De novo design of buttressed loops for diversifying protein functions

Hanlun Jiang

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

David Baker, Chair

Barry Stoddard

Philip Bradley

Program Authorized to Offer Degree:

Biochemistry

University of Washington

Abstract

De Novo Design of Buttressed Loops for Diversifying Protein Functions

Hanlun Jiang

Chair of the Supervisory Committee:
David Baker
Department of Biochemistry

Loops play essential roles in protein functions such as protein-protein interactions, ligand recognition and metal binding. However, designing structured and functional loops remains to be a long-standing challenge for protein engineering. In contrast to alpha helices and beta sheets, loops are substantially more disorganized and structurally less defined. Such intrinsic flexibility makes it difficult to predict or design the loop structures. One feasible strategy for designing loops is to stabilize them by creating extensive interactions between loops and the rest of the protein. In particular, loop buttressing, in which loops are interacting and stabilizing each other, is a trick that has been successfully used by nature. We aim to design and test buttressed loops that adopt rigid conformations and can be further engineered for functions. Inspired by the fold of ankyrin repeat proteins, we developed a computational design protocol that combines canonical secondary

structural motifs into structured loops with kinematic closure algorithms to diversify the fold of designed helical repeat proteins. The designed loop conformations were confirmed by both the predictions of AlphaFold and the crystal structures. The scaffold proteins, installed with buttressed loops, were then functionalized by designing new interfaces for mediating protein-protein and protein-peptide interactions. To demonstrate the generalizability of our method, we further applied the loop designing protocol to diversify a *de novo* TIM barrel. These designs demonstrated that the novel binding grooves or pockets formed between the buttressed loops and the original protein scaffolds provide a unique category of protein interfaces for engineering new functions.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

It would not be possible for me to complete the journey of graduate study without the kind support from so many people. Thanks to them, I not only learned so much about protein design and science in general, but also became more resilient to obstacles and more confident to take on challenging problems.

I would like to thank my thesis advisor David Baker, for encouraging me to work on the challenging but fascinating problem of loop design. I enjoyed every discussion we had and was constantly benefited from his suggestions on the project. I am truly grateful for his support and guidance for my scientific career.

A few colleagues in the Baker lab helped me greatly to kick start the project. Ian Haydon introduced me to the TIM barrels and we had great fun looking at the loop structures during my rotation. Daniel-Adriano Silva gave me great guidance towards the end of my rotation and at the beginning of my thesis project. Jorge Fallas and George Ueda walked me through the protocols of protein expression and characterizations. They were always happy to give me advice when I needed others' input for trouble shooting my failed experiments. TJ Brunette and Derrick Hicks taught me how to make DHRs. I enjoyed working with Shane Caldwell on diversifying TIM barrel designs and he gave me many helpful tips for crystallography.

I benefited so much by working with people with strong experimental expertise. Kevin Jude from the lab of Chris Garcia set up hundreds of crystal trays and solved all the structures I described in this thesis. I would not be able to finish my graduate study without Kevin. Aerin Yang in the Garcia lab performed yeast display assays to characterize my designed proteins. Alexis Courbet and Ryan Kibler helped me ship samples for native mass spectrometry and SAXS data collection.

Our collaborator Susan Tsutakawa and her colleagues at LBNL helped me run the SAXS samples and analyzed the scattering data. Xinting Li and Lauren Carter performed in-house mass spectrometry and SEC-MALS for all of my purified proteins. Michael Murphy, Kandise VanWormer and Austin Smith helped me numerous times when I needed to locate reagents or had problem with instruments in the lab.

Towards the end of my graduate study when I started developing applications for my designs, I learned a lot from the following people who happily shared with me their design expertise. Wei Yang, Buwei Huang, Nate Bennett and Brian Coventry taught me how to design protein-protein interfaces. Kejia Wu gave me many insightful suggestions on the design of repeat peptide binding proteins. Susana Vazquez, Isaac Lutz, Phil Leung and Florian Praetorius provided a lot of guidance and help with the experimental characterization of peptide binding proteins.

I would like to thank my thesis committee members: Phil Bradley, Frank DiMaio, Kelly Lee and Barry Stoddard, for their insightful advice on my project and helpful suggestions on my scientific career.

I want to thank my graduate program BPSD for the precious opportunity and their support for my study. Erin Kirschner helped me throughout my study at UW and she was always there for me whenever I needed help to navigate the life of graduate school. Christina Larmore and Ning Zheng organized BPSD student seminar, a great platform for me to learn the fascinating research conducted by fellow graduate students. Chip Asbury gave me some very helpful career advice. Michelle Matsunaga and Zari Magness helped me so much with all kinds of paper work and meeting arrangement.

I appreciate the lunch/dinner group with Longxing Cao, Qian Cong, Gyu Rie Lee, Hahnbeom Park, Minkyung Baek, Ivan Anishchanka, Vasilina Anishchanka, Wei Yang and Zhe Li, with whom I learned a lot about scientific research and life in general.

Finally, I would like to thank my family: my wife Fátima Pardo Avila, my parents Xuan Dong and Guoyuan Jiang, mis suegros Eusebio Pardo y Socorro Avila, tia Dolores Avila, Eren y Erik, for their unconditional support and love throughout this journey.

Chapter 1. Developing Computational methods for Designing Buttressed Loops on Repeat Proteins

1.1 Introduction

1.1.1 *Computational design of structured protein loops and their functions*

Loops play essential roles in protein functions such as protein-protein interactions, ligand recognition and catalysis[1-7]. Although most loops on protein surface are flexible, they often adopt specific conformations when interacting with proteins or small molecules. Therefore, designing structured loops that are organized in their functional conformations can potentially enhance the binding affinity by lowering the entropic cost.

Since most loops in natural proteins are short[8, 9], usually with less than five residues, principles for designing short loops have been formulated by mining the protein data bank (PDB)[10]. By mapping adjacent secondary structure elements to their tertiary features, Koga *et al.* generated a set of rules to describe the geometries of ideal loops connecting the secondary structures[11]. These rules were further extended by Lin *et al.* with more detailed description of the backbone torsion angles for each residue in the loops[12]. Guided by these principles, protein designers today can confidently link adjacent secondary structures with short structured loops at high success rate[13-17].

While the principles of short loops have led to great success in the design of many highly stable *de novo* proteins consisting of alpha helices or/and beta sheets, less progress was made in designing

long loops. An early success in long loop design was the engineering of a rigid seven-residue loop in monomeric triosephosphate isomerase (TIM)[18]. In this work, the originally flexible eight-residue loop was redesigned by iteratively generating new loop models with Monte Carlo simulations and manually selecting conformations from low-energy models. As one of the early attempts of automated computational loop design, Hu *et al.* redesigned a 12-residue loop in fibronectin type III (FN3) domain by grafting native protein fragments of same length from PDB[19]. The sequences of the loops were then redesigned to optimize the stability. Out of three designs that were experimentally tested, two were crystallized, one of which had the designed loop resolved in a conformation very close to the design model (RMSD=0.46 Å). Using a similar design strategy, MacDonald *et al.* successfully generated and inserted eight-residue *de novo* hairpins into synthetic beta-solenoid proteins[20]. Despite of these early successes, none of the loops designed in the aforementioned studies carry any functions.

There has been limited but important progress in designing functional loops. Murphy *et al.* demonstrated that the enzyme specificity can be altered by redesigning loops at the active site[21]. In particular, a four-residue loop was used to replace a seven-residue loop in wild-type human guanine deaminase, which resulted in a 2.5e6-fold change of substrate specificity from guanine to ammelide. Later the Fleishman group developed an antibody-specific design protocol[22]. This method takes the advantage of abundant structural information on antibody scaffolds and builds loops for complementarity determining regions (CDRs) based on the conformation clusters generated *a priori*. The designed antibodies were structurally verified by X-ray crystallography and shown to bind target protein at mid-nanomolar affinities[23]. However, being highly specific to the antibody scaffolds, this method is difficult to be generalized for loop design in other protein

scaffolds. More excitingly, Krivacic *et al.* recently developed hybrid loop design methods which led to reasonably accurate and functioning 11/12-residue loops[24].

1.1.2 *Computational methods for sampling and predicting loop conformations*

Most loop sampling algorithms can be divided into two categories: template-based methods, where native loop fragments that have desired shapes are grafted or recombined, and template-free methods, where a closure algorithm is used to generate loop conformations without the knowledge of native loop conformations[6, 25, 26]. Since the template-based methods use the loop fragments that already exist in PDB, the loop conformations generated are usually geometrically favorable and the original sequences of the fragments often encode the structures of loops[6, 27]. This is particularly helpful for the following sequence designs. However, limited by the available native fragments, these methods often suffer from insufficient sampling of conformational space. In contrast, template-free methods such as cyclic coordinate descent (CCD)[28] and kinematic closure (KIC)[29-31] excel in exploring loop conformational space uncharted by PDB, but usually demand a reliable scoring function and an efficient filtering scheme to select loop backbone conformations that are not only geometrically reasonable but also likely to be stabilized by the following sequence design.

Recent success in developing fast conformational sampling protocols and sequence design strategies have motivated us to revisit the problem of loop design. In particular, new hybrid algorithms that combine the advantages of both template-based and template-free methods have emerged and demonstrated superior sampling efficiency than both types of methods when used alone[24, 32, 33]. The fast search algorithm for native hydrophobic interactions[34] and exhaustive hydrogen bond network sampling protocol[35] provide the foundation for accurate

design of the interactions between loops and the rest of the protein, which is crucial for structural stabilization.

The latest advances in deep-learning-based structure prediction further enable effective *in silico* validation of loop designs. Both AlphaFold[36-38] and RoseTTAFold[39] have shown great promise in accurate prediction of protein folds with complex architecture. Many successfully predicted protein structures contain long loops that were hard to model by previous methods. This progress provides us with a great opportunity for rigorously assessing loop designs before experimental characterizations.

1.1.3 *Loop buttressing on repeat proteins as a strategy for designing multiple structured loops for new functions*

While redesigning a loop at the active site or molecular recognition interface can lead to new protein functions, *de novo* design of new protein scaffolds often requires placing multiple structured loops at the same site. Many native proteins, such as antibodies[1, 4, 23] and TIM barrels[40, 41] achieve versatile binding modes by organizing multiple loops on a highly conserved, stable scaffold. Crystal structures of these proteins reveal intricate hydrogen bond networks that lock the loops together[1, 41]. This suggests the problem of multiple loop design cannot be dissected into several tasks of single loop design, but should be instead treated as a combinatorial task where both individual loop stability and loop-loop compatibility should be considered at the same time. As a result, the combined sampling space can easily exceed the capability of the state-of-the-art supercomputers, making this problem particularly challenging.

One way to reduce the combined loop conformation space is to harness the internal symmetries of repeat proteins. A great example by nature is the ankyrin fold[42, 43]. Ankyrins contain 4-29

repeats of helix-loop-helix-turn units arranged in right-handed solenoid shapes[43, 44]. Due to their repetitive nature, loops in ankyrins are generally of same lengths and adopt a highly-conserved beta hairpin conformation. Each pair of neighboring loops are buttressed by sharing the same set of ASN- and/or HIS-facilitated bidentate hydrogen bond networks. These networks are well conserved across different species and ankyrins of various functions[43, 45]. Such unique buttressing feature inspires us to explore the possibility of solving the multiple loop design problem by installing buttressed loops of identical conformations in repeat proteins.

1.2 Results

Our repeat protein loop sampling method consists of two stages: single loop sampling and buttressed loop sampling.

For the single loop sampling stage, we developed a hybrid sampling method that combines selected motif fragments and kinematic closure algorithm to efficiently exploring hairpin-shape loop conformations. Two libraries of motif fragments, native beta turns and canonical helical capping motifs, were curated. By incorporating these fragments in the loop closure algorithm, we biased the sampling towards the conformations that resemble beta hairpins which are geometrically compatible with buttressing. A gap was created on one repeat unit of the scaffold repeat protein by deleting a short loop. During the sampling simulation, long loops were then generated to reconnect the gap. Each generated loop backbone structure was filtered by low steric clashes, but also selected by having at least two backbone-to-backbone hydrogen bonds. These hydrogen bonds often led to a hairpin-shape loop and provided the loop with interval stabilization.

At the buttressed loop sampling stage, the single loops generated from the previous stage were propagated to each repeat unit by grafting. Subsequently, the loops were selected based on if buttressing bidentate hydrogen bonds can be built between neighboring loops. Specifically, one round of fast sequence design was performed for each residue on the loop, using a set of selected residues compatible with bidentate hydrogen bonds (ASN, ASP, HIS, GLU and GLN). Loop conformations with interloop bidentate hydrogen bonds were used for full-scaffold sequence design.

To evaluate the efficiency of our method, in the following sections, we compared our method with template-based and template-free algorithms to test the efficiency of native beta hairpin loops. Using an idealized ankyrin as an example, we also demonstrated the enhancement of predicted folding by employing loop buttressing in repeat proteins.

1.2.1 *The hybrid method efficiently recapitulated the ankyrin-like loops*

To benchmark our method, we tested its performance on sampling and recovering a 13-residue loop from an idealized ankyrin structure[46]. Specifically, the buttressed, beta hairpin loop (G59-T71) was deleted and modeled using our method, a templated-based method or a template-free method. For the templated-based method, we used the DirectSegmentLookupMover from Rosetta[47, 48]. Given a gap in a protein structure, this mover computes the geometric transformation between the two residues on each side of the gap and quickly finds loops with the closest matches to that transformation via a precompiled hash table. For the template-free method, we used Rosetta's GeneralizedKIC[31, 48] module, a KIC-based algorithm that is generalized for loop perturbations, selections, filtering and incorporation of non-canonical amino acids for protein loop sampling.

For each method, 1000 loop conformations were generated and compared to the structure of the loop from the idealized ankyrin. As shown in Fig. 1.3.1, our hybrid method can efficiently sample the near-native conformation space with lowest RMSD < 1Å. In contrast, neither Lookup or GenKIC was able to sample loops within 2Å.

