Title
Characterization, Design and Application of Natural and Engineered Symmetric Protein Complexes

Author
Liu, Yuxi

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Characterization, Design and Application of Natural and Engineered Symmetric Protein Complexes

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy in Biochemistry, & Molecular Biology

by

Yuxi Liu

2018

ABSTRACT OF THE DISSERTATION

Characterization, Design and Application of Natural and Engineered Symmetric Protein Complexes

by

Yuxi Liu

Doctor of Philosophy in Biochemistry and Molecular Biology

University of California, Los Angeles, 2018

Professor Todd O. Yeates, Chair

We frequently find proteins exist in oligomeric forms in nature. The abundance of dimers, trimers and tetramers with cyclic or dihedral symmetries in the Protein Data Bank is a good testimony. Even more, it is not rare to find proteins form highly ordered, symmetric, large complexes. These oligomeric forms are usually essential for their functions. Ferritin forms an octahedral cage with 24 subunits to store iron; some virus capsid proteins assemble into icosahedral cages; vaults, which are large dihedral particles widely conserved in eukaryotes, have biological functions yet to be discovered. These fascinating structures inspire three types of questions: How do individual subunits interact form such symmetric complexes? How can we reproduce such complexes with protein engineering? How do we put engineered symmetric protein complexes to application? My thesis work consists of projects addressing all three questions.

My first project, described in Chapter 1, concerns bacterial microcompartments (MCP), which are large proteinaceous organelles enclosed by an icosahedral or pseudo-icosahedral shell. MCPs usually enclose special metabolic pathways that are inefficient or toxic in the cytosol. To do so, MCPs must form a sealed barrier with its shell proteins. It was hypothesized that at least

one type of the proteins forming the shell of MCPs has to be pentameric instead of hexameric. Indeed, we proved that the BMV proteins, a family of protein highly conserved in MCP operons, formed pentamers in solution. Together with other crystallographic evidence, we conclude BMV proteins form pentamers to cap and seal the MCP shell. In addition to MCPs, I worked on another natural oligomeric protein, bactofilin. Bactofilins are fiber-forming proteins that are widely conserved among bacteria. These proteins have roles in diverse biological functions including but not limited to cell motility, cell wall synthesis and modification. Chapter 2 describe my preliminary biochemical and structural work on bactofilins.

Next, I moved on to symmetry-based engineering protein complexes. In Chapter 3, I included a recent review paper on the theory and successes in symmetry-based protein engineering that I participated in preparing. Designed complexes need to be validated at high resolution with X-ray crystallography, but for a long time, the low yield and solubility of the designs complicated their validation. In Chapter 4, we show that mutating solvent-exposed side chains to charged amino acids improved the solubility of a previously low yield tetrahedral design and enabled validation by crystallography. Next, I advanced to a bigger challenge in designing symmetric nanoparticles—icosahedral particles. Icosahedral particles are made up of 60 asymmetric units, as compared to 12 in tetrahedral particles, making them much more difficult to design with accuracy. I was able to validate three different icosahedral design with crystallography, making them the largest designed protein assemblies ever crystallized to date. This work is described in Chapter 5. Additionally, I have made other independent design efforts, one to combine DNA and protein as building materials to design tetrahedral complexes, another to design protein sheets with layer group symmetry. These efforts are documented in Chapter 6.I

In the last chapter, I utilized the validated tetrahedral designs as a scaffold in cryo-electron microscope (cryo-EM) for small targets. Despite recent advancements in cryo-EM techniques, small targets remain difficult. By arranging small targets around tetrahedral particles, we can overcome the size limit and provide multiple views to alleviate the commonly seen orientation preference. My project used a type of versatile adaptor protein, designed ankyrin repeat proteins (DARPins), to connect the tetrahedral particles to the imaging targets. We show that the resulting construct is amenable to structural analysis by single particle cryo-EM, allowing us to identify and solve the structure of the attached DARPin at near-atomic detail. The result demonstrates that proteins considerably smaller than the theoretical limit of 50 kDa for cryo-EM can be visualized clearly when arrayed in a rigid fashion on a symmetric designed protein scaffold. Because the amino acid sequence of a DARPin can be chosen to confer tight binding to various other proteins, the system provides a future route for imaging diverse macromolecules, potentially broadening the application of cryo-EM to proteins of typical size in the cell.

In conclusion, my thesis work contributes to the understanding of natural oligomeric complexes, expands our capacity in designing symmetric assemblies, and puts forward an example of a useful application of the designed assemblies.

This dissertation of Yuxi Liu is approved.

David S. Eisenberg

Pascal Francois Egea

Todd O. Yeates, Committee Chair

University of California, Los Angeles

2018

To my loving family and my brilliant friends

You have inspired me

TABLE OF CONTENTS

List of Figures

List of Tables

Acknowledgements

The six years in graduate school had a lot of ups and downs. The work described in here would not have happened without the support of others. First and foremost, I would like to express my deepest gratitude towards my thesis advisor, Todd Yeates. Prof. Yeates has given me unwavering support throughout the years, even during difficult times. His confidence and optimism have always inspired me to push forward. I have enjoyed great freedom to pursue many different tops during my graduate career. The time I spent here at UCLA as a graduate student is one the best times of my life, and I owe that largely to Prof. Yeates.

I would like to thank my committee members, Prof. David Eisenberg, Prof. Pascal Egea, Prof. Jim Bowie, and Prof. Robert Gunsulus for their insights and advice. I'd like to give special thanks to Prof. Bowie, who is my undergraduate research mentor, for leading me to the invigorating world of scientific research, also for continuously support me with my various endeavors during graduate school. I also thank Prof. Eisenberg for his comments and suggestions during joint group meetings.

I would like to acknowledge my collaborators and co-authors, whose talent and diligent work motivated me to think more critically. I have had the privilege to work closely with some of them. I thank Dr. Nicole Wheatly for numerous conversions on microbiology and molecular biology, Dr. Jacob Bale and Prof. Neil King for suggestions and discussions on protein engineering and computation, and Dr. Shane Gonen for opening the door to cryo-electron microscopy.

Over the years I have come to learn the importance of a supportive environment, and the fact that such an environment is not always easy to find. I have been extremely fortunate to work with a group of brilliant people who are always there to help. I would like to acknowledge the

UCLA DOE Macromolecular Crystallization Core Technology Center, X-ray Crystallography Core Technology Center, and the Protein Expression Core Technology Center. Dr. Mark Arbing has given numerous advice and generously shared with us many materials. Dr. Duilio Cascio, Dr. Michael Sawaya, and Michael Collazo have given invaluable suggestions at each step in crystallography, from sample preparation, crystallization, data collection, data processing, to deposition. Not only in crystallography, Dr. Cascio goes out of his way to help, advice, accommodate, and promote my various efforts. His helped enabled me to venture further and wider. I would also like to acknowledge Prof. Hong Zhou and the staff at UCLA EICN for their advice on electron microscopy and for suggestions on data processing.

Every member of the Yeates lab, past or present, has generously helped me in various ways in research and in life. Some have become very close personal friends. I have benefitted greatly from the daily discussion we carry, where new ideas sprung. I would like to acknowledge the undergraduates/post-bachelor trainees who made substantial contributions to the research I present in the thesis: Joanna Ngo, Marianne Vo, and Duc Huynh.

I have been very fortunate to receive several supports for my graduate study. I would like to thank the UCLA MBI Whitcome Fellowship for supporting me during 2015-2017, the UCLA Dissertation Year Fellowship during 2017-2018. I also thank the Audree Fowler Fellowship in Protein Science for providing extra resources for my research.

Last but not least, I would like to thank my family and friends. One cannot become a good scientist before becoming a good person. You have shared with me your love for science and love for life. You have showed me by sample what is diligence, humbleness, kindness, and compassion. I am eternally grateful to have you in my life and I strive to become as good as you are.

Vita

Education

9/2012-9/2018 Graduate program in Biochemistry & Molecular Biology, UCLA

9/2008-6/2012 B.S. in Molecular, Cell, and Developmental Biology, UCLA

Publications

- **Yuxi Liu**[*], Shane Gonen[*], Tamir Gonen, Todd O. Yeates. (2018) Near-atomic cryo-EM imaging of a small protein displayed on a designed scaffolding system. *PNAS*:201718825.
- Jacob B. Bale, Shane Gonen[*], **Yuxi Liu**[*], William Sheffler, Daniel Ellis, Chantz Thomas, Duilio Cascio, Todd O. Yeates, Tamir Gonen,Neil P King, and David Baker. 2016 "Accurate design of megadalton-scale multi-component icosahedral protein complexes". Science 353, 389-394 (2016).
- Jacob B. Bale, Rachel U. Park, **Yuxi Liu**, Neil P. King, Todd O. Yeates, and David Baker. "Structure of a Designed Tetrahedral Protein Assembly Variant Engineered to Have Improved Soluble Expression." Protein Science 24, 1695-1701 (2015): 1695–1701.
- Wheatley, Nicole M., Soheil D. Gidaniyan[*], **Yuxi Liu**[*], Duilio Cascio, and Todd O. Yeates. 2013. "Bacterial Microcompartment Shells of Diverse Functional Types Possess Pentameric Vertex Proteins." *Protein Science* 22 (5): 660–665 (2013).
- Todd O. Yeates, **Yuxi Liu**, Joshua Laniado. 2016. "The Design of Symmetric Protein Nanomaterials Comes of Age in Theory and Practice" Current Opinion in Structural Biology *39*, 134–143 (2016).

[*]These authors contributed equally.

Selected Presentations and Awards

- Yuxi Liu, Shane Gonen, Tamir Gonen, Todd O. Yeates. "Near-Atomic Cryo-EM Imaging of a Small Protein Displayed on a Designed Scaffolding System". Feb 2018, Keystone symposia oral and poster presentation

  This presentation was sponsored by the Keystone Symposia Future of Science Fund scholarship.

- Yuxi Liu, Marianne Vo, Duilio Cascio, Todd Yeates. "A Modular Scaffold to Symmetrize Proteins for Cryo-EM Imaging" Mar 2017. West Coast Protein Crystallography Workshop oral presentation.

2018	Biochemistry Dissertation Award
2017-2018	Audree Fowler Fellowship in Protein Science
9/2017-6/2018	Dissertation Year Fellowship
7/2016-6/2017	Whitcome Fellowship
9/2015-8/2016	Whitcome Fellowship
6/1/2012	Graduated with Highest Departmental Honor from Department of Molecular, Cell & Developmental Biology Graduated with College Honors & Magna cum laude

Chapter 1. Bacterial Microcompartment Shells of Diverse Functional Types Possess Pentameric Vertex Proteins[1]

[1]Nicole M. Wheatley, [2,$]Soheil D. Gidaniyan, [3,$]Yuxi Liu, [2]Duilio Cascio,[1,2,3,*]Todd O. Yeates

[1] Molecular Biology Institute, University of California, Los Angeles, CA 90095

[2] UCLA-DOE Institute for Genomics and Proteomics, University of California, Los Angeles, CA 90095

[3] Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095

[$]These authors contributed equally

Keywords: OCAC, oligomeric state determination, EutN, bacterial microcompartments, protein assembly, capsid, glycyl radical, BMV, pentameric vertex protein, GrpN

Data deposition: PDB structure coordinates and diffraction data – ID 4I7A

Abbreviations:

BMC – bacterial microcompartment (hexameric) shell protein

BMV – bacterial microcompartment vertex protein (pentameric)

Grp – glycyl radical-based propanediol utilizing MCP

MCP – bacterial microcompartment

OCAC – oligomeric characterization by addition of charge

OCAM - oligomeric characterization by addition of mass

[1] The main body of this chapter is composed of the adapted manuscript of a published paper (1). An additional section, Supplementary Text, is included here to discuss some observation on the BMV structures. Reprint with permission from John Wiley and Sons (license number 4425161393588).

ABSTRACT

Bacterial microcompartments (MCPs) are large proteinaceous structures comprised of a roughly icosahedral shell and a series of encapsulated enzymes. Experimental studies have characterized MCPs carrying out three different metabolic functions, and bioinformatics studies have implicated other types, including one believed to perform glycyl radical-based metabolism of 1,2-propanediol (Grp). Here we report the crystal structure of a protein (GrpN), which is presumed to be part of the shell of a Grp-type MCP in *Rhodospirillum rubrum* F11. GrpN is homologous to a family of proteins (EutN/PduN/CcmL/CsoS4) whose members have been implicated in forming the vertices of MCP shells. Consistent with that notion, the crystal structure of GrpN revealed a pentameric assembly. That observation revived an outstanding question about the oligomeric state of this protein family: pentameric forms (for CcmL and CsoS4A) and a hexameric form (for EutN) had both been observed in previous crystal structures. To clarify these confounding observations, we revisited the case of EutN. We developed a molecular biology-based method for accurately determining the number of subunits in homo-oligomeric proteins, and found unequivocally that EutN is a pentamer in solution. Based on these convergent findings, we propose the name bacterial microcompartment vertex (BMV) for this special family of MCP shell proteins.

INTRODUCTION

Microcompartments (MCPs) are polyhedrally shaped supramolecular protein assemblies that physically encapsulate select metabolic pathways in prokaryotes (reviewed in refs. (2–6)). This encapsulation serves to concentrate and separate specific enzymes and their metabolic intermediates from the cytoplasm, thereby increasing reaction efficiency, retaining volatile intermediates, and/or protecting the cellular milieu from toxic intermediates.(6) The

carboxysome MCP increases carbon-fixation efficiency by encapsulating carbonic anhydrase and RuBisCO together, so the CO_2 produced by the first enzyme can be delivered at high concentration to the second enzyme.(5, 7) Cofactor B_{12}-dependent MCPs for ethanolamine utilization (Eut) and 1,2-propandiol utilization (Pdu) function by retaining volatile or toxic aldehyde compounds – propionaldehyde and acetaldehyde – that occur as intermediates in those pathways. (6, 8–12) Bioinformatic studies have suggested the existence of an MCP for glycyl radical-based 1,2-propanediol metabolism, referred to hereafter as Grp.(13) Sequence analysis suggests that the Grp MCP, like the Pdu MCP, houses enzymes for metabolizing 1,2-propanediol, but that it uses a glycyl radical enzyme for the key dehydration reaction rather than B_{12}-dependent enzymes.(13) The Grp MCP has not yet been studied biochemically.

Structural analyses have answered a number of questions regarding the geometry and mechanism of MCP assembly and function (reviewed in ref. (3)). Current models describe MCPs as being comprised of sheets of hexameric shell proteins (belonging to the BMC family of proteins) forming polyhedral facets (14–16), along with a much smaller number of special pentameric proteins (belonging to the EutN/PduN/CcmL/CsoS4 protein family) placed at the vertices of the polyhedral shell.(17) While crystallographic studies of (hexameric) BMC shell proteins have consistently supported this model of MCP organization, experiments on the putative vertex proteins have provided a less coherent picture. Two crystal structures of carboxysome vertex proteins, CcmL and CsoS4A, are indeed pentameric(17) (PDB accession code 2QW7 and 2RCF). However, a third structure of a presumptive vertex protein, EutN from *E. coli* (PDB code 2Z9H), revealed a hexameric quaternary assembly with nearly hexagonal shape (18) Likewise, multiple genetic studies have shown that deleting the presumptive vertex proteins compromises the formation of closed polyhedral shells (19–21), yet closed shells can apparently be formed

with lower efficiency in some cases (22). These equivocal results have complicated the interpretation of the architectural role of this protein family.

Electron microscopy studies indicate that carboxysomes tend to be more geometrically regular and more nearly icosahedral in shape compared to other types of MCPs that have been visualized (23, 24); this has raised the possibility that pentameric units required for symmetric icosahedral architecture might be present only in carboxysome MCPs. Here we provide evidence that this family of proteins serves as the pentameric vertex element across multiple divergent types of MCPs. Specifically, we report the pentameric crystal structure of GrpN from the Grp MCP, and show also that EutN is a pentamer in solution, contrary to previous crystallographic findings.

RESULTS AND DISCUSSION

An operon encoding enzymes and shell proteins for a presumptive Grp MCP has been described in *R. rubrum.*(13) One of the encoded proteins, referred to hereafter as GrpN (Gene ID: 12642037), was identified as a representative of the EutN/PduN/CcmL/CsoS4 family of shell proteins (Pfam domain: PF03319), hereafter described as bacterial microcompartment vertex (or BMV) proteins. GrpN was overexpressed in *E. coli* and purified by metal affinity chromatography on the basis of a polyhistidine tail added to the protein. Single crystals were obtained in space group I23, and an X-ray structure was determined by molecular replacement and refined at a resolution of 3.2 Å. The final atomic model is 96% complete; residues 65-70 are in a disordered loop and could not be modeled accurately.

As expected, the tertiary fold follows closely that reported earlier for other members of this protein family (Fig. 1.1). The core structure of GrpN is comprised of five antiparallel beta strands that curve to create a small beta barrel. The C-termini and loop regions between beta

strands extend away from the beta-barrels, making contacts with adjoining subunits. The atomic coordinates align well to CcmL, CsoS4A and EutN, with RMSD values of 0.73 Å, 0.81 Å, and 0.80 Å, respectively, over C-alpha positions. Stabilization of the quaternary structure is largely mediated by contacts between loop regions from adjacent beta-barrels; minor contributions are made by barrel-to-barrel contacts. Structural comparisons between BMV structures show that the N-termini and the small beta barrels are highly conserved, while the C-termini and regions adjoining the core beta sheets differ slightly in length and secondary structure content. For example, GrpN and CsoS4A lack C-terminal beta hairpins present in both CcmL and EutN.

GrpN crystallized as a pentamer in the asymmetric unit of the crystal (Table 1.1). The pentameric unit of GrpN matches those reported earlier for CcmL from the beta-type carboxysome and CsoS4A from the alpha-type carboxysome. One surprising aspect of the GrpN crystal structure is the arrangement of pentamers in the unit cell. Twelve crystallographically-related pentamers are arranged in a nearly icosahedral fashion. This 60-subunit icosahedral arrangement is unlikely to reflect the natural assembly of the much larger MCP; in native MCPs, pentameric units are presumably surrounded by BMC-type hexamers. In addition, the interactions between pentamers in the crystal involve relatively small sites of contact. Nonetheless, this spontaneous assembly could be an interesting starting point for engineering a novel icosahedral cage. The design of self-assembling protein cages by various methods has been discussed in a series of recent studies (25–29) The observation that GrpN is a pentamer establishes that other MCPs besides carboxysomes possess special pentameric shell proteins to explain the closure of an otherwise flat layer of hexameric units provided by the BMC-type proteins.

Our finding that GrpN forms a pentameric assembly left the earlier reported hexameric structure for *E. coli* EutN as an outlying observation. The hexameric state reported for EutN could either indicate an unusual structural role for this homolog, or it could represents a spurious or minor oligomeric form selected during crystallization. To distinguish between these two possibilities, we devised a novel method to determine the oligomeric state of EutN in solution (Fig 1.2).

The method we developed for determining oligomeric states is based on the change in charge-to-mass ratio that occurs after proteolytic cleavage of a small, charged N- or C-terminal sequence that is genetically appended at the end of the native protein subunit (Fig 1.2A). Following different degrees of exposure to protease, a homo-oligomer composed of *n* subunits can be partially proteolyzed to generate a total of *n+1* distinct charge forms. The distinct charge forms can be enumerated by native gel electrophoresis. To be consistent with the nomenclature of established methods, we refer to our approach as OCAC, an acronym for oligomeric characterization by addition of charge (Fig. 1.2B). A related method described earlier, OCAM (oligomeric characterization by addition of mass), was developed to determine the oligomeric state of membrane proteins.(30) OCAM determines oligomeric state based on changes in the mobility of protein complexes on blue native gels, which arise from differential removal of relatively massive domains fused to the native protein subunits. In contrast, OCAC requires addition of relatively small terminal extensions to the native protein: essentially a protease cleavage site and as few as 1 or 2 directly adjacent charged amino acids. As with OCAM, the success of OCAC depends on a slow rate of exchange between subunits in different oligomeric complexes. Under conditions where we applied the method, native gels were capable of distinguishing between oligomers with just one unit charge difference. Another method based on

counting distinct oligomeric forms was described several years ago, wherein the chemical modification of a genetically engineered cysteine residue by a charged reagent resulted in mobility shifts on SDS-PAGE. (31) Unlike this earlier method, OCAC uses limited proteolysis to create alternative charge forms and does not require the complex to be stable in the presence of sodium dodecyl sulfate. The OCAC method also avoids the potential challenges of using chemical approaches when multiple cysteine residues are present in a protein.

We applied the OCAC method to examine the oligomeric state of EutN, which we had earlier reported to be a hexamer based on its crystal structure. EutN from *E. coli* was cloned to include a C-terminal TEV site followed by LEKK-6His. The resulting net charge difference between uncleaved and cleaved EutN is one charge unit (at pH 8.6 of the native gel running buffer). After recombinant expression and purification, this construct was subjected to a series of increasing concentrations of TEV protease, and then run on a native gel.

Surprisingly, the results clearly indicate that EutN is a pentamer in solution (Fig 1.2C). Six sharp bands are obtained by separation on a native gel. These arise from oligomers having 0, 1, 2, 3, 4, or 5 tails cleaved. This pentameric behavior contradicts two independent crystal structures in the PDB showing EutN in a hexameric arrangement (18)(PDB ID 2Z9H and 2HD3). A comparison of conditions used to obtain crystals of EutN did not suggest a clear explanation for the observed difference between crystalline and solution oligomer states. Although we see no evidence in the OCAC method for hexameric assembly, it is possible that a minor hexameric species of EutN exists in solution, which happens to crystallize preferentially. If the protein can equilibrate between pentameric and (minor) hexameric forms in solution, a low energy crystal form could drive the protein to a hexameric configuration.

Our combined results demonstrating the pentameric states of both GrpN and EutN resolve a looming doubt regarding a particular aspect of microcompartment formation. While the role of BMC type hexamers in forming the flat facets of MCPs has been apparent since the first structural studies (3, 14), the universality of pentameric bacterial microcompartment vertex (BMV) proteins in MCPs has remained uncertain.(18, 22) Here we confirm the first pentameric assemblies of BMV proteins from MCPs besides the carboxysome: GrpN from *Rhodospirillum rubrum* by X-ray crystallography, and EutN from *E. coli* in solution by the OCAC method. These results help reunify ideas for how different types of MCPs are constructed.

MATERIALS AND METHODS

Cloning. A gene sequence was designed, using codons optimized for expression in *E. coli*, to encode GrpN with a C-terminal hexahistidine tag, using the online program DNAWorks.(32) The resulting nucleotide sequence was synthesized by Biomatik, and then transferred to pET22b(+) vector using the NdeI restriction site via Isothermal Assembly (a.k.a. Gibson Assembly)(33) We followed the isothermal protocol utilizing 20 nucleotide base pair complementary overhangs.

Full length EutN was amplified from *E. coli* genomic DNA. Synthetic DNA oligomers, purchased from IDT Inc., were used to add a TEV cleavage site and the positively charged tail onto the C-terminus of EutN with PCR. We refer to the resulting protein as EutN(+). Using Isothermal Assembly, EutN(+) was transferred into a pET22b(+) vector between NdeI and XhoI restriction sites in order to append a C-terminal 6xHis tag. The sequences of both GrpN and EutN(+) were verified by Laragen, Inc.

Expression and Purification. The expression of GrpN-6His and EutN(+) were induced with 1mM IPTG in BL21 cells, shaking at 250 rpm, for 3 to 5 hours at 37˚C. Cells were spun

down for 5 min at 6,000 rpm and stored at -20˚C. Cells were suspended in 50mM Tris-HCl pH 7.6, 300mM NaCl, 20 mM imidazole, with Protease Inhibitor Cocktail (Sigma, Cat # P8849) and lysed by sonication. Cells were spun down in rotor SS-34 at 16,500 rpm for 30 min, filtered through a 0.2um filter, and applied to a Hi-trap Nickel column by syringe at room temperature. Protein was eluted in one step with of 50mM Tris –HCl pH 7.6, 300mM NaCl, 400mM Imidazole. GrpN was dialyzed into 2L of 10mM Tris-HCl pH 7.6, 20 mM NaCl for 1 hour at 4˚C, and then again in 2L of fresh buffer overnight. GrpN-6His tended to precipitate even at 4˚C, so the protein solution was subjected to centrifugation at 20,800 rcf for 1-2 minutes throughout the protein preparation process to pellet protein precipitate.

Crystal Structure Determination. Initial crystallization screens were performed in 96-well, hanging drop trays, set up with the nanoliter liquid handling Mosquito from TTP LabTech. Upon optimization of condition G8 from Hampton Research screen HR2-110, cube-shaped crystals were obtained within the following condition ranges: 1.4 M - 1.6 M ammonium sulfate, 0.1 M NaCl, 0.1 M HEPES pH 7.0 – 7.6, at protein concentrations between 20 mg/mL and 40 mg/mL. Hanging drops were 1:1 well: protein with a total drop size between 2 and 4 uL. Crystals generally took between 1 and 10 days to grow.

Diffraction data extending to 3.2 Å resolution were collected at the Argonne National Laboratory, Advanced Photon Source (APS), beamline 24-ID-C. The structure of GrpN was phased by molecular replacement using the program PHASER (34) Coordinates for the CcmL pentamer (PDB accession code 2QW7) were used as the search model. The structure was built using the program COOT (35) and refined using PHENIX (36) and BUSTER (37) with a final R_{work} and R_{free} of 0.2671 and 0.2885 respectively. 95% of the backbone dihedral angles are

within the favored regions of a Ramachandran diagram. Coordinates and structure factors have been deposited with the PDB with ID code 4I7A.

OCAC Assay. Aliquots of purified EutN(+) (in 50mM Tris-HCl pH 7.6, 300mM NaCl and 300mM imidazole) were incubated on ice with fresh 1mM DTT for 30 minutes. Due to weak UV absorbance at 280 nm, concentrations were adjusted based on band intensity visualized by SDS-PAGE. TEV protease was added to EutN(+) aliquots in a series of dilutions, resulting in the following final concentrations: 0.0, 0.002, 0.01, 0.02, 0.1, 0.5 mg/mL TEV protease. These TEV: EutN(+) samples were incubated at room temperature for 15 minutes. Protease reactions were stopped with addition of 5X native loading dye containing 10mM iodoacetamide. Reactions were then run on a native gel (BioRad CAT# 456-1096) at 100V volts for 2 hours at room temperature.

ACKNOWLEDGMENTS

This work was supported by NIH grant R01AI08114 to TOY. NMW is supported by the Ruth L. Kirschstein National Research Service Award GM007185. The authors thank Michael Thompson, Julien Jorda, Thomas Bobik and members of the Yeates lab for helpful discussions. We thank Michael Sawaya and the staff of the UCLA-DOE X-ray Crystallography Core Facility, which is supported by DOE Grant DE-FC02-02ER63421. We thank M. Capel, K. Rajashankar, N. Sukumar, J. Schuermann, I. Kourinov and F. Murphy at NECAT beamlines 24-ID at APS, which are supported by NIH grants P41RR015301 and P41GM103403. Use of the APS is supported by the DOE under contract DE-AC02-06CH11357.

The authors declare that they have no competing interests related to this manuscript.

SUPPLEMENTARY TEXT

Interestingly, the majority of BMVs crystalized so far did so as pentamers, while E. coli EutN crystalized twice as hexamers (Fig 1.3A). In the meantime, our OCAC assay showed that the major form of EutN in solution is pentamer. This suggests that EutN has an equilibrium between pentamers and hexamers. While pentamers dominate in solution, hexamers crystalize more easily. The question then is, what feature in EutN enables this pentamer-hexamer transition? After observing the primary sequences of the crystallized BMVs, I found that all BMVs with pentameric crystal form have polar amino acids at residue 14 at the interface between subunits, which do not interact with the highly conserved hydrophobic residue 47. However, EutN has Val at residue 14, which forms a hydrophobic interaction with residue 47. This interaction possibly favors a tighter interface and introduces an angular twist to the adjacent subunit, allowing the insertion of an additional subunit (Fig 1.3B). Experiments to test this hypothesis are yet to be completed.

Table 1.1. X-ray data collection and model refinement statistics.

Statistics	**Value**
Wavelength (Å)	0.9791
Resolution range (Å)	19.8 - 3.2 (3.314 - 3.2)
Space group	I23
Unit cell (Å)	a=b=c=150.7
Total reflections recorded	49986 (3856)
Unique reflections	9248 (700)
Multiplicity	5.4 (5.5)
Completeness (%)	96.6 (99.6)
Mean I/sigma(I)	21.0 (3.0)
Wilson B-factor (Å^2)	84.4
R-meas[29]	7.3% (62%)
Model R-work	0.267 (0.340)
Model R-free	0.288 (0.381)
Number of non-hydrogen atoms	2930
macromolecules	2926
chloride	3
water	1
Protein residues	423
Geometric deviations (rms)	
bonds (Å)	0.004
angles (°)	0.69
Ramachandran favored (%)	95
Ramachandran outliers (%)	0
Clashscore[30]	13.35
Average B-factor (Å^2)	87.6
protein atoms	87.6
solvent (water)	82.2

Statistics for the highest-resolution shell are shown in parentheses.

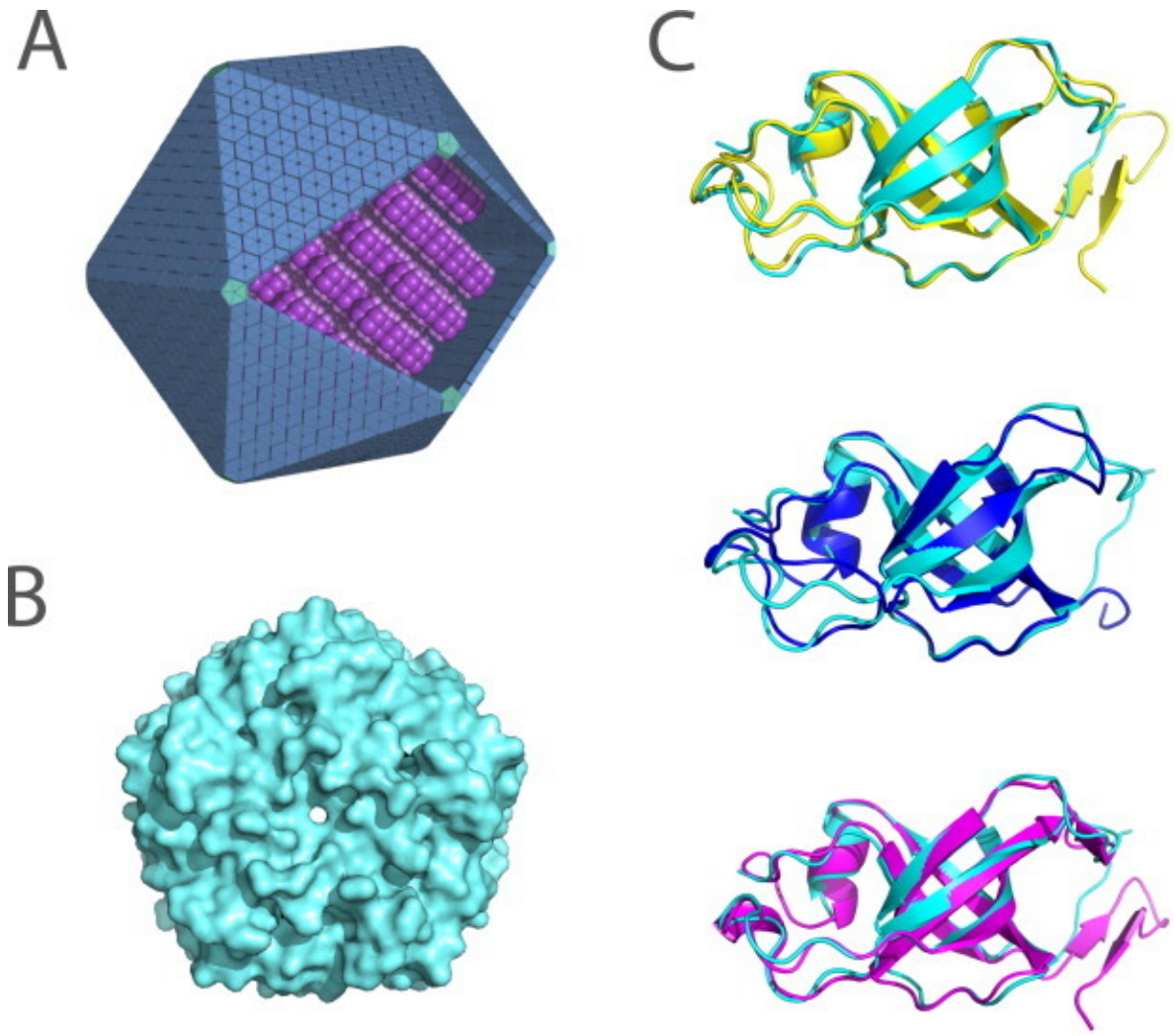


Figure 1.1 Comparison of the structure of GrpN with other bacterial microcompartment vertex (BMV) proteins.

(A) An idealized model of an MCP showing pentameric units at the vertices of the polyhedral shell. (B) Space-filling model of the pentameric structure of GrpN. (C) Superposition of a GrpN monomer with CcmL (top, yellow), CsoS4A (middle, blue), and EutN (bottom, magenta).

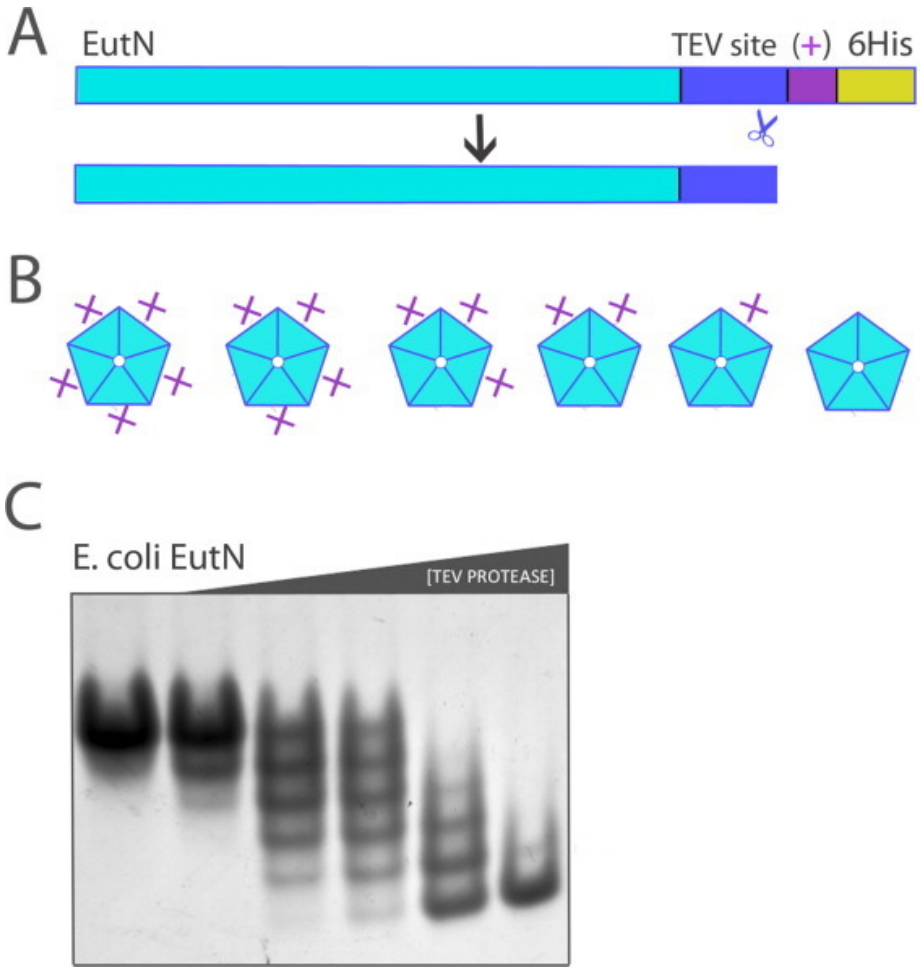


Figure 1.2 Application of the OCAC (oligomeric characterization by the addition of charge) method to the EutN shell protein.

(A) Diagram of primary structure of EutN before and after cleavage with TEV protease. The plus symbol "+" denotes one additional positive charge. (B) For an *n*-oligomer, *n+1* possible charge states exist. Shown are the six possible charge-states of a pentamer. (C) EutN OCAC native gel. The engineered EutN(+) oligomer was incubated with TEV protease for 15 minutes, quenched with iodoacetamide, and then run on a native gel. From left to right, TEV protease concentrations are 0.0, 0.002, 0.01, 0.02, 0.1, 0.5 mg/mL. The presence of six distinct bands shows that EutN is a pentamer.

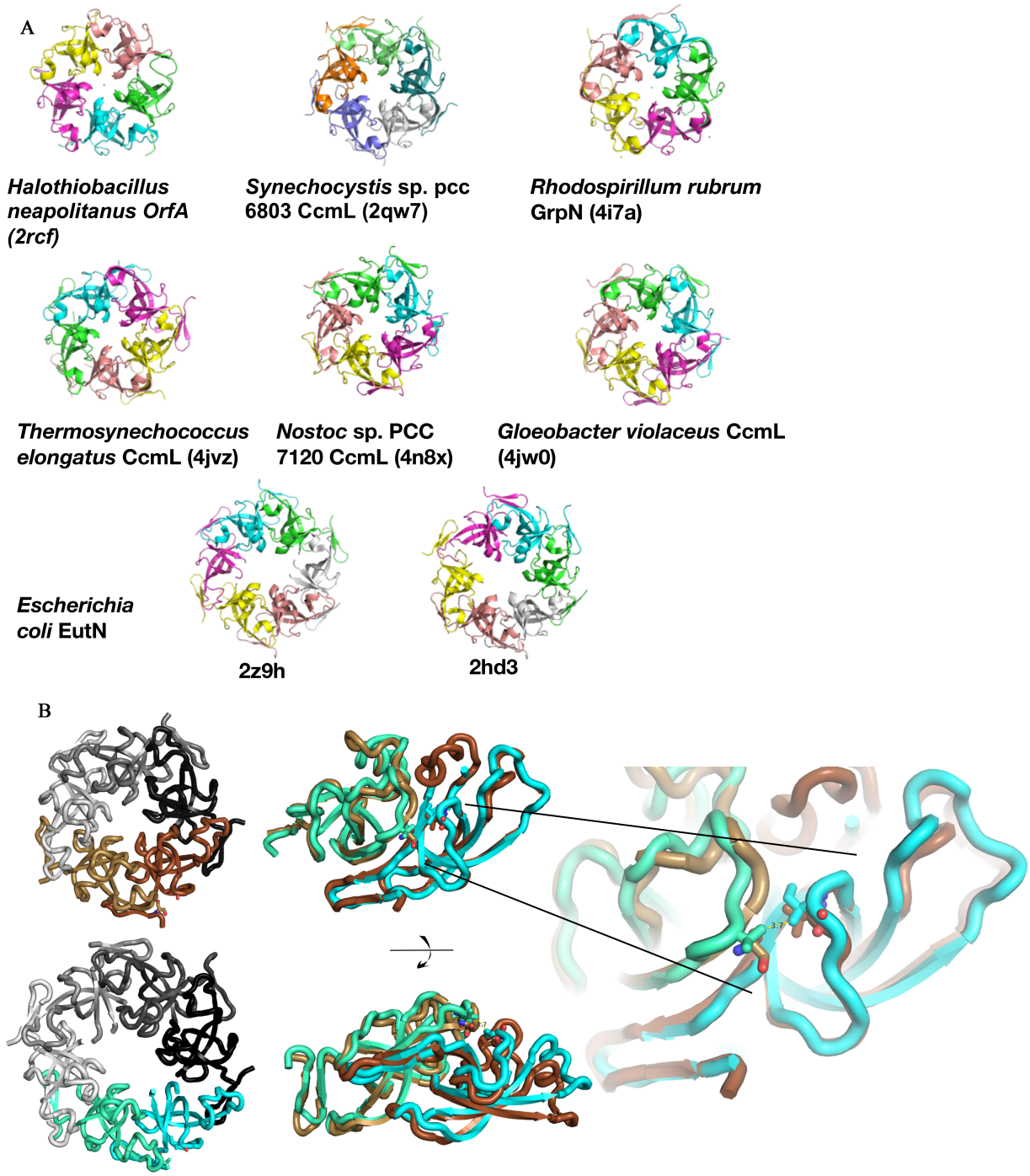


Figure 1.3. Comparison of BMV crystal structures in pentameric and hexameric forms.

A. List of current BMV crystal structures, shown in cartoon with each chain colored differently. Source organism, protein name, and PDB ID are listed below each structure (1, 17, 38–40). B. Detailed comparison between inter-subunit interactions between a pentameric BMV CcmL

(2qw7) and *E. coli* EutN (2z9h). Left column: CcmL (top) and EutN (bottom) viewed down the symmetry axes. The subunits that are used in comparison are colored in chocolate & brown in CcmL and green & cyan in EutN. The rest of the subunits are colored in different shades of grey. Middle column: CcmL subunit in chocolate and EutN subunits in green are aligned. Residues S14 on CcmL subunit (in chocolate), V47 on CcmL subunit (in brown), V14 on EutN subunit (in green) and I47 on EutN subunit (in cyan) are shown in sticks with O atoms in red and N atoms in blue. Left column: zoom in on the residues 14 and 47 surrounding area.

REFERENCES

1. Wheatley NM, Gidaniyan SD, Liu Y, Cascio D, Yeates TO (2013) Bacterial microcompartment shells of diverse functional types possess pentameric vertex proteins. *Protein Science* 22(5):660–665.

2. Kerfeld CA, Heinhorst S, Cannon GC (2010) Bacterial microcompartments. *Annu Rev Microbiol* 64:391–408.

3. Yeates TO, Thompson MC, Bobik TA (2011) The protein shells of bacterial microcompartment organelles. *Curr Opin Struct Biol* 21(2):223–231.

4. Yeates TO, Crowley CS, Tanaka S (2010) Bacterial microcompartment organelles: protein shell structure and evolution. *Annu Rev Biophys* 39:185–205.

5. Cannon GC, et al. (2001) Microcompartments in prokaryotes: carboxysomes and related polyhedra. *Appl Environ Microbiol* 67(12):5351–5361.

6. Cheng S, Liu Y, Crowley CS, Yeates TO, Bobik TA (2008) Bacterial microcompartments: their properties and paradoxes. *Bioessays* 30(11–12):1084–1095.

7. Price GD, Coleman JR, Badger MR (1992) Association of carbonic anhydrase activity with carboxysomes isolated from the cyanobacterium *Synechococcus* PCC7942. *Plant Physiol* 100(2):784–793.

8. Sampson EM, Bobik TA (2008) Microcompartments for B12-dependent 1,2-propanediol degradation provide protection from DNA and cellular damage by a reactive metabolic intermediate. *J Bacteriol* 190(8):2966–2971.

9. Penrod JT, Roth JR (2006) Conserving a Volatile Metabolite: a Role for Carboxysome-Like Organelles in *Salmonella enterica*. *J Bacteriol* 188(8):2865–2874.

10. Brinsmade SR, Paldon T, Escalante-Semerena JC (2005) Minimal Functions and Physiological Conditions Required for Growth of *Salmonella enterica* on Ethanolamine in the Absence of the Metabolosome. *J Bacteriol* 187(23):8039–8046.

11. Stojiljkovic I, Bäumler A, Heffron F (1995) Ethanolamine utilization in *Salmonella typhimurium*: nucleotide sequence, protein expression, and mutational analysis of the cchA cchB eutE eutJ eutG eutH gene cluster. *J Bacteriol* 177(5):1357–66.

12. Bobik T, Xu Y, Jeter R, Otto K, Roth J (1997) Propanediol utilization genes (pdu) of *Salmonella typhimurium*: three genes for the propanediol dehydratase. *J Bacteriol* 21(179):6633–9.

13. Jorda J, Lopez D, Wheatley NM, Yeates TO (2012) Using comparative genomics to uncover new kinds of protein-based metabolic organelles in bacteria. *Protein Sci*:n/a-n/a.

14. Kerfeld CA, et al. (2005) Protein structures forming the shell of primitive bacterial organelles. *Science* 309(5736):936–938.

15. Samborska B, Kimber MS (2012) A dodecameric CcmK2 structure suggests β-carboxysomal shell facets have a double-layered organization. *Structure* 20(8):1353–1362.

16. Tsai Y, et al. (2007) Structural analysis of CsoS1A and the protein shell of the *Halothiobacillus neapolitanus* carboxysome. *PLoS Biol* 5(6):e144.

17. Tanaka S, et al. (2008) Atomic-level models of the bacterial carboxysome shell. *Science* 319(5866):1083–1086.

18. Tanaka S, Sawaya MR, Yeates TO (2010) Structure and mechanisms of a protein-based organelle in *Escherichia coli*. *Science* 327(5961):81–84.

19. Price GD, Howitt SM, Harrison K, Badger MR (1993) Analysis of a genomic DNA region from the cyanobacterium *Synechococcus sp. strain* PCC7942 involved in carboxysome assembly and function. *J Bacteriol* 175(10):2871–2879.

20. Cheng S, Sinha S, Fan C, Liu Y, Bobik TA (2011) Genetic analysis of the protein shell of the microcompartments involved in coenzyme B12-dependent 1,2-propanediol degradation by *Salmonella*. *J Bacteriol* 193(6):1385–1392.

21. Parsons JB, et al. (2008) Biochemical and Structural Insights into Bacterial Organelle Form and Biogenesis. *J Bacteriol* 283(21):14366–14375.

22. Cai F, et al. (2009) The pentameric vertex proteins are necessary for the icosahedral carboxysome shell to function as a CO2 leakage barrier. *PLoS ONE* 4(10):e7521.

23. Schmid MF, et al. (2006) Structure of *Halothiobacillus neapolitanus* carboxysomes by cryo-electron tomography. *J Mol Biol* 364(3):526–535.

24. Iancu CV, et al. (2007) The structure of isolated *Synechococcus strain* WH8102 carboxysomes as revealed by electron cryotomography. *J Mol Biol* 372(3):764–773.

25. Lai Y-T, King NP, Yeates TO (2012) Principles for designing ordered protein assemblies. *Trends Cell Biol* 22(12):653–661.

26. Lai Y-T, Cascio D, Yeates TO (2012) Structure of a 16-nm cage designed by using protein oligomers. *Science* 336(6085):1129.

27. Yang Y, Burkhard P (2012) Encapsulation of gold nanoparticles into self-assembling protein nanoparticles. *J Nanobiotechnology* 10:42.

28. King NP, et al. (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* 336(6085):1171–1174.

29. Wörsdörfer B, Woycechowsky KJ, Hilvert D (2011) Directed evolution of a protein container. *Science* 331(6017):589–592.

30. Gandhi CS, Walton TA, Rees DC (2011) OCAM: A new tool for studying the oligomeric diversity of MscL channels. *Protein Sci* 20(2):313–326.

31. Gouaux J, et al. (1994) Subunit stoichiometry of staphylococcal alpha-hemolysin in crystals and on membranes: a heptameric transmembrane pore. *Proc Natl Acad Sci USA* 91(26):12828–31.

32. Hoover DM (2002) DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res* 30(10):43e – 43.

33. Gibson DG, et al. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6(5):343–345.

34. McCoy AJ, et al. (2007) Phaser crystallographic software. *J Appl Crystallogr* 40(Pt 4):658–674.

35. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60(Pt 12 Pt 1):2126–2132.

36. Adams PD, et al. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66(Pt 2):213–221.

37. Headd JJ, et al. (2012) Use of knowledge-based restraints in phenix.refine to improve macromolecular refinement at low resolution. *Acta Crystallogr D Biol Crystallogr* 68(4):381–390.

38. Sutter M, Wilson SC, Deutsch S, Kerfeld CA (2013) Two new high-resolution crystal structures of carboxysome pentamer proteins reveal high structural conservation of CcmL orthologs among distantly related cyanobacterial species. *Photosynth Res* 118(1–2):9–16.

39. Keeling TJ, Samborska B, Demers RW, Kimber MS (2014) Interactions and structural variability of β-carboxysomal shell protein CcmL. *Photosynth Res* 121(2–3):125–133.

40. Forouhar F, et al. (2007) Functional insights from structural genomics. *J Struct Funct Genomics* 8(2–3):37–44.

Chapter 2. Study of Bactofilin, A Novel Bacterial Cytoskeleton

ABSTRACT

Bacteria are complex organisms with extensive intracellular organizations. While they have many cytoskeleton proteins that are homologous to those present in eukaryotes, there are also many unique to the bacteria kingdom, among which are the bactofilins. Bactofilins are fiber-forming proteins that are widely conserved among bacteria. These proteins have roles in diverse biological functions including but not limited to cell motility, cell wall synthesis and modification. Of particular interest is the *Helicobacter pylori* bactofilin CcmA, which is important for maintaining the helical cell shape of *H. pylori* as well as its virulence. The biological role of *H. pylori* CcmA is not fully understood. Another interesting bactofilin *Myxococcus xanthus* BacO is one of the four bactofilin genes the organism carries. At the moment, there is only one atomic-level structure solved for the bactofilin family. There is also little understanding of the mechanism and regulation of bactofilin fiber formation. This project hoped to contribute to our understanding of how CcmA regulates *H. pylori* cell shape and virulence. A better understanding of CcmA and other bactofilins may elucidate a previously unknown aspect of bacterial physiology and open up new avenues for development of novel classes of antibiotics that interfere with bacteria-specific biologically processes, such as cell shape regulation and cell wall synthesis. The project described in the chapter originally aimed to use a combination of X-ray crystallography, computational structural biology, electron microscopy and other biochemical techniques to elucidate the bactofilin structures at the atomic level and to determine the structure of the CcmA polymer. Since I wasn't able to obtain a structure, this chapter will simply record a selected few of the interesting observations I made along the way without drawing any early conclusions.

INTRODUCTION

Bacterial cells are highly organized by filaments and cytoskeletons. Homologs to all three major groups of eukaryotic cytoskeletons are present in bacteria. FtsZ forms rings and is homologous to tubulin. The two proteins share a common 3D fold, but have distinct primary structure, kinetics, and biological functions (1). FtsZ forms a Z-ring at the site of separation during cell division, which serves as a scaffold, recruiting and positioning enzymes involved in peptidoglycan synthesis (2). Similarly, MreB is structurally and evolutionally related to actin (3). MreB plays a critical role in determining the rod shape of many species (4). CreS from *Caulobacter crescentus* has the same domain organization as intermediate filaments. In *C. crescentus*, CreS forms a membrane-associated sheet on one side of cell and reduces the insertion into the peptidoglycan by introducing mechanical stress on this side, therefore, giving rise to the curved cell shape characteristic of *C. crescentus* (5). In addition to the eukaryotic cytoskeleton homologs, several polymer-forming proteins unique to bacteria have been identified, but have not yet been studied in detail. The most recent addition to this group is the bactofilin family.

The bactofilin family is widely conserved among bacteria (Fig 2.1A) (6). Its homologs are found among all major branches of the bacteria kingdom. In addition, many organisms encode more than one copy of bactofilins in their genome. Bactofilins participate in a variety of processes in different organisms. For example, BacP from *Myxococcus xanthus* plays a role in social motility, which depends on the positioning of Type IV pili on the leading end of the cell pole. BacP forms fibers at the opposite ends of the cells and SofG small GTPase shuttles on these BacP fibers. The type IV pili motor ATPases PilB and PilT are sorted to the correct poles by the shuttling SofG and other GTPases (7). BacE and BacF from *Bacillus subtilis* are also involved in

cell mobility. They are essential for motility and are required for the establishment of flagellar hook- and filament structures of the flagellum, but not for the formation of basal bodies (8). In *Caulobacter crescentus*, BacA and BacB play a role in cell wall modification and cellular morphogenesis. BacA and BacB are recruited to the stalked pole at the onset of S phase, forming a membrane-associated sheet. Purified BacA forms fibers *in vitro* (Fig 2.1B) and overexpressed BacA can form extensive membrane-associated sheets, which eventually cause cell to swell and burst. Moreover, BacA and BacB co-immunnoprecipitation pulls down PbpC, a cell wall biosynthesis enzyme (6). These results indicate that BacA and BacB can mediate cell shape through direct interaction with proteins that regulate cell wall biosynthesis.

A bactofilin family protein CcmA was discovered in *Helicobacter pylori* to be required for cell morphology. *H. pylori* is the most common human bacterial pathogen; it colonizes the gastric mucosa in approximately 50% of the world population. Infection persists throughout the patient's life unless treated. *H. pylori* can cause a variety of diseases, ranging from mild conditions such as gastritis and peptic ulcers, to more severe illness such as gastric cancer and mucosal-associated lymphoma (9). *H. pylori* has a helical cell shape that is believed to be important in aiding the bacteria to penetrate the viscous gastric mucosa. Recently, a random transposon insertion screen identified a *csd1* mutant that was rod-shaped instead of helical (10). Csd1 is a homolog of LytM- type endopeptidases from Staphylococcus aureus. Based on this homology, it is predicted to catalyze the hydrolysis of the peptide crosslinks making up the peptidoglycan. A similar phenotype is observed for mutants of two other Csd1 homologs encoded in the H. pylori genome, Csd2 and Csd3. Interestingly, CcmA, a bactofilin, is encoded together in a predicted operon with Csd1 and Csd2. Additionally, the CcmA mutant displays phenotypes similar to the csd1 mutant (Fig 2.2). Furthermore, rod-shaped mutants csd1 and

CcmA are out-competed by helical wild type in mouse stomach colonization test, supporting the notion that the helicity plays a role in *H. pylori* virulence. Biophysical models show that a helical cell shape can be formed from rod-shaped cells by selective relaxation of the peptidoglycan crosslinks (11). While the finding that csd1-3 mutants lack the helical cell shape is consistent with this theory, how CcmA regulates the cell shape is still unknown.

Beyond the aforementioned studies about bactofilins, we know very little about this family. There is only one NMR structure of *Caulobacter vibrioides* BacA. The lack of extensive tertiary structure also limits our understanding of how monomers polymerize into fibers or sheets, and therefore, the regulation of polymerization *in vivo*. For example, *C. crescentus* BacA forms stable fibers in vitro. It is constantly present in the cytoplasm but only polymerizes at the stalk at the time of stalk initiation (6). How *C. crescentus* controls the BacA polymerization state is unclear. Particular to *H. pylori*, there is no direct evidence on the spatial and temporal arrangement of CcmA polymers. Additionally, though the CcmA gene is found within 300 bp of the endopeptidase gene in many organisms (12), it is unknown whether CcmA interacts with Csd1-3 or other enzymes involved in cell wall modification and cellular morphogenesis. This project aimed to elucidate the CcmA structure with atomic-level accuracy in both the monomeric and polymeric forms. Since there was no structural data obtained, the remaining of this chapter will focus on recording some of the interesting observation without drawing early conclusions. As the project developed, two papers studying the bactofilin structure with solid state NMR were published (13, 14)., which show that the *C. vibrioides* BacA fold into a beta solenoid structure with a tightly packed core formed by the conserved hydrophobic residues. My preliminary observations are consistent with the findings in these papers.

PRELIMINARY RESULTS

Computational predictions of bactofilin structures. The majority of bactofilins contains a conserved bactofilin domain flanked by unconserved, unstructured flexible loops (15, 16). While it is possible the flexible loops provide important contact points for diverse binding partners and enable bactofilins to participate in different cellular processes, this project mainly concerned with the conserved, folded bactofilin domain. The Hidden Markov Model (Fig. 2.3A) (17) for the bactofilin domain shows that there are 1) conserved hydrophobic residues at alternating positions; 2) highly conserved glycine residues. A few computational attempts were made to predict the bactofilin structure, including Rosetta (18) and HHpred (19), in combination with co-evolutionary analysis tools such as BCOV (20), PsiCOV (21), and evfold (22). The best results, or the results that have the most resemblance to folded proteins judged by eyes, were given by Robetta (Fig 2.3B-C) (23).

The overall fold for Robetta-generated models were quite similar to the solid state NMR structure solved later (13, 14). Compared to the Robetta models, the NMR structure had better packed core and its backbone hydrogen bonds were better satisfied by its secondary structure. However, all structures used the conserved hydrophobic residues to fill the inside of the beta solenoid and placed the highly conserved glycines on the turns between beta strands.

Crystallization attempts. The initial aim of this project was to solve a crystal structure of a bactofilin domain. No diffracting crystals were obtained. Here, I simply list the constructs tested as a guide for future studies (Table 2.1).

After the initial observations that His6 tagged *H. pylori* CcmA were invariably insoluble (pYL2-5) whether with or without the presences of flanking flexible loops (residues 1-16 on the N-terminus and residues 121-136 on the C-terminus), I introduced a series of single or double

mutations based on co-evolution analysis (results not shown) or based on the computationally predicted models (pYL11-15,18,19). The overall aim was to remove residues that were likely to make inter-subunit contacts, as polymeric and fibrous complexes hardly crystallize. Unfortunately, these constructs were not soluble either. They could be purified in the unfolded state, but aggregated upon refolding. Under negative stain EM, these aggregates appeared to be composed of regularly arranged protein subunits (Fig 2.4A) in two-dimension. This result hints that, at least under the *in vitro* purification condition I adopted, CcmA is able to form more than one type of interface among each other. A similar observation was made on *C. crescentus* BacA purified from a native source (Fig 2.4B) (13).

In the next round of attempts, I included the *M. thanxus* BacO in addition to CcmA and started to use the surface entropy reduced maltose binding protein (SER MBP) as a tag (pYL16,17, 20-24,26-42,43-51). The intention was to interfere with the fiber formation process by genetically fusing CcmA to a larger carrier protein that can sterically hinder CcmA polymerization. This idea is further supported by the finding that genetically fusing *M. xanthus* BacM to mCherry changed or disrupted its cellular localization (24). MBP is also frequently used as carrier protein to increase the solubility of a target protein. Furthermore, by mutating the surface-exposed residues to alanine and removing the flexible linker between SER MBP and the target protein, SER MBP is optimized for crystallization purposes (25). These constructs could be purified solubly (Fig 2.5A), but the purified species eluted in the void volume on size exclusion chromatography (Fig 2.5B), ran as a smear on native Tris-glycine agarose gel (Fig 2.5C), and appeared as fiber-like assemblies under negative stain EM (Fig 2.4C, D). These findings strongly suggest that adding a large tag on the N-terminus of bactofilins is not sufficient to completely block its polymerization.

In vivo split GFP screen: over-expressed CcmA formed folded aggregation in *E.coli*.

CcmA is a challenging target due to its tendency to form fibers *in vitro*. In addition to introducing selected mutations based on predicted structures, a parallel screen based on the split GFP for protein solubility was designed (Fig 2.6) to look for soluble bactofilin mutants from randomized library (26). In this screen, a super-folder GFP (27) is split into two different plasmids compatible for co-expression: one containing strand 11 fused to the C-terminus of the test protein, the other strands 1-10. Strands 1-10 alone will fold properly and remains in the cytoplasm, but will not fluorescence unless complemented by strand 11 (28). Strand 11 will only complement strand 1-10 if the test protein is soluble. After sequential induction, soluble GFP 11 fusions spontaneously bind GFP 1–10, and the resulting fluorescence is proportional to the amount of soluble, non-aggregated GFP 11–tagged protein. Desired clones will then be picked for propagation and further characterization.

As a first step I tested this screen with wild-type full-length CcmA. While negative controls of empty GFP11 vector or a known inclusion body protein fused to GFP11 both yielded no fluorescence, CcmA fused to GFP11 and maltose binding protein (MBP) fused to GFP11 both gave strong fluorescence, even to similar intensity (Fig 2.7). This result suggests that 1) the aggregation formed by CcmA under over expression condition is not inclusion body; 2) the C-terminus of CcmA is accessible to the rest of the cytoplasm.

Other biochemical studies and observations. Consistent with the CcmA *in vivo* split GFP screen result, *M. xanthus* BacO overexpressed in *E. coli* could be purified from the cell lysate by low speed centrifugation (6). The purified BacO formed either individual fibers or bundles of fiber (Fig 2.4E). At ~150 mM NaCl salt concentration, purified BacO fibers also formed a gel-like substance (results not shown). In addition, purified BacO fibers also showed a different fiber

diffraction pattern when suspended in water vs 20mM Tris pH 8.0, 150mM NaCl (Fig 2.8). These are consistent with the observation that *C. crescentus* BacA shows different inter-fiber organization at different salt level (6).

DISCUSSION

This project was carried out in 2013. From today's point of view, the bactofilin structure might be achievable through cryo-EM. It is also imaginable that under certain conditions, the bactofilin fiber bundles would be ordered enough for cryo-electron diffraction. As of today, there is a solid state NMR structure of the bactofilin domain. However, it is still unclear how the subunits polymerize into a fiber. Vasa, et al., proposed that the bactofilin domains of *C. crescentus* BacA come together in a head-to-tail fashion to form a straight fiber (13). Meanwhile, we cannot rule out that bactofilin domains have other ways to associate with each other. For example, they can polymerize in a head-to-head, tail-to-tail fashion. Furthermore, the fiber can have a twist, instead of being straight. Additionally, as I and others have observed, purified bactofilin fibers can form additional interfaces and make sheets/bundles under different conditions. These interactions are reversible to a certain extent. Do bactofilin fibers associate with each other and change the way of association *in* vivo? Do these observed reversible interactions represent the type of sheets/bundles formed *in vivo*?

More questions remain regarding the biological roles of the wide-spread bactofilins. So far, the characterized bactofilins tend to locate near cell membranes (6) and associate with cell wall modeling systems (6, 24) or cell mobility machineries (7, 8). Do they serve as a hub for other proteins to assemble and organize? Do they change the physical property of the surrounding cell membrane and cytoplasm? As observed, *C. crescentus* BacA and BacB are present in cytoplasm at stable concentration but only locate to the stalk during the swarmer-to-

stalked cell transition (6). Therefore, *C. crescentus* must have a way to regulate BacA and BacB polymerization states and/or sheet formation in synchronization with the cell cycle. This mechanism is yet to be discovered.

MATERIALS AND METHODS

Genes and sequences. This project generated a large number of constructs. The wild-type genes where synthesized as dsDNA from IDT and cloned into pET-22b vector. Point mutations were generated with quick change PCR reactions. MBP fusion constructs were cloned into pMALX(E) vector (25). I'll list here the protein sequences of the ones whose results are discussed in the chapter.

