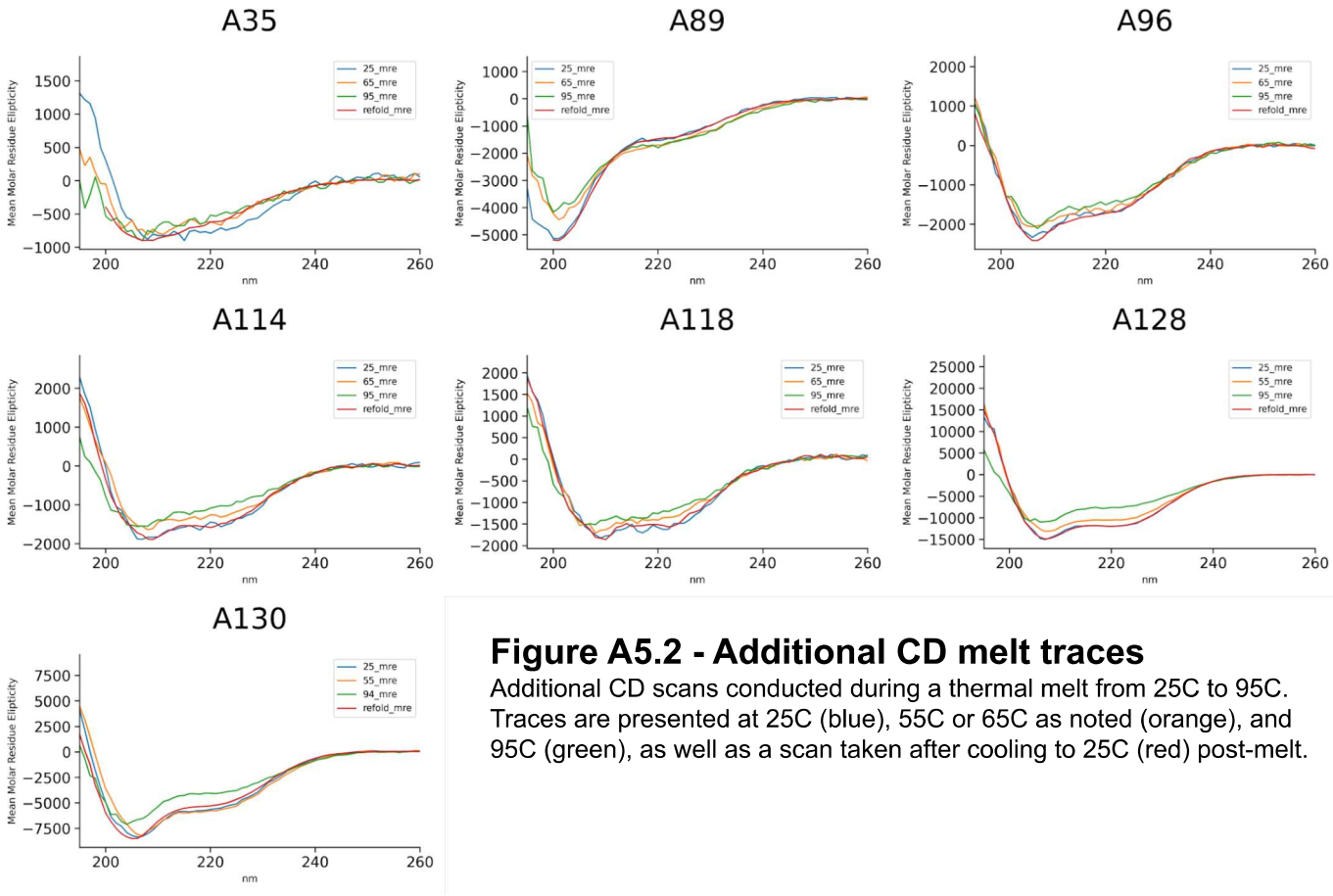


Figure A5.2 - Additional CD melt traces
Additional CD scans conducted during a thermal melt from 25C to 95C. Traces are presented at 25C (blue), 55C or 65C as noted (orange), and 95C (green), as well as a scan taken after cooling to 25C (red) post-melt.