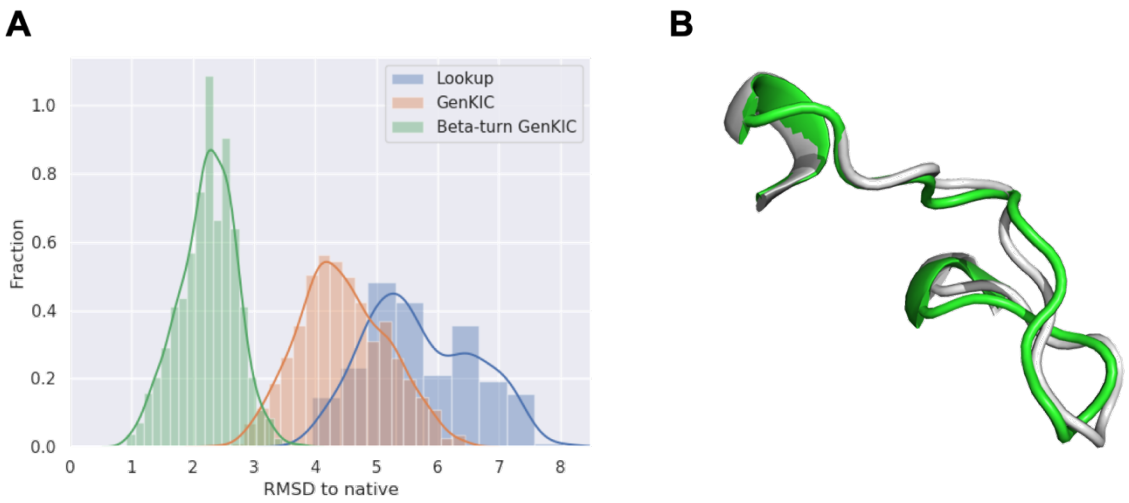

Fig. 1.2.1 Benchmark of loop sampling methods

(A) Comparison of conformational sampling among our hybrid method (Beta-turn GenKIC), a template-based method (Lookup) and a template-free method (GenKIC). Calpha RMSD to an idealized 13-residue ankyrin loop[46] was computed for each sampled loop conformations. (B) Superimposition of loop conformation (green) with lowest RMSD onto the native idealized ankyrin loop (white).

1.2.2 *Loop buttressing enhanced the predicted protein folding*

We used idealized ankyrin to test if the state-of-the-art protein structure prediction method can detect the contribution of loop buttressing interactions. The bidentate-hydrogen-bond forming ASNs (Fig. 1.3.2A) and HISs (Fig. 1.3.2B) at the loop region of ankyrin are evolutionarily conserved and shown to be important to the folding stability[46]. However, these hydrogen bonds are mostly at protein surface and might not contribute as greatly to the folding free energy as the buried hydrophobic interactions. To assess if current structure prediction method can detect the enhancement of stability by loop buttressing, we created sequence variants where the buttressing ASNs or/and HISs were mutated into ALAs. Structures predicted by AlphaFold for the variants were compared to the wild-type (WT) ankyrin design (Table 1.3.1). While AlphaFold was able to predict the WT structure for the NtoA or HtoA mutants, it predicted that the double mutant (NtoA_HtoA) do not fold into the WT structure, as indicated by low Local Distance Difference Test (LDDT) score and large RMSD in Table 1.3.1. As shown in Fig. 1.3.2D, the loop conformations of the predicted structure for the double mutant were disorganized.

	WT	NtoA	HtoA	NtoA_HtoA
LDDT	97.055	93.817	93.532	81.775
RMSD (Å)	0.480	0.567	0.586	3.191

Table 1.2.1 AlphaFold prediction of sequence variants of idealized ankyrin

LDDT indicates the confidence of AlphaFold for the predicted structures. RMSD was computed for all the Calpha atoms between predicted variant structures and the WT idealized ankyrin.

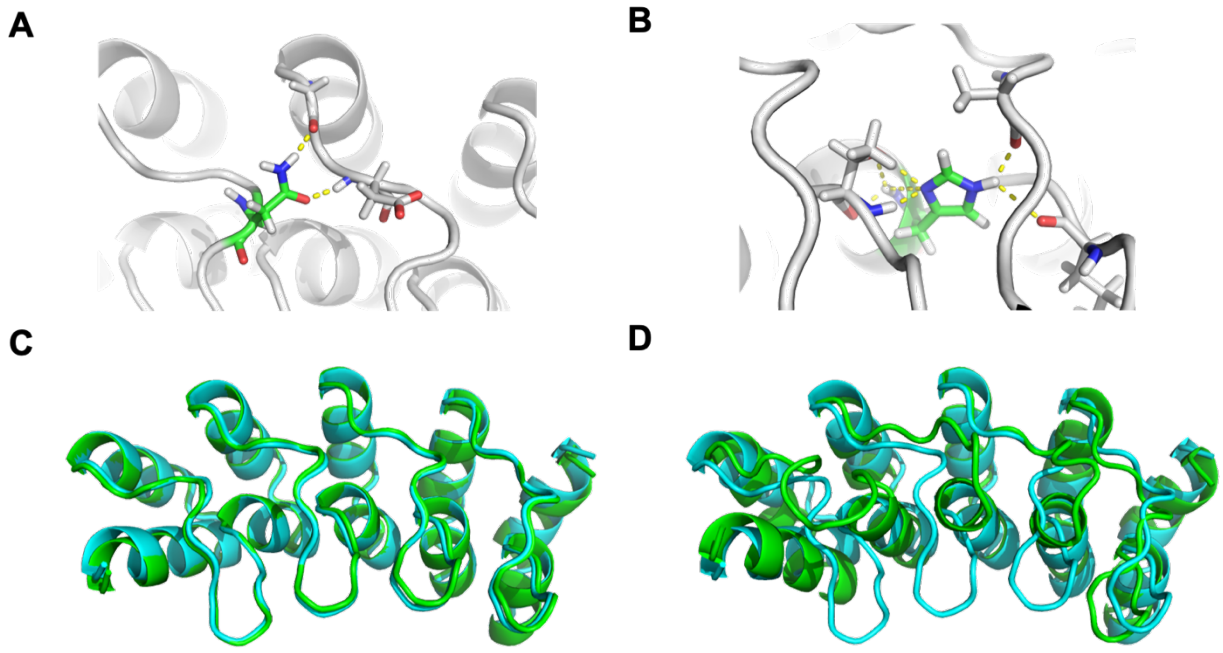

Fig. 1.2.2 Structure prediction of sequence variants of idealized ankyrin by AlphaFold (A) Loop buttressing bidentate hydrogen bond mediated by ASN. (B) Loop buttressing bidentate hydrogen bond mediated by HIS. (C, D) Superimposition of AlphaFold predicted structures (green) onto the idealized ankyrin (cyan). (C) WT (D) The NtoA_HtoA mutant.

1.3 Discussion

Using the idealized ankyrin, we benchmarked our loop sampling method and tested the impact of loop buttressing on folding via structure prediction. An interesting feature identified in native proteins is that beta hairpin loops are often used for molecular recognition. The beta turn motif in the hairpin contributes to the structural stability through the hydrogen bond. Because this hydrogen bond is formed between atoms on the protein backbone, the residues can be redesigned to facilitate protein interactions without disrupting the loop stability. To harness this advantage of

beta hairpin loops, we biased the single loop sampling by incorporating beta-turn motifs in the kinematic closure algorithm. This allowed us to efficiently sample hairpin-shape loops (Fig. 1.3.1). The successful *in silico* detection of the contribution to protein folding by loop buttress with AlphaFold allowed us to confidently apply our buttressed loop sampling protocol to the repeat proteins.

1.4 Materials and Methods

1.4.1 *The hybrid computational method for single loop modeling*

We developed a hybrid method that assembles native structural motifs via kinematic loop closure. To guide the sampling towards the hairpin-shape conformations, we constructed a motif library that consists of native beta turns. A beta turn motif is defined by having a backbone-to-backbone hydrogen bond between the carbonyl group of residue i and the amine group of residue i+3[49, 50]. Beta turns can be further divided in to four types: Type I, Type II, Type I' and Type II'[50]. In this work, we searched for native beta turn fragments by mining a set of selected PDB based on 90% maximum sequence identity and 1.6 Å resolution cutoff from PISCES[51]. The collected beta turns were further clustered by K-centers algorithm[52] into 180 motifs. Using the same approach, we compiled a library of native helical capping motifs to guide the sampling of loops connecting helices in the repeat proteins.

To connect the gap in the input protein scaffold, we used GeneralizedKIC (GenKIC)[31, 48] for loop closure. As shown in Fig. 1.5.1, a loop fragment was first constructed by stitching native helical capping motifs, beta turn motifs and KIC residues with randomized backbone torsion angles. In each step of GenKIC, kinematic loop closure was performed to connect the loop to the

intended insertion site. Loop conformations were filtered by backbone steric clashes. We further filtered the models by selecting loops with at least two backbone-to-backbone hydrogen bonds. To avoid the helical conformations, we removed the models that are predicted to have more than 5 consecutive helical residues by DSSP[53]. This ensured the extended beta-hairpin shape which contributed to the loop stability and compatibility for buttressing.

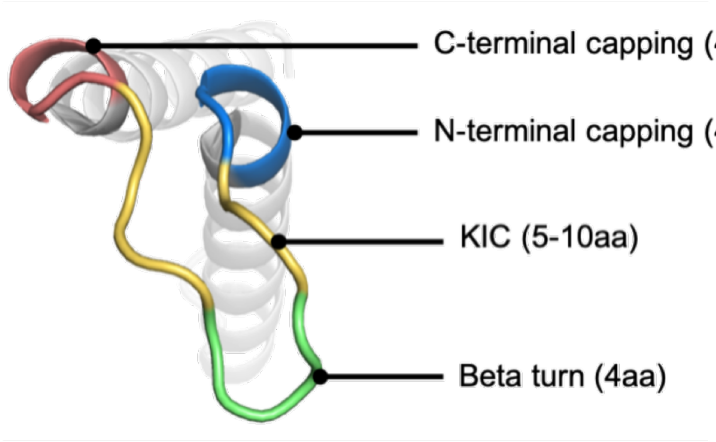

Fig. 1.4.1 Composition of a loop in the hybrid single loop sampling method

The motif residues (C-terminal capping, N-terminal capping and beta turn) are modeled by grafting native fragments from our motif libraries. The KIC residues are generated by Monte Carlo sampling of backbone torsion angles in Ramachandran probability distribution.

The sampling efficiency test was performed using the idealized ankyrin structure[46]. The loop insertion site was created by removing G59-T71 in the input structure. For each tested method, a 1000-pose sampling trajectory was launched. To evaluate the resulted loop conformations, we computed the RMSD of Calpha atoms between each loop models and the original loop on the idealized ankyrin structure.

1.4.2 *Loop propagation and buttressing design for repeat proteins*

To install the loops of the same conformation in each unit of repeat proteins, we used the RepeatPropagationMover in Rosetta[47, 48]. After filtering out the loops with steric clashes, we computed three metrics to help select the best loop conformations for buttressing: number of interloop backbone-to-backbone hydrogen bonds, loop motif scores and direction scores. We required at least one interloop backbone-to-backbone hydrogen bond between each pair of neighboring loops to maximize the sequence-independent loop buttressing effect. The motif scores were computed by matching the selected pairs of residues to the known native hydrophobic residue pairs in PDB[34, 47, 48]. The scores for each match of residue pairs in the loop regions were then summed to one total score. Only the loops with a negative total motif score were selected. The direction scores described the relative orientation of the loops from the rest of input repeat proteins. Specifically, we defined two vectors: vector ***a*** started from the center of mass of the two loop terminal residues, to the farthest Calpha atom of the loop; vector ***b*** started from the same point as ***a***, but pointed towards the center of mass of the repeat unit. The direction score was derived by computing the angle between the two vectors:

$$Direction\ score = \cos^{-1}\frac{a \cdot b}{|a||b|}$$

The accepted angles ranged from 45° to 135°.

Next, we performed a fast sequence design task to identify loop conformations compatible with interloop bidentate hydrogen bond networks. From each propagated set of loops, the loop on the second repeat unit was selected for sequence design. One packing step using PackRotamersMover[47, 48] was conducted separately for each residue on this loop using amino acids that are compatible with forming side-chain-to-backbone bidentate hydrogen bonds: ASN, ASP, GLN, GLU and HIS. We excluded amino acids with long side chains (ARG and LYS), as

their high entropic cost might diminish the free energy contribution of buttressing. After each packing step, hydrogen bonds between the packed residue and its neighboring residues were counted and one bidentate interaction was required to pass this step. Specifically, a bidentate hydrogen bond was defined as two separate hydrogen bonds forming between atoms in the functional group of the side chain and the atoms in the backbone of the neighboring repeat unit. Alternatively, we used a three-stage scheme to maximize the sampling efficiency of bidentate hydrogen bonds: identifying pseudo bidentate hydrogen bonds, constrained minimization for building hydrogen bonds and confirmation of bidentate hydrogen bonds. We defined that pseudo hydrogen bonds have donor-acceptor distances < 3Å and a hydrogen bond angle > 120°. After propagating the designed residue to all the repeat units, we imposed a harmonic distance constraint between each donor and acceptor atoms with target distance as 2Å and standard deviation as 0.5Å. At the minimization stage, we performed symmetric minimization of the loops to improve the interactions of potential hydrogen bonds. Finally, using the Rosetta score function, we searched for bidentate hydrogen bonds in the minimized loop conformations.

To test the sensitivity of state-or-the-art structure prediction method for loop buttressing bidentate hydrogen bonds, we made sequence variants of the idealized ankyrin design by mutating the residues participating in bidentate hydrogen bonds to ALAs: NtoA, HtoA and the double mutant NtoA_HtoA. AlphaFold[37] was used to predict structures for both WT and each of the sequence variants. Structure comparison was then performed by computing the RMSD of the Calpha atoms.

Chapter 2. Diversifying Designed Repeat Helical Proteins with Buttressed loops

2.1 Introduction

While antibodies are still playing the central role of protein therapeutics, major progress has been made in drug development using non-antibody binding proteins which show superior properties in thermal/pH stability, binding affinities, tissue delivery and industrial-scale manufacture[54, 55]. Among the most studied alternative protein scaffolds, ankyrin stands out for its multiple successes in recent pre-clinal studies, but also for its repetitive architecture that is well-suited for modular protein binder design[43, 56]. Native ankyrins contain 4-29 repeats of helix-loop-helix-turn structure units[44, 45]. The structured, hairpin-shape loops extend from the helices and form a beta-sheet-like structure that is stabilized by interloop buttressing via extensive hydrogen bond networks (see Fig. 2.1.1). This feature contributes to the formation of the long binding groove that is geometrically compatible with many globular protein targets. Using consensus sequence design and site-directed mutagenesis, the Plückthun group built libraries of Designed Ankyrin Repeat Proteins (DARPins) which have been routinely used to identify high-affinity binding proteins via ribosome display, phage display and yeast display[43]. While DARPins are proven candidates for the non-antibody-based protein therapeutics, these protein scaffolds are limited by the highly conserved ankyrin structures. Therapeutic targets that share poor shape complementarity with one DARPin are also likely to be incompatible with the other DARPins. Developing structurally

diverse repeat protein scaffolds that include but not are not limited to ankyrin fold remains a great challenge for drug discovery.