- *Helicobacter pylori* CcmA (JHP1457)

MAIFDNNNKSANAKTGPATIIAQGTKIKGELHLDYHLHVDGELEGVVHSKNTVVIGQTGSVVGEIFANKLVVNGKFTGTVEAEVVE
IMPLGRLDGKISTQELVVERKGILIGETRPKNIQGGALLINEQEKKIENK

- *Myxococcus xanthus* BacO (MXAN4636), same as in pYL25

MSFTPRTARHTPFERRTTLMANTVIGSSIVIDGEISGDEDLVIQGTVKGKISLKESLYVEGSGVVEADIETQNVEIAGRVTGNIVA
SDKVELKTDCRVVGDIKAPRILIADGASFKGNVDMDMKER

- pYL5

MATIIAQGTKIKGELHLDYHLHVDGELEGVVHSKNTVVIGQTGSVVGEIFANKLVVNGKFTGTVEAEVVEIMPLGRLDGKISTQEL
VVERKGILIGETRPKNIQLEHHHHHH

- pYL16

MKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLAEITPAAAF
QDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKY
AAGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSAVNYGVTVLPTFKGQPSKP
FVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAV
RTAVINAASGRQTVDAALAAAQTNAAAASATIIAQGTKIKGELHLDYHLHVDGELEGVVHSKNTVVIGQTGSVVGEIFANKLVVNG
KFTGTVEAEVVEIMPLGRLDGKISTQELVVERKGILIGETRPKNIQ

- pYL35

MKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLAEITPAAAF
QDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKY
AAGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSAVNYGVTVLPTFKGQPSKP
FVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAV
RTAVINAASGRQTVDAALAAAQTNAAAASMSFTPRTARHTPFERRTTLMANTVIGSSIVIDGEISGDEDLVIQGTVKGKISLKESL
YVEGSGVVEADIETQNVEIAGRVTGNIVASDKVELKADARVVGDIKAPRILIADGASFKGNVDMDMKER

Protein expression and purification methods. 25mL of overnight LB culture with BL21(DE3) cells were used to inoculated 1 L of LB at 37 °C. Expression was induced at $O.D._{600}$ 0.6 with 1mM IPTG. Induced cultures were shaken at 37 °C for four hours until harvest. His-tagged constructs were purified under denaturing condition following the protocol below:

Bactofilin Lysis buffer: 50mM sodium phosphate monobasic, 300mM NaCl, 5mM imidazole, pH 8.0, 1 mM PMSF (Invitrogen), DNase I

Bactofilin Wash buffer 1: 50mM sodium phosphate monobasic, 300mM NaCl, 5mM imidazole, final pH 8.0, 0.05% Tween-20
Bactofilin Wash buffer 2: 50mM sodium phosphate monobasic, 300mM NaCl, 5mM imidazole, 20 mM glycine, 6M GdnHCl, final pH 8.0
Bactofilin Ni2+ buffer A: 50 mM sodium phosphate monobasic, 300mM NaCl, 5mM imidazole, 6M GdnHCl, final pH 8.0
Bactofilin Ni2+ buffer B: 50 mM sodium phosphate monobasic, 300mM NaCl, 500mM imidazole, 6M GdnHCl, final pH 8.0
Bactofilin SEC buffer: 50mM sodium phosphate monobasic, 300mM NaCl, final pH 8.0
Bactofilin Refolding buffer:50mM sodium phosphate monobasic, 300 mM NaCl, final pH 8.0, 15% glycerol, 1mM β-cyclodextrin
Bactofilin Storage buffer: 50 mM sodium phosphate monobasic, 300mM NaCl, final pH 8.0

1. Resuspend cells in 20mL Bactofilin Lysis buffer per liter of culture. Lyse cells with sonication (50% output, 5sec pulse with 3sec break, total 4min)
2. Separate cell lysate and pellets by centrifuging at 20,000g for 30min.
3. Resuspend cell pellets in 30mL of Bactofilin Wash buffer 1 per 2L of culture (first break up the pellet with a spatula, then sonicate at 50% output, 5sec pulse with 3sec break, total 2min). Stir at RT for 1h
4. The suspension is centrifuged at 15,000rpm for 20 min, after which the inclusion bodies were washed once with Bactofilin Lysis buffer.
5. Resuspend the inclusion body in 30mL of Bactofilin Wash buffer 2 per 2L of culture (first break up the pellet with a spatula, then sonicate at 50% output, 5sec pulse with 3sec break, total 2min). Stir overnight at 4 °C
6. IMPORTANT: centrifuge at 15,000rpm for 20 min, then filter supernatant with 0.45µm filters.
7. Load filtered supernatant manually to 5mL HisTrap HP column (pre-equilibrated with Bactofilin Ni2+ buffer A). Wash with 5CV of Bactofilin Ni2+ buffer A, and elute with 10CV linear gradient to 100% Bactofilin Ni2+ buffer B.
8. Refold by dialyzing against 2L of Refolding buffer twice.

MBP tagged constructs are purified following the protocol below:

Bactofilin Maltose buffer: 50mM sodium phosphate monobasic, 300mM NaCl, 10mM maltose, pH 8.0, 2mM DTT
Resuspension Buffer: 20mM Tris-base, pH 8.0, 100mM NaCl, 2mM TCEP
SEC Buffer: 10mM Tris-base, pH 8.0, 50mM NaCl, 2mM TCEP, final pH 4.5

1. Resuspend the cell pellets in 50mL of Resuspension Buffer.
2. Use the regular tip on the sonicator to lyse the cells. Settings: 5s pulse, 3s break, total 5min, 50% output.
3. Spin down the lysate by centrifuging at 28,000rpm for 20min.
4. Clean up 6mL of amylose resin (NEB) in gravity column with one 50mL water wash and one 50mL resuspension buffer wash.
5. Mix the supernatant and the amylose resin, let rotate at 4˚C for 30min.
6. Remove the flow through. Wash with 50mL of resuspension buffer.
7. Elute protein with 5 fractions of 15mL of Bactofilin Maltose buffer.
8. Concentrate down the elutions to 20mL. On the day of size exclusion chromatography, concentrate down to 2.5mL

Negative stain EM. 5 uL of sample was applied to a formvar supported carbon film on 300-mesh copper grid. The excessive sample was blotted away with filter paper after 30 sec and stained with 5 uL of 2% uranyl acetate for 30 sec. Then the grid was washed with 5 uL of filtered water twice. Air-dried grids were imaged at room temperature with FEI Tecnai T12 electron microscope equipped with Gatan 2kX2k CCD camera.

In vivo split GFP screen. The *in vivo* split GFP screen was conducted according to the protocol described in (28) with the following minor discrepancies. The plasmid carrying the strands 1-10 of the split super folder GFP was in either pRSF1b or pRSF1b backbone with a p15A origin of replication. The fluorescence of each colony was recorded with a Bio-Rad FX Pro plus Fluorimager. The random library of full-length CcmA was generated with mutagenesis PCR using GeneMorph II Random Mutagenesis Kit (Agilent) with the target mutation rate at ~ 9 mutations/kbp. The primers used are 5'-ctggtgccgcgcggcagccatatgtctgctaatgcaaaaaccggt-3' and 5'-accagaccctccatcggatccctcattgatcagcagcgca-3'. This pair of primers annealed over the flexible linker flanking the bactofilin domain, therefore allowed mutations across the entire bactofilin domain.

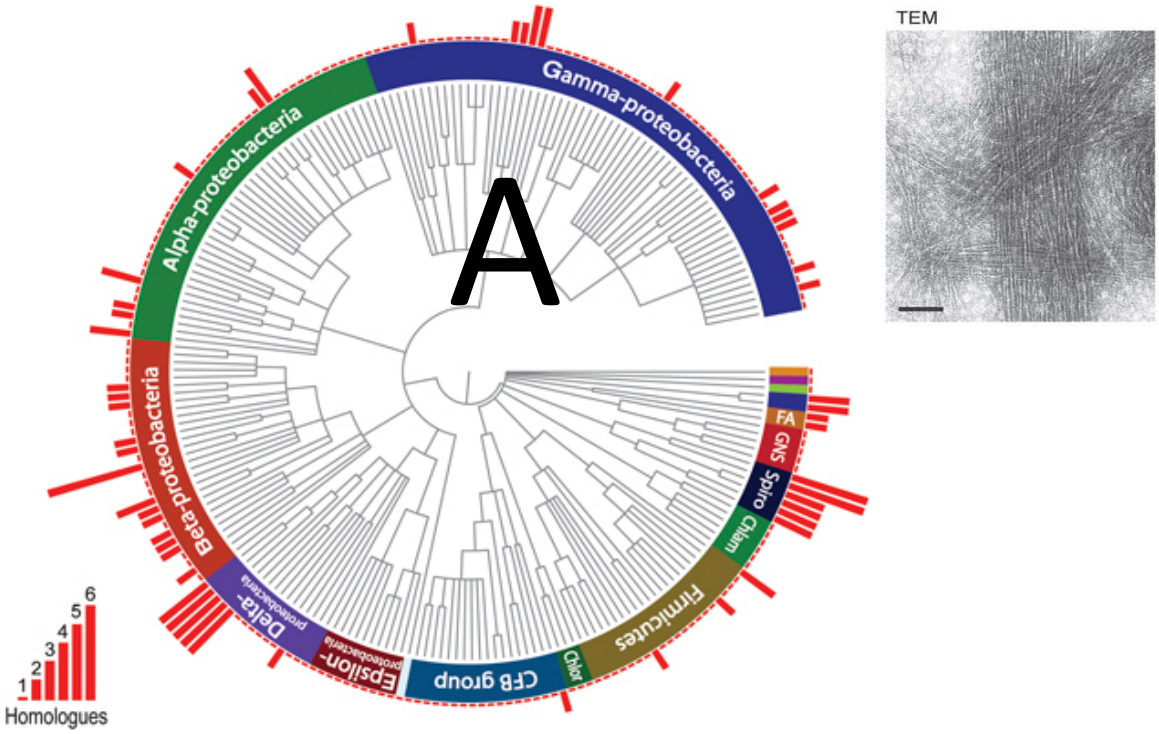


Figure 2.1. Bactofilin is a widely conserved fiber-forming protein in bacteria (6).

A. Phylogenetic distribution of bactofilin homologs among bacteria. The number of bactofilin homologues encoded in the genome is indicated by a red bar. B. Negatively stained filaments of the *C. crescentus* bactofilin homolog BacA visualized by transmission electron microscopy (TEM) (bar: 75 nm). Reprint with permission from John Wiley and Sons (license number 4422060640664).

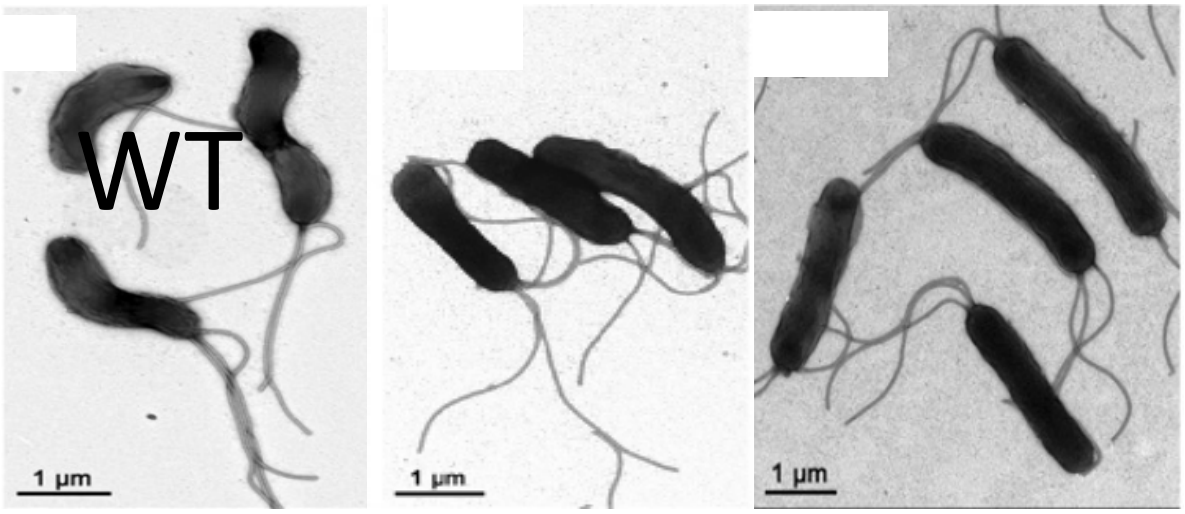


Figure 2.2. TEM images of wild-type (left), *csd1* (middle) and *CcmA* mutant (right) cells (10). *csd1* and *CcmA* mutant cells is rod-shaped rather than helical.

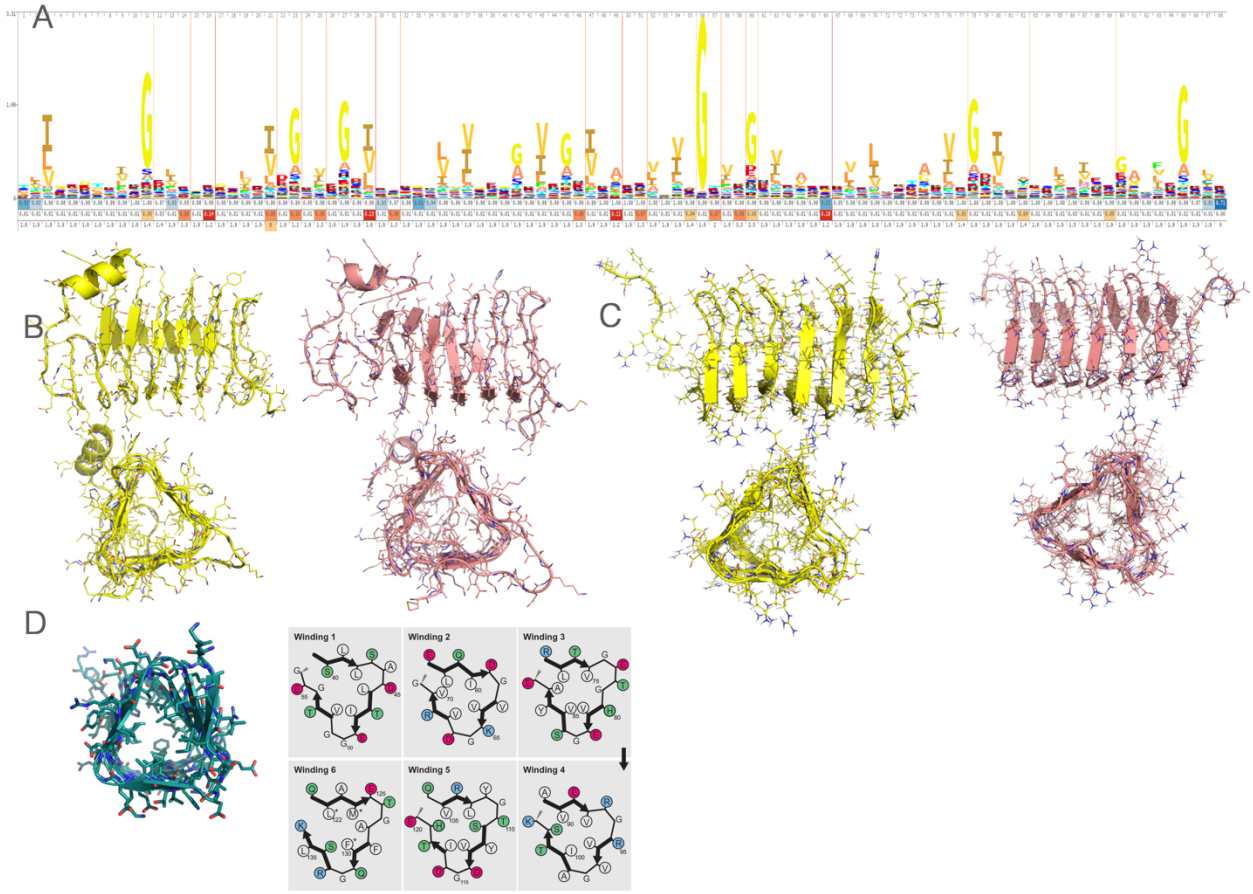


Figure 2.3. Preliminary computational results.

A. Hidden Markov Model of the bactofilin domain from the Pfam server (29). B. Lowest energy models of *H. pylori* CcmA generated by Robetta. C. Lowest energy models of *M. xanthus* BacO generated by Robetta. D. Left, solid state NMR structure (PDB ID 2N3D, one state in ensemble) of *C. vibrioides* BacA. Right, Schematic representation of the six windings. Hydrophobic residues are colored white, acidic residues are colored red, basic residues are colored blue, and others are colored green. Adapted from (14). Reprinted with permission from AAAS.

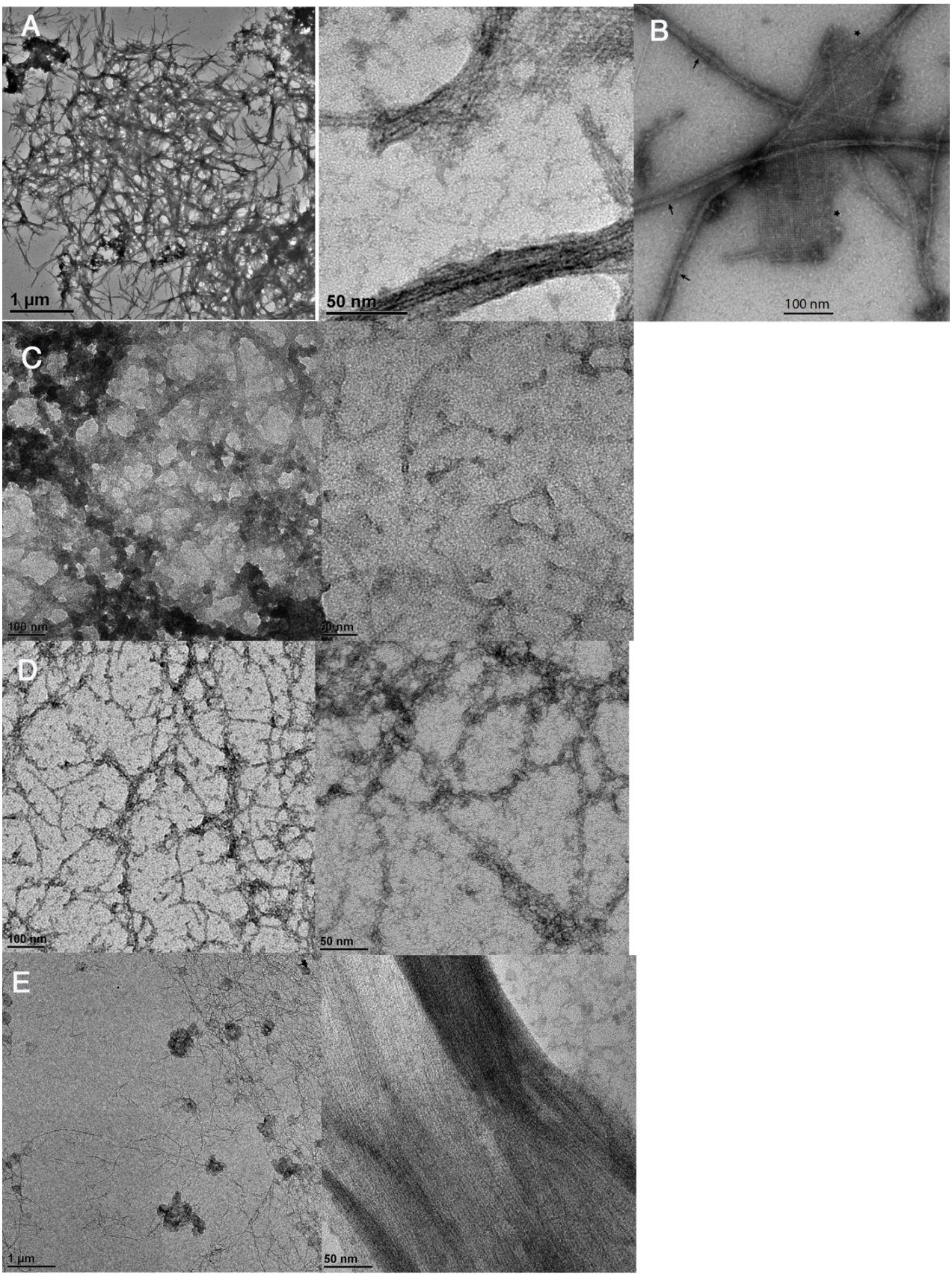

A
1 µm
50 nm
B
100 nm
C
100 nm
50 nm
D
100 nm
50 nm
E
1 µm
50 nm

Figure 2.4. Negative stain EM images of recombinantly purified bactofilin constructs.

A. Refolded pYL5 *H. pylori* J99 CcmA-His6 17-120. B. Purified *C. crescentus* BacA from native source. Adapted from (16). C. pYL16 fixed arm SER MBP-CcmA 17-120 D. pYL35 fixed arm SER MBP-BacO T94A C96A. E. Recombinantly purified BacO.

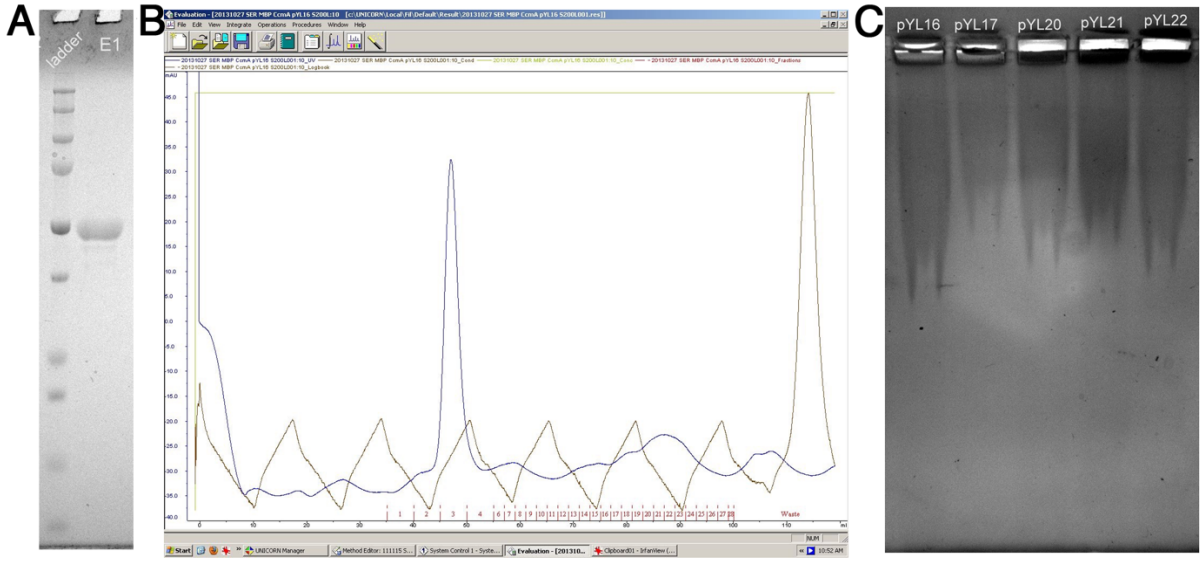


Figure 2.5. Examples of purification and preliminary biochemical analysis of SER MBP fused bactofilin constructs.

A. SDS-PAGE of pYL16 purification. Left, protein ladder. Right, first elution from amylose resin. The strongest band corresponds to pYL16. B. pYL16 size exclusion chromatography profile on Superdex 200 (GE Healthcare). C. Native agarose gel.

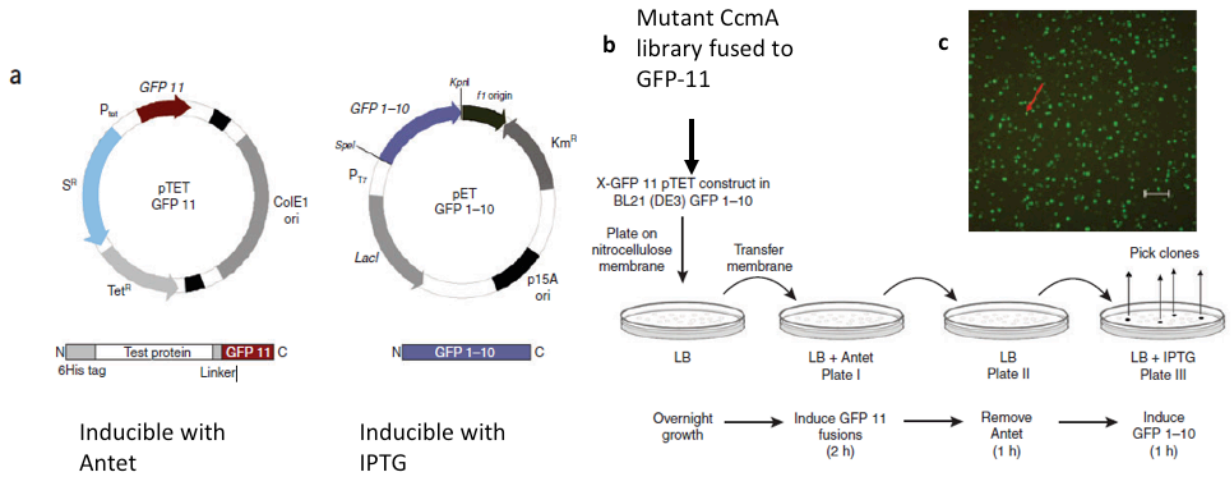


Figure 2.6. *In vivo* split GFP solubility screen (adapted from (26)).

A. The protein of interest is expressed, tagged with the GFP 11 on its C-terminus, under the control of the tet promoter (Ptet) from a pTET plasmid. GFP 1–10 is expressed under the control of the T7 promoter (PT7) from a pET plasmid. Antet, anhydrotetracyclin. B. *In vivo* solubility screening using sequential induction. c. Example of fluorescence image of *E. coli* colonies expressing protein fragments fused to GFP 11 upon complementation with GFP 1-10 after sequential induction. The red arrow indicates a sample of the colonies selected for further experiments.

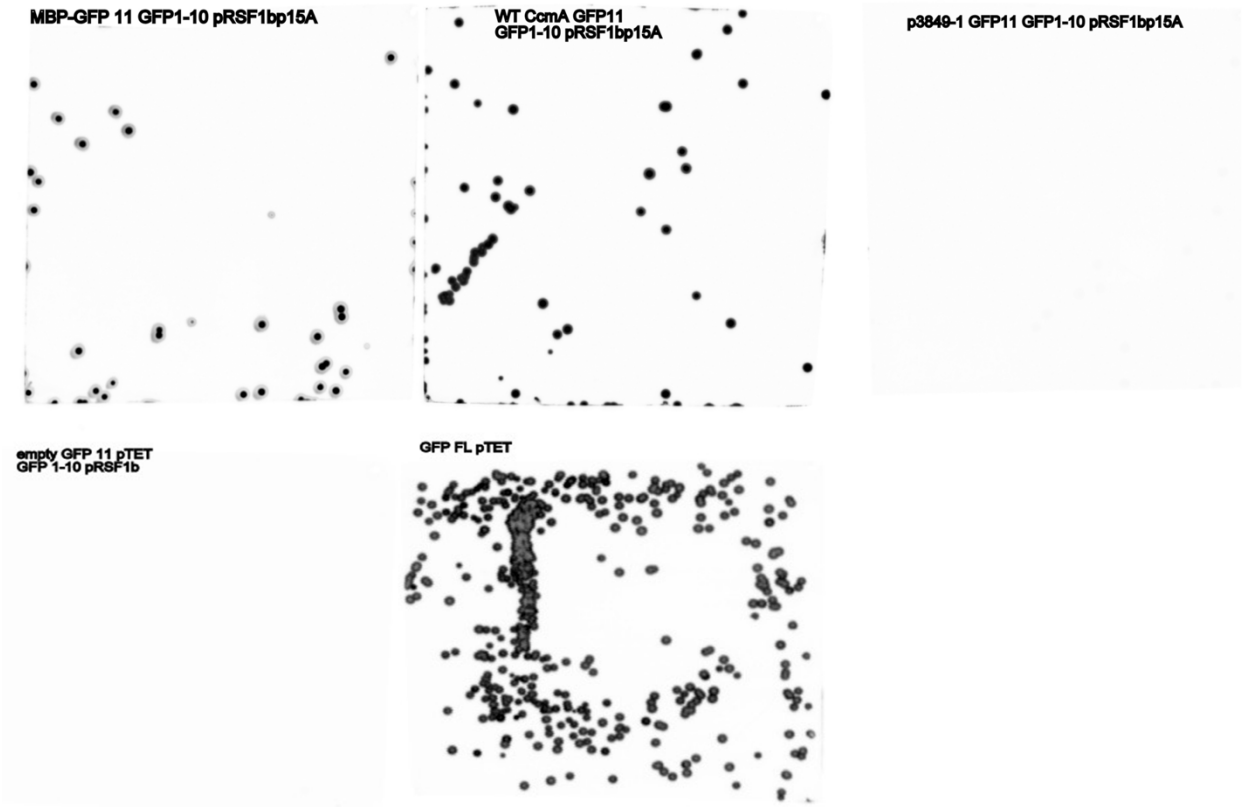


Figure 2.7. *In vivo* split GFP screen results.

Top row, from left to right: 1) positive control with MBP fused to GFP 11; 2) wild-type full-length CcmA fused to GFP 11; 3) negative control with a protein known to form inclusion bodies in *E. coli*. Bottom row, from left to right: 1) negative control with empty GFP11 vector; 2) full-length sfGFP on the GFP11 vector, which serves as an AnTet quality control.

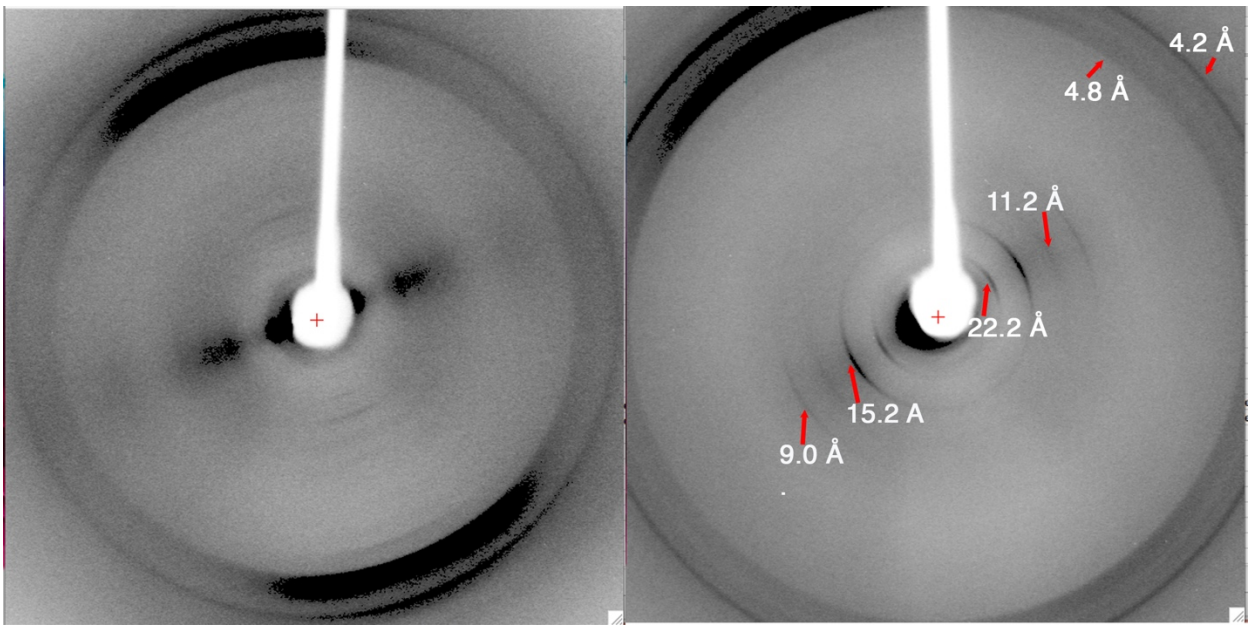


Figure 2.8. Fiber diffraction of full-length BacO.

Left: in water. Right: in 20mM Tris pH 8.0, 150mM NaCl.

Table 2.1. List of bactofilin constructs tested for crystallization purposes.

Construct name	Description	Solubility
pYL2	Helicobacter pylori J99 bactofilin CcmA-His6	No
pYL3	Helicobacter pylori J99 bactofilin CcmA-His6 17-136	No
pYL4	Helicobacter pylori J99 bactofilin CcmA-His6 1-120	No
pYL5	Helicobacter pylori J99 bactofilin CcmA-His6 17-120	No
pYL11	Helicobacter pylori J99 bactofilin CcmA-His6 17-120 I20A	No
pYL12	Helicobacter pylori J99 bactofilin CcmA-His6 17-120 D39S	No
pYL13	Helicobacter pylori J99 bactofilin CcmA-His6 17-120 V54A	No
pYL14	Helicobacter pylori J99 bactofilin CcmA-His6 17-120 F66A	No
pYL15	Helicobacter pylori J99 bactofilin CcmA-His6 17-120 E105S	No
pYL16	fixed arm SER MBP-CcmA 17-120	Yes
pYL17	fixed arm SER MBP-CcmA 17-120 F66A	Yes
pYL18	Helicobacter pylori J99 bactofilin CcmA-His6 17-120 F76A	No
pYL19	Helicobacter pylori J99 bactofilin CcmA-His6 17-120 F66A F76A	No
pYL20	fixed arm SER MBP-CcmA 17-120 H32A	Yes
pYL21	fixed arm SER MBP-CcmA 17-120 H48A	Yes
pYL22	fixed arm SER MBP-CcmA 17-120 Q57A	Yes
pYL23	fixed arm SER MBP-CcmA 17-136	Yes
pYL24	fixed arm SER MBP-BacO 1-126	Yes
pYL25	BacO 1-126	No
pYL26	fixed arm SER MBP-CcmA 17-120 V54K	Yes
pYL27	fixed arm SER MBP-CcmA 17-120 V54K Q57A	Yes
pYL28	fixed arm SER MBP-CcmA 17-120 H32A F66A	Yes
pYL29	fixed arm SER MBP-CcmA 17-120 I21K	Yes
pYL30	fixed arm SER MBP-CcmA 17-120 I119K	Yes
pYL31	fixed arm SER MBP-CcmA 17-120 I21K I119K	Yes
pYL32	fixed arm SER MBP-CcmA 17-120 H48A F66A	Yes
pYL33	fixed arm SER MBP-BacO-His61-126	Yes
pYL34	fixed arm SER MBP-BacO C96A 1-126	Yes
pYL35	fixed arm SER MBP-BacO T94A C96A 1-126	Yes
pYL37	fixed arm SER MBP-BacO T94K C96K1-126	Yes
pYL38	fixed arm SER MBP-BacO T94K C96K G61K G63K 1-126	Yes
pYL39	fixed arm SER MBP-BacO G61K G63K 1-126	Yes
pYL40	fixed arm SER MBP-BacO 21-126	Yes
pYL41	fixed arm SER MBP-BacO 21-126 T94A C96A	Yes
pYL42	fixed arm SER MBP-BacO 21-126 T94K C96K	Yes
pYL43	BacO T94K C96K 1-126	No
pYL48	fixed arm SER MBP-CcmA 1-120	Yes
pYL49	fixed arm SER MBP-CcmA FL Q57K G59K	Yes
pYL50	fixed arm SER MBP-CcmA FL P89K G91K	Yes
pYL51	fixed arm SER MBP-CcmA FL Q57K G59K P89K G91K	Yes

REFERENCES

1. Oliva MA, Cordell SC, Löwe J (2004) Structural insights into FtsZ protofilament formation. *Nat Struct Mol Biol* 11(12):1243–1250.

2. Errington J, Daniel RA, Scheffers D-J (2003) Cytokinesis in Bacteria. *Microbiol Mol Biol Rev* 67(1):52–65.

3. van den Ent F, Amos LA, Löwe J (2001) Prokaryotic origin of the actin cytoskeleton. *Nature* 413(6851):39–44.

4. Doi M, et al. (1988) Determinations of the DNA sequence of the mreB gene and of the gene products of the mre region that function in formation of the rod shape of Escherichia coli cells. *J Bacteriol* 170(10):4619–4624.

5. Ausmees N, Kuhn JR, Jacobs-Wagner C (2003) The Bacterial Cytoskeleton: An Intermediate Filament-Like Function in Cell Shape. *Cell* 115(6):705–713.

6. Kühn J, et al. (2010) Bactofilins, a ubiquitous class of cytoskeletal proteins mediating polar localization of a cell wall synthase in Caulobacter crescentus. *EMBO J* 29(2):327–339.

7. Bulyha I, et al. (2013) Two Small GTPases Act in Concert with the Bactofilin Cytoskeleton to Regulate Dynamic Bacterial Cell Polarity. *Developmental Cell* 25(2):119–131.

8. El Andari J, Altegoer F, Bange G, Graumann PL (2015) Bacillus subtilis Bactofilins Are Essential for Flagellar Hook- and Filament Assembly and Dynamically Localize into Structures of Less than 100 nm Diameter underneath the Cell Membrane. *PLoS One* 10(10). doi:10.1371/journal.pone.0141546.

9. Alm RA, et al. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. *Nature* 397(6715):176–180.

10. Sycuro LK, et al. (2010) Peptidoglycan Crosslinking Relaxation Promotes Helicobacter pylori's Helical Shape and Stomach Colonization. *Cell* 141(5):822–833.

11. Huang KC, Mukhopadhyay R, Wen B, Gitai Z, Wingreen NS (2008) Cell shape and cell-wall organization in Gram-negative bacteria. *PNAS*. doi:10.1073/pnas.0805309105.

12. Jensen LJ, et al. (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucl Acids Res* 37(suppl 1):D412–D416.

13. Vasa S, et al. (2014) β-Helical architecture of cytoskeletal bactofilin filaments revealed by solid-state NMR. *PNAS*:201418450.

14. Shi C, et al. (2015) Atomic-resolution structure of cytoskeletal bactofilin by solid-state NMR. *Sci Adv* 1(11). doi:10.1126/sciadv.1501087.

15. Punta M, et al. (2011) The Pfam protein families database. *Nucleic Acids Research* 40(D1):D290–D301.

16. Finn RD, et al. (2014) Pfam: the protein families database. *Nucl Acids Res* 42(D1):D222–D230.

17. Schuster-Böckler B, Schultz J, Rahmann S (2004) HMM Logos for visualization of protein families. *BMC Bioinformatics* 5(1):7.

18. DiMaio F, Leaver-Fay A, Bradley P, Baker D, André I (2011) Modeling Symmetric Macromolecular Structures in Rosetta3. *PLoS ONE* 6(6):e20450.

19. Zimmermann L, et al. (2018) A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology* 430(15):2237–2243.

20. Savojardo C, Fariselli P, Martelli PL, Casadio R (2013) BCov: a method for predicting β-sheet topology using sparse inverse covariance estimation and integer programming. *Bioinformatics*:btt555.

21. Jones DT, Buchan DWA, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2):184–190.

22. Marks DS, et al. (2011) Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS ONE* 6(12):e28766.

23. Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. *Nucl Acids Res* 32(suppl 2):W526–W531.

24. Koch MK, McHugh CA, Hoiczyk E (2011) BacM, an N-terminally processed bactofilin of Myxococcus xanthus, is crucial for proper cell shape. *Molecular Microbiology* 80(4):1031–1051.

25. Moon AF, Mueller GA, Zhong X, Pedersen LC (2010) A synergistic approach to protein crystallization: Combination of a fixed-arm carrier with surface entropy reduction. *Protein Science* 19(5):901–913.

26. Cabantous S, Waldo GS (2006) In vivo and in vitro protein solubility assays using split GFP. *Nat Meth* 3(10):845–854.

27. Pédelacq J-D, Cabantous S, Tran T, Terwilliger TC, Waldo GS (2006) Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotech* 24(1):79–88.

28. Cabantous S, Terwilliger TC, Waldo GS (2005) Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat Biotech* 23(1):102–107.

29. Finn RD, et al. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 44(D1):D279–D285.

Chapter 3. The Design of Symmetric Protein Nanomaterials Comes of Age in Theory and Practice[2]

Todd O. Yeates[1,2,3], Yuxi Liu[1], Joshua Laniado[3]

[1]UCLA Department of Chemistry and Biochemistry

[2]UCLA-DOE Institute for Genomics and Proteomics

[3]UCLA-Molecular Biology Institute

Key words: protein design, protein assembly, symmetry, protein cages. Bionanotechnology

[2] This chapter is the adapted version of a published review article of which I participated in writing (1). The content of this chapter should serve as a good introduction to the theory of symmetry-based protein engineering and some relevant current literatures. Chapters 4-6 all record my efforts in the protein engineering field, either led to published literature (Chapter 4 & 5) or stayed as preliminary results (Chapter 6).

ABSTRACT

In nature, protein molecules have evolved as building blocks for the assembly of diverse and complex structures, many of which exhibit a high degree of symmetry. This observation has motivated a number of recent engineering efforts in which the advantages of symmetry have been exploited to design novel self-assembling protein structures of great size. Materials ranging from cages to extended two and three-dimensional arrays have been demonstrated. Especially for extended arrays, a vast number of geometrically different design types are possible. A table of geometric rules is provided for designing a universe of novel materials by combining two component symmetries.

INTRODUCTION

Building blocks that have self-complimentary interfaces can self-assemble into elaborate structures. Nature serves as a rich source of inspiring specimens. At the macromolecular scale, viral capsids are quintessential examples, but other equally extraordinary macromolecular assemblies abound in nature (reviewed in [1-3]). The beauty and functional utility of these assemblies have long-motivated engineering efforts to create comparable structures in the laboratory. Beginning in the 1980's Ned Seeman pioneered ideas for using DNA molecules as building blocks for nanostructures [4]. Over the years, those ideas and various strategic variations led to the creation of elaborate supramolecular architectures and design patterns based on nucleic acids (reviewed in [5]). In nature, protein molecules have been the choice for the evolution of large assemblies with diverse form and function. But the engineering path to following Nature's lead has been challenged by the complexity of the rules that govern protein folding and assembly. To overcome those challenges, special strategies are needed.

In developing a strategic approach for building with protein molecules, Nature provides a major clue. Symmetry prevails in naturally evolved protein assembles. This is an empirical fact evident in the vast database of known macromolecular structures [6,7], but the prevalence of high symmetry in large protein assemblies was anticipated at least as far back as 1956 when Crick and Watson emphasized that viral capsids were likely able to evolve more easily in symmetric forms because symmetric assemblies require the fewest number of distinct interfacial contacts between individual subunits [8]. That key observation applies as well to designed structures, and indeed the early history of designing protein assemblies is rich with cases of relatively simple symmetric structures such as dimers and helical filaments [9-12] . The push in recent years to create very large protein assemblies has been guided even more strongly by principles of symmetry.

SYMMETRY-BASED DESIGN STRATEGIES

The symmetry of an object is fully described by the set of spatial operations (e.g. rotations) that leave the entire object unchanged except for an undetectable exchange of identical subunits. Because the symmetry of an object obeys the properties of a mathematical group, each specific type of symmetry is often referred to as a symmetry group. The symmetry group of a structure can be used to understand how many structurally distinct contact types are required to hold all the subunits together in one connected object. Certain simple types of architecture can be created from a building block that touches itself in just one way; i.e. using a single contact type. The possible outcomes are limited to structures like cyclic rings of subunits, or single filaments (Fig. 3.1A). More complex architectures require building blocks with more than one distinct interface.

A relatively simple group theory analysis explains the minimum number of distinct contact types required to achieve a given target symmetry. This was articulated first in the context of three-dimensional crystals [13] and then in the context of designed protein assemblies by Padilla

et al. [14]. For example, if all the elements of a symmetry group can be generated by repeated application of a single element of the group (i.e. a rotational operation), then one contact type is sufficient. The cyclic or single filament architectures noted above are examples of this type. If two elements from the symmetry group must be chosen in order to obtain the full symmetry group by repeated operations, then two contact types are required, and so on. Surprisingly, it turns out that a great many types of symmetry – including finite cages and many extended two and three-dimensional arrays – can be generated using just two properly chosen symmetry elements in combination (Fig. 3.1A). This key point frames the problem of designing novel protein assemblies by prescribing the number of distinct contact types that must be built into a protein building block in order for it to assemble into the desired architecture. More than the minimum number of contact types can be present in a final assembly, but not fewer. The specific geometry of the interfaces is of course crucial for obtaining the desired result. And molecular strategies are required for creating these oriented interfaces.

Various strategies have been developed for building the multiple distinct interfacial contact into a protein molecule in order to generate elaborate supramolecular structures (Fig. 3.1B). Padilla et al. [14] laid out a first strategy at a time when prospects for designing de novo interfaces into protein molecules were still remote. By necessity, naturally evolved interfaces were exploited by using simple natural protein oligomers (e.g. dimers and trimers) as a starting point. To create a single molecular building block containing two distinct interface types, a method was developed for genetically fusing two naturally oligomeric protein domains. In order to control relative geometry, only oligomeric domains having terminal alpha helices were considered, so that directly fusing two such proteins might create a geometrically predictable outcome if a continuous alpha helix was preserved between the two domains. The diversity of architectures possible by the

general approach was described, and a first demonstration was provided – a 12-subunit assembly in the form of a tetrahedral cage was designed from a dimer plus trimer fusion. This protein assembled into geometric structures consistent with the design, among a range of other polymorphic forms. Several years later, Lai et al [15,16] showed that introducing two or three amino acid mutations into the original designed protein was sufficient to produce 12-subunit assemblies in high yield, which could be crystallized and validated in atomic detail. Not surprisingly, some flexibility of the helix linker gave rise to assemblies that were flexed or deformed from perfect symmetry, but which otherwise conformed to the intended tetrahedral design [16]. A different, 24-subunit cubic cage in good agreement with its design was subsequently demonstrated using the same helix fusion strategy [17].

Major leaps forward in design strategy were made by King et al. [18] working with globular protein domains and Lanci et al. [19] working with coiled-coil polypeptides; they foresaw that computational methods for introducing novel interfaces into protein surfaces by amino acid sequence design had matured to the point where they might allow large symmetric assemblies to be created. Grueninger et al., took an earlier step in this direction by designing double-ring assemblies from naturally cyclic structures [20]. Following the symmetry ideas discussed above, starting with simple oligomeric proteins (e.g. dimers and trimers) means that only one additional interface needs to be designed into the protein in order to create complex architectures (Fig. 3.1B). In King et al., the procedure was enabled by a special algorithm written for the Rosetta-Design program to preserve overall symmetry while sampling the rigid body degrees of freedom available to the component oligomers [21]. From 41 initial designs, two cubic cages were produced in high yield and could be validated by crystallography. One was a tetrahedral cage built from four trimers situated at the vertices of a tetrahedron and contacting each other primarily via a designed interface

with two-fold symmetry. The other was a cubic/octahedral cage built from eight trimers at the corners of a cube, again with the trimers interacting primarily via a designed interface with two-fold symmetry. The polypeptide design work by Lanci et al. [19] relied on a trimeric coil-coil motif as the starting point; the introduction of lateral and vertical contacts gave a three dimensional crystalline material also validated by crystallography.

Within the bounds of symmetry-based methods of design, several strategic variations are possible beyond those noted above. Some of the possibilities are briefly described here (Fig. 3.1B). Sinclair et al. [22] introduced a variation on the oligomer fusion method that relaxed the requirement for a continuous alpha helical linker; it applies to certain extended two and three-dimensional assemblies where two oligomers can be fused in a way that preserves a rotational symmetry element they both share. A few cases of well-ordered layers were demonstrated with that approach. King et al. [23] introduced a two-component variation on the de novo interface design strategy. Two different natural oligomers comprise the starting materials, and computational sequence design is used to introduce a heterotypic interface between the two subunit types. By relying on two separate oligomeric components, the idea shares similarity with the helix fusion method. But the helix fusion is rendered unnecessary by the designed interface between the two oligomers. Furthermore, the non-covalent nature of the association between the two oligomeric components enables production and purification of separate components, with full assembly occurring upon mixing. In the first application of the two-component strategy, a series of approximately 60 designed cages was tested experimentally, and five were validated in detail by crystallography [23,24].

In other design approaches, metals or bivalent ligands have been used as a way to introduce a new interface or self-associating interaction between oligomeric components ([25-27], reviewed

in [28]). The addition of metals promotes assembly when metal binding half-sites (e.g. two suitably disposed histidine residues) are designed into a protein surface. Without further computational design of protein-protein interactions beyond the metal site, the metal site approach tends to give rise to assemblies whose outcomes are hard to predict in detail owing to alternate possible orientations of the metal ligands. But various reports [29-35] have shown that the approach can be used to create interesting and diverse materials ranging from small oligomers to helical structures to layers and three-dimensional arrays. A particularly intriguing variation was demonstrated recently by the introduction of a Zn-binding site at the three-fold symmetry axis of the natural, cubically symmetric ferritin cage. Adding a bivalent, metal-binding organic ligand joins the cubic protein cages into a three-dimensional body centered crystal lattice in a predictable fashion [35]. In a report by Sakai et al., the authors used bivalent organic ligands to create an additional dimeric subunit association for linking D2 tetramers together, in an unspecified orientation in this case. Changing the spacer length between the binding moieties on the organic ligand resulted in two different types of three-dimensional crystals conforming to space groups allowed by the combination of D2+C2 symmetry components [36].

Other strategic variations for combining symmetry elements are possible but have not been deeply explored yet (Fig. 3.1B). DNA (or RNA) provides a facile route for introducing a two-fold symmetry element using a palindromic nucleotide sequence. This could be exploited in combination with DNA binding proteins and either interface design or oligomeric fusion to create symmetric hybrid materials composed of proteins and nucleic acids. A related strategy for assembling protein and DNA components together has been taken by Mou et al. in creating linear or helical filaments [37]. With proteins, variations on the alpha helix fusion approach offer prospects of improved rigidity. A coiled-coil linker could be employed as the motif joining two

separate oligomeric units, with a single helix of a hetero coiled-coil motif extending from each of the oligomeric components. These are avenues for ongoing study.

DESIGNED PROTEIN CAGES

Designed assemblies of defined, finite size can take the form of shells or cages or more compact clusters following one of the three possible cubic symmetries in three dimensions. These are based on the symmetries of the five Platonic solids; the cube and the octahedron are duals of each other while the icosahedron and the dodecahedron are duals, so together with the tetrahedron (which is its own dual), there are three possible symmetries: T (order 12), O (order 24), and I (order 60). The symmetry rules for generating these have been articulated [14,38] (Fig. 3.2A). For each symmetry type, nearly every combination of two *intersecting* symmetry axes is a possible choice for generating the final structure, but there are some exceptions (Fig. 2 legend).

Using designed protein molecules, all three of the cubic symmetry types have been successfully produced and validated in atomic detail (Fig. 3.2B). Some of the target symmetries have been obtained using distinctly different symmetry combinations. Symmetry T has been obtained by combinations of 2-fold plus 3-fold symmetry components and by 3-fold plus 3-fold components [23]. In recent work, icosahedral symmetry has been obtained by all three possible combinations (2 plus 3, 2 plus 5, and 3 plus 5) (unpublished data: Baker, Bale, Liu, Yeates, et al). As a proof of concept, these novel designs cover most of the possible space for symmetric cages. The designed architectures range in number of subunits from 12 to 120, with diameters from 11 to 40 nm and masses from 276 kDa to 2.8 MDa. The numerous possible applications for designed protein cages have been reviewed elsewhere [39-42] and are not elaborated here other than to summarize that practical uses are likely to include both interior and exterior capabilities: (i)

encapsulation, delivery and release of molecular cargo [43-48], and (ii) polyvalent display of motifs for signaling or antigenic effects, as in synthetic vaccines [49,50].

DESIGN RULES FOR BUILDING EXTENDED MATERIALS IN TWO AND THREE DIMENSIONS

In contrast to the finite cage designs that arise from two intersecting symmetry axes as discussed above, when two component symmetries are combined in an arrangement where any of their axes are not intersecting, the result cannot be finite and must instead be an extended or unbounded material. Filaments are one possible outcome, arising from two non-intersecting 2-fold axes of symmetry [14]. But more complex outcomes are obtained by combining higher symmetries. There, unbounded materials that extend as either 2-dimensional layers or 3-dimensional arrays (i.e. crystals) are possible. The geometric rules for a few possibilities of this type were laid out earlier based on the combination of 2-fold and 3-fold axes of symmetry [14]. Beyond those, a vast range of possibilities arise from combinations of higher component symmetries. A few designs within that scope have been demonstrated in recent work [51,52]. A complete set of geometric rules for generating two and three-dimensional materials has not been articulated previously. In order to promote further studies, we provide a list of the allowable symmetry combinations here (Table 3.1).

Two-dimensional layers can be of two different classes depending on their sidedness, or lack thereof. When two symmetry axes that are both perpendicular to the layer are combined, and at least one of those axes is of higher order than a 2-fold, the result is a layer with distinguishable sides (i.e. a distinct top and bottom). In that sense such layers are oriented. The allowable symmetry combinations for oriented layers are 2+3, 2+4, 2+6, 3+3, 3+6, and 4+4 (Table 1). Layered structures of the other class, where the top and bottom of the layer are indistinguishable,

arise whenever one or both of the two symmetries being combined carries a 2-fold axis of symmetry in the plane of the layer. A total of 33 layer designs are possible (Table 1).

Much of the design space for designed layers is unexplored, but a few recent successes have been reported. Small two-dimensional assemblies with limited range order were described by Ringler et al. by doubly biotinylating the subunits of aldolase (a C4 tetramer) and then assembling those tetramers using streptavidin (a D2 tetramer with a biotin binding site in each subunit) [26]. If long range order had been achieved in this case, the result would have corresponded to layer symmetry p422. Longer range order in designed layers was demonstrated by Sinclair et al. [22]. There the best case was obtained by combining the D2 tetrameric streptavidin with a D4 octameric protein that had been biotinylated. The relative orientations of the components in that case were not specifically designed and could have produced other results, but a two-dimensional layer with p422 symmetry was obtained (Fig. 3.3). More recently, computational interface design was used to create specifically defined layers of a few different types [51]. A 2-fold interface was designed between natural C6 hexameric units to give layer symmetry p6, between natural C4 tetrameric units to give layer symmetry $p42_12$, and between natural C3 trimeric units to give layer symmetry p321 (Fig. 3.3). In another variation [52], two copies of a C6 hexameric protein were fused in tandem in such a way that the short linker between them, along with a designed 2-fold interface, led to a pseudo-p6 layer.

For designing extended three-dimensional crystalline materials, the possibilities are even more expansive, and the design rules are not so obvious. Defining the outcome from a combination of two separate symmetries relies on two considerations. The overall rotational symmetry of the resulting material is given by the (group) product of the rotational symmetries of the two components; this is a relatively straightforward issue. A somewhat more complex problem is

discerning the correct outcome among the possible space groups having a particular rotational symmetry. The correct choice can generally be ascertained from the standard tables of crystallographic space groups by identifying the one having Wyckoff positions with symmetries corresponding to those of the two components being combined. In total, more than 70 distinct types of 3D materials are possible within the scheme of combining two types of rotation (point group) symmetries. The majority of them have underlying cubic (T or O) symmetries and are therefore isotropic, while the others are dihedral.

On the experimental side, the space of designed protein crystals is mostly unexplored, but there are a few early examples. As discussed above, Lanci et al. [19] designed a coiled-coil peptide to form P6 space group symmetry, and Sontz et al. [35] combined a ferritin, into which a metal side had been engineered, with a bivalent bridging compound to form a crystal whose space group symmetry was pseudo-I432. Sinclair, et al. used their fusion method to form three-dimensional solid materials, but without sufficient order to confirm the intended crystalline packing by X-ray diffraction [22]. Beyond a handful of examples, the area of designed protein crystals remains open. The possible applications for such materials are diverse: creating materials with a very high density of reactive/catalytic groups or recognition motifs, and conferring specific physical properties on target proteins, including spacing, dimensionality, porosity, and solid-phase separability from solution components.

VARIATIONS, CHALLENGES AND FUTURE DIRECTIONS

The ideas and rules formulated here are somewhat narrowly constructed. They represent the simplest design routes (i.e. the minimum requirements) for construction by symmetric assembly. Broader outcomes are possible if interfaces between components are designed in different ways. For example, King's single-component designed interface method [18] allows for

symmetries not accounted for in Table 3.1 if a new contact between like oligomers creates a screw axis of symmetry instead of a pure rotation. Ultimately, in the absence of limits on designing protein-protein interfaces, any symmetric architecture could be designed, including those that require larger numbers of distinct interaction types. The three-dimensional crystal in space group P6 designed by Lanci et al [19] is a case in point using relatively simple building blocks. In addition, other approaches for designing large protein and peptide-based structures have been developed that rely less strictly, or in different ways, on symmetry. Fletcher et al. [53] combined a homotrimeric coiled-coil and a heterodimeric coiled-coil that interact with each other. This resulted in unilamellar spheres approximately 100 nanometers in diameter, with overall structures that were not exactly symmetric though assembly was driven by local symmetry. Doll et al. [54] also combined coiled-coil sequences with different symmetric properties (5-fold and 3-fold) to produce roughly spherical clusters having sizes consistent with icosahedral assembly. In a distinctly different line of attack, Gradišar et al. demonstrated the construction of a tetrahedral architecture on the basis of asymmetric interactions between coiled coil motifs in a long, designed protein molecule whose folding pattern traverses the edges of the entire polyhedron twice [55].

The different design strategies discussed here present their own advantages and challenges. The initial design strategy of helical fusions between oligomers presents a relatively low barrier in the sense of not needing to create de novo interfaces, but flexibility creates an obvious challenge. This sometimes leads to alternate assembly outcomes [17,56]. Strategies involving more rigid linkers could improve the reliability of this method, and a recent report demonstrates some success in rigidifying a continuous alpha helical linker between two protein components by a specific chemical cross-link between cysteines at positions i and i+11 [57].

The main challenges with methods based on de novo interface design relate to misfolding or unintended assembly – often insoluble aggregation – most likely caused by introduction of new regions of hydrophobicity in a protein surface. The success rates for geometrically specific interface design in the context of symmetric assemblies is currently in the range of about 10%. One case has been reported where a failed design could be rescued by increasing the charge on the protein molecule [24]. This suggests that optimizing certain design parameters and selection criteria might substantially increase the success rates. New strategies for high throughput selection or screening of designs for correct asembly could also be impactful.

As design strategies continue to improve, and with construction rules in hand for building wide-ranging types of symmetric architectures, the coming years should bring a rich diversity of new protein based materials with useful applications.

Acknowledgements

This work was supported by NSF grant CHE-1332907. The authors thank Dan McNamara, Yen-Ting Lai, Kevin Cannon and other members of the Yeates lab for their ideas, and members of the David Baker laboratory, including Jacob Bale and Neil King, for access to their designs ahead of publication. We thank Frank DiMaio and Will Sheffler for discussions on symmetry rules and Dek Woolfson for discussions related to coiled-coil polypeptides.

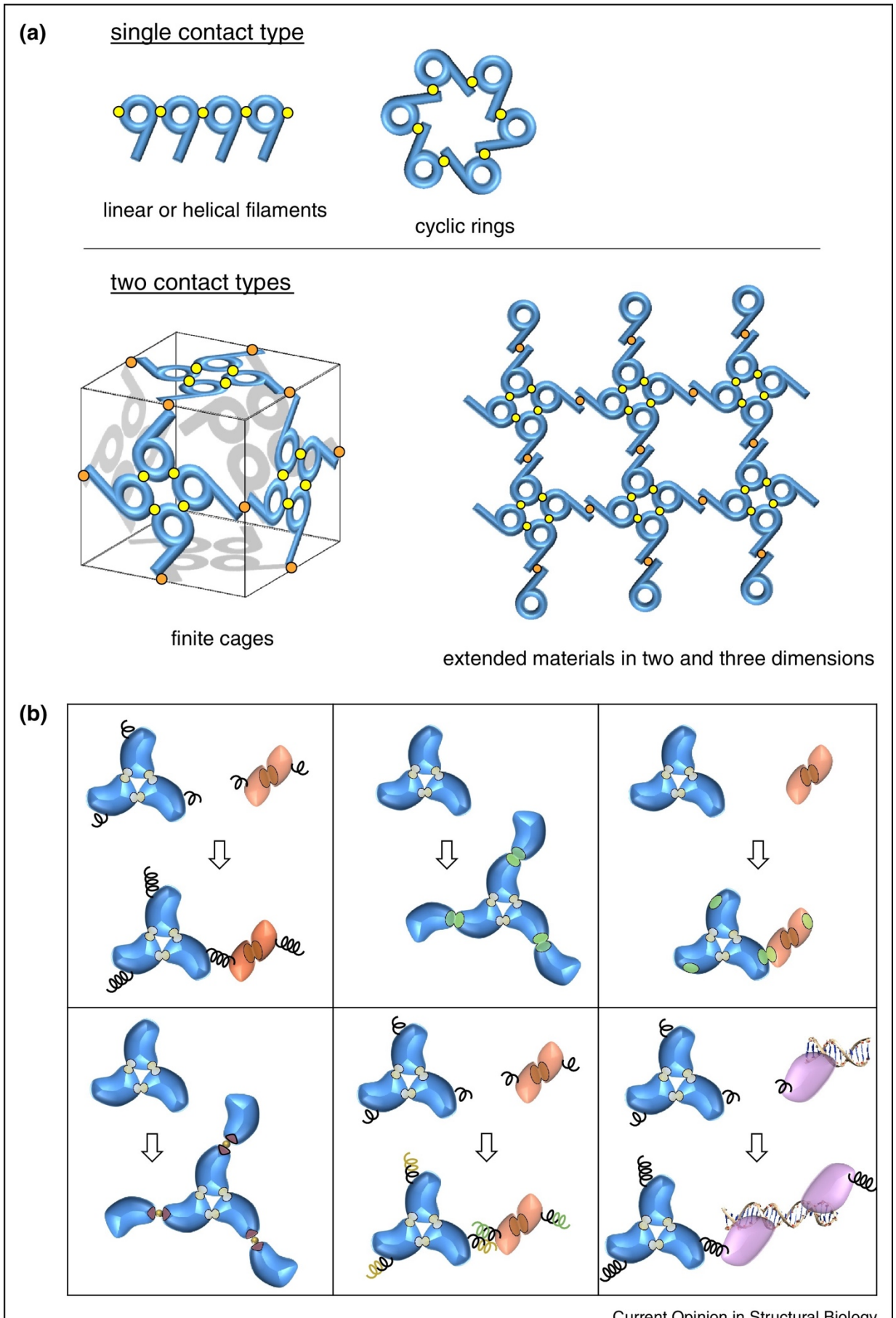
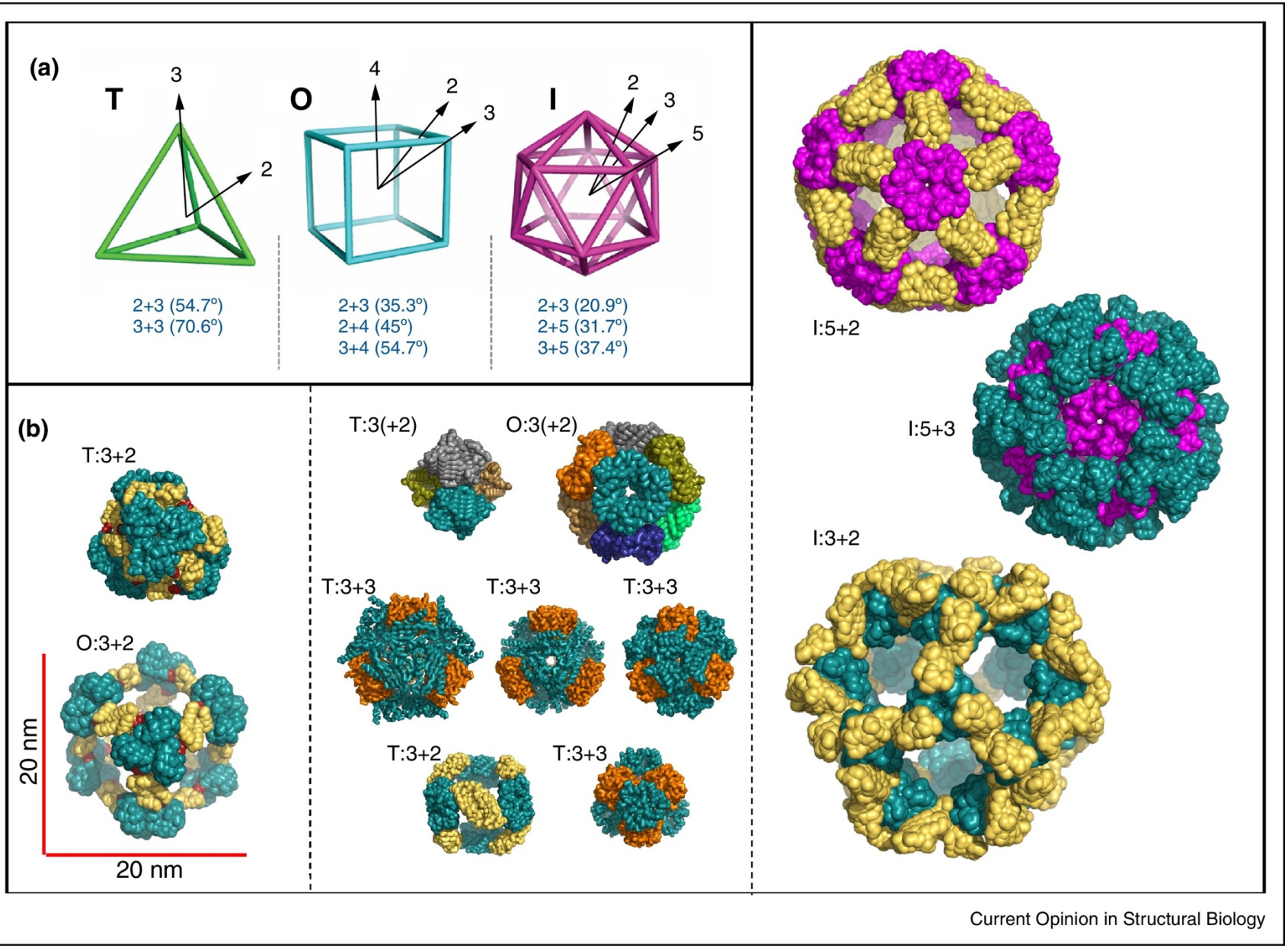

(a)
single contact type
linear or helical filaments
cyclic rings
two contact types
finite cages
extended materials in two and three dimensions
(b)
Current Opinion in Structural Biology

Figure 3.1. Assembly consequences and strategies for introducing multiple contact types into protein building blocks.

