

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Precise Design of Protein Structures

Permalink

<https://escholarship.org/uc/item/4bn8g9kk>

Author

Pan, Xingjie

Publication Date

2020

Peer reviewed|Thesis/dissertation

Precise Design of Protein Structures

by

Xingjie Pan

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

AND

UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:

DocuSigned by:

Tanja Kortemme

Tanja Kortemme

FA555ADA28F1439...

Chair

DocuSigned by:

William Degrado

William Degrado

DocuSigned by:

John Dueber

John Dueber

B016800A3FAF4B2...

Committee Members

Copyright 2020

by

Xingjie Pan

Acknowledgements

Five years ago, I flew across the Pacific Ocean for the dream of science. So many adventures happened in the past five years. So many people gave me love and help. First, I want to thank my family, my mother Peiyan Wang, my father Lisheng Pan and my girlfriend Wenbo Xu. Their unconditional love and support gave me courage to pursue the career that I want to devote myself to. I also want to thank my mentor Tanja Kortemme. She not only taught me how to be a good scientist, but also inspired me with her endless enthusiasm to scientific questions. Furthermore, I would like to thank all the Kortemme lab members, especially Kale Kundert, Yang (Sunny) Zhang, Kyle Barlow, Samuel Thompson, Tina Perica, Anum Azam Glasgow, James Lucas, Chris Mathy, Cody Krivacic, Amanda Loshbaugh and Simon Kretschmer. They made the lab such a friendly and vigorous environment for research. Finally, I want to thank all my friends. They are the stars that make the night sky so beautiful.

Precise Design of Protein Structures

Xingjie Pan

Abstract

Proteins perform most of biological functions. The ultimate approach to understand protein functions is designing novel protein functions from scratch. The ability to design novel functional proteins would also be revolutionary for biology research, medicine and synthetic biology. Proteins function by placing specific chemical groups at precise 3-dimensional (3D) positions and orientations. Therefore, *de novo* functional protein design relies on precise control of protein 3D structures. In this thesis, I review the recent advances in *de novo* protein design (chapter 1). Then I describe two of my research projects that aimed to extend the ability of computational methods to design precise protein 3D structures. In the first project (chapter 2), I developed a method termed loop-helix-loop unit combinatorial sampling (LUCS) that systematically samples geometries of protein secondary structures. I applied LUCS to design *de novo* protein fold families of two topologies. In the second project (chapter 3), I developed two computational methods termed fragment kinematic closure (FKIC) and loophash kinematic closure (LHKIC) for modeling protein local segments. FKIC predicts protein local segment structures with significantly higher accuracy and efficiency than previous methods. LHKIC designs local segment structures by combining the power of the loophash algorithm and the kinematic closure algorithm. I expect these new methods can bring the ideal of designing arbitrary protein functions closer to reach.

Contents

Chapter 1 Recent advances in <i>de novo</i> protein design principles and methods.....	1
Introduction.....	1
Scoring functions	2
Membrane scoring functions.....	2
Scoring non-protein molecules	3
TERM-based scoring	4
Protein scoring by machine learning models	4
Backbone generation.....	5
Helical bundles.....	5
<i>De novo</i> design by assembling local structures	6
Rational design.....	7
Symmetric design.....	8
Fold family design	8
Cyclic peptides.....	9
Machine learning backbone design.....	9
Side chain sampling	10
Side chain design with flexible backbones	11
Hydrogen bond networks	12
Machine learning sequence design	13
Side chain design by quantum computing	13

Function design	14
Ligand binding sites	14
Protein binders	15
Protein assemblies	16
Membrane proteins	17
Conformational changes	17
Protein switches	18
Future perspectives	19
Chapter 2 Computational design of <i>de novo</i> protein fold families	25
Introduction	25
Results and Discussion	27
Materials and Methods	35
Note	53
Chapter 3 Precise Local Segment Structure Prediction and Design	102
Introduction	102
FKIC and LHKIC algorithms	103
FKIC benchmarking	105
Discussion	108
Methods	109
Notes	121

References	153
------------------	-----

List of figures

Figure 1.1 Advances in scoring functions.....	21
Figure 1.2 Advances in backbone generation.....	22
Figure 1.3 Advances in sidechain sampling	23
Figure 1.4 Advances in functional design.....	24
Figure 2.1 LUCS sampling strategy to create <i>de novo</i> designed protein fold families with tunable geometries.	54
Figure 2.2 Diversity of geometries in naturally occurring fold topologies to enable distinct functions.....	55
Figure 2.3 Common loop connector elements in naturally occurring LHL units.....	56
Figure 2.4 Detailed flowchart of the LUCS design protocol.	57
Figure 2.5 Characterization of well-folded designs.....	58
Figure 2.6 ^{15}N - ^1H HSQC spectra of designs whose structures were solved by NMR.....	68
Figure 2.7 Close agreement between models and experimentally determined structures of designed proteins.	69
Figure 2.8 Conformations of core side chains overall agree between designed models and structures solved by NMR.	70
Figure 2.9 Comparison of loop conformations between designed models and structures solved by NMR.	71
Figure 2.10 Geometry space sampled by <i>de novo</i> designed fold families.....	72
Figure 2.11 Distributions of LHL geometries in designed <i>de novo</i> fold families.	74

Figure 2.12 Models of well-folded designs show considerable geometry differences to their best matches from TM-align.	75
Figure 2.13 Graphical representation of binning the helix geometry space.	76
Figure 2.14 Structural features encoding distinct protein geometries.	77
Figure 2.15 Residues that form knobs-into-sockets packing.	78
Figure 2.16 Folding simulations for selected designs and their mutations.	79
Figure 2.17 Identification of tertiary motifs for designed models.	82
Figure 2.18 Number of buried unsatisfied hydrogen bond acceptors found in the top8000 protein structure data set.	85
Figure 3.1 FKIC and LHKIC protocols.	127
Figure 3.2 FKIC improves structure prediction accuracy.	128
Figure 3.3 Examples of failures of FKIC.	129
Figure 3.4 Detailed FKIC/LHKIC protocol.	130

List of tables

Table 2.1 Loop libraries.....	86
Table 2.2 Number of models at each design stage.....	87
Table 2.3 Protein sequences of designs selected for experimental testing.	88
Table 2.4 DNA sequences of designs selected for experimental testing.....	92
Table 2.5 Experimental characterization of designs.....	98
Table 2.6 NMR statistics.....	100
Table 2.7 X-ray data reduction and model refinement.	101
Table 3.1 Datasets and performance summary.....	131
Table 3.2 Mixed Segment dataset detailed performance.....	132
Table 3.3 Multiple Segments dataset detailed performance.....	136
Table 3.4 Standard dataset detailed performance.....	140
Table 3.5 Failure cases, Standard dataset.....	146
Table 3.6 Multiple Segments (12 residues) dataset detailed performance.....	147
Table 3.7 Summary of energy function comparisons.....	149
Table 3.8 Multiple Segments (8 residues) dataset detailed performance.....	150
Table 3.9 Comparison of median RMSD of lowest energy models on perturbed structures	152

Chapter 1 Recent advances in *de novo* protein design

principles and methods

Introduction

There have been rapid developments in *de novo* protein design in the past few years (Huang, et al., 2016) (Marcos & Silva, 2018) (Kuhlman & Bradley, 2019) (Leman, et al., 2020). Designer proteins are changing biomedical and synthetic biology research. Successful applications include sensing molecules (Bick, et al., 2017) (Glasgow, et al., 2019) (Quijano-Rubio, et al., 2020), blocking viral infections (Chevalier, et al., 2017) (Cao, et al., 2020), tuning immune systems (Silva, et al., 2019) (Mohan, et al., 2019), assembling biomaterials (Gonen, et al., 2015) (Chen, et al., 2019) (Shen, et al., 2018) and building protein logic gates (Langan, et al., 2019) (Chen, et al., 2020). A fundamental driving force for these successful applications is the development of computational design principles and methods, which addressed several major challenges in protein design.

The challenges for designing *de novo* functional proteins come from several aspects including scoring functions, backbone generation, side chain sampling and function design. In this chapter, I talk about the development of design principles and methods in these aspects. I focus on the advances made in the past 5 years. For readers interested in the history of *de novo* protein design, I recommend a recent review written by Korendovych and DeGrado (Korendovych & DeGrado, 2020).

Scoring functions

A scoring function in computational protein design is a function that distinguishes protein sequences/structures with desired properties from incorrect “decoys”. Because proteins need to fold to perform their functions and the folded structure is the lowest free energy state (Anfinsen, 1973), most scoring functions estimate free energies of structures. The pioneer protein energy function developed by (Levitt & Lifson, 1969) used harmonic terms for bond energies and a Lennard-Jones potential for van der Waals interactions. Modern physics-based energy functions (Brooks, et al., 1983) (Jorgensen & Tirado-Rives, 1988) (Cornell, et al., 1995) account for more energy terms such as electrostatics and hydrogen-bonding. An alternative approach to physics-based energy terms is using statistics from known structures to derive potential functions (Tanaka & Scheraga, 1976). The Rosetta scoring function (Alford, et al., 2017) is widely used in modern *de novo* protein design. The first version of the Rosetta scoring function (Simons, et al., 1997) was a statistical potential function derived from structures in the Protein Data Bank (PDB) (Bernstein, et al., 1978) (Berman, et al., 2000) using Bayesian statistics. Physics-based terms were incorporated later, making the current version of the Rosetta scoring function a hybrid function with both physics-based and statistical terms (Park, et al., 2016) (Alford, et al., 2017).

Membrane scoring functions

Scoring functions for soluble proteins take advantage of the large number of solved structures in the PDB to fit and validate parameters (Jacobson, et al., 2002) (Leaver-Fay, et al., 2013). Transmembrane proteins make up about 30% of open reading frames in known genomes (Wallin & Heijne, 1998) (Tan, et al., 2008). However, the current PDB only has structures of about one

thousand unique membrane proteins. Thus, the lack of structural data is an obstacle for developing membrane protein scoring functions. An early version of the Rosetta membrane scoring function (Yarov-Yarovoy, et al., 2005) used statistics from 28 transmembrane proteins to fit parameters and was validated by *ab initio* structure prediction (Rohl, et al., 2004) of 12 multipass membrane proteins. Recently, a new membrane scoring model (Alford, et al., 2020) was developed, which aims to capture the heterogeneous membrane environment (**Figure 1.1A**). The interface between bulk water and bulk lipid is modeled as a continuous transition of hydration fraction. Water filling pores are calculated by a convex-hull algorithm (Leman, et al., 2017). The water-to-bilayer transfer energy is calculated using the hydration fraction and the Moon and Fleming hydrophobicity scale (Cp & Kg, 2011). The new membrane model improves membrane protein orientation predictions, $\Delta\Delta G$ calculations, native structure discrimination, and native sequence recovery.

Scoring non-protein molecules

Many protein functions involve interactions with other types of molecules such as DNA, RNA, saccharides or small molecules. Expanding the types of molecules supported by scoring functions is critical for designing such protein functions. Scoring functions for DNA (Havranek, et al., 2004) and RNA (Das & Baker, 2007) have been successfully applied to structure prediction and design (Ashworth, et al., 2006) (Das, et al., 2010). Recently a scoring function was developed for saccharide and glycoconjugate structures (Labonte, et al., 2017) (**Figure 1.1B**). The scoring function incorporates the CHI energy function (Nivedha, et al., 2014) for carbohydrates calculated by quantum mechanics. Benchmarking results on docking problems showed that the scoring function can predict glyco-ligand binding. In contrast to polymers, small molecules have highly diverse combinations of chemical groups, making it challenging to transfer parameters calculated

for representative molecules to other molecules. A new approach (Park, et al., 2020) simultaneously optimized all parameters in a small molecule energy function guided by thousands of small molecule crystal structures (**Figure 1.1B**). The resulting scoring functions significantly improved docking success rate.

TERM-based scoring

In many protein design projects, the estimated free energies of structures are used to evaluate the fitness of sequences. However, this approach does not consider if there are alternative states a sequence can fold into that have even lower free energies. One way to overcome this limitation is by directly calculating the fitness in protein sequence space. Protein structures can be broken up into 3 dimensional local pieces called tertiary structural motifs (TERMs) (Zheng, et al., 2015). Half of the structures in the PDB can be described by only about 600 TERMS (Mackenzie, et al., 2016), indicating that the sequence preferences of each TERM could be used to evaluate the fitness of a sequence for a given local structure (**Figure 1.1C**). A strong correlation (Zheng, et al., 2015) was observed between the TERM derived scores and protein structure model accuracies from the Critical Assessment of Structure Prediction (CASP). Recently the TERM score was used to predict protein-peptide binding energies and design peptide binders of anti-apoptotic proteins Bfl-1 and Mcl-1 (Frappier, et al., 2019).

Protein scoring by machine learning models

The power of machine learning models to learn the statistical representations underlying rich sequence and structural data provides new perspectives for protein structure prediction and design (Alley, et al., 2019) (Xu, 2019) (Senior, et al., 2020) (Yang, et al., 2020) (**Figure 1.1D**). Neural

network models trained with sequence evolutionary data and structures from the PDB outperform traditional methods in structure prediction (Xu, 2019) (Senior, et al., 2020) (Yang, et al., 2020). Neural networks that predict inter-residue orientations can be repurposed for assessing the probability of the desired structure for a given sequence, which works as a scoring function for protein design (Norm, et al., 2020). Another deep convolutional neural network scoring function predicts the probability distribution of amino acid types at each residue position conditioned on the local environment (Anand, et al., 2020).

Backbone generation

Backbone structures determine the overall shapes of proteins, and therefore play a critical role in protein functions. Normal sized proteins have hundreds to thousands of backbone degrees of freedom. It is impossible to sample the backbone structure space by brute force. Because folded proteins need to have well packed cores and satisfied hydrogen bonds, only a small fraction of the backbone structure space can stably exist (Li, et al., 1996) (Helling, et al., 2001). A workaround to the difficulty of *de novo* backbone design is redesigning native backbone structures from the PDB for new functions (Glasgow, et al., 2019). Small structural changes such as backrub moves (Smith & Kortemme, 2008) can also be introduced to diversify the native backbones.

Helical bundles

Helical bundles were the first type of protein fold designed at atomic accuracy (Harbury, et al., 1998). Due to the regularity and symmetry, backbone structures of coiled-coil helical bundles can be sampled near exhaustively by Crick's parameterization (Crick, 1953). The systematic backbone

sampling method and the high stability (Huang, et al., 2014) of the fold make helical bundles a good model system for designing a broad scope of functions such as ligand binding (Polizzi & DeGrado, 2020), ion transport (Joh, et al., 2014) and switches (Langan, et al., 2019). More details on recent progress in coiled-coil design can be found in a review by Woolfson (Woolfson, 2017).

***De novo* design by assembling local structures**

De novo backbones beyond helical bundles can be designed by a fragment assembly strategy originally used in structure prediction (Bowers, et al., 2000) (**Figure 1.2A**). The first step of the method is defining a blueprint that specifies the lengths and relative orientations of secondary structures. Short fragments with desired secondary structures are extracted from the PDB. Then the fragments are assembled to a 3D protein model. The Top7 protein designed by this method has a fold topology that was never observed in nature (Kuhlman, et al., 2003). This strategy was developed to design *de novo* backbones of a number of different fold topologies. Each fold topology requires specific design rules derived from native structures. Idealized alpha-beta fold proteins favor certain β -hairpin chirality, end-point strand residue directions and loop torsion bins (Koga, et al., 2012) (Lin, et al., 2015). Proteins with curved β sheets need bulges and register shifts to enhance their β sheet curvatures (Marcos, et al., 2017). The jelly roll fold topology is constrained by loop conformation, side chain directionality and β -strand length (Marcos & Silva, 2018). Folding of β -barrel proteins requires glycine kinks and β -bulges to reduce Lennard–Jones repulsive interactions (Dou, et al., 2018). Traditionally, *de novo* designed proteins were validated by low throughput assays. Recent developments in large-scale DNA synthesis (Kosuri & Church, 2014) enabled high throughput stability screening of *de novo* designed small proteins (Rocklin, et al., 2017). The screen identified thousands of sequences encoding stable designs. The buried nonpolar

surface area and the local sequence-structure compatibility are significantly correlated with the success rates.

Nature recombines pieces of existing proteins to make new structures without any blueprints defined beforehand. Computational methods have mimicked a few of nature's strategies to assemble local structures. Proteins with controllable curvatures can be designed by combinations of modular leucine-rich-repeat units (Park, et al., 2015) (**Figure 1.2A**). The SEWING method (Jacobs, et al., 2016) combines continuous or discontinuous helical building blocks from arbitrary proteins (**Figure 1.2A**). SEWING first extracts small substructures from existing structures. Substructures that share high similarity in local regions are overlapped and combined. Finally, loops are designed to close the gaps between discontinuous elements. Helical proteins designed by SEWING have higher structural variance than previous *de novo* helical proteins. A recent method called AbDesign (Lipsh-Sokolik, et al., 2020) mimics natural homologous recombination. AbDesign breaks each protein from a structure family into large segments and recombines segments into new structures. AbDesign can build large numbers of new structures even for moderately sized families of homologs.

Rational design

Human rationale can design the atomic details of *de novo* proteins (**Figure 1.2B**). The online game Foldit (Cooper, et al., 2010) posed a challenge for players to design *de novo* proteins (Koepnick, et al., 2019). The players were provided with a set of tools to mutate, move, and score protein structures. Starting from a fully extended peptide chain, players were able to design *de novo* proteins that closely agree with the experimentally solved structures. Like the wisdom from crowd

citizens, experts' knowledge is useful for protein design. The TopoBuilder (Yang, et al., 2020) protocol lets designers build proteins in a bottom-up approach starting from functional motifs. Designers define the sizes and 3D coordinates of secondary structure elements. The coordinates are then transformed into constraints for the Rosetta FunFolDes (Bonet, et al., 2018) method to build a full atom model. The TopoBuilder protocol successfully designed a few protein binders.

Symmetric design

The complexity of the backbone design problem can be reduced by symmetry (**Figure 1.2C**). A 4-fold symmetric TIM-barrel was designed using the blueprint fragment assembly strategy (Huang, et al., 2016). Experimental characterization of the designs revealed important hydrogen bonds defining the strand register between repeat units. Tandem repeat proteins made of a series of identical helix–loop–helix–loop structural motifs can be systematically sampled (Brunette, et al., 2015). The designed repeat proteins span a broad range of curvatures. By modulating the curvature, alpha tandem repeat proteins can form closed toroid structures (Doyle, et al., 2015). A large number of proteins with diverse shapes can be generated by designing rigid junctions to connect helical repeat proteins (Brunette, et al., 2020).

Fold family design

Native proteins with the same fold topology can have distinct functions due to their geometric differences (Fox, et al., 2014) (Dawson, et al., 2017). Exploring geometric variation within fold families is critical for function designs which require precise placement of active sites. The LUCS method systematically samples loop-helix-loop geometries in arbitrary protein folds by near exhaustive testing of combinations of short loops (Pan, et al., 2020) (**Figure 1.2D**). The generated

geometries had similar distributions to those in native structures but also included thousands of new structures. Experimentally solved structures spanned a wide range in the sampled distribution. Another enumerative algorithm was developed to sample diverse pocket structures of NTF2 fold proteins (Basanta, et al., 2020). Parameters like sheet curvatures, loop types and secondary structure lengths are sampled during a hierarchical backbone assembly process. Thousands of stable designs with diverse pocket geometries were identified by a high-throughput yeast surface display experiment.

Cyclic peptides

Natural constrained peptides have strong pharmacological activities. The GenKIC method (Bhardwaj, et al., 2016) generalized the KIC algorithm (Mandell, et al., 2009) from loop modeling to sample cyclic peptides and peptides constrained by disulfide bonds (**Figure 1.2E**). The designed peptides closely matched the experimentally solved structures and showed high stability against thermal and chemical denaturation. Large scale backbone generation using GenKIC showed that backbones of cyclic peptides with 7 to 10 residues can be enumerated near exhaustively (Hosseinizadeh, et al., 2017). In combination with TERMS, GenKIC was applied to design meso-size proteins stabilized by multivalent cross-linkers (Dang, et al., 2017).

Machine learning backbone design

Machine learning models trained with the rich structural data from the PDB are able to build novel protein backbone structures (**Figure 1.2F**). A recent work developed a generative adversarial network (GAN) (Anand, et al., 2019) model that generates protein structures represented as pairwise distances between all backbone atoms. Then a pre-trained deep convolutional neural

network recovers the 3D backbone structure from pairwise distances. Some of the generated designs were validated by fragment based folding prediction (Bradley, et al., 2005). Another study developed a variational auto-encoder model that focuses on generating immunoglobulin structures (Eguchi, et al., 2020). The model can generate immunoglobulins with defined complementarity determining regions through latent vector optimization. Another method uses the idea of neural network hallucination for protein design (Anishchenko, et al., 2020). The model repurposes the neural network from TR-Rosetta (Yang, et al., 2020). The TR-Rosetta network is a fast method to predict the inter-residue contact map of an arbitrary sequence. A loss function is defined as Kullback-Leibler divergence between the TR-Rosetta neural network predicted contact map and a background distribution representing unfolded proteins. Novel sequences and structures can be designed simultaneously by optimizing the loss function through Monte Carlo simulated annealing. Diverse structures were designed by the model and shown to be well folded by experimental characterization.

Side chain sampling

Amino acid sequences determine protein structures and functions. In *de novo* protein design, side chains types and conformations are sampled to stabilize backbone structures and perform functions. Early *de novo* design works used amino acids that favor specific secondary structure types (Ho & DeGrado, 1987) or locations relative to the protein surfaces (Kamtekar, et al., 1993) to define protein structures. Because the side chain conformations are clustered as rotamers (Shapovalov & Dunbrack, 2011), side chain design can be formulated as a discrete optimization problem, i.e. find a combination of rotamers that minimize the energy of the structure. The

complexity of the problem grows exponentially with the increase of the number of residues. Small scale side chain design problems can be solved deterministically by the dead-end elimination algorithm (Desmet, et al., 1992). The complexities of modern *de novo* protein side chain design problems are too big to be solved deterministically. Currently, side chain design problems are usually solved by Monte Carlo methods (Lee & Subbiah, 1991) (Kuhlman & Baker, 2000), which do not guarantee to find the global minimum but the solutions are often sufficiently accurate for applications. The efficiency of side chain sampling methods can be improved by constraining the amino acid types allowed at each residue position. LayeredDesign is a common strategy (Rocklin, et al., 2017) (Marcos, et al., 2017) (Marcos & Silva, 2018) (Pan, et al., 2020) to constrain designable amino acid types (**Figure 1.3A**). In LayeredDesign, residue positions are divided into 3 categories, namely, core, boundary and surface. The core region only allows hydrophobic amino acids; the surface region only allows polar amino acids; and the boundary region allows all amino acids. The LayeredDesign method increases sampling speed and reduces artefacts, such as buried polar residues, that may result from insufficient sampling or scoring errors. To further eliminate flawed designs, the results from Monte Carlo samplers are often filtered by a set of properties such as core packing (Sheffler & Baker, 2010) and hydrogen bond satisfaction (Pan, et al., 2020) (**Figure 1.3B**). A high throughput stability screen of designed small proteins showed that buried nonpolar surface area and local sequence-structure compatibility had strong correlations with the stabilities of designs (Rocklin, et al., 2017) .

Side chain design with flexible backbones

Solutions of fixed backbone side chain design problems are sensitive to the backbone structures. Because the Lennard-Jones potential term in score functions scales as 12^{th} power of distance when

two atoms are close to each other, a small adjustment of backbone structures may result in a large energy change. State-of-the-art side chain design methods often sample both side chain rotamers and local backbone conformations (**Figure 1.3C**). A study (Loshbaugh & Kortemme, 2020) benchmarked several flexible backbone side chain design methods including CoupledMoves, BackrubEnsemble and FastDesign as well as a fixed backbone design method. The methods that exploit backbone flexibility outperformed fixed backbone design in recapitulating observed protein sequence profiles. Among the flexible backbone methods, CoupledMoves had the best performance.

Hydrogen bond networks

Hydrogen bonds play an important role in the specificity of protein-ligand and protein-protein interactions. The formation of a hydrogen bond only allows narrow ranges of distances and orientations between donor groups and acceptor groups. Almost all the hydrogen bond donor or acceptor groups in a protein must form hydrogen bonds within the protein or with solvent molecules to avoid large energetic penalties of unsatisfied hydrogen bonds. The HBNet method addresses the challenges for hydrogen bond design by systematically searching possible hydrogen bond networks (Boyken, et al., 2016) (**Figure 1.3D**). HBNet constructs a graph whose nodes are rotamers that have hydrogen bond donors or acceptors. Two nodes are connected by an edge if the rotamers of the nodes can form hydrogen bonds. Hydrogen bond networks can be generated by traversing the graph. HBNet was successfully applied to design helical bundle homo-oligomers with specificity mediated by hydrogen bond networks. A Monte Carlo version of the HBNet method uses a stochastic algorithm to traverse the HBNet graph (Maguire, et al., 2018). This new approach significantly improved the sampling speed and makes larger design problems possible.

Machine learning sequence design

A number of machine learning methods for protein sequence design were developed recently (**Figure 1.3E**). Deep neural network methods were trained to predict probabilities of amino acids at each residue position of a backbone structure (Wang, et al., 2018) (Anand, et al., 2020). Generative models including generative adversarial networks (GAN) (Karimi, et al., 2019) (Repecka, et al., 2019), variational autoencoders (VAE) (Davidsen, et al., 2019) (Hawkins-Hooker, et al., 2020) , and graph-based (Ingraham, et al., 2019) (Strokach, et al., 2019) models were able to generate novel protein sequences with or without backbone structures. The structure prediction neural network from TR-Rosetta (Yang, et al., 2020) can also be repurposed for sequence design (Norn, et al., 2020). For a protein sequence, the TR-Rosetta neural network predicts distances, angles and dihedrals for every pair of residues. A loss function is defined as the difference between the prediction and the target structure. The gradient of the loss is back-propagated through the TR-Rosetta neural network to optimize the sequence. Since machine learning models and traditional Monte Carlo samplers use different information for sequence design, combining the two methods improved performance over each single method (Wang, et al., 2018) (Norn, et al., 2020).

Side chain design by quantum computing

Algorithms to deterministically solve the side chain design problem have exponential time complexity on electronic computers. Problems with exponential time complexity on electronic computers may be solved in polynomial time on quantum computers. The QPacker method (Mulligan, et al., 2019) maps the fixed backbone design problem to the D-Wave quantum computer (**Figure 1.3F**). QPacker was applied to design side chains of small peptides with up to tens of

rotamers. Both QPacker and classical computer algorithms were able to find global energy minima for the test problems. Within the achievable problem sizes, QPacker had nearly constant run time. Although current quantum computer hardware only works for side chain design problems with a small number of rotamers, future development of quantum computers may enable the solution of global energy minima of normal sized proteins.

Function design

Proteins perform functions by placing atoms with certain physical chemical properties at specific positions in the 3-dimensional space. Initial work on functional protein design directly borrowed native functional site motifs (Robertson, et al., 1994) (Choma, et al., 1994). Development of *de novo* protein structure design is gradually making the ideal of precise placement of arbitrary functional groups possible.

Ligand binding sites

Ligand binding is a common function for native proteins. *De novo* ligand binding site design requires high accuracy in sampling and scoring. Specificity of ligand binding is often realized by polar interactions which are highly sensitive to the positions and orientations of polar groups. A mis-aligned hydrogen bond could cause a few kcal/mol free energy penalty and reduce the binding affinity by an order of magnitude. Early work designed *de novo* binding sites by manually defining side chains that form favorable interactions with ligands (Tinberg, et al., 2013) (Polizzi, et al., 2017) (Bick, et al., 2017). An effort that uses HBNet and a Monte Carlo sequence design algorithm to design hydrogen bonds resulted in designs that bind to ligands with flipped orientations (Dou,

et al., 2017). The authors suggested that the flipped ligand orientation was because sampling methods failed to model subtle structural changes and the scoring function underestimated desolvation energies for the ligand. This negative result highlights the challenges inherent in sampling and energy evaluation in binding site designs. Recent development in binding site generation methods address these challenges. The RIF docking method (Dou, et al., 2018) generates an ensemble of billions of discrete amino acid side chains that make hydrogen-bonding and non-polar hydrophobic interactions with the target ligand. The ensemble of side chains are matched to the backbone scaffold during which an enumerative search of scaffold positions is performed. RIF docking was successfully applied to design a binding site for the fluorogenic compound DFHBI on a *de novo* beta barrel scaffold (Dou, et al., 2018). Two other methods utilize the structural information in the PDB to generate binding site ensembles (Lucas & Kortemme, 2020) (Polizzi & DeGrado, 2020). These methods break the ligand into small fragments and find residues that interact with the fragments from the PDB. The interacting residues are combined into binding sites by Monte Carlo simulated annealing (Lucas & Kortemme, 2020) or built onto backbone scaffolds by an algorithm called Convergent Motifs for Binding Sites (COMBS) (Polizzi & DeGrado, 2020) (**Figure 1.4A**). The COMBS method was applied to engineer *de novo* proteins that bind the drug apixaban with sub-micromolar affinity.

Protein binders

Like ligand binding site design, designing binders to target proteins requires high accuracy scoring and sampling. A workaround to these challenges is using binding motifs from known interfaces. Proteins that bind to influenza haemagglutinin and botulinum neurotoxin B (Chevalier, et al., 2017) were designed by building known helical motifs that bind to targets onto *de novo* designed

small protein scaffolds (Rocklin, et al., 2017). More than two thousand high-affinity binders were validated by a high throughput yeast surface display assay. Likewise, proteins that bind to the interleukin-2 (IL-2) and interleukin-15 (IL-15) receptors were designed by building a helical bundle from interface helices of native IL-2 and IL-15 (Silva, et al., 2019). Though difficult, the interface interactions can be designed without native motifs. Recently, the RIF docking method originally developed for small molecule binding site design was applied to design small proteins that bind to the human angiotensin-converting enzyme 2 (ACE2) (Cao, et al., 2020) (**Figure 1.4B**). The most potent design after experimental affinity optimization has picomolar affinity to ACE2.

Protein assemblies

Protein assembly is a type of protein-protein interface design, where both sides of each interface are designed (**Figure 1.4C**). Homo-oligomers with cyclic symmetries were designed by systematic enumeration of arrangements and motif guided interface design (Fallas, et al., 2017). A set of heterodimers that have orthogonal binding specificities were designed using parametric backbone generation and HBNet (Chen, et al., 2019). The orthogonal heterodimers can be used to design protein logic gates (Chen, et al., 2020). Self-assembled nano-cages with high order symmetries were designed by symmetric docking followed by Monte Carlo interface sequence design (King, et al., 2014) (Hsia, et al., 2016). Fusing the designed cages to membrane binding and endosomal sorting recruiting peptides induced the formation of nanocage-containing extracellular vesicles (Votteler, et al., 2016). The strategy of combining symmetric arrangement of protein chains and Monte Carlo interface sequence design was also successfully applied to design protein filaments (Shen, et al., 2018), amyloid like structures (Zhang, et al., 2018) and 2 dimensional materials (Gonen, et al., 2015) (Chen, et al., 2019).

Membrane proteins

Proteins that localize to phospholipid bilayer membranes have been designed since the emergence of *de novo* protein design (Lear, et al., 1988) (Whitley, et al., 1994). Membrane-spanning peptides that self-assemble into helical bundles were designed to perform functions such as cofactor binding (Korendovych, et al., 2010) and ion transport (Joh, et al., 2014). Recent advances significantly expanded the scope of membrane protein design. A study of the driving force of membrane protein stability showed that steric packing of nonpolar side chains alone is sufficient for the folding of membrane proteins (Mravic, et al., 2019). Using a steric packing code derived from the natural protein phospholamban, the authors were able to design a synthetic protein stabilized entirely by nonpolar side chains. Accurate multipass transmembrane proteins were designed (Lu, et al., 2018) utilizing a recently developed framework for membrane protein modeling (Alford, et al., 2015) (**Figure 1.4D**). In this work, the parametrically generated backbones were stabilized by HBNet designed hydrogen bond networks and Monte Carlo side chain design. Orientations of the designs were specified by incorporating a ring of amphipathic aromatic residues at the lipid-water boundary on the extracellular side and a ring of positively charged residues on the cytoplasmic side. This strategy was then applied to design transmembrane pores (Xu, et al., 2020). A designed 12-helix pore selectively passed potassium over sodium and a designed 16-helix pore enabled the passage of biotinylated Alexa Fluor 488.

Conformational changes

One of the most challenging functions to design is conformational changes between multiple states. A single state design would be successful as long as the designed state is a deep global

minimum, so that the scoring errors can be tolerated (Baker, 2019). However, multi-state design requires considerable accuracy in scoring relative stabilities, such that the probability distributions among multiple states can be modeled correctly. Additionally, multi-state design must simultaneously optimize several objectives, for example energies of each state and energy differences between states. This multiple objective optimization problem adds significant challenges to sequence design. A recently developed meta-multistate design (meta-MSD) protocol designed a protein that has a tryptophan side chain switching between defined conformational states on the millisecond timescale (Davey & Chica, 2017) (**Figure 1.4E**). Meta-MSD begins with generating a backrub ensemble of backbones. Side chains are designed by optimizing the Boltzmann-weighted average energy of all members from the ensemble using the FASTER algorithm (Allen & Mayo, 2009). The energy landscape of a designed sequence was estimated using energies of each backbone structure from the ensemble. Sequences with energy landscapes that support desired conformational dynamics were selected as final designs.

Protein switches

Protein switches change their conformations when triggered by external signals. It seems that there is an extra layer of complexity over designing proteins that adopt multiple conformations. However, designing switches is a more tractable problem (**Figure 1.4F**). The external trigger can introduce a large free energy bias toward one state, making the problem less sensitive to scoring errors. An early work designed a protein that switches between two distinct folds (Ambroggio & Kuhlman, 2006). The authors used a Monte Carlo side chain design method to optimize the sum of energies of the two folded states. The fold change was triggered by addition of Zn^{2+} ions. This success shows that it is possible to design protein switches by solving single objective optimization

problems. Following this principle, proteins were designed to change conformations in response to pH change (Boyken, et al., 2019) or Ca^{2+} ions (Wei, et al., 2020). A modular protein switch that senses a small molecule was designed through an induced dimerization mechanism (Glasgow, et al., 2019). A ligand binding site for farnesyl pyrophosphate was designed at the interfaces of a native heterodimer complex. The designed proteins dimerized in the presence of the ligand and transduced downstream signals such as fluorescence. LOCKR is another recently designed protein switch system (Langan, et al., 2019), which is made of one helical bundle and a helical peptide called key. The key peptide can replace a helix in the bundle and expose a signal on the replaced helix. The LOCKR system was used to induce protein degradation and localization (Langan, et al., 2019), target cells with precise combinations of surface antigens (Lajoie, et al., 2020) and detect viral proteins (Quijano-Rubio, et al., 2020).

Future perspectives

De novo protein design is at a rapidly developing stage. There are still many unsolved problems. In order to save computing time, scoring functions use a number of approximations such as implicit solvation models and pairwise decomposable scoring terms. Improving scoring accuracy and speed will continue to be an important direction. Current backbone geometry sampling methods are limited to certain secondary structures and fold topologies. Developing methods that cover the full designable backbone space will greatly expand reachable functions. The working size for the state-of-the-art side chain design methods are generally within tens to a few hundred residues without symmetry, while many proteins have over a thousand residues. This limit could be pushed by innovations in design algorithms or utilizing parallel computing architectures. Although a

variety of *de novo* protein functions have been designed, none of the functions can be designed routinely. Applying the design protocols on different problems or even testing the methods systematically can be valuable for standardizing the methods. The recent advances in designing basic functions including ligand binding, protein-protein interaction, membrane localization and induced switching are making it possible to design more complex functions such as cell surface receptors and motors. The emerging machine learning methods may potentially revolutionize the field in multiple aspects. There will be plenty of opportunities and challenges in this relatively new sub-field. The best design strategies for many problems might be combinations of machine learning models and traditional design methods. With that said, I foresee that *de novo* protein design will continue to develop rapidly in the future.

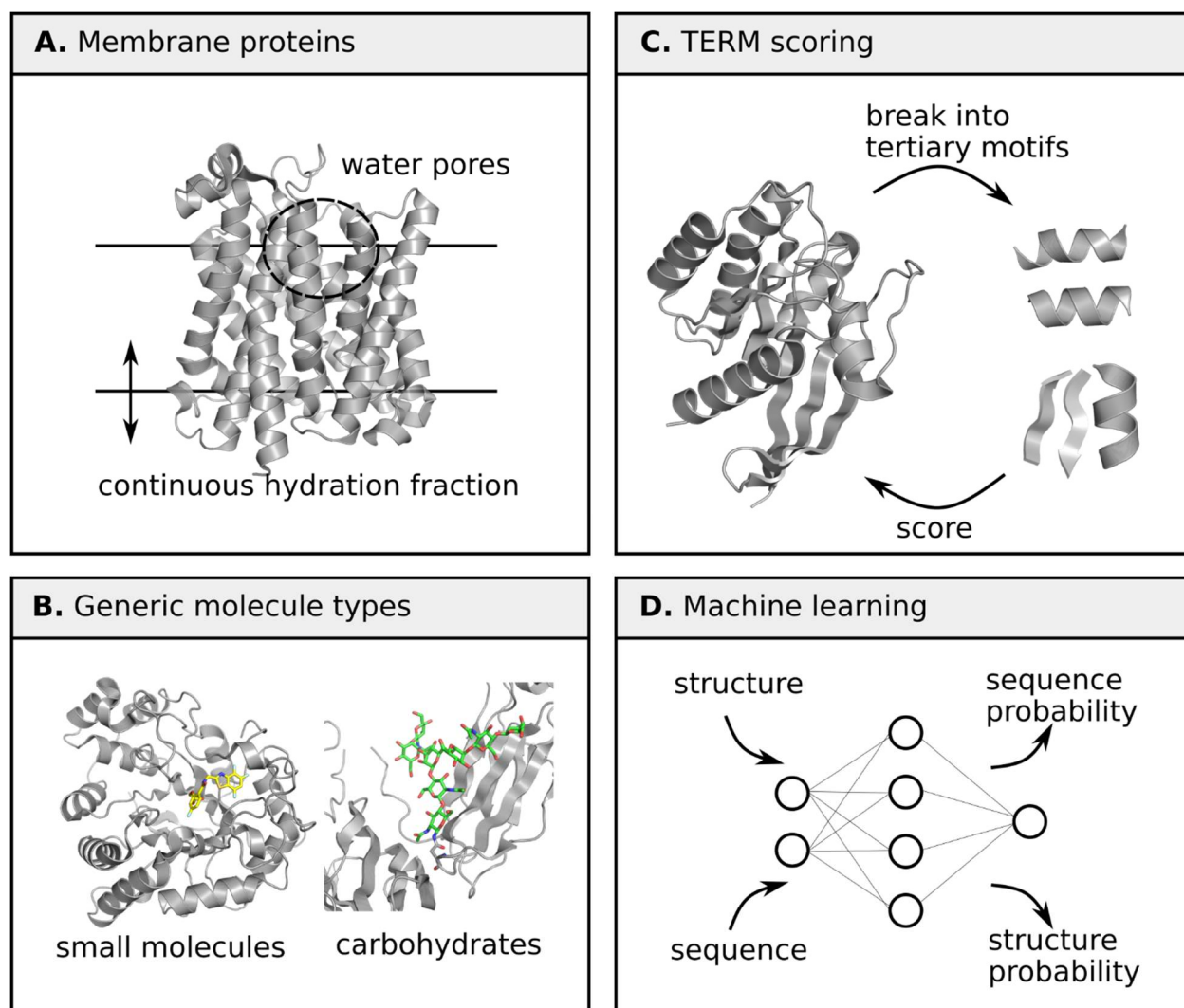


Figure 1.1 Advances in scoring functions

A. A membrane scoring function (Alford, et al., 2020) uses a continuous hydration fraction to calculate the free energy change of residues from water to the lipid environment. Water pores in membrane proteins are explicitly modeled. **B.** Protein design scoring functions are generalized to model small molecules (Park, et al., 2020) and carbohydrates (Nivedha, et al., 2014). **C.** The TERMS based scoring function (Zheng, et al., 2015) breaks proteins into tertiary structure motifs and evaluates the fitness of the sequence using the sequence profiles of the tertiary motifs. **D.** Machine learning methods predict the probability of sequences given a structure (Anand, et al., 2020) or the probability of structures given a sequence (Norn, et al., 2020). The predicted probabilities can be used as scores for the compatibility between sequences and structures.

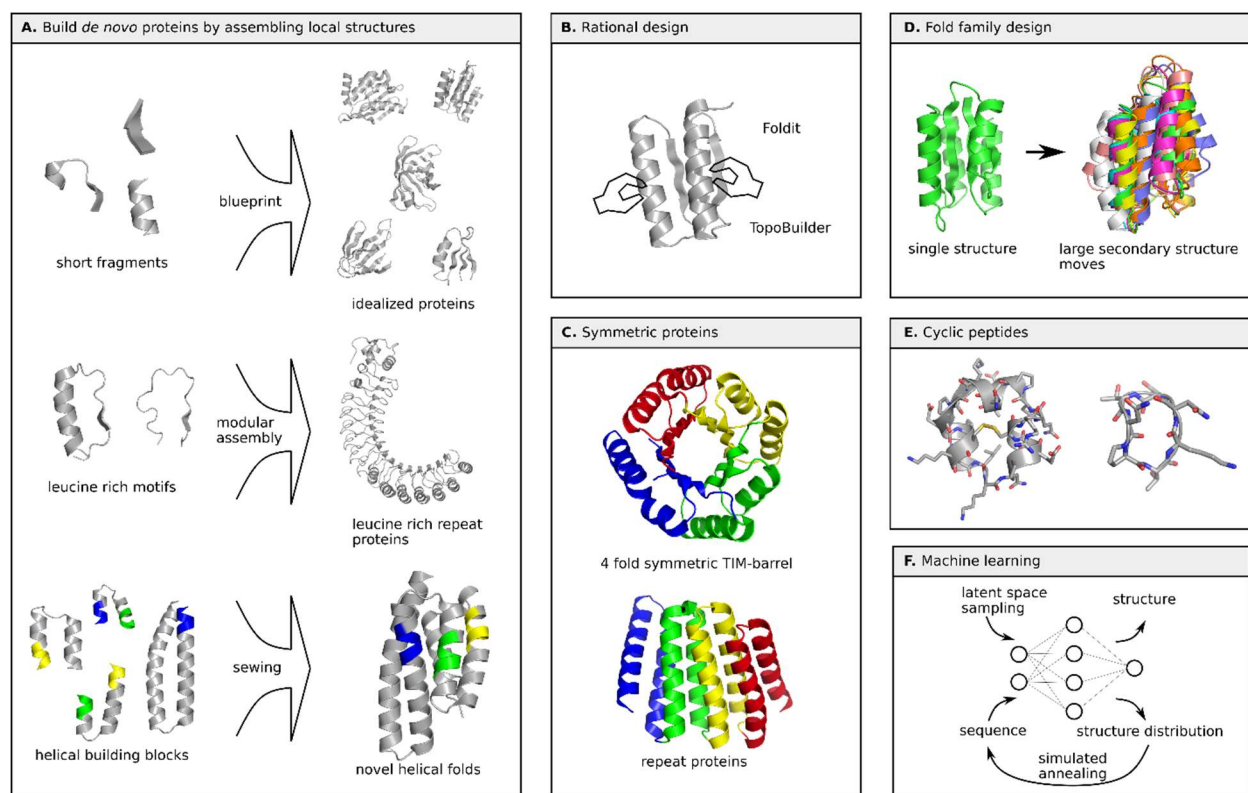


Figure 1.2 Advances in backbone generation

A. Methods build *de novo* proteins by assembling local structures. The blueprint method assembles fragments of 3 residues or 9 residues into idealized (Koga, et al., 2012) (Lin, et al., 2015) (Kuhlman, et al., 2003) (Marcos, et al., 2017) (Marcos, et al., 2018) (Dou, et al., 2018) (Rocklin, et al., 2017) structures with different fold topologies. Modular leucine rich motifs are connected into repeat proteins with defined curvatures (Park, et al., 2015). The SEWING method (Jacobs, et al., 2016) connects local structure pieces into helical proteins with novel folds. **B.** The Foldit game (Koepnick, et al., 2019) and TopoBuilder (Yang, et al., 2020) let players or experts rationally design the atomic details of backbone structures. **C.** Symmetry reduces the complexity of backbone design. Symmetry was used to design 4-fold symmetric TIM barrels and repeat proteins. **D.** *De novo* protein fold families can be generated by making large secondary structure moves (Pan, et al., 2020) (Basanta, et al., 2020). **E.** The GenKIC method (Bhardwaj, et al., 2016) designs cyclic peptides as well as peptides constrained by disulfide bonds. **F.** Generative machine learning methods build novel backbone structures by latent space sampling (Anand, et al., 2019) (Eguchi, et al., 2020). The hallucination method (Anishchenko, et al., 2020) uses the TR-Rosetta neural network to predict the structure distribution of a sequence. The sequence is optimized in a Monte Carlo simulated annealing simulation by maximizing the divergence between the predicted structure distribution and a background distribution that represents unfolded proteins.

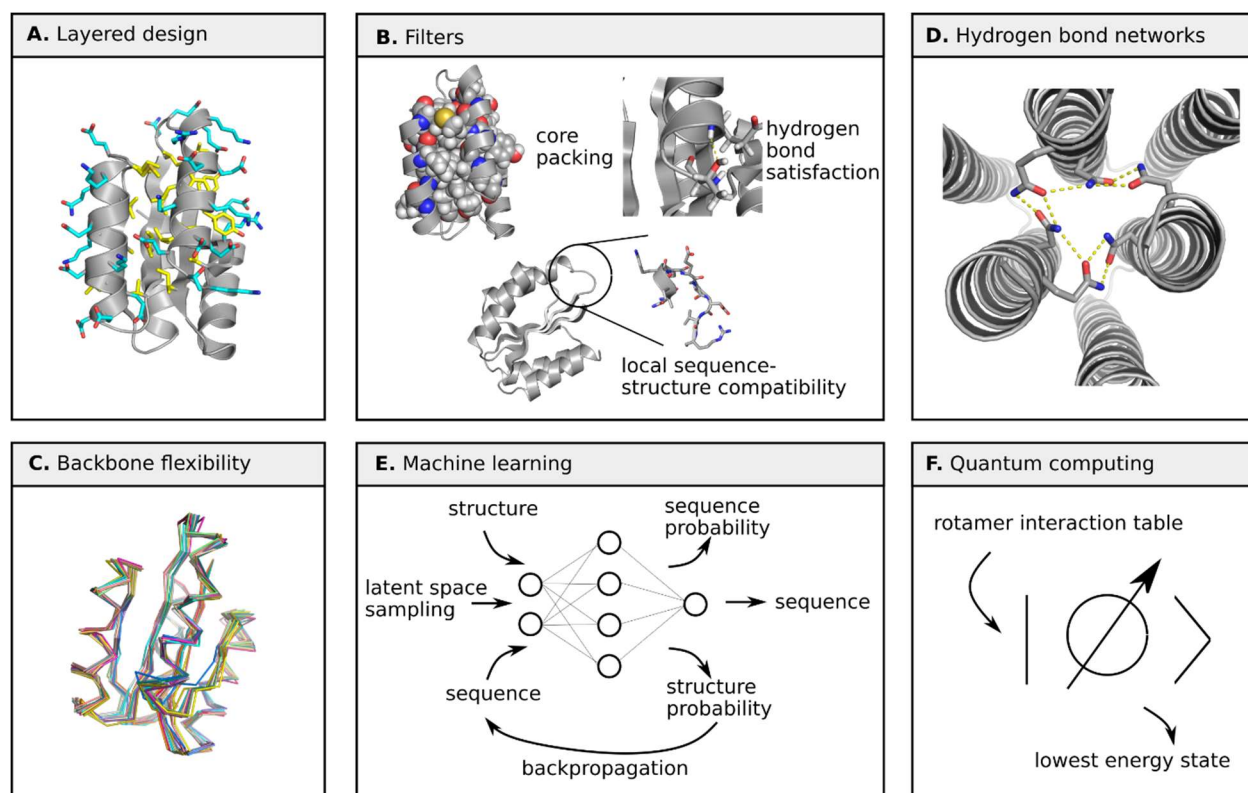


Figure 1.3 Advances in sidechain sampling

A. In layered design, polar residues (cyan) are only allowed at surface and boundary positions, while hydrophobic residues (yellow) are only allowed at boundary and core positions. **B.** Structures generated by side chain design methods can be evaluated by a set of filters, such as core packing quality, hydrogen bond satisfaction and local sequence structure compatibility. **C.** Side chain design methods that exploit backbone flexibility outperforms fixed backbone methods (Loshbaugh & Kortemme, 2020). **D.** The HBNNet method (Boyken, et al., 2016) designs hydrogen bond networks. **E.** Neural networks can predict the probabilities of sequences given a backbone structure (Wang, et al., 2018) (Anand, et al., 2020). Generative machine learning models design sequences by latent space sampling (Karimi, et al., 2019) (Repecka, et al., 2019) (Davidsen, et al., 2019) (Hawkins-Hooker, et al., 2020) (Ingraham, et al., 2019) (Strokach, et al., 2019). The TR-Rosetta neural network predicts the probability of the structure of a given sequence. The difference between the desired structure and the predicted structure can be backpropagated through the neural network to optimize the sequence (Norn, et al., 2020). **F.** The discrete optimization problem of side chain design can be solved by quantum computers. Given the table of rotamer interaction energies, a quantum algorithm (Mulligan, et al., 2019) can solve the lowest energy state in a constant time.

