Beyond single-protein de novo design: A generative algorithm for the NTF2-like superfamily

Benjamin Basanta

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

David Baker, Chair

Frank Dimaio

Philip Bradley

Program Authorized to Offer Degree:

Biochemistry

University of Washington

Abstract

Beyond single-protein *de novo* design: A generative algorithm for the NTF2-like superfamily

Benjamin Basanta

Chair of the supervisory committee:

Professor David Baker

Biochemistry

Natural proteins evolved over billions of years to regulate cellular growth, ward off infection and capture and store solar energy. Proteins thus serve as the molecular basis for life. The promise of protein design is to use nature's favorite toolbox to solve modern human problems without having to wait for the long and meandering path of natural selection. Protein structure determines function, so it is not surprising proteins sample a great variety of structures. Thus, it is reasonable to expect that designing proteins with functions not seen in nature would require access to comparable structural diversity, more specifically, diversity of active site structure. Despite the advances in *de novo* protein design, the systematic generation of proteins containing pockets that can harbor substrates has been lacking.

The use of a natural small alpha-beta fold, the Nuclear Transport Factor 2-like (NTF2-like) fold, to design a high affinity small-molecule binding protein, and the diversity observed in that family, have posed the idea that significant pocket structural diversity could be derived from this relatively simple, small, alpha-beta fold.

To explore this idea, we analyzed the structures of proteins belonging to the NTF2-like superfamily and other proteins with similar characteristics to understand the determinants of their structural diversity. The most salient feature of the NTF2-like superfamily is the curved beta-sheet

that forms most of their pockets in its concave face. Curved beta sheets depart from the classic beta pleated structure displaying bulges, tight kinks and irregular bending and twisting. The first step towards *de novo* design of a large variety of NTF2-like proteins is devising principles for designing curved beta sheets. In this work, we demonstrate we can generate a number of different curved sheets in the context of NTF2-like proteins. We then use these principles, along with additional information from native NTF2-like proteins, to create a generative algorithm that can widely sample structural diversity while still producing physically realistic models. We show this algorithm can produce a large variety of NTF2-like proteins, and through cycles of large-scale design and validation, we increase the diversity and success rate of its output. As a proof of principle, we design a binder of the mycotoxin aflatoxin B1, which could serve as starting material for devising aflatoxin-chelating or degrading materials.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge my undergraduate advisors, Cristian Solari and Visitación Conforti, for giving me the chance to work with them very early during my undergraduate studies. They provided critical support later on for my experiences at Janelia Farm, where I worked with Loren Looger and Jonathan Marvin on protein design for the first time. I would like to acknowledge Jonny and Loren for the amazing opportunity they gave me to work with them in exciting research, in a stimulating and friendly environment. This experience opened many doors for me, and solidified my passion for protein design. For this, I am ever grateful.

I would have not had the ability and confidence to pursue this thesis work without the solid foundations I learned at Universidad de Buenos Aires. I would like to acknowledge my professors at UBA, in particular, Julio Caramelo, Alejandro Nadra, Diego Ferreiro, Susana Silverstein, Fabiana Lo Nostro, Alberto Kornblihtt and Osvaldo Uchitel.

I would like to acknowledge the support from David Baker and everyone at the Baker lab, who provided invaluable technical help and expert advice. In particular, I would like to thank Enrique Marcos, with whom I worked closely and learned a lot from.

I would like to acknowledge the community from the Biological Physics Structure and Design (BPSD) PhD program, who provided support through hard patches and milestone celebrations. In particular, I would like to thank Erin Kirchner, BPSD program coordinator, who was always ready to help, lend an ear, or celebrate. I would like to thank BPSD alumnus and friend Ian Haydon for invaluable advice on scientific communication. I would like to acknowledge the support from the members of my cohort, Zibo Chen, Kiri Choi, Una Nattermann and Hannah Baughman, for their close friendship, which helped me navigate the difficult first year away from my family. In particular, I would like to acknowledge the unwavering support of Hannah Baughman, who has become more than a friend to me, and whom without my life would not be as happy as it is.

I would like to acknowledge the support from my Seattle family, Raymond and Jannie Lee, who were at my side ever since I arrived in Seattle.

I would like to acknowledge the emotional support from my Argentine friends and family. In particular, I would like to thank my grandmother, Lucy, and her sister Nilda, for their love, patience and help after school, when I was an undisciplined middle-school student.

Finally, I would like to thank my parents, Patricia Diaz and Daniel Basanta, for being at my side and helping me become who I am.

DEDICATION

I would like to dedicate this thesis to my professors from the Facultad de Ciencias Exactas y Naturales at Universidad de Buenos Aires, who teach and work with passion despite the difficulties of doing research in Argentina. The quality and dedication to your work are an inspiration to me.

CHAPTER 1. INTRODUCTION

1.1 Current state of functional protein design

Most cases to functional protein design are based on redesigning native proteins, which provide the structural scaffold on which to construct a binding site (1, 2). Depending largely on native proteins limits the set of available scaffolds to those for which structures have been solved, can accommodate a binding site, and tolerate mutations. The protocols used in these cases generally start by searching for placements of the small molecule or transition state of interest, along with key interacting side chains, in a scaffold pocket. The adjacent positions are then mutated to improve the interaction energy (or an equivalent score) between ligand and protein (1). This redesign process is done by modeling the protein backbone as a rigid or quasi-rigid structure. This limit on backbone movement is necessary for maintaining the native contacts outside of the binding site, limits or directly precludes optimization ligand/substrate-protein interactions. Furthermore, studies where optimization of binding affinity or enzyme activity are done by directed evolution, show that mutations in the second interaction shell and/or backbone rearrangements are a necessary step to attain native-like affinities or activities (2–4). Taken together, these observations suggest two main ways forward to computational design of proteins with native-like activities: A) Generate *de novo* scaffolds, aimed to expand the native repertoire and supplement backbone flexibility during design. B) Address the inconsistencies observed when only the residues adjacent to the ligand are redesigned in native scaffolds. A proof of principle of this approach indicates its feasibility, and suggests that with a larger diversity of *de novo* proteins the number of attainable functions would be much larger (5).

The grounds for generation of *de novo* scaffolds have been established in recent literature, in most cases using a set of rules and algorithms that guide the construction of protein models composed of canonical secondary-structure elements (6–11). Forward-looking statements about functionalization are a common denominator in this body of work, but for most of it, a clear strategy for function design is lacking. With few notable exceptions (5), the main obstacle in the way of functionalization is the limited or non-existent availability of structural features able to accommodate an active site. Most cases are small, globular scaffolds that only have very small concave depressions or, in other cases, bigger scaffolds where the active site residues would have to be located in loops, which emulates nature (12), but

presents a greater challenge for accurate modeling (13). Taking these observations into account, a next generation of *de novo* scaffolds, conceived for functionalization, should meet the following criteria: A) Have a pocket or pockets, where the binding site can be placed. B) Loops should not be part, or be a minority, of the secondary-structure elements that form these pockets. C) The size and shape of the pockets should be controllable and highly variable.

1.2 The Nuclear Transport Factor II as a model for functional protein de novo design

Defining a protein fold as the arrangement of its secondary structure elements (SSEs) relative to each other in space, a possible strategy to meet the above-mentioned criteria is using a fold observed in nature. This target fold would serve as a basic frame from which to generate the desired diversity, while making the scaffold design process tractable. For this work, we propose the Nuclear Transport Factor 2 (NTF2) fold as such starting point. The dominant motif of this fold is a central six-stranded beta-sheet with a high degree of curvature. This curved sheet is complemented with three helices that close over it, forming a cone that encloses a pocket, surrounded, in most cases, exclusively by SSEs different from loops (14). Furthermore, this protein fold has undergone substantial sequence and functional diversification during evolution, suggesting an equally broad potential for *de novo* proteins of the same family (15–19). The wide array of shapes and sizes observed in native NTF2 structures is also encouraging for the aims of this project, suggesting that even more variants than those found in nature can be created.

1.3 A generative algorithm for proteins of the NTF2-like superfamily

The main challenge for sampling and expanding the native NTF2 structural landscape is doing so in a systematic and tractable way. That is, devising algorithms that generate physically realistic models, but are not constrained to sampling only local variations of known structures. Suggestions for what structural features are determinant of the NTF2 fold can be taken from the observation of native structures. With

this information, exploring the NTF2 structural space in a sensible way should be feasible. We propose that this exploration is done through a generative algorithm; a set of instructions that takes a limited set of parameters and consistently generates protein models with an NTF2 fold.

One of the cornerstones of *de novo* protein design is the selection of sequences that favor the native state while disfavoring alternative states. This is achieved by selection of sequences with a strong local bias towards the designed backbone conformation (10). Incorporation of binding site design in this framework would impose constraints to the available sequence space, especially for positions directly involved in ligand interaction. Because the side chains in these positions must interact with the adjacent residues, the constraints would also affect these, and by the propagation of this effect, the whole protein. This view is in agreement with the challenges explained above, and highlights the importance of adopting design strategies that take into account the binding site structure during sequence optimization. The strategy I propose here is concurrent optimization of the binding site and overall protein structure, in combination with scaffold *de novo* design. This approach is fundamentally different from those used so far, where a binding site is designed in a preexisting scaffold and only the local interactions are optimized. The main challenge for the implementation of this strategy is the reduction of the available sequence space by the constraints imposed by the binding site, but current sequence space search methods are suitable for this problem if supplemented with backbone diversity, as proposed here.

The validation of the proposed design strategy requires a ligand test case, ideally, one that avoids adding complexity of its own. A particular kind of ligands has been identified as low-complexity: small molecules with few rotatable bonds, a high proportion of hydrophobic surface and some, but not numerous, hydrogen-bond acceptors and/or donors (20).

Taken together, the challenges presented above, and the strategies proposed to overcome them, delineate the aims and scope of this work:

1) Identify structural and sequence determinants of the NTF2 fold.

2) Using the information gathered in aim 1, devise algorithms that generate realistic computational models of proteins with an NTF2-like fold.

3) Design a small-molecule-binding protein as a poof of concept by integrating the design of a binding site in the algorithms developed in aim 2.

1.4 REFERENCES

1. Morin, A., Meiler, J., and Mizoue, L. S. (2011) Computational design of protein-ligand interfaces: potential in therapeutic development. *Trends Biotechnol.* **29**, 159–66
2. Khare, S. D., and Fleishman, S. J. (2013) Emerging themes in the computational design of novel enzymes and protein-protein interfaces. *FEBS Lett.* **587**, 1147–1154
3. Preiswerk, N., Beck, T., Schulz, J. D., Milovník, P., Mayer, C., Siegel, J. B., Baker, D., and Hilvert, D. (2014) Impact of scaffold rigidity on the design and evolution of an artificial Diels-Alderase. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8013–8
4. Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L., and Baker, D. (2013) Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*. **501**, 212–6
5. Dou, J., Vorobieva, A. A., Sheffler, W., Doyle, L. A., Park, H., Bick, M. J., Mao, B., Foight, G. W., Lee, M. Y., Gagnon, L. A., Carter, L., Sankaran, B., Ovchinnikov, S., Marcos, E., Huang, P.-S., Vaughan, J. C., Stoddard, B. L., and Baker, D. (2018) De novo design of a fluorescence-activating β-barrel. *Nature*. **561**, 485–491
6. Huang, P.-S., Feldmeier, K., Parmeggiani, F., Fernandez Velasco, D. A., Höcker, B., and Baker, D. (2016) De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34
7. Lin, Y., Koga, N., Tatsumi-koga, R., Liu, G., Clouser, A. F., and Montelione, G. T. (2015) Control over overall shape and size in de novo designed proteins. 10.1073/pnas.1509508112
8. Park, K., Shen, B. W., Parmeggiani, F., Huang, P.-S., Stoddard, B. L., and Baker, D. (2015) Control of repeat-protein curvature by computational protein design. *Nat. Struct. Mol. Biol.* **22**, 167–74
9. Voet, A. R. D., Noguchi, H., Addy, C., Simoncini, D., Terada, D., Unzai, S., Park, S.-Y., Zhang, K. Y. J., and Tame, J. R. H. (2014) Computational design of a self-assembling symmetrical β-propeller protein. *Proc. Natl. Acad. Sci.* **111**, 15102–15107
10. Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., and Baker, D. (2012) Principles for designing ideal protein structures. *Nature*. **491**, 222–7
11. MacDonald, J. T., Maksimiak, K., Sadowski, M. I., and Taylor, W. R. (2010) De novo backbone scaffolds for protein design. *Proteins Struct. Funct. Bioinforma.* **78**, 1311–1325
12. Bartlett, G. J., Porter, C. T., Borkakoti, N., and Thornton, J. M. (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**, 105–21
13. Hu, X., Wang, H., Ke, H., and Kuhlman, B. (2007) High-resolution design of a protein loop. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 17668–73
14. Eberhardt, R. Y., Chang, Y., Bateman, a, Murzin, a G., Axelrod, H. L., Hwang, W. C., and Aravind, L. (2013) Filling out the structural map of the NTF2-like superfamily. *BMC Bioinformatics*. **14**, 327
15. Li, K., Zhao, K., Ossareh-Nazari, B., Da, G., Dargemont, C., and Marmorstein, R. (2005) Structural basis for interaction between the Ubp3 deubiquitinating enzyme and its Bre5 cofactor. *J. Biol. Chem.* **280**, 29176–29185
16. Kerfeld, C. A. (2004) Structure and function of the water-soluble carotenoid-binding proteins of cyanobacteria. *Photosynth. Res.* **81**, 215–225
17. Sultana, A., Kallio, P., Jansson, A., Wang, J.-S., Niemi, J., Mäntsälä, P., and Schneider, G. (2004) Structure of the polyketide cyclase SnoaL reveals a novel mechanism for enzymatic aldol condensation. *EMBO J.* **23**, 1911–1921
18. Bullock, T. L., Clarkson, D. W., Kent, H. M., and Stewart, M. (1996) The 1.6 Å Resolution Crystal Structure of Nuclear Transport Factor 2 (NTF2). *J. Mol. Biol.* **260**, 422–431
19. Nagata, Y., Mori, K., Takagi, M., Murzin, a G., and Damborský, J. (2001) Identification of protein fold and catalytic residues of gamma-hexachlorocyclohexane dehydrochlorinase LinA. *Proteins*. **45**, 471–477

20. Allison, B., Combs, S., DeLuca, S., Lemmon, G., Mizoue, L., and Meiler, J. (2014) Computational design of protein-small molecule interfaces. *J. Struct. Biol.* **185**, 193–202

CHAPTER 2. STRUCTURAL ANALYSIS OF THE NTF2-LIKE PROTEIN SUPERFAMILY

2.1 Abstract

Named after the rat protein Nuclear Transport Factor II (NTF2), the NTF2 fold has been observed in a myriad of proteins in all domains of life, performing diverse functions. These small (~120 amino-acids) globular proteins and domains are shaped like a cone, and contain a pocket where substrates and ligands are bound. The idea of using large sets of NTF2-like proteins from which to select scaffolds for novel function design emerged from the successful redesign of these proteins for small molecule binding. Large and diverse sets of NTF2-like domains could be realized using *de novo* design to systematically sample structural diversity. In order to do this, a general understanding of the NTF2-like fold is paramount. Here we analyze the structures of native NTF2-like proteins in search of the determinants of this fold, the features necessary for folding and how backbone structure gives rise to pocket structural diversity. We identify patterns that repeat in most native NTF2-like proteins, suggesting their importance for folding, as well as backbone patterns in the areas close to the pocket that provide hints for how structural diversity could be systematically sampled.

2.2 Introduction

The NTF2-like fold was first observed in the scytalone dehydrogenase of *Magnaporthe grisea* (1). This enzyme, part of the melanin synthesis pathway, is a homo-trimer whose monomer is a cone-shaped protein with a large pocket that opens at the base (Figure 2.1 A). The cone is mostly formed by a large six-stranded, curved sheet, and finished by two helices. The longer four strands that form the sheet (S3-6, figure 2.1 B) are contiguous in sequence and are paired in an antiparallel way. The two additional strands (S1 and 2, figure 2.1 B) form a short antiparallel hairpin with strand 1 attached to the rest of the sheet by a parallel pairing to strand 6. The sides of the cone are closed by the two N-terminal helices (H1 and 2, figure 2.1 B), which pack against each other in an antiparallel way, and against the large sheet perpendicularly to the direction of the strands (Figure 2.1 A). Proteins from the NTF2-like superfamily share this basic architecture, with deviations in some particular cases.

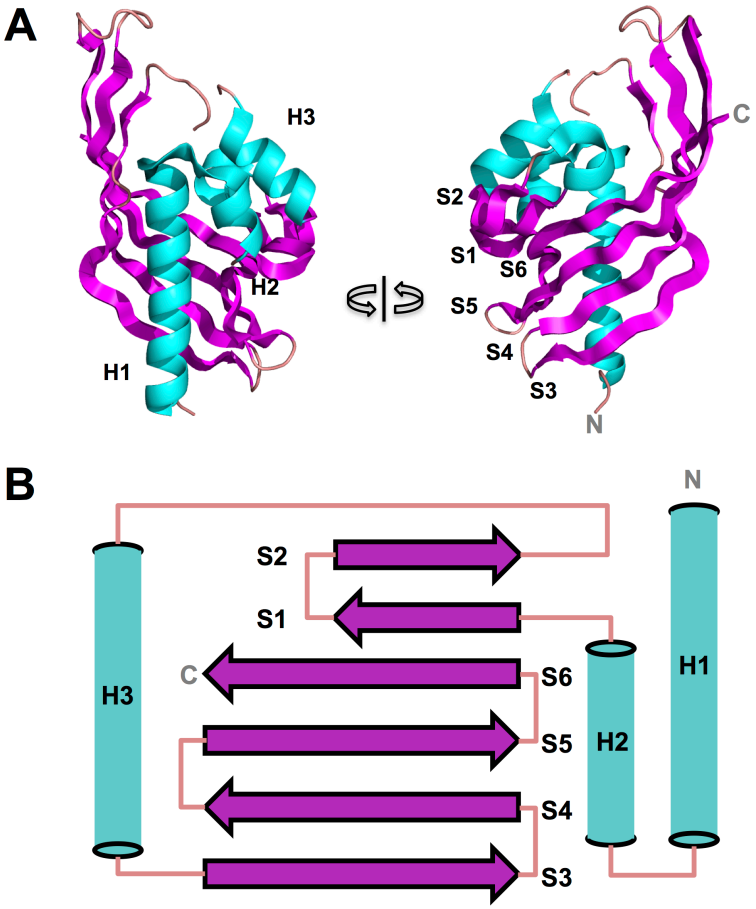


Figure 2.1: **NTF2 basic structural elements. A**. Cartoon representation of the scytalone dehydrogenase (PDB 1IDP) monomer in two perspectives, showing the 3D arrangement of secondary structure elements: Helices colored in cyan, strands in magenta, loops in tan. **B**. Secondary structure scheme of scytalone dehydrogenase showing connectivity and strand pairing.

In order to carry out a detailed structural analysis of NTF2-like domains it is worth using pre-established protein structure classification tools, such as the CATH Protein Structure Classification Database (2), the Structural Classification Of Proteins – extended (SCOPe) (3), or Pfam (4). All three databases have taxa grouping NTF2-like proteins or domains: The NTF2-like superfamily (d.17.4) in SCOPe, the homologous superfamily (3.10.450.50) in CATH, and the NTF2 clan (CL0051) in Pfam. Interestingly, the number of Pfam sequence entries for NTF2-like proteins is in the order of the 76000, almost an order of magnitude less than the number of entries for the most widespread enzyme structural family, the Triose Phosphate Isomerase clan (PFam clan CL0036). SCOPe is readily set up for obtaining

low-homology structure subsets, we therefore used its NTF2 classification (domain superfamily d.17.4) for the basis of our analysis.

NTF2-like proteins have a wide range of functions and have been observed in all life kingdoms, as seen in the CATH and Pfam NTF2 entries. Those with known function can be characterized in two groups: enzymatic and non-enzymatic (2, 5), with most representatives with known function belonging the first group. Non-enzymatic NTF2-like proteins have functions such as small-molecule binding (6), or form part of large structural complexes involved in vital cell functions (5). The majority enzymatically active and small-molecule binding NTF2-like domains for which structures are known have their functional site in the pocket formed by the curved sheet, highlighting the importance of the connection between the structural plasticity of the pocket and the functional variability of these domains. Understanding this connection, the main subject of this chapter, is paramount for the generation of useful *de novo* NTF2-like proteins.

A common feature of pockets in NTF2-like domains is that they are lined by both hydrophobic and hydrophilic interactions, with the latter including hydrogen bonds and charge-charge interactions with both protein side chains and ordered water molecules; examples of this can be found in the pockets of *Magnaporthe grisea* scytalone dehydrogenase (1), rat Nuclear Transport Factor II (7) and *Mycobacterium tuberculosis* epoxide hydrolase (8).

Most NTF2-like domains form some type of homo-oligomer, with few monomeric exceptions, dimers being the most common. The homo-oligomeric interface is often located on a flat part of the convex face of the large sheet, forming an elongated patch that runs perpendicular to the strand direction; examples of this can be seen in the structures of *Rhodococcus erythropolis* limonene-1,2-epoxide hydrolase (PDB 1NU3) (9), *Pseudomonas putida* ketosteroid isomerase (PDB 1E3V) (10) and the rat nuclear transport factor 2 (PDB 1OUN). The same interface configuration can be seen in homotrimers, such as the *Magnaporthe grisea* scytalone dehydrogenase (1), or the beta subunit of the benzoate 1,2-dioxygenase from *Burkholderia mallei* (PDB 3E99). Other, less common, interface conformations involve contacts between sheets and the N-terminal helix, or between elements adjacent the pocket opening; examples of this can be seen in the structures of putative enzymes deposited in the PDB with the IDs 2CHC and

3B8L. The frequency with which NTF2-like domains are part of homo-oligomeres and how dominant a particular type of interface is, may be indicative of a strong evolutionary advantage to this configuration.

We begin the following result section by analyzing protein structures from the NTF2-like superfamily and devising a system that provides a framework to understand its diversity. Then, using this framework, we identify constant structure and sequence features that can be used for creating a diverse set of *de novo* NTF2 structures. Finally, I map the relationship between both variable and constant features to pocket shape and size in order to understand how pocket diversity arises from secondary structure arrangement.

2.3 RESULTS

2.3.1 A systematic framework for analyzing NTF2-like protein structures

NTF2-like domains have a characteristic secondary structure arrangement that can be described as a series of interconnected elements, not in sequence space, but in local three-dimensional space. The variation in NTF2-like domains can then be analyzed using this system as a reference. In order to establish this systematic framework, we analyzed a non-redundant set of 92 NTF2-like domains from the SCOPe database (See section 2.5.1). In order to come up with a series of structural elements that are present on all NTF2-like domains, we compared all 92 structures to the SCOPe entry for the *Magnaporthe grisea* scytalone dehydrogenase (d1idpa_), using TMalign (11).

The non-redundant NTF2 domains contain different numbers of secondary structural elements, owning to extensions of either termini, insertions, or strand or helix breaks. To establish a general frame for analysis, it is worth coming up with a naming scheme that fits an average secondary structure composition. The alignment of domains shows that the scheme presented in figure 2.1 for *Magnaporthe grisea* scytalone dehydrogenase is an appropriate generalization for the NTF2-like domains. The elements aligned to the four C-terminal strands tend to contain breaks and non-ideal beta strand configurations in similar places that do not result in deviations from the average strand path, it is therefore reasonable to approximate these broken strands to a single one of those contained in the general

scheme. The same is true for the three helices and the short hairpin paired to strand 6. With a general scheme to refer to common elements in NTF2 domains, we can move on to analyzing the variations length and three-dimensional arrangement.

The large, curved, sheet formed by the four C-terminal strands is the most prominent structural element of the NTF2-like superfamily (Figure 2.2 A and B), and it forms most of the surface lining the pockets. This sheet can be divided in three sides or faces: a central flat base, delimited by strand breaks or other non-ideal strand configurations (Figure 2.2 C), and two faces, or arms, one on each side of the base, that twist in opposite directions (Figure 2.2 E). Since one of the arms is less structurally diverse and shorter on average than the other we named the arms “short” (Strands 5 and 6) and “long” (Strands 3, 4 and 5) to distinguish them.

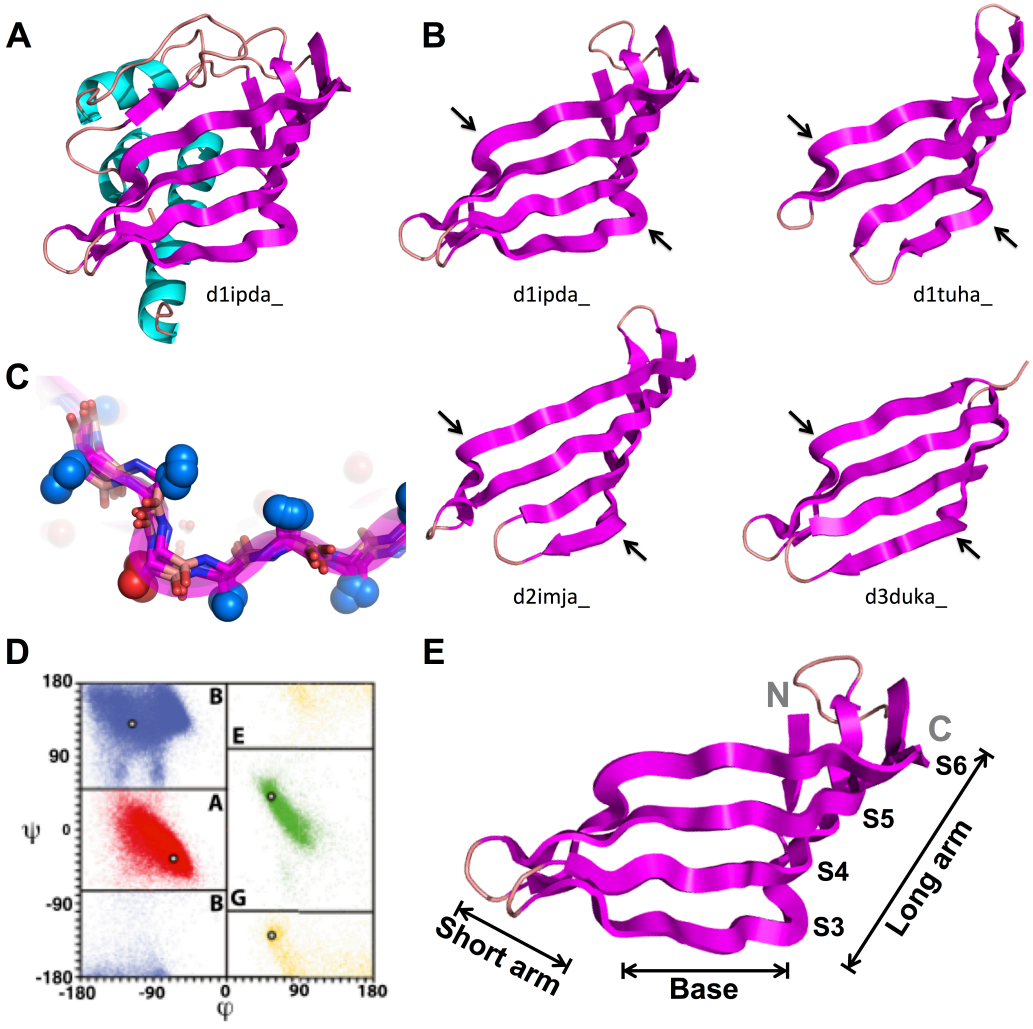


Figure 2.2: **Structural features of NTF2-like domain sheets. A**. Cartoon representation of the SCOPe entry d1idp1_ (*Magnaporthe grisea* scytalone dehydrogenase) showing the 3D arrangement of secondary structure elements: Helices colored in cyan, strands in magenta, loops in tan. **B**. Sheets from diverse SCOPe NTF2-like superfamily entries, with arrows pointing at beta bulges on strands 3 and 6. **C.** Close-up to strand 3 bulges on four aligned SCOPe entries for four NTF2-like proteins (d1gy6a_, d1of5b_, d1q40a_ and d1zo2a1) using TMAlign (11). **D.** Ramachandran plot binned and colored by general torsion populations, with A for "alpha" (helical conformations), "B" for beta (beta-sheet conformations), "E" for extended and "G" for glycine, called ABEGO bins as a group. **E.** d1idp1_ sheet with main sheet subdivisions and strand numbers labeled. Termini are labeled with bold gray "N" and "C".

The most common non-ideal beta structure delimiting the three faces of the large sheet is the beta bulge (Figure 2.2 C), which consists of a one-residue insertion with alpha-helical torsion resulting in a bulging of the strand (Figure 2.2 D) (12). Beta bulges and short loop regions mark an abrupt change in the direction of the sheet where the arms originate (Figure 2.2 E). These irregularities are specifically located in strands 3 and 6, with strands 4 and 5 having a more constant bend and twist. The relative spacing between irregularities in strand 3 and 6 dictates the width of the flat base, as exemplified by the different sheets in figure 2.2 B.

In all NTF2-like proteins, the base and long arm form a large portion of the inner surface of the pocket, thus, their relative orientation dictates pocket shape and size. In particular, the long arm forms a portion of the pocket entrance and can have an additional bulge close to the N-terminus of strand 3 that influence its curvature. The long arm is formed by 4 to 8 rows of paired residues in strands 3, 4 and 5, while the short arm is formed by either one of two rows of strands 5 and 6. Differently from the long arm, the short arm does not form part of the pocket and, given its smaller range of lengths, has less structural variation. The short arm interacts mainly with the two N-terminal helices.

The two N-terminal helices constitute another key structural element present in all NTF2-like domains. These two helices, H1 and H2, pack against the sheet and H3, forming the core of the domain and closing the cone (Figure 2.1 A). Helix 1 and 2 stack in an antiparallel fashion against each other, connected by a structurally conserved two-residue loop that forms the 360º turn connection (Figure 2.3 A and B). Helix 1 is on average longer than helix 2, and follows the direction of the long arm, packing against S3 near the helix C-term (Figure 2.3 C). Helix 2 is almost exclusively 7 residues long, except in few cases, when it is extended, or part of it unravels (Figure 2.3 D). The C-terminus of helix 1 and the C-terminus of helix 2 form part of the internal surface of the pocket.

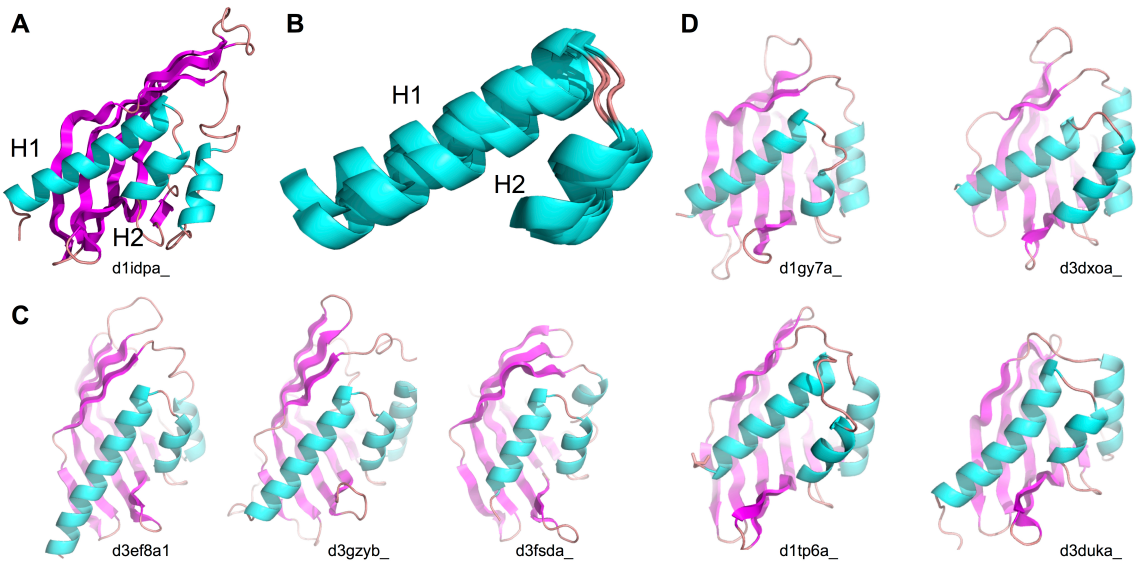


Figure 2.3: **Structural features of NTF2-like domain N-terminal helices A**. Helices 1 and 2 in labeled on the *Margnaporthe grisea* scytalone dehydrogenase strcutre. **B.** Superimposition of four H1-H2 pairs of different NTF2 SCOPe entries (d3gzyb_, d3owsa_, d3stda_, d4cdla_). **C.** Three different NTF2 SCOPe entries with different H1 lengths. **D.** Four less common configurations H1-H2 pairs, with different loop configurations and H2 lengths.

Helix 1 and helix 2 are followed in sequence by a short hairpin that pairs parallel to strand 6, which in turn is followed by a sharp turn and helix 3, connecting to strand 3 (Figure 2.1 B and 2.4 A-F). This hairpin and helix 3 form the majority of the pocket entrance, or mouth (Figure 2.4 A). Both elements are highly variable (Figure 2.4), with a minority of structures included in the NTF2-like SCOPe superfamily completely lacking the hairpin (Figure 2.4 F). Since the hairpin strand pairings limit the variability of the placement of the first residues of helix 3, its minimum length and loop connection to strand 3 are mainly dictated by the conformation of the main sheet. The complete replacement of the hairpin by irregular loops (Figure 2.4 B, C, D and E), long H3-S3 connections (Figure 2.4 D) and extensions of the S4-S5 loop (Figure 2.4 E) combine to produce highly diverse entrances to the pockets.

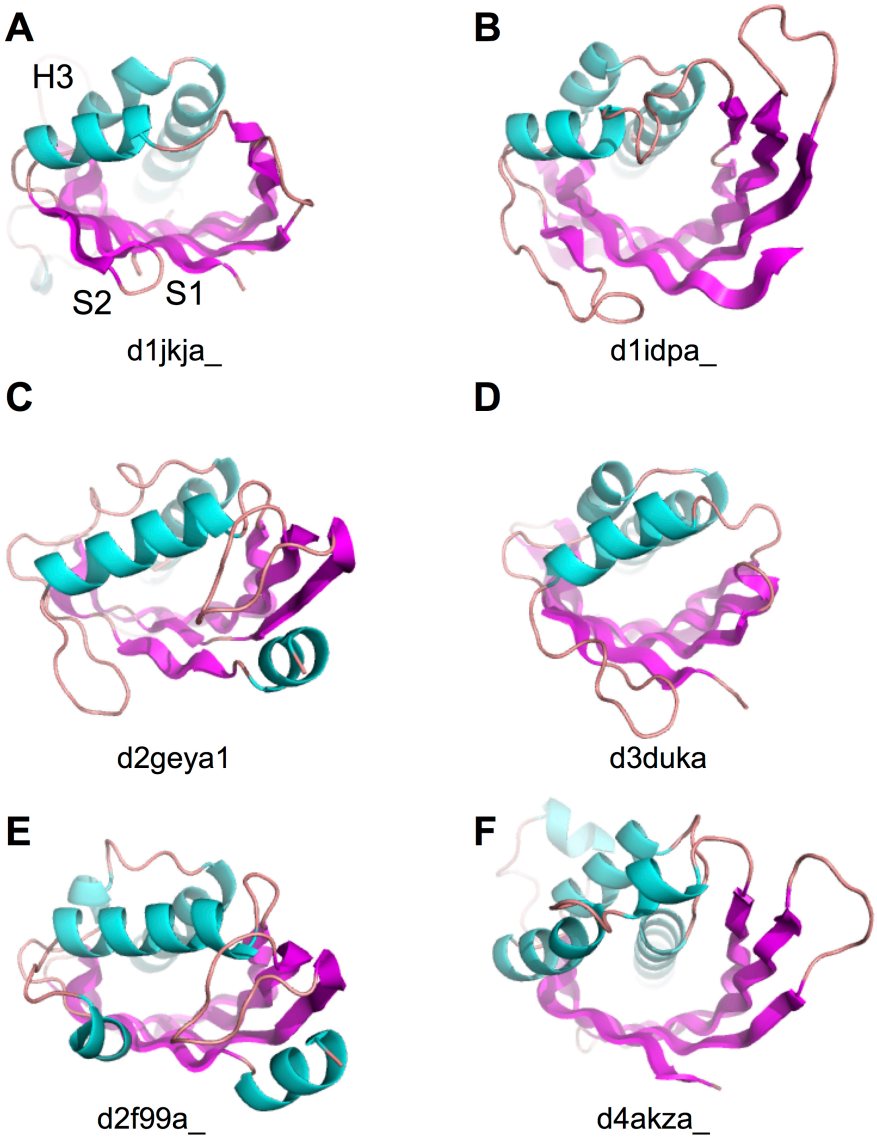


Figure 2.4: **Structural features of NTF2-like domain frontal elements A**. SCOPe entry d1jkja_, with H3 and the frontal hairpin strands labeled. This entry presents the canonical hairpin and helix 3 structures. **B**. An NTF2-like domain showing a long irregular connection between H3 and S3, as well as irregular loops that take the place of the frontal hairpin. **C**. Structure showing a canonical H3, and a short irregular loop that takes the place of the frontal hairpin. **D**. Structure showing an elongated loop in place of the frontal hairpin. The H3-S3 connection is a long structured loop that covers the distance between the C-terminal end of H3 and the N-terminus of S3 in the very short long arm of the main sheet. **E**. A structure showing yet another configuration of the irregular loop elements that frequently takes the place of the frontal hairpin. **F**. Example of a structure missing the frontal hairpin, along with other deviations from the canonical structure presented in Figure 1.1 B.

As exemplified by structures in Figure 2.4 B, C and E, NTF2-like domains can present extensions at the C-terminus that interact with the long arm. These elements are often helical, but can also be

extensions of S6 (Figure 2.4 B), or combinations of these and loops. These C-terminal elements, combined with extensions of the frontal hairpin and some specific configurations of H3, contribute to the shape and size of the pocket, and can sometimes occlude its opening, in which case it can be found in the space left between the H1-H2 connection and H3 (Figure 2.5).

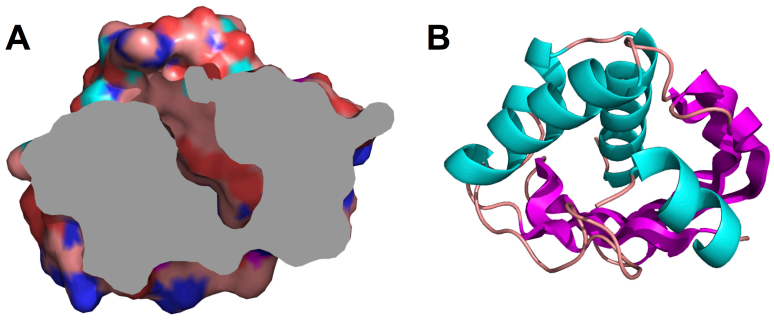


Figure 2.5: **Pocket opening features in NTF2-like domains A**. Solvent surface rendering of SCOPe entry d3g8za_, showing the pocket and its opening in the space between the H1-H2 connection and H3. **B**. Cartoon representation of the same structure, showing an extended frontal loop and a C-terminal helix that occlude the space where the pocket opening would normally be.

2.3.2 Sequence and structure motifs in NTF2-like proteins

As previously reported (13), conserved sequence-structure motifs contain information that can be used for *de novo* design, specially for connections between secondary structure elements, where the sequence can favor certain conformations. In this section we analyze conserved sequence-structure motifs in the NTF2-like superfamily to obtain information that can be used to build them from scratch.

In the previous section we noted that the structure of the N-terminal helices and their connection are highly conserved, here we evaluate if there is sequence similarity that underlies this structural similarity. From the non-redundant NTF2 domain set, we manually extracted 62 structures where the N-terminal helices are not kinked or unraveled, the connection between them is two amino-acids long, and its Ramachandran bin sequence is GB (See figures 2.2 D and 2.3 B). We produced a multiple sequence alignment using PROMAL3D (14), and used it to generate a sequence logo of the primary sequence neighborhood of the H1-H2 connection (See Materials and Methods 2.5.2). The loop between H1 and H2 has a conserved motif, with glycine and aspartate enriched at the first and second positions respectively

(Figure 2.6 A and B), in agreement with similar connections in a general set of protein structures (Figure 2.6 C).

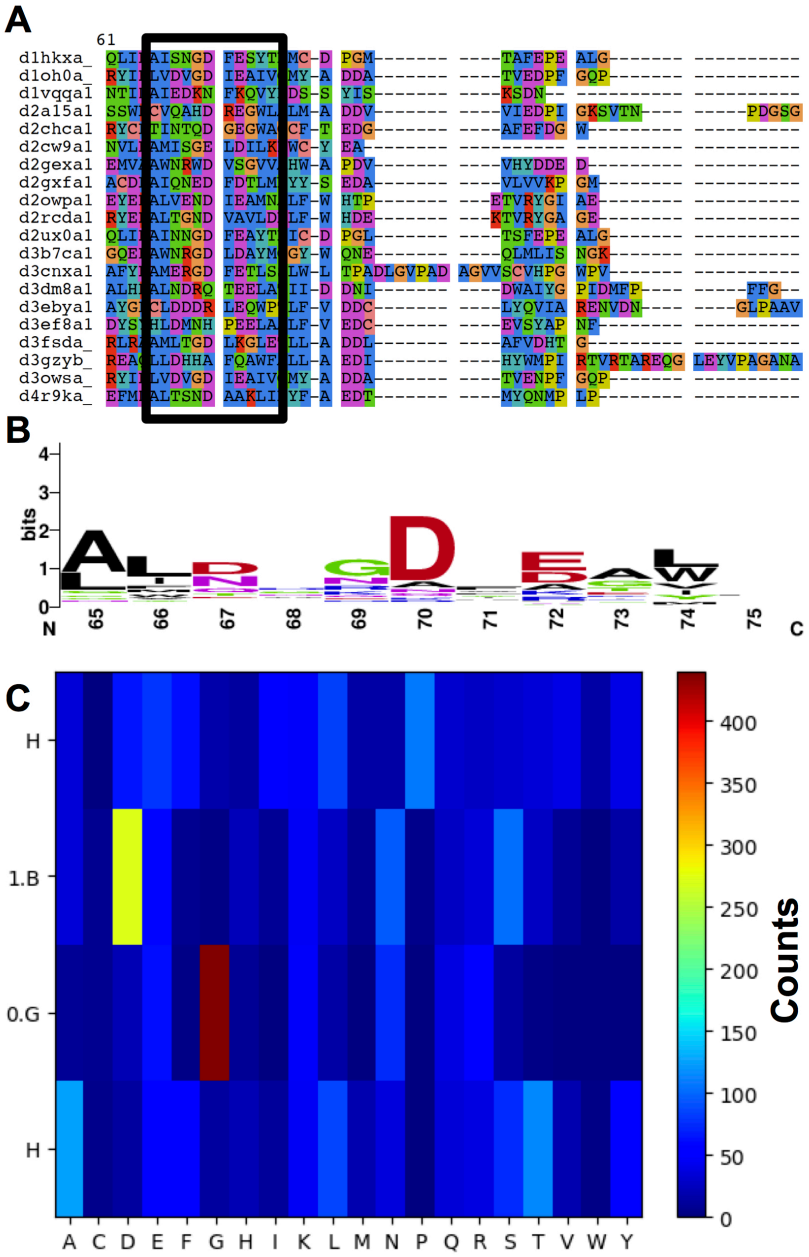


Figure **2.6**: **Conserved motifs in the H1-H2 connection A**. A random subset of 20 sequences from the MSA generated by PROMAL3D, with a framed 10-residue window around the sequences of the H1-H2 connection. **B**. Sequence logo of the 10-residue window around the H1-H2 connection (MSA positions 69 and 70). **C**. Heat map showing counts per amino-acid identity in a general set of PDB structures for 2-residue connections (and flanking helical residues) between helices, with torsional bins G-B.

Another conserved connection is the one between H2 and S1, a 5-residue loop with very conserved torsions (BBAAB Ramachandran bins – see figure 2.2 D) and hydrogen bond pattern (Figure 2.7 A and B). Furthermore, this motif interacts closely with the N-terminus of H3 and its connection with S2, forming a non-local helix cap (Figure 2.7 A and B). The S1-H3 connection also has conserved torsions in the E bin. A third of the structures in the non-redundant set show this highly conserved motif with little or no deviation, while other structures with small deviations have slightly different torsion bins (GBAAB instead of BBAAB), or do not interact with the N-terminus of H3.

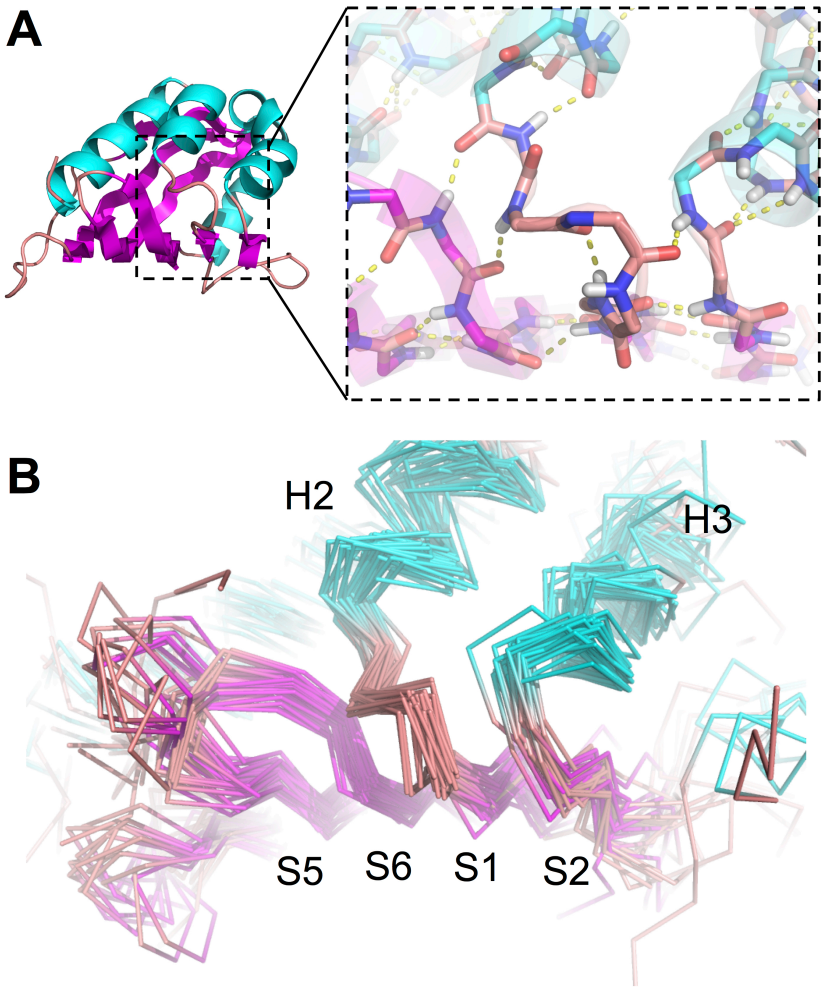


Figure **2.7**: **Conserved motifs in the H2 S1 connection A**. NTF2 domain d1z1sa1 (left) with a close-up to the connection between H2 and S1. Hydrogen bonds are presented as yellow dashed lines and backbone atoms as sticks. **B**. Ensemble of 33 NTF2 domain structures where the H2-S1 connection and its interaction with the N-terminus of H3 are highly conserved.

A multiple sequence alignment of the domains displaying this structural motif shows an underlying sequence motif: the third and fourth positions are enriched in negatively charged amino-acids, flanked by hydrophobic ones (Figure 2.8 A and B). The S1-H3 connection, intimately in contact with the 5-residue loop motif, also shows a conserved sequence motif: The one-residue connection is almost exclusively glycine, followed by a positive amino acid, then by a negative one (Figure 2.8 C and D).

Comparing these motifs to those observed in similar loops in the PDB (Figure 2.8 E) shows a number of differences: In the H2-S1 connection the only similarity between NTF2s and general PDB structures is the preference for aspartate in the fourth position, with proline in the second and third positions being much less represented in NTF2s than expected. In the third position NTF2s prefer a negatively charged amino-acid to proline. Interestingly, the S2-H3 connection shows a preference for glycine, very much in agreement with the PDB counts, but the positions following the loop show very strong preferences that are not apparent in the PDB data.

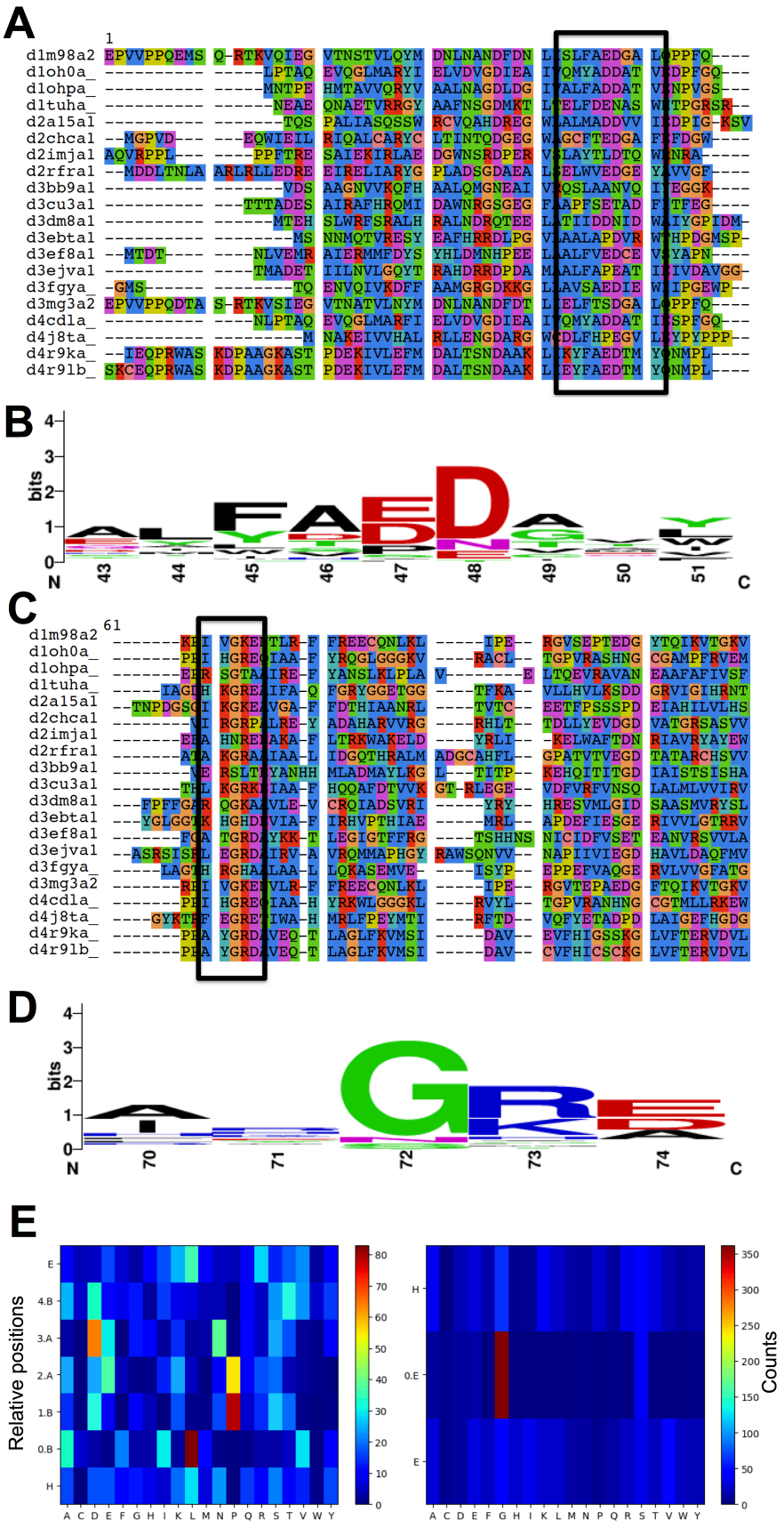

A
B
C
D
E
Relative positions
Counts

Figure 2.8: **Conserved sequence motifs in the H2 and S1 connection A**. Multiple sequence alignment of 33 NTF2 domains with highly conserved H2-S1 connections, with sequence window around connection framed. **B**. Sequence logo of the framed section in panel A. **C**. Same multiple sequence aliment as panel A, but the framed window is the sequence corresponding to the S2-H3 connection. **D**. Sequence logo of the framed section in panel C. **E**. Amino-acid counts from similar loops as the H2-S1 connection (right) and S2-H3 connection (left) in PDB sequences.

2.3.3 Relationship between backbone and pocket structure

Understanding how pocket structural diversity arises from backbone diversity is key for generating useful variability in *de novo* NTF2-like domains. Here we used the pocket detection algorithm CLIPPERS (15) to analyze the pocket structures of a non-redundant set of NTF2-like proteins (See Materials and methods 2.5.3). We then cross pocket structure information with overall structural similarity to extract simple modes in which backbone structure changes affect pocket structure.

In order to understand how backbone structure affects pocket structure, we must remove the effects of different sequences. We achieve this by generating dummy structures where all side-chains have been replaced by valine, (See Materials and methods 2.5.4) and running the pocket analysis algorithms on them. The resulting pocket volumes follow an intuitive trend: larger poly-valine pockets correlate with larger native-sequence pockets, and the size of the native pocket never exceeds that of poly-valine (Figure 2.9).

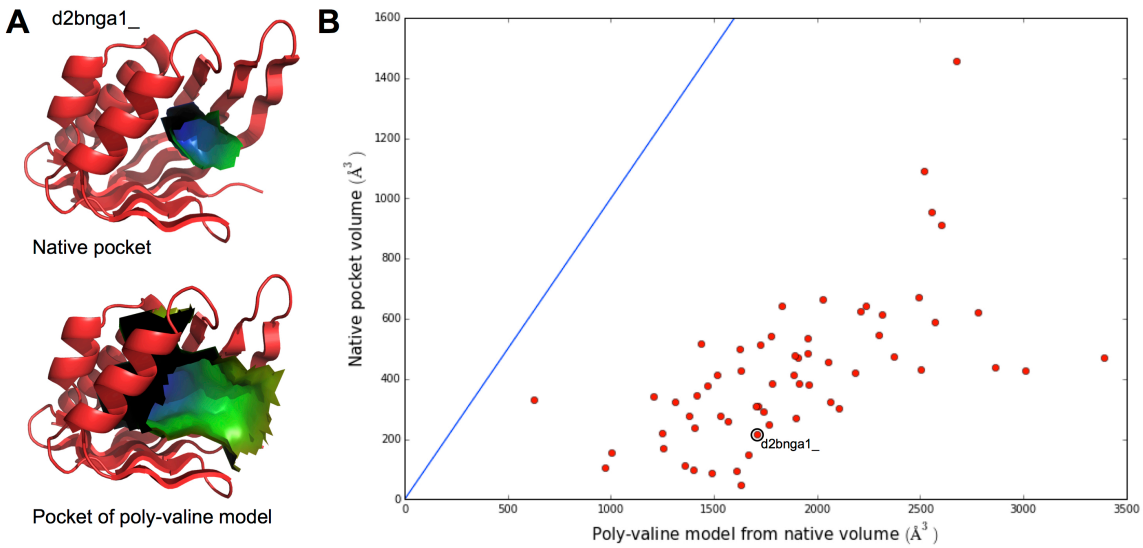


Figure 2.9: **Relationship between backbone and pocket structure A**. Comparison of pockets detected by CLIPPERS in the native structure and the poly-valine model of d2bnga1_. **B**. Scatter plot of native pocket volume versus poly-valine model pocket volume. A solid blue line represents

the diagonal.

Clustering backbones by structural similarity yields 15 clusters, with the two largest ones (13 and 14) spanning the whole range of pocket volumes (Figure 2.10 A) (See Materials and methods 2.5.5). Superimposing structures from each of these clusters provides a visual representation of how specific changes in backbone structure affect pocket structure (Figure 2.10 B). The diversity on both clusters is concentrated near the opening of the pocket, with the less variable part of the structure near the N-terminus of H1 and the main sheet strands it is in contact with. This indicates that backbone structure modulates pocket size mainly by extending or retracting the structural elements directly at the opening: The long arm S4-S5 hairpin, the frontal hairpin, the H3-S3 connection and presence or absence of C-terminal elements. Single pairwise comparisons from each cluster illustrate this more clearly: From cluster 13, the pocket of d3fsda_ is around 500$Å^3$ larger than d3duka_ thanks to an extended long arm, whose sharp curvature near the N-term of S3 makes it wrap around and extending the pocket near the C-terminus of H3 (Figure 2.10 C). Structures d3er7a_ and d1z1sa1, from cluster 14, have a similar difference in pocket volume, but display a mechanism of pocket extension that does not rely on the long arm (of similar length in both of them): d1z1sa1 has an additional 12-residue helix on its C-terminus, that packs against the long arm, and a long loop that extends the frontal hairpin (Figure 2.10 D).

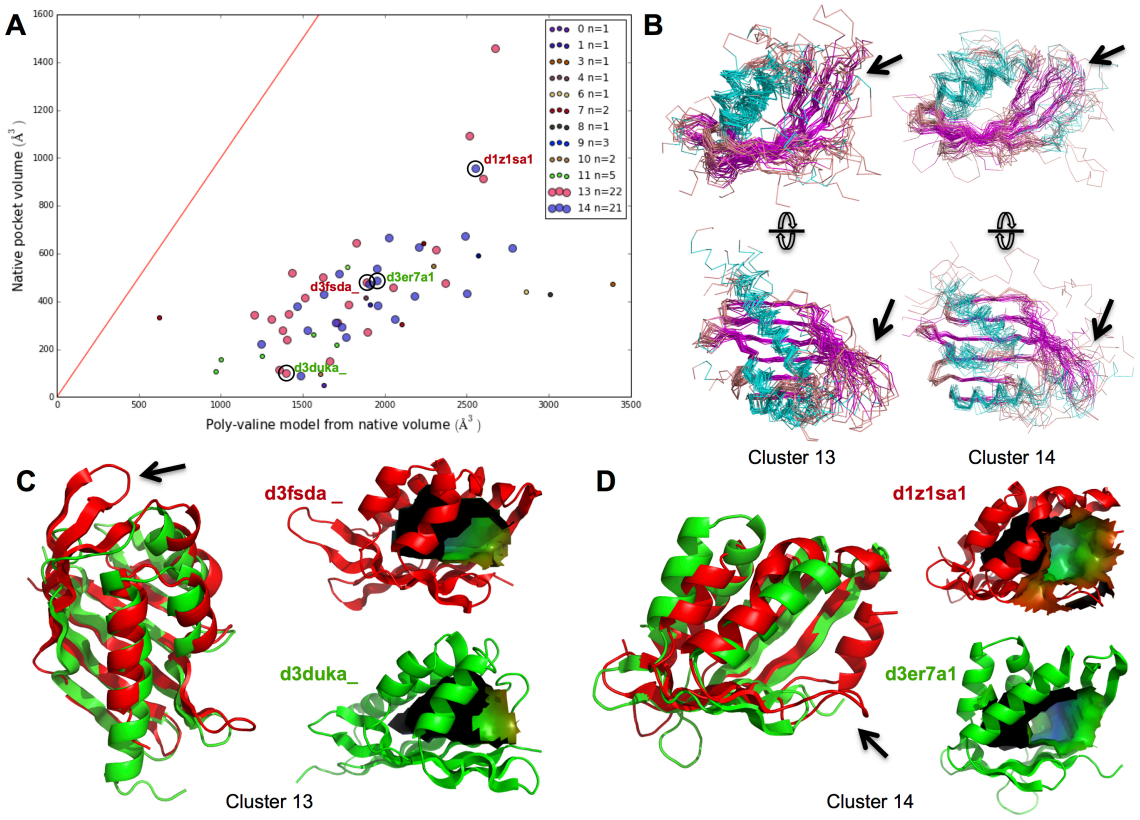


Figure 2.10: **Determinants of pocket volume A**. Pocket volume of native structure vs. volume of poly-valine model pocket, with different colors for each cluster. The largest clusters, 13 and 14, have larger markers. **B**. Superimpositions of all structures of clusters 13 and 14, with arrows pointing at highly variable regions. **C**. Example structures from cluster 13, with an arrow pointing at the long arm. **D**. Example structures from cluster 14, with an arrow pointing at the frontal hairpin and C-terminal helix.

2.4 Discussion

NTF2-like domains show a wide range functions and of active site structures. Here we analyzed their structures using a systematic framework in order to understand how the wide variety of pocket structures is realized in this these small domains. The final goal of this analysis is to enable the generation of a large and diverse library of de novo NTF2-like proteins that can be used for designing new functions.