A) Illustration of varied symmetric architectural forms and the number of distinct contact types required for connectivity between molecular building blocks. Two contact types are sufficient to create diverse assemblies. B) Different molecular strategies for creating a building block having two distinct contact types in a defined orientation. Left to right (top): alpha helical fusion; 1-component interface design; 2-component interface design. Left to right (bottom): metal or ligand bridging; coiled-coil helical fusions; designed symmetrization of DNA binding proteins.



Current Opinion in Structural Biology

Figure 3.2**.** Design and validation of self-assembling protein cages with high symmetry.

A) The three types of cubic symmetry (T, O, and I) are illustrated on the framework of the Platonic solids. Angles between pairs of rotational symmetry axes that can be combined to create a self-assembling building block with the target symmetry are listed. B) Engineered protein cages obeying all three possible cubic symmetries have been produced; a subset of structures that have been validated by X-ray crystallography are shown to scale. Left panel: a tetrahedral cage (PDB 4IQ4) [16] and an octahedral cage (PDB 4QCC) [17] designed using the alpha helical fusion strategy [14] . Natural trimers are in blue; natural dimers are in yellow; the alpha-helical linkers are in red. Right panel, tetrahedral cages and an octahedral cage engineered by de novo interface design with either one or two components. Top row are one-component designs (PDB 4EGG & 3VCD, left to right) [16], with each trimer shown in a different color for clarity. The middle and bottom rows (PDB 4NWR, 4NWP, 4NWO, 4NWN, & 4ZK7, left to right and top to bottom) are two-component designs [23,24]. Natural dimers are in yellow. Natural trimers are in blue and orange to differentiate the different trimers in the same design. Letters on the top left corner of the structure indicates the symmetry type (T: tetrahedral, O: octahedral, I: icosahedral). The numbers in the annotation indicate the component symmetry types. Where present, parenthetical values indicate the symmetry of the main designed interface. [Images of the icosahedral structures to be added upon publication (Bale, Baker, Yeates, et al., unpublished data)].

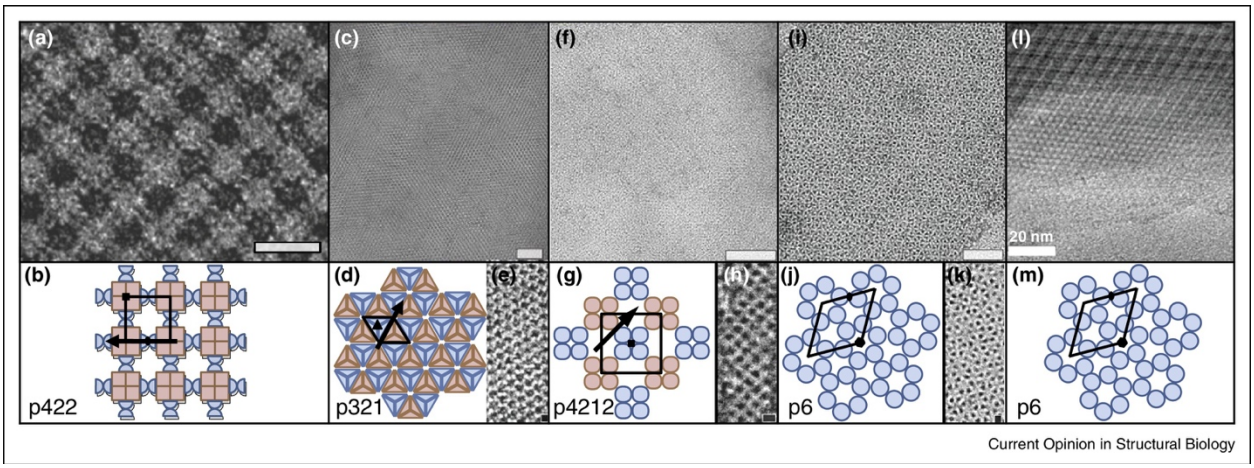


Figure 3.3**.** Electron micrographs of protein layers designed to assemble with high symmetry and showing long-range order.

Symmetry diagrams are shown under each micrograph, accompanied in some cases by enlarged images. Each symmetry diagram shows the repeating unit cell within which one instance of each of the component symmetry elements is indicated using standard symbols: black arrows, 2-fold symmetry axes in the plane of the layer; black ovals, black triangles, black squares, and black hexagons indicate 2-, 3-, 4-, and 6- fold axes perpendicular to the plane of the layer, respectively. A) a p422 layer formed by combining D4 and D2 symmetry components; C) a p321 layer formed by combining C3 trimers with a 2-fold de novo interface; I) a $p42_12$ layer formed by combining C4 tetramers with a 2-fold de novo interface; I) a p6 layer formed by combining C6 hexamers and a 2-fold de novo interface; L) a pseudo p6 layer formed by combining C6 hexamers using a covalent fusion and a pseudo 2-fold de novo interface. E, H, K: enlarged images of C, F, I, respectively. Scale bars: A – 20 nm; C – 50 nm; F – 50 nm; I – 50 nm; L – 20 nm. Images reproduced with permission from: panel A – Sinclair, J. C., et al. (2011) [22]; panels C, E, F, H, I, and K – Gonen, S., et al. (2015)[51]; panel L –Matthaei, J. F., et al. (2015) [52].

Table 3.1. Multiplication table for designing self-assembling protein materials from combinations of two simpler symmetric components or interfaces[#].

x	C2	C3	C4	C6	D2	D3	D4	D6	T	O
C2	‡	D3, T, O, I p6, p321 $I2_13$, $P4_132$	D4, O p4, $p42_12$ I432	D6 p6, p622	c222, p422, p622 $I4_122$, $P6_222$, I432, $I4_132$	p321, p622 R32, $P6_322$, $F4_132$, $I4_132$, I432, $P4_132$	p422 I422, P432, I432	p622 P622	P23, F23, $F4_132$	P432, F432, I432
C3		T p3 $P2_13$	O F432	p6	p622 P23, F432, $I4_132$	p321, p312 $P4_132$	P432	p622	F23	F432
C4			p4 P432		p422, $p42_12$ I432, F432	I432	p422 P432		F432	P432, F432
C6					p622	p622				
D2					p222, p622 F222, $P4_222$, $P6_222$, $P4_232$, $I4_132$	p622 P622, $P4_232$, $I4_132$	p422 P422, I422, I432	p622 P622	P23, F432, $P4_232$	F432, I432
D3						p321 P312, $P6_322$, $P4_232$, $F4_132$, $P4_132$	I432	p622 P622	$F4_132$	I432
D4							p422 P422, P432			P432
D6										
T									F23	F432
O										P432, F432

Finite assemblies (point group symmetries) are indicated in the blue font. 2-D layers are indicated in red, 3-D crystalline arrays in purple. In many cases, two component symmetries can be combined in distinct geometries that give rise to different symmetry types.

Gray boxes indicate symmetry combinations that are disallowed mathematically. A few symmetry combinations are not formally disallowed but are not amenable to design using compact building blocks.

Whenever a chiral space group appears (e.g. $P4_132$), its enantiomer (e.g. $P4_321$) is also possible but is not listed here for brevity.

[#] Additional possibilities exist but are not listed here for arrangements where more than two component symmetries are combined, or where one of the component symmetries is a screw axis of rotation.

[‡] Non-intersecting 2-fold axes give rise to linear or helical filaments.

REFERENCES

1. Goodsell DS, Olson AJ: **Structural symmetry and protein function**. *Annual Review of Biophysics and Biomolecular Structure* 2000, **29**:105-153.

2. Marsh JA, Teichmann SA: **Structure, Dynamics, Assembly, and Evolution of Protein Complexes**. *Annual Review of Biochemistry* 2015, **84**:551-575.

3. Yeates TO, Thompson MC, Bobik TA: **The protein shells of bacterial microcompartment organelles**. *Current Opinion in Structural Biology* 2011, **21**:223-231.

4. Seeman NC: **Nucleic acid junctions and lattices**. *Journal of Theoretical Biology* 1982, **99**:237-247.

5. Jones MR, Seeman NC, Mirkin CA: **Nanomaterials. Programmable materials and the nature of the DNA bond**. *Science (New York, N.Y.)* 2015, **347**:1260901.

6. Ispolatov I, Yuryev A, Mazo I, Maslov S: **Binding properties and evolution of homodimers in protein–protein interaction networks**. *Nucleic Acids Research* 2005, **33**:3629-3635.

7. Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA: **Evolution of protein complexes by duplication of homomeric interactions**. *Genome Biology* 2007, **8**:R51.

8. Crick FH, Watson JD: **Structure of small viruses**. *Nature* 1956, **177**:473-475.

9. Pandya MJ, Spooner GM, Sunde M, Thorpe JR, Rodger A, Woolfson DN: **Sticky-End Assembly of a Designed Peptide Fiber Provides Insight into Protein Fibrillogenesis**. *Biochemistry* 2000, **39**:8728-8734.

10. Ogihara NL, Ghirlanda G, Bryson JW, Gingery M, DeGrado WF, Eisenberg D: **Design of three-dimensional domain-swapped dimers and fibrous oligomers**. *Proceedings of the National Academy of Sciences* 2001, **98**:1404-1409.

11. Robertson DE, Farid RS, Moser CC, Urbauer JL, Mulholland SE, Pidikiti R, Lear JD, Wand AJ, DeGrado WF, Dutton PL: **Design and synthesis of multi-haem proteins**. *Nature* 1994, **368**:425-432.

12. Kuhlman B, O'Neill JW, Kim DE, Zhang KY, Baker D: **Conversion of monomeric protein L to an obligate dimer by computational protein design**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**:10687-10691.

13. Wukovitz SW, Yeates TO: **Why protein crystals favour some space-groups over others**. *Nature Structural Biology* 1995, **2**:1062-1067.

14. Padilla JE, Colovos C, Yeates TO: **Nanohedra: Using symmetry to design self assembling protein cages, layers, crystals, and filaments**. *Proceedings of the National Academy of Sciences* 2001, **98**:2217-2221.

15. Lai Y-T, Cascio D, Yeates TO: **Structure of a 16-nm Cage Designed by Using Protein Oligomers**. *Science* 2012, **336**:1129-1129.

16. Lai Y-T, Tsai K-L, Sawaya MR, Asturias FJ, Yeates TO: **Structure and Flexibility of Nanoscale Protein Cages Designed by Symmetric Self-Assembly**. *Journal of the American Chemical Society* 2013, **135**:7738-7743.

17. Lai Y-T, Reading E, Hura GL, Tsai K-L, Laganowsky A, Asturias FJ, Tainer JA, Robinson CV, Yeates TO: **Structure of a designed protein cage that self-assembles into a highly porous cube**. *Nature Chemistry* 2014, **6**:1065-1071.

• Using the alpha helical fusion strategy, the authors reported the design and crystallographic validation of a highly porous cubic assembly with a large central cavity, demonstrating the feasibility of controlling geometry over long distance while providing porosity. Helix flexibility admitted alternative assembly forms.

18. King NP, Sheffler W, Sawaya MR, Vollmar BS, Sumida JP, André I, Gonen T, Yeates TO, Baker D: **Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy**. *Science* 2012, **336**:1171-1174.

19. Lanci CJ, MacDermaid CM, Kang S-g, Acharya R, North B, Yang X, Qiu XJ, DeGrado WF, Saven JG: **Computational design of a protein crystal**. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**:7304-7309.

20. Grueninger D, Treiber N, Ziegler MOP, Koetter JWA, Schulze M-S, Schulz GE: **Designed protein-protein association**. *Science (New York, N.Y.)* 2008, **319**:206-209.

21. DiMaio F, Leaver-Fay A, Bradley P, Baker D, André I: **Modeling Symmetric Macromolecular Structures in Rosetta3**. *PLoS ONE* 2011, **6**:e20450.

22. Sinclair JC, Davies KM, Vénien-Bryan C, Noble MEM: **Generation of protein lattices by fusing proteins with matching rotational symmetry**. *Nature Nanotechnology* 2011, **6**:558-562.

23. King NP, Bale JB, Sheffler W, McNamara DE, Gonen S, Gonen T, Yeates TO, Baker D: **Accurate design of co-assembling multi-component protein nanomaterials**. *Nature* 2014, **510**:103-108.

•• Using improved algorithms in the Rosetta software suite specifically adapted for the design of symmetric protein assemblies, the authors illustrated the design and validation of a series of novel protein cages wherein each cage assembles from two distinct oligomeric components. A de novo interface between the two types of oligomers drives the symmetric assembly.

24. Bale JB, Park RU, Liu Y, Gonen S, Gonen T, Cascio D, King NP, Yeates TO, Baker D: **Structure of a designed tetrahedral protein assembly variant engineered to have improved soluble expression**. *Protein Science* 2015, **24**:1695-1701.

25. Dotan n, Arad n, Frolow n, Freeman n: **Self-Assembly of a Tetrahedral Lectin into Predesigned Diamondlike Protein Crystals**. *Angewandte Chemie (International Ed. in English)* 1999, **38**:2363-2366.

26. Ringler P, Schulz GE: **Self-Assembly of Proteins into Designed Networks**. *Science* 2003, **302**:106-109.

27. Salgado EN, Ambroggio XI, Brodin JD, Lewis RA, Kuhlman B, Tezcan FA: **Metal templated design of protein interfaces**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**:1827-1832.

28. Salgado EN, Radford RJ, Tezcan FA: **Metal-Directed Protein Self-Assembly**. *Accounts of Chemical Research* 2010, **43**:661-672.

29. Salgado EN, Lewis RA, Mossin S, Rheingold AL, Tezcan FA: **Control of Protein Oligomerization Symmetry by Metal Coordination – C2 and C3 Symmetrical Assemblies through Cu(II) and Ni(II) Coordination**. *Inorganic chemistry* 2009, **48**:2726-2728.

30. Brodin JD, Ambroggio XI, Tang C, Parent KN, Baker TS, Tezcan FA: **Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays**. *Nature Chemistry* 2012, **4**:375-382.

31. Salgado EN, Lewis RA, Faraone-Mennella J, Tezcan FA: **Metal-Mediated Self-Assembly of Protein Superstructures: Influence of Secondary Interactions on Protein**

Oligomerization and Aggregation. *Journal of the American Chemical Society* 2008, **130**:6082-6084.

32. Salgado EN, Faraone-Mennella J, Tezcan FA: **Controlling Protein−Protein Interactions through Metal Coordination: Assembly of a 16-Helix Bundle Protein**. *Journal of the American Chemical Society* 2007, **129**:13374-13375.

33. Laganowsky A, Zhao M, Soriaga AB, Sawaya MR, Cascio D, Yeates TO: **An approach to crystallizing proteins by metal-mediated synthetic symmetrization**. *Protein Science* 2011, **20**:1876-1890.

34. Leibly DJ, Arbing MA, Pashkov I, DeVore N, Waldo GS, Terwilliger TC, Yeates TO: **A Suite of Engineered GFP Molecules for Oligomeric Scaffolding**. *Structure* 2015, **23**:1754-1768.

35. Sontz PA, Bailey JB, Ahn S, Tezcan FA: **A Metal Organic Framework with Spherical Protein Nodes: Rational Chemical Design of 3D Protein Crystals**. *Journal of the American Chemical Society* 2015, **137**:11598-11601.

•• The authors constructed a three-dimensional protein crystal with a prescribed lattice by introducing metal–organic linker interactions between adjacent ferritin cages, which are naturally cubic/octahedral. This is the first report of a designed metal-organic mediated 3-D protein crystal. Its designed structure was shown to be accurate by x-ray diffraction at atomic resolution.

36. Sakai F, Yang G, Weiss MS, Liu Y, Chen G, Jiang M: **Protein crystalline frameworks with controllable interpenetration directed by dual supramolecular interactions**. *Nature Communications* 2014, **5**:4634.

37. Mou Y, Yu J-Y, Wannier TM, Guo C-L, Mayo SL: **Computational design of co-assembling protein-DNA nanowires**. *Nature* 2015, **525**:230-233.

38. Lai Y-T, King NP, Yeates TO: **Principles for designing ordered protein assemblies**. *Trends in Cell Biology* 2012, **22**:653-661.

39. Yeates TO, Padilla JE: **Designing supramolecular protein assemblies**. *Current Opinion in Structural Biology* 2002, **12**:464-470.

40. Doll TAPF, Raman S, Dey R, Burkhard P: **Nanoscale assemblies and their biomedical applications**. *Journal of the Royal Society Interface* 2013, **10**.

41. López-Sagaseta J, Malito E, Rappuoli R, Bottomley MJ: **Self-assembling protein nanoparticles in the design of vaccines**. *Computational and Structural Biotechnology Journal* 2015, **14**:58-68.

42. Howorka S: **Rationally engineering natural protein assemblies in nanobiotechnology**. *Current Opinion in Biotechnology* 2011, **22**:485-491.

43. Wörsdörfer B, Woycechowsky KJ, Hilvert D: **Directed evolution of a protein container**. *Science (New York, N.Y.)* 2011, **331**:589-592.

44. Patterson DP, Schwarz B, El-Boubbou K, Oost Jvd, Prevelige PE, Douglas T: **Virus-like particle nanoreactors: programmed encapsulation of the thermostable CelB glycosidase inside the P22 capsid**. *Soft Matter* 2012, **8**:10158-10166.

45. Champion CI, Kickhoefer VA, Liu G, Moniz RJ, Freed AS, Bergmann LL, Vaccari D, Raval-Fernandes S, Chan AM, Rome LH, et al.: **A Vault Nanoparticle Vaccine Induces Protective Mucosal Immunity**. *PLOS ONE* 2009, **4**:e5409.

46. Han M, Kickhoefer VA, Nemerow GR, Rome LH: **Targeted Vault Nanoparticles Engineered with an Endosomolytic Peptide Deliver Biomolecules to the Cytoplasm**. *ACS Nano* 2011, **5**:6128-6137.

47. Kar UK, Jiang J, Champion CI, Salehi S, Srivastava M, Sharma S, Rabizadeh S, Niazi K, Kickhoefer V, Rome LH, et al.: **Vault Nanocapsules as Adjuvants Favor Cell-Mediated over Antibody-Mediated Immune Responses following Immunization of Mice**. *PLOS ONE* 2012, **7**:e38553.

48. Zschoche R, Hilvert D: **Diffusion-Limited Cargo Loading of an Engineered Protein Container**. *Journal of the American Chemical Society* 2015, **137**:16121-16132.

49. Correia BE, Bates JT, Loomis RJ, Baneyx G, Carrico C, Jardine JG, Rupert P, Correnti C, Kalyuzhniy O, Vittal V, et al.: **Proof of principle for epitope-focused vaccine design**. *Nature* 2014, **507**:201-206.

50. Sliepen K, Ozorowski G, Burger JA, van Montfort T, Stunnenberg M, LaBranche C, Montefiori DC, Moore JP, Ward AB, Sanders RW: **Presenting native-like HIV-1 envelope trimers on ferritin nanoparticles improves their immunogenicity**. *Retrovirology* 2015, **12**:82.

51. Gonen S, DiMaio F, Gonen T, Baker D: **Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces**. *Science* 2015, **348**:1365-1368.

•• The authors reported three symmetric protein layers designed in each case by introducing a de novo interface between symmetry-related copies of a natural cyclic oligomer. These 2D protein layers were designed following precise rules of symmetry and showed long range order.

52. Matthaei JF, DiMaio F, Richards JJ, Pozzo LD, Baker D, Baneyx F: **Designing Two-Dimensional Protein Arrays through Fusion of Multimers and Interface Mutations**. *Nano Letters* 2015, 10.1021/acs.nanolett.5b01499.

53. Fletcher JM, Harniman RL, Barnes FRH, Boyle AL, Collins A, Mantell J, Sharp TH, Antognozzi M, Booth PJ, Linden N, et al.: **Self-Assembling Cages from Coiled-Coil Peptide Modules**. *Science* 2013, **340**:595-599.

• Mimicking the self-assembly strategies of proteins, the authors showed that an extended material can be designed by linking trimeric and dimeric coiled-coils through disulfide bonds.

54. Doll TAPF, Dey R, Burkhard P: **Design and optimization of peptide nanoparticles**. *Journal of Nanobiotechnology* 2015, **13**:73.

55. Gradišar H, Božič S, Doles T, Vengust D, Hafner-Bratkovič I, Mertelj A, Webb B, Šali A, Klavžar S, Jerala R: **Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments**. *Nature chemical biology* 2013, **9**:362-366.

56. Lai Y-T, Jiang L, Chen W, Yeates TO: **On the predictability of the orientation of protein domains joined by a spanning alpha-helical linker**. *Protein Engineering Design and Selection* 2015, 10.1093/protein/gzv035:gzv035.

57. Jeong WH, Lee H, Song DH, Eom J-H, Kim SC, Lee H-S, Lee H, Lee J-O: **Connecting two proteins using a fusion alpha helix stabilized by a chemical cross linker**. *Nature Communications* 2016, **7**:11031.

Chapter 4. Structure of a designed tetrahedral protein assembly variant engineered to have improved soluble expression[3]

Authors: Jacob B. Bale[1,2], Rachel U. Park[1], Yuxi Liu[3], Shane Gonen[1,4], Tamir Gonen[4], Duilio Cascio[5], Neil P. King[1,6], Todd O. Yeates[3,5], and David Baker[1,6,7*]

Affiliations:

[1] Department of Biochemistry, University of Washington, Seattle, WA 98195, USA.

[2] Graduate Program in Molecular and Cellular Biology, University of Washington, Seattle, WA 98195, USA.

[3] UCLA Department of Chemistry and Biochemistry, Los Angeles, CA 90095, USA.

[4] Janelia Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn VA 20147.

[5] UCLA-DOE Institute for Genomics and Proteomics, Los Angeles, CA 90095, USA.

[6] Institute for Protein Design, University of Washington, Seattle, WA 98195, USA.

[7] Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA.

**Correspondence to*: David Baker, University of Washington, Molecular Engineering and Sciences Building, Box 351655, Seattle, WA 98195-1655. E-mail: dabaker@uw.edu

[3] This chapter is the adapted version of a published article. Reprint with permission from John Wiley and Sons (license number 4425180572168). The project was a collaboration with the Baker lab and King lab at University of Washington, Seattle. I performed the crystallography work in this project. The whole manuscript is included here to provide an overall picture.

ABSTRACT

We recently reported the development of a computational method for the design of co-assembling, multi-component protein nanomaterials. While four such materials were validated at high-resolution by X-ray crystallography, low yield of soluble protein prevented X-ray structure determination of a fifth designed material, T33-09. Here we report the design and crystal structure of T33-31, a variant of T33-09 with improved soluble yield resulting from redesign efforts focused on mutating solvent-exposed side chains to charged amino acids. The structure is found to match the computational design model with atomic-level accuracy, providing further validation of the design approach and demonstrating a simple and potentially general means of improving the yield of designed protein nanomaterials.

Key Words: computational protein design; crystal structure; solubility; co-assembly; symmetry; tetrahedral; nanomaterial

INTRODUCTION

Symmetric homomeric and heteromeric protein complexes perform a broad range of functions in biological systems[1,2]. Inspired by these natural protein-based molecular machines and materials, many efforts have been undertaken to design novel supramolecular protein structures[3-19]. We recently described a design strategy that combines symmetric modeling with protein-protein interface design in order to generate novel protein assemblies with atomic-level accuracy[7,16]. Using this approach we were able to successfully design five novel tetrahedral protein nanomaterials formed through the co-assembly of multiple copies of two distinct protein subunits[16].

All five designs were confirmed to yield co-assembled nanoparticles of the expected size and shape by analytical size exclusion chromatography (SEC) and negative stain electron microscopy (EM). Crystal structures of four of the nanomaterials were found to match the design models with high accuracy, but we were unable to attempt crystallization of the fifth design, termed T33-09, due to low yield of soluble protein. In addition to the limited soluble yield of T33-09, the majority of unsuccessful designs exhibited low or undetectable amounts of soluble expression. This observation, combined with a lack of discernible differences in the calculated metrics of interface quality for successful and unsuccessful design models, indicated that developing methods to increase soluble expression of the designs is likely to be important for improving our design approach.

With this motivation, we designed and experimentally characterized variants of T33-09 in which a subset of the solvent-exposed side chains on each subunit were mutated to either positively or negatively charged amino acids. This approach, referred to as "supercharging" when taken to an extreme, has previously been shown to be effective at increasing protein solubility[20,21] and is an enticing option for improving our designed nanomaterials as it avoids the need to mutate core or interface residues, which are generally less tolerant of mutations than surface residues. Using a quick and simple cell lysate-based screen, this approach led to the successful production of a design variant with significantly increased soluble yield and to the determination of a high-resolution structure of the redesigned material. As intended, the designed interface and the overall structure of the nanomaterial were not changed during the redesign process and were found to match closely with the experimentally determined structure.

RESULTS AND DISCUSSION

T33-09 is comprised of multiple copies of two distinct protein subunits, referred to as A and B, each about 110 amino acids in length. Both subunit types are naturally trimeric, and the introduction of a *de novo* designed protein-protein interface between the two types of subunits gives rise to a symmetric, tetrahedral assembly comprised of four trimers of each type[16]. In an attempt to rescue the low solubility of this designed material, one positively charged and one negatively charged version of each protein subunit were designed using the Rosetta macromolecular modeling software package (see Materials and Methods), yielding four new variants with 4 to 8 mutations per subunit compared to the original design [Supporting Information Table 4.S1]. Synthetic genes encoding the four designed variants were cloned into the pET29b vector (Novagen) for inducible expression in *Eschericia coli* and the level of soluble expression and assembly state of all nine possible pairwise combinations of original, negatively, or positively charged A and B subunits was then assessed by mixing cell lysates containing the individually expressed subunits and analyzing the resulting soluble and insoluble fractions by polyacrylamide gel electrophoresis (PAGE). One combination of subunits, with a negatively charged A subunit and the original B subunit, was found to significantly increase the yield of the assembled state in the soluble fraction [Supporting Information Figure 4.1S]. We named this new design variant, which contains 5 mutations in the A component relative to the original design, T33-31.

SDS-PAGE analysis of individually expressed subunits showed a clear increase in soluble expression of the redesigned, negatively charged subunit A compared to the original design [Fig. 4.1(A)]. Purified T33-31, obtained by nickel affinity chromatography and size exclusion chromatography of co-expressed (data not shown) or *in vitro*-mixed hexahistidine-

tagged subunits, yielded a dominant peak by analytical SEC near the same elution volume as T33-09, matching the expected size of approximately 24 subunits [Fig. 4.1(B)]. SDS-PAGE analysis of the SEC peak fractions yielded two bands of approximately equal intensity near the expected molecular weights for subunits A and B [Fig. 4.1(C)]. Negative-stain electron microscopy of the purified assembly fractions revealed fields of monodisperse particles that closely resemble the design model at low resolution and are indistinguishable from previously obtained electron micrographs of T33-09 [Fig. 4.1(D)][16]. Taken together, these data provide strong evidence that T33-31 co-assembles to form a structure of similar size and shape to our design model and with the expected one to one stoichiometry of subunits A and B.

Facilitated by the increased yield, purified T33-31 was subsequently characterized by X-ray crystallography in order to confirm the accuracy of the design at high-resolution. T33-31 crystallized readily, leading to the determination of a 3.4 Ångstrom structure [Fig. 4.2]. The asymmetric unit of the crystal comprises one complete tetrahedron. The backbone atoms of the three subunits composing the interface in the design model (two subunits from component A and one subunit from component B) have an average root mean square deviation (r.m.s.d.) of 0.6 Å compared to the twelve non-crystallographically-related instances of the equivalent atoms in the crystal structure. The r.m.s.d. over all backbone atoms in the 24 subunits compared to the design model is only slightly higher at 0.7 Å [Fig. 4.2]. At positions where the electron density permitted side chain placement, the T33-31 design model also matches the crystal structure with high accuracy. While the backbone and side chain conformations do not match as well at the redesigned positions (W43E, Q44E, H62D, A73E, and T78E), this is not surprising because: 1) the backbone degrees of freedom (DOFs) were held fixed during the computational design protocol despite many of the mutated residues residing in loop regions and 2) the side chains are

highly exposed to solvent and expected to be able to adopt many conformations. Other than the five mutated side chains in subunit A and several additional non-mutated surface residues, the T33-09 and T33-31 design models are nearly identical and thus the original T33-09 design model matches the crystal structure equally well over both the backbone and the core and interface side chain conformations.

These results provide further validation of our approach to designing novel supramolecular protein complexes and highlight the potential utility of including residues distant from the protein-protein interface in the design process. The results also demonstrate the modularity and tunability of the designed materials; it is possible to change particular features of the designs, such as solubility, by modifying the different protein subunits (A or B) and/or different regions of the protein subunits (e.g. surface, core, or interface positions) independently of one another. In this case, five surface mutations to subunit A were sufficient to significantly increase the soluble yield of T33-09 without changing the overall structure of the design. This surface redesign approach bypasses the difficulties of adjusting sensitive interfaces and core interactions, providing a relatively simple means of improving the solubility of these materials. Given the many possible applications of designed protein nanomaterials, additional experiments and methods development aimed at improving solubility and other desirable properties of the designs are merited. The genetic basis and modular nature of this class of nanomaterials, combined with the wealth of previously developed methods for protein modification[22,23], should facilitate these efforts. In conjunction with computational redesign approaches, such as the one used in the present study, the development and utilization of methods for directed evolution[24-26] of protein nanostructures[27-29] should provide particularly powerful tools to help tailor these new nanomaterials for a wide variety of features and target applications.

MATERIALS AND METHODS

Computational design. All design calculations were performed within the Rosetta macromolecular modeling suite[30,31] using the original T33-09 design model as the starting point, with the same symmetric modeling setup and treatment of the backbone and rigid body DOFs as published previously[16]. Within this context, side chains with greater than 28 Å^2 of solvent accessible surface area, and not already possessing the desired charge state, were selected as designable positions. Two new design models were generated, one in which all designable residues in subunit A were allowed to mutate to aspartate or glutamate, while those in subunit B were allowed to mutate to arginine or lysine, and another in which all designable residues in subunit A were allowed to mutate to arginine or lysine, while those in subunit B were allowed to mutate to aspartate or glutamate. The resulting designs were refined and selected for experimental characterization based on Rosetta score metrics and visual inspection in PyMOL[32].

Protein expression, lysate screening, and purification. Codon-optimized genes encoding the designed variants of subunit A and B were purchased (Integrated DNA Technologies) and cloned into the pET29b expression vector between the NdeI and XhoI restriction endonuclease sites for individual expression. Two co-expression constructs were also generated in the pET29b expression vector, one expressing the negatively charged subunit A together with the positively charged subunit B and one expressing the positively charged subunit A together with the negatively charged subunit B. The pairs of genes for these co-expression constructs were cloned between the NdeI and XhoI restriction sites and connected by an intergenic region derived from the pETDUET-1 vector as described previously[16]. The pET29b encoded hexahistidine tag was appended to the C-terminus of each individual expression construct and to subunit B in the co-

expression constructs. Expression constructs for the wild-type proteins and the original T33-09 design were generated as described previously[16].

Expression plasmids were transformed into BL21 Star (DE3) *E. coli*. Cells were grown in LB medium supplemented with 50 mg L^{-1} of kanamycin at 37 °C until an OD_{600} of 0.8 was reached. Protein expression was induced by addition of 1.0 mM isopropyl-thio-β-D-galactopyranoside and allowed to proceed for 3 h at 37 °C before cells were harvested by centrifugation. Cells were lysed by sonication in 50 mM TRIS pH 8.0, 250 mM NaCl, 1 mM DTT, 20 mM imidazole supplemented with 1 mM phenylmethanesulfonyl fluoride.

For lysate-based screening experiments, a portion of the crude lysates of the original, negatively and positively charged versions of subunits A and B were mixed in all nine possible pairwise combinations in one-to-one volumetric ratios. Mixed and unmixed lysates were incubated at 4 °C for 1 hour followed by 22 °C for an additional hour. Insoluble material was then cleared by centrifugation and the samples analyzed by denaturing and non-denaturing PAGE. For comparison, the samples were analyzed together with clarified lysates of the unmixed subunits, the wild-type subunits, and co-expressed subunits of the original T33-09 design, negatively charged subunit A and positively charged subunit B, and positively charged subunit A and negatively charged subunit B.

For purification of T33-31, *in vitro*-mixed samples were obtained by mixing cells prior to lysis and subsequently incubating the crude lysates at 4 °C for 1 hour with gentle rocking followed by incubation at 22 °C for 1 hour with gentle rocking. Crude lysates of these *in vitro*-mixed samples, co-expressed T33-09 subunits, and individually expressed wild-type subunits were cleared by centrifugation and filtered through 0.22 μM filters. The filtered supernatants were purified by nickel affinity chromatography and eluted using a linear gradient of imidazole.

Fractions containing pure protein(s) of interest were pooled, concentrated, and further purified on a Superdex 200 10/300 gel filtration column using 25 mM TRIS pH 8.0, 150 mM NaCl, 1 mM DTT as running buffer. Gel filtration fractions containing pure protein in the desired assembly state were pooled, concentrated, and stored at room temperature or 4 °C for subsequent use in analytical size exclusion chromatography, electron microscopy, and X-ray crystallography.

Analytical size exclusion chromatography. Analytical SEC was performed on a Superdex 200 30/100 gel filtration column using the running buffer described above. Wild-type proteins and designed materials were each loaded onto the column at a concentration of 50 μM. The apparent molecular weights of the designed proteins were estimated by comparison to the corresponding wild-type proteins and previously determined nanocage standards.

Negative Stain Electron Microscopy. 3 μl of SEC purified T33-31 at 0.1 mg mL^{-1} was applied to glow discharged, carbon coated 200-mesh copper grids (Ted Pella, Inc.), washed with Milli-Q water and stained with 0.75% uranyl formate as described previously[33]. Grids were visualized on a 120 kV Tecnai Spirit T12 transmission electron microscope (FEI, Hillsboro, OR). All images were recorded using a bottom-mount Teitz CMOS 4k camera at 60,000x magnification at the specimen level. The contrast of all micrographs was enhanced in Fiji[34].

Crystallization. T33-31 was crystallized using the hanging drop vapor diffusion method at room temperature. Crystals grew in hanging drops containing 0.11 μL of protein at 13 mg mL^{-1} and 0.1 μL of a 100 mL well solution containing 100 mM HEPES buffer at pH 7.5, 9% (w/v) polyethylene glycol 8000, and 11.7% (v/v) ethylene glycol. Crystals with tetrahedral or octahedral morphology grew over the course of about two to three days and reached dimensions of about 50-100 μm. For X-ray data collection a crystal was cryo-protected using the well solution augmented with 33% glycerol.

Crystallographic data collection, structure determination, and refinement. Diffraction data sets were collected at the Advanced Photon Source (APS) beamline 24-ID-C equipped with a Pilatus-6M detector. All data were collected at 100 K. Data were collected at a detector distance of 602 mm, with 0.5° oscillations, and at 0.979100 Å wavelength. The crystals showed diffraction to 3.25 Å. The XDS/XSCALE package[35] was used to integrate, reduce, and scale the data. The data were reduced in $P2_12_12_1$ space group symmetry. Based on the crystal symmetry, it was expected that the asymmetric unit of the crystal would contain a complete tetrahedral assembly composed of 24 peptide chains, corresponding to a Matthews coefficient of 2.44 $Å^3$/Da and a 49.5% solvent content in the crystal. We used the PHASER program[36] to determine the structure by molecular replacement, with the full model of the designed tetrahedron as the search model. Molecular replacement yielded a single solution with log-likelihood (LLG) 334. The symmetry axes of the tetrahedron do not overlap with the symmetry axes of the space group. After the solution was obtained, the structure was refined in iterative runs using the BUSTER[37-40] program. In each run, a single translation libration screw-motion (TLS) group was assigned per peptide chain and TLS was switched on for the first and third big-cycles (TLSbasic). We also used the automatic setup for non-crystallographic symmetry (autoncs), and limited the refinement resolution range to 100-3.4 Å. At each step, the quality of the refined model was assessed by COOT[41], and adjustments were made when there was support based on Fo-Fc difference maps. The limited resolution did not support the addition of any bound water molecules during refinement. The final R and R_{free} values were 18.9% and 23.9%. The molecular replacement solution was further confirmed using omit maps (following simulated annealing in torsion angle space) generated around several regions of the protein using PHENIX[42]. Omit maps were calculated around the following regions: residues 18-25 in chains A-L, residues 32-51 in

chains A-L, residues 11-25 in chains M-X, residues 31-61 in chains M-X, residues 15-25 in chains A-L, and 11-25 in chains M-X. These fragments were chosen to be either in the core of one of the protein subunits, or at the designed interface between two proteins. In all cases, the density came back for each of the deleted fragments, validating the molecular replacement solution. Coordinates and structure factors have been deposited in the Protein Data Bank with accession code 4ZK7.

Acknowledgments. This work was supported by the Howard Hughes Medical Institute (TG and DB) and the JFRC visitor program (SG), the National Science Foundation under CHE-1332907 (DB and TOY), the Defense Advanced Research Projects Agency under W911NF-14-1-0162 (DB and NPK), and an NSF graduate research fellowship to JBB (grant no. DGE-0718124). TOY and YL also acknowledge support from the BER program of the DOE Office of Science. We thank Michael Collazo for assistance with protein crystallization and the staff at NECAT beamline 24 ID of the Argonne National Laboratory APS for assistance with X-ray data collection.

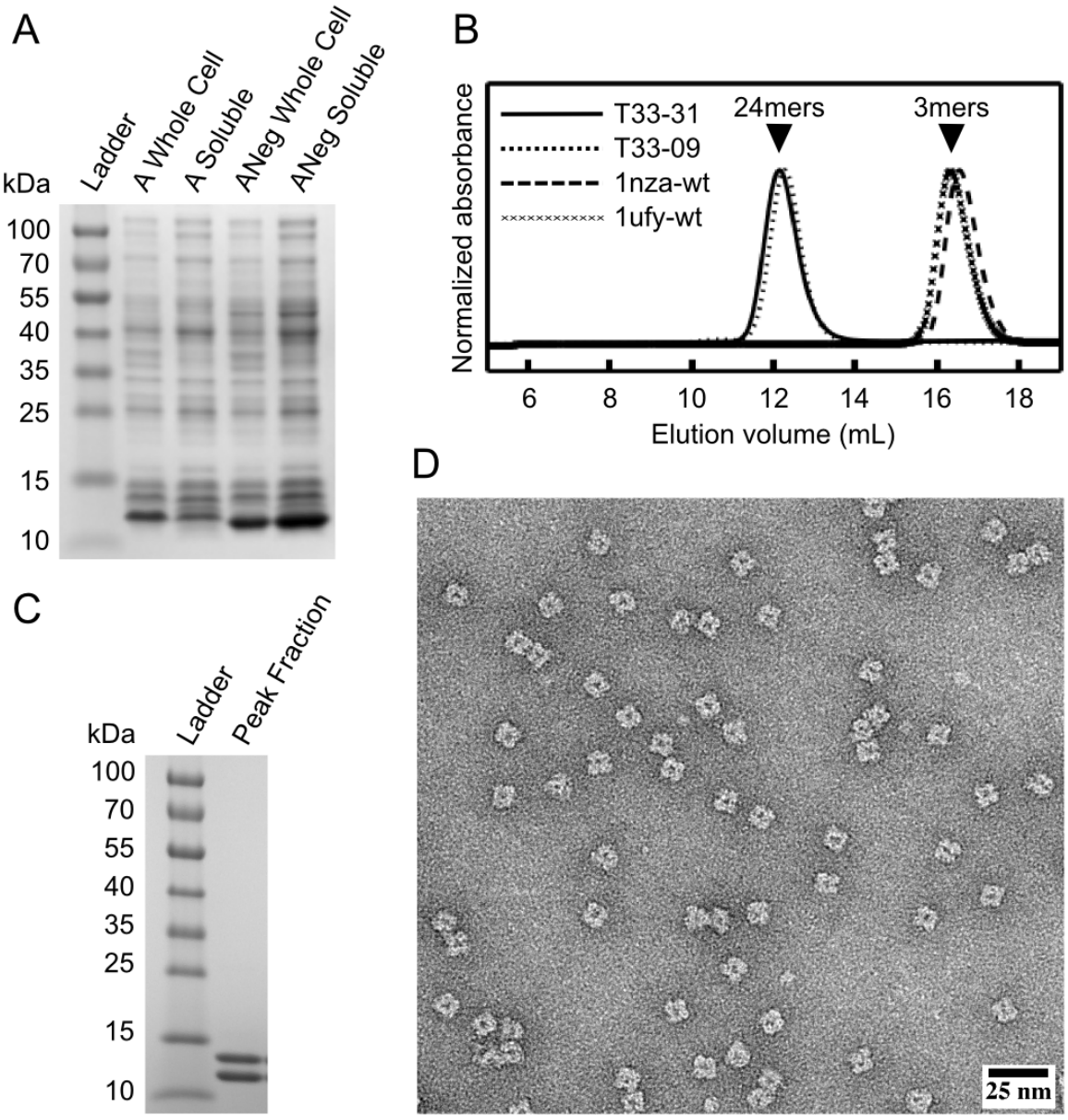


Figure 4.1. Experimental characterization of designed protein assembly T33-31 by SDS-PAGE, analytical SEC, and electron microscopy.

(A) SDS-PAGE analysis of whole cell and clarified lysates from cells expressing the original subunit A or the redesigned, negatively charged subunit A (ANeg). A strong band is observed near the expected molecular weight of 12.5 kDa in the clarified lysate of ANeg, but is only faintly visible in the subunit A sample. (B) SEC chromatograms of purified designs and wild-type oligomeric proteins from which they are derived. The A and B subunits are derived from

Protein Data Bank entries 1nza and 1ufy, respectively. The designed proteins elute near the expected volume for the target tetrahedral assembly (‘24mers’, arrow), while the wild-type proteins elute as trimers (‘3mers’, arrow). The T33-09 sample was produced from co-expressed subunits, while the T33-31 sample was produced through *in vitro* mixing as described in the Materials and Methods. (C) SDS-PAGE analysis of SEC-purified T33-31. Two bands, with approximately equal intensity, are observed near the expected molecular weights of 12.5 and 14.5 kDa. (D) Negative stain electron micrograph of *in vitro*-mixed, SEC-purified T33-31.

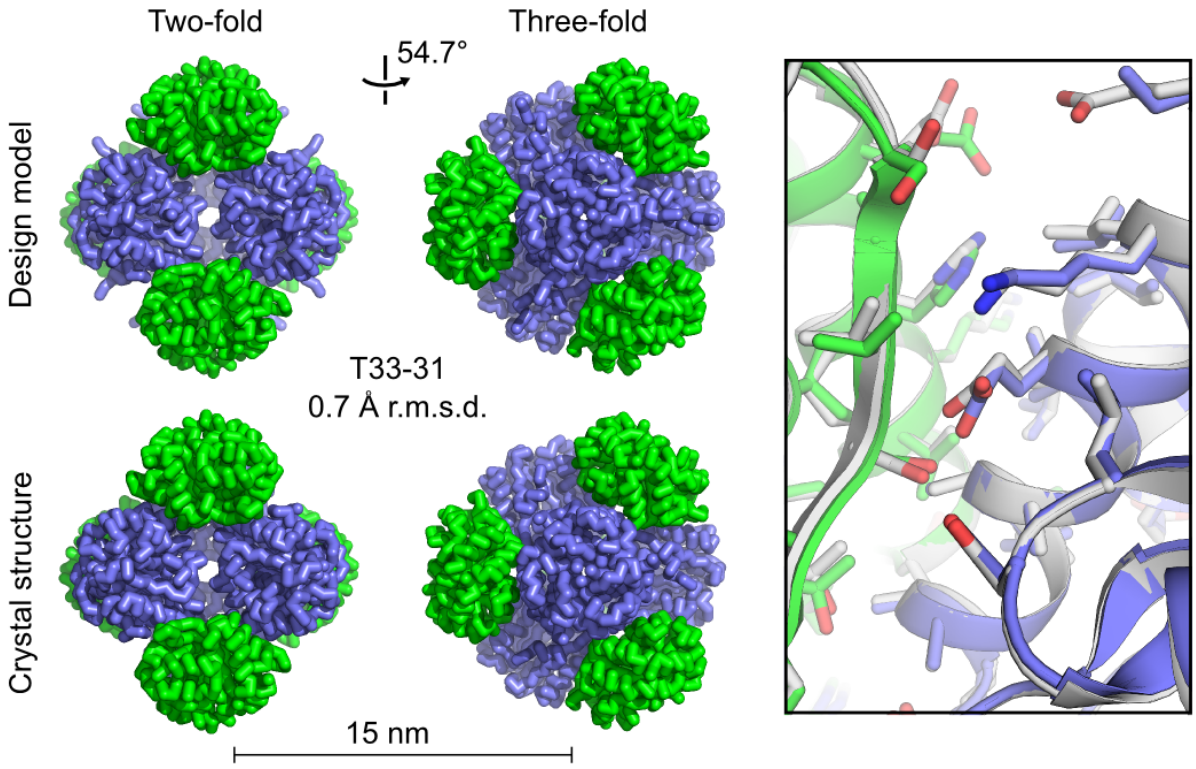


Figure 4.2. T33-31 crystal structure and design model.

At left, views along the two-fold and three-fold symmetry axes are shown for the T33-31 computational design model (top) and crystal structure (bottom, PDB ID 4ZK7, scale bar: 15 nm). The r.m.s.d. was calculated using the backbone atoms in all 24 chains of the design model compared to the asymmetric unit of the crystal structure. At right, an overlay is shown of the designed interface in the design model (white) and crystal structure (green and blue). Poor electron density prevented modeling beyond the beta or delta carbon for some amino acid side chains in the crystal structure. The subunits involved in the interface shown are respresented by protein chains S, A, and U in the deposited PDB structure. In the amino acid side chains shown, oxygen atoms are red, nitrogen atoms are blue, and sulfur atoms are orange.

Table 4.S1. Amino acid sequences of wild-type scaffolds and designed variants.

Mutated residues in the negatively and positively charged variants (relative to the original design) are shown in red and underlined.

Name	Sequence
1NZA[1]	MEEVVLITVPSEEVARTIAKALVEERLAACVNIVPGLTSIYRW QGEVVEDQELLLLVKTTTHAFPKLKERVKALHPYTVPEIVAL PIAEGNREYLDWLRENTG
T33-09A	MEEVVLITVPSALVAVKIAHALVEERLAACVNIVPGLTSIYRW QGSVVSDHELLLLVKTTTHAFPKLKERVKALHPYTVPEIVALP IAEGNREYLDWLRENTG
T33-09ANeg	MEEVVLITVPSALVAVKIAHALVEERLAACVNIVPGLTSIYR<u>E</u> <u>E</u>GSVVSDHELLLLVKTTT<u>D</u>AFPKLKERVK<u>E</u>LHPY<u>E</u>VPEIVALP IAEGNREYLDWLRENTG
T33-09APos	MEEVVLITVPSA<u>K</u>VAVKIAHALV<u>K</u>ERLAACVNIVPGLTSIYR<u>K</u> <u>K</u>GSVVSDHELLLLVKTTT<u>K</u>AFPKLKERVK<u>R</u>LHPY<u>K</u>VPEIVAL PIAEGNREYL<u>R</u>WLRENTG
1UFY[2]	MVRGIRGAITVEEDTPEAIHQATRELLLKMLEANGIQSYEELA AVIFTVTEDLTSAFPAEAARQIGMHRVPLLSAREVPVPGSLPR VIRVLALWNTDTPQDRVRHVYLREAVRLRPDLESAQ
T33-09B	MVRGIRGAITVEEDTPAAILAATIELLLKMLEANGIQSYEELA AVIFTVTEDLTSAFPAEAARLIGMHRVPLLSAREVPVPGSLPR VIRVLALWNTDTPQDRVRHVYLNEAVRLRPDLESAQ
T33-09BNeg	MVRGIRGAITVEEDTPAAILAATIELLLKMLEANGI<u>E</u>SYEELA AVIFTVTEDLTSAFPAEAARLIGMHRVPLLSAREVPVPGSLPR VIRVLALWNTDTPQD<u>E</u>VRHVYLNEAV<u>E</u>LRPDLES<u>D</u>Q
T33-09BPos	MVRGIRGAITVEEDTPAAILAATIELLLKML<u>K</u>ANGIQSY<u>K</u>ELA AVIFTVTEDLTSAFPAEAARLIGMHRVPLLSAREVPVPGSLPR VIRVLALWNT<u>K</u>TPQDRVRHVYLN<u>K</u>A<u>K</u>RLRPDL<u>K</u>S<u>K</u>Q

Footnotes:

1. Protein Data Bank entry for the protein from which the T33-09A sequence is derived

2. Protein Data Bank entry for the protein from which the T33-09B sequence is derived.

Table 4.S2. Crystallographic Statistics for Data Collection and Structure Refinement of T33-31 (PDB ID 4ZK7).

Data Collection	
Space group	$P2_12_12_1$
Cell dimensions	
a, b, c (Å)	121.1, 128.4, 204.7
α, β, γ (°)	90.0, 90.0, 90.0
Resolution (Å)	108.77-3.25
R_{merge} (%)	20.5 (60.6)
CC1/2 (%)	98.4 (73.3)
CC* (%)	99.6 (92.0)
Mean I/σ	5.5 (1.2)
Completeness (%)	96.3 (66.8)
Multiplicity	4.0 (2.0)
Wilson B-factor	57.5
Refinement	
Resolution range (Å)	88.10-3.40 (3.49-3.40)
No. reflections	44218 (3234)
R_{work}/R_{free} (%)*	19.0/23.9
No. atoms	20678
Protein	20678
Ligand/ion	0
Water	0
Average B factors	72.6
Protein	72.6
Ligand/ion	NA
Water	NA
Protein residues	2646
R.m.s. deviations	
Bond length (Å)	0.01
Bond angles (Å)	1.2
Ramachandran favored (%)	91.3
Ramachandran allowed (%)	8.3
Ramachandran generally allowed (%)	0.5
Ramachandran outliers (%)	0

Footnotes:

Statistics in parentheses refer to the highest resolution shell

* R_{free} calculated using 10% of the data.

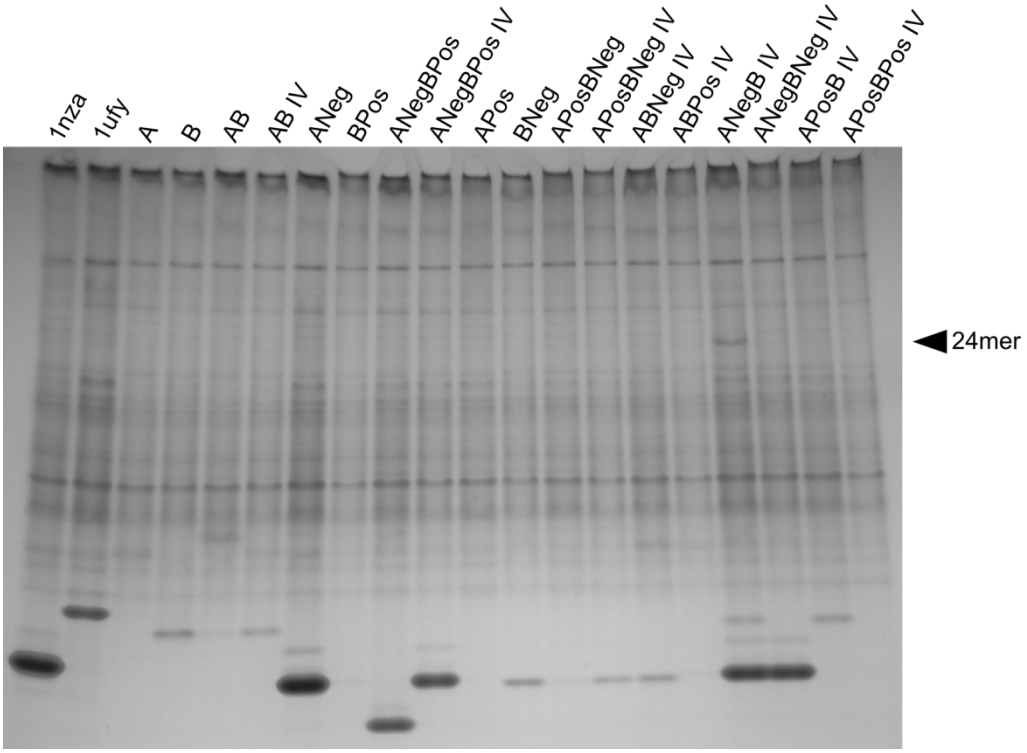


Figure 4.S1. Native PAGE analysis of wild-type proteins and designed variants.

Clarified cell lysates of the wild-type protein scaffolds from which the A and B subunits of T33-09 are derived (PDB IDs 1NZA and 1UFY); individually expressed original, negatively, and positively charged subunits; co-expressed subunits of original A and B, ANeg with BPos, and APos with BNeg; and *in vitro*-mixed samples of individually expressed subunits (indicated by "IV" in the labels above) were subjected to native PAGE and stained with GelCode Blue (Thermo Scientific). A slowly migrating band ('24mer', arrow), absent from the unmixed ANeg and B samples, is clearly observed in the ANegB IV (T33-31) sample. Such a band is not clearly detectable in the clarified lysates of the original T33-09 design (AB and AB IV) or any of the other designed variants. Such a band is detectable for the original T33-09 design only when subunit B possesses a peptide tag for fluorescence labeling instead of a polyhistidine tag[16]. *In vitro*-mixed samples were produced through mixing of equal volumes of crude lysates containing the individually expressed subunits as described in the Materials and Methods section.

REFERENCES

1. Goodsell DS, Olson AJ (2000) Structural symmetry and protein function. Annu Rev Biophys Biomol Struct 29:105-53.
2. Janin J, Bahadur RP, Chakrabarti P (2008) Protein-protein interaction and quaternary structure. Q Rev Biophys 41:133-80.
3. Lai YT, King NP, Yeates TO (2012) Principles for designing ordered protein assemblies. Trends Cell Biol 22:653-61.
4. King NP, Lai YT (2013) Practical approaches to designing novel protein assemblies. Curr Opin Struct Biol 23:632-8.
5. Sinclair JC (2013) Constructing arrays of proteins. Curr Opin Chem Biol 17:946-51.
6. Brodin JD, Ambroggio XI, Tang C, Parent KN, Baker TS, Tezcan FA (2012) Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays. Nat Chem 4:375-82.
7. King NP, Sheffler W, Sawaya MR, Vollmar BS, Sumida JP, Andre I, Gonen T, Yeates TO, Baker D (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science 336:1171-4.
8. Lanci CJ, MacDermaid CM, Kang SG, Acharya R, North B, Yang X, Qiu XJ, DeGrado WF, Saven JG (2012) Computational design of a protein crystal. Proc Natl Acad Sci U S A 109:7304-9.
9. Stranges PB, Machius M, Miley MJ, Tripathy A, Kuhlman B (2011) Computational design of a symmetric homodimer using beta-strand assembly. Proc Natl Acad Sci U S A 108:20562-7.

10. Sinclair JC, Davies KM, Venien-Bryan C, Noble ME (2011) Generation of protein lattices by fusing proteins with matching rotational symmetry. Nat Nanotechnol 6:558-62.
11. Lai YT, Cascio D, Yeates TO (2012) Structure of a 16-nm cage designed by using protein oligomers. Science 336:1129.
12. Der BS, Machius M, Miley MJ, Mills JL, Szyperski T, Kuhlman B (2012) Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. J Am Chem Soc 134:375-85.
13. Fletcher JM, Harniman RL, Barnes FR, Boyle AL, Collins A, Mantell J, Sharp TH, Antognozzi M, Booth PJ, Linden N and others (2013) Self-assembling cages from coiled-coil peptide modules. Science 340:595-9.
14. Boyle AL, Bromley EH, Bartlett GJ, Sessions RB, Sharp TH, Williams CL, Curmi PM, Forde NR, Linke H, Woolfson DN (2012) Squaring the circle in peptide assembly: from fibers to discrete nanostructures by de novo design. J Am Chem Soc 134:15457-67.
15. Grigoryan G, Kim YH, Acharya R, Axelrod K, Jain RM, Willis L, Drndic M, Kikkawa JM, DeGrado WF (2011) Computational design of virus-like protein assemblies on carbon nanotube surfaces. Science 332:1071-6.
16. King NP, Bale JB, Sheffler W, McNamara DE, Gonen S, Gonen T, Yeates TO, Baker D (2014) Accurate design of co-assembling multi-component protein nanomaterials. Nature 510:103-8.
17. Gradisar H, Bozic S, Doles T, Vengust D, Hafner-Bratkovic I, Mertelj A, Webb B, Sali A, Klavzar S, Jerala R (2013) Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments. Nat Chem Biol 9:362-6.

18. Lai YT, Reading E, Hura GL, Tsai KL, Laganowsky A, Asturias FJ, Tainer JA, Robinson CV, Yeates TO (2014) Structure of a designed protein cage that self-assembles into a highly porous cube. Nat Chem 6:1065-71.
19. Voet AR, Noguchi H, Addy C, Simoncini D, Terada D, Unzai S, Park SY, Zhang KY, Tame JR (2014) Computational design of a self-assembling symmetrical beta-propeller protein. Proc Natl Acad Sci U S A 111:15102-7.
20. Der BS, Kluwe C, Miklos AE, Jacak R, Lyskov S, Gray JJ, Georgiou G, Ellington AD, Kuhlman B (2013) Alternative computational protocols for supercharging protein surfaces for reversible unfolding and retention of stability. PLoS One 8:e64363.
21. Lawrence MS, Phillips KJ, Liu DR (2007) Supercharging proteins can impart unusual resilience. J Am Chem Soc 129:10110-2.
22. Stephanopoulos N, Francis MB (2011) Choosing an effective protein bioconjugation strategy. Nat Chem Biol 7:876-84.
23. Spicer CD, Davis BG (2014) Selective chemical protein modification. Nat Commun 5:4740.
24. Jackel C, Kast P, Hilvert D (2008) Protein design by directed evolution. Annu Rev Biophys 37:153-73.
25. Badran AH, Liu DR (2015) In vivo continuous directed evolution. Curr Opin Chem Biol 24:1-10.
26. Currin A, Swainston N, Day PJ, Kell DB (2015) Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. Chem Soc Rev 44:1172-239.
27. Worsdorfer B, Woycechowsky KJ, Hilvert D (2011) Directed evolution of a protein container. Science 331:589-92.

28. Song WJ, Tezcan FA (2014) A designed supramolecular protein assembly with in vivo enzymatic activity. Science 346:1525-8.
29. Dalkara D, Byrne LC, Klimczak RR, Visel M, Yin L, Merigan WH, Flannery JG, Schaffer DV (2013) In vivo-directed evolution of a new adeno-associated virus for therapeutic outer retinal gene delivery from the vitreous. Sci Transl Med 5:189ra76.
30. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W and others (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 487:545-74.
31. DiMaio F, Leaver-Fay A, Bradley P, Baker D, Andre I (2011) Modeling symmetric macromolecular structures in Rosetta3. PLoS One 6:e20450.
32. The PyMOL Molecular Graphics System. 1.5: Schrödinger, LLC 2012.
33. Nannenga BL, Iadanza MG, Vollmar BS, Gonen T (2013) Overview of electron crystallography of membrane proteins: crystallization and screening strategies using negative stain electron microscopy. Curr Protoc Protein Sci Chapter 17:Unit17 15.
34. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B and others (2012) Fiji: an open-source platform for biological-image analysis. Nat Methods 9:676-82.
35. Kabsch W (2010) Xds. Acta Crystallogr D Biol Crystallogr 66:125-32.
36. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. J Appl Crystallogr 40:658-674.
37. Bricogne G (1997) [23] Bayesian statistical viewpoint on structure determination: Basic concepts and examples. 276:361-423.