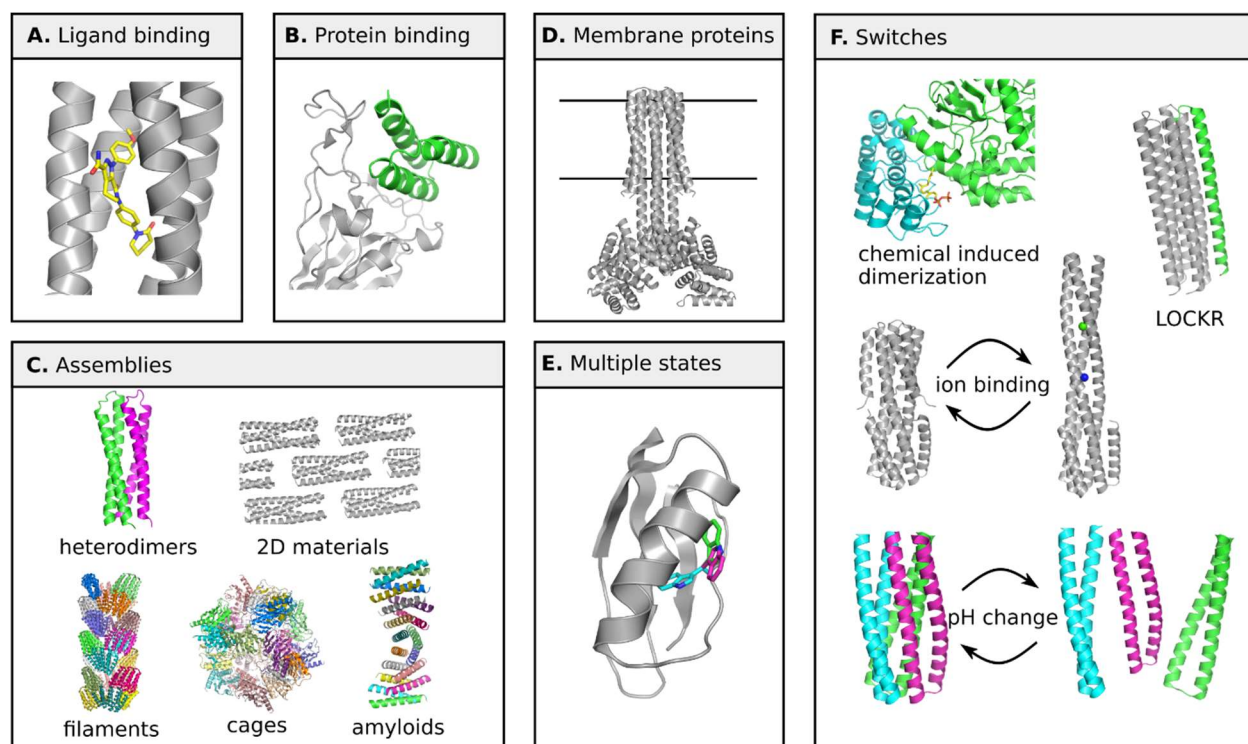


Figure 1.4 Advances in functional design

A. An apixaban (yellow) binder designed by the Convergent Motifs for Binding Sites (COMBS) algorithm (Polizzi & DeGrado, 2020). **B.** A *de novo* protein (green) binds the human angiotensin-converting enzyme 2 (ACE2) (grey) at pico-molar affinity (Cao, et al., 2020). **C.** *De novo* proteins self-assemble into heterodimers (Chen, et al., 2019), 2D materials (Chen, et al., 2019), filaments (Shen, et al., 2018), cages (King, et al., 2014) and alpha amyloids (Zhang, et al., 2018). **D.** A *de novo* designed multipass transmembrane protein that has a defined orientation (Lu, et al., 2018). **E.** The designed DANCER protein has a tryptophan side chain that switches between predicted conformational states on the millisecond timescale (Davey & Chica, 2017). **F.** Designed switch proteins change conformations in response to various signals. A designed chemical induced dimerization system (Glasgow, et al., 2019) dimerizes upon binding to a farnesyl pyrophosphate molecule. In the LOCKR system, a helical peptide called key can displace and expose the signal peptide (green) (Langan, et al., 2019). A designed helical bundle protein changes conformation upon binding to a calcium ion (green) and a chloride ion (blue) (Wei, et al., 2020). A designed helical trimer changes its oligomerization state in response to pH changes (Boyken, et al., 2019).

Chapter 2 Computational design of *de novo* protein fold families

This chapter is adapted from the published article:

Pan, X., Thompson, M. C., Zhang, Y., Liu, L., Fraser, J. S., Kelly, M. J. S., & Kortemme, T. (2020). Expanding the space of protein geometries by computational design of *de novo* fold families. *Science*, 369(6507), 1132–1136.

Author contributions:

X.P. conceived the idea for the project; X.P. and T.K. conceived the computational and experimental approach; X.P. developed and performed the computational design; X.P. and Y.Z. performed the majority of the experimental characterization; X.P. and M.J.S.K. determined the NMR structures; X.P., M.C.T., L.L., and J.S.F. determined the crystal structure; J.S.F., M.J.S.K., and T.K. provided guidance, mentorship, and resources; and X.P. and T.K. wrote the manuscript with contributions from the other authors.

Introduction

Design of proteins with new and useful architectures and functions requires precise control over molecular geometries (Baker, 2010) (Kundert & Kortemme, 2019). In nature, proteins adopt a limited set of protein fold topologies (Dawson, et al., 2017) (Fox, et al., 2014) (Hou, et al., 2005) that are reused and adapted for different functions. Here we define “topology” as the identity and connectivity of secondary structure elements (**Figure 2.1A**). Within a given topology, geometric features including length and orientations of secondary structure elements are often highly variable

(Dawson, et al., 2017) (Fox, et al., 2014). These considerable geometric differences between proteins with the same topology are necessary as they define the exquisite shape and physicochemical complementarity characteristic of protein functional sites. Creating proteins with new functions *de novo* therefore requires the ability to design proteins not only with different topologies, but also distinct custom-shaped geometries within these topologies optimal for each function (**Figure 2.1A**).

Computational design has been successful in mimicking the ability of evolution to generate diverse protein structures spanning alpha-helical (Huang, et al., 2014) (Jacobs, et al., 2016) (Thomson, et al., 2014) (Hill, et al., 2000) (Harbury, et al., 1998) alpha-beta (Rocklin, et al., 2017) (Koga, et al., 2012) (Huang, et al., 2016) and beta-sheet (Dou, et al., 2018) (Marcos, et al., 2018) fold topologies, including novel folds (Kuhlman, et al., 2003). However, most design methods do not include explicit mechanisms to vary geometric features within a topology. For instance, successful design methods assemble protein structures from peptide fragments using a definition of the desired fold and topological rules derived from naturally occurring structures (Koga, et al., 2012). Subsequent iterative cycles of fixed-backbone sequence optimization and fixed-sequence structure minimization (Kuhlman, et al., 2003) refine atomic packing interactions, but do not create substantial changes in geometry. An exception are methods that use parametric equations to sample backbone variation (Crick, 1953) or take advantage of modular protein elements, but these methods are restricted to helical bundles (Huang, et al., 2014) (Thomson, et al., 2014) (Harbury, et al., 1998) or repeat protein (Brunette, et al., 2015) architectures, respectively.

Here we sought to develop a generalizable computational design approach that mimics the ability of evolution to create geometric variation within a given fold topology (**Figure 2.1**). When analyzing geometric variation in protein fold families, we found that 84% of naturally occurring fold families contain variations in loop-helix-loop (LHL) elements (**Figure 2.2**). We hence reasoned that a method that systematically samples geometric variation in these units would not only be able to recapitulate a large fraction of geometric diversity in naturally occurring structures but also to create fold families of *de novo* designed proteins with tunable geometries (**Figure 2.1B**).

Results and Discussion

To develop a generalizable method that systematically samples geometries of LHL, we first examined the individual connecting loop elements in native LHL units. For all LHL elements from all CATH superfamilies (Dawson, et al., 2017) of non-redundant structures, 72.8% of the loops contained 5 residues or less (**Figure 2.3A**). We extracted 313,072 loops of length 2 to 5 connecting to helices from the Rosetta non-redundant fragment database (Gront, et al., 2011) and sorted loops into 12 libraries based on loop length and type of adjacent secondary structure (**Table 2.1**). For each library, only non-redundant loops were kept (**Methods**); this procedure yielded between 224 and 5,826 loops per library. The loop libraries had degeneracies (total number of loops divided by the number of non-redundant loops in each library) ranging from 4.4 to 202 (**Figure 2.3B**), indicating that evolution frequently used similar loop structures in different proteins. This suggests that the identified loop element libraries could also be used to computationally sample novel protein structures that have not been explored by nature.

We developed a protocol called loop-helix-loop unit combinatorial sampling (LUCS, **Figure 2.1C**, **Figure 2.4**). LUCS starts with an input protein fold, which can be naturally occurring or as in our case *de novo* designed (**Methods**), and a definition of gaps in which to insert LHL units. The first step systematically samples all individual loop elements from our libraries (**Table 2.1**). For each gap, loops are inserted at each end of the gap and any loops that clash with the input structure are removed. In a second step all pairs of remaining loops are tested for supporting LHL units by growing helices from each loop. If helices grown from the two ends meet in the middle, excess residues are removed in the third step and the gap closed by energy minimization with a chain-break penalty and hydrogen bond restraints. Closed LHL units with distorted hydrogen bonds geometries, steric clashes or suboptimal interactions between designed backbones and the environment are discarded (**Methods**). In a fourth step, combinations of LHL units at different positions can be screened to yield final structures that have multiple compatible LHL units with systematically sampled lengths and orientations.

To validate the ability of LUCS to generate distinct geometries within given fold topologies, we applied the method to three design problems (**Figure 2.1D**). In the first two design problems, we varied one (RO1) or two (RO2) LHL units of a *de novo* designed protein (Koga, et al., 2012) (PDB:2LV8) with a Rossmann fold topology. In the third problem, we varied two LHL units of a *de novo* designed protein (Marcos, et al., 2017) (PDB:5TPJ) with a nuclear transport factor 2 (NTF2) fold topology (NT). In principle, LUCS can sample topologies with an arbitrary number of LHL units. For the systems we tested, systematic geometry sampling generated approximately 10^4 LHL elements for each gap. To limit the required computing power, we screened 10^6 random combinations of LHL units and generated between 10^4 - 10^5 final backbone structures for each

design problem (**Table 2.2**). We then applied the Rosetta FastDesign protocol (**Methods**) to optimize sequences for all residue positions within 10 Å from the new LHL elements. The number of designed residues for each backbone was between 33 and 87. We note that Rosetta FastDesign also introduces structural changes outside the reshaped LHL elements of the designed fold through gradient-based torsion minimization, although these changes are small (backbone heavy atom root-mean-square deviation (RMSD) < 1 Å). Following sequence design, we filtered the design models computationally using a set of quality criteria that included a minimal number of buried unsatisfied hydrogen bond donors/acceptors, tight atomic packing interactions in the protein core, and compatibility between sequences and local structures (**Methods**).

For each of the three design problems, we selected 50 low Rosetta energy (Park, et al., 2016) designs from models that passed the quality filters and had diverse conformations for further computational characterization. The Rosetta FastDesign simulations optimized low-energy sequences given a desired structure. To determine the converse, whether the desired structure is also a low energy conformation given the sequence, we conducted *ab initio* protein structure prediction simulations in Rosetta (Bradley, et al., 2005). For the Rossman fold designs, we required the lowest energy predicted structure to be within 1 Å C α RMSD of the design model. For the NTF2 fold designs, we used a less strict criterion requiring a number of low-energy models to be close to the design model, to account for the more difficult problem of sampling native-like structures for proteins larger than 100 amino acids. 10, 25 and 10 designs that passed these tests were chosen for experimental characterization for each of the three design problems, respectively (**Figure 2.1D**, **Table 2.3**, **Table 2.4**). The designed proteins were recombinantly expressed in *E. coli* and purified using His-tag affinity and size exclusion chromatography. 5/10, 8/25 and 4/10

designs were found to be monomeric and well folded for each of the three design problems, respectively, as determined by far-UV circular dichroism and one dimensional ^1H and 2-dimensional ^{15}N HSQC nuclear magnetic resonance (NMR) spectroscopy (**Figure 2.1D**, **Figure 2.5**, **Table 2.5**).

To assess whether the designed structures adopted their intended geometries, we solved structures for three designs (RO2-1, RO2-20, and RO2-25) that sampled two LHL units in the Rossmann fold topology using NMR spectroscopy, and one structure for the NTF2 fold topology designs (NT-9) by X-ray crystallography (**Methods**, **Figure 2.6**, **Table 2.6** **Table 2.7**). The experimentally solved RO2 design structures closely matched the computational models (**Figure 2.7 A-C**), with backbone heavy atom RMSDs between models and solved structures within 1.3 Å, core hydrophobic side chains in good agreements with the designed models (**Figure 2.8**) and 5 of the loops in designed LHL units well converged (**Figure 2.9**). In the crystallographic electron density map obtained at 1.5 Å resolution for the NTF2 fold design (NT-9), strong signal was clearly identifiable inside a surface pocket (**Figure 2.7D**), which was interpreted as a bound phospholipid (1,2-diacyl-sn-glycero-3-phosphoethanolamine, see **Methods**). The two N- and C-terminal helices (residues 1-20 and 113-128), which had not been reshaped by LUCS, were pushed apart to accommodate the ligand, leading to an overall backbone heavy atom RMSD between design and model of 2.7 Å. However, when excluding the N- and C-termini helices and aligning the remainder of the design, the backbone heavy atom RMSD between the model and the solved structure was 1.4 Å (**Figure 2.7E**). Moreover, the designed side chain packing interactions between the reshaped helices were in excellent agreement with the design (**Figure 2.7F**). Taken together, our structural analysis confirmed the designed geometry in the reshaped regions for all 4 designs. The presence

of a ligand in the NT-9 design is consistent with the known ability of the NTF2 fold to bind to diverse hydrophobic small molecules, and highlights the exciting possibility to introduce new functions such as ligand binding by reshaping protein geometries.

We next analyzed the magnitude of the geometric differences between our designs. We first compared the backbone heavy atom RMSDs between the reshaped helices of all well folded designs (**Figure 2.1D**) after aligning the non-reshaped regions using both the design models and experimentally solved structures (**Figure 2.10A, Figure 2.11**). For the designs with one LHL unit reshaped, 18 out of 20 off-diagonal differences are more than 3 Å (**Figure 2.10A, left**). For the designs with two LHL units reshaped, 55 out of 68 off-diagonal differences are more than 4 Å (**Figure 2.10A, middle and right**). This scale of variation exceeds the backbone changes generated by existing flexible backbone design methods (Davey & Chica, 2017) (Ollikainen, et al., 2013) that are typically smaller than 2 Å RMSD. For each well-folded design, we also identified the closest structures in the protein data bank (PDB) using TM-align (Zhang & Skolnick, 2005). 15 out of the 17 designed LHL units were significantly different (RMSD > 3 Å for one LHL reshaped designs and RMSD > 4 Å for two LHL reshaped designs) from their closest match in the PDB (**Figure 2.10A, Figure 2.12**), indicating that the design protocol not only generates stable structures with considerable conformational divergence, but also geometries not observed in known structures.

We further analyzed the distribution of sampled geometries and their coverage of designable backbone structure space, where a structure is defined as designable if at least one sequence folds into that structure. As a computational approximation, we defined the models that passed the

quality filters after the first iteration of sequence design (**Methods**) as designable because they had good core packing, hydrogen bond satisfaction and local sequence structure compatibility with the designed sequence. We projected the center and directions of the helices onto the underlying beta sheets (**Figure 2.10B**). The sampled helices from designable models at each position encompassed the distributions derived from native protein structures in the PDB (**Figure 2.10B, right panels**). For the NTF2 fold, the distributions sampled in the designs were slightly shifted to the upper left when compared to the distributions in known structures (**Figure 2.11**). This difference could be a result of the presence of a C-terminal helix in our designs occupying the region shown in the right of the space projection, whereas C terminal helices were often missing in the ensemble of known structures. Overall, since the number of known protein structures for a given topology is limited, the structure space covered by the known structures is much sparser than the space covered by the sampled structures. We quantified the size of structure space by dividing the 6-dimensional space of helix centers and orientations into bins (**Methods, Figure 2.13**). For the geometries sampled in this work, the known structures covered between 12 and 26 bins, while LUCS generated structures covered between 63 and 221 bins (**Figure 2.10C**); the smaller number of geometries in the NT designs (relative to the RO designs) could be a consequence of the additional C-terminal helix present on our NT designs restricting the accessible space of the two sampled helices). The 17 well folded designs (**Figure 2.1D**) sampled between 3 and 7 bins for each helix, respectively, and the majority (18/22) of these bins were not covered by known structures (**Figure 2.10D**). All but one of the well folded designs had at least one helix in a novel bin. Five well folded designs had both helices in novel bins (**Figure 2.10E**). Taken together, these results show that LUCS generates highly diverse geometries encompassing those found in nature but also exceeding known structure space, indicating that a large part of designable protein structure remains unexplored.

We next sought to understand in more detail how the backbone geometries of the designed proteins were defined by the precise details of their non-covalent intramolecular interactions. The three experimentally solved Rossmann fold topology structures had distinct sequence patterns (**Figure 2.14A**) resulting in distinct packing arrangements (**Figure 2.14B, C**) in their hydrophobic cores. The beta sheets favored beta branched residues as expected, but the side chain sizes varied across different designs and resulted in differential hydrophobic packing. In particular, we observed previously described knob-socket type packing motifs (Joo, et al., 2012) (**Figure 2.14C, Figure 2.15**) where nonpolar side chains fit into pockets formed by three residues on helices. These arrangements result in matched geometries between the side chains from sheets and helices that likely contribute to specifying the three-dimensional arrangement of the helices (**Figure 2.16**). We also applied tertiary motif analysis using MASTER (Zhou & Grigoryan, 2015). For all well-folded designs, we were able to match tertiary motifs to both the designed loops and interacting secondary structure elements (**Figure 2.17**). Moreover, we identified side chains mediating helix-helix, helix-sheet and helix-loop interactions that are similar in our designs and the corresponding matched tertiary motifs (**Figure 2.14D**). Despite the close match between the local structures in the design and the tertiary motifs, the source proteins of the motifs had overall structures very different from the designs (**Figure 2.17**). Since tertiary motif information was not used directly in LHL backbone sampling or sidechain design, we conclude that recurrent tertiary motifs can be recapitulated solely by our LUCS sampling protocol and the Rosetta energy function (Park, et al., 2016).

Previous key achievements in *de novo* design (Rocklin, et al., 2017) (Koga, et al., 2012) (Huang, et al., 2016) (Dou, et al., 2018) (Marcos, et al., 2017) (Marcos, et al., 2018) focused on designing

one or a few structures for diverse non-helical-bundle topologies by deriving design rules for specific topologies to identify the most favorable “idealized” geometries. This topology-centric strategy typically finds deep energy minima and thereby succeeds in overcoming errors in energy functions to create highly stable *de novo* folds. In contrast, natural and LUCS generated structure families adopt non-ideal geometric features such as diverse helix positions, orientations, lengths and conformations of connector elements, and exploring these non-ideal regions presents extra challenges (Baker, 2019). Nevertheless, we show here that LUCS achieves accurate atom-level control over diverse geometries, and our designs are not notably less stable than their *de novo* designed starting points (**Figure 2.5**). This success could at least partially be explained by the ability of LUCS to recover three-dimensional packing arrangements that are recurrent in nature (**Figure 2.14D**, **Figure 2.17**), but without using this information as input.

We envision many applications for LUCS to precisely tune protein geometries for new protein functions that require atom-level control. The generalizable strategy underlying LUCS (**Figure 2.1C**) does not require prior definition of structural variation based on design rules identified in native structures (Marcos, et al., 2017) (Lin, et al., 2015). New protocols could exploit this ability to flexibly tune protein geometries during design simulations while simultaneously building new functional sites for ligand binding or protein-protein recognition. The systematic sampling of protein geometries should also enable designing dynamic proteins (Davey, et al., 2017) that can switch between multiple distinct *de novo* designed conformations. Methods such as LUCS bring control over designable protein geometry space for arbitrary functions within reach.

Materials and Methods

Analysis of native loop-helix-loop (LHL) units in naturally occurring structures

Protein domain structures of all 2737 CATH superfamilies were downloaded from the CATH database (V4.1.0) (Dawson, et al., 2017). Secondary structures for each protein were assigned using the DSSP algorithm (Kabsch & Sander, 1983) integrated in Rosetta (Leaver-Fay, et al., 2011) and used to identify LHL units in each structure. The number of LHL units in each superfamily was defined as the median number of LHL units in all structures from that superfamily. More than 83% of superfamilies had at least one LHL unit (**Figure 2.2A**). The LHL units in each superfamily were clustered. Two LHL units were clustered together if they connected the same types of secondary structure elements and the distances between the C α atoms of their starting and ending residues were within 3 Å. The helix RMSDs (see section: RMSD calculation between helices with different lengths) between all pairs of LHL units in the same clusters are shown in **Figure 2.2B**.

Generation of loop libraries (Table 2.1)

Each loop library contains loops with a defined length and connecting two secondary structure elements with defined types. Loop libraries were created by scanning the non-redundant protein structure database VALL (Gront, et al., 2011). All loops in VALL with a given length and connecting defined secondary structure types were selected. Loops were discarded if the helices they connected to had less than 6 residues or the strands they connected to had less than 3 residues. The number of loops in each library was recorded for subsequent redundancy calculation. If a loop had 2 Å or lower backbone heavy atom RMSD with other loops and the RMSD of the C terminal

residues was lower than 1.5 Å, the loop was considered redundant and removed. The redundancy of each loop library was defined as the number of all loops divided by the number of non-redundant loops.

LUCS loop-helix-loop sampling

LUCS sampling is generally applicable to both *de novo* scaffolds and naturally occurring proteins, and the sampled geometries in our designs have similar distributions as those as in natural protein families (**Figure 2.10B**). However, we expect the success rate to be higher if the starting scaffold is more stable, and since *de novo* proteins are often highly stable, they are good starting points.

For each LHL unit to be sampled on a given scaffold structure, two insertion points were chosen as inputs. Residues between insertion points in the input structure were removed. Compatible loops with 2, 3, 4 or 5 residues were selected for each insertion point and inserted. All protein residues were mutated to alanine. Clashes between loop residues and the scaffold were detected. Two heavy atoms were defined to be clashing if the distance between atoms was smaller than the sum of their van der Waals radii times a scale factor of 0.6. Loops that did not clash with the scaffold were kept. All pairs of non-clashing loops were screened to test if a helix can be built to bridge the gap. For each pair of loops, 10-residue helices were grown from each of the ends of the loops using the ideal alpha helix dihedrals ($\phi=-57^\circ$, $\psi=-47^\circ$). The directions of half helices were calculated as the normalized average of all vectors pointing from atom N to atom C of each helix residue. If the dot product of directions of half helices was within 0.5, harmonic angle restraints were applied to the angle formed by the C α atoms of the helix start residue, front helix break residue and the helix end residue as well as the angle formed by the C α atoms of the helix start residue, back helix break

residue and the helix end residue. Rosetta energy (only omega, rama_prepro and restraint terms were enabled) minimization was applied to minimize the restrained angles to align the two halves of the helix. During the minimization, phi, psi and omega torsions of the LHL unit residues were movable degrees of freedom. After aligning the helices, if there was a pair of residues on the half helices that had a backbone heavy atom RMSD within 3 Å, the excess residues were trimmed, and Rosetta energy minimization was applied to the LHL unit residue phi, psi and omega torsions to close the gap. During minimization, distances between the atom O and atom H in helix backbone hydrogen bonds as well as angles formed by atom N, atom H and atom O were restrained to maintain the helical structure. Quality filters were applied to the closed helix:

- Rosetta backbone hydrogen bond scores for all reshaped helix residues were lower than -0.7.
- There were no clashes (same definition as above) involving the reshaped helices after mutating all residues to valine (residues were mutated from ALA to VAL to check there was space between backbones for subsequent side chain design).
- The median contact degree (the number of C α atoms within 10 Å from the C α atom of a residue) of helix residues was greater than 1.
- The number of buried unsatisfied hydrogen bonds within the reshaped LHL units was less than 4.

If more than one pair of insertion points was sampled, two LHL units at different insertion points were deemed compatible if there was no clash (same definition as before) between them. A final ensemble of models was produced by applying a group of compatible LHL units at each insertion point. This step resulted in 13,421 models for RO1 designs, 290,836 models for RO2 designs, and

136,683 models for NT designs (**Table 2.2**). The protocol was developed using PyRosetta (Chaudhury, et al., 2010) and is available at:

https://github.com/Kortemme-Lab/loop_helix_loop_resaping/releases/tag/1.0.0

The PyRosetta package used in this study was compiled from the Rosetta source code (commit: 3135d32229f5ebd35c8a716af00dcdffbf81805).

Sequence design

The spatial positions of residues were defined as the positions of their C β atoms (C α for glycine). The sidechain directions were defined as the vector pointing from the C α atoms to the C β atoms. A residue on the scaffold was defined as pointing toward the reshaped backbone region if there was a residue in the reshaped region such that the cosine of the angle between the vector from the surrounding residue C α atom to the reshaped residue C α atom and the sidechain direction vector of the surrounding residue was greater than 0.5. Residues within 10 Å (C α -C α distance) from the backbone of the reshaped region and pointing toward the reshaped region were designed (i.e. allowed to change amino acid residue type and rotamer conformation). Residues within 8 Å from any designable residue and pointing toward the designable region were repackable (i.e. allowed to change rotamer conformation but keeping amino acid residue type). Residue types allowed for designable residues were determined by the extent of residue burial using the Rosetta LayeredDesign task operation (Koga, et al., 2012). Cysteine and histidine were disallowed as designable residue types to avoid issues with disulfide bond formation and pH dependency. The Rosetta FastDesign [https://www.rosettacommons.org/docs/latest/scripting_documentation/RosettaScripts/Movers/movers_pages/FastDesignMover] and RotamerTrial [https://www.rosettacommons.org/docs/latest/scripting_documentation/RosettaScripts/Movers/m

overs_pages/RotamerTrialsMover] protocols were applied to design sequences that stabilize the backbone structures. The first iteration of design was done using the Dunbrack rotamer library without extra rotamers (Shapovalov & Dunbrack, 2011). 58626, 432735 and 409101 sequences were designed for RO1, RO2 and NT. Designed structures that had less than 2 buried unsatisfied hydrogen bond donors or acceptors, had Rosetta hole scores (Sheffler & Baker, 2010) for the designed residues smaller than 0 and had fragment qualities (Marcos, et al., 2017) better than 2 Å were selected for the next iteration of designs. 1163, 9934 and 5715 designs passed the filters for RO1, RO2 and NT. For RO2 and NT, the selected designs were further optimized by 3 repeats of Rosetta FastDesign with extra rotamers enabled by the ex1 and ex2 options in the ExtraRotamersGeneric task operation. 49578 and 22831 sequences were designed for RO2 and NT. Final designs were filtered:

- Fragment qualities for the reshaped regions were better than 1 Å.
- Rosetta hole scores for the designable and repackable residues were lower than 0.
- Rosetta helix complementarity scores (Lawrence and Coleman shape complementarity) (Lawrence & Colman, 1993) for the reshaped helices were better than 0.6.
- There were no buried unsatisfied hydrogen bonds according to the custom buried unsatisfied hydrogen bond filter (next section).
- There were no oversaturated hydrogen bonds, i.e. hydrogen bond acceptors receiving hydrogen bonds from more than the allowed number of donors, for designable and repackable residues.
- The ratio of hydrophobic solvent accessible surface area (SASA) over the total SASA was lower than 58%.

722 and 98 designs passed the filters for RO2 and NT. The custom buried unsatisfied hydrogen bond filter sometimes underestimated the number of buried unsatisfied hydrogen bonds and the holes filter allowed large hydrophobic holes which can exist in native proteins as binding sites. Therefore, designs that passed the filters with lowest residue average Rosetta scores were examined manually to check for possible issues such as hydrophobic voids and buried unsatisfied hydrogen bonds missed by the automatic filter. For the RO2 designs, we selected 50 top-scoring designs passing these manual criteria.

For the NT designs, we attempted to introduce cavities as potential binding pockets into the protein core. The 98 NT designs that passed the second iteration had hydrophobic cavities because the Rosetta layered design task operation assigned the residues surrounding the cavities to the core layer. We manually selected 10 low energy designs that passed iteration 2 (picking the lowest energy designs that were substantially different from each other) for further sequence design simulations to introduce pockets with polar residues into the proteins. We used the same sequence design protocol as in iteration 2, but with the restriction that 1-3 pocket residues had to be polar residues. For each of the 10 selected designs, 100 new sequences were designed. We then filtered the new designs with the same automatic filters followed by manual inspection as the previous round. 202 designs passed the automatic filters, of which 50 were selected by their Rosetta energy. For each of the 10 designs selected from iteration 2, at least one design derived from it was selected. The sequence design code is available at:

https://github.com/Kortemme-Lab/local_protein_sequence_design/releases/tag/1.0.0

Custom buried unsatisfied hydrogen bond filter

The standard Rosetta buried unsatisfied hydrogen bond filter (Maguire, et al., 2018) overestimated the number of buried unsatisfied hydrogen bond donors and acceptors in many native structures (**Figure 2.18**) because it identified hydrogen bonds based on one static model. However, due to modeling errors and protein conformational flexibility, some of the identified buried unsatisfied hydrogen donors or acceptors present in the models may nevertheless form proper hydrogen bonds. In order to find only the buried unsatisfied hydrogen bond donors or acceptors that cannot be compensated by allowing some conformational flexibility, we developed a structure quality filter called the backrub ensemble consensus buried unsatisfied hydrogen bond (BECBUBH) filter. For each protein residue, except for the N- and C-terminal residues, the BECBUBH filter generated 5 structures by applying a local backrub (Smith & Kortemme, 2008) move on the residues and each of their two neighboring residues. The filter then determined the buried unsatisfied hydrogen bond donors or acceptors for each structure. Only buried unsatisfied donors or acceptors that were consistently present in all 5 structures were recorded. We used the BECBUBH filter to filter designed models.

***Ab initio* structure prediction**

Rosetta *ab initio* structure prediction simulations were run for the top 50 selected designs for each design problem (**Table 2.2**). Fragments for structure predictions were generated using the `make_fragments.pl` script (Gront, et al., 2011) distributed with Rosetta. The command for fragment generation was

```
fragments.pl -verbose -id design_id -frag_sizes 3,9 -n_frgs 200 -n_candidates 1000  
sequence.fasta
```

The structures of designs were predicted using the AbinitioRelax application in Rosetta. 20,000 models were generated for each designed sequence. The command was

```
AbinitioRelax.linuxgccrelease -abinitio:relax -use_filters true -abinitio::increase_cycles 10 -  
abinitio::rg_reweight 0.5 -abinitio::rsd_wt_helix 0.5 -abinitio::rsd_wt_loop 0.5 -relax::fast -  
in:file:fasta sequence.fasta -in:file:frag3 fragments_3mer_file -in:file:frag9 fragments_9mer_file  
-psipred_ss2 ss2_file_from_frag_generation -nstruct num_output -out:sf score_file_output -  
out:file:silent silent_file_output
```

For NT designs, standard Rosetta *ab initio* structure prediction simulations failed to sample models sufficiently close to the target model. The target model typically had a significantly lower Rosetta energy than any of the prediction models, indicating a sampling issue. To overcome this problem, we biased *ab initio* structure prediction simulations to favor low RMSD fragments by setting the weight of the score term FragmentCrmsd to be 5 and priority to be 800 in the fragment generation step. We then run standard *ab initio* structure prediction using the biased fragment set. We confirmed that designs folded into the desired target conformation within 1.5 Å when biasing the input fragments used during the structure prediction calculations.

Projecting helices to underlying beta sheets

Models with the Rossmann fold or NTF2 fold were aligned to PDB:2LV8 or PDB:5TPJ, respectively. The 3-dimensional (3D) helix centers were calculated by averaging the C α atom coordinates for all residues in a given helix. The 3D helix directions were defined as the average of C=O bond directions of the helix residues. The 3D centers and directions were projected onto a 2-dimensional (2D) plane by Cartesian projections. The beta sheet peptide bonds were shown as parallelograms where the C α atoms were located at the center of horizontal edges.

Binning the helix geometry space

Each helix was assigned a 6-dimensional vector whose first 3 coordinates were the helix center position and the last 3 coordinates were the direction of the helix. As above, the center of a helix was the average of the C α atom coordinates. The direction was the average of C=O bond directions normalized to a unit vector. The Cartesian space of helix centers was divided into 2Å cubes and the directions were divided into 8 octants. A 6-dimensional bin was assigned to a helix based on the cube and octant that the helix belonged to (**Figure 2.13**).

RMSD calculation between helices with different lengths

To calculate RMSD between two helices with different lengths, the longer helix was truncated. Only the middle part of the longer helix that had the same length as the shorter helix was kept. RMSDs were calculated for backbone heavy atoms (N, C α and C), enforcing that all corresponding backbone heavy atoms were included.

TM-align analysis

The designed models were submitted to the COFACTOR server (Zhang, et al., 2017), which used TM-align to find the 10 closest structures from the PDB. Except for the design RO2_5, the PDB structure 2KPO is the closest structure for all Rossmann fold designs and it ranked 2nd for the design RO2_5. The PDB structure 5TPJ is the closest for all NTF2 fold designs. The difference between the designed geometries and known structures was then quantified by calculating the RMSD between the designed helices and the corresponding helices on the closest known structures with the same topologies found by TM-align (**Figure 2.10A**), as described in the “RMSD calculation between helices with different lengths” section above.

Analysis of tertiary structure motifs

Tertiary structure motif analysis was performed using the MASTER (Zhou & Grigoryan, 2015) program. Small pieces of tertiary structures were specified manually from the well folded designs. The tertiary pieces included designed loops and fragments of interacting secondary structure elements involving the backbone reshaped regions (**Figure 2.17**). The helical element sizes were between 1 to 2 turns (5-9 residues) and the strand element sizes were between 3 to 5 residues. The tertiary pieces were extracted from the design models and saved as query pdb files. The query pdb files were converted to MASTER input query (pds) files by

```
createPDS --type query --pdb query.pdb
```

The query structures were searched against the standard MASTER database with

```
master --query query.pds --targetList MASTER/database/list --rmsdCut 0.5 --matchOut  
query.match --seqOut query.seq --bbRMSD --structOut query.struct
```

Protein expression tests

Plasmids (pET-28a(+)) encoding the designed proteins were ordered from Twist Bioscience. The DNA sequences of the designed proteins were inserted between the NdeI and XhoI restriction sites, which added the DNA coding sequence for an N-terminal MGSSHHHHHHSSGLVPRGSHM tag to the designed proteins. *Escherichia coli* BL21(DE3) cells were transformed with the plasmid. Proteins were expressed by overnight cell culture in 5mL autoinduction medium (ZY medium, 10 g/L tryptone, 5 g/L yeast extract) supplemented with the following stock mixtures: 20xNPS (1M Na₂HPO₄, 1 M KH₂PO₄, and 0.5 M (NH₄)₂SO₄), 50x 5052 (25% glycerol, 2.5% glucose, and 10% α-lactose monohydrate), 1000x trace metal mixture (50 mM FeCl₃, 20 mM CaCl₂, 10 mM each of MnCl₂ and ZnSO₄, and 2 mM each of CoCl₂, CuCl₂, NiCl₂, Na₂MoO₄, Na₂SeO₃, and H₃BO₃ in 60 mM HCl)) (Studier, 2005) with 50 µg/ml kanamycin at 37°C. Cell cultures were aliquoted into 1mL aliquots. Cells were spun down by centrifuging at 20,000 g for 3 min. Cell pellets were resuspended in 1mL Phosphate-buffered saline (PBS) buffer (8g/L NaCl, 0.2g/L KCl, 1.44g/L Na₂HPO₄, 0.24g/L KH₂PO₄, pH=7.4) and lysed by sonication. Soluble and insoluble parts were separated by centrifuging at 21,000 g for 5 minutes. The solubility of a designed protein was assessed by Coomassie-stained SDS-PAGE (BIO-RAD Cat. #456-1095).

Protein expression and purification

Escherichia coli BL21(DE3) cells were transformed with the plasmids encoding the N-terminally His-tagged designed proteins. Colonies were inoculated into 5mL LB medium and cultured at

37°C for 12 hours. Seed cultures were inoculated into 1 L autoinduction medium and cultured at 225 RPM shaking speed at 37°C overnight. Cell cultures were centrifuged at 5,000 g for 5 minutes to spin down the cells. Cell pellets were resuspended in 30mL equilibration buffer (50mM Tris pH=7.5, 300mM NaCl, 10 mM imidazole) and lysed using a microfluidizer. Cell lysate was centrifuged at 39,000 g for 30 minutes to separate the soluble and insoluble fractions. The soluble fraction was mixed with 1mL Ni-resin beads (Thermo Scientific #88222) to pull down the His-tagged proteins. Ni-resin beads were washed 3 times with the wash buffer (50mM Tris pH=7.5, 300mM NaCl, 25mM imidazole). Proteins were eluted 3 times with 1mL elution buffer (50mM Tris pH=7.5, 300mM NaCl, 250mM imidazole). For crystallization, His-tags were cleaved by 10 NIH unit bovine thrombin (Sigma-Aldrich T4648-10KU) overnight at room temperature. Then the samples were further purified using a HiLoad® 16/600 Superdex® 75 pg size exclusion column (GE) with PBS buffer. The monomeric fraction was collected for subsequent characterization.

Analytical size exclusion chromatography

Protein samples purified by His-tag pull down (if sufficiently pure) or after additional purification with the HiLoad® 16/600 Superdex® 75 pg size exclusion column were analyzed using a Superdex® 75 10/300 GL size exclusion column from GE with PBS buffer. The relation between elution time and log molecular weight was fitted using a linear regression model with the BioRad Gel Filtration Standard (Catalog #151-1901).

Circular dichroism spectroscopy

Circular dichroism (CD) data were collected on a Jasco J-710 spectrometer. Purified RO1 proteins were diluted into 50mM phosphate buffer (pH 7.0). Purified RO2 and NT proteins were diluted

into PBS buffer (8g/L NaCl, 0.2g/L KCl, 1.44g/L Na₂HPO₄, 0.24g/L KH₂PO₄, pH=7.4). The concentrations of diluted samples ranged from 2μM to 5μM. Protein concentrations were determined using the absorbance at 280 nm using a NanoDrop (Thermo Scientific). CD spectra were measured using a 1mm cuvette at 25°C. Melting curves at 220nm were measured by increasing temperature from 25°C to 95°C using a rate of 1°C/min.

1D-¹H NMR spectra

Purified proteins were exchanged into 50mM phosphate buffer (pH 7.0). Samples were concentrated to 15-200 μM. 56 mL D₂O was added to 500 mL samples such that the final volume had 10% D₂O and 90% H₂O. The 1D ¹H spectra (pulse program: zgpg30) were measured at 297.9 K using a Bruker Avance I 800 MHz spectrometer with a Z-gradient TXI cryo-probe. The temperature was calibrated with 4% MeOD using the following coefficients of $T(K) = (4.109 - D) * 0.008708$ where D is the chemical shift difference between the CH₃ and OH protons in methanol. The spectra were processed with the program NMRPipe (Delaglio, et al., 1995).

Protein expression for NMR structure determination

¹⁵N and ¹³C labeled proteins were expressed by growing *E. coli* in M9 minimal medium that included ¹³C-glucose and ¹⁵NH₄Cl (6g/L Na₂HPO₄, 3g/L KH₂PO₄, 0.5g/L NaCl, 0.5g/L ¹⁵NH₄Cl, 50mg/L EDTA, 8.3mg/L FeCl₃ x 6 H₂O, 0.84mg/L ZnCl₂, 0.13mg/L CuCl₂ x 2 H₂O, 0.1mg/L CoCl₂ x 6 H₂O, 0.1mg/L H₃BO₃, 0.016mg/L MnCl₂ x 6 H₂O, 0.2% (w/v) ¹³C-glucose, 1mM MgSO₄, 0.3mM CaCl₂, 1mg/L Biotin, 1mg/L Thiamine). Single bacterial colonies were first inoculated into 5mL seed culture and grown at 37 °C overnight. The seed culture was inoculated into 1L ¹⁵N ¹³C labeled M9 minimal medium and grown at 37 °C until OD₆₀₀ reached 0.5-0.7.

Then 1mL 1M IPTG was added to induce protein expression at 18 °C overnight. The expressed proteins were purified following the Ni-resin pull down protocol described in the protein purification section. The His-tag purified proteins were further purified using the HiLoad® 16/600 Superdex® 75 pg size exclusion column from GE with 50mM phosphate buffer at pH 7.0. The monomeric fractions were collected and concentrated to 0.5-1 mM for NMR experiments.

Structure determination by NMR

Proteins labeled with ^{15}N and ^{13}C were exchanged into 50mM phosphate buffer (pH 7.0). 10% D_2O was added to samples and the final protein concentrations were 0.74-0.8 mM. NMR spectra were measured at 297.9 K. Two dimensional (2D) ^{15}N -HSQC (pulse program: fhsqcf3gpqh), 2D ^{13}C -HSQC (pulse program: hsqcetgpsisp2), 16 ms 3D HCCH-TOCSY (pulse program: hcchdigp3d) and 120 ms 3D simultaneous $^{13}\text{C}/^{15}\text{N}$ -NOESY-HSQC (pulse program: noesyhsqcgpsismsp3d) spectra were measured using a Bruker Avance I 800 MHz spectrometer with a Z-gradient TXI cryo-probe. 3D CACB(CO)NH (pulse program: hncocacbgpwg3d) and 3D CACBNH (pulse program: hncacbgpwg3d) spectra were measured using a Bruker Avance DRX500 spectrometer with a Z-gradient QCI cryo-probe at 297.6 K (same temperature calibration as described above). NMR spectra were processed using the program NMRPipe (Delaglio, et al., 1995) and indirect referencing to an external DSS standard was used (Wishart, et al., 1995). Resonances were assigned using the program CCPN Analysis (Vranken, et al., 2005). Backbone resonances were assigned using the 2D ^{15}N -HSQC, 3D CACB(CO)NH and 3D CACBNH spectra. Sidechain resonances were assigned using the 2D ^{13}C -HSQC and 3D HCCH-TOCSY spectra. Distance restraints were generated using CCPN Analysis. Dihedral restraints were generated using the program DANGLE (Cheung, et al., 2010). Hydrogen bond restraints included in the structure

calculations were based on secondary structures predicted by DANGE and nOe patterns typical of alpha-helical secondary structure. The programs ARIA version 2.3.2 (Rieping, et al., 2007) and CNS version 1.2 .1 (Brünger, et al., 1998) were used to calculate the NMR structures. To solve the structures, 9 iterations of simulated annealing were performed using CNS. For the first 8 rounds of simulated annealing, the n_structures parameter was set to 100 and the n_best_structures parameter was set to 35. For the 9th round, the n_best_structures parameter was set to 20. Finally, a refinement in water was performed on the 20 structures from the 9th iteration. Otherwise, the default values were used for the remaining ARIA parameters. The ensemble of the refined structures was validated using the PDB validation server (Berman, et al., 2003) and the Protein Structure Validation Suite (PSVS) server (Bhattacharya, et al., 2006). The agreement between structures and NMR data were assessed using the program PyRPF (Skinner, et al., 2015) integrated into the CCPN suite.

Protein crystallization

We concentrated the NT_9 protein to 25.2 mg/mL in buffer containing 20 mM Tris Buffer (pH=7.5) and 150 mM sodium chloride, and carried out initial crystallization trials using the JCSG I-IV commercial crystallization screen (Qiagen). Crystallization drops were prepared in 96-well sitting drop format by mixing 100 nl of protein solution with 100 nl of the mother liquor using a Mosquito liquid handling robot (TTP Labtech). Drops were sealed inside a reservoir containing an additional 100 µl of the mother liquor solution. Crystals were obtained from mother liquor containing 0.1M MES Buffer at pH=6, 30% PEG-600, 5% PEG-1000, and 10% Glycerol. Crystals from the initial screen were used for data collection without further optimization.

X-ray data collection and processing

Prior to X-ray data collection, crystals were flash-cooled by rapid plunging into liquid nitrogen. The high concentrations of polyethylene glycols and glycerol in the crystallization mother liquor allowed the crystals to be harvested and frozen directly without additional cryoprotection. We collected single-crystal X-ray diffraction data on beamline 8.3.1 at the Advanced Light Source. The beamline was equipped with a Pilatus3 S 6M detector (Dectris), the X-ray energy was set to 11111 keV, and the crystals were maintained at a cryogenic temperature (100 K) throughout the course of data collection.

We processed the X-ray data using the Xia2 system (Winter, 2010), which performed indexing, integration, and scaling with XDS and XSCALE (Kabsch, 2010), followed by merging with Pointless (Evans, 2006). A resolution cutoff (1.50 Å) was taken where the completeness of the data fell to a value of approximately 90%. Although other metrics of data quality (such as CC1/2 and $\langle I/\sigma I \rangle$) suggest that a more aggressive resolution cutoff would be acceptable, we were limited by the data completeness that could be obtained with the minimum accessible sample-to-detector distance. Further information regarding data collection and processing is presented in **Table 2.7**. The reduced diffraction data were analyzed with phenix.xtriage (http://www.ccp4.ac.uk/newsletters/newsletter43/articles/PHZ_RWGK_PDA.pdf) to check for common crystal pathologies, none of which were identified.

X-ray structure determination

We obtained initial phase information for calculation of electron density maps by molecular replacement using the program Phaser (McCoy, et al., 2007), as implemented in the PHENIX suite

(Adams, et al., 2012). We identified a single copy of the protein in the asymmetric unit using the coordinates from our design model, consistent with an analysis of Matthews probabilities for the observed unit cell and molecular weight of the protein (Kantardjieff & Rupp, 2003) (Weichenberger & Rupp, 2014).