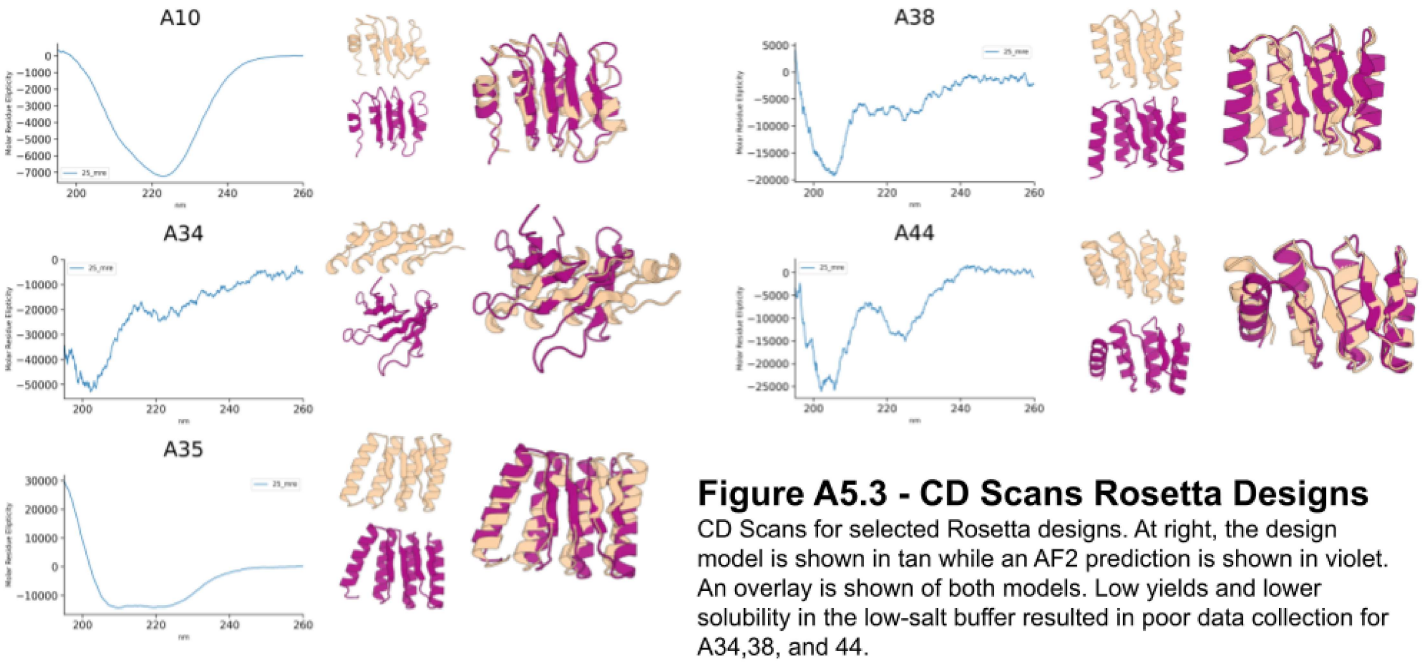


Figure A5.3 - CD Scans Rosetta Designs
CD Scans for selected Rosetta designs. At right, the design model is shown in tan while an AF2 prediction is shown in violet. An overlay is shown of both models. Low yields and lower solubility in the low-salt buffer resulted in poor data collection for A34,38, and 44.

Appendix 6 - **Rosetta Score Trends**

Data Presentation Preface

For all box and whisker plot figures in this appendix: central line shows median, box shows interquartile range, and whiskers show range, excluding outliers. All outliers are shown as diamonds.