38. Blanc E, Roversi P, Vonrhein C, Flensburg C, Lea SM, Bricogne G (2004) Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. Acta Crystallogr D Biol Crystallogr 60:2210-21.

39. Roversi P, Blanc E, Vonrhein C, Evans G, Bricogne G (2000) Modelling prior distributions of atoms for macromolecular refinement and completion. Acta Crystallogr D Biol Crystallogr 56:1316-23.

40. Bricogne G (1993) Direct phase determination by entropy maximization and likelihood ranking: status report and perspectives. Acta Crystallogr D Biol Crystallogr 49:37-60.

41. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. Acta Crystallogr D Biol Crystallogr 66:486-501.

42. Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW and others (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr 66:213-21.

Chapter 5. Accurate design of megadalton-scale two-component icosahedral protein complexes[4]

One Sentence Summary:

A computational approach enables design of 120-subunit icosahedral protein cages capable of packaging macromolecular cargo.

Authors:

Jacob B Bale[1,2], Shane Gonen[1,3†], Yuxi Liu[4†], William Sheffler[1], Daniel Ellis[5], Chantz Thomas[6], Duilio Cascio[4,7,8], Todd O Yeates[4,7], Tamir Gonen[3], Neil P King[1,5,*], and David Baker[1,5,9,*]

Affiliations:

[1] Department of Biochemistry, University of Washington, Seattle, WA 98195, USA.

[2] Graduate Program in Molecular and Cellular Biology, University of Washington, Seattle, WA 98195, USA.

[3] Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA.

[4] UCLA Department of Chemistry and Biochemistry, Los Angeles, CA 90095, USA.

[5] Institute for Protein Design, University of Washington, Seattle, Washington 98195.

[6] Department of Chemistry, University of Washington, Seattle, WA 98195, USA.

[7] UCLA-DOE Institute for Genomics and Proteomics, Los Angeles, CA 90095, USA.

[4] This chapter is the adapted version of a published article. The project was a collaboration with the Baker lab and King lab at University of Washington, Seattle. I performed the crystallography work in this project. The whole manuscript is included here to provide an overall picture.

[8] UCLA Department of Biological Chemistry and The Molecular Biology Institute, Los Angeles, CA 90095, USA.

[9] Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA.

† These authors contributed equally to this work.

* Corresponding author. E-mail: neilking@uw.edu (NPK); dabaker@uw.edu (DB)

ABSTRACT

Nature provides many examples of self- and co-assembling protein-based molecular machines, including icosahedral protein cages that serve as scaffolds, enzymes, and compartments for essential biochemical reactions and icosahedral virus capsids, which encapsidate and protect viral genomes and mediate entry into host cells. Inspired by these natural materials, we report the computational design and experimental characterization of co-assembling two-component 120-subunit icosahedral protein nanostructures with molecular weights (1.8-2.8 MDa) and dimensions (24-40 nm diameter) comparable to small viral capsids. Electron microscopy, SAXS, and X-ray crystallography show that ten designs spanning three distinct icosahedral architectures form materials closely matching the design models. *In vitro* assembly of independently purified components reveals rapid assembly rates comparable to viral capsids and enables controlled packaging of molecular cargo via charge complementarity. The ability to design megadalton-scale materials with atomic-level accuracy and controllable assembly opens the door to a new generation of genetically programmable protein-based molecular machines.

MAIN TEXT

The remarkable forms and functions of natural protein assemblies have inspired many efforts to engineer self- and co-assembling protein complexes (*1-24*). A common feature of these approaches, as well as the structures that inspired them, is symmetry. By repeating a small number of interactions in geometric arrangements consistent with the formation of regular structures, symmetry reduces the number of unique interactions and subunits required to form higher order assemblies (*2, 25*). Symmetric complexes can be designed to form through self-assembly of a single type of protein subunit or co-assembly of two or more distinct types of

protein subunits. Multi-component materials possess several important advantages, including the potential to control initiation of assembly by mixing independently prepared components. This property could allow, for example, assembly to be performed in the presence of cargo molecules in order to package the cargo inside the designed nanomaterial. Thus far, only relatively small, 24-subunit two-component tetrahedra have been designed with high accuracy (*20, 26*). Packaging substantial amounts of cargo will require larger assemblies, such as those with icosahedral symmetry; icosahedra possess the highest possible symmetry of any polyhedron in three-dimensional space and, consequently for the purpose of packaging, generate the maximum enclosed volume for symmetric assemblies formed from a given size protein subunit (*27, 28*).

We set out to design two-component icosahedral protein complexes capable of packaging macromolecular cargo via controlled *in vitro* assembly. The two-fold, three-fold, and five-fold rotational axes present within icosahedral symmetry provide three possible ways to construct such complexes from pairwise combinations of oligomeric building blocks; we refer to these architectural types as I53, I52 and I32 (fig. S1). The I53 architecture is formed from a combination of twelve pentameric building blocks and twenty trimeric building blocks aligned along the five-fold and three-fold icosahedral symmetry axes, respectively (Fig. 5.1, A-E; I53 = Icosahedral assembly constructed from 5mers and 3mers). Similarly, the I52 architecture is formed from twelve pentamers and thirty dimers (Fig. 5.1F), and the I32 architecture is formed from twenty trimers and thirty dimers, each aligned along their corresponding icosahedral symmetry axes (Fig. 5.1G). To generate novel icosahedral assemblies, 14,400 pairs of pentamers and trimers, 50,400 pairs of pentamers and dimers, and 276,150 pairs of trimers and dimers derived from X-ray crystal structures (tables 5.S1-3) were arranged as described above, with each building block allowed to rotate around and translate along its five-fold, three-fold, or two-

fold symmetry axis. These degrees of freedom (DOFs) were systematically sampled to identify configurations suitable for interface design, as assessed by several parameters, including the size and secondary structure content of the newly formed interface, as well as the backbone geometry between pairs of contacting residues. Protein-protein interface design calculations were then carried out on the resulting 66,115 designs of type I53, 35,468 designs of type I52, and 161,007 designs of type I32. The designs were filtered based on a variety of metrics, including interface area, predicted binding energy, and shape complementarity (*29*). 71 designs of type I53, 44 of type I52, and 68 of type I32, derived from 23 distinct pentameric, 57 distinct trimeric, and 91 distinct dimeric protein scaffolds, were selected for experimental characterization (fig. S2-5, table 5.S4).

Codon-optimized genes encoding each pair of designed sequences were cloned into a vector for inducible co-expression in *E. coli*, with a hexahistidine tag appended to the N- or C-terminus of one subunit in each pair. The proteins were expressed at small scale, purified by immobilized metal-affinity chromatography (IMAC), and clarified lysates and purification products subjected to gel electrophoresis under denaturing conditions to screen for soluble expression and co-purification of the hexahistidine-tagged and non-tagged subunits (fig. S6A). Designs appearing to co-purify were subsequently analyzed by non-denaturing gel electrophoresis to screen for slowly migrating species as an additional indication of assembly to higher order materials (fig. S6B). Those found to both co-purify and assemble were expressed at larger scale and purified by IMAC followed by size exclusion chromatography (SEC, fig. S7). Ten pairs of designed proteins, four I53 (I53-34, I53-40, I53-47, and I53-50), three I52 (I52-03, I52-32, and I52-33) and three I32 designs (I32-06, I32-19, and I32-28), yielded major peaks by SEC near the elution volumes expected based on the diameters of the design models (Fig. 5.2,

table 5.S4). Two other designs, I53-51 and I32-10, also appeared to form large, discrete assemblies, but their structures could not be verified by subsequent experiments (Supplementary Text, fig. S8 and S9).

Small-angle X-ray scattering (SAXS) performed on the SEC-purified samples indicated all ten designs form assemblies similar to the intended three-dimensional configurations in solution. The experimentally measured SAXS profiles are feature-rich and distinct, with multiple large dips in scattering intensity in the region between 0.015 $Å^{-1}$ and 0.15 $Å^{-1}$, each of which is closely recapitulated in profiles calculated from the design models (Fig. 5.2). In order to further evaluate how accurately and uniquely the design models match the experimental data, each was compared to a set of alternative models generated by systematically perturbing the radial displacements and rotations of the building blocks in each design by +/- 10 Å and/or 20 degrees, respectively. The vast majority of alternative configurations were found to produce worse fits to the experimental data than the original design models (Fig. 5.2), suggesting that the materials assemble quite precisely in solution.

The information provided by SAXS about the overall ensemble of structures observed in solution for each design was complemented and corroborated by visualization of individual particles by negative stain electron microscopy (EM). Micrographs of I53-34, I53-40, I53-47, I53-50, I52-03, I52-33, I32-06, and I32-28 show fields of particles with the expected size and shape of the design models, and particle averaging yields distinct structures clearly matching the models (Fig. 5.3). The large trimeric and pentameric voids observed in the I52 and I32 averages, for instance, closely resemble the cavities in projections generated from the corresponding design models when viewed down the three-fold and five-fold symmetry axes, respectively. The turreted morphology of the I53-50 and I52-33 design models and projections, resulting from

pentameric and dimeric components that protrude away from the rest of the icosahedral shell, are also readily apparent in the corresponding class averages. Although the results from SEC and SAXS strongly indicate I52-32 and I32-19 form assemblies closely matching the design models in solution, both appear to be unstable under the conditions encountered during grid preparation, yielding broken particles not suitable for further EM analysis (fig. S10).

To further evaluate the accuracy of our designs, X-ray crystal structures were determined for one material from each of the three different architectural types: I53-40, I52-32, and I32-28 (Fig. 5.4 and table 5.S5). Although the resolution of the structures (3.5 to 5.6 Å) is insufficient to permit detailed analysis of the side chains at the designed interfaces, backbone-level comparisons show the inter-building block interfaces were designed with high accuracy, giving rise to 120-subunit complexes that match the computational design models remarkably well. Comparing pairs of interface subunits from each structure to the design models yields backbone root mean square deviations (r.m.s.d.) between 0.2 and 1.1 Å, while the r.m.s.d. over all 120 subunits in each material ranges from 0.8 to 2.7 Å (Fig. 5.4, A-C and table 5.S6). With diameters between 26 and 31 nm, over 130,000 heavy atoms, and molecular weights greater than 1.9 megadaltons, these structures are comparable in size to small viral capsids and, to our knowledge, the largest designed biomolecular nanostructures to date to be verified by X-ray crystallography (fig. S11).

The multi-component composition of the materials presents the possibility of controlling their assembly through *in vitro* mixing of independently produced building blocks (*20*). Taking advantage of this feature, the assembly kinetics of an I53-50 variant (fig. S12A) with improved individual subunit stability was investigated by light scattering (Supplementary Materials). SEC-purified components were mixed at concentrations of 64, 32, 16, or 8 μM and the change in light

scattering monitored over time (Fig. 5.4D). Assembly is roughly halfway complete within 1 minute at 64 and 32 uM, 3 minutes at 16 uM, and 10 minutes at 8 uM. Similar assembly time scales have been observed for several viral capsids (*30, 31*). Since our design process focused exclusively on structure without any consideration of kinetics, these results raise the interesting possibility that the rate of assembly of these viral capsids has not been highly optimized during evolution.

The ability to assemble the materials *in vitro* potentially enables the controlled packaging of macromolecular cargoes. To investigate this possibility, the trimeric and pentameric components of an I53-50 variant with several mutations to positively charged residues on the interior surfaces of the two components (Supplementary Materials) were successively mixed with a supercharged GFP with a net charge of -30 (*32*), and encapsulation was evaluated using SEC followed by SDS-PAGE of relevant fractions (Fig. 5.4E and Supplementary Materials). When both the packaging reaction and SEC were performed in a buffer containing low (65 mM) NaCl, GFP(-30) and both I53-50 components co-eluted from the column at the same elution volume previously observed for unmodified I53-50 (Fig. 5.2D). Mixtures of GFP(-30) with only one of the two components eluted at later volumes, indicating that the observed co-elution requires assembly of I53-50 (fig. S12, B-D). When the packaging reaction was carried out with buffer containing high (1 M) NaCl or using a variant of the trimeric component lacking the positively charged residues on the interior surface, little to no co-elution was observed (Fig. 5.4E), demonstrating that packaging is driven by the engineered electrostatic interactions between the I53-50 interior and GFP(-30). High salt incubation resulted in disassociation of packaged GFP (fig. S12E), as also observed for an evolved variant of a naturally occurring protein container that packages cargo via electrostatic complementarity (*33, 34*). Based on

measurements of fluorescence intensity and UV/Vis absorbance, we estimate approximately 7 to 11 GFPs are packaged per icosahedral assembly in 65 mM NaCl, occupying roughly 11 to 17% of the interior volume (Supplementary Materials).

How do the architectures of our designs compare to those of virus capsids and other icosahedral protein complexes found in nature? In the nomenclature introduced by Caspar and Klug (*27*), our designs can be considered T=1 assemblies in which the asymmetric unit is a heterodimer comprising one subunit from each of the two components. The most similar naturally occurring structures of which we are aware are Cowpea Mosaic Virus (CPMV) and related 120-subunit capsids with pseudo T=3 symmetry. Like our I53 designs, CPMV is composed of 60 copies each of two distinct protein subunits, with one type of subunit arranged around the icosahedral 5-folds and a second type of subunit arranged around the 3-folds (fig. S13). However, the two subunits of CPMV are composed of three similar domains occupying spatially equivalent positions to those found in T=3 assemblies formed from 180 copies of a single type of protein subunit (*35, 36*). Our I53 designs display no such underlying pseudosymmetry and therefore cannot be considered pseudo T=3. Furthermore, we are not aware of any natural protein complexes characterized to date that exhibit I52 or I32 architectures. Our designs thus appear to occupy new regions of the protein assembly universe, which have either not yet been explored by natural evolution or are undiscovered at present in natural systems.

The size and complexity of the materials presented herein, together with the accuracy with which they assemble, push the boundaries of biomolecular engineering into new and exciting territory. The large lumens of our designed materials, combined with their multi-component nature and the ability to control assembly via mixing of purified components, makes

them well suited for encapsulation of a broad range of materials including small molecules, nucleic acids, polymers, and other proteins. These features, along with their potential for precisely engineered chemical or genetic modifications, make them attractive starting materials for the design of functional protein nanomaterials for applications in targeted drug delivery, vaccine design, and bioenergy.

ACKNOWLEDGMENTS

We thank Michael Sawaya and Michael Collazo for their assistance with crystallography, conducted at the UCLA-DOE X-ray Crystallization and Crystallography Core Facilities, which are supported by DOE Grant DE-FC02-02ER63421. We thank M. Capel, K. Rajashankar, N.Sukumar, J.Schuermann, I. Kourinov and F. Murphy at NECAT beamlines 24-ID at APS, which are supported by grants from the National Center for Research Resources (5P41RR015301-10) and the National Institute of General Medical Sciences (8 P41 GM103403-10) from the National Institutes of Health. Use of the APS is supported by DOE under Contract DE-AC02-06CH11357. We thank the staff at the SIBYLS beamline at Lawrence Berkeley National Lab, including Kathryn Burnett, Gregory Hura, Michal Hammel, Jane Tanamachi, and John Tainer for the services provided through the mail-in SAXS program, supported by the Department of Energy (DOE) Integrated Diffraction Analysis (IDAT) grant contract number DE-AC02-05CH11231. We also thank Una Nattermann for help with electron microscopy, Yang Hsia for assistance with light scattering experiments, Chris Stafford for mass spectroscopy, Brooke Nickerson for assistance with *in vitro* assembly experiments, and Gabriel Rocklin for providing scripts used in data analysis. This work was supported by the Howard Hughes Medical Institute (SG, DC, TG, and DB) and the Janelia Research Campus visitor program (SG), the Bill and Melinda Gates Foundation (DB and NPK), Takeda Pharmaceutical Company (NPK),

the National Science Foundation (DB and TOY, grant no. CHE-1332907) and the Defense Advanced Research Projects Agency (DB and NPK, grant no. W911NF-14-1-0162). YL was supported by a Whitcome Fellowship through the UCLA Molecular Biology Institute and JBB by an NSF graduate research fellowship (grant no. DGE-0718124). Coordinates and structure factors were deposited in the Protein Data Bank with accession codes 5IM5 (I53-40), 5IM4 (I52-32), and 5IM6 (I32-28). JBB, WS, NPK, DE, and DB have filed a non-provisional patent application, U.S. 14/930,792, related to the work presented herein.

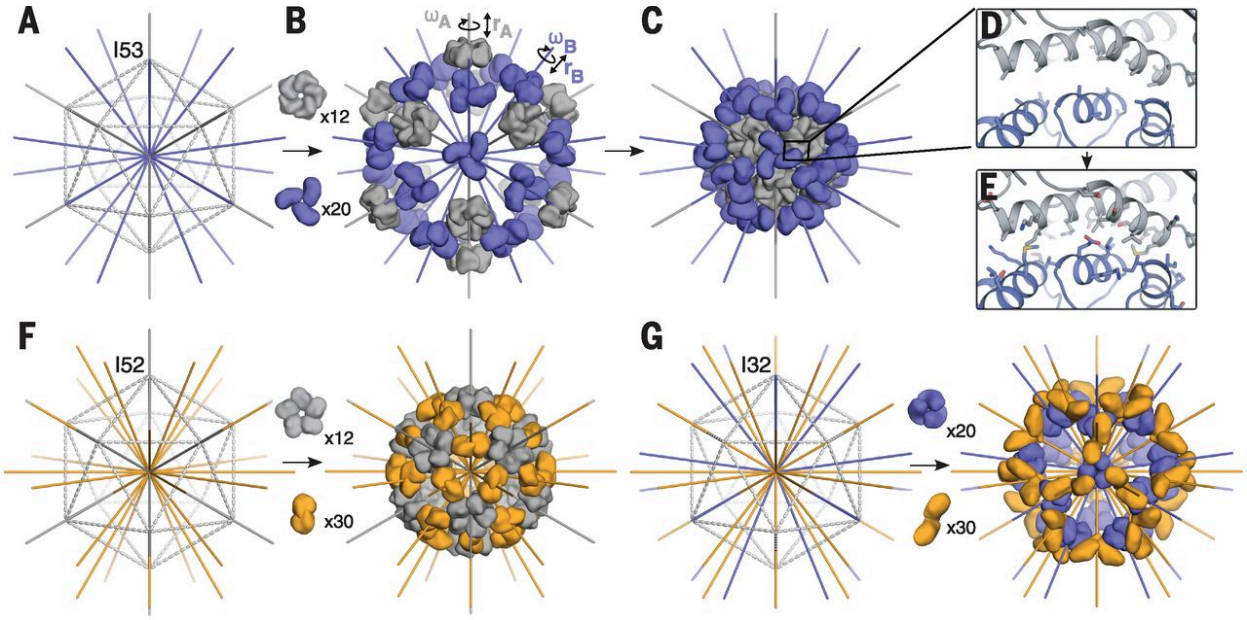


Fig. 5.1. Overview of the design method and target architectures.

(**A-E**) A schematic of the design process illustrated with the I53 architecture. (**A)** An icosahedron is outlined with dashed lines, with the five-fold symmetry axes (grey) going through each vertex and three-fold symmetry axes (blue) going through each face of the icosahedron. (**B)**12 pentamers (grey) and 20 trimers (blue) are aligned along the 5-fold and 3-fold symmetry axes, respectively. Each oligomer possesses two rigid body DOFs, one translational (r) and one rotational (ω) that are systematically sampled to identify configurations (**C**) with a large interface between the pentamer and trimer (**D**) suitable for protein-protein interface design; only the backbone structure and beta carbons of the oligomers are taken into account during this procedure. (**E)** Amino acid sequences are designed at the new interface to stabilize the modeled configuration. **(F)** The I52 architecture comprises 12 pentamers (grey) and 30 dimers (orange) aligned along the five-fold and two-fold icosahedral symmetry axes. **(G)** The I32 architecture comprises 20 trimers (blue) and 30 dimers (orange) aligned along the three-fold and two-fold icosahedral symmetry axes.

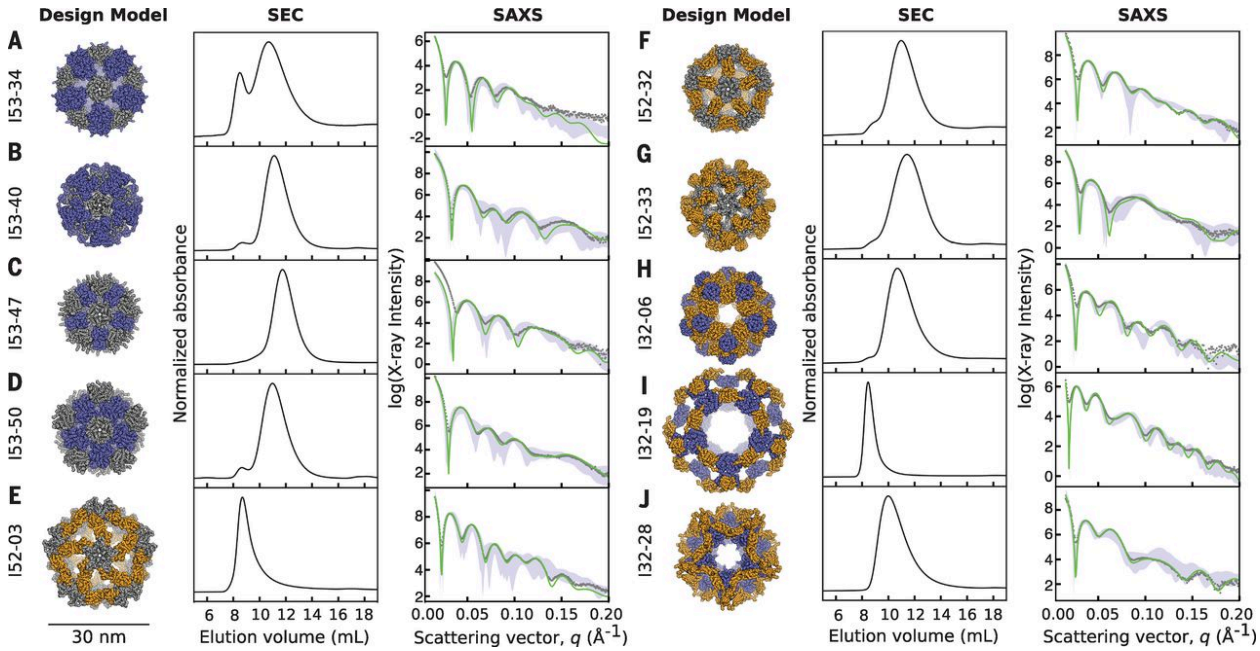


Fig. 5.2. Experimental characterization by size exclusion chromatography and small-angle X-ray scattering.

Computational design models (left), SEC chromatograms (middle), and SAXS profiles (right) are shown for (**A**) I53-34, (**B**) I53-40, (**C**) I53-47, (**D**) I53-50, (**E**) I52-03, (**F**) I52-32, (**G**) I52-33, (**H**) I32-06, (**I**) I32-19, and (**J**) I32-28. Design models (shown to scale relative to the 30 nm scale bar) are viewed down one of the 5-fold symmetry axes with ribbon-style renderings of the protein backbone (pentamers are shown in grey, trimers in blue, and dimers in orange). Co-expressed and purified designs yield dominant SEC peaks near the expected elution volumes for the target 120-subunit complexes and X-ray scattering intensities (grey dots) that match well with profiles calculated from the design models (green). Alternative configurations of the designs, generated by translating and/or rotating the oligomeric building blocks in the design models about their aligned symmetry axes by +/- 10 Å and/or 20 degrees, respectively, generally fit worse with the SAXS data than the original design models (the range of values obtained from fitting the alternative configurations is shown with light blue shading).

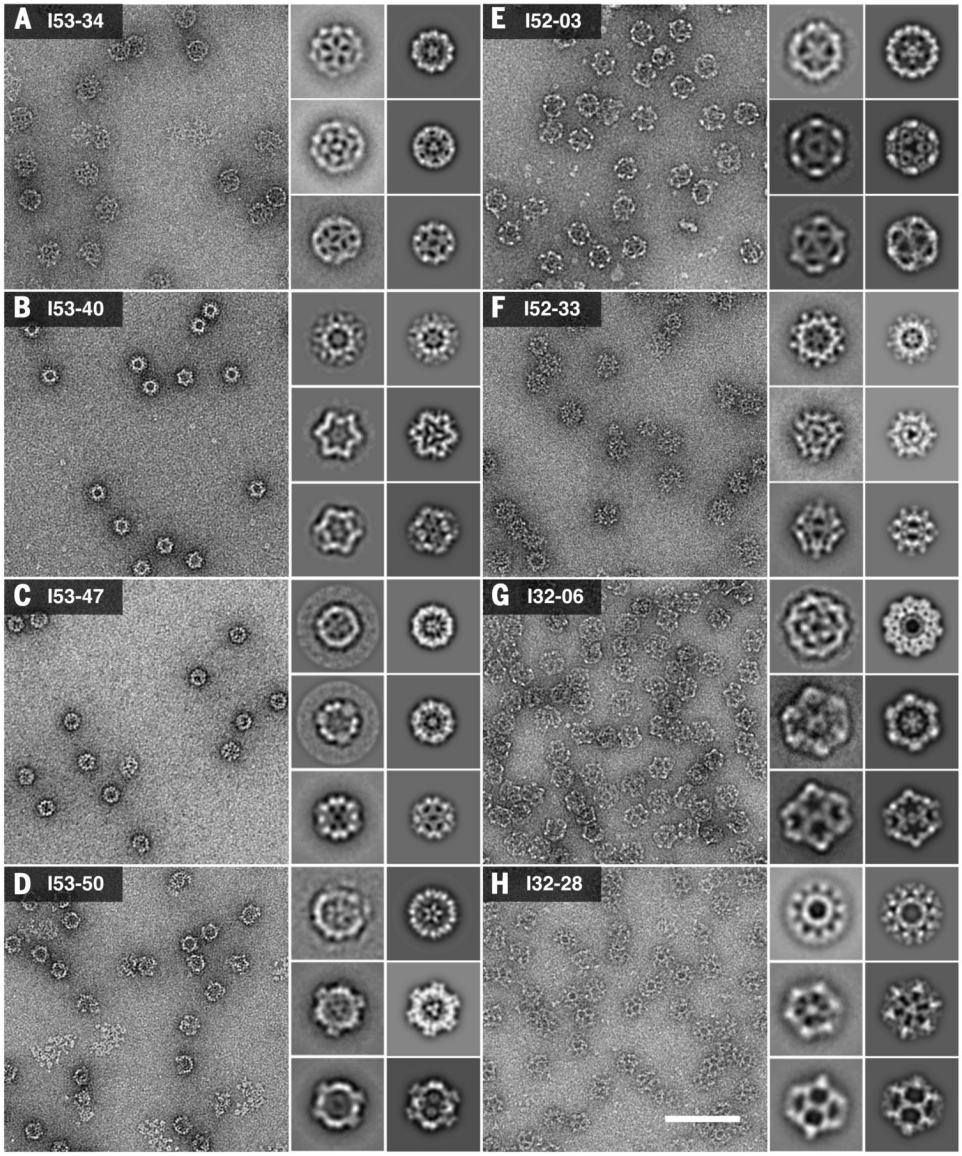


Fig. 5.3. Characterization of the designed materials by electron microscopy.

Left: raw negative stain electron micrographs of co-expressed and purified (**A**) I53-34, (**B**) I53-40, (**C**) I53-47, (**D**) I53-50, (**E**) I52-03, (**F**) I52-33, (**G**) I32-06, and (**H**) I32-28. All raw micrographs shown to scale relative to 30 nm scale bar in panel (**H**). Insets: experimentally computed class averages (roughly corresponding to the five-fold, three-fold, and 2-fold icosahedral symmetry axes; left) along with back projections calculated from the design models (right).

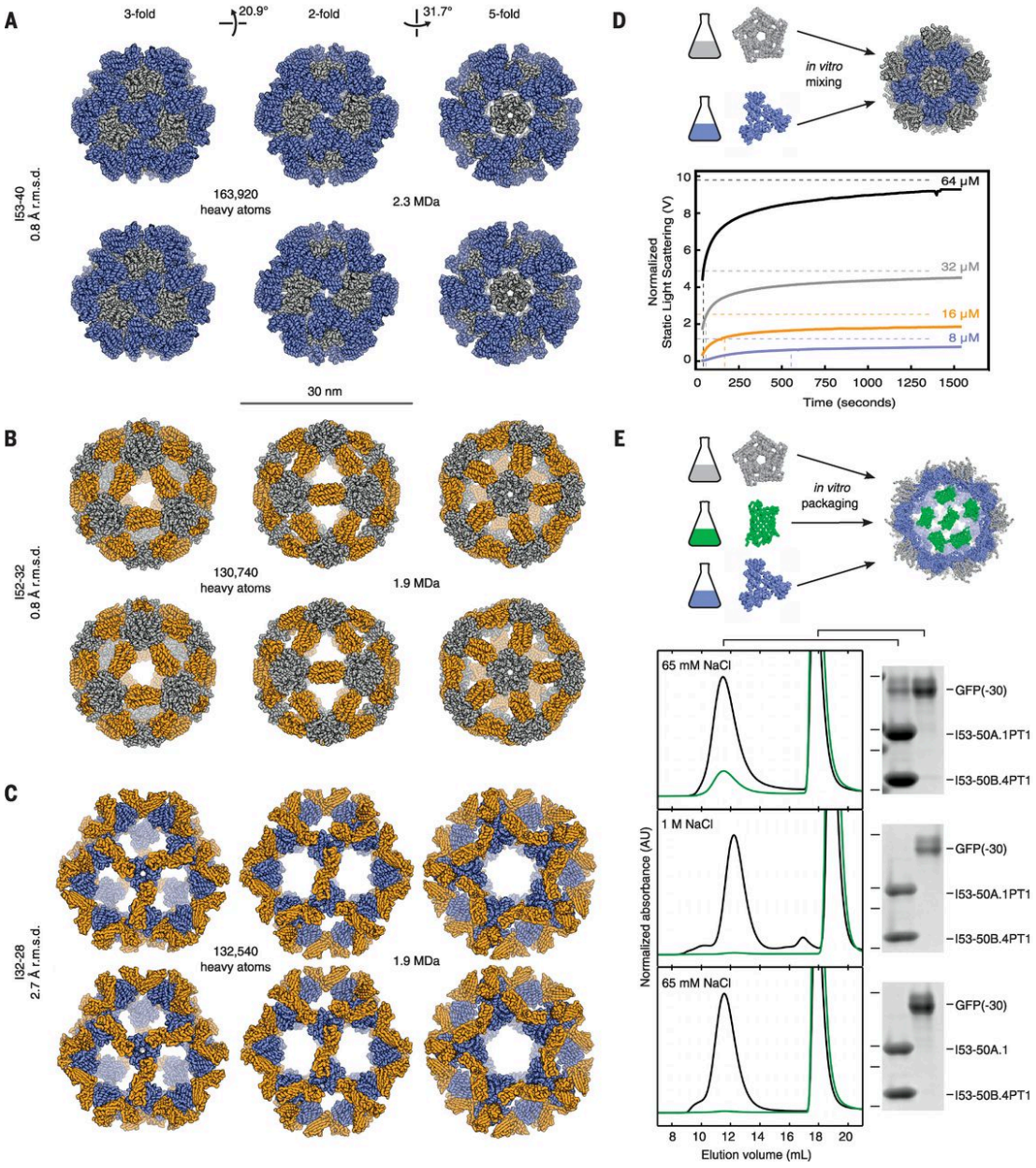


Fig. 5.4. Crystal structures, assembly dynamics, and packaging.

Design models (top) and X-ray crystal structures (bottom) of (**A**) I53-40, (**B**) I52-32, and (**C**) I32-28. Views shown to scale along the 3-fold, 2-fold, and 5-fold icosahedral symmetry axes. Pentamers shown in grey, trimers blue, and dimers orange. R.m.s.d.s are between crystal structures and design models over all backbone atoms in all 120 subunits. (**D**) *In vitro* assembly dynamics of I53-50. (Top) Schematic illustration. (Bottom) Normalized static light scattering

intensity (detector voltage, solid lines) plotted over time after mixing independently expressed and purified variants of the I53-50 trimer and pentamer in a 1:1 molar ratio at final concentrations of 8, 16, 32, or 64 μM each (blue, orange, grey, and black lines, respectively). Intensities measured from SEC-purified assembly at 8, 16, 32, or 64 μM concentrations indicated with dashed horizontal lines and used as the expected endpoint of each assembly reaction. The midpoint of each reaction is marked with a dashed vertical line. (**E**) Encapsulation of supercharged GFP in a positively charged I53-50 variant. (Top) Schematic illustration. (Bottom) Superose 6 chromatograms and SDS-PAGE analysis of packaging/assembly reactions performed in buffer containing: (Top Panel) 65 mM NaCl, (Middle Panel) 1 M NaCl, or (Bottom Panel) 65 mM NaCl with a trimer variant without mutations to positively charged residues. In each case, the same buffer used in the packaging/assembly reaction was also used during SEC. Absorbance measurements at 280 nm (black) and 488 nm (green) are shown. Each SEC chromatogram was normalized relative to the 280 nm peak near 12 mL elution volume. Locations of 37, 25, 20, and 15 kDa molecular weight markers on SDS-PAGE gels are indicated by horizontal lines.

REFERENCES AND NOTES

1. Y. T. Lai, N. P. King, T. O. Yeates, Principles for designing ordered protein assemblies. Trends Cell Biol 22, 653-661 (2012).
2. N. P. King, Y. T. Lai, Practical approaches to designing novel protein assemblies. Curr Opin Struct Biol 23, 632-638 (2013).
3. J. E. Padilla, C. Colovos, T. O. Yeates, Nanohedra: using symmetry to design self assembling protein cages, layers, crystals, and filaments. Proc Natl Acad Sci U S A 98, 2217-2221 (2001).
4. P. Ringler, G. E. Schulz, Self-assembly of proteins into designed networks. Science 302, 106-109 (2003).
5. S. Raman, G. Machaidze, A. Lustig, U. Aebi, P. Burkhard, Structure-based design of peptides that self-assemble into regular polyhedral nanoparticles. Nanomedicine 2, 95-102 (2006).
6. D. Grueninger, N. Treiber, M. O. Ziegler, J. W. Koetter, M. S. Schulze, G. E. Schulz, Designed protein-protein association. Science 319, 206-209 (2008).
7. S. Raman, Machaidze G., Lustig A., Olivieri V., Aebi U., Burkhard P., Design of Peptide Nanoparticles Using Simple Protein Oligomerization Domains. The Open Nanomedicine Journal 2, 15-26 (2009).
8. K. Usui, T. Maki, F. Ito, A. Suenaga, S. Kidoaki, M. Itoh, M. Taiji, T. Matsuda, Y. Hayashizaki, H. Suzuki, Nanoscale elongating control of the self-assembled protein filament with the cysteine-introduced building blocks. Protein Sci 18, 960-969 (2009).
9. E. N. Salgado, R. J. Radford, F. A. Tezcan, Metal-directed protein self-assembly. Acc Chem Res 43, 661-672 (2010).
10. G. Grigoryan, Y. H. Kim, R. Acharya, K. Axelrod, R. M. Jain, L. Willis, M. Drndic, J. M. Kikkawa, W. F. DeGrado, Computational design of virus-like protein assemblies on carbon nanotube surfaces. Science 332, 1071-1076 (2011).
11. J. C. Sinclair, K. M. Davies, C. Venien-Bryan, M. E. Noble, Generation of protein lattices by fusing proteins with matching rotational symmetry. Nat Nanotechnol 6, 558-562 (2011).
12. P. B. Stranges, M. Machius, M. J. Miley, A. Tripathy, B. Kuhlman, Computational design of a symmetric homodimer using beta-strand assembly. Proc Natl Acad Sci U S A 108, 20562-20567 (2011).
13. A. L. Boyle, E. H. Bromley, G. J. Bartlett, R. B. Sessions, T. H. Sharp, C. L. Williams, P. M. Curmi, N. R. Forde, H. Linke, D. N. Woolfson, Squaring the circle in peptide assembly: from fibers to discrete nanostructures by de novo design. J Am Chem Soc 134, 15457-15467 (2012).
14. J. D. Brodin, X. I. Ambroggio, C. Tang, K. N. Parent, T. S. Baker, F. A. Tezcan, Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays. Nat Chem 4, 375-382 (2012).
15. B. S. Der, M. Machius, M. J. Miley, J. L. Mills, T. Szyperski, B. Kuhlman, Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. J Am Chem Soc 134, 375-385 (2012).
16. N. P. King, W. Sheffler, M. R. Sawaya, B. S. Vollmar, J. P. Sumida, I. Andre, T. Gonen, T. O. Yeates, D. Baker, Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science 336, 1171-1174 (2012).

17. Y. T. Lai, D. Cascio, T. O. Yeates, Structure of a 16-nm cage designed by using protein oligomers. Science 336, 1129 (2012).
18. C. J. Lanci, C. M. MacDermaid, S. G. Kang, R. Acharya, B. North, X. Yang, X. J. Qiu, W. F. DeGrado, J. G. Saven, Computational design of a protein crystal. Proc Natl Acad Sci U S A 109, 7304-7309 (2012).
19. J. M. Fletcher, R. L. Harniman, F. R. Barnes, A. L. Boyle, A. Collins, J. Mantell, T. H. Sharp, M. Antognozzi, P. J. Booth, N. Linden, M. J. Miles, R. B. Sessions, P. Verkade, D. N. Woolfson, Self-assembling cages from coiled-coil peptide modules. Science 340, 595-599 (2013).
20. N. P. King, J. B. Bale, W. Sheffler, D. E. McNamara, S. Gonen, T. Gonen, T. O. Yeates, D. Baker, Accurate design of co-assembling multi-component protein nanomaterials. Nature 510, 103-108 (2014).
21. Y. T. Lai, E. Reading, G. L. Hura, K. L. Tsai, A. Laganowsky, F. J. Asturias, J. A. Tainer, C. V. Robinson, T. O. Yeates, Structure of a designed protein cage that self-assembles into a highly porous cube. Nat Chem 6, 1065-1071 (2014).
22. A. R. Voet, H. Noguchi, C. Addy, D. Simoncini, D. Terada, S. Unzai, S. Y. Park, K. Y. Zhang, J. R. Tame, Computational design of a self-assembling symmetrical beta-propeller protein. Proc Natl Acad Sci U S A 111, 15102-15107 (2014).
23. S. Gonen, F. DiMaio, T. Gonen, D. Baker, Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. Science 348, 1365-1368 (2015).
24. Y. Hsia, J. B. Bale, S. Gonen, D. Shi, W. Sheffler, K. K. Fong, U. Nattermann, C. Xu, P. Huang, R. Ravichandran, S. Yi, T. N. Davis, T. Gonen, N. P. King, D. Baker, Design of a hyperstable 60-subunit protein icosahedron. Nature, in press (included as reference material).
25. D. S. Goodsell, A. J. Olson, Structural symmetry and protein function. Annu Rev Biophys Biomol Struct 29, 105-153 (2000).
26. J. B. Bale, R. U. Park, Y. Liu, S. Gonen, T. Gonen, D. Cascio, N. P. King, T. O. Yeates, D. Baker, Structure of a designed tetrahedral protein assembly variant engineered to have improved soluble expression. Protein Sci 24, 1695-1701 (2015).
27. D. L. Caspar, A. Klug, Physical principles in the construction of regular viruses. Cold Spring Harb Symp Quant Biol 27, 1-24 (1962).
28. R. Zandi, D. Reguera, R. F. Bruinsma, W. M. Gelbart, J. Rudnick, Origin of icosahedral symmetry in viruses. Proc Natl Acad Sci U S A 101, 15556-15560 (2004).
29. M. C. Lawrence, P. M. Colman, Shape complementarity at protein/protein interfaces. J Mol Biol 234, 946-950 (1993).
30. A. Zlotnick, J. M. Johnson, P. W. Wingfield, S. J. Stahl, D. Endres, A theoretical model successfully identifies features of hepatitis B virus capsid assembly. Biochemistry 38, 14644-14652 (1999).
31. A. Zlotnick, R. Aldrich, J. M. Johnson, P. Ceres, M. J. Young, Mechanism of capsid assembly for an icosahedral plant virus. Virology 277, 450-456 (2000).
32. M. S. Lawrence, K. J. Phillips, D. R. Liu, Supercharging proteins can impart unusual resilience. J Am Chem Soc 129, 10110-10112 (2007).
33. B. Worsdorfer, K. J. Woycechowsky, D. Hilvert, Directed evolution of a protein container. Science 331, 589-592 (2011).
34. R. Zschoche, D. Hilvert, Diffusion-Limited Cargo Loading of an Engineered Protein Container. J Am Chem Soc 137, 16121-16132 (2015).

35. T. Lin, Z. Chen, R. Usha, C. V. Stauffacher, J. B. Dai, T. Schmidt, J. E. Johnson, The refined crystal structure of cowpea mosaic virus at 2.8 A resolution. Virology 265, 20-34 (1999).
36. T. Lin, A. J. Clark, Z. Chen, M. Shanks, J. B. Dai, Y. Li, T. Schmidt, P. Oxelfelt, G. P. Lomonossoff, J. E. Johnson, Structural fingerprinting: subgrouping of comoviruses by structural studies of red clover mottle virus to 2.4-A resolution and comparisons with other comoviruses. J Virol 74, 493-504 (2000).
37. The PyMOL Molecular Graphics System v. 1.5. (Schrödinger, LLC 2012).
38. E. Krissinel, K. Henrick, Inference of macromolecular assemblies from crystalline state. J Mol Biol 372, 774-797 (2007).
39. S. J. Fleishman, A. Leaver-Fay, J. E. Corn, E. M. Strauch, S. D. Khare, N. Koga, J. Ashworth, P. Murphy, F. Richter, G. Lemmon, J. Meiler, D. Baker, RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. PLoS One 6, e20161 (2011).
40. F. DiMaio, A. Leaver-Fay, P. Bradley, D. Baker, I. Andre, Modeling symmetric macromolecular structures in Rosetta3. PLoS One 6, e20450 (2011).
41. A. Leaver-Fay, M. J. O'Meara, M. Tyka, R. Jacak, Y. Song, E. H. Kellogg, J. Thompson, I. W. Davis, R. A. Pache, S. Lyskov, J. J. Gray, T. Kortemme, J. S. Richardson, J. J. Havranek, J. Snoeyink, D. Baker, B. Kuhlman, Scientific benchmarks for guiding macromolecular energy function improvement. Methods Enzymol 523, 109-143 (2013).
42. M. J. O'Meara, A. Leaver-Fay, M. D. Tyka, A. Stein, K. Houlihan, F. DiMaio, P. Bradley, T. Kortemme, D. Baker, J. Snoeyink, B. Kuhlman, Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. J Chem Theory Comput 11, 609-622 (2015).
43. B. Kuhlman, D. Baker, Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 97, 10383-10388 (2000).
44. W. Sheffler, D. Baker, RosettaHoles2: a volumetric packing measure for protein structure refinement and validation. Protein Sci 19, 1991-1995 (2010).
45. S. J. Fleishman, T. A. Whitehead, E. M. Strauch, J. E. Corn, S. Qin, H. X. Zhou, J. C. Mitchell, O. N. Demerdash, M. Takeda-Shitaka, G. Terashi, I. H. Moal, X. Li, P. A. Bates, M. Zacharias, H. Park, J. S. Ko, H. Lee, C. Seok, T. Bourquard, J. Bernauer, A. Poupon, J. Aze, S. Soner, S. K. Ovali, P. Ozbek, N. B. Tal, T. Haliloglu, H. Hwang, T. Vreven, B. G. Pierce, Z. Weng, L. Perez-Cano, C. Pons, J. Fernandez-Recio, F. Jiang, F. Yang, X. Gong, L. Cao, X. Xu, B. Liu, P. Wang, C. Li, C. Wang, C. H. Robert, M. Guharoy, S. Liu, Y. Huang, L. Li, D. Guo, Y. Chen, Y. Xiao, N. London, Z. Itzhaki, O. Schueler-Furman, Y. Inbar, V. Potapov, M. Cohen, G. Schreiber, Y. Tsuchiya, E. Kanamori, D. M. Standley, H. Nakamura, K. Kinoshita, C. M. Driggers, R. G. Hall, J. L. Morgan, V. L. Hsu, J. Zhan, Y. Yang, Y. Zhou, P. L. Kastritis, A. M. Bonvin, W. Zhang, C. J. Camacho, K. P. Kilambi, A. Sircar, J. J. Gray, M. Ohue, N. Uchikoga, Y. Matsuzaki, T. Ishida, Y. Akiyama, R. Khashan, S. Bush, D. Fouches, A. Tropsha, J. Esquivel-Rodriguez, D. Kihara, P. B. Stranges, R. Jacak, B. Kuhlman, S. Y. Huang, X. Zou, S. J. Wodak, J. Janin, D. Baker, Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol 414, 289-302 (2011).
46. G. L. Hura, A. L. Menon, M. Hammel, R. P. Rambo, F. L. Poole, 2nd, S. E. Tsutakawa, F. E. Jenney, Jr., S. Classen, K. A. Frankel, R. C. Hopkins, S. J. Yang, J. W. Scott, B. D. Dillard, M. W. Adams, J. A. Tainer, Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nat Methods 6, 606-612 (2009).

47. P. S. Huang, Y. E. Ban, F. Richter, I. Andre, R. Vernon, W. R. Schief, D. Baker, RosettaRemodel: a generalized framework for flexible backbone protein design. PLoS One 6, e24109 (2011).
48. D. Schneidman-Duhovny, M. Hammel, J. A. Tainer, A. Sali, Accurate SAXS profile computation and its assessment by contrast variation experiments. Biophys J 105, 962-974 (2013).
49. D. Schneidman-Duhovny, M. Hammel, A. Sali, FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic Acids Res 38, W540-544 (2010).
50. B. L. Nannenga, M. G. Iadanza, B. S. Vollmar, T. Gonen, Overview of electron crystallography of membrane proteins: crystallization and screening strategies using negative stain electron microscopy. Curr Protoc Protein Sci Chapter 17, Unit17 15 (2013).
51. G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees, S. J. Ludtke, EMAN2: an extensible image processing suite for electron microscopy. J Struct Biol 157, 38-46 (2007).
52. M. van Heel, G. Harauz, E. V. Orlova, R. Schmidt, M. Schatz, A new generation of the IMAGIC image processing system. J Struct Biol 116, 17-24 (1996).
53. J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J. Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, A. Cardona, Fiji: an open-source platform for biological-image analysis. Nat Methods 9, 676-682 (2012).
54. W. Kabsch, Xds. Acta Crystallogr D Biol Crystallogr 66, 125-132 (2010).
55. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, Phaser crystallographic software. J Appl Crystallogr 40, 658-674 (2007).
56. A. Vagin, A. Teplyakov, MOLREP: an Automated Program for Molecular Replacement. Journal of Applied Crystallography 30, 1022-1025 (1997).
57. P. D. Adams, P. V. Afonine, G. Bunkoczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr 66, 213-221 (2010).
58. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. Acta Crystallogr D Biol Crystallogr 66, 486-501 (2010).
59. C. Jackel, J. D. Bloom, P. Kast, F. H. Arnold, D. Hilvert, Consensus protein design without phylogenetic bias. J Mol Biol 399, 541-546 (2010).

APPENDIX

Supplementary Materials and Methods

Scaffold preparation

Input homodimeric, homotrimeric, and homopentameric scaffolds for design were derived from crystal structures deposited in the Protein Data Bank (PDB, http://www.rcsb.org/pdb/) and from crystal structures and design models of a small set of de novo designed homooligomeric structures not yet deposited in the PDB (data unpublished). Coordinates of all the biological assemblies in the PDB (ftp://ftp.wwpdb.org/pub/pdb/data/biounit/coordinates) and the de novo designed oligomers were processed as described below to standardize them for input into Rosetta and detect whether or not they possessed C2, C3, or C5 symmetry. Several structures possessing dihedral symmetry were included as scaffolds as well, with the intention that the unwanted 2-fold interfaces would be disrupted during the design process.

Assemblies containing multiple models were converted to a single model containing all chains in all the models. Alternative side chains and HETATM records were removed, selenomethionines replaced with methionines, and the chain with the lowest average r.m.s.d. (as calculated by the super command in PyMOL (37)) to all other chains was selected to be the input chain for design. Residues with missing main chain atoms were removed from the design input chain and its residues renumbered starting from 1. Copies of the design input chain were iteratively superimposed onto the other chains in the assembly until superimposed onto all other chains and an attempt made with each iteration to detect a rotational axis of symmetry. Assemblies were discarded that were found not to possess cyclic symmetry or to be too asymmetric, as assessed by the dispersion of symmetry axes implied by each tuple of symmetrically related atoms. Each passing assembly was assigned the highest cyclic symmetry detected and its symmetry axis aligned along the vector [0,0,1] and its center of mass translated to the origin. The resulting PDB-derived structures were then filtered according to the criteria detailed below; wherein the stringency of the criteria was adjusted relative to the number of structures available for each type of cyclic symmetry (fewer trimers are available than dimers, and fewer pentamers are available than trimers or dimers, so the criteria used to select pentamers was the least stringent, followed by the criteria used to select trimers). The de novo designed scaffolds were not subjected to these additional selection steps.

PDB structures determined to possess C2 or C3 symmetry were cross-validated with the PISA database (38) by filtering out any that did not match the symmetry detected by PISA or meet the default PISA criteria for dissociation energy, accessible surface area, buried surface area, percent buried surface area, and average chain length. For dimers, the resulting PDB IDs were input into the advanced search tool in the PDB to selected proteins clustered at 90% sequence identity with: 1) X-ray resolution less than 2 Å, 2) chain lengths between 125 and 250 amino acids, and 3) E. coli as the host organism for protein expression. For trimers, the resulting PDB IDs were input into the advanced search tool and clustered at 90% sequence identity with: 1) X-ray resolution less than 2.5 Å, chain lengths between 75 and 250 amino acids, and 3) E. coli as the host organism for expression. For structures determined to possess C5 symmetry, the PDB IDs were input into the advance search tool and clustered at 90% sequence identity with: 1) X-ray resolution less than 3.5 Å and 2) chain lengths between 40 and 400 amino acids. The C2, C3, and C5 scaffolds passing these automated filtering criteria were also manually inspected in PyMOL (37) with regard to the quality of the secondary structure elements available on the

surface of the scaffolds. Lastly, PDB ID 1jml, a crystal structure of a de novo designed dimer, was added to the C2 scaffolds despite not passing our automated filter criteria, as well as three additional C3 scaffolds that did not pass the automated filter criteria, but were deemed scaffolds of high interest due to their structural roles in biology: PDB IDs 3i87, 4fay, and 1gcm.
The resulting homopentameric scaffold set is listed in Table 5.S1, homotrimeric scaffold set in Table 5.S2, and homodimeric scaffold set in Table 5.S3.

Symmetric docking

Symmetric docking was carried out as described previously (20), with the following changes to the score function and criteria used to the select configurations for design. The score function used to measure the suitability of a given configuration for design (i.e., the "designability" of the configuration) was modified to favor protein backbone configurations matching those of commonly observed interaction motifs found in high-resolution crystal structures in the PDB. By biasing the docked configurations in such a manner, we hoped to increase the percentage of designs passing our criteria for experimental testing and thereby improve the efficiency of the design pipeline. Up to 50 top scoring configurations were output for each pair of scaffolds and filtered according to the following criteria. I53 configurations were removed if they had fewer than 35 or greater than 70 contacting residues (residues with β-carbon atoms within 12 Å) at the interface and if they had a score less than 180. The number of contacting residues was used as a proxy for interface size and the range of 35 to 70 contacting residues was chosen in order to select designs with similar interface sizes to the previously successful two-component tetrahedra (20). I52 configurations were removed that had a score less than 200, less than 35 or greater than 70 contacting residues, fewer than 180 identified interaction motifs, and an average score per contacting residue less than 4.5. I32 configurations were removed that had a score less than 220, less than 35 or greater than 70 contacting residues, fewer than 200 identified interaction motifs, and an average score per contacting residue less than 5. Of the remaining I52 and I32 configurations, the top 5 scoring designs from each scaffold pair were selected for design. In total, 66,115 docked configurations of type I53, 35,468 of type I52, and 161,007 of type I32 were carried forward for design, as described below (Fig. S2).

Protein-protein interface design

The design protocols used in the present study were based on our original two-component protein-protein interface design methods (20) implemented within the RosettaScripts framework (39), with a number of modifications aimed at improving success rates and decreasing the need for manual intervention. The process was split into three different stages: I) interface design, II) automated reversion, and III) resfile-based refinement. The protocols used in each stage require as input a symmetry definition file and a PDB file containing a single subunit of both scaffold proteins (produced by concatenating the two scaffold protein PDB files used as input for docking, changing the chain of the second subunit to "B", and renumbering the residues to continue across the two chains).

All design calculations were performed on the two independent subunits and propagated symmetrically (40). Except where specified otherwise, all calculations were performed with the standard talaris2013 scorefunction (41, 42). Throughout the design process, interface residues were selected that satisfied the following three criteria: 1) the residue's beta carbon atom (alpha carbon in the case of glycine) was within 10 Å of a beta carbon (alpha carbon in the case of glycine) in a different oligomeric building block, 2) the residue possessed non-zero solvent accessible surface area (SASA, calculated with a 2.2 Å radius probe) when the protein subunits

were in the unbound state, and 3), with the exception of residues with high Lennard-Jones repulsive scores (fa_rep), the residue did not make contacts (any heavy atoms within 5 Å) with other subunits in the same oligomeric building block. Residues matching all three criteria were considered designable. An expanded set of residues, in which criteria 3 was not enforced, was also used in certain portions of the protocols. Throughout the following text, we refer to the residues fulfilling criteria 1, 2, and 3 as "design positions" and those fulfilling criteria 1 and 2 as "interface positions"; all design positions were therefore also interface positions, but not all interface positions were design positions. These positions were updated at multiple points throughout stages I through III; appending any positions that newly satisfied the selection criteria to the previously defined sets. All residues not in the selected sets or otherwise specified as designable via user-defined resfiles, remained fixed throughout the design process. During certain stages of the design process, the design positions were further divided into three categories based on their SASA (calculated with a 2.2 Å radius probe with the protein subunits in the bound, icosahedral assembly state); positions with SASA less than 25 Å2 were designated as "core", positions with SASA between 25 and 35 Å2 designated as "boundary", and positions with SASA greater than 35 Å2 were designated as "surface" positions. Except where specified otherwise, the chi2 angle for aromatic side chains being repacked or designed was restricted to between 70 and 110 degrees during all RosettaDesign (43) steps in all stages.

Stage I: Interface design. Ten independent design trajectories were run for each of the selected I53, I52, and I32 docked configurations (vide supra). In order to more finely sample nearby configurations during design, the translational and rotational rigid body DOFs of each component were perturbed slightly at the start of each trajectory by sampling randomly from Gaussian distributions with standard deviations of 0.25-0.50 Å and 0.5-1.0 degrees, respectively (0.25 Å and 0.5 degrees standard deviations were used with the I52 and I32 designs and 0.50 Å and 1.0 degree standard deviations were used with the I53 designs). Perturbed configurations yielding less than 30 or more than 90 design positions, or more than 4 clashing backbone atoms (distance between backbone amide nitrogen and carbonyl oxygen atoms <= 2.6 Å; distance between all other backbone/beta carbon atom pairs <= 3.5 Å), were removed from further consideration. In addition, during the I52 and I32 design process, perturbed configurations were removed from further consideration if they did not possess at least one interface position that contacted >= 5 other interface positions across the design interface, wherein each of those contacting interface positions also contacted >=4 interface positions across the design interface (contact being defined as a distance between beta carbon (alpha carbon in the case of glycine) atoms of <= 8.9 Å). This additional filter, henceforth referred to as the two-way neighbor count, was intended as a sequence-independent measure of interface interconnectivity in order to favor configurations with at least one large, contiguous, connected interface patch likely to yield a substantial hydrophobic core upon design.

An initial round of design was carried out on the remaining configurations as follows. For the I52 and I32 designs, the RosettaDesign algorithm was used to sample the amino acid identities and side chain conformations of the core design positions using: 1) a version of the talaris2013 scorefunction in which the Lennard-Jones repulsive term (fa_rep) was down-weighted to favor tightly packed interfaces and 2) a version of the Rosetta side chain rotamer library that was modified to include the side chain conformations of the interaction motifs identified during docking. The scorefunction was then set back to the standard talaris2013 scorefunction and the Rosetta energy minimized through a series of small changes to the design position side chain conformations (i.e., the side chains were "minimized") (40). For the I53, I52,

and I32 designs, the RosettaDesign algorithm was then employed to sample the amino acid identities and side chain conformations at all design positions, followed by one round of minimization of the design position side chains and the rigid body DOFs, repacking of the design position side chains (i.e., the conformations of the side chains were sampled using the RosettaDesign algorithm), and minimization of the design position side chains and the rigid body DOFs. A reduced amino acid set was employed during this design step such that only the current amino acid, the wild-type amino acid or mutation to the following amino acids were allowed: AFILMSTV (F allowed only in I52 and I32 designs).

Designs with contacting interface areas between 1,100 and 2,000 Å2 were carried forward to another round of design, divided into three parts according to each position's SASA as follows. First, the RosettaDesign algorithm was used to sample the amino acid identities and side chain conformations of the core design positions with a reduced amino acid set such that only the current amino acid, the wild-type amino acid or mutation to the following amino acids were allowed: AFILMSTV (one-letter codes; F allowed only in I52 and I32 designs). Second, the RosettaDesign algorithm was used to sample the amino acid identities and side chain conformations of the boundary design positions with a reduced amino acid set such that only the current amino acid, the wild-type amino acid or mutation to the following amino acids were allowed: ADEHIKLMNQRSTV. Third, the RosettaDesign algorithm was used to sample the amino acid identities and side chain conformations of the surface design positions with a reduced amino acid set such that only the current amino acid, the wild-type amino acid and mutation to the following amino acids were allowed: ACDEGHKNPQRST (a modified version of the Rosetta scorefunction, with the fa_elec score term up-weighted to favor low energy electrostatic interactions, was used during this step in the I52 and I32 designs). The side chain conformations of the inteface positions and the rigid body DOFs were minimized, interface position side chains repacked, and the side chain conformations of the inteface positions and the rigid body DOFs minimized one more time.

The resulting I53 designs were filtered at the end of Stage I to remove those in which the contacting interface area was less than 1,100 Å2, the shape complementarity of the designed interface was less than 0.65 (29) or the predicted binding energy of the designed interface was greater than -15 Rosetta energy units (REUs). The I52 designs were filtered to remove those that had a contacting interface area less than 1,100 Å2, a shape complementarity score less than 0.62, predicted binding energy greater than -20 REUs, 55 or greater mutations, more than 2 buried unsatisfied hydrogen bonds at the designed interface, an average Rosetta energy greater than -1.6 for the interface positions, more than 3 methionines, 3 phenylalanines, or 9 alanines at the designed positions, more than 3 mutated residues with fa_rep scores greater than 2.5 REU, more than 3 positions in which the hbond_bb_sc score was more than 0.5 REU higher in the design than the native scaffold, more than 2 positions in which the fa_atr score was more than 4.6 REU higher in the design than the native scaffold, 0 positions passing the two-way neighbor count filter, or an interface atomic contact count score less than 45 (defined as the number of side chain carbon atom pairs within 4.5 Å of eachother across the designed interface, where pairs were only counted between the following amino acid types: ACFGILMPTVWY). The I32 designs were filtered to remove those that had a contacting interface area less than 1,100 Å2, a shape complementarity score less than 0.65, predicted binding energy greater than -20 REUs, more than 50 mutations, more than 2 buried unsatisfied hydrogen bonds at the designed interface, an average Rosetta energy greater than -1.6 for the interface positions, more than 3 methionines, 3 phenylalanines, or 9 alanines at the designed positions, more than 3 mutated residues with fa_rep

scores greater than 2.5 REU, more than 3 positions in which the hbond_bb_sc score was more than 0.5 REU higher in the design than the native scaffold, more than 2 positions in which the fa_atr score was more than 4.6 REU higher in the design than the native scaffold, 0 positions passing the two-way neighbor count filter, or an interface atomic contact count score less than 50. For each design, the values of the final rigid body DOFs were output to a scorefile along with the filter metric values and the standard talaris2013 score terms, and a standard Rosetta resfile was output containing each of the design positions and their amino acid identities. The resulting 4,219 I53, 1,760 I52, and 6,584 I32 designs were subjected to the automated reversion protocol described in Stage II below (Fig. S2).

Stage II: Automated reversion. In order minimize the number of mutations made to the scaffold proteins and reduce the amount of manual refinement required in Stage III, the goal of the second stage in the design process was to identify and attempt to revert, via an automated computational protocol, mutated residues predicted not to be critical for assembly and/or which resulted in substantial losses to: 1) core packing interactions within the native protein scaffolds or 2) hydrogen bonding interactions between backbone and side chain atoms within the native protein scaffolds.

The first step in the automated reversion process was to regenerate each design from the input scaffolds using the rigid body DOFs and resfiles output from stage I. For the I53 designs, regeneration was accomplished as follows: 1) the rigid body DOFs were used to reposition the subunits in the assembled state, 2) the resfiles were used as input to the RosettaDesign algorithm to reintroduce the initial design mutations, 3) design position side chains were subjected to one round of minimization, repacking, and minimization, 4) the core positions were repacked, followed by the boundary, and then the surface positions, and 5) interface position side chains were subjected to one round of minimization, repacking, and minimization. For the I52 and I32 designs, regeneration was accomplished as follows: 1) the rigid body DOFs were used to reposition the subunits in the assembled state, 2) the resfiles were used as input to the RosettaDesign algorithm to reintroduce the initial design mutations, 3) the surface positions were repacked using a version of the talaris2013 score function with the fa_elec term up-weighted to favor low energy electrostatic interactions, 4) interface position side chains were subjected to one round of minimization, repacking, and minimization, and 5) interface position side chains and rigid body DOFs were minimized, interface position side chains repacked, and then interface position side chains and rigid body DOFs minimized one more time.