Next, we attempted to rebuild our model of the protein using the electron-density maps calculated using the model phases derived from molecular replacement. We immediately noticed that a cavity in the protein surface, inherent to the designed topology, was occupied with a large electron density feature present in both 2mFo-DFc and mFo-DFc electron density maps. This feature was interpreted as phosphatidylethanolamine, an *E. coli* phospholipid that we suspect binds to the protein during recombinant expression and remains bound throughout the purification. In addition to modeling the bound phospholipid, we rebuilt parts of the design model using the initial electron-density maps calculated from molecular replacement. We then performed additional, iterative refinement of atomic positions, individual atomic displacement parameters (B-factors) with a TLS model, and occupancies, using riding hydrogen atoms and automatic weight optimization, until the model reached convergence. Our tentative modeling of phosphatidylethanolamine bound to the protein was supported by the overall flatness of mFo-DFc electron density maps in the vicinity of the ligand, and by the reduction in R-free obtained when adding the ligand to the model. All model building was performed using Coot (Afonine, et al., 2012) and refinement steps were performed with phenix.refine (v1.16-3549) within the PHENIX suite (Adams, et al., 2012) (Afonine, et al., 2012). Restraints for the phosphatidylethanolamine (PEE) ligand were calculated using phenix.elbow (Moriarty, et al., 2009). The final model coordinates were deposited in the Protein

Data Bank (PDB (Bernstein, et al., 1978)) under accession code 6W90. Further information regarding model building and refinement is presented in **Table 2.7**.

Note

Analysis of the specific interactions in defining protein 3D geometries

To test the relevance of the specific packing interactions highlighted in **Figure 2.14C** for specifying the targeted geometries, we made *in silico* single and double mutations at these residue positions to either smaller (alanine) or larger (phenylalanine) side chains. We then ran *ab initio* folding simulations on these variants (**Figure 2.16**) and quantified the energy gap between correctly and incorrectly folded models using a structure prediction decoy discrimination score described in (Conway & DiMaio, 2016). We found that all mutations worsened the decoy discrimination score, indicating that the mutated residues play a role in specifying the designed structure. This behavior was particularly drastic for four of the simulated variants (RO2_20_L29A, RO2_20_L29A_I81A, RO2_25_V82A and RO2_25_F3A_V82A), where the distributions of RMSD between predicted models and designed states were shifted significantly (Kolmogorov-Smirnov statistic > 0.2 and p-value $< 10^{-300}$). For three other variants (RO2_1_I39A_V55A, RO2_25_F3A and RO2_25_V82F), the *ab initio* folding simulations sampled conformations that have lower predicted energies than the average energies of the relaxed designed models, but have an RMSD of more than 1 Å from the designed structures. In these cases, the mutated variants could populate alternative conformations different from the originally designed geometries, although we caution that experimental structure determination of these variants would be required for more detailed analysis. Overall, our simulation results highlight the importance of the highlighted specific packing interactions in defining the three-dimensional geometries of the designs. Other factor such as loop preferences are also likely to contribute to specifying the three-dimensional structure.

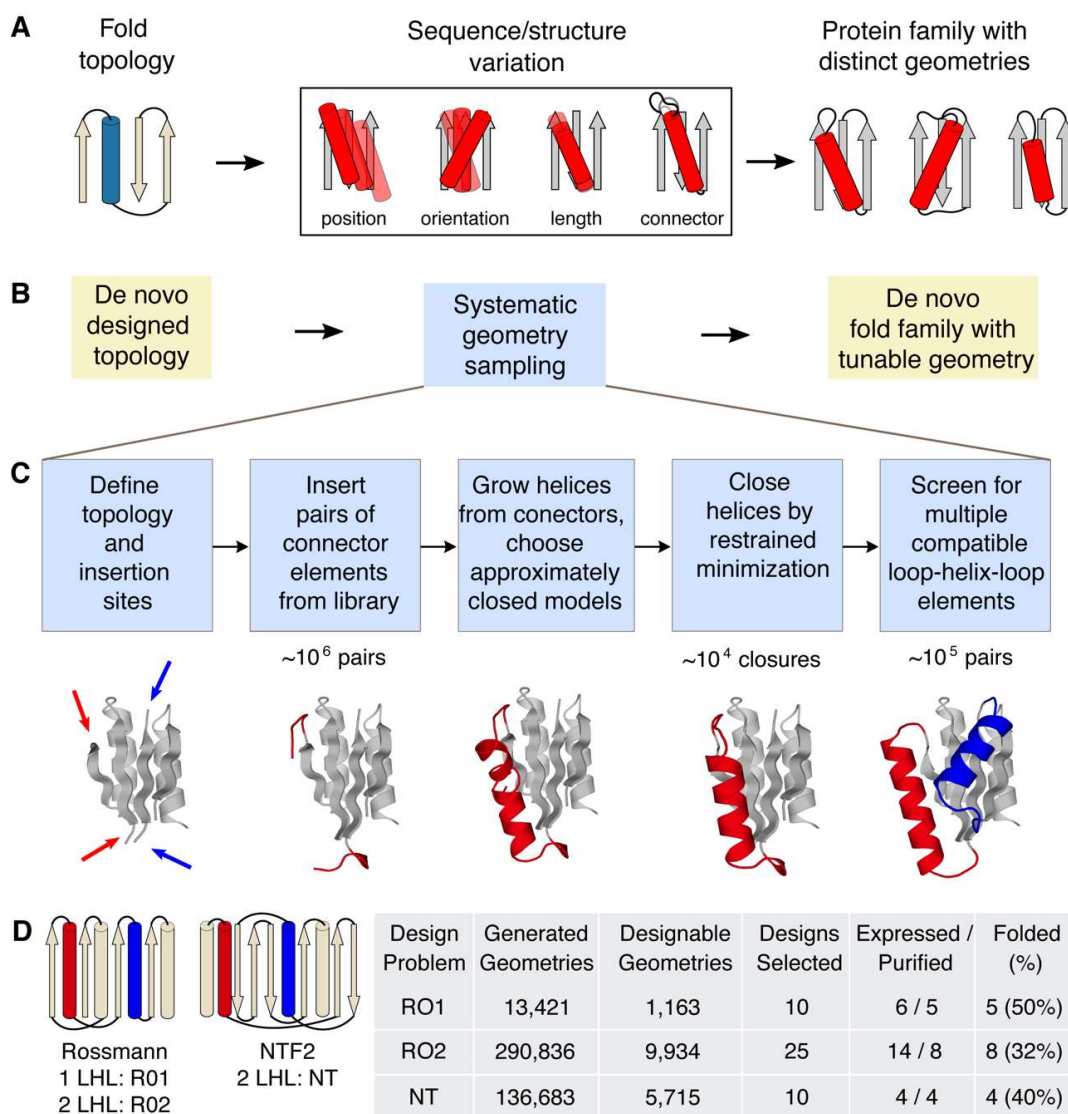


Figure 2.1 LUCS sampling strategy to create *de novo* designed protein fold families with tunable geometries.

A. In nature, protein fold topologies (left) are diversified to create families of proteins with distinct geometries (right) optimized for function. Alpha-helices are shown as cylinders and beta-strands as arrows. The box shows schematic representations of common types of geometric variation. **B.** The LUCS computational design protocol seeks to mimic the ability of evolution to diversity protein geometries to generate *de novo* designed fold families. **C.** Schematic of the LUCS protocol for sampling LHL geometries. The reshaped LHL units are colored in red and blue. Typical numbers of models generated at major stages of the protocol are indicated. **D.** Designed fold families. Schematic shows fold topologies and design problems (Rossman fold with 1 or 2 reshaped LHL units, and NTF2 fold with 2 reshaped LHL units). Also shown are numbers for geometries generated by LUCS, designed models that passed quality filters, and experimentally characterized designs for three design problems. % folded indicates the fraction of experimentally tested designs that adopted folded structures.

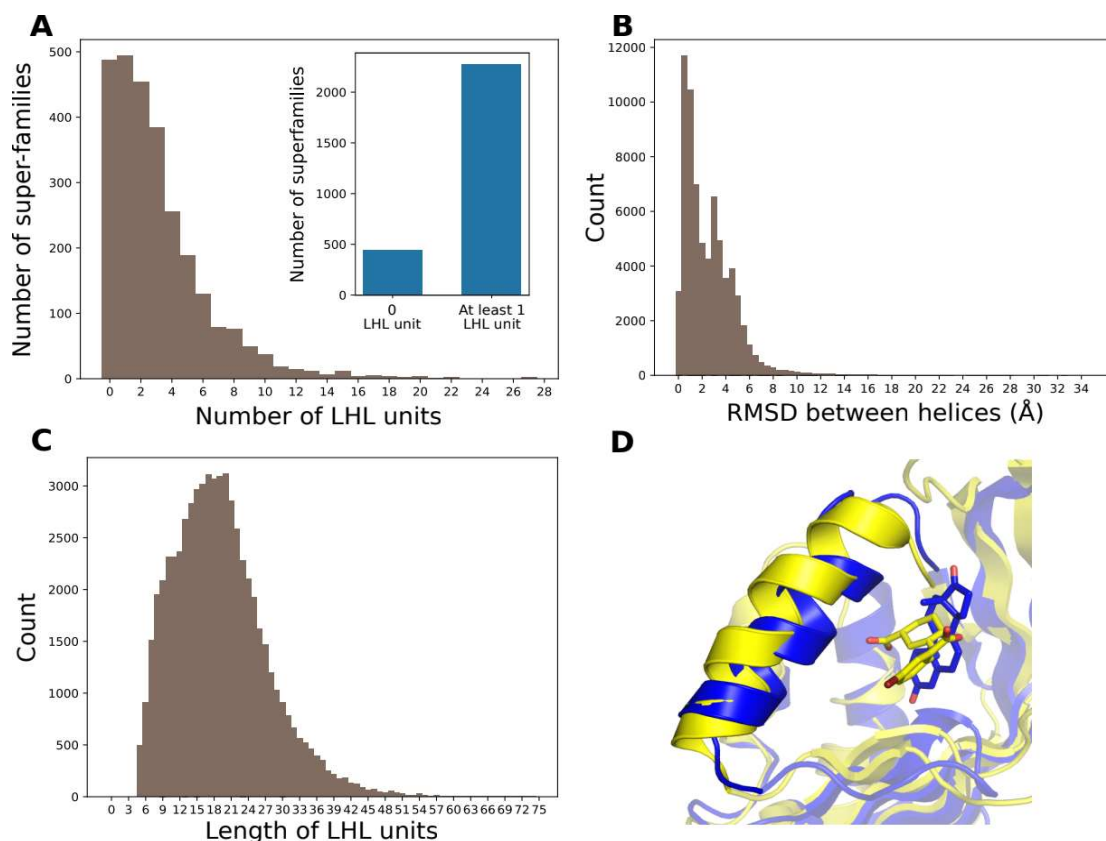


Figure 2.2 Diversity of geometries in naturally occurring fold topologies to enable distinct functions.

A. Distribution of the number of LHL units contained in each CATH protein superfamily. Inset: numbers of superfamilies that have no or at least one LHL units. 83.7% of all CATH superfamilies have at least one LHL unit. **B.** Diversity of the geometries of LHL units in CATH superfamilies. Shown are the helix RMSDs between all pairs of LHL units in the same clusters computed as described in Methods. **C.** Distribution of LHL lengths in all CATH structure superfamilies. **D.** Example where a change of the LHL geometry at the active site alters ligand specificity. The blue LHL element is from ketosteroid isomerase (PDB:1OH0) that binds a equilenin and the yellow LHL element is from Phenazine biosynthesis protein A/B that binds a 5-bromo-2- $\{[(1S,3R)\text{-}3\text{-carboxycyclohexyl}]\text{amino}\}$ benzoic acid (PDB:3JUM).

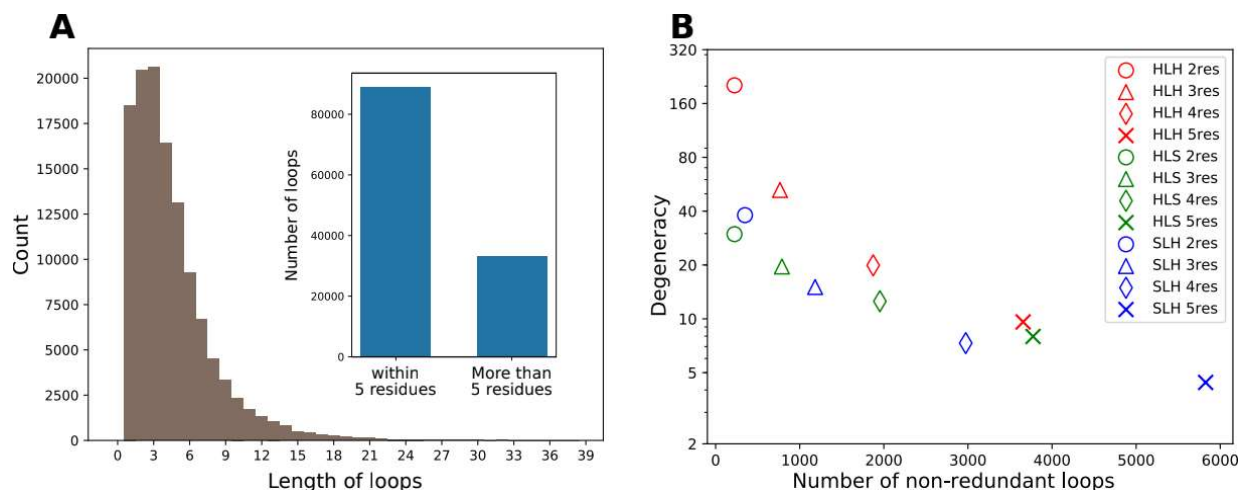


Figure 2.3 Common loop connector elements in naturally occurring LHL units.

A. Distribution of loop lengths in LHL units. Inset: numbers of loops that have at most 5 residues or more than 5 residues. **B.** Sizes of non-redundant loop libraries versus the degeneracy defined as the total number of loops divided by the number of non-redundant loops. Loops connecting two helices are shown in red. Loops connecting a helix and a strand are shown in green. Loops connecting a strand and a helix are shown in blue. Loops with different numbers of residues are indicated by different markers.

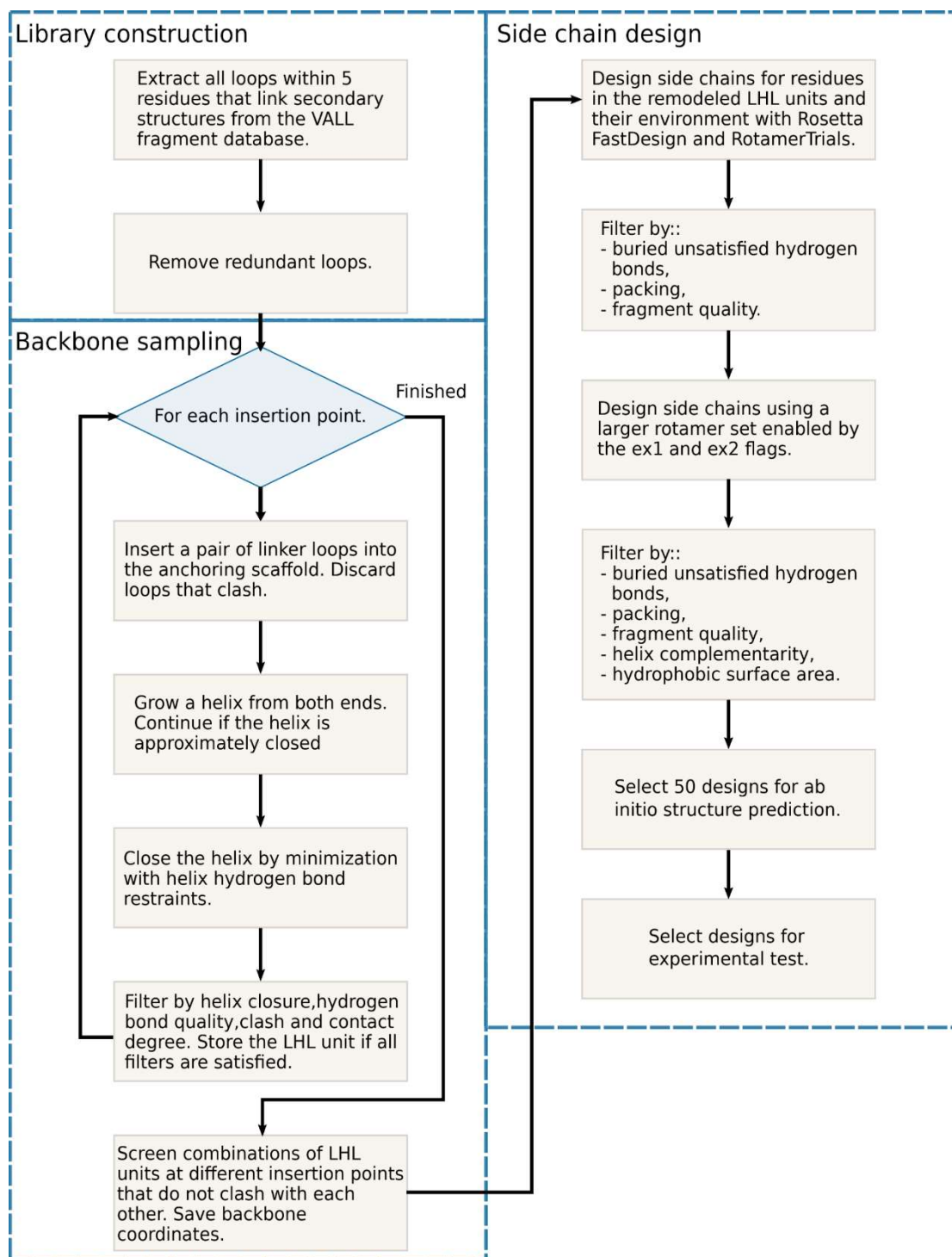


Figure 2.4 Detailed flowchart of the LUCS design protocol.

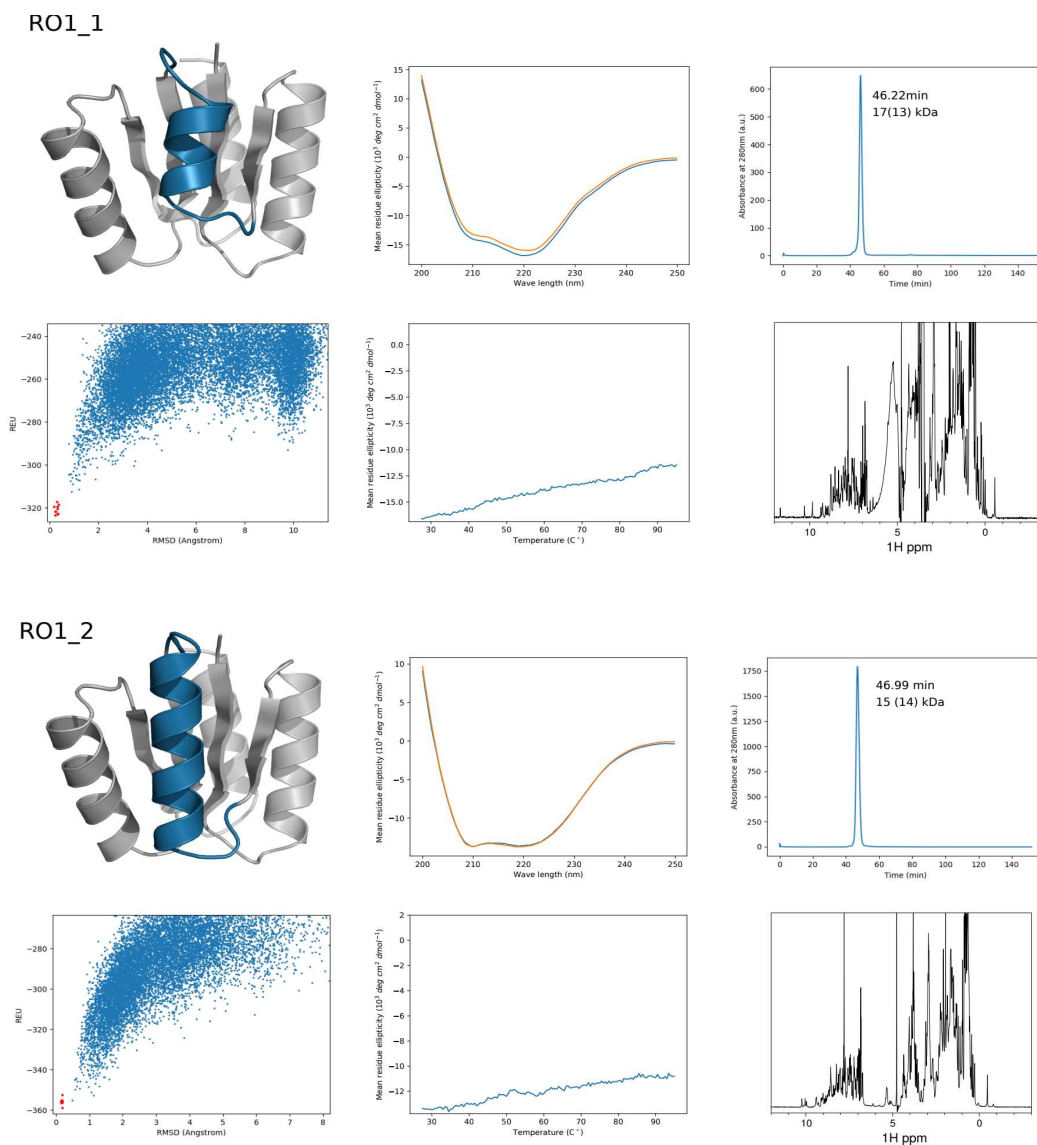
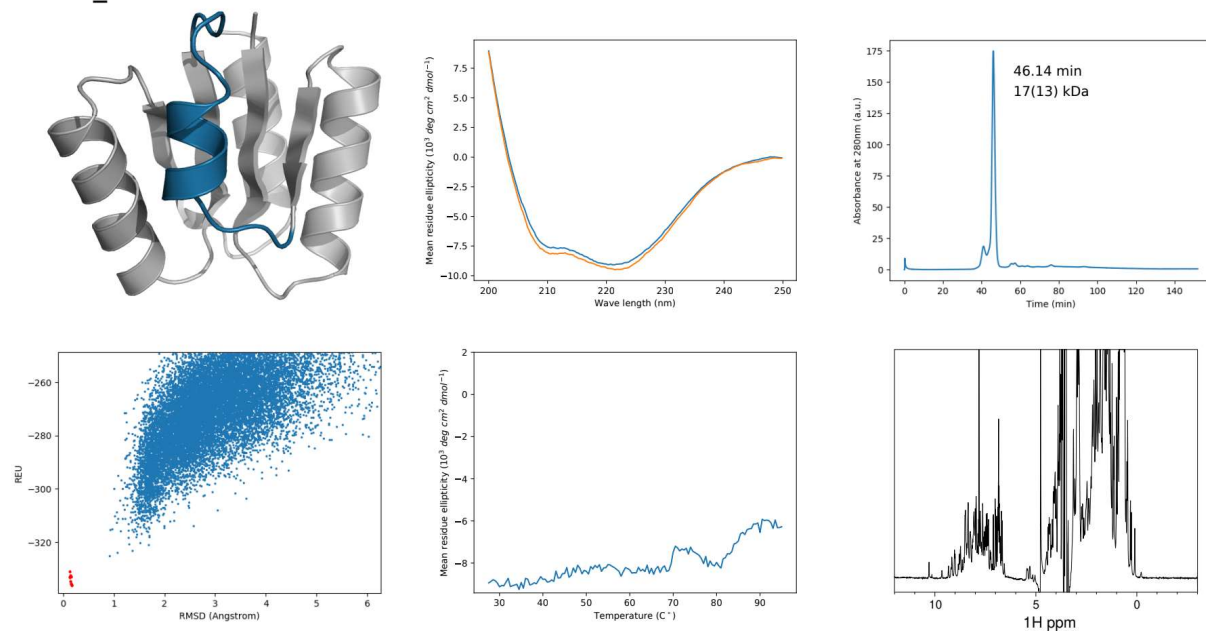


Figure 2.5 Characterization of well-folded designs.

For each design, the design model is shown in the upper left panel with the reshaped LHL units shown in blue. The lower left panels show the results of Rosetta *ab initio* structure prediction simulations. Each blue point represents a model from the prediction simulations and each red point represents a relaxed design model. For the NTF2 fold designs, the *ab initio* structure prediction simulations used a biased fragment set (see Methods). The upper middle panels are the CD spectra at 25°C before (blue) and after measuring a melting curve (orange). The lower panels are CD melting curves measured at 220nm. Most designed proteins do not show a melting transition below 95°C, indicating that they are not notably destabilized relative to their *de novo* designed parent proteins (Koga, et al., 2012) (Marcos, et al., 2017). The upper right panels show the size exclusion chromatograms. The peak positions and molecular weights calculated from peak positions (molecular weights calculated from amino acid sequences) are shown next to the monomer peak. For designs RO1_8, RO1_9, RO2_5, RO2_6, RO2_9, RO2_10, RO2_20 and RO2_25, the

chromatograms were measured using samples purified by gel filtration. The chromatograms for the remainder of the designs were measured directly after His-tag purification. The lower right panels show the 1D- ^1H NMR spectra.

RO1_5



RO1_8

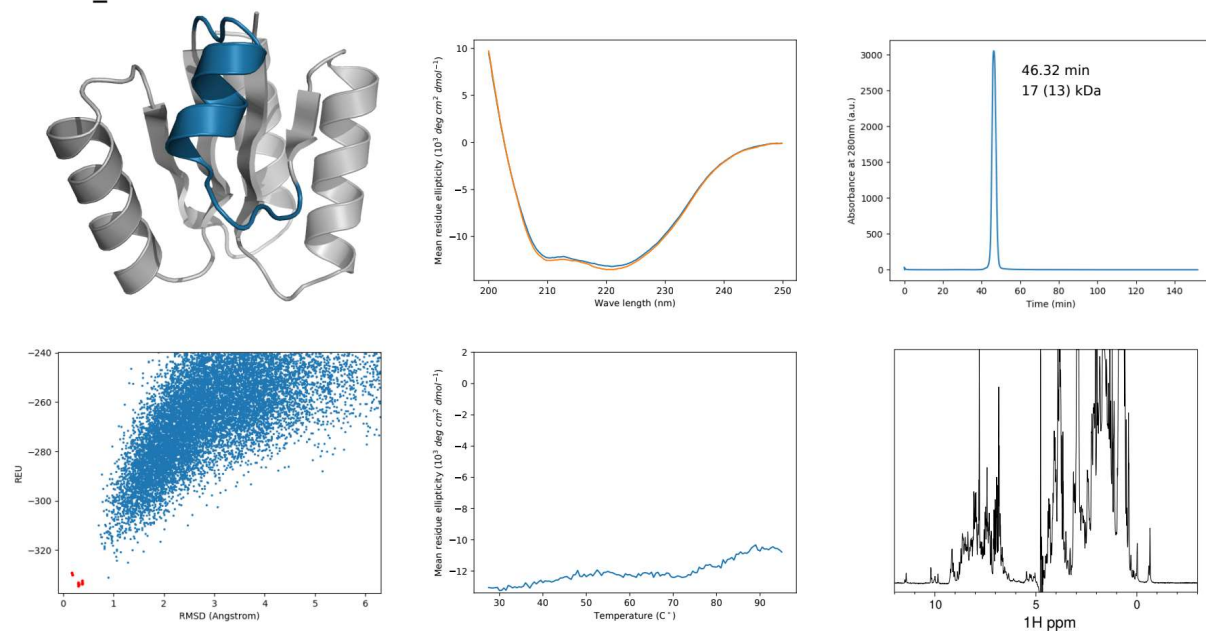
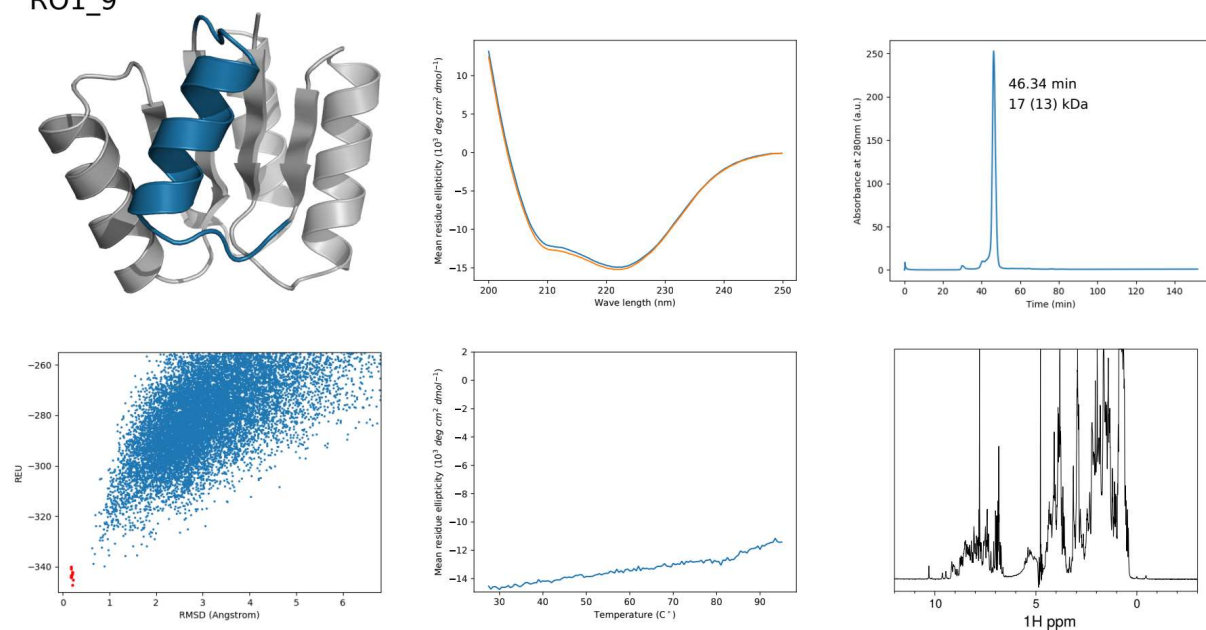


Figure 2.5. Characterization of well folded designs cont.

RO1_9



RO2_1

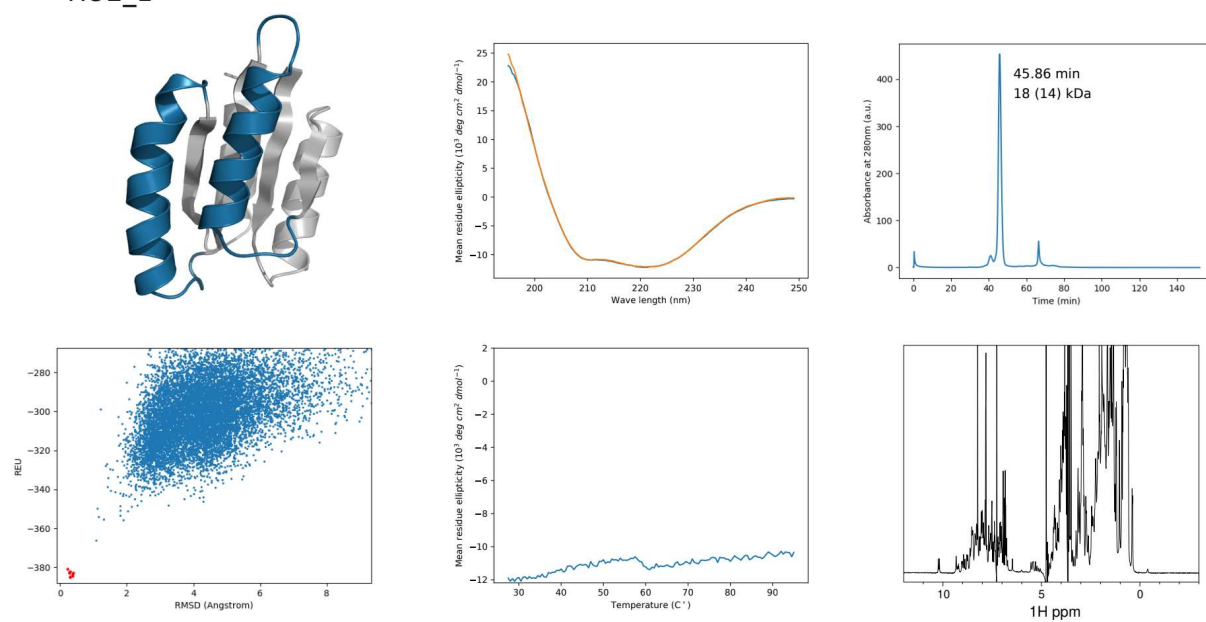
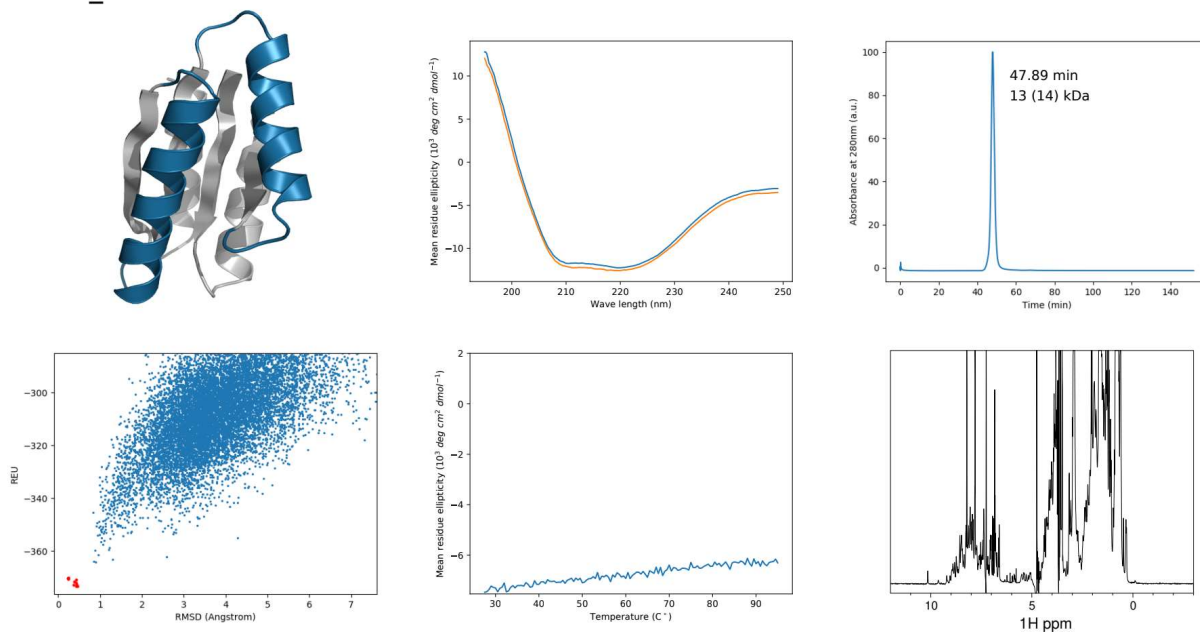


Figure 2.5. Characterization of well folded designs cont.

RO2_5



RO2_6

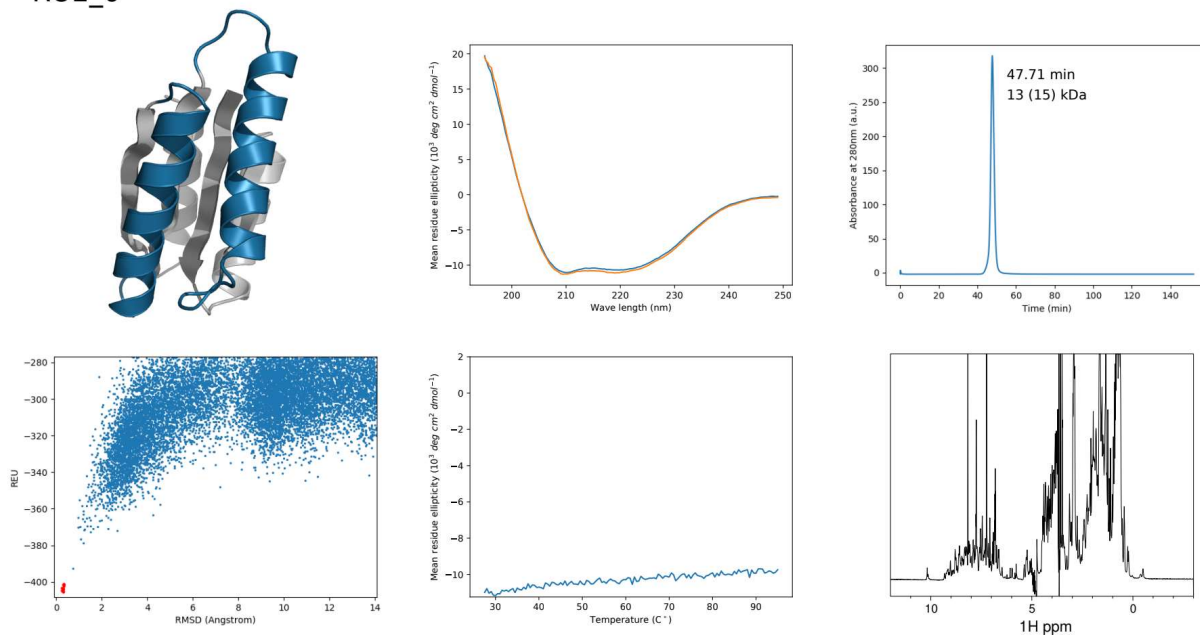
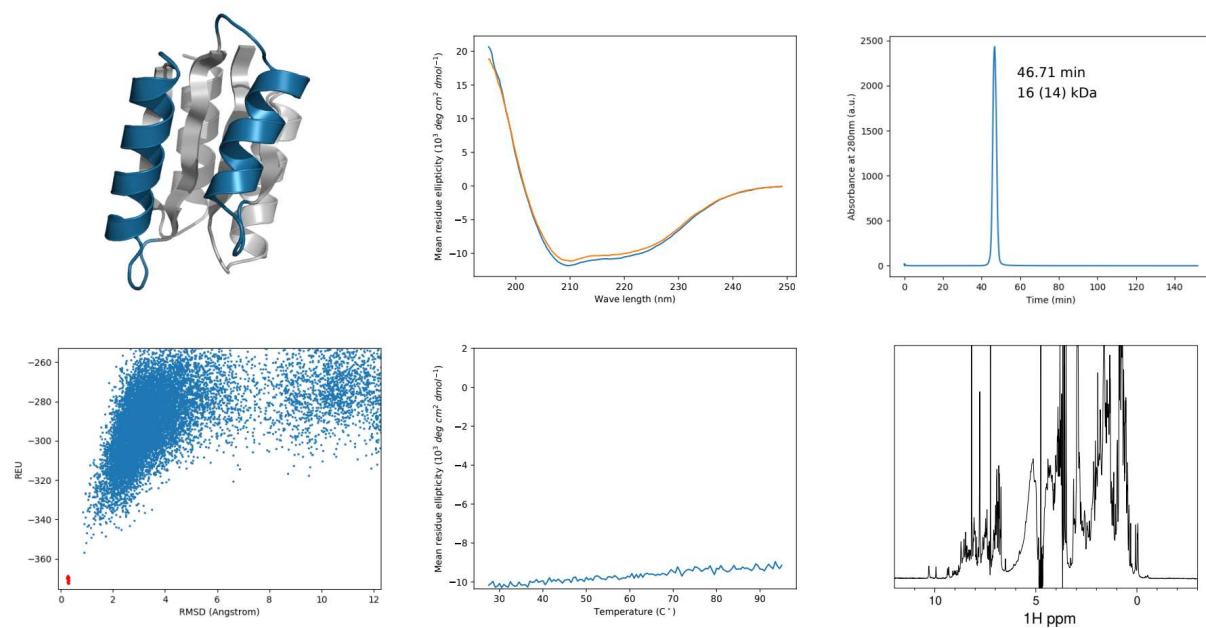


Figure 2.5. Characterization of well folded designs cont.

RO2_9



RO2_10

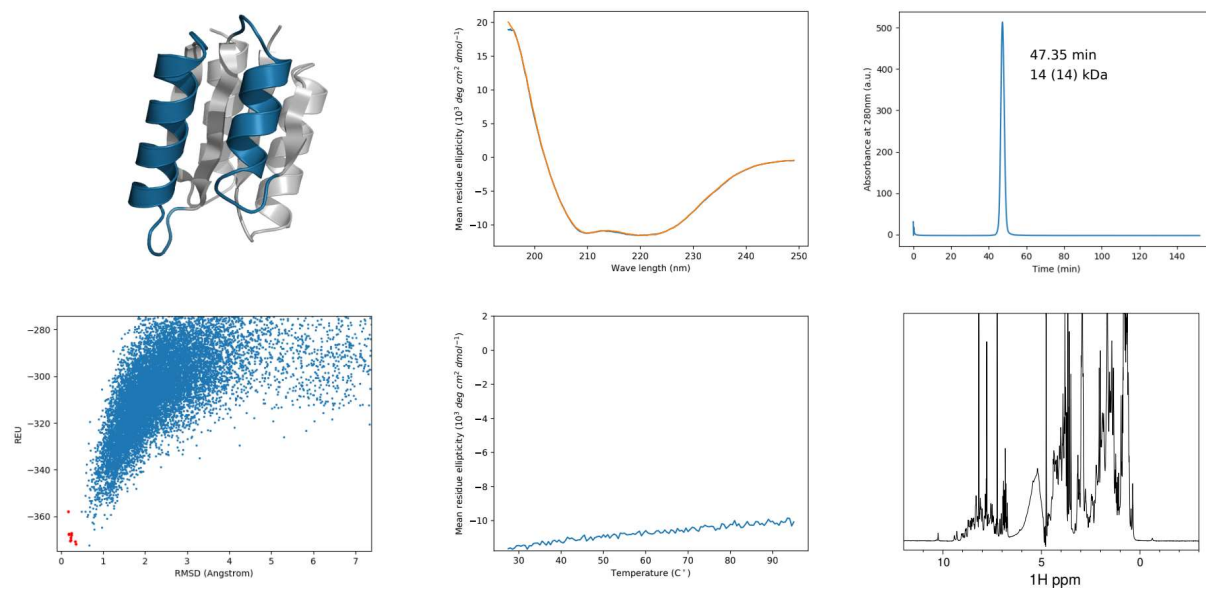
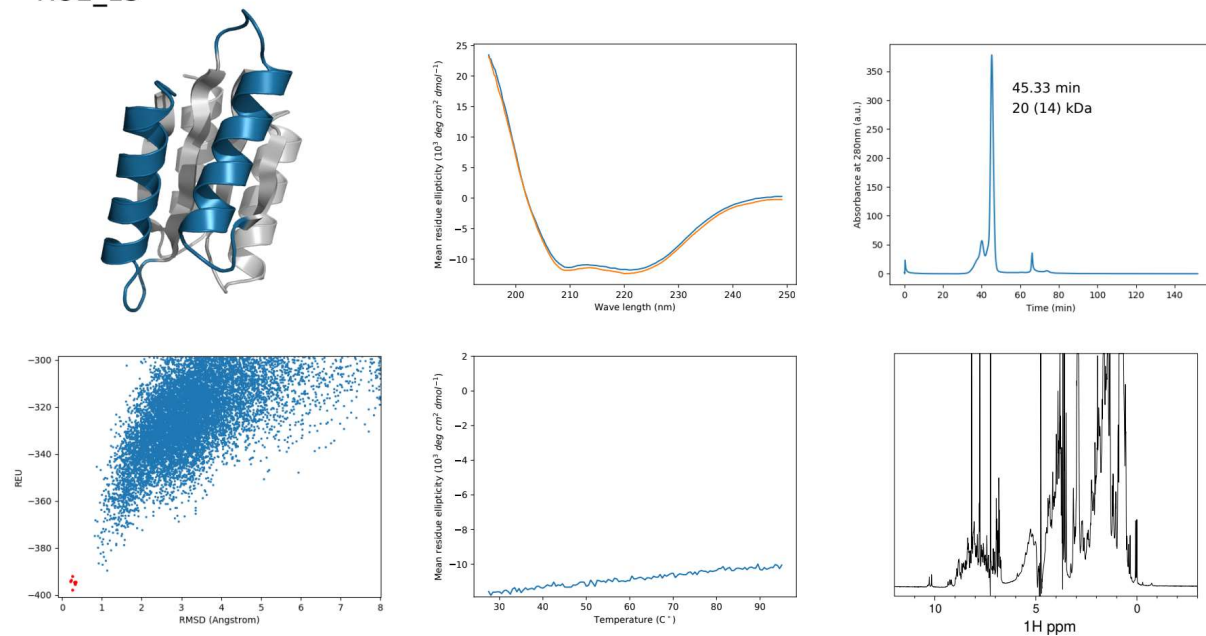


Figure 2.5. Characterization of well folded designs cont.

RO2_15



RO2_20

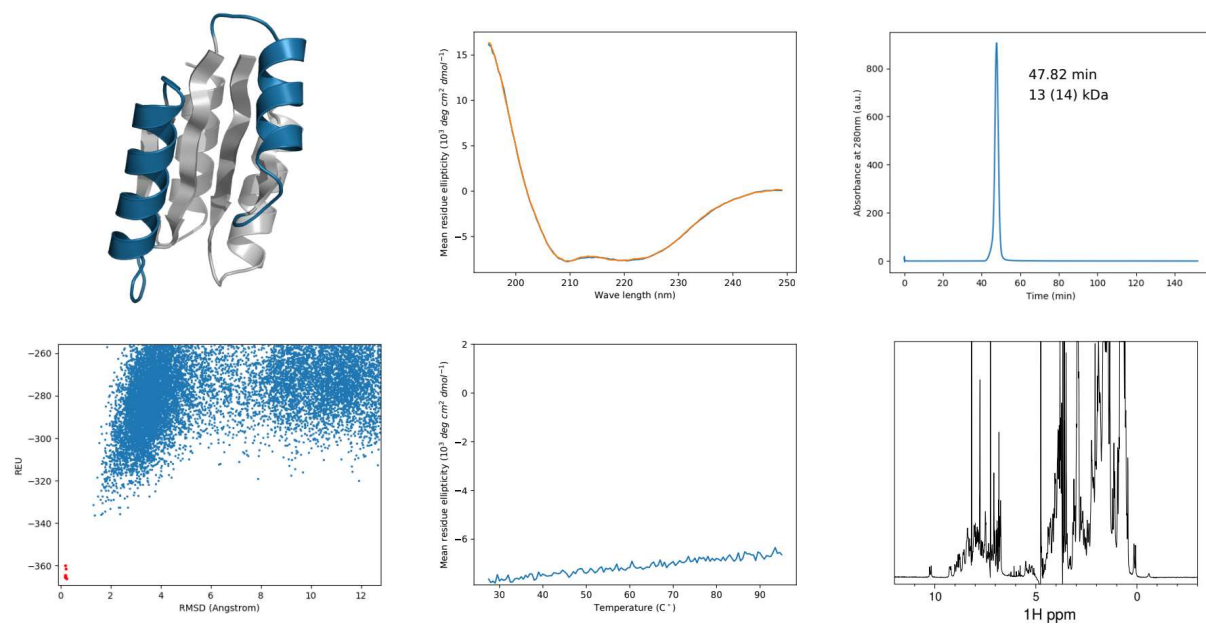
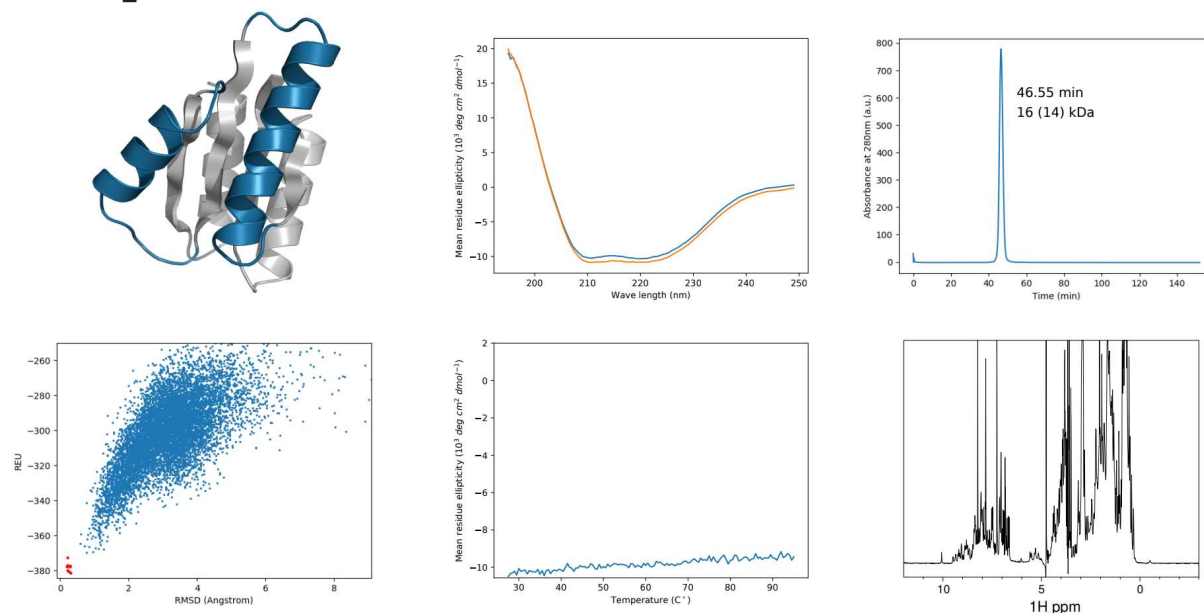


Figure 2.5. Characterization of well folded designs cont.