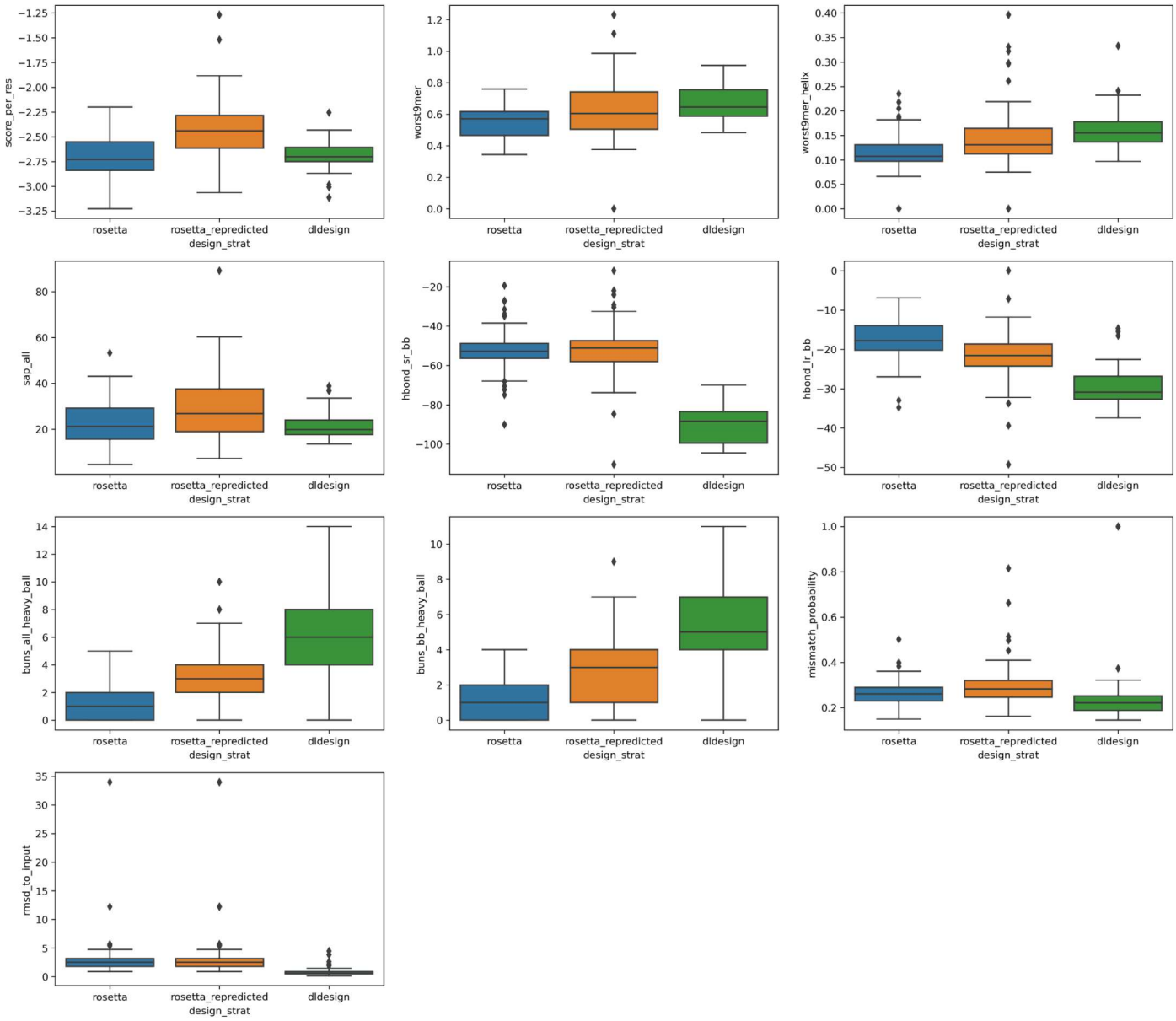


Figure A6.1 - Rosetta Scores for Designs by Model Type

“rosetta” represents the pool of designs (all constraint groups together) designed entirely without DL methods, while dldesign represents models designed with DL. “Rosetta repredicted” are the Rosetta model sequences predicted with AF2, and the best pLDDT model scored with Rosetta. Of particular note, hbond_lr_bb and hbond_sr_bb are notably better in dldesigns, metrics of long-range and short range hydrogen bonds, respectively, even though fragment quality and sequence choice appears to suffer (worst9mer and mismatch probability). Overall rosetta score is roughly equivalent between dldesigns and rosetta designs, but AF2 predictions generate models which score more poorly on both pLDDT and Rosetta Energy than dldesigns overall. Notably, buns_all (buried, unsaturated polar atoms) is higher for both repredictions and dldesign.