An initial round of greedy optimization was then used to revert mutations to the native amino identities as follows. During the first stage of the algorithm, reversions were tested individually and ranked by the change in shape complementarity, if the individual reversions did not: 1) increase the total score of the reverted position by more than 0.1 REUs relative to the native scaffold, 2) increase the predicted binding energy by more than 1.5 REUs, 3) increase the number of buried unsatisfied hydrogen bonds at the interface, or 4) decrease the shape complementarity of the interface by more than 0.02. During the second stage of the algorithm, reversions that passed the first stage were combined one at a time proceeding from the best ranked to the worst ranked individual reversions, only accepting those that still passed the four criteria above in the context of all previously accepted reversions and which also did not cause the predicted binding energy to be greater than -15 REU or the shape complementarity to be less than 0.60. During this combining stage, the reference structure for measuring the change in shape complementarity and predicted binding energy was reset after each accepted mutation. This portion of the automated reversion process was split into two parts, reversion of the core

and boundary positions, followed by reversion of the surface positions, wherein reversions to the following amino acid identities were not allowed at surface positions: FGILMPVWY (alanine reversions were also disallowed at surface positions in the I52 and I32 designs).

Following this initial round of greedy optimization-based reversion, a second round was carried out focused on reverting remaining mutations that caused significant losses of native hydrogen bonding between backbone and side chain atoms. For each position in which the hbond_bb_sc score was more than 0.5 REU higher in the design than the native scaffolds, reversions were tested individually and ranked by the change in shape complementarity, if the individual reversions improved the hbond_bb_sc score by more than 0.4 REU and did not: 1) increase the total score of the reverted position by more than 0.1 REUs relative to the native scaffold, 2) increase the predicted binding energy by more than 3.0 REUs, 3) increase the number of buried unsatisfied hydrogen bonds at the interface, or 4) decrease the shape complementarity of the interface by more than 0.03. Reversions that passed the first stage were combined one at a time proceeding from the best ranked to the worst ranked individual reversions, only accepting those that still passed the above criteria in the context of all previously accepted reversions and which also did not cause the predicted binding energy to be greater than -15 REU or the shape complementarity to be less than 0.60. During this combining stage, the reference structure for measuring the change in shape complementarity, predicted binding energy, and change in hbond_bb_sc score was reset after each accepted mutation.

Following the second round of greedy optimization-based reversion, a third and final round was carried out focused on reverting remaining mutations that caused significant losses of native packing interactions. For each position in which the fa_atr score was more than 4.6 REU higher in the design than the native scaffold, reversions were tested individually and ranked by the change in shape complementarity, if the individual reversions improved the fa_atr score by more than 3.8 REU and did not: 1) increase the total score of the reverted position by more than 0.1 REUs relative to the native scaffold, 2) increase the predicted binding energy by more than 3.0 REUs, 3) increase the number of buried unsatisfied hydrogen bonds at the interface, or 4) decrease the shape complementarity of the interface by more than 0.03. Reversions that passed the first stage were combined one at a time proceeding from the best ranked to the worst ranked individual reversions, only accepting those that still passed the above criteria in the context of all previously accepted reversions and which also did not cause the predicted binding energy to be greater than -15 REU or the shape complementarity to be less than 0.60. During this combining stage, the reference structure for measuring the change in shape complementarity, predicted binding energy, and change in fa_atr score was reset after each accepted mutation.

During all three stages of greedy optimization, all interface positions were subjected to one round of minimization, repacking, and minimization prior to evaluating the effects of each reversion. Following the last round of greedy optimization, interface position side chains and rigid body DOFs were minimized, interface position side chains repacked, and then interface position side chains and rigid body DOFs minimized one more time. During the second and third stage, the side chain conformation of each reverted position was restricted to the conformation present in the native scaffold rather than Rosetta's standard side chain conformation library.

In addition to the standard Rosetta scores, the following metrics were used to assess the quality of the designs resulting from the automated reversion protocol: 1) the contacting interface area, 2) the shape complementarity of the design interface, 3) the predicted binding energy of the designed interface, 4) the number of mutations, 5) the number of buried unsatisfied hydrogen

bonds at the designed interface, 6) the average Rosetta energy of the interface positions, 7) the number of methionines, phenylalanines, or alanines at the design positions, 8) the number of mutated residues with fa_rep scores greater than 2.5 REU, 9) the number of positions for which the hbond_bb_sc score was more than 0.5 REU higher in the design than the native scaffold, 10) the number of positions for which the fa_atr score was more than 4.6 REU higher in the design than the native scaffold, 11) the number of positions passing the two-way neighbor count filter, 12) the interface atomic contact count score, 13) the RosettaHoles (44) packing score, 14) and the average degree of connectivity (45) of the interface positions within the context of the unassembled oligomeric building blocks. These metrics, in combination with visual inspection in PyMOL (37), were used to select 88 I53, 57 I52, and 93 I32 designs subsequently subjected to resfile-based refinement as outline in Stage III below (Fig. S2).

Stage III: Resfile-based refinement. The final stage of the design process involved one or more cycles of resfile-based redesign with user-guided mutations. In each iteration of the process, a combination of visual inspection and analysis of the design metrics was used to generate modified resfiles for each design containing a small number of user-defined mutations relative to the resfiles output from Stage II. The starting configuration for each redesign was generated from the two input scaffolds using the rigid body DOFs output from Stage II and then RosettaDesign was applied with the residue identities and side-chain packing behavior specified in the input resfile. For the I52 and I32 designs, interface positions with SASA greater than 35 Å2 were repacked using a version of the talaris2013 score function with the fa_elec term up-weighted, and then the interface position side chains were minimized, repacked, and minimized. For all designs, the symmetric rigid body DOFs and the side chains specified in the input resfile were then minimized, side chains repacked, and minimized prior to calculating the full suite of design metrics. This process was iterated until designs were obtained which were deemed suitable for experimental testing or no longer worth pursuing. In total, 71 I53, 44 I52, and 68 I32 designs were carried forward for experimental characterization (Fig. S2-5).

Source code, examples, and design models

Source code is freely available to academic users through the Rosetta Commons agreement (http://www.rosettacommons.org). Examples of each stage of the docking and interface design process can be found in the zipped archives docking.zip and design.zip. Design models of I53-34, I53-40, I53-47, I53-50, I53-51, I52-03, I52-32, I52-33, I32-06, I32-10, I32-19, and I32-28 in PDB format, along with corresponding symmetry definition files, are provided in the zipped archive design_models.zip. The files ending in ".pdb.gz" contain the full 120-subunit icosahedral assembly, split into 60 different models for ease of viewing in molecular graphics packages such as PyMOL (37). The files ending in "_asu.pdb" contain the asymmetric unit (i.e., one subunit from each of the two components). The full icosahedral assemblies contained within the files ending in ".pdb.gz" can be viewed in PyMOL by issuing the command "set all_states, on". The full icosahedral assemblies can also be generated in Rosetta using the provided symmetry definition files, as in the following example command-line for I53-50:

path_to_rosetta_score_application_executable –database path_to_rosetta_database –s I53-50_asu.pdb –symmetry_definition I53.sym –out:output

Small-scale expression, purification, and screening

Genes encoding the 71 pairs of I53 sequences were synthesized and cloned into a variant of the pET29b expression vector (Novagen, Inc.) between the NdeI and XhoI endonuclease

restriction sites. Genes encoding the 44 pairs of I52 sequences and 68 pairs of I32 sequences were synthesized and cloned into a variant of the pET28b expression vector (Novagen, Inc.) between the NcoI and XhoI endonuclease restriction sites.

The two protein coding regions in each DNA construct are connected by an intergenic region. The intergenic region in the I53 designs was derived from the pETDuet-1 vector (Novagen, Inc.) and includes a stop codon, T7 promoter/lac operator, and ribosome binding site. The intergenic region in the I52 and I32 designs only includes a stop codon and ribosome binding site. The sequences of the I53, I52 and I32 intergenic regions are as follows:

I53 intergenic region DNA sequence:
5'-TAATGCTTAAGTCGAACAGAAAGTAATCGTATTGTACACGGCCGC
ATAATCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCC
ATCTTAGTATATTAGTTAAGTATAAGAAGGAGATATACTT-3'

I52 intergenic region DNA sequence:
5'-TAAAGAAGGAGATATCAT-3'

I32 intergenic region DNA sequence:
5'-TGAGAAGGAGATATCAT-3'

The constructs for the I53 protein pairs thus possess the following set of elements from 5′ to 3′: NdeI restriction site, upstream open reading frame (ORF), intergenic region, downstream ORF, XhoI restriction site. The constructs for the I52 and I32 protein pairs possess the following set of elements from 5′ to 3′: NcoI restriction site, upstream ORF, intergenic region, downstream ORF, XhoI restriction site. In each case, the upstream ORFs encode components denoted with the suffix "A"; the downstream ORFs encode the "B" components (Table 5.S4). This allows for co-expression of the designed protein pairs in which both the upstream and downstream ORFs have their own ribosome binding site, and in the case of the I53 designs, both ORFs also have their own T7 promoter/lac operator.

For purification purposes, each co-expression construct included a 6x-histidine tag (HHHHHH) appended to the N or C terminus of one of the two ORFs.
Expression plasmids were transformed into BL21(DE3) E. coli cells. Cells were grown in LB medium supplemented with 50 mg L-1 of kanamycin (Sigma) at 37° C until an OD600 of 0.8 was reached. Protein expression was induced by addition of 0.5 mM isopropyl-thio-β-D-galactopyranoside (Sigma) and allowed to proceed for either 5 h at 22 °C or 3 h at 37 °C before cells were harvested by centrifugation.

The designed proteins were screened for soluble expression and co-purification as follows. Cells collected from 2 to 4 mL expression cultures were lysed by sonication in 25 mM TRIS pH 8.0, 250 mM NaCl, 1 mM DTT, 20 mM imidazole supplemented with 1 mM phenylmethanesulfonyl floride and the lysates cleared by centrifugation. A portion of each soluble fraction was saved for analysis by SDS-PAGE. The remaining portion of each soluble fraction was applied to His MultiTrap FF nickel-coated filter plates pre-equilibrated with 25 mM TRIS pH 8.0, 250 mM NaCl, 1 mM DTT, 20 mM imidazole running buffer (GE Healthcare). Wells were washed three times with running buffer before eluting with 25 mM TRIS pH 8.0, 250 mM NaCl, 1 mM DTT, 400 mM imidazole buffer, followed by a second elution with 100 mM EDTA, pH 8.0. The soluble fractions from the clarified cell lysates and two elution fractions

from each sample were then analyzed by SDS-PAGE to identify those containing species near the expected molecular weight of both protein subunits (indicating co-purification). Elution fractions from those samples were subsequently subjected to native (non-denaturing) PAGE to identify slow migrating species further indicating assembly to higher order materials.

Large-scale expression and purification

Those designs appearing to co-purify and yielding slowly migrating species by native PAGE were subsequently expressed at larger scale (1 to 12 liters of culture) and purified as follows. Cells were lysed by sonication or microfluidization in 25 mM TRIS pH 8.0, 250 mM NaCl, 1 mM DTT, 20 mM imidazole supplemented with 1 mM phenylmethanesulfonyl fluoride, and the lysates were cleared by centrifugation and filtered through 0.22 μm filters (Millipore). The proteins were purified from the filtered supernatants by immobilized metal-affinity chromatography (IMAC) via gravity columns with nickel-NTA resin (Qiagen) or HisTrap HP columns (GE Healthcare) using 25 mM TRIS pH 8.0, 250 mM NaCl, 1 mM DTT, 20 mM imidazole running/wash buffer and 25 mM TRIS pH 8.0, 250 mM NaCl, 1 mM DTT, 500 mM imidazole elution buffer. Elution fractions containing pure protein(s) of interest were pooled, concentrated using centrifugal filter devices (Sartorius Stedim Biotech), and further purified on a Superose 6 10/300 gel filtration column (GE Healthcare) using 25 mM TRIS pH 8.0, 150 mM NaCl, 1 mM DTT as running buffer. Gel filtration fractions containing pure protein in the desired assembly state were pooled, concentrated, and stored at room temperature or 4 °C for subsequent analyses.

Based on initial results from analytical size exclusion chromatography (SEC) and electron microscopy, additional buffer conditions were explored for several of the designs, including I53-34, I53-51, I52-32, I52-33, and I32-19. The analytical SEC, small-angle X-ray scattering (SAXS), and electron microscopy (EM) data reported here are from samples prepared in the buffer conditions described above, except as follows: 1) 5 % (v/v) glycerol was added to all purification buffers for I53-34, 2) 5 % (v/v) glycerol was added and the NaCl concentration increased to 300 mM for all I53-51 purification buffers, and 3) the NaCl concentration was increased to 500 mM for all I52-33 purification buffers used to prepare the samples for SAXS and EM. Although adding 5% (v/v) glycerol to all buffers used for purification of I32-19 moderately improved the results from electron microscopy, it had little effect on analytical SEC or SAXS; all I32-19 data reported here was collected from I32-19 samples purified in the standard buffers (without glycerol), except for the EM data reported in Figure S10c.

Analytical size exclusion chromatography

The analytical SEC data reported here was performed on a Superose 6 10/300 gel filtration column (GE Healtchare) using 25 mM TRIS pH 8.0, 150 mM NaCl, 1 mM DTT as the running buffer, with the following exceptions: 1) 5% (v/v) glycerol was added to the buffer for I53-34 and 2) 5% (v/v) glycerol was added and the NaCl concentration increased to 300 mM in the I53-51 buffer. The designed materials were loaded onto the column with each component present at a subunit concentration of 20-50 μM.

Small-angle X-ray scattering

Scattering measurements were performed at the SIBYLS 12.3.1 beamline at the Advanced Light Source, LBNL, on 20 microliter samples loaded into a helium-purged sample chamber (46). Purified samples were rerun over gel filtration with running buffering containing 25 mM TRIS pH 8.0, 150 mM NaCl, 2% (v/v) glycerol, 1 mM DTT (the glycerol was added to the gel filtration buffer in order to reduce radiation damage during X-ray data collection), with the following exceptions: 1) the I53-34 buffer contained 5% (v/v) glycerol instead of 2% (v/v)

glycerol, 2) the I53-51 buffer contained 5% (v/v) glycerol instead of 2% (v/v) glycerol and 300 mM NaCl instead of 150 mM NaCl, and 3) the I52-33 buffer contained 500 mM NaCl instead of 150 mM NaCl and did not contain glycerol. Data were collected on the resulting gel filtration fractions and on samples concentrated ~2-10x from the gel filtration fractions, with the gel filtration buffer or concentrator eluates used for buffer subtraction. Sequential exposures ranging from 0.5 to 6 seconds were taken at 12 keV, with visual checks for radiation-induced damage to the protein. The data used for fitting were selected for having higher signal to noise ratio and lack of radiation-induced aggregation.

The versions of the design models used for SAXS comparison were obtained as follows: 1) residues were removed that were present in the crystal structures used as the design scaffolds, but absent in the experimentally tested design constructs (e.g. residues from expression tags used during crystallography of the scaffold proteins, but not included in the tested designs) and 2) residues were added that were absent in the crystal structures used as the design scaffolds, but present in the experimentally tested design constructs (e.g. residues from expression tags included in the tested designs, but absent in the crystal structures used as the design scaffolds). For each design, the missing residues were added by running 100 independent Monte Carlo fragment insertion trajectories in RosettaRemodel (47) followed by backbone and side chain torsion angle minimization in RosettaScripts (39). The results were then clustered in Rosetta with a radius of 2 Å and the center of the largest cluster used as the model for comparison with the experimental data. These extra steps of removing or adding residues were necessary to ensure the design models matched the sequences of the experimentally produced constructs on which the SAXS data was collected. The FOXS algorithm (48, 49) was used to calculate scattering profiles from each of these design models and fit them to the experimental data (fitted profiles shown as green lines and experimental data shown as grey dots in Figure 2).
In order to further evaluate how accurately and uniquely these models match the experimental data, each was also compared to a set of alternative models generated by systematically perturbing the radial displacements and rotations of the building blocks in each design by +/- 10 Å and/or 20 degrees, respectively. While maintaining icosahedral symmetry, each component was translated by -10.0, 0.0, or 10.0 Å along its symmetry axis and rotated by -20.0, 0.0, or 20.0 degrees about its symmetry axis. All 81 possible combinations of translations and rotations were sampled for each design. For each configuration without clashing backbones (distance between backbone amide nitrogen and carbonyl oxygen atoms <= 2.6 Å; distance between all other backbone/beta carbon atom pairs <= 3.5 Å) FOXS was then used to calculate a scattering profile from the model and fit it to the experimental data. The range of intensities calculated from each set of alternative models at each scattering vector, q (Å-1), is shown as light blue shaded regions in the SAXS profile plots in Figure 2.

Transmission electron microscopy (TEM)

2-3µl of purified I32-06, I32-19, I32-28, I52-03, I52-32, I52-33, I53-34, I53-40, I53-47, I53-50 and I53-51 samples at concentrations ranging from ~0.01-10 mg mL-1 were applied to glow discharged, carbon coated 200-mesh copper grids (Ted Pella, Inc.), washed with Milli-Q water or appropriate buffer (same as used in purification, with or without added fresh dithiothreitol) or not washed and stained with 0.75% uranyl formate as described previously (50). Grids were visualized for assembly validation and stability and subsequently optimized for data collection by alteration of buffer composition, protein concentration or grid washing. Screening and data collection was performed on a 120 kV Tecnai Spirit T12 transmission electron microscope (FEI, Hillsboro, OR) with a bottom-mount TVIPS F416 CMOS 4k camera.

All the data used to calculate averages were collected at 30,000x magnification at the specimen level.

TEM Data Analysis, Calculation of Averages and Back Projections

Coordinates for 6052 (I32-06), 5851 (I32-28), 6588 (I52-03), 7616 (I52-33), 2974 (I53-34), 5715 (I53-40), 3976 (I53-47) and 6329 (I53-50) unique particles were obtained for averaging using EMAN2 (51). Extracted frames of these particles were used to calculate class averages by refinement in IMAGIC (52) using multiple rounds of MSA (multivariate statistical analysis) and MRA (multi-reference alignment). Representative 5-, 3- and 2-fold back-projections were calculated using the Rosetta design model .pdb files in IMAGIC using the appropriate pixel size and filtered to 30Å resolution. The contrast of all micrographs was enhanced in Fiji (53) and some back projections in the figure were rotated and magnified for clarity.

Crystallization of I53-40, I52-32, and I32-28 icosahedral cages

These three protein assemblies were crystallized using the hanging drop vapor diffusion method at room temperature. The original buffer was 25mM Tris, 150mM NaCl, 1mM DTT at pH 8.0. 2% glycerol was added to stabilize I52-32. I53-40 formed crystals at 10.6 mg/mL protein concentration in a drop containing 0.11 µL of protein solution plus 0.1 □L of well solution (100 mM sodium acetate buffer at pH 4.6, 0.4M calcium chloride, and 3.2% (v/v) isopropanol). The crystals grew over the course of about two to four days and reached dimensions of about 50-100 µm. I53-32 crystallized at 14.5 mg/mL in 0.267 □L of protein solution plus 0.267 □L of well solution containing 0.17M potassium sodium tartrate tetrahydrate. In about two to four days, crystals reached dimensions of about 100-150 µm. Crystals of I32-28 grew in hanging drops containing 0.11 µL of protein solution with 14.5 m/mL I32-28 and 0.1 µL of well solution containing 12.3% PEG 1500, and 18% glycerol. Crystals reached dimensions of about 50-100 µm in about two weeks. For X-ray data collection, crystals were protected from freezing damage using the corresponding well solution augmented with 33% glycerol for I53-30 and I52-32, and 20% glycerol for I32-28.

Crystallographic data collection, structure determination, and refinement

Diffraction data sets were collected at the Advanced Photon Source (APS) beamline 24-ID-C equipped with a Pilatus-6M detector at 100 K. Data for I53-40, I52-32, and I32-28 were collected with detector distances of 550nm, 600nm, and 650nm, respectively. The x-ray wavelength was 0.9791 Å for I53-40 and 0.9792 Å for I52-32 and I32-28. Oscillations were set to 0.25˚ for I53-40 and 0.2° for I52-32 and I32-28. We obtained data at resolutions of 3.7 Å for I53-40, 3.5 Å for I52-32, and 5.6 Å for I32-28.

The general workflow for structure determination and refinement was as following. We first used the XDS/XSCALE package (54) to integrate, reduce, and scale all three sets of data. We then used the PHASER program (55) to determine the structures by molecular replacement (MR) using appropriate search models for each assembly. The MR solutions were further confirmed with self-rotation functions generated by the MOLREP (56) program. Following molecular replacement, atomic models were refined in iterative runs using the PHENIX program (57) followed by assessment using COOT (58) after each run. The limited resolutions did not support the addition of any bound water molecules or ligands during refinement for any of the three structures.

The I53-40 diffraction data set was reduced in space group symmetry I222. Using complete pentamers and complete trimers as independent search ensembles, MR yielded a single solution with a log-likelihood (LLG) value of 15952 after automatic placement of three copies of

the pentamer and five copies of the trimer in the asymmetric unit (AU). The AU of the crystal contains 30 polypeptide chains, which corresponds to a quarter of the complete icosahedral assembly consisting of 120 protein molecules. The crystal has a Matthews coefficient of 4.64 Å3/Da and a 73.5% solvent content. Three of the two-fold symmetry axes of the icosahedron overlap with the principle symmetry axes of the I222 space group. In addition to the three two-fold crystallographic symmetry axes, the self-rotation function shows clear non-crystallographic symmetry axes (12 two-fold axes, 10 three-fold axes, and six five-fold axes) at the expected orientations, consistent with the MR solution. In each refinement run, each polypeptide chain was refined as a rigid body and a single translation libration screw-motion (TLS) group. We implemented the group adp strategy with one B-factor per residue. The 30 chains in AU were divided into two non-crystallographic symmetry (NCS) groups, one group containing all the pentameric subunits and the other all the trimeric subunits. At each step, adjustments were made in COOT (58) when there was support based on Fo-Fc difference maps. The final R and Rfree values were 24.0% and 24.7%.

The I52-32 data set was reduced in space group symmetry R3 (indexed as H3). The ensemble used for the MR search was extracted from the designed model, containing a complete pentamer and one dimeric subunit adjacent to each pentameric subunit. MR yielded a single solution with a log-likelihood (LLG) value of 22297 after automatic placement of four copies of the ensemble in the AU. The AU of the crystal contains 40 polypeptide chains, which corresponds to a third of the complete icosahedral assembly consisting of 120 protein molecules. The crystal has a Matthews coefficient of 6.23 Å3/Da and a 80.2% solvent content. One of the three-fold symmetry axes of the icosahedron overlaps with the crystal c-axis, the principle symmetry axis of the H3 space group. In addition to the three-fold crystallographic symmetry axis, the self-rotation function shows clear non-crystallographic symmetry axes (15 two-fold axes, nine three-fold axes, and six five-fold axes) at the expected orientations, consistent with the MR solution. In each refinement run, each peptide chain was refined as a rigid body and a single TLS group. The 40 chains in the AU were divided into two NCS groups, one group containing all the pentameric subunits and the other all the dimeric subunits. At each step, adjustments were made when there was support based on Fo-Fc difference maps; regions with weak densities and high B-factors were removed during COOT assessment. The final R and Rfree values were 22.7% and 23.3%.

The I32-28 data set was reduced in space group symmetry R3 (indexed as H3). The ensemble used for MR was extracted from the designed model. It was composed of five trimeric subunits and five adjacent dimeric subunits located around a five-fold asymmetry axis. MR yielded a solution with a log-likelihood (LLG) value of 3230 after automatic placement of four copies of the ensemble in the AU. The AU of the crystal contains 40 polypeptide chains, which corresponds to a third of the complete icosahedral assembly consisting of 120 protein molecules. The crystal has a Matthews coefficient of 7.29 Å3/Da and a 83.1% solvent content. One of the three-fold symmetry axes of the icosahedron falls on the crystal c-axis, the principle symmetry axis of the H3 space group. In addition to the three-fold crystallographic symmetry axis, the self-rotation function shows clear non-crystallographic symmetry axes (15 two-fold axes, nine three-fold axes, and six five-fold axes) at the expected orientations, consistent with the MR solution. In each refinement run, each peptide chain was refined as a rigid body and a single TLS group. We implemented the group adp strategy with one B-factor per residue and put the trimeric subunits into one NCS group while allowing the dimeric subunits to move in the absence of NCS restraints. Because of the low resolution in this case (5.6 Å), minimal

adjustments were made to the starting molecular replacement solution, and only when there was support based on Fo-Fc difference maps during COOT (58) assessment, and where suitable alternative conformations exi sted in the dimer structure from which the designed model was derived (PDB: 3NQN). The final R and Rfree values were 22.6% and 24.0%.

Quantitative comparsion of crystal structures and design models

Root mean square deviations (r.m.s.d.) over backbone atoms (N, Cα, C, O) were calculated between each design model and corresponding crystal structure using the pair_fit command in PyMOL (37, Table 5.S6). Chains in each design model were renamed such that there were 60 different chains, each of which comprised a pair of contacting subunits, one subunit from each component. One complete 120-subunit cage was then generated from each crystal structure by applying crystal lattice symmetry and the chains renamed to match the corresponding chains in the design models. R.m.s.d. calculations were carried out for the full icosahedral assemblies or with individual pairs of subunits, one from each component.

Design and characterization of I53-50 variants

Details of the design of the I53-50 variants used in the in vitro assembly and GFP packaging experiments will be described elsewhere (D. Ellis and N.P. King, unpublished results). Briefly, a consensus design approach was used to first identify surface mutations predicted to enhance the stability and/or solubility of the I53-50 components. Using the Rosetta macromolecular modeling suite, the I53-50 computational design model was redesigned by allowing optimization of the identities of relatively exposed residues (defined as having a solvent accessible surface area of greater than 20 square Ångstroms), excepting polar residues (aspartate, glutamate, histidine, lysine, asparagine, glutamine, and arginine) and residues near the designed protein-protein interface between the pentameric and trimeric components. Mutations that resulted in losses of significant atomic packing interactions or side chain-backbone hydrogen bonds were discarded. A position-specific scoring matrix (PSSM) based on homologous protein sequences was used to augment the default Rosetta score function to favor residues that appear frequently at a given position in homologous proteins, a design approach referred to as consensus protein design (59). Multiple design trajectories were performed with varying weights on the contribution of the PSSM, and mutations to polar residues that appeared favorable across all design trajectories were selected for inclusion in the variant proteins.

Subsequently, manually selected amino acid positions were computationally scanned using Rosetta to generate an I53-50 variant nanoparticle with a highly positively charged interior surface. Amino acid identities allowed during sequence design calculations were limited to arginine, lysine, or the native amino acid at each position. Each position was designed independently, and relevant score metrics were assessed to select mutations for inclusion in the variant proteins.

Amino acid sequences for the three variant proteins used in this study (I53-50A.1, I53-50A.1PT1, and I53-50B.4PT1) are included in Table 5.S4. Mutations identified by consensus design intended to enhance stability and/or solubility are highlighted in green, while those included to provide electrostatic interactions with negatively charged cargoes are highlighted in blue.

The I53-50A.1, I53-50A.1PT1, and I53-50B.4PT1 variants were cloned individually into pET29b using the NdeI and XhoI restriction sites as described above. The I53-50A variants were expressed for 3 h at 37 °C as described above, while I53-50B.4PT1 was expressed for 5 h at 18 °C. IMAC purification was carried out as described above, with the exception that buffers contained 500 mM NaCl and 0.75% 3-[(3-Cholamidopropyl)dimethylammonio]-1-

propanesulfonate (CHAPS), which was found to minimize precipitation and aggregation of the individual protein components. After IMAC, the individual components were purified by gel filtration on a Superdex 200 10/300 GL (GE Life Sciences) using 25 mM TRIS pH 8.0, 500 mM NaCl, 1 mM DTT as running buffer, with 0.75% CHAPS included for I53-50A.1PT1 and I53-50B.4PT1. The ability of I53-50A.1 and I53-50B.4PT1 to assemble to the designed icosahedral architecture upon mixing in vitro was analyzed by mixing purified components in a 1:1 molar ratio with each component present at a subunit concentration of 50 μM or 100 μM. Mixtures were allowed to incubate at room temperature overnight, and were then purified on a Superose 6 Increase 10/300 gel filtration column (GE Life Sciences) using 25 mM TRIS pH 8.0, 500 mM NaCl, 1 mM DTT, 0.375% CHAPS as running buffer (Fig. S12A).

Static light scattering

Static light scattering experiments were conducted at 22 °C using a DynaPro Nanostar with a 1 μL quartz cuvette (Wyatt Technology Corp.). The buffer composition of all analyzed samples was 25 mM TRIS pH 8.0, 500 mM NaCl, 1 mM DTT, 0.375% CHAPS.
For each in vitro assembly reaction, 10 μL of I53-50A.1 and 10 μL of I53-50B.4PT1, independently expressed and purified as described above (Design and characterization of I53-50 variants), were transferred to a clean microcentrifuge tube and mixed by pipetting up and down 10 times. 10 uL of the mixed sample was then loaded into the cuvette and static light scattering intensity recorded over time. In each experiment, data collection was initiated 35 to 40 seconds after mixing began and both the cuvette and the Nanostar were pre-equilibrated to 22 °C prior to adding sample. Reaction mixtures contained a 1:1 molar ratio of I53-50A.1 to I53-50B.4PT1 with each component present at a subunit concentration of 8, 16, 32, or 64 μM.

As a reference for the scattering intensity expected for complete assembly, SEC-purified nanoparticles of I53-50A.1 and I53-50B.4PT1 (see Design and characterization of I53-50 variants) were also analyzed at concentrations of 8, 16, 32, or 64 μM. These endpoint values are plotted as dashed horizontal lines in Figure 4D. As a reference for the scattering intensity expected from the unassembled components, intensities were measured from solutions of I53-50B.4PT1 mixed with the wildtype I53-50 trimer (I53-50A-wt, Table 5.S4)) lacking the designed interface mutations necessary to mediate assembly; as with the assembly reactions, I53-50A-wt and I53-50B.4PT1 were mixed in a 1:1 molar ratio with each component present at a subunit concentration of 8, 16, 32, or 64 μM. Normalized static light scattering intensities for each assembly reaction were obtained by subtracting the intensity measured from the unassembled components (I53-50A-wt together with I53-50B.4PT1) at the corresponding concentrations.

GFP encapsulation

A codon-optimized gene encoding GFP(-30) (32) was purchased from Integrated DNA Technologies, and was cloned and expressed using the same procedures mentioned above, with protein expression proceeding for 5 h at 18 °C. IMAC was carried out as described above, with the exception that buffers contained 500 mM salt. Fractions containing GFP(-30) were diluted into 25 mM TRIS-HCl pH 7.5, 1 mM DTT and loaded onto a HiTrap Q HP column (GE Life Sciences) using the same buffer as running buffer. The protein was eluted using a linear gradient of NaCl (0–1 M), and GFP(-30)-containing fractions were concentrated and dialyzed against 25 mM TRIS pH 8, 1 mM DTT.

For GFP packaging reactions, purified I53-50A.1 or I53-50A.1PT1 was mixed with GFP(-30) followed ~30 s later by addition of I53-50B.4PT1. Each protein in the reaction was present at a final concentration of 24 μM, and the final buffer consisted of 25 mM TRIS pH 8, 1 mM DTT with either 0.065 or 1 M NaCl. The packaging reactions were incubated for 6-16 h,

and were then analyzed by gel filtration on a Superose 6 Increase 10/300 gel filtration column (GE Life Sciences) using 25 mM TRIS pH 8.0, 1 mM DTT, with either 0.065 or 1 M NaCl as running buffer. Retention of packaged GFP in the presence of high ionic strength was evaluated by collecting GFP-containing I53-50A.1PT1/I53-50B.4PT1 nanoparticle fractions from a 0.065 M NaCl packaging reaction, adding NaCl to 1 M concentration, incubating for 1 h, and analyzing the sample by gel filtration on a Superose 6 10/300 GL using 25 mM TRIS pH 8, 1 M NaCl, and 1 mM DTT as running buffer (Fig. S12E). Control experiments in which GFP(-30) was mixed with individual components in buffer with 0.065 M NaCl were also analyzed by gel filtration in low salt buffer (Fig. S12B-D).

Packaged GFP was quantified using three measurements: integration of peak areas on the gel filtration chromatograms, absorbance measurements at 280 and 488 nm, and measurements of fluorescence intensity. For peak integration, absorbance was monitored at both 280 and 488 nm during gel filtration and peak areas were integrated using UNICORN version 6.3.2.89 (GE Life Sciences). Absorbance measurements at 280 and 488 nm were obtained using a NanoDrop 8000 spectrophotometer. The absorbance of pure GFP(-30) at 280 and 488 nm in various concentrations of NaCl was measured using a NanoDrop 8000. These measurements were used to calculate the absorbance at 280 nm due to GFP(-30) in gel filtration fractions containing GFP(-30) packaged in I53-50 nanoparticles by either absorbance-based method. The relative amounts of absorbance at 280 nm due to GFP(-30) and the I53-50 components then allowed calculation of the molar ratio of the proteins using calculated extinction coefficients. Quantification of packaged GFP by measurement of fluorescence intensity was performed by comparing the intensity of gel filtration fractions containing GFP(-30) packaged in I53-50 nanoparticles to a standard curve generated using pure GFP(-30). Measurements were obtained using an Spectramax M3 plate reader (Molecular Devices). Fluorescence measurements yielded an estimate of ~7 GFPs per I53-50 nanoparticle (i.e., 60 subunits of each component), while both absorbance-based methods yielded an estimate of ~11 GFPs per I53-50 nanoparticle. Estimation of the internal volume of I53-50 occupied by packaged GFP assumed that GFP was a sphere with radius 2 nm and the interior volume of I53-50 was a sphere with radius 8 nm.

Supplementary Text

In addition to the ten successful designs (I53-34, I53-40, I53-47, I53-50, I52-03, I52-32, I52-33, I32-06, I32-19, and I32-28), SDS gels, native gels, and SEC data indicate an eleventh design, I53-51, is capable of forming co-assembled complexes similar in size to the design model, but it was found to be highly unstable under the conditions tested, yielding only partial assemblies by EM and a SAXS profile devoid of the large scale features expected from the design model (Fig. S8). A twelfth design, I32-10, was also found to yield large co-assembled complexes with roughly the expected shape, as determined by EM, but SEC, SAXS, and EM indicate the structure is significantly larger than intended (Fig. S9).

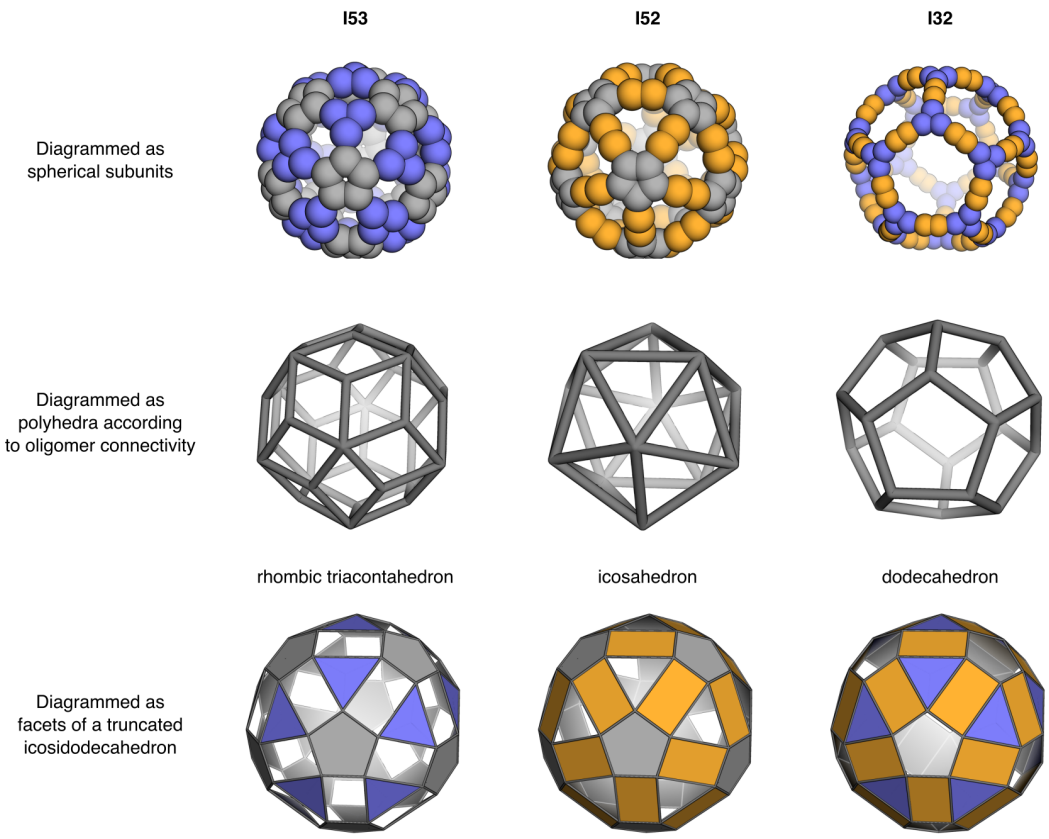


Fig. 5.S1. Design architecture diagrams.
The three types of design architectures targeted in the present study, I53, I52, and I32, each possess icosahedral symmetry, but are constructed using different pairs of oligomeric building blocks. In the top row, the architectures are illustrated using spheres to represent the individual protein subunits comprising each oligomeric building block; pentamers are shown in grey, trimers in blue, and dimers in orange, with each arranged around the corresponding 5-fold, 3-fold, and 2-fold icosahedral symmetry axes. In the middle row, the architectures are illustrated according to the manner in which the subunits are connected. In this representation, the I53, I52, and I32 architectures correspond to three different types of polyhedra—rhombic triacontahedra, icosahedra, and dodecahedra—each of which possess underlying icosahedral symmetry. The design architectures can also be represented as facets of a truncated icosidodecahedron, as illustrated in the bottom row. When viewed in this manner, the I53, I52, and I32 architectures are formed by keeping the pentagonal and triangular faces, pentagonal and rectangular faces, or the triangular and rectangular faces, respectively, with holes left at each of the other faces.

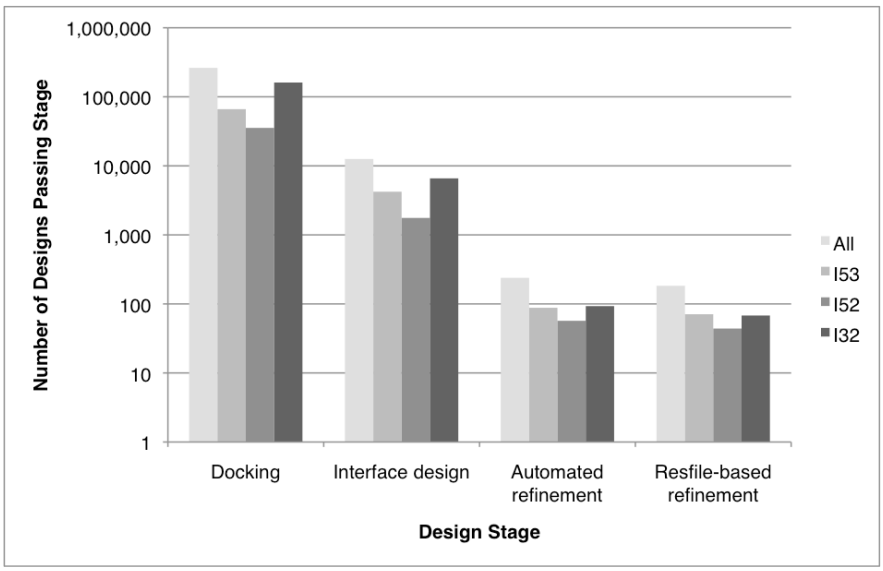


Fig. 5.S2. Number of designs passing each stage of protocol.
The number of design models passing each stage of the design process (docking, interface design, automated refinement, and resfile-based refinement; see the Symmetric Docking and Protein-Protein Interface Design sections above for details about each stage) is shown across all three architectures combined (data labeled as "All" in the legend) and individually (data labeled "I53", "I52", and "I32" in the legend). Data are plotted using a log scale for the y-axis.

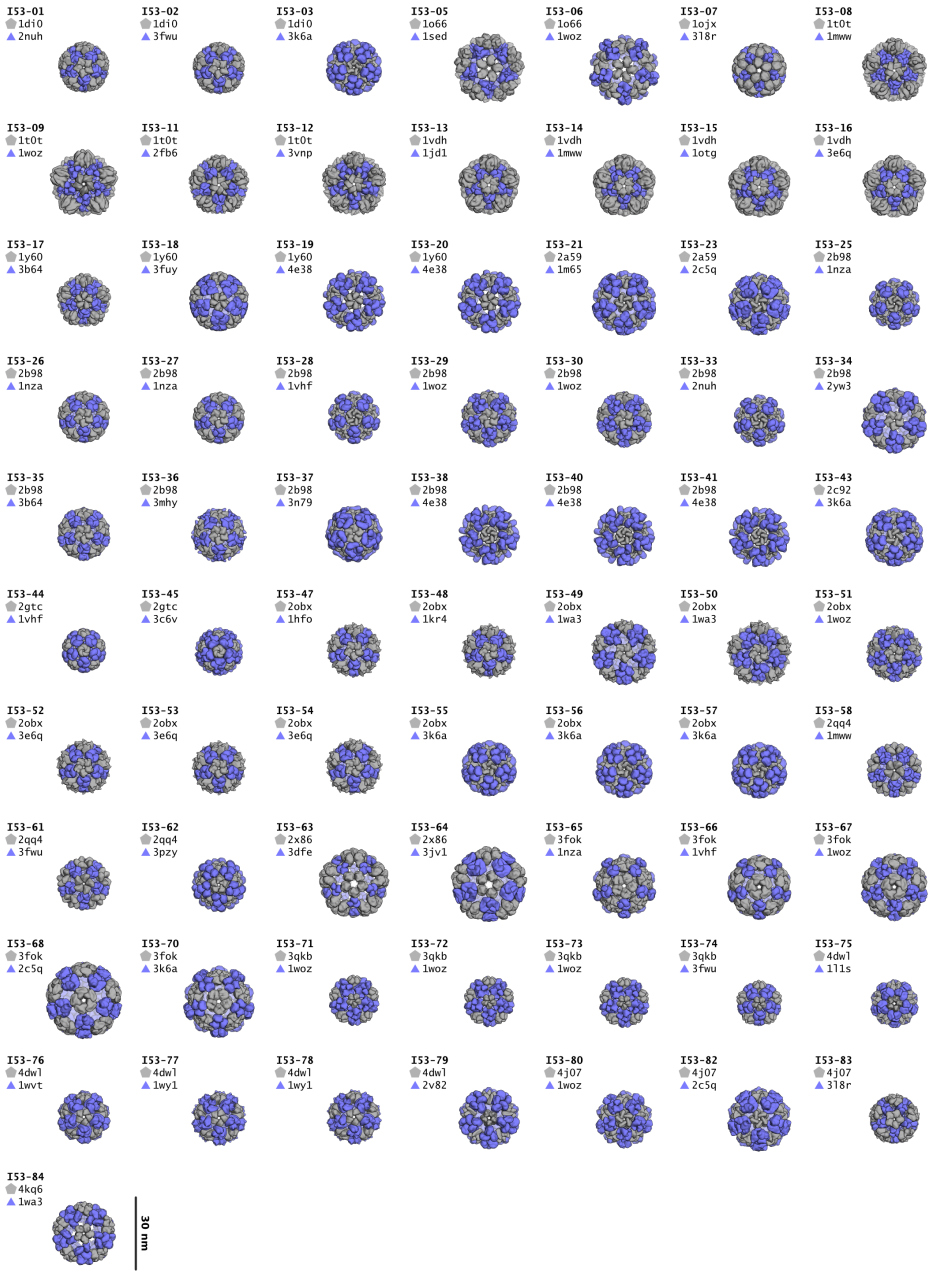


Fig. 5.S3. Models of 71 I53 designs selected for experimental characterization.
Smoothed surface representations are shown of each of the 71 I53 designs selected for experimental testing (rendered to scale relative to the 30 nm scale bar). Each is viewed down one of the icosahedral 5-fold symmetry axes, with the pentameric component of each design shown in grey and the trimeric component in blue. Each design is named according to its symmetric architecture (I53) followed by a unique identification number. The pairs of scaffold proteins from which the designs are derived are indicated directly below each design ID.

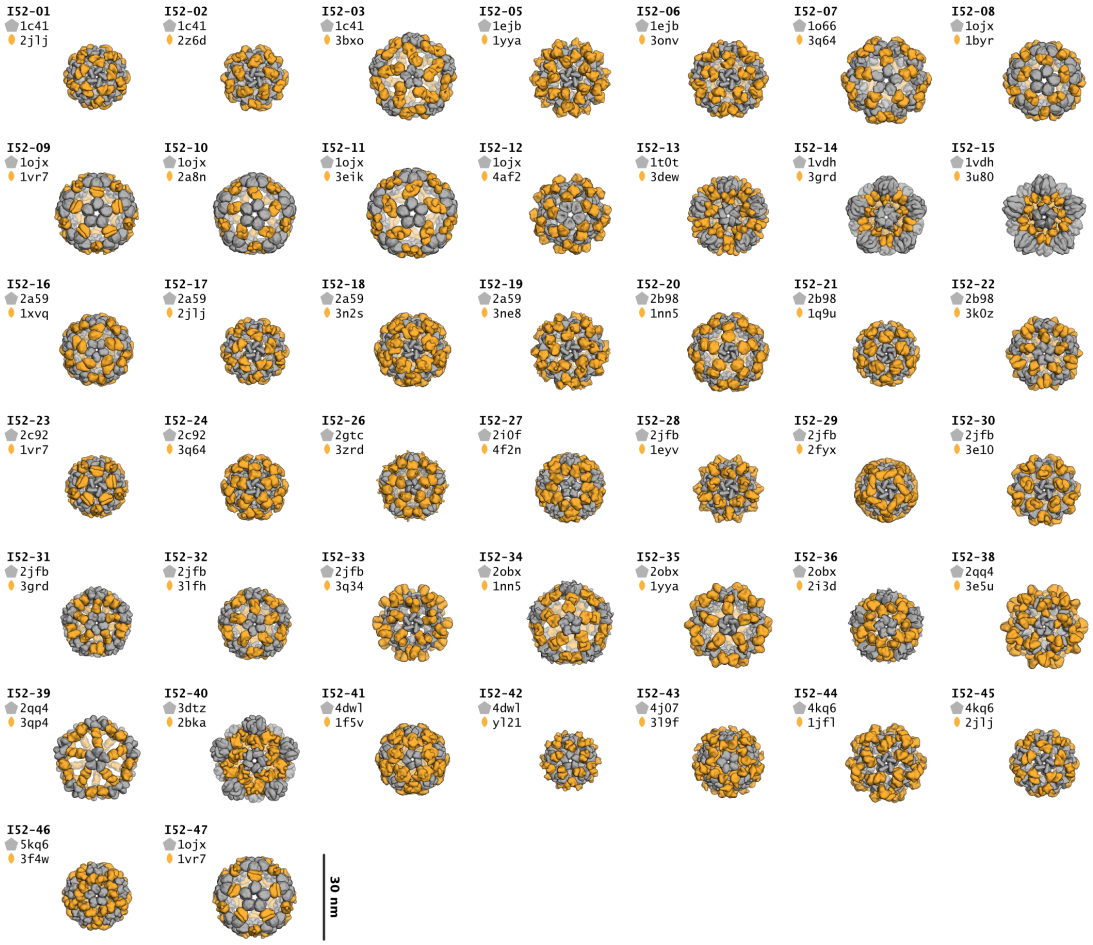


Fig. 5.S4. Models of 47 I52 designs selected for experimental characterization. Smoothed surface representations are shown of each of the 47 I52 designs selected for experimental testing (rendered to scale relative to the 30 nm scale bar). Each is viewed down one of the icosahedral 5-fold symmetry axes, with the pentameric component of each design shown in grey and the trimeric component in blue. Each design is named according to its symmetric architecture (I52) followed by a unique identification number. The pairs of scaffold proteins from which the designs are derived are indicated directly below each design ID.

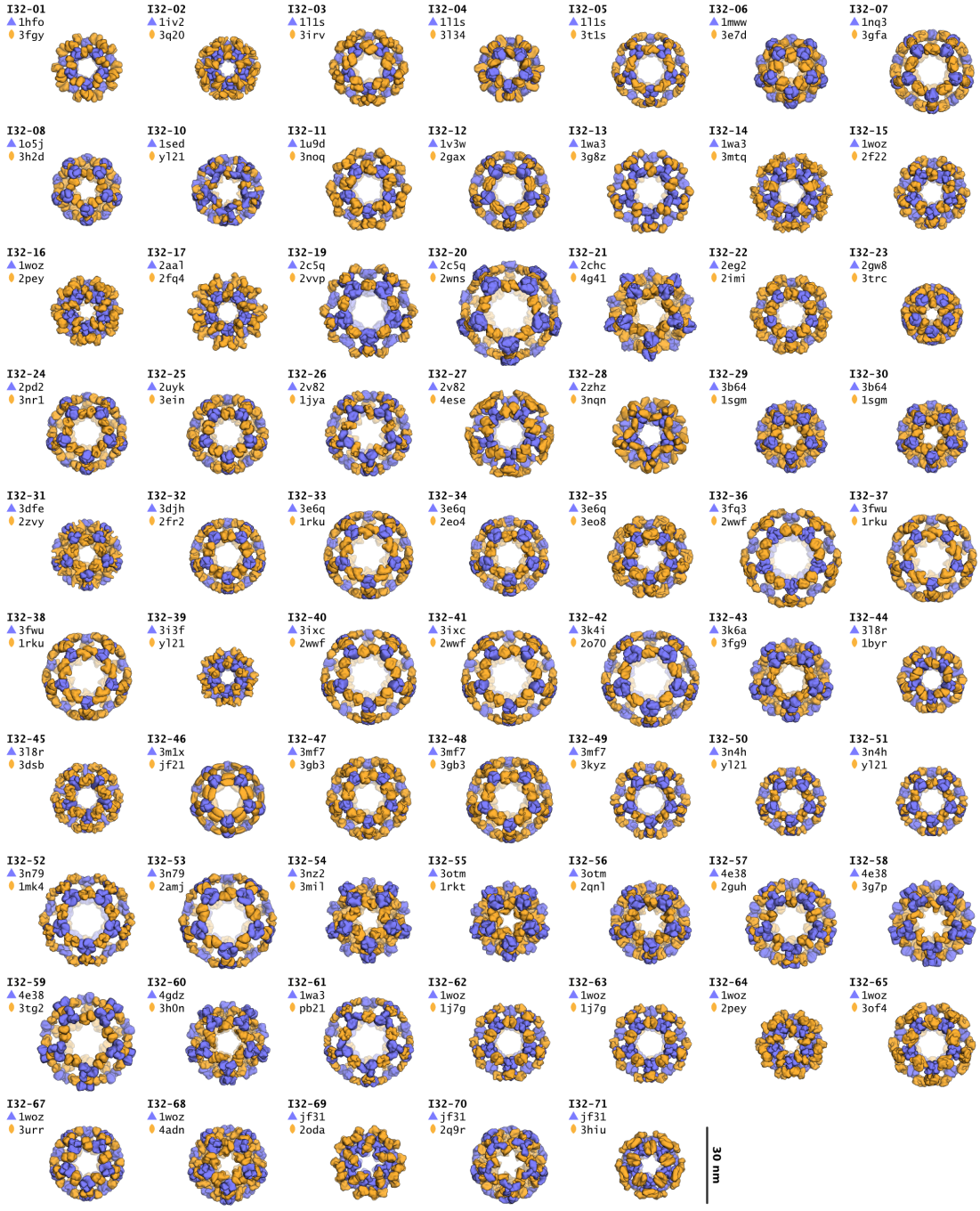
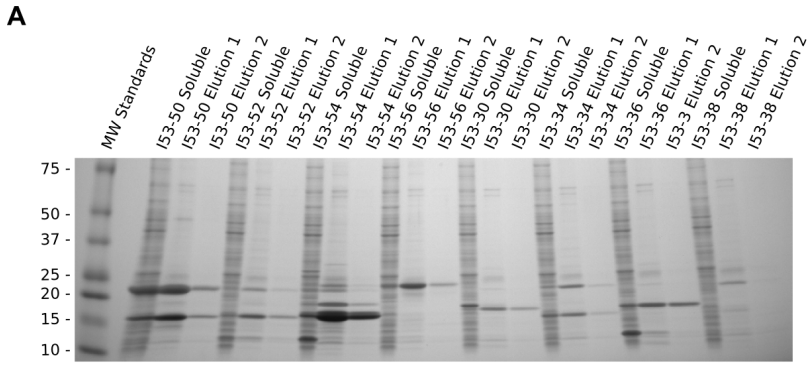


Fig. 5.S5. Models of 68 I32 designs selected for experimental characterization. Smoothed surface representations are shown of each of the 68 I32 designs selected for experimental testing (rendered to scale relative to the 30 nm scale bar). Each is viewed down one of the icosahedral 5-fold symmetry axes, with the pentameric component of each design shown in grey and the trimeric component in blue. Each design is named according to its symmetric architecture (I32) followed by a unique identification number. The pairs of scaffold proteins from which the designs are derived are indicated directly below each design ID.



Design	MW A (kDa)	MW B (kDa)	Tagged Component
I53-50	22.0	18.0	B
I53-52	13.2	18.1	B
I53-54	13.2	18.3	B
I53-56	20.4	17.1	A
I53-30	18.4	20.1	A
I53-34	22.9	17.2	A
I53-36	18.4	12.2	A
I53-38	17.2	23.2	B

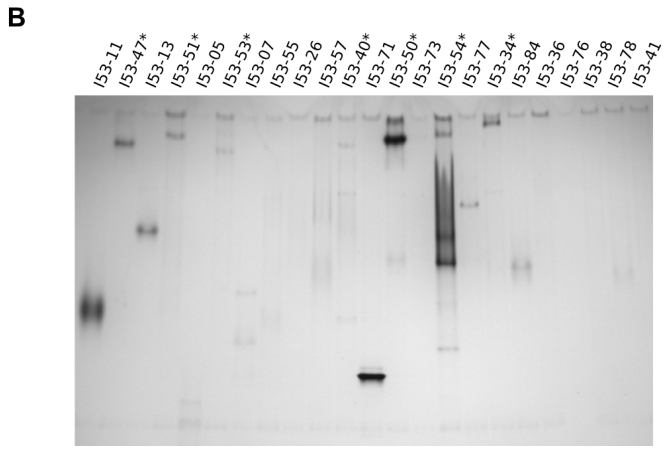


Fig. 5.S6. Example SDS and native PAGE gels from small-scale screening.
(A) An example SDS-PAGE gel from the initial screening of designs via small-scale expression and purification. Soluble fractions of cell lysates and elution fractions resulting from IMAC are shown for 8 of the I53 designs, along with molecular weight standards in the first lane of the gel (the approximate molecular weights in kilodaltons are indicated directly to the left of each band). The expected molecular weights of each designed component is shown in the table to the right (MW A = expected molecular weight of component A, MW B = molecular weight of component B) and the component containing the hexahistidine tag is indicated in the far right column. Two prominent bands, corresponding closely with the expected molecular weights are observed in the elution fractions of several of the designs, including I53-34 and I53-50, indicating possible co-assembly. (B) An example Native PAGE gel performed with the "Elution 1" fractions of those designs appearing to yield two-bands near the expected molecular weights by SDS-PAGE. Sharp bands near the top of the gel indicate potential assembly to higher order materials, such as the target 120-subunit complexes (designs yielding such species are marked with an asterisk).

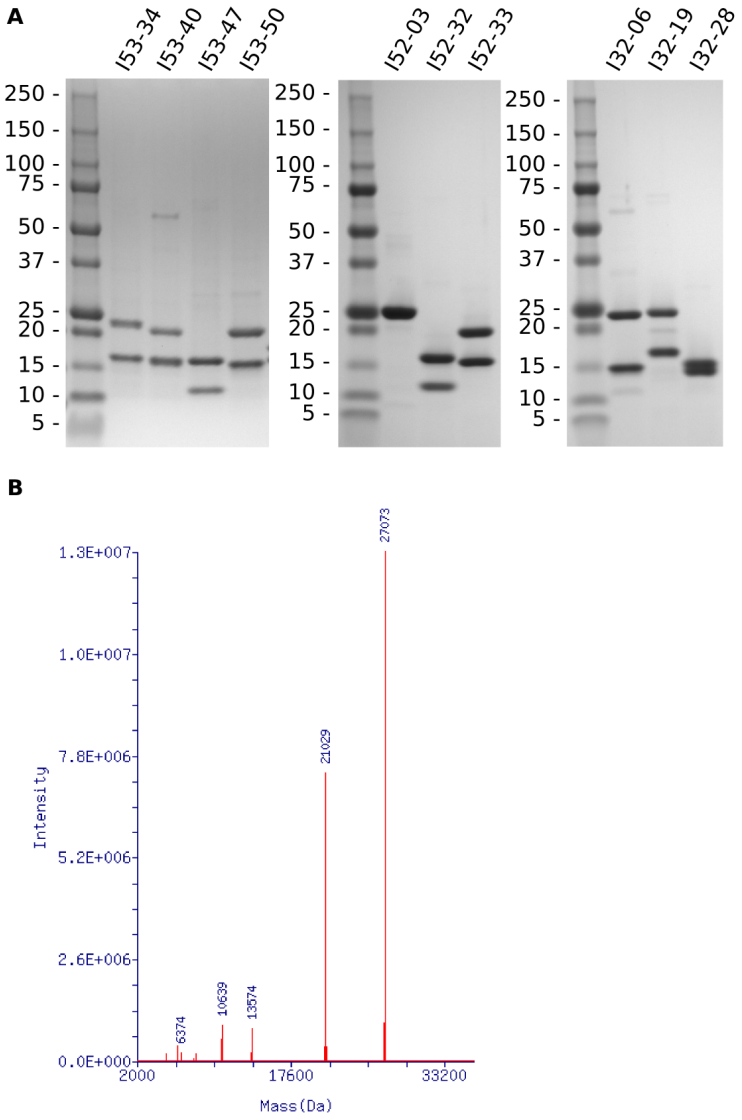


Fig. 5.S7. SDS-PAGE and mass spectrometry analysis of SEC purified samples.
(A) Results from SDS-PAGE analysis of SEC purified samples. The left lane in each panel contains protein molecular weight standards; the approximate molecular weights in kilodaltons are indicated directly to the left of each band. The right lanes in each panel contain the purified samples. For all of the materials except I52-03, clear bands, of similar staining intensity and near the expected molecular weights of each protein subunit, are present for each of the two proteins comprising the purified materials. (B) While only one band (near the expected molecular weight of 27 kDa for the dimer subunit) is clearly distinguishable for I52-03 via SDS-PAGE, mass spectrometry analysis shows that both protein subunits are present in the sample; the peak at 21,029 Da matches closely with the expected molecular weight of 21,026 Da for the pentamer subunit with loss of the initiator methionine.

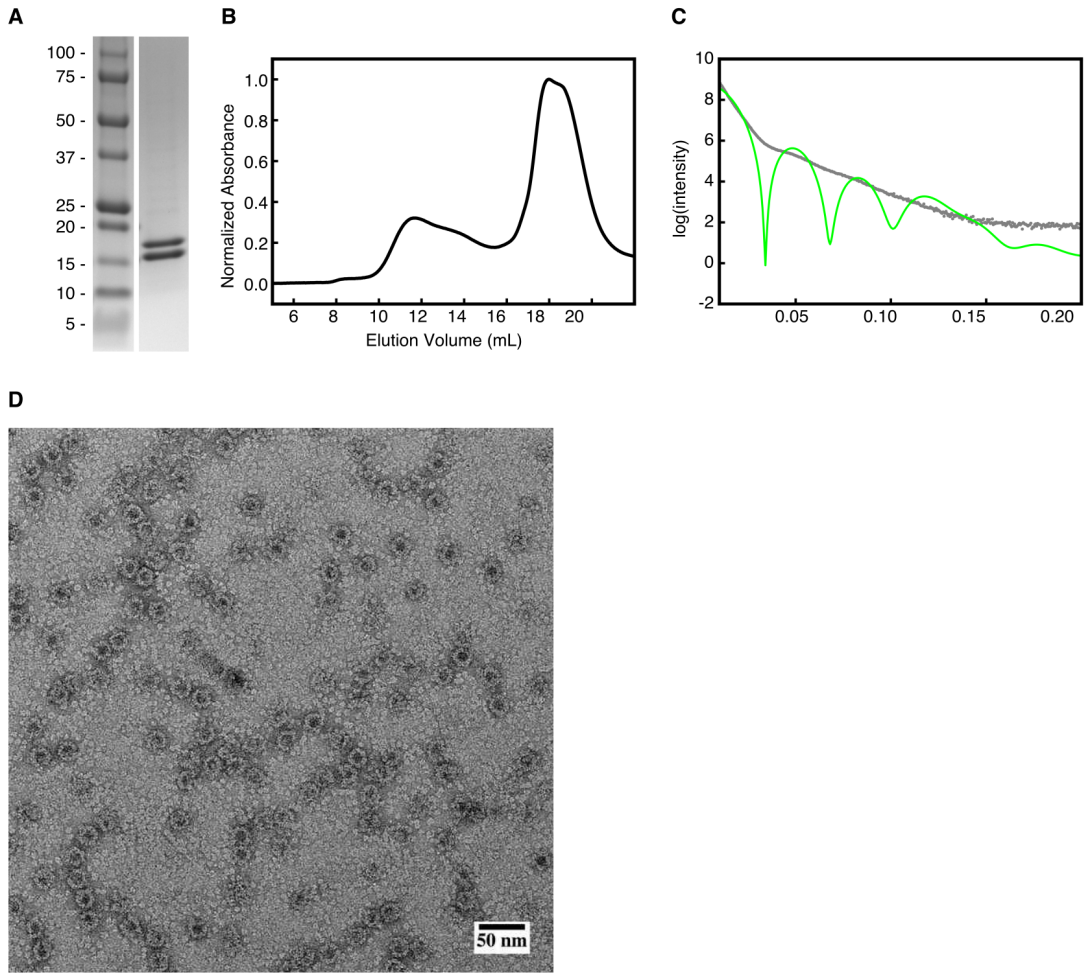


Fig. 5.S8. Experimental characterization of SEC purified I53-51.
(A) SDS-PAGE analysis of SEC purified I53-51 protein shows two bands near the expected molecular weights of 18.3 and 20.1 kDa (molecular weight standards are shown on the left, with the approximate weights in kilodaltons indicated to the left of each band). (B) Analytical SEC yields a small peak near the expected elution volume of 11 to 12 mL, but the peak is tailed heavily toward later elution volumes and a second larger peak is observed near 18 mL. (C) SAXS data (grey dots) does not match well with the profile calculated from the design model (green), nearly completely lacking the large dips in the intensity expected for the assembled material. (D) A representative negative stain electron micrograph is shown of SEC purified I53-51. Particles similar to the design models in shape and size are present, but many appear to be only partially assembled and many unassembled bulding blocks are also visible.