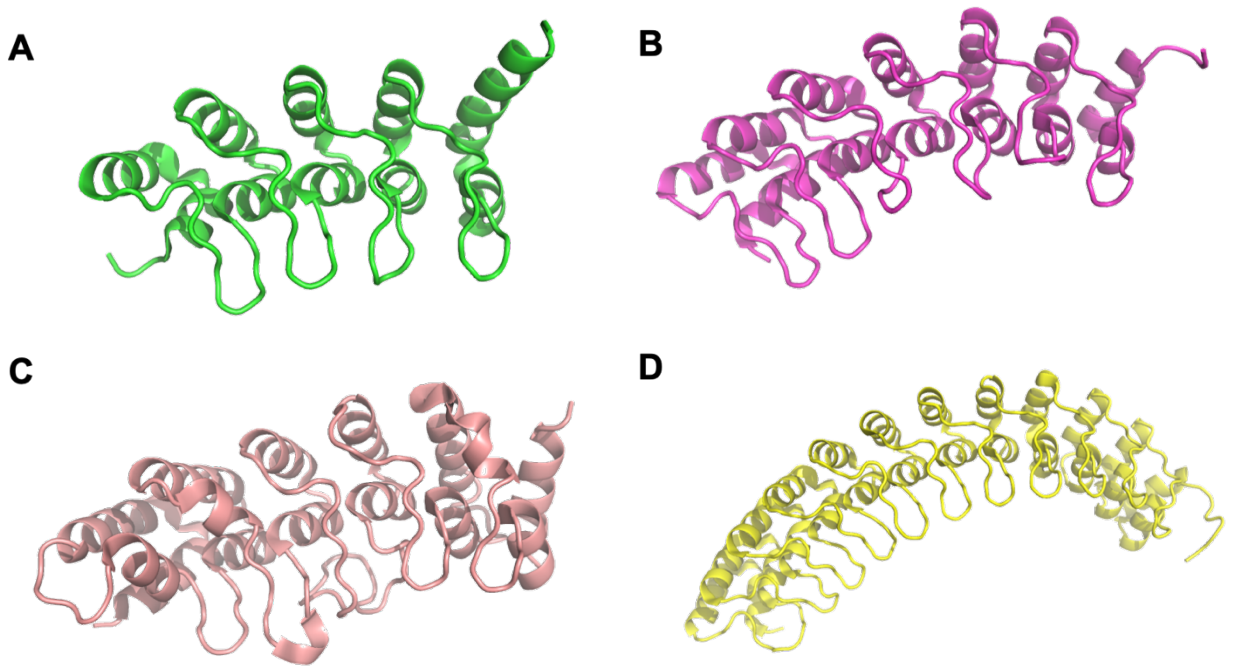

Fig. 2.1.1 Crystal structures of native ankyrins
PDB IDs: (A) 3SO8[57], (B) 4UUC, (C) 4N5Q[58] and (D) 1N11[59].

We set out to design repeat proteins with long, structured loops by generating diverse *de novo* helical repeat proteins (DHRs) from scratch and building structured loops that are stabilized by buttressing. To achieve this goal, we developed two computational protocols: 1. A fast and reproducible method for generating DHRs that are geometrically compatible with the insertion of long loops; 2. An efficient algorithm for building long, buttressed loops for a given DHR. For DHR generation, we developed two complementary methods. The first method is adapted from the previous work which used fragment assembly to generate DHRs with diverse helical

parameters (radius, twist and rise)[60]. We fine-tuned this method by specifying the helical parameters that are close to those of ankyrins and formulated geometric constraints to guide the fragment assembly. Meanwhile, we developed an alternative method that does not rely on native protein fragments. Specifically, we enumerated the possible geometries based on the given ranges of helical parameters, and constructed DHRs accordingly, using ideal alpha helices made from scratch. To build structured long loops by buttressing, we adapted the method described in Chapter 1, by specifying the position of beta turn motifs in the loops such that the loops generated are mostly extended beta hairpins.

2.2 Results

2.2.1 *A divide-and-conquer strategy for designing buttressed loops on DHRs*

We dissected the problem of multiple-loop design on DHRs into two sub-problems: 1. Designing DHRs with diverse shapes; 2. Designing loops that are compatible with the given DHRs and loop buttressing. To explore the wide range of DHRs, we developed two separate but complementary methods. The first method was based on previous work which used Rosetta Monte Carlo fragment assembly and guided the sampling of DHRs with tandem repeat symmetry [60, 61]. In order to efficiently explore the DHR structures that are potentially compatible with loop buttressing, we further constrained the sampling by helical parameters (radius, twist and rise)[60, 62] derived and expanded from native ankyrin proteins. While this protocol provided with a variety of DHR structures, it did not provide total control of the DHR assembly. For example, the helical parameters did not specify the relative orientation between the two helices within each repeat unit; neither did the parameters describe the rotations between the repeat units along the tangent line

with respect to Z axis. These problems can be tackled by formulating more geometric constraints based on inter-atom distances, angles, and dihedral angles. However, employing extra constraints often required fine-tuning of their weights during the fragment assembly for the maximized sampling efficiency. As an alternative approach, we developed a new protocol which used idealized helices and performed parametric assembly of DHRs. In particular, the protocol allowed specification of the geometric transformations between the helices in each repeat unit as well as the transformations between the repeat units. This method complemented the fragment assembly protocol well, in that a DHR structure of interest that was rarely sampled in fragment assembly, can be quickly reproduced and perturbed.

By combining the two DHR assembly methods with the buttressed loop sampling protocol described in Chapter 1, we were able to construct DHRs with multiple loops (Fig. 2.2.1). We filtered the assembled backbone models by their compatibility with loop-loop bidentate hydrogen bonds, helix-loop bidentate hydrogen bonds and loop-helix hydrophobic interactions. The models that contained all the three categories of interactions were subjected to two rounds of full-atom sequence design. In the first round, we extracted backbone fragments from the model and compared them to the database containing native PDB fragments, based on which we computed the sequence propensity for each fragment. Using this propensity, we constructed sequence profiles that were then used to guide sequence design with Rosetta all-atom score function and constraints based of repeat symmetry. In the second round of sequence design, we focused on the terminal repeat units that contained solvent-exposed hydrophobic residues due to the symmetric sequence design from the first round. We redesigned these residues with polar amino acids (Glu, Gln, Lys and Arg) in the absence of symmetric constraints to improve the solubility and folding.

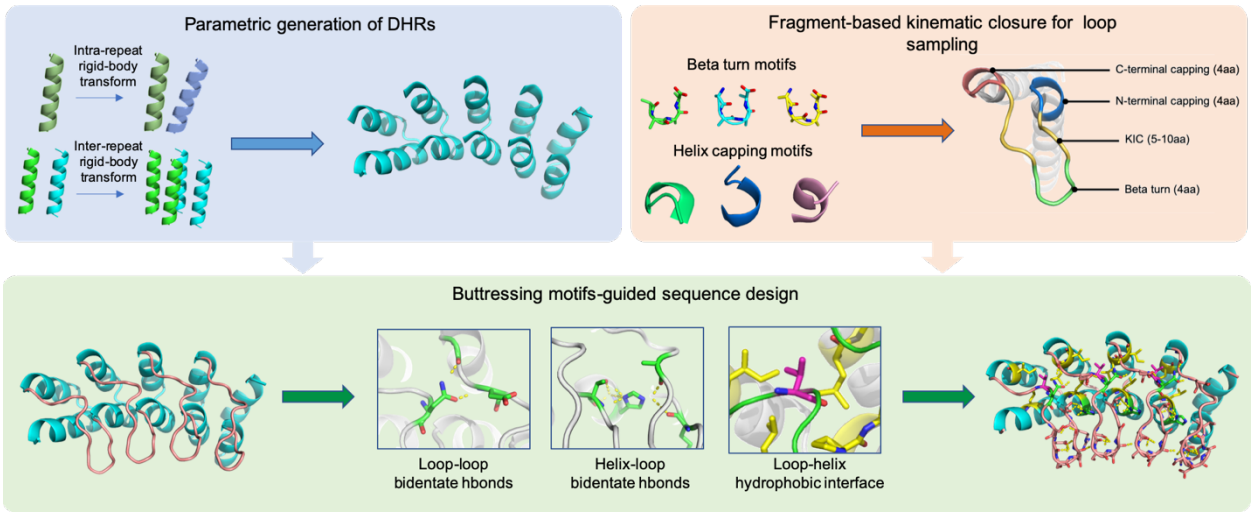

Fig. 2.2.1 Design strategy for buttressed loops on DHRs

DHRs are generated by systematically placing helices based on given parameters. For each DHR, a single long loop is built using our hybrid fragment-based kinematic closure method. The single loop is then propagated to every repeat unit of the DHR. Motifs that contribute to loop buttressing (loop-loop bidentate hydrogen bonds, helix-loop bidentate hydrogen bonds and loop-helix hydrophobic interactions) are searched and designed to optimize the stability of the loops.

Designs were tested by *in silico* structural prediction and evaluation methods such as GenKIC loop forward folding, molecular dynamics simulations, AlphaFold and RoseTTAFold structure predictions, before they were experimentally tested for expression, solubility, thermal stability, and structural characterization by X-ray crystallography.

2.2.2 *Designs span a wide range of geometries*

The geometries of DHRs were previously described by three parameters: radius, twist and rise[60, 62]. As shown in Fig. 2.2.2, our designs well covered the geometries of native ankyrins.

Specifically, the radius of our designs spans from <10Å to 100Å and their omega values (omega is the absolute value of twist) span from close to 0 to ~3. A selection of designs that are validated by our *in-silico* tests were shown in Fig. 2.2.3. Furthermore, our designs discovered loop buttressing interactions that were not used in native ankyrin proteins (see Fig. 2.2.4). The diverse overall shapes of DHRs and non-native buttressed loops together suggested that there might be a large space of possible multi-loop repeat helical proteins that have not been explored by ankyrin and other native repeat proteins.

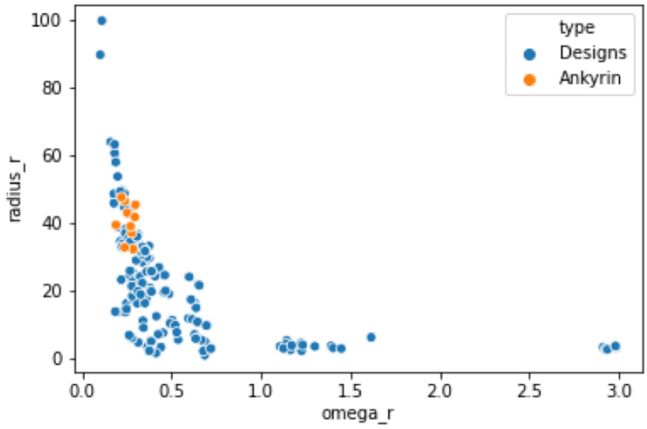

Fig. 2.2.2 Geometries of designs described by helical parameters

The geometries of our designs and representative native ankyrin structures are projected onto the plane defined by radius and omega of DHRs based on *Brunette et al. (2015)[60]*.

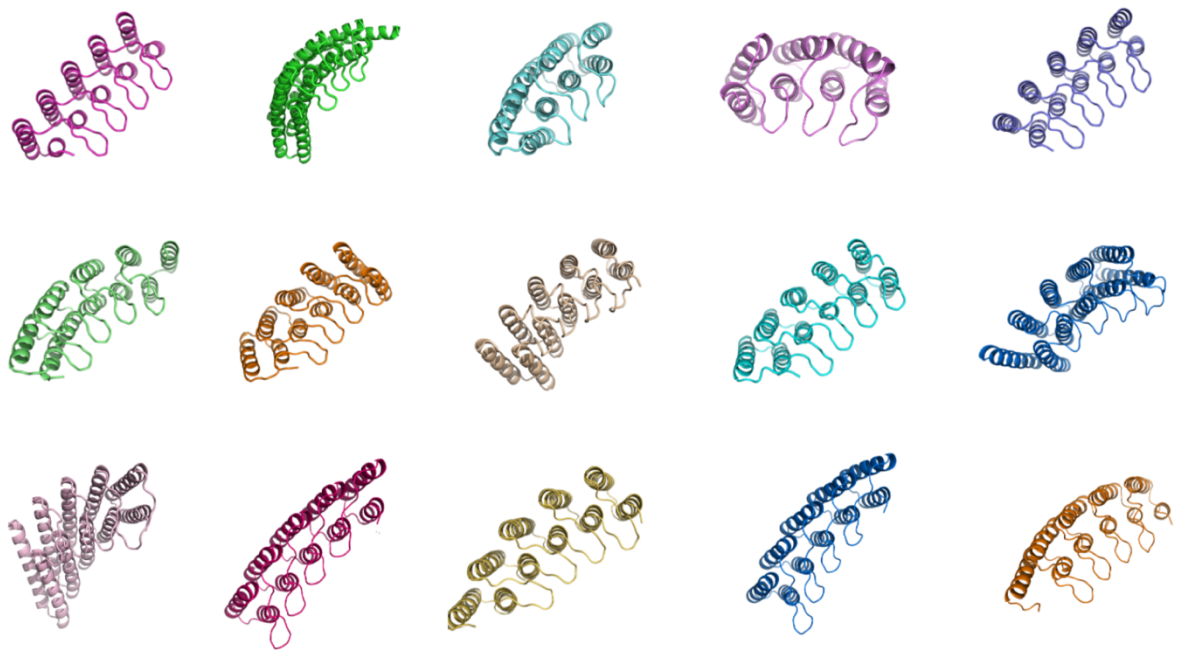

Fig. 2.2.3 A gallery of diverse designed proteins that passed the *in silico* filters

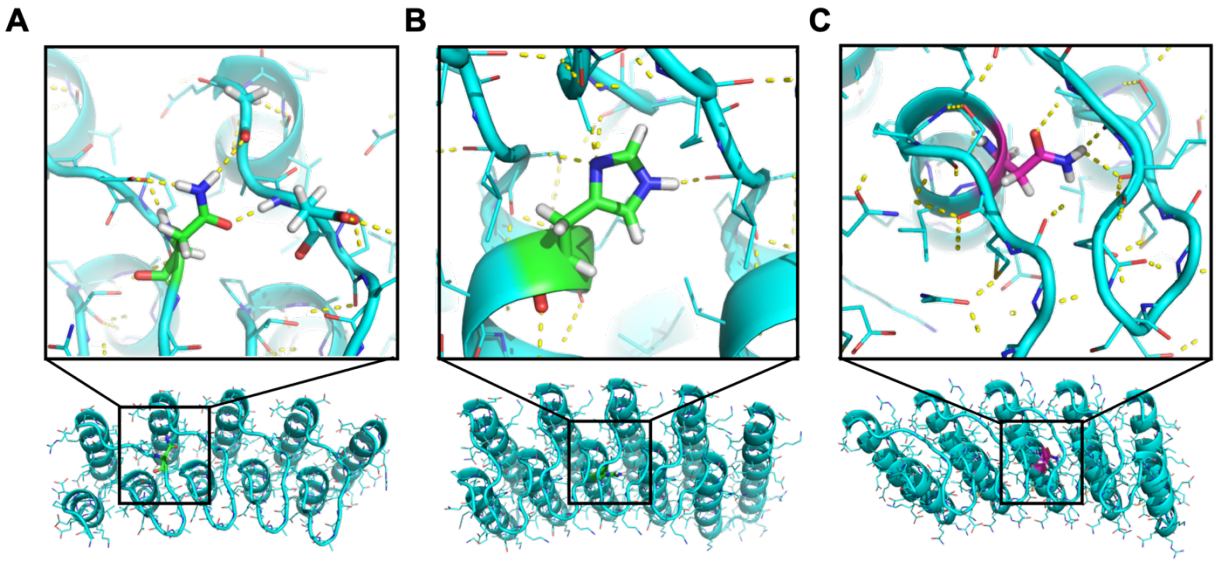

Fig. 2.2.4 Loop buttressing interactions in the designs

(A and B) The bidentate hydrogen bonds that exist in native ankyrins. (C) Newly discovered bidentate hydrogen bond interactions.