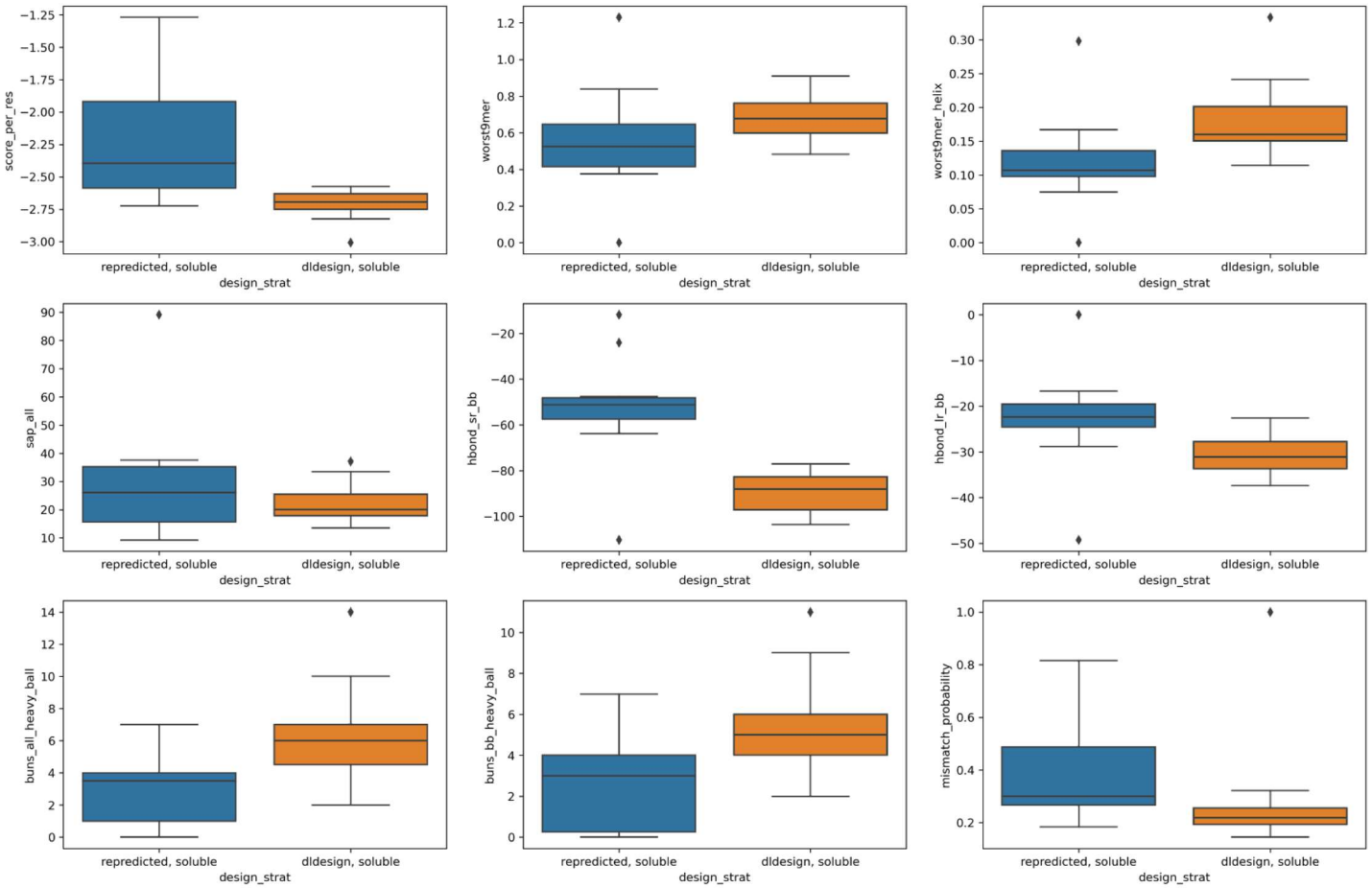


Figure A6.2 - Rosetta Scores of soluble models

Rosetta scores of soluble dldesign models and soluble Rosetta designed models are shown. Mismatch probability of single sequence psipred prediction is substantially smaller for soluble designs, and short and long range backbone hydrogen bond energies are improved (hbond_lr_bb and hbond_sr_bb).

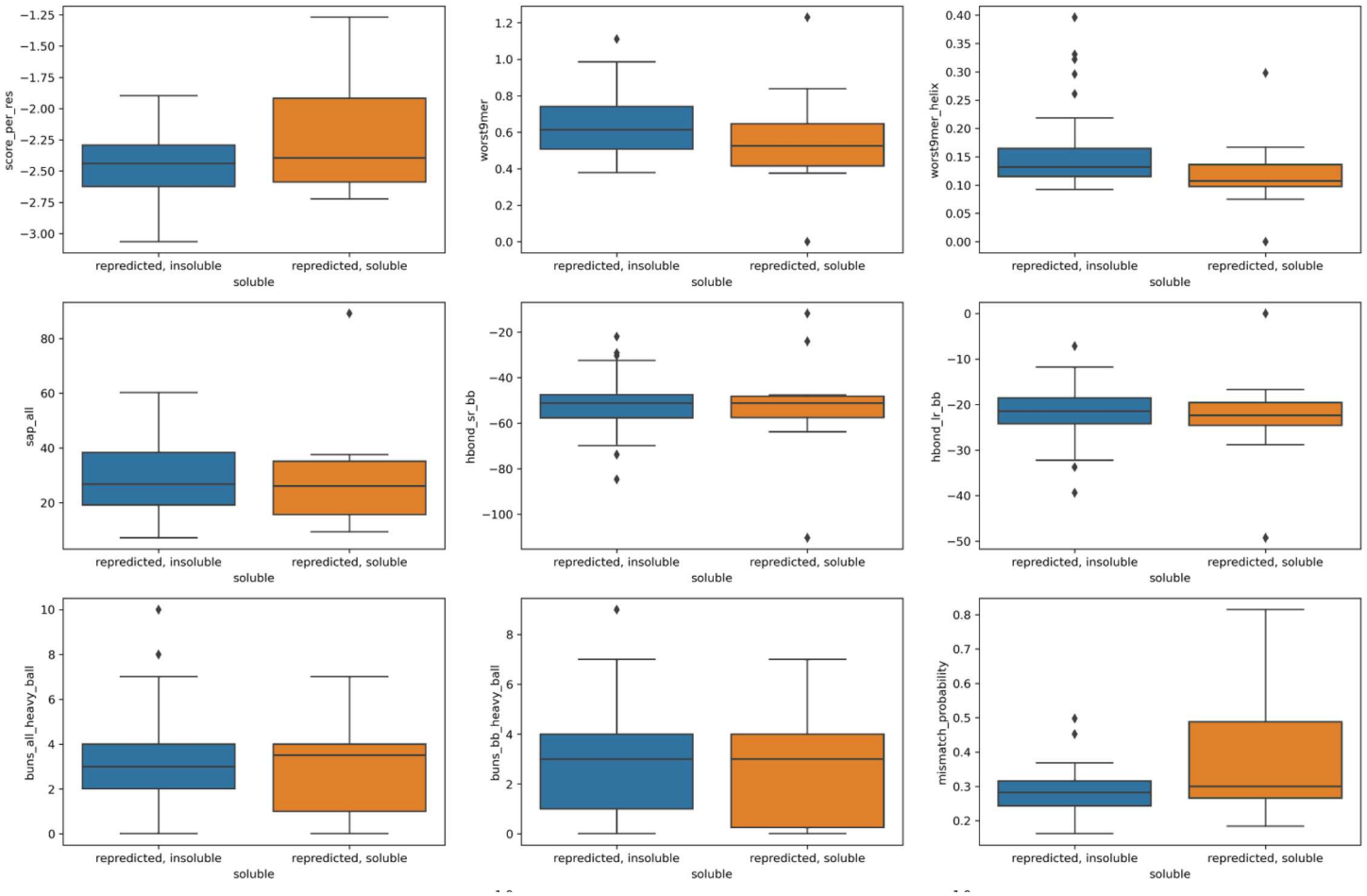


Figure A6.3 - Rosetta Scores for AF2 predictions of Rosetta designs found to be insoluble vs soluble