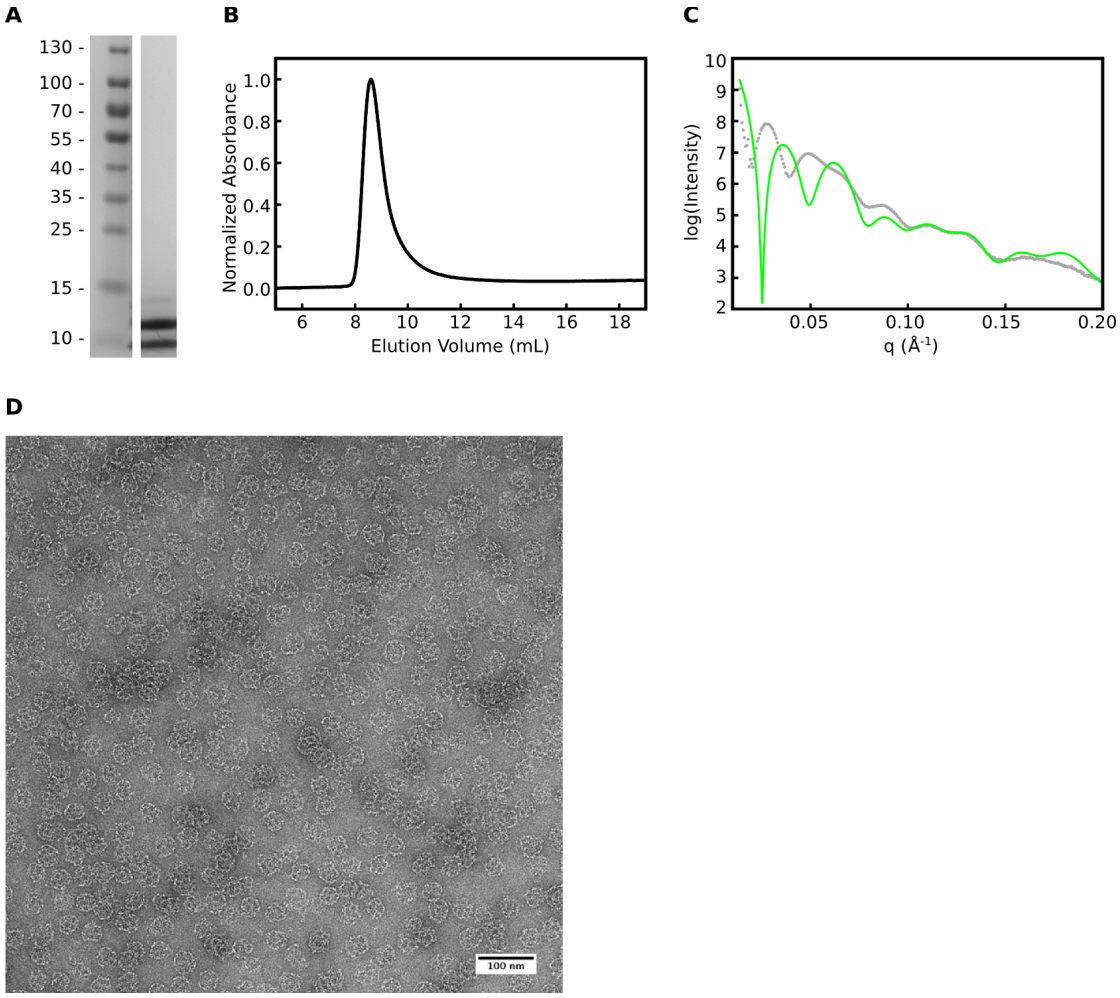


Fig. 5.S9. Experimental characterization of SEC purified I32-10.
(A) SDS-PAGE analysis of SEC purified I32-10 protein shows two bands near the expected molecular weights of 8.3 and 14.3 kDa (molecular weight standards are shown on the left, with the approximate weights in kilodaltons indicated to the left of each band). (B) Analytical SEC yields a single peak near 9 mL, significantly earlier than the elution volume expected based on the diameter of the design model. (C) SAXS data (grey dots) does not match well with the profile calculated from the design model (green); while large dips are observed in the signal, similar to those calculated from the design model, the first two dips are shifted toward lower q values. (D) A representative negative stain electron micrograph is shown of SEC purified I32-10. Spindly, cage-like particles are observed, but appear to be significantly larger than the 29 nm diameter of the design model.

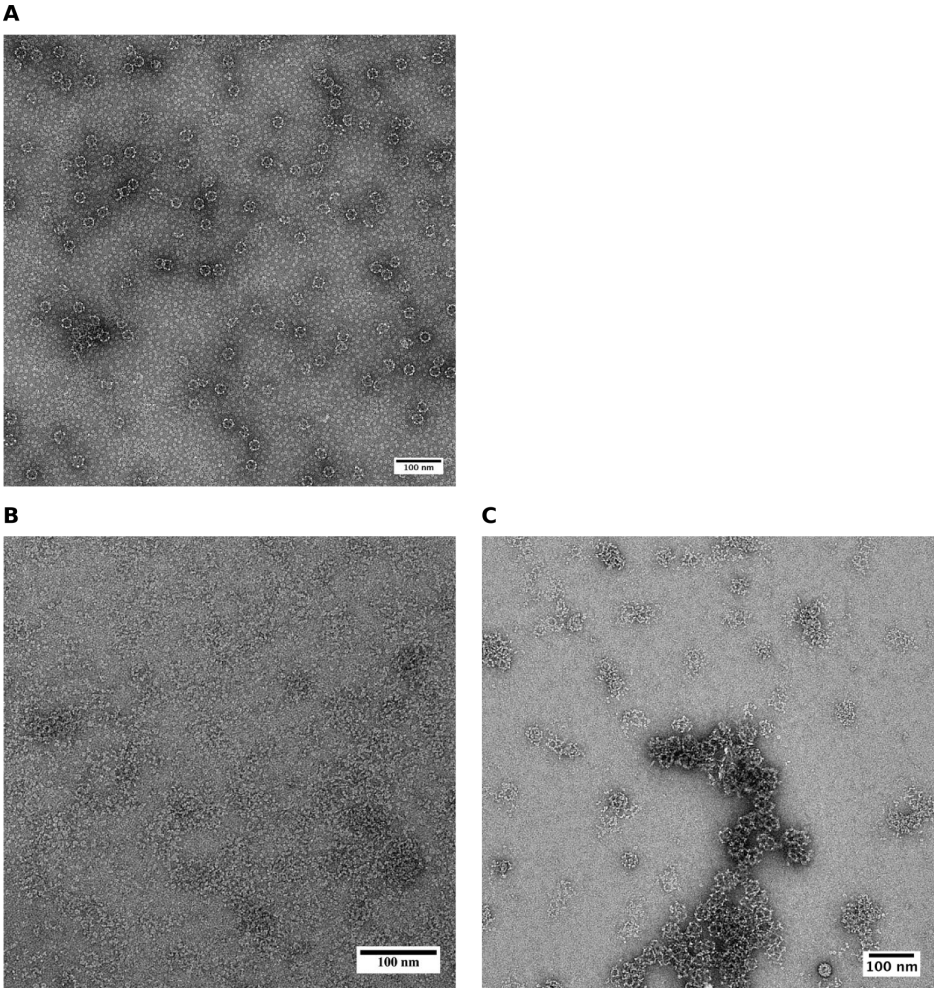


Fig. 5.S10. Electron micrographs of the I52-32 and I32-19 designs.
Representative negative stain electron micrographs are shown for SEC purified samples of the I52-32 (panel A) and I32-19 designs (panels B and C). (A) Assemblies similar in size and shape to the I52-32 design model were observed, along with partially assembled materials and unassembled building blocks, but were too heterogeneous for averaging. (B) In our standard buffer conditions, only aggregates and unassembled building blocks were observed for I32-19. (C) Images collected from sample purified with the addition of 5 percent (v/v) glycerol to all buffers displayed fewer unassembled building blocks and yielded some nanoparticles similar in size and shape to the design model, but also yielded a lot of aggregation and were not suitable for averaging. 100nm scale bars are shown in the lower right of each image.

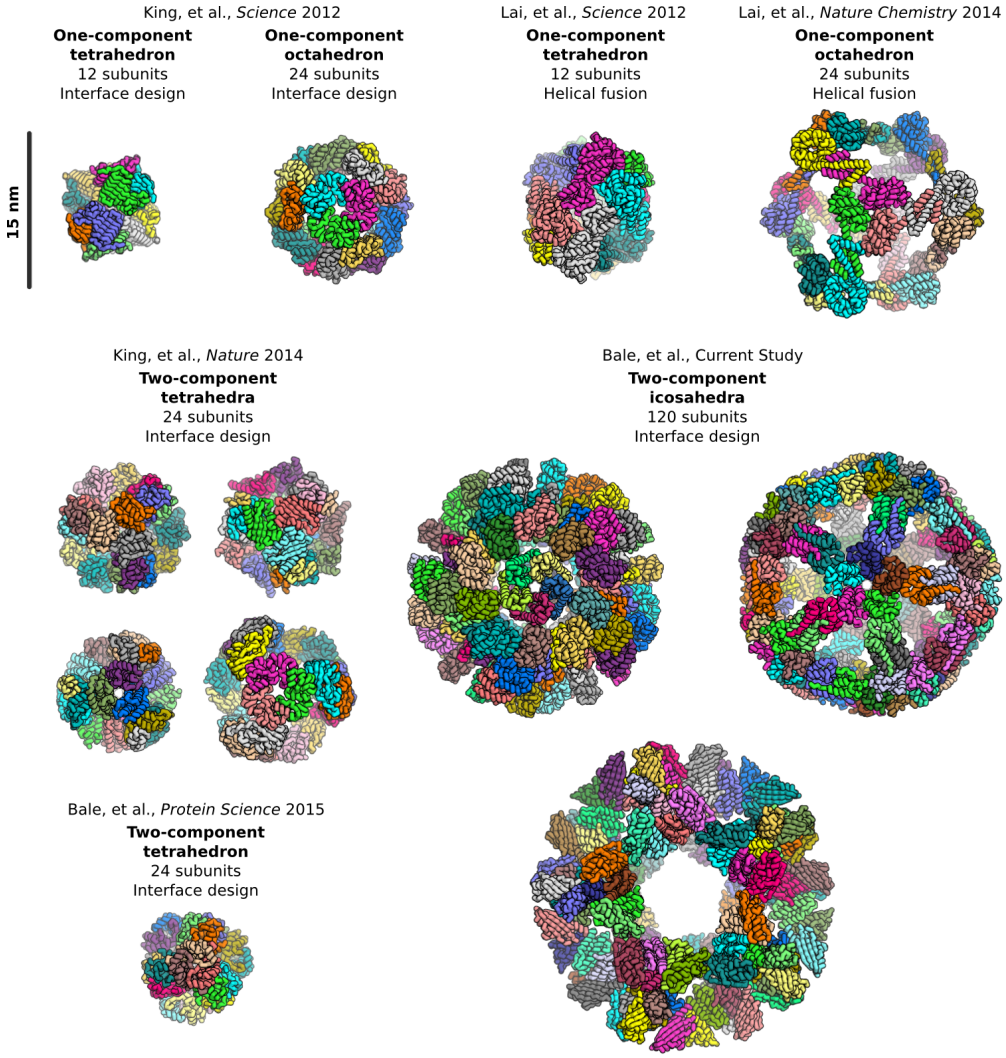


Fig. 5.S11. Comparison of designed protein cages confirmed by X-ray crystallography. Ribbon-style representations of are shown to scale of all the designed protein cages confirmed to date by X-ray crystallography (scale bar: 15 nm). Subunits comprising one whole cage were extracted from each crystal structure and views shown down one of the 2-fold, 3-fold, or 5-fold symmetry axes, with each chain assigned a different color. The source publication, number of distinct protein subunits (one-component versus two-component), symmetry, number of subunits per assembly, and design method (interface design versus helical fusion) are indicated for each structure.

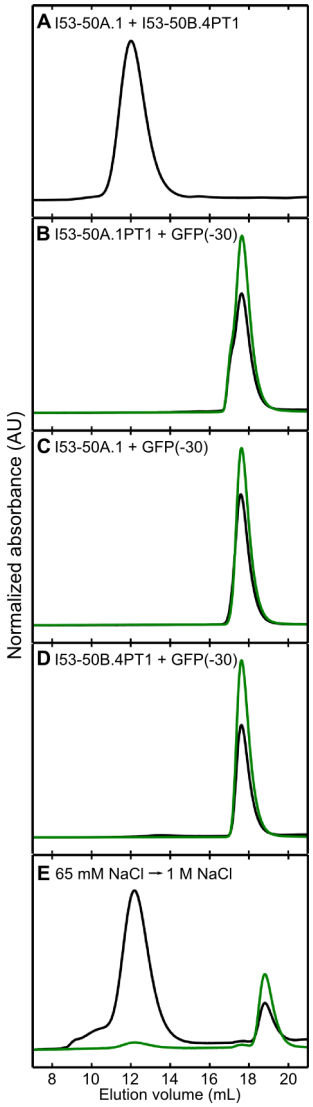


Fig. 5.S12. Analysis of I53-50 variant proteins by SEC.
(A) A mixture of I53-50A.1 and I53-50B.4PT1 (each component at 100 μM) in 25 mM TRIS pH 8, 0.5 M NaCl, 1 mM DTT, 0.375% CHAPS yielded a single peak at the same elution volume observed for co-expressed I53-50 (Fig. 2D). This result was obtained in the same conditions used to analyze assembly kinetics, and demonstrates that the conditions yield the designed icosahedral material. (B-D) Mixtures of GFP(-30) and (B) I53-50A.1PT1, (C) I53-50A.1, or (D) I53-50B.4PT1 did not yield peaks near that observed for assembled I53-50. Both components are required for packaging of GFP(-30). 25 mM TRIS pH 8, 65 mM NaCl, 1 mM DTT was used as running buffer. (E) After transfer of GFP(-30)-containing I53-50 obtained by packaging reactions and SEC in low (65 mM) salt conditions to high (1 M) salt conditions, most of the GFP(-30) no longer co-eluted with the assembled I53-50 (near 12 mL). Instead, released GFP(-30) eluted at ~19 mL. All gel filtration was performed using a Superose 6 Increase 10/300 GL.

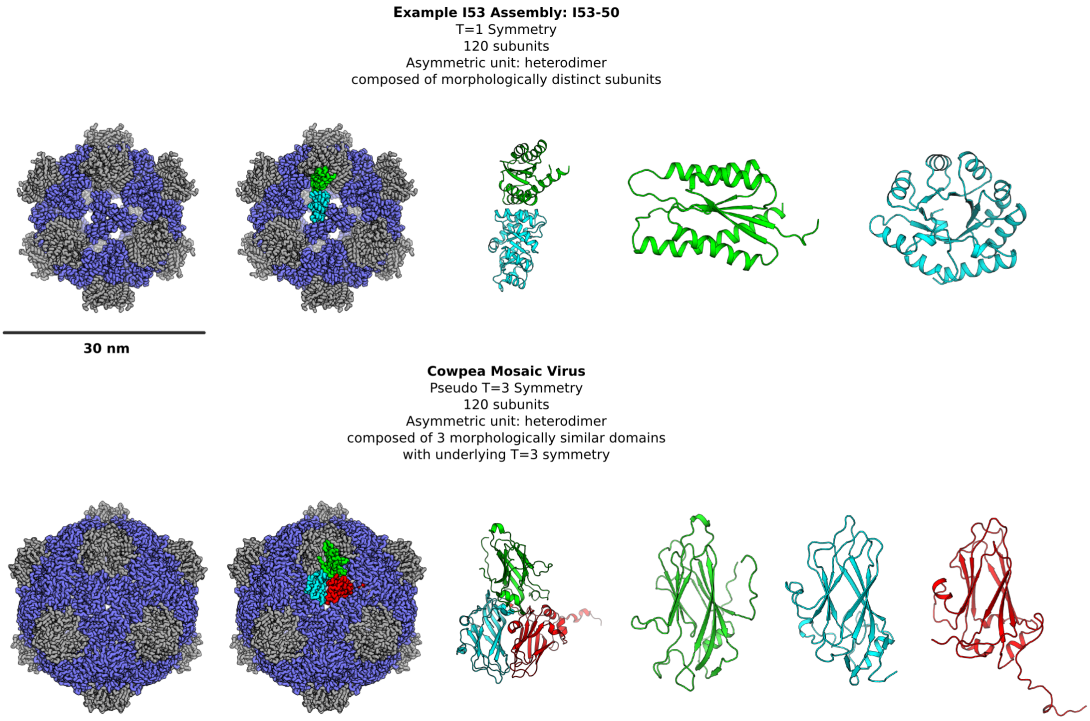


Fig. 5.S13. Comparison of Cowpea Mosaic Virus to the I53 architecture.
The I53 architecture (using the I53-50 design model as an example) is compared to that of the Cowpea Mosaic Virus (CPMV, PDB ID 1ny7). While CPMV meets the criteria of the I53 architecture, it also possesses higher order, pseudo T=3 symmetry. On the left, views are shown down the icosahedral 3-fold symmetry axis with the pentamer forming subunits colored grey and trimer forming subunits colored blue. Both I53-50 and CPMV are comprised of 12 pentamers aligned along the icosahedral 5-fold symmetry axes and 20 trimers aligned along the icosahedral 3-fold symmetry axes, with 120 subunits total. In both cases the asymmetric unit (middle panels, colored green, light blue, and red) is a heterodimer comprised of one pentamer forming subunit (green) and one trimer forming subunit (colored light blue in I53-50, colored light blue and red in CPMV) connected by a non-covalent protein interface. The trimer forming subunit of CPMV contains two jelly roll domains (light blue and red). The pentamer forming subunit of CPMV also contains a jelly roll domain. The full structure thus contains 180 jelly roll domains arranged similarly to a T=3 assembly. However, because the domains do not all possess the same sequence and two domains are fused together in each trimer subunit, the structure does not possess true T=3 symmetry, but rather pseudo T=3 symmetry. On the right, views of the individual domains making up the asymmetric unit are shown for I53-50 and CPMV, highlighting the structural similarity of the CPMV domains and dissimilarity of the I53-50 domains.

Table 5.S1. List of homopentameric PDB entries used as scaffolds for design (PDB ID and biological unit number, separated by an underscore).

1c41_1 1di0_1 1ejb_1 1jg5_1 1k5j_1 1nlq_1 1o66_1 1ojx_1 1qb5_1 1t0t_1 1vdh_1 1vpn_1 1xe0_1 1y2i_1 1y60_1 2a59_1 2b98_1 2c92_1 2gtc_1 2i0f_1 2jfb_1 2obx_1 2p1b_1 2qq4_1 2qw7_1 2rcf_1 2x86_1 3bwr_1 3by7_1 3dtz_1 3dwa_1 3fok_1 3hsa_1 3mxg_1 3nxg_1 3qkb_1 3s7v_1 3s7x_1 3sxp_1 3t30_1 4dmi_1 4dwl_1 4exw_1 4fmg_1 4i7a_1 4ind_1 4j07_1 4kq6_1

Table 5.S2. List of homotrimeric PDB entries used as scaffolds for design (PDB ID and biological unit number, separated by an underscore).

```
1avq_1 1c28_1 1c9k_1 1ca4_1 1czd_1 1dbf_1 1dg6_1 1di6_1 1duc_1 1el6_1 1f23_1 1f7l_1
1fth_1 1gcm_1 1ge8_1 1gr3_1 1gu9_1 1gx1_1 1h7z_1 1h9m_1 1hfo_1 1idp_1 1iv2_1 1jd1_1
1jlj_1 1jq0_1 1khx_1 1knb_1 1kr4_1 1krr_1 1l1s_1 1m65_1 1mvl_1 1mww_1 1n2m_1 1nog_1
1nq3_1 1nza_1 1o51_1 1o5j_1 1o91_1 1oni_1 1otg_1 1ox3_1 1p1l_2 1p9h_1 1pf5_1 1pg6_2
1pwb_1 1q23_1 1q5h_1 1q5x_1 1qre_1 1qu1_1 1rhy_1 1rj8_1 1rlh_2 1rty_1 1s55_1 1sed_1
1seh_1 1sjn_1 1t0a_1 1tcz_1 1td4_1 1u5x_1 1u9d_2 1ufy_1 1uiz_1 1uku_1 1uuy_1 1uxa_1
1v3w_1 1ve0_1 1vfj_1 1vhf_2 1vmf_1 1vmh_1 1vph_1 1wa3_1 1wck_1 1woz_1 1wp8_1 1wvt_1
1wy1_1 1wyy_1 1x25_1 1xhd_2 1xho_1 1xrg_1 1ygs_1 1yox_1 1yq5_1 1yqf_1 2a7k_1 2aal_1
2ah6_1 2arh_1 2b33_1 2bdd_2 2brj_1 2bsf_1 2bt9_1 2bzv_1 2c0a_1 2c5q_1 2chc_1 2cu5_1
2cvl_1 2dch_2 2dj6_1 2dt4_1 2e2a_1 2e7a_1 2ed6_1 2eg2_1 2f0c_1 2fb6_2 2fvh_1 2g2d_1
2gdg_1 2gr7_1 2gr8_1 2gw8_1 2h6l_1 2hx0_1 2i9d_1 2ibl_1 2idx_1 2ieq_1 2ig8_1 2is8_1
2ium_1 2j2j_1 2j9c_1 2jb7_1 2jjl_1 2nt8_1 2nuh_2 2oj6_1 2ol1_1 2otm_1 2p2o_1 2p6c_1
2p6h_1 2p6y_1 2p9o_1 2pd2_1 2pii_1 2pmp_1 2q35_1 2qg8_1 2qlk_1 2r6q_1 2re9_1 2rfr_1
2rie_1 2tnf_1 2uyk_1 2uzh_1 2v82_1 2vnl_1 2wds_1 2wh7_1 2wkb_1 2wld_1 2wq4_1 2x29_1
2x4j_1 2xcz_1 2xx6_1 2y8c_1 2yad_1 2yw3_1 2yzj_1 2zhz_1 3a76_1 3aa8_1 3b64_1 3b6n_1
3b93_1 3bsw_1 3bzq_1 3c19_1 3c6v_1 3ce8_1 3ci3_1 3cj8_1 3cnc_1 3cp1_1 3d01_1 3d9x_1
3da0_1 3de9_1 3dfe_1 3dho_1 3djh_1 3dli_1 3e6q_1 3eby_1 3ehw_1 3ejc_1 3ejv_1 3emf_1
3exv_1 3f09_1 3f0d_1 3f4f_1 3fq3_3 3ftt_1 3fuy_1 3fwt_1 3fwu_1 3gqh_1 3gtz_1 3gud_1
3h5i_1 3h6x_1 3htn_1 3hwu_1 3hyk_1 3hza_1 3hzs_1 3i3f_1 3i7t_1 3i82_1 3i87_1 3ifv_1
3ixc_1 3jv1_1 3k4i_1 3k6a_1 3k93_1 3k9a_1 3kan_1 3ke4_1 3kjj_1 3kwe_1 3kxr_1 3l60_1
3l7q_1 3l8r_1 3laa_1 3lgi_1 3lqw_1 3m1x_1 3mc3_1 3mci_1 3mdx_1 3mf7_1 3mhy_1 3mko_1
3mlc_1 3mqh_1 3n4h_1 3n79_1 3nfd_1 3nhv_1 3ntn_1 3nz2_3 3o46_1 3opk_1 3ot6_2 3otm_1
3p48_1 3pzy_1 3qc7_1 3qr7_1 3qr8_1 3quw_1 3qv0_1 3r1w_1 3r3r_2 3r6h_1 3r8y_1 3rwn_1
3so2_1 3syy_1 3t5s_1 3ta2_1 3tio_1 3tq5_1 3tqz_1 3txt_1 3v4d_1 3vbj_1 3vcr_1 3vnp_1
3zw0_1 4a0t_1 4aff_1 4e38_1 4e98_1 4ea7_1 4fay_1 4fur_1 4g2k_1 4gb5_1 4gdz_1 jf31_1
```

Notes:
1. the following 37 were included in the I53 design process, but not in the I32 design process:

```
1duc_1 1f23_1 1gcm_1 1gr3_1 1jq0_1 1o91_1 1ox3_1 1p9h_1 1qu1_1 1sjn_1 1td4_1 1wp8_1
1wyy_1 1yq5_1 2bsf_1 2ed6_1 2f0c_1 2ibl_1 2ieq_1 2ium_1 2jjl_1 2ol1_1 2vnl_1 2wh7_1
2wld_1 3c19_1 3cp1_1 3d9x_1 3ejc_1 3k9a_1 3laa_1 3mko_1 3qc7_1 3qr7_1 3qr8_1 4a0t_1
4g2k_1
```

2. jf31_1 is a *de novo* designed trimer (data unpublished)

Table 5.S3. List of homodimeric PDB entries used as scaffolds (PDB ID and biological unit number, separated by an underscore).

1a3c_1 1a8l_1 1alu_1 1alv_1 1b4p_1 1bkj_1 1byf_1 1byr_1 1c02_1 1coz_1 1cxq_1 1d6j_1
1dad_1 1dnl_1 1dqn_1 1dug_1 1ecs_1 1ep0_1 1eyv_1 1f1e_1 1f1g_1 1f1m_1 1f3a_1 1f5v_1
1f9z_1 1fit_1 1fj2_3 1fux_1 1fw1_1 1g0s_1 1g2i_1 1g2q_1 1g57_1 1h1y_1 1h99_1 1hgx_1
1hw1_1 1i0r_1 1i12_1 1i3c_1 1i52_1 1iq6_1 1is6_1 1ix9_1 1ixl_2 1izm_1 1j24_1 1j2r_2
1j3m_1 1j3q_1 1j7g_1 1j98_1 1jay_1 1jc4_1 1jfl_1 1jlv_1 1jml_2 1jya_1 1jzt_3 1k2e_1
1k3y_1 1k4i_1 1k66_1 1kll_1 1kqc_1 1ks2_1 1l1q_1 1l6r_1 1lj9_1 1ly1_1 1m0s_1 1m0u_1
1m4i_1 1mjh_1 1mk4_1 1mka_1 1mp9_1 1mqe_1 1msc_1 1mxi_1 1my6_1 1mzh_1 1n2a_1 1n99_1
1ney_1 1nf9_1 1nki_1 1nn5_1 1nox_1 1np6_1 1ns5_1 1nsj_1 1nu3_1 1nxm_1 1nxz_1 1nzn_2
1o22_1 1o3u_1 1o4t_1 1o50_1 1o5x_1 1o63_1 1o6d_2 1oe8_1 1oh0_1 1ohp_1 1oi6_1 1oiv_1
1oki_1 1on2_1 1ooe_1 1oqc_1 1oru_1 1oyj_1 1p6o_1 1pbj_2 1pdo_1 1pn9_1 1prx_1 1pvm_1
1q98_1 1q9u_1 1qb7_1 1qou_1 1qwi_1 1r29_1 1r9c_1 1rkt_1 1rku_1 1rxq_2 1s99_1 1sd4_1
1sgm_1 1sh8_1 1sjy_1 1sk4_2 1snd_1 1snn_1 1sqs_1 1sw0_1 1t5b_1 1t82_1 1t9m_1 1tc1_1
1tc5_1 1tcd_1 1tfe_1 1tks_1 1to4_1 1tu1_1 1tuh_1 1tw9_1 1twu_1 1ty9_1 1u3i_1 1u69_2
1u7i_1 1ues_1 1ukk_1 1upi_1 1usc_1 1usp_1 1uww_1 1v5x_1 1v8y_1 1v96_1 1v9y_1 1va0_1
1vcv_1 1ve2_1 1vf1_1 1vfr_1 1vh5_1 1vhq_1 1vi0_1 1via_1 1vj2_1 1vje_1 1vkc_1 1vki_1
1vl7_1 1vr7_1 1vzg_1 1w2y_1 1wc3_1 1wc9_1 1wkq_1 1wlt_1 1wov_1 1wpn_1 1wr8_1 1wwi_2
1x82_1 1xe7_1 1xfs_1 1xhn_1 1xi3_1 1xpc_1 1xre_1 1xs0_1 1xso_1 1xsq_1 1xuq_1 1xv2_1
1xvq_2 1xw6_1 1y0b_1 1y5h_1 1y7r_1 1y9w_1 1yfu_1 1yki_1 1ylk_1 1ylm_1 1ym3_1 1yoa_1
1yr0_1 1yuz_1 1yya_1 1z4e_1 1z72_1 1z9n_1 1z9p_1 1zb9_1 1zhv_2 1zjr_2 1zn8_1 1zo2_1
1zop_1 1zps_1 1zrn_1 1ztd_1 1zwy_1 2a15_1 2a2r_1 2a35_1 2a67_1 2a8n_1 2a9s_1 2ab0_1
2aef_1 2akp_3 2amj_1 2aps_1 2asf_1 2auw_2 2avd_1 2b06_1 2b0a_1 2b0c_1 2b0v_2 2b18_1
2b5g_1 2b9a_1 2bdr_1 2bka_1 2bnl_1 2bsj_1 2bz1_1 2c0z_1 2c2i_1 2c3q_1 2c4j_1 2car_1
2cl3_1 2cvd_1 2cw2_1 2cwz_1 2cyy_1 2czd_1 2d0j_1 2d2r_1 2d37_1 2d4p_2 2d4u_1 2d5m_1
2d7v_1 2dc1_1 2dc3_1 2dc4_1 2dd7_1 2ddc_1 2dm9_1 2dsc_1 2dtr_1 2dvk_1 2dxq_1 2dxu_1
2e8e_1 2eb1_1 2ecu_1 2een_1 2egv_1 2eh3_1 2ehp_1 2eix_1 2ejn_1 2eo4_1 2ess_1 2ev1_1
2f22_1 2f4p_1 2f5g_1 2f5t_1 2f62_1 2f6g_1 2f6u_1 2f99_2 2f9h_1 2fa1_1 2fa5_1 2fbh_1
2fbq_1 2fck_1 2fd5_1 2fex_1 2fhq_1 2fjt_1 2fl4_1 2fno_1 2fpr_1 2fq4_1 2fr2_2 2fre_1
2ft0_1 2fur_1 2fvu_1 2fyq_1 2fyx_1 2g0i_1 2g3a_1 2g3b_1 2g40_1 2g7s_1 2g84_1 2gau_1
2gax_3 2gen_1 2gfn_1 2gk4_1 2glz_1 2goj_1 2gpc_1 2gpu_1 2gpy_1 2gqr_1 2guh_1 2gux_1
2gvi_1 2gxg_1 2gyq_1 2gz4_1 2h0u_1 2ha8_1 2hbo_1 2hcm_1 2hhz_1 2hku_1 2hkv_1 2hl0_1
2hlj_1 2hng_1 2hnl_1 2hoq_1 2hpv_1 2hq9_1 2hsb_1 2htd_1 2huh_1 2hxw_1 2hyt_1 2i02_1
2i2o_1 2i3d_1 2i51_1 2i7a_1 2i7d_1 2i8b_1 2i8t_1 2ia1_1 2iai_1 2ibd_1 2id6_1 2ig6_1
2igi_1 2ihf_1 2ikk_1 2imf_1 2imi_1 2imj_1 2iml_1 2ims_1 2inb_1 2isy_1 2iu5_1 2ixk_1
2j27_1 2j8m_1 2jar_1 2jba_1 2je3_1 2jk2_1 2jlj_1 2lig_1 2no4_1 2nr4_1 2nrk_1 2nx4_1
2nx8_1 2nyb_1 2nyc_1 2nyi_1 2o08_1 2o28_1 2o6f_1 2o70_1 2o7m_1 2o95_1 2oa2_1 2ob5_1
2ocz_1 2oda_1 2oer_1 2oez_1 2ofx_1 2ogi_1 2oik_1 2okf_1 2oku_1 2omk_1 2onf_1 2ooj_1
2ook_1 2oqm_1 2oso_1 2ou3_1 2ou5_1 2ou6_1 2ov9_1 2owp_1 2ozh_1 2p12_1 2p25_1 2p5q_1
2p7o_1 2p84_1 2p8g_1 2p8j_1 2p92_1 2pa7_1 2pey_1 2pfb_1 2pfi_1 2pn0_1 2pn2_1 2pq7_1
2pqv_1 2prx_1 2ps1_1 2pvq_1 2pwo_1 2pyt_1 2q03_1 2q0y_1 2q24_1 2q2h_1 2q3t_1 2q3x_1
2q4n_1 2q4o_1 2q82_1 2q8o_1 2q9k_1 2q9r_1 2qe9_1 2qec_1 2qg3_1 2qgs_1 2qhk_1 2qib_1
2qjw_1 2qkp_1 2ql8_1 2qmm_1 2qni_1 2qnl_1 2qnt_1 2qqz_1 2qsx_1 2qtq_1 2qtr_1 2qud_1
2qx0_1 2r01_1 2r0x_1 2r1i_1 2r47_1 2r6u_1 2r6v_1 2raf_1 2ras_1 2rbb_1 2rc3_1 2rcv_1
2rh0_1 2rh7_1 2rhm_1 2rk3_1 2rk9_1 2rkh_1 2uv4_1 2v2g_1 2v57_1 2vez_1 2vg0_1 2vns_1
2vvp_1 2vvw_1 2vzx_1 2w2a_1 2w31_1 2w3q_1 2w43_1 2w4e_1 2w53_1 2w7w_1 2wag_1 2wb6_1
2wcr_1 2wcu_1 2wcw_1 2wfc_1 2wns_1 2wp7_1 2wqf_1 2wra_1 2wte_1 2wtg_1 2wwf_1 2wzo_1
2x5c_1 2xbu_1 2xhf_1 2xlg_1 2xme_1 2xpw_1 2xsq_1 2xwl_1 2y0o_1 2y7p_1 2yc3_1 2ycd_1
2yfd_1 2ykz_1 2ysk_1 2yvo_1 2yvs_1 2ywl_1 2ywr_2 2yyv_1 2yzk_1 2z0j_1 2z10_1 2z6d_1
2z8u_1 2z98_1 2zcm_1 2zcw_1 2zej_1 2zgl_1 2znd_1 2zo7_1 2zvy_1 3acd_1 3aia_1 3ajx_1
3b02_1 3b47_1 3bb9_1 3bby_1 3bem_1 3bfm_1 3bhn_1 3bhq_1 3bkw_1 3bl6_1 3bln_1 3bm1_1
3bmz_1 3bos_1 3bpk_1 3bpv_1 3bqx_1 3bqy_1 3bxo_1 3c3m_1 3c3p_1 3c3y_1 3c97_1 3can_1
3cb0_1 3cbg_1 3cc8_1 3ce1_1 3cex_1 3cjd_1 3cje_1 3cjn_1 3cjw_1 3clv_1 3cm3_1 3cng_1
3cp3_1 3ct6_1 3cu3_1 3cvo_1 3d00_1 3d0j_1 3d34_1 3d5p_1 3d7a_1 3db7_1 3dcm_1 3ddh_1
3dew_1 3dlo_1 3dm8_1 3dmc_1 3dn7_1 3dnx_1 3do8_1 3dpj_1 3dqp_1 3dsb_1 3dsh_1 3dtt_1
3duw_1 3dz8_1 3e10_1 3e2c_1 3e39_1 3e4v_1 3e5h_1 3e5t_1 3e5u_1 3e7d_1 3e97_1 3ebt_1
3ec6_1 3ec9_1 3ecf_1 3eei_1 3eer_1 3eik_1 3ein_1 3ejk_1 3ek3_1 3en8_1 3eo8_1 3eof_1
3er6_1 3er7_1 3es1_1 3esm_1 3eup_1 3ew1_1 3exq_1 3ey8_1 3f13_1 3f2i_1 3f2v_1 3f3x_1
3f4w_1 3f6d_1 3f6f_1 3f6v_1 3f7c_1 3f7e_1 3f7l_1 3f8h_1 3f8m_1 3f8x_1 3f9s_1 3fcd_1
3ff0_1 3fg9_1 3fge_1 3fgy_1 3fh1_1 3fiu_1 3flj_1 3fm2_1 3fm5_1 3fqm_1 3frc_1 3frq_1
3fv6_1 3fwz_1 3fxh_1 3fy3_1 3fyn_1 3g0k_1 3g13_1 3g14_1 3g16_1 3g46_1 3g6i_1 3g7p_1

3g7r_1 3g8k_1 3g8z_1 3gag_1 3gb3_1 3gby_1 3gdw_1 3ge6_1 3gfa_1 3ggq_1 3ghj_1 3giu_1
3glv_1 3gm5_1 3gpv_1 3gr3_1 3grd_1 3grz_1 3guz_1 3gyd_1 3gzr_1 3h05_1 3h07_1 3h0n_1
3h1s_1 3h2d_1 3h3l_1 3h4o_1 3h4y_1 3h51_1 3h8u_1 3h95_1 3ha2_1 3hiu_1 3hj9_1 3hm4_1
3hmz_1 3ho7_1 3hoi_1 3ht1_1 3huh_1 3hup_1 3hvv_1 3hzp_1 3i1j_1 3i24_1 3i3g_1 3ia1_1
3ia8_3 3ibm_1 3igr_2 3iis_1 3ijm_1 3ik7_1 3ilx_1 3inq_1 3ir3_1 3irv_1 3iso_1 3itf_1
3itq_1 3ix3_1 3jr2_1 3jtf_1 3jtw_1 3jum_1 3jx9_1 3k0z_1 3k1e_1 3k21_1 3k2v_1 3k67_1
3k69_1 3k86_1 3kbe_1 3kbq_1 3kby_1 3kdw_1 3keb_1 3keo_1 3kg0_2 3kgz_1 3kk4_1 3kkg_1
3kky_1 3kl1_1 3kmh_1 3kol_1 3kos_1 3ksh_1 3ksv_1 3kuv_1 3kvh_1 3kwk_1 3kyz_1 3kzp_1
3l18_1 3l34_1 3l3b_1 3l3u_1 3l7x_1 3l8u_1 3l9f_1 3l9y_1 3la7_1 3las_1 3lb5_1 3lby_1
3lf6_1 3lfh_1 3lfr_1 3lhq_1 3lio_1 3ll5_1 3llv_2 3lm2_1 3lnc_1 3lqn_1 3lqy_2 3lr0_1
3lte_1 3lv8_1 3lva_1 3lw3_1 3lwd_1 3lx7_1 3lyd_1 3lyh_1 3lyp_1 3lza_1 3lzl_1 3m0f_1
3m3h_1 3m3m_2 3m4i_1 3m6j_1 3m9l_1 3m9z_1 3mcw_1 3mdk_1 3mdp_1 3mgd_1 3mgk_1 3mgm_1
3mil_1 3mio_1 3mmh_1 3mms_1 3mng_1 3mnl_1 3mti_1 3mtq_1 3mvp_1 3mxj_1 3n2s_1 3n4j_1
3n6t_2 3nad_1 3nbc_1 3ndo_1 3ne8_1 3nfw_1 3nj2_1 3njc_1 3nl9_1 3nm6_1 3noq_1 3nqn_1
3nr1_1 3nrp_1 3ntv_1 3nua_1 3nym_1 3nzr_1 3o0m_1 3o10_1 3o1c_1 3o2r_1 3o4v_1 3o76_1
3o7b_1 3oa4_1 3of4_1 3of5_1 3oga_1 3ogh_1 3ohe_1 3oji_1 3okx_1 3oms_1 3on4_1 3onp_1
3onv_1 3oqp_1 3oru_1 3ovp_1 3oxp_1 3p0t_1 3pg6_1 3pib_2 3pjl_1 3pmd_1 3pp9_1 3pr8_2
3pss_1 3pu7_1 3pu9_1 3q0w_1 3q18_1 3q20_1 3q34_1 3q58_1 3q62_1 3q63_1 3q64_1 3q6a_1
3q80_1 3q90_1 3qao_1 3qbm_1 3qnc_1 3qoo_1 3qop_1 3qp4_1 3qp8_1 3qs2_1 3qsq_1 3qta_1
3qu1_1 3qxh_1 3qzx_1 3r0n_1 3r2q_1 3r2v_1 3r5g_1 3r6a_1 3r6f_1 3r77_1 3rjt_1 3rkc_1
3rmh_1 3rmu_1 3rnr_1 3rob_1 3rpe_1 3rpp_1 3rqi_1 3rv1_1 3ryk_1 3s6f_1 3s8i_1 3s9f_1
3sb1_1 3sj3_1 3sjs_1 3sk2_1 3sl7_1 3slz_1 3smd_2 3son_1 3soy_1 3sxm_1 3sxy_1 3t1s_1
3t43_1 3t8r_1 3t90_1 3t9y_1 3tem_1 3tg2_1 3tgn_1 3tgv_1 3tj8_1 3tjt_1 3tnj_1 3tqu_1
3tr0_1 3trc_1 3typ_1 3u1d_1 3u2a_1 3u6g_1 3u7i_1 3u80_1 3ub6_1 3uh9_1 3uie_1 3ups_1
3urr_1 3uw1_1 3vjz_1 3vln_1 3vp5_1 3zrd_1 3zve_1 3zw5_1 4a1i_1 4a5n_1 4adn_1 4ae7_1
4ae8_1 4af2_1 4ag7_1 4agh_1 4alg_1 4atm_1 4au1_1 4avm_1 4ax2_1 4ay0_1 4d9o_1 4di0_1
4dmb_1 4dn2_1 4ds3_1 4e08_1 4e2g_1 4eae_1 4ecj_1 4edh_1 4em8_1 4ese_1 4eun_1 4ew7_1
4ezg_1 4f2n_1 4f82_1 4f8y_1 4fak_1 4flb_1 4g41_1 4g6x_1 4g9b_1 4gak_1 4gci_1 4gdh_1
4gmk_1 4go7_1 6gsv_1 jf21_1 pb21_1 yl21_1

Notes:
1. jf21_1, pb21_1, and yl21_1 are *de novo* designed dimers (data unpublished)

Table 5.S4. Amino acid sequences.

Name	Sequence	Scaffold ID
I53-34A	MRGSHHHHHHGMEGMDPLAVLAESRLLPLLTVRGGEDLAGLATVLELMGVGALEITL RTEKGLEALKALRKSGLLLGAGTVRSPKEAEAALEAGAAFLVSPGLLEEVAALAQAR GVPYLPGVLTPTEVERALALGLSALKFFPAEPFQGVRVLRAYAEVFPEVRFLPTGGI KEEHLPHYAALPNLLAVGGSWLLQGDLAAVMKKVKAAKALLSPQAPG	2yw3 trimer
I53-34B	MTKKVGIVDTTFARVDMAEAAIRTLKALSPNIKIIRKTVPGIKDLPVACKKLLEEEG CDIVMALGMPGKAEKDKVCAHEASLGLMLAQLMTNKHIIEVFVHEDEAKDDDELDIL ALVRAIEHAANVYYLLFKPEYLTRMAGKGLRQGREDAGPARE	2b98 pentamer
I53-40A	MTKKVGIVDTTFARVDMASAAILTLKMESPNIKIIRKTVPGIKDLPVACKKLLEEEG CDIVMALGMPGKAEKDKVCAHEASLGLMLAQLMTNKHIIEVFVHEDEAKDDAELKIL AARRAIEHALNVYYLLFKPEYLTRMAGKGLRQGFEDAGPARE	2b98 pentamer
I53-40B	MSTINNQLKALKVIPVIAIDNAEDIIPLGKVLAENGLPAAEITFRSSAAVKAIMLLR SAQPEMLIGAGTILNGVQALAAKEAGATFVVSPGFNPNTVRACQIIGIDIVPGVNNP STVEAALEMGLTTLKFFPAEASGGISMVKSLVGPYGDIRLMPTGGITPSNIDNYLAI PQVLACGGTWMVDKKLVTNGEWDEIARLTREIVEQVNPGSLEHHHHHH	4e38 trimer
I53-47A	MPIFTLNTNIKATDVPSDFLSLTSRLVGLILSKPGSYVAVHINTDQQLSFGGSTNPA AFGTLMSIGGIEPSKNRDHSAVLFDHLNAMLGIPKNRMYIHFVNLNGDDVGWNGTTF	1hfo trimer
I53-47B	MNQHSHKDYETVRIAVVRARWHADIVDACVEAFEIAMAAIGGDRFAVDVFDVPGAYE IPLHARTLAETGRYGAVLGTAFVVNGGIYRHEFVASAVIDGMMNVQLSTGVPVLSAV LTPHRYRDSAEHHRFFAAHFAVKGVEAARACIEILAAREKIAAGSLEHHHHHH	2obx pentamer
I53-50A	MKMEELFKKHKIVAVLRANSVEEAIEKAVAVFAGGVHLIEITFTVPDADTVIKALSV LKEKGAIIGAGTVTSVEQCRKAVESGAEFIVSPHLDEEISQFCKEKGVFYMPGVMTP TELVKAMKLGHTILKLFPGEVVGPQFVKAMKGPFPNVKFVPTGGVNLDNVCEWFKAG VLAVGVGSALVKGTPDEVREKAKAFVEKIRGCTE	1wa3 trimer
I53-50B	MNQHSHKDYETVRIAVVRARWHAEIVDACVSAFEAAMADIGGDRFAVDVFDVPGAYE IPLHARTLAETGRYGAVLGTAFVVNGGIYRHEFVASAVIDGMMNVQLSTGVPVLSAV LTPHRYRDSDAHTLLFLALFAVKGMEAARACVEILAAREKIAAGSLEHHHHHH	2obx pentamer
I53-51A	MFTKSGDDGNTNVINKRVGKDSPLVNFLGDLDELNSFIGFAISKIPWEDMKKDLERV QVELFEIGEDLSTQSSKKKIDESYVLWLLAATAIYRIESGPVKLFVIPGGSEEASVL HVTRSVARRVERNAVKYTKELPEINRMIIVYLNRLSSLLFAMALVANKRRNQSEKIY EIGKSW	1woz trimer
I53-51B	MNQHSHKDYETVRIAVVRARWHADIVDQCVRAFEEAMADAGGDRFAVDVFDVPGAYE IPLHARTLAETGRYGAVLGTAFVVNGGIYRHEFVASAVIDGMMNVQLSTGVPVLSAV LTPHRYRSSREHHEFFREHFMVKGVEAAAACITILAAREKIAAGSLEHHHHHH	2obx pentamer
I52-03A	MGHTKGPTPQQHDGSALRIGIVHARWNKTIIMPLLIGTIAKLLECGVKASNIVVQSV PGSWELPIAVQRLYSASQLQTPSSGPSLSAGDLLGSSTTDLTALPTTTASSTGPFDA LIAIGVLIKGETMHFEYIADSVSHGLMRVQLDTGVPVIFGVLTVLTDDQAKARAGVI EGSHNHGEDWGLAAVEMGVRRRDWAAGKTE	1c41 pentamer
I52-03B	MYEVDHADVYDLFYLGRGKDYAAEASDIADLVRSRTPEASSLLDVACGTGTHLEHFT KEFGDTAGLELSEDMLTHARKRLPDATLHQGDMRDFQLGRKFSAVVSMFSSVGYLKT VAELGAAVASFAEHLEPGGVVVVEPWWFPETFADGWVSADVVRRDGRTVARVSHSVR EGNATRMEVHFTVADPGKGVRHFSDVHLITLFHQREYEAAFMAAGLRVEYLEGGPSG RGLFVGVPALEHHHHHH	3bxo dimer
I52-32A	MGMKEKFVLIITHGDFGKGLLSGAEVIIGKQENVHTVGLNLGDNIEKVAKEVMRIII AKLAEDKEIIIVVDLFGGSPFNIALEMMKTFDVKVITGINMPMLVELLTSINVYDTT ELLENISKIGKDGIKVIEKSSLKM	3lfh dimer
I52-32B	MKYDGSKLRIGILHARWNLEIIAALVAGAIKRLQEFGVKAENIIIETVPGSFELPYG SKLFVEKQKRLGKPLDAIIPIGVLIKGSTMHFEYICDSTTHQLMKLNFELGIPVIFG VLTCLTDEQAEARAGLIEGKMHNHGEDWGAAAVEMATKFNLEHHHHHH	2jfb pentamer
I52-33A	MAVKGLGEVDQKYDGSKLRIGILHARWNRKIILALVAGAVLRLLEFGVKAENIIIET VPGSFELPYGSKLFVEKQKRLGKPLDAIIPIGVLIKGSTMHFEYICDSTTHQLMKLN FELGIPVIFGVLTCLTDEQAEARAGLIEGKMHNHGEDWGAAAVEMATKFN	2jfb pentamer

I52-33B	MGANWYLDNESSRLSFTSTKNADIAEVHRFLVLHGKVDPKGLAEVEVETESISTGIP LRDMLLRVLVFQVSKFPVAQINAQLDMRPINNLAPGAQLELRLPLTVSLRGKSHSYN AELLATRLDERRFQVVTLEPLVIHAQDFDMVRAFNALRLVAGLSAVSLSVPVGAVLI FTARLEHHHHHH	3q34 dimer
I32-06A	MGSHHHHHHGMTDYIRDGSAIKALSFAIILAEADLRHIPQDLQRLAVRVIHACGMVD VANDLAFSEGAGKAGRNALLAGAPILCDARMVAEGITRSRLPADNRVIYTLSDPSVP ELAKKIGNTRSAAALDLWLPHIEGSIVAIGNAPTALFRLFELLDAGAPKPALIIGMP VGFVGAAESKDELAANSRGVPYVIVRGRRGGSAMTAAAVNALASERE	3e7d dimer
I32-06B	MITVFGLKSKLAPRREKLAEVIYSSLHLGLDIPKGKHAIRFLCLEKEDFYYPFDRSD DYTVIEINLMAGRSEETKMLLIFLLFIALERKLGIRAHDVEITIKEQPAHCWGFRGR TGDSARDLDYDIYV	1mww trimer
I32-10A	MEMDIRFRGDDLEALLKAAIMMIKAALKMGATITLSLDGNDLEIRITGVPEAARKAL ATIAEVLAKTFGITVTRTIR	yl21 dimer
I32-10B	MDSMDHRIERLEYYIQLLVKTVDMDRYPFYALLIDKGLSKEEGESVMRICQALSVAL ETLKALGQVTFDELLKIFAGALNEKLDVHETIFALYEQGLYQELMEVFIDIMKHFDL EHHHHHH	1sed trimer
I32-19A	MGSDLQKLQRFSTCDISDGLLNVYNIPTGGYFPNLTAISPPQNSSIVGTAYTVLFAP IDDPRPAVNYIDSVPPNSILVLALEPHLQSQFHPFIKITQAMYGGLMSTRAQYLKSN GTVVFGRIRDVDEHRTLNHPVFAYGVGSCAPKAVVKAVGTNVQLKILTSDGVTQTIC PGDYIAGDNNGIVRIPVQETDISKLVTYIEKSIEVDRLVSEAIKNGLPAKAAQTARR MVLKDYI	2c5q trimer
I32-19B	MSGMRVYLGADHAGYELKQAIIAFLKMTGHEPIDCGALRYDADDDYPAFCIAAATRT VADPGSLGIVLGGSGNGEQIAANKVPGARCALAWSVQTAALAREHNNAQLIGIGGRM HTLEEALRIVKAFVTTPWSKAQRHQRRIDILAEYERTHEAPPVPGAPALEHHHHHH	2vvp dimer
I32-28A	MGDDARIAAIGDVDELNSQIGVLLAEPLPDDVRAALSAIQHDLFDLGGELCIPGHAA ITEDHLLRLALWLVHYNGQLPPLEEFILPGGARGAALAHVCRTVCRRAERSIKALGA SEPLNIAPAAYVNLLSDLLFVLARVLNRAAGGADVLWDRTRAH	2zhz trimer
I32-28B	MILSAEQSFTLRHPHGQAAALAFVREPAAALAGVQRLRGLDSDGEQVWGELLVRVPL LGEVDLPFRSEIVRTPQGAELRPLTLTGERAWVAVSGQATAAEGGEMAFAFQFQAHL ATPEAEGEGGAAFEVMVQAAAGVTLLLVAMALPQGLAAGLPPALEHHHHHH	3nqn dimer
I53-50A.1	MKMEELFKKHKIVAVLRANSVEEAIEKAVAVFAGGVHLIEITFTVPDADTVIKALSV LKEKGAIIGAGTVTSVEQCRKAVESGAEFIVSPHLDEEISQFCKEKGVFYMPGVMTP TELVKAMKLGHDILKLFPGEVVGPQFVKAMKGPFPNVKFVPTGGVNLDNVCEWFKAG VLAVGVGDALVKGDPDEVREKAKKFVEKIRGCTEGSLEHHHHHH	I53-50A trimer
I53-50A.1PT1	MKMEELFKKHKIVAVLRANSVEEAIEKAVAVFAGGVHLIEITFTVPDADTVIKALSV LKEKGAIIGAGTVTSVEQCRKAVESGAEFIVSPHLDEEISQFCKEKGVFYMPGVMTP TELVKAMKLGHDILKLFPGEVVGPQFVKAMKGPFPNVKFVPTGGVNLDNVCKWFKAG VLAVGVGKALVKGKPDEVREKAKKFVKKIRGCTEGSLEHHHHHH	I53-50A.1 trimer
I53-50B.4PT1	MNQHSHKDHETVRIAVVRARWHAEIVDACVSAFEAAMRDIGGDRFAVDVFDVPGAYE IPLHARTLAETGRYGAVLGTAFVVNGGIYRHEFVASAVINGMMNVQLNTGVPVLSAV LTPHNYDKSKAHTLLFLALFAVKGMEAARACVEILAAREKIAAGSLEHHHHHH	I53-50B pentamer
I53-50A-wt	MHHHHHHGGMKMEELFKKHKIVAVLRANSVEEAKEKALAVFEGGVHLIEITFTVPDA DTVIKELSFLKEKGAIIGAGTVTSVEQCRKAVESGAEFIVSPHLDEEISQFCKEKGV FYMPGVMTPTELVKAMKLGHTILKLFPGEVVGPQFVKAMKGPFPNVKFVPTGGVNLD NVCEWFKAGVLAVGVGSALVKGTPDEVREKAKAFVEKIRGCTE	1wa3 trimer

Table 5.S5. X-ray diffraction data collection and refinement statistics. Statistics in parentheses refer to the highest resolution shell.

	I53-40 (PDB ID 5IM5)	I52-32 (PDB ID 5IM4)	I32-28 (PDB ID 5IM6)
Data Collection			
Space group	I 2 2 2	H 3	H 3
Cell dimensions			
a, b, c (Å)	265.62, 279.81, 301.33	258.82, 258.82, 641.73	284.17, 284.17, 640.47
α, β, γ (˚)	90.0, 90.0, 90.0	90.0, 90.0, 120.0	90.0, 90.0, 120.0
Resolution (Å)	199.3-3.7	213.9-3.5	213.2-5.6
Rmerge (%)	16.8 (84.4)	13.2 (67.4)	7.7 (97.3)
CC1/2 (%)	99.2 (64.3)	99.1 (65.7)	99.9 (61.1)
CC* (%)	99.8 (94.4)	99.8 (86.9)	100.0 (86.1)
Mean I/σ	8.02 (2.73)	7.31 (1.90)	13.46 (1.57)
Completeness (%)	99.0 (95.1)	99.1 (97.5)	99.5 (95.2)
Multiplicity	4.2 (3.5)	3.6 (3.6)	5.9 (5.1)
Wilson B-factor	70.7	75.0	341.8
Refinement			
Resolution range (Å)	199.3-3.7 (3.74-3.70)	213.9-3.50 (3.54-3.50)	213.2-5.6 (5.65-5.59)
No. reflections	118004 (3560)	200301 (5738)	53801 (141)
Rwork/Rfree (%)	24.0/24.7	22.7/23.3	22.6/24.0
No. atoms			
Protein	39198	40538	44240
Ligand/ion	0	0	0
Water	0	0	0
Average B factors			
Protein	100.4	104.8	317.4
Ligand/ion	NA	NA	NA
Water	NA	NA	NA
Protein residues	5245	5302	6000
R.m.s. deviations			
Bond length (Å)	0.011	0.008	0.009
Bond angles (Å)	1.3	1.1	1.1
Ramachandran favored (%)	92.3	95.5	97.9
Ramachandran allowed (%)	7.7	4.5	1.7
Ramachandran outliers (%)	0.00	0	0.4

Table 5.S6. Root mean square deviations (r.m.s.d.) between crystal structures and design models.

	I53-40	I52-32	I32-28
Crystal Structure	5IM5	5IM4	5IM6
Global r.m.s.d. (Å)[1]	0.806	0.822	2.703
Min, max, ave two-subunit r.m.s.d. (Å)[2]	0.233, 0.352, 0.278	0.647, 0.717, 0.699	0.425, 1.095, 0.608
Contents of asymmetric unit	15 chains of each component (30 subunits)	20 chains of each component (40 subunits)	20 chains of each component (40 subunits)
Structure used for superposition[3]	One cage generated from crystallographic 2-folds	One cage generated from crystallographic 3-folds	One cage generated from crystallographic 3-folds

[1]Global backbone r.m.s.d. values were calculated over all 120 subunits of each design model and corresponding subunits in each crystal structure.
[2]Two-subunit backbone r.m.s.d. values were calculated over pairs of interface subunits in each design model and corresponding subunits in each crystal structure. All 60 possible two-subunit r.m.s.d. values were calculated for each design model and crystal structure. The minimum (min), maximum (max), and average (ave) values are reported for each design.
[3]120 subunits comprising one complete cage were derived from each crystal structure as indicated.

SUPPLEMENTARY REFERENCES AND NOTES

1. Y. T. Lai, N. P. King, T. O. Yeates, Principles for designing ordered protein assemblies. Trends Cell Biol 22, 653-661 (2012).
2. N. P. King, Y. T. Lai, Practical approaches to designing novel protein assemblies. Curr Opin Struct Biol 23, 632-638 (2013).
3. J. E. Padilla, C. Colovos, T. O. Yeates, Nanohedra: using symmetry to design self assembling protein cages, layers, crystals, and filaments. Proc Natl Acad Sci U S A 98, 2217-2221 (2001).
4. P. Ringler, G. E. Schulz, Self-assembly of proteins into designed networks. Science 302, 106-109 (2003).
5. S. Raman, G. Machaidze, A. Lustig, U. Aebi, P. Burkhard, Structure-based design of peptides that self-assemble into regular polyhedral nanoparticles. Nanomedicine 2, 95-102 (2006).
6. D. Grueninger, N. Treiber, M. O. Ziegler, J. W. Koetter, M. S. Schulze, G. E. Schulz, Designed protein-protein association. Science 319, 206-209 (2008).
7. S. Raman, Machaidze G., Lustig A., Olivieri V., Aebi U., Burkhard P., Design of Peptide Nanoparticles Using Simple Protein Oligomerization Domains. The Open Nanomedicine Journal 2, 15-26 (2009).
8. K. Usui, T. Maki, F. Ito, A. Suenaga, S. Kidoaki, M. Itoh, M. Taiji, T. Matsuda, Y. Hayashizaki, H. Suzuki, Nanoscale elongating control of the self-assembled protein filament with the cysteine-introduced building blocks. Protein Sci 18, 960-969 (2009).
9. E. N. Salgado, R. J. Radford, F. A. Tezcan, Metal-directed protein self-assembly. Acc Chem Res 43, 661-672 (2010).
10. G. Grigoryan, Y. H. Kim, R. Acharya, K. Axelrod, R. M. Jain, L. Willis, M. Drndic, J. M. Kikkawa, W. F. DeGrado, Computational design of virus-like protein assemblies on carbon nanotube surfaces. Science 332, 1071-1076 (2011).
11. J. C. Sinclair, K. M. Davies, C. Venien-Bryan, M. E. Noble, Generation of protein lattices by fusing proteins with matching rotational symmetry. Nat Nanotechnol 6, 558-562 (2011).
12. P. B. Stranges, M. Machius, M. J. Miley, A. Tripathy, B. Kuhlman, Computational design of a symmetric homodimer using beta-strand assembly. Proc Natl Acad Sci U S A 108, 20562-20567 (2011).
13. A. L. Boyle, E. H. Bromley, G. J. Bartlett, R. B. Sessions, T. H. Sharp, C. L. Williams, P. M. Curmi, N. R. Forde, H. Linke, D. N. Woolfson, Squaring the circle in peptide assembly: from fibers to discrete nanostructures by de novo design. J Am Chem Soc 134, 15457-15467 (2012).
14. J. D. Brodin, X. I. Ambroggio, C. Tang, K. N. Parent, T. S. Baker, F. A. Tezcan, Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays. Nat Chem 4, 375-382 (2012).
15. B. S. Der, M. Machius, M. J. Miley, J. L. Mills, T. Szyperski, B. Kuhlman, Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. J Am Chem Soc 134, 375-385 (2012).
16. N. P. King, W. Sheffler, M. R. Sawaya, B. S. Vollmar, J. P. Sumida, I. Andre, T. Gonen, T. O. Yeates, D. Baker, Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science 336, 1171-1174 (2012).
17. Y. T. Lai, D. Cascio, T. O. Yeates, Structure of a 16-nm cage designed by using protein oligomers. Science 336, 1129 (2012).