We identified the main sheet as one of the key elements making up NTF2 domain structures and dictating pocket structure. These long, highly curved sheets present non-canonical beta structures (bulges), or directly short breaks on beta sheets, at points where the sheet takes a sharp turn from the strand direction. From their ubiquity, we propose these deviations from ideality play a key role in allowing

and controlling beta sheet curvature, as previously suggested (16, 17). Furthermore, we propose their placement relative to each other fine-tunes overall sheet structure, and therefore, pocket structure.

In our review of NTF2 domain structure and function we noted their tendency to form homo-oligomers. It is not clear to us if this tendency is a result of the bias of symmetric proteins to crystallize more readily (18), or an evolutionary advantage to homo-oligomers. Evolutionary advantages to homo-oligomers include enhanced stability (or even requirement for stability), or allosterism. It should be noted that the effect of oligomerization could be a composite of enhanced stability and symmetry, as stability and monodispersity have also been linked to ease of crystallization (18). Homo-oligomerization may well be an evolutionary crutch that enables plasticity in active site size and structure. It reasonable to expect that *de novo* design of these small domains will require careful optimization of well packed cores that provide the driving force for folding, specially when part or it is removed to make room for an active site.

We explored the sequence determinants of the few highly conserved structural motifs in NTF2-like domains, and found sequence stretches with high information content. The 360º turn from H1 to H2 is has highly conserved structure and sequence, with the latter agreeing with similar connections in the PDB. The direct encoding of the sequence to the structure is valuable information for designing *de novo* NTF2-like proteins.

The long loop connecting H2 and S1 has conserved sequence and structure. Interestingly, this loop is the hub for a network of main chain and side chain contacts with other non-local structure elements: S6, and the connection between S1 and H3. The main chain of the 5-residue loop has hydrogen bond contacts that cap the S6 beta bulge, with itself, the N-terminus of H3 (this could also be interpreted as a capping interaction) and S2. The sequence information of this loop and the interacting S2-H3 connection provides additional insight: The H2-S1 connection sequence shows enrichment in negative amino-acids that is nor observed in similar connections in the PDB, and the S2-H3 connection shows a similar trend, but positively charged amino acids in the N-terminus of H3. Positive amino-acid enrichment is not expected in proximity of positive charge density such as a helical N-terminus. Visual inspection of structures displaying both sequence motifs provides the final piece of the puzzle: The H2-S1 loop negative amino-acids swing towards the H3 N-terminus, away from the H2 C-terminus and S6 beta-bulge carbonyls, and the positive amino-acid that is the N-terminus of H3 reaches straight out and wedges

between the H2 C-terminus and a main chain carbonyl of the H2-S1 loop. The overall result is a charge helix cap swap that stabilizes the configuration of all elements involved. The specific and extensive nature of this interaction network makes it a key element to be reproduced in *de novo* designed NTF2-like proteins.

Finally, we analyzed structures in search of simple ways in which the main chain controls pocket shape and size. We found that altering the elements near the opening of the pocket can lead to large changes in pocket size. Specifically, the length and curvature of the long arm, the extension of the frontal hairpin, and C-terminal elements, could be parameters used to control pocket shape and size.

In conclusion, we gathered information that allows us to propose a strategy for *de novo* design of a diverse set of NTF2-like proteins. In the following chapters we will show that by understanding the rules for designing curved sheets, and using the sequence information from this analysis we are able to create *de novo* NTF2-like proteins. We then use this as a base for generating large, diverse sets of *de novo* NTF2-like proteins.

2.5 Materials and methods

2.5.1 Non-redundant set of NTF2-like domain structures

A non-redundant (<95% sequence identity) set of domain crystal structures was downloaded from the SCOPe database from http://scop.berkeley.edu/astral/pdbstyle/ver=2.05 on September 2015. From this set, NTF2-like domains (d.17.4 SCOPe v2.05 superfamily) were extracted by selecting only *.ent files where the domain record line matched d.17.4 at least partially.

2.5.2 Structure-based multiple sequence alignment using PROMALS3D and PDB analysis of secondary structure connections

Given the diversity of NTF2-like superfamily sequences (5) and our goals for producing a multiple sequence alignment (MSA), it is of key importance to include structural information in the process.

PROMALS3D integrates secondary structure prediction and 3D structure of homologous proteins to produce a MSA. We submitted the sequences of selected NTF2-like domains to the PROMALS3D server (http://prodata.swmed.edu/promals3d/promals3d.php), and obtained a MSA using the default options. This MSA was then plotted using SeaView (19), and used as input to generate a local sequence logos (20).

To obtain the amino-acid frequencies given secondary structure and torsion bins, we parsed secondary structure, sequence and torsion (ABEGO) bins from a non-redundant (<90% sequence ID) set of PDB structures, and summarized them in heat maps. The script for this can be found at https://github.com/basantab/NTF2analysis with the name read_stats_HH_cap.py.

2.5.3 Pocket structure analysis using CLIPPERS

An inventory of protein pockets was obtained running CLIPPERS (15) with default options for each of the 92 non-redundant NTF2 domains and their poly-valine counterparts. We then scanned through these inventories searching for the largest pockets using travel depth to define their boundaries in a sequence-agnostic way: We trimmed the pocket tree (done by starting with the deepest, group=1, and walking back with through parents, capping it at group # 120) using a mean_TD cutoff defined as: pocket mean_TD ~ max_TD - (max_TD - lowest_mean_TD) * X, with X = 0.75. The python code for this (pocketDetect_lines_TD_CLIPPERS.py) can be found at https://github.com/basantab/NTF2analysis. After detecting pockets for native and poly-valine models, structures where the pocket was not in the canonical location (sheet concave side), spanned micro-pockets on the surface or native and poly-valine pockets spanned different amino-acid position subsets, were discarded, leaving 62 native/poly-valine pairs for analysis. The structures discarded this way tended to have obliterated pockets or completely lack them, indicating that the pocket detection method produces intuitive results.

Pocket surface rendering was done using the rendering tools available in CLIPPERS for Pymol (The Pymol Molecular Graphics System, Version 2.0 Schrödinger, LLC).

2.5.4 Generating poly-valine models with Rosetta

Poly-valine models were generated using the MakePolyX mover available in the RosettaScripts application (21).

2.5.5 Clustering NTF2-like domains by structural similarity using CLANS and TM-align

In order to generate structurally similar clusters, we did an all vs. all comparison using TM-align (11), obtaining average TM-scores for each structure pair. This TM-score was then turned to a pseudo p-value by calculating pseudo-p-value=10^(-4*(1-(1-TMScore))). These pseudo-p-values for each structure pair were used as input for CLANS (22), where clustering was done with the "network based" method using p-values lower than 0.004.

2.5.6 Visualization of protein structures and image rendering

Images of protein structures were created with PyMOL (23)

2.6 ACKNOWLEDGEMENTS

We thank Sergey Ovchinnikov, Enrique Marcos, Javier Castellanos, Daniel Adriano-Silva, Gabriel Rocklin and Tom Lindsky for providing tools and suggestions for the structural and sequence analysis of NTF2-like domain, as well as helpful comments and discussions in general. We thank Darwin Alonso for technical support.

2.7 LITERATURE

1. Lundqvist, T., Rice, J., Hodge, C. N., Basarab, G. S., Pierce, J., and Lindqvist, Y. (1994) Crystal structure of scytalone dehydratase — a disease determinant of the rice pathogen, Magnaporthe grisea. *Structure*. **2**, 937–944
2. Knudsen, M., and Wiuf, C. (2010) The CATH database. *Hum. Genomics*. **4**, 207–12
3. Fox, N. K., Brenner, S. E., and Chandonia, J. M. (2014) SCOPe: Structural Classification of Proteins - Extended, integrating SCOP and ASTRAL data and classification of new structures.

Nucleic Acids Res. **42**, 1–6
4. Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L., and Bateman, A. (2008) The Pfam protein families database. *Nucleic Acids Res.* **36**, D281-8
5. Eberhardt, R. Y., Chang, Y., Bateman, a, Murzin, a G., Axelrod, H. L., Hwang, W. C., and Aravind, L. (2013) Filling out the structural map of the NTF2-like superfamily. *BMC Bioinformatics*. **14**, 327
6. Kerfeld, C. A. (2004) Structure and function of the water-soluble carotenoid-binding proteins of cyanobacteria. *Photosynth. Res.* **81**, 215–225
7. Bullock, T. L., Clarkson, D. W., Kent, H. M., and Stewart, M. (1996) The 1.6 Å Resolution Crystal Structure of Nuclear Transport Factor 2 (NTF2). *J. Mol. Biol.* **260**, 422–431
8. Johansson, P., Unge, T., Cronin, A., Arand, M., Bergfors, T., Jones, T. A., and Mowbray, S. L. (2005) Structure of an Atypical Epoxide Hydrolase from Mycobacterium tuberculosis Gives Insights into its Function. *J. Mol. Biol.* **351**, 1048–1056
9. Arand, M., Hallberg, B. M., Zou, J., Bergfors, T., Oesch, F., Van der Werf, M. J., De Bont, J. A. M., Jones, T. A., and Mowbray, S. L. (2003) Structure of Rhodococcus erythropolis limonene-1,2-epoxide hydrolase reveals a novel active site. *EMBO J.* **22**, 2583–2592
10. Ha, N.-C., Kim, M.-S., Lee, W., Choi, K. Y., and Oh, B.-H. (2000) Detection of Large p K_a Perturbations of an Inhibitor and a Catalytic Group at an Enzyme Active Site, a Mechanistic Basis for Catalytic Power of Many Enzymes. *J. Biol. Chem.* **275**, 41100–41106
11. Zhang, Y., and Skolnick, J. (2005) TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309
12. Craveur, P., Joseph, A. P., Rebehmed, J., and De Brevern, A. G. (2013) β-Bulges: Extensive structural analyses of β-sheets irregularities. *Protein Sci.* **22**, 1366–1378
13. Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., and Baker, D. (2012) Principles for designing ideal protein structures. *Nature*. **491**, 222–7
14. Pei, J., and Grishin, N. V. (2014) PROMALS3D: Multiple Protein Sequence Alignment Enhanced with Evolutionary and Three-Dimensional Structural Information. in *Methods in molecular biology (Clifton, N.J.)*, pp. 263–271, **1079**, 263–271
15. Coleman, R. G., and Sharp, K. a. (2010) Protein pockets: Inventory, shape, and comparison. *J. Chem. Inf. Model.* **50**, 589–603
16. Salemme, F. R. (1983) Structural properties of protein beta-sheets. *Prog. Biophys. Mol. Biol.* **42**, 95–133
17. Richardsont, J. S., Getzofft, E. D., and Richardsont, D. C. (1978) *The beta bulge: A common small unit of nonrepetitive protein structure*, [online] https://www.pnas.org/content/pnas/75/6/2574.full.pdf (Accessed January 6, 2019)
18. Gieg, R. A historical perspective on protein crystallization from 1840 to the present day. 10.1111/febs.12580
19. Gouy, M., Guindon, S., and Gascuel, O. (2010) SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Mol. Biol. Evol.* **27**, 221–224
20. Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004) WebLogo: A Sequence Logo Generator. *Genome Res.* **14**, 1188–1190
21. Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E. M., Khare, S. D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., Meiler, J., and Baker, D. (2011) Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling suite. *PLoS One*. 10.1371/journal.pone.0020161
22. Frickey, T., and Lupas, A. (2004) CLANS: A Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*. **20**, 3702–3704
23. DeLano, W. L. (2002) The PyMOL Molecular Graphics System, Version 1.1. *Schrödinger LLC*. 10.1038/hr.2014.17

CHAPTER 3. PRINCIPLES FOR DESIGNING PROTEINS WITH CAVITIES FORMED BY CURVED BETA-SHEETS

A version of this chapter has been previously published as:

Marcos*, Enrique, **Basanta*, Benjamin**, Tamuka M. Chidyausiku, Yuefeng Tang, Gustav Oberdorfer, Gaohua Liu, G.V.T. Swapna, Rongjin Guan, Daniel-Adriano Silva, Jiayi Dou, Jose Henrique Pereira, Rong Xiao, Banumathi Sankaran, Peter H. Zwart, Gaetano T. Montelione, David Baker "Principles for designing proteins with cavities formed by curved beta-sheets." *Science* 355.6321 (2017): 201-206.

* These authors contributed equally to this work

3.1 Abstract

Protein beta-sheets are rarely completely regular structures. The twist and bending of sheets provides critical structural diversity that enables proteins to carry out their functions effectively. Understanding the determinants of irregular sheet structures would allow protein design to tap on this diversity and attain protein functions not seen in nature. Here we use simple design principles to generate proteins with large pockets formed by curved beta-sheets, a fundamental stepping-stone towards designing functional proteins.

3.2 Introduction

Proteins with curved beta-sheets that form a pocket, as in the NTF2-like, beta-barrel, and jelly-roll folds, play key roles in molecular recognition, metabolic pathways and cell signaling. Approaches to designing small molecule binding proteins and enzymes to date have started by searching for native protein scaffolds with ligand binding pockets with roughly the right geometry, and then redesigning the surrounding residues to optimize interactions with the small molecule. While this approach has yielded new binding proteins and catalysts (1–5), it is not optimal: there may be no naturally occurring scaffold with a pocket with the correct geometry, and introduction of mutations in the design process may change the pocket structure (6, 7). Building *de novo* proteins with custom-tailored binding sites could be a more effective strategy, but this remains an outstanding challenge (8–11). *De novo* protein design has recently focused on proteins with ideal backbone structures (12–16) (straight helices, uniform beta-strands and short loops; see reference (17) for a an exception) and optimal core side-chain packing, but the binding pockets of naturally occurring proteins lie on concave surfaces formed by non-ideal features such as kinked helices, curved beta-sheets or long loops. The design of proteins with concave surfaces requires examination of how such irregular structural features can be programmed into the amino acid sequence.

As we described in Chapter 2, curved sheets are a key feature of NTF2-like proteins. We noted that in NTF2-like domains strand breaking and bulging tend to correspond with a sharp turn in the direction of the sheet, akin a hinge point. Furthermore, the relative distance between hinge points controls the width

of the sheet base. Beta-sheets have a general tendency to twist in the right-hand direction, when twist is measured in the direction parallel to beta stands (left-handed if measured in the direction perpendicular to strands, on the sheet plane), with the degree of twisting depending on sheet size and pairing composition: Sheets with more strands tend to be flatter, and sheets composed of antiparallel strand pairings tend to be more curved than those composed of parallel pairings. The underlying cause of these phenomena is a combination of steric and hydrogen-bond interactions (18, 19). Previous work reports beta-bulges correspond to areas with accentuated right-handed twist, on the edge strands of antiparallel sheets (20). Strand curvature has been found to be directed by hydrophobic packing, with the energetic cost of less favorable sheet conformations covered by more favorable side-chain interactions (21).

Overall, current knowledge of beta-sheet conformations suggests beta bulges and hydrophobic packing play a key role in stabilizing the conformation of the highly twisted sheets seen in NTF2-like proteins. In this chapter we contribute additional information to understand the forces underlying sheet curvature and twisting, and synthesize it in a set of principles that enable us to design curved beta-sheets from scratch.

3.3 RESULTS

3.3.1 Determinants of sheet curvature

We begin by analyzing how classic (20, 22) beta-bulges (irregularities in the pleating of edge strands) and register shifts (local termination of strand pairing) coupled with intrinsic beta-strand geometry induce curvature in antiparallel beta-sheets (23). We quantify the curvature of an edge strand making an antiparallel pairing with a second strand by the bend angle (Figure 3.1 A). The absolute value of the bend angle (α) at residue *i* is the angle between vectors from the $C_{\alpha(i)}$ atom to $C_{\alpha(i-2)}$ and $C_{\alpha(i+2)}$. The bend angle sign is a function of the relative orientation of a vector *c* describing the concave face of the edge strand (Figure 3.1 A, left), a vector S_{21} between the edge strand and the second strand direction (Figure 3.1 A, right). We analyzed the bend angle of two-stranded antiparallel beta-sheets in naturally occurring protein structures and in Rosetta folding simulations (Figures 3.1 B and S.3.1), and found that uniform strands

tend to have positive bend angles (due to steric interactions between paired beta-strands, Figure S.3.2), while strands containing beta-bulges tend to have negative bend angles (due to the different hydrogen bond pairing of beta-bulges; Figures 3.1. B, S.3.2 and S.3.3). For beta-sheets of three strands or more, we found that the type of strand pairing determines the magnitude of beta-sheet curvature (Figure 3.1 C). In uniform 3-stranded antiparallel beta-sheets, the bend directions of the two edge strand segments point in opposite directions, constraining the bend angle of the inner strand to close to zero and flattening the beta-sheet (Figure 3.1 D, top). In contrast, in 3-stranded beta-sheets with a beta-bulge in one of the edge strands, the two edge strand segments bend in the same direction, leading to increased overall bending of the beta-sheet (Figure 3.1 D, middle). In uniform beta-sheets, register shifts enhance bending by terminating pairing between strand segments that would otherwise have opposite bending directions and flatten the beta-sheet. (Fig. 3.1 D, bottom). Beta-sheet curvature can hence be programmed by combining beta-bulges and register shifts. For example, a number of naturally occurring proteins contain a three-strand beta-sheet core with beta-bulge derived curvature complemented by additional strands with register shifts propagating the curvature (Figure 3.1 E).

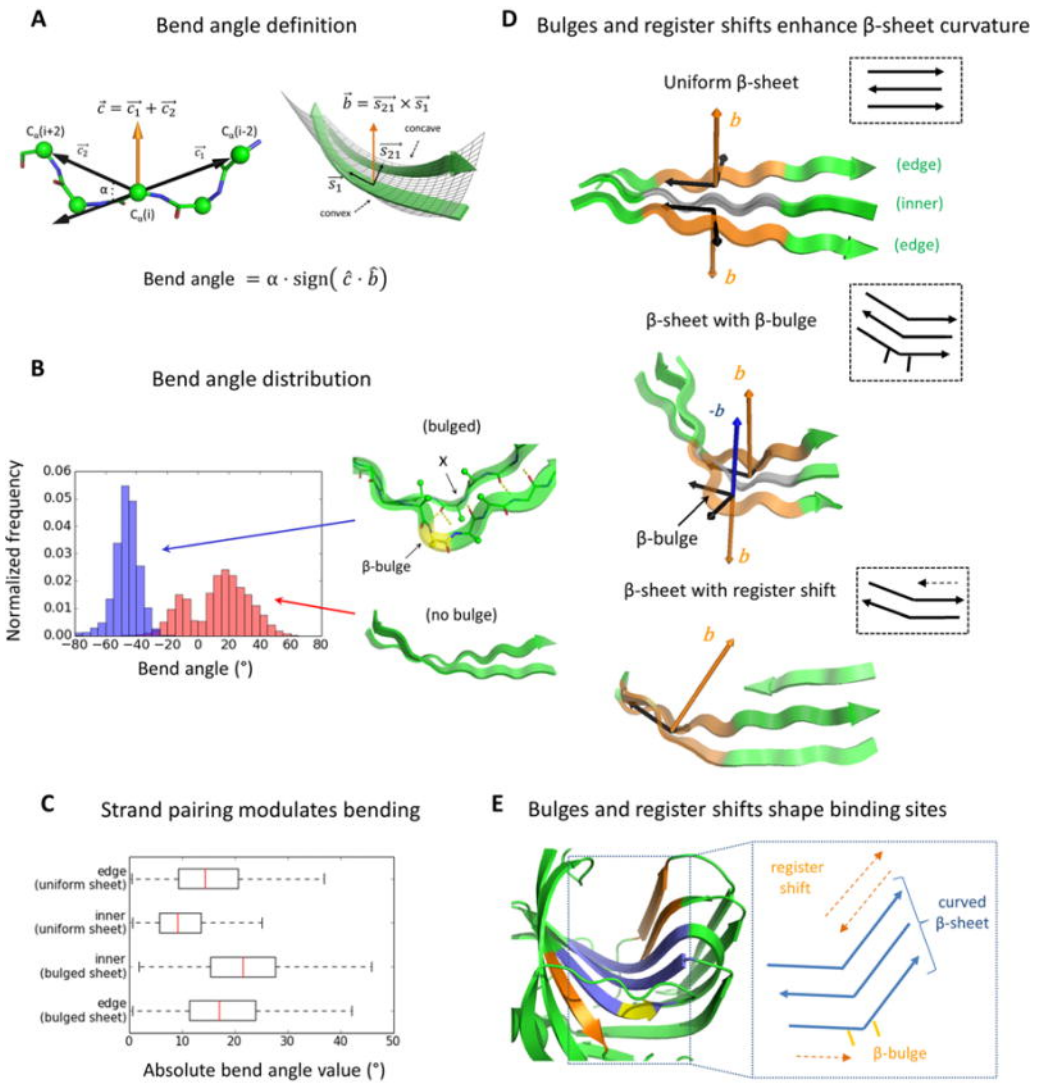


Figure 3.1: **Determinants of sheet curvature A**. Bend angle definition. **B**. Bend angle distributions for strand pairs formed by uniform (red) and bulged (blue) strands. The local hydrogen bonding and offset in side-chain directionality at the beta-bulge position are shown. The bulge and the residue following donate two backbone hydrogen bonds to the same residue X. **C**. Bend angle (absolute value) box plots of strands with different pairing types in native 3-stranded beta-sheets. The edge strand distribution in the bulged beta-sheet case (bottom) is for the strand that does not contain the bulge. **D**. Representation of the *b* vector in edge strand pairs for three types of 3-stranded beta-sheets. Beta-sheet with beta-bulge (middle) shows the -*b* vector for the bulged strand pair to indicate the natural bend direction resulting from a negative bend angle. **E**. On the left, cartoon representation of the binding site formed by a curved beta-sheet in a native xylanase (PDB entry 2B45). The curved 3-stranded beta-sheet core is shown in blue, the beta-bulge in yellow and the extra strands in orange. On the right, schematic representation of strand pairings in the curved beta-sheet formed by a beta-bulge and register shift.

3.3.2 De novo design of proteins with curved beta-sheets

Using these relationships between beta-bulges, register shifts and the direction and magnitude of beta-sheet curvature, we designed six protein folds (labeled from A to F, Figure 3.3) inspired by the naturally occurring cystatin and NTF2-like superfamilies with a 4-stranded antiparallel beta-sheet, beta-bulges at the edge strands and strand lengths ranging between 10 and 14 residues. The width of the beta-sheet central base (along the strand direction) is controlled by the relative position between beta-bulges (folds A, D and B have central bases of increasing width), while the depth (perpendicular to the strand direction) is controlled by the number of strand pairs (folds C, D, E and F increase the depth of folds A and B by adding on two extra strands; Figure 3.2). We complemented the beta-sheets with one (fold A), three (folds B, C and D) or four (folds E and F) α-helices to form overall cone-shaped structures (Figure 3.2 folds B, C and D have wide cone bases, while fold E partially occludes the cone base with the fourth helix), which provide a concave structural niche where functional sites could be designed, with an opening at the base of the cone.

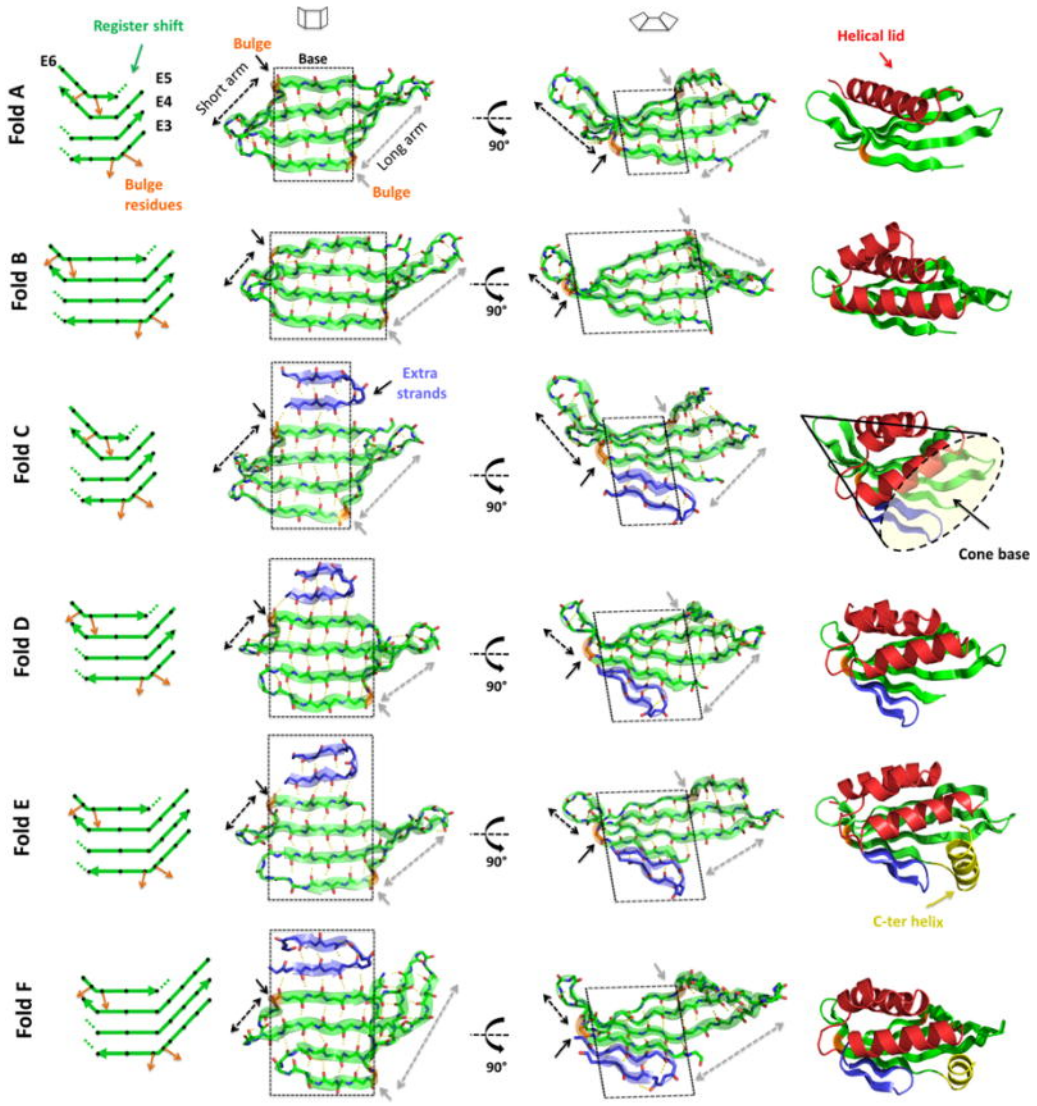


Figure 3.2: **De novo designed curved sheets**. On the left, diagrams of the 4-stranded antiparallel beta-sheets. Black diamonds represent residues with side-chains pointing to the convex face of the beta-sheet and orange arrows highlight the beta-bulge offset in side-chain directionality. Dotted lines show the local termination of strand pairing due to register shift between paired strands. Second and third columns show two views of the designed beta-sheets. Black and gray dashed arrows show the length of the short and long arms, respectively, that emerge from the flat central base (highlighted by a black dashed square). On the right, examples of each designed protein fold containing 4-stranded antiparallel beta-sheets (green), helical lids (red), extra strands (blue) and a C-terminal helix capping the pocket entrance (yellow). The concave base of these conical folds is well suited for small molecule binding site design.

We constructed the protein backbones with a stepwise Monte Carlo fragment assembly protocol (24) that sequentially adds elements of secondary structure (strands and helices), beta-bulges and loops (See sections 3.5 1 and 2). Hairpins were designed with two-residue loops following the beta/beta-rule (25), which requires beta-bulges to be at even and odd positions from the following and previous hairpin loops,

respectively (due to the offset in side-chain directionality of beta-bulges, Figure S.3.3). We then carried out RosettaDesign calculations (26) to favor amino acid identities and side-chain conformations with low-energy, tight packing and high sequence-structure compatibility (See sections 3.5 3 and 4). We hypothesized that beta-bulge positions could be specified at the sequence level solely by changing the normal alternating pattern of polar and hydrophobic amino acids (more complex patterns are observed in native structures (22, 27)) — in a beta-bulge, unlike regular strands, two successive residues point in the same direction (Figure 3.1 B). We relied on side-chain packing to drive strand bending in strands without beta-bulges (21). Loops were designed with sequence profiles obtained from protein fragments with similar backbone torsion angles (See section 3.5 3). In the case of *de novo* NTF2-like folds (Folds C, D, E and F), we enforced the hydrogen-bonding patterns in conserved structural elements described in Chapter 2, as well as the sequence features that enable them.

We characterized the folding energy landscape of the designs by Rosetta *ab initio* structure prediction calculations (28, 29) preceded by a fast initial screen to eliminate designs incapable of folding even with local bias towards the native structure (See section 3.5 4). We chose for experimental characterization designs with funnel-shaped energy landscapes ranging between 74 and 120 amino acids (Table S.3.1) (design names are dcs_X_n; where “dcs” stands for designed curved beta-sheet, “X” the fold type and “n” the design number; and a “_ss” suffix if disulfide bonds are present). Blast searches (30, 31) indicated that the designed sequences had weak or no similarity with native proteins (E-values ranging from 0.00002, for two of the nine fold D designs, to > 10; Table S.3.2); TM-align searches (32) identified structures with global fold similarity, but little sequence similarity (E-values > 10, except for the two designs of fold D with low E-value, where the top Blast hit was re-identified) and differences in the relative orientation of secondary structure elements and loop connections (Figure S.3.6).

We obtained synthetic genes encoding 37 designs, expressed the proteins in *Escherichia coli* and purified them by affinity chromatography. Thirty-three of the designs had far-ultraviolet circular dichroism spectra (CD) at 25 °C characteristic of alpha/beta proteins, and were monomeric by size-exclusion chromatography coupled with multi-angle light scattering (SEC-MALS; Figure 3.3, S.3.7 – 10, Table S.3.1). Thirty-one of the designs have a melting temperature (T_m) above 95 °C and 24 unfold

cooperatively in guanidinium chloride (GdmCl). Two-dimensional $^{1}H–^{15}N$ heteronuclear single quantum coherence (HSQC) spectra suggest that twelve designs fold into well-ordered structures. Fold E designs, which have a long C-terminal helix as a lid capping the cone base, were the most stable (with T_m > 95 °C and denaturation midpoints up to 6M GdmCl at 25 °C; figure S.3.11 and Table S.3.3). Fold F designs were also stable at high temperatures, but in some cases their non-cooperative unfolding and poor HSQC spectra (Figure S.3.10) suggest imperfect design of the short C-terminal helix interaction with the long arm.

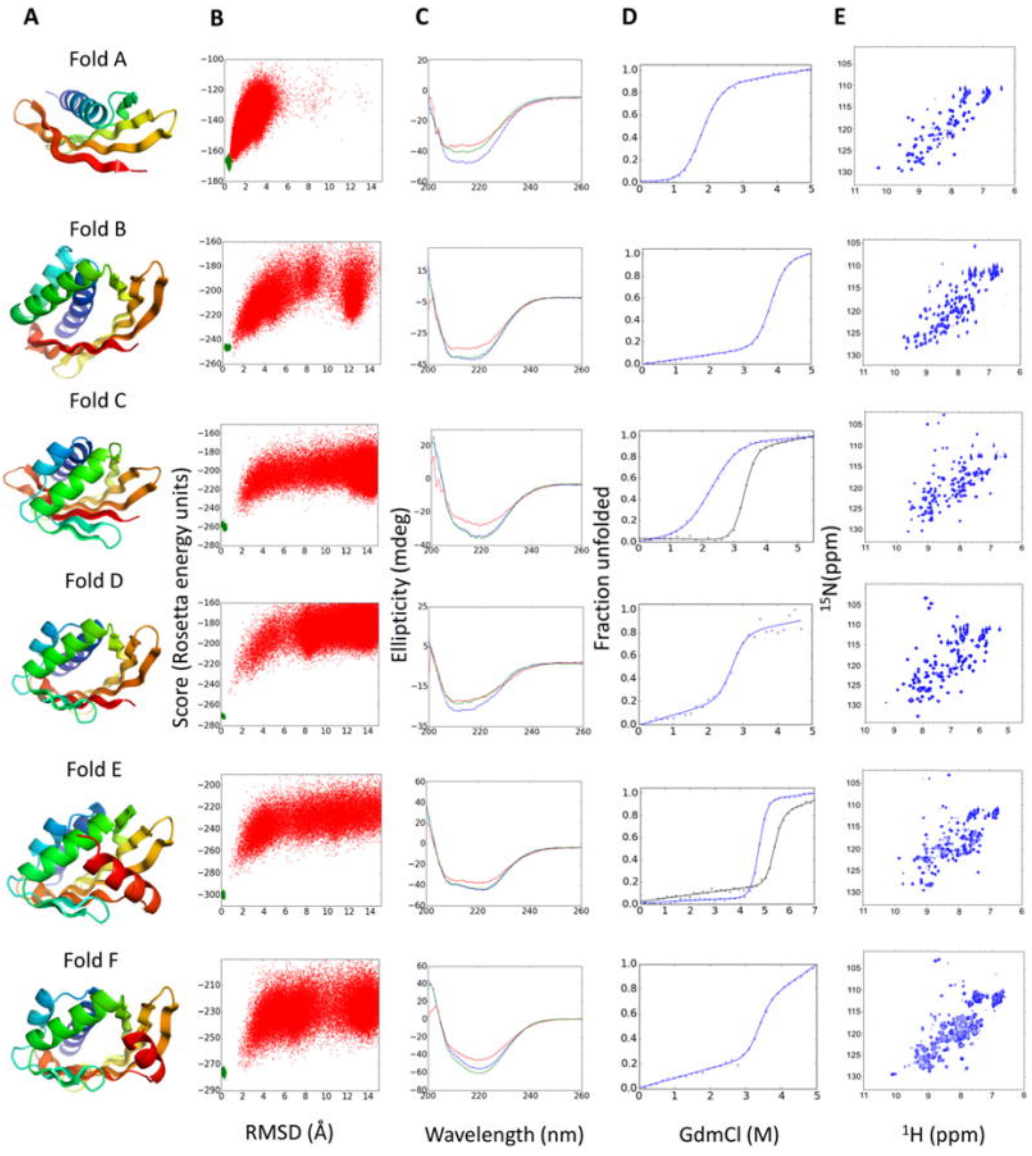


Figure 3.3: De novo designed folds with curved sheets A. Examples of design models for each fold. **B**. Folding energy landscapes generated by *ab initio* structure prediction calculations. Each dot represents the lowest energy structure identified in an independent trajectory starting from an extended

chain (red dots) or from the design model (green dots); x-axis shows the Cα-root mean squared deviation (RMSD) from the designed model; the y-axis shows the Rosetta all-atom energy. **C**. Far-ultraviolet circular dichroism spectra (blue: 25 °C, red: 95 °C, green: 25 °C after cooling). **D**. Chemical denaturation with GdmCl monitored with circular dichroism at 220 nm and 25 °C. For folds C and D the denaturation curves for designs stabilized by a disulfide bond or a dimer interface are shown in black lines. **E**. $^{1}H–^{15}N$ HSQC spectra obtained at 25 °C.

We reasoned that when designing function into these *de novo* scaffolds, the proximity between the active site and the protein core could compromise protein stability. We therefore decided to explore homo-dimer design, which, as discussed in Chapter 2, could be a source of stability in native NTF2 domains. We designed homo-dimers of fold E designs with shape complementary low energy interfaces formed by the convex face of the curved beta-sheet (See section 3.5.7). Nine designs with deep global energy minima at the designed dimer configuration in docking calculations were selected for experimental characterization and three were found to form soluble dimers; the best expressed design, dcs_E_4_dim9, is 1.4 kcal·mol^{-1} more stable than the original monomer (Figure S.3.11).

We solved the structures of nine designs by NMR spectroscopy or X-ray crystallography. These experimental structures span five different folds (from A to E) (Figure 3.4) and are in close agreement with the computational models (Cα-RMSDs from 1.0 to 2.1 Å). The overall beta-sheet curvatures were accurately recapitulated and beta-bulge positions were as predicted, supporting our hypothesis about local encoding of beta-bulges. Crystal contacts in the structures of dcs_C_1_ss, dcs_D_2, dcs_E_3, dcs_E_4 and dcs_A_4 support the idea that beta-bulges minimize edge-to-edge strand pairing (33): hydrogen bond pairing is restricted to the regular strand segments (Figures S.3.12 and S.3.12).

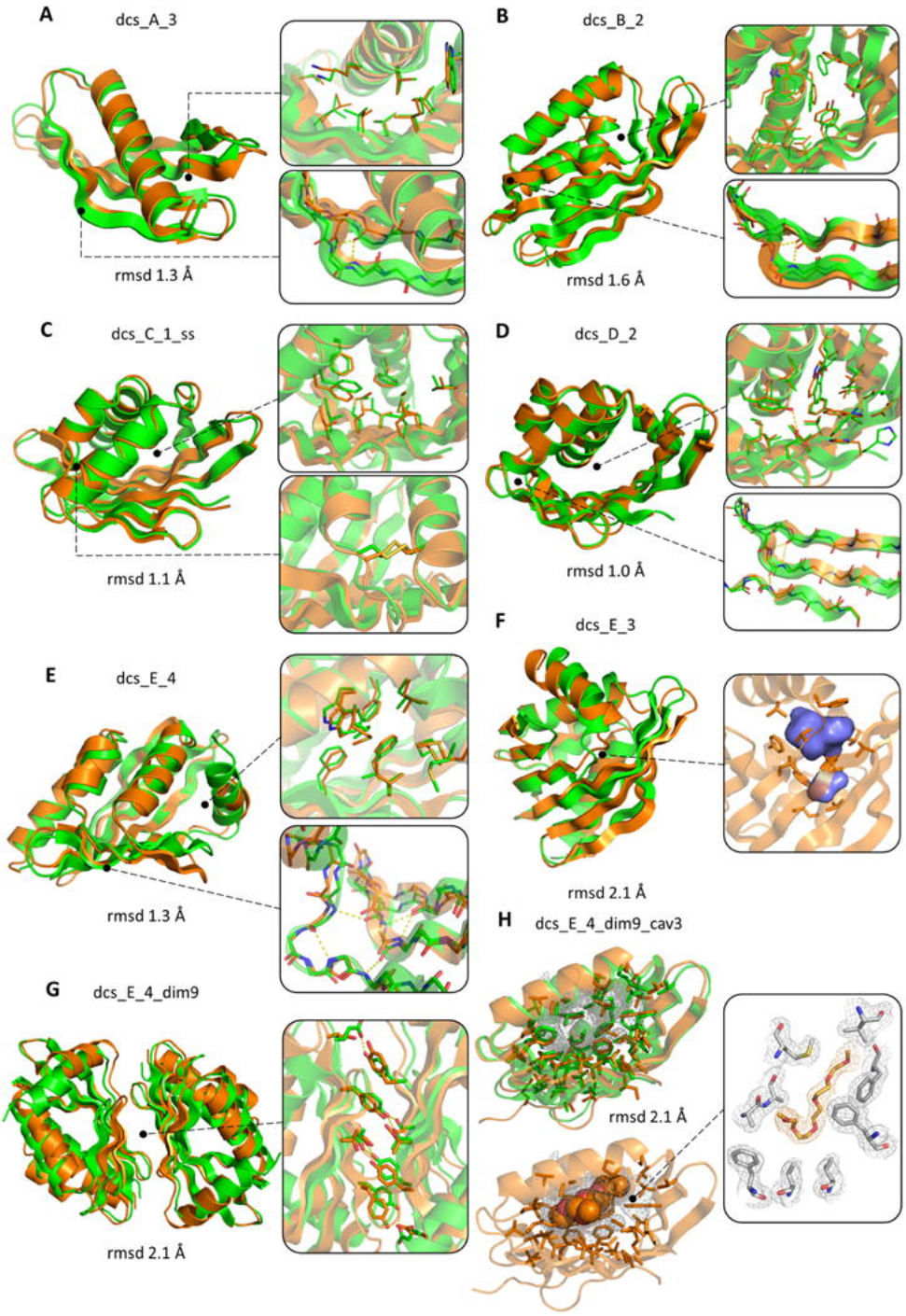


Figure 3.4: In each panel the experimental structure and the design model are superimposed and colored in orange and green, respectively. Insets show comparisons of side-chain rotamers, beta-bulge geometry and cavities; and designed side-chain and beta-bulge hydrogen bonds are shown in yellow dashed lines. The RMSD calculated over all C_a atoms is shown in each panel. **A** dcs_A_3 and **B.** dcs_B_2 were solved by NMR (comparisons utilized the lowest energy NMR model). **C**. dcs_C_1_ss (3.0 Å resolution) with designed disulfide bond in inset. **D**. dcs_D_2 (2.0 Å resolution). **E**. dcs_E_4 (2.9 Å resolution). **F.** dcs_E_3 (3.1 Å resolution); an internal hydrophobic cavity forms in both the design and the crystal structure (volume 192 $Å^3$). **G**. dcs_E_4_dim9 (2.4 Å resolution); the interface aromatic stacking and hydrogen bonding interactions are very similar in the crystal structure and design model (right inset).

H. dcs_E_4_dim9_cav3 (1.8 Å resolution). A large (520 $Å^3$) cavity is filled with a pentaethylene glycol molecule in the crystal structure (bottom left; electron density map is on right and design model on upper left). The C-terminal helix and the dimer interface are not shown for better visualization of the cavity.

The experimental structures for folds A (dcs_A_3 by NMR, Figure 3.4 A; and dcs_A_4 by X-ray crystallography, figure S.3.12) and B (dcs_B_2 by NMR; Figure 3.4 B) are in close agreement with the design models in the core of the beta-sheet and the helices. The designed side-chain packing between the tips of the two beta-sheet arms and the helix was better recapitulated in dcs_A_3 and dcs_A_4 than in dcs_B_2 (compare figures 3.4 A and B, right insets) where the long arm is more twisted in the NMR structure than in the design model; full control over beta-sheet geometry in these folds likely requires control over side-chain packing between the beta-sheet and the helical lid.

The crystal structures of fold C and D (Figures 3.4 C and D) are very close to the design models with designed aromatic packing and hydrogen bonding interactions bridging the protein core and the cone base; a designed disulfide bond is also correctly recapitulated (Figure 3.4 C, bottom inset). The two crystal structures for fold E monomeric designs also closely match the design models (Figures 3.4 E and F) with the cone base capped by the C-terminal helix in two different orientations. A buried cavity designed in one of these (dcs_E_3) expands toward the cone base in the crystal structure (Figure 3.4 F). We explored the ability of the fold C and D designs to support cavities by reducing the size and increasing the polarity of side-chains at the cone base (Figure S.3.15). Five of the nine redesigns tested (with up to 19 mutations) were soluble and monomeric (Figure S.3.16 and Table S.3.3).

The crystal structure of the designed homo-dimer dcs_E_4_dim9 closely matches the computational model over both the individual subunits and the designed beta-sheet interface (Figure 3.4 G and S.3 19). We designed large cavities by truncating side-chains at the cone base (Figure S.3.15, Tables S.3.2 and 3). The crystal structure of one such design revealed a large (520 $Å^3$) cavity very similar to that in the design model, lined by the curved curved-sheet (Figures 3.4 H and S.3.18; a pentaethylene glycol fills the cavity). This is the largest *de novo* designed cavity to date, and illustrates how large core packing vacancies can be programmed by designing curved beta-sheets topped by helices.

3.4 Discussion

The NMR and crystal structures show that NTF2-like proteins can be accurately designed *de novo* with the principles we have identified and using the information obtained from NTF2 domain analysis (Chapter 2). The designed proteins exhibit a rich combination of structural features: curved beta-sheets with beta-bulges and register shifts, loops of variable length, helices, disulfide bonds, beta-sheet interfaces and cavities. The hydrogen bond pattern we designed in the H2-S1 5-residue loop was recapitulated in the crystal structures, showing how sequence information can be used create irregular, yet rigid, secondary structure elements. We explored the role of NTF2 domain homo-oligomerization in our *de novo* system, and showed that it increases stability without interfering with the accessibility of potential binding pockets. The *de novo* NTF2 dimer tolerates substitutions of large hydrophobic side-chains to smaller or polar residues to line large pocket.

Computational methods have been used to design enzyme catalysts by defining an ideal active site ("theozyme") and then searching for placements of the theozyme in native protein scaffolds. This approach has yielded catalysts for a number of chemical reactions, including reactions not catalyzed by naturally occurring enzymes, but the resulting activities tend to be well below those of natural enzymes. This likely result from two shortcomings in the design strategy: the detailed theozyme geometry cannot be perfectly realized in any pre-existing scaffold, and the sequence changes introduced in the design process can produce unpredictable changes in structure (6, 7). The principles we describe in this work should pave the way to overcoming these issues by opening the design space of irregular beta-sheets.

Future design efforts should focus on expanding the diversity presented here, by combining the different sources of diversity explored here, and adding new ones. Only a subset of the variables in NTF2-like domains is sampled here, in an attempt to focus on deviations of beta-sheet regular structures. But, as described in Chapter 2, NTF-2 like domains sample a large variety of pocket sizes, which are a result of equally diverse backbone changes. This exploration is the focus of the next chapter.

3.5 Materials and methods

3.5.1 Protein backbone construction

Protein backbones were generated by Monte Carlo fragment assembly using 9 and 3-residue fragments with the target secondary structure and torsion bins (ABEGO), using the Blueprint Builder mover (34) implemented in RosettaScripts (35). We restricted regular strand and bulge residues to the "B" and "A" ABEGO bins, respectively. These Rosetta folding simulations use a sequence-independent centroid representation of the protein, as well as a scoring function that includes a hydrogen bonding term for backbone atoms, a soft Van der Waals term to avoid steric clashes, an omega angle term to ensure planarity of the peptide bond, and a radius of gyration term to favor compact structures. Thousands of independent folding trajectories are performed and subsequently filtered.

When building backbones involving non-local contacts, adding a constraint term to the scoring function increases the efficiency of the folding simulations. Due to the non-local character of beta-sheet contacts, we used distance and angle constraints to favor the ideal geometry of the backbone-backbone hydrogen bonds between the paired strand residues. For bulged strand pairs both the bulge and the residue following donate hydrogen bond to the same residue ("X") in the paired strand, but with different hydrogen bond distances according to distributions from native protein structures. The hydrogen bond distances to residue X from the bulge and the residue following were constrained to 2.9 and 3.4 Å, respectively. Conveniently, once the register shift between paired strands and the strand pairing type (parallel or antiparallel) are defined all pairings between strand residues and their corresponding constraints are determined. Additionally, when building flexible elements such as N- or C-terminal helices or loops with a particular hydrogen bond pattern (such are the cases for H1-H2 and H2-S1 loops), the use of constraints allows sampling structures closer to the target with more efficiency.

Strand fragments with low bending and twist are overrepresented in the fragment library derived from the PDB and, as a consequence, constraints are necessary to favor the construction of backbones with increased strand bending and twist. We used angle constraints between C-alpha atoms of residues with the same pleating at different separation levels, i.e. $C_\alpha(i\text{-}2n)$-$C_\alpha(i)$-$C_\alpha(i\text{+}2n)$ where i is the central residue and n is the separation level. Similarly, for twist we used dihedral constraints $C_\beta(i)$-$C_\alpha(i)$-$C_\alpha(i\text{+}2n)$-

$C_\beta(i+2n)$. The separation level provides control on the degree of locality of strand curvature. In addition, the inter-strand twist for an antiparallel pairing can also be controlled with dihedral constraints for $C_\alpha(k+2)$-$C_\alpha(k)$-$C_\alpha(p_k)$-$C_\alpha(p_k-2)$, where p_k is the strand residue paired to residue k.

3.5.2 Stepwise backbone building

The introduction of constraints in the fragment sampling trajectory can rapidly increase the ruggedness of the energy landscape, leading to its frustration, i.e. the trajectory sticking at a local energy minimum. This is a general limitation of the fragment-based approach that we circumvented by building backbones stepwise and constraining non-local contacts. We divided the construction of the target folds in several steps. For instance, for Fold E: (1) central 4-stranded antiparallel beta-sheet with a beta-bulge in each edge-strand. This step is further subdivided in three substeps: fist the two non-bulged, central strands are built, and twist and bend are imparted here. Then the two additional bulged flanking strands are build, using only the strands pairings for sampling guidance. (2) Helix 3 and hairpin-interdomain connection. (3) Helices 1 and 2 added at the N-terminus. (4) Addition of C-terminal helix. See figure 3.5.2. We used constraints for building the beta-sheet (strand pairings, bending and twist), the inter-domain connection and positioning helices 1 and 2. Helix 1 is a flexible element at the N-terminus that we constrained at interacting distance from the edge strand of the beta-sheet. The loop connecting helix 2 and the inter-domain connection was constrained to hydrogen bond the backbone of the edge strand. Helix 4 in Fold E designs was constrained to pack onto helix 3 at the entrance of the pocket. A complete example and working code for building all six folds are available at https://github.com/basantab/DeNovoCurvedSheetDesign.

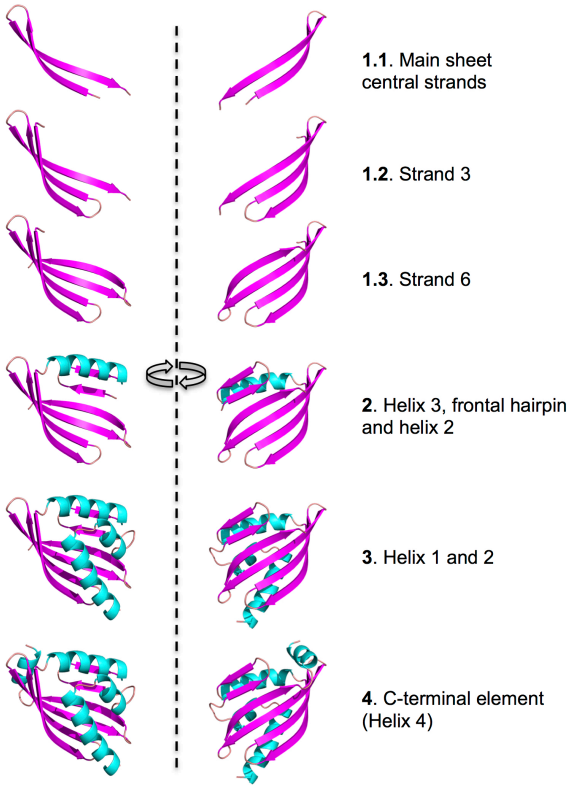


Figure 3.5.2: Graphical description of backbone assembly process for Fold E.

We have used four criteria to filter protein backbones at each step:

1. *Target topology:* protein models are filtered according to the match between the blueprint and the detected secondary structure, ABEGO sequence and topology (strand and helix pairings) of the built model.
2. *Native-like backbones:* to favor native-like backbones, protein models are also filtered on the basis of backbone hydrogen bonding energy (lr_hb score), Cβ-average degree (average number of Cβ-Cβ contacts between residues within 10 Å) and balance between exposed and buried SASA to favor compact structures. Additionally, we checked for deviations between backbone fragments of the designed structures and native fragments (FragmentLookup filter), which is indicative of local backbone strain.
3. *Geometrical features defining target structure:* depending on the protein topology to be built, additional filters are considered to evaluate the geometry of secondary structure elements as well as

their relative orientation, such as the strand twist/bending or the distance/angle between helix and strand.

Canonical loops: The conformations of loops connecting two secondary structure elements can be discretized by the sequence of their torsion bins (ABEGO). Previous works (14, 25, 36) have mined the PDB for information on the relationship between loop length and ABEGO, and the orientation and type of the secondary structure elements they bridge; we used the information obtained in Chapter 2 regarding NTF2-like domain loops to select the length and ABEGO bins of all loops. Only using the most frequent loop ABEGOs facilitates the design of their amino acid sequences, as explained below. In the case of the H3-S3 connection, which shows significant diversity in NTF2-like domains, we selected a one-residue loop with a B ABEGO conformation, that resembles a bulge at the beginning of S3, as seen in a few NTF2-like domains. The main reason for this selection is its short length and simple hydrogen bond pattern, but it also fits the orientation transition between H3 and S3. For Fold F a slightly conformation is chosen, as it fits better the orientation transition of H3 and S3 in this fold.

3.5.3 Sequence design

Thousands of backbones are subjected to RosettaDesign calculations (26, 37) with the full-atom Talaris2013 (38, 39) scoring function to favor amino acid identities and side-chain conformations with low-energy and tight packing. The design calculation corresponds to cycles of fixed backbone design followed by backbone relaxation, and the designs were filtered based on three independent criteria:

- Low total energy
- Tight packing: RosettaHoles (40), shape complementarity between secondary structure elements, packstat, and core side-chain average degree. Side-chain average degree is the average number of hydrophobic side-chain heavy-atom contacts within 4 Å. We developed this filter to improve the packing in the core of protein folds with large pockets, which are difficult to pack efficiently. This minimized the number of alanine residues in helices and valine residues in strands, while increasing the number of large hydrophobic side-chains.

- High sequence-structure compatibility: match between secondary structure of the designed structure and Psipred (41) secondary structure prediction from the designed amino acid sequence.

To achieve very low energy sequences with tight packing, for each backbone we ran multiple Generic Monte Carlo trajectories of the design protocol, optimizing simultaneously total energy and side-chain average degree, and subsequently applied all filters. The design calculations are performed using a restricted set of amino acids and rotamers for each position. The restrictions were such that hydrophobic amino acids were allowed in the core and polar amino acids in the surface. To improve the local sequence-structure compatibility in loops and beta-bulges we restricted their amino acid identities to the subset of amino acids most frequently observed in similar fragments in the PDB. This was done by the creation of sequence profiles for loops that shared the same ABEGO bins and adjacent secondary structure elements (See Chapter 2, section 2.5.2). The top 5 most frequent amino acids in each position were the only ones allowed, unless there was a strong preference for a particular amino acid. For the conserved connections described in Chapter 2, since in most cases the profiles extracted from NTF2-like domains were similar to those in a general set of protein structures, we enforced the profiles obtained from the general structure set. Additionally, amino acids identities conflicting with the expected hydrophobicity pattern were excluded. The loop ABEGO classification in combination with the corresponding sequence profile allows the automatic identification of well-known local sequence-structure motifs, such as N-terminal helix capping residues (D, N, S and T) or proline residues that restrict the Phi/Psi angles of the residue immediately before. These sequence motifs are seldom identified by the score function, thus giving poor local sequence-structure compatibility. For beta-bulges we built sequence profiles for positions *b*-1, *b*, *b*+1, *b*+2 and X; where *b* is the bulge position and X is the strand residue paired to the bulge. In general, positions *b* and *b*+1 were restricted to RKEQ, and *b*-1 and *b*+2 to ILVFY. To minimize the aggregation propensity, we incorporated polar residues at inward-pointing positions of edge strands and removed surface exposed hydrophobic residues. Due to the large size of the pockets of the target folds, efficient core packing was achieved by a high number of aromatic side-chains. As part of the protein core is solvent-exposed we preserved well-packed exposed aromatics that hydrogen bond polar residues at the surface (especially Trp-Glu and Tyr-Glu interactions).

3.5.4 Evaluation of sequence-structure compatibility

The compatibility between sequence and backbone structure is assessed in three steps:

1) *Fragment quality assessment.* The designed model sequence is spliced in overlapping 9-residue fragments, and two hundred 9-residue fragments with the same sequence and secondary structure are picked from a PDB-derived fragment database for each position. The RMSDs between all picked 9-mer fragments and the corresponding 9-mer of the designed structure are calculated. Two metrics evaluating the overall structural similarity between the ensemble of picked fragments and the designed structure are calculated to rank designs based on fragment quality. First, the percentage of fragments with RMSD < 1.5 Å and, second, the RMSD of the best fragment at the worst position. The quality of these fragments tests compatibility of the sequence and backbone structure at the local level.

2) *Biased Forward Folding.* After verifying the fragment quality, the sequence-structure compatibility is assessed at the global level by characterizing the folding energy landscape with Rosetta *ab initio* folding simulations starting from an extended chain (29, 42), on the Rosetta@home server. This is the most stringent computational test and those designs with funnel-shaped energy landscapes are selected for experimental characterization. In general, hundreds of designs pass the fragment quality filter and their folding energy landscape should be assessed. However, these simulations are too computationally demanding. The high contact order of the protein folds targeted in this work complicated the identification of designs with funnel-shaped energy landscapes and required to screen by *ab initio* folding too many designs with good fragment quality. We developed a new method, *Biased Forward Folding*, to quickly assess the folding energy landscape and select the most promising candidates for unbiased *ab initio* structure prediction. The standard Rosetta *ab initio* structure prediction method starts with a fragment picking process in which at each residue position 9- and 3-residue fragments are selected from the fragment library on the basis of similarity in sequence and secondary structure prediction. The top scoring fragments are then subjected to a Monte Carlo assembly process using a low resolution scoring function and, in a second step, the lowest energy structures are relaxed with a high-resolution scoring function. The fragment assembly process performs the large-scale conformational sampling, while the high-resolution relaxing step is limited to local backbone perturbations allowing side-chains to repack and find

low energy structures. Therefore the selection of fragments and their assembly process are the two primary limiting factors in sampling conformations close to the designed structure and obtain funnel-shaped energy landscapes. We hypothesized that those picked fragments structurally similar to the designed structure fragments are the main contributors to sampling near the designed structure during *ab initio*. Biasing *ab initio* folding simulations using a small subset of fragments close in RMSD to the design structure is therefore expected to have predictive power of the funnel character of the energy landscape near the design structure. If under this bias, sampling trajectories do not reach the target structure it is very unlikely that the standard *ab initio* simulation will sample closer. With a smaller set of fragments the number of folding trajectories necessary to map the energy landscape available gets dramatically reduced. We selected the three lowest-rmsd fragments (9 and 3 residues long) picked at each position and ran a low number of *ab initio* folding trajectories (between 30 and 50). This allows screening 10-100 times more designs than with *ab initio* folding simulations.

3) *Ab initio structure prediction*. Those designs having funnel-shaped energy landscapes in Biased Forward Folding simulations are then subjected to standard *ab initio* structure prediction simulations on Rosetta@home. For an energy landscape obtained from Biased Forward Folding or *ab initio* structure prediction to be funnel shaped we required to get sampling below 2 Å RMSD to the relaxed structure and a large energy gap with alternative structures to ensure that the designed structure is achievable and lower in energy to alternate states.

3.5.5 Design of disulfide bonds

We used the *Disulfidize* mover implemented in RosettaScripts to screen for pairs of residue positions with proper geometry for disulfide bond formation. We favored disulfide bonds between residues distant in primary sequence (at least a 6-residue separation) and with a disulfide score < -1.0. To increase the likelihood of finding good geometries for disulfide bond we locally perturbed the backbone structure with small moves (42) using the *Small* mover in RosettaScripts.

3.5.6 Cavity-creating mutations

We rationally selected amino acid positions close to the cone base and restricted the design to amino acid identities with smaller hydrophobic or polar side-chains. For dcs_E_4_dim9 mutants we selected those positions within contact distance of the diethylene glycol molecule bound to chain A.

3.5.7 Computational design of homo-dimers

We used the Residue Pair Transform method (43) to generate docking configurations with C2 symmetry suitable for designing the homo-dimer interface. We restricted the docking process to configurations that exclude helices from the dimer interface and maximize the number of β-sheet contacts. The top 50 scoring docked configurations were subjected to interface design calculations. Those beta-sheet residues at the convex face with the C_β atom within 10 Å of a C_β atom of the other subunit were selected for design. The possible amino acid identities at each design position were restricted based on the solvent accessible surface area (SASA). Designs were filtered based on buried SASA, shape complementarity and binding energy. Designs passing these criteria were subjected to asymmetric docking simulations and those with funnel-shaped energy landscapes were selected for experimental characterization.

3.5.8 Visualization of protein structures and image rendering

Images of protein structures were created with PyMOL (44) and Chimera (45).