2.2.3 *Designs are thermodynamically stable, monodisperse and well-folded*

To experimentally test the designs, we generated reverse-translated DNA sequences and obtained 106 genes that encoded the repeat proteins. The designed proteins were then expressed in *E. coli* and purified by Histag-immobilized metal affinity chromatography. In total 102 proteins were expressed, of which 77 were purified in soluble form. 52 proteins were monodispersed and 46 of them were monomeric, as suggested by the data from multi-angle light scattering coupled with size exclusion chromatography (SEC-MALS). 44 of these proteins showed expected alpha-helical circular dichroism (CD) spectrum at 25°C, remained at least partially folded at 95°C and recovered nearly all the signals when cooled down to 25°C. 14 designs were further validated by small-angle X-ray scattering (SAXS). Fig. 2.2.5 displays four representative designs that are monomeric, thermodynamically highly stable and validated by SAXS.

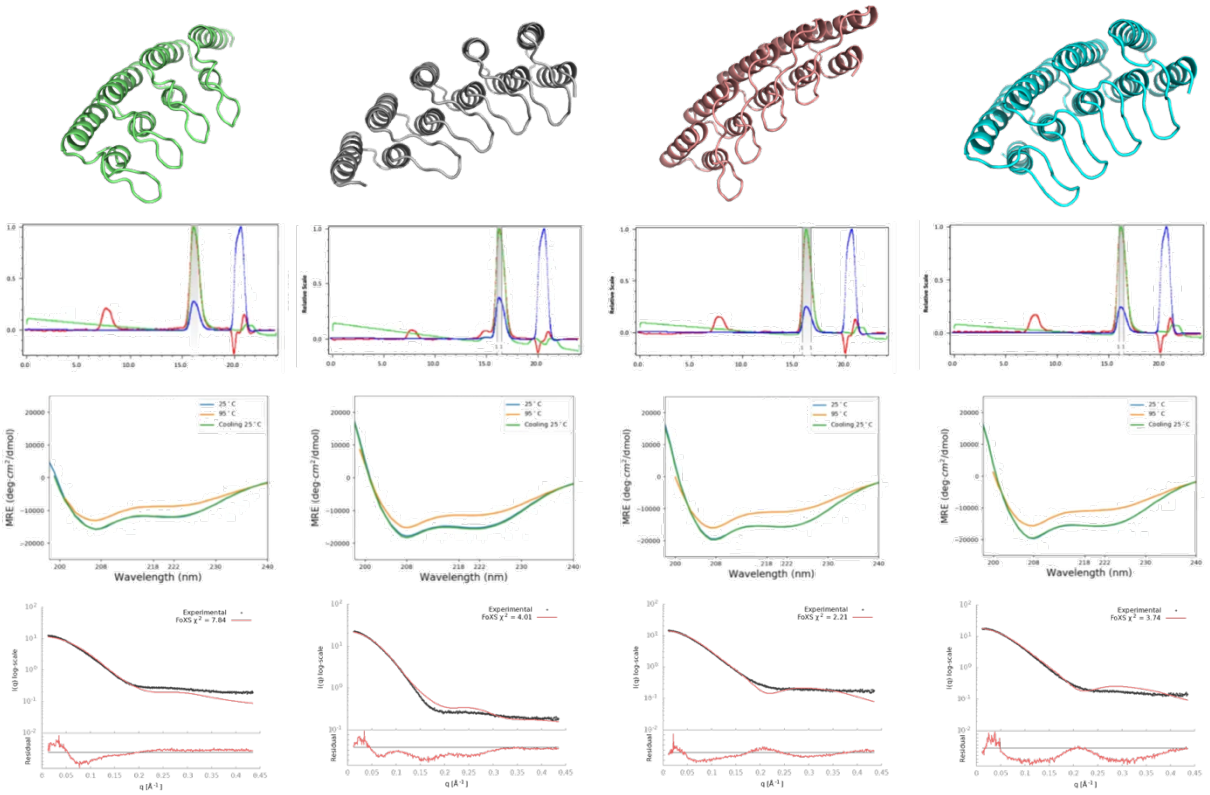

Fig. 2.2.5 Experimental characterization of designed DHRs with buttressed loops

For each design, a structure model is shown (top row). In the second row we list the traces of multi-angle light scattering coupled with size exclusion chromatography (SEC-MALS), with light scattering (LS) signals colored in red, UV280 signals colored in green and differential refractive index (dRI) colored in blue. In the third row we display the traces from circular dichroism (CD). In the bottom row are the alignment of small-angle X-ray scattering (SAXS) profiles to the expected profile computed from the designed models.

2.2.4 *Validation of designs by X-ray crystallography*

We obtained crystal structures for three designs, two of which had the loops well resolved. Crystal structure of design PDL_0_4 was solved at the resolution of 1.8Å. The designed model showed

overall agreement with the crystal structure with a Calpha RMSD of 1.7Å (Fig. 2.2.6 A). By computing the helical parameters, we noticed that the primary discrepancy in the crystal structure resulted from the inter-repeat transformation. In particular, the designed model was slightly curved (smaller radius), while the crystal structure was nearly flat (larger radius). Indeed, the individual repeat units between the design and the crystal structure varied only by Calpha RMSD ranging from 0.48-0.61Å, suggesting the design was highly accurate within each repeat unit (Fig. 2.2.6 B). Importantly, the designed loop buttressing interactions, the bidentate interloop hydrogen bonds (Fig. 2.2.6 C) and loop-helix salt bridges (Fig. 2.2.6 D), were accurately displayed in the crystal structure.

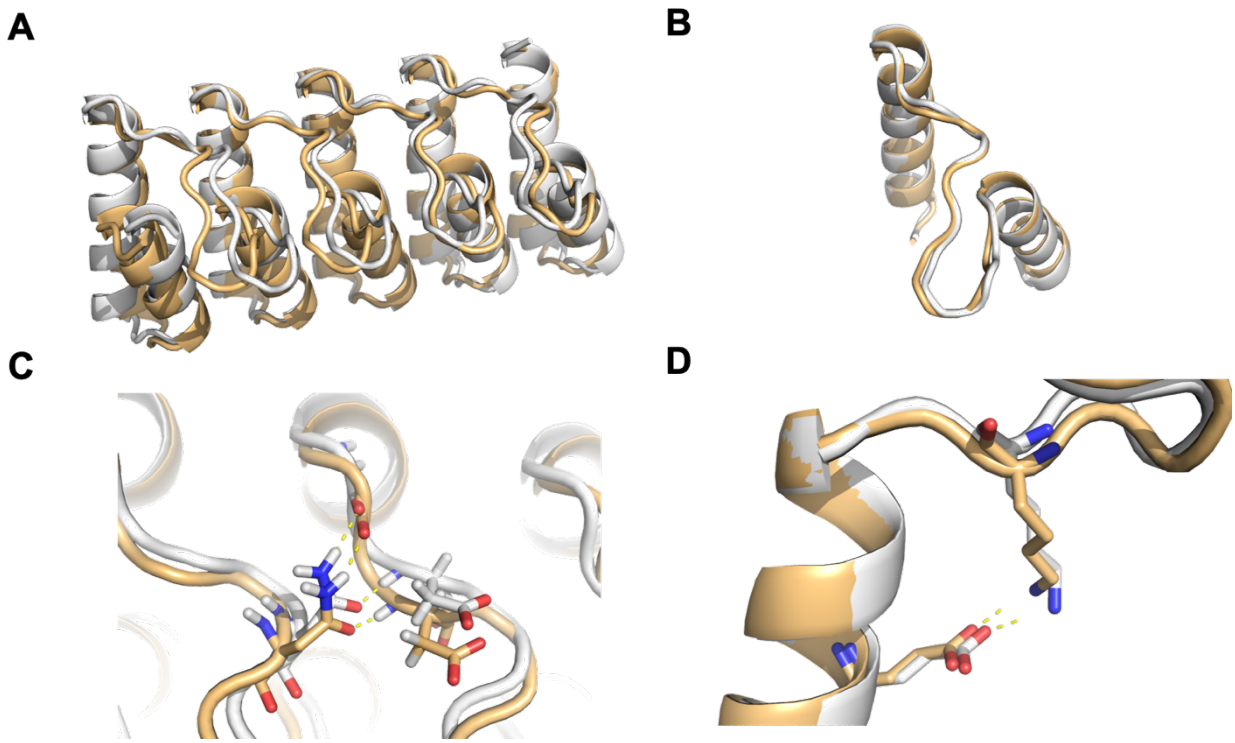

Fig. 2.2.6 Structure of design PDL_0_4

(A) Superimposition of crystal structure (yellow) onto the design model of PDL_0_4 (grey). (B) Alignment of single repeat units. (C, D) Comparison of designed loop buttressing interactions.

Design PDL_0_7 was experimentally tested to be highly stable, monomeric and validated by SAXS. We obtained the crystals which unfortunately diffracted poorly with the highest resolution at 4.2Å. Previous studies suggested that synthetic oligomerization can sometimes assist crystallization[63]. We therefore redesigned PDL_0_7 into a homodimer by introducing a surface hydrophobic dimer interface. The designed dimer behaved well experimentally and was later crystallized successfully. We solved the crystal structure of the dimer at 3Å. As shown in Fig. 2.2.7, the crystal structure closely matched the designed model, with a Calpha RMSD of 2.7Å. The monomer unit of crystal structure showed ever better agreement with the designed model with Calpha RMSD of 1.2Å. The main regions of disagreement between the structures were the terminal helices, as the N-terminal helices in the crystal structure tilted away from the neighboring repeat unit and the crystal structure contained a kink in the middle of each C-terminal helix.

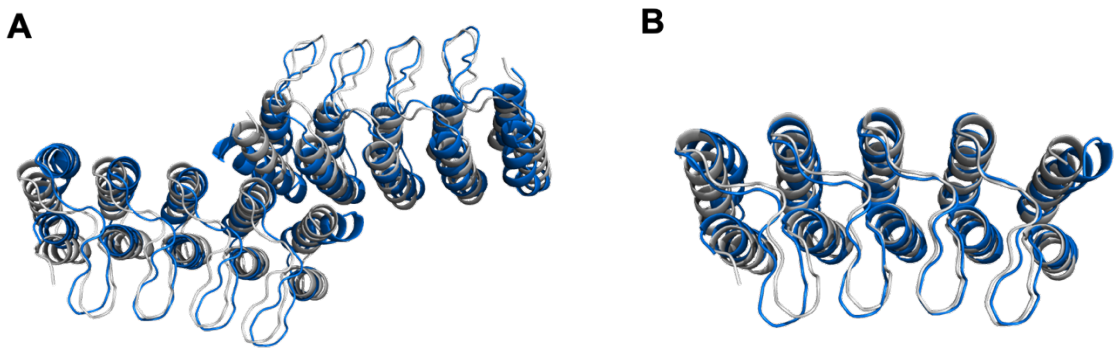

Fig. 2.2.7 Structure of a homodimerized design PDL_0_7

(A) Superimposition of crystal structure (blue) onto the design model of dimerized PDL_0_7 (grey). (B) Superimposition of a monomer in the crystal structure (blue) onto the designed model of PDL_0_7.

2.3 Discussion

In natural proteins, structured loops often play the central roles in molecular recognition, signal transduction and enzyme catalysis. However, organizing multiple structured loops at the functional sites has been a long-standing challenge for protein design. Here we harnessed the tandem repeat symmetry of the helical repeat proteins, and designed structured loops with symmetric buttressing. We found that the designs are highly soluble, thermally stable, folded and monodisperse, which makes them ideal protein scaffolds for functionalization. Crystal structures not only showed close agreement with our computational model, but also clearly resolved the structured loops that were buttressed accurately as designed. Our approach provided a promising strategy to tackle the general loop design problem for engineering new protein functions.

2.4 Materials and Methods

2.4.1 *Parametric DHR generation*

Ideal helices were generated using MakeBundleHelix mover in Rosetta[47, 48]. In the beginning of our protocol, an ideal helix with specified length, H1, is generated and placed away from the Z axis with given radius and angles corresponding to its orientation. A second helix, H2, was then modeled and placed according to the specification of the geometry transformation from H1 to H2. By combining H1 and H2 into one pose, we built the first repeat unit R1. Subsequently, we used user-specified helical parameters, twist and rise, together with parameters that define the rotations

between repeat units to perform geometric transformation to obtain the second unit R2. Based on the number of repeats desired, we propagated the repeat units to generate the complete DHR with the absence of loops. Next, using ConnectChainMover, we modeled the short loops to connect the helices to complete the pose.

2.4.2 *Design of buttressed loops*

For each of the generated DHRs, we applied the protocol described in Chapter 1 to build a single loop and propagated the loop to each repeat unit. We filtered the loops by requiring each repeat unit to have at least one interloop bidentate hydrogen bond, one backbone-to-backbone hydrogen bond and one loop-helix hydrophobic interactions.