18. C. J. Lanci, C. M. MacDermaid, S. G. Kang, R. Acharya, B. North, X. Yang, X. J. Qiu, W. F. DeGrado, J. G. Saven, Computational design of a protein crystal. Proc Natl Acad Sci U S A 109, 7304-7309 (2012).
19. J. M. Fletcher, R. L. Harniman, F. R. Barnes, A. L. Boyle, A. Collins, J. Mantell, T. H. Sharp, M. Antognozzi, P. J. Booth, N. Linden, M. J. Miles, R. B. Sessions, P. Verkade, D. N. Woolfson, Self-assembling cages from coiled-coil peptide modules. Science 340, 595-599 (2013).
20. N. P. King, J. B. Bale, W. Sheffler, D. E. McNamara, S. Gonen, T. Gonen, T. O. Yeates, D. Baker, Accurate design of co-assembling multi-component protein nanomaterials. Nature 510, 103-108 (2014).
21. Y. T. Lai, E. Reading, G. L. Hura, K. L. Tsai, A. Laganowsky, F. J. Asturias, J. A. Tainer, C. V. Robinson, T. O. Yeates, Structure of a designed protein cage that self-assembles into a highly porous cube. Nat Chem 6, 1065-1071 (2014).
22. A. R. Voet, H. Noguchi, C. Addy, D. Simoncini, D. Terada, S. Unzai, S. Y. Park, K. Y. Zhang, J. R. Tame, Computational design of a self-assembling symmetrical beta-propeller protein. Proc Natl Acad Sci U S A 111, 15102-15107 (2014).
23. S. Gonen, F. DiMaio, T. Gonen, D. Baker, Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. Science 348, 1365-1368 (2015).
24. Y. Hsia, J. B. Bale, S. Gonen, D. Shi, W. Sheffler, K. K. Fong, U. Nattermann, C. Xu, P. Huang, R. Ravichandran, S. Yi, T. N. Davis, T. Gonen, N. P. King, D. Baker, Design of a hyperstable 60-subunit protein icosahedron. Nature, in press (included as reference material).
25. D. S. Goodsell, A. J. Olson, Structural symmetry and protein function. Annu Rev Biophys Biomol Struct 29, 105-153 (2000).
26. J. B. Bale, R. U. Park, Y. Liu, S. Gonen, T. Gonen, D. Cascio, N. P. King, T. O. Yeates, D. Baker, Structure of a designed tetrahedral protein assembly variant engineered to have improved soluble expression. Protein Sci 24, 1695-1701 (2015).
27. D. L. Caspar, A. Klug, Physical principles in the construction of regular viruses. Cold Spring Harb Symp Quant Biol 27, 1-24 (1962).
28. R. Zandi, D. Reguera, R. F. Bruinsma, W. M. Gelbart, J. Rudnick, Origin of icosahedral symmetry in viruses. Proc Natl Acad Sci U S A 101, 15556-15560 (2004).
29. M. C. Lawrence, P. M. Colman, Shape complementarity at protein/protein interfaces. J Mol Biol 234, 946-950 (1993).
30. A. Zlotnick, J. M. Johnson, P. W. Wingfield, S. J. Stahl, D. Endres, A theoretical model successfully identifies features of hepatitis B virus capsid assembly. Biochemistry 38, 14644-14652 (1999).
31. A. Zlotnick, R. Aldrich, J. M. Johnson, P. Ceres, M. J. Young, Mechanism of capsid assembly for an icosahedral plant virus. Virology 277, 450-456 (2000).
32. M. S. Lawrence, K. J. Phillips, D. R. Liu, Supercharging proteins can impart unusual resilience. J Am Chem Soc 129, 10110-10112 (2007).
33. B. Worsdorfer, K. J. Woycechowsky, D. Hilvert, Directed evolution of a protein container. Science 331, 589-592 (2011).
34. R. Zschoche, D. Hilvert, Diffusion-Limited Cargo Loading of an Engineered Protein Container. J Am Chem Soc 137, 16121-16132 (2015).
35. T. Lin, Z. Chen, R. Usha, C. V. Stauffacher, J. B. Dai, T. Schmidt, J. E. Johnson, The refined crystal structure of cowpea mosaic virus at 2.8 A resolution. Virology 265, 20-34 (1999).

36. T. Lin, A. J. Clark, Z. Chen, M. Shanks, J. B. Dai, Y. Li, T. Schmidt, P. Oxelfelt, G. P. Lomonossoff, J. E. Johnson, Structural fingerprinting: subgrouping of comoviruses by structural studies of red clover mottle virus to 2.4-A resolution and comparisons with other comoviruses. J Virol 74, 493-504 (2000).
37. The PyMOL Molecular Graphics System v. 1.5. (Schrödinger, LLC 2012).
38. E. Krissinel, K. Henrick, Inference of macromolecular assemblies from crystalline state. J Mol Biol 372, 774-797 (2007).
39. S. J. Fleishman, A. Leaver-Fay, J. E. Corn, E. M. Strauch, S. D. Khare, N. Koga, J. Ashworth, P. Murphy, F. Richter, G. Lemmon, J. Meiler, D. Baker, RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. PLoS One 6, e20161 (2011).
40. F. DiMaio, A. Leaver-Fay, P. Bradley, D. Baker, I. Andre, Modeling symmetric macromolecular structures in Rosetta3. PLoS One 6, e20450 (2011).
41. A. Leaver-Fay, M. J. O'Meara, M. Tyka, R. Jacak, Y. Song, E. H. Kellogg, J. Thompson, I. W. Davis, R. A. Pache, S. Lyskov, J. J. Gray, T. Kortemme, J. S. Richardson, J. J. Havranek, J. Snoeyink, D. Baker, B. Kuhlman, Scientific benchmarks for guiding macromolecular energy function improvement. Methods Enzymol 523, 109-143 (2013).
42. M. J. O'Meara, A. Leaver-Fay, M. D. Tyka, A. Stein, K. Houlihan, F. DiMaio, P. Bradley, T. Kortemme, D. Baker, J. Snoeyink, B. Kuhlman, Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. J Chem Theory Comput 11, 609-622 (2015).
43. B. Kuhlman, D. Baker, Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 97, 10383-10388 (2000).
44. W. Sheffler, D. Baker, RosettaHoles2: a volumetric packing measure for protein structure refinement and validation. Protein Sci 19, 1991-1995 (2010).
45. S. J. Fleishman, T. A. Whitehead, E. M. Strauch, J. E. Corn, S. Qin, H. X. Zhou, J. C. Mitchell, O. N. Demerdash, M. Takeda-Shitaka, G. Terashi, I. H. Moal, X. Li, P. A. Bates, M. Zacharias, H. Park, J. S. Ko, H. Lee, C. Seok, T. Bourquard, J. Bernauer, A. Poupon, J. Aze, S. Soner, S. K. Ovali, P. Ozbek, N. B. Tal, T. Haliloglu, H. Hwang, T. Vreven, B. G. Pierce, Z. Weng, L. Perez-Cano, C. Pons, J. Fernandez-Recio, F. Jiang, F. Yang, X. Gong, L. Cao, X. Xu, B. Liu, P. Wang, C. Li, C. Wang, C. H. Robert, M. Guharoy, S. Liu, Y. Huang, L. Li, D. Guo, Y. Chen, Y. Xiao, N. London, Z. Itzhaki, O. Schueler-Furman, Y. Inbar, V. Potapov, M. Cohen, G. Schreiber, Y. Tsuchiya, E. Kanamori, D. M. Standley, H. Nakamura, K. Kinoshita, C. M. Driggers, R. G. Hall, J. L. Morgan, V. L. Hsu, J. Zhan, Y. Yang, Y. Zhou, P. L. Kastritis, A. M. Bonvin, W. Zhang, C. J. Camacho, K. P. Kilambi, A. Sircar, J. J. Gray, M. Ohue, N. Uchikoga, Y. Matsuzaki, T. Ishida, Y. Akiyama, R. Khashan, S. Bush, D. Fouches, A. Tropsha, J. Esquivel-Rodriguez, D. Kihara, P. B. Stranges, R. Jacak, B. Kuhlman, S. Y. Huang, X. Zou, S. J. Wodak, J. Janin, D. Baker, Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol 414, 289-302 (2011).
46. G. L. Hura, A. L. Menon, M. Hammel, R. P. Rambo, F. L. Poole, 2nd, S. E. Tsutakawa, F. E. Jenney, Jr., S. Classen, K. A. Frankel, R. C. Hopkins, S. J. Yang, J. W. Scott, B. D. Dillard, M. W. Adams, J. A. Tainer, Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nat Methods 6, 606-612 (2009).
47. P. S. Huang, Y. E. Ban, F. Richter, I. Andre, R. Vernon, W. R. Schief, D. Baker, RosettaRemodel: a generalized framework for flexible backbone protein design. PLoS One 6, e24109 (2011).

48. D. Schneidman-Duhovny, M. Hammel, J. A. Tainer, A. Sali, Accurate SAXS profile computation and its assessment by contrast variation experiments. Biophys J 105, 962-974 (2013).
49. D. Schneidman-Duhovny, M. Hammel, A. Sali, FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic Acids Res 38, W540-544 (2010).
50. B. L. Nannenga, M. G. Iadanza, B. S. Vollmar, T. Gonen, Overview of electron crystallography of membrane proteins: crystallization and screening strategies using negative stain electron microscopy. Curr Protoc Protein Sci Chapter 17, Unit17 15 (2013).
51. G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees, S. J. Ludtke, EMAN2: an extensible image processing suite for electron microscopy. J Struct Biol 157, 38-46 (2007).
52. M. van Heel, G. Harauz, E. V. Orlova, R. Schmidt, M. Schatz, A new generation of the IMAGIC image processing system. J Struct Biol 116, 17-24 (1996).
53. J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J. Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, A. Cardona, Fiji: an open-source platform for biological-image analysis. Nat Methods 9, 676-682 (2012).
54. W. Kabsch, Xds. Acta Crystallogr D Biol Crystallogr 66, 125-132 (2010).
55. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, Phaser crystallographic software. J Appl Crystallogr 40, 658-674 (2007).
56. A. Vagin, A. Teplyakov, MOLREP: an Automated Program for Molecular Replacement. Journal of Applied Crystallography 30, 1022-1025 (1997).
57. P. D. Adams, P. V. Afonine, G. Bunkoczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr 66, 213-221 (2010).
58. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. Acta Crystallogr D Biol Crystallogr 66, 486-501 (2010).
59. C. Jackel, J. D. Bloom, P. Kast, F. H. Arnold, D. Hilvert, Consensus protein design without phylogenetic bias. J Mol Biol 399, 541-546 (2010).

Chapter 6. Computational Design of DNA-Protein Hybrid Cages and Infinite 2D Layers

ABSTRACT

In nature, a wide variety of self-interacting proteins form ordered symmetric assemblies. Viral capsids, cytoskeleton proteins, and bacterial microcompartments are elegant examples. A long-standing goal in nanotechnology is to create such symmetric assemblies through accurate and controllable engineering of natural proteins. New ideas and methods have led to recent exciting successes in making novel finite assemblies like cube-shaped cages (1–8), but many domains of designs are still left blank. This chapter describes two projects I conducted during my PhD period. The first one aimed to introduce a different type of macromolecules, DNA, into protein cages. The second project aimed to design infinite 2D-crystalline layer by engineering self-interacting symmetric subunits. This chapter will focus on recording the experiments I have carried out and some lessons in engineering I learned over the years for the reference of any future research to be conducted in similar areas.

§6.1 DNA-Protein Hybrid Cages

INTRODUCTION

Recently, protein molecules have been successfully designed for self-assembly based on genetically fused natural oligomers (2, 9). Additionally, computational tools such as Rosetta are bringing the accuracy of these designs to atomic level (3, 5–8). These protein based self-assembling complexes designed from dimeric, trimeric, and tetrameric proteins follow a recently summarized (10, 11) (Fig 6.1) design principle. Specifically, an interaction between a natural dimeric and a trimeric protein can be engineered such that their symmetry axes are maintained at a specific angle. Depending on the geometric relationship between the symmetry axes, the

combined oligomerization of the dimer and trimer subunits thereby drives the formation of an ordered assembly such as an infinite layer or an octahedral cage. Meanwhile, multiple laboratories demonstrated that DNA can serve as the building block for self-assembling biomaterials that form marvelous complexes (12–15). The rigidity and programmability of double-stranded DNA enabled much success in this area. This project incorporated these recent advancements from both fields in order to design ordered hybrid assemblies composed of both DNA and protein elements with the aid of computational tools.

By producing the first materials in which DNA and protein are incorporated together into well-ordered hybrid assembles, successful designs from this project will allow us to exploit their separate advantages. DNA allows better prediction of ideal structure based on sequence than protein (10). This is advantageous because the relative orientation of two proteins joined by a DNA fragment and the distance between them can be predicted without crystal structures. With the advancements in computational technologies, we can start with more diverse proteins as the building block to make a broad range of assemblies with diverse functions, including catalysis and molecular recognition. We expect our new class of designed assemblies to open new avenues for applications in drug delivery, vaccine design, and bioactive materials.

RESULTS AND DISCUSSION

The design targets of the project were hybrid DNA-protein tetrahedra. In order to generate a tetrahedral symmetry, the building blocks needed to have a 2-fold and a 3-fold symmetry with the symmetry axes intersecting at 54.7° (Fig 6.1). Starting from monomeric DNA-binding proteins structures available in PDB, we introduced a 2-fold symmetry using a palindromic DNA sequence. The location and the orientation of the 2-fold symmetry axis was controlled by the length of the double stranded DNA (dsDNA) (Fig 6.2A). Then, a 3-fold

symmetric interface was designed onto the DNA binding protein using Rosetta (3, 16) in a way such that if the designed trimeric interface formed, the final assembly with the palindromic DNA would obey tetrahedral symmetry (Fig 6.2B).

In order to predict the orientation of the 2-fold symmetry axis reliably, one of the requirements for the starting structure was that the DNA must not overwind or unwind and be in near-perfect B-form. After manually curating the available structures from PDB, only three DNA-binding proteins were chosen (Fig 6.2A). Although limited by the number of starting structures, each structure presented multiple opportunities to design the three-fold interface because the relative position between the protein and the 2-fold axis changed by the location of the palindromic site. Following the Rosetta protocol developed by King et. al (3) with slight modifications (Fig 6.2B), a total of 18 constructs were selected for experimental characterization (Table 6.1). Six of the them expressed in soluble forms. Among them, DP11 showed a shift towards slower migrating species on size exclusion chromatography compared to the undersigned protein sequence. However, DP11 did not form a larger species upon mixing with its corresponding palindromic DNA sequence (Fig 6.3).

To further investigate the absence of designed tetrahedral species, I crystallized DP11 without its corresponding palindromic DNA (Table 6.2). To my disappointment, DP11 formed a dimer instead of the designed trimer. The Rosetta designed trimeric interface involves a long α-helix on one subunit and on a loop region from the adjacent subunit. Instead of forming a heterotypic trimeric interface, DP11 formed a homotypic dimeric interface involving only the α-helices (Fig. 6.4). This dimeric interface also blocked the DNA binding site, which was consistent with the observation that DP11 does not bind to its corresponding palindromic DNA. I tested two sets of mutations (DP11R1 & DP11R2), which included charged residues that are

close to each other in the dimeric interface, in the hope that breaking the dimers would favor the trimer formation. However, both sets of mutations behaved the same as DP11, DP11R1 also crystallized in the dimeric form (data not shown).

Looking back on the results from this project, it is not so surprising that DP11 formed dimers instead of trimers. In fact, the protocol developed by King, et al. (14) has higher success rate with interfaces formed by regions with secondary structures. Of the complexes verified by X-ray crystallography or EM, most have two α-helices at the interface (3, 5–8), while only one has two β-sheets (17). The higher success rate with well-ordered interface may be a result of imperfect algorithms to predict unsatisfied H-bonds in buried interfaces. Secondary structures satisfy the backbone H-bond requirement and effectively the mitigate this problem. Future designs should consider incorporating aspects of more recent protocols where H-bonds are predicted with higher accuracy (18) and/or favor interfaces with well-ordered interface (7).

The assembly of the final tetrahedra also relies on the high affinity between the DNA binding protein and its cognate DNA sequence. To check if the wild-type DNA binding protein could bind to the palindromic DNA, I performed an Electrophoretic mobility shift assay (EMSA) with 1XPX and 3W2A and four different palindromic sequences (Fig 6.5). 3W2A bound to its palindromic DNA sequence A102 quite well, causing an obvious shift for most of the DNA. It was also possible to distinguish A102 with one or two 3W2A bound. At perfect molar ratio (3W2A:A102=2:1), the A102 with two 3W2A bound dominated the population. However, this was not true for 1XPX. 1XPX only bound to A104 weakly, while almost does not bind to A101 and A103. Since 1XPX displayed different affinity depending on the position of the palindromic site, it is important in the future to test the affinity of each DNA binding protein to each palindromic DNA sequence. Additionally, the set of designs characterized all contained very

short palindromic sequences. The concern was that longer DNA sequences would lead to highly porous tetrahedra that were unsuitable for crystallization. However, with current advancement in single particle cryo-EM, larger, more porous designs can be readily characterized. This opens new opportunities for future design work.

Another factor important to the success of this project which I did not get an opportunity to explore fully is the ideality of the B-form DNA. If a palindromic DNA deviates from the ideal B-form, the angle of the 2-fold symmetry axis relative to the 3-fold symmetry axis of the DNA binding protein can change a lot. I tried to tackle this issue by including only the structures with near-ideal B-form DNA. However, there are two major draw backs to this strategy: 1) it severely limits the number of good starting structures; 2) the B-form DNA in the selected structures can be a crystallographic artifact. In fact, in all three starting structures, the DNA among adjacent asymmetric units forms pi stacking with each other, and straight DNA helices run through the unit cells. This interaction could favor B-form DNA in crystals and mask DNA distortion by protein binding in solution. Future experiment should consider more sophisticated methods to predict the 2-fold symmetry axis on palindromic DNA in various forms.

MATERIALS AND METHODS

Computational design with Rosetta. DNA-protein hybrid structures with one protein chain or with two proteins chains on an existing palindromic DNA were pulled from the PDB and manually curated for structures with near-ideal B-form DNA. Additionally, only structures with protein chains with 100-180 amino acids and expressed in *E. coli* were selected. The selected structures served as starting points for Rosetta symmetry designs. First, for each starting structure, a series of palindromic DNA sequences were generated to 1) keep the DNA as short as possible for higher rigidity 2) avoid clashes between DNA binding proteins on the same DNA

helix. The predicted 2-fold symmetry axis was translated and rotated to z-axis. Then, the DNA coordinates were removed and only the proteins were used as inputs to the Rosetta matdes_dock protocol using T2.sym (see below) as the symmetry definition file. The design steps follow the established protocols (3, 5) with minor changes. Each docking conformations with more than 100 Cβ contacts were visually checked for clashes between DNA helices as well as between DNA helices and proteins. A total of 433 docked conformations had no visible clashes and were passed to matdes_design. 364 matdes_design results with binding energy (ddG_filt) lower than -12.0, surface complimentary (sc1) higher than 0.55, and interface area (total_sasa) larger than 600 were passed onto an auto_reversion protocol (5) modified for one component designs. The goal of this step was to take away Rosetta introduced mutations that did not contribute strongly to the binding energy, surface complimentary, or the interface area size in order to increase the solubility of the designs and ultimately the design success rate. Output from the auto_reversion step was filtered with the same criteria with the addition that the designs cannot have more than two unsatisfied hydrogen bonds (uhb) or 12 mutations. The filtered output was checked again visually for any unidentified clashes involving DNA. A list of designs to cover diverse starting structures, palindromic sequences, and docked conformations were selected for experimental characterization (Table 6.1).

T2.sym:

```
symmetry_name T2
subunits 12
number_of_interfaces 11
E =
2*B1_1+B1_1:B1_2+B1_1:B2_1+B1_1:B2_2+B1_1:B3_1+B1_1:B3_2+B1_1:B4_1+B1_1:B4_2+B1_1:B5_1
+B1_1:B5_2+B1_1:B6_1+B1_1:B6_2
anchor_residue COM
virtual_coordinates_start
xyz C1 0.000000000000000,0.000000000000000,1.000000000000000 -1.000000000000000,-
0.000000000000000,0.000000000000000 0,0,0
xyz P1 0.000000000000000,0.000000000000000,1.000000000000000 -1.000000000000000,-
0.000000000000000,0.000000000000000 0,0,0
xyz B1_1 0.000000000000000,0.000000000000000,1.000000000000000 -1.000000000000000,-
0.000000000000000,0.000000000000000 0,0,0
```

```
xyz B1_2 0.000000000000000,0.000000000000000,1.000000000000000
1.000000000000000,0.000000000000000,0.000000000000000 0,0,0
xyz C2 1.000000000000000,0.000000000000000,0.000000000000000 0.000000000000000,-
1.000000000000000,0.000000000000000 0,0,0
xyz P2 1.000000000000000,0.000000000000000,0.000000000000000 0.000000000000000,-
1.000000000000000,0.000000000000000 0,0,0
xyz B2_1 1.000000000000000,0.000000000000000,0.000000000000000 0.000000000000000,-
1.000000000000000,0.000000000000000 0,0,0
xyz B2_2 1.000000000000000,0.000000000000000,0.000000000000000
0.000000000000000,1.000000000000000,0.000000000000000 0,0,0
xyz C3 0.000000000000000,1.000000000000000,0.000000000000000 0.000000000000000,-
0.000000000000000,1.000000000000000 0,0,0
xyz P3 0.000000000000000,1.000000000000000,0.000000000000000 0.000000000000000,-
0.000000000000000,1.000000000000000 0,0,0
xyz B3_1 0.000000000000000,1.000000000000000,0.000000000000000 0.000000000000000,-
0.000000000000000,1.000000000000000 0,0,0
xyz B3_2 0.000000000000000,1.000000000000000,0.000000000000000 0.000000000000000,-
0.000000000000000,-1.000000000000000 0,0,0
xyz C4 0.000000000000000,0.000000000000000,-1.000000000000000 1.000000000000000,-
0.000000000000000,0.000000000000000 0,0,0
xyz P4 0.000000000000000,0.000000000000000,-1.000000000000000 1.000000000000000,-
0.000000000000000,0.000000000000000 0,0,0
xyz B4_1 0.000000000000000,0.000000000000000,-1.000000000000000 1.000000000000000,-
0.000000000000000,0.000000000000000 0,0,0
xyz B4_2 0.000000000000000,0.000000000000000,-1.000000000000000 -1.000000000000000,-
0.000000000000000,0.000000000000000 0,0,0
xyz C5 -1.000000000000000,0.000000000000000,0.000000000000000
0.000000000000000,1.000000000000000,0.000000000000000 0,0,0
xyz P5 -1.000000000000000,0.000000000000000,0.000000000000000
0.000000000000000,1.000000000000000,0.000000000000000 0,0,0
xyz B5_1 -1.000000000000000,0.000000000000000,0.000000000000000
0.000000000000000,1.000000000000000,0.000000000000000 0,0,0
xyz B5_2 -1.000000000000000,0.000000000000000,0.000000000000000 0.000000000000000,-
1.000000000000000,0.000000000000000 0,0,0
xyz C6 0.000000000000000,-1.000000000000000,0.000000000000000 0.000000000000000,-
0.000000000000000,-1.000000000000000 0,0,0
xyz P6 0.000000000000000,-1.000000000000000,0.000000000000000 0.000000000000000,-
0.000000000000000,-1.000000000000000 0,0,0
xyz B6_1 0.000000000000000,-1.000000000000000,0.000000000000000 0.000000000000000,-
0.000000000000000,-1.000000000000000 0,0,0
xyz B6_2 0.000000000000000,-1.000000000000000,0.000000000000000 0.000000000000000,-
0.000000000000000,1.000000000000000 0,0,0
virtual_coordinates_stop
connect_virtual JP1 C1 P1
connect_virtual JP1_1 P1 B1_1
connect_virtual JB1_1 B1_1 SUBUNIT
connect_virtual JP1_2 P1 B1_2
connect_virtual JB1_2 B1_2 SUBUNIT
connect_virtual JC2 C1 C2
connect_virtual JP2 C2 P2
connect_virtual JP2_1 P2 B2_1
connect_virtual JB2_1 B2_1 SUBUNIT
connect_virtual JP2_2 P2 B2_2
connect_virtual JB2_2 B2_2 SUBUNIT
connect_virtual JC3 C1 C3
connect_virtual JP3 C3 P3
connect_virtual JP3_1 P3 B3_1
connect_virtual JB3_1 B3_1 SUBUNIT
connect_virtual JP3_2 P3 B3_2
connect_virtual JB3_2 B3_2 SUBUNIT
connect_virtual JC4 C1 C4
connect_virtual JP4 C4 P4
connect_virtual JP4_1 P4 B4_1
```

```
connect_virtual JB4_1 B4_1 SUBUNIT
connect_virtual JP4_2 P4 B4_2
connect_virtual JB4_2 B4_2 SUBUNIT
connect_virtual JC5 C1 C5
connect_virtual JP5 C5 P5
connect_virtual JP5_1 P5 B5_1
connect_virtual JB5_1 B5_1 SUBUNIT
connect_virtual JP5_2 P5 B5_2
connect_virtual JB5_2 B5_2 SUBUNIT
connect_virtual JC6 C1 C6
connect_virtual JP6 C6 P6
connect_virtual JP6_1 P6 B6_1
connect_virtual JB6_1 B6_1 SUBUNIT
connect_virtual JP6_2 P6 B6_2
connect_virtual JB6_2 B6_2 SUBUNIT
set_dof JP1 x(20) angle_x
set_jump_group JGP JP1 JP2 JP3 JP4 JP5 JP6
set_jump_group JGB JB1_1 JB1_2 JB2_1 JB2_2 JB3_1 JB3_2 JB4_1 JB4_2 JB5_1 JB5_2 JB6_1
JB6_2
```

Protein sequences. All genes were synthesized by IDT and inserted between the NcoI and BamHI sites on pET-M11, retaining a cleavable TEV site on the N-terminus of the proteins with the exception of DP17 and DP18. DP17 and DP18 are inserted between the BamHI and XhoI sites on pET-SUMO (p4955), a gift from the UCLA Protein Expression Core Technology Center.

```
DP01:
MKHHHHHHPMSDYDIPTTENLYFQGASSTLTPMHLRKAKLMFFWSRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQME
KTARQMVTEGIKTPDDLLIAGDSELYRVLNLHYNRNNHIEVPENFRQVVMSTLASFFLAIAGGKDTEQSWKKEIQKIISTLDTPVP
EYFKSPNFLEQ
DP02:
MKHHHHHHPMSDYDIPTTENLYFQGASSTLTPMHLRKAKLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQME
KYARQAVTEGIKTPDDLLIAGDSEVYRVLNEHYNRNNHIVVPMAFLMVVAITLLSFFRAIQGGKDTEQSWKKEIYKIISRLDQPVP
EYFKSPNFLEQ
DP03:
MKHHHHHHPMSDYDIPTTENLYFQGASSTLTPMHLRKAKLMFFWVRYPSSAVLKMYFPDIEFNKNNTAQLVKWFSNFREFYYIQME
KYARQAVTEGIKTPDDLLIAGDSSLYRVLNEHYNRNNHIVVPLNFLMVVAATLLEFFIAISGGKDTEQSWKKSIYKLISRLDAPVP
EYFKSPNFLRQ
DP04:
MKHHHHHHPMSDYDIPTTENLYFQGASSTLTPMHLRKAKLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQME
KYARQAVTEGIKTPDDLLIAGDSTLYRVLNLHYNRNNHIVVPLHFLLVVEATLASFKAAIQGGKDTEQSWKKSIYKLISAMDAPVP
EYFKSPNFLEQ
DP05:
MKHHHHHHPMSDYDIPTTENLYFQGASRTLTPMHLRKAKLMFFWVRYPSSAVLKMYFPDIVFEKNNTAQLVKWFSNFREFYYIQME
KYARQAVTEGIKTPDDLLIAGDSELYRVLNLHYNRNNHITVPANFLEVVDSTLASFFKAIQGGRDTEQSWKKSIYKIISEMDDPVP
EYFKSPSFLSA
DP06:
MKHHHHHHPMSDYDIPTTENLYFQGASSTLTPMHLRKAKLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQME
KYARQAVTEGIKTPDDLLIAGDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLTEFFLAIAGGADTANSWKKIIYMEISRMDDPVP
EYFKSPNFLEQ
DP07:
MKHHHHHHPMSDYDIPTTENLYFQGASIRELGIGLLKLRALGMSYKDIALLENLSRAKVTRAFQAASVPQAIISLFPIASELNFND
YKILFNYAKGLTKANEALRSTLPILKEEIKDLDTNLPPDIYKKEILNIIKKSKN
DP08:
MKHHHHHHPMSDYDIPTTENLYFQGASIRELGIGLLKKKALGMSYKEIALLENLSRAKVTRAFQAASVPQAIISLFPIASELNFND
YKILFNYAKGLTKANESLKSTLPILKEEIKDLDTNLPPDIYKKEILNIIKKSKN
```

DP09:
MKHHHHHHPMSDYDIPTTENLYFQGASSTLTPMHLRKAKLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQME
KYARQAVTEGIKTPDDLLIAGDSELYIVLAKHYNRNMHISIPANFLAVVETTLLSFFMAIQGGKDTEQSWKKSIYKIISRADDPVP
EYFKSPNFLEQ
DP10:
MKHHHHHHPMSDYDIPTTENLYFQGASSTLTPMHLRKAKLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQME
KYARQAVTEGIKTPDDLLIAGDSELMMVLNTHYNRENHISVPANFLAVVVTTLLSFMAAIQGGKDTEQSWKKKIYKIISRADDPVP
EEFKSPNFLEQ
DP11:
MKHHHHHHPMSDYDIPTTENLYFQGASSTLTPMHLRKAKLMFFWMRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQME
KYARQAVTEGIKTPDDLLIAGDSELYRVLNLHYNRNNHIEVPQNFRTVVLITLLLFFKAIQGGKDTEQSWKKSIYDIISTMDDPVP
EYFKSPNFLEQ
DP11R1:
MKHHHHHHPMSDYDIPTTENLYFQGASSTLTPMHLRKAKLMFFWKRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQME
KYARQAVTEGIKTPDDLLIAGDSELYRVLNLHYNRNNHIEVPQNFRTVVLITELLFFKAIQGGKDTEQSWKKSIYDIISTMDDPVP
EYFKSPNFLEQ
DP11R2:
MKHHHHHHPMSDYDIPTTENLYFQGASSTLTPMHLRKAKLMFFWKRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQME
KYARQAVTEGIKTPDDLLIAGDSELYRVLNLHYNRNNHIEVPQNFRTVVDITLDLFFKAIQGGKDTEQSWKKSIYDIISTMDDPVP
EYFKSPNFLEQ
DP12:
MKHHHHHHPMSDYDIPTTENLYFQGASIRELGIGLNFLKVSGMSYKDIAKKENLSRAKVTRAFQAASVPQEIISLFPIASELNFND
YKILFNYYKGLEKINASLSSTLPLLKLAIADLDTNLPPDIYKKEILNLIKKSKM
DP13:
MKHHHHHHPMSDYDIPTTENLYFQGASIRELGIGLNFLKVSGMSYKDIAKKENLSRAKVTRAFQAASVPQEIISLFPIASELNFND
YKILFNYYKGLEKANLSLSSTLPLLKLAIEDLDTNLPPDIYKKEILNLIKKSAM
DP14:
MKHHHHHHPMSDYDIPTTENLYFQGASIRELGIGLNFLKVSGMSYKDIAKKENLSRAKVTRAFQAASVPQEIISIFPKTSELNFND
YKILFNYYKGLEKANLLLSAMIPKLMEEIKDLDANLPPDIYKKEILNIIKKSKN
DP15:
MKHHHHHHPMSDYDIPTTENLYFQGASIRELGIGLNFLKVSGMSYKDIAKKENLSRAKVTRAFQAASVPQEIISLFPKASELNFND
YKILFNYYKGLEKANLLLSTTLPKLMEEILDLDTNLPPDIYKKEILNIIKKSKN
DP16:
MKHHHHHHPMSDYDIPTTENLYFQGASIRELGIGLNFLKVSGMSYKDIAKKENLSRAKVTRAFQAASVPQEIIALFPKASELNFND
YKILFNYYKGLEKANLRLETTLPKLMEEILDLDTNLPPDIYKKEILNIIKKSKN
DP17:
MGSSHHHHHHSSGLVPRGSHMASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSL
RFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGTVRADEEELIVAALASHTQGGESETFAKRAIESLVKKLKEKKDELDSLITA
ITTNGAHPSKCVTIQRTLDGRLQVAGRKGFPHVIYARLWTMPRLRKNSLKHVKYCQYAFDLKCDSVCVNPYHYEVKAQ
DP18:
MGSSHHHHHHSSGLVPRGSHMASMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSL
RFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGTVSNDDEKLIVASLMSHRQGGESETFAKRAIESLVKKLKEKKDELDSLITA
ITTNGAHPSKCVTIQRTLDGRLQVAGRKGFPHVIYARLWTMPMLEKNSLKHVKYCQYAFDLKCDSVCVNPYHYEVQAQ

Protein expression and purification 10mL of overnight LB culture with BL21(DE3) cells were used to inoculated 1 L of TB withn 5052 autoinduction sugar (19) supplemented with 0.4% glycerol and the appropriate antibiotics. Cultures were shaken at 37 °C for four hours and then 20 °C for two days before harvesting. Harvest cell pellets were resuspended in 50mM Tris–HCl, 500mM NaCl, final pH 7.0 and lysed by sonication. Cell lysate was cleared by centrifugation at 20,000 g for 20 min. Cleared lysate was loaded onto HisTrap columns (GE healthcare) equilibrated with the same buffer and eluted with a linear gradient over 10 column volumes to

the same buffer supplemented with 500 mM imidazole. Size exclusion chromatography (SEC) was performed on Superdex 200 10/30 GL columns (GE Healthcare) in 50mM Tris pH 7.0, 500mM NaCl. All purification steps were performed at 4 °C.

Crystallographic data collection, molecular replacement, and model refinement. Post-SEC DP11 fractions were concentrated and filtered with 0.22 um filter immediately before setting up crystallization trays at 4 °C. The best DP11 crystal was obtained in a 200 nL drop with 630 uM of DP11 and 0.2M NH_4Cl, 0.1M sodium citrate pH 5.0, 20% (v/v) PEG 6000 mixed at 1:1 ratio. Diffraction data sets were collected at the Advanced Photon Source (APS) beamline 24-ID-C equipped with a Pilatus-6M detector. The crystals showed diffraction to 2.54 Å. The XDS/XSCALE package (20) was used to integrate, reduce, and scale the data. The data were reduced in $P3_121$ space group symmetry. We used the PHASER program (21) to determine the structure by molecular replacement, with one 1XPX chain as the search model. Molecular replacement yielded a single solution. After the solution was obtained, the structure was refined in iterative runs using the PHENIX program (22) and manual curation in COOT (23). The final R and R_{free} values were 22.4% and 25.0%.

EMSA. Palindromic DNA sequences were synthesized as single stranded DNA oligos from IDT. They were resuspended to 100 uM in TE buffer and annealed to double stranded DNA in water bath slowly cooling from 100 °C to room temperature.10 uL of annealed DNA was mixed with 1XPX or 3W2A at a gradient of concentration (ranging from protein being in excess to DNA being in excess). Mixture of protein and DNA was kept at 4 °C overnight and loaded onto Tris-glycine native PAGE gel and ran at 200 V until DNA loading dye reached the gel front. The native gel is then stained with Stains-All (Sigma) solution.

§6.2 Infinite 2D Protein Layers

INTRODUCTION

Current success in symmetry based protein design is mainly limited to finite structures (reviewed in (15)), while extended materials like two-dimensional (2D) layers have enjoyed much less success (1, 17, 24). They have many potential applications, including as crystallization scaffolds and assembly platforms. Such materials represent the next targets for designed assemblies. The fundamental principles we learn from making 2D crystalline layers, such as how to choose building blocks and control geometry over long ranges, will also advance other important goals in nanotechnology, such as engineering 3D crystals.

A particularly important application of ordered symmetric assemblies in structural biology is to use them as scaffolds in electron microscopy (EM) samples, especially for membrane proteins. Membrane proteins account for 20-30% of all gene products (25), and they play essential roles in many biological functions. However, structural information is particularly hard to obtain for membrane proteins. Today, only a small percentage of all the structures deposited in the Protein Data Bank are membrane proteins. NMR studies are limited by protein size, and growing well-ordered 3D membrane protein crystals for X-ray crystallography is notoriously difficult. EM allows for the structural study of membrane proteins in a natural lipid environment, and is becoming an attractive alternative to X-ray methods.

Previously, 2D crystalline layers based on DNA Holliday junctions have been developed and examined by EM (26). It is therefore possible to utilize such 2D arrays to selectively order particles in certain views, especially those underrepresented in free-flowing particles, to overcome the common preferred orientation issue in cryo-EM. Another possible application for 2D crystalline array is to use them as scaffolds for forming 2D membrane protein crystals. To

date, a few membrane proteins have been imaged with near-atomic resolution based on samples with varying degrees of order (27). Samples with increasing long-range order open up the possibility of electron crystallography, which offers better resolutions (28). However, electron crystallography depends on screening for a condition that leads to membrane protein 2D crystalline layers, generally limiting its application to membrane proteins with a natural tendency to form high density packings or 2D crystalline layers *in vivo* (28–30). Otherwise, the screening process for 2D crystalline layers can be as labor-intensive as that for 3D crystals (31). This project designed a range of 2D crystalline layers built upon protein symmetry, which can be used as scaffolds to generate symmetrical arrays of membrane proteins (Fig. 6.6). In this way, the project has the potential to overcome a substantial bottleneck in structural biology by providing a facile and powerful means to high resolution structural analysis of membrane proteins.

RESULTS AND DISCUSSION

Based on ideas developed earlier in our laboratory, self-assembling proteins can be designed to make ordered assemblies such as symmetric cages, infinite 2D crystalline layers, or even 3D crystals (Table 3.1). To do so, the protein building block must have (at least) two types of symmetric interfaces in a specific geometric orientation. For example, for a subunit with 4-fold and 2-fold symmetries to form a finite cube-shaped assembly, its 4-fold axis must intersect with its 2-fold axis at 45˚(10, 32). If the two axes are perpendicular but non-intersecting, then the subunit will assemble to make a 2D crystalline array in the $p42_12$ layer group symmetry (Fig 6.7A).

First, I utilized a computational approach of designing symmetric protein assemblies originated in the Yeates laboratory. We computationally screened for pairs of natural oligomers with two different types of symmetries, looking for those that fulfill specific geometric

requirements of the target symmetry group when genetically fused together through their terminal α-helices (2, 4). This approach has been proved successful in making finite symmetric assemblies-- 12-subunit tetrahedra and 24-subunit octahedra (10, 33). Compared to other existing computational methods, our strategy does not require intensive prior knowledge in programing and scripting. This method is applicable to a wide range of starting materials, as it can utilize all oligomers with at least one terminal α-helix. Additionally, the designs can be generated through easy genetic fusion of natural oligomers, minimizing the changes to their wild-type properties. The above characters of our strategy make it more adoptable by biologists and compatible with various downstream biological experiments, which frequently rely on proteins being close to their native states.

Building on the experience we have gained from making finite symmetric self-assemblies, I pursued two different strategies in parallel to build protein-based 2D crystalline layers. 1) I modified the geometry restrains in the programs originally written for the helix-fusion strategy for cage designs to create 2D arrays. To increase screening efficiency, I added a step to automatically identify clashes between oligomers, and re-formatted the output file to generate more subunits for visual inspection. The case of fusing a cyclic (C_4) tetramer to a dimer is illustrated in (Fig 6.7B, i). 2) In a new innovation, I used protein-DNA hybrid complexes to build symmetric assemblies. To make a 2D layer, the first symmetry was introduced by fusing a monomeric DNA binding protein to a cyclic oligomer, and the required 2-fold symmetry via a palindromic DNA sequence (Fig 6.7B, ii). To make the 2D arrays applicable to diverse membrane proteins, I included in my designs a wide range of building blocks, and engineered 2D layers in as many symmetry groups as possible with a wide range of dimensions (Fig 6.8). Many

designs were quite porous and had extensive space for adding other moieties in downstream applications.

Thirteen constructs were selected for experimental test based on the following criteria: 1) cover a wide range of layer groups; 2) use diverse starting oligomers 3) contain short α-helical linkers (Table 6.3 and Fig 6.9A). Among them, hDP07 and hDP19 had good solubility and showed structures corresponding to one unit cell (Fig 6.9B & C). hDP07 was further optimized in two rounds. First, I took a closer check at the fusion design visually and identified a potential clash at Ala 112 between the N-terminal tetrameric subunit (PDB 1CUK) and the α-helical linker. Ala 112 was mutated to Gly on hDP07R1 (Fig 6.9D). Second, through literature searching I found that 1CUK has two domains that are flexible relative to each other. The conformational change affects the geometry in hDP07. To fix this, I mutated the hydrophilic residues lining the interface between the two domains to hydrophobic residues, in order to favor the closed conformation used in designing the layer. After this set of mutations, structures corresponding to one unit cell were spotted more frequently and can extend beyond one unit cell (Fig 6.9E). After another round of mutations to stabilize the connecting α-helix (hDP07R4 and hDRP07R5NL), I performed a nickelated monolayer lipid assisted 2D crystallization to help His-tagged layer designs to form 2D layers (34). However, there was still no visible layers under negative stain EM (data not shown).

Based on the preliminary results I described above, I make three suggestions for future efforts for this project. First, all starting materials should go through close vetting in literature to avoid proteins with known alternative conformations. Even a small population of alternative conformations can greatly reduce the chances of obtaining a 2D layer, since it takes many individual binding events for extended assemblies to form. Second, future designs should

consider using more than one contact points between adjacent oligomers. The designs I tested were all based on cyclic oligomers and there is only one connection through an α-helix between adjacent oligomers. Potentially this gave a lot of room for distortions around the α-helix. One way to limit this issue is to use dihedral oligomers in designs, which provides two helical linkers between adjacent oligomers. In fact, in favorable cases, short linkers without designed symmetry restraints can confine dihedral oligomers into 2D layers (1). Additionally, less porous designs may have less conformational flexibility.

Last but not least, we are in dire need for a more efficient and high-through put way of screening designed layers. Layer group symmetries usually require the symmetry axes of component cyclic oligomers to be *not* intersecting, which is a much less stringent requirement compared to being intersecting for forming cubic symmetries. This results in a much larger number of oligomer pairs that fulfill the geometry requirements for most layer group symmetries. Testing purified proteins one by one by searching negative stain grids severely limits the number of constructs that can be screened. A high throughput method would then increase the chance for finding a good layer simply by screening larger libraries of designs. Furthermore, most initial hits are unlikely to show large patches of 2D layers on negative stain grids. It then became time-consuming to screen large areas on each grid. A self-check mechanism, perhaps lowered stability or susceptibility to proteolysis, should be adopted to screen out unassembled proteins and enrich for correctly assembled 2D layers.

MATERIALS AND METHODS

Computational design. The symmetry based design process for 2D protein layers are similar to designing cubic cages (2, 4, 32) with some changes. C2, C3, and C4 homo-oligomers with subunits that had chain length 80-320 amino acids long, and that expressed *E. coli*, were

downloaded from the PDB. Each structure was checked against the DSSP database for presence of α-helical termini. An α-helical terminus must be at least six residues long and had no more than five unstructured amino acids beyond it. The symmetry axes of the oligomers with terminal α-helices were calculated as previously described (35). Next, pair-wise fusion between oligomers with compatible helical termini were performed computationally in Chimera (36) as previously described (2, 32). Specifically, the first six amino acids in an ideal poly-alanine α-helix was aligned to the terminal six residues on the stationary oligomer based on CA atoms. Then CA atoms of the terminal six residues on the moving oligomer were aligned onto the ideal α-helix and slid one residue at a time. At each position, the angle and distance between the two symmetry axes were recorded.

For designs with p4 layer group symmetry, C4 tetramer and C2 dimer fusion pairs with symmetry axes within 2° of being parallel and at least 30 Å in distance were selected. Among these, if there were no CA atoms within 5 Å of each other in the fusion pair, then pair was translated and rotated such that the origin of the p4 layer group sits at the origin of the P4 space group with the same unit cell length in a & b and the 4-fold symmetry axes align with each other. Clashes between asymmetric units were checked manually be expanding the layer in PyMol. Designs with p42(1)2, p6, and p321 were selected similarly. p42(1)2 designs composed of C4 tetramers and C2 dimers whose symmetry axes were within 0.5° of being orthogonal and at least 30 Å in distance. The origin of p42(1)2 layer group was positioned at (1/2, 1/2, 0) in P42(1)2 space group. p6 designs composed of C3 trimers and C2 dimers whose symmetry axes were within 2° of being parallel and at least 30 Å in distance. The origin of p6 layer group was positioned at the origin of P6 space group. p321 designs composed of C3 trimers and C2 dimers

whose symmetry axes were within 0.5° of being orthogonal and at least 30 Å in distance. The origin of p321 layer group was positioned at the origin of P321 space group.

Cloning and protein sequences. All constructs were synthesized from IDT and cloned - into the NdeI and HindIII sites on pET-28a, except for hDP07(R1/R2/R4/R5NL) and hDP12, which were cloned into pET-22b. hDP18 & hDP19 were cloned into pET-M11.

hDP03:
MGSSHHHHHHSSGLVPRGSHMAKEHSIRELGIGLNFLKVSGMSYKDIAKKENLSRAKVTRAFQAASVPQEIISLFPIASELNFNDY
KILFNYYKGLEKANESLSSTLPILKEEIKDLDTNLPPDIYKKEILNIIKKSAQANDLEDNMETLNDNLKVIEKADNAAQVKDALTK
MRAAALDAQKATPPKLEDKSPDSPEMHDFRHGFDILVGQIHDALHLANEGKVKEAQAAAEQLKTTCNACHQKYR

hDP06:
MGSSHHHHHHSSGLVPRGSHMAKEHSIRELGIGLNFLKVSGMSYKDIAKKENLSRAKVTRAFQAASVPQEIISLFPIASELNFNDY
KILFNYYKGLEKANESLSSTLPILKEEIKDLDTNLPPDIYKKEILNIIKKSAQRKEQRKEQRKNHFDVISAFIKSIRGSDPDATLY
WLANMVEAGEDPNFIFRRLLISACEDIGLADPNAIVVVQSCCDAFDRVGFPEGLFFLSQASLYLAISPKSNSTKSIFKAMEAIKAT
NVSLVPNHLKNNASNYLNPHNYQGKWLQQEYLPTDLQGIKFWKPKDSGWEKNKYED

hDP07:
MIGRLRGIIIEKQPPLVLIEVGGVGYEVHMPMTCFYELPEAGQEAIVFTHFVVREDAQLLYGFNNKQERTLFKELIKTNGVGPKLA
LAILSGMSAQQFVNAVEREEVGALVALPGIGKKTAERLIVEMKDRFKGLHGDLFTPAADLVLTSPASPATDDAEQEAVARLVALGY
KPQEASRMVSKIARPDASSETLIREALRAALQRAEQRAEATLEQHLEDTMKNPSIVGVLCTDSQGLNLGCRGTLSDEHAGVISVLA
QQAAKLTSDPTDIPVVCLESDNGNIMIQKHDGITVAVHKMAKLAAALEHHHHHH

hDP07R1:
MIGRLRGIIIEKQPPLVLIEVGGVGYEVHMPMTCFYELPEAGQEAIVFTHFVVREDAQLLYGFNNKQERTLFKELIKTNGVGPKLA
LAILSGMSAQQFVNAVERSEVGALVGLPGIGKKTAERLIVEMKDRFKGLHGDLFTPAADLVLTSPASPATDDAEQEAVARLVALGY
KPQEASRMVSKIARPDASSETLIREALRAALQRAEQRAEATLEQHLEDTMKNPSIVGVLCTDSQGLNLGCRGTLSDEHAGVISVLA
QQAAKLTSDPTDIPVVCLESDNGNIMIQKHDGITVAVHKMAKLAAALEHHHHHH

hDP07R2:
MIGRLRGIIIEKQPPLVLIEVGGVGYEVHMPMTCFVELPEAGQEAIVFTHFVVREDAQLLYGFNNKQERTLFKELIKTNGVGPKLA
LAILSGMSAQQFVNAVERSEVGALVGLPGIGKKTAERLIVEMKDRFKGLHGDLFTPAADLVLTSPASPATDDAEQEAVARLVALGY
KPQEASRMVSKIARPDASSLTLIIEALRAALQRAEQRAEATLEQHLEDTMKNPSIVGVLCTDSQGLNLGCRGTLSDEHAGVISVLA
QQAAKLTSDPTDIPVVCLESDNGNIMIQKHDGITVAVHKMAKLAAALEHHHHHH

hDP07R4:
MIGRLRGIIIEKQPPLVLIEVGGVGYEVHMPMTCFVELPEAGQEAIVFTHFVVREDAQLLYGFNNKQERTLFKELIKTNGVGPKLA
LAILSGMSAQQFVNAVERSEVGALVGLPGIGKKTAERLIVEMKDRFKGLHGDLFTPAADLVLTSPASPATDDAEQEAVAALVALGY
KPQEASRMVSKIARPDASSLTLVIEALRAALQRAEQRVEQELEQHLEDTMKNPSIVGVLCTDSQGLNLGCRGTLSDEHAGVISVLA
QQAAKLTSDPTDIPVVTLESDNGNIMIQKHDGITVAVHKMAKLAAALEHHHHHH

hDP07R5NL:
MIGRLRGIIIEKQPPLVLIEVGGVGYEVHMPMTCFVELPEAGQEAIVFTHFVVREDAQLLYGFNNKQERTLFKELIKTNGVGPKLA
LAILSGMSAQQFVNAVERSEVAALVGLPGIGKKTASRLIVEMKDRFKGLHGDLFTPAADLVLTSPASPATDDAEQEAVAALVALGY
KPQEASRMVSKIARPDASSLTLVIEALRAALQRAEQRAEQELEQHLEDTMKNPSIVGVLCTDSQGLNLGCRGTLSDEHAGVISVLA
QQAAKLTSDPTDIPVVTLESDNGNIMIQKHSGITVAVHKMAKLAAALEHHHHHH

hDP09:
MGSSHHHHHHSSGLVPRGSHMDYPRDLIGYGNNPPHPHWPGDARIALSFVLNYEEGGERCVLHGDKESEAFLSEMVAAQPLQGVRH
MSMESLYEYGSRAGVWRLLKLFKRRNVPLTVFAVAMAAQRNPEVIRAMVADGHEICSHGYRWIDYQYMDEAQEREHMLEAIRILTE
LTGQRPVGWYTGRTGPNTRRLVMEEGGFLYDSDTYDDDLPYWDPASTAEKPHLVIPYTLDTNDMRFTQVQGFNNGEQFFQYLKDAF
DVLYEEGATAPKMLSIGLHCRLIGRPARMAALERFIQYAQSHDKVWFARREDIARHWHRHAAAMENQKMQEPLVYRRILLTVDEDD
NTSSERAFRYATTLAHDYDVPLGICSVLESEDINIFDSLTPSKIQAKRKHVEDVVAEYVQLAEQRGVNQVEPLVYEGGDVDDVILE
QVIPEFKPDLLVTGADTEFPHSKIAGAIGPRLARKAPISVIVVR

hDP10:
MGSSHHHHHHSSGLVPRGSHMLEDLKRQVLEANLALPKHNLVTLTWGNVSAVDRERGVFVIKPSGVDYSIMTADDMVVVSIETGEV
VEGAKKPSSDTPTHRLLYQAFPSIGGIVHTHSRHATIWAQAGQSIPATGTTHANYFYGTIPCTRKMTDAEINGEYEWETGNVIVET
FEKQGIDAAQMPGVLVHSHGPFAWGKNAEDAVHNAIVLEEVAYMGIFCRQLAPQLPDMQQTLLNKHYLAAQQAAELVRDRQELIDA

RKKELKAYMMMGVTAIKPLYDSDVNGSNKQAAKEILKAMRFESDGYFFAYDSQGINTLHAIKPSLEGKNLYDLKDENGVAVIAGLI
DASQKGDGFLYFSWHKPTINAQAPKLGYAEYLQKWDWVLGTGIYIDD

hDP12:
MIGRLRGIIIEKQPPLVLIEVGGVGYEVHMPMTCFYELPEAGQEAIVFTHFVVREDAQLLYGFNNKQERTLFKELIKTNGVGPKLA
LAILSGMSAQQFVNAVEREEVGALVKLPGIGKKTAERLIVEMKDRFKENKQKQEKQEQDRKLADAHDQMLELAELLTDVLIKNVPG
LSEKHAEDASIYMAKNRAVFAAAFKNNATALSELSEPKLAAALEHHHHHH

hDP13:
MGSSHHHHHHSSGLVPRGSHMMNIVLARIDDRFIHGQILTRWIKVHAADRIIVVSDDIAQDEMRKTLILSVAPSNVKASAVSVSKM
AKAFHSPRYEGVTAMLLFENPSDIVSLIEAGVPIKTVNVGGMRFENHRSQITKSVSVTEQDIKAFETLSDKGVKLELRQLPSDASE
DFVQILRNVTKEQAKENAKALKDIMHILEDMKVGVFATLDEYGNPHARHAHITAANEEGIFFMTSPETHFYDQLMGDQRVAMTAIS
EEGYLIQVVRVEGTARPVENDYLKTVFADNPYYQHIYKDESSDTMQVFQIYAGHGFYHSLTQGHKYIFSIGQGEHSEVRAL

hDP14:
MGSSHHHHHHSSGLVPRGSHMDRMYELEYPSPEVSGQTAGGPTLIVALQGYADAGHAVESSSSHLMDALDHRLIASFNNDELIDYR
SRRPVVVIEHNEVTSMDELNLGLHVVRDNDNKPFLMLSGPEPDLRWGDFSNAVVDLVEKFGVENTICLYAAPMTVPHTRPTVVTAH
GNSTDRLKDQVSLDTRMTVPGSASLMLEKLLKDKGKNVSGYTVHVPHYVSASPYPAATLKLLQSIADSADLNLPLLALERDAEKVH
RQLMEQTEESSEIQRVVGALEQQYDSELERYRNRHAAASDDEAVTALALSAAKGNGRALEAFIKATQQDVWRFVAYLSDVGSADDL
TQETFLRAIGAIPRFSARSSARTWLLAIARHVVADHIRHVRSRPRTTRGARPEHLIDGD

hDP15:
MGSSHHHHHHSSGLVPRGSHMMNIVLARIDDRFIHGQILTRWIKVHAADRIIVVSDDIAQDEMRKTLILSVAPSNVKASAVSVSKM
AKAFHSPRYEGVTAMLLFENPSDIVSLIEAGVPIKTVNVGGMRFENHRSQITKSVSVTEQDIKAFETLSDKGVKLELRQLPSDASE
DFVQILRNVTFEVIASKIKDSINRDEYKTGMLMPNETALQEIYSSSRTTIRRAVDLLVEEGLVVRKNGVGLYVQPKLTAQNILEMT
GVMKNDTNENLKKDIKDFYIRKAGKFYAEIFGMKENELVYSIKFVQKSEHGATLDRLILPLGLYPDLQAKDFQIINIIELVNSGKY
KLFELEQELQLILAGNEQIKNMHLNENDPVFKLSSVFYAENDMPIAIQYHYEDAESTKYVVDFN

hDP16:
MGSSHHHHHHSSGLVPRGSHMPFITVGQENSTSIDLYYEDHGTGQPVVLIHGFPLSGHSWERQSAALLDAGYRVITYDRRGFGQSS
QPTTGYDYDTFAADLNTVLETLDLQDAVLVGFSTGTGEVARYVSSYGTARIAKVAFLASLEPFLLKTDDNPDGAAPQEFFDGIVAA
VKADRYAFYTGFFNDFYNLDENLGTRISEEAVRNSWNTAASGGFFAAAAAPTTWYTDFRADIPRIDVPALILHGTGDRTLPIENTA
RVFHKALPSAEYVEVEGAPHGLLWTHAEEVNTALLAFLAKALEAQKQKASFGSRMEESIRKTVTENTVVIYSKTWCSYCTEVKTLF
KRLGVQPLVVELDQLGPQGPQLQKVLERLTGQHTVPNVFVCGKHIGGCTDTVKLNRKGDLELMLAE

hDP17:
MGSSHHHHHHSSGLVPRGSHMDSSSRQYREKLKQVEEYMQYRKLPSHLRNKILDYYEYRYRGKMFDERHIFREVSESIRQDVANYN
CRDLVASVPFFVGADSNFVTRVVTLLEFEVFQPADYVIQEGTFGDRMFFIQQGIVDIIMSDGVIATSLSDGSYFGEICLLTRERRV
ASVKCETYCTLFSLSVQHFNQVLDEFPAMRKTMEEIAVRRAQAAEEAQREKDTRISKKMETMGIYFATPEWVALNGHRGPSPGQLK
YWQNTREIPDPNEDYLDYVHAEKSRLASEEQILRAATSIYGAPGQAEPPQAFIDEVAKVYEINHGRGPNQEQMKDLLLTAMEMKH

hDP18:
MKHHHHHHPMSDYDIPTTENLYFQGAMVRRIEDHISFLEKFINDVNTLTAKLLKDLQTEYGISAEQSHVLNMLSIEALTVGQITDK
QGVNDAAVSRRVKKLLNAELVKLEKPDSNTDQRLKIIKLSNKGKKYIKERKAIMSHIASDMTSDFDSKEIEKVRQVLEIIDYRIQS
YTSKLAQNMRARLYAAFRQVGEDLFAQGLISATAGNFSVRTSGGFLITKSGVQKARLTPEDLLEVPLSGPIPEGASVESVVHREVY
RRTGAKALVHAHPRVAVALSFHLSRLRPLDLEGQHYLKEVPVLAPKTVSATEEAALSVAEALREHRACLLRGHGAFAVGLKDAPEE
ALLEAYGLMTTLEESAQILLYHRLWQGAGPAL

hDP19:
MKHHHHHHPMSDYDIPTTENLYFQGAMESPLGSDLARLVRIWRALIDHRLKPLELTQTHWVTLHNIHQLPPDQSQIQLAKAIGISQ
PSLVRTLDQLEDKGLISRQTCASDRRAKRIKLTEKADALIAEMEEVIHKTRGEILAGISSEEIELLIKLIAKLEHNIMELAENDRN
KLARQIIDTCLEMTRLGLNQGTAGNVSVRYQAGMLITPTGIPYEKLTESHIVFIDGNGNHEEGKLPQSEWRFHMAAYQSRPDANAV
VHNHAVHCTAVSILNRSIPAIHYMIAAAGGNSIPCAPYATFGTRELSEHVALALKNRKATLLQHHGLIACEVNLEKALWLAHEVEV
LAQLYLTTLAITDPVPVLSDEEIAVVLEKFKTFGLRIEE

Protein expression and purification were carried out similarly as in §6.1. with slight modifications. Cells were resuspended in 50mM Tris pH 7.1, 300 mM NaCl, 5% glycerol, 1mM β-cyclodextrin and lysed by passing through Emulsiflex three times. Elutions from HisTrap columns (GE Healthcare) in a linear gradient to 500mM imidazole were pulled and dialyzed overnight into 50mM Tris pH 7.1, 50mM NaCl, 1mM DTT, 5% glycerol, 1mM β-cyclodextrin,

1mM DTT. Dialyzed sample was loaded onto HiTrap Q column (GE Healthcare) equilibrated with 50mM Tris pH 7.1, 1mM DTT, 5% glycerol, 1mM β-cyclodextrin and eluted with a linear gradient to 1M NaCl. HiTrap Q eluent was concentrated and loaded onto HiLoad 16/600 Superdex 200 column equibrated in 10mM Tris pH 7.0, 300mM NaCl,1mM β-cyclodextrin, 1mM DTT. Fractions right after the void were collected for negative stain EM analysis.

Negative stain EM. 5 uL of sample was applied to a formvar supported carbon film on 300-mesh copper grid. The excessive sample was blotted away with filter paper after 1min, washed twice with filtered water and then stained with 2% uranyl acetate for 30 sec. Air-dried grids were imaged at room temperature with FEI Tecnai T12 electron microscope equipped with Gatan 2kX2k CCD camera.

Nickelated monolayer lipid assisted 2D crystallization was performed similarly as described previously (34, 37) with minor modifications. Nickel-charged 1,2-dioleoyl-sn-glycero-3-[(N-(5-amino-1-carboxypentyl)iminodiacetic acid)succinyl] (DOGS-NTA Ni^{2+}) and *L*-α-phosphatidylcholine (PC) was mixed freshly in 1:1 hexane:chloroform to final concentrations of 12.5 ug/mL and 50 ug/mL, respectively. 15 uL of freshly purified hDP07R4 (5 mg/mL) and hDP07R5NL (0.2 mg/mL) in 10 mM Tris pH 7.0, 300 mM or 600 mM NaCl, 1mM DTT, 1mM β-cyclodextrin was spotted onto a Teflon plate and overlaid with 0.3 uL of the lipid mixture. The lipid layers were lifted onto formvar supported carbon film without glow discharge, once after 3 hours in a wetted chamber at RT and again after transferring to 4°C for another 15 hours. Grids were then stained with 2% uranyl acetate for 1 min prior to imaging.

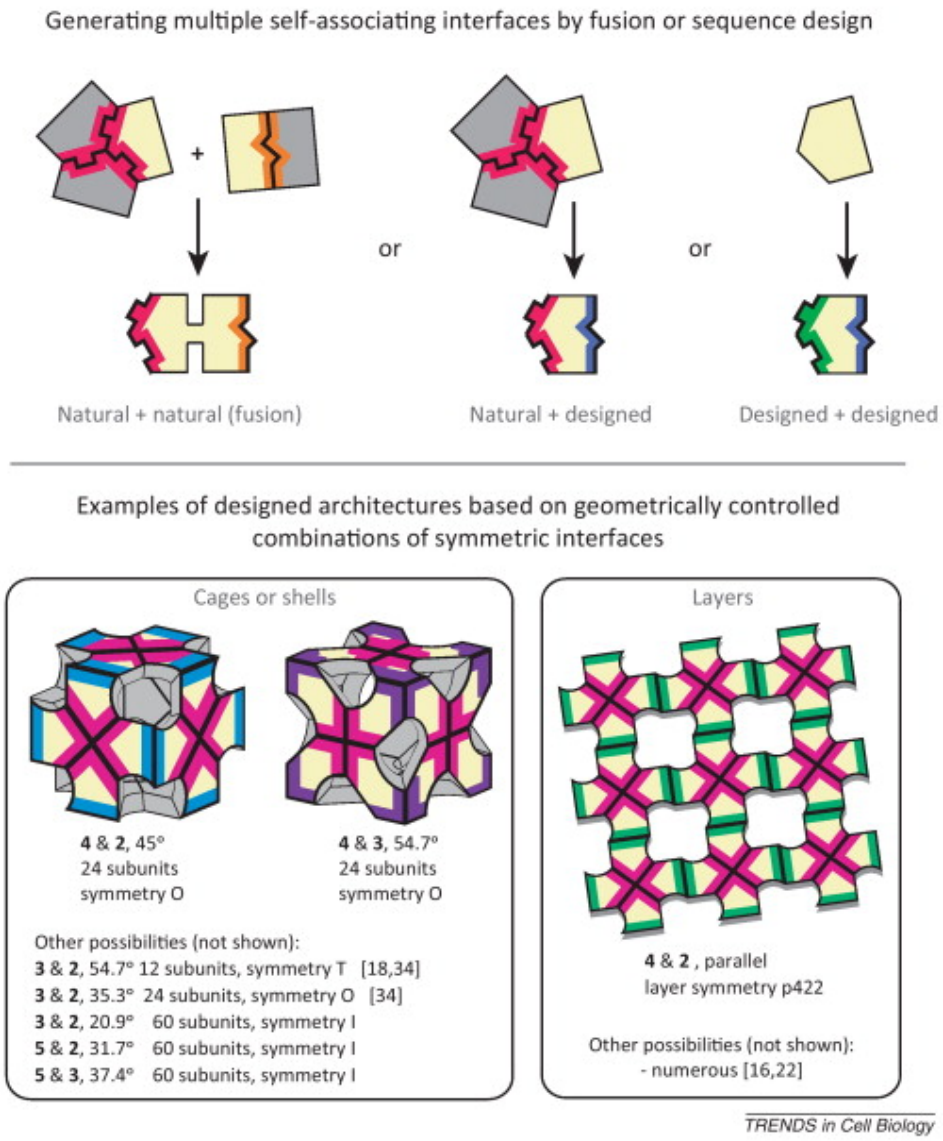


Figure 6.1 Principles for designing ordered protein assemblies.