3.5.9 Protein expression and purification

Genes encoding the designed protein sequences were obtained from Genscript and cloned into pET21_NESG (46, 47) (with C-terminal 6xHis tag) or pET-28b+ (with N-terminal 6xHis tag and a thrombin cleavage site) expression vectors. Plasmids were transformed into chemically competent *Escherichia coli* BL21 Star (DE3) cells from Invitrogen. Starter cultures were grown at 37°C in Luria-Bertani (LB) medium

overnight with antibiotic (50 µg/ml carbenicillin for pET21-NESG expression or 30 µg/ml kanamycin for pET-28b+ expression). For expression of non-isotopically-labeled proteins, overnight cultures were used to inoculate 500 ml of LB medium supplemented with antibiotic. To express ^{15}N-labelled proteins for NMR spectroscopy, starter cultures were transferred to 40 mL of MJ9 minimal media (48) with antibiotic, grown overnight and used to inoculate 500 ml of minimal media. After inoculation, cells were grown at 37 °C and 225 r.p.m until an optical density (OD_{600}) of 0.5-0.7 was reached. Protein expression was then induced with 1mM of isopropyl beta-D-thiogalactopyranoside (IPTG) at 18 °C. After overnight expression, cells were collected by centrifugation (at 4 °C and 4400 r.p.m for 10 minutes) and resuspended in 25 ml of lysis buffer (20 mM imidazole and phosphate buffered saline, PBS - 137 mM NaCl, 12 mM Phosphate, 2.7 mM KCl, pH 7.4). Resuspended cells were lysed by sonication or microfluidizer in the presence of lysozyme, DNAse and protease inhibitors. Lysates were centrifuged at 4 °C and 20,000 r.c.f. for 30 minutes; and the supernatant was filtered and loaded to a nickel affinity gravity column pre-equilibrated in lysis buffer for purification. The column was washed with three column volumes of PBS+30 mM imidazole and the purified protein was eluted with three column volumes of PBS+250 mM imidazole. The eluted protein solution was dialyzed against PBS buffer overnight. The expression of purified proteins was assessed by SDS-polyacrylamide gel electrophoresis and mass spectrometry; and protein concentrations were determined from the absorbance at 280 nm measured on a NanoDrop spectrophotometer (ThermoScientific) with extinction coefficients predicted from the amino acid sequences. Proteins were further purified by FPLC size-exclusion chromatography using a Superdex 75 10/300 GL (GE Healthcare) column.

3.5.10 Site-directed mutagenesis

Single-point mutations were obtained by QuikChange site-directed mutagenesis using 0.75 µl of the pET-28b+ constructs as templates, 1 µl of Phusion high-fidelity DNA polymerase (New England BioLabs), 10 µl of 5X Phusion buffer (New England BioLabs), 1.25 µl of a 10 mM deoxynucleotides (dNTP) solution mix and 1 µl of the designed forward and reverse primers solutions at 125 ng/µL. Primers were ordered from Integrated DNA Technologies. Full-length gene product was assembled by 1 cycle of PCR (95 °C

1.5 min), 18 cycles of PCR (95 °C 30 s, 55 °C 30 s, 72 °C 4 min) and 1 cycle of PCR (72 °C 6 min). Mutations were confirmed by sequencing.

3.5.11 Circular dichroism (CD)

Far-ultraviolet CD measurements were carried out with an AVIV spectrometer, model 420. Wavelength scans were measured from 260 to 195 nm at temperatures between 25 and 95 °C. Temperature melts monitored absorption signal at 220 nm in steps of 2 °C/min and 30 s of equilibration time. For wavelength scans and temperature melts a protein solution in PBS buffer (pH 7.4) of concentration 0.2-0.4 mg/ml was used in a 1 mm path-length cuvette.

Chemical denaturation experiments with guanidium chloride (GdmCl) were done with an automatic titrator using a protein concentration of 0.02-0.04 mg/ml and a 1 cm path-length cuvette with stir bar. PBS buffer (pH 7.4) was used for the cuvette solution and PBS+GdmCl for the titrant solution at the same protein concentration. GdmCl concentration was determined by refractive index. The denaturation process monitored absorption signal at 220 nm in steps of 0.2 M GdmCl with 1 min mixing time for each step and at 25 °C. The denaturation curves were fitted by non-linear regression to a two-state unfolding model to extract six parameters: slope and intercept for pre- and post-transition baselines, *m* value and the folding free energy (ΔG_{H2O}) (49, 50). The deviation of the fitted *m* value from its expected value given protein size was computed using the empirical correlation between the number of protein residues and the protein *m* value for denaturation with GdmCl (51).

3.5.12 Size exclusion chromatography combined with multiple angle light scattering (SEC-MALS)

SEC-MALS experiments were performed using a Superdex 75 10/300 GL (GE Healthcare) column combined with a miniDAWN TREOS multi-angle static light scattering detector and an Optilab T-rEX refractometer (Wyatt Technology). One hundred microliter protein samples of 1-3 mg/ml were injected to the column equilibrated with PBS (pH 7.4) or TBS (pH 8.0) buffer at a flow rate of 0.5 ml/min. The

collected data was analyzed with ASTRA software (Wyatt Technology) to estimate the molecular weight of the eluted species.

3.5.13 Nuclear magnetic resonance spectroscopy

^{15}N-HSQC screening

To evaluate whether the designed proteins fold into well-ordered structures ^{15}N-HSQC screening was carried out at 20 or 25 $^{\circ}$C using a 1.7 mm micro cryoprobe with automatic sample changer at 600 MHz. The spectra were generally recorded in multiple buffers, using standard protocols that have been published previously (46, 52). The buffers and temperatures providing the best quality spectra were used for the analyses provided in this study.

NMR structure determination of dcs_A_3 and dcs_B_2

The selected designs (dcs_A_3, NESG target OR485; dcs_B_2, NESG target OR664) were expressed and purified by following the standard NESG protocols (46). Synthetic genes (Genscript) cloned into the pET21_NESG expression vector (46, 47) were expressed in *E. coli* BL21 (DE3) pMGK cells as *U*-^{15}N, 5%^{13}C-enriched, and *U*-^{15}N, *U*-^{13}C-enriched proteins, using MJ9 minimal media (48), ^{13}C-glucose and $^{15}NH_3Cl$ as the sole sources of carbon and nitrogen, respectively. *U*-^{15}N, 5%^{13}C-labeled proteins were generated for stereo-specific assignments of isopropyl methyl groups of valines and leucines (53). Samples were determined to be homogeneous (>95%) by SDS-PAGE, and monomeric by size exclusion chromatography. The molecular weights of ^{13}C,^{15}N-enriched OR485 and ^{13}C,^{15}N-enriched OR664 were confirmed as 10.61 kDa and 14.31 kDa by MALDI-TOF, respectively, in good agreement with theoretical values (10.64 kDa and 14.33 kDa, respectively). The yields were 20 mg and 15 mg per liter culture, respectively.

All NMR spectra were recorded at 25 °C using Bruker AVANCE NMR spectrometer systems with cryogenic NMR probes at 600 and 800 MHz. The NMR structures were determined using standard NMR structure determination protocols, as previously described (54). NMR structures were determined in a “blind” fashion; i.e. without knowledge of the design structure. Structure quality assessment was done

using the Protein Structure Validation Software (PSVS) software suite (55, 56). Chemical shifts data and final structure coordinates were deposited in the Biological Magnetic Resonance Bank and Protein Data Bank, respectively. (NESG ID, BMRB and PDB IDs: OR485, BMRB 30139, 5kph for dcs_A_3; and OR664, BMRB 30128, 5kpe for dcs_B_2). The refinement statistics for the final structures are summarized in Table S.3.6.

3.5.14 Crystallization, data collection and structure determination

dcs_A_4 (NESG target OR486)

A DNA fragment encoding dcs_A_4 was synthesized and cloned into the bacterial expression vector pET21_NESG (46, 47), with a short C-terminal purification tag "LEHHHHHH". The plasmid was then transformed into *E. coli.* BL21(DE3) cells (Strategene) and grown in LB media (1L) at 37 °C to 0.8 OD_{600}, and induced with 1 mM IPTG over night at 17 °C. The bacteria were pelleted by centrifugation at 8000 r.c.f, and resuspended in PBS buffer, mild sonication was used to lyse cells. The lysate was clarified by centrifugation at 20000 r.c.f., then the supernatant was applied to a 5 ml His-tag affinity column (GE Healthcare), and eluted with PBS with 500 mM imidazole added. Further purification was carried out by size exclusion chromatography using a HighLoad 26/60 Superdex S75 column (GE Healthcare). The purified protein was over 95% pure based on SDS PAGE, and was also validated by MALDI-TOF mass spectrometry.

The purified dcs_A_4 (NESG target OR486) was concentrated to 10 mg/ml in 100 mM NaCl, 5 mM DTT, 0.02% NaN_3, 10 mM Tris-HCl at pH 7.5 and stored at -80 ˚C prior to crystallization. The initial crystallization screening was carried out at the high-throughput screening (HTS) facility at Hauptman-Woodward Institute (HWI) located in Buffalo, NY, where 1536 crystallization conditions were screened using the microbatch method (57). Initial crystallization hits were further optimized manually to obtain diffraction quality crystals. The addition of detergents in this screen was key to improving the crystals' quality. Optimal conditions for crystallization were obtained at room temperature in 0.1 M NaH_2PO4, 0.1 M Na Acetate, pH 5.5 and 28% PEG 400. Diffraction of OR486 crystals was first tested using a home X-ray facility with a Rigaku RAXV ++ detector. The crystals were harvested directly from the drops and

flash-frozen in liquid nitrogen. Diffraction data set to 2.44 Å was collected at the National Synchrotron Light Source, with beamline X4C, and the data were processed with HKL-2000 (HKL Research, Inc.). The structure was determined by molecular replacement using Phaser (58), with a preliminary NMR model of OR485 as initial search model. The refinement was carried out using Phenix (59, 60), and model adjusting was done in Coot (61). The statistics for the final structure refinement and model geometry are summarized in Table S4.

dcs_C_1_ss, dcs_D_2, dcs_E_3, dcs_E_4, dcs_E_4_dim9 and dcs_E_4_dim9_cav3

To prepare protein samples for X-ray crystallography, the buffer of choice was 25 mM Tris, 300 mM NaCl, pH 8.0. Proteins were expressed from pET28b+ constructs to cleave the 6xHis tag with thrombin. Dialyzed proteins were incubated with thrombin (1:5000 dilution) overnight at room temperature and cleaved samples were loaded to a column of benzamidine resin pre-equilibrated in lysis buffer. Resin was resuspended and nutated for 30-60 minutes to remove thrombin from solution. Flow-through was collected and washed with 3-5 mL of lysis buffer. Protease inhibitor (phenylmethylsuphonyl fluoride, PMSF) was added to the eluted sample, which was then applied to a nickel affinity column pre-equilibrated in lysis buffer to remove the cleaved 6xHis tag from solution. Flow-through was collected and washed with 1-2 column volumes. Proteins were further purified by FPLC as described above and specific cleavage of the 6xHis tag was tested by mass spectrometry.

Purified proteins were concentrated to approximately 10-20 mg/ml for screening crystallization conditions. Commercially available crystallization screens were tested in 96-well sitting or hanging drops with different protein:precipitant ratios (1:1, 1:2 and 2:1) using a mosquito robot. When possible, initial crystal hits were grown in larger 24-well hanging drops. Obtained crystals were flash-frozen in liquid nitrogen. X-ray diffraction data sets were collected at the Lawrence Berkeley National Laboratory (LBNL). Crystal structures were solved by molecular replacement with Phaser (58) using the design models as the initial search models. The structures were built and refined using Phenix (59, 60) and Coot (61).
The crystallization conditions for the solved crystal structures are the following:

- **dcs_C_1_ss:**
    - Protein solution: 15 mg/ml, 25 mM Tris hydrochloride (pH 7) and 0.1 M sodium chloride

- Reservoir solution: 0.1 M Tris hydrochloride, pH 8.5 and 25% PEG 3,350
- 20% glycerol as a cryoprotection solution

- **dcs_D_2:**
  - Protein solution: 16 mg/ml, 25 mM Tris hydrochloride (pH 8) and 0.3 M sodium chloride
  - Reservoir solution: 0.1 M sodium MOPS/HEPES, pH 7.5, 12.5% PEG 1000, 12.5% PEG 3350 and 12.5% 2-methyl-2,4-pentanediol and 0.2 M of amino acids (sodium glutamate, DL-alanine, glycine, DL-lysine HCl and DL-serine).
  - No cryoprotection added
- **dcs_E_3:**
  - Protein solution: 11 mg/ml, 25 mM Tris hydrochloride (pH 8) and 0.1 M sodium chloride
  - Reservoir solution: 0.2 M ammonium citrate dibasic and 30% PEG 3350
  - No cryoprotection added
- **dcs_E_4:**
  - Protein solution: 27 mg/ml, 25 mM Tris hydrochloride (pH 8) and 0.3 M sodium chloride
  - Reservoir solution: 0.1 M bicine/Trizma base, pH 8.5, 10% PEG 20 000, 20% PEG MME 550 and 0.03 M of each ethylene glycol (diethyleneglycol, triethyleneglycol, tetraethyleneglycol and pentaethyleneglycol).
  - No cryoprotection added
- **dcs_E_4_dim9:**
  - Protein solution: 8 mg/ml, 25 mM Tris hydrochloride (pH 8) and 0.3 M sodium chloride
  - Reservoir solution: 0.1 M potassium thiocyanate, pH 8 and 30% PEG MME 2000
  - 32% PEG MME 2000 and 10% glycerol as a cryoprotection solution
- **dcs_E_4_dim9_cav3:**
  - Protein solution: 8 mg/ml, 30 mM Tris hydrochloride (pH 8) and 0.1 M sodium chloride
  - Reservoir solution: 0.1 M sodium MOPS/HEPES, pH 7.5, 10% PEG 20 000, 20% PEG MME 550 and 0.3 M of halides (sodium fluoride, sodium bromide and sodium iodide).

No cryoprotection added.

3.6 ACKNOWLEDGEMENTS

We thank L. Carter for assistance with SEC-MALS and protein production; J. Nguyen, A. Young-Seug, Z. Wang, M. Bick, S. Jayaraman and P. O'Connell for assistances in X-ray crystallography; all Baker lab members for discussions, and Rosetta@Home volunteers for computing resources used in *ab initio* structure prediction calculations. Work carried out at the Baker laboratory was supported by the Howard Hughes Medical Institute and the Defense Threat Reduction Agency (Funding HDTRA 1-11-1-0041). X-ray diffraction data was collected at the National Synchrotron Light Source with beam line X4C (Brookhaven National Laboratory, Upton, U.S. Department of Energy) and the Advance Light Source (Lawrence Berkeley National Laboratory, Berkeley, California Department of Energy). The Berkeley Center for Structural Biology is supported in part by the National Institutes of Health, National Institute of General Medical Sciences, and the Howard Hughes Medical Institute. The Advanced Light Source is supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Work done by Enrique Marcos was supported by a Marie Curie International Outgoing Fellowship (FP7-PEOPLE-2011-IOF 298976). Work done by Gustav Overdorfer was supported by a Marie Curie International Outgoing Fellowship (332094 ASR-CompEnzDes FP7-PEOPLE-2012-IOF). This work was supported as a Community Outreach Activity of NIGMS PSI grant U54 GM094597 (to G.T.M). We thank Darwin Alonso for technical support. This work was facilitated though the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system and funded by the STF at the University of Washington.

3.7 LITERATURE

1. Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L., and Baker, D. (2013) Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*. **501**, 212–6
2. Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S., and Baker, D. (2008) Kemp elimination catalysts by computational enzyme design. *Nature*. **453**, 190–195
3. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas III, C. F., Hilvert, D., Houk, K. N., Stoddard, B. L., and Baker, D. (2008) De novo computational design of retro- aldol enzymes. *Science (80-.).* **319**, 1387–1391
4. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St.Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E., and Baker, D. (2010)

Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science (80-.).* **329**, 309–313
5. Rajagopalan, S., Wang, C., Yu, K., Kuzin, A. P., Richter, F., Lew, S., Miklos, A. E., Matthews, M. L., Seetharaman, J., Su, M., Hunt, J. F., Cravatt, B. F., and Baker, D. (2014) Design of activated serine-containing catalytic triads with atomic-level accuracy. *Nat. Chem. Biol.* **10**, 386–91
6. Richter, F., Blomberg, R., Khare, S. D., Kiss, G., Kuzin, A. P., Smith, A. J. T., Gallaher, J., Pianowski, Z., Helgeson, R. C., Grjasnow, A., Xiao, R., Seetharaman, J., Su, M., Vorobiev, S., Lew, S., Forouhar, F., Kornhaber, G. J., Hunt, J. F., Montelione, G. T., Tong, L., Houk, K. N., Hilvert, D., and Baker, D. (2012) Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. *J. Am. Chem. Soc.* **134**, 16197–206
7. Giger, L., Caner, S., Obexer, R., Kast, P., Baker, D., Ban, N., and Hilvert, D. (2013) Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498
8. Joh, N. H., Wang, T., Bhate, M. P., Acharya, R., Wu, Y., Grabe, M., Hong, M., Grigoryan, G., and DeGrado, W. F. (2014) De novo design of a transmembrane Zn $^{2+}$ -transporting four-helix bundle. *Science (80-.).* **346**, 1520–1524
9. Thomson, A. R., Wood, C. W., Burton, A. J., Bartlett, G. J., Sessions, R. B., Brady, R. L., and Woolfson, D. N. (2014) Computational design of water-soluble α-helical barrels. *Science (80-.).* **346**, 485–488
10. Doyle, L., Hallinan, J., Bolduc, J., Parmeggiani, F., Baker, D., Stoddard, B. L., and Bradley, P. (2015) Rational design of α-helical tandem repeat proteins with closed architectures. *Nature*. **528**, 585–588
11. Burton, A. J., Thomson, A. R., Dawson, W. M., Brady, R. L., and Woolfson, D. N. (2016) Installing hydrolytic activity into a completely de novo protein framework. *Nat. Chem.* **8**, 837–844
12. Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., and Baker, D. (2012) Principles for designing ideal protein structures. *Nature*. **491**, 222–7
13. Huang, P.-S., Oberdorfer, G., Xu, C., Pei, X. Y., Nannenga, B. L., Rogers, J. M., DiMaio, F., Gonen, T., Luisi, B., and Baker, D. (2014) High thermodynamic stability of parametrically designed helical bundles. *Science (80-.).* **346**, 481–485
14. Lin, Y., Koga, N., Tatsumi-koga, R., Liu, G., Clouser, A. F., and Montelione, G. T. (2015) Control over overall shape and size in de novo designed proteins. 10.1073/pnas.1509508112
15. Brunette, T., Parmeggiani, F., Huang, P.-S., Bhabha, G., Ekiert, D. C., Tsutakawa, S. E., Hura, G. L., Tainer, J. A., and Baker, D. (2015) Exploring the repeat protein universe through computational protein design. *Nature*. **528**, 580–584
16. Huang, P.-S., Feldmeier, K., Parmeggiani, F., Fernandez Velasco, D. A., Höcker, B., and Baker, D. (2016) De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34
17. Jacobs, T. M., Williams, B., Williams, T., Xu, X., Eletsky, A., Federizon, J. F., Szyperski, T., and Kuhlman, B. (2016) Design of structurally distinct proteins using strategies inspired by evolution. *Science (80-.).* **352**, 687–690
18. Salemme, F. R. (1981) Conformational and geometrical properties of β-sheets in proteins. *J. Mol. Biol.* **146**, 143–156
19. Chothia, C. (1973) Conformation of twisted β-pleated sheets in proteins. *J. Mol. Biol.* **75**, 295–302
20. Richardsont, J. S., Getzofft, E. D., and Richardsont, D. C. (1978) *The beta bulge: A common small unit of nonrepetitive protein structure*, [online] https://www.pnas.org/content/pnas/75/6/2574.full.pdf (Accessed January 6, 2019)
21. Fujiwara, K., Ebisawa, S., Watanabe, Y., Fujiwara, H., and Ikeguchi, M. (2015) The origin of β-strand bending in globular proteins. *BMC Struct. Biol.* **15**, 1–12
22. Chan, A. W. E., Hutchinson, E. G., Harris, D., and Thornton, J. M. (1993) Identification, classification, and analysis of beta-bulges in proteins. *Protein Sci.* **2**, 1574–1590
23. Salemme, F. R. (1983) Structural properties of protein beta-sheets. *Prog. Biophys. Mol. Biol.* **42**, 95–133
24. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. a, Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. A., Fleishman, S. J., Corn, E., Kim, D. E., Lyskov, S., Berrondo, M., Havranek, J. J., Mentzer, S., Popovic, Z., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. (2011) ROSETTA 3 : An Object-Oriented Software Suite

for the Simulation and Design of Macromolecules. *Methods Enzymol. Vol. 487*. **487**, 545–574
25. Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., and Baker, D. (2012) Principles for designing ideal protein structures. *Nature*. **491**, 222–7
26. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science*. **302**, 1364–8
27. Craveur, P., Joseph, A. P., Rebehmed, J., and De Brevern, A. G. (2013) β-Bulges: Extensive structural analyses of β-sheets irregularities. *Protein Sci.* **22**, 1366–1378
28. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., and Baker, D. (2004) Protein Structure Prediction Using Rosetta. 10.1016/S0076-6879(04)83004-0
29. Bradley, P., Misura, K. M. S., and Baker, D. (2005) Toward high-resolution de novo structure prediction for small proteins. *Science (80-.).* **309**, 1868–1871
30. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402
31. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*. **10**, 421
32. Zhang, Y., and Skolnick, J. (2005) TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309
33. Richardson, J. S., and Richardson, D. C. (2002) Natural NL -sheet proteins use negative design to avoid edge-to-edge aggregation
34. Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., and Baker, D. (2012) Principles for designing ideal protein structures. *Nature*. **491**, 222–7
35. Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E. M., Khare, S. D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., Meiler, J., and Baker, D. (2011) Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling suite. *PLoS One*. 10.1371/journal.pone.0020161
36. Wintjens, R. T., Rooman, M. J., and Wodak, S. J. (1996) Automatic classification and analysis of alpha alpha-turn motifs in proteins. *J. Mol. Biol.* **255**, 235–53
37. Kuhlman, B., and Baker, D. (2000) Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci.* **97**, 10383–10388
38. Leaver-Fay, A., O'Meara, M. J., Tyka, M., Jacak, R., Song, Y., Kellogg, E. H., Thompson, J., Davis, I. W., Pache, R. A., Lyskov, S., Gray, J. J., Kortemme, T., Richardson, J. S., Havranek, J. J., Snoeyink, J., Baker, D., and Kuhlman, B. (2013) Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol.* **523**, 109–143
39. O'Meara, M. J., Leaver-Fay, A., Tyka, M. D., Stein, A., Houlihan, K., Dimaio, F., Bradley, P., Kortemme, T., Baker, D., Snoeyink, J., and Kuhlman, B. (2015) Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *J. Chem. Theory Comput.* **11**, 609–622
40. Sheffler, W., and Baker, D. (2010) RosettaHoles2: A volumetric packing measure for protein structure refinement and validation. *Protein Sci.* **19**, 1991–1995
41. Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202
42. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., and Baker, D. (2004) Protein Structure Prediction Using Rosetta. *Methods Enzymol.* **383**, 66–93
43. J. A. Fallas, G. Ueda, W. Sheffler, V. Nguyen, D. E. McNamara, B. Sankaran, J. H. Pereira, F. Parmeggiani, TJ Brunette, D. Cascio, T. R. Yeates, P. Zwart, D. B. (2016) Computational design of self-assembling cyclic protein homooligomers. *Nat. Chem.*
44. DeLano, W. L. (2002) The PyMOL Molecular Graphics System, Version 1.1. *Schrödinger LLC*. 10.1038/hr.2014.17
45. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–12
46. Acton, T. B., Xiao, R., Anderson, S., Aramini, J., Buchwald, W. A., Ciccosanti, C., Conover, K., Everett, J., Hamilton, K., Huang, Y. J., Janjua, H., Kornhaber, G., Lau, J., Lee, D. Y., Liu, G., Maglaqui, M., Ma, L., Mao, L., Patel, D., Rossi, P., Sahdev, S., Shastry, R., Swapna, G. V. T., Tang, Y., Tong, S., Wang, D., Wang, H., Zhao, L., and Montelione, G. T. (2011) Preparation of

protein samples for NMR structure, function, and small-molecule screening studies. *Methods Enzymol.* **493**, 21–60
47. Xiao, R., Anderson, S., Aramini, J., Belote, R., Buchwald, W. A., Ciccosanti, C., Conover, K., Everett, J. K., Hamilton, K., Huang, Y. J., Janjua, H., Jiang, M., Kornhaber, G. J., Lee, D. Y., Locke, J. Y., Ma, L. C., Maglaqui, M., Mao, L., Mitra, S., Patel, D., Rossi, P., Sahdev, S., Sharma, S., Shastry, R., Swapna, G. V. T., Tong, S. N., Wang, D., Wang, H., Zhao, L., Montelione, G. T., and Acton, T. B. (2010) The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. *J. Struct. Biol.* **172**, 21–33
48. Jansson, M., Li, Y.-C., Jendeberg, L., Anderson, S., Montelione, G., and Nilsson, B. (1996) High-level production of uniformly 15N-and 13C-enriched fusion proteins in Escherichia coli. *J. Biomol. NMR*. 10.1007/BF00203823
49. Santoro, M. M., and Bolen, D. W. (1992) A test of the linear extrapolation of unfolding free energy changes over an extended denaturant concentration range. *Biochemistry*. **31**, 4901–4907
50. Scholtz, J. M., Grimsley, G. R., and Pace, C. N. (2009) Solvent denaturation of proteins and interpretations of the m value. in *Methods in enzymology*, pp. 549–565, **466**, 549–565
51. Geierhaas, C. D., Nickson, A. a, Lindorff-Larsen, K., Clarke, J., and Vendruscolo, M. (2006) BPPred: A Web-based computational tool for predicting biophysical parameters of proteins. *Protein Sci.* **16**, 125–134
52. Rossi, P., Swapna, G. V. T., Huang, Y. J., Aramini, J. M., Anklin, C., Conover, K., Hamilton, K., Xiao, R., Acton, T. B., Ertekin, A., Everett, J. K., and Montelione, G. T. (2010) A microscale protein NMR sample screening pipeline. *J. Biomol. NMR*. **46**, 11–22
53. Neri, D., Szyperski, T., Otting, G., Senn, H., and Wüthrich, K. (1989) Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional 13C labeling. *Biochemistry*. **28**, 7510–6
54. Liu, G. H., Shen, Y., Atreya, H. S., Parish, D., Shao, Y., Sukumaran, D. K., Xiao, R., Yee, A., Lemak, A., Bhattacharya, A., Acton, T. A., Arrowsmith, C. H., Montelione, G. T., and Szyperski, T. (2005) NMR data collection and analysis protocol for high-throughput protein structure determination. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 10487–10492
55. Bhattacharya, A., Tejero, R., and Montelione, G. T. (2007) Evaluating protein structures determined by structural genomics consortia. *Proteins Struct. Funct. Genet.* **66**, 778–795
56. Huang, Y. J., Powers, R., and Montelione, G. T. (2005) Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. *J. Am. Chem. Soc.* **127**, 1665–74
57. Luft, J. R., Collins, R. J., Fehrman, N. A., Lauricella, A. M., Veatch, C. K., and DeTitta, G. T. (2003) A deliberate approach to screening for initial crystallization conditions of biological macromolecules. *J. Struct. Biol.* **142**, 170–179
58. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674
59. Zwart, P. H., Afonine, P. V, Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., McKee, E., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K., Storoni, L. C., Terwilliger, T. C., and Adams, P. D. (2008) Automated structure solution with the PHENIX suite. *Methods Mol. Biol.* **426**, 419–35
60. Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 213–221
61. Emsley, P., and Cowtan, K. (2004) Coot: Model-building tools for molecular graphics. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60**, 2126–2132
62. Roy, A., Yang, J., and Zhang, Y. (2012) COFACTOR: An accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.* 10.1093/nar/gks372
63. Voss, N. R., and Gerstein, M. (2010) 3V: Cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Res.* 10.1093/nar/gkq395
64. Snyder, D. A., Chen, Y., Denissova, N. G., Acton, T., Aramini, J. M., Ciano, M., Karlin, R., Liu, J., Manor, P., Rajan, P. A., Rossi, P., Swapna, G. V. T., Xiao, R., Rost, B., Hunt, J., and Montelione,

G. T. (2005) Comparisons of NMR spectral quality and success in crystallization demonstrate that NMR and X-ray crystallography are complementary methods for small protein structure determination. *J. Am. Chem. Soc.* **127**, 16505–16511

65. Tejero, R., Snyder, D., Mao, B., Aramini, J. M., and Montelione, G. T. (2013) PDBStat: a universal restraint converter and restraint analysis software package for protein NMR. *J. Biomol. NMR*. **56**, 337–51

66. Hyberts, S. G., Goldberg, M. S., Havel, T. F., and Wagner, G. (1992) The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with X-ray structures. *Protein Sci.* **1**, 736–51

3.8 Supplementary Material

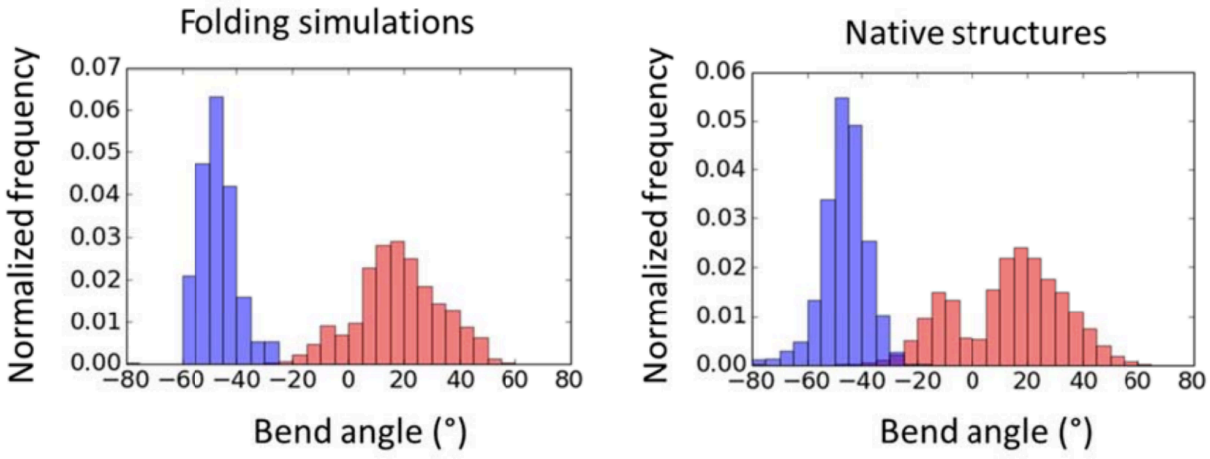


Figure S.3.1. Comparison of bend angle distributions from Rosetta folding simulations and native protein structures. Distributions for strand pairs formed by uniform and bulged strands are shown in red and blue colors respectively. Simulation distributions were obtained from two-stranded antiparallel beta-sheets built by fragment assembly. Right panel shows the same distribution as in figure. 3.1 B for comparison. Both folding simulations and native structural analysis show that uniform and bulged strand pairs favor positive and negative bend angles respectively.

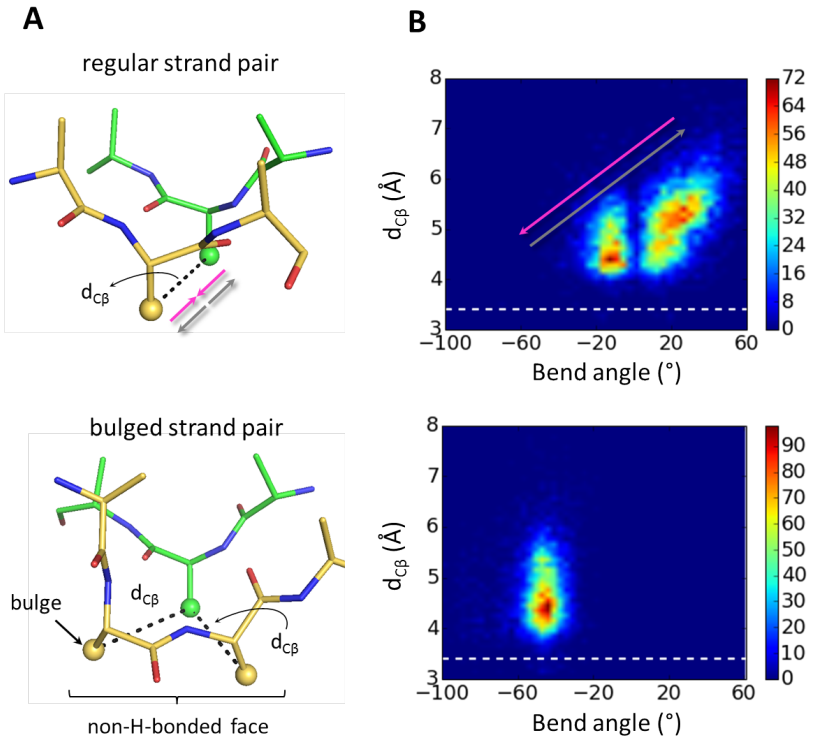


Fig. S.3.2. Effect of bulges on local strand bending A. Local geometry of regular and bulged strand pairs indicating the C_β-C_β inter-atomic distance ($d_{C\text{-}beta}$) between paired residues (dashed lines). Gray and pink arrows show $d_{C\text{-}beta}$ changes correlated with positive and negative bend angles, as shown in panel B. For the bulged pair, due to the offset in side-chain directionality, the $d_{C\text{-}beta}$ is also considered for the C_{beta} of the residue following the bulge. The different hydrogen bond pairing of bulges prevents strand pairing in one face of the strand as indicated. **B**. Distribution of bend angle sign and $d_{C\beta}$ for the two strand pair types. The white dashed line at 3.4 Å shows the steric clash limit between two carbon atoms (sum of Van der Waals radii). For regular strand pairs, the increase of bend angle tends to increase $d_{C\text{-}beta}$. Bulged strand pairs achieve more negative bend angles than regular strand pairs without decreasing $d_{C\text{-}beta}$ further. While the local geometry of bulges minimizes steric effects favoring negative bend angles, it disallows positive bend angles by preventing hydrogen bond pairing in one face of the strand. The low frequency of perfectly flat regular strands (see gap close to 0°) is due to partial contribution of intra-strand twist to the bend angle calculation.

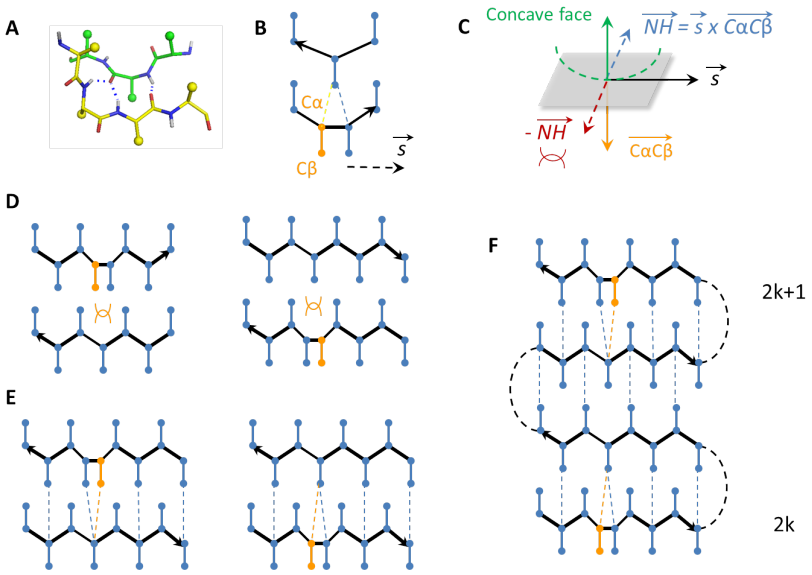


Figure S.3.3. C_{alpha}-C_{beta} vector patterns in curved sheets A. Local geometry of a bulged strand pair and **B**, its diagram representation. Bulges are highlighted in orange, regular strand residues are shown in blue and the vector *s* indicates the bulged strand direction. **C**. Description of the hydrogen bonding orientation of the bulge with respect to the concave face of the bulge local bend. Blue and red arrows indicate the directions where hydrogen bond is allowed and disallowed respectively. **D**. Diagram representation of incompatible strand pairings in the presence of a bulge. **E**. Diagram representation of compatible strand pairings in the presence of a bulge. Antiparallel hydrogen bonding between paired residues is drawn with dashed lines. **F**. Diagram of the strand pairing arrangement of a 4-stranded antiparallel beta-sheet compatible with two bulges at the edge strands. Bulges must be located at even, *2k*, and odd positions, *2k+1*, from the following and previous hairpin connections, respectively.

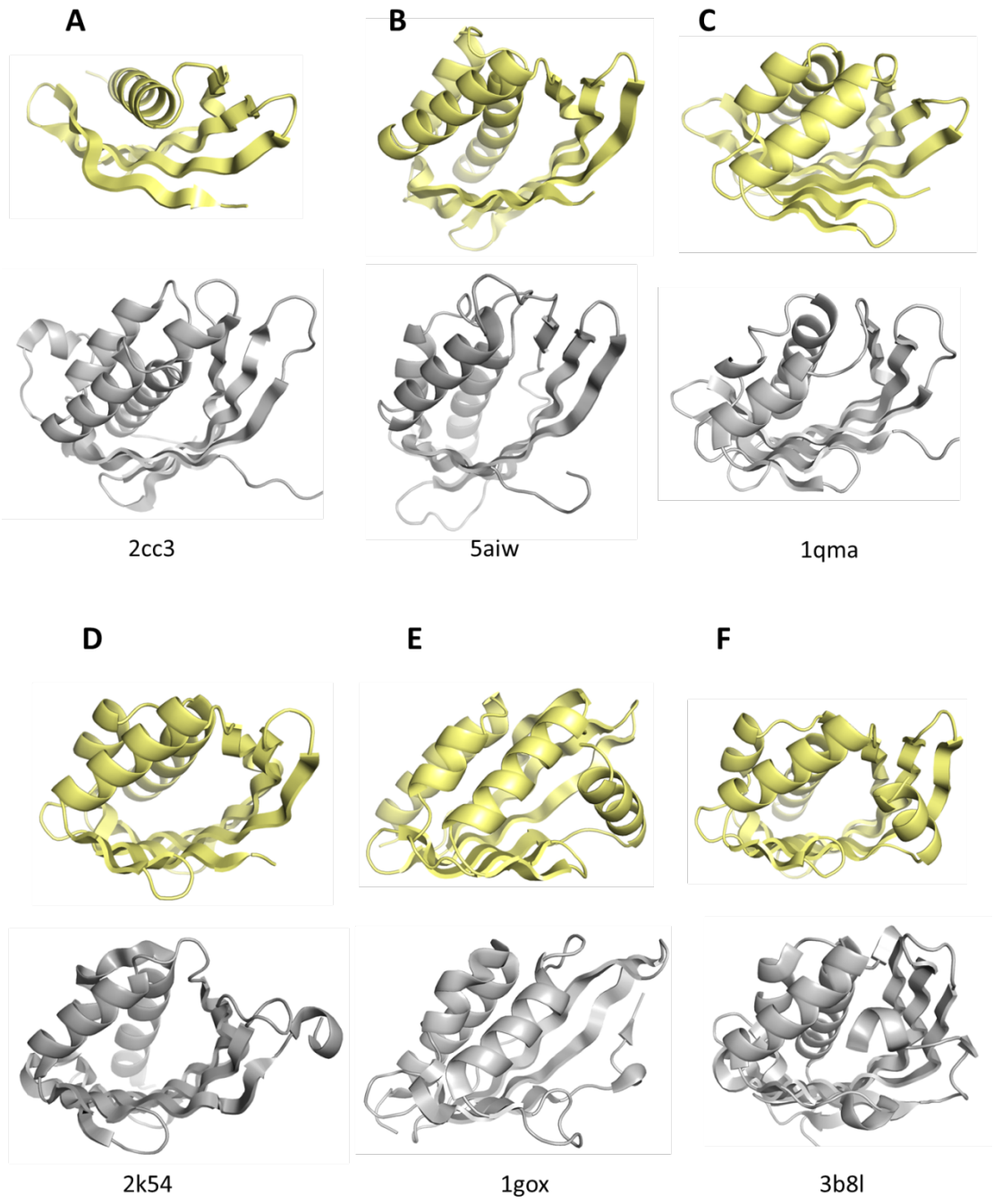


Fig. S.3.4. Best NTF2-like domain matches to *de novo* designed proteins. **A** designed structure representative of each fold (**A** to **F**) is compared with the closest structural analog, as determined by a TM-align search (32, 62). **A**. TM-score 0.80, sequence id. 6.8%. **B**. TM-score 0.78, sequence id. 9.3%. **C**. TM-score 0.82, sequence id. 6.6%. **D**. TM-score 0.86, sequence id. 19.2%. **E**. TM-score 0.74, sequence id. 14.4%. **F**. TM-score 0.79, sequence id. 6.9%. The top structural hits belong to the cystatin and NTF2-like superfamilies.

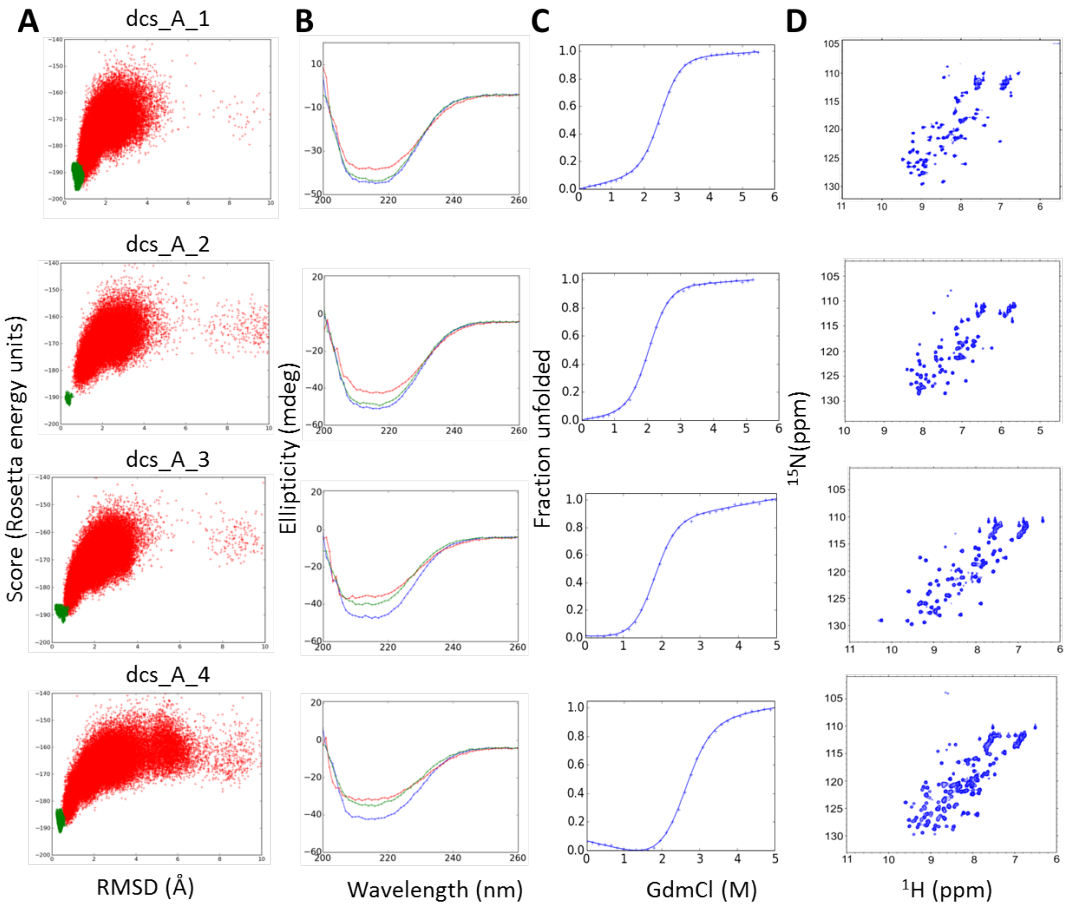


Fig. S.3.5. A. **Folding funnels and biophysical characterization of fold A designs**. Folding energy landscapes generated by *ab initio* structure prediction calculations. Each dot represents the lowest energy structure identified in an independent trajectory starting from an extended chain (red dots) or from the design model (green dots); *x*-axis shows the Cα-root mean squared deviation (RMSD) from the designed model; the *y*-axis shows the Rosetta all-atom energy. **B**. Far-ultraviolet circular dichroism spectra (blue: 25 °C, red: 95 °C, green: 25 °C after cooling). **C**. Chemical denaturation with GdmCl monitored with circular dichroism at 220 nm and 25 °C. **D**. ^{1}H-^{15}N HSQC spectra obtained at 25 °C.

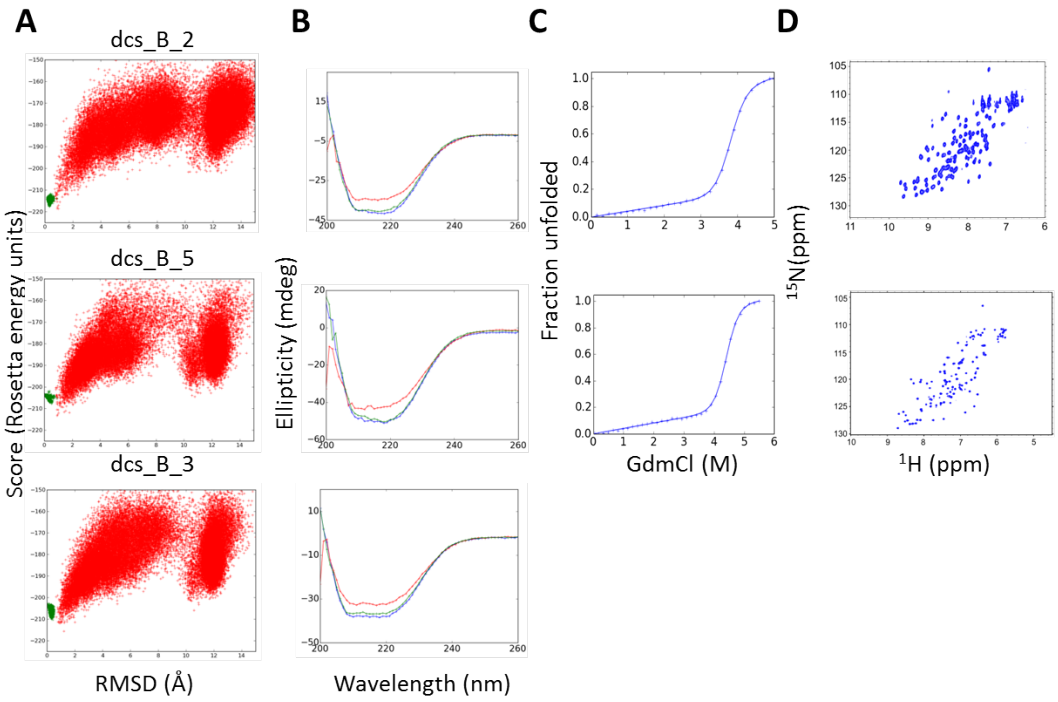


Fig. S.3.6. Figure S.3.5 Folding funnels and biophysical characterization of fold B designs. **A**. Folding energy landscapes generated by *ab initio* structure prediction calculations. Each dot represents the lowest energy structure identified in an independent trajectory starting from an extended chain (red dots) or from the design model (green dots); *x*-axis shows the Cα-root mean squared deviation (RMSD) from the designed model; the *y*-axis shows the Rosetta all-atom energy. **B**. Far-ultraviolet circular dichroism spectra (blue: 25 °C, red: 95 °C, green: 25 °C after cooling). **C**. Chemical denaturation with GdmCl monitored with circular dichroism at 220 nm and 25 °C. **D**. ^{1}H-^{15}N HSQC spectra obtained at 25 °C.

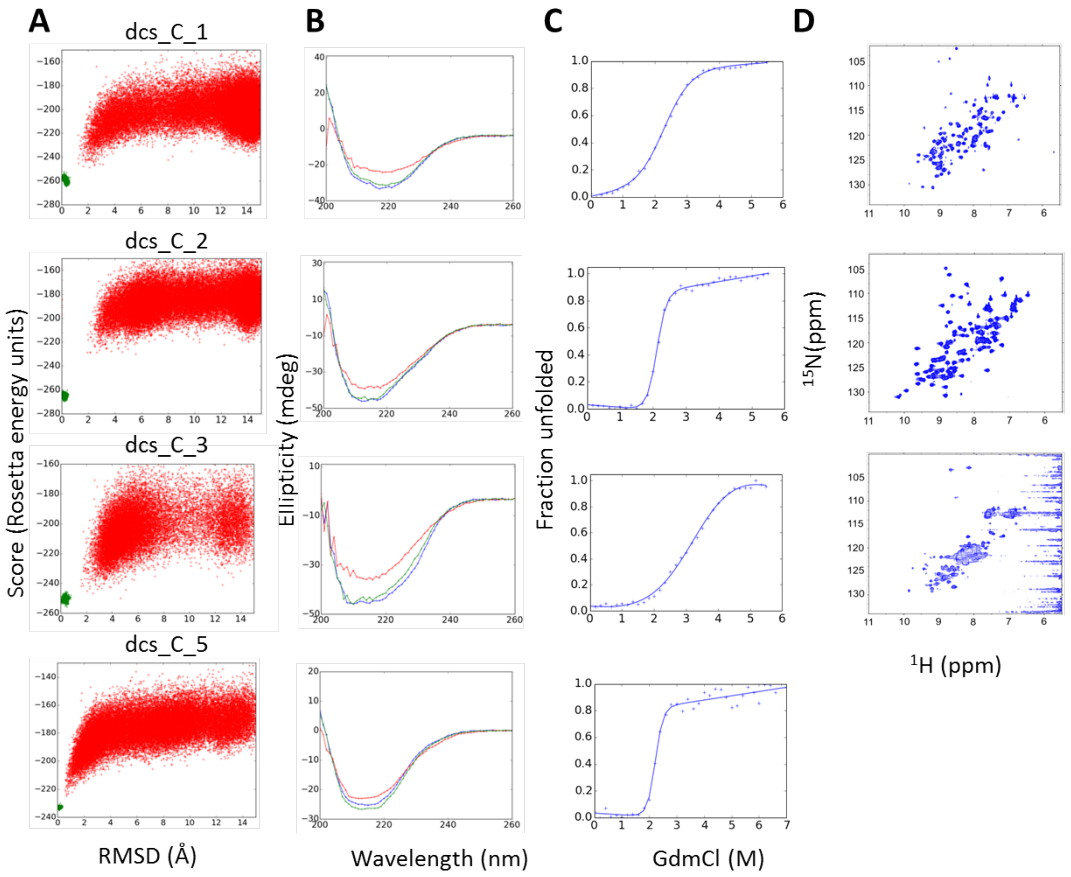


Figure S.3.7: Folding funnels and biophysical characterization of fold C designs. **A**. Folding energy landscapes generated by *ab initio* structure prediction calculations. Each dot represents the lowest energy structure identified in an independent trajectory starting from an extended chain (red dots) or from the design model (green dots); *x*-axis shows the Cα-root mean squared deviation (RMSD) from the designed model; the *y*-axis shows the Rosetta all-atom energy. **B** Far-ultraviolet circular dichroism spectra (blue: 25 °C, red: 95 °C, green: 25 °C after cooling). **C** Chemical denaturation with GdmCl monitored with circular dichroism at 220 nm and 25 °C. **D** ^{1}H-^{15}N HSQC spectra obtained at 25 °C.

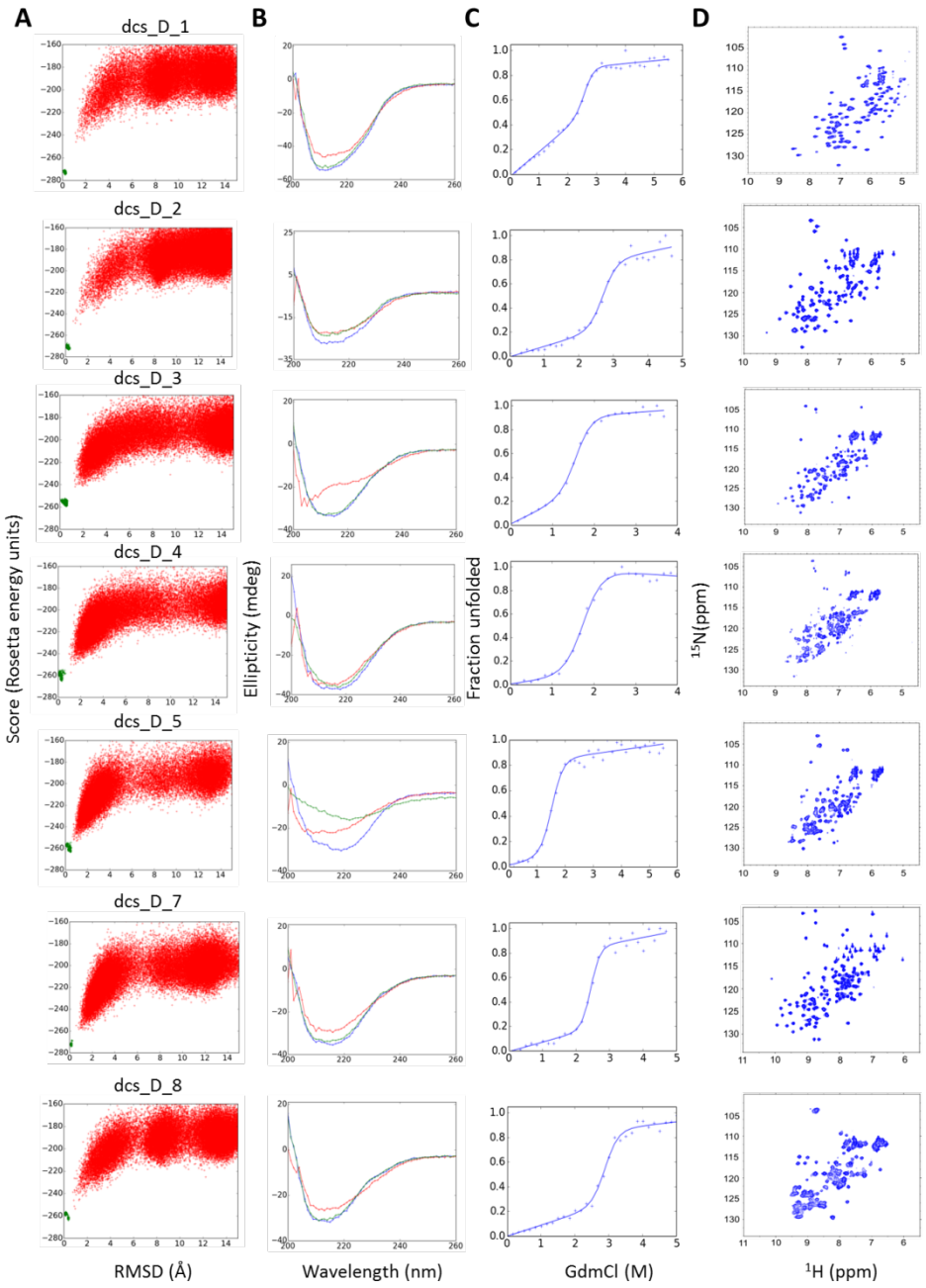


Figure S.3.8. Folding funnels and biophysical characterization of fold D designs A. Folding energy landscapes generated by *ab initio* structure prediction calculations. Each dot represents the lowest energy structure identified in an independent trajectory starting from an extended chain (red dots) or from the design model (green dots); *x*-axis shows the Cα-root mean squared deviation (RMSD) from the designed model; the *y*-axis shows the Rosetta all-atom energy. **B**. Far-ultraviolet circular dichroism spectra (blue: 25 °C, red: 95 °C, green: 25 °C after cooling). These proteins are more sensitive to temperature than others from different folds due to the high solvent accessibility of the pocket. **C**. Chemical denaturation with GdmCl monitored with circular dichroism at 220 nm and 25 °C. (D) ^{1}H-^{15}N HSQC spectra obtained at 25 °C.

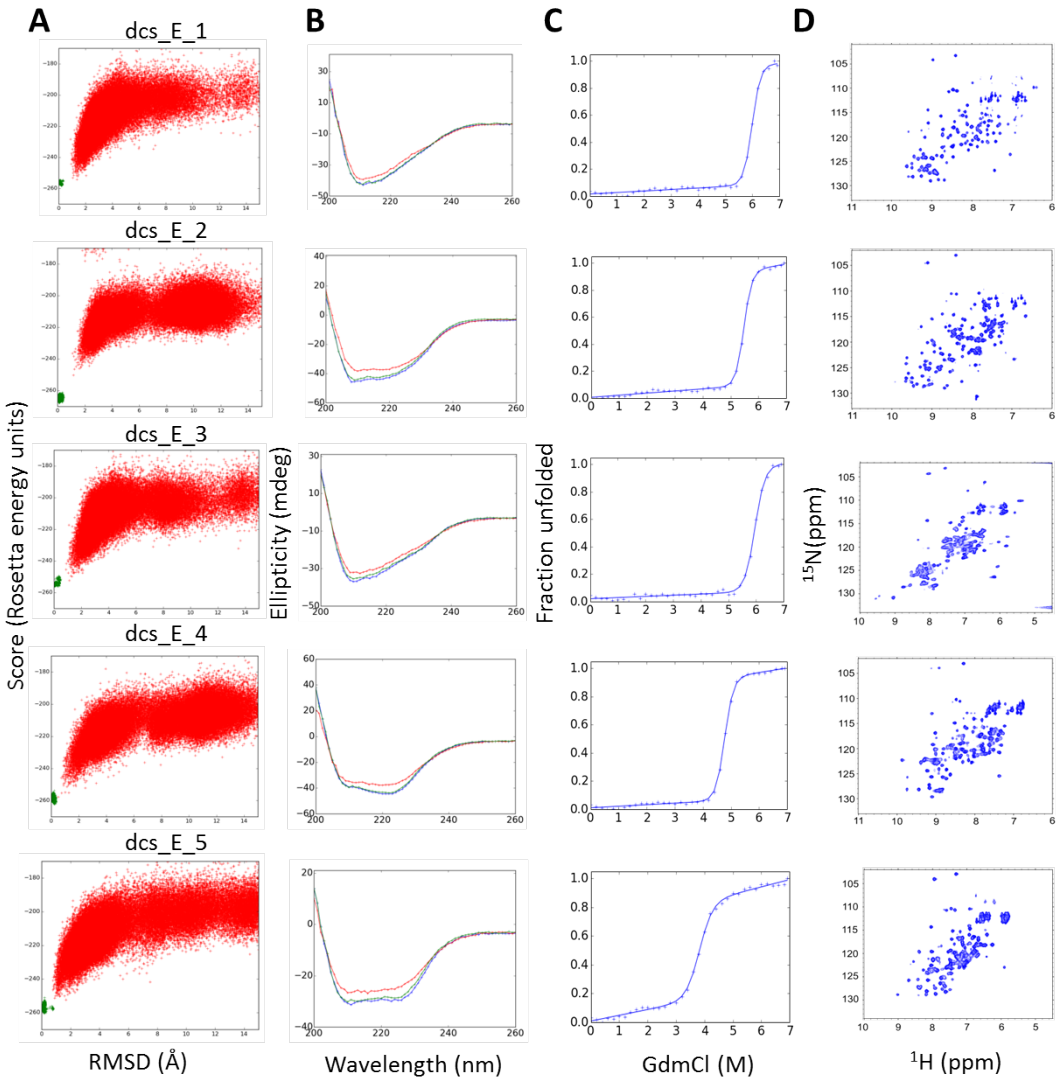


Figure S.3.9: Folding funnels and biophysical characterization of fold E designs **A**. Folding energy landscapes generated by *ab initio* structure prediction calculations. Each dot represents the lowest energy structure identified in an independent trajectory starting from an extended chain (red dots) or from the design model (green dots); *x*-axis shows the Cα-root mean squared deviation (RMSD) from the designed model; the *y*-axis shows the Rosetta all-atom energy. **B**. Far-ultraviolet circular dichroism spectra (blue: 25 °C, red: 95 °C, green: 25 °C after cooling). **C**. Chemical denaturation with GdmCl monitored with circular dichroism at 220 nm and 25 °C. **D**. ^{1}H-^{15}N HSQC spectra obtained at 25 °C.