2.4.3 *Sequence design*

For each backbone model generated and filtered from loop modeling step, we first compared the fragments at each residue to the native PDB fragments to construct a structure-based sequence profile. We incorporated this profile by FavorSequenceProfile mover to constrain the choices of amino acids for each residue. To further guide the choices of amino acids, we used LayerSelector to define the core, the boundary and the surface layers, and specified the allowed amino acids for each layer. We added residue type constraints to fix the identity of the residues participating loop buttressing bidentate hydrogen bonds, so the stabilizing interactions obtained during loop sampling would be maintained throughout sequence design. Next, we performed four rounds of full-structure sequence design using FastDesign mover under the repeat symmetric constraints to ensure the repeat units had the same structures and sequences. To improve the solubility and

folding of the designs, we next performed one round of FastDesign for the solvent-exposed hydrophobic residues on the terminal repeat units. Only polar residues such as Glu, Gln, Lys and Arg were used for this round of design. The designed structures were then refined by minimization in Cartesian space and subsequently filtered by total score, percentage of hydrophobic residues, backbone quality, backbone torsion angles, hydrogen bond content in loop regions. Top 10% scoring structures were further tested by *in silico* validation methods such as molecular dynamics simulations (RMSD < 3Å), AlphaFold (PLDDT > 80, RMSD < 3Å) and RoseTTAFold (PLDDT > 80, RMSD < 3Å).

2.4.4 *Protein expression and characterization*

Genes encoding the *in silico* validated designs were synthesized (IDT) and cloned into pET-29b expression vectors. The plasmids were transformed into BL21 (DE3) expression *E. coli* strain. Protein expression was performed using auto-induction protocol[64] at 37°C for 24h. To extract the proteins, we lysed the cells by sonication. The soluble proteins were obtained by centrifugation of lysate at 16,000g for 30min. The designs were then purified from the supernatant via Ni-NTA affinity resin and the monodisperse designs were obtained by performing size-exclusion chromatography using Akta Pure FPLC device using the buffer of 25mM pH8 Tris and 125mM NaCl. To perform multi-angle light scattering coupled with size exclusion chromatography, we prepared the purified protein at ~2mg/ml and injected 100ul of sample into a Superdex 200 10/300GL column and measured the light scattering signals using a miniDAWN TREOS device. To measure the circular dichroism (CD) signals, we first prepared the sample at ~0.2mg/ml in 25mM phosphate buffer in 1mm cuvette. A Jasco J-1500 CD spectrometer was used for all CD measurements. We set the range of wavelength from 190nm to 260nm and scanned over a three-

temperature (25°C, 95°C and cooling back at 25°C) set for each sample. We submitted all samples for small-angle X-ray scattering (SAXS) to Advanced Light Source, LBNL for data collection at the SIBYLS 12.3.1 beamline. The crystals were submitted to Advanced Photon Source beamline facility for the collection of diffraction data.

Chapter 3. Functionalization of Protein Scaffolds Installed with Buttressed loop

3.1 Introduction

Natural repeat proteins play crucial roles in molecular recognition, signal transduction and molecular scaffolding[60, 65]. Consisting of multiple copies of similar structural units, repeat proteins form elongated structures with extended, often concaved, surfaces that are particularly suitable for binding interactions[43]. Moreover, the modular composition of repeat proteins makes them great scaffolds for function reengineering and recombination. Several repeat proteins have been since engineered into protein binders[43, 46].

One of the most successfully engineered repeat protein scaffold is ankyrin. Natural ankyrins have 4-29 repeats of 33-residue helix-loop-helix-loop units[42, 44]. The interface between the long beta-hairpin loops and their neighboring helices forms an elongated groove that is frequently used for molecular recognition. By consensus sequence design and trinucleotide-based site-directed mutagenesis, the Plückthun group built the designed ankyrin repeat protein (DARPin) library[43, 56]. Combined with high-throughput screening technologies such as ribosome display, phage display, and yeast display, these libraries have been routinely used to produce high-affinity binders to therapeutic targets including but not limited to hepatocyte growth factor, vascular endothelial growth factor and human epidermal growth factor receptor[43, 56].

Since our designed DHRs with buttressed loops are inspired by the ankyrin fold and thermodynamically highly stable, we expected them to have equivalent potential for

functionalization as DARPins. In this chapter, we first demonstrated the advantage of the buttressed loops over the canonical DHRs by performing *in silico* protein interface design evaluations. Next, we applied our designs to target therapeutic sites on Apolipoprotein E. Finally, to demonstrate the modular and repetitive feature of our scaffolds, we designed repeat-protein-repeat-peptide binding pairs that can be potentially used for synthetic biology and proteomics studies.

3.2 Results

3.2.1 *In silico protein interface design experiments*

The buttressed loops expand the binding interface of DHRs and have the potential to provide greater shape complementarity to the globular protein targets. To demonstrate the advantage of our designs in the application of protein interface design, we compiled a library of 128 computationally validated scaffold designs and performed *in silico* docking experiment against a selected set of therapeutic targets. We followed the recently published interface design protocol[66] and performed sequence design for the best-scoring docks. As control, we removed the buttressed loops from the scaffolds and repeated the docking and design experiment. To evaluate and compare the quality of the designed binders, we computed and plotted the scores of three metrics: contact molecular surface (the buried area between the binder and the target protein), score per residue (the total Rosetta score normalized by number of residues in the binder) and ddG (the relative change of free energy due to binding based on Rosetta's score function). As shown

in Fig. 3.2.1, DHRs installed with buttressed loops receive better scores in all three metrics than their DHR-only version of designs.

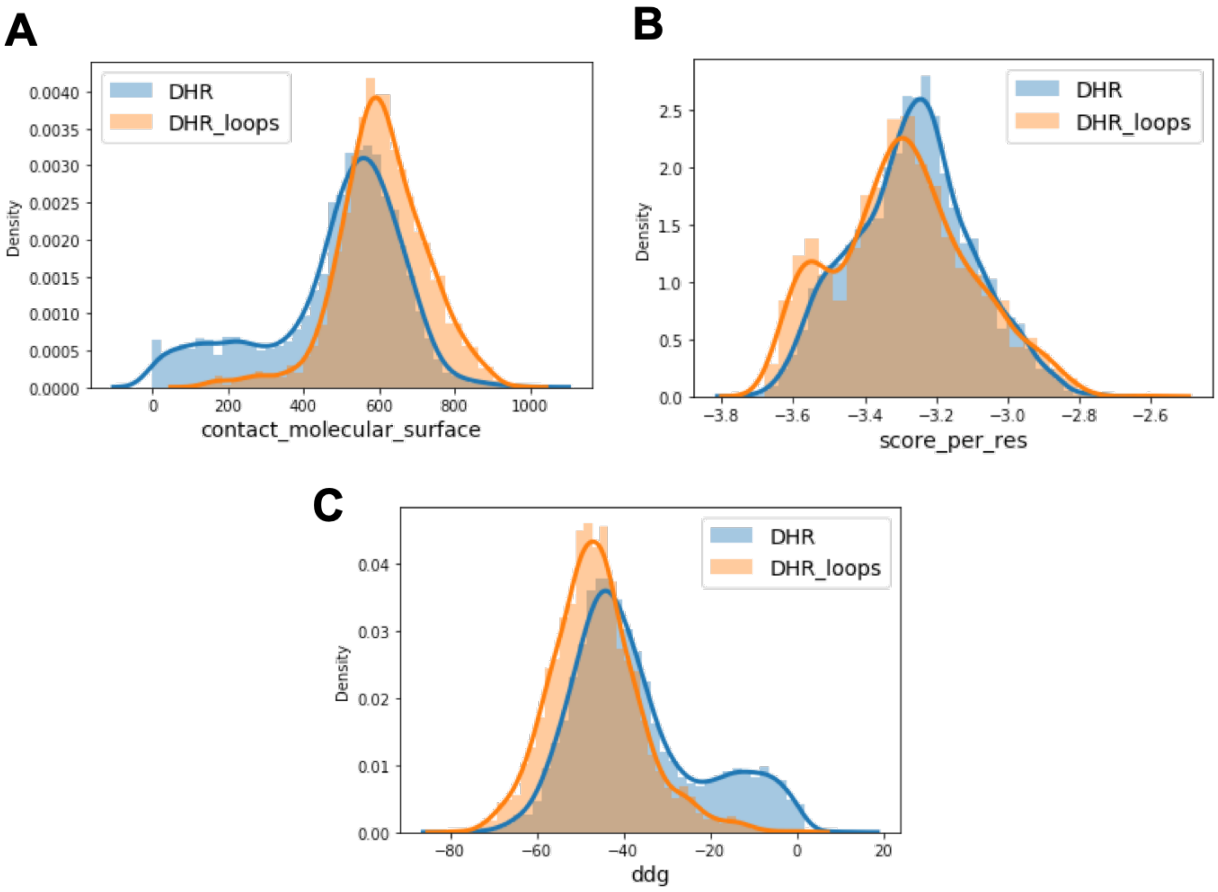

Fig. 3.2.1 Comparison of metrics for designed binders based on DHRs vs. binders based on DHRs with long loops

Three metrics, (A) contact molecular surface (the higher the better) (B) score per residues (the lower the better) and (C) ddG of binding (the lower the better), are computed and plotted for binder designs using DHRs inserted with long, buttressed loops (orange) and the DHRs only (blue).

3.2.2 *Designing binders against ApoE peptide*

Apolipoprotein E (ApoE) is a major lipoprotein that plays crucial role in cholesterol transport, neuronal signaling and inflammation in the brain[67, 68]. Due to its polymorphic nature, different alleles coexist with allele E3, the most common type, while allele E2 is associated with reducing the risk of Alzheimer's disease (AD) and E4 shown to increase the risk of AD [68]. These ApoE proteins are therefore of great interest for therapeutic target development.

We set out to design binding proteins against a helical ApoE fragment (QARLGADMEDVCGRLVQ) that was previously identified as a therapeutic target site. Using the same scaffold library and design protocol as in Section 3.2.1, we generated 11,506 binder designs and selected 1,000 top-scoring designs for *in silico* validation via AlphaFold complex structure prediction. As shown in Fig. 3.2.2A, our top-scoring designs contained a hydrophobic interface that shows reasonable shape complementarity with the ApoE peptide, and the residues on the beta turn of one buttressed loop form multiple hydrogen bonds and salt bridges for potentially target specific recognition. Such binding mode was well predicted by AlphaFold (Fig. 3.2.2B), suggesting the high quality of the binder design.

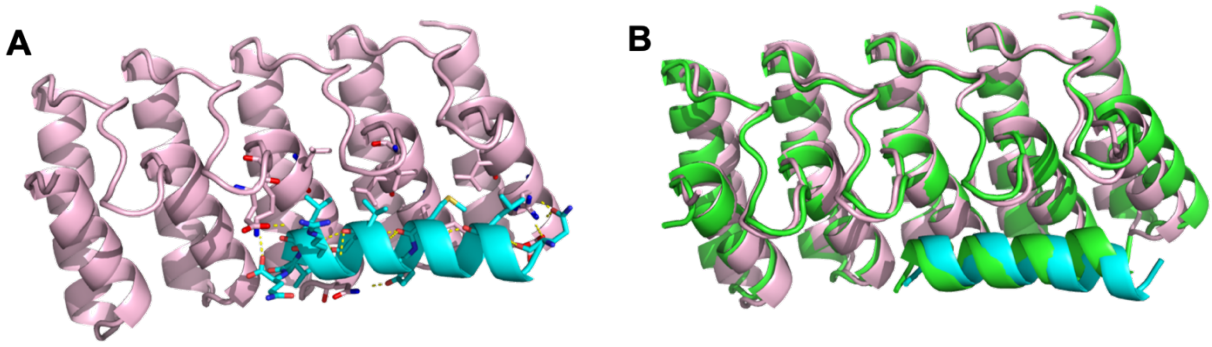

Fig. 3.2.2 A top-scoring design targeting ApoE peptide.

(A) ApoE binding design (pink) in complex with ApoE peptide (cyan). (B) Superimposition of the design in (A) with the AlphaFold-predicted complex model.

To experimentally characterize the binding affinities, we expressed and purified the computationally validated designs and performed fluorescence polarization assay using TAMRA-labeled ApoE peptide. As shown in Fig. 3.2.3A, the preliminary measurement of our designs suggest that they bind ApoE peptide at micromolar affinities. As control, we also measured the binding between our designs and parathyroid hormone peptide, which showed micromolar affinity, but generally weaker than those of ApoE peptides.

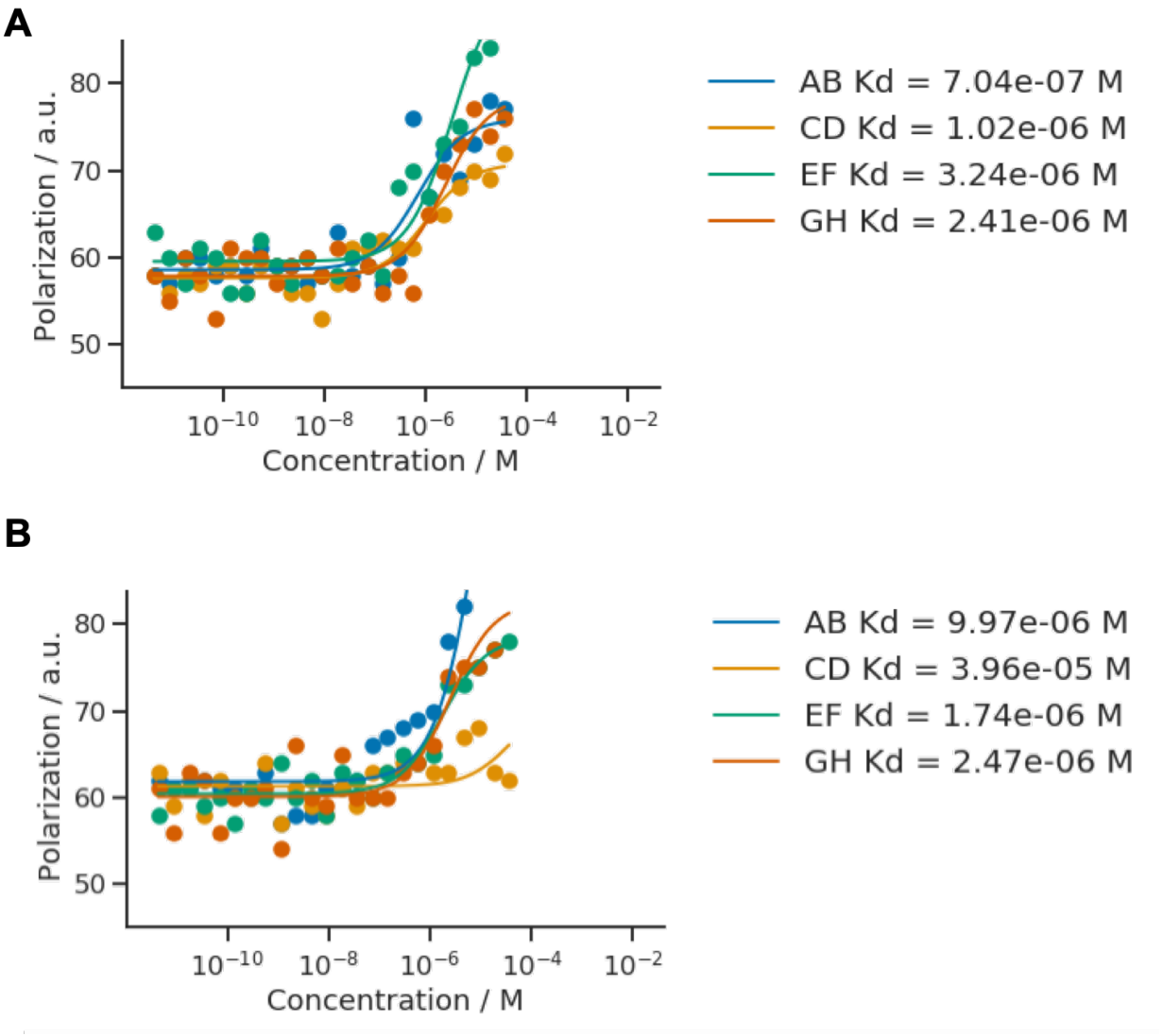

Fig. 3.2.3 Binding affinities between designed binders and peptide targets measured by fluorescence polarization assay

(A) Binding affinities measured of four designed binders against TAMRA-labeled ApoE peptide. (B) Binding affinities measured of the same four binders against TAMRA-labeled parathyroid hormone peptide.