RO2_25



NT_1

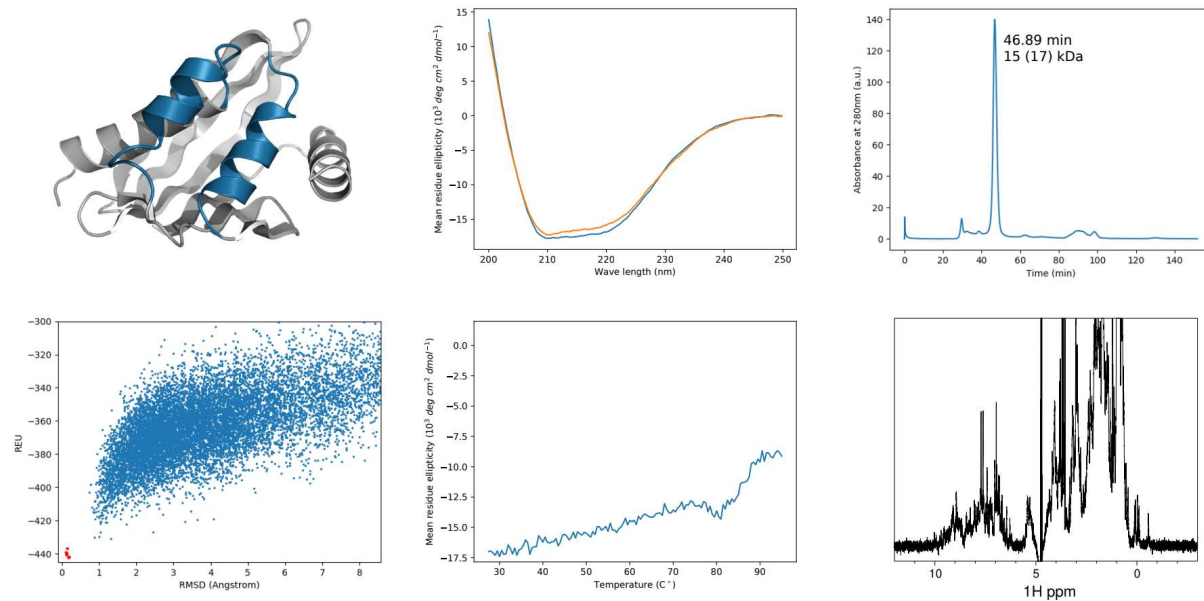
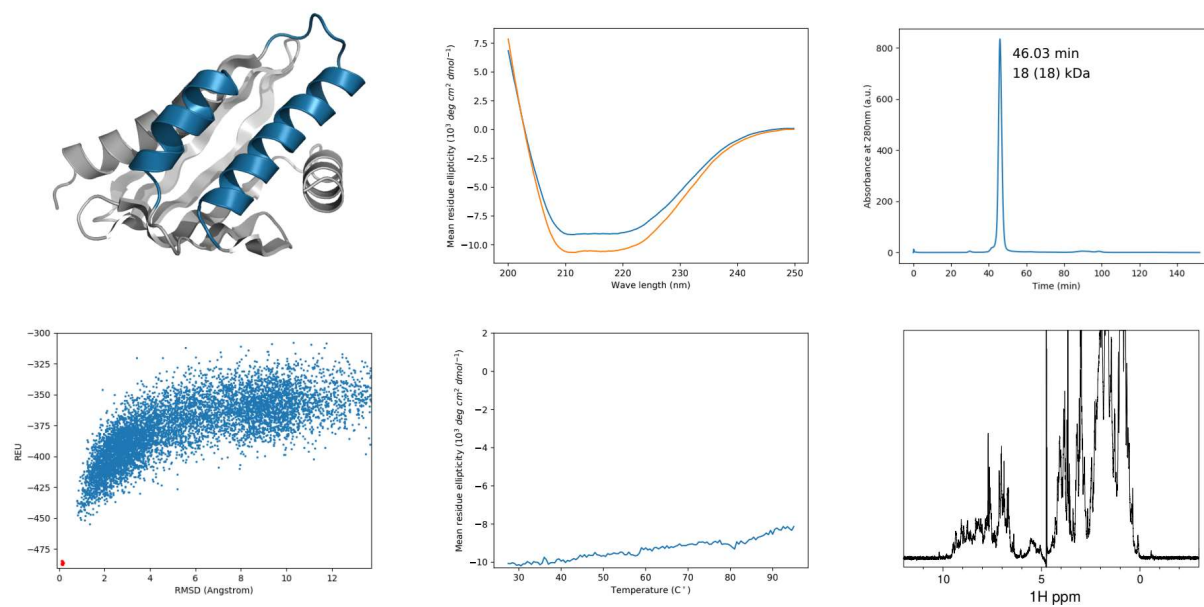


Figure 2.5. Characterization of well folded designs cont.

NT_8



NT_9

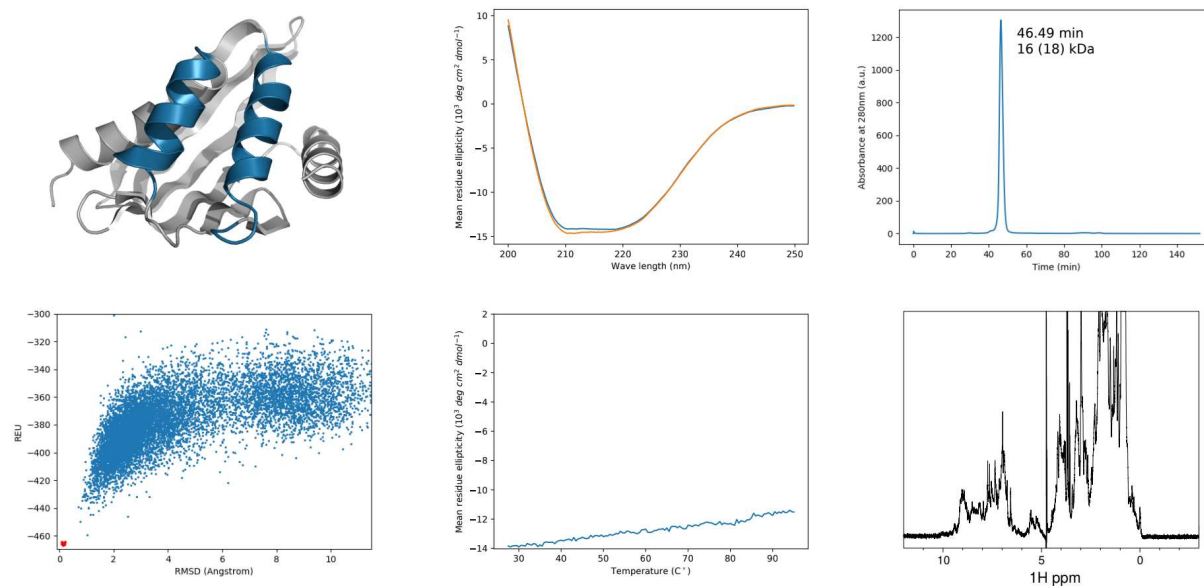


Figure 2.5. Characterization of well folded designs cont.

NT_10

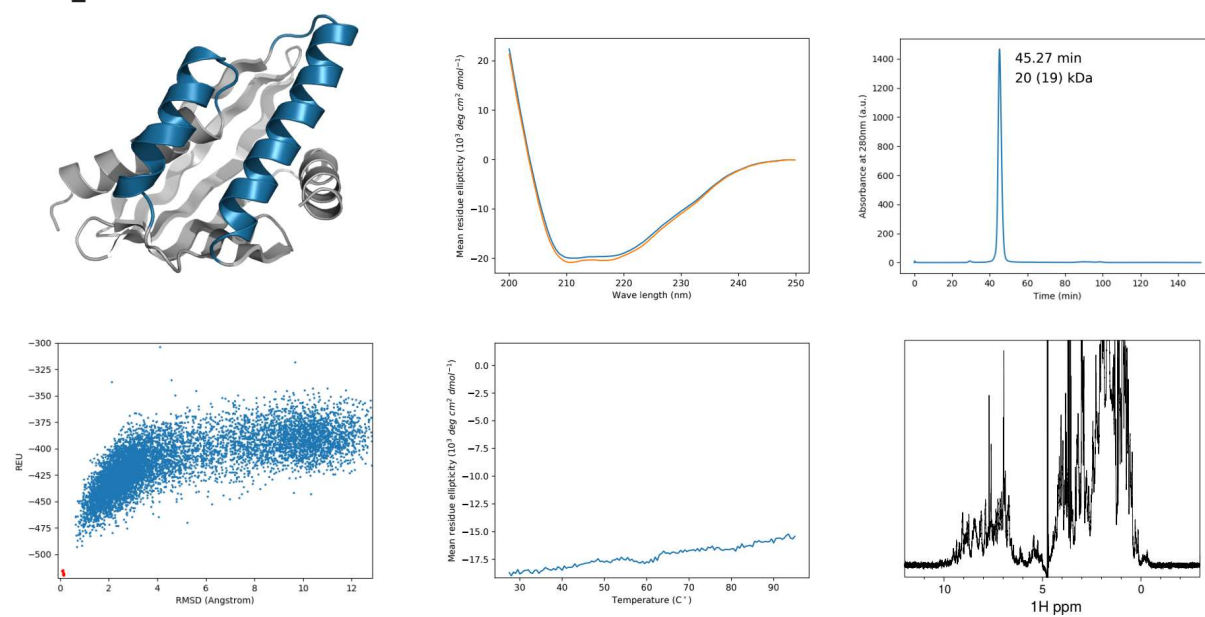


Figure 2.5. Characterization of well folded designs cont.

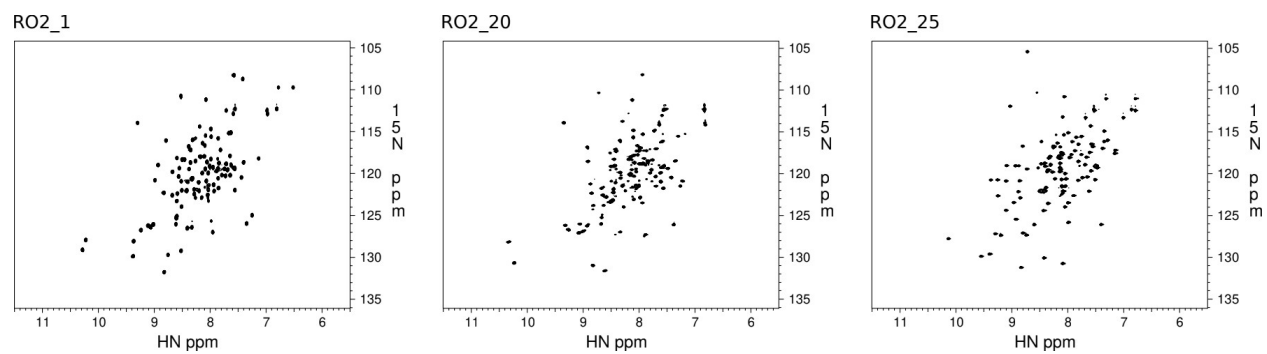


Figure 2.6 ^{15}N - ^1H HSQC spectra of designs whose structures were solved by NMR.

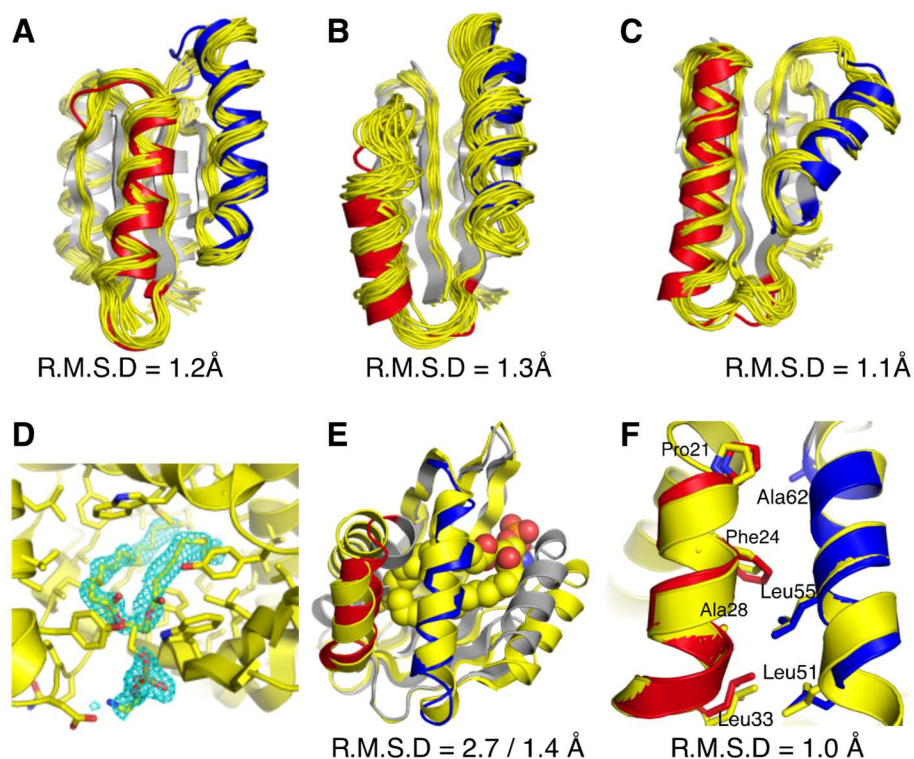


Figure 2.7 Close agreement between models and experimentally determined structures of designed proteins.

A-C, designs for the Rossmann fold topology and **D-F**, design for the NTF2 fold topology. Experimentally determined structures are shown in yellow and design models in grey with the reshaped LHL elements highlighted in red and blue. **A-C**. Comparison between computational models and NMR structures for designs RO2_1(**A**), RO2_20(**B**) and RO2_25(**C**). Also shown are the backbone heavy atom RMSDs calculated using the lowest energy structure from the NMR ensemble. **D**. The binding pocket of a phosphatidylethanolamine ligand. The 2Fo - Fc electron density map (cyan) for the ligand molecule is shown at 1.0 σ level. **E**. Comparison between computational model and X-ray crystal structure for the design NT_9. The phosphatidylethanolamine ligand is shown in space fill representation (carbon atoms in yellow, oxygen atoms in red, phosphorus atoms in orange, and nitrogen atoms in blue). Also shown are the backbone heavy atom RMSDs calculated including or excluding the terminal helices, respectively. **F**. Alignment between the designed helices in the computational model and the experimentally solved structure for design NT-9. The hydrophobic residues at the packing interface are shown in stick representation. The RMSD shown includes the helix backbone heavy atoms and side chain heavy atoms displayed as sticks.

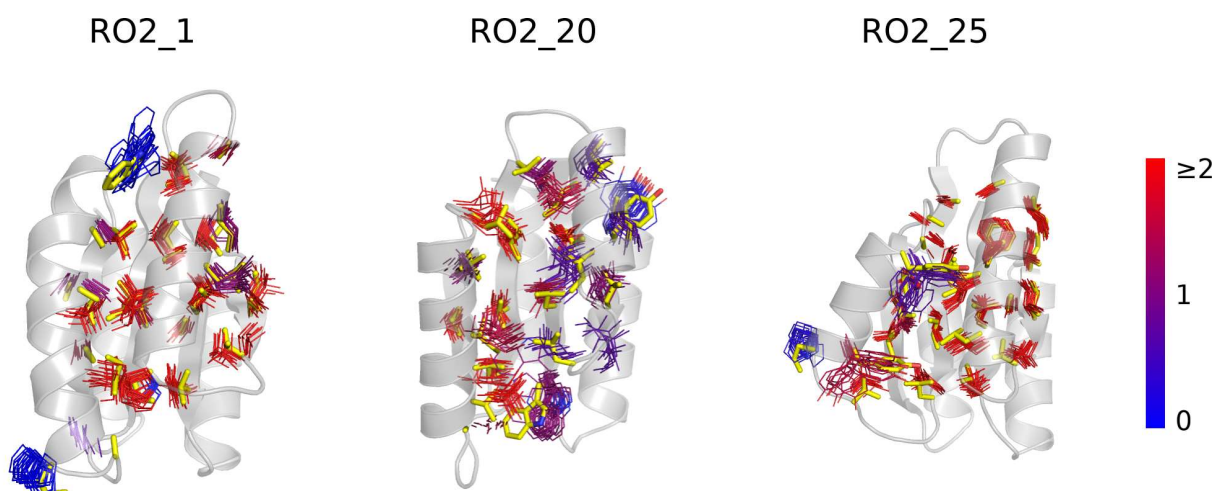


Figure 2.8 Conformations of core side chains overall agree between designed models and structures solved by NMR.

The backbones of the designed models are shown as cartoons. The hydrophobic core side chains of models are shown as yellow sticks. The conformations of the side chain in the NMR models are shown as lines and colored according to the ratio of the number of long range (distance in primary sequence larger than 3) NOEs involving its side chain atoms over the number of its side chain atoms.

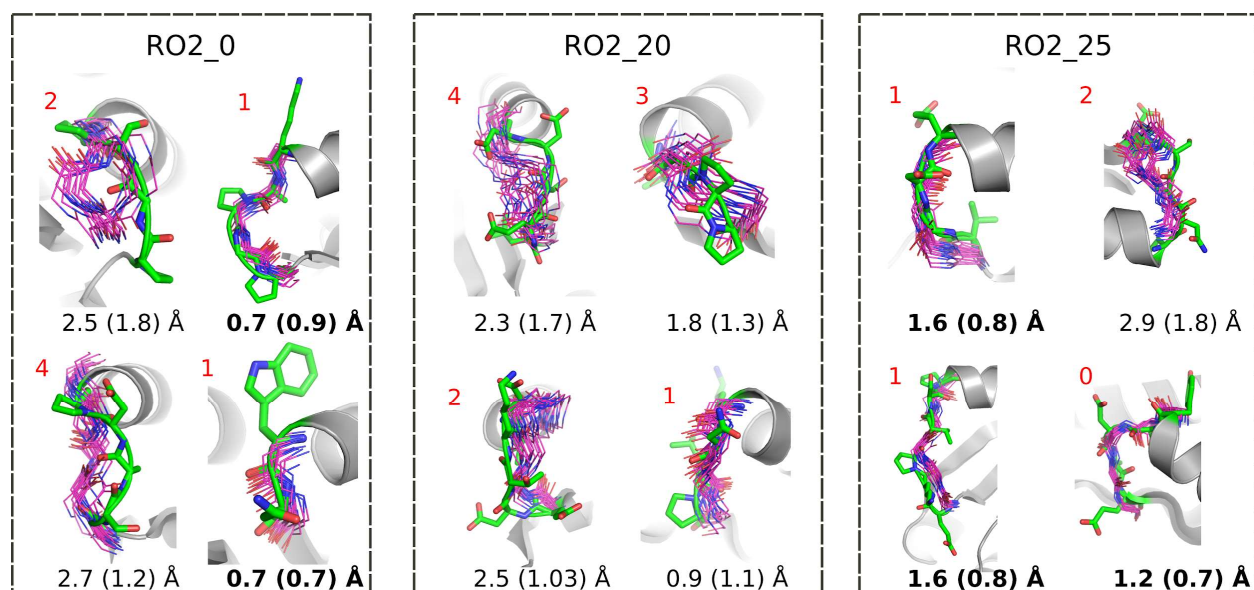


Figure 2.9 Comparison of loop conformations between designed models and structures solved by NMR.

The loops of designed models are shown as green sticks. The backbone heavy atoms of experimentally solved loops are shown in magenta. The backbone heavy atom RMSDs between the designs and lowest energy NMR models are shown below each loop and the max loop residue ensemble backbone RMSDs calculated by CCPN analysis are shown in parentheses. The 5 converged loops (pairwise backbone RMSD within the ensemble of NMR models within 1 Å) have RMSDs within 1.6 Å (shown in bold). The number of residues that lack long range (distance in primary sequence larger than 3) NOE restraints are shown at the top left of each loop (red).

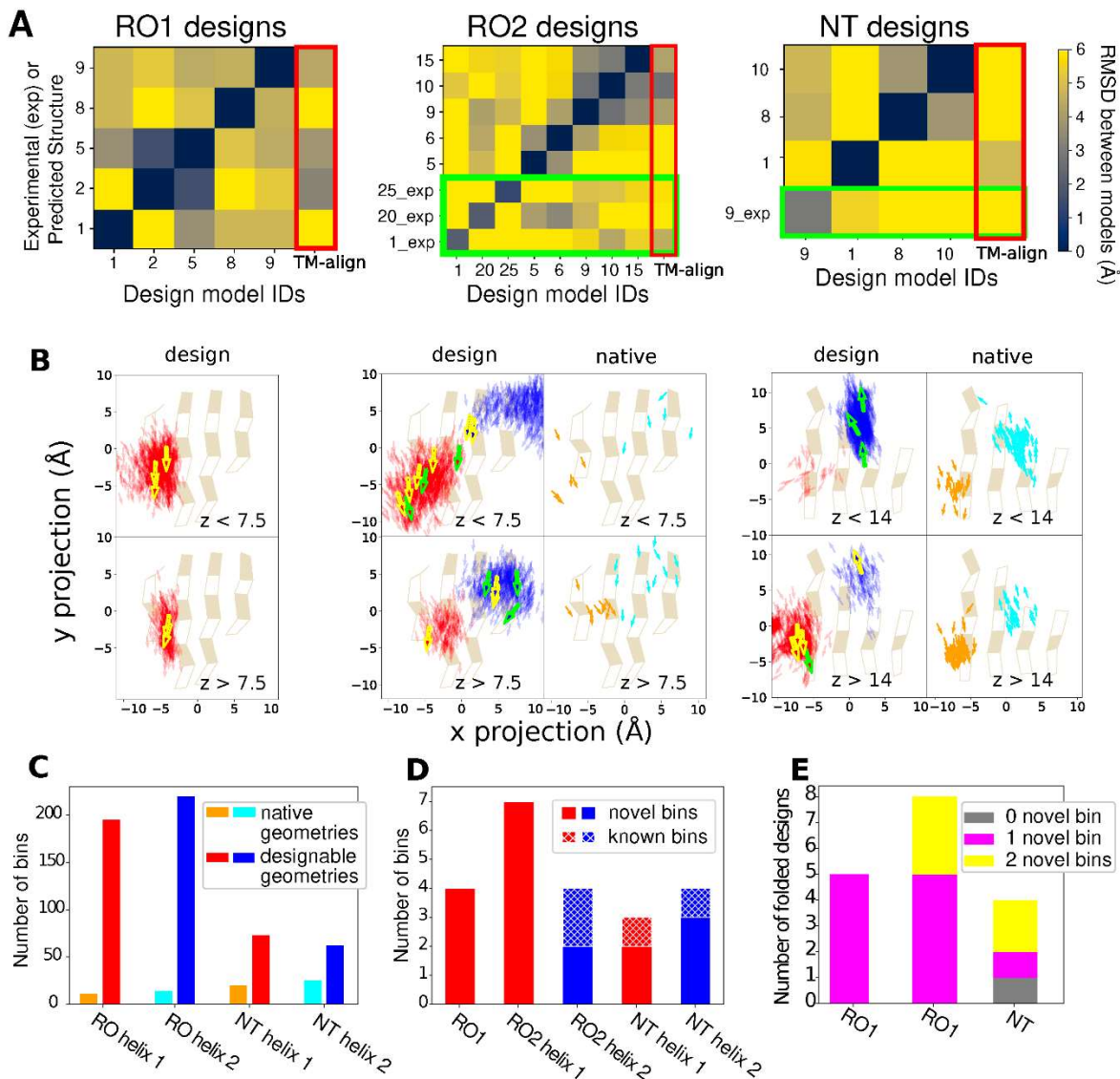


Figure 2.10 Geometry space sampled by *de novo* designed fold families.

In **A** and **B**, the columns show the 3 design problems: Left, Rossmann fold with one designed LHL unit (RO1); middle, Rossmann fold with two designed LHL units (RO2); right: NTF2 fold with two designed LHL units (NT). **A**. Heatmaps showing backbone RMSDs between the reshaped LHL-regions of well-folded designs, comparing design models (x axis) with experimentally determined structures (_exp) or lowest-scoring models from Rosetta structure prediction (y axis). Green boxes show RMSDs calculated using experimentally solved structures. Red boxes (right columns) show the RMSDs between designs and the closest known structures found by TM-align. **B**. Projection of centers and directions of designed helices (arrows) onto the underlying beta sheets. For the RO2 (middle) and NT (right) columns, panels show distributions in designable models (**Figure 2.1D**) on the left (helices colored red and blue), and in known naturally occurring structures on the right (corresponding helices in orange and cyan). The two rows show helices on two z-level planes based on their distances from the beta-sheet projection plane. For planes that

have more than 1000 sampled structures, only 1000 randomly selected helices are shown. For the designs, experimentally confirmed folded designs are represented as bold arrows with yellow boundaries and designs with experimentally solved structures as bold arrows with green boundaries. For the natural proteins, the Rossmann fold structures are from the CATH superfamily 3.40.50.1980 and the NTF2 fold structures are from the CATH superfamily 3.10.450.50. **C.** Number of structure bins occupied by known structures (orange, cyan) and sampled by designable models generated by LUCS (red, blue). **D.** Structure bins occupied by well folded designs. **E.** Classification of the well folded structures by the number of novel structure bins they occupy.

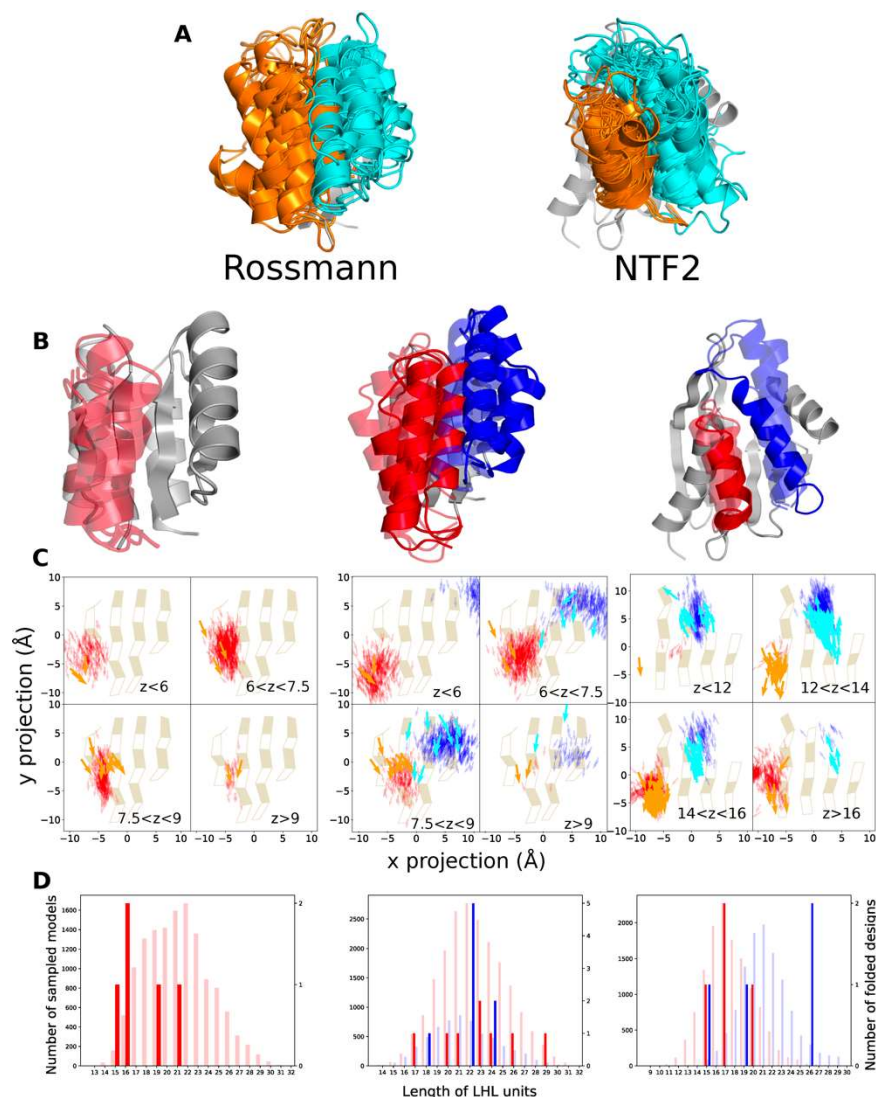


Figure 2.11 Distributions of LHL geometries in designed *de novo* fold families.

A. Structures of native Rossmann fold (left) and NTF2 fold (right) proteins (the two helices corresponding to the designed regions are shown in orange and cyan). The Rossmann fold structures are from the CATH superfamily 3.40.50.1980 and the NTF2 fold structures are from the CATH superfamily 3.10.450.50. In **B**, **C** and **D**, columns show the 3 design problems: Left, Rossmann fold with one designed LHL unit; middle, Rossmann fold with two designed LHL units; right: NTF2 fold with two designed LHL units. **B.** Superimposition of well-folded designs. The designed LHL units are colored in red or blue. Designs with experimentally solved structures are shown in solid colors and all other well-folded designs are shown in transparent colors. **C.** Projection of centers and directions of helices from designable LUCS models (red, blue) and known structures (orange, cyan) onto the underlying beta sheets, as shown in **Figure 2.10** but overlaid, depicted in 4 z-levels, and orange / cyan arrows are thicker / longer for visual clarity. **D.** The distributions of LHL lengths of the sampled models (light colors) and the well-folded designs (dark colors).

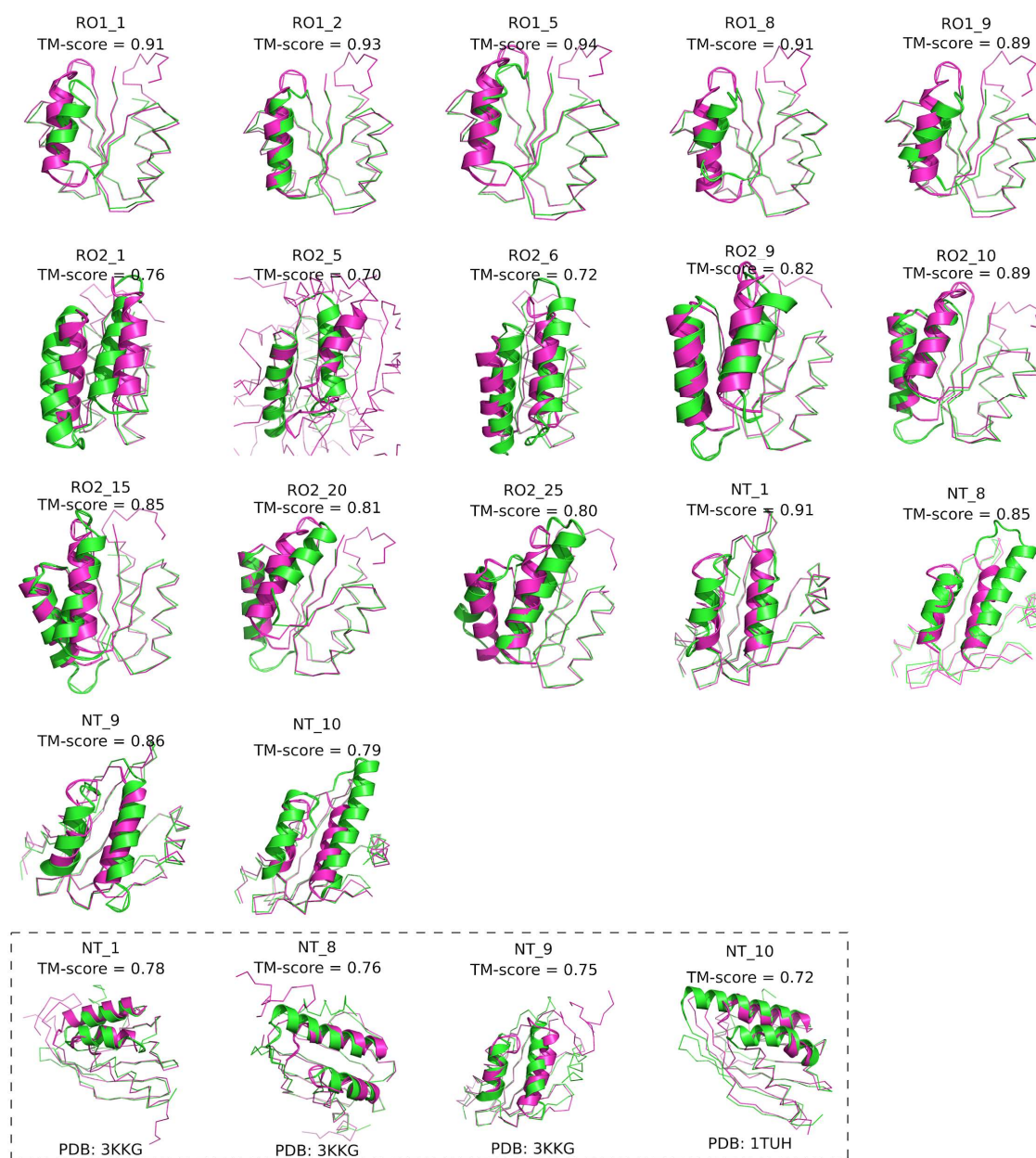


Figure 2.12 Models of well-folded designs show considerable geometry differences to their best matches from TM-align.

Design models are shown in green and the matched structures are shown in magenta. Except for RO2_5, all Rossmann fold designs matched best to PDB:2KPO which is the *de novo* designed input scaffold PDB:2LV8. All natural proteins in the top 10 TM-align matches to Rossmann fold designs have different topologies from the designed topology. All NTF2 fold designs matched best to the *de novo* designed input scaffold PDB:5TPJ. The last row shows the best natural protein matches to the NTF2 fold designs.

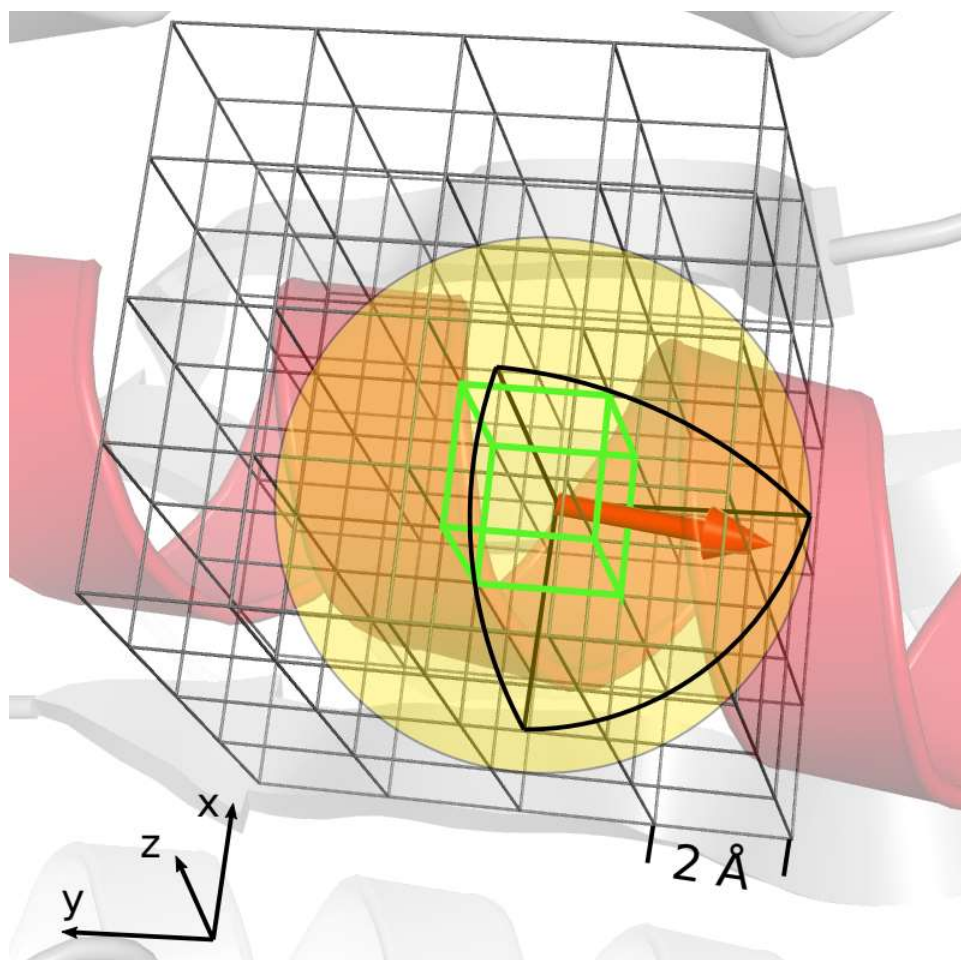


Figure 2.13 Graphical representation of binning the helix geometry space.

The helix in question (red) is assigned a vector which has 3 dimensions of position degrees of freedom and 3 dimensions of direction degrees of freedom. The position space is divided into 2 Å cubes and the direction space is divided into 8 octants. The helix vector is binned by the cube (green) and octant (black boundary) that it belongs to.

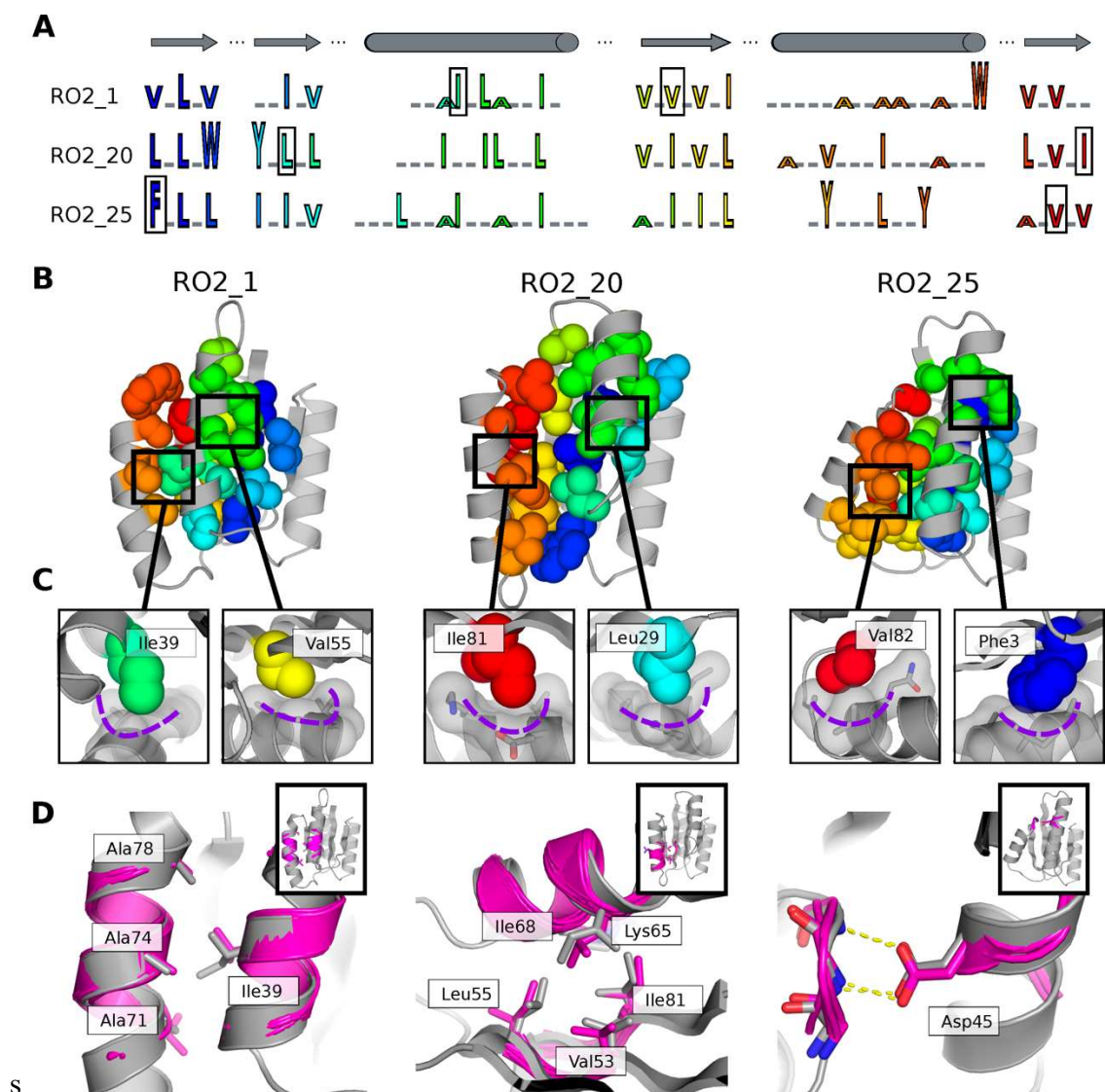


Figure 2.14 Structural features encoding distinct protein geometries.

A. Sequence patterns of the hydrophobic cores in three designed models for the Rossman fold, aligned by corresponding secondary structure elements (top). Hydrophobic residues are shown as letters in rainbow colors ordered by position in the primary protein sequence and scaled by side chain size. Grey underlines indicate positions of surface exposed polar residues. The residues in the boxes are the knob residues shown in **(C)**. **B.** Atomic packing of hydrophobic cores in the three experimentally determined structures for the Rossman fold (**Figure 2.7**). The hydrophobic side chains in the designed cores are shown as spheres. **C.** Knob-socket packing motifs found in the designs. Three residues on a helix (grey sticks and surfaces) form a socket accommodating a knob residue shown as colored spheres. **D.** Examples of tertiary motifs matching the designed LHL structures. The designed structures are shown in grey and the matched motifs are shown in magenta. Sidechains of the best matched tertiary motifs and design models are shown as sticks. Insets indicate location of the tertiary motif in the structure in the same orientation as in **B**.

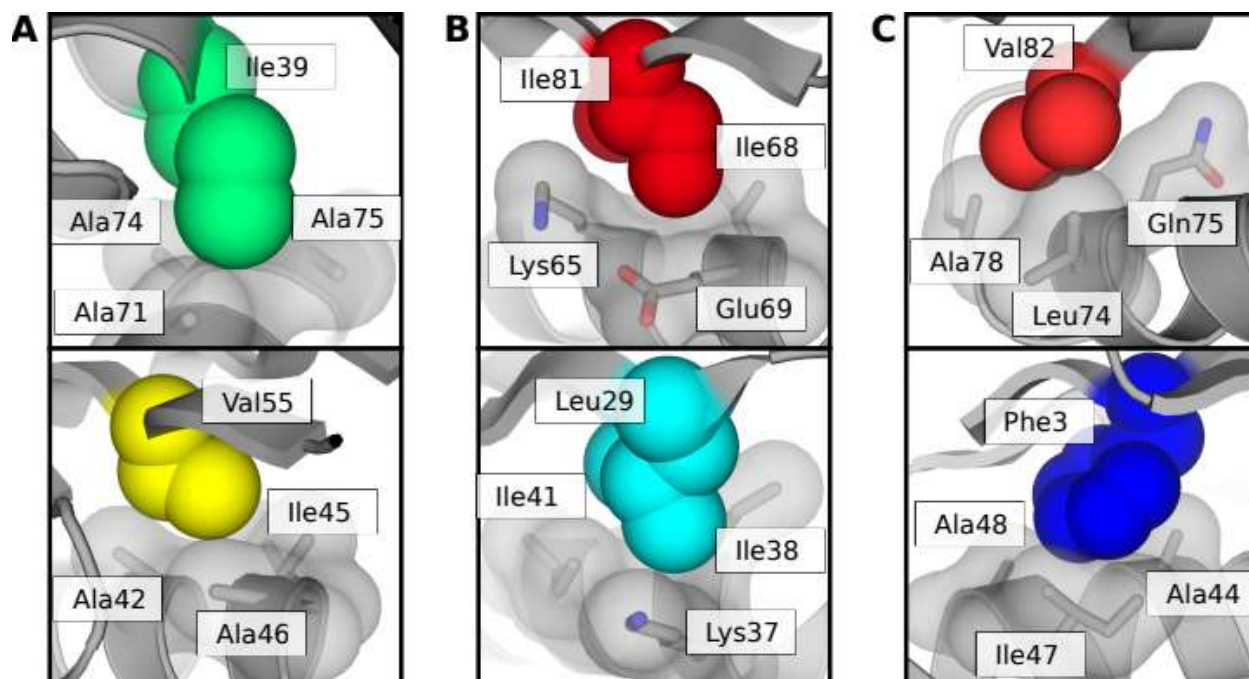


Figure 2.15 Residues that form knobs-into-sockets packing.

The knob residues shown as colored spheres and the socket residues are shown as grey sticks with transparent surfaces. **A.** RO2_1, **B.** RO2_20 **C.** RO2_25.