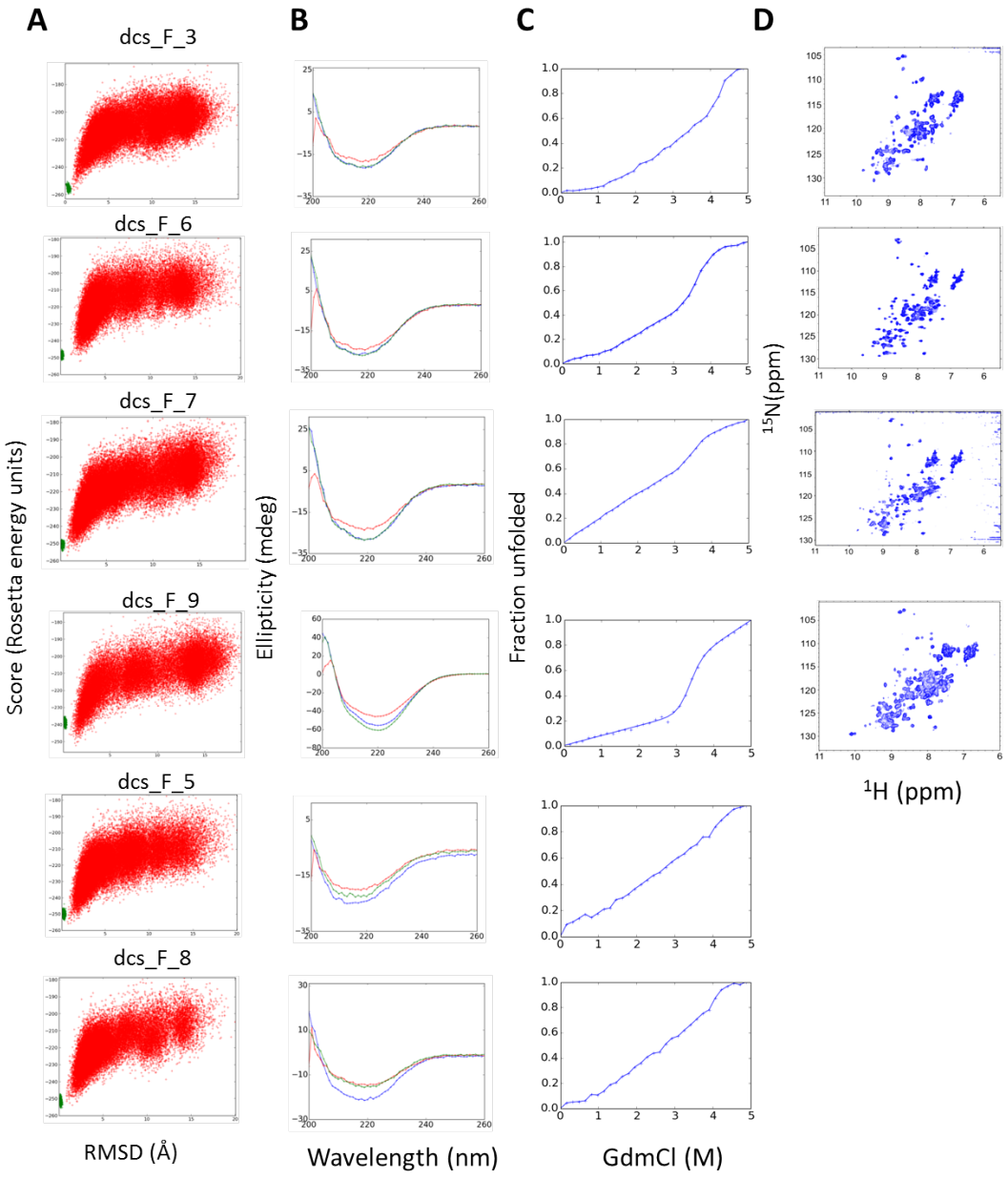


Figure S.3.10: Folding funnels and biophysical characterization of fold F designs. A. Folding energy landscapes generated by *ab initio* structure prediction calculations. Each dot represents the lowest energy structure identified in an independent trajectory starting from an extended chain (red dots) or from the design model (green dots); *x*-axis shows the Cα-root mean squared deviation (RMSD) from the designed model; the *y*-axis shows the Rosetta all-atom energy. **B**. Far-ultraviolet circular dichroism spectra (blue: 25 °C, red: 95 °C, green: 25 °C after cooling). **C**. Chemical denaturation with GdmCl monitored with circular dichroism at 220 nm and 25 °C. The non-sigmoidal transitions suggest molten globule character for these proteins. **D**. ^{1}H-^{15}N HSQC spectra obtained at 25 °C.

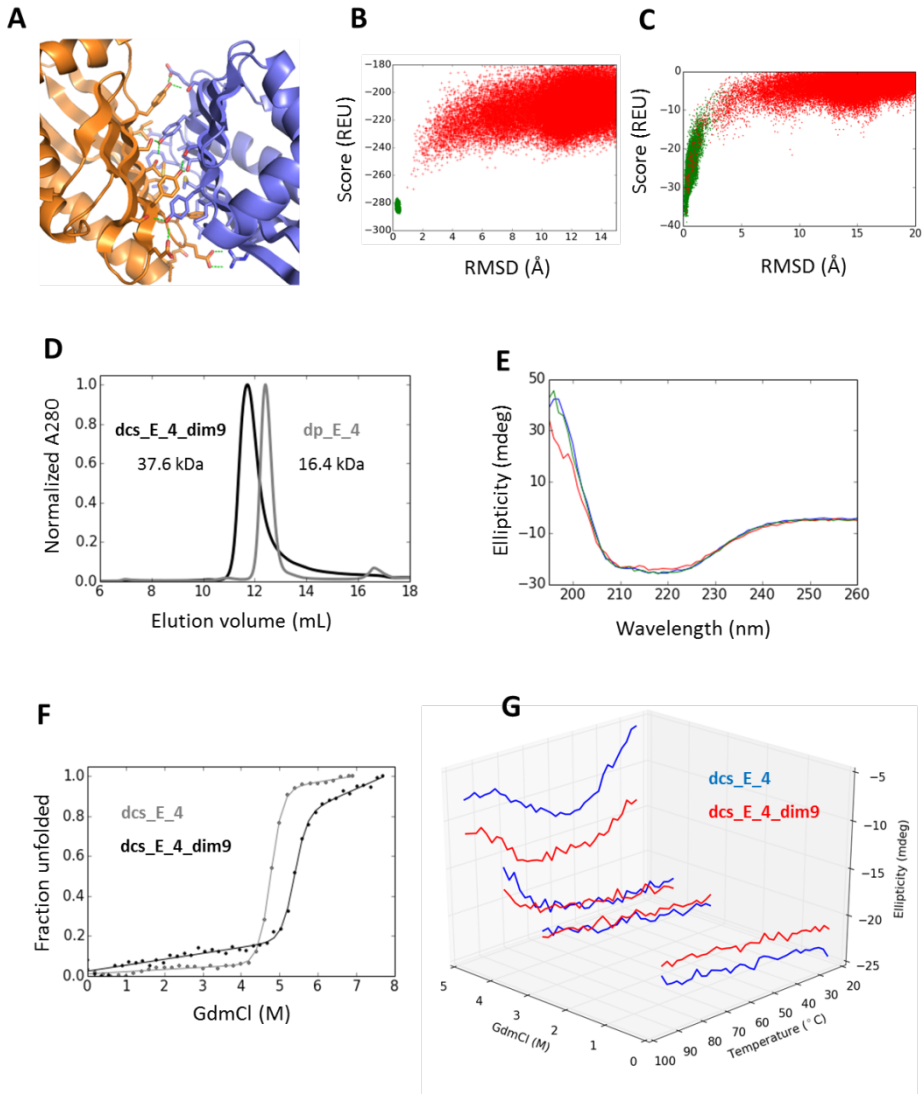


Figure S.3.11: Design of homodimerc de novo NTF2-like proteins A. Designed interface interactions involving hydrogen bonding and aromatic stacking. Hydrogen bonds are highlighted in green dashed lines. **B**. Folding energy landscape of the monomer subunit simulated with *ab initio* structure prediction. **C**. Asymmetric docking simulations of dcs_E_4_dim9 predict stable formation of the designed homo-dimer interface. **D**. Size-exclusion chromatograms monitoring UV absorbance at 280 nm. The shift in elution volume is consistent with dimer formation as assessed by multiple angle light scattering. **E.** CD wavelength scans for dcs_E_4_dim9 at 25°C (blue), 95°C (red) and 25°C after cooling (green). **F**. Chemical denaturation with GdmCl monitored with circular dichroism at 220 nm and 25°C. Continuous lines represent data fits to a two-state unfolding model. The higher C_m and lower folding free energy for dcs_E_4_dim9 indicate that the dimer interface provides additional stability (ΔΔG estimated in -1.4 kcal·mol-1). **G**. Chemical and thermal denaturation experiment on the monomer and dimer. At 4M GdmCl and ~90°C the monomer unfolds, whereas the dimer remains folded.

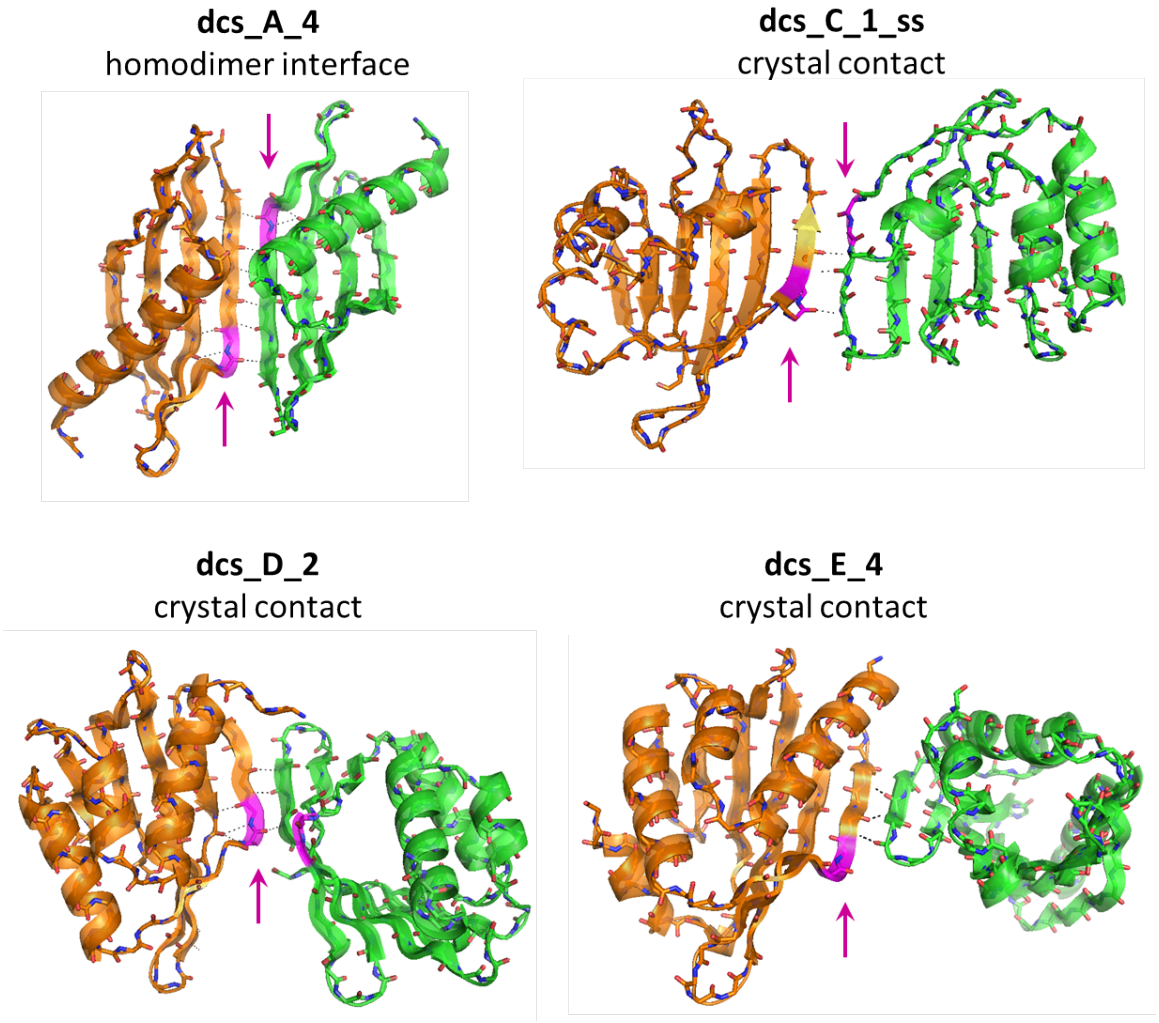


Figure S.3.12. Crystal contacts in de novo designed NTF2-like structures. The non-hydrogen bonding face of bulges (in magenta) restricts the hydrogen-bonded pairing between edge strands to regular segments, as observed in the homo-dimer interface of dcs_A_4 and in crystal contacts of dcs_C_1_ss, dcs_D_2 and dcs_E_4.

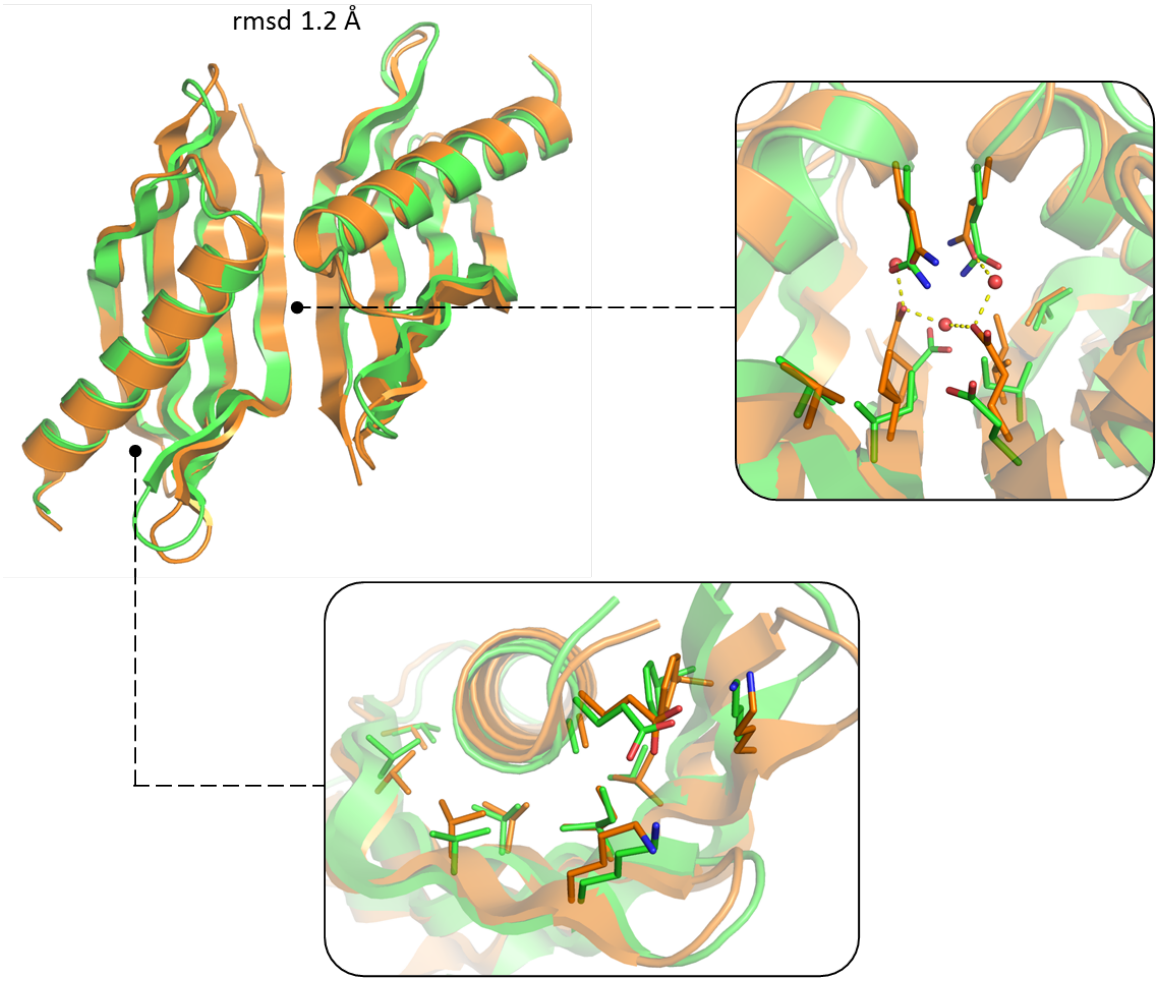


Fig. S.3.13. Crystal contacts in Fold A crystal structure. The monomeric design model is superimposed to each chain of the crystallized dimer for comparison. The experimental structure (2.4 Å resolution) and the design model are colored in orange and green, respectively; insets show comparisons of side-chain rotamers (right, homo-dimer interface; bottom, packing between beta-sheet long arm and helix). Water mediated hydrogen bonds formed at the interface are shown in yellow dashed lines. The RMSD is calculated over all Cα atoms of each chain: RMSD (chain A) 1.22 Å and RMSD (chain B) 1.21 Å.

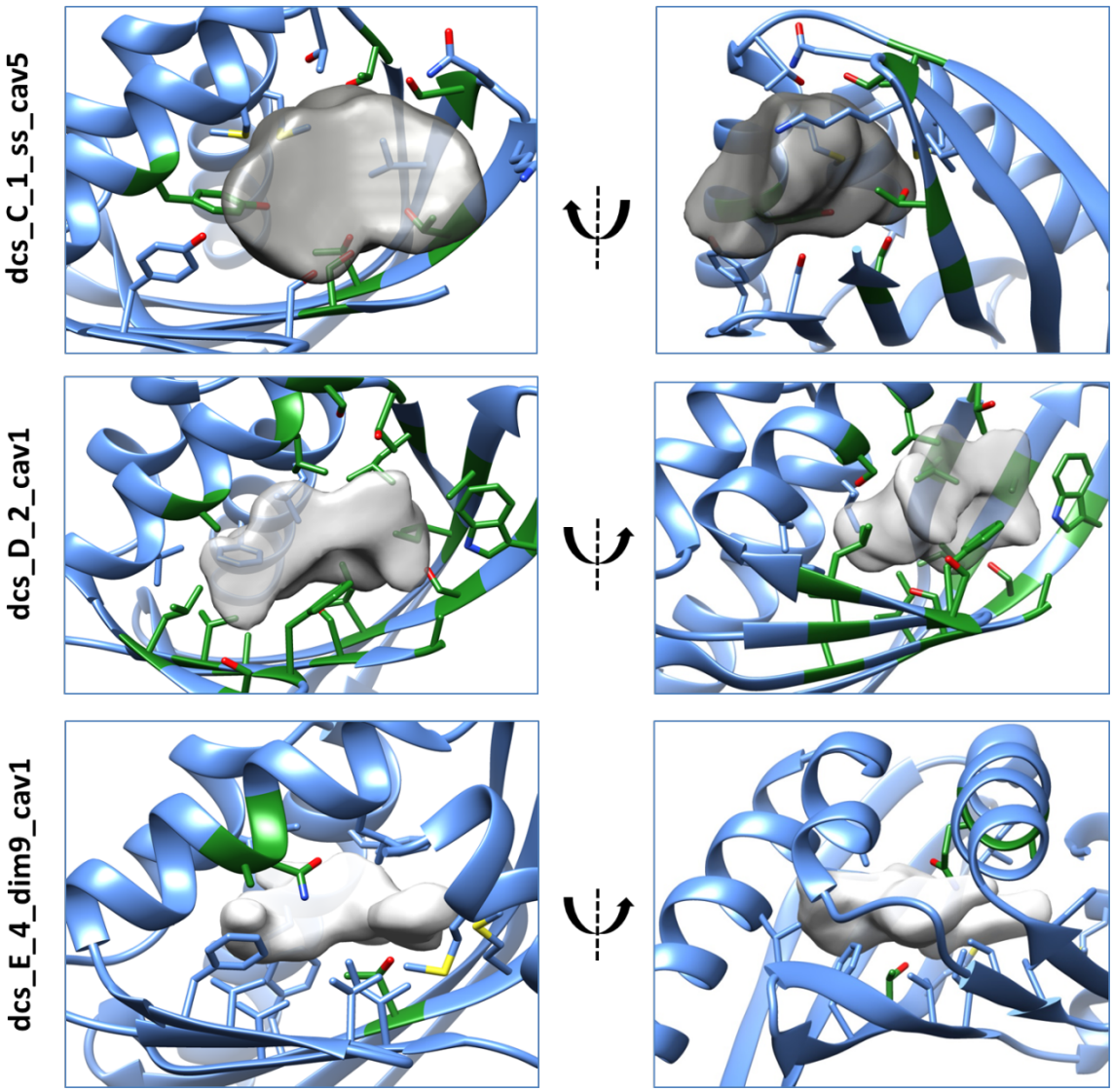


Fig. S.3.14. Cavity-forming mutant models. Side-chains of residues lining the cavities are shown and the incorporated mutations are colored in green. Cavities were calculated with the 3V webserver (63) using different probe radii depending on the degree of cavity burial (outer probe radii from 4 to 6 Å and inner probe radii from 1 to 2 Å) and a grid size of 0.5 Å.

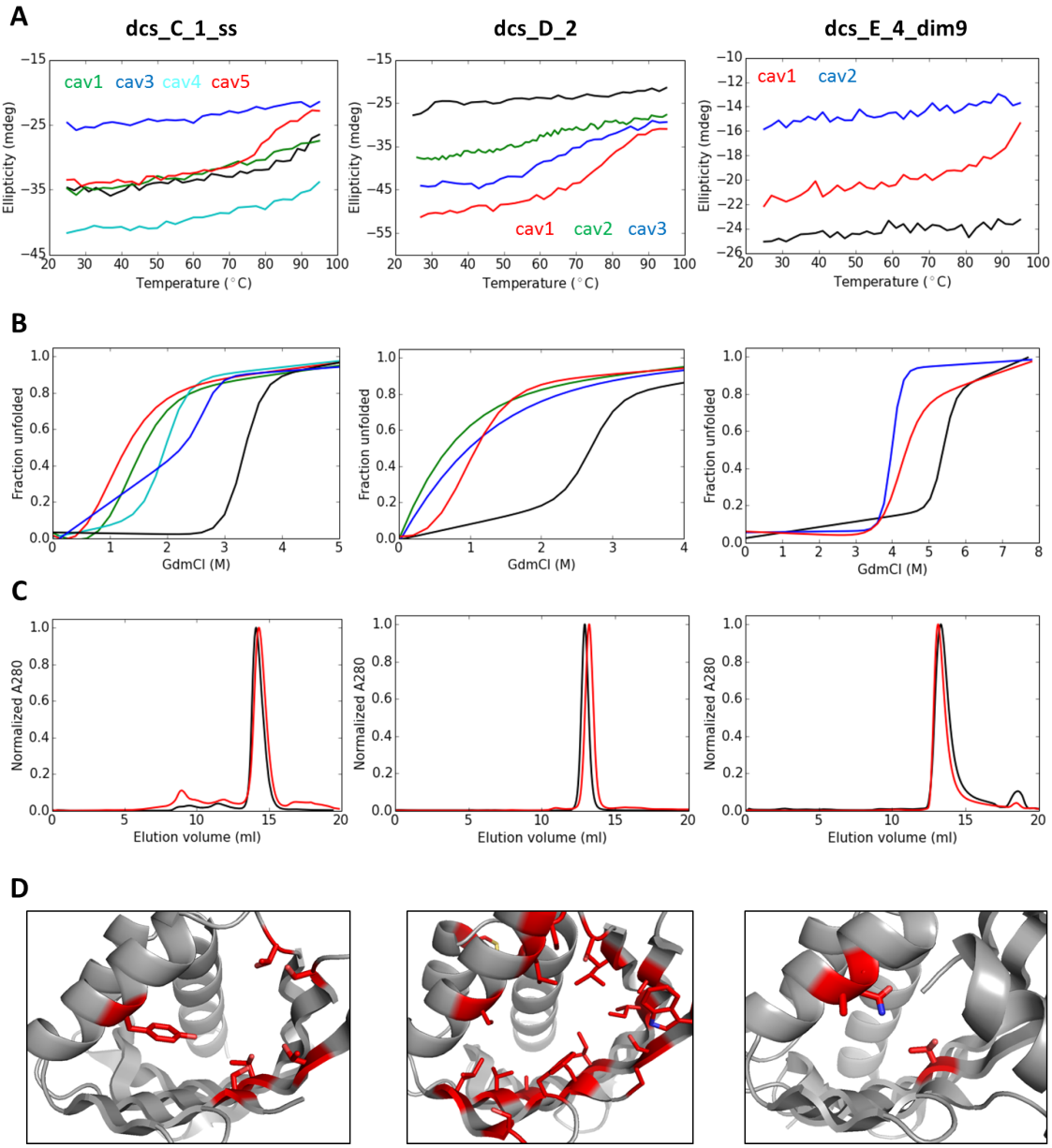


Fig. S.3.15. Experimental characterization of cavity-forming mutants. Characterization of cavity mutations in dcs_C_1_ss, dcs_D_2 and dcs_E_4_dim9. **A**. Temperature melting monitored with circular dichroism for the best expressed cavity-mutants and the corresponding parent designs (in black). Same colors identifying each mutant are used in the other panels. **B.** Chemical denaturation with GdmCl monitored with circular dichroism at 220 nm and 25°C. Continuous lines represent data fits to a two-state folding model. **C.** Size-exclusion chromatograms of a cavity mutant and its parent design (in black). **D.** Design model representations coloring the incorporated mutations of each mutant.

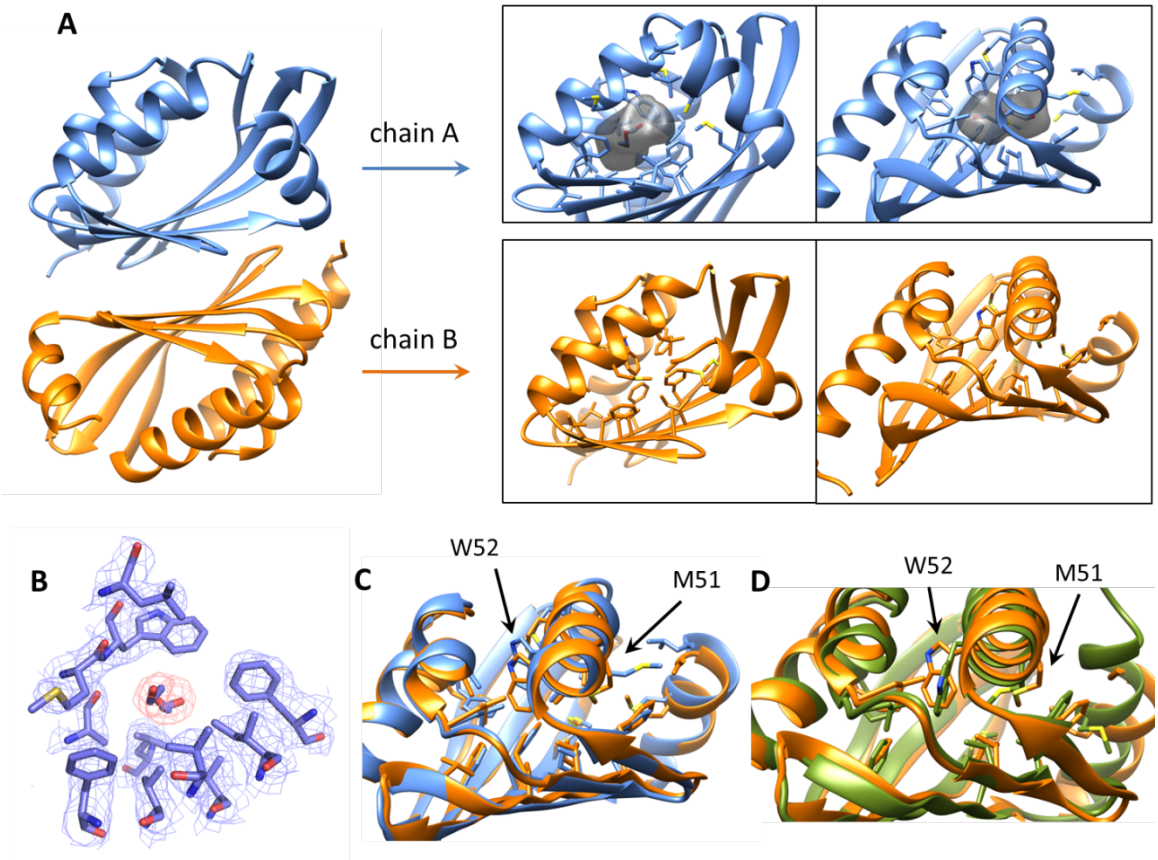


Fig. S.3.16. Formation of a binding cavity in the crystal structure of dcs_E_4_dim9. A. Chain A undergoes reorganization in the side-chains of M51 and W52 that enables binding of a diethylene glycol molecule from the crystallization solution. In chain B, the two side-chains have a different conformation that occludes the cavity, providing tighter hydrophobic packing. The cavity formed in chain A was calculated with 3V webserver (63) using an outer probe radius of 2.5 Å, an inner probe radius of 1.5 Å and a grid size of 0.5 Å. The calculated cavity volume is 191 $Å^3$. **B.** The internal cavity formed in chain A binds a diethylene glycol molecule as shown by the electron density. **C.** Superimposition of both chains highlights the differences in M51 and W52 side-chain conformations. **D.** Superimposition of chain B from the crystal structure (orange) with one of the two symmetric chains of the design model (green) shows that the designed conformation of M51 closely matches that from the chain B crystal structure, which provides better hydrophobic packing.

Table S.3.1. Summary of design experimental characterization.

Design name	Expressed	Soluble	CD spectra (25 °C)	T_m(°C)	Two-state GdmCl unfolding §	Oligomeric state†	HSQC quality*
dcs_A_1	Y	Y	αβ	> 95°C	Y	M	2
dcs_A_2	Y	Y	αβ	> 95°C	Y	M	2
dcs_A_3	Y	Y	αβ	> 95°C	Y	M	2
dcs_A_4	Y	Y	αβ	> 95°C	Y	M	2
dcs_B_1	N						
dcs_B_2	Y	Y	αβ	> 95°C	Y	M	2

dcs_B_3	Low	Y	αβ	> 95°C	N/A	N/A	N/A
dcs_B_4	N						
dcs_B_5	Y	Y	αβ	> 95°C	Y	M	2
dcs_C_1	Y	Y	αβ	> 95°C	Y	M	3
dcs_C_2	Y	Y	αβ	> 95°C	Y	M	2
dcs_C_3	Low	Y	αβ	> 95°C	Y	M	4
dcs_C_4	N						
dcs_C_5	Y	Y	αβ	> 95°C	Y	M	N/A
dcs_D_1	Y	Y	αβ	> 95°C	Y	M	2
dcs_D_2	Y	Y	αβ	> 95°C	Y	M	2
dcs_D_3	Y	Y	αβ	~85°C	Y	M	3
dcs_D_4	Y	Y	αβ	> 95°C	Y	M	4
dcs_D_5	Y	Y	αβ	> 95°C	Y	M	3
dcs_D_6	Y	Y	αβ	> 95°C	Y	M	3
dcs_D_7	Y	Y	αβ	> 95°C	Y	M	1
dcs_D_8	Y	Y	αβ	> 95°C	Y	M	3
dcs_D_9	N						
dcs_E_1	Y	Y	αβ	> 95°C	Y	M	2
dcs_E_2	Y	Y	αβ	> 95°C	Y	M	2
dcs_E_3	Y	Y	αβ	> 95°C	Y	M	3
dcs_E_4	Y	Y	αβ	> 95°C	Y	M	3
dcs_E_5	Y	Y	αβ	> 95°C	Y	M	3
dcs_F_1	Y	Y	αβ	> 95°C	N/A	M	N/A
dcs_F_2	Y	Y	αβ	> 95°C	N/A	M	N/A
dcs_F_3	Y	Y	αβ	> 95°C	N	M	3
dcs_F_4	Y	Y	αβ	> 95°C	N/A	M	N/A
dcs_F_5	Y	Y	αβ	> 95°C	N	M	N/A
dcs_F_6	Y	Y	αβ	> 95°C	N	M	3
dcs_F_7	Y	Y	αβ	> 95°C	N	M	3
dcs_F_8	Y	Y	αβ	90°C	N	M	N/A
dcs_F_9	Y	Y	αβ	> 95°C	Y	M	3
Disulfide variants							
dcs_C_1_ss	Y	Y	αβ	> 95°C	Y	M	1
dcs_C_2_ss	Y	Y	αβ	> 95°C	Y	M	N/A
dcs_C_4_ss	N						
dcs_C_5_ss	Y	Y	αβ	> 95°C	Y	M	N/A
dcs_D_4_ss1	Y	Y	αβ	> 95°C	Y	M	N/A
dcs_D_4_ss2	Y	Y	αβ	> 95°C	Y	M	N/A
dcs_D_4_ss12	Y	Y	αβ	> 95°C	Y	M	N/A
dcs_D_8_ss	Y	Y	αβ	> 95°C	Y	M	N/A
Homodimers							
dcs_E_4_dim1	Y	Y	N/A	N/A	N/A	D	N/A
dcs_E_4_dim2	Y	Y	N/A	N/A	N/A	M/D‡	N/A
dcs_E_4_dim3	Y	Y	N/A	N/A	N/A	M/D‡	N/A

dcs_E_4_dim4	Y	Y	N/A	N/A	N/A	M/D‡	N/A
dcs_E_4_dim5	Y	Y	N/A	N/A	N/A	D	N/A
dcs_E_4_dim6	N						N/A
dcs_E_4_dim7	N						N/A
dcs_E_4_dim8	N						N/A
dcs_E_4_dim9	Y	Y	αβ	> 95°C	Y	D	4
Cavity mutants							
dcs_C_1_ss_cav1	Y	Y	αβ	~85°C	Y	M	N/A
dcs_C_1_ss_cav2	Y	Y	αβ	N/A	N/A	A	N/A
dcs_C_1_ss_cav3	Y	Y	αβ	> 95°C	Y	M	N/A
dcs_C_1_ss_cav4	Y	Y	αβ	> 95°C	Y	M	N/A
dcs_C_1_ss_cav5	Y	Y	αβ	> 95°C	Y	M	N/A
dcs_C_1_ss_cav6	Y	Y	αβ	N/A	N/A	A	N/A
dcs_D_2_cav1	Y	Y	αβ	75°C	Y	M	N/A
dcs_D_2_cav2	Y	Y	αβ	65°C	N	M	N/A
dcs_D_2_cav3	Y	Y	αβ	65°C	N	M	N/A
dcs_E_4_dim9_cav1	Y	Y	αβ	> 95°C	Y	D	N/A
dcs_E_4_dim9_cav2	Y	Y	αβ	> 95°C	Y	M	N/A
dcs_E_4_dim9_cav3	Y	Y	αβ	> 95°C	Y	D	N/A

§ The denaturation curve was sigmoidal and could be fitted to a two-state folding mechanism.
† Oligomeric state of the dominant species based on SEC-MALS (M, monomer ; D, dimer). A, denotes dominant aggregate species
‡ The error in the molecular weight estimate is too high to determine whether the main peak corresponds to a monomer or dimer species.
***** HSQC quality was ranked from 1 to 4 based on the peak dispersion and intensity (64): 1, excellent; 2, good; 3, promising; 4, poor.

Table S.3.2. Designed protein sequences. The lowest E-value obtained from BLAST (30, 31) searches (against the NCBI nr database of non-redundant protein sequences) is shown.

Design name	Amino acid sequence	E-value
dcs_A_1	KSDELQKRVVEYAKEVILRQKGDPTLDIQVKRVQTTGNT LRVELEIRTGNTTRQYQIEVEIRGDTFQVRRVQETGGS	>10
dcs_A_2	KDDELQKRVVEYAKEVLLRQKGDPTTDIQVKRVQTTGN TVRVELELRVGNETTQMQIEVEIQGDTFQVRRVQKTGG S	>10
dcs_A_3	PSEEEEKRQVKQVAKEKLLEQSPNSKVQVRRVQKQGN TIRVELELRTNGKKENYTVEVERQGNTWTVKRITRTVGS	>10
dcs_A_4	PSEEEEKRRAKQVAKEKILEQNPSSKVQVRRVQKQGNT IRVELEITENGKKTNITVEVEKQGNTFTVKRITETVGS	5.4
dcs_B_1	QDIVEAAKQAAIAIFQLWKNPTDPKAQKLLKKILSPDLLK QMEKHARKLQKQGIHFEVKRVEVEKTGNTVQVTVEIEK TTGGTRQRRTYQMRFEVDGDTIRRVTVTEVGS	>10
dcs_B_2	QDIVEAAKQAAIAIFQLWKNPTDPEAQELLNKILSPDVLD QVREHARELQKQGIHFEVKRVEVTTDGNTVNVTVELEE TTGGTTTNTTYELRFEVDGDTIRRVTVTQNGS	0.81
dcs_B_3	QDIVEAAKQAAIAYFQLLKNPTDPEAQNLLNKILSPDVLD QVKEHAKKLQKQGIHFEVKRVEVETTGNTVKVKVELEK ETGGTRQRKRYTLRFEVDGDTIKRVTTTQTGSWS	2.3
dcs_B_4	QDIVEAAKQAVIAYFQLLKNPTDPDAQNLLRKILSPDLLE	0.75

	QIKRHARQLQKQGIHFEVKRVEVETTGNTVKVTVEIEKK TGGTRTRKRYKLRFEVDGDTIKRVTVTQTGSWS	
dcs_B_5	QDIVEAAKQAAIAYFQLLKNPTDPDAQNLLRKILSPDVLE QIKRHARQLQKQGIHFEVKRVEVTTTGNTVQVTVEIEET TGGTTTQTTYKLRFEVDGDTIKRVTVTQTGSWS	>10
dcs_C_1	SEEAKIAIELFKEAMKDPERFKEMVSPDTRIESNGQEYR GSEEAKKFAEEMKKTHPWEVRVERYRSDGDRFEIELRV NFNGKTFRMEIRMRKVNGEFRIEEMRLHG	0.72
dcs_C_2	QPDEVKKIAQEWWERMMRNPRQIEELIDPNTRLRDGNT ELTGREVQEYMKEWVTKVRFEVKEVTKEGNVYRVRLK VEENGKTKEMEIRLEDDNGRMRFKEIEIRG	>10
dcs_C_3	DKEEAKKLAELIERAYRNPDVAREVFSPNTRFEDNGRE THDVEEWMEEIKRQGRPVEVRVKEITRDGNEMRIRLRIR YNGEEYEMEIRFRHEDGQWKIEEMRWRG	0.43
dcs_C_4	DDIEKMMKKFVQWMRDGNPEYVERMVSPNTKFRHNG QETKGSDIVREWMKKLLNMRVEVKRYRIKNGELELEIEF ETGDRTSTVTFRLRLENGQMHLEEMEFRN	1.0
dcs_C_5	SEDDVRREVQRVWEEIRNNPEALREYVDPNTHLHDGN QQYSGEEVQEYMRELVTRVEFRVRRVEKKGNTWKVEV EVRENGQEKEMHIEFEEDNGKFKFKRIEIRG	1.1
dcs_D_1	PEEEKMARLFIEAVEKGDPELMRKVISPDTRVEDNGREF TGDEVSEWVKEIQKRGEQWHLRRYTKEGNSWRFELQV DNNGQTEQWEVQIEVRNGRIKRVTVTHV	0.00002
dcs_D_2	PEEEKAARLFIEALEKGDPELMRKVISPDTRMEDNGREF TGDEVVEYVKEIQKRGEQWHLRRYTKEGNSWRFEVQV DNNGQTEQWEVQIEVRNGRIKRVTITHV	0.00002
dcs_D_3	SPEKEESKLVEEFMKLMEQGDPEEMLKLISPDTRLEKD GEEYNGEEVRQYWEKEMREGTKFQVREVTTQGNKVRI RVQVQQNGTTTQEQYEVEMRDGRIRRITVHTRG	0.026
dcs_D_4	SPEKEESKLVEEFMKLMEQGDPEEMKKLISPDTRLERD GEEYNGEEVRQFWEEEMRQGLKFQVREVTTQGNKVRI RVQVQKNGTTTQVQFEVEMRDGRIRRITVHERG	0.003
dcs_D_5	SEEESKVAQEMMKMISKGDPDEIRKHMSPDTRVDFNG EEYSGEEVARMWEKERRKGRQYEVKRYQSKGNEVQF ELEVQDNGKTETIQIRVRVENGRVKEVQITTH	>10
dcs_D_6	SEEEEKVAQEMMKAIQKGDPDEIRKYLSPDVRVKVNGE EYSGEEVVRYWEKERRKGRRWEVKRYQTDGNEVQFE LQVEDNGKTEQYEIRVRVENGRVKEIQITTH	0.087
dcs_D_7	SEEEERVAKEMMEAIQKGDPDEIRKYLSPDVRVKVNGE EYSGEEVVRYWEKEKRKGRRWEVKRYQTKGNEVQFE LQVEDNGKTEQWEIRVRVENGRVKEIQITQH	0.003
dcs_D_8	PEVVKVWKRIMEALQKGDPELLKKMISPDTRMEVNGQT FTGEEVVRYWEEEIRRGRQWTVKRYTEKGNEVEFEVE QQDGDETRTYRVQVRVRNGQVEEIQVTQV	0.53
dcs_D_9	SEHEKHARQIEKAWKKGNPEELKKVVSPDTRMDFNGE EYRGKERIEEMMRRKKERGVEITLERVQHKGNELQLRV QFTEGNQTKQYEFRFEFENGQVRRVEVREN	0.022
dcs_E_1	SREEIRKVVEEMLRSLKQGSPEDISKYLSPDVRLEVGNY TFEGSEQVTKFWRMWTKFVDRVEVRKVQVDGNHVRV EMEVEWNGKRWTFEMEVEVRNGKIKRIRLQVDPEFKK VVQNIWNLL	0.007
dcs_E_2	TKDEVKKMVEILKKAFEEGDPEKIVSLLSPNVRLEMGNY TWEGSEQVEEFLRYLMEIVDRVEVRRIKVRPNHIEVEVE MEFNGKSFEVEWRFEIENGKVRRVEVRVTPEMKKIVEK VYRKA	0.23

dcs_E_3	SREEIRKVVEEMVRKLKQGSPEDISKYLSPDVRLEVGNY TFEGSEQVTKFWRMLTKFVDRVEVRKVQVDGNHVRVE VEVEWNGKKWTFEVEVEVRNGKIKRIRLQVDPEFKKVV QNIWNLL	0.002
dcs_E_4	TQEEVRKIMEKLKKAFKQGNPEQIVSLLSPDVRVKVGN QEFSGSEEAEKMWRKLMKFVDRVEVRRVKVDENRVEI EVEFEVNGQRYSMEFHFEVENGKVRRVEIRISPTMKKL MKQILNYG	1.1
dcs_E_5	TKKEVEKMARTFKEAMNQGNPEQLTSKLSPDVRLRIGN QEFEGSEEVEKWLRRWFNLVDRVEVRRIKVEDNHVEV EVEVELNGKNVEIEFRFEIRNGKVERMEIRVTPDMKKFA EKINKYG	0.001
dcs_F_1	DENEKMKKMVRQFLELIEKEDPDEIRKLLSPDTRVTFNG RTFTGPEEFAKELQELRKQGIRFQFTEAEIQTDNGKLQI RVEVTLTVNGQEYRSEVTFTIRVENGVIKEVTIQFSPKLQ EALKGGS	0.48
dcs_F_2	DENEKMKEAVRQFLELIEKEDPDEIRKLLDPNTRVTFNG KTFTGPEEFAKELQELRKQGIRFQFTVKEIQTDNGKLQI RVEVTLTVNGQEYRSEVTFTIRVENGVIKEVTIQFSPKLQ EALKGGS	0.52
dcs_F_3	DENEKMKEMVREFLEIIEKRDPNEIRKLLDPNTRVTFDG RTYTGPEEFAKELQELEKQGIEFQFTIKEIQTDNGVLQIR VEVTLTVNGQEYRSEVTFTIRVENGVIKEVTIQFSPKLQE ALKGGS	0.82
dcs_F_4	DENEKMKEMVREFLELIEKRDPEEMRKLLSPDTRVTFD GKTFTGPEEFAKELQELEKQGIEMQYTVKEIQTDNGVLQ IRVEVTLTVNGQEYRSEVTFTIRVENGTIKEVTIQYSPKL QEALKGGS	0.18
dcs_F_5	DEDEKMKEIVKQFLELIKREDPEELRKLLSPDTRVTFNG RTYTGPEEFAKELQEMRKRGVRFQFTIKEVRTVNGVMK IRFEVQVTVNGVTYRSEVTIQIRVENGVIKEVTIQFSPKL QEAIEGGS	0.067
dcs_F_6	DENEKMKEIVKQFLELIKREDPEELRKLLSPDTRVTFDG RTFTGPEEFAKELQEMRKRGVRFQFTEAEVQTDNGKLK IRFEVQVTVNGQTYRSEVTIQIRVENGVIKEVTIQFSPKL QEAIEGGS	0.007
dcs_F_7	DEDEKMKEIVKQFLELIKRRDPEELRKLLDPNTRVTFNG KTFTGPEEFAKELQELEKRGVEMQYTIKEVQTDNGKMKI RFEVQVTVNGQTYRSEVTIQIRVENGVIKEVTIQYSPKLQ EALEGGS	6.5
dcs_F_8	DEDEKMKEIVKQFLELMKRRDPEEMRKLLDPNTRVTFN GKTFTGPEEFAKELQEMEKRGVEFQFTIKEVRTVNGVM KIRFEVQVTVNGVTYRSEVTIQIRVENGVIKEVTIQFSPKL QEAIEGGS	1.4
dcs_F_9	DPAEQAREIVRQFLELIQRRDPEELRRLLSPDTRVTFNG RTFTGPERFAEALQELERRGVEMQYTIQEVQTENGRMS IRFEVQVTVNGQTYRSEVTIQIRVENGRIREVTIQYSPRL QEALEGGSGW	0.02
Disulfide variants		
dcs_C_1_ss	SEEAKIAIELFKEAMKDPERFKEMCSPDTRIESNGQEYR GSEECKKFAEEMKKTHPWEVRVERYRSDGDRFEIELRV NFNGKTFRMEIRMRKVNGEFRIEEMRLHG	0.84
dcs_C_2_ss	QPDEVKKIAQEWWERMMRNPRQIEELIDPNTRCRDGN TELTGRECQEYMKEWVTKVRFEVKEVTKEGNVYRVRL KVEENGKTKEMEIRLEDDNGRMRFKEIEIRG	>10

dcs_C_3_ss	DKEEAKKLCELIERAYRNPDVAREVFSPNTRFEDNGRE THDVEEWMEEIKRQGRPVECRVKEITRDGNEMRIRLRI RYNGEEYEMEIRFRHEDGQWKIEEMRWRG	1.2
dcs_C_5_ss	SEDDVRREVQRVWEEIRNNPEALCEYVDPNTHLHDGN QQYSGEEVCEYMRELVTRVEFRVRRVEKKGNTWKVEV EVRENGQEKEMHIEFEEDNGKFKFKRIEIRG	1.6
dcs_D_4_ss1	SPCKEESKLVEEFMKLMEQGDPEEMKKLISPDTRLERD GEEYNGEEVRQFWEEEMRQGLKFQVREVTTQGCKVRI RVQVQKNGTTTQVQFEVEMRDGRIRRITVHERG	0.006
dcs_D_4_ss2	SPAKEESKLVEEFMKLMEQGDPEEMCKLISPDTRLERD GEEYNGEEVCQFWEEEMRQGLKFQVREVTTQGAKVRI RVQVQKNGTTTQVQFEVEMRDGRIRRITVHERG	0.002
dcs_D_4_ss12	SPCKEESKLVEEFMKLMEQGDPEEMCKLISPDTRLERD GEEYNGEEVCQFWEEEMRQGLKFQVREVTTQGCKVRI RVQVQKNGTTTQVQFEVEMRDGRIRRITVHERG	0.021
dcs_D_8_ss	PECVKVWKRIMEALQKGDPELLKKMISPDTRMEVNGQT FTGEEVVRYWEEEIRRGRQWTVKRYTEKGNECEFEVE QQDGDETRTYRVQVRVRNGQVEEIQVTQV	0.38
Homo-dimer designs		
dcs_E_4_dim1	TEEEVRKIMEKLKKAFKQGNPEQIVSLLSPDVRVQVGN QEFSGSEEAEKMWRKLMKFVDRVEVRRVSVFENVVVIE VEFEVNGQRYSMIFVFFVENGKVSMVIIYISPTMAKLMK QILNYG	0.061
dcs_E_4_dim2	TREEVRKIMEKLKKAFKQGNPEQIVSLLSPDVVVVVGNQ DFKGSEEAEKMWRKLMKFVDRVEVKKVQVYENIVIIEVE FEVNGQRYEMLFTFYVENGKVKMVSIFISPTMKKLMKQI LNYG	0.037
dcs_E_4_dim3	TEEEVRKIMEKLKKAFKQGNPEQIVSLLSPDVVVVVGNQ SFSGSEEAEKMWRKLMKFVDRVEVRKVRVFENIVLIEV EFEVNGQRYSMFFTFYVENGKVAAVSIWISPTMKKLMK QILNYG	0.4
dcs_E_4_dim4	TAEEVRKIMEKLKKAFKQGNPEQIVSLLSPDVFVMVGNQ SFSGSEEAEKMWRKLMKFVDRVEVKKVQVYENIVIIEVE FEVNGQRYAMLFTFYVENGKVKAVSIFISPTMKKLMKQI LNYG	1.2
dcs_E_4_dim5	TEEEVRKIMEKLKKAFKQGNPEQIVSLLSPDVAVQVGNQ EFSGSEEAEKMWRKLMKFVDRVEVRDVRVAENIVVIFV EFEVNGQRYVMAFVFFVENGKVSQVVIYISPTMKKLMK QILNYG	0.75
dcs_E_4_dim6	SREEIRKVVEEMLRSLKQGSPEDISKYLSPDVRLEVGNY TFEGSEQVTKFWRMWTKFVDRVEVKEVKVAGNYVIVV MSVEWNGKRWEATMIVTVRNGKIKRIILAVDEEFKKVVQ NIWNLL	0.001
dcs_E_4_dim7	SREEIRKVVEEMLRSLKQGSPEDISKYLSPDVFLLVGNY TFEGSEQVTKFWRMWTKFVDRVEVRRVEVAGNAVVVL MEVEWNGKRWTFYMLVVVRNGKIKRIALAVDPEFSKVA QNIWNLL	0.16
dcs_E_4_dim8	TREEARKIMEKLKKAFKQGNPEQIVSLLSPDVRVVVGNQ EFKGSEEAEKMWRKLMKFVDRVEVARVRVDENMVVIA VEFEVNGQRYVMFFAFVVENGKVKAVFIFISEEAMKLMK QILNYG	1.8
dcs_E_4_dim9	TEEEVRKIMEKLKKAFKQGNPEQIVSLLSPDVKVDVGNQ SFSGSEEAEKMWRKLMKFVDRVEVRDVRVFEAAVMIA VEFEVNGQRYEMIFTFYVENGKVSMVSIYISPTMKKLMK QILNY	0.37

Cavity mutants		
dcs_C_1_ss_cav1	SEEAKIAIELFKEAMKDPERFKEMCSPDTRIESNGQEYR GSEECKKFAEEMKKTHPWEVRVERYRSDGDRFEIELRV NFNGKTTRTEIRMRKVNGEFRIEEMRSHG	1.4
dcs_C_1_ss_cav2	SEEAKIAIELFKEAMKDPERFKEMCSPDTRIESNGQEYR GSEECKKSAEEMKKTHPWEVRVERYRSDGDRFEIELR VNFNGKTTRTEIRMRKVNGEFRIEEMRSHG	2.3
dcs_C_1_ss_cav3	SEEAKIAIELFKEAMKDPERFKEMCSPDTRIESNGQEYR GSEECKKYAEEMKKTHPTEVRVERYRSDGDRFEIELRV NSNGKTFRMEIRMRKVNGEFRIEEMRLHG	7.0
dcs_C_1_ss_cav4	SEEAKIAIELFKEAMKDPERFKEMCSPDTRIESNGQEYR GSEECKKSAEEMKKTHPTEVRVERYRSDGDRFEIELRV NSNGKTFRMEIRMRKVNGEFRIEEMRLHG	1.7
dcs_C_1_ss_cav5	SEEAKIAIELFKEAMKDPERFKEMCSPDTRIESNGQEYR GSEECKKYAEEMKKTHPTEVRVERYRSDGDRFEIELRV NSNGKTTRTEIRMRKVNGEFRIEEMRSHG	2.5
dcs_C_1_ss_cav6	SEEAKIAIELFKEAMKDPERFKEMCSPDTRIESNGQEYR GSEECKKSAEEMKKTHPTEVRVERYRSDGDRFEIELRV NSNGKTTRTEIRMRKVNGEFRIEEMRSHG	0.85
dcs_D_2_cav1	PEEEKAARLFIECLEKGDPECMRKVISPDTRVEFNGSEL TGDEVVESVKELQKSGTQLHLRRYTKEGNSWRFEIQAD NNGQTWQSEIQIEVRNGRIKRATSTA	1.5
dcs_D_2_cav2	PEEEKAARLFIEALEKGDPELCRKVISPDTRAEINGSEYT GDEVVESCKELQKSGTQIHLRRYTKEGNSWRFEVQAD NNGQTYQSEIQIEVRNGRIKRATSTA	2.7
dcs_D_2_cav3	PEEEKACRLFIEALEKGDPELMRKVISPDTRAEINGREFT GDEVVESVKEMQKRGVQAHLRRYTKEGNSCRFEVQTD INGQTEQSEIQIEVRNGRIKRATTTA	0.00005
dcs_E_4_dim9_cav1	TEEEVRKIMEKLKKAFKQGNPEQIVSLLSPDVKVDVGNQ SFSGSEEAEKAQRKLMKFVDRVEVRDVRVFENAVMIAV EFEVNGQRYKMITTFYVENGKVSMVSIYISPTMKKLMKQ ILNYG	2.9
dcs_E_4_dim9_cav2	TEEEVRKIMEKLKKAFKQGNPEQIVSLLSPDVKVDVGNQ SFSGSEEAEKAARKLMKFVDRVEVRDVRVFENAVMIAV EFEVNGQRYKVIVTFYVENGKVSMVSIYISPTMKKLMKQ ILNYG	0.72
dcs_E_4_dim9_cav3	TEEEVRKIMEKLKKAFKQGNPEQIVSLLSPDVKVDVGNQ SFSGSEEAEKAARKLMKFVDRVEVRDVRVFENAVMIAV EFEVNGQRYKMIFTFYVENGKVSMVSIYISPTMKKLMKQ ILNYG	0.21

Table S.3.3. Parameters fitted to GdmCl denaturation curves for designed proteins. Denaturation curves were measured for those proteins with soluble expression. N/A indicates data that is not available due to the lack of sigmoidal character in the denaturation curves (highly linear). In those cases the slope of the native baseline was calculated from a linear fit, which are indicated by an asterisk (*). For designs with disulfides, denaturation curves in the presence of the reducing agent Tris(2-carboxyethyl)phosphine (TCEP) are also shown.

Design name	slope native baseline (M^{-1})	m-value ($kcal \cdot mol^{-1} \cdot M^{-1}$)	m-value deviation (%)	ΔG ($kcal \cdot mol^{-1}$)	C_m (M)
dcs_A_1	0.053	2.0	-3.6	-5.0	2.5
dcs_A_2	0.033	2.0	-4.5	-4.0	2.1
dcs_A_3	-0.010	2.1	5.5	-3.8	1.8
dcs_A_4	-0.066	1.7	-13.4	-4.5	2.7

dcs_B_2	0.043	2.8	-14.6	-10.5	3.8
dcs_B_5	0.043	3.1	-5.3	-13.4	4.5
dcs_C_1	0.036	1.4	-55.5	-3.2	2.3
dcs_C_2	-0.019	4.0	30.2	-8.5	2.2
dcs_C_3	-0.039	0.8	-72.9	-2.8	3.3
dcs_C_5	-0.013	4.2	34.2	-9.2	2.3
dcs_D_1	0.197	3.7	20.3	-9.7	2.6
dcs_D_2	0.057	2.8	-8.0	-7.7	2.8
dcs_D_3	0.167	3.8	19.0	-6.0	1.6
dcs_D_4	0.047	2.7	-14.4	-4.8	1.8
dcs_D_5	0.052	3.0	-4.1	-4.5	1.6
dcs_D_6	0.064	2.6	-16.8	-6.1	2.4
dcs_D_7	0.068	4.4	42.1	-10.8	2.5
dcs_D_8	0.089	3.0	-1.2	-8.8	2.9
dcs_E_1	0.011	3.9	6.7	-23.1	6.0
dcs_E_2	0.016	4.2	16.3	-23.0	5.5
dcs_E_3	0.008	3.3	-8.4	-19.7	6.0
dcs_E_4	0.011	3.8	6.3	-18.4	4.8
dcs_E_5	0.040	2.5	-29.9	-9.7	3.9
dcs_F_3	0.213*	N/A	N/A	N/A	N/A
dcs_F_5	0.203*	N/A	N/A	N/A	N/A
dcs_F_6	0.125*	N/A	N/A	N/A	N/A
dcs_F_7	0.200*	N/A	N/A	N/A	N/A
dcs_F_8	0.221*	N/A	N/A	N/A	N/A
dcs_F_9	0.078	3.5	-4.7	-11.8	3.4
Disulfide variants					
dcs_C_1_ss	-0.005	3.5	9.6	-11.5	3.4
dcs_C_1_ss + TCEP	0.067	2.1	-33.5	-5.7	2.7
dcs_C_5_ss	-0.013	2.3	-25.9	-6.7	3.0
dcs_D_4_ss1	0.066	2.9	-8.5	-7.1	2.5
dcs_D_4_ss2	0.121	3.5	10.2	-9.7	2.8
dcs_D_4_ss12	0.065	3.3	3.0	-11.7	3.6
dcs_D_4_ss12 + TCEP	-0.032	1.4	-56.3	-2.5	3.2
dcs_D_8_ss	-0.003	3.6	17.1	-16.4	4.6
dcs_D_8_ss + TCEP					
Homodimer designs					
dcs_E_4_dim9	0.030	3.67	1.46	-19.77	5.44
Cavity mutants					
dcs_C_1_ss_cav1	-0.212	1.75	-44.74	-2.16	1.41
dcs_C_1_ss_cav3	0.223	4.0	26.49	-10.73	2.62
dcs_C_1_ss_cav4	0.061	3.02	-4.49	-5.94	1.98
dcs_C_1_ss_cav5	-0.547	1.5	-52.46	-1.10	1.06

dcs_D_1_cav1					
dcs_E_4_dim9_cav1	-0.008	2.09	-42.24	-8.78	4.24
dcs_E_4_dim9_cav2	0.003	4.55	25.59	-18.13	4.03
dcs_E_4_dim9_cav3	0.029	3.02	-16.39	-13.19	4.66

Table S.3.4. X-ray crystallography data collection and refinement statistics

Design name	**dp_A_4**	**dp_D_2**	**dp_C_1_ss**
Data collection			
Space group	C2	P 21 21 21	C 2 2 21
Cell dimensions			
a, b, c (Å)	56.14, 70.62, 41.04	28.25, 34.36, 100.39	81.31, 101.54, 101.58
α, β, γ (°)	90, 113.16, 90	90, 90, 90	90, 90, 90
Wavelength (Å)	0.97916	0.97625	1.0
Resolution (Å)	2.44 (2.44-2.48)	2.0 (2.0-2.05)	3.0
R_{sym} or R_{merge} (%)	5.2 (6.0)	4.7 (23.3)	11 (101)
CC1/2	0.986	0.99 (0.98)	0.91 (0.64)
I/σI	33.87 (18.23)	24.6 (7.7)	16 (1.1)
Completeness (%)	93.1 (42.8)	92.3 (97.0)	98 (87)
Redundancy	6.7 (6.2)	7.8 (7.8)	8 (7)
Refinement			
Resolution (Å)	2.44	2.0	3.0
No. reflections	5311	6503	8364
R_{work} (%) / R_{free} (%)	21.8 / 25.6	17.2/20.1	27.9/29.1
No. atoms			
Protein	1216	918	2760
Water	59	69	21
B-factors (Å^2)			
Protein	27.6	31.1	103.6
Water	29.8	42.0	65.3
R.m.s. deviations			
Bond lengths (Å)	0.003	0.007	0.003
Bond angles (°)	0.584	0.875	0.808
Ramachandran statistics (%)			
Favored	99	99	98
Outliers	0	0	0
Rotamer outliers (%)	0	0	4.2

*Values in parentheses are for highest-resolution shell.

Table S.3.5. X-ray crystallography data collection and refinement statistics

Design name	**dcs_E_3**	**dcs_E_4**	**dcs_E_4_dim9**
Data collection			
Space group	P 41 21 2	P 42 21 2	P 1 21 1
Cell dimensions			
a, b, c (Å)	49.81, 49.81, 113.1	75.53, 75.53, 50.07	38.21, 32.79, 86.48
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 92.11, 90
Wavelength (Å)			
Resolution (Å)	3.10 (3.31-3.10)	2.91 (3.09-2.91)	2.47 (2.57-2.47)
R_{sym} or R_{merge} (%)	7.2 (26.8)	3.5 (20.8)	3.5 (18.4)
CC1/2	0.999 (0.99)	0.999 (0.98)	0.999 (0.98)
I/σI	21.7 (8.8)	75.2 (7.9)	22.1 (7.0)
Completeness (%)	100 (100)	96	100
Redundancy	11.7 (12.4)	4.5 (4.5)	3.7 (3.8)

Refinement			
Resolution (Å)	3.10	2.91	2.47
No. reflections	2891	3314	7943
R_{work} (%) / R_{free} (%)	22.6/ 25.2	24.7/26.8	21.3/25.9
No. atoms			
Protein	966	912	1807
Water	0	2	69
B-factors (Å^2)			
Protein	68.4	73.7	52.4
Water	-	75.3	55.6
R.m.s. deviations			
Bond lengths (Å)	0.006	0.002	0.004
Bond angles (°)	0.866	0.458	0.549
Ramachandran statistics (%)			
Favored	96	97	97
Outliers	0	0	0
Rotamer outliers (%)	0	0	0

*Values in parentheses are for highest-resolution shell.