3.2.3 *Designing repeat peptide binding proteins*

The repetitive structures of our designs make them particularly suited for engineering modular interactions. One application is to design repeat-protein-repeat-peptide binding pairs that can be used for engineered cellular signaling. Indeed, some native ankyrins were found to bind peptides with the PxLPxI/L (x can be any amino acid) sequence motif[57]. We reasoned that such sequence motif can be generalized to XYZ(n), where n is the number of repeats, X and Z are hydrophobic residues interacting with the helices and the helix-loop joint of our scaffolds, and Y can be a polar residue with long side chain interacting with residues at the beta turns of the buttressed loops on the scaffolds.

We set out to dock randomly generated repeats of tripeptide backbones to the potential binding interfaces on our designed scaffolds. Specifically, we compiled a list of Calpha atom coordinates for the target hydrophobic residues at the binding interface for the ankyrin and DARPin complexes in PDB. The geometric transformations from the helix-loop joints to these Calpha coordinates in the native ankyrin or DARPins were computed. Using these transformations, we guided the docking procedure by first aligning our generated tripeptide repeats to the target Calpha coordinates and subsequently performed rigid-body perturbation to diversify the docked poses. We next performed sequence design to improve the interactions between the repeat peptides and the repeat scaffolds.

The designs were validated *in silico* by AlphaFold. The predicted structures agreed well with the designs, with the exception that the predicted peptide structures in some designs shifted by one register in the complex (Fig. 3.2.4A). However, the interactions in the predicted structures recapitulated those in our designs. In most binding modes, the hydrophobic interactions (Fig.

3.2.4B) are likely to be the primary contributors to the binding affinity, while the hydrophilic interactions (Fig. 3.2.4C) improve the binding specificity between the peptides and the repeat proteins.

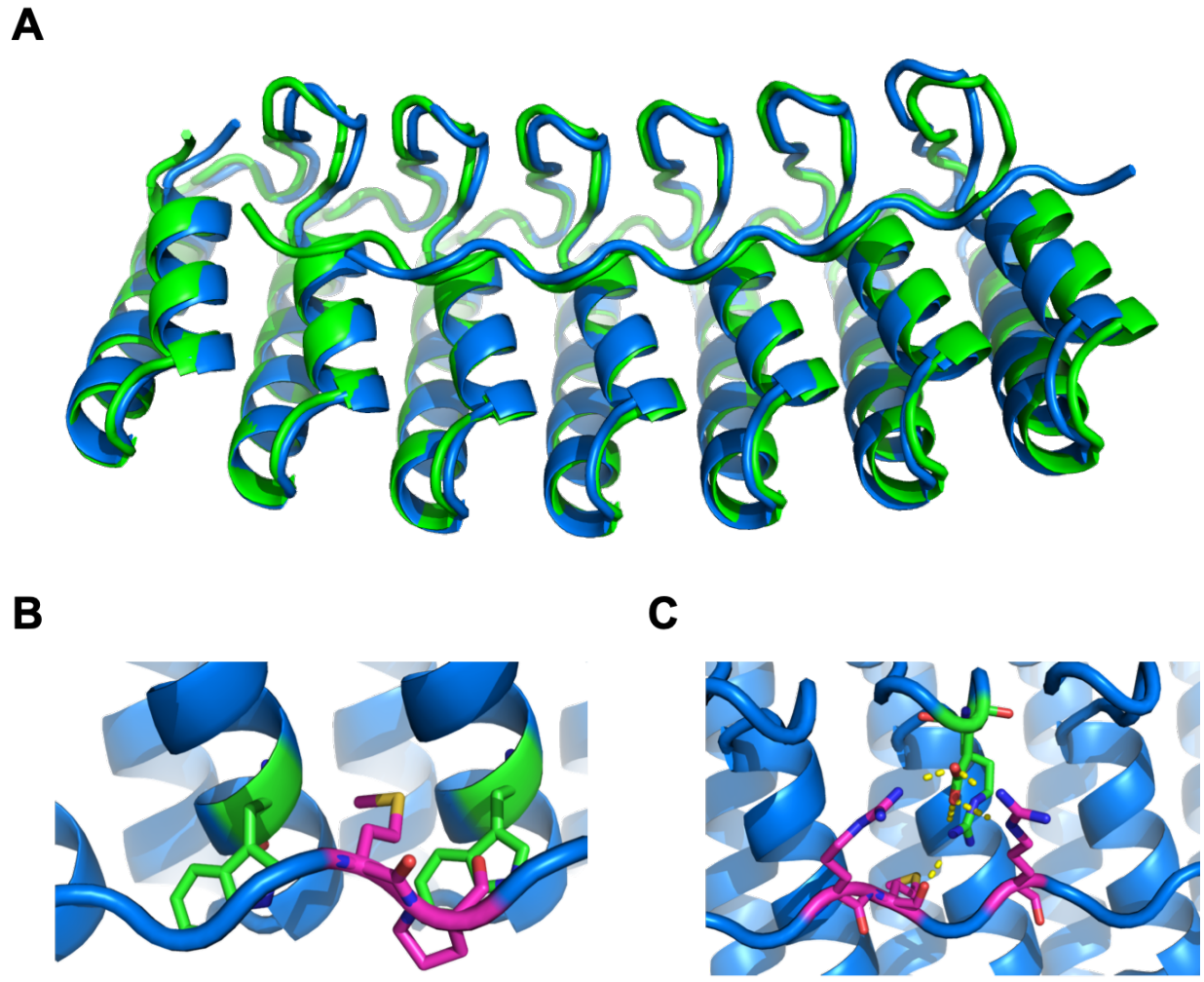

Fig. 3.2.4 A representative designed repeat protein in complex with repeat peptide. (A) Superimposition of designed repeat-protein-repeat-peptide complex (blue) with the AlphaFold-predicted complex (green). (B) Residues contributing to hydrophobic interactions between repeat protein (green) and repeat peptide (magenta). (C) Residues contributing to polar interactions between repeat protein (green) and repeat peptide (magenta).

We experimentally expressed and tested 25 designed repeat proteins, of which 4 designs showed detectable binding signals from the initial screening. From these designs, we selected one design targeting DLPx6 repeat peptide (Fig. 3.2.5A) and performed fluorescence polarization measurements for binding affinity. As shown in Fig. 3.2.5D, our designed binder displayed strongest binding towards its cognate peptide target DLPx6 with a KD of 1.32nM. To test the sequence specificity of the binding, we also measured its affinity towards three related repeat peptides: KLPx6, PLPx6 and DLSx6. The significantly weaker binding affinities against these non-cognate peptides, even though each peptide shares the identities for two of three amino acids with the cognate peptide, suggested that our designed binder was highly specific towards its target. The nearly 5000-fold difference of binding affinity between DLPx6 and DLSx6 suggested that the hydrophobic interactions involving proline residues (Fig. 3.2.5B), together with rigidity of peptide due to proline residues, contribute greatly to the protein-peptide affinity. In contrast, mutating aspartate residues of the peptide into lysine or proline led to only 40-fold or 10-fold decrease of binding affinity respectively. This indicated that the loop-peptide salt bridges (Fig. 3.2.5C) contribute mainly to the sequence-specific binding between the repeat protein and DLPx6 peptide.

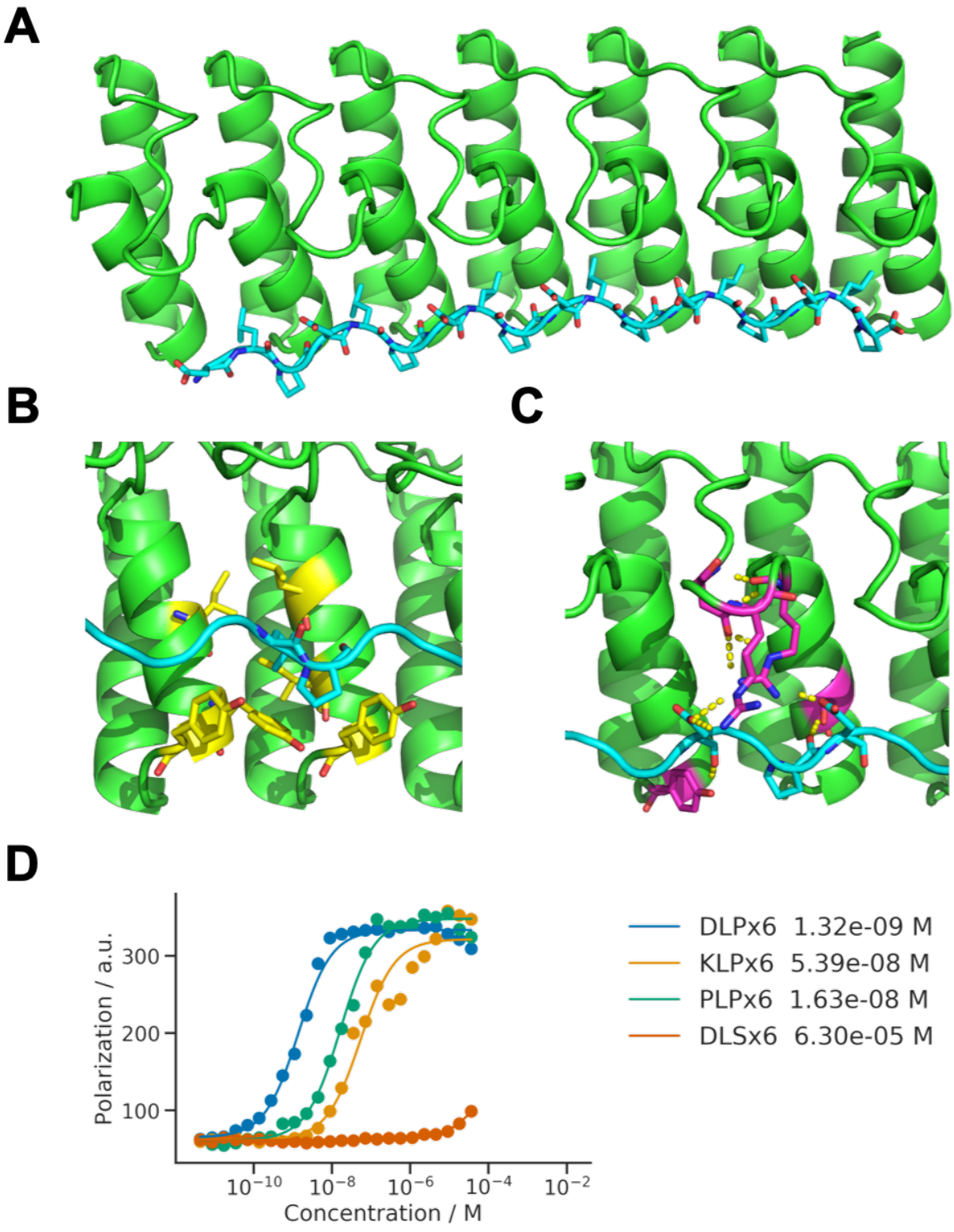

Fig. 3.2.5 A repeat protein binding DLPx6 peptide.

(A) Structural model of a repeat protein-DLPx6 complex where the repeat protein is colored in green and DLPx6 is colored in cyan. (B) Hydrophobic interactions where residues from the repeat protein are colored in yellow. (C) Hydrophilic interactions where residues from the repeat protein are colored in magenta. (D) Fluorescence polarization assay of binding affinity between the repeat protein and repeat peptides.

3.3 Discussion

In this chapter we described the application of the designed DHRs with buttressed loops in three tasks: binding protein against native complete protein targets, binding protein against a helical fragment of a therapeutic target and engineered repeat-peptide-repeat-protein pairs. Through the *in-silico* design and complex prediction experiments, we demonstrated that the buttressed loops can systematically improve the scores in the metrics predictive to binding affinity. We used a 128-scaffold library to design binding protein against a therapeutic targeting site on ApoE. The binders achieved modest binding affinities. We expected that further improvement of the design can be made via the expansion of the scaffold library. Due to its short length, the ApoE peptide fragment might have a complex conformational ensemble which contains both helical and unstructured forms. Consequently, a greater entropic cost is needed to achieve binding ApoE peptide fragment in its helical conformations. Given that this fragment exists in the helical form in the complete ApoE structures[69], we reasoned that an extended version of the fragment can assist the stabilization of its helical conformation and improve the measured binding affinity. Our design and characterization of repeat-peptide-repeat-protein pairs showed that it is possible to extend and

idealize the native sequence motifs of ankyrin binding peptides. This approach shed light on the generalization of using modular features from repeat proteins for engineered signaling networks.

3.4 Materials and Methods

We used the recently developed protein interface design method for *in-silico* binder docking and design experiments as well as binder design targeting the ApoE peptide fragment[66]. Docking of repeat peptides to the binder scaffold was guided by the geometric transformation between binding proteins of native ankyrins or DARPins to their targets in the crystal structures from PDB. Symmetric sequence design was performed for each docked peptide-protein pair following the same protocol as described in Chapter 2. All the designed complexes were computationally tested by AlphaFold[37, 38] before experimental characterization.