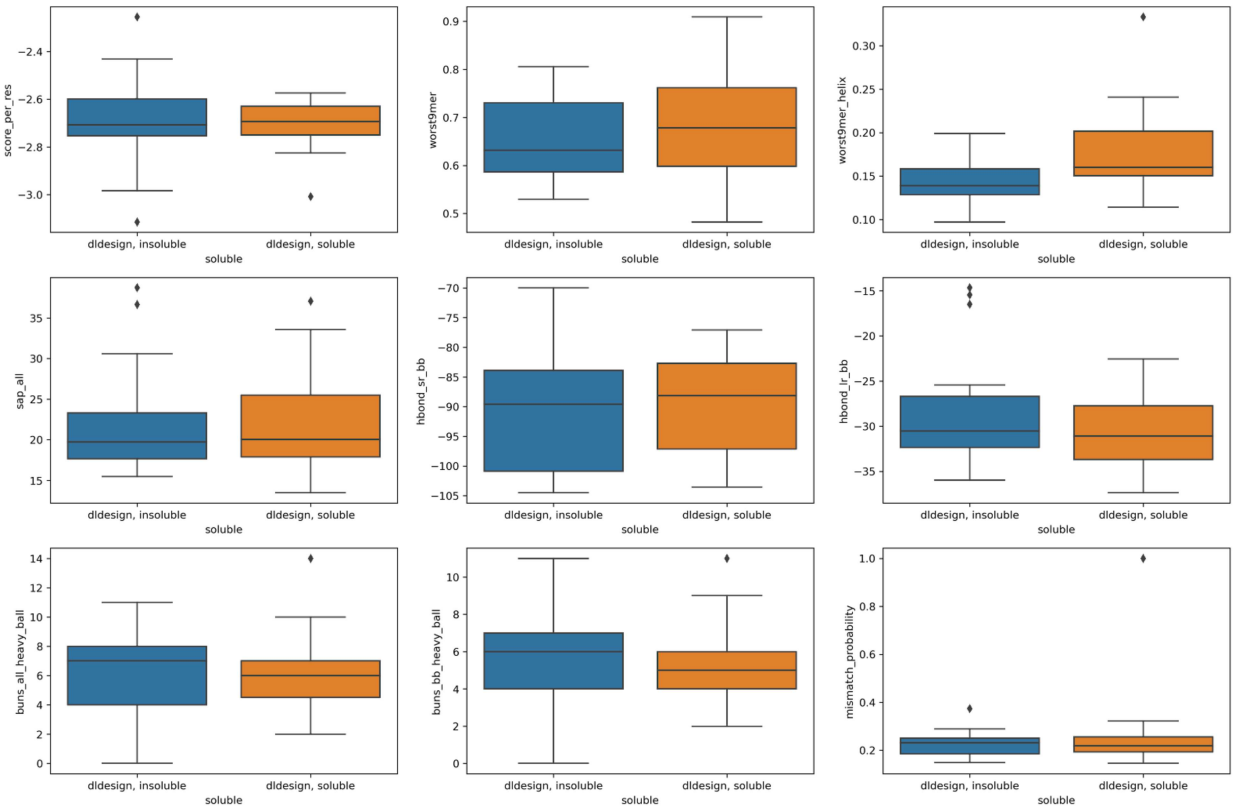


Figure A6.4 - Rosetta Scores for AF2 predictions of dldesigns found to be insoluble vs soluble