Table S.3.6. NMR and refinement statistics for protein structures.

Design name	**dcs_A_3**	**dcs_B_2**
NESG ID	**OR485**	**OR664**
PDB ID	**5kph**	**5kpe**
NMR distance and dihedral constraints		
Distance constraints		
Total NOE	2012	3395
Intra-residue	553	673
Inter-residue		
Sequential (|i-j| = 1)	505	865
Medium-range (|i-j| ≤ 4)	301	655
Long-range (|i-j| ≥ 5)	653	1202
Intermolecular		
Hydrogen bonds	76	56
Total dihedral angle restraints	139	186
Phi	35	93
Psi	35	93
Structure statistics		
Violations		
RMS of distance violation/constraint¶ (Å)	0.01	0.01
RMS of dihedral angle violation/constraint (°)	0.88	0.93
Max distance constraint violation (Å)	0.66	0.40
Max dihedral angle violation (°)	7.80	974
Average medoid r.m.s.d.** (Å)		
Heavy	0.5±0.15	0.5±0.19
Backbone	1.1±0.10	0.9±0.11
RPF Scores		
Recall	0.977	0.963
Precision	0.929	0.973
F-measure	0.952	0.968
DP-scores	0.786	0.886
Structure quality factors (raw/Z-score⌘)		
Procheck G-factor (phi / psi only)**	-0.42/1.34	-0.20/-0.47
Procheck G-factor (all dihedral angles)**	0-.19/-1.12	-0.14/-0.83
Verify3D	0.34/-1.93	0.407/0.16

ProsaII (-ve)	0.79/0.58	1.08/1.78
MolProbity clashscore	15.34/-1.11	13.11/-0.72
Ramachandran plot summary from Richardson's lab		
Most favored regions (%)	97.1	98.6
Allowed regions (%)	2.8	1.4
Disallowed regions (%)	0.1	0

***** Analyzed for the 20 lowest energy refined structures for each designed protein, which are deposited in the PDB: OR485 (5kph, residues 1-85), DI_7S, OR664 (5kpe, residues 1-120) using PDBSTAT (65) and PSVS 1.4 (55, 56).

§ PEG and phage were used as alignment media 1 and 2.

¶ Calculated by using sum over r^{-6}.

⌘ With respect to mean and standard deviation for a set of 252 X-ray structures with sequence lengths < 500, resolution ≤ 1.80 Å, R-factor ≤ 0.25 and R-free ≤ 0.28; a positive value indicates a 'better' score.

** Calculated among 20 refined structures for ordered residues that have sum of phi and psi order parameters (66) S(phi)+S(psi)>1.8 (55). The ordered residues of OR485: 4-48, 50-75; OR664: 4-52 55-82, 84-108.

4.1 Abstract

Proteins from the NTF2-like superfamily exhibit a wide range of pocket shapes and sizes in a relatively small scaffold. This structural diversity is based on the configurations of a curved beta-sheet combined with at least three alpha helices, providing a variety of conformations that support a wide range of functions. Systematic generation of pocket structural diversity remains an outstanding challenge in protein *de novo* design. The relatively simple system that gives rise to wide structural diversity in NTF2-like proteins could be adapted to be used in *de novo* design. Here we build, test and improve a generative algorithm for proteins of the NTF2-like superfamily. We generate thousands of models, which we screen experimentally to improve and generalize the algorithm, resulting in an algorithm that creates diverse and stable proteins. This algorithm covers a significant part of the native NTF2-like structural space, and generates solutions not seen in nature. Finally, we use this generative algorithm to design an aflatoxin B1 binding protein.

4.2 Introduction

In their broadest definition, generative algorithms are sets of instructions that take a limited set of input parameters to generate solutions that meet certain criteria. The design and architecture communities have developed and used generative algorithms in recent times to explore design space in search of aesthetically and technically novel solutions (1–3). A unique aspect of generative algorithms is that, in order to generate a wide variety of productive solutions, they must capture the essential attributes of the objects they generate. Furthermore, differently from mere copying or memorization, these algorithms are able to produce novel solutions by rapidly sampling design space using a minimal set of instructions and requirements, limiting biases from history, human intervention and circumstantial constraints.

The value of developing a generative algorithm for protein design is two-fold: On one hand, it enables sampling of novel solutions not seen in nature, on the other, it can capture essential attributes of protein structure. In its most general implementation, a generative algorithm for protein structures contains a

minimal set of instructions that encode the physical constraints of the system, but is not biased by evolution, facilitating the generation of solutions distant from those found in nature. Conversely, structural space that is inaccessible to the generative algorithm and is not sampled by nature, may be off-limits due to physical constraints rather than evolutionary history.

The generative algorithm presented in this work produces atomic models of proteins belonging to the NTF2-like superfamily in two stages, first the protein backbone is assembled, then, the sequence is optimized to fold into that structure, similarly to the strategy described in Chapter 3. We begin with a rudimentary implementation that produces a set of predefined NTF2-like subfamilies (4), and build up from it to a more general version. In order to improve the conformational diversity and stability of the proteins generated, we cycle between stages of design, in vitro testing and algorithm refinement. The main qualitative jump during algorithm development is the transition from a “pre-defined fold” strategy, to a “decision tree” strategy. The “pre-defined fold” strategy consists in finding a set of parameters that consistently and efficiently produce NTF2-like proteins with similar overall shape, and narrow local sampling. The process of finding such parameters is time consuming and must be guided by a human intervention. The “decision tree” strategy focuses on rules that dictate the flow of input to output during backbone assembly, ensuring the building steps are always productive. This strategy makes human intervention unnecessary for diversity search, as at each step of the decision tree, the algorithm measures key features of the input, and provides a set of compatible output options.

In the final stage of this process we are able to produce a large variety of proteins, some of them not seen in nature before. We use this diversity and the knowledge we gained though massive design and testing to design an Aflatoxin B1 binding protein, as an example of the possible applications of the generative algorithm.

4.3 Results

4.3.1 An NTF2-like generative algorithm composed of discrete subfamilies

The combination of a backbone generation algorithm and sequence design, as described in Chapter 3, can be considered a rudimentary implementation of an NTF2 generative algorithm. Furthermore, we can bundle different versions of its backbone generation component, as well as creating new ones, and obtain more variable output. Such is the initial version of the NTF2 generative algorithm: We modified the parameters of the algorithms described in 3.5.1 and 3.5.2 to generate nine different backbone subfamilies (Methods 4.5.1). Each of these versions samples the local backbone structural space, and the same algorithm is used to design the sequence for all of them (Methods 4.5.1).

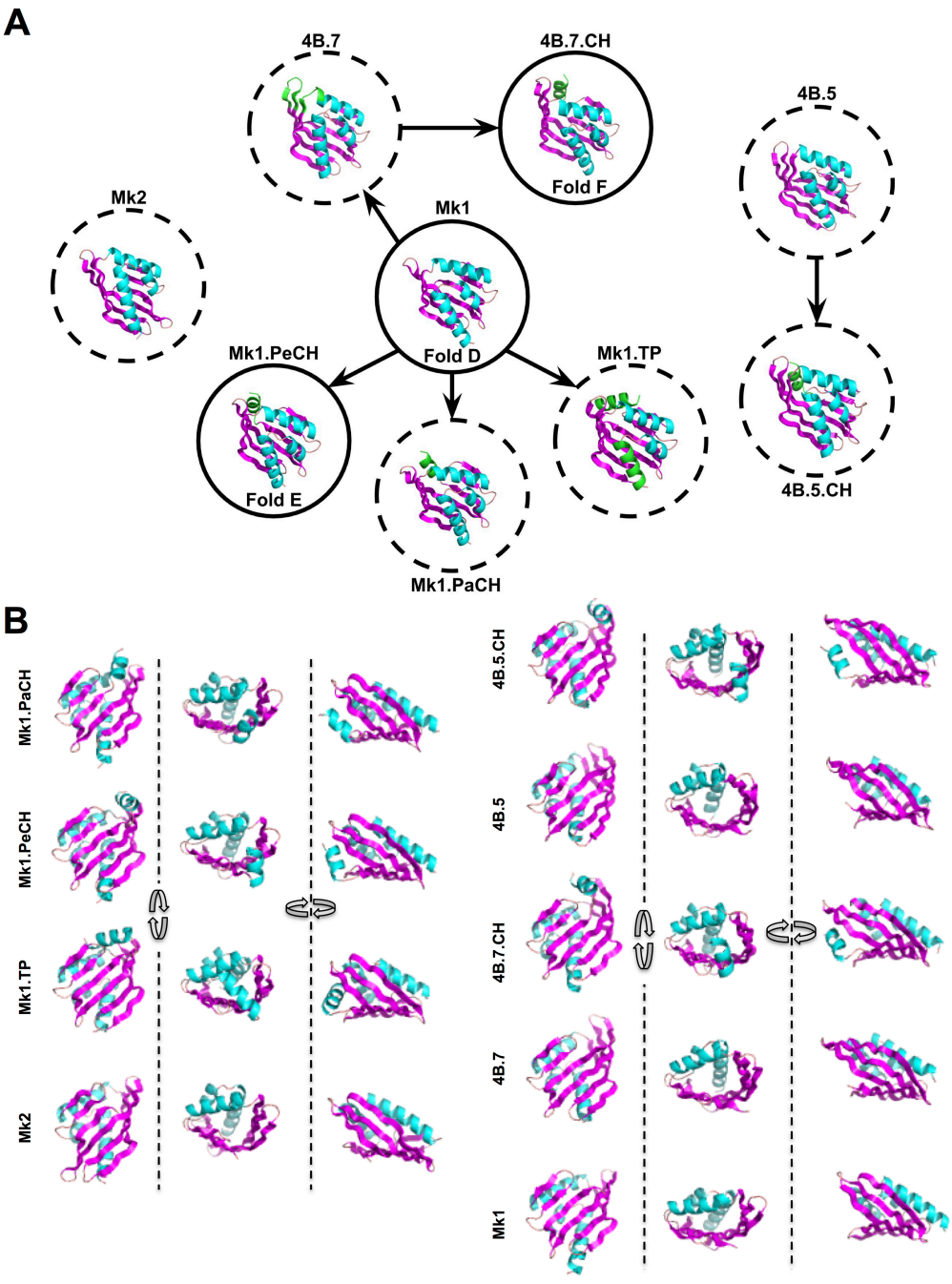


Figure 4.1: Nine different subfamilies generated by the first version of the generative algorithm A. Nine different subfamilies generated by the first version of the generative algorithm. Helices are colored cyan, strands magenta and loops tan. Solid circles indicate subfamilies described in Chapter 3, dashed circles indicate new subfamilies. The direction of the arrows indicates derivation of subfamilies from each other, with modifications colored green. **B**. Different perspective for each of the nine sampled subfamilies showing sheet from the bottom (left), side (right) and pocket entrance (middle). The same coloring scheme as A is used, excepts changes are not highlighted.

Since the final goal of this work is to generate proteins that can accommodate an active/binding site, we focused on subfamilies amenable to this. We discarded subfamilies (referred to as "folds" in Chapter 3) A-C and generated new ones, some of them based on subfamily D (Figure 4.1 A and B). These subfamilies sample several base widths, long and short arm lengths, C-terminal helix addition, and different opening placements (See Table 4.1 and Methods 4.5.1). We evaluated the range of pocket volumes sampled, and compared them to the distribution of volumes in native NTF2-like domains (Figure 4.2 A). From the volume distributions, it is clear that proteins designed without a particular function tend to have small pockets, likely due to the maximization of hydrophobic contacts by the Rosetta score function. To address this issue and sample more native-like volume values, we designed binding sites for a relatively large steroid molecule (Methods 4.5.2), cortisol (Figure 4.2 B), in 492 proteins from the Mk1.PeCH subfamiliy (Figure 4.2 C). More than half of the designs for cortisol binding have pocket volumes above 450 $Å^3$ (measured as described in 2.5.3), in a range that overlaps with native NTF2-like domains. Measured pocket volumes below 450$Å^3$ in binder designs are likely a result of the pocket not being open to solvent (cavity), a consequence of the design method not taking into account this opening feature and placing amino-acids in conformations that block the pocket opening.

Subfamily	Base width	Long arm length	Short arm length	C-helix	Opening	# Generated
Mk1	5	4	2	No	Typical	600
Mk1.PaCH	5	4	2	Yes – parallel to long arm	Typical	225
Mk1.PeCH	5	4	2	Yes – perpendicular to long arm	Typical	224
Mk1.TP	5	4	2	Yes – Occludes classic pocket entrance	Between H1-H2 loop and H3	597
Mk2	3	6	4	No	Typical	600
4B.5	7	4	2	No	Typical	600
4B.5.CH	7	4	2	Yes – parallel to long arm	Typical	600
4B.7	5	6	2	No	Typical	600
4B.7.CH	5	6	2	Yes – parallel to long arm	Typical	600

Table 4.1: **Structural characteristics of each of the nine folds produced by the first version of the generative algorithm**. Base width is measured as the number of residues between the bulges in S3 and S6. Long and short arm lengths are measured by the number of residues between the N-terminus of the strand (S3 for the long arm, S6 for the short arm) and the "A" ABEGO bulge residue.

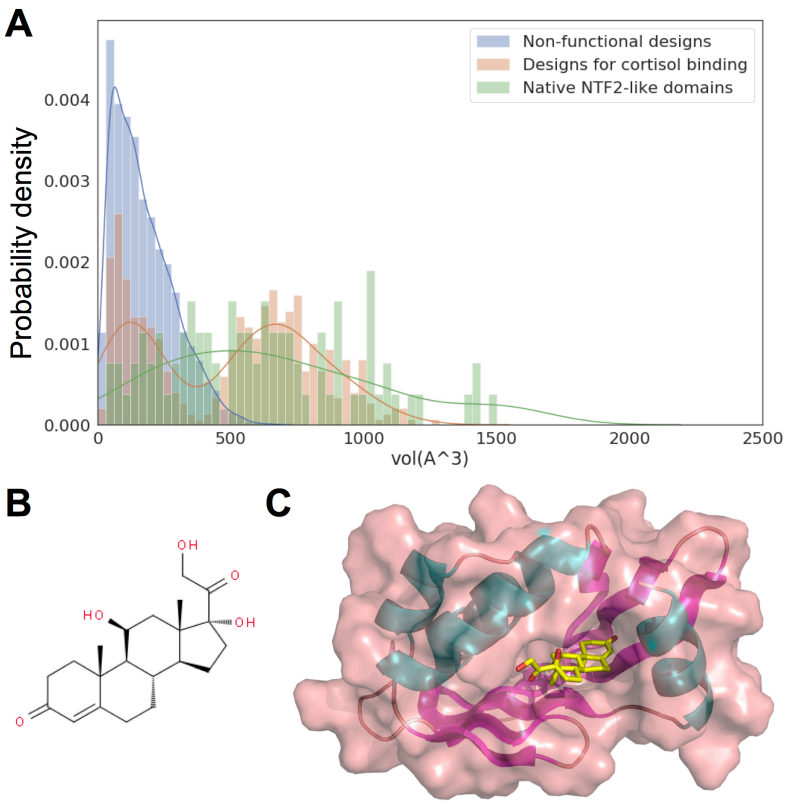


Figure 4.2: Pocket size distribution in de novo NTF2-like proteins A. Normalized histogram of pocket volumes of 4696 designs generated without specific function (non-functional, blue), 492 designs for cortisol binding (orange), and 92 native NTF2-like domains (green). **B**. Diagram of the cortisol molecule. **C**. Cartoon and surface representations of a design for cortisol binding with a ~650Å^3 pocket. Helices are colored cyan, strands magenta and loops tan. Surface is transparent, colored tan. The single cortisol molecule bound in the pocket is represented as yellow (carbon) and red (oxygen) sticks.

We evaluated the diversity of the proteins generated by the generative algorithm by comparing them to each other and to structures of the non-redundant NTF2-like domain set (see Chapter 2 – section 2.5.1). Because we focus on overall-protein diversity, in contrast to localized deviations, like loops, we employed TM-align (5) to produce protein-protein alignments, a method widely used in template modeling (see methods section 4.5.3 for more details). The reported metric for alignment quality by TM-align, is the TM-score. We quantified the diversity within the group of 92 non-redundant native NTF2-like domains, and within a set of 10 randomly selected models generated for each of the 9 sub-families (90 models total). Although there is no theoretical limit to the number of *de novo* models we can generate for each of the subfamilies, given there is a limited number of subfamilies, we decided to cap the number of models per subfamily to reach the number of structures in the native set in order carry out conservative

calculations for diversity. Figure 4.3.A shows the TM-scores of alignments within native group are in general lower than those within the *de novo* group, although both sample values above 0.5 almost exclusively, as expected for proteins in the same superfamily. This indicates that the native set is in general more diverse than the *de novo* set. With this in mind, we can ask to what extent does the *de novo* set sample the native set. To this end, we aligned each of the structures in the native set to all the *de novo* structures, and obtained the *de novo* models that best match each of them. We then calculate the z-score of each of the alignments with respect to the native vs. native TM-score distribution, and produce a ranking (red dots, figure 4.3 B). Most of the best *de novo* matches to natives (65%), have a z-score > 1, indicating they are at least one standard deviation better than the average native-native alignment – i.e., better than random. The same calculations can be done for quantifying how well do natives cover the *de novo* set, a measure of the novelty of the designed proteins (green dots, figure 4.3 B). In this case, close to 90% of the best matches have z-scores above 1. Using the same criterion as above, the *de novo* space is well covered by the known NTF2-like domain structures.

To better understand the alignment scores and coverage of the native space, it is worth looking at a series of alignments in a range of z-scores (Figure 4.3 C). For the native structure best recapitulated by a *de novo* design (d1tuha_), we can see almost perfect overlap of all strands in the main sheet, as well as helices 1 and 3. The frontal hairpin is extended by a loop in the native structure, which is not covered by the *de novo* match, helix 2 is slightly unraveled, the connection to helix 4 (C-terminal helix), and the helix itself do not overlap as well as the main sheet strands. Despite local differences, the *de novo* backbone traces the native structure closely. In contrast, the native structure for which the best *de novo* match has the lowest z-score (d2cw9a1) shows large sections of non-overlapping segments, with the native structure completely lacking the frontal hairpin, and having an unusually large number of additional secondary structure elements in the N-terminus. Additionally, the C-terminus of H3 and the long arm overlap poorly with the *de novo* match. Structures with intermediate match z-scores follow a similar trend: the z-scores get lower as the number of non-canonical elements (e.g., extended loops) increase, and *de novo* structures fail to sample twisted long arm conformations and H3-E3 connections paired to them.

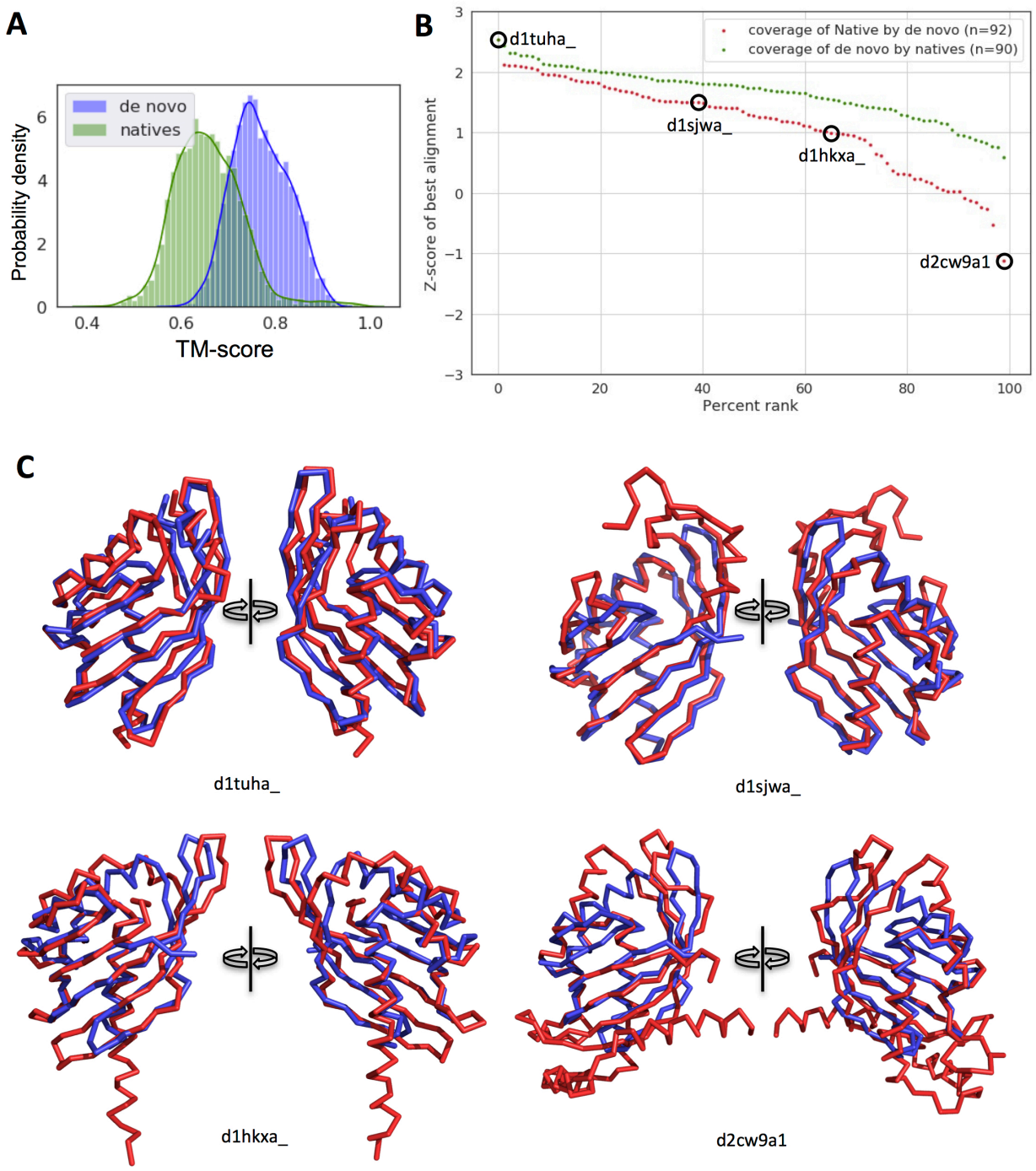


Figure 4.3: Structural comparison between native and *de novo* NTF2-like proteins generated by the first version of the generative algorithm A. Normalized histogram of all *vs*. all TM-scores within native and *de novo* sets. **B.** Z-score ranking of the best matching de novo designs to natives and vice-versa. **C.** Backbone ribbon representations of aligned structures for different native NTF2-like domains and the best matching *de novo* models.

Despite natives covering most of the *de novo* space, roughly 10% of the best matches to the *de novo* set have a z-score below 1. Interestingly, this group contains all the models of the Mk1.TP subfamily, indicating this subfamily is a significant departure from the typical NTF2-like domain. Mk1.TP is the only *de novo* subfamily where we purposefully occluded the frontal opening with a long C-terminal helix that runs from the C-terminus of E6 to the N-terminus of H3. We adopted this strategy as a simplification over how native structures close the frontal opening: an elongated loop in the frontal hairpin and a short C-terminal helix. We should clarify that through manual inspection we found an NTF2-like domain structure (PDB ID 3MSO), not recorded in the SCOPe database, where a long C-terminal helix has a similar position as in Mk1.TP designs, also occluding the frontal opening, and with an opening between the H1-H2 helix connection and H3. Although the C-terminal helix of 3MSO is similar to that in Mk1.TP structures, the TM-score between 3MSO and the lowest ranking Mk1.TP *de novo* protein was marginally better than the previously found best match, due to significant difference in the placement of the N-terminal helices.

The diversity analysis done so far reveals the general coverage and novelty of the *de novo* set. We can gain a more nuanced understanding of the diversity structure of the *de novo* set by constructing a dendrogram using 1-TM-score as pairwise distances. Figure 4.4 shows a dendrogram containing both sets of *de novo* and native NTF2-like domains, and a heat map depicting pairwise distances (See Methods 4.5.4). Interestingly, most native and *de novo* structures cluster separately, with one small isolated group of natives in a branch of *de novo* structures near the center of the dendrogram. As expected by the TM-score distributions (figure 4.3 A), *de novo* models cluster more tightly than natives, but these tight *de novo* clusters are not completely separated from natives, they are rather inserted in them. As expected from the comparison method and backbone building protocol, *de novo* models with similar sheets cluster together (i.e., Mk1, 4B and Mk2), with the presence of the C-terminal helix being less important for clustering.

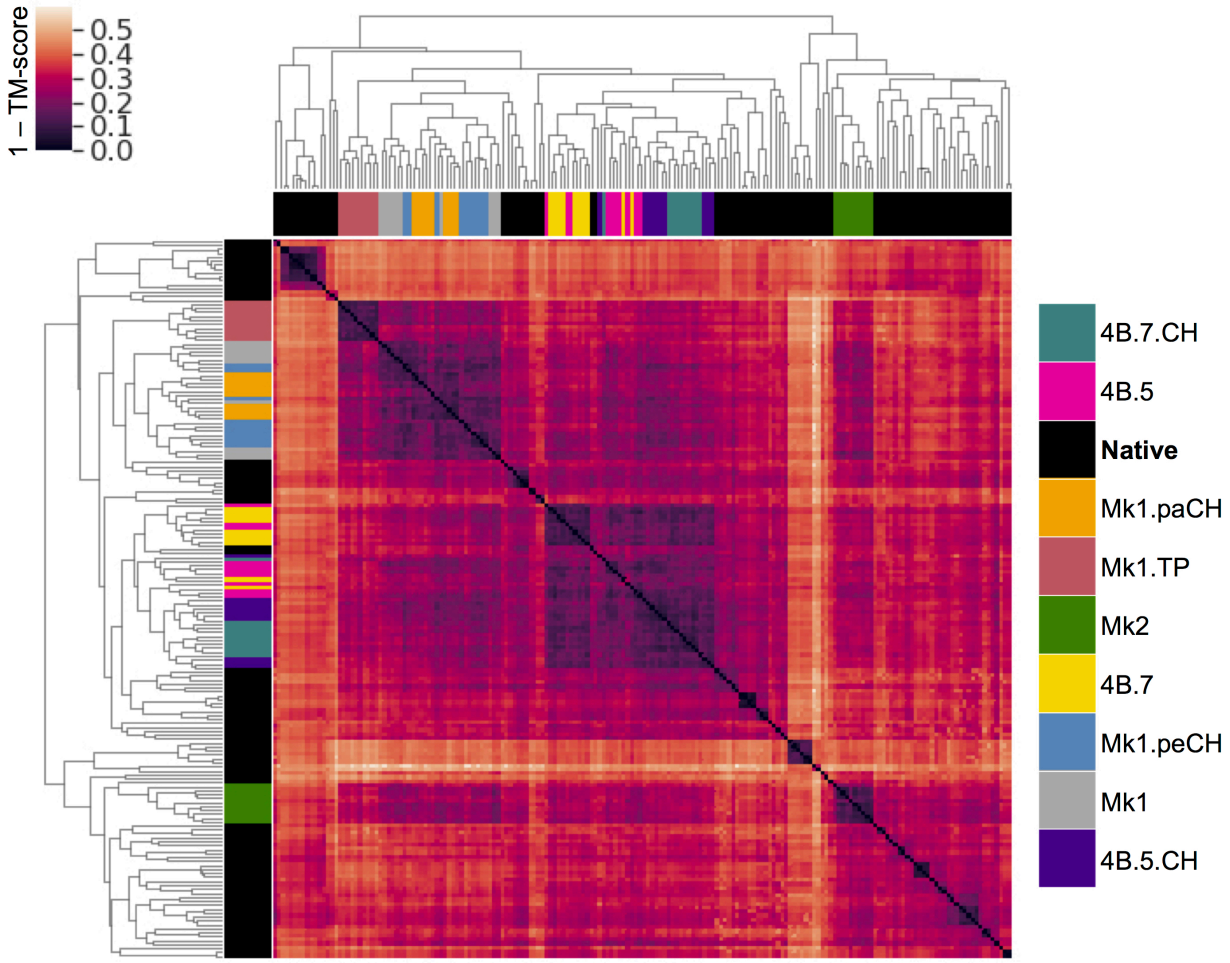


Figure 4.4: **A heat map of clustered NTF2-like domains**, along with the dendrogram (two identical ones on upper and left sides of the heat map) resulting from clustering. The distance metric used for clustering, 1-TMscore (upper left color bar), is depicted in the heat map. Different protein *de novo* subfamilies are represented as different colored bars at the tips of the dendrogram (color legends on the bar on the right).

The first version of the generative algorithm generates a limited number of NTF2-like subfamilies that sample different values of a set of NTF2-like superfamily fold parameters (Table 4.1). Despite their limited variability, they can accommodate a range of pocket sizes comparable to native NTF-like domains, at least *in silico*. Since pocket size is likely not the only parameter critical for active/binding site design, we expect further efforts to increase diversity to be necessary. Analysis of the diversity generated by the first version of the generative algorithm shows that *de novo* designs sample a few regions of the native space with high density, with limited sampling outside of that. Further efforts to increase diversity will include

addressing the shortcomings highlighted by the alignments shown in figure 4.3 C: more variable sheet structures, H3-E3 connections and hairpin extension.

The following sections focus on determining what fraction of the proteins generated by the algorithm actually fold to the designed structure, and to what extent designing pockets in them is detrimental for stability. We then use that information to increase the number and diversity of stable proteins generated by the algorithm.

4.3.2 High-throughput screening of de novo NTF2-like proteins generated by the first version of the algorithm

In Chapter 3 we showed the applicability of the proposed curved sheet design principles by designing and characterizing a few examples for each subfamily. In the same spirit, we sought to show the generative algorithm produces sequences that fold in the designed structure by measuring stability for thousands of them using a high-throughput assay based on protease resistance (6): Briefly, genes encoding for thousands of different *de novo* NTF2 sequences are transformed in yeast for surface display in a one-pot fashion. Different aliquots of this yeast culture are then subject to increasing concentrations of proteases, cells still displaying full proteins after this treatment are isolated by Fluorescence Activated Cell Sorting (FACS). Deep-sequencing of the sorted populations reveals which sequences are protease resistant and to what degree, providing an estimate for folding free energy. The metric reported by this assay is the stability score, an estimate of how much protease is necessary to degrade a protein over that expected if the protein was completely unfolded. A stability score of 0 indicates that half of the protein is degraded by the same amount of protease as expected if it was unfolded, i.e., it is probably unfolded. A stability score of 1 indicates that 10X more protease is required to degrade half of the protein than expected if it was unfolded.

We prepared a library of designs for screening by generating 3000-7000 proteins for each of the nine subfamilies, and selecting subset of 225-600 for each of them (See table 4.1 and Methods 4.5.5). In addition to these, we redesigned 492 models from the Mk1.PeCH subfamily to have a binding site for cortisol (See figure 4.2, Methods 4.5.2). As a negative control, we included 2570 design sequences

where we scrambled the amino-acid identities keeping glycine, proline in place, but exchanging positions within hydrophobic and hydrophilic groups. These scrambles sequences were proportionally derived from all subsets, and their purpose is to establish the baseline for distinguishing folded and unfolded sequences, as done in (6). To generate the genes encoding for these proteins, we followed the protocol described in (7). Gene fragment design is described in Methods section 4.5.6.

Of the 7706 candidate sequences, 5403 were observed at least once in the naïve expressing sorted population, by deep sequencing (as described in (6)). From those 5403, we obtained reliable stability scores for 3728 (Stability score CI width < 0.5, stability scores obtained as described in, and software provided with (6)), 2708 designs and 1020 scrambles, which we focus on for analysis. The reported stability score is the base 10 logarithm of the fold concentration of protease required to degrade half of the proteins (EC50) in comparison to what would be expected if that protein was unfolded, i.e., a stability score value of 1 means that it takes 10 times the protease concentration to degrade a protein than what would be expected if that protein was unfolded. As described in (6), the unfolded state model was derived from thousands of scrambled sequences and is mainly based on the cut-sites found in a given sequence. The distributions of stability scores for designs and scrambles show clear differences, with the scrambles having values around 0, and designs having what seems like a bimodal distribution, with maxima around 0 and 1 (Figure 4.5).

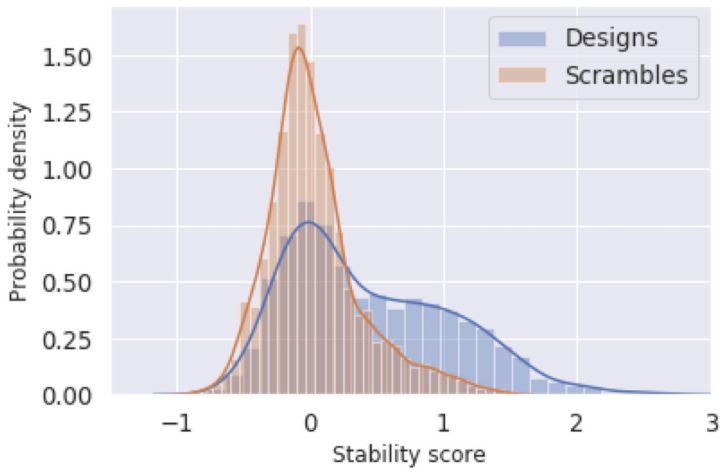


Figure 4.5: Stability score distributions for designs (blue) and scrambles (orange).

The subpopulation of designs with stability scores above 1 (578 designs, 21%) is composed of representatives of all subfamilies, and only 2% of scrambles (Figure 4.6). The proportion of stable designs in each subfamily is variable, with Mk1.* designs having the highest percentages of stable sequences, and 4B.5.* having the lowest. Despite these differences, all designs subfamilies have substantially more representatives with stability score above 1 than scrambles.

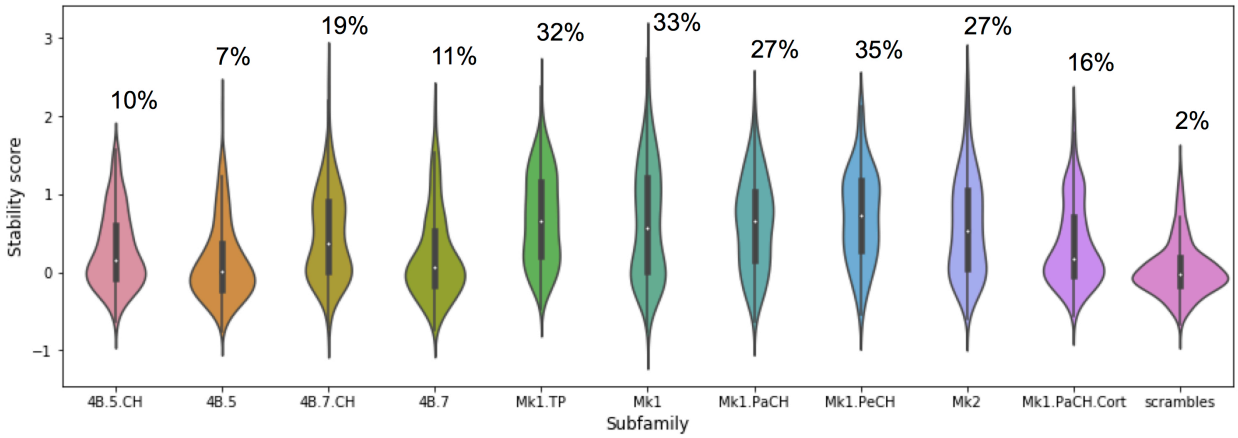


Figure 4.6: Violin plot of stability scores separated by subfamily. Percentages of members with stability score >1 are displayed above each subfamily distribution.

Looking to understand the determinants of stability in *de novo* NTF2-like proteins, we applied LASSO logistic regression (8, 9) to our dataset, searching for features (Tables in section 4.5.7) that predict weather an NTF2 model generated by our algorithm would have a stability score above (stable) or below (unstable) 1. Briefly, a logistic regression model predicts the probability of a binary outcome using a logistic function that depends on a weighted summation of features. By sampling a series of L1 regularization values, we obtain models with varying degrees of parsimony, and for each of those L1 values we also generate different random partitions of our dataset (See methods 4.5.8). This way, for each L1 value we obtain models with a spread on accuracy, which we use for selecting an L1 regularization value that maximizes accuracy and minimizes complexity - i.e., the number of features with weight different from 0. As seen in figure 4.7 A, no substantial accuracy is gained for L1 values above 0.02, we therefore focus on those models for our analysis, as we assume they contain the least number of features that can be used to produce the most accurate predictions. The simplest measure of the

importance of each feature is its assigned coefficient. Figure 4.7 B shows the average and standard deviation of t all assigned coefficients for each of the 16 features with the highest weights. The simplest model with L1=0.02 has 12 features that have low correlation values to each other, except for 3 of them, that are strongly linked to the number of hydrophobic amino-acids (Figure 4.7 C).

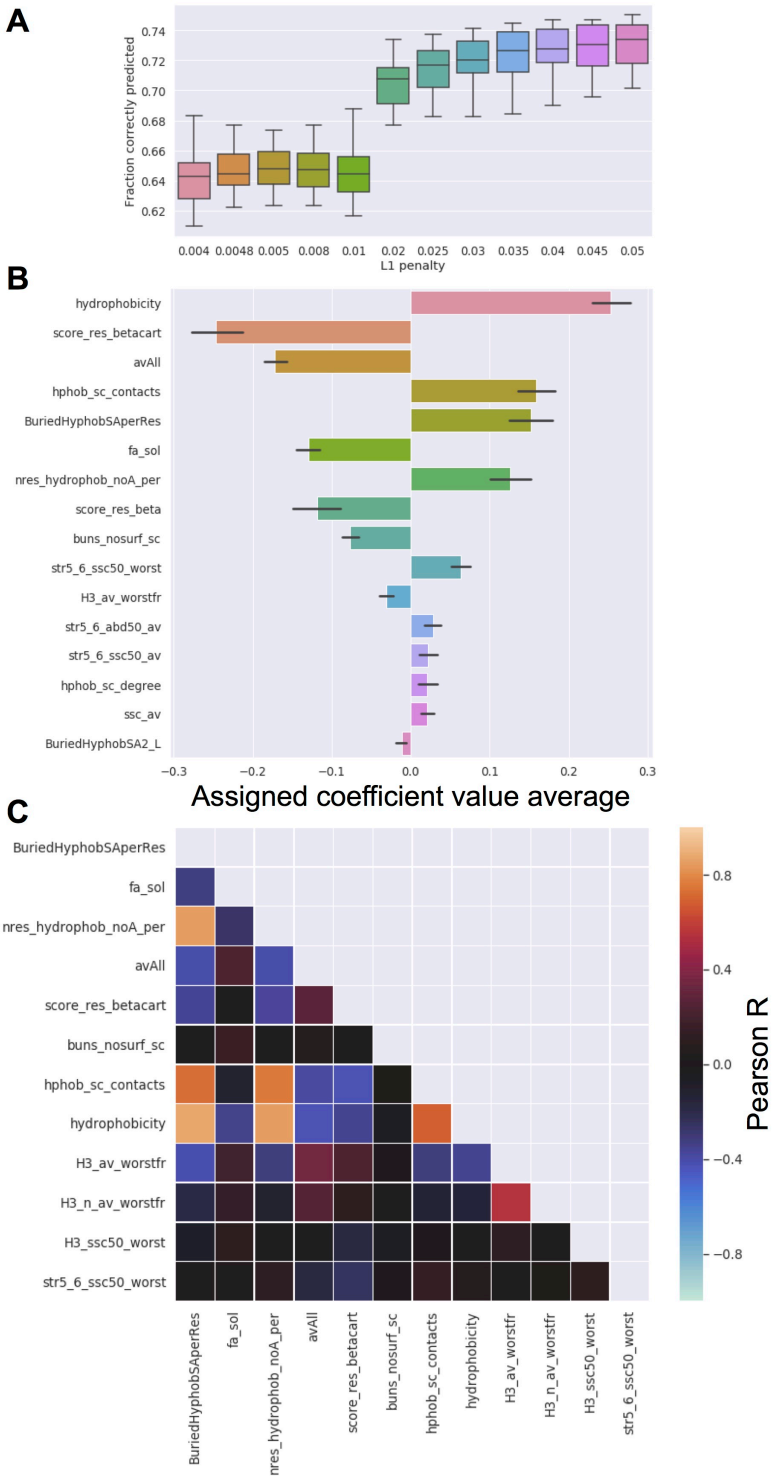


Figure 4.7: Logistic regression on designs generated by the first version of the generative algorithm. **A**. Boxplot of model accuracy on test set for different values of L1 penalty. Each box represents 40 different random partitions of the dataset, with one third of it as test set in each case. **B**. Absolute weights of the 15 features with the highest average weights in the 40 dataset partitions. **C**. Correlation matrix for features in the simplest model out of the 40.

The three main contributors to the summation of terms that predicts stability score above 1, *hydrophobicity*, *score_res_betacart* and *avAll*, are hydrophobicity, Rosetta score, and local degree of disagreement between sequence and structure, respectively. We should note that the sign of the terms indicates whether predicted stability increases for lower (negative coefficient) or higher (positive coefficient) values of the feature. In our case, higher hydrophobicity, lower (better) Rosetta score, and lower disagreement (RMSD) between local sequence and structure, are predictive of stability score above 1. Other features that correlate with stability are related tertiary structure motifs (TERMs, (10)), sequence-structure agreement in specific stretches (*H3_av_worstfr*), and buried side-chain polar atoms with no hydrogen bonds (*buns_nosurf_sc*). Predictive TERM-related features relate to the whole protein (*ssc_av*), or to the short arm hairpin specifically (*str5_6_ssc50_worst, str5_6_ssc50_av* and *str5_6_abd50_av*).

Because we used a range of values for the L1 regularization parameter and selected the best models with the lowest (most restrictive) regularization value, we expect features in any specific models to show little or no correlation, as it is expected that features carrying the same information to be represented by a single one within their group (implicit feature selection by L1 regularization). Figure 4.7 C shows the correlation (Pearson R) between variables in the simplest (12 features with non-zero coefficients) model with L1=0.02. As expected, most correlation values are around 0.0, with few exceptions. The features showing high correlation can be grouped under those related to hydrophobicity (*hydrophobicity, nres_hydrophob_noA_per*, *hphob_sc_contacts and BuriedHyphopSAperRes*). Despite being correlated, these features can carry significantly different information. Take hydrophobicity and the residue-normalized buried hydrophobic surface area (*BuriedHyphobSAperRes*) as examples: the latter is an organic-solubility-weighted sum of the different amino-acids types in the sequence, it uses no structural information, while *BuriedHyphobSAperRes* is strictly dependent on the 3D structure of the protein. Although correlated, both quantities can carry different information.

Given the degree of structural similarity among the tested proteins, we were able to do a per-position enrichment analysis (See methods 4.5.9 for details), where we select a subset of structurally homologous positions (Figure 4.8 A) in all tested models, and calculated the enrichment values of different amino-acid identities in stable (stability score>1) designs (Figure 4.8 B).

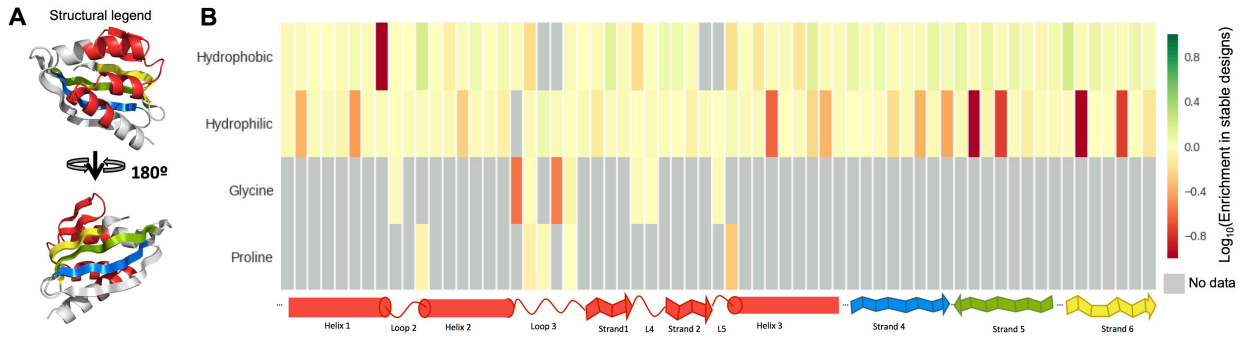


Figure 4.8: Heat map of enrichment or depletion by position. **A**. Positions structurally homologous in all ordered models, selected for comparison. The selection is depicted over an Mk1 model. Colors depict stretches continuous in sequence. **B**. Heat map of enrichment or depletion by position, with amino-acid types grouped in four categories: hydrophobic (AFILMVWY), hydrophilic (DEHKNQRST), proline and glycine. Upward-pointing pleating in strands points towards the core of the protein. Yellow cells indicate no enrichment or depletion, or a non-significant difference.

Figure 4.8 B shows depletion of hydrophilic residues in positions the point towards the core. This is especially evident in strands, where the expected hydrophobic/hydrophilic alternating patters follows the alternation of C_{alpha}-C_{beta} vectors, even with a break at the bulge in strand 6. Interestingly, hydrophobic amino-acids are enriched in multiple positions at outward-facing positions of strands, potentially indicating some hydrophobic interactions in the middle of exposed flat sheets can contribute to stability (11). Depletion of glycine in certain loop positions could indicate preference to other specific identities, observed in loops with similar structure. Depletion of proline on the N-terminus of H3 suggest the hydrogen-bond network involving this residue, which includes its N backbone atom, is necessary for proper folding, as identified on Chapter 2. Overall, the enrichment of amino-acid identities on stable designs is in agreement with the models, indicating that stable proteins fold into structures closely resembling, if not identical, to those modeled.

In the following section we show the results of the biochemical characterization of a subset of designs, the crystal structures we obtained, and how these could guide the refinement of the generative algorithm.

4.3.3 Structural validation of proteins generated by the first version of the algorithm

To verify the reliability of the stability score derived from the protease assay as a measure of structural stability or folding free energy, we selected a relatively small number of designs to be characterized experimentally (Table 4.2). The proteins we selected have stability score above 1, and sample a range of pocket volumes and subfamilies.

Design name	Subfamily	Pocket volume (Å^3)	Stability score
BB45nHm0313	4B.5	351	2.17
BBM1TPm0012	Mk1.TP	63	2.38
BBM2nHm0111	Mk2	58	2.20
BBM2nHm0481	Mk2	149	2.51
BBM2nHm0589	Mk2	101	2.51
BBMHCYm0000*	Mk1.PaCH	159/494	1.76
BBMHCYm0098*	Mk1.PaCH	54/583	1.27
BBMHCYm0099*	Mk1.PaCH	385/733	1.15
BBMHCYm0118*	Mk1.PaCH	295/814	1.22
BBMHCYm0142*	Mk1.PaCH	279/756	1.98
BBMHCYm0257*	Mk1.PaCH	129/690	1.09
BB45nHm0217	4B.5	305	2.01
BB45nHm0313	4B.5	351	2.17
BB45nHm0520	4B.5	275	1.61
BB47nHm0104	4B.7	122	2.08
BB47nHm0234	4B.7	326	1.96
BB47nHm0512	4B.7	253	2.07

Table 4.2: Designs selected for biochemical characterization. *: Denotes models where a binding site for cortisol was designed. For cortisol binding designs, pocket volume is reported for the Rosetta-relaxed *apo* model (left) and the *holo* designed model (right).

Table 4.3 is a summary of the experimental results for the 17 selected proteins (See Methods 4.5.10). More than half of the tested proteins are not soluble when expressed in *E. coli*, or form soluble aggregates. It is possible this is an artifact of high intracellular expression, which would not affect the system used for screening in *S. cerevisiae*. We did not investigate whether denaturation and refolding would recover the insoluble or aggregated samples. Despite this, most cases where soluble protein can be produced in *E. coli*, have a circular dichroism spectrum consistent with regular secondary structure, and titration with guanidine hydrochloride shows unfolding in a two-state transition, a hallmark of proteins that fold into a single structure (Figure 4.9). Table 4.4 shows the biochemical parameters obtained by fitting a sigmoid curve to the guanidine hydrochloride titration data (See Methods 4.5.10). For all designs except BBM2nHm0481, the observed m-values, a measure of buried surface area upon folding, are within the expected range for the buried surface area calculated from the models (12). Differently from

previous reports (6), the stability score derived from high-throughput experiments seems to have no correlation with $\Delta G_{unfolding}$ (Figure 4.10), although the low number of samples and the narrow range of stability scores sampled makes it hard to derive any conclusions.

Design Name	Soluble expression	Within expected SEC EV	Quaternary structure	Folded protein	Denaturation curve
BB45nHm0313	Yes	Yes	Tetramer	-	-
BBM1TPm0012	Yes	No	-	-	-
BBM2nHm0111	Yes	Yes	Monomer	Yes	Two-state
BBM2nHm0481	Yes	Yes	Monomer	Yes	Two-state
BBM2nHm0589	Yes	Yes	Monomer	Yes	Two-state
BBMHCYm0000	Yes	Yes	Monomer	Yes	Two-state
BBMHCYm0098	Yes	Yes	Monomer	Yes	Two-state
BBMHCYm0099	No	-	-	-	-
BBMHCYm0118	No	-	-	-	-
BBMHCYm0142	Yes	Yes	Monomer	Yes	Two-state
BBMHCYm0257	Low	No	-	-	-
BB45nHm0217	Yes	No	-	-	-
BB45nHm0313	Yes	No	-	-	-
BB45nHm0520	No	-	-	-	-
BB47nHm0104	Yes	Yes	-	Yes	Two-state
BB47nHm0234	Yes	No	-	-	-
BB47nHm0512	Yes	Yes	-	Yes	Gradual

Table 4.3: Summary of experimental results on 17 proteins selected for biochemical characterization.

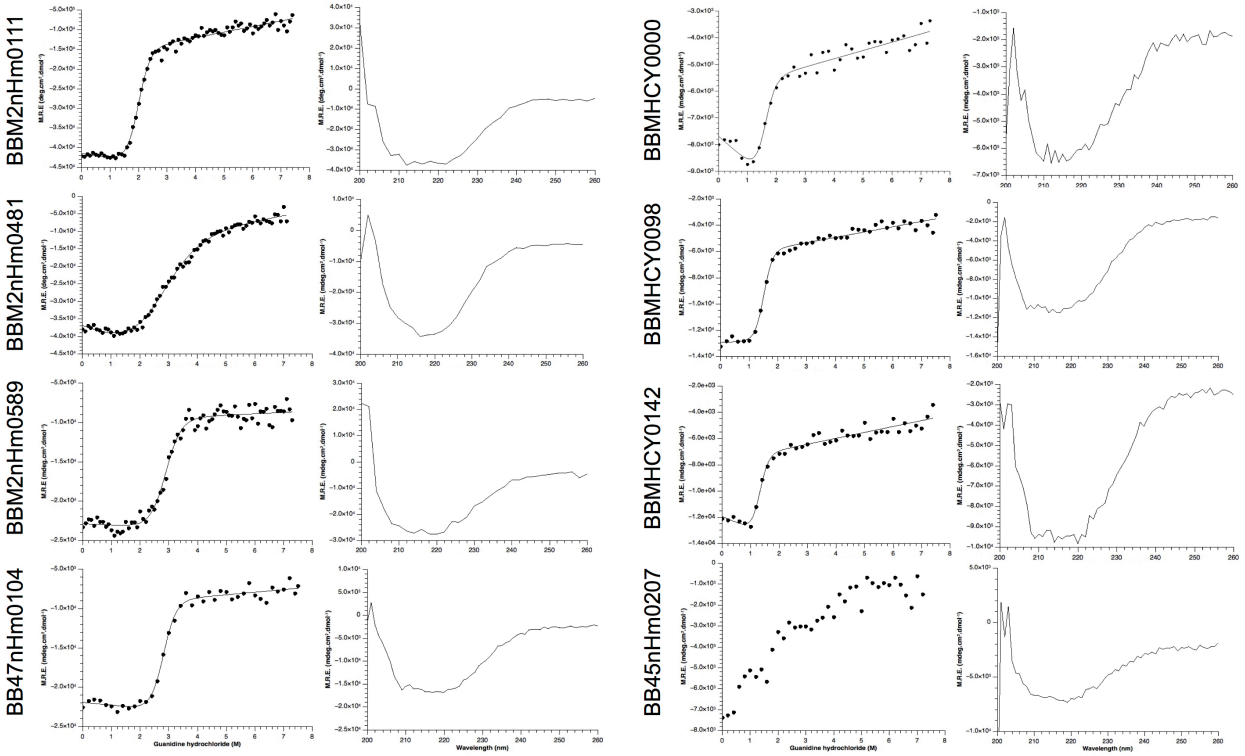


Figure 4.9: Circular dichroism spectra (right) and denaturation curves (left) for soluble, monomeric designs. Curve fits for chemical denaturation experiments are shown in solid lines.

Design Name	$\Delta G_{folding}$ (kcal/mol)	Stability score	Experimental m-value ($kcal.mol^{-1}.M^{-1}$)	Expected m-value ($kcal.mol^{-1}.M^{-1}$)
BBM2nHm0111	-5.71	2.2	2.87	4.86
BBM2nHm0481	-2.46	2.51	0.92	4.98
BBM2nHm0589	-6.42	2.51	2.2	4.92
BBMHCYm0000	-5.24	1.76	3.29	5.00
BBMHCYm0098	-5.82	1.27	3.86	5.03
BBMHCYm0142	-5.13	1.98	3.89	4.99
BB47nHm0104	-7.52	2.08	2.67	4.79

Table 4.4: **Thermodynamic parameters obtained from fitting guanidine hydrochloride denaturation curves**. Expected m-values are calculated using the buried surface area based on protein models and the equations described in (12).

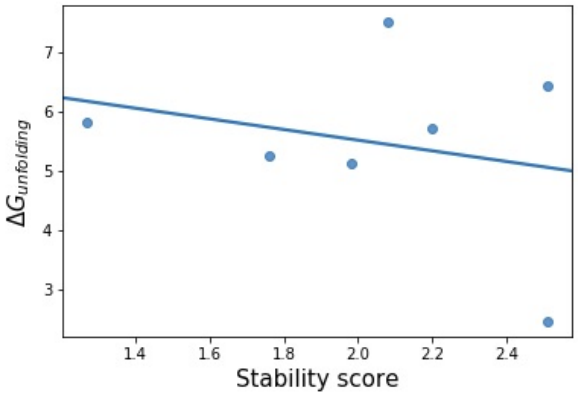


Figure 4.10: Free energy of unfolding as a function of stability score for 7 stable designs. A solid line shows the linear fit to the data. No significant correlation (Pearson R: -0.25, p-value: 0.58) is observed.

We tested cortisol binding in the proteins designed to bind it, by isothermal titration calorimetry, but we observed no signal. Figure 4.11 A-C shows isothermal titration calorimetry (Methods 4.5.11) results for BBMHCYm0000 (H1), BBMHCYm0098 (H2) and BBMHCYm0142 (H5), respectively, as well as a positive control (anti-cortisol monoclonal antibody - mAb), and a negative control (BBM2nHm0481 – M22).

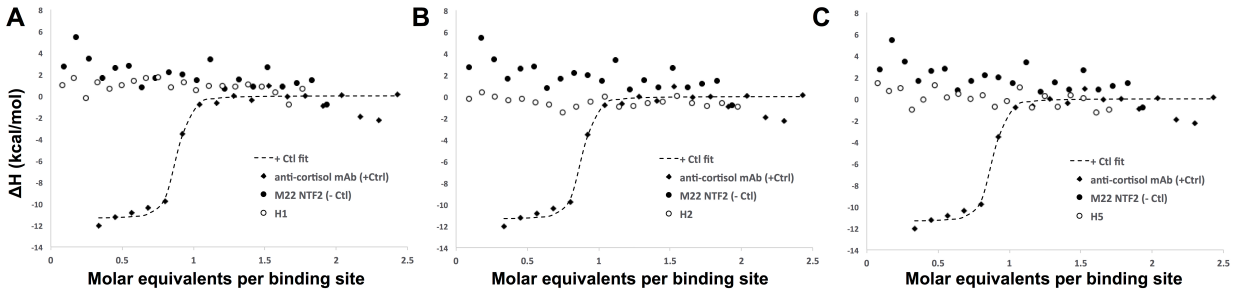


Figure 4.11: Isothermal titration calorimetry results for titration of BBMHCYm0000, BBMHCYm0098, and BBMHCYm0142 with cortisol in aqueous solution.

Although less than half of the proteins we selected for biochemical characterization displayed behaviors consistent with the models (globular, monomeric proteins with structured hydrophobic cores), we obtained folded designs for the Mk2 subfamily, which we had not been able to successfully design before. We also obtained folded designs for proteins with large binding pockets, without resorting to additional stabilizing strategies described in Chapter 3. To confirm the stable designs are folded in the modeled conformations, we pursued crystals structures for all of them.

We obtained crystals structures for BBM2nHm0481 (1.62 Å resolution, figure 4.12) and BBM2nHm0589 (1.38 Å resolution, figure 4.13). See Methods 4.5.12 for data collection and analysis metrics.

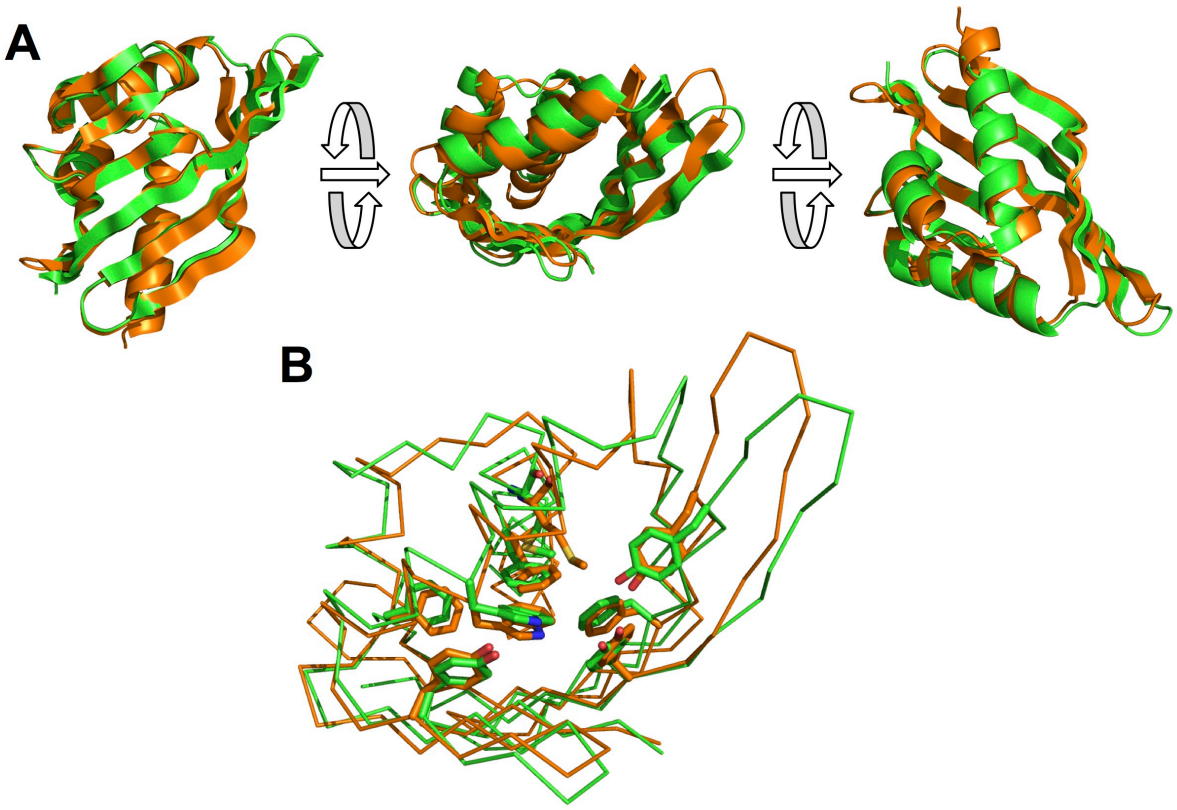


Figure 4.12. Crystal structure of BBM2nHm0481 A: Overlay of backbone atoms from BBM2nHm0481 model (orange) and crystal structure chain A (green) from three different perspectives. **B**: Overlay of model (orange) and crystal structure (green) with backbone shown as ribbon and large core side-chains highlighted as sticks.