Designed proteins were expressed and purified following the same protocol described in Chapter 2. To conduct the fluorescence polarization binding assays, we synthesized ApoE peptide fragment that contained 5'-TAMRA labels. Serial dilutions of binder-peptide mixture were performed in the 96-well assay plates. The polarization signals were measured in a Synergy Neo2 hybrid multi-mode plate reader.

Chapter 4. Diversifying *De Novo* Designed TIM barrels with Buttressed Loops

4.1 Introduction

The TIM barrel folds represent the protein structure of approximately 10% of all enzymes and perform a great variety of catalytic functions[70]. Each TIM barrel consists of eight β/α folds, giving eight sites for loop insertion on the top of the barrel. In native TIM barrels, multiple extended loops are involved in shaping the pockets responsible for substrate recognition and enzyme catalysis (Fig. 4.1.1). While many efforts were made to redesign native TIM barrels for new functions[18, 71-74], progress in *de novo* design of functional TIM barrels has been thwarted by lacking highly stable base scaffolds and the challenging task of combinatorial loop sampling and design.

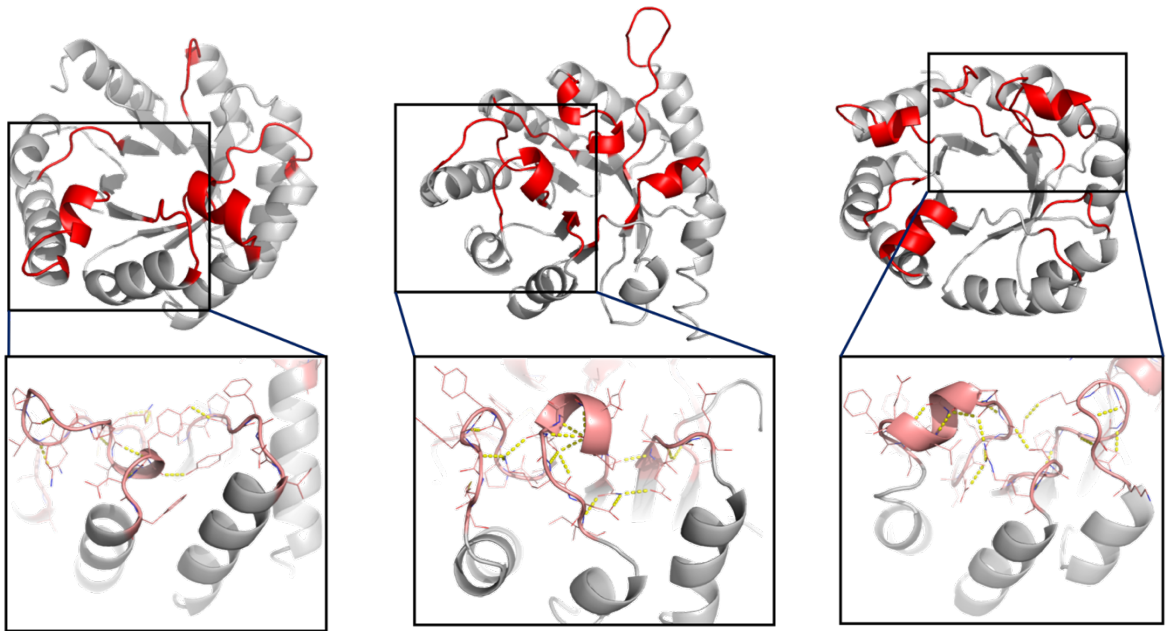

Fig. 4.1.1 Crystal structures of three native TIM barrels

The βα loops are colored in red with the buttressing interactions highlighted in the inset panels.

One of the recent breakthroughs in TIM barrel engineering is the success of a *de novo* designed TIM barrel scaffold[13]. This design was shown to be highly stable both thermodynamically and chemically. One unique feature of this novo TIM barrel is that the structure contains a C4 symmetry axis. The four identical repeats of βαβα unit enable the modular redesign of the TIM barrel. Within each repeat, the two βα structures offer two different insertion sites for loop installation.

To demonstrate that our method for designing buttressed loops can be generalized to all repeat proteins, we adapted the protocol for the C4 symmetry of the *de novo* TIM barrel and diversified the scaffold with four-/eight-loop insertions. Experimental characterizations suggested the designs were thermodynamically stable and monodisperse. These designs provide a new platform for engineering novel enzyme activities.

4.2 Results

We set out to adapt the loop sampling and design protocol described in Chapter 1 and Chapter 2 for loop design problem in the symmetric *de novo* TIM barrel. We observed that natural TIM barrels frequently used small secondary structural motifs (beta turns and short helices) in their βα loops. Inspired by this feature, we constructed a library of short helices (4-10aa) and combined it with the previously generated beta turn library for the hybrid loop sampling protocol. Another modification we made to the single loop composition was to optionally add a Gly residue at the end of the beta strand in each βα unit (Fig. 4.2.1). The initial sampling suggested that a Gly residue at this position can permit more loop conformations that would otherwise not be possible due to the steric clashes and high-energy backbone torsion angles.

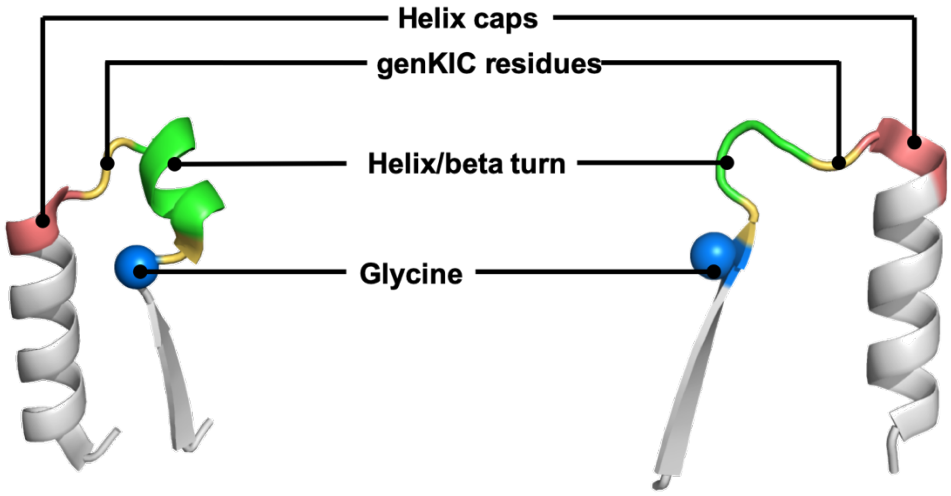

Fig. 4.2.1 Composition of a single loop in the loop sampling method

Using this method, we were able to generated diverse loop conformations which can be stabilized through buttressing. We expressed and purified the designed proteins. Of 12 characterized proteins, 7 were soluble expressed. Two designs were monomeric as suggested by size-exclusion chromatography (Fig. 4.2.3). The characteristic mixture of alpha helix and beta sheet signals in the circular dichroism spectrum at both 25°C and 95°C indicated that the designs were well folded and highly stable (Fig. 4.2.3).

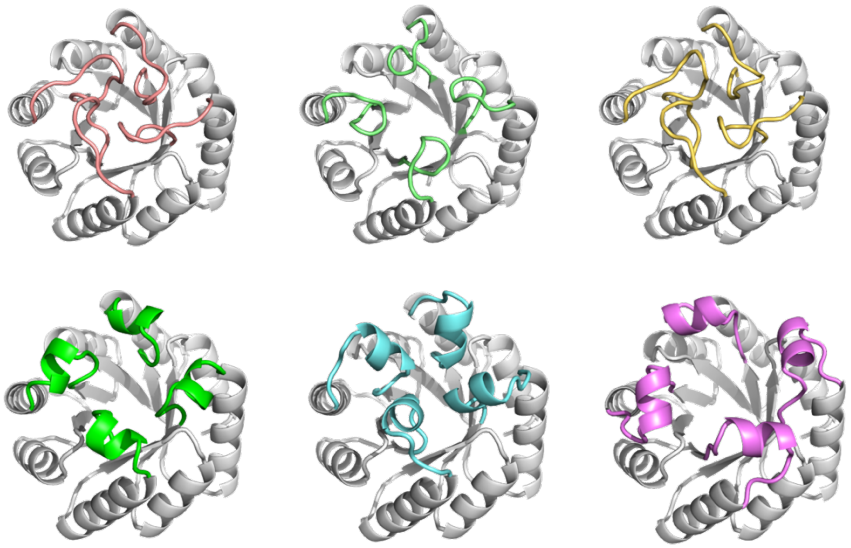

Fig. 4.2.2 Representative designs with long loops

The beta-turn fragment was used for the designs in the upper panel and a short helical fragment was used for the designs in the lower panel.

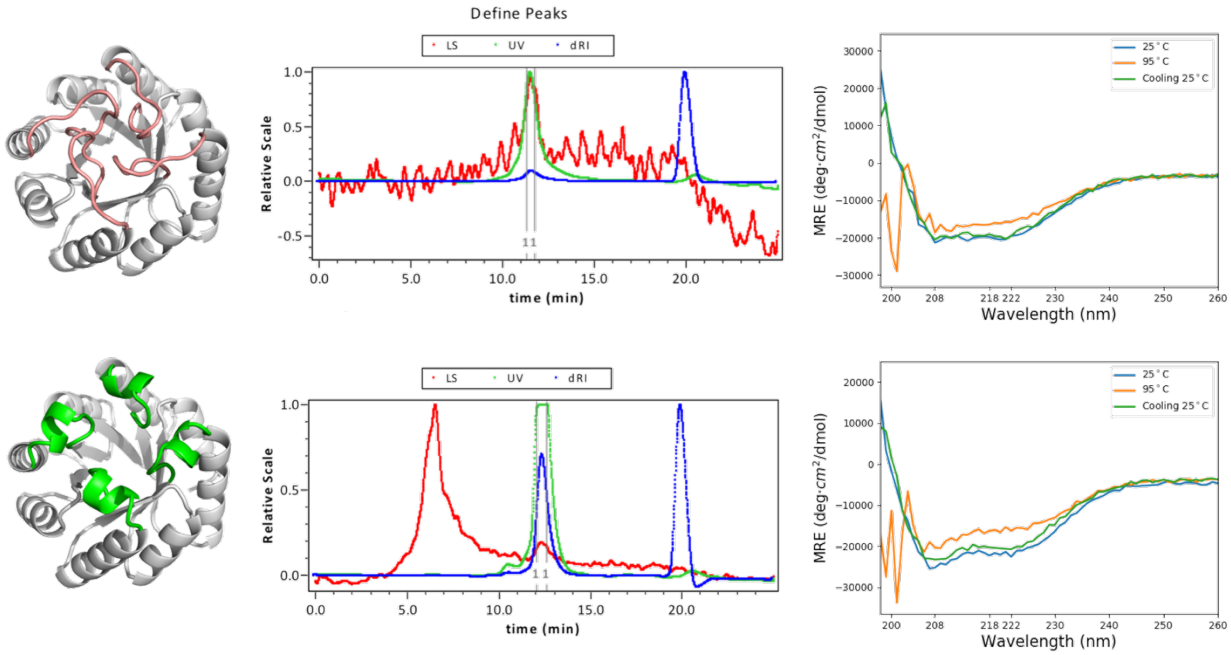

Fig. 4.2.3 Experimental characterization of designs.

In the middle column are the traces of multi-angle light scattering coupled with size exclusion chromatography (SEC-MALS), with light scattering (LS) signals colored in red, UV280 signals colored in green and differential refractive index (dRI) colored in blue. In the right column displays the traces from circular dichroism (CD).

To further explore the combinatorial loop conformation space, we designed and selected a set of loops for each unique βα connection. Of 20 designs experimentally tested, one was monomeric and displayed expected CD spectrum with high stability (Fig. 4.2.4).

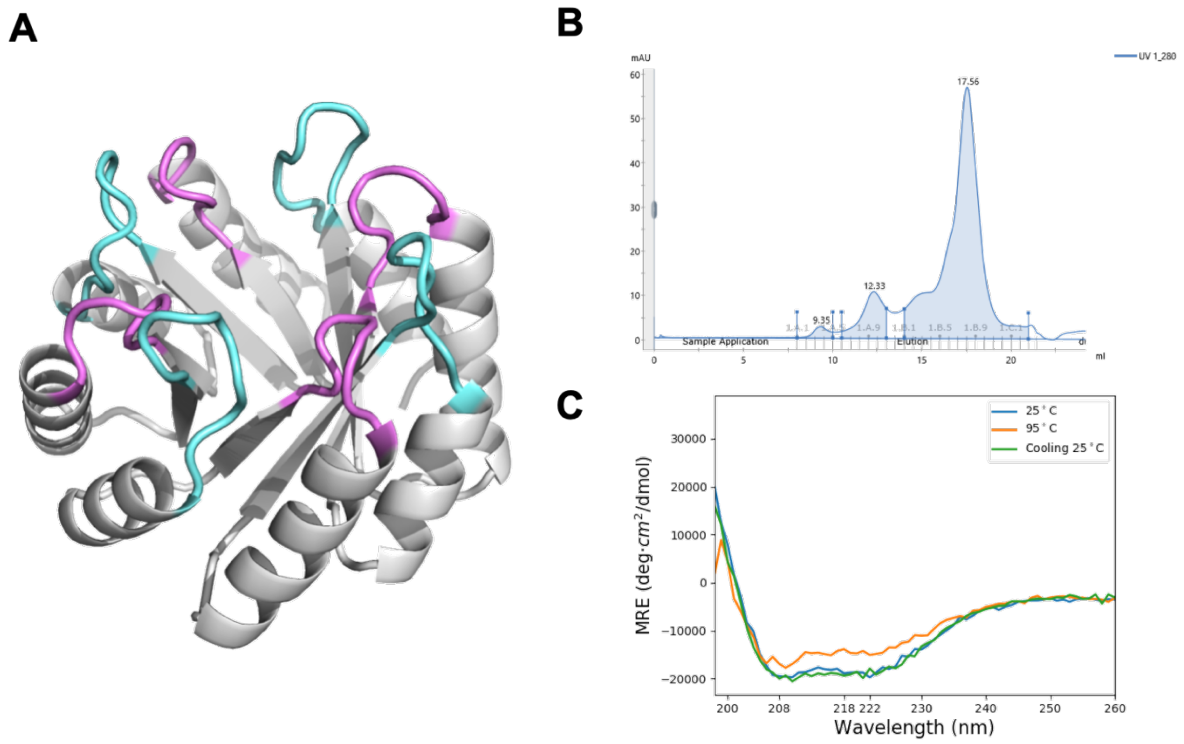

Fig. 4.2.4 Experimental characterization of *de novo* TIM barrel with eight buttressed loops (A) Designed model of *de novo* TIM barrel with two sets of loops colored in cyan and magenta. (B) Size exclusion chromatography of the design. (C) Circular dichroism of the design.