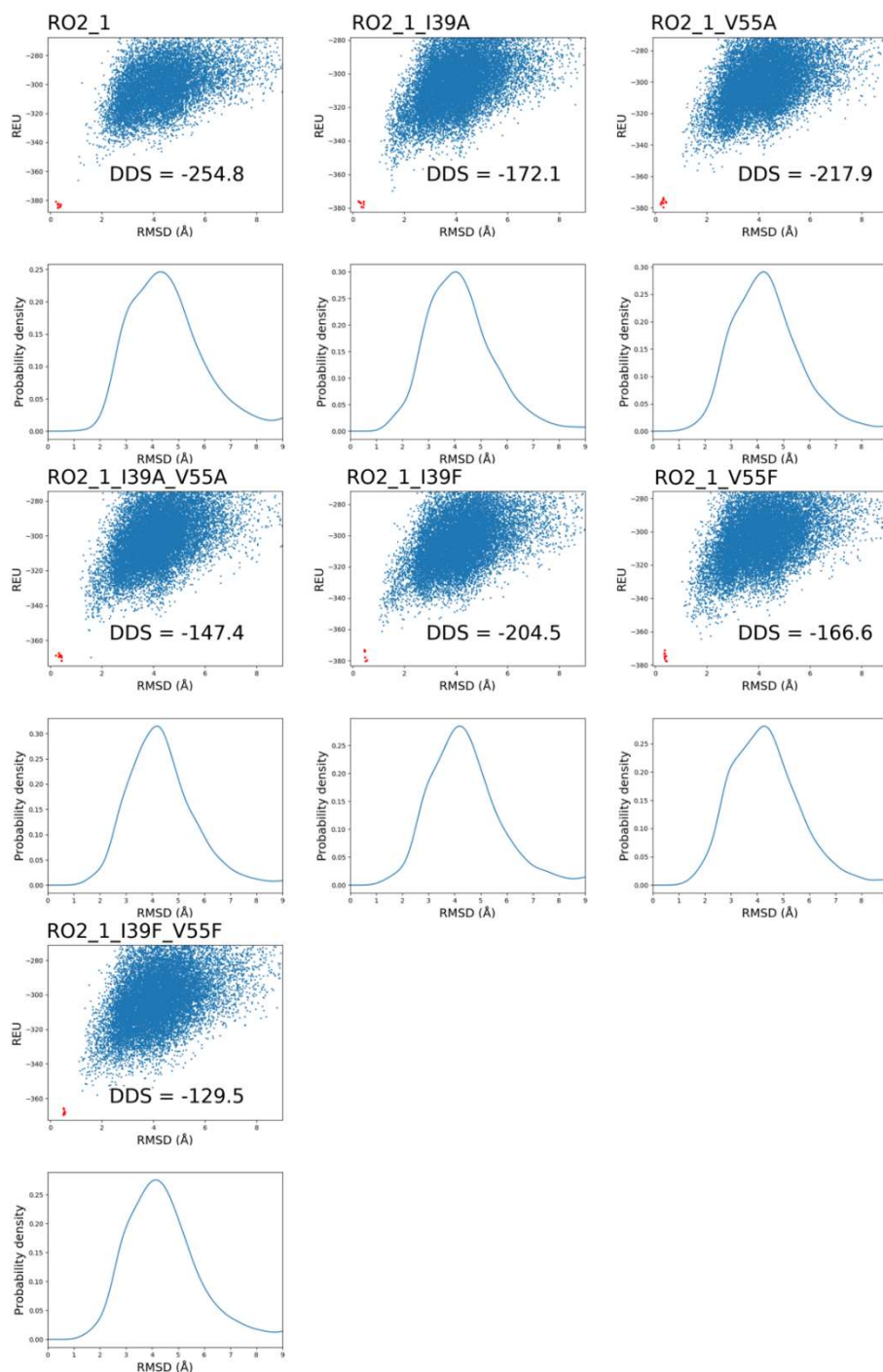


Figure 2.16 Folding simulations for selected designs and their mutations.

In the Rosetta energy unit (REU) vs. RMSD scatter plots, each blue point represents a model predicted by *ab initio* folding simulations and each red point represents a relaxed designed state model. The decoy discrimination scores (DSS) computed as in (Conway & DiMaio, 2016) are also shown. The distributions of RMSDs between predicted models and designed models are shown below the scatter plots.

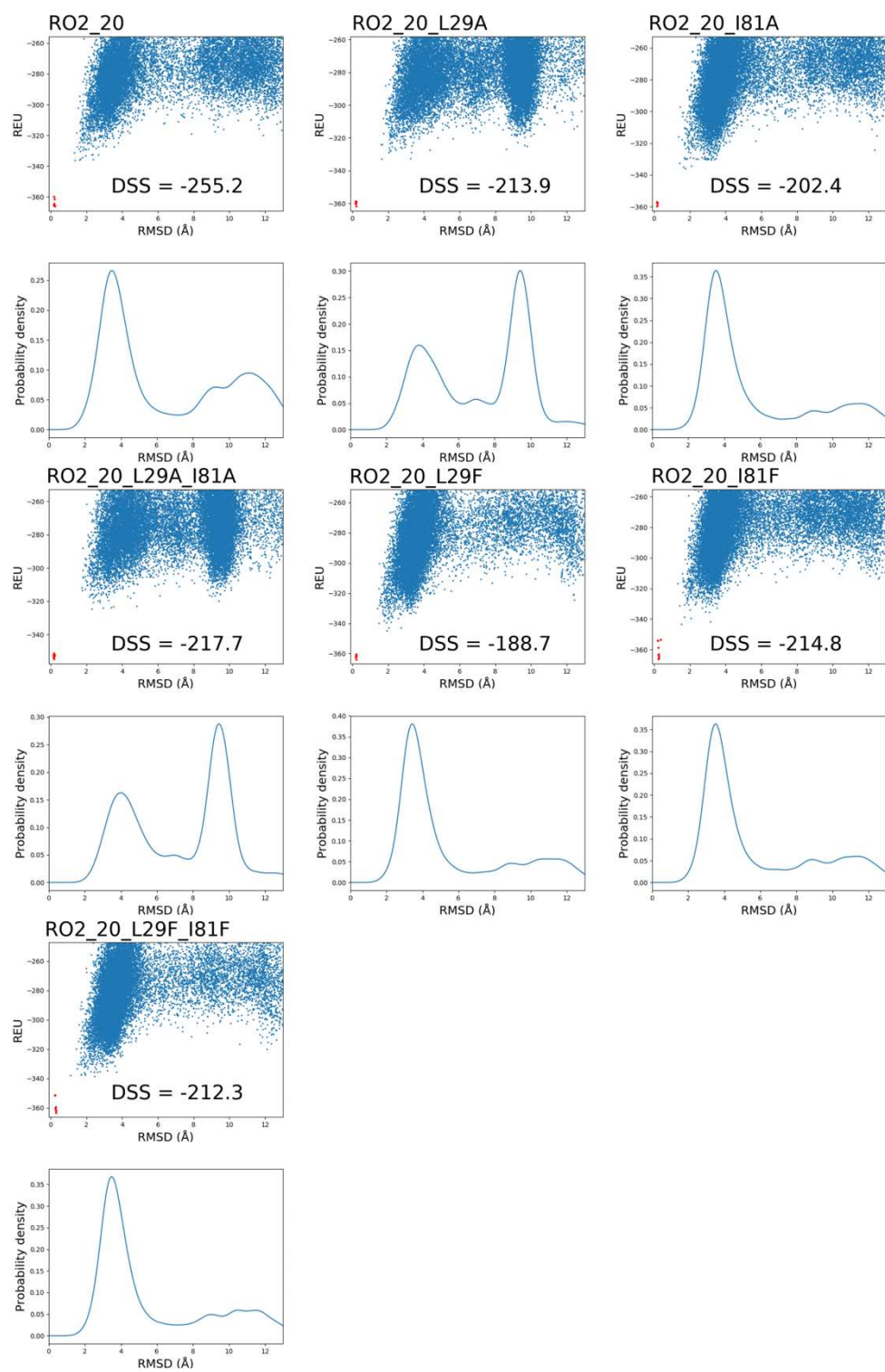


Figure 2.16. Folding simulations for selected designs and their mutations cont.

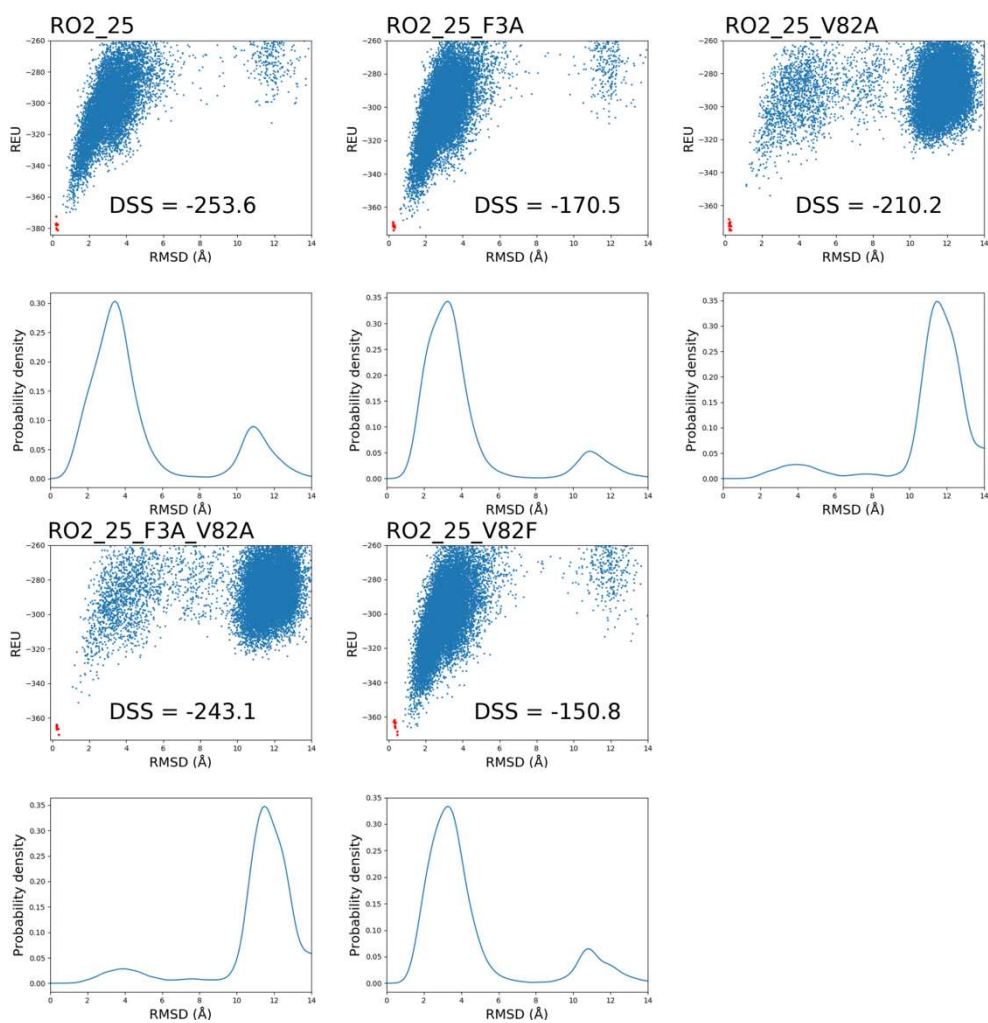


Figure 2.16. Folding simulations for selected designs and their mutations cont.

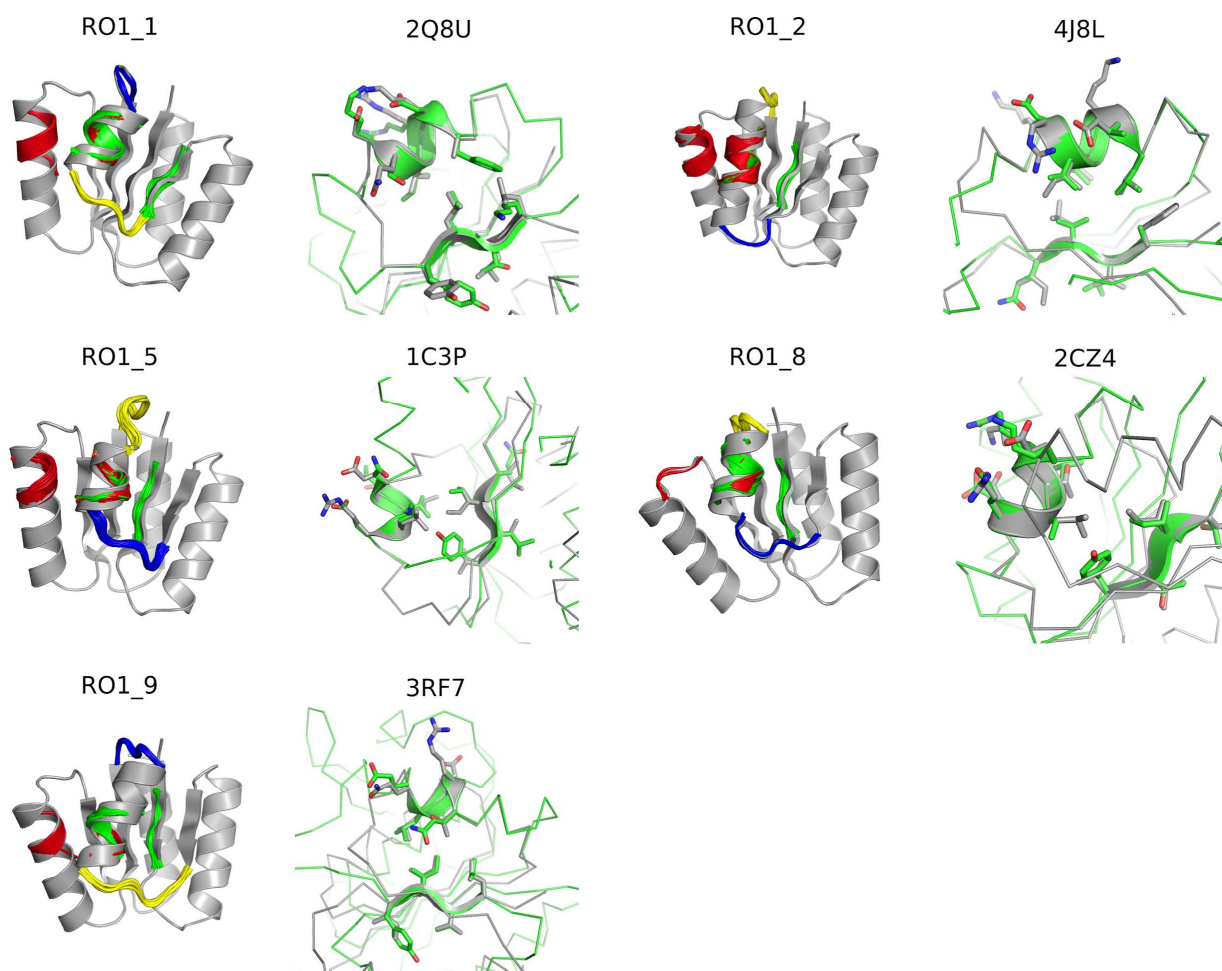


Figure 2.17 Identification of tertiary motifs for designed models.

For each design, the left panel shows the designed model in grey and the matched tertiary motifs in color. The right panel shows the top match of the green motif. The designs are colored in grey and the source structure of one of the motifs in green. The two structures are superimposed at the motif region. The backbones of the matched region are shown as cartoons and the side chains as sticks. The PDB code of the motif source structures are shown at the top of the right panels.

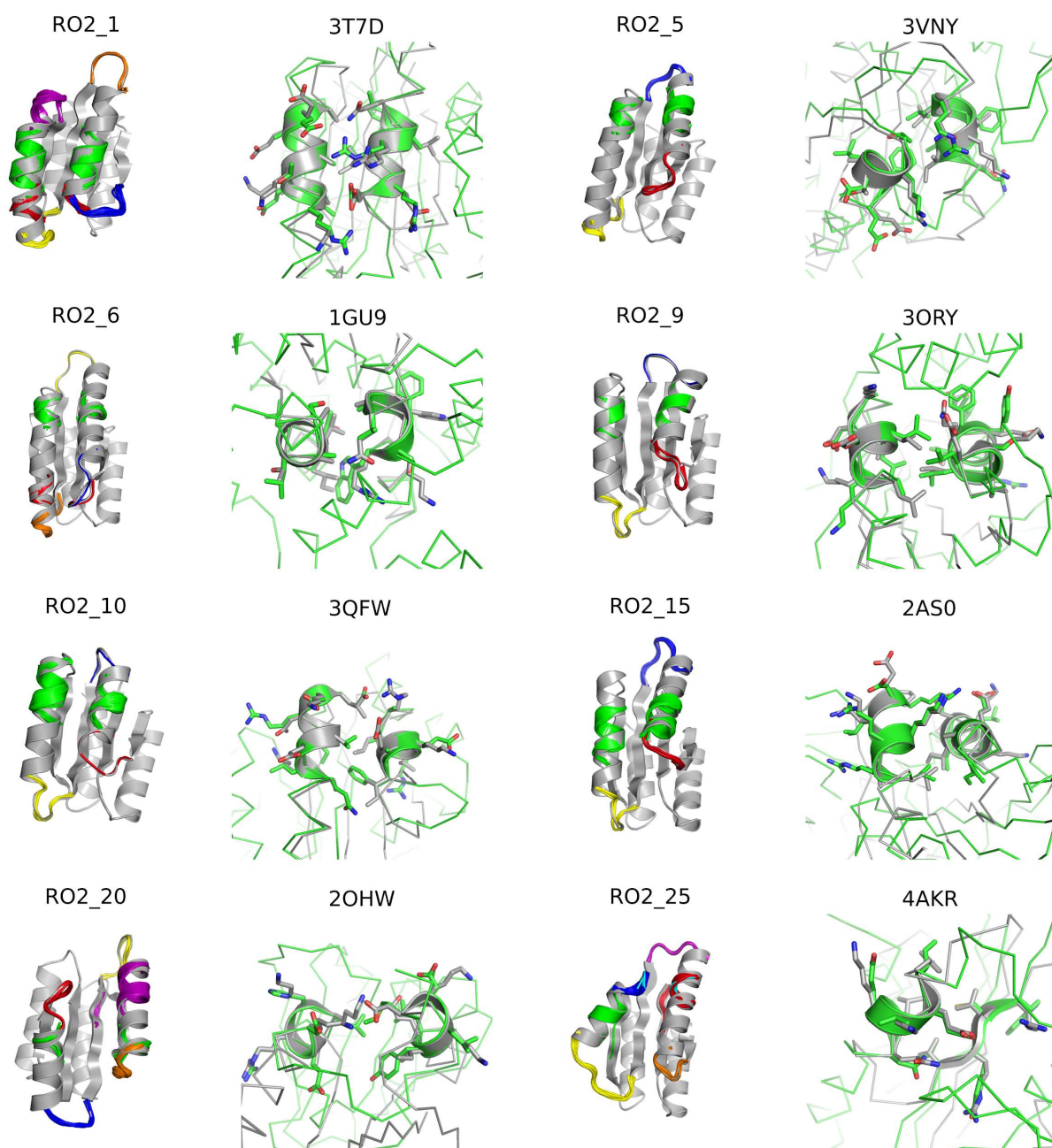


Figure 2.17. Identification of tertiary motifs for designed models cont.

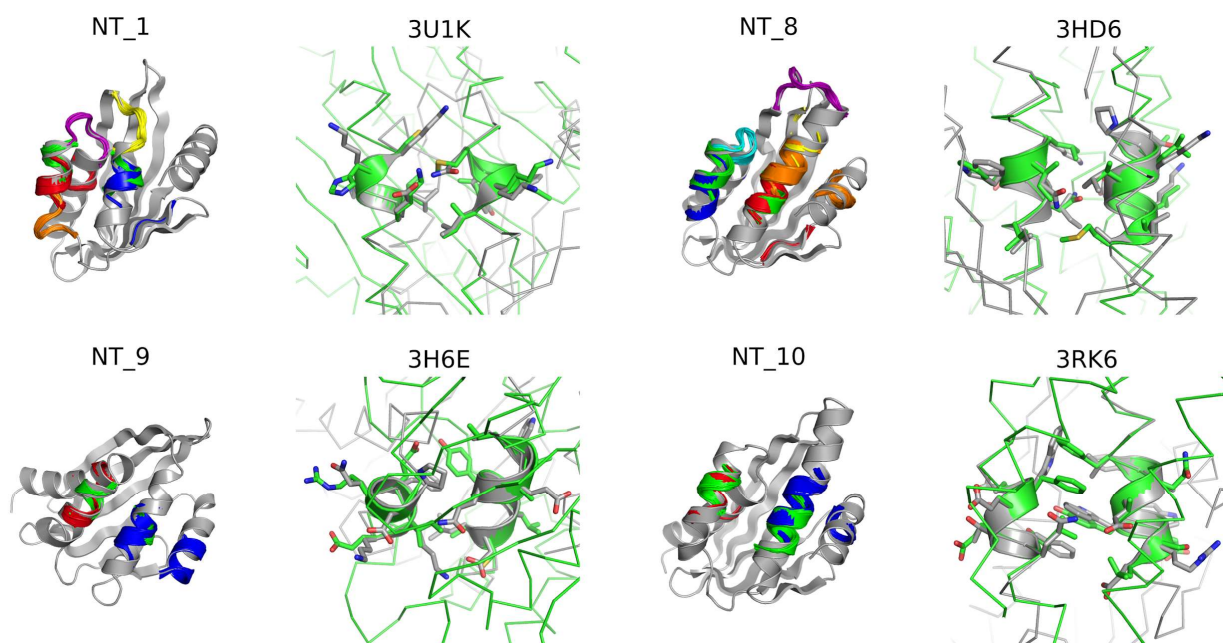


Figure 2.17. Identification of tertiary motifs for designed models cont.

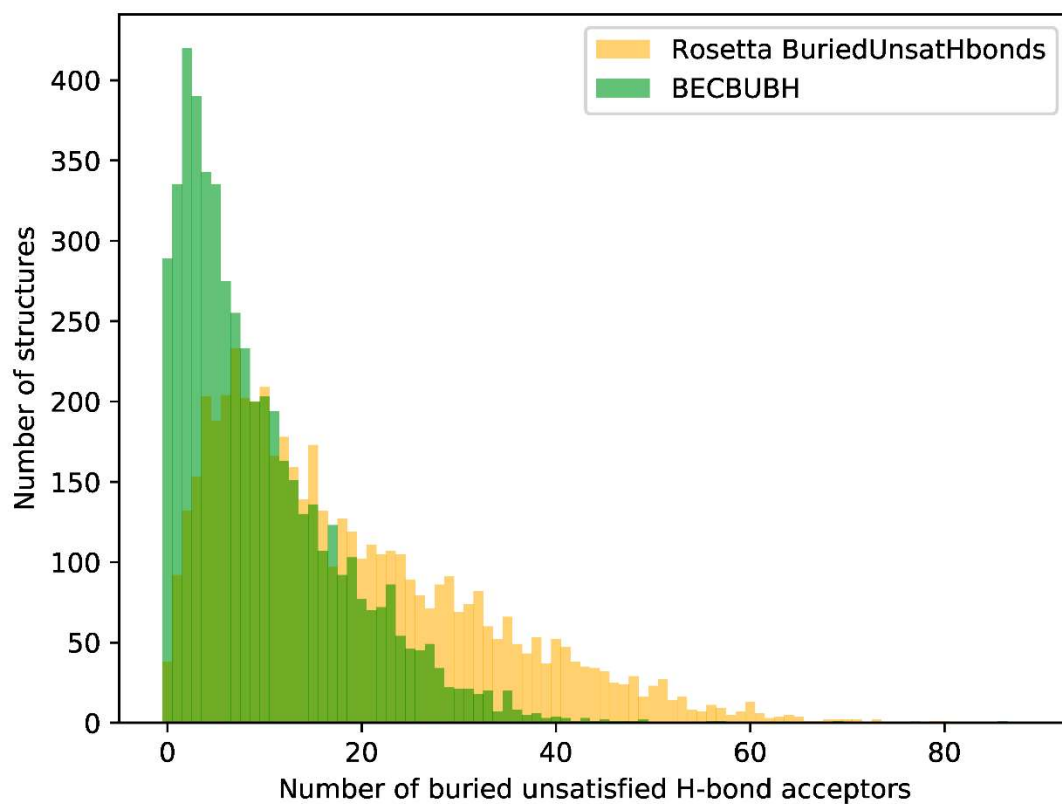


Figure 2.18 Number of buried unsatisfied hydrogen bond acceptors found in the top8000 protein structure data set.

The unsatisfied acceptors found by the Rosetta BuriedUnsathbonds filter are shown in yellow and the unsatisfied acceptors found by the BECBUBH filter are shown in green. The top8000 structure data set is downloaded from <http://kinemage.biochem.duke.edu/databases/top8000.php>.

Table 2.1 Loop libraries

Type	Residue number	Number of all loops	Number of non-redundant loops	Degeneracy
Helix-loop-helix	2	45273	224	202
	3	40194	765	52.5
	4	37295	1873	19.9
	5	35118	3656	9.61
Helix-loop-strand	2	6663	224	29.7
	3	15465	788	19.6
	4	24487	1953	12.5
	5	30061	3774	7.97
Strand-loop-helix	2	13263	349	38.0
	3	17819	1183	15.1
	4	21750	2974	7.31
	5	25684	5826	4.41

Table 2.2 Number of models at each design stage.

Design stage	RO1	RO2	NT
Backbones generated	13421	290836	136683
Sequences designed (iteration 1)	58626	432735	409101
Designs passed filters (iteration 1)	1163	9934	5715
Designs selected for <i>ab initio</i> folding	50	n/a	n/a
Sequences designed (iteration 2)		49578	22831
Designs passed filters (iteration 2)		722	98
Designs selected for <i>ab initio</i> folding		50	n/a
Designs selected for refinement (iteration 3)			10 (manual)
Sequences designed (iteration 3)			1000
Designs passed filters (iteration 3)			202
Designs selected for <i>ab initio</i> folding			50
Designs passed <i>ab initio</i> folding	22	25	10
Designs experimentally tested	10	25	10

Table 2.3 Protein sequences of designs selected for experimental testing

Design ID	Protein Sequence
RO1_1	KLVVVIDSNDKKLIEEAKKMAEKANLLLLYDVDEDQVRKAAGNARILVL VSNDEQLDKWKEWAQRLELDVRTRKVTSPDEAKRWIKEFSEE
RO1_2	KLFVLILSNDKKLIEEAKKMAEKANLELYVSSSEEDAKRILKELKDRNADS VLVLVSNDEQLDTAKEWAQRLELNVTRKVTSPDEAKRWIKEFSEE
RO1_3	TLVVVIISNDKKLIEEAKKMAEKANLQLYTDLPDQAVKLAKKLNADKVL VLVSNDEDLDKAKEAAQRAELDVRIRKVTSPDEAKRWIKEFSEE
RO1_4	KLVVIIISNDKKLIEEAKKMAEKANLELYVVDSEELKKLLKKIADENPNTK VLILVSNDEQLDLAKEIAQRLELDVRTRKVTSPDEAKRWIKEFSEE
RO1_5	TLVVIIDSNDKKLIEEAKKMAEKANLELYYERDIEDLLRKLKDADRILILVS NDEQLDKAKEIAQRLEVPRTRKVTSPDEAKRWIKEFSEE
RO1_6	KLSVIISNDKKLVEEAKKMAEKANLELYVVTDPDQAEKIIRKLIKEDPTIRI LVVLVSNDEALDWVKELAQKLEVDLRTRKVTSPDEAKRWIKEFSEE
RO1_7	TLVVIISNDKKLIEEAKKMAEKANLYLFEVTSDEDWKKAIKTAKEIAKKE QRPLRILVLVSNDEQLDKAKEIAQRQELDVRTRKVTSPDEAKRWIKEFSEE
RO1_8	TLVVIIDSNDKKLIEEAKKMAEKANLILIESSPDPEKTLRDLNADRVLVLVS NDEQLDTWKWEAQRWELPIRTRKVTSPDEAKRWIKEFSEE
RO1_9	TLLVIISNDKKLVEEAKKMAEKANLILINSPLSPEQLERTVKSVNADRVL LVSNDEQLDQAKETAQRAELPIRTRKVTSPDEAKRWIKEFSEE
RO1_10	TLVVIISNDKKLIEEAKKMAEKANLILLVVDNPPEALERAYRLNADKILVL VSNDEQLDWAKEAAQRWELPVRVRKVTSPDEAKRWIKEFSEE
RO2_1	RLVVLIVSNDKKLIEEARKMAEKANLELITVPGSPPEAIRLAQEIAEKAPGP VKVLVLITGSADPDEKTKAKKAAEEARKWNVRVRTVTSPDEAKRWIKEFS EE
RO2_2	QLYVIISNDKKLIEEARKMAEKANLNLLTADVDEAYELAKKLIDKAGSA KVLILITGSADPSQKKKIKELAEKARSLNVRIRIVTSPDEAKRWIKEFSEE
RO2_3	TLVVIIIVSKDKKLIEEARKMAEKANLLL VVYEPGEDEEAAKEASRRLKESL NNNQPAKVLVLISSSLSPSLAETAQQLAPDAEVRIRTVTSPDEAKRWIKEF SEE
RO2_4	TLVVFIVSKDKKLIEEARKMAEKANLKLYTAPVSPSIAEKVAKEAKKKNQP AKFLFLVDGTDPTAREIATKLAKYASTVANAENVIREVTSPDLAKRWIKEF SEE
RO2_5	GLLVIIIVSKDKKLIEEARKMAEKANLLLITAPTDPRELETAIKLLQKSNTPIK ILILSDGTDPTAEKIAKKLAKEAATKANAEYRIRKVTSPDQAKRWIKEFSEE

Table 2.3 Protein sequences of designs selected for experimental testing cont.

Design ID	Protein Sequence
RO2_6	GLVVIIIVSKDKKLIEEARKMAEKANLYLFTLEPNADPSQLDTRLRKWAQEIL KRDGPSKLKVLVLSGDGTDPTAQKLAKLIAKIVATAANAIEVRIRSVTSPDQA KRWIKEFSEE
RO2_7	TLIVVIVSKDKKLIEEARKMAEKANLLLFTGDLTNEQEKTAEAAARDGGS AKILILSDGTDPDARDKATKAACKLATKLNAEFRIREVTSPPDQAKRWIKF SEE
RO2_8	TLVVIIISDDKKLIEEARKMAEKANLILVTKSEIDDAIREIKKKAKDRPAKILI LSDGTNPEAEKIAKKIAEKIAKILNAEVRIRKVTSPDQAKRWIKFSEE
RO2_9	NLIVFIWSNDKKLIEEARKMAEKANLYLFTLGDNAEKVLQEAVEKVAGDN VKILVLIEDTKDADKLAKKLKEIADKKNWDIRIRKVTSPDEAKRWIKFSE E
RO2_10	RLIVIILSNDKKLIEEARKMAEKANLELITVRSDEIDIEKVLRKAGNAKVLILLI EDTKDADKLAKKAKEAADKLNVDLRIRKVTSPDEAKRWIKFSEE
RO2_11	SLFVIIFSNDKKLIEEARKMAEKANLILITVEGSPSAVQEAIKIAVEIARKQN AESIKILLVENTKDAEKVKKLAKEAADKLNVDIRIRTVTSPDEAKRWIKF FSEE
RO2_12	SLFVIIYSNDKKLIEEARKMAEKANLNLYTVSGDWREVKKLIEELIKRAKD KNPSEEVKVLVVLDPRATEAAKKLEKNAPPNVRIRTVTSPDEAKRWIKF SEE
RO2_13	ELIVLILSNDKKLIEEARKMAEKANLELYTLEGDDDEQIKKWIKKLAKTALS RNPSEAKILVLVEDTKDADKKIKIKKAADKANIEIRIRKVTSPDEAKRWIKF FSEE
RO2_14	QLFVIIVSNDKKLIEEARKMAEKANLELYTADLDTAVKIAKELLKKAEGPA KVLILVSGSASPDQKTKLDKIAKKLRSYNIRLREVTSPPDQAKRWIKFSEE
RO2_15	KLVLVILSNDKKLIEEARKMAEKANLELLTLDGSPEQLKKILKTLLDKAGD RPLKILVLIEDTKDADKWAKAIEAAKELNIDVRIRKVTSPDEAKRWIKFSE EE
RO2_16	TLIVIIISNDKKLIEEARKMAEKANLNLYTWDDEDKAKKALKDATKYENV KLLFLIENTKDAEKIEKKIKDTAKKLNLDVRVRLVTSPDEAKRWIKFSEE
RO2_17	TLIVVIWSNDKKLIEEARKMAEKANLLLLTVTSDEDLKKAAKIAQSAPGEV KVLVVIEDTKDADKIADKAKKIFKANVDIRIRKVTSPDEAKRWIKFSEE
RO2_18	TLVVLIWSNDKKLIEEARKMAEKANLYLITVGDDKALEKAIRTAEKIAKD NNADSFKILILIEDTKDADKISKKAKDIASKLNIEIRVRKVTSPDEAKRWIKF FSEE
RO2_19	TLIVLIISNDKKLIEEARKMAEKANLLLYTLEPNQDPSIEKEIKTIQKRADPR DLKILVLIENTKDAEKIATEIKRKA EKNNLVNVRIRLVTSPDEAKRWIKFSE E

Table 2.3 Protein sequences of designs selected for experimental testing cont.

Design ID	Protein Sequence
RO2_20	GLLVLIWSNDKKLIEEARKMAEKANLYLLTLETDDKKIEDILKSLGPPVKIL VLLEDTKDADKVKKEIEKKARKKNLPVRIRKVTSPDEAKRWIKFSEE
RO2_21	TLIVIIISNDKKLIEEARKMAEKANLVLTDEGSPSAEEKLKKTTITDAKRKDP TDPVKILVLIEDTKDADKIAEEIKRKADKANWDVRIRKVTSPDEAKRWIKF FSEE
RO2_22	TLVVLIIFSNDKKLIEEARKMAEKANLELYTRSELDPNIVTKLRDNAENAKL LVLIEDTKDADKLAEKIKKALDKNNIDVRIRKVTSPDEAKRWIKFSEE
RO2_23	SLVVFIIWSNDKKLIEEARKMAEKANLELITVSSIDQAIKLAREIAKKQKRPA KFLILVSGSLDPSQKKKVDEIAKEARKDNIRVRTVTSPDEAKRWIKFSEE
RO2_24	KLLVVILSNDKKLIEEARKMAEKANLELITVTSLEEAKKAAEKALKEANG NAKVLVLITGSADPTQKKKATEWAKKAKDYNIRVRTVTSPDEAKRWIKF SEE
RO2_25	TLFVLILSNDKKLIEEARKMAEKANLILITVGDEEELKKAIKKADDIAKKQN SSEAKILILLEKPVSPYEYKKLQKYADAEVRVRTVTSPDEAKRWIKFSEE
NT_1	SREEIRKVVETFLRAANSQDKKKLEEAAKNILSPDVRLEVGNVTWTSIEQM LKFYQLSEIDRVEIRKVQVDGNHVRVEIEVERNGKKWTWEVEVEVRNGLI KRIRNQVDPEYKKDQVNIWNNT
NT_2	SREEIRKVVEEFIRAQEDPDKLEKVASKALSPDVRVEIGNFTLEDKKQVIK WQKAFYKVLQEKAGKDAFRYEIRKVQVDGNHVRVEVEVETNGKKWTY EIELEVRNGKIKRIRLQVDPEYKEIVQLAWNRT
NT_3	SREEIRKVVETWVRLFNSGDPRDRKKYEKAQKELLSPDVRTEIGNYTIIEPG TLERFVQAYWKVLDELWPNVPIRVEIRKVQVDGNHVRVEVEVEINGKKYT FEIEVEVRNGKIKRIRIQRDPEMKELIQIAWNRT
NT_4	SREEIRKVVETYVRISLSSEETKKILRDLLSPDARLEFGNYTIESGDIWKFM QLFWKYYAGDAPLRLEIRKVQVDGNHVRVEVEVETKGKKWTYEIEVEVR NGKIKRVRTQVDPEYKKALQYAWNAT
NT_5	SREEIRKIVELFVKAWDNPDAREKFEKNKDKVLSPDVRLEIGNFTLENKDK LESFYRVLIKLWQEKAGPNVRIEIRKVQVDGNHVRVEVEVETNGKKWTYE IEVEVRNGTIKRIRTQYDPEYKKDIQQAWNLS
NT_6	SREEIRKIVETIVRANRDTSLFEKLAKELNLFSPDTRIEIGNYTFEGDAIKVIK AYIEANLQFAKKVSKDAPVRIEIRKVQVDGNHVRVEVEIELAGKKFTTEIE VEVRNGVVKRIRIQVDPEFKKLQYAWNKT
NT_7	SREEIRKVVEIFIRLQSLDPSQLEKALKDLNILSPDVRLEVGNITLNSADKLIR FLALITEILIRLWTGKPAPLRVEIRKVQVDGNHVRVEIEQEINGRKWTYEIEF EVRNGVIKRIRVQLDPSTKEAVQRAWNLT

Table 2.3 Protein sequences of designs selected for experimental testing cont.

Design ID	Protein Sequence
NT_8	SREEIRKVVETFVRAKQDPREFTKALSLLSPDVRMEIGNYTLTSIRDIKRFFE ALVEIWKRKNLTDWRYEIRKVQVDGNHVRVIEVETQTDGKKWTWEIEIEVR NGKIKRIREQYDPEYKKDVQLAWNLT
NT_9	SREEIRKVVEEYIRLLYTDPDQFKKAARDKLLSPDVRVIEIGNYTFDSRNLDLDR FLDAMQEWASRYDRVEIRKVQVDGNHVRVEIELESNGKKWTFEIEVEVRN GKIKRIRQQVDPEYKKVVQNLWNNT
NT_10	SREEIRKVVETWIRLFYSSDPNDWETFQKAKKDLLSPDVRVEIGNYTLNSE QVDRWWEAWVKIIQKELEEKNEPLRTEIRKVQVDGNHVRVEIEQEKNKG KWTFEVEVEVRNGKIKRIRQQVDPEYKKEVQAAWNNT

Table 2.4 DNA sequences of designs selected for experimental testing

Design ID	DNA Sequence
RO1_1	AAACTGGTGGTGGTGAATTGATAGCAACGATAAAAACTGATTGAAGAAGCGAAAAAATGGCGGAAAAA GCGAACCTGCTGCTGCTGTATGATGTGGATGAAGATCAGGTGCGTAAAGCGGCGGGCAACGCGCTATTC TGGTGTGGTGAAGCAACGATGAACAGCTGGATAAATGGAAGAATGGGCGCAGCGTCTGGAAGTGGATG TGCGTACCCGTAAAGTGACCAGCCCGGATGAAGCGAAACGTTGGATTAAAGAATTTAGCGAAGAATAATA AGGCAGCTAAGGCAGCTAAGGC
RO1_2	AAACTGTTTGTGCTGATTCTGAGCAACGATAAAAACTGATTGAAGAAGCGAAAAAATGGCGGAAAAAG CGAACCTGGAAGTGTATGTGAGCAGCAGCGAAGAAGATGCGAAACGTATTCTGAAAGAACTGAAAGATCG TAACGCGGATAGCGTGTGGTGTGGTGAAGCAACGATGAACAGCTGGATACCGCGAAAGAATGGGCGCA GCGTCTGGAAGTGAACGTGCGTACCCGTAAAGTGACCAGCCCGGATGAAGCGAAACGTTGGATTAAAGAA TTAGCGAAGAATAATAAGGC
RO1_3	ACCTGGTGGTGGTGAATTATTAGCAACGATAAAAACTGATTGAAGAAGCGAAAAAATGGCGGAAAAAG CGAACCTGCAGCTGTATACCGATCTGGATCCGGATCAGGCGGTGAAACTGGCGAAAAAATGAACGCGGA TAAAGTGTGGTGTGGTGAAGCAACGATGAAGATCTGGATAAAGCGAAAGAAGCGGCGCAGCGTGTGGGA ACTGGATGTGCGTATTCGTAAAGTGACCAGCCCGGATGAAGCGAAACGTTGGATTAAAGAATTTAGCGAA GAATAATAAGGCAGCTAAGGC
RO1_4	AAACTGGTGGTGAATTATTCTGAGCAACGATAAAAACTGATTGAAGAAGCGCAGAAAAATGGCGGAAAAAG CGAACCTGGAAGTGTATGTGGTGGATAGCGAAGAAGTGAAGAACTGCTGAAAAAATTTGCGGATGAAAA CCCGAACACCAAAGTGTGATTCTGGTGAAGCAACGATGAACAGCTGGATCTGGCGAAAGAAATTTGCGCAG CGTCTGGAAGTGGATGTGCGTACCCGTAAAGTGACCAGCCCGGATGAAGCGAAACGTTGGATTAAAGAAT TTAGCGAAGAATAATAAGGC
RO1_5	ACCTGGTGGTGAATTATTGATAGCAACGATAAAAACTGATTGAAGAAGCGAAAAAATGGCGGAAAAAG CGAACCTGGAAGTGTATTATGAACGTGATATTGAAGATCTGCTGCGTAAACTGAAAGATGCGGATCGTATT CTGATTCTGGTGAAGCAACGATGAACAGCTGGATAAAGCGAAAGAAATTTGCGCAGCGTCTGGAAGTGCCGG TGCGTACCCGTAAAGTGACCAGCCCGGATGAAGCGAAACGTTGGATTAAAGAATTTAGCGAAGAATAATA AGGCAGCTAAGGCAGCTAA
RO1_6	AAACTGAGCGTGATTATTCTGAGCAACGATAAAAACTGGTGAAGAAGCGAAAAAATGGCGGAAAAA GCGAACCTGGAAGTGTATGTGGTGACCGATCCGGATCAGGCGGAAAAAATTTCTGTAAGTGAATTAAAGA AGATCCGACCATTCGTATTCTGGTGTGGTGAAGCAACGATGAAGCGCTGGATTGGGTGAAAGAACTGGCG CAGAACTGGAAGTGGATCTGCGTACCCGTAAAGTGACCAGCCCGGATGAAGCGAAACGTTGGATTAAAG AATTTAGCGAAGAATAATAA
RO1_7	ACCTGGTGGTGAATTATTCTGAGCAACGATAAAAACTGATTGAAGAAGCGAAAAAATGGCGGAAAAAG CGAACCTGTATCTGTTGAAGTGACCAGCGATGAAGATTGAAAAAAGCGATTAAACCGCGAAAGAAAT TGCGAAAAAAGAACAGCGTCCGCTGCGTATTCTGGTGTGGTGAAGCAACGATGAACAGCTGGATAAAGCG AAAGAAATTTGCGCAGCGTCAGGAAGTGGATGTGCGTACCCGTAAAGTGACCAGCCCGGATGAAGCGAAAC GTTGGATTAAAGAATTTAGCGAAGAATAA
RO1_8	ACCTGGTGGTGAATTATTGATAGCAACGATAAAAACTGATTGAAGAAGCGAAAAAATGGCGGAAAAAG CGAACCTGATTCTGATTGAAAGCAGCCCGGATCCGAAAAAACCTGCGTGATCTGAACGCGGATCGTGTG CTGGTGTGGTGAAGCAACGATGAACAGCTGGATACCTGGAAGAATGGGCGCAGCGTTGGGAAGTCCCG ATTCGTACCCGTAAAGTGACCAGCCCGGATGAAGCGAAACGTTGGATTAAAGAATTTAGCGAAGAATAATA AGGCAGCTAAGGCAGCTAA

Table 2.4 DNA sequences of designs selected for experimental testing cont.

Design ID	DNA Sequence
RO1_9	ACCCTGCTGGTGATTATTCTGAGCAACGATAAAAACTGGTGGAAGAAGCGAAAAAATGGCGGAAAAAG CGAACCTGATTCTGATTAACAGCCCGCTGAGCCCGGAACAGCTGGAACGTACCGTGAAAAGCGTGAACGC GGATCGTGTGCTGATTCTGGTGAGCAACGATGAACAGCTGGATCAGGCGAAAGAAACCGCGCAGCGTGCG GAACTGCCGATTCTACCCGTAAAGTGACCAGCCCGGATGAAGCGAAACGTTGGATTAAAGAATTTAGCG AAGAATAATAAGGCAGCTAA
RO1_10	ACCCTGGTGGTGATTATTATTAGCAACGATAAAAACTGATTGAAGAAGCGAAAAAATGGCGGAAAAAG CGAACCTGATTCTGCTGGTGGTGATAACCCGGAAGAAGCGCTGGAACGTGCGTATCGTCTGAACGCGGA TAAATTTCTGGTGCTGGTGAGCAACGATGAACAGCTGGATTGGGCGAAAGAAGCGGCGCAGCGTTGGGA ACTGCCGGTGCGTGTGCGTAAAGTGACCAGCCCGGATGAAGCGAAACGTTGGATTAAAGAATTTAGCGAA GAATAATAAGGCAGCTAAGGC
RO2_1	CGCCTTGTTGTATTGATCGTAAGTAATGACAAGAAGTTGATCGAAGAGGCCCGCAAGATGGCTGAGAAGG CTAATTTGGAGTTGATCACGGTTCAGGTAGTCTGAGGAAGCCATCCGCTTGGCTCAAGAGATCGCCGAG AAGGCTCTGGGCCCGTTAAGGTATTGGTCTTAATCACAGTTTCAGCCGACCCGACGAGAAGACGAAGG CCAAGAAGGCAGCAGAGGAAGCTCGCAAGTGGAATGTCCGCGTTCGCACGTTACATCTCTGACGAGGC AAAGCGCTGGATCAAGGAGTTCTCAGAGGAGTGA
RO2_2	CAGCTGTATGTGATTATTTCCAGTAACGATAAGAACTGATTGAAGAGGCGCGTAAAATGGCGGAAAAAGG CAAACCTGAACCTGCTTACAGCCGATGTGGATGAAGCGTATGAACTGGCGAAGAAATTGATTGATAAAGCA GGGAGCGCGAAAGTACTTATTCTGATTACCGGCAGTGCGGATCCCTCTCAGAAGAAGAAAATTAAGAAGT CGCGGAAAAGGCACGTAGCTTAAACGTACGTATTCTGATTGTTACCAGCCCGGATGAAGCGAAACGTTGGA TTAAAGAATTTTCGGAAGAATAA
RO2_3	ACACTCGTTGTCATCATCGTATCGAAGGACAAGAAGCTTATCGAAGAGGCCCGCAAGATGGCTGAGAAGG CTAATTTGCTCTTGGTCGTTTACGAGCCAGGTGAGGACGAAGAGGCAGCCAAAGAGGCCAGTCGCCGCTTA AAGGAGTCTCTCAATAATAATCAACCTGCAAAGGTCTTGGTTCTCATCTCAAGTTCCCTCTCACCTAGTTAG CCGAGACGGCAGCCAAGCAATTGGCCCCAGACGCTGAGGTACGCATCCGCACGGTAACATCTCCAGACGA GGCAAAGCGCTGGATCAAGGAGTTCTCTGAGGAGTGA
RO2_4	ACTTTAGTCGTTTTTCATCGTTAGTAAGGACAAGAAGCTCATCGAGGAAGCCCGCAAGATGGCCGAGAAGGC CAATCTTAAGCTCTACACAGCTCCAGTTTCCCCTCTATCGCCGAGAAGGTCGCTAAAGAGGCCAAAGAAGA AGAATCAACCAGCAAAGTTCCTCTTCTCGTTGACGGTACGGACCCACAGCACGCGAGATCGCCACAAAG TTAGCCAAGTACGCATCAACTGTCGCAATGCCGAGGTCCGCATCCGCGAGGTTACTTCTCCTGACCTTGCA AAGCGCTGGATCAAGGAGTTCTCAGAGGAGTGA
RO2_5	GGTCTCCTCGTAATCATCGTTAGCAAGGACAAGAAGTTAATTGAAGAAGCTCGCAAGATGGCCGAGAAGG CAAATTTATTGTTAATCACAGCACCTACGGACCCACGCGAGTTGGAGACTGCAATCAAGTTACTTCAAAAGT CAAATACGCCAATCAAGATCTTAATCTTATCAGACGGAACGGACCCACGGCTGAGAAGATCGCAAAGAAG TTGGCCAAGGAAGCAGCAACGAAGGCCAATGCCGAGTACCGCATCCGCAAGGTTACTTACCCGACCAAG CAAAGCGCTGGATCAAGGAGTTCTCTGAGGAGTGA
RO2_6	GGGCTCGTAGTCATCATCGTTTCTAAGGACAAGAAGTTAATCGAAGAGGCCCGCAAGATGGCCGAGAAGG CCAATTTATACTTATTCAGTTAGAGCCAAATGCCGACCAAGTCAATTGGACACGCTCCGCAAGTGGGCC AAGAGATCCTTAAGCGGACGGGCCCTCAAAGCTCAAGGTATTGGTTCTCTCGGACGGTACGGACCCACA GCACAAAAGTTGGCCAAGCTTATCGCAAAGATCGTAGCAACGGCAGCCAATGCCGAGGTTTCGCATCCGCTC CGTAACTTCACCTGACCAAGCCAAGCGCTGGATCAAGGAGTTCTCCGAGGAGTGA

Table 2.4 DNA sequences of designs selected for experimental testing cont.