The structure of BBM2nHm0481 is in close agreement with the model (Figure 2.12 A), with a C_{alpha} RMSD of 1.53Å (the C_{alpha} RMSD between the two chains forming the asymmetric unit is 0.78Å). With the exception of the connection between strand 5 and 6, all the secondary structure elements distal to the "mouth" overlay almost perfectly. The main discrepancies between the model and the structure are in the area surrounding the mouth, i.e., the distal end of the long arm and the C-terminus of helix 3. Discrepancies in this area are more likely since the average number of contacts is lower in this region. Figure 4.12 shows the close agreement in core rotamer conformations between structure and model. Core rotamer recapitulation in the crystal structure of *de novo* designed proteins is key for functional design, as it demonstrates a degree of structural control fine enough to accurately design interactions with small molecules and substrates. Overall, the designed model of BBM2nHm0481 is successfully recapitulated in its crystal structure.

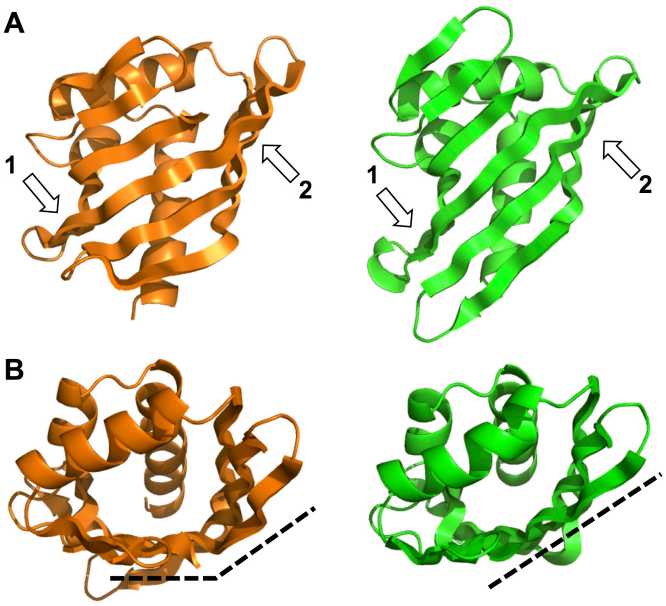


Figure 4.13 A: **Comparison between BBM2nHm0589 model (orange) and crystal structure (green).** Arrows point at the strand register shift locations that differ between the model and crystal structure. 1 and 2 indicate homologous sites in both structures where the main differences occur. **B**. Model and crystal structure of BBM2nHm0589 in perspectives that highlight the difference in the sheet curvature.

The crystal structure of BBM2nHm0589 shows significant differences from the model (Figure 4.13). Mainly, the register shift between strand 5 and 6 shortens from 6 to 4 residues (Figure 4.13 A) and the whole sheet flattens (Figure 4.13 B). This change in allowed by strand 5 shortening by 2 residues, which become part of the loop that connects strand 5 to 6. Overall, both arms keep their initial length, but the flat section (the base) in the middle of the strand disappears and the loop between strand 5 and 6 becomes one residue longer.

We hypothesized the angle between the base and the long arm (Figure 4.14 B, left) is energetically unfavorable, and the flatter, more favorable, sheet conformation observed (Figure 4.14 B, right) requires the register shift to change for helix 3 to successfully connect both extremes of it. Based on these assumptions, we proposed two sets of mutations: One that would stabilize the designed conformation of the short arm, preventing the register shift by forcing strand 5 to have the designed length (D101K, L106W, Figure 4.15 A, left); another one that would stabilize the sharp bending on strand 5 at the base of the long arm, and would therefore prevent strand 6 from shifting (I64F, A80G, T94P, Figure 4.15 A, right).

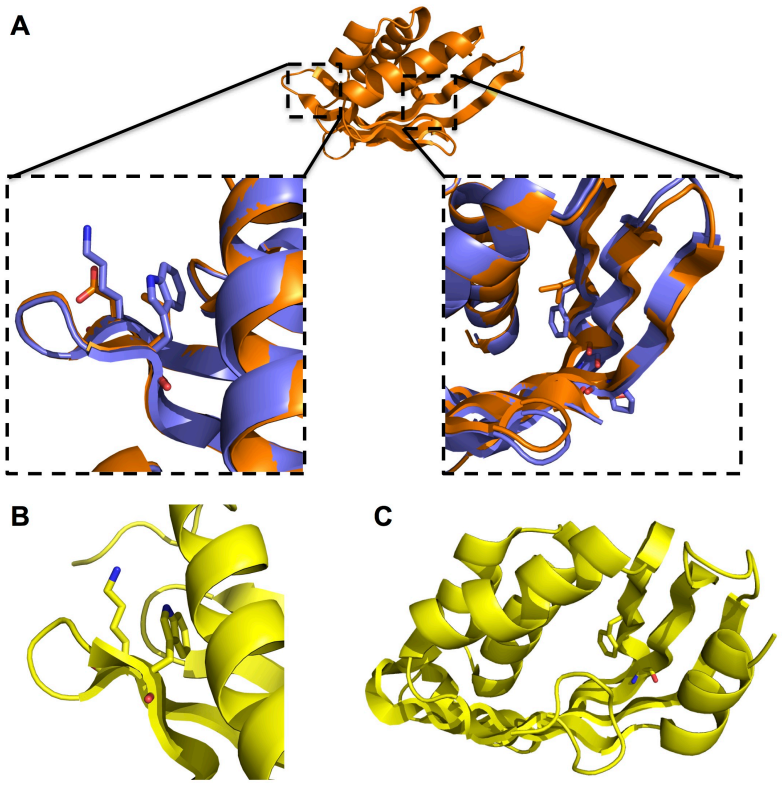


Figure 4.14 A: Comparison between original BBM2nHm0589 model (orange) and models with both sets of mutations (purple). Top: cartoon representation of original model. Left: overlay of set 1 mutations and original model sequence (mutated side-chains are represented as sticks). Right: overlay of set 2 mutations and original model sequence (mutated side-chains are represented as sticks). **B**: Close-up of NTF2 PDB 3DUK, with short arm displaying interactions homologous to those designed in set 1 mutations. **C**: NTF2 PDB 3EC9, side chains involved in similar interactions to set 2 mutations are presented as sticks.

Both suggested structural motifs have been observed in native NTF2-like domains (Figures 4.15 B and C), and the glycine-phenylalanine pair is reminiscent of interactions previously reported to be necessary for constructing beta-barrels (13). Modeling the proposed mutations in the context of the original model shows they are compatible with the rest of the structure and side chains, and, for set 2 mutations in particular, clearly enhance sheet curvature (Figure 4.14 A, left inset).

We tested both sets of mutations independently and combined. All three variants were soluble and monomeric when expressed in *E. coli*, and had similar stability to the parent design (Figure 4.16).

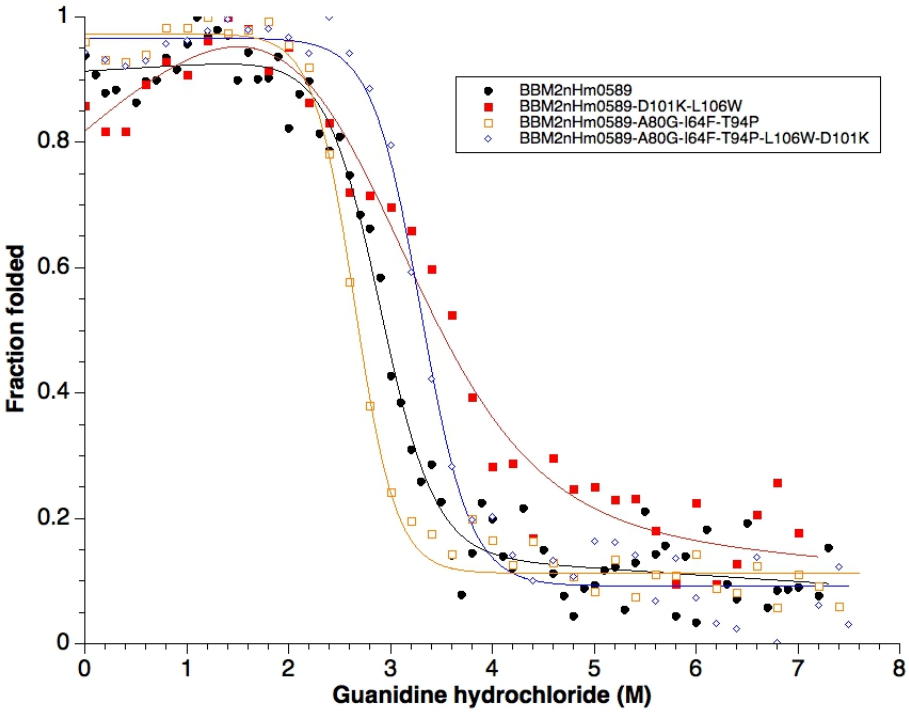


Figure 4.15 Denaturation curves in guanidine hydrochloride for BBM2nHm0589 mutants. Denaturation curves in guanidine hydrochloride for BBM2nHm0589 (solid black circles) and mutants D101K L106W (solid red squared), A80G I64F T94P (open orange squared) and their combination (D101K L106W A80G I64F T94P – open blue diamonds). Fits to data are plotted as solid lines in the corresponding colors.

We obtained a crystal structure for the variant combining set 1 and 2 mutations (Figure 4.16). The structure of the BBM2nHm0589 5-fold mutant is in close agreement with the model (C_{alpha} RMSD between model and chain A: 1.55Å, C_{alpha} RMSD between chains in asymmetric unit: 0.45Å). Differently from the parent design, the 5-fold mutant structure displays the expected strand register shift, as well as expected strand and loop lengths (Figure 4.16 A). Furthermore, the rotamers of mutated side-chains are in the designed configuration, suggesting the mutations have a role in the structural rearrangement. As in the structure of BBM2nHm0481, most core side chains in the structure are in agreement with the model (Figure 4.16 C), except for Y113, which is flipped outwards almost 180º in the crystal structure, enlarging the pocket (Figure 4.16 B). Despite minor core side-chain conformational differences, the crystal structure recapitulates the large pocket displayed by the model.

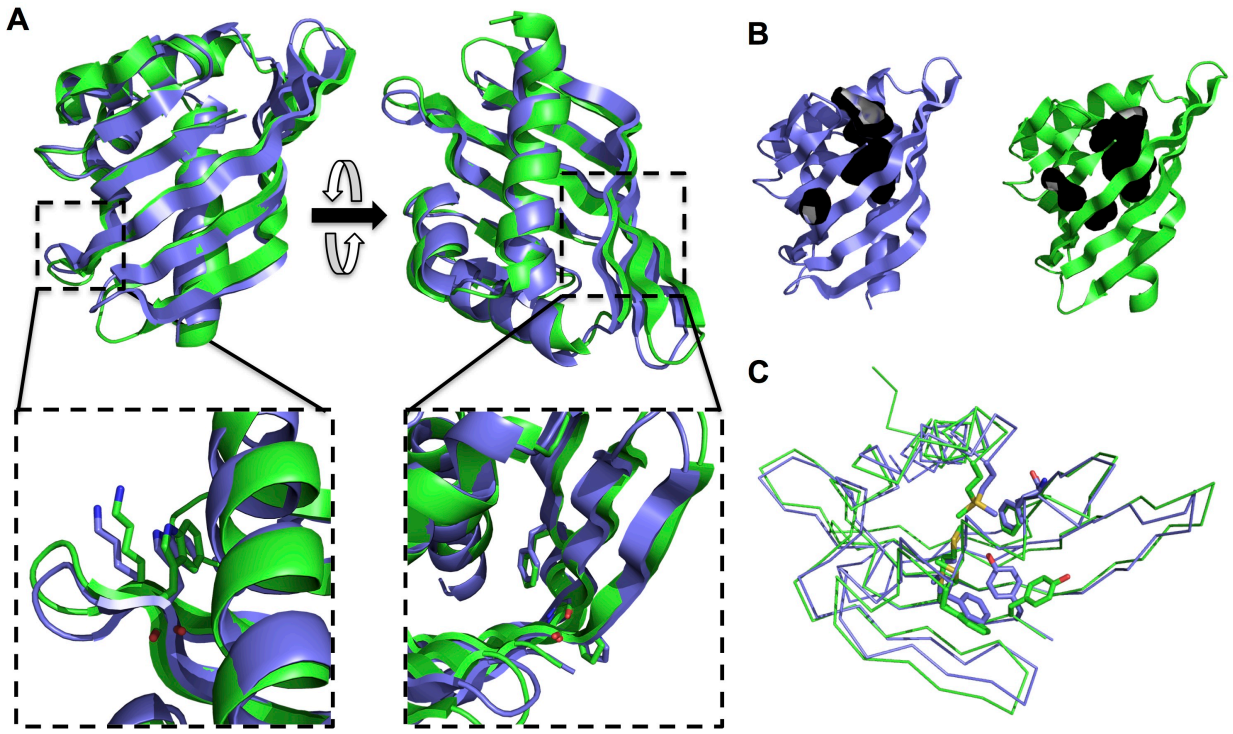


Figure 4.16 Crystal structure of BBM2nH0589 5-fold mutant **A**. cartoon representation of design model (purple) and structure (green) of BBM2nHm0589 5-fold mutant, showing close agreement in backbone structure (top, two perspectives), and the side-chain configuration of the mutated residues (bottom, set 1: left, set 2: right). **B**. Comparison of pocket (black surface) in design model and structure. **C**. Overlay of model and structure, with backbone represented as ribbon, and large core side-chains as sticks.

The successful redesign of BBM2nHm0589 provides critical information for design of NTF2-like proteins, indicating certain sheet configurations require specific sequence features to be stabilized. This information, along with the features used by the logistic regression model (Section 4.3.2), shows a clear avenue of improvement in success rate and structural diversity of *de novo* NTF2-like proteins. In the following sections, we show how combining these new insights with a more general backbone generation algorithm unlock additional orders of magnitude of NTF2 subfamily diversity. We also explore functionalization of stable proteins detected by high-throughput screening and design new ones using the information gained from the large-scale analysis.

4.3.4 Reimplementation of the generative algorithm for generalization and incorporation of lessons from the first version

The first version of the generative NTF2 algorithm is a bundle of similar backbone-generation algorithms with specific parameters for each case. Differently from a general algorithm, the input parameters of these backbone-generation algorithms cannot be exchanged. A general NTF2 generative algorithm should be able to take a large space of parameters as inputs, detect incompatibilities and produce output accordingly. In this section we go over the process of unifying and generalizing the NTF2 backbone generative algorithm with a new implementation.

Similarly to the first version of the generative NTF2 algorithm, the new version is based on biased fragment assembly (See methods section of Chapter 3). The code for the new implementation and technical details can be found in the method section of this chapter (4.5.13).

The previous version of the generative algorithm splits the backbone generation process in at least five stages (See Chapter 3, Section 3.5.2), with the initial three devoted to building the main sheet. The bend and twist of the sheet are independently constrained in the first stage, on strands 4 and 5, with the following two stages placing the flanking bulged strands, 3 and 6, that are in turn just constrained to the correct hydrogen bond pairing with the initial two strands. In this context, different base width and arm lengths are encoded by at least six variables: relative bulge placement, strand lengths (one per strand) and register shift between strands. Moreover, there is no way to independently adjust base width and arm lengths. Additional bulges on the long arm are not considered, and arbitrary combinations of parameters can be selected without checking for compatibility or consideration of later steps.

In the reimplementation we sought to add a layer of control and simplification between the user and the lowest-level structural parameters: The main sheet is constructed in a single step, and input parameters are base width, arm lengths, degree of curvature between the long arm and the base and, if an additional bulge is added to the long arm, the curvature around this second bulge. Additional parameters control sheet features for specific cases, such as extensions of strand 6 and 4, that are only compatible with certain combinations of arm length, curvature and width length. Table 4.5 is a comprehensive list of sheet parameters and their units, and figure 4.17 shows a few examples of parameter combinations.

Parameter (short name)	Units (Allowed values)	Explanation
Base width (base_width)	Residues (3,5)	Number of residues between relative positions of main bulges on strand 3 and 6, including the B

		ABEGO bulge residue.
Long arm length (long_arm_l)	Pairs of residues (2,3,4)	Number of residue pairs between the main bulge in strand 3 and the N-terminus of strand 3, does not take into account additional bulge residues when strand 3 has additional bulges.
Short arm length (short_arm_l)	Pairs of residues (1,2)	Number of residue pairs between the main bulge in strand 6 and the C-terminus of strand 6.
Additional bulge on the long arm (Second_bulge_E3)	Boolean (True,False)	Presence or not of a second bulge on strand 3.
Placement of second bulge on strand 3 (Second_b_place)	Pairs of residues (1,2,null)	Position of the second bulge on strand 3, relative to the main bulge, towards the N-terminus of strand 3. If Second_bulge_E3 is False, then Second_b_place is null.
Degree of curvature angle between long arm and base (E3_MainBulgeCurve)	(160-10*X)º, where X is one of the allowed values (1,2,3)	Average value for harmonic angle constraint centered in the strand 4 residue that is paired to the main strand 3 bulge.
Degree of curvature angle centered at the E3 second bulge (E3_SecBulgeCurve)	Null or (160-10*X)º, where X is one of the allowed values (1,2,3,null)	Average value for harmonic angle constraint centered in the strand 4 residue that is paired to the second strand 3 bulge. If Second_bulge_E3 is False, then E3_SecBulgeCurve is null.
Extension of strand 6 (ExtendedE6)	Boolean (True,False)	If True, strand 6 is extended by 2 residues on its C-terminus. This is only compatible with a low degree of strand curvature on the main bulge.
Extension of strand 4 (ExtendedE4)	Boolean (True,False)	If True, strand 4 is extended by 2 residues on its N-terminus. This is only compatible with a short arm of length 2.
Small degree of curvature on the long arm (CurvedLongArm)	Boolean (True,False)	If True, impose a 150º angle constraint on the central residues of the long arm on strands 3 and 4 to impart a small curvature in the absence of a second bulge on strand 3.

Table 4.5: Comprehensive list of sheet parameters, their units and brief explanation.

The implementations of sheet curvature and twist are also simplified respect to the previous version: we split the range of strand curvatures observed in native NTF2-like domains in three representative values: 155º, 145º and 135º, with 5º of standard deviation. The user selects only a "degree" of curvature: 1, 2 or 3, corresponding to constraints centered in 155º, 145º or 135º (Table 4.5). Constraints favoring these angles are placed on the C_{alpha} carbons of the positions of strand 4 flanking the residue paired with the bulge on strand 3, and the upstream position of strand 5, implicitly generating a degree of sheet twisting proportional to the curvature value (Figure 4.18, yellow lines). For the short arm, twist and bending are solely imparted by the bulge in strand 6, consistently with the narrow degree of variability seen in native NTF2-like domains.

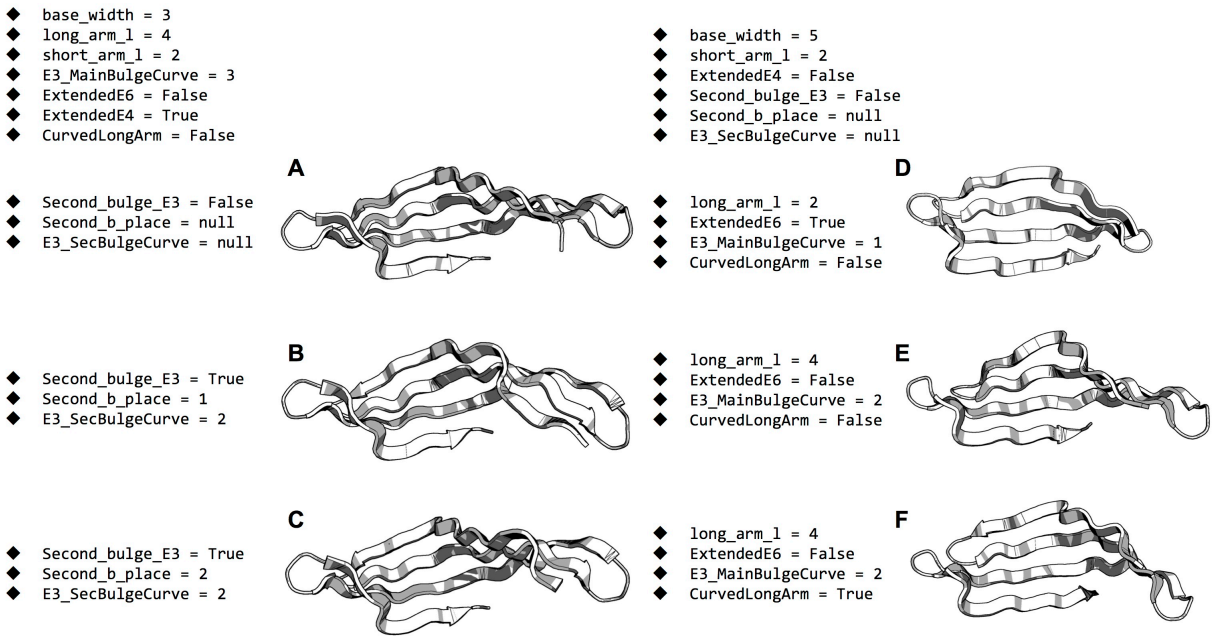


Figure 4.17: **Exemplar sheets produced with six different parameter combinations.** Common parameters for each column are enumerated at the top. On the left, elongated sheets where the main changes are related to the strand 3 second bulge placement. On the right, sheets with a wide base, with different degrees of long arm length and curvature, as well as extension of strand 6.

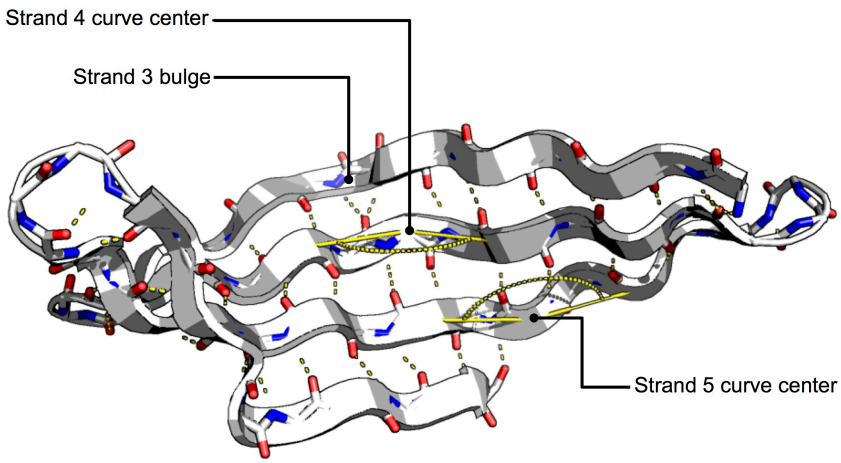


Figure 4.18: **Sheet example with labeled angle constraints** (yellow lines and arches). Constraints are placed only on the middle strands, at shifted positions to generate twist.

Differently from the previous implementation, sheet curvature is focused on key positions in each strand. This hardly represents a decrease in sheet conformation variability, as the base curvature is already constrained by the strand pairings, and the number of long arm conformations is greatly

increased by the addition of a secondary bulge. Nonetheless, since certain long arm lengths can accommodate some bending with little twisting (as seen in PDB 4CDL), we introduced an input variable to impart such curvature in long arms that do not have an additional bulge (Compare sheet structures E and F on Figure 4.17, where CurvedLongArm = False and True, respectively).

Another departure from the first version of the generative algorithm is the possibility of extending strand 6 by 2 residues in the cases where the main bulge curvature degree is small enough to allow it (E3_MainBulgeCurve = 1, Figure 4.17 D). This strand extension, right at the entrance of the pocket, can extend the pocket in additional ways than those described in Chapter 2. The necessity for a small degree of curvature in order to allow extension of strand 6 stems from the principles described in Chapter 3.

The crude combination of all parameter values generates 6912 different variants, but a large number of these are not logical (e.g., Second_bulge_E3 = False and E3_SecBulgeCurve = 2), and a portion of the logical ones would produce unrealistically curved sheets (e.g., E3_MainBulgeCurve = 3 and ExtendedE6 = True). Further restraints are imposed on the available sheet space by the need to complete later steps in the backbone generation: Sheets that cannot be properly connected and/or packed with helices should not be produced. To account for these unproductive cases, we introduced a number of checkpoints before beginning sheet construction to ensure the input parameter values are compatible and productive (See Table 4.6). These checks are based on of manual inspection of results from sheet building trajectories, and limit the number of possible sheet parameter combinations to 68.

Logic check	Explanation
All variables must have allowed values	Unexpected values outside of those described in Table 4.5 are not allowed.
If E3_MainBulgeCurve = 1, then ExtendedE6 must be False	As explained in Chapter 3, sheets with high degree of curvature require bulges to alleviate the clashes caused by bending. Therefore, it is only possible to extend strand 6 past the curvature center on strand 5 if the degree of curvature is low.
If Second_bulge_E3 = True, then Second_b_place and E3_SecBulgeCurve must be integers	If a second bulge is placed on strand 3, a position and curvature for it must be selected
If long_arm_l = 2, then Second_bulge_E3 must be False	An additional bulge on strand 3 cannot be placed is the long arm is not long enough.
If Second_bulge_E3 = True and long_arm_l = 3, then Second_b_place must be 1	An additional bulge on strand 3 cannot be placed on a position beyond the N-terminus of strand 3
If base_width = 5, and long_arm_l = 4, then Second_bulge_E3 must be True	A sheet with base width 5 and long arm length 4 (the highest lengths for both sheet components) would be impossible to connect by helix 3, even at the highest degree of main bulge curvature. It is therefore required for sheets these length to have an additional bulge on strand 3 to curve the N-

	terminus of strand 3 back towards the center of the sheet where it can be connected by helix 3.
If base_width = 3, then ExtendedE4 must be True, else, ExtendedE4 must be False	The extension of strand 4 is only compatible with base width 3, and base width 5 is only compatible with non-extended strand 4. This is a rule derived from manual inspection of native NTF2-like domains.
If base_width = 3, then short_arm_l must be 2	When the base width is the shortest, the short arm must provide additional interactions with H1.
If long_arm_l > 3 (same as = 4), and Second_bulge_E3 = False, then E3_MainBulgeCurve must be 3	If the length of the long arm is 4, and there is no second bulge on strand 3, then, in order to avoid making a sheet that is too elongated, the curvature degree at the main bulge must be 3.
If long_arm_l > 2, then E3_MainBulgeCurve must be higher than 1	Regardless of the presence of a second bulge on strand 3, if the long arm has length 3 or 4, its degree of curvature must be 2 or 3 in order to avoid an excessively elongated sheet
If long_arm_l = 4, Second_bulge_E3 = True, and Second_b_place = 1, then E3_SecBulgeCurve must be exactly 2.	In the case that we have the longest possible long arm with second bulge on strand 3, and this bulge is close to the main strand 3 bulge, then the degree of curvature at the second bulge must be 2 to avoid the sheet extending too far (E3_SecBulgeCurve = 1), or folding back onto itself (E3_SecBulgeCurve = 3).
If CurvedLongArm = True, then long_arm_l must be 3 or 4	Imparting a small degree of curvature on a long arm without a second bulge requires it to be a minimal length of 6 residues, otherwise constraint vertices will target atoms outside the logical range.
If Second_bulge_E3 = True, then CurvedLongArm must be False	When a second bulge is present on the long arm, curvature is dictated solely by E3_ SecBulgeCurve.

Table 4.6: Implemented logic controls to ensure input sheet parameters result in productive construction of sheets.

To prevent the long arm from extending to unproductive distance, in addition to the checks shown in table 4.6, we implemented a simple distance constraint during building of all sheets that biases the sheet arch distance (vector J in figure 4.19 A) to be below 27Å.

We should note that the base width takes only two possible values (3 or 5, not 7) in the new version of the backbone-generating algorithm, excluding the longer values seen in designs from subfamilies 4B.5 and 4B.5.CH. There are two reasons for this choice: The relatively low numbers of stable designs obtained for this fold (Figure 4.6 and table 4.3), and the ability of the new algorithm to capture similar sheet conformations with a combination of base width length, long arm length and strand 3 bulge curvature.

After sheet construction, the following stage is connection of the N-terminus of strand 3 with the bulge on strand 6 (Figure 4.19 A) by the combination of helix 2, the frontal hairpin and helix 3. In the first version

of the generative algorithm, the choice of helix 3 length and H3-S3 connection loop is part of the input variables, preventing the possibility of adapting these elements to each sheet instance, and therefore undermining diversity and efficiency. In the new version we implement logic that, based on a small number of sheet geometric features, discards unproductive sheets and selects productive helix 3/H3-S3 loop combinations. These features are: 1) the distance between the strand 6 bulge and the strand 3 N-terminus (Figure 4.19 A), 2) the protrusion of the long arm over strand 6 (Figure 4.19 B), and 3) the angle between the planes formed by the tip of the long arm and the base (Figure 4.19 C).

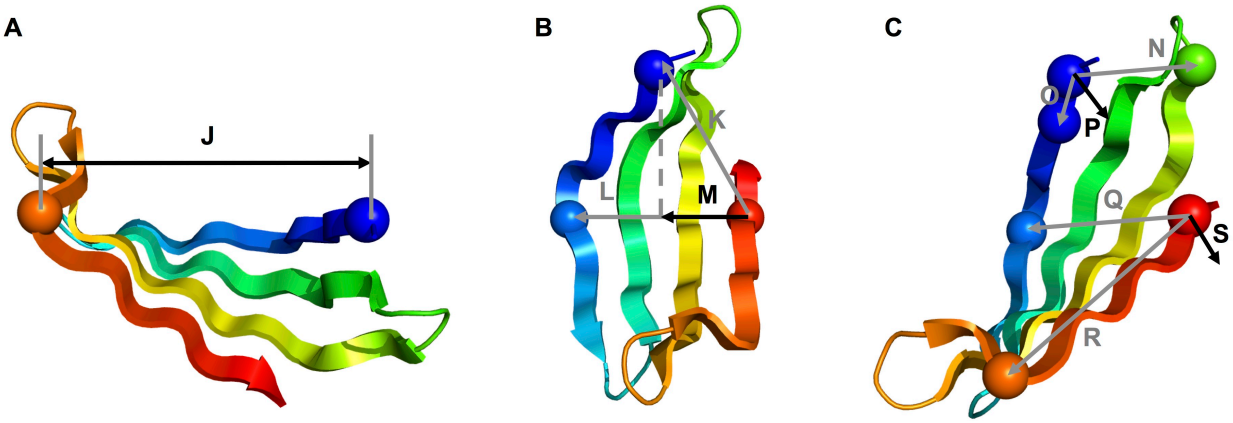


Figure 4.19 Sheet geometrical features: A. Exemplar *de novo* sheet with distance vector J defined as the one connecting the C_{alpha} atoms of residue #1 of strand 6 bulge, and the N-terminus of strand 3. **B**. Exemplar sheet with protrusion vector M defined as the projection of vector K on L. **C**. Exemplar sheet with labeled vectors P and S, which are perpendicular to the planes formed by O and N, and Q and R, respectively. In all cases, C_{alpha} atoms used to calculate vectors are depicted as spheres.

The sheet distance (length of vector J on figure 4.19 A) is intuitively related to the length of helix 3 plus H3-S3 loop, and it is calculated as the vector from the C_{alpha} of the residue on the N-terminus of strand 3 to the C_{alpha} of the "A" ABEGO residue on the strand 6 bulge. The sheet distance is important for detecting exceedingly elongated sheets and discarding them. The protrusion distance is the length of vector M (projection of K on L) in figure 4.19 B, and takes negative values if M points in the same direction as L, and positive values otherwise. The protrusion value is a measure of how much the long arm protrudes over the end of the base (Figure 4.19 B), and it is especially important for determining the length and torsion of the H3-S3 loop. The "long arm angle" is the angle between the planes formed by the strands at the tip of the long arm (plane formed by vectors O and N in figure 4.19 C), and the plane formed by the base (plane formed by vectors Q and R in figure 4.19 C), which is also the angle between

their perpendicular vectors P and S (Figure 4.19 C). This angle is important for detecting sheets that fold outwards and discarding them.

Although constraining the sheet distance (J, figure 4.19 A) limits the overextension of the long arm, when a small degree of curvature is imparted on the long arm in the absence of a bulge (CurvedLongArm = True), the sheet can adopt an outward bending that would prevent the construction of the next step. To prevent wasting resources on trying to continue backbone building with such sheets, we abort trajectories where sheet distance, protrusion and long arm angle are outside of a certain range. Other checks at the end of the construction trajectory are: proper hydrogen bonding between strand pairs, no large deviations from peptide bond planarity, and no Phi/Psi angles outside of the allowed Ramachandran space.

During the second stage in backbone building we use the sheet protrusion to select a loop connection from a predefined subset. We then select a helix 3 length, (10, 11, 14 and 15 residues), based on the selected loop connection. For the H3-S3 connection, there are 6 predefined loops with lengths from 0 (no loop, H3 connected directly to E3) to 5 residues and different torsions. Table 4.7 provides additional information for these connections.

Short name	Length (# residues)	ABEGO string	Compatible H3 lengths
BA	2	BA	10,14
GBA	3	GBA	11,15
GB	2	GB	11,15
ClassicDirect	0	-	10,14
BulgeAndB	4	GBAB	11,15
BBGB	4	BBGB	10,14

Table 4.7: H3-S3 loop connection information.

The compatible H3 lengths for each loop type depend exclusively on the first ABEGO bin of the connection (for ClassicDirect the first ABEGO bin of the connection is “B”, as it is a typical strand residue). The reason for this is the fixed periodicity of helix 3, as it always begins at the S2-H3 connection with the same relative orientation: As shown in figure 4.20, loops with beginning ABEGO B provide ideal directionality and hydrogen-bonding for helices of length 10 and 14, and loops with beginning ABEGO G provide ideal directionality and hydrogen-bonding for helices of length 11 and 15.

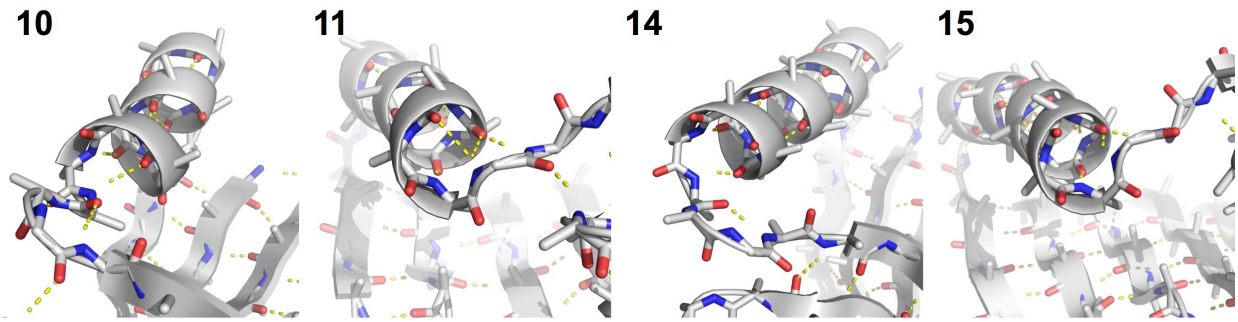


Figure 4.20: **Examples of each possible H3 length** (top left), with the following H3-S3 loop. For 10 and 14, the H3-S3 loops are from the BBGB type, and for 11 and 15, from the GBAB type. Main-chain hydrogen bonds are shown as yellow dashes.

The allowed H3-S3 loop connections (Table 4.7) were chosen by either their presence in native NTF2-like domains, or their prevalence in similar connections in the PDB (14). Additionally, this subset is large enough to contain the necessary range of connection lengths, and small enough to make sheet and H3-legth matching simple. As previously mentioned, the main criterion for matching a connection type to a given sheet, is its protrusion: Long connections (BulgeAndB, BBGB and GBA) are better for connecting sheets with negative protrusion, since they allow H3 to be placed directly on top of the frontal hairpin by working as a spacer between the H3 C-terminus and the S3 N-terminus (Figure 4.21 A). Conversely, short connections (AB, GB and ClassicDirect) are better for sheets with large positive protrusions, since extension outwards by a long connection would make packing against H3 impossible in later steps (Figure 4.21 C). A second bulge on strand 3 imparts additional twist on the long arm, and its combination with a long connection would push H3 too close to the C-terminus of strand 6, therefore, when a second bulge on strand 3 is present, short connections must be used almost exclusively (Figure 2.21 B). All connection types are allowed in cases outside of the extremes discussed above.

The frontal hairpin is constructed during the same stage as H3 and its connection to S3, as it serves as an anchor point on the other side of the arch described by the sheet. The short length and simple hydrogen-bond pattern of the frontal hairpin make it easy to construct, as it has few degrees of freedom. The main variation within this part of the structure is its extension: hairpin strands are most commonly 4 residues long, but if the distance between the strand 6 bulge and its C-terminus is 6 residues or more, the hairpin can be extended to have 6 residues per strand (Figure 4.22).

The main variables describing this stage of the backbone building process are summarized in table 4.8.

Parameter (short name)	Units (Allowed values)	Explanation
H3 length (h_len)	Residue number (10,11,14,15)	Length, in residues, of H3
H3-S3 loop connection type (connection_type)	Categorical - Loops with defined ABEGO strings (See table 4.7)	See table 4.7
Frontal hairpin length (hairpin_len)	Strand residue pair number (4,6)	Length, in residue pairs, of the hairpin formed by S1 and S2 (See figure 4.22)

Table 4.8: NTF2 backbone construction stage 2 parameters.

The final quality controls at the end of this stage are: proper hydrogen bonding between newly built strand pairs, no large deviations from peptide bond planarity, no Phi/Psi angles outside of the allowed Ramachandran space, and no large deviations from straight helical structure for helix 3.

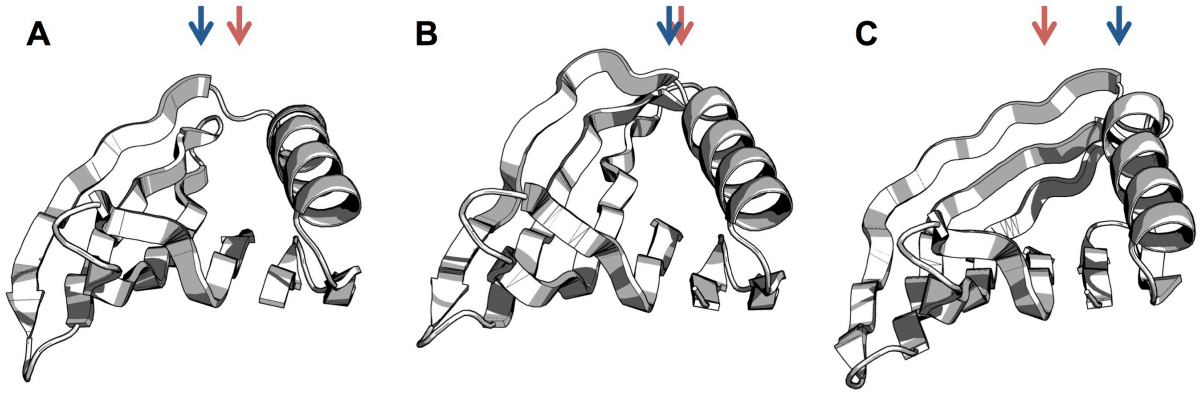


Figure 4.21 Interaction between protrusion, H3-S3 loop length and H3: A. Sheet with negative protrusion value, with long H3-S3 connection (BulgeAndB). Red arrows indicate the position on the C-terminal strand 6 C_{alpha}, blue arrows indicate the position of the N-terminal strand 3 C_{alpha}. A blue arrow at the left of the red indicated negative protrusion values, and vice versa. **B**. Sheet with small negative protrusion value, with second bulge on strand 3, featuring a short connection (ClassicDirect). **C**. Sheet with large positive protrusion, with short connection (BA).

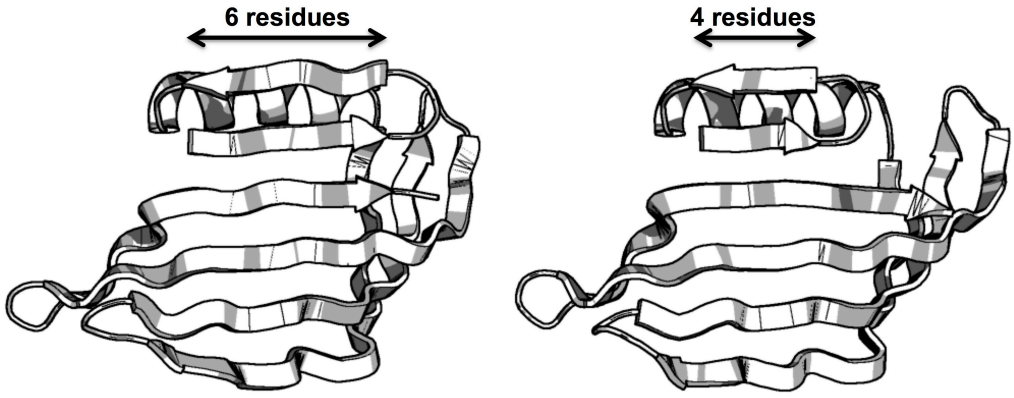


Figure 4.22: Different hairpin lengths in sheets constructed with otherwise identical parameters.

The final stage in backbone generation is closing the cone described by the sheet with the two N-terminal helices. As in the previous step, this one is also adapted to fit the parts of the structure produced in previous steps. The first decision made based on the previous stage is the length of helices 1 and 2: We first compute the distance between the S6 bulge to the H3-S3 connection (Figure 4.23 A, length of vector *i*), if this distance is below 25Å, then the H1 will have a length of 19 residues, and H2, 7 (Figure 4.23 D). Otherwise, an extra turn is added to both helices (lengths 23 and 11, respectively) (Figure 4.23 E). The H1-H2 connection is always a 2-residue loop with ABEGO bins GB, no matter the *i* distance.

To properly place both H1 and H2 relative to the sheet and H3, we implement constraints based on their geometry. If the sheet has a second bulge on strand 4, and a protrusion value (Figure 4.19 B) below 6, the C-terminal of H1 is packed against the residue immediately after the strand 3 bulge, with the strand residue C_{alpha}-C_{beta} vector pointing towards the socket formed by the last H1 turn (see figure 4.23 B and E), as described in (15). In all other cases, the N-terminal S3 residue is the one used to fill the H1 C-terminal socket (see figure 4.23 B and D). A similar logic is used to constrain the H1 N-terminus: If the short arm length is 2, then the second residue of strand 6 is used as a knob to fit in a socket on the side of H1, right next to the short arm (Figure 4.23 C). Otherwise, the residue immediately before the N-terminus of strand 6 is used as knob. Additional hydrogen-bond constraints are added to the H2-S1 connection to enforce its conserved structure. To bias helical structure towards canonical, straight, conformations, we add Phi/Psi constraints, but only during a minimization stage where the backbone is slightly moved in place. This simple distance constraint scheme is general enough to efficiently complete the final step in basic NTF2-like protein backbone construction for a large variety of inputs.

The main variables describing this stage of the backbone building process are summarized in table 4.9.

Parameter (short name)	Units (Allowed values)	Explanation
Opening placement (Opening)	Categorical (Classic,Alternative)	The pocket opening on NTF2s can be placed either between the frontal hairpin and H3 (Classic), or between H3 and the H1-H2 connection (Alternative).
H1 length (h1_len)	Residue number (23,19,14)	Length, in residues, of H1
H2 length (h2_len)	Residue number (11,7)	Length, in residues, of H2
C-terminal helix (has_cHelix)	Boolean (True,False)	Presence or not of a C-terminal helix

Table 4.9: NTF2 backbone construction stage 3 parameters.

As previously described, native NTF2-like domains often exhibit C-terminal elements that extend pocket surface outwards. The last stage of NTF2 backbone generation is adding such C-terminal element (exclusively a helix, termed C-helix), if requested by the user and allowed by the geometry of the structure constructed so far. To add a C-terminal helix, the fist step is to ensure the sheet conformation would allow it to have reasonable structural contacts with the rest of the structure: Sheets where the distance between the C_{alpha} carbon of the S5 N-terminal residue and S6 C-terminal residue is below 15Å or above 18.5Å will be discarded. This distance is a measure of the pocket mouth size, and extremely low or high values could result in the placement of the C-helix that either closes the pocket or has little interaction surface with the rest of the protein, respectively. In all cases, the C-terminal helix is 8 residues long, and is connected to S6 by a one-residue loop with ABEGO bin B. To ensure the C-terminal helix is interacting with the rest of the structure without closing the pocket, constraints between the C-terminal of the C-helix and the tip of the long arm are imposed during backbone construction.

The main variables describing this optional stage of the backbone building process are summarized in table 4.10.

Parameter (short name)	Units (Allowed values)	Explanation
C-helix length (h_len)	Residue number (11,8)	Length, in residues, of the C-terminal helix.

Table 4.10: NTF2 backbone construction optional stage 4 parameters.

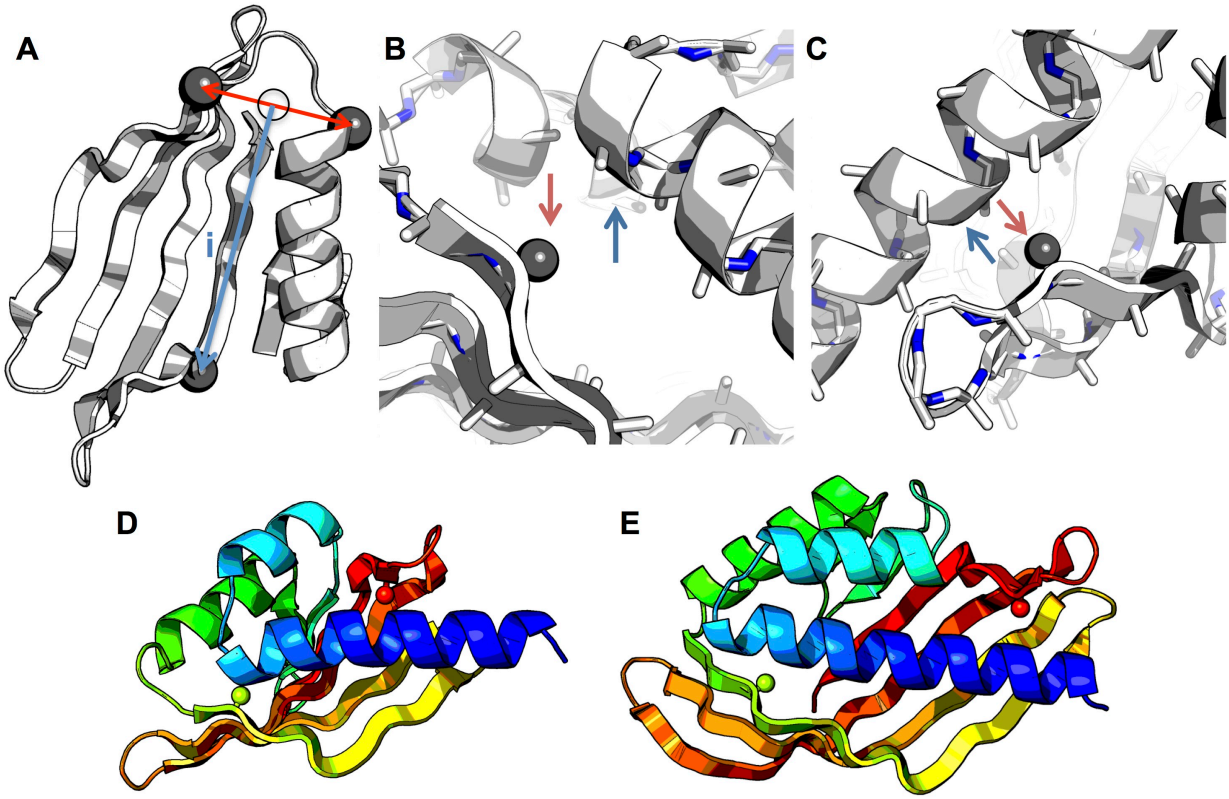


Figure 4.23 N-terminal helix placement relative to sheet: A. Distance between the H3-S3 connection and the strand 6 bulge is measured as the length of vector *i*, which goes from the middle point between the N-terminus of S3 and the C-terminus of H3 (black circle), to the C_{alpha} of the first residue of the S6 bulge. **B**. Schematic of knob-socket interaction on the C-terminus of H1, the red arrow points at the know residue (C_{beta} represented as sphere), the blue arrow points at the socket. **C**. Schematic of knob-socket interaction on the N-terminus of H1, similarly to B. **D**. Example of *de novo* NTF2 backbone with short H1 and H2 helices. Strand knob residues are represented as spheres. **E**. Example of *de novo* NTF2 backbone with long H1 and H2 helices. Strand knob residues are represented as spheres.

So far we have described the process of generating *de novo* NTF2-like structures where the opening of the pocket is on the canonical, frontal, position. As part of the efforts to generate structures with pockets that are as diverse as possible, we also implemented a variant within the generative algorithm that enables the user to place the opening of the pocket between H3 and the H1-H2 connection, similar to the Mk1.TP subfamily in the initial version of the generative algorithm (Figure 2.24).

Building a backbone with the alternative opening placement requires the trajectory to diverge from the previously described at the stage of H1-H2 building. The first step in the “alternative opening” path is checking sheet geometry: the opening of the pocket must have a diameter (distance between the C_{alpha} carbon of the S5 N-terminal residue and S6 C-terminal residue) between 13Å and 16Å, and the distance

between the H3-S3 connection to the S6 bulge (Figure 4.23 A, distance *i*) must be below 22Å. If the sheet passes these geometric filters, then the construction of H1 and H2 can be set up: All backbones with alternative openings have a 7-residue H2, and a 14-residue H1, shorter than those with a canonical opening. In order to place H1 and H2 on the sheet while maintaining an opening between these helices and H3, we bias backbone construction using constraints. Similarly to the "canonical opening" path, the N-terminus of H1 is constrained to interact with the short arm (Figure 4.23 C), but the C-terminus is constrained to be near the main bulge of S3. In order to prevent the H1 C-terminal constraints to drag the N-terminal helices too far from H3, constraints are placed between H2 and the middle of H3.

The final step in the generation of backbones with alternative openings is occluding the place where the canonical opening would be, with a C-terminal helix. Differently from the C-helix used to extend the pocket outwards, this C-terminal helix is 11 residues long, and must be placed optimally between the tip of the long arm and the frontal hairpin. The connection between S6 and the C-terminal helix is a 2-residue loop with ABEGO bins AB, which allows a sharp turn in backbone direction to point the helix back towards the pocket opening. To bias the placement of the C-terminal helix, we use distance constraints that maintain the C-terminus of the C-helix near the N-terminus of H3, and near the midpoint between the C-terminus of S2 and the N-terminus of S4 (Figure 4.24).

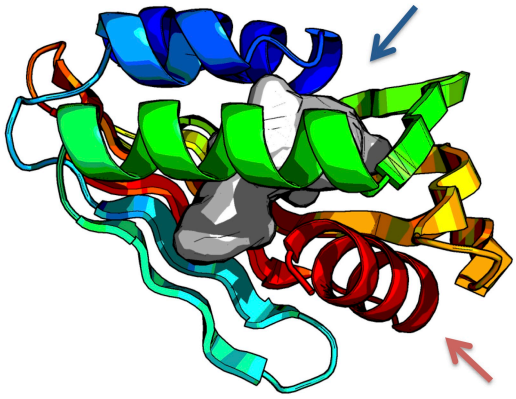


Figure 4.24 *De novo* NTF2 design with alternative pocket opening: Schematic of a finished (optimized sequence) *de novo* NTF2 protein with pocket opening in the space between H1 and H3 (blue arrow), instead of the canonical placement at the front, where the C-helix has taken its place (red arrow). The pocket surface is rendered in grey.

Once the backbone is generated, we design its sequence, optimized to fold into that structure. To design this new generation of proteins, we incorporated the lessons from the logistic regression model trained on high-throughput screening data, and the crystal structures obtained for some of those proteins. We added requirements for the minimal number of hydrophobic residues, extended the optimization of secondary structure propensity, and devised logic to detect when to allow glycine on highly curved strand positions (See methods 4.5.13).

Several of the most important features detected by the simplest logistic regression model are strongly linked to hydrophobic residue content (*BuriedHyphobSAperRes*, *nres_hydrophob_noA_per, fa_sol, hydrophobicity* and *hphob_sc_contacts*, figure 4.7 C and table 4.5.s). We decided to optimize all these features simultaneously by biasing the overall composition of the sequence to be between 27% and 33% non-alanine hydrophobic residues (FILMVWY). This resulted in a significant improvement of hydrophobicity-related metrics in the new designs (Figure 4.25). To maintain realistic core/surface hydrophobic character, we also biased surface residues to be 90% non-hydrophilic, and core residues to be 90% hydrophobic. The composition of positions at the interface was not biased.

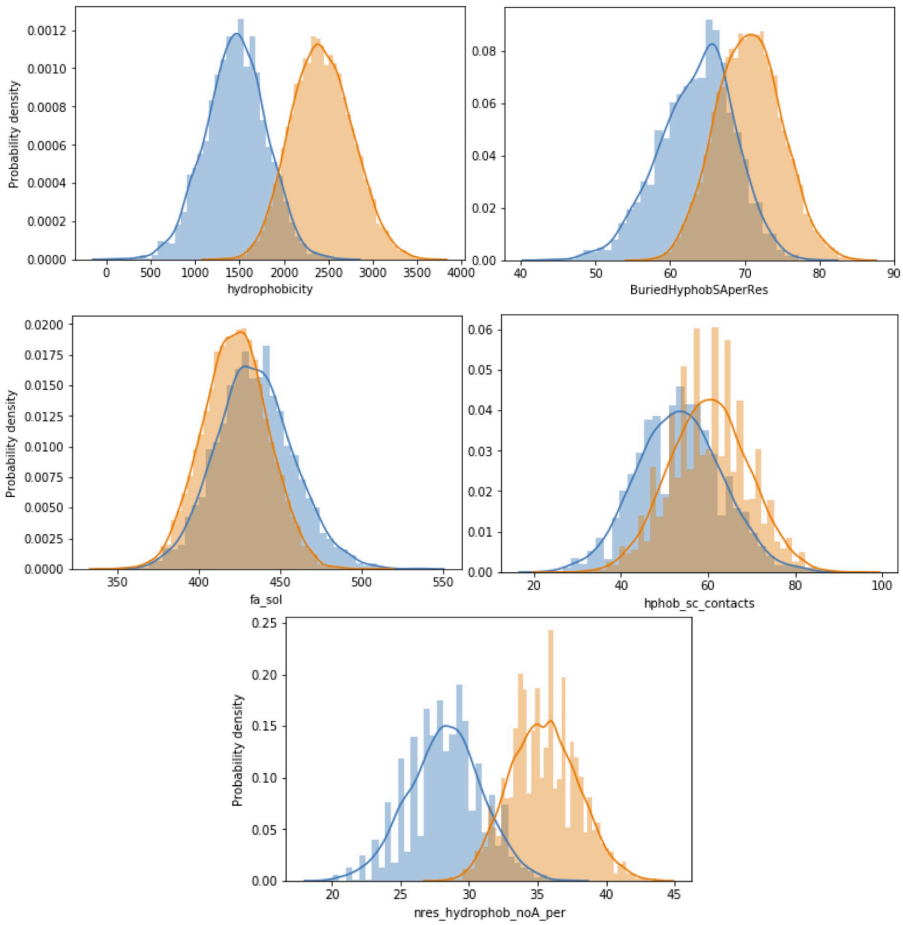


Figure 4.25: Distributions of hydrophobicity-related features with high weights in the logistic regression model. Blue: Distributions for designs made by the first version of the generative algorithm. Orange: Distributions for designs made by the second version of the generative algorithm.

We attempted to optimize other important features related to local sequence-structure agreement (*AvAll, H3_av_worstfr* and *H3_n_av_worstfr*), and buried unsatisfied side-chain polar groups (*buns_nosurf_sc*), which are not well represented by the Rosetta score function, by randomly mutating residues and keeping mutations that improve these features. For the specific case of sequence-structure agreement features, calculation during the design process is prohibitive, so we optimized for low secondary structure mismatch probability instead (See SSmismatch in Table 4.5.7). Although some of these features show a small shift towards worse values with the new protocol (Figure 4.26), some improve, and it is possible the optimization process we implemented prevents their deterioration as a side effect of hydrophobicity optimization.

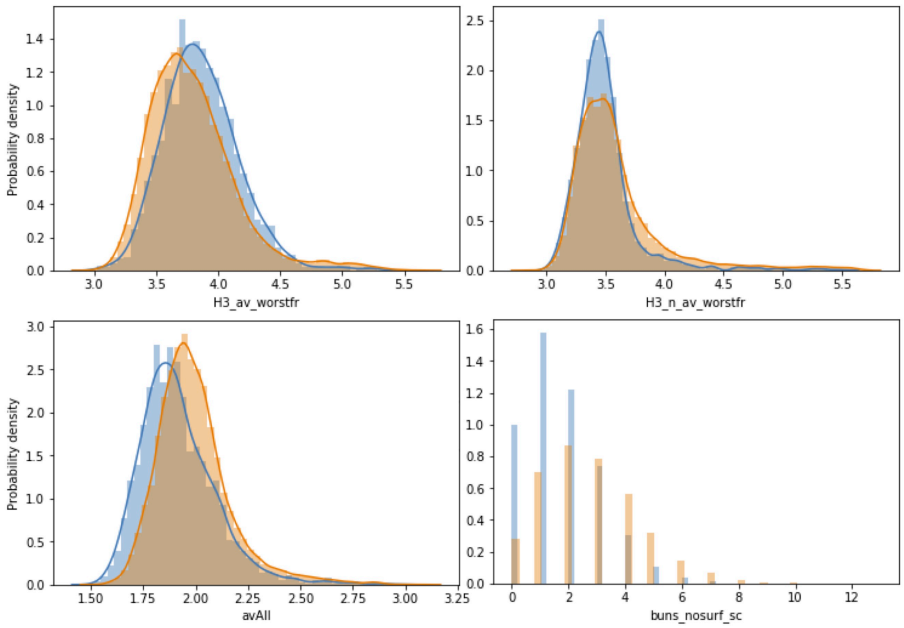


Figure 4.26: **Distributions of local sequence-structure agreement metrics** (*H3_av_worstfr*, *H3_n_av_worstfr*, and *avAll* - lower values mean better agreement) and *buns_nosurf_sc* (buried unsatisfied polar atoms in non-surface side chains). Blue: Designs made with the first version of the generative algorithm. Orange: Designs made with the new version od the generative algorithm.

Rosetta score per residue (*score_res_betacart*) is optimized by default on all design calculations; we therefore only introduced penalties to worse scores during the random mutation stage. Figure 4.27 shows only a small shift towards worse score values for new designs. Finally, for features related to tertiary structural motifs (TERMS, (16), *H3_ssc50_worst*, *str5_6_ssc50_worst*, see Tables at section 4.5.7) we did not apply any direct optimization strategy, but see improvement on *str5_6_ssc50_worst*, and deterioration on *H3_ssc50_worst* (Figure 2.27).

As described in the previous section, placing glycine on strand positions with high degrees of curvature is of key importance to stabilize their conformation. In the case described on figure 4.16, the position of S4 paired to the main S3 bulge has a high degree of curvature, which we stabilized by mutating it to glycine and engineering other positions to pack in the space left by the alanine side-chain. It has been reported in the literature (17) that glycine is more common on strands positions with curvature angles below 150º, which is consistent with the curvature observed in the original BBM2nHm0589 design (Figure 4.13). The curvature angle of 146.5º on the residue paired to the main S3 bulge (measured between C_{alpha} atoms of positions *i*-2, *i* and *i*+2) in the original BBM2nHm0589 design, becomes more pronounced (128º) when mutated to glycine. To systematize this “glycine rescue” in positions paired to

main and second bulges on S3, we decided to allow only glycine on S4 positions paired to S3 bulges that have curvature angles below 147.5, leaving it to Rosetta rotamer packing and score function to design additional interactions with the enforced glycine. Figure 4.28 shows examples of the implemented strategy.