Appendix 7- **pLDDT of Capped DL-Design ABRs**

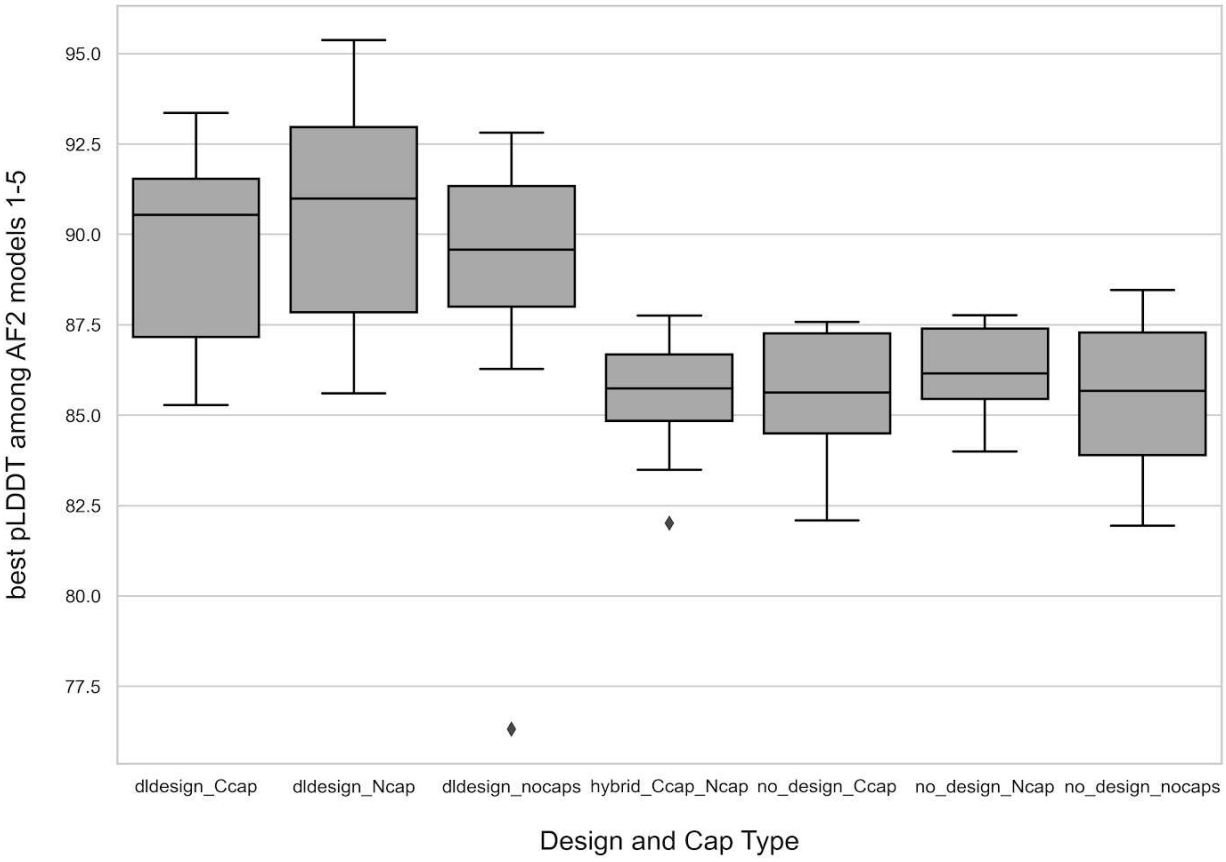


Figure A7.1 - pLDDT comparison for cap deletion experiment

between fully MPNN designed models, MPNN & Rosetta hybrid designed, capped models, and capless hybrid design models. Fully MPNN designed models have higher pLDDT overall. Presence of caps does not independently affect pLDDT.