Design ID	DNA Sequence
RO2_7	ACTCTCATCGTCGTTATCGTATCTAAGGACAAGAAGTTAATCGAGGAAGCCCGCAAGATGGCCGAGAAGGC CAATTTACTCTTATTCAGTGGAGACCTCACGAATGAGCAAGAGAAGACAGCAAAAGAGGCTGCCGACCGCG ACGGATCAGCTAAGATCCTTATCCTTAGTGACGGTACGGACCCTGACGCTCGCGACAAGGCTACGAAGGCC GCCAAGAAGCTCGCAACAAAGCTTAATGCTGAGTTCGCGATCCGCGAGGTAACGTACCTGACCAAGCCAA GCGCTGGATCAAGGAGTTCAGTGAGGAGTGA
RO2_8	ACATTGGTTGTTATCATCATCTCAGACGACAAGAAGTTGATCGAGGAAGCACGCAAGATGGCTGAGAAGG CAATTTAATCTTGGTTACGAAGAGTGAGATCGACGACGCAATCCGCGAGATCAAGAAGAAGGCCAAGGA CCGCCCTGCCAAGATCTTAATCTTAAGTGACGGGACGAATCCAGAGGCAGAGAAGATCGCTAAGAAGATC GCAGAGAAGATCGCCAAGATCCTCAATGCCGAGGTACGCATCCGCAAGGTAACGAGTCTGACCAAGCAA AGCGCTGGATCAAGGAGTTCAGTGAGGAGTGA
RO2_9	AATTTGATCGTTTTATCTGGTCCAATGACAAGAAGCTTATCGAAGAGGCCCGCAAGATGGCAGAGAAGGC AAATTTGTACCTTTCACGTTGGGAGACAATGCTGAGAAGGTTTTACAAGAGGCCGTTGAGAAGGTTGCCG GTGACAATGTAAAGATCTTGGTTTTGATCGAGGACACGAAGGACGCTGACAAGCTTGCAAAGAAGTTAAA GGAGATCGCAGACAAGAAGAATTGGGACATCCGCATCCGCAAGGTAACCTCGCCAGACGAGGCTAAGCGC TGGATCAAGGAGTTCCTGAGGAGTGA
RO2_10	CGCTTGATCGTAATCATTCTGTCCAATGACAAGAAGCTTATCGAAGAGGCCCGCAAGATGGCCGAGAAGGC AAATCTTGAGCTTATCACAGTTCGAGTGACGAGGACATCGAGAAGGTATTACGCAAGGCCGGGAATGCA AAGTCTTTTTGCTTATCGAGGACACGAAGGACGCCGACAAGTTGGCCAAGAAGGCCAAGGAAGCCGCCG ACAAGTTGAATGTAGACCTTCGCATCCGCAAGGTAACGTCTCCTGACGAGGCTAAGCGCTGGATCAAGGAG TTCAGTGAGGAGTGATAA
RO2_11	TCGTTGTCGTCATCATCTTCTCAAATGACAAGAAGCTTATCGAGGAAGCCCGCAAGATGGCAGAGAAGGC CAATTTGATCTTGATCACAGTCGAGGGTCTCCTTCTGCAGTCCAAGAGGCAATCAAGATCGCCGTTGAGAT CGCACGCAAGCAAAATGCAGAGTCGATCAAGATCTTGCTTTTAGTCGAGAATACAAAGGACGCAGAGAAG GTAAAGAAGCTCGCCAAAGAGGCCGCCGACAAGTTAAATGTTGACATCCGCATCCGCACTGTTACGTCTCC AGACGAGGCTAAGCGCTGGATCAAGGAGTTCCTCGAGGAGTGA
RO2_12	AGTCTTTTCGTTATCATCTACTCGAATGACAAGAAGCTCATCGAGGAAGCACGCAAGATGGCAGAGAAGGC TAATCTTAATTTATACACGTTTCGGGAGACTGGCGCGAGGTTAAGAAGCTTATCGAGGAGTTGATCAAGC GCGCCAAGGACAAGAATCCATCTGAGGAAGTTAAGGTCTTACTTTTAGTCAAGGACCCTCGCGCCACTGAG GCTGCCAAGAAGTTAGAGAAGAATGCACCCCAATGTTTCGCATCCGCACTGTTACGTCTCCCGACGAGGC CAAGCGCTGGATCAAGGAGTTCAGAGGAGTGA
RO2_13	GAGTTGATCGTTTTGATCCTTTTGAATGACAAGAAGCTCATCGAGGAAGCACGCAAGATGGCAGAGAAGG CCAATCTCGAATTGTACACGTTGGAAGGTGACGACGAGCAATCAAGAAGTGGATCAAGAAGCTCGCCAA GACTGCCTTGCTCGCAATCCAGTGAGGCAAAGATCTTAGTCCTTGTTGAGGACACTAAGGACGCAGACA AGAAGATCAAGATCATCAAGAAGGCCGCCGACGAGGCAAATATCGAGATCCGCATCCGCAAGGTTACATC ACCCGACGAGGCAAAGCGCTGGATCAAGGAGTTCAGAGGAGTGA
RO2_14	CAACTCTTCGTCATCATCGTATCAAATGACAAGAAGCTTATTGAAGAGGCACGCAAGATGGCAGAGAAGGC CAATCTCGAATTTACACGGCAGACTTAGACACTGCAGTAAAGATCGCCAAGGAGTTGTTGAAGAAGGCCG AGGGGCCAGCTAAGGTCCTCATCTTAGTCTCAGGGTCAGCATCGCCAGACCAAAAGACAAAGTTAGACAA GATCGCCAAGAAGCTTCGCTCGTACAATATCCGCTTACGCGAGGTTACGTCTCCAGACGAGGCAAAGCGCT GGATCAAGGAGTTCAGCGAGGAGTGA

Table 2.4 DNA sequences of designs selected for experimental testing cont.

Design ID	DNA Sequence
RO2_15	AAGCTGGTCGTCTTGATCTTGAGCAATGACAAGAAGTTGATTGAAGAGGCCCGCAAGATGGCCGAGAAGG CAAATTTGGAGTTGCTTACGTTAGACGGTTCACCTGAGCAACTCAAGAAGATCCTTAAGACGCTTTTAGACA AGGCCGGAGACCGCCATTGAAGATCTTGGTCTTGATCGAGGACACGAAGGACGCCGACAAGTGGGCCAA GGCAATCAAAGAGGCCGCCAAGGAGCTCAATATCGACGTCCGCATCCGCAAGGTCACATCCCCTGACGAG GCCAAGCGCTGGATCAAGGAGTTCTCGGAAGAGTGA
RO2_16	ACACTTATCGTCATCATCTCGAATGACAAGAAGCTCATCGAGGAAGCCCGCAAGATGGCCGAGAAGGC CAATTTAAATCTCTACACGTGGGACGACGAGGACAAGGCTAAGAAGGCATTAAAGGACGCAACTAAGTAC GAGAATGTTAAGTTATTCTCATCGAGAATACTAAGGACGCCGAGAAGATCGAGAAGAAGATCAAGG ACACGGCAAAGAAGCTCAATTTAGACGTCCGCGTTCGTTGGTTACGTGCGCAGACGAGGCCAAGCGCTGG ATCAAGGAGTTCTCAGAGGAGTGA
RO2_17	ACATTAATCGTCGTAATCTGGTCCAATGACAAGAAGTTAATCGAAGAGGCACGCAAGATGGCTGAGAAGG CAAATCTCTCTTATTGACAGTTACTTCCGACGAGGACTTGAAGAAGGCAGCAAAGATCGCACAAAGTGCC CCAGGAGAGGTAAGGTCCTTCTCTTGTGAGGACACAAAGGACGCTGACAAGATCGCCGACAAGGCCA AGAAGATCTTCAAGAAGGCCAATGTAGACATCCGCATCCGCAAGGTTACGTGCGCTGACGAGGCCAAGCG CTGGATCAAGGAGTTCTCCGAGGAGTGA
RO2_18	ACACTTGTAGTCTTAATCTGGTCAATGACAAGAAGCTTATCGAAGAGGCACGCAAGATGGCCGAGAAGG CCAATCTCTACCTTATCACGGTCGGTGACGACAAGGCCTTGGAGAAGGCTATCCGCACGGCAGAGAAGATC GCAAAGGACAATAATGCCGACTCTTTCAAGATCTTAATCCTTATCGAGGACACAAAGGACGCCGACAAGAT CAGTAAGAAGGCAAAGGACATCGCCAGCAAGCTCAATATCGAGATCCGCGTTCGCAAGGTAACATCTCTTG ACGAGGCTAAGCGCTGGATCAAGGAGTTCTCTGAGGAGTGA
RO2_19	ACTTTAATCGTTTTGATCATCTCGAATGACAAGAAGTTGATCGAGGAAGCCCGCAAGATGGCCGAGAAGGC TAATTTATTACTTTACACTCTTGAGCCTAATCAAGACCCATCCATCGAGAAGGAGATCAAGACGATCCAAAA GCGCGTGACCCACGCGACTTAAAGATCCTTGTCTTATCGAGAATACTAAGGACGACAGAGAAGATCGCCA CGGAGATCAAGCGCAAGGCAGAGAAGAATAATTTAAATGTACGCATCCGCTTGTACGTGCCCCGACGA GGCCAAGCGCTGGATCAAGGAGTTCAAGTGAAGGAGTGA
RO2_20	GGATTATTGGTCCTCATTTGGTCAATGACAAGAAGCTCATCGAGGAAGCTCGCAAGATGGCTGAGAAGG CCAATCTCTACCTCTTGACGCTCGAACTGACGACAAGAAGATCGAGGACATCTTAAAGTCGCTCGGGCCG CCCGTTAAGATCCTCGTTCTCTTAGAGGACACAAAGGACGCCGACAAGGTCAAGAAGGAGATCGAGAAGA AGGCCCGCAAGAAGAATTTACCCGTACGCATCCGCAAGGTAACCTTCGCCAGACGAGGCCAAGCGCTGGAT CAAGGAGTTCAGTGAGGAGTGA
RO2_21	ACTTTGATTGTATCATCATCAGTAACGACAAGAAGTTGATCGAGGAAGCTCGCAAGATGGCTGAGAAGGC CAATCTCGTCTTGATCACTGACGAGGGAAGTCCAGTGACAGAGGAGAAGCTCAAGAAGACAATCACGGAC GCCAAGCGCAAGGACCCAACGGACCCAGTAAAGATCTTAGTATTAATCGAGGACACTAAGGACGCCGACA AGATCGCCGAGGAGATCAAGCGCAAGGCTGACAAGGCAAATTGGGACGTCCGCATCCGCAAGGTAACATC CCCAGACGAGGCCAAGCGCTGGATCAAGGAGTTCTCCGAGGAGTGA
RO2_22	ACTTTGGTTGTTTTGATCTTCTCGAATGACAAGAAGTTGATCGAAGAGGCACGCAAGATGGCAGAGAAGGC AAATCTCGAACTTTACACTCGAGTGAGTTGGACCCAAATATCGTAACAAAGCTCCGCGACAATGCAGAGA ATGCCAAGCTTCTTGATTGATCGAGGACACGAAGGACGACAGACAAGCTCGCCGAGAAGATCAAGAAGGC CCTCGACAAGAATAATATCGACGTACGCATCCGCAAGGTAACGTGCGCTGACGAGGCCAAGCGCTGGATCA AGGAGTTCTCTGAGGAGTGA

Table 2.4 DNA sequences of designs selected for experimental testing cont.

Design ID	DNA Sequence
RO2_23	TCATTAGTTGCTTCATCTGGTCCAATGACAAGAAGTTGATCGAGGAAGCTCGCAAGATGGCAGAGAAGGC TAATTTGGAGTTGATCACGGTCAGTTCAATCGACCAAGCAATCAAGTTGGCCCGCAGATCGCAAAGAAGC AAAAGCGCCCTGCCAAGTTCTTGATCTTGGTCTCAGGGAGTTTAGACCCCTCTCAAAAGAAGAAGGTTGAC GAGATCGCTAAAGAGGCCCCGCAAGGACAATATCCGCGTTCGCACAGTCACTTCGCCTGACGAGGCCAAGC GCTGGATCAAGGAGTTCAGTGAGGAGTGA
RO2_24	AAGTTGCTCGTTGTTATCCTTAGTAATGACAAGAAGTTGATCGAGGAAGCACGCAAGATGGCAGAGAAGG CAATTTAGAGTTAATCACGGTTACATCGCTTGAAGAGGCCAAGAAGGCCGCGAGAGAAGGCCTTGAAAGA GGCAAATGGTAATGCCAAGGTAATCGTTTTAATCACTGGGTCCGCCGACCCAACGCAAAAGAAGAAGGCCA ACTGAGTGGGCAAAGAAGGCTAAGGACTACAATATCCGCGTACGCACAGTAACATCTCCAGACGAGGCCA AGCGCTGGATCAAGGAGTTCAGCGAGGAGTGA
RO2_25	ACGCTTTTCGTACTCATCTTATCAAATGACAAGAAGCTTATCGAGGAAGCCCGCAAGATGGCAGAGAAGGC CAATTTGATCCTCATCACGGTCGGGGACGAAGAGGAGTTAAAGAAGGCCATCAAGAAGGCCGACGACATC GCTAAGAAGCAAAATTCGTACAGAGGCCAAGATCCTCATCTTGCTTGAGAAGCCAGTCTCGCCTGAGTACGA GAAGAAGTTACAAAAGTACGCAGACGCAGAGGTTGCGGTTGCGACAGTTACGTCACCAGACGAGGCCAAG CGCTGGATCAAGGAGTTCCTGAGGAGTGA
NT_1	TCGCGCGAGGAGATCCGTAAAGTAGTCGAAACCTTTCTGCGTGCTGCGAATAGCCAGGACAAAAAAGAAC TCGAAGAGGCGGCGAAAAATATTCTGTACCTGATGTTCTGCTGGAAGTCGGCAACTATACCTGGACCAGC ATTGAACAAATGCTTAAGTTTTATCAACTGTCTGAGATTGATCGCGTCGAAATTCGCAAAGTGCAGGTCGAT GGCAACCACGTGCGTGTGGAAATTGAAGTAGAACGTAATGGCAAAAAGTGGACCTGGGAAGTTGAGGTG GAAGTACGTAATGGTTTGATTAAACGCATTGTAATCAGGTCGATCCGGAATATAAAAAAGATGTGCAAAA TATCTGGAATAATACCTAA
NT_2	AGCCGCGAAGAAATCCGTAAAGTAGTGGAAGAATTTATCCGTGCGCAAGAAGATCCCGACAAACTCGAGA AAGTAGCGTCAAAGGCGCTGTCCCGGACGTCCGCGTCGAAATTGGCAATTCACATTGGAAGATAAGAA GCAGGTCATCAAATGGCAAAAGGCCTTTTATAAAGTACTGCAGGAAAAAGCGGGGAAGGATGCCTCGTTT CGCTATGAAATTCGCAAAGTACAAGTAGACGGTAATCATGTTCCGGTTCGAAGTAGAAGTGGAAACGAACG GGAAAAAATGGACCTATGAAATCGAACTGGAGGTCCGTAATGGTAAAAATCAAACGTATTGCGCTGCAAGT GGATCCGGAATATAAAGAAATCGTGCAACTGGCATGGAATCGTACATAA
NT_3	TCTCGCGAGGAAATCCGCAAAGTCGTGAAACGTGGGTTTCGCTTATTCAACAGCGGCGACCCGCGCGACCG CAAAAAATATGAAAAAGCTCAGAAGGAGCTGCTGTCCCTGACGTTGCGACCGAAATTGGAAATTATACGA TCGAACCCGGAACCTGGAGCGTTTTGTGCAGGCATACTGGAAGTACTTGATGAGCTTTGGCCGAACGTC CCTATCCGCGTAGAGATCCGCAAAGTGAAGTGGATGGCAACCATGTGCGTATTGAGGTGGAAGTAGAAA TCAACGGTAAAAAATATACGTTTGAAATTGAAGTCGAGGTGCGTAATGGAAAGATCAAGCGTATTGCTATT CAGCGCGATCCCGAAATGAAGGAGCTGATCCAGATCGCGTGGAAACCGTACCTGA
NT_4	AGTCGTGAAGAAATCCGCAAGGTTGTTGAAACGTACGTTGCGATCTCACTGAGCTCTAGCGAAGAAACCAA AAAAATTCTGCGTGACTTGCTGTGCGCTGACGCACGCTTAGAATTTGGTAACTATACGATCGAATCCGGTGA TATTTGGAAATTCATGCAGTTGTTTTGGAATACTACGCCGGCGACGCACCGCTCCGCTGGAAATCCGTAA AGTGCAGGTTGATGGTAATCATGTTCTGATCGAAGTCGAAGTCGAGACCAAGGGAAAAAATGGACCTAT GAAATTGAAGTGGAAGTCCGCAACGGTAAAAATTAACGTGTTGCGACCCAAGTTGATCCGGAATATAAGAA AGCATTGCAATATGCCTGGAACGCAACCTAA

Table 2.4 DNA sequences of designs selected for experimental testing cont.

Design ID	DNA Sequence
NT_5	TCTCGTGAAGAGATTTCGCAAAATCGTGGAGCTCTTTGTAAAAGCATGGGATAATCCTGACGCTCGCGAAAA GTTTGAAAAGAATAAAGACAAGGTATTATCTCCCGATGTCCGTCTGGAAATTGGGAATTTTACCCTGGAAA ATAAAGATAAACTTGAATCTTTCTACCGTGTCTGATCAAACCTTTGGCAGGAGAAAAGCTGGTCCCAACGTCC GTATCGAGATCCGCAAGGTGCAAGTCGATGGTAACCATGTGCGTGTGCAAGTTGAAGTGGAACGAATGG CAAAAAATGGACCTACGAAATTGAAGTAGAAGTTCGTAACGGTACGATTAAACGCATCCGCACTCAGTATG ACCCTGAATATAAAAAGGACATCCAGCAGGCTTGGAACTCTCATAA
NT_6	AGCCGGGAAGAAATCCGCAAAATTGTGGAACGATCGTACGTGCGAATCGCGACACGAGTCTCTTTGAGA AGTTAGCCAAAGAAGTGAATCTTTAGTCCGGACACCCGATTGAAATTGGCAATTATACTTTTGAAGGGG ATGCGATCAAAGTTATCAAAGCGTACATTGAAGTAATCTGCAGTTCGCGAAAAAGGTTTCGAAAAGACGCG CCGGTGC GTATTGAAATCCGTAAAGTGCAAGTGGATGGTAACCACGTTCTGTGAGGTTGAAATCGAACT GGCTGGAAAGAAATTTACAACAGAGATCGAGGTTGAAGTGCGCAACGGGGTGGTTAAGCGCATTCTGATT CAAGTCGACCCCGAGTTCAAAAACTGGTTCACTACGCTTGGAAATAAACGTAA
NT_7	AGCCGCGAGGAAATTCGGAAGGTGGTAGAAATTTTCATTCTCTTCAAAGCCTGGACCCGAGCCAGTTAGA AAAAGCTTTAAAAGATTTGAACATTCTTTCGCCGGATGTGCGCTGGAAGTCGGAATATCACGCTGAACT CCGCGGATAAACTCATTCTGTTTCTTAGCGCTCATTACGGAGATCCTGATTGCGCTGTGGACGGGCAAACCCG CGCTCTGCGTGTAGAAATCCGTAAAGTGCAAGTAGATGGTAATCACGTGCGCGTGGAGATTGAACAAGA AATTAACGGGCGTAAGTGACGTACGAAATTGAATTTGAAGTACGCAATGGCGTTATTAAACGTATTCTGTG TGCAGTGGACCCATCCACCAAGAGGCTGTCCAGCGTGCATGGAACCTTACCTAA
NT_8	AGTCGCGAGGAAATTCGTAAAGTTGTGGAGACCTTCGTCCGTGCAAAGCAGGATCCGCGTGAATTCACCAA AGCGTTGTCCCTGCTCAGCCCTGATGTTCCGGATGGAAATCGGCAACTACACTCTGACCTCCATCCGGGACAT TAAACGCTTTTTTCGAGGCCTTAGTAGAAATTTGGAACGTAAAACTTGACTGACTGGCGTTACGAAATTCG CAAGGTGCAGGTCGATGGCAATCACGTCCGCATTGAGGTGGAACACAGACCGACGGCAAAAAATGGACT TGGGAAATTGAAATCGAAGTACGCAATGGGAAGATCAAACGTATCCGCGAGCAATATGACCCTGAATACA AAAAGGATGTACAGCTGGCATGGAACCTGACTTGA
NT_9	AGCCGCGAAGAGATCCGCAAAGTTGTGAGGAGTATATCCGTCTGCTGTATACAGACCCTGATCAGTTCAA GAAAGCGGCCCGGATAAATTGCTGAGTCCGGATGTGCGTATTGAAATCGGTAATTATACGTTTGATTCCC GCAACCTCGATCGCTTTCTGGACGCAATGCAGGAATGGGCGAGCCGTTACGATCGTGTGGAATTCGCAAG GTTCAAGTTGACGGGAACCATGTGCGTGTGAGATTGAGTTGGAATCGAACGGTAAAAATGGACGTTTG AAATCGAGGTTGAGGTTGCAATGGCAAAATTAAGCGCATCCGTCAACAGGTGGATCCGGAATATAAAAA AGTTGTGCAGAATCTGTGGAATAACACGTGA
NT_10	TCCCGGGAAGAGATTTCGGAAGGTCGTGGAGACTTGGATCCGCCTTTTTTATAGCTCGGATCCGAACGACTG GGAAACGTTCCAGAAAGCGAAGAAAGATCTGCTCTACCAGATGTGCGCGTAGAAATCGGAAATTACACG CTTAATAGTGAACAAGTCGATCGTTGGTGGGAAGCCTGGGTGAAAAATCATCCAGAAAGAATTGGAAGAGA AAAACGAACCGCTCCGCACGGAAATTCGCAAGGTTCAAGTGACGGCAATCACGTACGCGTGGAAATCGA GCAGGAAAAAACGGCAAGAAGTGACCTTTGAAGTTGAAGTGGAAGTTCGTAATGGAAAAATCAAGCGT ATTCGTCAACAGGTAGACCCAGAGTACAAAAAGGAGGTCCAGGCGGCTTGGAAACAACACCTAA

Table 2.5 Experimental characterization of designs

Design ID	Soluble	Relative dimer peak size in SEC*	Alpha-beta protein CD spectrum	Well resolved 1D-NMR	¹⁵ N HSQC with good dispersion
RO1_1	Y	no dimer peak	Y	Y	
RO1_2	Y	no dimer peak	Y	Y	
RO1_3	N				
RO1_4	N				
RO1_5	Y	0.11	Y	Y	
RO1_6	Y				
RO1_7	N				
RO1_8	Y	no dimer peak	Y	Y	
RO1_9	Y	<0.05	Y	Y	
RO1_10	N				
RO2_1	Y	0.06	Y	Y	Y
RO2_2	Y				
RO2_3	N				
RO2_4	N				
RO2_5	Y	no dimer peak	Y	Y	
RO2_6	Y	no dimer peak	Y	Y	
RO2_7	N				
RO2_8	N				
RO2_9	Y	no dimer peak	Y	Y	
RO2_10	Y	no dimer peak	Y	Y	
RO2_11	N				
RO2_12	Y				
RO2_13	Y				
RO2_14	Y				
RO2_15	Y	0.15	Y	Y	
RO2_16	N				
RO2_17	N				
RO2_18	N				
RO2_19	N				

Table 2.5 Experimental characterization of designs cont.

Design ID	Soluble	Relative dimer peak size in SEC*	Alpha-beta protein CD spectrum	Well resolved 1D-NMR	¹⁵ N HSQC with good dispersion
RO2_20	Y	no dimer peak	Y	Y	Y
RO2_21	N				
RO2_22	N				
RO2_23	Y				
RO2_24	Y				
RO2_25	Y	no dimer peak	Y	Y	Y
NT_1	Y	0.09	Y	Y	
NT_2	N				
NT_3	N				
NT_4	N				
NT_5	N				
NT_6	N				
NT_7	N				
NT_8	Y	<0.05	Y	Y	
NT_9	Y	<0.05	Y	Y	
NT_10	Y	<0.05	Y	Y	

Table 2.6 NMR statistics

	RO2_1	RO2_20	RO2_25
Number of residues	105	100	104
Distance restraints			
Total NOE	1288	1542	2326
intra-residue [$i = j$]	385	455	577
sequential [$ i - j = 1$]	375	426	595
medium range [$1 < i - j < 5$]	203	305	492
long range [$ i - j \geq 5$]	325	356	662
Hydrogen bonds	36	96	36
Dihedral	226	222	230
Violations			
Distance constraints (Å)	0.008±0.001	0.02±0.002	0.014±0.0009
Dihedral angle constraints (°)	1.1±0.7	1.3±0.1	2.2±0.17
Max. distance constraint violation (Å)	0.38	0.37	0.46
Max. dihedral angle restraint violation (°)	14.4	14.2	15.8
Num. Distance violations > 0.3 Å	0.1±0.4	0.6±0.7	1.1±0.7
Num dihedral violations between 5-10°	2.3±2	3.6±1	5.6±2
Num dihedral violations > 10°	0.85±1	0.25±0.5	3.8±1
Validate peaks vs structures (PyRPF)			
Recall	0.985	0.95	0.944
Precision	0.746	0.687	0.793
F-score	0.849	0.798	0.862
DP-score	0.87	0.855	0.9
Structure validation (PSVS)			
Deviations from idealized geometry			
Bond lengths (Å)	0.004	0.004	0.005
Bond angles (°)	0.6	0.7	0.8
Ramachandran plot			
Most favored regions	95.20%	90.90%	86.80%
Additionally allowed regions	4.80%	9.10%	12.40%
Generously allowed regions	0	0.10%	0.60%
Disallowed regions	0	0	0.20%
Average pairwise r.m.s.d. (Å)			
Heavy	1.4	1.4	1.1
Backbone	0.8	0.8	0.6
Structure Quality Factors (raw/Z-scores)			
Procheck G-factor (phi/psi)	0.13/0.83	-0.09/-0.04	-0.17/-0.35
Procheck G-factor (all)	0.01/0.06	-0.15/-0.89	-0.18/-1.06
Verify3D	0.29/-2.73	0.27/-3.05	0.26/-3.21
MolProbity clashscore	8.27/0.11	9.32/-0.07	12.46/-0.61

Table 2.7 X-ray data reduction and model refinement.

Wavelength	1.116Å
Resolution Range	40.94-1.50 Å (1.53-1.50 Å)
Unit Cell	a=33.50Å , b=51.99Å , c=66.41Å $\alpha=\beta=\gamma=90^\circ$
Space Group	$P2_12_12_1$
Unique Reflections	19203 (895)
Multiplicity	22.7 (12.4)
Completeness	99.9% (92.9%)
$\langle I/\sigma I \rangle$	18.5 (1.8)
CC _{1/2}	0.999 (0.719)
R _{pim}	0.019 (0.283)
R _{work}	0.1864
R _{free}	0.2104
Total Refined Atoms	1344
Protein Residues	128
Solvent Molecules	81
Refined Ligand Atoms	48
Average B-factor	33.81Å ²
RMSD _{bonds}	0.013Å
RMSD _{angles}	1.31°
Rama. Plot:	
Favored	98.44%
Allowed	1.56%
Outliers	0.0%
Molprobit Clashscore ^d	3.84
PDB ID	6W90

Chapter 3 Precise Local Segment Structure Prediction and Design

This chapter is adapted from a part of a manuscript in preparation by Cody Krivacic, Kale Kundert, Xingjie Pan, Roland A. Pache, Lin Liu, Shane O Conchúir, Jeliazko Jeliazkov, Jeffrey J. Gray, Michael C. Thompson, James S. Fraser and Tanja Kortemme.

Author contributions:

XP, RR, and TK developed the concepts for the project. RR and XP developed and benchmarked the FKIC and LHKIC methods, with contributions from KK, SOC, JJ, and JYG. JYG and TK provided guidance, mentorship and resources. CK, KK, XP, RR, and TK wrote the manuscript with contributions from the other authors.

Introduction

Proteins use intricate 3-dimensional geometries in their active sites to carry out a vast array of biological functions, including catalysis, ligand sensing, signal transduction, and protein recognition. Protein design algorithms to engineer new functional site geometries of custom-made enzymes or novel binding proteins (Feldmeier & Höcker, 2013) thus have many important practical applications but require building 3-dimensional conformations of local active sites with atom-level accuracy (Baker, 2010) (Blomberg, et al., 2013). Several advances have been made with the difficult problem of reconstructing loop regions without regular secondary structure that are prevalent in protein functional sites (Canutescu & Dunbrack, 2003) (Jacobson, et al., 2004)

(Mandell, et al., 2009) (Stein & Kortemme, 2013) (Zhao, et al., 2011) (Sellers, et al., 2008). Related computational methods have been used successfully to graft new loops into naturally occurring proteins (Hu, et al., 2007), to connect secondary structure elements in synthetic proteins (MacDonald, et al., 2016) and to remodel an existing loop to alter enzyme specificity (Murphy, et al., 2009). However, designing active site segments is still a highly challenging problem due to limitations of accuracy and efficiency of existing structure prediction and design methods (Kundert & Kortemme, 2019). Algorithms that overcome these limitations are critically needed.

FKIC and LHKIC algorithms

To increase the efficiency of robotics-inspired sampling methods (Mandell, et al., 2009) to predict or design structures, we developed two new methods (FKIC and LHKIC) that model the geometries of local regions in proteins. FKIC and LHKIC integrate two concepts that have separately led to considerable advances in protein modeling: building protein conformations by insertion of peptide fragments frequently observed in known structures (Simons, et al., 1997) (Bradley, et al., 2005) (Raveh, et al., 2010) (Tyka, et al., 2012), and improved sampling of loop regions with an inverse kinematic closure algorithm termed ‘KIC’ (Mandell, et al., 2009) borrowed from the field of robotics (Coutsias, et al., 2004).

KIC determines ‘mechanically accessible’ conformations for protein segments of given lengths by sampling the phi/psi torsion degrees of freedom in the segment. Three C α atoms of an N-residue segment are designated as pivots, leaving N-3 non-pivot C α atoms. In the standard implementation of KIC in Rosetta (Mandell, et al., 2009), non-pivot torsions are sampled from a residue type-specific Ramachandran map; KIC then finds all values for the pivot torsions that close the loop, if

any exist. A major bottleneck for KIC is that the number of possible conformations (combinations of torsion angles) increases dramatically with increasing numbers of residues to be modeled. This problem becomes particularly limiting when predicting larger contiguous segments or entire functional sites which often contain both loop and secondary structure regions in several interacting segments. To enable these more challenging but also more realistic problems, we integrated KIC with fragment insertion (Lee, et al., 2010) in the Rosetta protein structure prediction and design suite (Leaver-Fay, et al., 2011). In the new FKIC structure prediction method, non-pivot degrees of freedom are taken from peptide fragments that are picked using the sequence of the target loop (Gront, et al., 2011); KIC is then used to determine the values of the pivot torsions (**Figure 3.1a**). We reasoned that FKIC would combine the improvements of KIC demonstrated previously (Mandell, et al., 2009) with the reduction of degrees of freedom by using coupled torsion angles from fragments (in contrast to sampling all non-pivot torsions independently from Ramachandran space in KIC).

In loop design problems, the sequences of the target loops are variable during the sampling. We therefore developed an additional method, LHKIC, which uses the loophash protocol (Tyka, et al., 2012) to pick fragments that simultaneously sample structures and sequences of the target loops. The loophash protocol uses the 6D transformation between the residue before the first pivot and the residue after the last pivot as a query key to find fragments that approximately close the gap between these two residues (**Figure 3.1a**). After insertion of a fragment, KIC determines the pivot torsions that close the gap and LHKIC mutates remodeled residues to the amino acids from the inserted fragment to improve local sequence-structure compatibility. Individual FKIC and LHKIC sampling moves (**Figure 3.1a**) are then followed by optimization of side chain conformations in

and around the altered backbone region and integrated into a Monte Carlo minimization protocol (**Figure 3.1b**); sampled conformations are evaluated with Rosetta's all-atom energy function (Park, et al., 2016) (Alford, et al., 2017).

FKIC benchmarking

We tested the ability of FKIC to recapitulate the local conformations of protein segments, given their sequences, on a benchmark of 45 12-residue loops, which were used previously to evaluate KIC ("Standard" set). We also constructed two new datasets designed to address the challenges of modeling active site regions. The first new set contained 30 16-residue-segments where each segment contains both regular secondary structure elements and loop regions ("Mixed Segment" set). We reasoned that FKIC should improve the sampling of regular secondary structure elements by building them from fragments, whereas these elements might be difficult to model with KIC alone, since KIC samples each non-pivot torsion angle independently. The second new set contained 30 pairs of interacting 10-residue segments ("Multiple Segments" set). At the beginning of each simulation, we discarded information on the native geometry of all target segments by building starting conformations with extended phi/psi torsion angles and idealized bond lengths and angles. All side chains within 10Å from the target segments were deleted and replaced with side chain conformations from a rotamer library (Dunbrack & Cohen, 1997) during FKIC simulations. In all benchmark simulations, we also excluded fragments from structures homologous to the benchmark cases (**Methods**) to test FKIC's ability to predict novel structures. For each test protein in each set, we generated 500 models with FKIC and calculated the backbone heavy atom root mean square deviation (RMSD) of each target segment after aligning the protein without the modeled segment to its crystal structure. We also applied control methods that use KIC

and fragment insertion (CCD) (Canutescu & Dunbrack, 2003) alone to the same datasets using an otherwise identical protocol in Rosetta. For most comparisons, we replaced KIC with an improved KIC protocol called next-generation KIC (NGK) (Stein & Kortemme, 2013). We used two performance metrics: The first quantifies prediction accuracy by determining the RMSD of the model with the best (lowest) Rosetta energy to each native structure and then taking the median RMSD value across each dataset. The second quantifies sampling performance by measuring the fraction of native-like (correct) models generated for each protein case, where native-like is defined as $<1\text{\AA}$ (“sub- $\text{\AA}$ ”) RMSD to the native structure, and again taking the median for each dataset. We also measured the median run time to determine whether any increased sampling performance increases computational cost (**Methods** and **Table 3.1a**).

The Rosetta KIC method has previously been shown (Mandell, et al., 2009) to lead to improvements over the state-of-the-art molecular mechanics method (Sellers, et al., 2008) , with additional improvements implemented in the NGK update (Stein & Kortemme, 2013) that perform comparably to methods GalaxyLoop-PS2 (Park, et al., 2014), RCD+ (López-Blanco, et al., 2016), Sphinx, FREAD and LEAP (Marks, et al., 2018) when tested on identical datasets. Here we show that FKIC improves structure prediction accuracy over CCD, KIC and NGK, with the largest changes for the two new datasets (**Figure 3.2a**, top and **Table 3.1a**). On the 16-residue Mixed Segment dataset, which tests the ability of FKIC to predict conformations of protein segments with arbitrary secondary structure composition, the median accuracy improved to $0.53\text{\AA}$ RMSD with FKIC compared to $1.29\text{\AA}$ and $1.07\text{\AA}$ with CCD and NGK alone, respectively. In 23/30 cases FKIC was able to identify conformations very close to the crystallographic structure ($<1\text{\AA}$ RMSD) as the lowest scoring model, compared to only 13 cases with CCD and 15 cases with NGK (**Table 3.2**).

For the Multiple Segments dataset, which tests the ability of FKIC to predict conformations of discontinuous interacting segments (a common feature of protein active sites), FKIC was the only method that yields atomic ($1\text{\AA}$) median accuracy, compared to $1.97\text{\AA}$ and $1.29\text{\AA}$ with CCD and NGK alone, respectively (**Figure 3.2a, top** and **Table 3.1a**). Representative examples where FKIC correctly reconstructs protein geometries while NGK fails are shown for the Mixed Segment and Multiple Segments datasets in **Figure 3.2b, c** and details are given in **Table 3.2** and **Table 3.3**. The improvements on the Standard dataset are smaller (median RMSD is $0.62\text{\AA}$ with FKIC compared to $0.64\text{\AA}$ for NGK, **Table 3.1a**), but for 35/45 proteins FKIC finds lower energy structures than NGK (Error! Reference source not found.). Cases where FKIC predictions did not lead to the identification of sub- $\text{\AA}$ accuracy lowest-scoring models can be attributed to both sampling and energy function limitations (**Table 3.5** and **Figure 3.3**).

The second notable advance with FKIC, in addition to the improvements in accuracy described above, is how frequently FKIC generates conformations that are $<1\text{\AA}$ RMSD from the crystallographic conformation (**Figure 3.2a, bottom**). For the Mixed Segment set, the median fraction of sub- $\text{\AA}$ predictions for FKIC is 52.3%, which is 45- and 105-fold higher than for NGK and CCD, respectively. For the Multiple Segments dataset, the median fraction of sub- $\text{\AA}$ predictions is 28.5% with FKIC, which is 5-fold higher than with NGK (5.5%) and 143-fold higher (0.2%) than with CCD (**Table 3.1a**). Notably, the improvement of the fraction of sub- $\text{\AA}$ predictions is negatively correlated with the mean 3-mer fragment distance from the native structure (**Figure 3.2d, Methods**), showing that high quality fragments focus the sampling on native-like conformations. In several cases, FKIC can find correct solutions for even larger conformational sampling problems such as a set with 2 interacting 12-residue segments (**Note 1** and **Table 3.6**).

In all benchmark simulations described above, we excluded fragments from structures homologous to the benchmark cases. With fragments from homologs present in the database, both prediction accuracy and the median fraction of sub-Å predictions improved further, as expected (**Table 3.1b**). FKIC might therefore be able to accurately sample even larger regions in cases where homologous structures are available. Taken together, we conclude that FKIC drastically increases sampling efficiency in the prediction of protein segments from sequence.

Discussion

We have introduced two new robotics-inspired sampling methods, FKIC and LHKIC, that model the structures of backbone segments with high accuracy and efficiency. FKIC leads to considerable improvements over the two approaches it combines, the fragment-independent loop reconstruction method NGK and the fragment-insertion based prediction approach CCD. In addition to sub-Å structure prediction accuracy, our results demonstrate that FKIC provides up to ~140- fold improvement in sampling native-like conformations on the challenging problems of modeling local protein geometries with multiple segments and arbitrary secondary structure composition. This key advance in sampling performance paves the way to integrate FKIC with other methods in a variety of applications. We expect that FKIC will be most useful in applications that predict or design the conformations of local regions in proteins where the remainder of the protein stays relatively fixed or undergoes only slight adjustments. As such, applications to homology modeling may require integration of FKIC with more aggressive remodeling in the entire protein, not just a local region (**Note 4**). In contrast, FKIC should be suited to accurately predict the conformations of local protein regions from limited or low-resolution experimental data, and, in combination with LHKIC, to design new backbone geometries not seen in nature. LHKIC, while performing slightly

worse than FKIC when predicting conformations given their sequences (**Table 3.1**), is well-suited to design because of the ability to sample both new sequences and new structures simultaneously.

Methods

FKIC

We developed FKIC as a new protocol for modeling local protein geometries in Rosetta. FKIC is based on the KIC protocol (Mandell, et al., 2009), but, instead of sampling non-pivot ϕ/ψ torsions probabilistically from Ramachandran space, FKIC uses coupled $\phi/\psi/\omega$ degrees of freedom from consecutive residues of protein fragments of size nine, three or one to sample conformational space. During the low- and high-resolution sampling stages (**Figure 3.1**), each KIC move in the original KIC protocol is replaced by an FKIC move (**Figure 3.1B**). An FKIC move consists of the following sequence of steps: (i) a fragment library is chosen at random from all available libraries (i.e. 9mers, 3mers and 1mers), (ii) the chosen fragment library is searched for fragment alignment frames that (at least partially) overlap with the given target sub-segment, (iii) one of the alignment frames is chosen at random, (iv) one of the 200 fragments contained in the given alignment frame is chosen at random, (v) the $\phi/\psi/\omega$ torsions of that fragment are applied to the respective overlapping region of the given target sub-segment, and (vi) the segment is closed using kinematic closure. We ran the FKIC protocol using the same fragment libraries as for the CCD protocol.

LHKIC

We developed LHKIC for designing local protein geometries in Rosetta. LHKIC and FKIC share the same simulation pipeline (**Figure 3.1**). In LHKIC, the non-pivot $\phi/\psi/\omega$ degrees of freedom are sampled from fragments picked by the loophash algorithm (Tyka, et al., 2012). At each KIC

sampling step, we calculate the 6D transformation from the residue before the first pivot to the residue after the last pivot. We use the 6D transformation to query a pre-compiled loophash database (**Method**). One 6D transformation query can return multiple loops. Torsions of a random loop from the returned loops are applied to the residues between the pivot residues. By default, LHKIC does not mutate the sequence of the loop. When the `loophash_perturb_sequence` option is set to true, LHKIC applies the sequence of the returned loop to the pivot residues and the residues between the pivots.

Rosetta

FKIC and NGK benchmarking simulations were performed using the Rosetta macromolecular modeling and design suite (<https://www.rosettacommons.org/software>), revision 59052. The LHKIC method was developed later and used Rosetta revision 60022. KIC simulation results reported in **Figure 3.1a** in the main text were taken from (Conchúir, et al., 2015). The Rosetta energy function has recently undergone several revisions, including the changes described in the “talaris2013” and “talaris2014” versions (O’Meara, et al., 2015) and the latest improvements made in the “ref2015” version (Park, et al., 2016). The ref2015 energy function was used for all benchmarks unless otherwise noted. Compared to the other energy functions, ref2015 showed a consistent performance improvement (**Table 3.7**). (Conchúir, et al., 2015)

12-Residue “Standard” benchmark dataset

We first tested FKIC on the 12-residue loop benchmark dataset used in previous work (Conchúir, et al., 2015) (Mandell, et al., 2009) (Stein & Kortemme, 2013). This benchmark dataset consists of 45 protein structures from the Protein Data Bank (PDB) containing non-redundant 12-residue

target segments without regular secondary structure, curated from two previously described datasets (Sellers, et al., 2008) (Wang, et al., 2007). We used this dataset even though it is not ideal (for example, the conformation of several segments might be influenced by crystal contacts, see **Note 2** and **Table 3.5**) to facilitate comparison of FKIC with previous protocols. For each loop, we retained the N and C α atoms of the N-terminal residue, as well as the C α , C and O atoms of the C-terminal residue, which serve as loop anchor points for kinematic closure (as in (Mandell, et al., 2009)).

16-Residue “*Mixed Segment*” benchmark dataset

The *Mixed Segment* benchmark dataset consists of 30 structures from the PDB containing 16-residue target segments. The target segments are derived from structures in the *Standard* benchmark dataset above using the following criteria:

- The crystallographic resolution of the experimentally determined structure is equal to or better than 2Å.
- Each segment has 5 to 11 residues that contain alpha helices or beta strands defined using DSSP (Joosten, et al., 2011) and the remainder of the segment is designated as loop.
- Residues in the segment are at least 4Å away from any chains or copies of the molecule in other asymmetric units, to avoid crystal contacts.
- The segment is at least 5 residues away from the chain termini.

The segment that satisfies all criteria with the lowest distance from the protein surface is selected.

10-Residue “*Multiple Segments*” benchmark dataset

The *Multiple Segments* benchmark dataset consists of 30 structures from the PDB each containing a pair of 10-residue interacting target segments. Structures are derived from the top8000 dataset (Chen, et al., 2010) using the following criteria:

- The crystallographic resolution of the experimentally determined structure is equal to or better than 2Å. Each segment has less than 3 residues that have regular secondary structure, defined as above.
- Residues in the segment are at least 4Å away from any chains or copies of the molecule in other asymmetric units, to avoid crystal contacts.
- Each segment is at least 5 residues away from the chain termini.
- Segments in each pair are within 4Å and are separated by at least 5 residues in primary sequence.

The pair of segments that satisfies all criteria with the lowest distance from the protein surface is selected. We also constructed two analogous sets that contain either two 8-residue segments or two 12-residue segments (**Note 1**; **Table 3.6** and **Table 3.8**).

Preparation of benchmark input structures

To exclude information on the native conformation of the target segment(s) for all benchmark datasets, all side chains in the segment(s) as well as side chains within 10Å of the segment(s) (based on all-atom pairwise distance measurements) are removed. The backbone information is removed by changing the segment into an extended conformation with idealized bond lengths and angles. The datasets are constructed with an openly available script: https://github.com/Kortemme-Lab/benchmark_set_construct.

Generation of fragment libraries

We generated libraries of 9-mer and 3-mer fragments for all benchmark cases using the fragment picking method described in (Gront, et al., 2011). The method selects fragments from a representative database of 16,801 protein chains extracted from the PDB and culled such that any two chains have at most 60% sequence identity. The fragment database is part of the Rosetta software and located in `rosetta/tools/fragment_tools/vall.jul19.2011.gz`. The fragments selected for each segment sequence position span a 3- or 9-residue frame, which overlaps with neighboring frames. Moreover, we allow sampling of 1-mer fragments, which consist of a single triplet of $\phi/\psi/\omega$ torsions that are generated on the fly by the respective modeling protocol (see below) based on the 3-mer fragment library for the given position. The `make_fragments.pl` script in the `rosetta/tools/fragment_tools/` directory integrates several data sources to maximize fragment quality, including sequence similarity and the detection of homologs using PsiBLAST (Altschul, et al., 1997), predicted secondary structure similarity using PsiPred (Ward, et al., 2003) and prediction of preferred ϕ/ψ torsions and solvent accessibility using SPARKS-X (Yang, et al., 2011). Importantly, for benchmarking purposes, we ran simulations using fragment libraries that exclude homologs to the given query sequence, by providing the `-nohoms` flag to the `make_fragments.pl` script, which excludes all protein chains with a PsiBLAST E-value < 0.05 (Altschul, et al., 1997) from the fragment picking process.