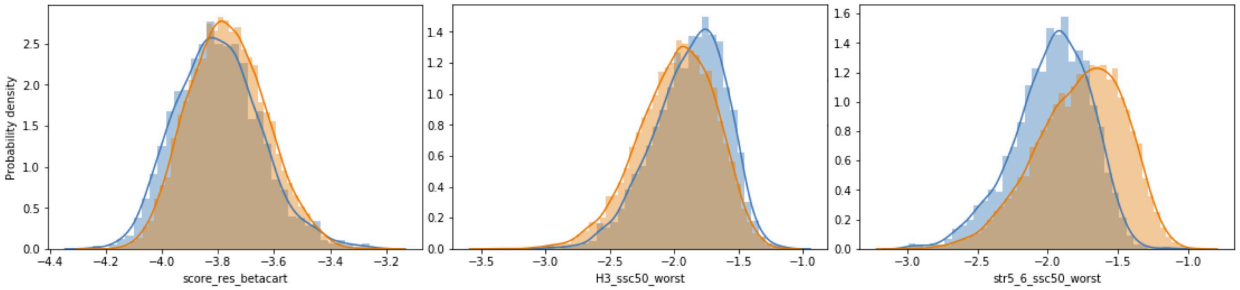


Figure 4.27: **Distribution of TERM-related features** H3_ssc50_worst and str5_6_ssc50_worst (higher values indicate higher model quality), and Rosetta score per residue, score_res_betacart, in designs made by the first version of the generative algorithm (blue), and the new version (orange).

Given the relatively low diversity of conformations of the short arm, we decided to also enforce the lysine-tryptophan pair when short_arm_l=2, as done as part out efforts to re-fold BBM2nHm0589 (See figure 4.16 A, lower left panel).

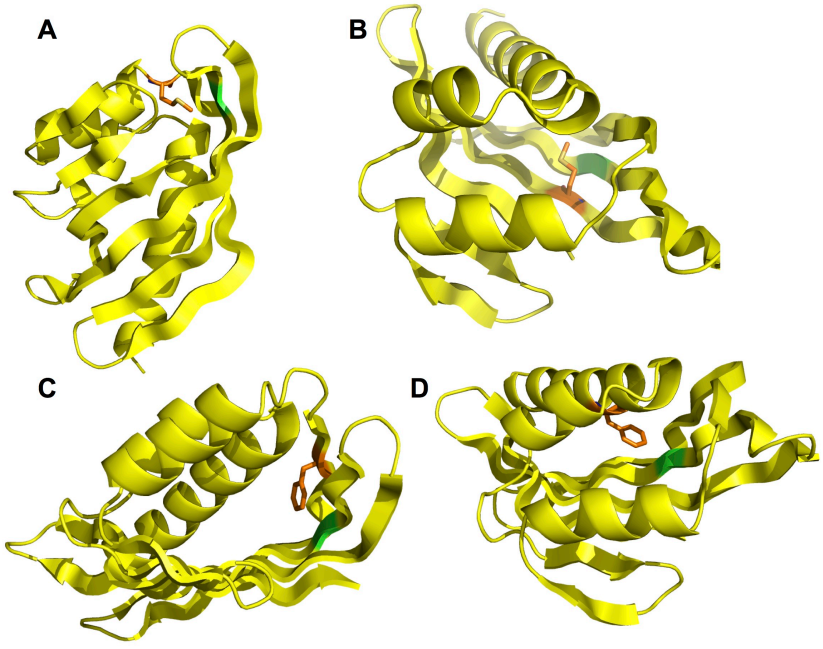


Figure 4.28 ***De novo*** **and native NTF2-like proteins with glycine in highly-curved sheet positions**: **A.** Glycine rescue applied to second bulge on S3, a packing solution is found where methionine (orange) interacts with glycine (green). **B.** Glycine rescue applied to main bulge on S3, a packing solution is found where methionine (orange) interacts with glycine (green). **C.** Glycine rescue applied to main bulge on S3, a packing solution is found where phenylalanine (orange) interacts with glycine (green) in a similar way as in the BBM2nHm0589 mutant. **D.** Glycine rescue applied to main bulge on S3, a packing solution is found

where phenylalanine (orange), from a position far in primary structure, interacts with glycine (green).

4.3.5 Evaluation of diversity generated by the new NTF2 generative algorithm

In the previous section we described a new generative algorithm for proteins of the NTF2-like superfamily. This algorithm is composed of a more general backbone-generating algorithm and a design algorithm that implements the lessons from the previous version. In this section we analyze the designs it produces to evaluate their diversity and compare them to native NTF2-like domains.

Despite each backbone generation step having a number of variables, and each of these variables, several possible values, not all combinations will produce models that pass al quality control filters. To evaluate the diversity of the generative algorithm, we launched backbone-building trajectories that randomly select a starting set of sheet parameters of the possible 68, and try to build the rest of the structure in a limited number of attempts. From 36102 structures produced, the resulting number of unique parameter combinations is 1503, more than an order of magnitude of those produced by the first version of the generative algorithm.

As shown in Chapter 2, the main sheet is determinant for pocket size and structure, and the most variable subelement of the canonical NTF2 structure. To provide an idea of the systematic variation in sheet structure, we constructed a simplified category system of 59 sheet types that sample a grid of sheet length and curvature values and analyzed their frequency and features among the 36102 designs produced (Figure 4.29). The differences in sampling for each of the sheets observed in figure 4.29 can be attributed to two possible reasons: A trivial one, which is that sheet types with more sampling simply concentrate a higher number of “subtypes”, and therefore get more construction attempts - e.g., type 43 contains sheets where the second bulge placement can be in two different positions, effectively making 43 a sum of at least two sheet subtypes. A more interesting reason is that certain sheets are more strained or sample conformations that cannot be completed with the available options at subsequent steps, resulting in early termination of the trajectory. A combination of both explanations is also possible.

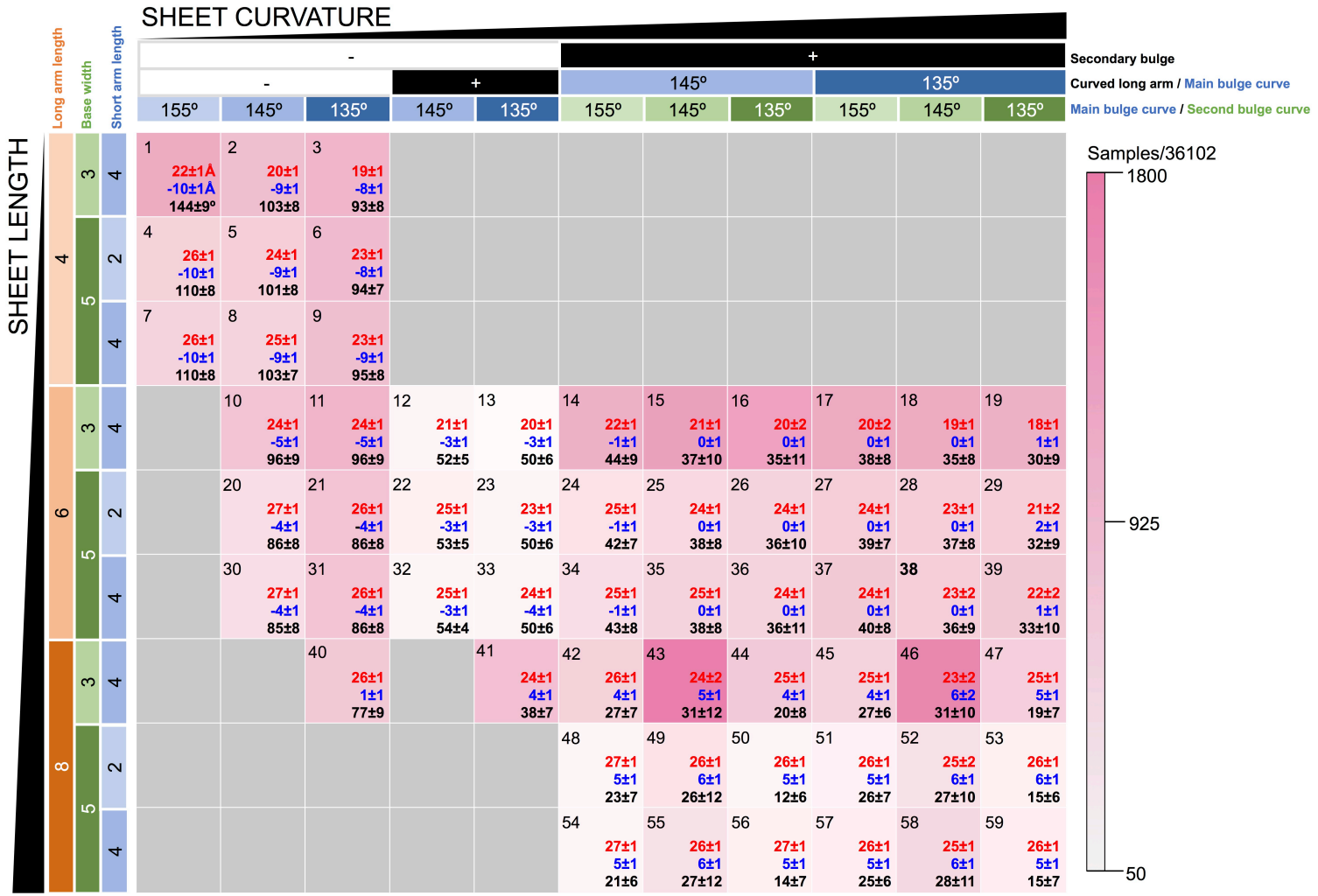


Figure 4.29: Simplified NTF2 category system based on a subset of sheet parameters that map to sheet length (Y axis, orange, green and blue boxes) and curvature (X axis, black, white, blue and green boxes). Each cell is colored with intensity proportional to the number of times their parameter combinations are observed in the set of 36102 generated structures. The number at the top left of each cell is the designated ID for that parameter combination. The three bold numbers on the bottom right are the mean and standard deviation of: sheet arch distance (in Å, red, defined in figure 4.19 A), protrusion (in Å, blue, defined in figure 4.19 B) and long arm angle (in degrees, black, defined in figure 4.19 C).

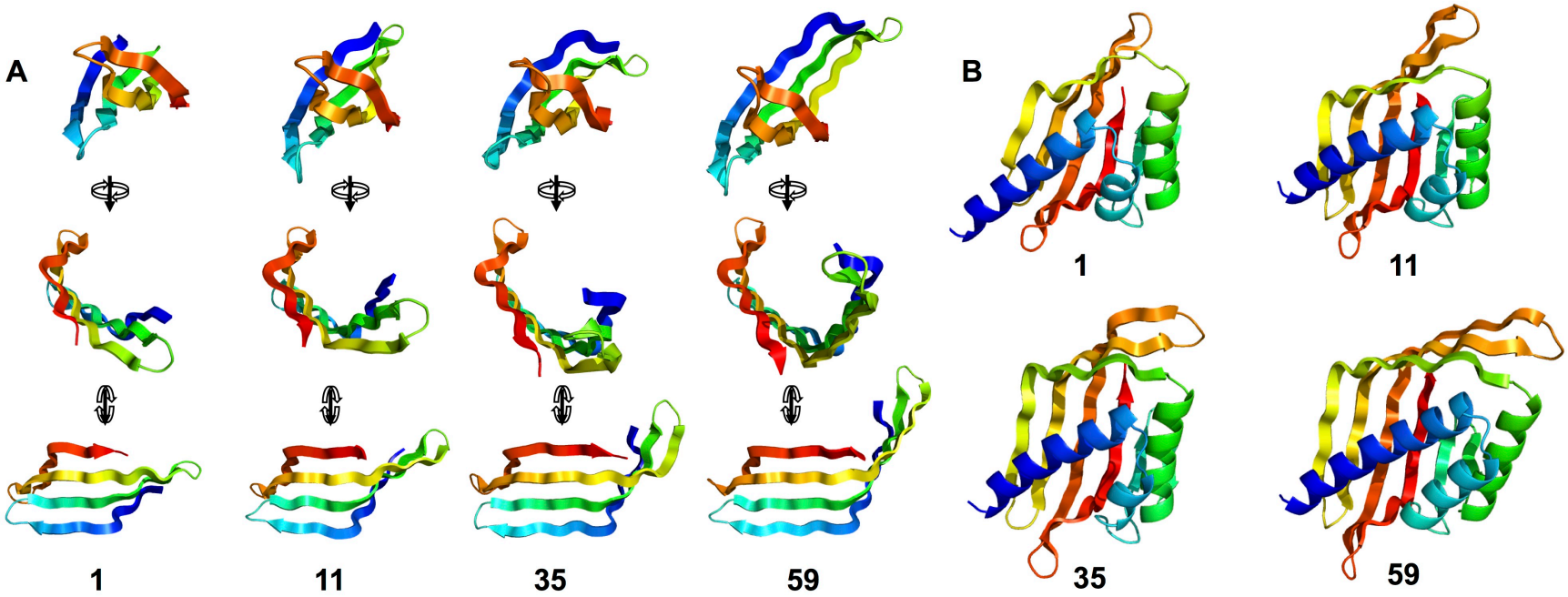


Figure 4.30 Examples of structures generated by the generative algorithm: **A**. Sheet conformation examples for types across sheet length and curvature space (types 1, 11, 35 and 59 from figure 4.29). **B**. Completed *de novo* NTF2-like proteins featuring each of the sheet types highlighted in A.

Figure 4.29 shows how the three main-sheet geometrical parameters change as a function of sheet length and curvature. Sheet arch distance (Figure 4.29, red values in each cell) tends to increase with sheet length at equal values of sheet curvature, mainly due to increases in base width and long arm length. The contrary is observed for increasing sheet curvature for a given length value. Both of these trends are in line with the intuitive understanding of how sheet length and curvature are generated, and how the sheet arch distance is calculated. The only cases where these trends break is types with a curved long arm (CurvedLongArm +, types 12, 13, 22, 23, 32, 33 and 41), that tend to have shorter arch distances than their bulged counterparts, at least when the former have a low degree or curvature in the second S3 bulge (Types 14, 24, 34 and 42). The protrusion and long arm angle follow inverse trends, protrusion increases with the length of the long arm and sheet curvature, while angle decreases. This is linear relation is expected, as the plane formed at the tip od the long arm rotates as the long arm elongates and curves, as seen in the sheets of types 1, 11, 35 and 59 (Figure 4.30 A). Despite their large structural differences, all these sheets can be used to construct full NTF2-like proteins using the rules described in the previous section (Figure 4.30 B).

Similarly as we showed in Section 4.3.1, we can quantify the coverage of the native NTF2-like domain space by the generative algorithm and vice versa. Since a great advantage of the new generative algorithm is its ability to produce orders of magnitude more unique parameter combinations (9 vs. 1503), we considered a fair comparison with natives would include at least one member of each parameter combination.

Figure 4.31 A shows a comparison of all vs. all TM-scores for natives and the proteins generated by the two versions of the generative algorithm. Although the new version is more diverse than the original one, it still does not reach the range of natives. We attribute the spread of TM-scores in natives to the higher proportion of irregular elements in their structures, and sampling of topologies that diverge significantly from the NTF2 canonical structure by, for example, not having a frontal hairpin. The cases where native structures lack the frontal hairpin make up a large part of the native space that is poorly covered by the latest version of the generative algorithm (Figure 4.31 B – red dots near Best alignment Z-score = 0, and Figure 4.31 C, bottom right), largely because we purposefully avoided creating such types of structures with shallow or non-existent pockets.

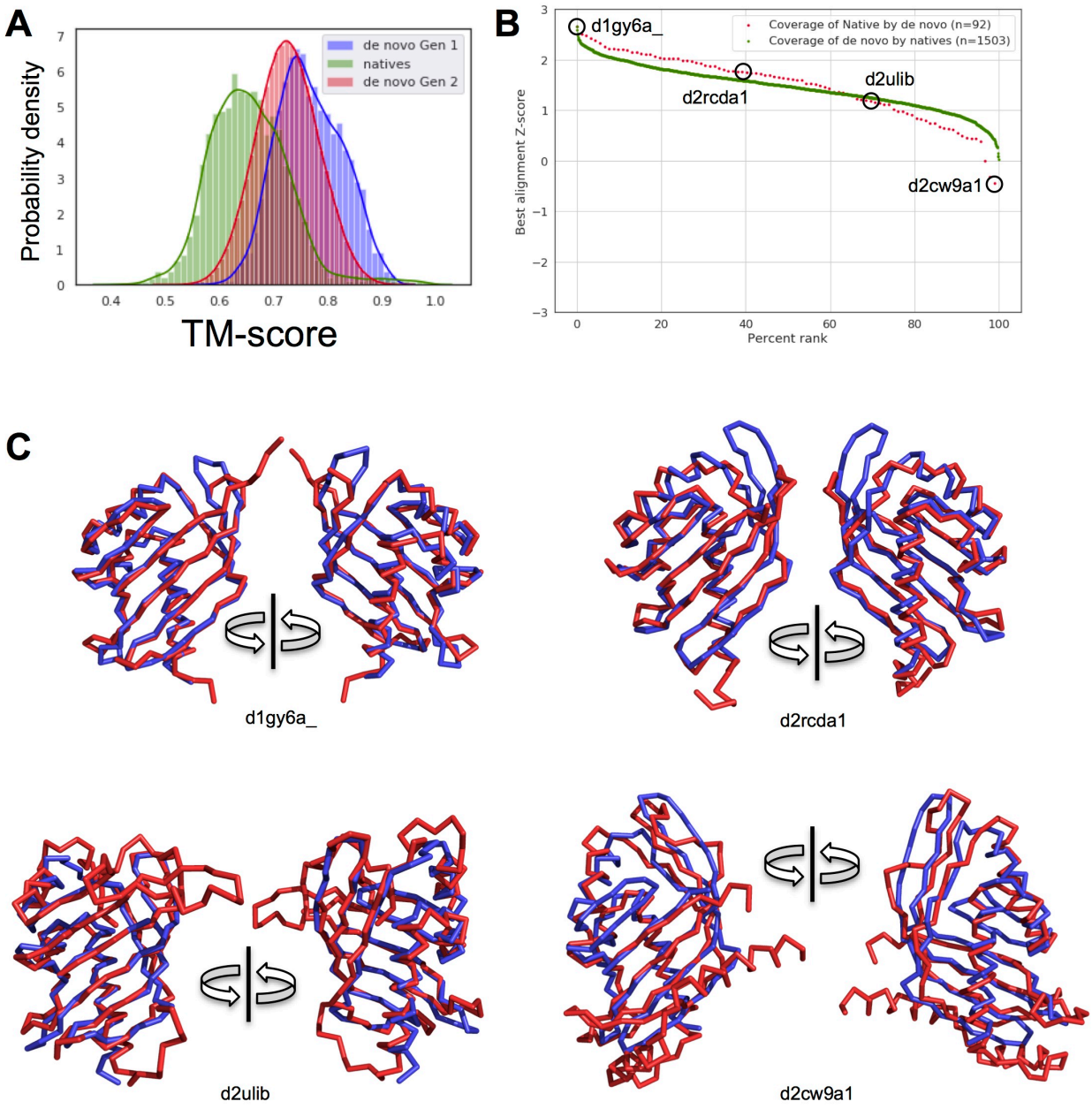


Figure 4.31 Structural diversity of proteins generated by the second version of the generative algorithm compared to native NTF2-like domains: A. Normalized histogram of all *vs*. all TM-scores within native and *de novo* sets. Gen. 1: first version of the generative algorithm, Gen. 2: new version of the generative algorithm. **B.** Z-score ranking of the best matching *de novo* designs to natives and vice-versa. **C.** Backbone ribbon representations of aligned structures for different native NTF2-like domains and the best matching *de novo* models.

As shown in figure 4.31 B, the 10% lower-ranked native matches to *de novo* (green dots, percent rank from 90 to 100) are not well recapitulated by natives (Best alignment Z-score < 1). This represents

nearly 150 parameter combinations that produce proteins significantly different from those seen in nature, a number that nearly doubles the non-redundant set of NTF2-like domain structures.

Another interesting comparison between natives *de novo* designs is the distribution of pocket sizes. We measured the pocket volume of the 36102 designs generated with the new version of the generative algorithm and compared that to the distribution of native pocket volumes. Although the distribution of de novo designs is highly skewed to small pocket sizes, it samples the whole range observed in natives (Figure 4.32). We attribute the skewedness of the design distribution to the fact that we did not designed any kind of binding or active site in them, which is arguably the reason why the volume distribution in natives is centered around larger pocket volumes. When comparing the numbers of designs with large pocket volumes (>750Å^3), to the number of natives in the same range, we see we produced significantly more designs with large pocket volumes (Figure 4.32 inset).

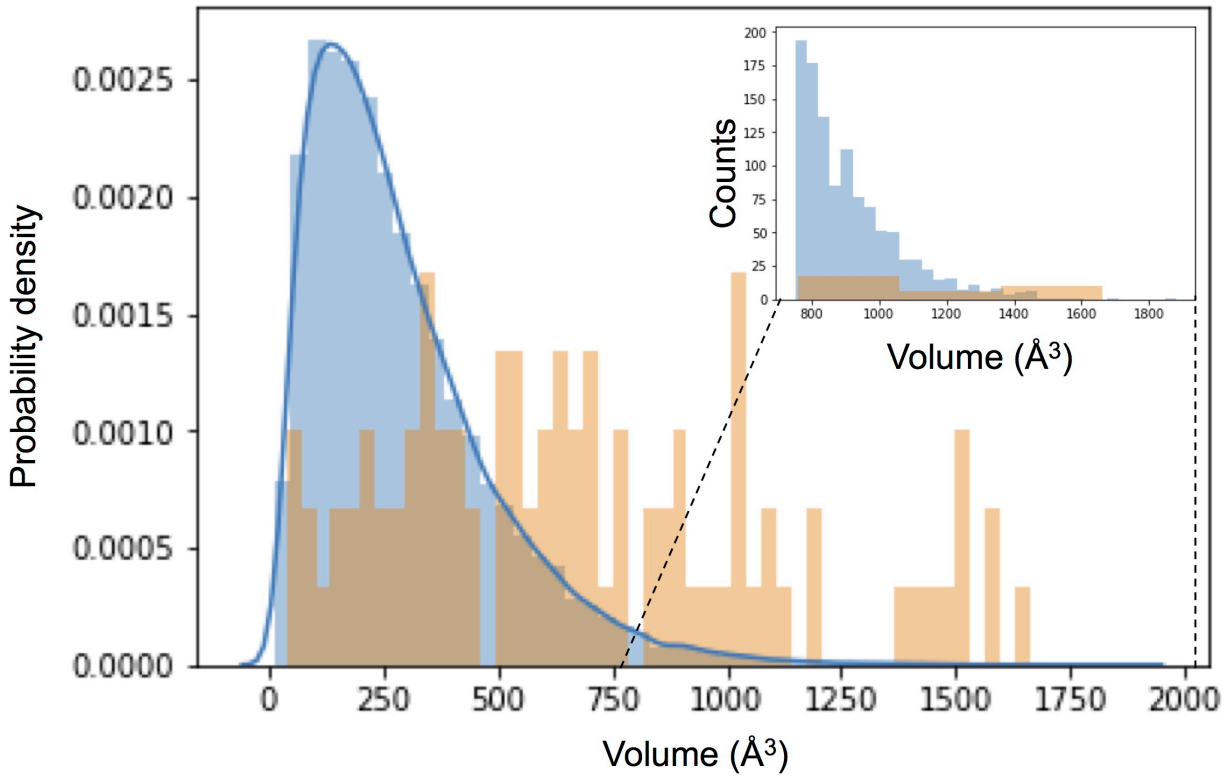


Figure 4.32: Normalized distribution of pocket sizes in de novo designs (blue) and native NTF2-like domains (orange). Inset: Histogram of pocket volumes above 750 Å^3, same coloring scheme.

To gain a more nuanced understanding of how native and *de novo* NTF2 space relate, we constructed a dendrogram and heat map using 1-TMscore as pairwise distance measure (Figure 4.33),

comparing the 1503 examples for each different parameter combination and 92 natives. Given the large numbers of de novo designs, we expected the natives to be interspersed among the designs, but this is not the case, they cluster together in the same groups as in figure 4.4. In particular, a small branch of natives has large pairwise distances with all the rest, native or designed (top left). This small branch is composed by NTF2-like domains lacking the frontal hairpin, indicating these are an exclusive group even among natives.

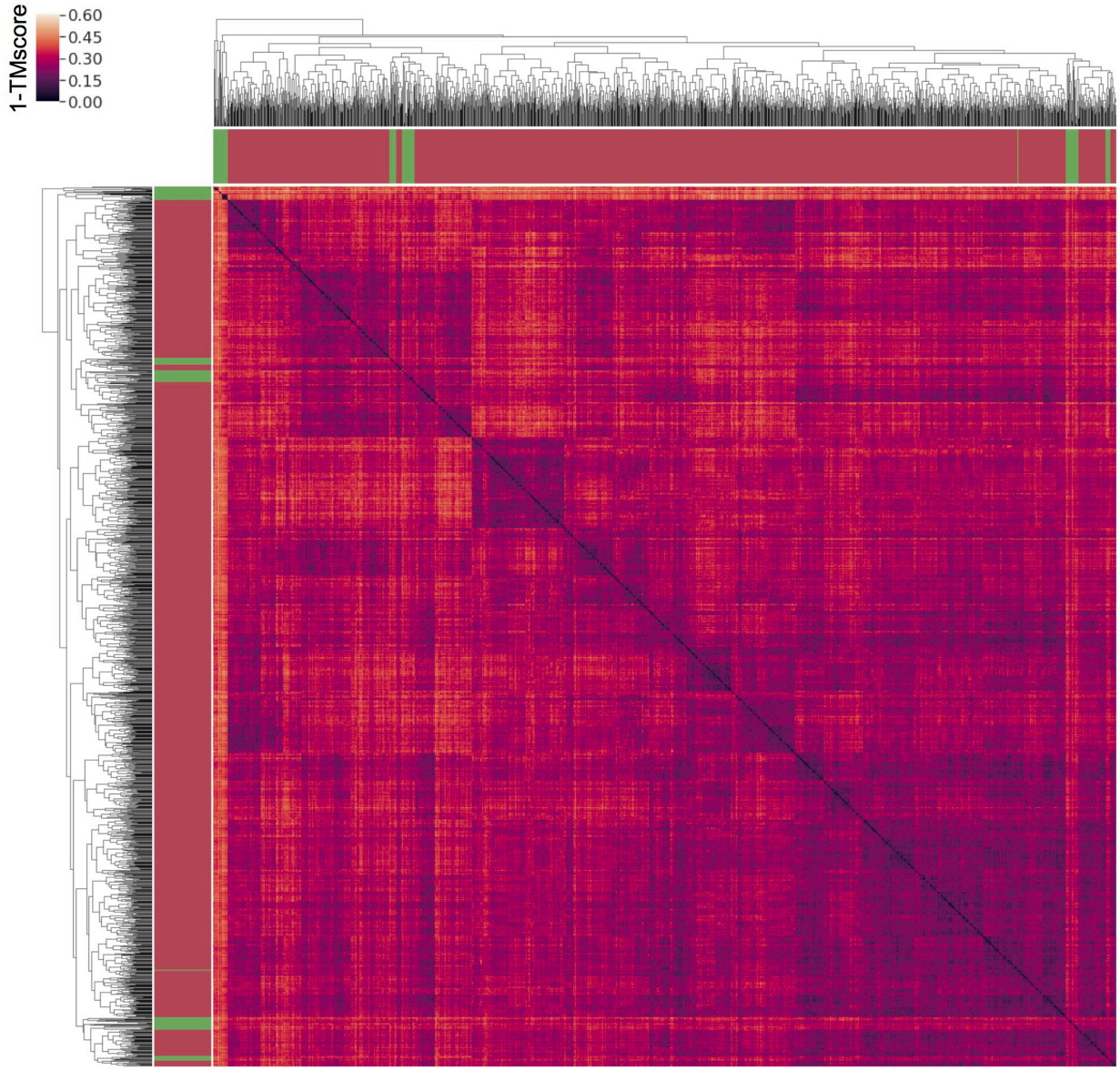


Figure 4.33: **Heat map of clustered NTF2-like domains, along with the dendrogram** (two identical ones on upper and left sides of the heat map) resulting from clustering. The distance metric used for clustering, 1-TMscore (upper left color bar), is depicted in the heat map. *De novo* designs are labeled as red bars at the tips of the dendrogram, while natives are green.

The analysis of *de novo* designs diversity shows the new version of the generative algorithm samples a large space outside of observed native structures, with a comparable range of pocket sizes. Although part of the native space is not sample by the generative algorithm, we argue this space is not particularly productive, as it includes large deviations of the canonical NTF2-like structures that do not provide significant diversity in pocket structure. The diversity generated by the new generative algorithm focuses on sheet structure, which forms a large proportion of the pocket structure.

4.3.6 High-throughput screening of de novo NTF2-like proteins generated by the second version of the algorithm

To evaluate the performance of the new generative algorithm we used the same strategy described in 4.3.2. From the 1503 unique parameter combinations, only 527 produce NTF2-like proteins short enough to be synthesized in two pieces as part of oligonucleotide arrays (6, 7). Using the logistic regression model trained on the data from the first large scale experiment, we predicted the probability of having a stability score above 1 for an initial set of 11548 designs made with the new generative algorithm. We selected a subpopulation of them (10073) enriched in designs with high chance of being stable. In parallel, we clustered all by several fold-related features (See methods 4.5.14), and obtained 6550 clusters. From each cluster, we chose a representative found in the “stable” design population. We collected 6025 designs this way (some clusters had no representatives in the 10073 “stable” design list). The process of generating DNA sequences for these designs further trimmed the number of designs to 5678, across 323 unique NTF2 parameter combinations.

We included controls to test the effect of glycine and tryptophan/lysine pair placements in protein stability. We mutated the glycine residues in highly curved strand positions to other amino-acids, as well as the residues around them to optimize packing (245 mutants). We did the same for tryptophan/lysine pairs (248 mutants). As in the large scale screening done in section 4.3.2, we included scrambled controls (1000 sequences). To ensure the results of this experiment are comparable to the previous one, we included 59 sequences for which we obtained stability scores in the previous round.

The new stability scores obtained for control proteins evaluated in the last round of high-throughput screening match closely the previously reported values (Figure 4.34).

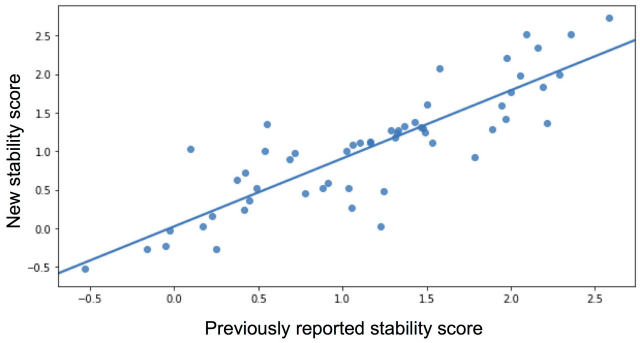
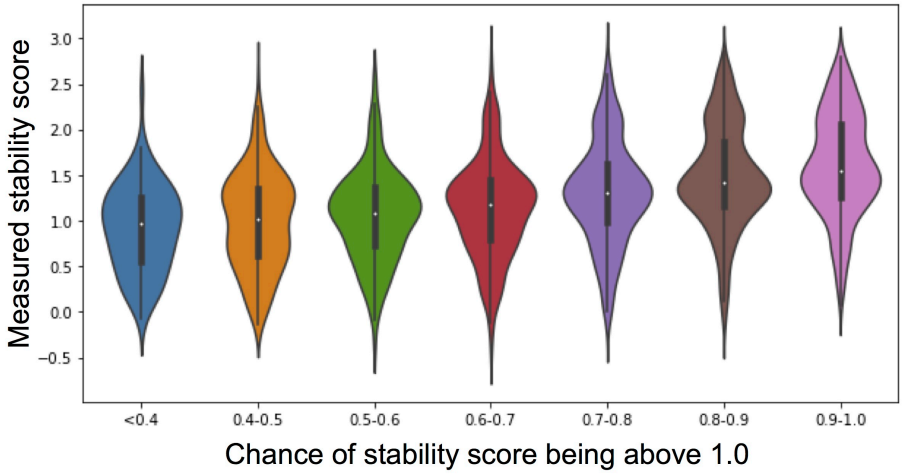


Figure 4.34: **Scatter plot of stability scores measured in the current set of experiments vs. previously reported stability score** (n=55). Blue line is the linear fit to the data (Slope = 0.88, intercept = 0.02, Pearson R: 0.86)

We obtained reliable stability scores for 5188 designs (not counting controls of any kind), of which 3744 (72%) have stability score above 1. This is an ostensibly larger fraction of designs with high stability score than before (21%), indicating the changes included in the design strategy were effective. When we compare the design stability scores with the predicted chances of them being above 1, we see a clear trend indicating the logistic regression model generalizes to this new set of proteins (Figure 4.35).



Bin	Actual fraction with S.S. > 1
<0.4	0.44
0.4-0.5	0.52
0.5-0.6	0.58
0.6-0.7	0.62
0.7-0.8	0.74
0.8-0.9	0.83
0.9-1.0	0.88

Figure 4.35: **Comparison of predicted and measured stability scores**. Designs are binned by the predicted probability of their stability score being above 1, and for each bin the distribution of measured stability scores is plotted. The table on the left shows the fraction of designs with measured stability score above 1 for each bin.

Although the fraction of designs with stability score above 1 is larger than in the previous large scale screening, the distribution of stability scores of scrambles has also moved to larger values (Figure 4.36).

A large fraction of scrambles now have stability score above 1 (26%), it is therefore adequate to move the threshold that defines stability to higher stability score values. For this threshold we choose stability score 1.5, as only 3% of scrambles are above this value. The number of designs defined as stable with this new threshold is 1641, 32% of all designs, a modest, yet significant, increase from the fraction of stable designs in the previous high-throughput screening (21%).

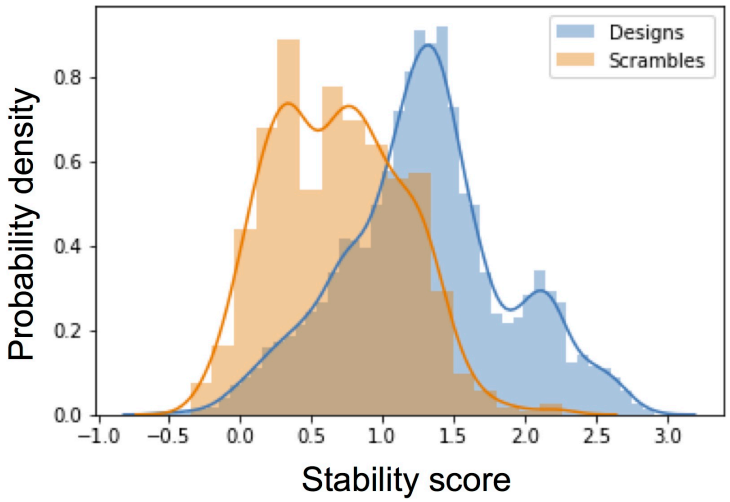


Figure 4.36: Stability score distributions for designs and scrambles for proteins designed by the new generative algorithm.

Among the stable proteins, as defined by the new, more stringent threshold, there are representatives of 242 unique NTF2 parameter combinations, 75% of all tested combinations.

To evaluate the effect of tryptophan/lysine pairs on the short arm, and glycine at highly curved strand positions on protein stability, we compare the stability scores of designs where these features where replaced by other compatible amino-acids (Methods 4.5.16). We obtained stability scores for 100 glycine control pairs, and 198 for tryptophan/lysine controls (Figure 4.37). Replacement of glycine residues does not seem to have a specific effect on protein stability: in most cases there is no effect, and when there is, it can be in favor or not of the designed glycine (Paired T-test p-value: 0.9). There seems to be an increase in stability score for designs with the tryptophan/lysine pair compared to those where it was removed (Paired T-test p-value: $0.2.10^{-8}$), although the magnitude is effect is very small. Despite the lack significant effect from the sequence features introduced to ensure designed strand register on stability score, we consider they may still be important for structural control.

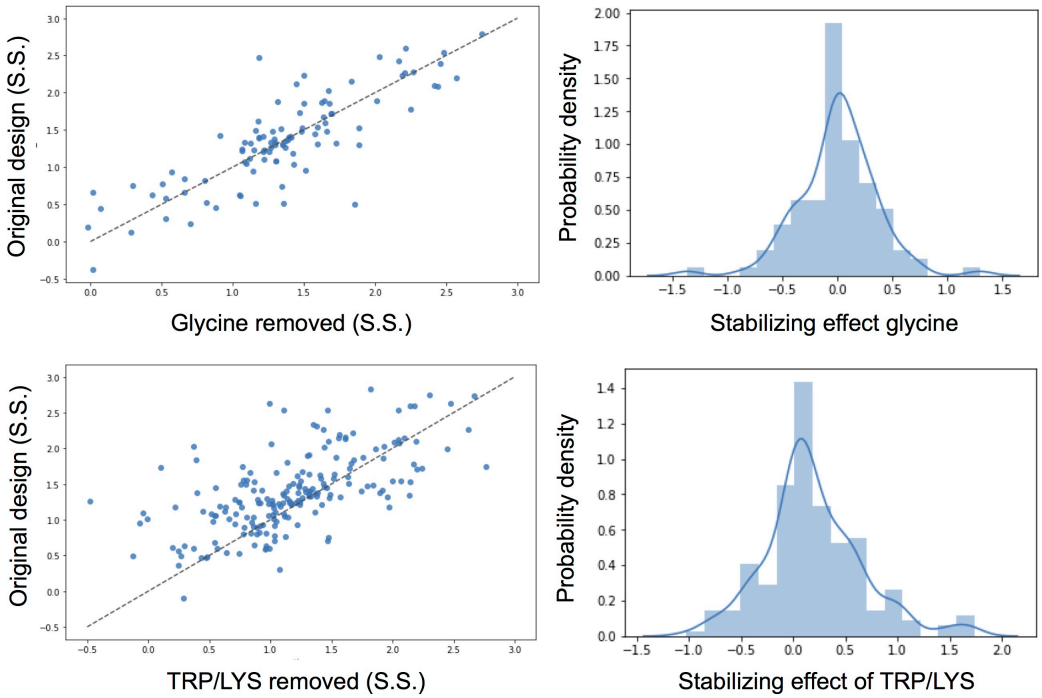


Figure 4.37 Stabilization effect of tryptophan-lysine pairs and strand glycines: Left: Scatter plot of original designs and control stability scores for high-curvature strand positions with glycine (top) and TRP/LYS pairs at short arm (bottom). The dashed line shows the diagonal. Right: Distributions for stabilizing effect (Stability score of original design minus Stability score of control) of high-curvature strand positions with glycine (top) and TRP/LYS pairs at short arm (bottom).

In an effort to understand how to further improve designs, we trained a logistic regression model to distinguish between designs with stability score above or below 1.5. We followed the same methodology as described in Section 4.3.2, and obtained a ranking of features by their weight (Figure 4.38). Surprisingly, the top features continue to be related to hydrophobicity (*hphob_sc_contacts* and *BuriedHyphobSAperRes*), but now focus not on the number but in contact surface between hydrophobic residues. The third feature in terms of importance, *dsc_av*, is derived from TERM analysis (16), and it is correlated to the frequency to which certain amino-acids are observed in the center of a given local region of a protein 3D structure. In the previous experiment, the most important features related to TERM analysis focused on helix 3 and the S5-S6 hairpin, *dsc_av* focuses on sequence instead of backbone conformation abundance, and is an average over the whole protein instead of focused on any particular structural element. Rosetta score remains among the most important features, although now lower in the ranking (*score_res_betacart*). The following terms (*loop3_flank_av_worstfr* and *av_all_strand*) are related to local sequence-structure agreement, specifically in the primary sequence stretches that form strands.

This contrasts with the relatively higher importance of average RMSD of all fragments in all positions (*avAll*) in the previous experiment. Buried unsatisfied polar atoms (*buns_nosurf_sc*) are again detected by the logistic regression model as important. *T1_absq*, the absolute number of charged residues at the first 3 residues of all helices, not related to any feature detected as important in the previous experiment, is now the 8th in importance. The rest of the terms, which have very small weights, are also related to hydrophobic side-chain contacts.

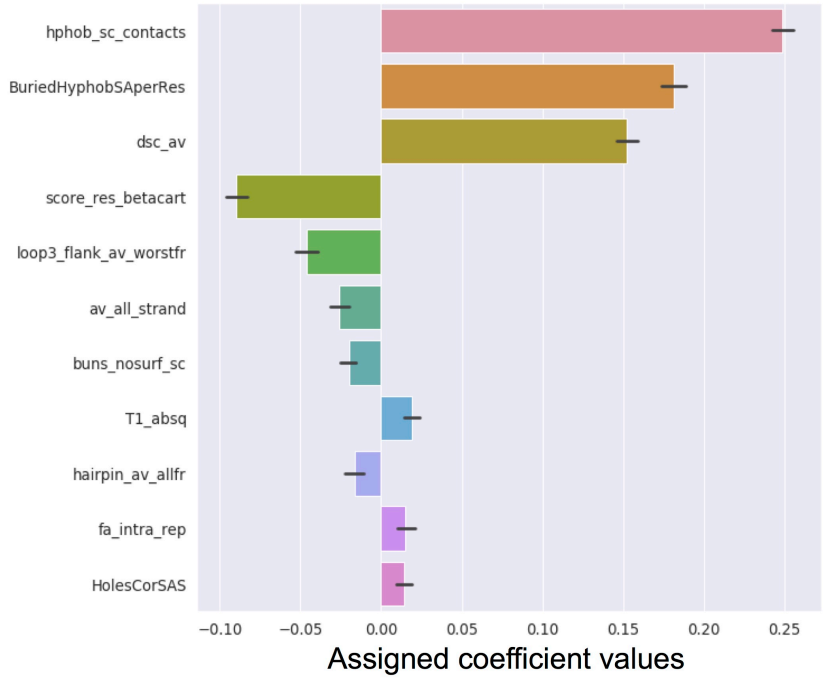


Figure 4.38 Average weights of top 10 features: Bar plot showing average weights of top 10 features with the highest weight absolute values used by 40 different logistic regression models trained with different dataset partitions.

In section 4.3.2 we showed per-enrichment positions were in agreement with the expected hydrophobicity patterns for comparable secondary structure stretches. Figure 4.39 shows the results of the same type of analysis (Methods 4.5.9) on the latest high-throughput screening dataset.

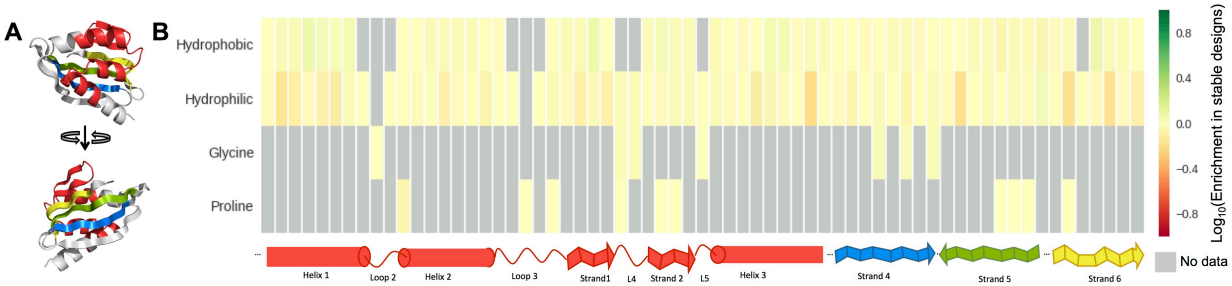


Figure 4.39 Per-position enrichments: **A.** Positions structurally homologous in all tested models, selected for comparison. Colors depict stretches continuous in sequence. **B**. Heat map of enrichment or depletion by position, with amino-acid types grouped in four categories: hydrophobic (AFILMVWY), hydrophilic (DEHKNQRST), proline and glycine. Upward-pointing pleating in strands points towards the core of the protein. Yellow cells indicate no enrichment or depletion, or a non-significant difference.

Differently from the clear per-position enrichment patterns we see for the proteins from the first high-throughput experiment (Figure 4.8), the proteins made with the new generative algorithm show little to no enrichment or depletion in most positions (Figure 4.39). There are a number of possible explanations for this lack of high enrichment values. It is possible the larger structural diversity in this set results in a proportional diversity of environments around the selected positions, blurring the distinction between core, interface and surface positions. Another possible explanation is that the increased stability of these designs allows a higher diversity of residue identities to be tolerated at core and surface positions. Different degrees of combinations of the two proposed explanations are also possible. In any case, a less restricted sequence landscape is advantageous for design of active and binding sites, as they impose sequence and structural restrictions of their own.

To discard the possibility that lack of high enrichment/depletion values in the sequence profile stems from overwhelmingly hydrophobic pockets (a realistic concern, as hydrophobicity was optimized for) we analyzed the polar residue content in pocket positions - defined as inward pointing side-chains near the pocket opening. Figure 4.40 shows the distribution of polar residue fractions in pocket positions of all ordered designs (blue), and in stable designs (orange). More than a quarter of pocket residues (on average, 24 residues are labeled as “pocket” in each design) are polar in most designs, suggesting the reason why they are not depleted is that they are well tolerated.

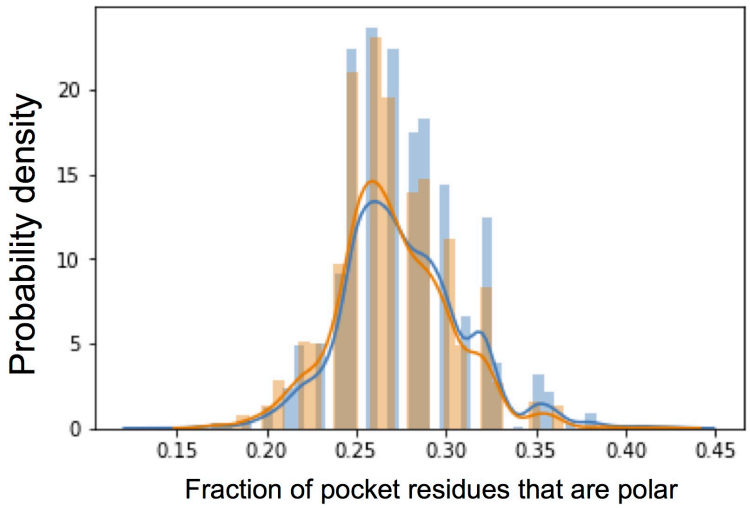


Figure 4.40: Distribution of polar residue fractions in pocket positions of all ordered designs (blue), and in stable designs (orange).

As mentioned in the previous section, the new generative algorithm can generate models with pocket sizes covering the whole range seen in natives. Although we are able to test only a subset of the possible NTF2 parameter combinations, specifically those that result in proteins less than 119 amino acids long, it is worth examining the pocket size of proteins detected as stable by the second high-throughput experiment (Figure 4.41). Designs tested cover most of the range of native pocket sizes and, more interestingly, designs found to be stable also cover this range. For pocket sizes up to 800Å^3, we obtain two orders of magnitude more stable designs than there are natives in the non-redundant set. Between 800Å^3 and 1200 Å^3, we identified a similar number of designs as there are natives in the non-redundant set. We did not test designs with cavity volumes above these values: even though the generative algorithm can produce such designs (Figure 4.32), they are probably too long to be tested with the high-throughput methodology (See methods 4.5.6).

The high-throughput screening of designs generated by the second version of the generative algorithm shows it can design diverse and stable NTF2-like proteins from scratch. Stability score distributions indicate we have successfully improved the design methodology in such a way that it generalizes to a wide set of parameter combinations. It is possible the improved stability of these designs allows more diverse amino-acid composition in is pockets, which is advantageous for functional design. The subset of stable designs obtained from this assay is more than an order of magnitude larger than the non-redundant set of NTF2-like domains, and covers most of its pocket size range.

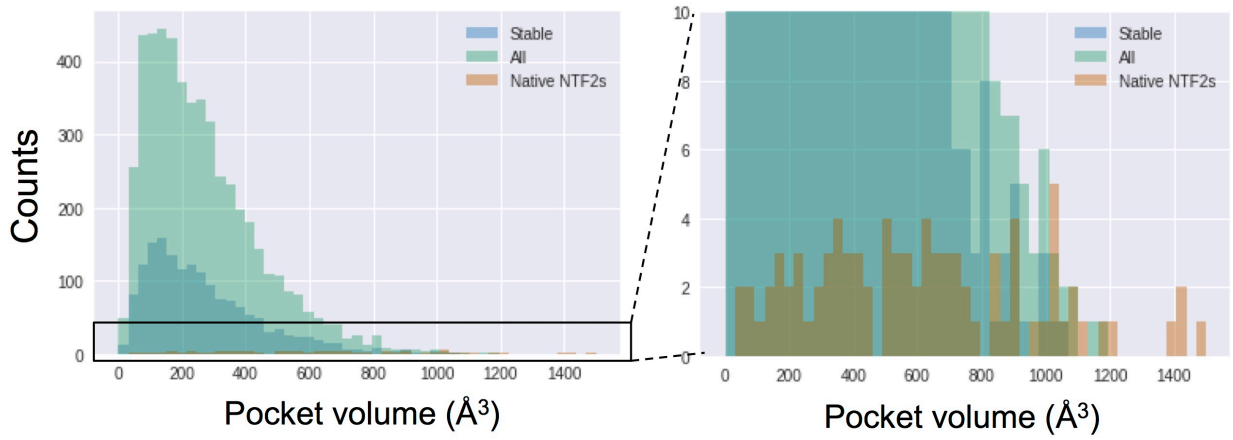


Figure 4.41: Distribution of pocket volumes among all designs tested (All), those detected as stable (Stable), and native NTF2-like domains (Native NTF2s).

In the following section, we biochemically characterize a set of designs and obtain further insights on how to improve our methodology, as we plan to move on to designing functional designs.

4.3.7 Structural validation of proteins generated by the second version of the generative algorithm.

To further characterize the proteins detected as stable by the second round of high throughput screening, we selected a subset of them that have a combination of characteristics we have not designed before, and/or a large pocket. Table 4.11 shows a limited set of parameters for each design that highlight their novelty.

Design Name (short name)	Base width	Long arm length	H3-S3 ABEGO	Hairpin length	Second bulge	Second bulge place	Pocket volume (Å^3)	Stability score
APXUALRM (MC1)	5	3	-	4	True	1	79	2.06
CNOCZZYN (MC3)	3	4	-	2	True	2	277	2.16
IPQZYEHY (MC6)	3	3	-	2	True	1	398	2.1
MQGQLKLY (MC7)	5	3	-	2	True	1	313	1.7
NPHNECCY (MC8)	3	4	-	2	True	2	596	1.55
PVNDHOOV (MC9)	5	2	GBAB	4	False	null	556	1.83
QLNTLIPS (MC10)	3	4	-	2	True	1	507	2.7
QPAJWNJL (MC11)	3	4	-	2	True	1	96	1.7

RWBLOJXV (MC12)	3	4	-	2	True	2	100	2.35
BMZQQOSL (MC2_1)	3	3	GB	2	False	null	383	2.19
CFRZAXWD (MC2_2)	3	3	-	2	True	1	87	2.2
CGBTHRRH (MC2_3)	3	3	BA	4	False	null	561	2.7
DEZFDZKN (MC2_5)	3	4	-	4	True	2	182	1.5
CQXWMZNN (MC2_4)	3	4	BA	4	True	2	283	2.4
DFEBCGLM (MC2_6)	5	3	-	2	True	1	83	2.4
JZXIQIRH (MC2_7)	5	3	-	4	True	1	657	2.7
ODCAZTIO (MC2_9)	5	2	GBAB	2	False	null	151	1.6
UTEWRJFN (MC2_11)	3	3	BBGB	2	False	null	332	1.6
WMNMRJMU (MC2_12)	3	3	BBGB	2	True	1	327	2.5
KVGAMRYX (CAV1)	3	2	GBBA	2	False	null	904	2
MTNNCMGU (CAV2)	3	3	BBGB	2	False	null	831	2.33
OBJWKGFB (CAV3)	3	3	BBGB	2	True	1	981	2.4
QZFIQMXG (CAV4)	5	3	BA	2	False	null	771	2.1
VJZGDPLE (CAV5)	5	3	-	2	True	1	812	2.14
VMXPYKBP (CAV6)	3	4	-	2	True	1	784	2.3

Table 4.11: 25 designs selected for biochemical characterization based on novelty and pocket size.

Table 4.12 summarizes the results of the biochemical characterization (Methods 4.5.10). From the 25 designs considered, 10 are soluble when expressed in *E. coli*, have the correct quaternary structure and are folded proteins that denature in a two-state transition. Those that do not display this behavior are mostly not soluble when expressed in *E. coli*. Figure 4.42 shows the denaturation curves and CD spectra for the 10 stable designs, and Table 4.13 shows the values obtained by fitting sigmoid curves to the data (See methods 4.5.10). The folding free energy values obtained this way are on average higher than for proteins selected from the first high-throughput screening experiment (Figure 4.43, blue dots). The expected m-values for proteins selected from the second high-throughput experiment are also higher on average then for proteins from the previous round, but the measured m-values are not. Figure 4.43 shows free energy of unfolding as a function of stability score for designs from this round (red dots) and the previous one (blue dots). A linear fit to the data returns no significant correlation.

Design Name (short name)	Soluble expression	Within expected SEC EV	Quaternary structure	Folded protein	Denaturation curve
APXUALRM (MC1)	No	-	-	-	-
CNOCZZYN (MC3)	No	-	-	-	-
IPQZYEHY (MC6)	Yes	Yes	Monomer	Yes	Two-state
MQGQLKLY (MC7)	No	-	-	-	-
NPHNECCY (MC8)	No	-	-	-	-
PVNDHOOV (MC9)	Yes	Yes	Monomer	Yes	Two-state
QLNTLIPS (MC10)	Yes	Yes	Monomer	Yes	Two-state
QPAJWNJL (MC11)	Yes	No	-	-	-
RWBLOJXV (MC12)	No	-	-	-	-
BMZQQOSL (MC2_1)	Yes	Yes	Monomer	Yes	Two-state
CFRZAXWD (MC2_2)	Yes	Yes	Monomer	Yes	Two-state
CGBTHRRH (MC2_3)	Yes	Yes	Monomer	Yes	Two-state
DEZFDZKN (MC2_5)	No	-	-	-	-
CQXWMZNN (MC2_4)	Yes	Yes	Monomer	Yes	Two-state
DFEBCGLM (MC2_6)	Yes	Yes	Dimer	-	-
JZXIQIRH (MC2_7)	Yes	Yes	Monomer	Yes	Two-state
ODCAZTIO (MC2_9)	No	-	-	-	-
UTEWRJFN (MC2_11)	No	-	-	-	-
WMNMRJMU (MC2_12)	Yes	Yes	Monomer	Yes	Two-state
KVGAMRYX (CAV1)	Yes	Yes	Dimer	-	-
MTNNCMGU (CAV2)	Yes	No	-	-	-
OBJWKGFB (CAV3)	Yes	No	-	-	-
QZFIQMXG (CAV4)	Yes	Yes	Dimer	-	-
VJZGDPLE (CAV5)	Yes	No	-	-	-
VMXPYKBP (CAV6)	Yes	Yes	Monomer	Yes	Two-state

Table 4.12: Results of biochemical characterization for new designs selected from among those detected as stable in the second high-throughput screening experiment.

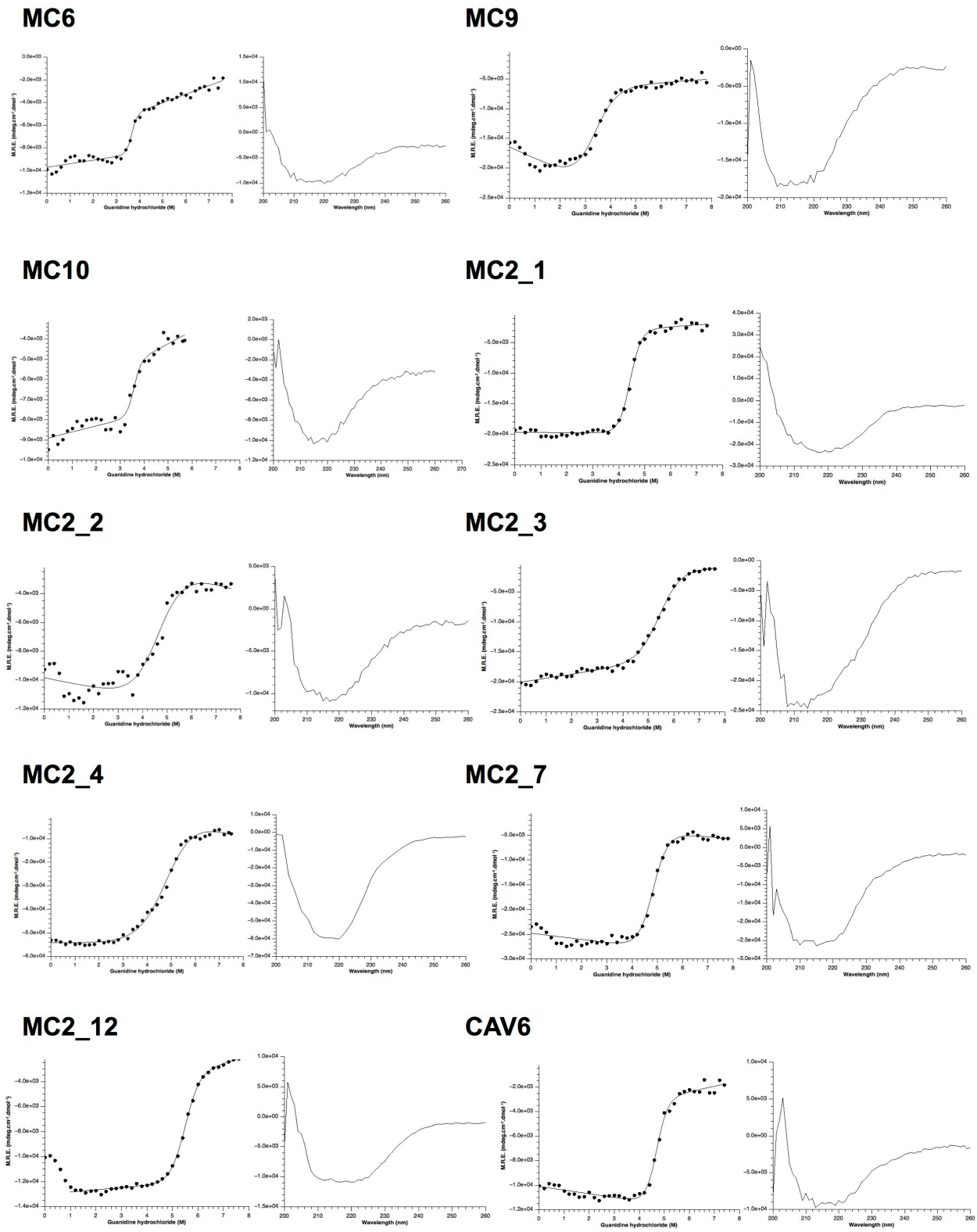


Figure 4.42 Denaturation curves and circular dichroism spectra: Denaturation curves and circular dichroism spectra for proteins selected from the second round of high-throughput screening that showed two-state unfolding behavior in titrations with guanidine hydrochloride.

Design Name	$\Delta G_{folding}$ (kcal/mol)	Stability score	Experimental m-value ($kcal.mol^{-1}.M^{-1}$)	Expected m-value ($kcal.mol^{-1}.M^{-1}$)
IPQZYEHY (MC6)	-18	2.1	4.9	5.01
PVNDHOOV (MC9)	-4.6	1.8	1.33	5.03
QLNTLIPS (MC10)	-13	2.7	3.8	5.23
BMZQQOSL (MC2_1)	-12	2.19	2.8	5.16
CFRZAXWD (MC2_2)	-5.3	2.2	1.13	5.01
CGBTHRRH (MC2_3)	-7.6	2.7	1.39	5.3
CQXWMZNN (MC2_4)	-5	2.4	1.03	5.22
JZXIQIRH (MC2_7)	-10	2.7	2	5.17
WMNMRJMU (MC2_12)	-11	2.5	2.1	5.01
VMXPYKBP (CAV6)	-12.7	2.3	2.7	5.08

Table 4.13: Thermodynamic parameters obtained from fitting guanidine hydrochloride denaturation curves. Expected m-values are calculated using the buried surface area based on protein models and the equations described in (12).