Appendix 8 - **Superhelical Parameters of Repeats**

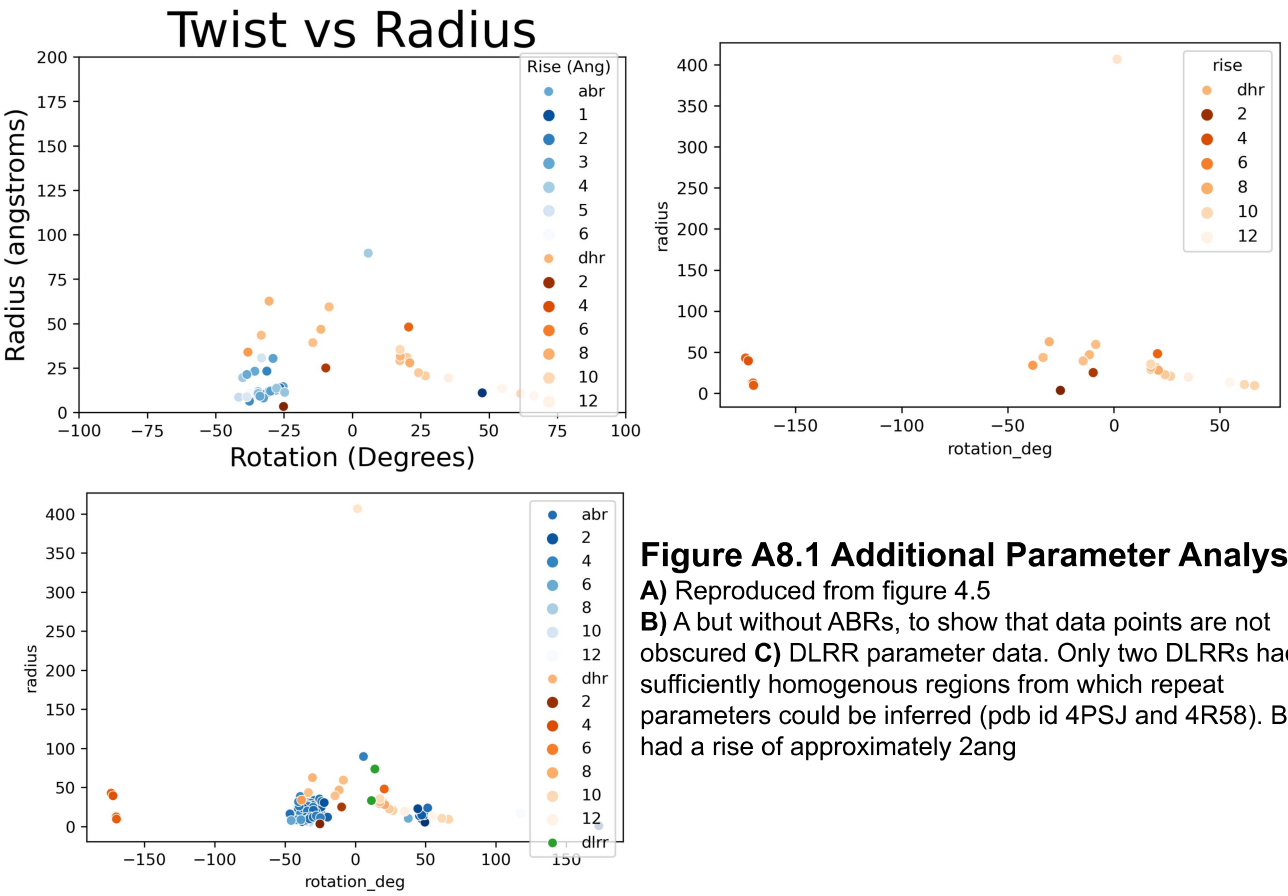


Figure A8.1 Additional Parameter Analysis
A) Reproduced from figure 4.5
B) A but without ABRs, to show that data points are not obscured **C)** DLRR parameter data. Only two DLRRs had sufficiently homogenous regions from which repeat parameters could be inferred (pdb id 4PSJ and 4R58). Both had a rise of approximately 2ang

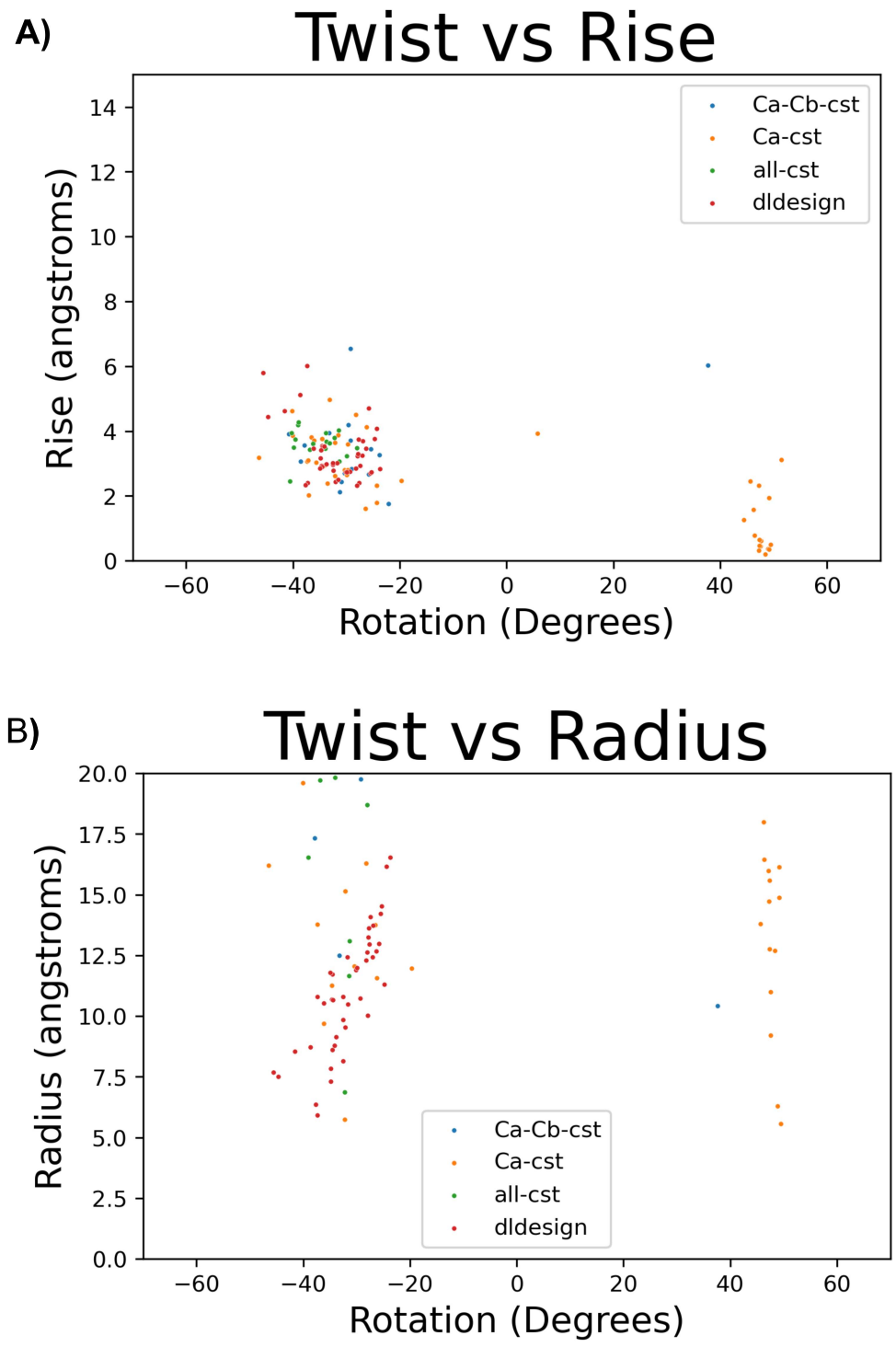


Figure A8.2 Constraints reduce the search space

Increasing geometric constraints visibly reduce the designed parameters of the whole design pool when plotted as A) rotation vs Rise B) Rotation vs Radius