4.3 Discussion

We adapted the method developed for designing buttressed loops for DHRs to build loops for *de novo* symmetric TIM barrel. In particular, we demonstrated that the same method can successfully design four loops or eight loops without negative impact on the folding and thermal stability of the TIM barrel. The adaptions we made, incorporating short helical fragments in the loops and adding N-terminal Gly residues, were important for improving conformational diversity. We noted that, compared to buttressed loop designs in DHRs, fewer models were generated for the *de novo* TIM barrel. This was possibly resulted from the four-fold symmetric constraints. While the symmetry enabled the propagation of the loops, it limited the loop conformations, as few loop structures were compatible with the buttressing interactions. One strategy to further improve the diversity is to used two-fold symmetry which would allow combination of different loop conformations without significantly increasing the complexity of the loop sampling problem. The relatively low ratio of monomeric and folded proteins among all the tested designs can be attributed to the insufficient stabilization of the loops, but also the potentially sub-optimal stability of the *de novo* scaffold. Since natural TIM barrels have tendency to dimerize for improved stability, an approach to aid the functional engineering of the *de novo* TIM barrel is to design its oligomerized versions.

4.4 Materials and Methods

The *de novo* TIM barrel was initially designed by Huang *et al.*[13]. and optimized for improved stability by Romero-Romero *et al.*[73]. For four-loop insertion we removed the turns connecting the long βα units and for eight-loop insertion we removed the turns of both long βα and short βα units. Libraries of native helical capping motifs and native beta turns described in Chapter 1,

together with a newly constructed libraries of short idealized helical fragments (4-10 amino acids in length), were used for the hybrid loop sampling protocol. After being propagated to every repeat unit, the loops were filtered based on their compatible buttressing interactions. For full-length sequence design, we selected the loops that contain at least one backbone-to-backbone hydrogen bond or one side-chain-to-backbone bidentate hydrogen bond. We followed the same protocol and instruments described in previous chapters for expression, purification and experimental characterization of the selected designs.

BIBLIOGRAPHY

1. Tiller, K.E. and P.M. Tessier, *Advances in Antibody Design.* Annu Rev Biomed Eng, 2015. **17**: p. 191-216.

2. Zhang, J., M. Palangat, and R. Landick, *Role of the RNA polymerase trigger loop in catalysis and pausing.* Nat Struct Mol Biol, 2010. **17**(1): p. 99-104.

3. Wheatley, M., et al., *Lifting the lid on GPCRs: the role of extracellular loops.* Br J Pharmacol, 2012. **165**(6): p. 1688-1703.

4. Tsuchiya, Y. and K. Mizuguchi, *The diversity of H3 loops determines the antigen-binding tendencies of antibody CDR loops.* Protein Sci, 2016. **25**(4): p. 815-25.

5. Kadumuri, R.V. and R. Vadrevu, *LoopX: A Graphical User Interface-Based Database for Comprehensive Analysis and Comparative Evaluation of Loops from Protein Structures.* J Comput Biol, 2017. **24**(10): p. 1043-1049.

6. Kundert, K. and T. Kortemme, *Computational design of structured loops for new protein functions.* Biol Chem, 2019. **400**(3): p. 275-288.

7. Kuhlman, B. and P. Bradley, *Advances in protein structure prediction and design.* Nat Rev Mol Cell Biol, 2019. **20**(11): p. 681-697.

8. Fiser, A., R.K. Do, and A. Sali, *Modeling of loops in protein structures.* Protein Sci, 2000. **9**(9): p. 1753-73.

9. Li, Y., *Conformational sampling in template-free protein loop structure modeling: an overview.* Comput Struct Biotechnol J, 2013. **5**: p. e201302003.

10. Berman, H.M., et al., *The Protein Data Bank.* Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 6 No 1): p. 899-907.

11. Koga, N., et al., *Principles for designing ideal protein structures.* Nature, 2012. **491**(7423): p. 222-7.

12. Lin, Y.R., et al., *Control over overall shape and size in de novo designed proteins.* Proc Natl Acad Sci U S A, 2015. **112**(40): p. E5478-85.

13. Huang, P.S., et al., *De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy.* Nat Chem Biol, 2016. **12**(1): p. 29-34.

14. Rocklin, G.J., et al., *Global analysis of protein folding using massively parallel design, synthesis, and testing.* Science, 2017. **357**(6347): p. 168-175.

15. Chevalier, A., et al., *Massively parallel de novo protein design for targeted therapeutics.* Nature, 2017. **550**(7674): p. 74-79.

16. Marcos, E., et al., *Principles for designing proteins with cavities formed by curved beta sheets.* Science, 2017. **355**(6321): p. 201-206.

17. Marcos, E., et al., *De novo design of a non-local beta-sheet protein with high stability and accuracy.* Nat Struct Mol Biol, 2018. **25**(11): p. 1028-1034.

18. Thanki, N., et al., *Protein engineering with monomeric triosephosphate isomerase (monoTIM): the modelling and structure verification of a seven-residue loop.* Protein Eng, 1997. **10**(2): p. 159-67.

19. Hu, X., et al., *High-resolution design of a protein loop.* Proc Natl Acad Sci U S A, 2007. **104**(45): p. 17668-73.

20. MacDonald, J.T., et al., *Synthetic beta-solenoid proteins with the fragment-free computational design of a beta-hairpin extension.* Proc Natl Acad Sci U S A, 2016. **113**(37): p. 10346-51.

21. Murphy, P.M., et al., *Alteration of enzyme specificity by computational loop remodeling and design.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9215-20.

22. Lapidoth, G.D., et al., *AbDesign: An algorithm for combinatorial backbone design guided by natural conformations and sequences.* Proteins, 2015. **83**(8): p. 1385-406.

23. Baran, D., et al., *Principles for computational design of binding antibodies.* Proc Natl Acad Sci U S A, 2017. **114**(41): p. 10900-10905.

24. Krivacic, C., et al., *Accurate positioning of functional residues with robotics-inspired computational protein design.* Proc Natl Acad Sci U S A, 2022. **119**(11): p. e2115480119.

25. Gipson, B., et al., *Computational models of protein kinematics and dynamics: beyond simulation.* Annu Rev Anal Chem (Palo Alto Calif), 2012. **5**: p. 273-91.

26. Papaleo, E., et al., *The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery.* Chem Rev, 2016. **116**(11): p. 6391-423.

27. Karami, Y., et al., *DaReUS-Loop: accurate loop modeling using fragments from remote or unrelated proteins.* Sci Rep, 2018. **8**(1): p. 13673.

28. Canutescu, A.A. and R.L. Dunbrack, Jr., *Cyclic coordinate descent: A robotics algorithm for protein loop closure.* Protein Sci, 2003. **12**(5): p. 963-72.

29. Coutsias, E.A., et al., *A kinematic view of loop closure.* J Comput Chem, 2004. **25**(4): p. 510-28.

30. Mandell, D.J., E.A. Coutsias, and T. Kortemme, *Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling.* Nat Methods, 2009. **6**(8): p. 551-2.

31. Bhardwaj, G., et al., *Accurate de novo design of hyperstable constrained peptides.* Nature, 2016. **538**(7625): p. 329-335.

32. Marks, C., et al., *Sphinx: merging knowledge-based and ab initio approaches to improve protein loop prediction.* Bioinformatics, 2017. **33**(9): p. 1346-1353.

33. Barozet, A., et al., *A reinforcement-learning-based approach to enhance exhaustive protein loop sampling.* Bioinformatics, 2020. **36**(4): p. 1099-1106.

34. Fallas, J.A., et al., *Computational design of self-assembling cyclic protein homo-oligomers.* Nat Chem, 2017. **9**(4): p. 353-360.

35. Boyken, S.E., et al., *De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity.* Science, 2016. **352**(6286): p. 680-7.

36. Jumper, J., et al., *Applying and improving AlphaFold at CASP14.* Proteins, 2021. **89**(12): p. 1711-1721.

37. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.

38. Jumper, J. and D. Hassabis, *Protein structure predictions to atomic accuracy with AlphaFold.* Nat Methods, 2022. **19**(1): p. 11-12.

39. Baek, M., et al., *Accurate prediction of protein structures and interactions using a three-track neural network.* Science, 2021. **373**(6557): p. 871-876.

40. Kadumuri, R.V. and R. Vadrevu, *Diversity in alphabeta and betaalpha Loop Connections in TIM Barrel Proteins: Implications for Stability and Design of the Fold.* Interdiscip Sci, 2018. **10**(4): p. 805-812.

41. Wierenga, R.K., *The TIM-barrel fold: a versatile framework for efficient enzymes.* FEBS Lett, 2001. **492**(3): p. 193-8.

42. Kumar, A. and J. Balbach, *Folding and Stability of Ankyrin Repeats Control Biological Protein Function.* Biomolecules, 2021. **11**(6).

43. Pluckthun, A., *Designed ankyrin repeat proteins (DARPins): binding proteins for research, diagnostics, and therapy.* Annu Rev Pharmacol Toxicol, 2015. **55**: p. 489-511.

44. Bork, P., *Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally?* Proteins, 1993. **17**(4): p. 363-74.

45. Mosavi, L.K., D.L. Minor, Jr., and Z.Y. Peng, *Consensus-derived structural determinants of the ankyrin repeat motif.* Proc Natl Acad Sci U S A, 2002. **99**(25): p. 16029-34.

46. Parmeggiani, F., et al., *A general computational approach for repeat protein design.* J Mol Biol, 2015. **427**(2): p. 563-75.

47. Leaver-Fay, A., et al., *ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules.* Methods Enzymol, 2011. **487**: p. 545-74.

48. Leman, J.K., et al., *Macromolecular modeling and design in Rosetta: recent methods and frameworks.* Nat Methods, 2020. **17**(7): p. 665-680.

49. Marcelino, A.M. and L.M. Gierasch, *Roles of beta-turns in protein folding: from peptide models to protein engineering.* Biopolymers, 2008. **89**(5): p. 380-91.

50. Venkatachalam, C.M., *Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units.* Biopolymers, 1968. **6**(10): p. 1425-36.

51. Wang, G. and R.L. Dunbrack, Jr., *PISCES: a protein sequence culling server.* Bioinformatics, 2003. **19**(12): p. 1589-91.

52. Gonzalez, T.F., *Clustering to minimize the maximum intercluster distance.* Theor. Comput. Sci., 1985(38): p. 293-306.

53. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.

54. Yu, X., et al., *Beyond Antibodies as Binding Partners: The Role of Antibody Mimetics in Bioanalysis.* Annu Rev Anal Chem (Palo Alto Calif), 2017. **10**(1): p. 293-320.

55. Simeon, R. and Z. Chen, *In vitro-engineered non-antibody protein therapeutics.* Protein Cell, 2018. **9**(1): p. 3-14.

56. Stumpp, M.T., K.M. Dawson, and H.K. Binz, *Beyond Antibodies: The DARPin((R)) Drug Platform.* BioDrugs, 2020. **34**(4): p. 423-433.

57. Xu, C., et al., *Sequence-specific recognition of a PxLPxI/L motif by an ankyrin repeat tumbler lock.* Sci Signal, 2012. **5**(226): p. ra39.

58. Shi, D.J., et al., *Crystal structure of the N-terminal ankyrin repeat domain of TRPV3 reveals unique conformation of finger 3 loop critical for channel function.* Protein Cell, 2013. **4**(12): p. 942-50.

59. Michaely, P., et al., *Crystal structure of a 12 ANK repeat stack from human ankyrinR.* EMBO J, 2002. **21**(23): p. 6387-96.

60. Brunette, T.J., et al., *Exploring the repeat protein universe through computational protein design.* Nature, 2015. **528**(7583): p. 580-4.

61. Huang, P.S., et al., *RosettaRemodel: a generalized framework for flexible backbone protein design.* PLoS One, 2011. **6**(8): p. e24109.

62. Kobe, B. and A.V. Kajava, *When protein folding is simplified to protein coiling: the continuum of solenoid protein structures.* Trends Biochem Sci, 2000. **25**(10): p. 509-15.

63. Banatao, D.R., et al., *An approach to crystallizing proteins by synthetic symmetrization.* Proc Natl Acad Sci U S A, 2006. **103**(44): p. 16230-5.

64. Studier, F.W., *Protein production by auto-induction in high density shaking cultures.* Protein Expr Purif, 2005. **41**(1): p. 207-34.

65. Marcotte, E.M., et al., *A census of protein repeats.* J Mol Biol, 1999. **293**(1): p. 151-60.

66. Cao, L., et al., *Design of protein binding proteins from target structure alone.* Nature, 2022.

67. Lynch, J.R., et al., *APOE genotype and an ApoE-mimetic peptide modify the systemic and central nervous system inflammatory response.* J Biol Chem, 2003. **278**(49): p. 48529-33.

68. Williams, T., D.R. Borchelt, and P. Chakrabarty, *Therapeutic approaches targeting Apolipoprotein E function in Alzheimer's disease.* Mol Neurodegener, 2020. **15**(1): p. 8.

69. Chen, J., Q. Li, and J. Wang, *Topology of human apolipoprotein E3 uniquely regulates its diverse biological functions.* Proc Natl Acad Sci U S A, 2011. **108**(36): p. 14813-8.

70. Goldman, A.D., J.T. Beatty, and L.F. Landweber, *The TIM Barrel Architecture Facilitated the Early Evolution of Protein-Mediated Metabolism.* J Mol Evol, 2016. **82**(1): p. 17-26.

71. Lapidoth, G., et al., *Highly active enzymes by automated combinatorial backbone assembly and sequence design.* Nat Commun, 2018. **9**(1): p. 2780.

72. Nagarajan, D., G. Deka, and M. Rao, *Design of symmetric TIM barrel proteins from first principles.* BMC Biochem, 2015. **16**: p. 18.

73. Romero-Romero, S., et al., *The Stability Landscape of de novo TIM Barrels Explored by a Modular Design Approach.* J Mol Biol, 2021. **433**(18): p. 167153.

74. Romero-Romero, S., et al., *Evolution, folding, and design of TIM barrels and related proteins.* Curr Opin Struct Biol, 2021. **68**: p. 94-104.