CCD commands

The Rosetta CCD loop modeling method using fragment insertion and the cyclic coordinate descent (CCD) closure technique (Canutescu & Dunbrack, 2003) is described in (Wang, et al., 2007). We used the following Rosetta scripts to run the simulations:

```
<ROSETTASCRIPTS>
```

```
  <TASKOPERATIONS>
```

```
    <RestrictToLoops name="loop" loops_file="%%loop_file%%"/>
```

```
  </TASKOPERATIONS>
```

```
  <MOVERS>
```

```
    <LoopmodelWrapper name="modeler" loops_file="%%loop_file%%"
```

```
    fast="%%fast%%"/>
```

```
  </MOVERS>
```

```
  <PROTOCOLS>
```

```
    <Add mover_name="modeler"/>
```

```
  </PROTOCOLS>
```

```
</ROSETTASCRIPTS>
```

Let the home directory of Rosetta be */path/to/rosetta/main*; the input structure be *my_structure.pdb*; the loop file be *my_structure.loop*; the fragment files be *my_structure.200.9mers.gz* and *my_structure.200.3mers.gz*; and the Rosetta script XML file be *loopmodel.xml*. Then the command to run one simulation is:

```

/path/to/rosetta/main/source/bin/rosetta_scripts.linuxgccrelease -database
/path/to/rosetta/main/database -in:file:s my_structure.pdb -parser:protocol loopmodel.xml -
parser:script_vars loop_file=my_structure.loop fast=no -out:prefix prefix -overwrite -
loops:remodel quick_ccd -loops:refine refine_ccd -ex1 -ex2 -corrections:beta -loops:frag_sizes 9
3 1 -loops:frag_files my_structure.200.9mers.gz my_structure.200.3mers.gz

```

NGK commands

The NGK loop modeling method is described in (Conchúir, et al., 2015). We used the following Rosetta script to run the simulations:

```

<ROSETTASCRIPTS>

  <TASKOPERATIONS>

    <RestrictToLoops name="loop" loops_file="%%loop_file%%"/>

  </TASKOPERATIONS>

  <MOVERS>

    <LoopModeler name="modeler" config="kic" loops_file="%%loop_file%%"
fast="%%fast%%" />

  </MOVERS>

  <PROTOCOLS>

    <Add mover_name="modeler"/>

  </PROTOCOLS>

</ROSETTASCRIPTS>

```

Let the home directory of Rosetta be */path/to/rosetta/main*; the input structure be *my_structure.pdb*; the loop file be *my_structure.loop*; and the Rosetta script XML file be *loopmodel.xml*. Then the command line to run one simulation is:

```
/path/to/rosetta/main/source/bin/rosetta_scripts.mysql.linuxgccrelease -database  
/path/to/rosetta/main/main/database -in:file:s my_structure.pdb -parser:protocol loopmodel.xml  
-parser:script_vars loop_file=my_structure.loop fast=no -out:prefix prefix -overwrite -  
corrections:beta
```

FKIC commands

We used the following Rosetta script to run FKIC simulations:

```
<ROSETTASCRIPS>  
  
  <TASKOPERATIONS>  
  
    <RestrictToLoops name="loop" loops_file="%%loop_file%%"/>  
  
  </TASKOPERATIONS>  
  
  <MOVERS>  
  
    <LoopModeler  
      name="modeler"  
  
      config="kic_with_fragments"  
  
      loops_file="%%loop_file%%"  
  
      fast="%%fast%%">  
  
    </LoopModeler>
```

</MOVERS>

<PROTOCOLS>

<Add mover_name="modeler"/>

</PROTOCOLS>

</ROSETTASCRIPTS>

Let the home directory of Rosetta be */path/to/rosetta/main*; the input structure be *my_structure.pdb*; the loop file be *my_structure.loop*; the fragment files be *my_structure.200.9mers.gz* and *my_structure.200.3mers.gz*; and the Rosetta script XML file be *loopmodel.xml*. Then the command line for one simulation is:

```
/path/to/rosetta/main/source/bin/rosetta_scripts.mysql.linuxgccrelease -database  
/path/to/rosetta/main/database -in:file:s my_structure.pdb -parser:protocol loopmodel.xml -  
parser:script_vars loop_file=my_structure.loop fast=no -out:prefix prefix -overwrite -  
corrections:beta -loops:frag_sizes 9 3 1 -loops:frag_files my_structure.200.9mers.gz  
my_structure.200.3mers.gz
```

Simulations with native bond lengths and angles

For control simulations that use native bond lengths and angles as input, we replaced the input structure *my_structure.pdb* with the native structure *my_structure_native.pdb* and replaced the Rosetta script XML file with

<ROSETTASCRIPTS>

```

<TASKOPERATIONS>

  <RestrictToLoops name="loop" loops_file="%%loop_file%%"/>

</TASKOPERATIONS>

<MOVERS>

  <LoopModeler

    name="modeler"

    config="kic_with_frgs"

    loops_file="%%loop_file%%"

    fast="%%fast%%">

    <Build skip="True" />

  </LoopModeler>

</MOVERS>

<PROTOCOLS>

  <Add mover_name="modeler"/>

</PROTOCOLS>

</ROSETTASCRIPTS>

```

Fragment distance calculation

The chord distance (North, et al., 2011) was calculated between pairs of residue 3mers. The chord distance between two angles is defined as: $D^2(\theta_1, \theta_2) = 2 - 2 \cos(\theta_2 - \theta_1)$. In our case, this value was calculated for both backbone dihedral angles and summed over paired residues between

fragments and target loops: $\langle D \rangle = \frac{1}{n} \sum_i \left(\frac{1}{2} D^2(\phi_1^i, \phi_2^i) + \frac{1}{2} D^2(\psi_1^i, \psi_2^i) \right)$, with defining 3mers. $\langle D \rangle$ will

have a minimum of 0 if the angles match exactly and a maximum of 4 if the angles differ by 180 degrees.

Generation of loophash databases

We generated the loophash databases using the `loophash_createfiltereddb` application and the VALL database distributed with Rosetta. The command line is:

```
mpirun -np 32 loophash_createfiltereddb.mpi.linuxgccrelease -lh:db_path loophash_db/ -  
in:file:vall path_to_rosetta_tools_repository/tools/fragment_tools/vall.jul19.2011.gz -  
lh:loopsizes 3 4 5 6 7 8 9 10 11 12 13 14 -lh:num_partitions 32 -lh:createdb_rms_cutoff 4.5 6  
7.5 9 10.5 12 13.5 15 16.5 18 19.5 21
```

where *loophash_db* is the output path.

Loop structure prediction with LHKIC

We use the following Rosetta script to run the simulations:

```
<ROSETTASCRIPS>
```

```
<MOVERS>
```

```
<LoopModeler name="modeler" config="loophash_kic" loops_file="%%loop_file%%"  
fast="%%fast%%" />
```

```
</MOVERS>
```

```
<PROTOCOLS>
```

```

    <Add mover_name="modeler"/>

</PROTOCOLS>

</ROSETTASCRIPTS>

```

Let the home directory of Rosetta be */path/to/rosetta/main*; the input structure be *my_structure.pdb*; the loop file be *my_structure.loop*; the Rosetta script XML file be *loopmodel.xml*; and the path to the loophash database be *path_to_loophash_db*. Then the command line to run one simulation is:

```

/path/to/rosetta/main/source/bin/rosetta_scripts.mysql.linuxgccrelease -database
/path/to/rosetta/main/main/database -in:file:s my_structure.pdb -parser:protocol loopmodel.xml
-parser:script_vars loop_file=my_structure.loop fast=no -out:prefix prefix -overwrite -
lh:loopsizes 6 8 10 -lh:db_path path_to_loophash_db

```

Simulation Time

The simulations were run on the shared QB3 Cluster at UCSF. The major CPU types of the cluster are Intel Xeon X5650 (2.66GHz, x86_64), Intel Xeon E5430 (2.66Ghz, x86_64), Intel Xeon E5345 (2.33GHz, x86_64) and AMD Opteron 275 (2.2GHz, x86_64).

The median simulation times of FKIC for a single trajectory were 3456 seconds, 7196 seconds and 8834 seconds for the *Standard*, *Mixed Segment* and *Multiple Segments* datasets, respectively. For NGK, the corresponding simulation times were 3642 seconds, 7341 seconds and 9472 seconds.

For CCD, the simulation times were 2299 seconds, 4309 seconds and 5204 seconds. Thus, FKIC is consistently faster than NGK but slower than CCD by an average factor of 1.62.

Because 500 models are generated per protein, the median simulation times for each protein were 480, 999 and 1226 CPU hours for the *Standard*, *Mixed Segment* and *Multiple Segments* datasets, respectively. However, because individual simulation trajectories are independent of each other, if sufficient CPU cores are available then FKIC is bound only by the time needed to perform a single trajectory.

Notes

Note 1. 8-Residue and 12-residue *Multiple Segments* datasets

To further benchmark the performance of FKIC on multiple interacting segments, we constructed a dataset of 8-residue interacting segments and a dataset of 12-residue interacting segments in a similar manner to the construction of the 10-residue *Multiple Segments* dataset (see **Methods**). On the 8-residue interacting segment dataset, FKIC has 0.65Å median accuracy and 59.9% median fraction of sub-Å prediction; NGK has 0.79Å median accuracy and 35.6% median fraction of sub-Å prediction. On the 12-residue interacting dataset, FKIC has 1.53Å median accuracy and 0.21% median fraction of sub-Å prediction; NGK has 1.94Å median accuracy and 0% median fraction of sub-Å prediction. Thus, FKIC improves the prediction accuracy for multiple interacting segments consistently on different segment lengths and is able to find correct solutions for large conformational search problems, such as the set with two interacting 12-residue segments where previous methods such as NGK and CCD frequently fail (**Table 3.6** and **Table 3.8**).

Note 2. Analysis of failure cases

Despite the performance improvements with FKIC, in particular for the challenging *Mixed Segment* and *Multiple Segments* datasets, FKIC failed to accurately model 12 of the 45 segments in the *Standard* 12-residue benchmark dataset. In four cases (1cs6, 1msc, 2tgi and 4i1b) no sub-Å model was generated. In the other eight cases (1arb, 1bhe, 1cyo, 1m3s, 1onc, 1qlw, 1t1d and 1thg), sub-Å models were generated but could not be identified by energy (the RMSDs of lowest energy structures were larger than 1.1Å). These failures could result from deficiencies in sampling near-native conformations, from inaccuracies in the energy function and/or from problems with the crystal structure conformation such as effects of crystal packing. Sampling and energy function errors are often coupled, as the energy function guides sampling during the simulations. To gain insights into potential reasons for the failures we observed, we ran simulations of the failed proteins starting from their native structures as inputs. In these simulations, we skipped the first build stage (yellow in **Figure 3.4**) so that the native bond angles and bond lengths were kept. The results of these simulations allowed us to classify failure cases into 4 categories (**Figure 3.3**):

(1) Only the native-input simulations generate sub-Å models, which are correctly identified by energy. This occurred in two of the 12 failure cases (1cs6 and 2tgi). As the energies of native-like models are much lower than the non-native decoys (**Figure 3.3a**), failures in these cases are most likely due to the insufficient sampling.

(2) Both standard and native-input FKIC generate sub-Å models, but these models are only correctly identified by lowest energy in the native-input simulations (three of the 12 failure cases: 1bhe, 1onc and 1t1d). As the native-input simulations generated a larger number of correct models

with lower energies (**Figure 3.3b**), the failures are likely caused by the failure of Rosetta to efficiently sample near-native energy minima. One of the possible explanations is that the standard simulation idealizes bond lengths and bond angles in standard FKIC. Because of the rugged energy landscape, small conformational changes can result in significant energy differences.

(3) As in (2), both simulations generate sub-Å models and native-input simulations correctly identify these models by energy, but standard FKIC generates incorrect models with lower energies (**Figure 3.3c**, two of the 12 failure cases: 1arb and 1qlw). This behavior indicates linked scoring and sampling deficiencies.

(4) Neither standard nor native input simulations generate sub-Å models (**Figure 3.3d**; five of 12 failure cases: 1cyo, 1m3s, 1msc, 1thg and 4i1b). While these simulations start from the native backbones, they do not include crystal contacts. Because crystal packing affecting loop conformations is well known (Rapp & Pollack, 2005) (Eyal, et al., 2005), there is a formal possibility that the failures of 1cyo, 1m3s, 1msc and 4i1b are due to crystal packing. For the lowest energy models for 1cyo and 4i1b, the incorrectly modeled loops would have unfavorable contacts in the crystal lattice. For 1m3s and 1msc, the native loop conformations make contacts with another monomer in the crystal. For 1thg, the RMSD of the lowest energy structure improved from 1.86Å to 1.12Å when including native bond lengths and angles, so there might be both sampling and energy function problems.

In sum, in particular for categories (1) and (2), it could be beneficial to incorporate sampling of bond lengths and bond angles, which we kept to their idealized values to reduce the conformational

space to be sampled. Category (3) is indicative of energy function failures although we note that sampling and scoring are coupled in our simulations that accept or reject models based on their energies. Category (4) identifies a few cases where crystal packing may influence the conformation of the modeled segment in the experimentally determined structure.

Note 3. Loop modeling on template-based models and perturbed datasets

To test the performance of FKIC in contexts where the environment of the remodeled loop is non-native, we benchmarked FKIC on several datasets from (Park, et al., 2014), including a template based model and three sidechain/backbone perturbed loop datasets (**Table 3.9**). On the side chain perturbed datasets, FKIC performs similarly to NGK, and outperforms the other methods. This behavior is likely because surrounding residues are repacked during FKIC or NGK simulations, indicating that these methods can account for slight imperfections in the environment that can be resolved by altering side chain conformations. On backbone perturbed datasets, the performance of FKIC is comparable to the reported results of GalaxyLoop-PS2, with a median RMSD of lowest-scoring models of 1.68 Å and 1.65 Å for FKIC and Galaxy-PS2, respectively. On the template-based model dataset, none of the methods performs well, with the median RMSD of lowest scoring models above 3 Å for all methods.

It should be noted that the prediction and evaluation approach described here is not suited to appropriately assess the accuracy of loop modeling methods in environments where the surrounding backbone is perturbed, since the backbone in the environment is not allowed to change during the simulation. In our study, we compare a predicted loop structure to the native loop structure after aligning the surrounding environment of the loop. The native loop is the correct

answer when the surrounding environment is unperturbed. However, the native loop conformation may not be compatible with the perturbed backbone and can therefore not be identified as the lowest scoring model in simulations where the backbone of the environment remains in its (unchanged) perturbed conformation. Our analysis on the dataset containing backbone perturbations after MD simulations from (Park, et al., 2014) supports this argument. For each protein in this dataset, we defined the residues within 10 Å from the native loop as surrounding residues. We then superimposed the native structure and the perturbed structure by the backbone heavy atoms of the surrounding residues and calculated heavy atom steric clashes between the native loop and the perturbed backbone atoms. Two heavy atoms were defined as clashing if their interatomic distance was within 2.5 Å. This analysis revealed that twelve out of the twenty native loops in the dataset contained clashes with their perturbed environments that cannot be resolved with simulations that do not relax the surrounding backbone. When excluding these cases, the FKIC and GalaxyLoop-PS2 median RMSDs improved to 1.3 Å and 1.4 Å, respectively.

These considerations highlight that modeling loops in perturbed environments such as homology models remains an important unsolved problem. To adapt the design-centric loop modeling methods presented here to the problem of homology modeling should include simultaneous or iterative refinement of both loop structures and the environment.

Note 4. Benchmark LHKIC on structure prediction

The LHKIC method was developed for loop design, but it can also be used for structure prediction. We benchmarked the prediction performance of LHKIC on the Standard 12 residue dataset (**Table 3.1**). The performance of LHKIC is comparable to NGK and FKIC in terms of median RMSD of

lowest scoring models. The median sub- Å fraction of LHKIC is 24.35%, better than NGK but worse than FKIC. This result indicates that sequence independent fragments (as in LHKIC) can improve sampling in structure predictions over non fragment-based methods such as NGK, but the improvement is smaller than when using fragments picked with sequence information as in FKIC. Note that the absolute energies for LHKIC cannot be directly compared to the other methods since LHKIC was developed in a newer version of Rosetta (revision 60022, see **Methods**).

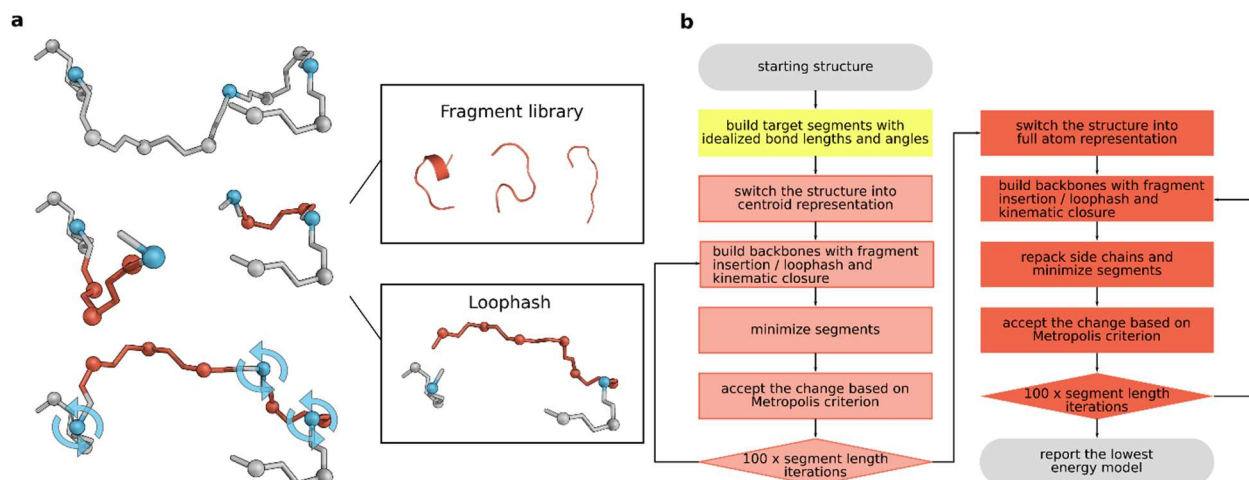


Figure 3.1 FKIC and LHKIC protocols.

(a) Individual FKIC/LHKIC move. Three Ca atoms (blue) on the target segment to be modeled are picked randomly as pivots. Fragment insertion or loop hash is applied to sample torsion degrees of freedom at non-pivot atoms (red), which breaks the chain. The KIC algorithm is then used to close the chain by determining appropriate values for the pivot torsions. **(b)** Integration of FKIC/LHKIC moves into a structure prediction protocol in Rosetta. Two stages of Monte Carlo simulated annealing - a beginning low resolution stage (centroid, light red) followed by a high-resolution stage (all-atom, red) - are used to find the lowest Rosetta energy conformation of target segments. For benchmarking, the native conformations of target segments and surrounding side chains are removed at the beginning of the simulation (**Methods**).

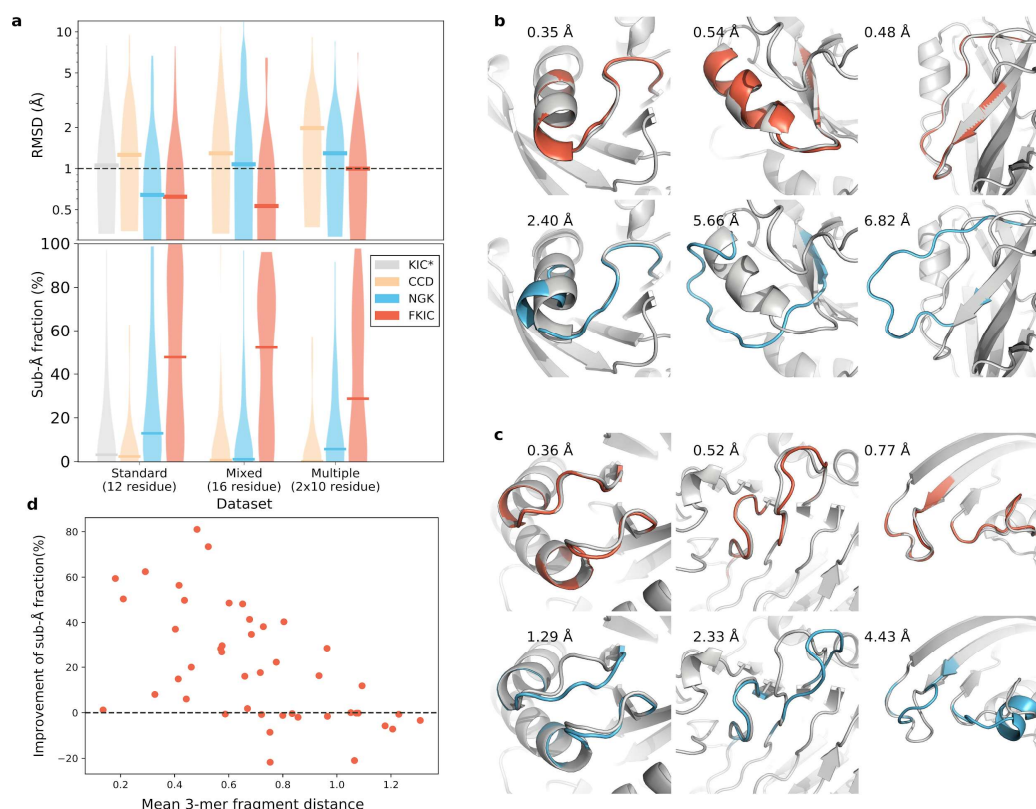


Figure 3.2 FKIC improves structure prediction accuracy.

(a) Comparison of performance of different methods for three datasets: (i) Standard dataset described in (Conchúir, et al., 2015), and 2 new sets: (ii) A “Mixed Segment” dataset with 30 16-residue regions that contain both loops and segments of regular secondary structure and (iii) a “Multiple Segments” dataset of 30 cases with 2 separate 10-residue regions that are interacting. KIC (Mandell, et al., 2009): grey; CCD (Canutescu & Dunbrack, 2003): orange; NGK (Stein & Kortemme, 2013): blue; FKIC: red. Upper panel: violin plot of RMSD of lowest energy (best) model across each dataset. FKIC is the only method that provides predictions with atomic accuracy (≤ 1 Å median RMSD) for all datasets. Lower panel: violin plot of fraction of predicted models in each dataset that have sub-Å accuracy. FKIC leads to considerable improvements over previous methods. Asterisk indicates data from (Conchúir, et al., 2015); all other simulations were run with the ref2015 (also called beta_nov15 when we ran the benchmark) Rosetta energy function (Park, et al., 2016); methods using fragments (CCD and FKIC) used identical fragment libraries that excluded fragments from structural homologs to the target proteins. **(b,c)** FKIC correctly reconstructs geometries where the previous state-of-the-art method, NGK, fails. Shown are examples from the Mixed Segment **(b)** and Multiple Segments dataset **(c)**. Experimentally determined structures: grey; predictions from FKIC: red, top; predictions from NGK: blue, bottom. RMSDs to the experimentally determined structures are given in each panel in Angstroms. **(d)** The fraction of sub-Å predictions is negatively correlated with the mean 3-mer fragment distance (**Methods**). Each data point represents a protein from the standard 12 residue dataset.

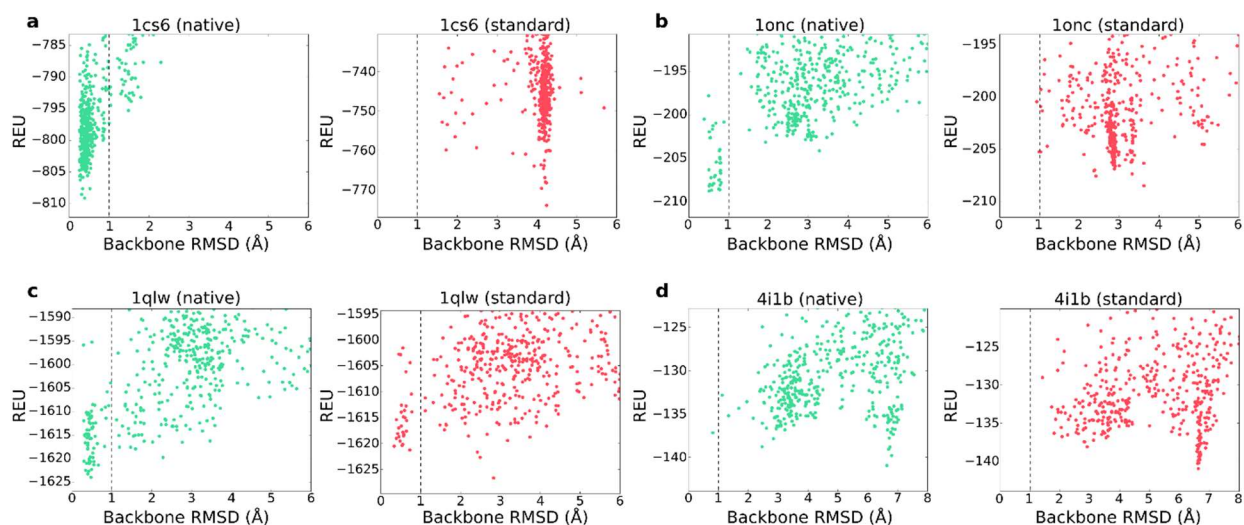


Figure 3.3 Examples of failures of FKIC.

Results of standard FKIC are shown in red (right in each panel) and results of FKIC with native input information and native bond lengths and angles are shown in green (left in each panel). Each point represents a Rosetta generated model. REU, Rosetta energy units. **(a)** Sub-Å models are generated only with native inputs. **(b)** Standard FKIC generates a few sub-Å models but they are not identified by energy. Using native inputs generates a larger number of near-native solutions that can be correctly identified by energy. **(c)** The simulation with native inputs correctly identifies native-like models, but a model generated by standard FKIC has lower energy. **(d)** Neither standard nor native-input simulations correctly identify sub-Å models.

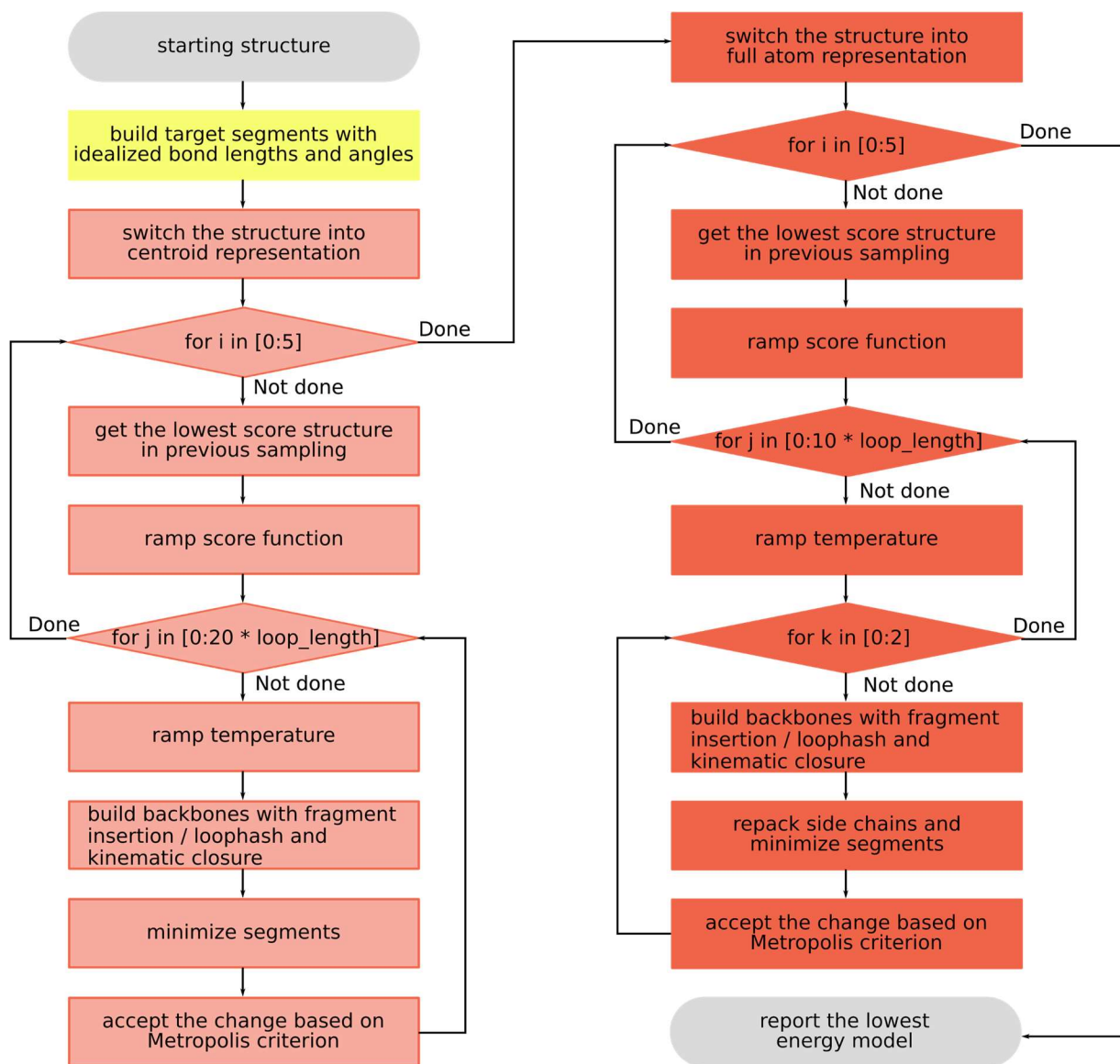


Figure 3.4 Detailed FKIC/LHKIC protocol.

The FKIC / LHKIC modeling protocol has a build stage (yellow), a centroid sampling stage (light red) and a full atom sampling stage (red). Both the centroid stage and the full atom stage perform simulated annealing which ramp the *rama* and *fa_rep* terms of the Rosetta energy function (Park, et al., 2016) (Alford, et al., 2017) in outer cycles and ramp the temperature in inner cycles.

Table 3.1 Datasets and performance summary**a) Comparison of methods**

Dataset	Sampling method	Rosetta energy function	Median RMSD of lowest energy model (Å)	Median RMSD of lowest RMSD model (Å)	Median RMSD all models (Å)	Median sub-A fraction	Median lowest energy (REU)	Median time (s)
<i>Standard</i>	KIC*	score12	1.05	NA	NA	4.30%	NA	NA
<i>Standard</i>	CCD	ref2015	1.26	0.47	3.27	2.00%	-709.18	2299
<i>Standard</i>	NGK	ref2015	0.64	0.37	2.70	13.00%	-712.65	3642
<i>Standard</i>	FKIC	ref2015	0.62	0.32	1.16	47.80%	-716.78	3456
<i>Standard</i>	LHKIC	ref2015	0.55**	0.34	2.66	24.35%	-655.18 ***	3057
<i>Mixed</i>	CCD	ref2015	1.29	0.67	3.46	0.50%	-719.42	4309
<i>Mixed</i>	NGK	ref2015	1.07	0.45	4.65	1.15%	-728.18	7341
<i>Mixed</i>	FKIC	ref2015	0.53	0.34	1.46	52.30%	-739.38	7196
<i>Multiple</i>	CCD	ref2015	1.97	0.90	2.95	0.20%	-557.06	5204
<i>Multiple</i>	NGK	ref2015	1.29	0.52	2.35	5.50%	-573.60	9472
<i>Multiple</i>	FKIC	ref2015	1.00	0.41	1.82	28.50%	-581.42	8834

REU, Rosetta energy units

* taken from (Conchúir, et al., 2015); all other simulations were run using the latest Rosetta energy function “ref2015” as described in (Park, et al., 2016)

** bold numbers denote best performance for given dataset

*** The LHKIC is benchmarked on a different revision from the other methods. The absolute energy value should not be compared with the other methods.

b) Inclusion of fragments from homologous structures

Dataset	Sampling method	Rosetta energy function	Median RMSD of lowest scoring model (Å)	Median RMSD of lowest RMSD model (Å)	Median RMSD all models (Å)	Median sub-A fraction	Median lowest energy (REU)	Median time (s)
<i>Standard</i>	FKIC	talaris2013	0.70	0.36	1.19	44.89%	-274.50	1646
<i>Standard</i>	FKIC (+ homologs)	talaris2013	0.59	0.33	0.85	66.80%	-273.29	1919

Table 3.2 Mixed Segment dataset detailed performance

			CCD					NGK					FKIC				
PDB name	Target segment residues		RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models
1dqz	103-118		0.50														
		46-61	5.27	0.63	2.77	-691.08	0.00	0.68	-706.51	0.50	-699.10	0.03	0.61	-709.42	0.34	-770.26	0.16
			-779.57	-662.32													
			0.87	0.47	1.33	-524.80	0.00	0.28	-706.51	0.30	-699.10	0.03	0.61	-709.42	0.34	-770.26	0.16
			-766.75	-651.31													
			0.00	0.28	0.00	-928.40	0.00	0.00	-928.40	0.44	-941.02	0.03	0.97	-709.42	0.34	-770.26	0.16
			1.34	0.68	3.57	-691.08	0.00	0.68	-706.51	0.50	-699.10	0.03	0.61	-709.42	0.34	-770.26	0.16
			-789.74	-668.11													
			0.74	0.30	0.50	-524.80	0.00	0.30	-706.51	0.50	-699.10	0.03	0.97	-709.42	0.34	-770.26	0.16
			-781.50	-663.53													
			0.00	0.97	0.03	-928.40	0.00	0.00	-928.40	0.44	-941.02	0.03	0.97	-709.42	0.34	-770.26	0.16
			1.27	0.61	0.29	-691.08	0.00	0.61	-706.51	0.50	-699.10	0.03	0.61	-709.42	0.34	-770.26	0.16
			-796.08	-683.38													
			0.51	0.34	0.29	-524.80	0.00	0.34	-706.51	0.50	-699.10	0.03	0.51	-709.42	0.34	-770.26	0.16
			-770.26	-663.29													
			0.16	0.96	0.67	-928.40	0.00	0.96	-928.40	0.44	-941.02	0.03	0.96	-709.42	0.34	-770.26	0.16
			2.40	2.94	2.94	-928.40	0.00	2.94	-928.40	0.44	-941.02	0.03	2.94	-928.40	0.38	-974.06	0.04
			-1255.72	-528.00													
			0.45	1.17	0.77	-524.80	0.00	1.17	-524.80	0.53	-507.84	0.00	0.77	-507.84	0.38	-974.06	0.04
			-1222.43	-507.84													
			0.01	0.00	0.01	-928.40	0.00	0.01	-928.40	0.53	-507.84	0.00	0.01	-507.84	0.38	-974.06	0.04
			0.35	0.42	0.42	-928.40	0.00	0.42	-928.40	0.53	-507.84	0.00	0.42	-507.84	0.38	-974.06	0.04
			-1261.68	-531.35													
			0.24	2.31	0.38	-524.80	0.00	2.31	-524.80	0.53	-507.84	0.00	0.38	-507.84	0.38	-974.06	0.04
			-1198.37	-509.84													
			0.84	0.00	0.04	-928.40	0.00	0.00	-928.40	0.53	-507.84	0.00	0.04	-507.84	0.38	-974.06	0.04

Table 3.2 Mixed Segment dataset detailed performance cont.

		CCD						NGK						FKIC					
	PDB name		lcm	lzm	lf46	li7p	lms9	loth	loyc										
	Target segment residues	76-91	252-267	24-39	45-60	155-170	431-446	272-287	330-345										
	RMSD of lowest energy model (Å)	0.38	1.45	6.88	5.48	0.68	10.98	1.31	1.35										
	Lowest energy (REU)	-694.76	-977.38	-654.25	-705.82	-647.33	-2721.15	-612.54	-557.96										
	Lowest RMSD (Å)	0.38	0.66	3.31	1.07	0.68	4.21	0.74	0.48										
	Energy of lowest RMSD model (REU)	-694.76	-969.81	-556.10	-701.81	-647.33	-2392.51	-510.97	-137.82										
	Fraction sub-Å models	0.03	0.01	0.00	0.00	0.00	0.00	0.01	0.08										
	RMSD of lowest energy model (Å)	0.26	4.62	0.50	7.10	1.59	11.83	2.24	5.66										
	Lowest energy (REU)	-710.85	-983.36	-684.52	-723.09	-654.72	-2744.95	-685.23	-536.60										
	Lowest RMSD (Å)	0.19	1.38	0.40	0.99	1.51	3.10	0.45	0.59										
	Energy of lowest RMSD model (REU)	-689.15	-962.42	-668.73	-708.18	-648.48	-2427.70	-555.42	126.42										
	Fraction sub-Å models	0.37	0.00	0.00	0.00	0.00	0.00	0.02	0.00										
	RMSD of lowest energy model (Å)	0.28	4.38	6.00	1.51	0.56	6.42	0.71	0.54										
	Lowest energy (REU)	-719.71	-991.09	-681.88	-735.38	-657.64	-2749.58	-673.30	-593.99										
	Lowest RMSD (Å)	0.21	0.38	1.37	1.08	0.37	0.70	0.38	0.39										
	Energy of lowest RMSD model (REU)	-706.81	-983.50	-611.59	-704.28	-652.65	-2079.87	-559.34	-471.11										
	Fraction sub-Å models	0.72	0.21	0.00	0.00	0.46	0.01	0.80	0.65										

Table 3.2 Mixed Segment dataset detailed performance cont.

PDB name	Target segment residues	CCD					NGK					FKIC				
		RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models
1pbe	200-215	0.35	-803.31	0.35	-803.31	0.17	0.59	-814.05	0.31	-807.25	0.25	0.57	-816.86	0.25	-813.62	0.58
1qlw	284-299	0.46	-1557.59	0.39	-1547.37	0.09	0.52	-1566.76	0.34	-1561.36	0.06	0.35	-1573.51	0.18	-1569.30	0.73
1srp	418-433	0.53	-733.70	0.53	-733.70	0.00	0.67	-733.27	0.65	-729.35	0.01	0.42	-743.39	0.29	-739.06	0.32
1tca	163-178	2.33	-829.93	1.05	-815.78	0.00	0.28	-850.81	0.28	-850.81	0.08	0.36	-852.46	0.27	-850.07	0.17
1thg	359-374	3.78	-1290.91	0.98	-798.31	0.00	7.31	-836.82	2.20	-782.91	0.00	0.87	-820.01	0.76	-793.95	0.28
1thw	97-112	0.50	-365.34	0.50	-365.34	0.00	0.33	-404.36	0.28	-361.20	0.35	0.34	-406.53	0.27	-361.89	0.61
1tib	166-181	0.49	-469.26	0.49	-469.26	0.07	0.30	-513.50	0.25	-488.46	0.71	0.29	-512.87	0.25	-499.93	0.81
1tml	26-41	0.83	-733.03	0.69	-729.62	0.02	0.35	-756.99	0.34	-745.83	0.04	0.35	-764.53	0.26	-756.61	0.80

Table 3.2 Mixed Segment dataset detailed performance cont.

		CCD					NGK					FKIC				
PDB name																
Target segment residues																
		RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models
3hsc	2sil	1xif	2ebn	2exo	2pia											
125-140	209-224	50-65	11-26	48-63	43-58											
0.33	0.65	1.28	0.50	3.51	1.08											
-874.75	-440.99	-815.74	-686.47	-615.43	-666.90											
0.31	0.35	0.47	0.50	1.17	0.37											
-869.11	-366.67	-813.47	-686.47	-602.03	-594.66											
0.99	0.13	0.02	0.03	0.00	0.33											
0.81	6.82	0.45	0.31	4.79	1.52											
-868.57	-523.38	-820.65	-715.52	-654.47	-683.56											
0.81	0.37	0.37	0.25	2.30	0.39											
-868.57	-370.74	-819.84	-708.04	-602.47	-581.39											
0.01	0.01	0.01	0.09	0.00	0.43											
0.30	0.48	0.59	0.29	4.88	0.45											
-878.17	-551.19	-829.72	-719.70	-680.12	-682.60											
0.27	0.29	0.32	0.24	0.43	0.34											
-873.40	-365.12	-820.15	-711.08	-626.40	-600.69											
0.80	0.67	0.35	0.74	0.07	0.96											

Table 3.3 Multiple Segments dataset detailed performance

		CCD						NGK						FKIC					
PDB name	Target segment residues	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models			
1a3a	11-20 87-96	5.65	-340.13	0.83	-284.99	0.00	0.84	-316.80	0.45	-246.79	0.13	0.84	-325.15	0.48	-254.18	0.58			
1deu	170-179 225-234	6.49	16511.10	1.81	16591.50	0.00	2.21	16114.00	1.56	16135.00	0.00	2.57	16115.50	0.97	16143.30	0.00			
1dqz	148-157 252-261	2.62	-623.02	1.45	-555.00	0.00	1.93	-671.27	0.41	-637.28	0.12	2.07	-667.08	0.55	-626.29	0.06			
1euv	529-538 570-579	2.17	-500.01	1.56	-474.99	0.00	1.88	-521.07	0.96	-513.85	0.00	1.10	-523.17	0.93	-517.57	0.01			
1fo9	239-248 261-270	1.09	-774.02	1.09	-774.02	0.00	0.84	-804.34	0.43	-795.79	0.67	0.81	-811.44	0.40	-797.06	0.83			
1fr	211-220 278-287	9.14	-700.66	2.23	-670.62	0.00	8.51	-719.42	2.41	-696.58	0.00	3.02	-714.95	0.77	-712.83	0.00			
1h1n	116-125 153-162	0.49	-651.04	0.43	-637.42	0.17	1.29	-670.91	0.25	-633.85	0.36	0.36	-671.26	0.23	-668.07	0.60			
1h6u	170-179 202-211	0.37	-728.01	0.36	-724.78	0.45	0.37	-732.42	0.32	-729.53	0.28	0.42	-734.81	0.28	-730.74	0.38			

Table 3.3 Multiple Segments dataset detailed performance cont.

		CCD					NGK					FKIC				
PDB name	Target segment residues	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models
1kzq																
44-53 83-92	85-94 118-127															
3.58		3.33														
-444.31			-308.26													
1.87				1.79												
-407.29					-295.82											
0.00						0.00										
2.61							3.11									
-466.18								-326.13								
1.72									1.51							
-458.54										-317.13						
0.00											0.00					
1.99												1.95				
-463.25													-325.51			
0.85														1.06		
-450.18															-308.24	
0.00																0.00

Table 3.3 Multiple Segments dataset detailed performance cont.

		CCD					NGK					FKIC				
PDB name	Target segment residues	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models
1m0z	71-80 49-58	0.72					0.73					0.66				
1nzm	12-21 98-107	0.40					0.32					0.24				
1qwd	18-27 93-102	2.97					1.17					0.90				
1t6g	140-149 197-206	1.00					1.17					1.10				
1u09	240-249 293-302	1.85					2.13					3.31				
1w0d	275-284 315-324	1.84					1.61					0.75				
1xdw	49-58 27-36	0.89					0.68					0.68				
1xg2	184-193 218-227	0.89					-783.62					-811.82				
3.59																
-594.97																
0.93																
-554.27																
0.00																
2.33																
-637.06																
0.49																
-614.76																
0.01																
0.52																
-638.36																
0.41																
-621.65																
0.56																

Table 3.3 Multiple Segments dataset detailed performance cont.

		CCD					NGK					FKIC				
PDB name																
Target segment residues																
2b0a	102-111 123-132	RMSD of lowest energy model (Å)					RMSD of lowest energy model (Å)					RMSD of lowest energy model (Å)				
		Lowest energy (REU)					Lowest energy (REU)					Lowest energy (REU)				
		Lowest RMSD (Å)					Lowest RMSD (Å)					Lowest RMSD (Å)				
		Energy of lowest RMSD model (REU)					Energy of lowest RMSD model (REU)					Energy of lowest RMSD model (REU)				
		Fraction sub-Å models					Fraction sub-Å models					Fraction sub-Å models				
2a4a	37-46 18-27	RMSD of lowest energy model (Å)					RMSD of lowest energy model (Å)					RMSD of lowest energy model (Å)				
		Lowest energy (REU)					Lowest energy (REU)					Lowest energy (REU)				
		Lowest RMSD (Å)					Lowest RMSD (Å)					Lowest RMSD (Å)				
		Energy of lowest RMSD model (REU)					Energy of lowest RMSD model (REU)					Energy of lowest RMSD model (REU)				
		Fraction sub-Å models					Fraction sub-Å models					Fraction sub-Å models				
1xsz	77-86 119-128	RMSD of lowest energy model (Å)					RMSD of lowest energy model (Å)					RMSD of lowest energy model (Å)				
		Lowest energy (REU)					Lowest energy (REU)					Lowest energy (REU)				
		Lowest RMSD (Å)					Lowest RMSD (Å)					Lowest RMSD (Å)				
		Energy of lowest RMSD model (REU)					Energy of lowest RMSD model (REU)					Energy of lowest RMSD model (REU)				
		Fraction sub-Å models					Fraction sub-Å models					Fraction sub-Å models				
1xwt	333-342 369-378	RMSD of lowest energy model (Å)					RMSD of lowest energy model (Å)					RMSD of lowest energy model (Å)				
		Lowest energy (REU)					Lowest energy (REU)					Lowest energy (REU)				
		Lowest RMSD (Å)					Lowest RMSD (Å)					Lowest RMSD (Å)				
		Energy of lowest RMSD model (REU)					Energy of lowest RMSD model (REU)					Energy of lowest RMSD model (REU)				
		Fraction sub-Å models					Fraction sub-Å models					Fraction sub-Å models				
1yif	176-185 144-153	RMSD of lowest energy model (Å)					RMSD of lowest energy model (Å)					RMSD of lowest energy model (Å)				
		Lowest energy (REU)					Lowest energy (REU)					Lowest energy (REU)				
		Lowest RMSD (Å)					Lowest RMSD (Å)					Lowest RMSD (Å)				
		Energy of lowest RMSD model (REU)					Energy of lowest RMSD model (REU)					Energy of lowest RMSD model (REU)				
		Fraction sub-Å models					Fraction sub-Å models					Fraction sub-Å models				
1zvt	659-668 716-725	RMSD of lowest energy model (Å)					RMSD of lowest energy model (Å)					RMSD of lowest energy model (Å)				
		Lowest energy (REU)					Lowest energy (REU)					Lowest energy (REU)				
		Lowest RMSD (Å)					Lowest RMSD (Å)					Lowest RMSD (Å)				
		Energy of lowest RMSD model (REU)					Energy of lowest RMSD model (REU)					Energy of lowest RMSD model (REU)				
		Fraction sub-Å models					Fraction sub-Å models					Fraction sub-Å models				
240a	102-111 123-132	RMSD of lowest energy model (Å)					RMSD of lowest energy model (Å)					RMSD of lowest energy model (Å)				
		Lowest energy (REU)					Lowest energy (REU)					Lowest energy (REU)				
		Lowest RMSD (Å)					Lowest RMSD (Å)					Lowest RMSD (Å)				
		Energy of lowest RMSD model (REU)					Energy of lowest RMSD model (REU)					Energy of lowest RMSD model (REU)				
		Fraction sub-Å models					Fraction sub-Å models					Fraction sub-Å models				

Table 3.4 Standard dataset detailed performance

PDB name	Target segment residues	CCD					NGK					FKIC				
		RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models
1cs6	1a8d	1a8d	1arb	1bhe	1bn8	1c5e	1cb0	1cnv	1cs6	1cs6	1cs6	1cs6	1cs6	1cs6	1cs6	1cs6
145-156	155-166	155-166	182-193	121-132	298-309	82-93	33-44	188-199	33-44	33-44	33-44	33-44	33-44	33-44	33-44	33-44
3.60	2.83	2.83	2.37	2.09	0.70	0.40	0.41	3.09	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.41
-731.51	-1008.38	-1008.38	-554.35	-923.69	-987.69	-742.69	-679.12	-691.00	-679.12	-679.12	-679.12	-679.12	-679.12	-679.12	-679.12	-679.12
1.71	2.14	2.14	0.23	1.84	0.33	0.30	0.30	1.24	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30
-704.86	-973.52	-973.52	-530.49	-913.87	-982.93	-742.50	-671.31	-665.03	-671.31	-671.31	-671.31	-671.31	-671.31	-671.31	-671.31	-671.31
0.00	0.00	0.00	0.01	0.00	0.02	0.09	0.10	0.00	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
4.26	0.38	0.38	0.54	0.68	0.75	0.36	0.25	1.06	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
-777.82	-1024.27	-1024.27	-562.94	-908.66	-993.21	-745.93	-679.95	-701.43	-679.95	-679.95	-679.95	-679.95	-679.95	-679.95	-679.95	-679.95
0.99	0.38	0.38	0.37	0.30	0.43	0.28	0.16	0.47	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16
-760.81	-1024.27	-1024.27	-539.98	-903.45	-986.78	-742.69	-651.78	-685.34	-651.78	-651.78	-651.78	-651.78	-651.78	-651.78	-651.78	-651.78
0.00	0.03	0.03	0.38	0.18	0.09	0.41	0.36	0.01	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36
4.24	0.42	0.42	2.06	1.91	0.48	0.38	0.36	1.02	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36
-773.94	-1022.05	-1022.05	-563.70	-915.76	-1000.64	-748.17	-680.65	-700.87	-680.65	-680.65	-680.65	-680.65	-680.65	-680.65	-680.65	-680.65
1.55	0.31	0.31	0.39	0.31	0.30	0.31	0.14	0.39	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
-745.58	-1019.72	-1019.72	-526.90	-906.99	-991.29	-744.34	-675.44	-699.85	-675.44	-675.44	-675.44	-675.44	-675.44	-675.44	-675.44	-675.44
0.00	0.01	0.01	0.17	0.13	0.44	0.90	0.84	0.03	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84

Table 3.4 Standard dataset detailed performance cont.