Appendix 9 - **ABR 10 Additional Data**

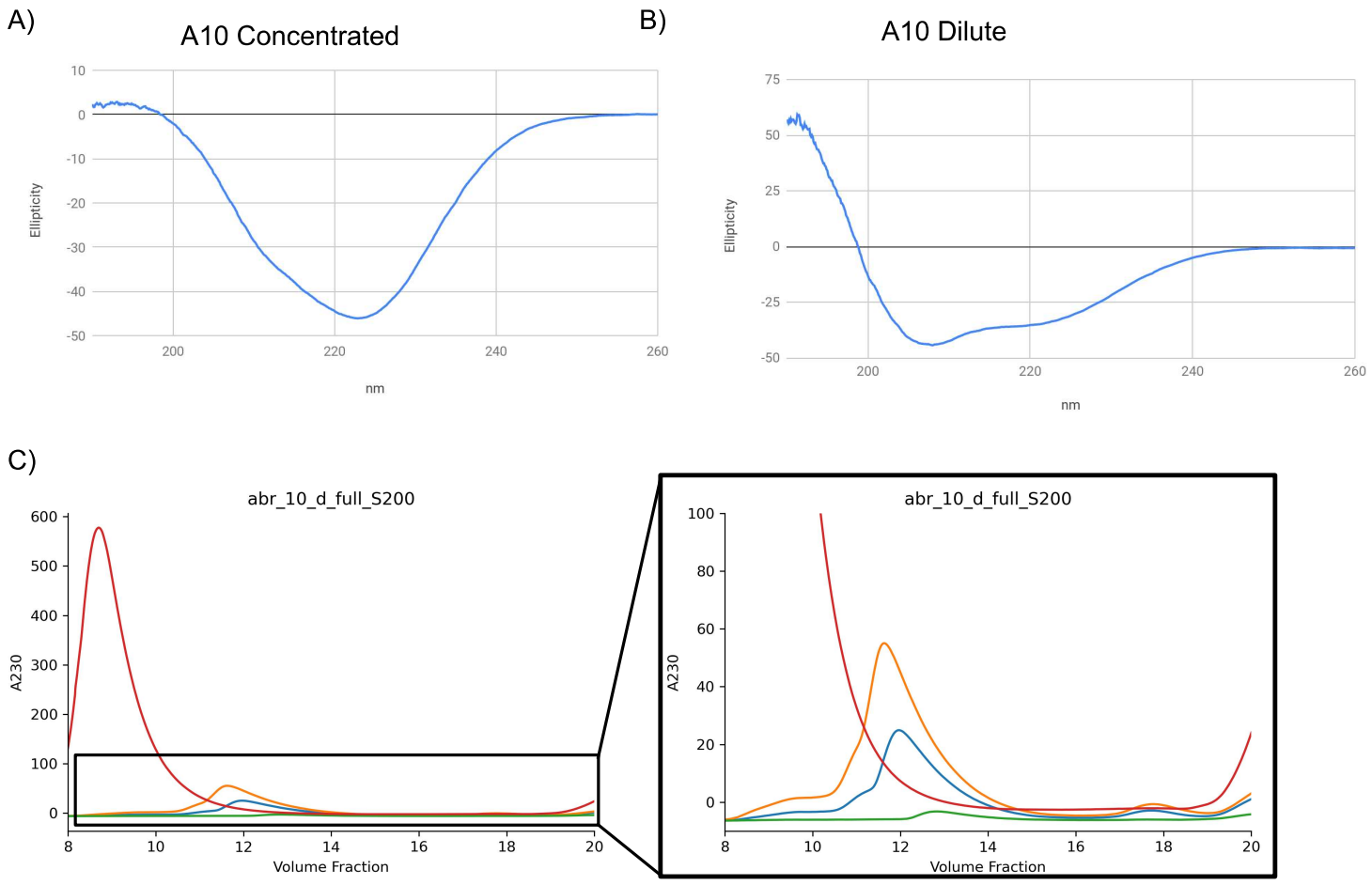


Figure A9.1 - ABR 10 concentration dependent behavior

A) CD spectrum at 25C for ABR10 harvested at high concentration (retention ml 10 on S75) vs at low concentration **B)** (retention ml 12). **C)** A sample of ABR 10 collected and concentrated until it formed a soluble aggregate (the void volume of S75 columns is roughly at 8ml) then was serially diluted to 1/5th, 1/10th, and 1/100th of the original concentration. The dilution series' SEC traces are shown overlaid as red, orange, blue, and green respectively.