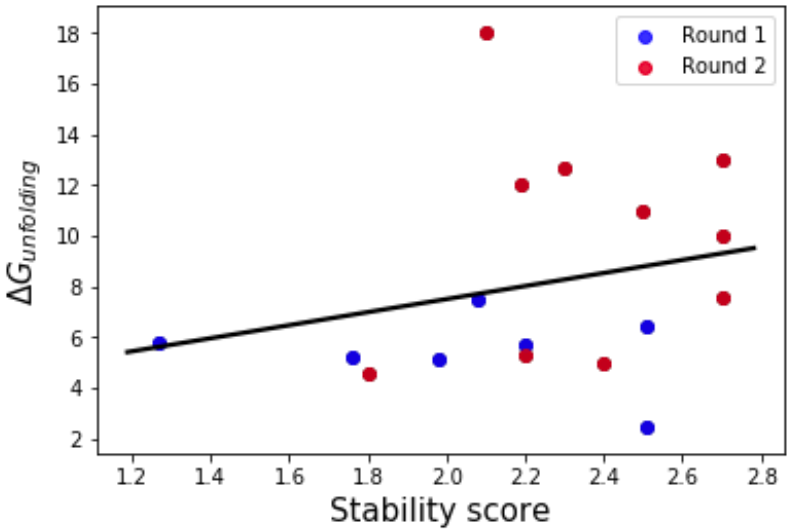


Figure 4.43: Free energy of unfolding as a function of stability score for designs from the first (blue dots) and second round (red dots). The black line is the best linear fit to the data (Pearson R: 0.25). The correlation is not significant (p-value: 0.346).

In summary, we show that proteins detected as stable in the second round of high-throughput screening, when able to express then in *E. coli*, have the characteristics of well-folded proteins. Furthermore, the well-folded proteins sample a diverse set of parameters that had not been sampled before. Structural studies will be necessary to evaluate to what extent these proteins recapitulate their modeled structure. In the next section we present results showing the utility of *de novo* NTF2-like proteins for design of functional proteins.

4.3.7 Using de novo NTF2-like proteins as scaffolds to design an aflatoxin-B1-binding protein.

As we mentioned in the introduction to this thesis, the main motivation of this work is devising new methods for generating protein structural diversity at a large scale, such that arbitrary catalytic and binding sites can be designed. In this section we explore the *de novo* NTF2-like functional proteins.

An interesting challenge for protein design that could benefit from a larger diversity of small stable scaffolds with pockets is small-molecule binding design (18). Here we use the knowledge gained from successive rounds of high-throughput screening to design proteins completely from scratch that have a binding site for a small molecule. Despite of this approach being potentially challenging for not relying on a scaffold with validated stability, successful small-molecule binder design in this context would show that the large diversity of *de novo* NTF2 scaffolds generated by the generative algorithm can be used without going through burdensome experimental characterization.

We focus on Aflatoxin B1 as a target for our small-molecule binding efforts (Figure 4.44). Aflatoxin B1 belongs to a family of mycotoxins produced by the *Aspergillus* fungi that can cause chronic and acute liver problems in mammals. Aflatoxin contamination of livestock feed results in large economical losses, especially in tropical locations where proper storage of feed and detection of the toxin are lacking (19). Small, easily produced stable proteins able to bind Aflatoxin B1 could be used in specific sensors to detect it, or as remediation agents by encapsulation.

To design an Aflatoxin B1-binding protein, we produced docked conformations of it on a set of 500 scaffolds longer then 119 amino acids (As described for the first design of cortisol binders in 4.5.2). This set was obtained by clustering all long (>119 amino acids) scaffolds produced by their descriptive sheet features (Figure 4.19) and construction parameters (See Section 4.3.4). From nearly 5000 docked conformations, we selected those where the docked Aflatoxin molecule had a channel of entry to the protein, and had better interaction scores better than a standard deviation. We gathered the unique parameter combinations observed in the selected subset and generated more scaffolds with those same parameter combinations. We went through the same docking process and obtained nearly 7300 docked

configurations where the small molecule was exposed to solvent to some extent. We then optimized local interactions with the small molecule.

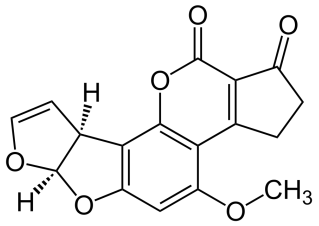


Figure 4.44: Skeletal formula of Aflatoxin B1. For modeling purposes, only carbonyl oxygen atoms are treated as hydrogen-bond acceptor polar atoms, ether oxygen atoms are treated as non-polar.

The optimization of the binding site is done in several stages. Briefly: First, we select the subset of residues surrounding the small molecule and use the Rosetta score function and packing to find interactions that optimize protein stability and interaction with the ligand. Mutations introduced by the docking protocol are not necessarily kept during this process, as they may contain clashes with the ligand or other residues. The information provided by the docking process is mainly the location of the molecule inside the protein. During the second stage, fully-satisfied hydrogen-bond networks are detected, involving polar atoms in the ligand that are not exposed to solvent. This step is critical to ensure polar atoms are not shielded from solvent without being hydrogen bonded, as this would make such conformation highly unfavorable. The last stages design the outer interaction shells optimizing the features detected as important by logistic regression (for the full protocol see methods 4.5.16).

Figure 4.45 shows a design generated using the proposed method. Figure 4.45 A shows polar interactions with the ligand, embedded in a large hydrogen-bond network. Figure 4.45 is a surface rendering of the protein showing the ligand lodged in a pocket with high shape complementarity.

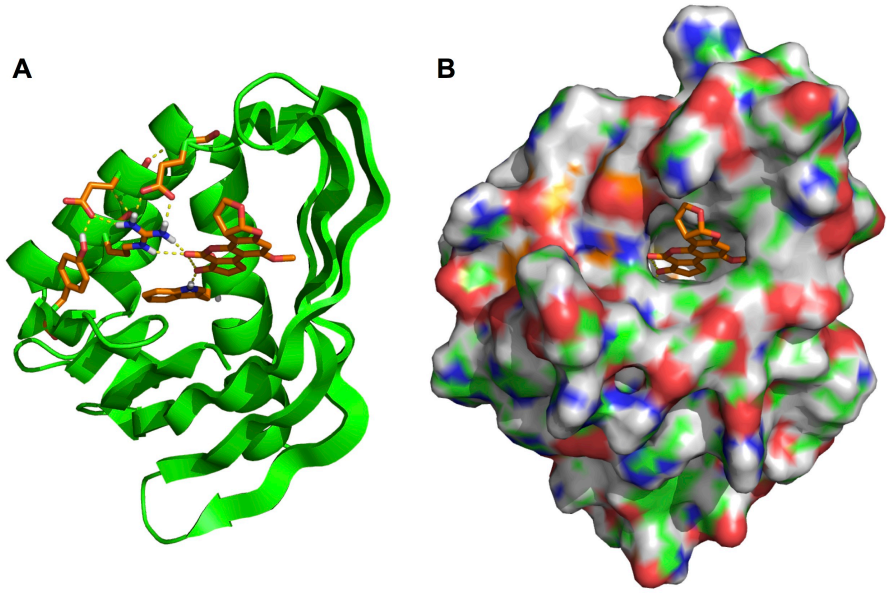


Figure 4.45 A: Model for design LAFL5, showing backbone as cartoon (green), ligand and side-chains involved in hydrogen bond network as sticks (orange). Hydrogen bonds are shown as yellow dotted lines. **B**. Surface rendering of design LAFL5, showing ligand lodged in binding pocket, partially exposed to solvent.

From all the proteins designed using the proposed methodology, we chose 8 of them that had the best Rosetta score, no unsatisfied buried polar atoms, and good shape complementarity with the ligand. As the expectation for these proteins is to eventually be part of a device with long shelf life, or be exposed to harsh environments, we designed single disulfide bridges in all 8 of them. We expressed these proteins in *E. coli* and characterized them. The results of this characterization are presented in table 4.14. Two of the tested proteins are soluble, folded, and have the correct quaternary structure. One of them has a gradual denaturation curve despite of being folded at room temperature. Figure 4.46 shows the denaturation curves and CD spectra for LAFL5 and LAFL6.

Design Name	Soluble expression	Within expected SEC EV	Quaternary structure	Folded protein	Disulfide is formed	Denaturation curve
LAFL1	No	-	-	-	-	-
LAFL2	No	-	-	-	-	-
LAFL3	No	-	-	-	-	-
LAFL4	No	-	-	-	-	-
LAFL5	Yes	Yes	Monomer	Yes	Yes	Two-state
LAFL6	Yes	Yes	Monomer	Yes	Yes	Gradual
LAFL7	No	-	-	-	-	-
LAFL8	No	-	-	-	-	-

Table 4.14: Experimental characterization of designs for Aflatoxin B1 binding.

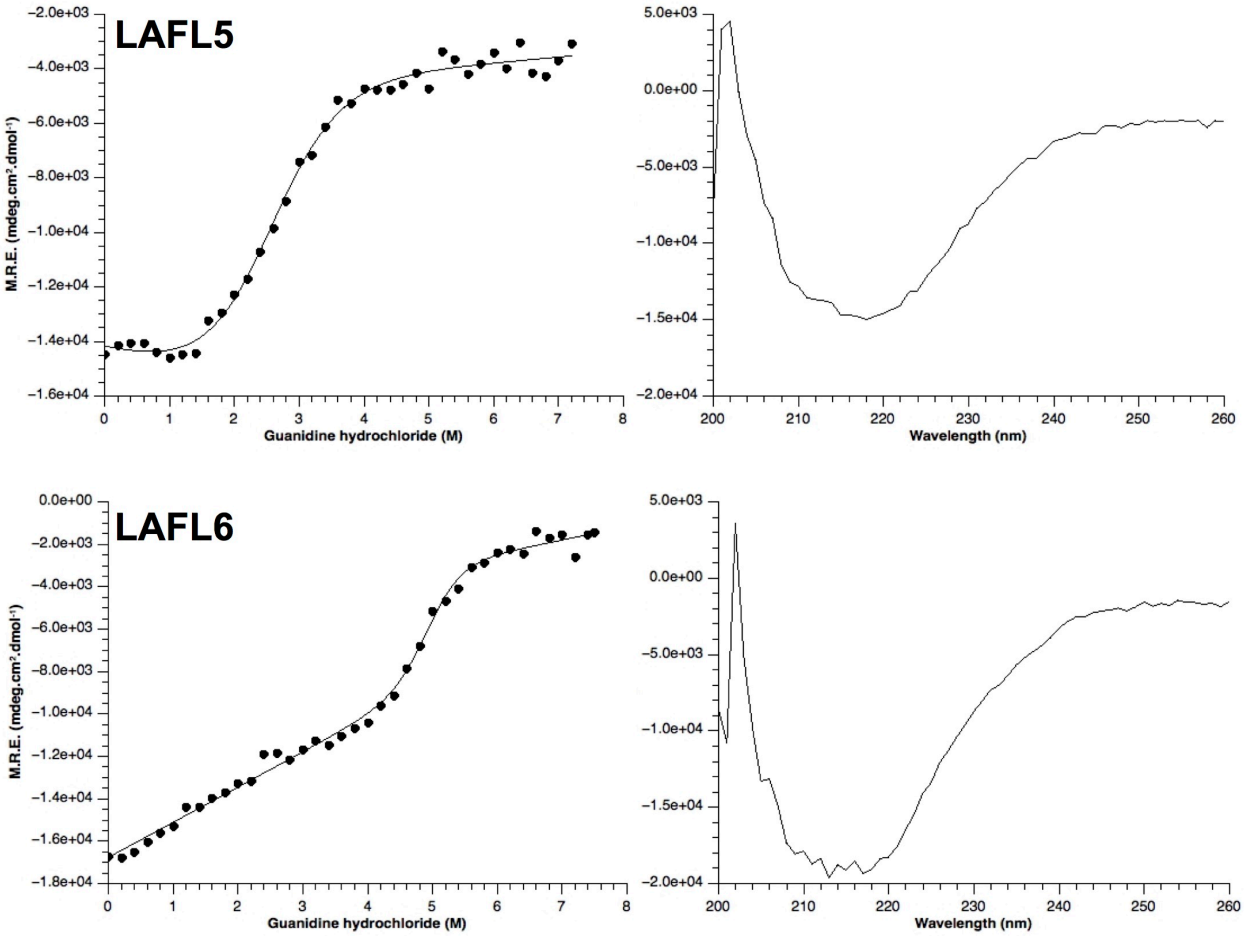


Figure 4.46: Denaturation curved and CD spectra for designs for Aflatoxin B1 binding LAFL5 and LAFL6. Fits to denaturation curves yield free energy of folding of -2.8 kcal/mol and -11kcal/mol for LAFL5 and LAFAL6 respectively. The m-values obtained are 1.12 and 2.24 for LAFL5 and LAFAL6 respectively.

We tested Aflatoxin B1 binding on LAFL5 and LAFL6. We chose to test LAFL6 for binding despite its gradual denaturation curve, as we hypothesized a protein with a large pocket may not necessary have a large well-packed core. We tested Aflatoxin binding by equilibrium dialysis (See methods 4.5.17). LAFL5 did not show signs of binding. As shown in figure 4.47, LAFL6 binds Aflatoxin B1 with a dissociation constant of around 50μM. Although this affinity is not immediately useful for field applications, it provides a starting point for improvement. Further insight on how to improve affinity should come from mutagenesis studies, as well as empirical structures obtained from X-ray crystallography. More importantly, this result shows small-molecule binding proteins can be obtained from scaffolds generated by the second version of the generative algorithm, directly from computational modeling, without relying on previous scaffold validation.

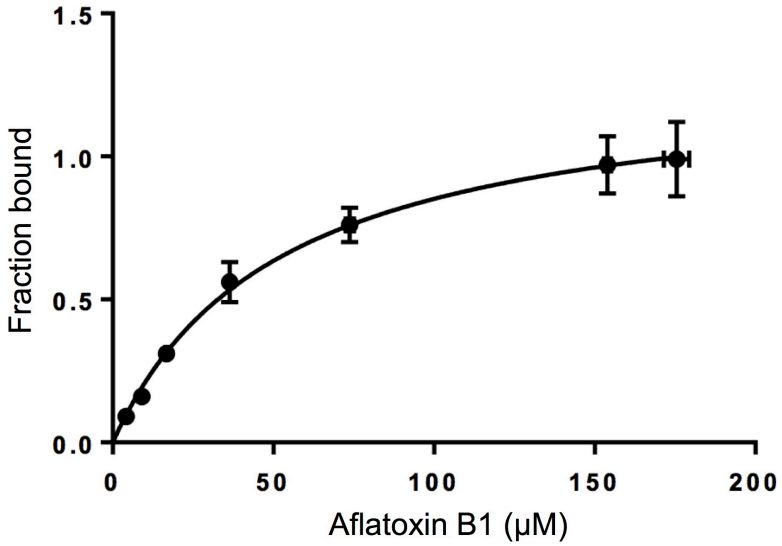


Figure 4.47: Aflatoxin B1 binding curve to LAFL6 obtained from equilibrium dialysis experiments. The solid line is the fit of a single site binding equation to the data, with K_d 51μM.

4.4 Discussion

In the previous chapter we described the principles for designing curved beta-sheets, and showed their applicability by designing a number of complex folds, including several of the NTF2-like superfamily. Although these designs explore a range of curved sheet conformations, the wide and systematic generation of diverse structures necessary for functional design was lacking. In this chapter we gather information that enables design of large numbers of stable NTF2-like proteins, and apply this knowledge in the construction of an NTF2-like protein generative algorithm that produces a wide variety of NTF2-like proteins.

The first rudimentary version of the generative algorithm is a collection of backbone-building algorithms, like the ones described in Chapter 3, with additional parameter combinations, totaling 9 subfamilies (Figure 4.1). Although the proteins generated by these algorithms are not a substantial departure from those seen in nature in terms of structure (Figures 4.3 and 4.4), they serve as a starting point to understand the determinants of stability in NTF2-like proteins. Using high-throughput screening methods, we evaluate the stability of nearly 6000 *de novo* NTF2-like proteins, and obtain insights on what features distinguish between stable and unstable designs. Hydrophobic side-chain packing and Rosetta score are the most predictive features, followed by agreement between local structure and sequence, and

hydrogen-bond satisfaction of buried polar side-chain atoms (Figure 4.7). The detection of these features as predictive of stability is in line with our understanding of protein biochemistry, but points at a lack of fine-tuning of these properties in our designs: Although the proteins tested were designed to have hydrophobic cores and optimized side-chain packing, the extent to which this was done was insufficient in many cases. This is particularly true for buried unsatisfied polar atoms, which are explicitly penalized by the Rosetta score function, but were still present in designs. We hypothesize the importance of features related to local sequence-structure agreement (fragment-related and TERM-related) is associated to the entropic cost of folding, as sequences with strong tendencies to form a given type of secondary structure are potentially more likely to be preordered during folding.

The large numbers of sequences tested in this first high-throughput screening experiment enabled per-position analysis of amino-acid type enrichments. The enrichment or depletion of hydrophobic and hydrophilic amino-acids in structurally homologous positions in stable designs when compared to the whole population (Figure 4.8) shows a clear pattern: hydrophilic amino-acids are strongly depleted in positions pointing inwards, towards the concave side of the sheet, while hydrophobic amino-acids are generally well tolerated in most positions, and even enriched on the outward-facing flat side of the sheet. This pattern is even visible around the S6 bulge, where the alternating directionality of strand side-chains is broken. The low tolerance to polar residues at inward (pocket) positions reveals, on one hand, that most of the stable proteins are likely folded with the expected positions pointing towards the core, as designed. On the other, it reveals it would likely be difficulty to install any active or binding sites involving polar interactions in these proteins.

Characterization of single stable designs obtained from the first high-throughput screening experiment offered further insights on determinants of NTF2 stability and structure. Almost all designs that expressed solubly in *E. coli* are folded, monomeric proteins with a single optimal structure, as shown by the two-state transition during denaturation (Figure 4.9). Four of the designs characterized this way belong to subfamilies we have not obtained stable designs for before, Mk2 and 4B.7. The lack of soluble expression in *E. coli* for over a half of the designs is puzzling, and could be attributed to a number of reasons: It is possible the expression pathway for chimeric expression on Aga II during high-throughput screening on yeast surface offers a better environment for folding than high intracellular expression in *E.*

coli. Or, there is a hidden “false positive mode” of the screening system, in which unfolded designs, but not scrambles, appear as protease-resistant with high stability scores. We believe the first explanation is more parsimonious than the second.

We obtained crystal structures for two designs belonging to the Mk2 subfamily. One of them recapitulated the model to great detail (4.12); the other (BBM2nHm0589) had significant deviations from the model (Figure 4.13). The unexpected crystal structure of BBM2nHm0589 offers critical insight on determinants of NTF2 structure. In the crystal structure, a 2-residue register shift between strands 5 and 6 not present in the model brings the beta-bulges closer together and flattens the sheet structure, with all residues originally modeled as part of the core still pointing towards the concave side of the sheet (Figure 4.13). We hypothesized the introduction of glycine at a critical position, in combination with other mutations, would stabilize the conformation originally designed, as observed in beta-barrels (13) and some native NTF2-like domains. We showed that modeling such mutations indeed improves sheet bending (Figure 4.14). The crystal structure of the 5-fold mutant recapitulated the structure of the model to great detail, with all introduced side-chains in the modeled conformations. These results highlight the importance of implementing strategies to stabilize modeled conformations over other possible ones (i.e., negative design), and inform the design for the next generation of *de novo* NTF2-like proteins.

With a better understanding of the determinants of stability and structure of NTF2-like proteins we focused on the systematic generation of diverse proteins from this fold. We developed a generative algorithm that explores all productive structural degrees of freedom seen in native NTF2-like domains, and expands on them using known protein-design principles (Section 4.3.4). The proteins generated by this algorithm cover a large space of the natives and, present significant departures from it, even by a conservative measure (Figure 4.31). If we consider the generative algorithm can create arbitrary numbers of any of the 1503 different parameter combinations, it is possible to assume the coverage of native space is even better, and the novel structures more numerous than our conservative calculations. All the information gathered in the first high-throughput experiment was integrated in the sequence design protocol, including logic that places glycine residues at highly bent strand positions.

Using the same high-throughput techniques as before, we tested the stability of proteins representing a large subset of structural parameter combinations that were compatible with the gene assembly

technology. We see a dramatic increase in the number of stable proteins, as predicted by a logistic regression model based on design features (Figures 4.35 and 4.36). Interestingly, the stability of unfolded sequences also increased, making it necessary to change the threshold over which a protein is labeled as stable. Even with a higher bar to clear, we see a 50% increase in the percentage of stable proteins compared to the initial round, which cover 75% of the tested 323 different structural parameter combinations. The design features that predict stability above the new threshold are similar to those detected in the previous experiment, with less influence from fragment-quality metrics (Figure 4.38).

The hydrophobicity pattern of per-position enrichments shows increased tolerance to polar amino-acids in the pockets of the proteins designed using the new generative algorithm (Figure 4.39). Given the significant proportion of polar amino acids in pocket positions (Figure 4.40), we attribute their tolerance to increased design stability, which enables replacement of hydrophobic interactions with polar side chains. The ability to place a significant number of polar side-chains in the pocket of *de novo* NTF2-like proteins opens the door to the design of complex active and binding sites.

We compared the distribution of pocket sized in the proteins designed using the new generative algorithm to native NTF2-like domains (Figure 4.41), and verify that we cover the majority of the range seen in natives. Furthermore, the number of new designs labeled as stable in the high-throughput experiment covers a similar range of sizes, and exceeds the number of native NTF2-like domains in the non-redundant structure set by an order of magnitude for more than a half of the size range (Figure 4.41, right).

As done for proteins designed by first version of the generative algorithm, we selected a few interesting cases for biochemical characterization. All selected proteins have stability scores above 1.5, and at least one structural feature not sampled before (Table 4.11). We found most proteins solubly expressed in *E. coli* are folded and have a single energy minimum (Figure 4.42).

The final goal of designing highly diverse, stable proteins is generating the conformational space required to construct precise binding and active sites. Using the knowledge gained by high-throughput screening and validated algorithms we designed a protein from scratch that is able to bind Aflatoxin B1, a widespread food contaminant, with a K_d of 50μM (Figure 4.47). Although we obtained modest binding

from one of the two folded proteins we designed, most of the ones we designed did not express well in *E. coli* or aggregated, highlighting there are still challenges to be addressed.

Further improvements to the generative algorithm should come from the elucidation of structures generated by the second version, and implementing methods to improve the design features correlated with stability.

To better understand the causes of inconsistent *E. coli* expression across designs it may be worth pursuing refolding studies, to distinguish between unstable proteins and those for which an aggregation-prone folding intermediate is captured in inclusion bodies.

Although resistance to protease digestion has been linked to folding free energy, we see no correlation between stability score as calculated in this study, and folding free energy (Figure 4.43). We consider this lack of correlation stems from the narrow range of stability scores of the proteins for which we determined free folding energy. To address this issue it might be worth determining the folding free energy of proteins with a wider range of stability scores.

The strategy we employ to construct the more general version of the generative algorithm, where we break up a protein fold into well-understood pieces, and organize them in a system that makes sure they come together harmonically in as many ways as possible, can potentially be adapted to other folds, specially now that principles for designing curved sheets have been described. Folds as simple as the G-protein fold, or as complex as the alpha/beta hydrolase fold can be broken into simple components and their structural degrees of freedom explored systematically. The guidelines of this process, implicitly enunciated in this dissertation, are: 1) Understand the common structural elements that are part of all proteins in the target family. 2) Identify the axle/s or spine/s of the fold, which work as a hub to connect all the basic elements, and establish the assembly strategy bases on it. 3) Analyze how elements relate and change together. 4) At all previous points reduce and adapt pieces and connections to simple structures and rules that are well understood.

4.5 MATERIALS AND METHODS

4.5.1 De novo NTF2 backbone generation and sequence design for the first round of high-throughput screening

Backbones were constructed as described in Chapter 3, Methods sections 3.5.1 and 3.5.2. For subfamilies not described in said sections, the same backbone construction algorithms were used, but parameters were changed accordingly. Scripts for producing these backbones can be found at https://github.com/basantab/NTF2Analysis/NewSubfamiliesGeneration.

The files to design the sequence on the designs for the first round can be found in the above-mentioned Git repository. Briefly: The design protocol begins by generating 4 different possible sequences using the Rosetta FastDesign mover in core, interface and surface layers separately. Then random mutations are tested, accepting only those that improve secondary structure prediction without worsening score, introducing Ramachandran outliers or worsening the shape complementarity between helices and the rest of the protein.

4.5.2 Design of cortisol binding sites in proteins of the Mk1.PeCH subfamily

Cortisol was docked in designs of the Mk1.PeCH subfamily using RIFDOCK as described in (13). Briefly: A Rotamer interaction field was generated around a model cortisol molecule and then used to place cortisol, along with favorable polar and hydrophobic interactions, in the pocket of predesigned Mk1.PeCH proteins. Because all proteins from the same subfamily have the same position numbering (i.e., same length), positions pointing towards the concave side of the sheet were selected by hand, and the same position was used to dock the ligand in all scaffolds. A similar design method as shown in 4.5.1 was used afterwards to redesign the protein to better accommodate the designed pocket.

4.5.3 Protein-protein alignment by TM-align

For each alignment, TM-align optimizes and reports TM-score, a measure of the distance between C_a carbons of aligned residues in target and template, normalized by protein length. The optimization

algorithm used by TM-align results in alignments where superposition of segments with similar local structure is optimized over superposition of segments with disparate local structure. Because TM-score is normalized by target length, and we align proteins with similar, but not equal, lengths, for any given alignment, the TM-score we report is the average between two values.

4.5.4 Dendrogram generation for structural comparison

To construct a dendrogram using protein-protein distances, we first computed the TM-score between all pairs of proteins (See Methods 4.5.3). As dendrogram construction requires distances, we used 1-TM-score, which is higher as proteins are structurally more different. The code for generating figures 4.3, 4.4, 4.31 and 4.33 can be found at https://github.com/basantab/NTF2analysis/tree/master/Dendrogram_generation

Briefly, pairwise distances (TMscore-1) are used to generate a linkage matrix by hierarchical clustering, using the average method. This linkage matrix is then provided to the plotting function to draw the dendrogram.

4.5.5 Sequence clustering by similarity using Clustal Omega and the scipy Python library

Sequence clustering was done based on sequence similarity as measured by Clustal Omega (20). This sequence alignment software is widely used for multiple sequence alignment. As part of its functionality, Clustal Omega produces a sequence similarity matrix that can be used for clustering. Here is a command line example using the *.fasta file “example.fasta”, containing several fasta-formatted sequences:

```
clustalo -i example.fasta --outfmt clustal --outfile out.clustal --output-order tree-order --seqtype protein --distmat-out=out.dst --full
```

This will produce the distance matrix “`out.dst`”, which can then be read by the cluster.hierarchy.linkage function in the scipy Python library to produce a linkage matrix and, finally, clusters. We chose to use the “average” method for generating the linkage matrix, and generated flat

clusters from it using only the upper bound in the number of clusters as a criterion. Code for making the cluster based con Clustal Omega distance matrices can be found at https://github.com/basantab/NTF2analysis/tree/master/SequenceSimCluster/get_clusters_pairwise_dists.py.

4.5.6 Design of gene fragments for multiplex gene assembly

In order to obtain full-length genes from fragments synthetized in DNA microarrays, they must be assembled from halves, as described in (7). To generate highly orthogonal overlaps, we generated DNA sequences using DNAWorks (21), then split the gene in half and altered the composition of the around 20 overlapping nucleotides to have as low homology as possible with other halves in the pool, while maintaining an adequate melting temperature and GC content, and staying below the maximum oligonucleotide length (230 nucleotides). An optimized version of the algorithm described in (7) can be found at https://github.com/basantab/OligoOverlapOpt.

4.5.7 Features calculated for de novo NTF2 design stability prediction

Scripts for extracting design features used in logistic regression model training can be found in the public GitHub repository: https://github.com/basantab/NTF2analysis in the feature_extraction folder. For features described in (6) extracted with specialized code, refer to the supplementary material of that publication.

Features calculated using Rosetta filters and score function ref2015 (when dependent on score function):

Feature name	Explanation
Holes	Rosetta filter “Holes”, described in (22), using default values.
HolesCorSCN	Rosetta filter “Holes”, but only for core atoms, with core defined by the number of side-chain neighbors.
HolesCorSCNnBB	Rosetta filter “Holes”, but only for core atoms, with core defined by the number of side-chain neighbors, and not taking into account backbone atoms
HolesCorSAS	Rosetta filter “Holes”, but only for core atoms, with core defined by the solvent accessible surface area of side-chains.
HolesCorSASnBB	Rosetta filter “Holes”, but only for core atoms, with core defined by the

	solvent accessible surface area of side-chains, not including backbone atoms.
nres	Length of the protein in amino-acids
cavity_vol	Rosetta “CavityVolume” filter with default values
BuriedHyphobSA	Buried surface area of all atoms in hydrophobic residues (FAMILYVW) as calculated by the “BuriedSurfaceArea” Rosetta filter.
BuriedHyphobSA_H	Buried surface area of all atoms in hydrophobic residues (FAMILYVW) as calculated by the “BuriedSurfaceArea” Rosetta filter. In helices only.
BuriedHyphobSA_E	Buried surface area of all atoms in hydrophobic residues (FAMILYVW) as calculated by the “BuriedSurfaceArea” Rosetta filter. In strands only.
BuriedHyphobSA_L	Buried surface area of all atoms in hydrophobic residues (FAMILYVW) as calculated by the “BuriedSurfaceArea” Rosetta filter. In loops only.
BuriedHyphobSA2_H	Buried surface area of all atoms as calculated by the “BuriedSurfaceArea” Rosetta filter. In helices only.
BuriedHyphobSA2_E	Buried surface area of all atoms as calculated by the “BuriedSurfaceArea” Rosetta filter. In strands only.
BuriedHyphobSA2_L	Buried surface area of all atoms as calculated by the “BuriedSurfaceArea” Rosetta filter. In loops only.
nres_aro	Number of aromatic residues in the protein
nres_aro_E	Number of aromatic residues in the protein strands
nres_aro_H	Number of aromatic residues in the protein helices
nres_aro_L	Number of aromatic residues in the protein loops
nres_H	Number of residues in the protein helices
nres_E	Number of residues in the protein strands
nres_L	Number of residues in the protein loops
nres_aro_per_res	Number of aromatic residues in the protein, divided by its length
nres_charge	Number of charged residues in the protein
nres_hydrophob	Number of hydrophobic residues in the protein
nres_hydrophob_noA	Number of hydrophobic residues in the protein, not counting alanine
nAla	Number of alanine residues in the protein
nres_H_per	Fraction of residues in helices
nres_E_per	Fraction of residues in strands
nres_L_per	Fraction of residues in loops
nres_charge_per	Number of charged residues in the protein divided by its length *100
nres_hydrophob_per	Number of hydrophobic residues in the protein divided by its length*100
nres_hydrophob_noA_per	Number of non-alanine hydrophobic residues in the protein divided by its length*100
nAla_per	Number of alanine residues in the protein divided by its length*100
BuriedHyphobSAperRes	Buried surface area of all atoms in hydrophobic residues (FAMILYVW) as calculated by the “BuriedSurfaceArea” Rosetta filter, divided by protein length
total_score	Total Rosetta score (calculated by ref2015)
scoreRes	Total Rosetta score (calculated by ref2015) divided by protein length
ramaRes	Total rama Rosetta score term (calculated by ref2015) divided by protein length
fa_atr	Total fa_atr Rosetta score term (calculated by ref2015)
fa_atrRes	Total fa_atr Rosetta score term (calculated by ref2015) divided by protein length
fa_repRes	Total fa_rep Rosetta score term (calculated by ref2015) divided by protein length
charge	Absolute protein charge (Assuming typical amino-acid behavior at pH7)
hx_sc	Shape complementarity between helice and the rest of the protein (See SSShapeComplementarityFilter filter in Rosetta)
longestPS	Length of the longest continuous stretch of polar amino-acids
longestPS_H	Length of the longest continuous stretch of polar amino-acids in helices
longestPS_E	Length of the longest continuous stretch of polar amino-acids in strands

longestPS_L	Length of the longest continuous stretch of polar amino-acids in loops
exposedHyphob	Number of solvent-exposed hydrophobic residues (See ExposedHydrophobics filter in Rosetta)
SSmismatch	Nth root of the productory of all residue probabilities of NOT being in the modeled secondary structure state, as calculated by PSIPRED (23), where N is the length of the protein. See the SSPrediction filter in Rosetta.
hb_lr_bb_per_res	hb_lr_bb Rosetta score term, divided by protein length
hb_lr_bb_per_sheet	hb_lr_bb Rosetta score term, divided by the number of residues in sheets
hb_sr_bb_per_helix	hb_sr_bb Rosetta score term, divided by the number of residues in helices
av_loop_rama_prepro	rama_prepro Rosetta score term in loops, divided by the number of residues in loops
av_loop_p_aa_pp	p_aa_pp Rosetta score term in loops, divided by the number of residues in loops
av_rama_pp_loop	av_loop_rama_prepro+ av_rama_pp_loop
geom_res	Number of residues with large deviations of Omega angle from planarity
AvDeg	Average number of residues in contact with each position in the protein (See AverageDegree filter in Rosetta)
arom_in_core_SS_SCN	Number of aromatic amino-acids in core positions of non-loop secondary structure elements, with core defined by the number of side-chain neighbors
arom_in_core_SS_SASA	Number of aromatic amino-acids in core positions of non-loop secondary structure elements, with core defined by solvent accessible surface area
hyphob_in_core_SS_SCN	Number of hydrophobic amino-acids in core positions of non-loop secondary structure elements, with core defined by the number of side-chain neighbors
hyphob_in_core_SS_SASA	Number of hydrophobic amino-acids in core positions of non-loop secondary structure elements, with core defined by solvent accessible surface area
core_SCN	Number of core residues, with core defined by the number of side-chain neighbors
core_SASA	Number of core residues, with core defined by solvent accessible surface area

Table 4.5.7.1: Design features based on Rosetta filters with score function ref2015

Features calculated using Rosetta filters and beta_nov16 score function (when dependent on score function):

Feature name	Explanation
score_res	Total Rosetta score divided by protein length.
score_res_betacart	Total Rosetta score divided by protein length, with score calculated taking into account deviations from ideal covalent bonds angles and lengths.
hyphob_contact	Number of carbon-carbon atomic contacts between hydrophobic residues.
hphob_sc_contacts_rta	Number of carbon-carbon atomic contacts between hydrophobic residues, not counting alanine.
hyphob_Aro_contact	Number of carbon-carbon atomic contacts between aromatic residues.
hyphob_contact_norm	Number of carbon-carbon atomic contacts between hydrophobic residues divided by protein length
hyphob_Aro_contact_norm	Number of carbon-carbon atomic contacts between aromatic residues divided by protein length

Table 4.5.7.2: Design features based on Rosetta filters with score function beta_nov16

Features calculated using Rosetta filters related to burial of unsatisfied polar atoms:

Feature name	Explanation
buns_all	Total number of residues with at least one buried polar unsatisfied atom
buns_nosurf_all	Total number of residues with at least one buried polar unsatisfied atom, except in exposed residues
buns_nosurf_sc	Total number of residues with at least one buried polar unsatisfied side-chain atom, except in exposed residues
buns_nosurf_bb	Total number of residues with at least one buried polar unsatisfied backbone atom, except in exposed residues

Table 4.5.7.3: Design features based on Rosetta filters related to burial of unsatisfied polar atoms

Features calculated using CLIPPERS (24) pocket detection and inventory software. The main pocket is detected with the logic explained in Chapter 2 methods:

Feature name	Explanation
pckt_vol	Volume of main detected pocket in $Å^3$
mouth_n	Number or mouths or openings of the main detected pocket
mouth_area	Area of the largest mouth of the main detected pocket
pckt_maxTD	Shortest distance (Å) between a point in the outer hull and the deepest part of the main detected pocket.

Table 4.5.7.4: Design features based on CLIPPERS pocket detection and inventory software.

Overall protein fragment metrics calculated for protein fragments with similar sequence and secondary structures to 9-mer sequence stretches (protein length-9) in the target protein. For each 9-mer, 200 structure fragments are derived, as described in (4, 25).

Feature name	Explanation
low_rms_worst	Maximum RMSD among the subset of fragments with the lowest RMSD for all positions
avBest	Average RMSD of all fragments with the lowest RMSD for all positions
avAll	Average RMSD of all collected fragments
av_all_loop	Average RMSD of all collected fragments for loop positions
av_best_loop	Average RMSD of all fragments in all positions, in the loop with the lowest average RMSD
max_av_loop	Average RMSD of all fragments in all positions, in the loop with the highest average RMSD
max_av_best_loop	Maximum average RMSD of all loop positions
point_loop_av_all	Average of all fragments RMSD starting at the first position of all loops
point_loop_av_worst	Average of maximum RMSD of fragments starting at the first position of all loops
av_all_strand	Average RMSD of all collected fragments for strand positions
av_best_strand	Average RMSD of all fragments in all positions, in the strand with the lowest average RMSD
max_av_strand	Average RMSD of all fragments in all positions, in the strand with the highest average RMSD
max_av_best_strand	Maximum average RMSD of all strand positions
av_all_helix	Average RMSD of all collected fragments for helix positions
av_best_helix	Average RMSD of all fragments in all positions, in the helix with the lowest average RMSD
max_av_helix	Average RMSD of all fragments in all positions, in the helix with the highest average RMSD
max_av_best_helix	Maximum average RMSD of all helix positions

Table 4.5.7.5: Protein-wide fragment-related features.

Overall protein TERM metrics are calculated based on the output of the scripts provided with (26). TERM-based metrics were calculated based on the per-positions abundance_50, design_score_50 and structural score.

Feature name	Explanation
abd_w	Worst abaundance_50 value among all protein positions
abd_b	Best abaundance_50 value among all protein positions
abd_av	Average of all abaundance_50 values for all protein positions
dsc_w	Worst design_score_50 value among all protein positions
dsc_b	Best design_score_5value among all protein positions
dsc_av	Average of all design_score_5values for all protein positions
ssc_b	Worst structural score value among all protein positions
ssc_w	Best structural score value among all protein positions
ssc_av	Average of all structural score values for all protein positions

Table 4.5.7.6: Protein-wide TERM-related features

To obtain insight regarding specific parts of the proteins, we divided the protein in continuous sequence stretches that form local structures (sometimes with overlapping positions), and calculated different fragment and TERM features in each of them:

Structure stretch name	Explanation
N-term_helix	Residues from the N-terminus up to the second to last helix 1 turn
H1H2_link	Last H1 turn, loop connection to H2 and first 4 residues of H2
loop3_flank	Loop 3 and flanking residues
hairpin	S1 and 2, and connections
H3_n	4 residues of N-terminus of H3 and 3 previous residues
H3	All of H3
H3C_str3	Last 4 residues of H3, connection to S3 and 5 first residues of S3
str3_4	5 last residues of S3 and 5 first residues of S4
str4_5	5 last residues of S4 and 5 first residues of S5
str5_6	5 last residues of S5 and 5 first residues of S6
str6c_ch	C-terminus of S6 to the C-terminus of the protein, when a C-terminal helix is present.

Table 4.5.7.7: Different local structural domains for TERM and fragment local feature calculation.

For each of the above stretches, TERM and fragment metrics were calculated, and the final name of the features calculated this way are <stretch name>_<metric>.

Metric name	Explanation
av_allfr	Average of all fragments at all positions
av_bestfr	Average of only the fragments with the lowest RMSD at all positions
av_worstfr	Average of only the fragments with the highest RMSD at all positions
best_at_worstfr	Highest RMSD among the lowest RMSD fragments of all positions
abd50_av	Average of all abundance_50 values
dsc50_av	Average of all design_score_50 values
ssc50_av	Average of all structure score values
abd50_worst	Worst abundance_50 value among all positions
dsc50_ worst	Worst design_score_50 value among all positions
ssc50_ worst	Worst structural score value among all positions

Table 4.5.7.8: Different ways of calculating TERM and fragment local features.

Features Tminus1_netq, Tend_netq, T1_absq, Tminus1_absq, Tend_absq, abego_res_profile, abego_res_profile_penalty, largest_hphob_cluster, n_hphob_clusters, hphob_sc_contacts, hphob_sc_degree, n_charged, hydrophobicity, contig_not_hp_internal_max, contig_not_hp_avg, contig_not_hp_avg_norm, tryp_cut_sites, chymo_cut_sites, chymo_with_LM_cut_sites, nearest_chymo_cut_to_Nterm, nearest_chymo_cut_to_Cterm, nearest_tryp_cut_to_Nterm, nearest_tryp_cut_to_Cterm, nearest_tryp_cut_to_term and nearest_chymo_cut_to_term, were calculated using the enchance_score_file.py script provided with (6), and are thoroughly explained in their supplementary materials.

4.5.8 LASSO logistic regression model training on stability data

To identify features that predict stability, we trained LASSO logistic regression models using the features described in the previous section. The data and code for analysis of data derived from the first high-throughput experiment can be found at: https://github.com/basantab/NTF2analysis/blob/master/ProteaseAnalysisExp1/LassoLogisticRegression.ipynb

Analysis of data from the second high-throughput experiment can be found at: https://github.com/basantab/NTF2analysis/blob/master/ProteaseAnalysisExp2/LassoLogisticRegression_new_version.ipynb

4.5.9 Hydrophobicity enrichment sequence profile

Designs were split between stable and unstable depending on the threshold selected for each experiment (see Results), and the enrichment was calculated based on the whole population frequencies vs. the frequencies in the stable population. Code for these calculations, figures and derivation of sequence data from designs on the first high-throughput experiment can be found at https://github.com/basantab/NTF2analysis/tree/master/Exp1_SeqProfile

For designs tested on the second experiment:

https://github.com/basantab/NTF2analysis/tree/master/Exp2_SeqProfile

4.5.10 Experimental characterization of designs

Protein expression and purification in *E. coli:* Genes encoding the designed protein sequences were obtained from IDT already cloned in pET29b+ or pET21b+ (with N-terminal 6xHis tag followed by a TeV cut-site) expression vectors. Plasmids were transformed into chemically competent *Escherichia coli* Lemo21 cells from Invitrogen. Starter cultures were grown at 37°C in Luria-Bertani (LB) medium overnight with antibiotic (50 µg/ml carbenicillin for pET21b+ expression or 30 µg/ml kanamycin for pET-28b+ expression). For expression, overnight 5mL LB cultures were used to inoculate 500 mL of Auto-induction medium supplemented with antibiotic (27). After overnight expression, cells were collected by centrifugation (at 4 °C and 4400 r.p.m for 10 minutes) and resuspended in 25 ml of lysis buffer (30 mM imidazole and phosphate buffered saline, PBS - 137 mM NaCl, 12 mM Phosphate, 2.7 mM KCl, pH 7.4). Resuspended cells were lysed by sonication or microfluidizer in the presence of lysozyme, DNAse and protease inhibitors. Lysates were centrifuged at 4 °C and 20,000 r.c.f. for 30 minutes; and the supernatant was filtered and loaded to a nickel affinity gravity column pre-equilibrated in lysis buffer for purification. The column was washed with three column volumes of PBS+30 mM imidazole and the purified protein was eluted with three column volumes of PBS+300 mM imidazole. The eluted protein solution was dialyzed against PBS buffer overnight. The expression of purified proteins was assessed by SDS-polyacrylamide gel electrophoresis; and protein concentrations were determined from the absorbance at 280 nm measured on a NanoDrop spectrophotometer (ThermoScientific) with extinction coefficients predicted from the amino acid sequences. Proteins were further purified by FPLC size-exclusion chromatography using a Superdex 75 10/300 GL (GE Healthcare) column.

Circular dichroism (CD)*:* Far-ultraviolet CD measurements were carried out with an AVIV spectrometer, model 420. Wavelength scans were measured from 260 to 200 nm at temperatures between 25 and 95 °C. For wavelength scans and temperature melts a protein solution in PBS buffer (pH

7.4) of concentration 0.2-0.4 mg/ml was used in a 1 mm path-length cuvette, or 10 times more dilute for 1cm path-length cells.

Chemical denaturation experiments with guanidine hychloride were done with an automatic titrator using a protein concentration of 0.02-0.04 mg/ml and a 1 cm path-length cuvette with stir bar. PBS buffer (pH 7.4) was used for the cuvette solution and PBS+GdmCl for the titrant solution at the same protein concentration. GdmCl concentration was determined by refractive index. The denaturation process monitored absorption signal at 222 nm in steps of 0.1 or 0.2 M GdmCl with 1 min mixing time for each step and at 25 °C. The denaturation curves were fitted by non-linear regression to a two-state unfolding model to extract six parameters: slope and intercept for pre- and post-transition baselines, *m* value and the folding free energy (ΔG_{H2O}) (28, 29). The deviation of the fitted *m* value from its expected value given protein size was computed using the empirical correlation between the number of protein residues and the protein *m* value for denaturation with GdmCl (12).

Size exclusion chromatography combined with multiple angle light scattering (SEC-MALS)*:* To evaluate protein quaternary structure, SEC-MALS experiments were performed using a Superdex 75 10/300 GL (GE Healthcare) column combined with a miniDAWN TREOS multi-angle static light scattering detector and an Optilab T-rEX refractometer (Wyatt Technology). One hundred microliter protein samples of 1-3 mg/ml were injected to the column equilibrated with PBS (pH 7.4) or TBS (pH 8.0) buffer at a flow rate of 0.5 ml/min. The collected data was analyzed with ASTRA software (Wyatt Technology) to estimate the molecular weight of the eluted species.

4.5.11 Isothermal titration calorimetry

Microcal ITC with AutoITC2000 autosampler was used to carry out the titrations and measure the heat generated by binding. Adapted software based on Origin was used to analyze results.

Buffer used: 10 mM Tris-HCl pH 8, 150 mM NaCl

Positive control: Anti-Cortisol monoclonal antibody XM210 (ab1949), reported K_d: 600fM

Negative control: BBM2nHm0481

Cortisol concentration in syringe: 46μM

Concentration of protein in cell: 5μM

4.5.12 Crystallography data collection and analysis metrics

To prepare protein samples for X-ray crystallography, the buffer of choice was 25 mM Tris, 50 mM NaCl, pH 8.0. Proteins were expressed from pET29b+ constructs to cleave the 6xHis tag with TeV. Proteins were incubated with TeV (1:100 dilution) overnight at room temperature and cleaved samples were loaded to a Ni-NTA column pre-equilibrated in PBS+30mM Imidazole. Flow-through was collected and washed with 1-2 column volumes. Proteins were further purified by FPLC as described above and specific cleavage of the 6xHis tag was verified by SDS-PAGE.

Purified proteins were concentrated to approximately 10-20 mg/ml for screening crystallization conditions. Commercially available crystallization screens were tested in 96-well sitting or hanging drops with different protein:precipitant ratios (1:1, 1:2 and 2:1) using a mosquito robot. When possible, initial crystal hits were grown in larger 24-well hanging drops. Obtained crystals were flash-frozen in liquid nitrogen. X-ray diffraction data sets were collected at the Advanced Light Source (ALS). Crystal structures were solved by molecular replacement with Phaser (30) using the design models as the initial search models. The structures were built and refined using Phenix (31, 32) and Coot (33).

The crystallization conditions for the solved crystal structures are the following:

BBM2nHm0589: (His-tag not cleaved): Protein solution concentration: 56mg/L

1:1 dilution in 0.09M Sodium fluoride; 0.09M Sodium bromide; 0.09M Sodium iodide, 0.1M Tris/BICINE pH 8.5, 50% v/v of 40% v/v PEG 500 MME; 20 % w/v PEG 20000. (Morpheus-HT96 B9 (34))

BBM2nHm0589_I64F_A80G_T94P_D101K_L106W: (His-tag cleaved): Protein solution concentration: 7.7mg/mL

1:1 dilution in 0.09M Sodium nitrate, 0.09 Sodium phosphate dibasic, 0.09M Ammonium sulfate, pH 6.5 0.1M Imidazole/MES monohydrate (acid), %50v/v of 40% v/v PEG 500 MME; 20 % w/v PEG 20000 (Morpheus-HT96 C1 (34))

BBM2nHm0481: (His-tag cleaved): Protein solution concentration: 48.7mg/mL

1:1 dilution in 0.12M 1,6-Hexanediol; 0.12M 1-Butanol; 0.12M 1,2-Propanediol; 0.12M 2- Propanol; 0.12M 1,4-Butanediol; 0.12M 1,3-Propanediol, 0.1M Imidazole/MES monohydrate (acid), pH 6.5, and 50% v/v of 40% v/v PEG 500 MME; 20 % w/v PEG 20000 (Morpheus-HT96 D1 (34)).

	BBM2nHm0589	**BBM2nHm0481**	**BBM2nHm0589_I64F_A80G_T94P_D101K_L106W**
Wavelength	1	1	0.9786
Resolution range	28.27 - 1.38 (1.429 - 1.38)	44.33 - 1.62 (1.678 - 1.62)	44.72 - 1.83 (1.896 - 1.83)
Space group	C 1 2 1	P 21 21 21	P 31
Unit cell	60.076 30.498 61.099 90 97.837 90	32.578 36.814 177.303 90 90 90	38.51 38.51 134.148 90 90 120
Total reflections	100427 (9900)	132857 (13262)	81860 (8151)
Unique reflections	22794 (2258)	28087 (2762)	19589 (1956)
Multiplicity	4.4 (4.4)	4.7 (4.8)	4.2 (4.2)
Completeness (%)	99.65 (99.78)	93.63 (79.12)	88.44 (66.92)
Mean I/sigma(I)	19.06 (0.87)	19.37 (0.96)	11.07 (0.68)
Wilson B-factor	20.24	25.13	29.86
R-merge	0.0366 (1.774)	0.03966 (1.586)	0.06516 (2.12)
R-meas	0.04167 (2.02)	0.04474 (1.779)	0.07492 (2.439)
R-pim	0.01965 (0.9543)	0.02027 (0.7912)	0.03643 (1.194)

CC1/2	1 (0.319)	1 (0.322)	0.999 (0.432)
CC*	1 (0.695)	1 (0.698)	1 (0.777)
Reflections used in refinement	22779 (2254)	26376 (2187)	17345 (1309)
Reflections used for R-free	1997 (196)	1897 (156)	1702 (122)
R-work	0.1825 (0.3084)	0.2109 (0.3592)	0.2163 (0.3734)
R-free	0.2158 (0.3547)	0.2403 (0.3462)	0.2546 (0.4206)
CC(work)	0.952 (0.690)	0.958 (0.647)	0.951 (0.722)
CC(free)	0.932 (0.612)	0.947 (0.681)	0.955 (0.650)
Number of non-hydrogen atoms	1008	1879	1770
macromolecules	951	1776	1715
solvent	57	103	55
Protein residues	116	217	229
RMS(bonds)	0.019	0.005	0.007
RMS(angles)	1.78	0.63	0.84
Ramachandran favored (%)	100.00	99.04	100.00
Ramachandran allowed (%)	0.00	0.96	0.00
Ramachandran outliers (%)	0.00	0.00	0.00

Rotamer outliers (%)	0.00	0.00	0.00
Clashscore	8.59	3.21	5.76
Average B-factor	29.64	41.31	48.40
macromolecules	29.12	41.11	48.55
solvent	38.31	44.71	43.69
Number of TLS groups		13	12

Table 4.5.9: Data collection and refinement statistics. Statistics for the highest-resolution shell are shown in parentheses.

4.5.13 Generative algorithm for proteins from the NTF2-like superfamily

All code can be downloaded from GitHub at: https://github.com/basantab/NTF2Gen

The NTF2Gen repository contains all the tools for *de novo* design of NTF2-like proteins. The main script is CreateBeNTF2_backbone.py, which manages the construction of NTF2 backbones, followed by DesignBeNTF2.py, which designs sequence on a given backbone generated by the previous script. To generate backbones from a specific set of parameters, use CreateBeNTF2PDBFromDict.py.

The fundamental building blocks of the backbone generation protocol are Rosetta XML protocols (included in the repository) that are specialized instances of the BlueprintBDRMover Rosetta fragment assembly mover. All checks and filters mentioned in the result section previous to design are implemented either in the XML files or the python scripts. The design script is also based on a set of XML protocols, one for each of the described stages. Glycine placement in highly curved strand positions and selection of pocket positions is managed by DesignBeNTF2.py (BeNTF2seq/Nonbinding). Pocket positions are selected by placing a virtual atom in the midpoint between the H3-S3 connection and the S6 bulge, and choosing all positions whose C_{alpha}-C_{beta} vector is pointing towards the virtual atom (the V_{atom}-C_{alpha}-C_{beta} angle is smaller than 90º), and their C_{alpha} is closer than 8Å.

4.5.14 Selection of designs for second high-throughput experiment

Using the logistic regression model trained on data derived from the first high-throughput experiment, we predicted the probability of being stable of 11548 models designed by the new generative algorithm, and selected a subset of 10073 with chance > 0.75. A mistake in the calculation of these values resulted in the selection of designs not being strictly above 0.75, but biased towards higher values, which turned out to be beneficial for verifying the applicability of the model to the second round of designs. In parallel, we clustered all designs by their fold features, and for each cluster we searched for a representative in the 10073 designs subset. All code and values can be found in the Git repository https://github.com/basantab/NTF2analysis, in the Exp2_selection folder.

4.5.15 Comparison of stability controls for tryptophan and glycine sequence features

To evaluate the effect of enforced tryptophan/lysine pairs in the short arm, and glycine in curved strand positions, we included controls where these modifications were replaced by other amino-acid identities. We measured stabilities for designs and their relative controls and evaluated their differences. All data and code to produce the figures presented are available in the public Git repository https://github.com/basantab/NTF2analysis, in the folder GlycineTryp_control_analysis.

4.5.16 Small-molecule binding protein design

As described above, RIDOCK (13) was used to dock Aflatoxin B1 in models of *de novo* NTF2-like proteins. The designable positions for this process were selected during the generation of these scaffolds, as described in 4.513. The following stages are similar to those described for sequence design in the new generative algorithm in the results section, except we added an intermediate step where we generate an ensemble of structures and look for fully satisfied hydrogen bond networks that include the ligand.

Detailed protocols for each stage can be found at https://github.com/basantab/NTF2Gen (BeNTF2seq/Binding).

4.5.17 Equilibrium dialysis for Aflatoxin B1 binding detection

Equilibrium dialysis was performed in Thermo Fisher 8K MWCO RED plates, incubating aflatoxin B1 solutions with and without the tested protein, for 4hs at 300rmp orbital shaking, at 37ºC. Protein is tested at 25μM against increasing concentrations of aflatoxin. At the end of the incubation period aflatoxin concentrations on each side are measured by fluorescence. The fraction of aflatoxin bound is calculated from the difference to the expected equilibrium concentrations in the absence of binding.

4.6 Acknowledgements

We thank Enrique Marcos, Daniel Adriano-Silva, Gabriel Rocklin, Anastassia Vorobieva, Ralph Cacho, Marc Lajoie, Scott Boyken, Ian Haydon, Derrick Hicks, Brian Coventry, Alexander Ford, Allan Ferrari, Matthew Bick, Hahnbeom Park and Yakov Kipnis for helpful comments and discussions in general. We thank Adam Moyer, William Sheffler for providing some of the computational tools for protein design. We thank Matthew Bick and Alex Kang for assistance with protein sample preparation for crystallization, data collection and analysis. We thank Inna Goreshnick, Jorgen Nelson and Cassie Brian for experimental assistance regarding deep sequencing, yeast work and multiplex assembly. We thank Philip Leung and Ted Baughman for their work testing Aflatoxin B1 binding proteins. We thank the Rosetta community for general input and help with troubleshooting. We thank Darwin Alonso, Luki Goldschmidt and Patrck Vecchiato for technical support. This work was facilitated though the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system at the University of Washington.

4.7 Literature

1. Terzidis, K. (2004) Algorithmic Design : A Paradigm Shift in Architecture? in *Architecture in the Network Society [22nd eCAADe Conference Proceedings / ISBN 0-9541183-2-4] Copenhagen (Denmark) 15-18 September 2004, pp. 201-207*, pp. 201–207
2. Bessa, M. (2009) Algorithmic Design. *Archit. Des.* **79**, 120–123
3. Schumacher, P. (2009) Parametricism: A New Global Style for Architecture and Urban Design. *Archit. Des.* **79**, 14–23
4. Marcos, E., Basanta, B., Chidyausiku, T. M., Tang, Y., Oberdorfer, G., Liu, G., Swapna, G. V. T., Guan, R., Silva, D.-A., Dou, J., Pereira, J. H., Xiao, R., Sankaran, B., Zwart, P. H., Montelione, G. T., and Baker, D. (2017) Principles for designing proteins with cavities formed by curved β sheets. *Science*. **355**, 201–206
5. Zhang, Y., and Skolnick, J. (2005) TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309
6. Rocklin, G. J., Chidyausiku, T. M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., Carter, L., Ravichandran, R., Mulligan, V. K., Chevalier, A., Arrowsmith, C. H., and Baker, D. (2017) Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*. **357**, 168–175
7. Klein, J. C., Lajoie, M. J., Schwartz, J. J., Strauch, E.-M., Nelson, J., Baker, D., and Shendure, J. (2016) Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res.* **44**, e43
8. WALKER, S. H., and DUNCAN, D. B. (1967) Estimation of the probability of an event as a function of several independent variables. *Biometrika*. **54**, 167–179
9. Tibshirani, R., and Tibshirani, R. (1994) Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B*. **58**, 267--288
10. Zheng, F., Zhang, J., and Grigoryan, G. (2015) Tertiary Structural Propensities Reveal Fundamental Sequence/Structure Relationships. *Structure*. **23**, 961–971
11. Yan, S., Gawlak, G., Makabe, K., Tereshko, V., Koide, A., and Koide, S. (2007) Hydrophobic Surface Burial Is the Major Stability Determinant of a Flat, Single-layer β-Sheet. *J. Mol. Biol.* **368**, 230–243
12. Myers, J. K., Pace, C. N., and Scholtz, J. M. (1995) Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci.* **4**, 2138–48
13. Dou, J., Vorobieva, A. A., Sheffler, W., Doyle, L. A., Park, H., Bick, M. J., Mao, B., Foight, G. W., Lee, M. Y., Gagnon, L. A., Carter, L., Sankaran, B., Ovchinnikov, S., Marcos, E., Huang, P.-S., Vaughan, J. C., Stoddard, B. L., and Baker, D. (2018) De novo design of a fluorescence-activating β-barrel. *Nature*. **561**, 485–491
14. Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., and Baker, D. (2012) Principles for designing ideal protein structures. *Nature*. **491**, 222–7
15. Fraga, K. J., Joo, H., and Tsai, J. (2016) An amino acid code to define a protein's tertiary packing surface. *Proteins Struct. Funct. Bioinforma.* **84**, 201–216
16. Zheng, F., Zhang, J., and Grigoryan, G. (2014) Tertiary Structural Propensities Reveal Fundamental Sequence/Structure Relationships. *Structure*. **23**, 961–971
17. Fujiwara, K., Ebisawa, S., Watanabe, Y., Fujiwara, H., and Ikeguchi, M. (2015) The origin of β-strand bending in globular proteins. *BMC Struct. Biol.* **15**, 1–12
18. Huang, P.-S., Boyken, S. E., and Baker, D. (2016) The coming of age of de novo protein design. *Nature*. **537**, 320–327
19. Kumar, P., Mahato, D. K., Kamle, M., Mohanta, T. K., and Kang, S. G. (2016) Aflatoxins: A Global Concern for Food Safety, Human Health and Their Management. *Front. Microbiol.* **7**, 2170
20. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D., and Higgins, D. G. (2014) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539–539
21. Hoover, D. M., and Lubkowski, J. (2002) DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.* **30**, e43
22. Sheffler, W., and Baker, D. (2010) RosettaHoles2: A volumetric packing measure for protein structure refinement and validation. *Protein Sci.* **19**, 1991–1995
23. Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring

matrices. *J. Mol. Biol.* **292**, 195–202
24. Coleman, R. G., and Sharp, K. a. (2010) Protein pockets: Inventory, shape, and comparison. *J. Chem. Inf. Model.* **50**, 589–603
25. Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., and Baker, D. (2012) Principles for designing ideal protein structures. *Nature*. **491**, 222–7
26. Zhou, J., and Grigoryan, G. (2014) Rapid search for tertiary fragments reveals protein sequence-structure relationships. *Protein Sci.* 10.1002/pro.2610
27. Studier, F. W. (2005) Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–34
28. Santoro, M. M., and Bolen, D. W. (1992) A test of the linear extrapolation of unfolding free energy changes over an extended denaturant concentration range. *Biochemistry*. **31**, 4901–4907
29. Scholtz, J. M., Grimsley, G. R., and Pace, C. N. (2009) Solvent denaturation of proteins and interpretations of the m value. in *Methods in enzymology*, pp. 549–565, **466**, 549–565
30. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674
31. Zwart, P. H., Afonine, P. V, Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., McKee, E., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K., Storoni, L. C., Terwilliger, T. C., and Adams, P. D. (2008) Automated structure solution with the PHENIX suite. *Methods Mol. Biol.* **426**, 419–35
32. Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 213–221
33. Emsley, P., and Cowtan, K. (2004) Coot: Model-building tools for molecular graphics. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60**, 2126–2132
34. Gorrec, F. (2009) The MORPHEUS protein crystallization screen. *J. Appl. Crystallogr.* **42**, 1035–1042