PDB name	Target segment residues	CCD					NGK					FKIC				
		RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models
1f46																
64-75																
2.27																
-709.18																
0.80																
-697.38																
0.03																
3.09																
-712.65																
0.78																
-705.47																
0.14																
0.44																
-716.78																
0.32																
-710.87																
0.30																

Table 3.4 Standard dataset detailed performance cont.

PDB name		CCD					NGK					FKIC				
Target segment residues		RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models
1oyc	1oth	1onc	1my7	1mse	1ms9	1m3s	1i7p									
203-214	69-80	23-34	254-265	9-20	529-540	68-79	63-74									
2.99	0.52	2.88	0.49	7.68	2.48	5.40	0.51									
-776.66	-837.57	-200.66	-504.92	151.93	-3018.90	-672.68	-669.97									
0.89	0.38	1.70	0.35	2.93	0.77	0.68	0.36									
-691.71	-836.26	-190.59	-496.96	188.30	-2998.79	-666.48	-661.74									
0.00	0.18	0.00	0.11	0.00	0.04	0.01	0.11									
0.30	0.45	3.63	0.61	6.54	0.27	5.05	0.42									
-818.69	-837.27	-209.76	-511.55	148.08	-3041.97	-674.66	-674.13									
0.28	0.31	0.46	0.40	1.27	0.22	0.44	0.31									
-719.38	-834.46	-208.03	-507.90	162.88	-3041.75	-664.02	-651.77									
0.13	0.34	0.04	0.75	0.00	0.91	0.31	0.53									
0.32	0.50	3.62	0.53	7.85	0.35	5.66	0.38									
-819.70	-843.18	-208.46	-512.36	147.53	-3044.20	-673.18	-672.19									
0.25	0.27	0.93	0.45	1.08	0.23	0.42	0.33									
-727.24	-833.69	-200.42	-507.57	155.56	-3030.97	-664.04	-669.48									
0.54	0.97	0.00	0.81	0.00	0.69	0.51	0.90									

Table 3.4 Standard dataset detailed performance cont.

PDB name		CCD						NGK						FKIC					
Target segment residues		RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models			
1thw	178-189	1.06	-394.74	1.06	-394.74	0.00	0.65	-413.28	0.49	-411.02	0.05	0.61	-414.92	0.46	-409.31	0.23			
1thg	127-138	2.88	-1419.28	1.64	-1385.22	0.00	1.05	-1434.24	0.71	-1411.26	0.07	1.86	-1434.72	0.28	-1423.84	0.07			
1tea	305-316	9.52	-846.00	2.37	-806.46	0.00	0.66	-859.34	0.26	-854.54	0.12	0.29	-864.98	0.24	-848.65	0.24			
1tld	127-138	0.42	-206.93	0.34	-204.69	0.14	2.79	-217.61	0.32	-207.60	0.09	2.94	-215.11	0.30	-209.60	0.39			
1srp	311-322	0.48	-782.41	0.31	-773.57	0.05	0.31	-877.44	0.25	-772.31	0.99	0.37	-879.03	0.25	-759.90	0.98			
1rro	17-28	0.44	-272.26	0.34	-270.03	0.08	0.64	-277.16	0.39	-270.25	0.10	0.63	-279.08	0.32	-269.39	0.48			
1qlw	31-42	4.34	-1611.33	2.07	-1604.34	0.00	4.87	-1624.91	0.30	-1605.29	0.13	2.83	-1626.69	0.33	-1620.54	0.06			
1pbe	129-140	2.60	-833.07	0.43	-790.79	0.06	0.45	-820.65	0.32	-801.51	0.32	0.38	-824.99	0.33	-804.46	0.72			

Table 3.4 Standard dataset detailed performance cont.

PDB name	Target segment residues	CCD					NGK					FKIC				
		RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models
2m2																
90-101																
1.41																
-283.79																
1.41																
-283.79																
0.00																
0.64																
-297.34																
0.32																
-289.54																
0.40																
0.68																
-298.25																
0.29																
-290.57																
0.69																

Table 3.5 Failure cases, Standard dataset

PDB name	Target segment residues	FKIC				FKIC native input				Reason for failure
		RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	
1arb	182-193	2.06	-563.70	0.39	-526.90	0.40	-557.64	0.21	-537.67	insufficient sampling
1bhe	121-132	1.91	-915.76	0.31	-906.99	0.24	-940.55	0.24	-928.05	insufficient sampling
1cs6	145-156	4.24	-773.94	1.55	-745.58	0.39	-809.13	0.25	-804.74	insufficient sampling
1cyo	12-23	5.09	-196.90	0.51	-191.38	1.06*	-194.88	0.32	-188.95	crystal packing
1m3s	68-79	5.66	-673.18	0.42	-664.04	5.39	-680.83	0.43	-669.76	crystal packing
1msc	9-20	7.85	147.53	1.08	155.56	8.74	146.72	0.86	163.39	crystal packing
1onc	23-34	3.62	-208.46	0.93	-200.42	0.51	-208.74	0.39	-200.49	insufficient sampling
1qlw	31-42	2.83	-	0.33	-1620.54	0.48	-	0.29	-1595.85	insufficient sampling / energy function deficiency
1tld	127-138	2.94	-215.11	0.30	-209.60	0.33	-220.36	0.22	-217.79	insufficient sampling
1thg	127-138	1.86	-	0.28	-1423.84	1.12	-	0.23	-1434.68	insufficient sampling / energy function deficiency
2tgi	48-59	3.14	-163.39	1.36	-87.48	0.41	-189.26	0.31	-182.44	insufficient sampling
4ilb	46-57	6.63	-140.98	1.04	-97.29	6.64	-140.95	0.81	-137.17	crystal packing

* bold numbers indicate cases where FKIC native input simulations also failed to correctly identify sub-Å models

Table 3.6 Multiple Segments (12 residues) dataset detailed performance

PDB name	Target segment residues	NGK					FKIC				
		RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models
lceo	10-21, 53-64	3.74	-786.28	2.57	-763.87	0.00	2.47	-801.47	1.60	-774.43	0.00
ldeu	47-58, 117-128	2.07	16572.20	1.27	16605.60	0.00	1.05	16537.10	0.53	16568.70	0.06
ldqz	146-157, 250-261	2.78	-657.72	0.89	-627.68	0.00	0.80	-656.04	0.80	-656.04	0.00
leuv	553-564, 527-538	1.35	-497.87	1.16	-480.99	0.00	2.59	-502.24	1.58	-495.34	0.00
lfr	209-220, 276-287	9.25	-717.41	2.57	-698.25	0.00	7.33	-723.89	2.19	-708.61	0.00
lh6u	168-179, 190-201	1.01	-715.67	1.01	-715.67	0.00	0.40	-745.35	0.26	-733.33	0.64
li7k	138-149, 66-77	1.80	-322.00	0.78	-312.85	0.00	0.47	-327.72	0.36	-321.18	0.30
lj7d	33-44, 67-78	2.30	-329.13	1.35	-323.97	0.00	1.98	-343.43	0.92	-333.10	0.01
ljfu	32-43, 134-145	5.49	-409.27	1.49	-391.57	0.00	1.29	-415.60	0.54	-400.58	0.05
lku1	677-688, 729-740	4.37	567.65	1.65	576.01	0.00	1.11	560.84	0.67	563.89	0.02
lm0z	199-210, 244-255	3.35	-576.11	1.60	-126.88	0.00	2.17	-583.89	1.12	27.73	0.00
lqsl	264-275, 333-344	1.98	-957.12	1.77	-923.44	0.00	1.67	-970.92	0.90	-957.65	0.00
lryl	54-65, 118-129	2.19	-363.80	1.73	-334.16	0.00	2.23	-362.49	1.75	-348.55	0.00
lsuu	741-752, 780-791	6.21	-711.24	2.66	-698.90	0.00	3.11	-708.22	2.53	-692.72	0.00
lt6g	195-206, 220-231	1.10	-765.81	0.85	-747.86	0.01	1.87	-771.86	0.62	-756.32	0.17
lu09	238-249, 291-302	2.59	-945.68	1.25	-916.78	0.00	2.91	-952.12	0.60	-929.50	0.00
lv5d	282-293, 308-319	6.08	-938.87	4.75	-895.23	0.00	7.20	-945.93	5.15	-920.17	0.00
lw0d	65-76, 274-285	5.47	-661.63	1.25	-534.61	0.00	0.65	-676.38	0.31	-650.39	0.44
lwko	38-49, 109-120	1.52	-353.30	0.97	-321.36	0.00	1.19	-357.76	0.46	-321.97	0.08
lxwt	367-378, 38-49	5.47	-974.75	1.52	-954.86	0.00	2.62	-995.65	1.07	-950.23	0.00
lxyz	792-803, 813-824	7.55	-786.12	3.18	-737.98	0.00	1.99	-800.04	0.65	-767.73	0.00
lyif	142-153, 174-185	5.15	-773.41	2.34	-712.48	0.00	3.84	-781.24	1.91	-702.07	0.00
lzvt	657-668, 714-725	0.35	-545.67	0.32	-545.44	0.03	1.44	-542.72	1.41	-526.26	0.00

Table 3.6 Multiple Segments (12 residues) dataset detailed performance cont.

		NGK					FKIC				
PDB name	Target segment residues	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models
2ahf	182-193, 208-219	6.52	-707.08	3.03	-269.19	0.00	4.51	-775.93	0.80	-371.41	0.00
2b49	674-685, 652-663	0.55	-574.78	0.55	-574.78	0.04	2.88	-579.10	0.81	-572.12	0.00
2c61	410-421, 164-175	3.31	-981.07	0.74	-973.54	0.00	5.14	-987.68	1.29	-970.39	0.00
2cyg	248-259, 294-305	0.72	-728.09	0.72	-728.09	0.02	1.22	-753.92	0.58	-723.71	0.55
2e01	87-98, 256-267	2.02	-909.67	1.61	-894.78	0.00	0.44	-946.66	0.31	-922.67	0.75
2g30	790-801, 822-833	1.45	-500.35	1.18	-486.01	0.00	3.35	-486.53	1.61	-483.88	0.00
2hju	299-310, 326-337	1.78	-390.05	1.08	-377.63	0.00	1.44	-393.43	1.35	-385.84	0.00

Table 3.7 Summary of energy function comparisons

Dataset	Sampling method	Rosetta energy function	Median RMSD of lowest energy model (Å)	Median RMSD of lowest RMSD model (Å)	Median RMSD all models (Å)	Median sub-Å fraction	Median time (s)
<i>Standard</i>	NGK	talaris2013	0.74	0.37	2.71	11.40%	2281
<i>Standard</i>	NGK	ref2015	0.64	0.37	2.70	13.00%	3642
<i>Standard</i>	FKIC	talaris2013	0.70	0.36	1.19	44.89%	1646
<i>Standard</i>	FKIC	talaris2014	0.64	0.35	1.27	46.20%	1813
<i>Standard</i>	FKIC	ref2015	0.62	0.32	1.16	47.80%	3456
<i>Mixed</i>	NGK	talaris2013	1.94	0.45	4.66	1.90%	3788
<i>Mixed</i>	NGK	ref2015	1.07	0.45	4.65	1.15%	7341
<i>Mixed</i>	FKIC	talaris2013	0.61	0.34	1.74	34.60%	3654
<i>Mixed</i>	FKIC	ref2015	0.53	0.34	1.46	52.30%	7196

Table 3.8 Multiple Segments (8 residues) dataset detailed performance

		NGK					FKIC				
PDB name	Target segment residues	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models
1a8d	278-285,303-310	1.15	-998.03	0.35	-993.09	0.10	0.94	-983.98	0.73	-978.66	0.01
1a8u	225-232,251-258	0.94	-	1073.51	-1073.51	0.00	0.65	1081.86	0.58	-1036.74	0.42
1ako	150-157,173-180	0.45	-668.54	0.26	-666.92	0.66	0.43	-670.50	0.29	-662.28	0.68
1bhe	282-289,344-351	0.51	-877.73	0.28	-837.52	0.93	0.68	-875.49	0.38	-865.22	0.24
1bn8	250-257,338-345	1.98	-963.94	0.62	-959.38	0.05	2.45	-967.40	0.68	-949.27	0.00
1brt	28-35,205-212	0.76	-510.31	0.73	-500.12	0.21	0.81	-609.85	0.33	-488.96	0.94
1c5e	49-56,97-104	0.45	-702.21	0.29	-695.39	0.66	0.44	-703.36	0.26	-697.43	0.67
1cil	39-46,80-87	0.57	178.15	0.46	199.82	0.96	0.64	187.24	0.38	217.96	1.00
1cs6	126-133,158-165	1.10	-741.75	0.33	-719.69	0.57	0.72	-742.82	0.43	-741.61	0.73
1dqz	135-142,111-118	0.23	-	1264.84	-1221.96	0.98	0.23	1262.80	0.21	-1149.68	0.99
1ede	52-59,283-290	0.79	-689.00	0.69	-680.74	0.46	0.47	-694.65	0.32	-684.81	0.84
1exm	305-312,255-262	0.95	-998.75	0.30	-980.32	0.22	0.38	-994.71	0.31	-970.07	0.57
1gai	282-289,401-408	0.65	-306.61	0.26	-292.27	0.25	0.33	-294.19	0.28	-288.39	0.63
1gof	7-14,31-38	0.80	-	1253.43	-1216.17	0.22	0.87	1228.70	0.66	-1174.50	0.41
1jev	451-458,147-154	0.79	-	1258.90	-1242.82	0.86	0.88	1247.44	0.37	-1239.03	0.97
1ms9	433-440,468-475	2.27	-	2629.19	-2392.50	0.14	2.74	2411.81	0.29	-2332.46	0.02
1nif	64-71,221-228	1.49	-722.30	0.93	-643.85	0.00	1.50	-726.32	0.51	-651.66	0.14

Table 3.8 Multiple Segments (8 residues) dataset detailed performance cont.

		NGK					FKIC				
PDB name	Target segment residues	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models	RMSD of lowest energy model (Å)	Lowest energy (REU)	Lowest RMSD (Å)	Energy of lowest RMSD model (REU)	Fraction sub-Å models
1oth	219-226,280-287	0.30	-671.55	0.22	-640.91	1.00	0.27	-654.61	0.22	-649.22	1.00
1qlw	130-137,67-74	0.40	-	1600.69	-1560.94	1.00	0.37	-	1596.93	-1592.87	1.00
1srp	260-267,294-301	0.26	-760.01	0.19	-758.04	0.55	0.25	-760.73	0.19	-758.14	0.45
1tad	109-116,159-166	0.65	-	1951.28	-1945.22	0.98	0.64	-	1964.78	-1947.35	0.76
1thg	351-358,307-314	2.63	-979.60	2.07	-928.81	0.00	3.48	-977.31	1.80	-935.77	0.00
1tib	171-178,211-218	1.08	-466.52	0.67	-440.39	0.37	0.54	-471.50	0.29	-460.77	0.53
1udc	32-39,78-85	0.62	-668.79	0.54	-660.04	0.06	0.59	-667.57	0.46	-660.73	0.19
3bto	256-263,280-287	1.15	-	3377.16	-3110.83	0.19	1.07	-	3405.40	-3274.16	0.06
3grs	131-138,292-299	1.06	-998.99	0.29	-916.54	0.34	1.52	-	1000.24	-929.86	0.67
4pga	112-119,216-223	0.60	-	1315.25	-1309.43	0.97	0.57	-	1380.06	-1313.87	1.00
6cel	132-139,367-374	1.36	-588.32	0.97	-568.66	0.00	1.70	-584.39	0.81	-573.68	0.00

Table 3.9 Comparison of median RMSD of lowest energy models on perturbed structures

Method	8 residue side chain perturbed (Å)	12 residue side chain perturbed (Å)	12 residue backbone perturbed (Å)	template based models (Å)
HLP*	2.2	2.25	-	-
HLP-SS*	0.85	1.15	-	-
NGK*	0.4	0.75	-	3.9
Galaxy-PS1*	1.45	3.05	-	3.5
Galaxy-PS2*	1.05	1.55	1.65	3.3
FKIC	0.45	0.64	1.68	3.9
FKIC with homologous fragments	0.42	0.54	-	-

* Values reported by (Park, et al., 2014)

References

- Adams, P. D. et al., 2012. Phenix - a comprehensive python-based system for macromolecular structure solution. In: s.l.:s.n., pp. 539-547.
- Afonine, P. V. et al., 2012. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D-biological Crystallography*, 68(4), pp. 352-367.
- Alford, R. F., Fleming, P. J., Fleming, K. G. & Gray, J. J., 2020. Protein Structure Prediction and Design in a Biologically Realistic Implicit Membrane.. *Biophysical Journal*, 118(8), pp. 2042-2055.
- Alford, R. F. et al., 2017. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, 13(6), pp. 3031-3048.
- Alford, R. F. et al., 2015. An Integrated Framework Advancing Membrane Protein Modeling and Design.. *PLOS Computational Biology*, 11(9).
- Allen, B. D. & Mayo, S. L., 2009. An efficient algorithm for multistate protein design based on FASTER.. *Journal of Computational Chemistry*, 31(5), pp. 904-916.
- Alley, E. C. et al., 2019. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16(12), pp. 1315-1322.
- Altschul, S. F. et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.. *Nucleic Acids Research*, 25(17), pp. 3389-3402.
- Ambroggio, X. I. & Kuhlman, B. A., 2006. Computational design of a single amino acid sequence that can switch between two distinct protein folds.. *Journal of the American Chemical Society*, 128(4), pp. 1154-1161.

- Anand, N., Eguchi, R. & Huang, P.-S., 2019. *Fully differentiable full-atom protein backbone generation*. s.l., s.n.
- Anand, N. et al., 2020. Protein Sequence Design with a Learned Potential. *bioRxiv*.
- Anfinsen, C. B., 1973. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096), pp. 223-230.
- Anishchenko, I. et al., 2020. De novo protein design by deep network hallucination. *bioRxiv*.
- Ashworth, J. et al., 2006. Computational redesign of endonuclease DNA binding and cleavage specificity.. *Nature*, 441(7093), pp. 656-659.
- Baker, D., 2010. An exciting but challenging road ahead for computational enzyme design.. *Protein Science*, 19(10), pp. 1817-1819.
- Baker, D., 2019. What has de novo protein design taught us about protein folding and biophysics. *Protein Science*, 28(4), pp. 678-683.
- Basanta, B. et al., 2020. An enumerative algorithm for de novo design of proteins with diverse pocket structures.. *Proceedings of the National Academy of Sciences of the United States of America*, 117(36), pp. 22135-22145.
- Berman, H., Henrick, K. & Nakamura, H., 2003. Announcing the worldwide Protein Data Bank.. *Nature Structural & Molecular Biology*, 10(12), pp. 980-980.
- Berman, H. M. et al., 2000. The Protein Data Bank. *Nucleic Acids Research*, 28(1), pp. 235-242.
- Bernstein, F. C. et al., 1978. The Protein Data Bank: a computer-based archival file for macromolecular structures.. *Archives of Biochemistry and Biophysics*, 185(2), pp. 584-591.
- Bhardwaj, G. et al., 2016. Accurate de novo design of hyperstable constrained peptides.. *Nature*, 538(7625), pp. 329-335.

Bhattacharya, A., Tejero, R. & Montelione, G. T., 2006. Evaluating protein structures determined by structural genomics consortia.. *Proteins*, 66(4), pp. 778-795.

Bick, M. J. et al., 2017. Computational design of environmental sensors for the potent opioid fentanyl.. *eLife*, Volume 6, pp. 1-23.

Blomberg, R. et al., 2013. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature*, 503(7476), pp. 418-421.

Bonet, J. et al., 2018. Rosetta FunFoldes — A general framework for the computational design of functional proteins. *PLOS Computational Biology*, 14(11).

Bowers, P. M., Strauss, C. E. & Baker, D., 2000. De novo protein structure determination using sparse NMR data.. *Journal of Biomolecular NMR*, 18(4), pp. 311-318.

Boyken, S. E. et al., 2019. De novo design of tunable, pH-driven conformational changes. *Science*, 364(6441), pp. 658-664.

Boyken, S. E. et al., 2016. De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science*, 352(6286), pp. 680-687.

Bradley, P., Misura, K. M. S. & Baker, D., 2005. Toward high-resolution de novo structure prediction for small proteins.. *Science*, 309(5742), pp. 1868-1871.

Brooks, B. R. et al., 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2), pp. 187-217.

Brunette, T. et al., 2020. Modular repeat protein sculpting using rigid helical junctions.. *Proceedings of the National Academy of Sciences of the United States of America*, 117(16), pp. 8870-8875.

Brunette, T. J. et al., 2015. Exploring the repeat protein universe through computational protein design.. *Nature*, 528(7583), pp. 580-584.

- Brünger, A. et al., 1998. Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Crystallographica Section D-biological Crystallography*, 54(5), pp. 905-921.
- Canutescu, A. A. & Dunbrack, R. L., 2003. Cyclic coordinate descent: A robotics algorithm for protein loop closure.. *Protein Science*, 12(5), pp. 963-972.
- Cao, L. et al., 2020. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors.. *Science*, 370(6515), pp. 426-431.
- Chaudhury, S., Lyskov, S. & Gray, J. J., 2010. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26(5), pp. 689-691.
- Chen, V. B. et al., 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D-biological Crystallography*, 66(1), pp. 12-21.
- Chen, Z. et al., 2019. Programmable design of orthogonal protein heterodimers. *Nature*, 565(7737), pp. 106-111.
- Chen, Z. et al., 2019. Self-Assembling 2D Arrays with de Novo Protein Building Blocks.. *Journal of the American Chemical Society*, 141(22), pp. 8891-8895.
- Chen, Z. et al., 2020. De novo design of protein logic gates. *Science*, 368(6486), pp. 78-84.
- Cheung, M.-S., Maguire, M. L., Stevens, T. J. & Broadhurst, R. W., 2010. DANGLE: A Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure.. *Journal of Magnetic Resonance*, 202(2), pp. 223-233.
- Chevalier, A. et al., 2017. Massively parallel de novo protein design for targeted therapeutics. *Nature*, 550(7674), pp. 74-79.
- Choma, C. T. et al., 1994. Design of a heme-binding four-helix bundle. *Journal of the American Chemical Society*, 116(3), pp. 856-865.

Conchúir, S. Ó. et al., 2015. A Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling and Design. *PLOS ONE*, 10(9).

Conway, P. & DiMaio, F., 2016. Improving hybrid statistical and physical forcefields through local structure enumeration.. *Protein Science*, 25(8), pp. 1525-1534.

Cooper, S. et al., 2010. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307), pp. 756-760.

Cornell, W. D. et al., 1995. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, 117(19), pp. 5179-5197.

Coutsias, E. A., Seok, C., Jacobson, M. P. & Dill, K. A., 2004. A kinematic view of loop closure.. *Journal of Computational Chemistry*, 25(4), pp. 510-528.

Cp, M. & Kg, F., 2011. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proceedings of the National Academy of Sciences of the United States of America*, 108(25), pp. 10174-10177.

Crick, F. H. C., 1953. The Fourier transform of a coiled-coil. *Acta Crystallographica*, 6(8), pp. 685-689.

Dang, B. et al., 2017. De novo design of covalently constrained mesosize protein scaffolds with unique tertiary structures.. *Proceedings of the National Academy of Sciences of the United States of America*, 114(41), pp. 10852-10857.

Das, R. & Baker, D., 2007. Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences of the United States of America*, 104(37), pp. 14664-14669.

- Das, R., Karanicolas, J. & Baker, J. D., 2010. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature Methods*, 7(4), pp. 291-294.
- Davey, J. A. & Chica, R. A., 2017. Multistate Computational Protein Design with Backbone Ensembles.. *Methods of Molecular Biology*, Volume 1529, pp. 161-179.
- Davey, J. A., Damry, A. M., Goto, N. K. & Chica, R. A., 2017. Rational design of proteins that exchange on functional timescales.. *Nature Chemical Biology*, 13(12), pp. 1280-1285.
- Davidson, K. et al., 2019. Deep generative models for T cell receptor protein sequences. *eLife*, Volume 8.
- Dawson, N. L. et al., 2017. CATH: an expanded resource to predict protein function through structure and sequence.. *Nucleic Acids Research*, Volume 45.
- Delaglio, F. et al., 1995. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR*, 6(3), pp. 277-293.
- Desmet, J., Maeyer, M. D., Hazes, B. & Lasters, I., 1992. THE DEAD-END ELIMINATION THEOREM AND ITS USE IN PROTEIN SIDE-CHAIN POSITIONING. *Nature*, 356(6369), pp. 539-542.
- Dou, J. et al., 2017. Sampling and Energy Evaluation Challenges in Ligand Binding Protein Design. *Protein Science*, 26(12), pp. 2426-2437.
- Dou, J. et al., 2018. De novo design of a fluorescence-activating β -barrel.. *Nature*, 561(7724), pp. 485-491.
- Doyle, L. et al., 2015. Rational design of α -helical tandem repeat proteins with closed architectures. *Nature*, 528(7583), pp. 585-588.
- Dunbrack, R. L. & Cohen, F. E., 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science*, 6(8), pp. 1661-1681.

- Eguchi, R. R., Anand, N., Choe, C. A. & Huang, P.-S., 2020. IG-VAE: Generative Modeling of Immunoglobulin Proteins by Direct 3D Coordinate Generation. *bioRxiv*.
- Evans, P., 2006. Scaling and assessment of data quality. *Acta Crystallographica Section D-biological Crystallography*, 62(1), pp. 72-82.
- Eyal, E. et al., 2005. The limit of accuracy of protein modeling: influence of crystal packing on protein structure.. *Journal of Molecular Biology*, 351(2), pp. 431-442.
- Fallas, J. A. et al., 2017. *Computational design of self-assembling cyclic protein homo-oligomers.. s.l.:s.n.*
- Feldmeier, K. & Höcker, B., 2013. Computational protein design of ligand binding and catalysis.. *Current Opinion in Chemical Biology*, 17(6), pp. 929-933.
- Fox, N. K., Brenner, S. E. & Chandonia, J.-M., 2014. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, Volume 42, pp. 304-309.
- Frappier, V. et al., 2019. Tertiary Structural Motif Sequence Statistics Enable Facile Prediction and Design of Peptides that Bind Anti-apoptotic Bfl-1 and Mcl-1.. *Structure*, 27(4), p. 606.
- Glasgow, A. A. et al., 2019. Computational design of a modular protein sense-response system. *Science*, 366(6468), pp. 1024-1028.
- Gonen, S., DiMaio, F., Gonen, T. & Baker, D., 2015. Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces.. *Science*, 348(6241), pp. 1365-1368.
- Gront, D. et al., 2011. Generalized fragment picking in Rosetta: design, protocols and applications.. *PLOS ONE*, 6(8).
- Harbury, P. B. et al., 1998. High-Resolution Protein Design with Backbone Freedom. *Science*, 282(5393), pp. 1462-1467.

Havranek, J. J., Duarte, C. M. & Baker, D., 2004. A Simple Physical Model for the Prediction and Design of Protein–DNA Interactions. *Journal of Molecular Biology*, 344(1), pp. 59-70.

Hawkins-Hooker, A. et al., 2020. Generating functional protein variants with variational autoencoders. *bioRxiv*.

Helling, R. et al., 2001. The designability of protein structures.. *Journal of Molecular Graphics & Modelling*, 19(1), pp. 157-167.

Hill, R. B., Raleigh, D. P., Lombardi, a. A. & DeGrado, W. F., 2000. De novo design of helical bundles as models for understanding protein folding and function.. *Accounts of Chemical Research*, 33(11), pp. 745-754.

Ho, S. P. & DeGrado, W. F., 1987. Design of a 4-helix bundle protein: synthesis of peptides which self-associate into a helical protein. *Journal of the American Chemical Society*, 109(22), pp. 6751-6758.

Hosseinzadeh, P. et al., 2017. Comprehensive computational design of ordered peptide macrocycles.. *Science*, 358(6369), pp. 1461-1466.

Hou, J., Jun, S.-R., Zhang, C. & Kim, S.-H., 2005. Global mapping of the protein structure space and application in structure-based inference of protein function. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10), pp. 3651-3656.

Hsia, Y. et al., 2016. Design of a hyperstable 60-subunit protein icosahedron. *Nature*, 535(7610), pp. 136-139.

Huang, P.-S. et al., 2016. De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nature Chemical Biology*, 12(1), pp. 29-34.

Huang, P. S., Boyken, S. E. & Baker, D., 2016. The coming of age of de novo protein design. *Nature*, 537(7620), pp. 320-327.

- Huang, P. S. et al., 2014. High thermodynamic stability of parametrically designed helical bundles.. *Science*, 346(6208), pp. 481-485.
- Hu, X., Wang, H., Ke, H. & Kuhlman, B. A., 2007. High-resolution design of a protein loop. *Proceedings of the National Academy of Sciences of the United States of America*, 104(45), pp. 17668-17673.
- Ingraham, J., Garg, V. K., Barzilay, R. & Jaakkola, T. S., 2019. *Generative Models for Graph-Based Protein Design*. s.l., s.n., pp. 15820-15831.
- Jacobson, M. P., Kaminski, G. A., Friesner, R. A. & Rapp, C. S., 2002. Force Field Validation Using Protein Side Chain Prediction. *Journal of Physical Chemistry B*, 106(44), pp. 11673-11680.
- Jacobson, M. P. et al., 2004. A hierarchical approach to all-atom protein loop prediction.. *Proteins*, 55(2), pp. 351-367.
- Jacobs, T. M. et al., 2016. Design of structurally distinct proteins using strategies inspired by evolution.. *Science*, 352(6286), pp. 687-690.
- Joh, N. H. et al., 2014. De novo design of a transmembrane Zn²⁺-transporting four-helix bundle.. *Science*, 346(6216), pp. 1520-1524.
- Joo, H. et al., 2012. An amino acid packing code for α -helical structure and protein design.. *Journal of Molecular Biology*, 419(3), pp. 234-254.
- Joosten, R. P. et al., 2011. A series of PDB related databases for everyday needs.. *Nucleic Acids Research*, 39(1), pp. 364-368.
- Jorgensen, W. L. & Tirado-Rives, J., 1988. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin.. *Journal of the American Chemical Society*, 110(6), pp. 1657-1666.

- Kabsch, W., 2010. XDS. *Acta Crystallogr. D*, Volume 66, pp. 125-132.
- Kabsch, W. & Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), pp. 2577-2637.
- Kamtekar, S. et al., 1993. Protein design by binary patterning of polar and nonpolar amino acids. *Science*, 262(5140), pp. 1680-1685.
- Kantardjieff, K. A. & Rupp, B., 2003. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein–nucleic acid complex crystals. *Protein Science*, 12(9), pp. 1865-1871.
- Karimi, M., Zhu, S., Cao, Y. & Shen, Y., 2019. De Novo Protein Design for Novel Folds using Guided Conditional Wasserstein Generative Adversarial Networks (gcWGAN). *bioRxiv*, p. 769919.
- King, N. P. et al., 2014. Accurate design of co-assembling multi-component protein nanomaterials.. *Nature*, 510(7503), pp. 103-108.
- Koepnick, B. et al., 2019. De novo protein design by citizen scientists.. *Nature*, 570(7761), pp. 390-394.
- Koga, N. et al., 2012. Principles for designing ideal protein structures. *Nature*, 491(7423), pp. 222-227.
- Korendovych, I. V. & DeGrado, W. F., 2020. De novo protein design, a retrospective.. *Quarterly Reviews of Biophysics*, Volume 53.
- Korendovych, I. V. et al., 2010. De novo design and molecular assembly of a transmembrane diporphyrin-binding protein complex.. *Journal of the American Chemical Society*, 132(44), pp. 15516-15518.

- Kosuri, S. & Church, G. M., 2014. Large-scale de novo DNA synthesis: technologies and applications. *Nature Methods*, 11(5), pp. 499-507.
- Kuhlman, B. A. et al., 2003. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science*, 302(5649), pp. 1364-1368.
- Kuhlman, B. & Baker, D., 2000. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences of the United States of America*, 97(19), pp. 10383-10388.
- Kuhlman, B. & Bradley, P., 2019. Advances in protein structure prediction and design.. *Nature Reviews Molecular Cell Biology*, 20(11), pp. 681-697.
- Kundert, K. & Kortemme, T., 2019. Computational design of structured loops for new protein functions.. *Biological Chemistry*, 400(3), pp. 275-288.
- Labonte, J. W., Adolf-Bryfogle, J., Schief, W. R. & Gray, J. J., 2017. Residue-centric modeling and design of saccharide and glycoconjugate structures. *Journal of Computational Chemistry*, 38(5), pp. 276-287.
- Lajoie, M. J. et al., 2020. Designed protein logic to target cells with precise combinations of surface antigens.. *Science*, 369(6511), pp. 1637-1643.
- Langan, R. A. et al., 2019. De novo design of bioactive protein switches. *Nature*, 572(7768), pp. 205-210.
- Lawrence, M. C. & Colman, P. M., 1993. Shape Complementarity at Protein/Protein Interfaces. *Journal of Molecular Biology*, 234(4), pp. 946-950.
- Lear, J., Wasserman, Z. & DeGrado, W., 1988. Synthetic amphiphilic peptide models for protein ion channels. *Science*, 240(4856), pp. 1177-1181.

- Leaver-Fay, A. P. et al., 2013. Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. In: s.l.:s.n., pp. 109-143.
- Leaver-Fay, A. et al., 2011. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules.. In: s.l.:s.n., pp. 545-574.
- Lee, C. & Subbiah, S., 1991. *Prediction of protein side-chain conformation by packing optimization*. s.l.:s.n.
- Lee, J. et al., 2010. Protein loop modeling by using fragment assembly and analytical loop closure.. *Proteins*, 78(16), pp. 3428-3436.
- Leman, J. K., Lyskov, S. & Bonneau, R., 2017. Computing structure-based lipid accessibility of membrane proteins with mp_lipid_acc in RosettaMP.. *BMC Bioinformatics*, 18(1), p. 115.
- Leman, J. K. et al., 2020. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods*, 17(7), pp. 665-680.
- Levitt, M. & Lifson, S., 1969. Refinement of protein conformations using a macromolecular energy minimization procedure.. *Journal of Molecular Biology*, 46(2), pp. 269-279.
- Li, H., Helling, R., Tang, C. & Wingreen, N. S., 1996. Emergence of Preferred Structures in a Simple Model of Protein Folding. *Science*, 273(5275), pp. 666-669.
- Lin, Y.-R. et al., 2015. Control over overall shape and size in de novo designed proteins.. *Proceedings of the National Academy of Sciences of the United States of America*, 112(40), p. 201509508.
- Lipsh-Sokolik, R., Listov, D. & Fleishman, S. J., 2020. The AbDesign computational pipeline for modular backbone assembly and design of binders and enzymes.. *Protein Science*.
- López-Blanco, J. R., Canosa-Valls, A. J., Li, Y. & Chacón, P., 2016. RCD+: Fast loop modeling server. *Nucleic Acids Research*, Volume 44.

Loshbaugh, A. L. & Kortemme, T., 2020. Comparison of Rosetta flexible-backbone computational protein design methods on binding interactions.. *Proteins*, 88(1), pp. 206-226.

Lucas, J. E. & Kortemme, T., 2020. New computational protein design methods for de novo small molecule binding sites.. *PLOS Computational Biology*, 16(10).

Lu, P. et al., 2018. Accurate computational design of multipass transmembrane proteins.. *Science*, 359(6379), pp. 1042-1046.

MacDonald, J. T. et al., 2016. Synthetic beta-solenoid proteins with the fragment-free computational design of a beta-hairpin extension.. *Proceedings of the National Academy of Sciences of the United States of America*, 113(37), pp. 10346-10351.

Mackenzie, C. O., Zhou, J. & Grigoryan, G., 2016. Tertiary alphabet for the observable protein structural universe. *Proceedings of the National Academy of Sciences of the United States of America*, 113(47), p. 201607178.

Maguire, J. B., Boyken, S. E., Baker, D. & Kuhlman, B. A., 2018. Rapid Sampling of Hydrogen Bond Networks for Computational Protein Design.. *Journal of Chemical Theory and Computation*, 14(5), pp. 2751-2760.

Mandell, D. J., Coutsiaris, E. A. & Kortemme, T., 2009. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature Methods*, 6(8), pp. 551-552.

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

Marcos, E. et al., 2018. De novo design of a non-local beta-sheet protein with high stability and accuracy. *Nature Structural & Molecular Biology*, 25(11), pp. 1028-1034.

Marcos, E. & Silva, D., 2018. Essentials of de novo protein design: Methods and applications. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(6).

- Marks, C., Shi, J. & Deane, C. M., 2018. Predicting loop conformational ensembles.. *Bioinformatics*, 34(6), pp. 949-956.
- McCoy, A. J. et al., 2007. Phaser crystallographic software. *Journal of Applied Crystallography*, 40(4), pp. 658-674.
- Mohan, K. et al., 2019. Topological control of cytokine receptor signaling induces differential effects in hematopoiesis.. *Science*, 364(6442).
- Moriarty, N. W., Grosse-Kunstleve, R. W. & Adams, P. D., 2009. electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta Crystallographica Section D-biological Crystallography*, 65(10), pp. 1074-1080.
- Mravic, M. et al., 2019. Packing of apolar side chains enables accurate design of highly stable membrane proteins.. *Science*, 363(6434), pp. 1418-1423.
- Mulligan, V. K. et al., 2019. Designing Peptides on a Quantum Computer. *bioRxiv*, p. 752485.
- Murphy, P. M. et al., 2009. Alteration of enzyme specificity by computational loop remodeling and design.. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23), pp. 9215-9220.
- Nivedha, A. K. et al., 2014. Importance of ligand conformational energies in carbohydrate docking: Sorting the wheat from the chaff.. *Journal of Computational Chemistry*, 35(7), pp. 526-539.
- Norn, C. et al., 2020. Protein sequence design by explicit energy landscape optimization. *bioRxiv*.
- North, B., Lehmann, A. & Dunbrack, R. L., 2011. A new clustering of antibody CDR loop conformations. *Journal of Molecular Biology*, 406(2), pp. 228-256.

O'Meara, M. J. et al., 2015. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *Journal of Chemical Theory and Computation*, 11(2), pp. 609-622.

Ollikainen, N., Smith, C. A., Fraser, J. S. & Kortemme, T., 2013. Flexible backbone sampling methods to model and design protein alternative conformations.. In: s.l.:s.n., pp. 61-85.

Pan, X. et al., 2020. Expanding the space of protein geometries by computational design of de novo fold families.. *Science*, 369(6507), pp. 1132-1136.

Park, H. et al., 2016. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *Journal of Chemical Theory and Computation*, 12(12), pp. 6201-6212.

Park, H., Lee, G. R., Heo, L. & Seok, C., 2014. Protein loop modeling using a new hybrid energy function and its application to modeling in inaccurate structural environments.. *PLOS ONE*, 9(11).

Park, H. et al., 2020. Learning a force field from small-molecule crystal lattice predictions enables consistent sub-Angstrom protein-ligand docking. *bioRxiv*.

Park, K. et al., 2015. Control of repeat-protein curvature by computational protein design.. *Nature Structural & Molecular Biology*, 22(2), pp. 167-174.

Polizzi, N. F. & DeGrado, W. F., 2020. A defined structural unit enables de novo design of small-molecule-binding proteins. *Science*, 369(6508), pp. 1227-1233.

Polizzi, N. F. et al., 2017. De novo design of a hyperstable non-natural protein-ligand complex with sub-Å accuracy. *Nature Chemistry*, 9(12), pp. 1157-1164.

Quijano-Rubio, A. et al., 2020. De novo design of modular and tunable allosteric biosensors.. *bioRxiv*.

- Rapp, C. S. & Pollack, R. M., 2005. Crystal packing effects on protein loops.. *Proteins*, 60(1), pp. 103-109.
- Raveh, B., London, N. & Schueler-Furman, O., 2010. Sub-angstrom modeling of complexes between flexible peptides and globular proteins.. *Proteins*, 78(9), pp. 2029-2040.
- Repecka, D. et al., 2019. Expanding functional protein sequence space using generative adversarial networks. *bioRxiv*, p. 789719.
- Rieping, W. et al., 2007. ARIA2: Automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics*, 23(3), pp. 381-382.
- Robertson, D. E. et al., 1994. Design and synthesis of multi-haem proteins. *Nature*, 368(6470), pp. 425-432.
- Rocklin, G. J. et al., 2017. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347), pp. 168-175.
- Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D., 2004. Protein structure prediction using Rosetta.. In: s.l.:s.n., pp. 66-93.
- Sellers, B. D. et al., 2008. Toward better refinement of comparative models: Predicting loops in inexact environments. *Proteins*, 72(3), pp. 959-971.
- Senior, A. W. et al., 2020. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792), pp. 706-710.
- Shapovalov, M. V. & Dunbrack, R. L., 2011. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6), pp. 844-858.
- Sheffler, W. & Baker, D., 2010. RosettaHoles2: A volumetric packing measure for protein structure refinement and validation. *Protein Science*, 19(10), pp. 1991-1995.

- Shen, H. et al., 2018. De novo design of self-assembling helical protein filaments.. *Science*, 362(6415), pp. 705-709.
- Silva, D.-A. et al., 2019. De novo design of potent and selective mimics of IL-2 and IL-15. *Nature*, 565(7738), pp. 186-191.
- Simons, K. T., Kooperberg, C., Huang, E. & Baker, D., 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, 268(1), pp. 209-225.
- Skinner, S. P. et al., 2015. Structure calculation, refinement and validation using CcpNmr Analysis. *Acta Crystallographica Section D-biological Crystallography*, 71(1), pp. 154-161.
- Smith, C. A. & Kortemme, T., 2008. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction.. *Journal of Molecular Biology*, 380(4), pp. 742-756.
- Stein, A. & Kortemme, T., 2013. Improvements to Robotics-Inspired Conformational Sampling in Rosetta. *PLOS ONE*, 8(5).
- Strokach, A. et al., 2019. Designing real novel proteins using deep graph neural networks. *bioRxiv*, p. 868935.
- Studier, F. W., 2005. Protein production by auto-induction in high-density shaking cultures. *Protein Expression and Purification*, 41(1), pp. 207-234.
- Tanaka, S. & Scheraga, H. A., 1976. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins.. *Macromolecules*, 9(6), pp. 945-950.
- Tan, S., Tan, H. T. & Chung, M. C. M., 2008. Membrane proteins and membrane proteomics. *Proteomics*, 8(19), pp. 3924-3932.

- Thomson, A. R. et al., 2014. Computational design of water-soluble α -helical barrels. *Science*, 346(6208), pp. 485-488.
- Tinberg, C. E. et al., 2013. Computational design of ligand-binding proteins with high affinity and selectivity.. *Nature*, 501(7466), pp. 212-216.
- Tyka, M. D., Jung, K. & Baker, D., 2012. Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers.. *Journal of Computational Chemistry*, 33(31), pp. 2483-2491.
- Votteler, J. et al., 2016. Designed proteins induce the formation of nanocage-containing extracellular vesicles. *Nature*, 540(7632), pp. 292-295.
- Vranken, W. F. et al., 2005. The CCPN data model for NMR spectroscopy: Development of a software pipeline. *Proteins*, 59(4), pp. 687-696.
- Wallin, E. & Heijne, G. v., 1998. GENOME-WIDE ANALYSIS OF INTEGRAL MEMBRANE PROTEINS FROM EUBACTERIAL, ARCHAEAN, AND EUKARYOTIC ORGANISMS. *Protein Science*, 7(4), pp. 1029-1038.
- Wang, C., Bradley, P. & Baker, D., 2007. Protein–Protein Docking with Backbone Flexibility. *Journal of Molecular Biology*, 373(2), pp. 503-519.
- Wang, J., Cao, H., Zhang, J. Z. H. & Qi, Y., 2018. Computational Protein Design with Deep Learning Neural Networks. *Scientific Reports*, 8(1), pp. 6349-6349.
- Ward, J. J., McGuffin, L. J., Buxton, B. F. & Jones, D. T., 2003. Secondary structure prediction with support vector machines.. *Bioinformatics*, 19(13), pp. 1650-1655.
- Weichenberger, C. X. & Rupp, B., 2014. Ten years of probabilistic estimates of biocrystal solvent content: new insights via nonparametric kernel density estimate.. *Acta Crystallographica Section D-biological Crystallography*, 70(6), pp. 1579-1588.

- Wei, K. Y. et al., 2020. Computational design of closely related proteins that adopt two well-defined but structurally divergent folds.. *Proceedings of the National Academy of Sciences of the United States of America*, 117(13), pp. 7208-7215.
- Whitley, P., Nilsson, I. & Heijne, G. v., 1994. De novo design of integral membrane proteins.. *Nature Structural & Molecular Biology*, 1(12), pp. 858-862.
- Winter, G., 2010. xia2: an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography*, 43(1), pp. 186-190.
- Wishart, D. S. et al., 1995. ¹H, ¹³C and ¹⁵N chemical shift referencing in biomolecular NMR.. *Journal of Biomolecular NMR*, 6(2), pp. 135-140.
- Woolfson, D. N., 2017. Coiled-Coil Design: Updated and Upgraded. *Sub-cellular biochemistry*, Volume 82, pp. 35-61.
- Xu, C. et al., 2020. Computational design of transmembrane pores.. *Nature*, 585(7823), pp. 129-134.
- Xu, J., 2019. Distance-based protein folding powered by deep learning.. *Proceedings of the National Academy of Sciences of the United States of America*, 116(34), pp. 16856-16865.
- Yang, C. et al., 2020. A bottom-up approach for the de novo design of functional proteins. *bioRxiv*.
- Yang, J. et al., 2020. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences of the United States of America*, 117(3), pp. 1496-1503.
- Yang, Y., Faraggi, E., Zhao, H. & Zhou, Y., 2011. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-

dimensional structural properties of query and corresponding native properties of templates.

Bioinformatics, 27(15), pp. 2076-2082.

Yarov-Yarovoy, V., Schonbrun, J. & Baker, D., 2005. Multipass membrane protein structure prediction using Rosetta.. *Proteins*, 62(4), pp. 1010-1025.

Zhang, C., Freddolino, P. L. & Zhang, Y., 2017. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Research*, Volume 45.

Zhang, S.-Q.et al., 2018. Designed peptides that assemble into cross-alpha amyloid-like structures.. *Nature Chemical Biology*, 14(9), pp. 870-875.

Zhang, Y. & Skolnick, J., 2005. TM-align: a protein structure alignment algorithm based on the TM-score.. *Nucleic Acids Research*, 33(7), pp. 2302-2309.

Zhao, S., Zhu, K., Li, J. & Friesner, R. A., 2011. Progress in super long loop prediction.. *Proteins*, 79(10), pp. 2920-2935.

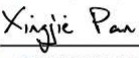
Zheng, F., Zhang, J. & Grigoryan, G., 2015. Tertiary Structural Propensities Reveal Fundamental Sequence/Structure Relationships. *Structure*, 23(5), pp. 961-971.

Zhou, J. & Grigoryan, G., 2015. Rapid search for tertiary fragments reveals protein sequence–structure relationships. *Protein Science*, 24(4), pp. 508-524.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

296355AC8C134E5... Author Signature

11/30/2020
Date