A general requirement for engineering complex self-assembling structures is to create a protein molecule bearing multiple distinct self-associating interfaces (top). This can be achieved through the use of natural oligomeric proteins or computational sequence design, separately or in combination. The multiple interfaces must be combined according to specific geometric rules if defined structures are to be created (bottom). The geometric design requirements are given for some example assemblies. In each case, the two types of symmetry elements are noted in bold, together with the angle they must form. The symmetries of the resulting assemblies are given (T:

tetrahedral; O: octahedral; I: icosahedral). For 2D layers and 3D crystals (not shown), rules for constructing the full range of possible symmetries are articulated in Chapter 3 (Table 3.1). This figure is adapted from (12). Reprint with permission from Elsevier (license number 4425181171954).

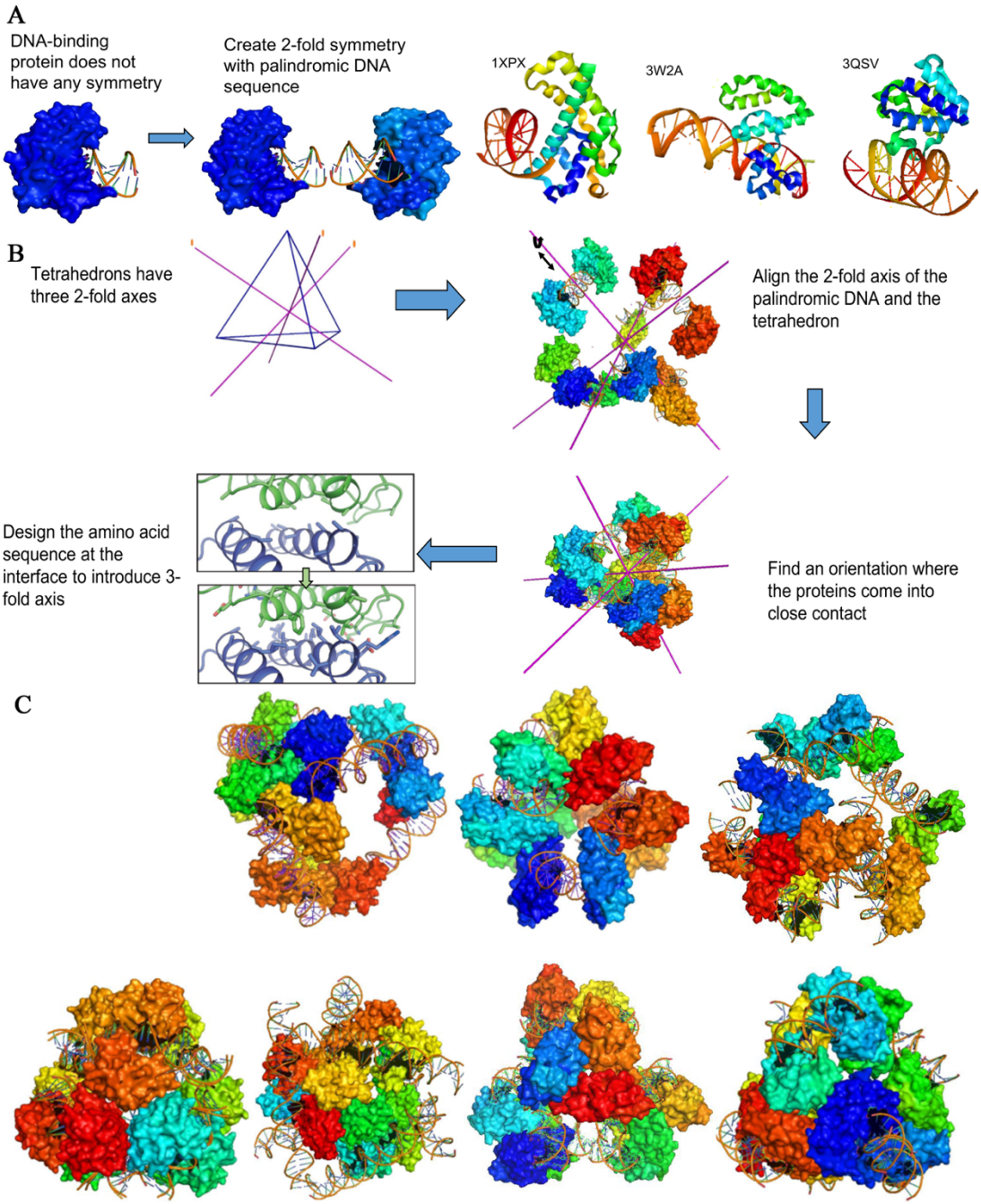


Figure 6.2. Design process of DNA-protein hybrid tetrahedra.

A. 2-fold symmetry axes are introduced onto a DNA binding protein with its cognate DNA when a palindromic site is introduced into the DNA sequence. The three DNA binding proteins chosen as starting points are shown with their PDB IDs. B. Symmetry based Rosetta design protocol. The 2-fold symmetry axis of each palindromic DNA sequence is aligned onto the 2-fold axes of

a tetrahedron. The DNA binding proteins are allowed to slide and rotate around the 2-fold axes. Tetrahedral symmetry is satisfied at each of these radial displacement and rotation angle. Once a designable position (enough residues on one subunit are in close proximity to the other subunit) is found, Rosetta then calculates for the amino acid sequence at the interface that would in theory favor the formation of the new 3-fold interface. C. Partial list of designed DNA-protein tetrahedra. Each protein chain in the tetrahedra is shown in a different color.

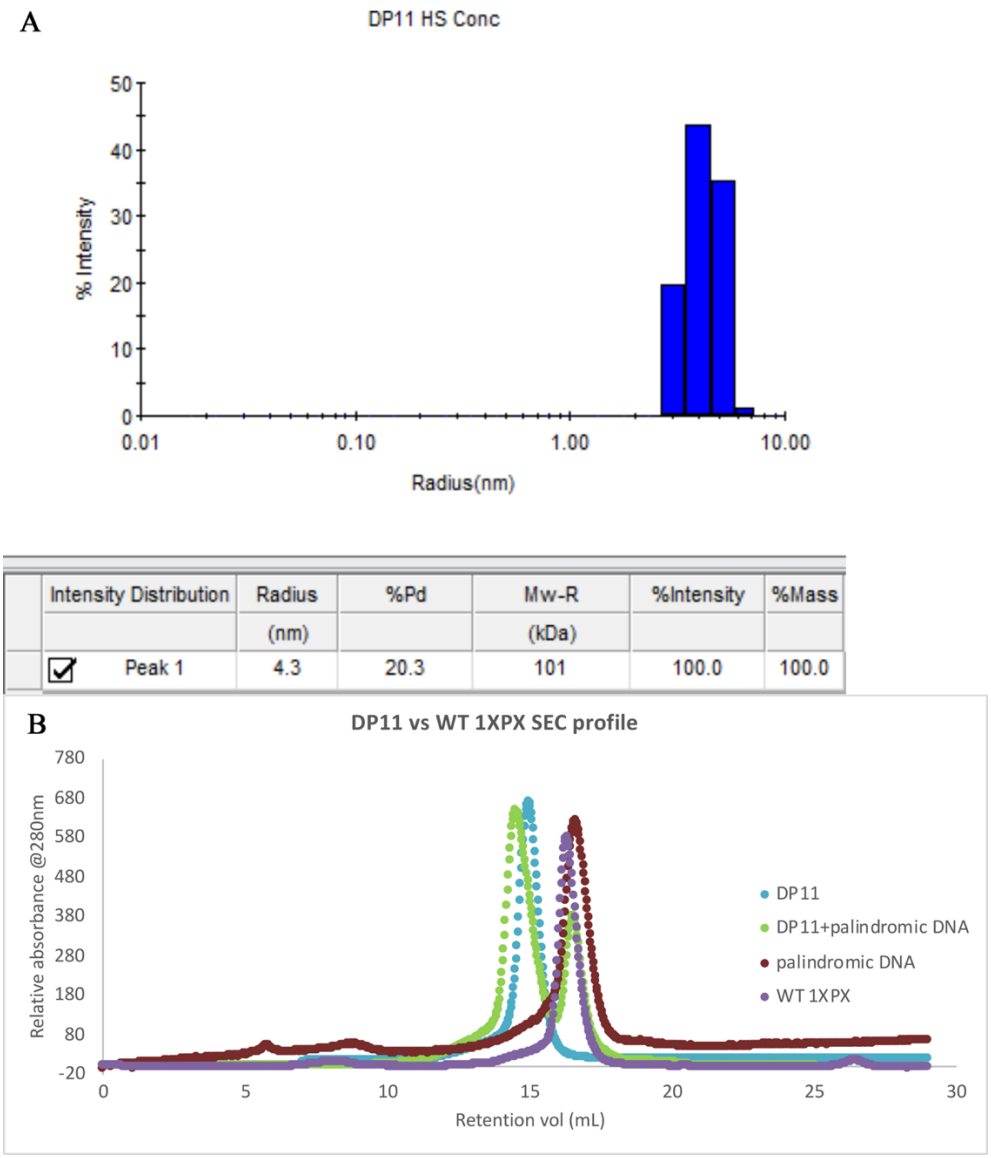


Figure 6.3. Biochemical characterization of DP11.

A. Dynamic light scattering experiment of purified DP11 alone, plotted based on mass percentage. B. Size exclusion chromatography profile. DP11+palindromic DNA: DP11 subunit: dsDNA = 2:1. WT 1XPX is the starting point of DP11 Rosetta design and contains no Rosetta introduced mutations.

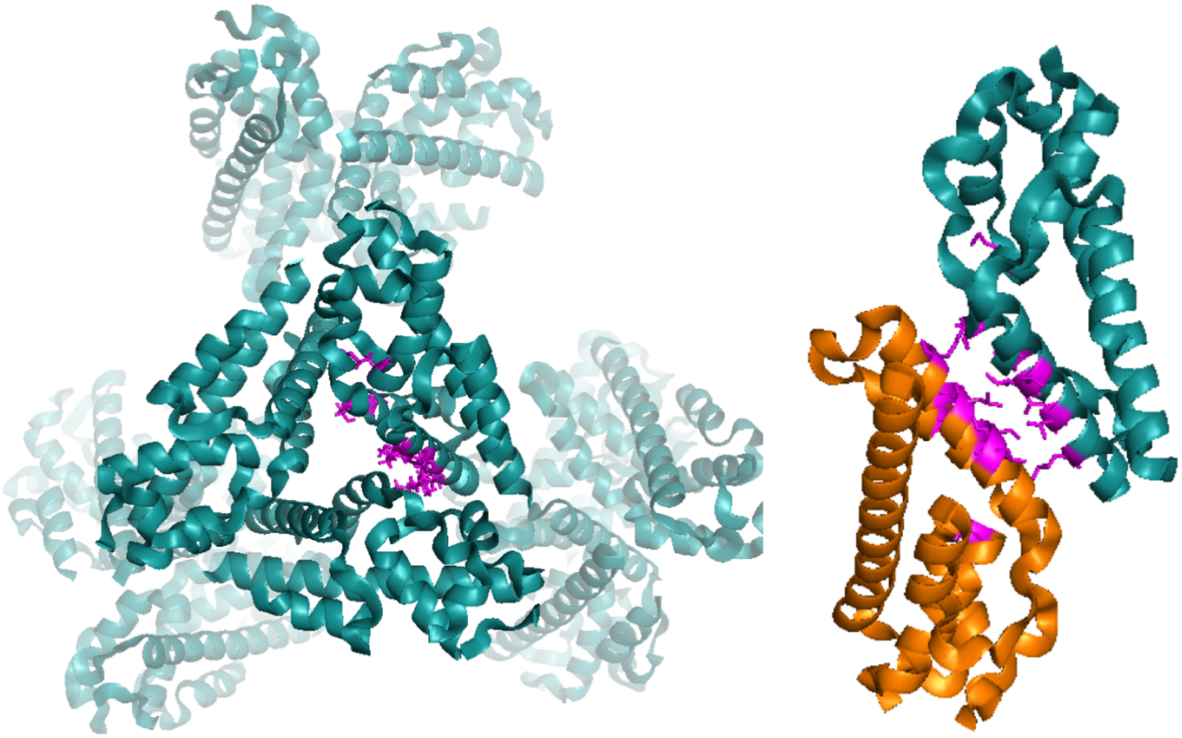


Figure 6.4. Designed DP11 trimeric interface (left) vs crystalized dimeric interface (right, each subunit shown in different color).

Mutations introduced by Rosetta are labeled in magenta and shown in sticks.

Protein	DNA	sequence
1XPX	A101	CAGGCATGCTAGCATGCCTG
3W2A	A102	ATACTTCATTTCATACTGAATTCAGTATGAAATGAAGTAT
1XPX	A103	CAGGCATGCATGCCTG
1XPX	A104	AGCATGCCTGCAGGCATGCT

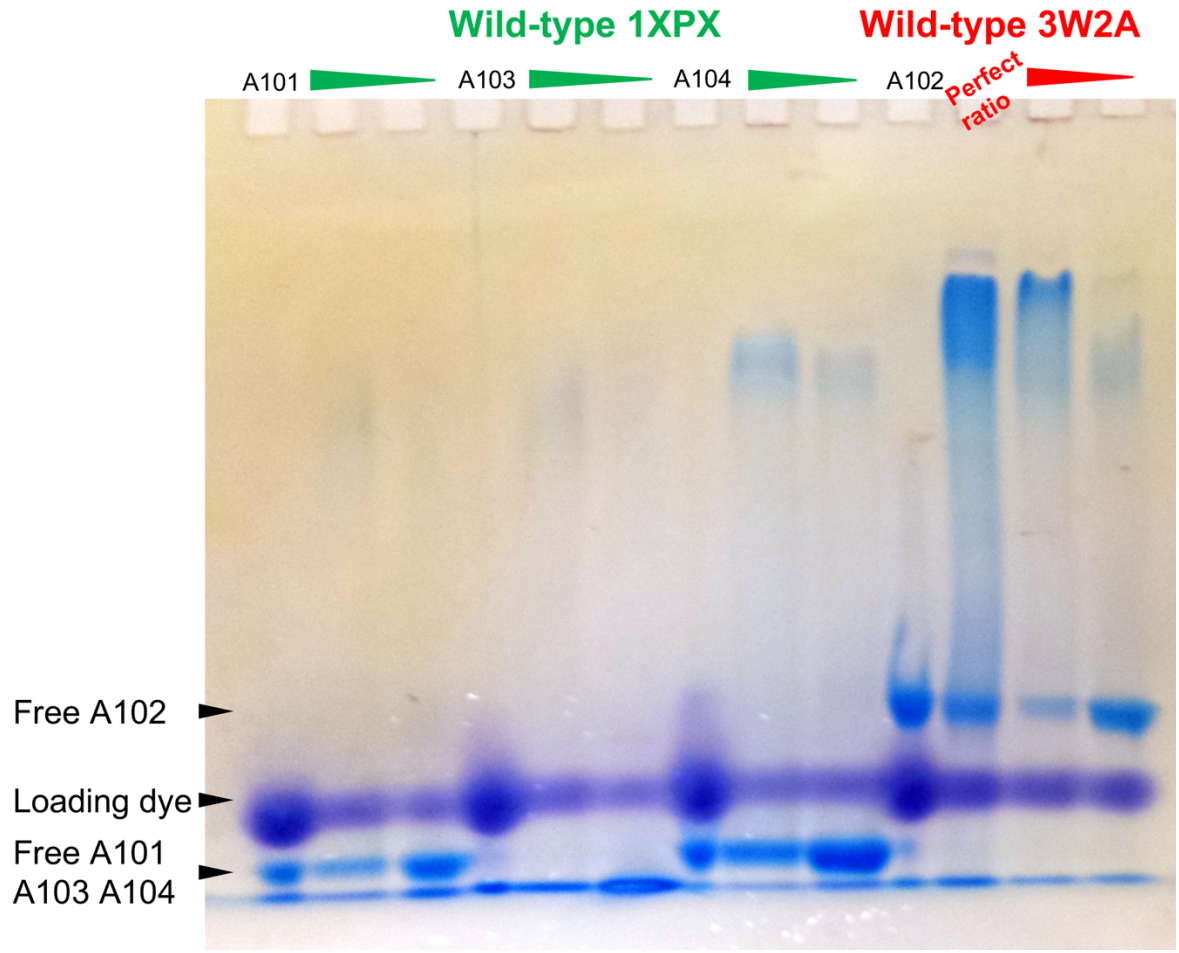


Figure 6.5 EMSA with wild-type 1XPX and 3W2A.

The palindromic site on the DNA sequence is registered with a red star sign. Green and red wedges represent 1XPX and 3W2A protein concentrations, respectively.

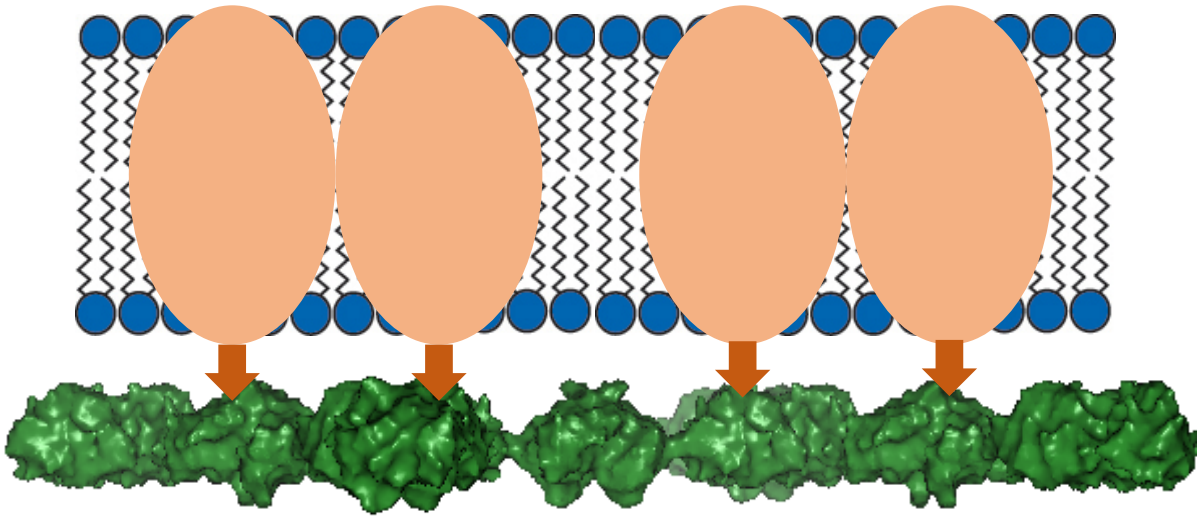


Figure 6.6. Examples of using designed 2D layer as scaffolds for membrane proteins. Generic representation. Membrane proteins are shown as orange ovals in lipid membrane. Specific attachments (brown arrows) are designs between membrane protein and 2D crystalline layer (green).

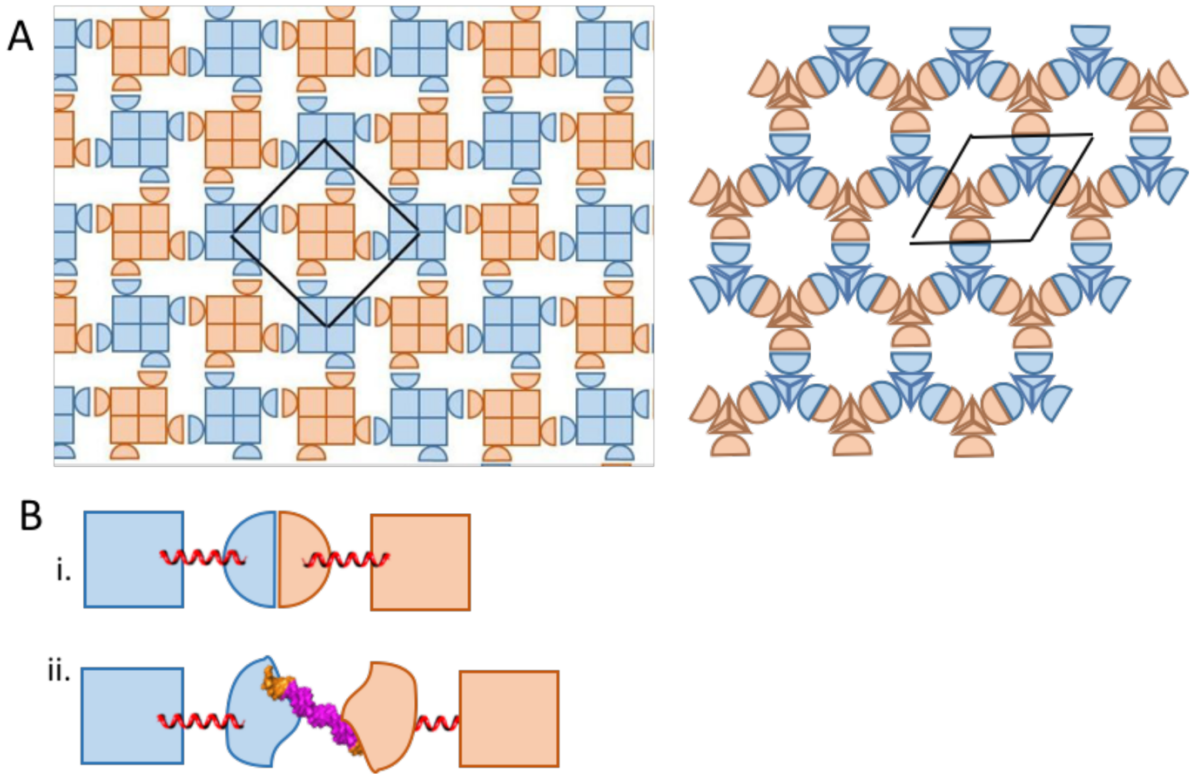


Figure 6.7. Design principles for 2D protein layers.

A) Left: A $p42_12$ layer symmetry group formed with tetrameric (squares) and dimeric (semicircles) interactions. Right: A p321 layer symmetry group formed with trimeric (triangles) and dimeric (semicircles) interactions. The two types of symmetry axes in both layer symmetry groups are perpendicular but not intersecting. Subunits related by 2-fold axes are shown with different colors. The unit cells are outlined in black. B) Two approaches for generating the kinds of components shown in A. The cases using tetramers are illustrated. The same strategy applies to trimers. Genetically fuse a natural tetramer and i) a dimer or ii) a monomeric DNA-binding protein through terminal α-helices (red). In ii, the 2-fold axis is introduced by palindromic DNA (magenta and orange).

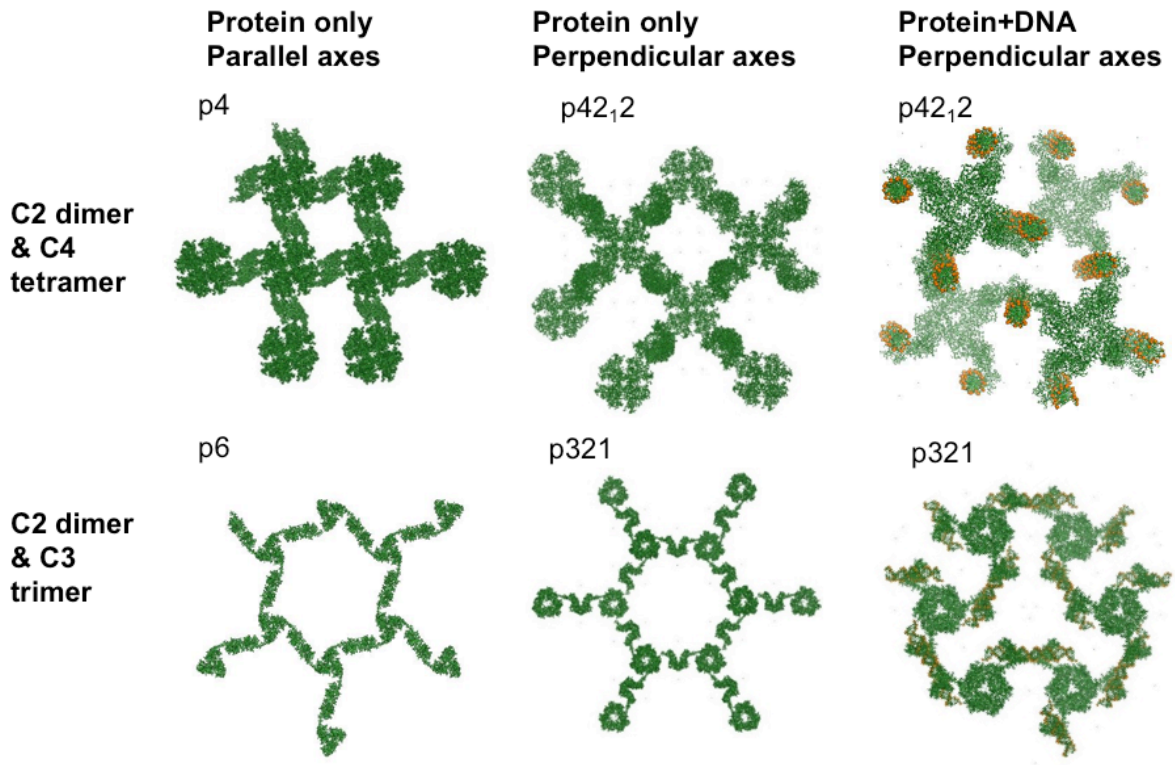


Figure 6.8. Examples of designed 2D layer based on helix-fusion strategy.

Only part of the layer is shown for each example. Using different combinations of protein oligomers or protein oligomers plus palindromic DNA, we can design 2D layer in several layer groups. There is a wide range of unit cell sizes and porousness of the designs. For clarity the structures are not shown to scale. The unit cell lengths from left to right in the top row are 11.7 nm, 19.8 nm, and 17.5 nm. The bottom row from left to right: 31.7 nm, 30.4 nm, and 13.9 nm. Protein is shown in green. DNA is shown in green with the backbone phosphate in orange.

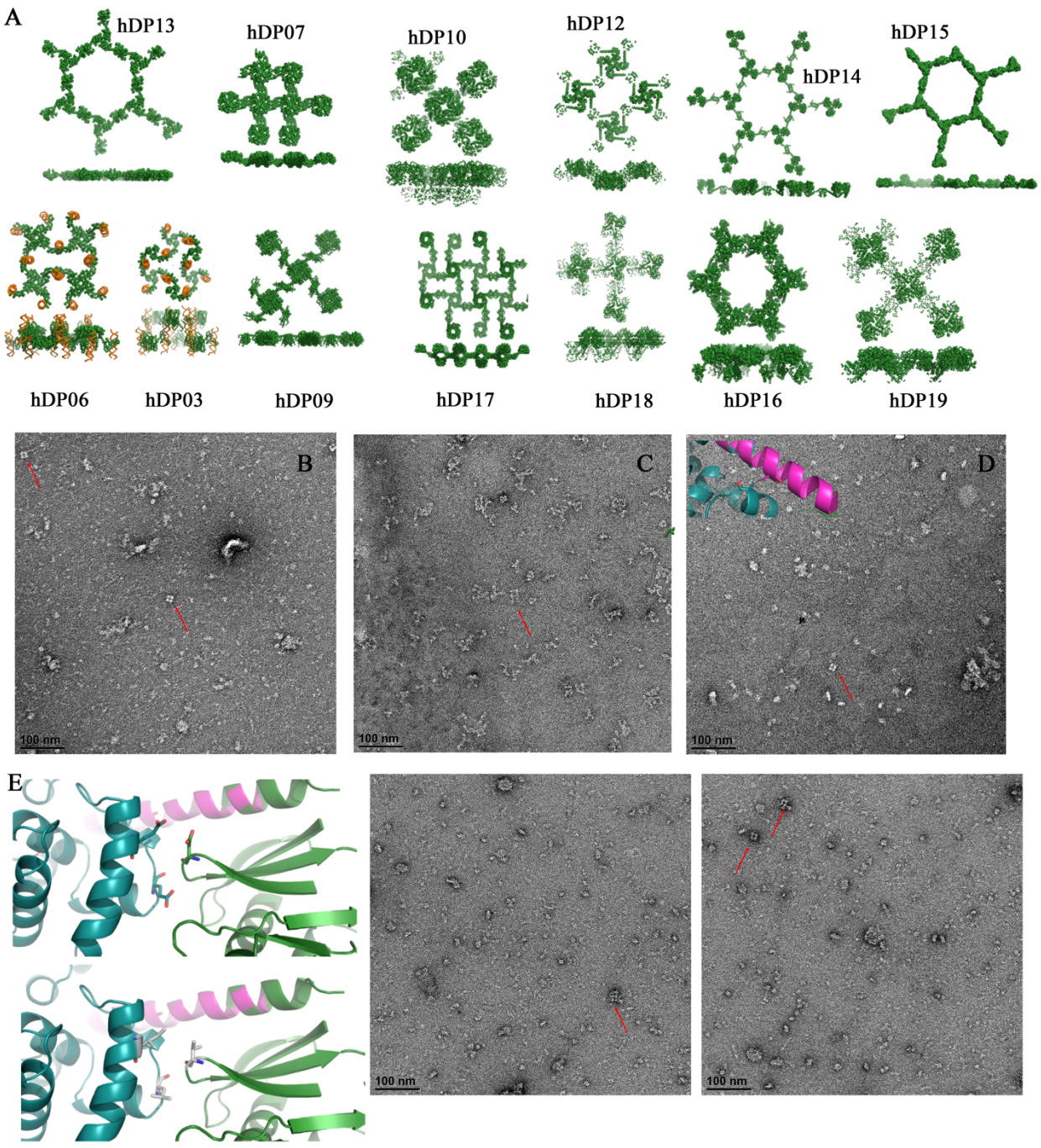


Figure 6.9. Experimentally tested 2D designs.

A. Experimentally tested designs. Color scheme is the same as in Fig. 5.8. For each design, a view through the layer plane is shown above a view along the layer plane. Not shown to scale. B & C. Negative stain EM of hDP07 and hDP19, respectively. D. hDP07R1 negative stain. Inset

structure shows a potential clash site on computationally generated design, between the connecting α-helix (magenta) and the N-terminal tetramer (1CUK, teal). Residues in close contact are in sticks. E. Left top, wild type hydrophilic residues (in sticks) line the interface between two domains (teal and green) on the tetrameric subunit 1CUK. Left bottom, hDP07R2 construct with hydrophobic mutations at the same interface (in sticks). Middle and right, negative stain images of hDP07R2. Red arrows mark particles of interest.

Table 6.1 List of experimentally tested DNA-protein hybrid tetrahedra.

Design name	Starting PDB ID	Palindromic DNA sequence
DP01	1XPX	CAGGCATGCTAGCATGCCTG
DP02	1XPX	CAGGCATGCATGCCTG
DP03	1XPX	CAGGCATGCATGCCTG
DP04	1XPX	CAGGCATGCATGCCTG
DP05	1XPX	CAGGCATGCATGCCTG
DP06	1XPX	AGCATGCCTGCAGGCATGCT
DP07*	3W2A	ATACTTCATTTCATACTGAATTCAGTATGAAATGAAGTAT
DP08*	3W2A	ATACTTCATTTCATACTGAATTCAGTATGAAATGAAGTAT
DP09	1XPX	CAGGCATGCATGCCTG
DP10	1XPX	CAGGCATGCATGCCTG
DP11*	1XPX	CAGGCATGCTAGCATGCCTG
DP12*	3W2A	ATACTTCATTTCATACTGATCAGTATGAAATGAAGTAT
DP13	3W2A	ATACTTCATTTCATACTGATCAGTATGAAATGAAGTAT
DP14	3W2A	AAAGGGATTTCAGTATGAAATTTCATACTGAAATCCCTTT
DP15	3W2A	AAAGGGATTTCAGTATGAAATTTCATACTGAAATCCCTTT
DP16	3W2A	AAAGGGATTTCAGTATGAAATTTCATACTGAAATCCCTTT
DP17*	3QSV	TGCAGTCTAGACTGCA
DP18*	3QSV	TGCAGTCTAGACTGCA

* Soluble constructs

Table 6.2 DP11 crystallography statistics

	DP11		
Data Collection			
Space group	P 31 2 1		
Cell dimensions			
a, b, c (Å)	104.17	104.17	56.93
α, β, γ (˚)	90	90	120
Resolution (Å)	2.54 (2.60)		
Rmerge (%)	10.5 (72.5)		
Mean I/σ	10.61 (1.97)		
Completeness (%)	99.1 (97.6)		
Multiplicity	7.42 (6.88)		
Refinement			
Resolution range (Å)	100-2.54 (2.78-2.54)		
No. reflections	11596 (2790)		
Rwork/Rfree (%)	22.4/25.0		
No. atoms	1177		
Average B factors	65.81		
Protein residues	143		
Ramachandran favored (%)*	94.6		
Ramachandran outliers (%)*	5.4		

* Values given by PROCHECK

Table 6.3 List of experimentally tested 2D layer designs

Name	Trimer/tetramer	Dimer	+/-DNA	Symmetry	Unit cell (nm)	Folded?
hDP03	3DE9	3W2A	+	p321	13.9	√
hDP06	3CTD	3W2A	+	p42(1)2	17.5	X
hDP07	1CUK	3MSH	-	p4	11.7	√
hDP09	1Z7A	3FG9	-	p4	14	X
hDP10	1K0W	4EXO	-	p4	8.9	√
hDP12	1D8L	3ERM	-	p42(1)2	16.8	√
hDP13	1BLE	2HHZ	-	p6	25.4	√
hDP14	2P90	2O7G	-	p6	24.4	X
hDP15	1BLE	3EDP	-	p6	31.7	X
hDP16	1BRT	3RHC	-	p6	17	√
hDP17	2PTM	2Z7J	-	p42(1)2	18.7	X
hDP18	2FLF	3KP7	-	p4	10	√
hDP19	1E4C	3DEU	-	p4	10.6	√

REFERENCES

1. Sinclair JC, Davies KM, Vénien-Bryan C, Noble MEM (2011) Generation of protein lattices by fusing proteins with matching rotational symmetry. *Nat Nano* 6(9):558–562.

2. Lai Y-T, Cascio D, Yeates TO (2012) Structure of a 16-nm Cage Designed by Using Protein Oligomers. *Science* 336(6085):1129–1129.

3. King NP, et al. (2012) Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. *Science* 336(6085):1171–1174.

4. Lai Y-T, et al. (2014) Structure of a designed protein cage that self-assembles into a highly porous cube. *Nat Chem* 6(12):1065–1071.

5. King NP, et al. (2014) Accurate design of co-assembling multi-component protein nanomaterials. *Nature* 510(7503):103–108.

6. Bale JB, et al. (2015) Structure of a designed tetrahedral protein assembly variant engineered to have improved soluble expression. *Protein Science* 24(10):1695–1701.

7. Bale JB, et al. (2016) Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* 353(6297):389–394.

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

9. Lai Y-T, Tsai K-L, Sawaya MR, Asturias FJ, Yeates TO (2013) Structure and Flexibility of Nanoscale Protein Cages Designed by Symmetric Self-Assembly. *J Am Chem Soc* 135(20):7738–7743.

10. Lai Y-T, King NP, Yeates TO (2012) Principles for designing ordered protein assemblies. *Trends in Cell Biology* 22(12):653–661.

11. Yeates TO, Liu Y, Laniado J (2016) The design of symmetric protein nanomaterials comes of age in theory and practice. *Current Opinion in Structural Biology* 39:134–143.

12. Douglas SM, et al. (2009) Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature* 459(7245):414–418.

13. Seeman NC (2010) Nanomaterials Based on DNA. *Annual Review of Biochemistry* 79(1):65–87.

14. Bai X, Martin TG, Scheres SHW, Dietz H (2012) Cryo-EM structure of a 3D DNA-origami object. *PNAS* 109(49):20012–20017.

15. Martin TG, et al. (2016) Design of a molecular support for cryo-EM structure determination. *PNAS* 113(47):E7456–E7463.

16. DiMaio F, Leaver-Fay A, Bradley P, Baker D, André I (2011) Modeling Symmetric Macromolecular Structures in Rosetta3. *PLoS ONE* 6(6):e20450.

17. Gonen S, DiMaio F, Gonen T, Baker D (2015) Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. *Science* 348(6241):1365–1368.

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

19. Studier FW (2005) Protein production by auto-induction in high-density shaking cultures. *Protein Expression and Purification* 41(1):207–234.

20. Kabsch W (2010) XDS. *Acta Crystallogr D Biol Crystallogr* 66(Pt 2):125–132.

21. McCoy AJ, et al. (2007) Phaser crystallographic software. *J Appl Crystallogr* 40(Pt 4):658–674.

22. Adams PD, et al. (2010) *PHENIX* : a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D Biological Crystallography* 66(2):213–221.

23. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66(Pt 4):486–501.

24. Matthaei JF, et al. (2015) Designing Two-Dimensional Protein Arrays through Fusion of Multimers and Interface Mutations. *Nano Lett*. doi:10.1021/acs.nanolett.5b01499.

25. Krogh A, Larsson B, Heijne G von, Sonnhammer ELL (2001) Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *J MOL BIOL* 305:567–580.

26. Malo J, et al. (2005) Engineering a 2D Protein–DNA Crystal. *Angewandte Chemie International Edition* 44(20):3057–3061.

27. Goldie KN, et al. (2014) Cryo-electron Microscopy of Membrane Proteins. *Electron Microscopy*, Methods in Molecular Biology., ed Kuo J (Humana Press), pp 325–341.

28. Raunser S, Walz T (2009) Electron Crystallography as a Technique to Study the Structure on Membrane Proteins in a Lipidic Environment. *Annual Review of Biophysics* 38(1):89–105.

29. Henderson R, et al. (1990) Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *Journal of Molecular Biology* 213(4):899–929.

30. Sabra MC, Uitdehaag JC, Watts A (1998) General model for lipid-mediated two-dimensional array formation of membrane proteins: application to bacteriorhodopsin. *Biophys J* 75(3):1180–1188.

31. Iacovache I, et al. (2010) The 2DX robot: A membrane protein 2D crystallization Swiss Army knife. *Journal of Structural Biology* 169(3):370–378.

32. Padilla JE, Colovos C, Yeates TO (2001) Nanohedra: Using symmetry to design self assembling protein cages, layers, crystals, and filaments. *PNAS* 98(5):2217–2221.

33. Lai Y-T, et al. (2014) Structure of a designed protein cage that self-assembles into a highly porous cube. *Nat Chem* 6(12):1065–1071.

34. Dryden KA, Crowley CS, Tanaka S, Yeates TO, Yeager M (2009) Two-dimensional crystals of carboxysome shell proteins recapitulate the hexagonal packing of three-dimensional crystals. *Protein Sci* 18(12):2629–2635.

35. Tai C-H, Paul R, Dukka KC, Shilling JD, Lee B (2014) SymD webserver: a platform for detecting internally symmetric protein structures. *Nucleic Acids Res* 42(Web Server issue):W296-300.

36. Pettersen EF, et al. (2004) UCSF Chimera—A visualization system for exploratory research and analysis. *J Comput Chem* 25(13):1605–1612.

37. Barklis E, et al. (1997) Structural analysis of membrane-bound retrovirus capsid proteins. *The EMBO Journal* 16(6):1199–1213.

Chapter 7. Engineering of symmetric scaffolds for small proteins in cryo-EM[5]

Yuxi Liu[1,*] Shane Gonen[2,3,*], Tamir Gonen[4], Todd O. Yeates[1,5,6$]

[1]UCLA Department of Chemistry and Biochemistry

[2]Howard Hughes Medical Institute, Janelia Research Campus

[3]Howard Hughes Medical Institute, UCSF

[4]Howard Hughes Medical Institute, Departments of Physiology and Biological Chemistry, David Geffen School of Medicine, UCLA

[5]UCLA-DOE Institute for Genomics and Proteomics

[6]UCLA Molecular Biology Institute

* Equal author contributions

ABSTRACT

Current single particle electron cryo-microscopy (cryo-EM) techniques can produce images of large protein assemblies and macromolecular complexes at atomic level detail without the need for crystal growth. However, proteins of smaller size, typical of those found throughout the cell, are not presently amenable to detailed structural elucidation by cryo-EM. Here we use protein design to create a modular, symmetrical scaffolding system to make protein molecules of typical size amenable to cryo-EM. Using a rigid continuous alpha-helical linker, we connect a

[5] This chapter is the adapted version of a published article. The project was a collaboration with the Gonen lab at the HHMI Janelia Research Campus. I performed the computational design as well as the biochemical and preliminary structural characterization on the designed constructs. Dr. Shane Gonen performed the cryo-EM data collection and processing.

small 17 kDa protein (DARPin) to a protein subunit that was designed to self-assemble into a cage with cubic symmetry. We show that the resulting construct is amenable to structural analysis by single particle cryo-EM, allowing us to identify and solve the structure of the attached small protein at near-atomic detail, ranging from 3.5 to 5 Å resolution. The result demonstrates that proteins considerably smaller than the theoretical limit of 50 kDa for cryo-EM can be visualized clearly when arrayed in a rigid fashion on a symmetric designed protein scaffold. Furthermore, because the amino acid sequence of a DARPin can be chosen to confer tight binding to various other protein or nucleic acid molecules, the system provides a future route for imaging diverse macromolecules, potentially broadening the application of cryoEM to proteins of typical size in the cell.

Significance statement: New electron microscopy methods are making it possible to view the structures of large proteins and nucleic acid complexes at atomic detail, but the methods are difficult to apply to molecules smaller than about 50 kDa, which is larger than the size of the average protein in the cell. The present work demonstrates that a protein much smaller than that limit can be successfully visualized when it is attached to a large protein scaffold designed to hold 12 copies of the attached protein in symmetric and rigidly defined orientations. The small protein chosen for attachment and visualization can be modified to bind to other diverse proteins, opening a new avenue for imaging cellular proteins by cryo-EM.

Keywords: Cryo-electron microscopy (CryoEM), protein design, DARPin, protein cage, protein scaffold

INTRODUCTION

Recent advancements have brought single particle electron microscopy techniques to the forefront of structural biology (1–3). In favorable cases, three-dimensional cryo-EM image reconstruction methods are able to produce structures of macromolecular complexes at atomic

level detail (4–9). In such studies, very large macromolecular assemblies offer important advantages in signal processing and imaging, and this advantage is enhanced in systems that are highly symmetric – e.g. composed of many repeating copies of one or a few protein building blocks. For those reasons, viral capsids are quintessential examples for favorable cryo-EM reconstruction. At the other end of the spectrum, however, individual protein molecules of typical size (e.g. 50 kDa or smaller), which lack the aforementioned advantages, remain extremely difficult to visualize at atomic detail by electron microscopy. This critical size limitation represents a singular impediment to the universal application of electron microscopy for elucidating structures of most proteins in the human genome.

Recent studies have shown that small proteins can be computationally re-designed so that multiple copies of the protein subunit will self-assemble into large, symmetric cages with shapes resembling regular geometric solids: e.g., a tetrahedron, cube, or icosahedron (10–16). The structures resulting from some of these designed assembly approaches have sufficiently large mass and high symmetry that they can be analyzed readily by cryo-EM. However, current methods for designing protein assemblies are laborious and unpredictable, often requiring substantial trial-and-error experiments and prior structural knowledge to achieve success. Those challenges have made it impractical to take a given target protein of interest, whose structure may not be known, and engineer it to assemble into a large symmetric assembly that would be amenable to cryo-EM.

It would advance cryo-EM applications tremendously if it were possible to easily attach a protein of interest to a symmetric scaffold in a rigid way, so that many copies of the target protein would be displayed in well-defined, symmetric orientations. Being able to turn a given protein into its own kind of capsid structure would confer on it the features of large size and symmetry that are

critically advantageous for cryo-EM imaging. In the present study, we demonstrate a route towards that goal (Fig. 7.1A).

RESULTS

For designing a modular cryo-EM scaffold, we took as a starting point a set of protein cages designed by King et al.(14), specifically those built from 24 subunits, four trimers of two different subunit types. These assemble with the different trimer types sitting at alternating corners of a cube, in arrangements that obey tetrahedral symmetry. In the current study, we focused our attention on designed protein cages where one or both component subunit types contain at least one alpha helical terminus. Through further design, we extended the alpha helical terminus of the cage protein by genetic fusion to join the alpha helical terminus of a small protein target of only ~17 kDa known as DARPin (Designed Ankyrin Repeat Protein) (17). This design element, fusing two proteins with terminal helices, is intended to create a semi-rigid and geometrically predictable helical connection spanning between two proteins that would otherwise be flexibly joined (Fig. 7.1A). This idea was developed by Padilla et al. and subsequently expanded upon and validated in various contexts, including in various DARPin designs (10, 11, 18–22).

The choice of a DARPin as the first fusion partner to the cage is critical, as DARPins have been developed as a general platform for binding other protein molecules. Through genetic selection techniques, amino acid sequence changes in loop regions of the DARPin protein can be identified for conferring tight binding to various target proteins of interest (23–26). In addition, their largely alpha-helical nature makes DARPins suitable for fusion to other proteins by the continuous alpha helical fusion approach. Taken together, the essence of our scaffolding system is that a rigid protein cage forming a core structure will present (as genetic fusions) 12 rigid and symmetrically disposed DARPin proteins projecting outward (Fig. 7.1). In the future, loop

sequences specific for binding some other target protein can be readily exchanged into the basic DARPin structure, thereby enabling the facile capsid-like assembly of varied target proteins. Importantly, this strategy ultimately circumvents the need to perform engineering experiments on future targets themselves by restricting design efforts to the protein cage and its fused DARPin. We experimentally tested several variations in the amino acid sequence and length of the helical connection between the cage subunit and the DARPin based on computationally generated fusion models (see Methods and SI Text). We devoted our efforts to specific design choices that disposed the DARPin binding surfaces in highly accessible orientations for subsequent utility in binding cognate target proteins. Among the designs investigated, five could be purified in soluble form from a bacterial overexpression system and were shown to self-assemble into structures of the expected size and shape by negative-stain electron microscopy (Fig. 7.S1).

Next, we pursued a full structural elucidation for one of the scaffold designs, referred to here as DARP14, by 3D cryo-EM reconstruction (Fig. 7.2 & 7.3, Table 7.S1). DARP14 was imaged on a Titan Krios using a K2 direct electron detector (see Methods). A total of 3665 movies were recorded for motion correction and after reference-free 2D classification 229,953 particles were selected for 3D analysis. In the raw cryo-EM images, the core of the protein cage was discernible but the individual DARPin components appeared weaker or were practically invisible (Fig. 7.2A). This was expected, and further reinforced the well-known challenge of imaging small protein molecules on their own. Subsequent 2-D class averages and 3-D reconstruction showed the powerful advantage of being able to locate and apply symmetry-averaging to the smaller DARPin components when displayed on the engineered scaffold (Fig. 7.2 & 7.3). 3-D analysis of the cage based on a subset of 34,650 particles produced a reconstruction with the majority of the core at atomic resolutions of ~2.5 Å with an overall resolution of ~3 Å (Fig. 7.2C, Table 7.S1).

The side chains of the amino acids in the core of the cage are clearly discernable in the resulting density maps and are consistent with the designed protein. This is the first example of an atomic resolution structure of a designed protein determined by single particle cryo-EM.

Importantly, much of the attached DARPin was also visible in the 2D class images (Fig 7.3A) and reconstructions. To account for the possibility of slight variations in the orientations of the attached DARPins, which would compromise their resolution, we applied subsequent classification to 183,753 particles after masking out the B type subunit of the cage in order to focus on the DARPin component (see Methods). This substantially improved the structural details visible for the DARPin, resulting in a ~3.5 Å resolution structure overall (Fig 7.3B-C) and allowing us to clearly model the helical secondary structural elements within the density (Fig. 7.3D) in configurations consistent with the known crystal structure of the DARPin (PDB 3ZU7) (Table 7.S1).

The DARPin protein we attached to the cage is comprised of five repeats of a common structural motif (the ankyrin repeat). In our final 3-D reconstruction of the DARPin, the first four repeats could be resolved at near-atomic detail, with the local resolution worsening from 3.5 Å to 5 Å toward the tip of the structure (Fig. 7.3C). This worsening of the resolution toward the tips of cryoEM structures has been observed in other cryoEM studies (27). Moreover, we suspect that the fifth DARPin repeat in our designed scaffold may be flexible and partially unwound, further contributing to its weakness in the final image. Consistent with this explanation, the thermal vibration parameters (B factors) in previous crystal structures of DARPins are higher for this region of the protein (28, 29) (Fig. 7.S2). We note that this tendency toward terminal unwinding is not an impediment to forming a well-ordered complex between a DARPin and its cognate target, as demonstrated in multiple previous crystal structures where the DARPin and its target are well-

ordered when bound together (24, 30–32). Notwithstanding the loss of resolution at the end of the attached DARPin, the result represents the first example of a small protein being visualized at near-atomic resolution by a cryo-EM scaffolding approach.

DISCUSSION

Our analysis demonstrates that the alpha helical fusion scheme used here provides a connection between the symmetric cage and the DARPin that is rigid enough to enable near atomic-resolution imaging. This is a critical result as it was not known in advance whether the alpha helical fusion would hold the DARPin in a sufficiently ordered configuration. The ordered nature of the DARPin was evident in preliminary 2D averaging (Fig. 7.3A) even before 3D reconstruction and application of symmetry to optimize the imaging of the cage. In comparing our final structure to the initial computational model, a minor reorientation of the DAPRin component (by approximately 13 degrees) is evident (Fig. 7.3E and Fig. 7.S3). Among the several designs we explored (Fig. 7.S1), the structure of the one analyzed here appears to be influenced, beyond our designed continuous alpha-helical fusion, by a few additional atomic contacts between the DARPin and the cage subunits. These contacts likely help stabilize the DARPin in a well-defined orientation on the scaffolding cage. The relatively high orientational rigidity we obtained for the DARPin promises good prospects for similarly rigid attachment of other proteins to the DARPin for their visualization in subsequent studies. Past studies of DARPin complexes indicate stable and rigid binding to their cognate protein targets (*21*, *26–28).*

Our results emphasize two major points. First, the DARPin component is a small protein (17 kDa) whose separate structure would otherwise be impossible to resolve by single particle cryo-EM methods. Yet it can be visualized in near atomic detail when its image is reconstructed in the context of rigid assembly on a large symmetric protein cage. Recent work by Coscia et al.

(33) was able to image a larger (40 kDa) target protein fused to a natural protein scaffold at lower resolution (local resolution between 6 and 10 Å) and only after extensive biochemical analysis and optimization of linker lengths. We show here that a rational design of a continuous alpha helical attachment to a cubically symmetric designed protein cage can provide the rigidity required to achieve near atomic resolutions even for a small 17 kDa attached target protein. Since our present scaffold was the best among only a relatively small number of candidates investigated, it is likely that further design efforts could improve the degree of rigidity, making it possible to reach an even better spatial resolution. Second, our development of a DARPin as the fused protein component introduces a critical element of modularity. Building on this system, the challenging molecular engineering required to create symmetric architectures will not need to be repeated for each application to a new target protein to be imaged. In principle, no modification to a target protein is required, because the loop sequences of the DARPin carried on the scaffold can be mutated to bind various target proteins in their native forms.

The ease of attachment and rigidity of cognate target proteins bound to the scaffold are key issues for future studies. It is possible that attached proteins could exhibit higher flexibility, or on the contrary they could help rigidify the DARPin. Also, very large target proteins could create steric challenges for full-occupancy attachment. For the scaffold explored here, the closest approach between the centers of the 12 DARPin binding surfaces is about 65 Å, which we expect would allow the scaffold to accommodate proteins as large as 200 kDa without collisions (Fig. 7.S4). This is notable given that proteins larger than this size can be imaged directly without scaffolding by cryoEM. Thus, we expect our scaffold to be compatible with most imaging targets below the current cryo-EM size limit. Finally, different target proteins may be more or less suitable for the symmetric scaffolding approach. Proteins that naturally self-associate are likely to be

problematic, for example. Indeed, we observed that the ERK2 protein that binds to the DARPin we used in this initial study self-associates, and as a result the scaffold could not be maintained in solution upon addition of the target protein in this case (data not shown).

Ultimately, it could prove important to develop a suite of distinct scaffolding system using variations on the design theme developed here. Each such scaffold could provide a distinct opportunity for obtaining a high-resolution structure of a target protein irrespective of how small that target protein is as demonstrated in this study. Further developments on this scaffolding approach should ultimately enable the facile imaging of large numbers of cellular proteins whose structures have previously been beyond the reach of cryoEM.

MATERIALS AND METHODS

Computational α-helix fusion methods. Computational α-helix fusion models were generated similarly to our previous work (10, 11). As a test case for fusing to a protein cage, we used a DARPin whose sequence was selected to bind to the extracellular signal-regulated kinase 2 (ERK2), and whose structure in complex with its cognate partner is known (PDB 3ZU7). In choosing a protein cage as the fusion partner, we restricted our attention to those that have a protein with a terminal α-helix at least six amino acids long and with no more than ten unstructured amino acids beyond it. The set of protein cages that satisfied this criteria included six protein assemblies designed in previous work (11, 14, 15, 34, 35). Next, we tested the feasibility of pair-wise joining between the protein cage subunit and the DARPin subunit. To do so, we first aligned an ideal α-helix to the last six helical residues on the cage subunit. Then we aligned the DARPin terminal helix to the ideal α-helix. The aligned position of the DARPin on the ideal α-helix was slid one residue at a time. The range of sliding was from a six-residue overlap to a 15-residue insertion relative to the helical termini of the DARPin and the cage

subunit. We inspected the models at each aligned position and removed those with excessive clashes. If the fusion model had overlapping helical termini, the amino acid sequence within the overlap was chosen to maintain good native contacts within each subunit. If the fusion model required an insertion between the helical termini, ER/K-rich helix segments (19, 36) were used. The final set of experimentally tested models were chosen to give different DARPin orientations relative to the cage subunit while providing large space for attachment of imaging targets. The construct with the shortest linker for each DARPin orientation was selected. In total, nine constructs were judged to be suitable for experimental characterization. These were based on the single DARPin noted earlier fused to one subunit of two different two-component cages, T33-21 (14) and T33-31 (35). Based on different helical lengths for connection to the DARPin, there were three candidate fusions to cage protein T33-31 and six fusions to cage protein T33-21.

Cloning, Expression, and Purification. Constructs named DARP10, DARP11, DARP12, DARP14, and DARP16 were expressed and purified under conditions similar to those used for the cage proteins alone (14) with slight modifications. We purchased *E. coli* codon optimized gene fragments (Integrated DNA Technologies) and inserted the sequences encoding both cage subunit A and subunit B into a pET-22b vector, separated by the intergenic region of pET-DUET. Proteins were expressed in auto-induction media at 20 °C for two days.

Cells were suspended and lysed in Lysis Buffer (50mM Tris pH 8.0 250mM NaCl, 20mM imidazole) supplemented with DNase, lysozyme, and protease inhibitor (Thermo Scientific Pierce). Cleared lysate was loaded onto a HisTrap column (GE Healthcare) and eluted with a linear gradient of Elution Buffer (50mM Tris pH 8.0 250mM NaCl, 500mM imidazole). Pooled and concentrated fractions were then further purified with size-exclusion chromatography

on a Superose 6 Increase column (GE Healthcare). Fractions corresponding to intact tetrahedral assemblies were used in further analysis.

Negative stain electron microscopy. Freshly purified proteins at about 50 μg/mL were applied onto glow-discharged 200 or 300 mesh copper formvar supported carbon grids (Ted Pella, Inc.), washed with Milli-Q water and stained with 2% uranyl acetate or 0.75% uranyl formate. Micrographs were collected using a Tecnai T12 with a bottom mount TVIPS F416 4K x 4K CMOS camera at a nominal magnification of 49,000x at the specimen level.

Electron cryo-microscopy (cryo-EM). *DARP14 grid screening* – Purified, concentrated DARP14 was screened for ice thickness, stability and particle distribution using a FEI TF20 microscope equipped with a bottom mount TVIPS F416, 4K x 4K CMOS camera.

DARP14 grid freezing for data collection – Superose 6 Increase column (GE Healthcare) purified, concentrated DARP14 was diluted to ~0.5 mg/mL using 10mM Tris pH 8.0, 500mM NaCl supplemented with 1mM of freshly prepared dithiothreitol (DTT) (Acros) and 4μL was pipetted on to C-Flat, carbon-coated, 1.2/1.3 200 Mesh copper grids (Electron Microscopy Sciences). Grids were blotted and frozen in liquid ethane using a Vitrobot Mark IV (FEI) and stored for data collection under liquid nitrogen.

Data Collection – Super-resolution movies were collected using a FEI titan krios (Thermo Fisher) microscope equipped with a Gatan K2 Summit direct electron detector at 22500X magnification at the specimen level with a physical pixel size of 1.31 Å/pixel (0.655 Å/pixel super-resolution).

Data Processing *Cryo reconstructions* – Super-resolution movies of frozen DARP14 were corrected for beam-induced motion using MotionCor2 (37). Particles were picked using the XMIPP software package (38). All co-ordinates were imported into and all unbinned

micrographs analyzed using the RELION 2 software pipeline (39). An initial model used for all stages of 3D reconstructions was calculated *de-novo* using the Stochastic Gradient Descent algorithm in RELION 2.1-beta-0 using a subset of the calculated 2D classes. 3D classification and refinement were performed with enforced T symmetry. The final refined map containing DARPins was made while masking out all B-subunits. Local resolution was estimated using ResMap (40) and RELION. All masks were created from the reconstructions using combinations of both RELION and UCSF Chimera (41). All RELION calculations were done using different versions of RELION 2 except for the final refinements, post-processing and local resolution estimations (including the final fourier shell correlation calculations), which were done using RELION version 2.1-beta-1.

Structure Analysis – All reconstructions were analyzed using UCSF Chimera and COOT (42). The design model was initially fit using UCSF Chimera, followed by structure relaxation in Rosetta (43, 44) without enforced symmetry. Refined models were analyzed using UCSF Chimera, PyMOL (45) and COOT.

Acknowledgements:

The authors thank Marianne Vo for her contribution to protein production, James Evans for preliminary assistance with the EM studies and Johan Hattne (Gonen Lab HHMI Janelia) for help with th.e computational cluster. This work was funded by the BER program of the Department of Energy Office of Science, award DE-FC02-2ER63421, and by the UCLA Whitcome Fellowship to YL. The Gonen laboratory is supported by the Howard Hughes Medical Institute.

Author Contributions: TOY and TG helped design and supervise the research. YL performed the protein design and characterization experiments. SG performed and SG and TG analyzed the cryo-electron microscopy studies. All authors contributed to writing the manuscript and to the preparation of the figures.

Competing Interests: The authors declare no competing financial interests

Data Deposition: EM density maps and model coordinates are deposited in the Electron Microscopy Data Bank (EMDB) and Protein Data Bank (PDB) with accession codes: EMD-7403, 7436, 7437 and PDB - 6C9I and 6C9K.

SUPPORTING INFORMATION

Protein Sequences

Protein sequences of soluble genetic fusions between protein cages and the chosen DARPin. The regions corresponding to the DARPins are underlined. Additional residues designed between the cage subunit and its joining DARPin are double underlined.

- DARP10:

Subunit A:

MEEVVLITVPSALVAVKIAHALVEERLAACVNIVPGLTSIYREEGSVVSDHELLLLVKTTTDAFPKLKERVKELHPYEVPEIVALP
IAEGNREYLDWLRENMERARQELGKKLLEAARAGQDDEVRILMANGADVNAHDDQGSTPLHLAAWIGHPEIVEVLLKHGADVNARD
TDGWTPLHLAADNGHLEIVEVLLKYGADVNAQDAYGLTPLHLAADRGHLEIVEVLLKHGADVNAQDKFGKTAFDISIDNGNEDLAE
ILQKLN

Subunit B:

MVRGIRGAITVEEDTPAAILAATIELLLKMLEANGIQSYEELAAVIFTVTEDLTSAFPAEAARLIGMHRVPLLSAREVPVPGSLPR
VIRVLALWNTDTPQDRVRHVYLNEAVRLRPDLESAQLEHHHHHH

- DARP11

Subunit A:

MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIYKIMGEIGSKGKIEGISEERIAW
LLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKVLTVTREFGIGAEAAAYLLALSDLLFLLARVIEIEKNDLGKKLL
EAARAGQDDEVRILMANGADVNAHDDQGSTPLHLAAWIGHPEIVEVLLKHGADVNARDTDGWTPLHLAADNGHLEIVEVLLKYGAD
VNAQDAYGLTPLHLAADRGHLEIVEVLLKHGADVNAQDKFGKTAFDISIDNGNEDLAEILQKLN

Subunit B:

MPHLVIEATANLRLETSPGELLEQANKALFASGQFGEADIKSRFVTLEAYRQGTAAVERAYLHACLSILDGRDIATRTLLGASLCA
VLAEAVAGGGEEGVQVSVEVREMERLSYAKRVVARQRLEHHHHHH

- DARP12

Subunit A:

MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIYKIMGEIGSKGKIEGISEERIAW
LLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKVLTVTREFGIGAEAAAYLLALSDLLFLLARVIEIEKNQDLGKKL
LEAARAGQDDEVRILMANGADVNAHDDQGSTPLHLAAWIGHPEIVEVLLKHGADVNARDTDGWTPLHLAADNGHLEIVEVLLKYGA
DVNAQDAYGLTPLHLAADRGHLEIVEVLLKHGADVNAQDKFGKTAFDISIDNGNEDLAEILQKLN

Subunit B:

MPHLVIEATANLRLETSPGELLEQANKALFASGQFGEADIKSRFVTLEAYRQGTAAVERAYLHACLSILDGRDIATRTLLGASLCA
VLAEAVAGGGEEGVQVSVEVREMERLSYAKRVVARQRLEHHHHHH

- DARP14:

Subunit A:

MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIYKIMGEIGSKGKIEGISEERIAW
LLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKVLTVTREFGIGAEAAAYLLALSDLLFLLARVIEIELGKKLLEAA
RAGQDDEVRILMANGADVNAHDDQGSTPLHLAAWIGHPEIVEVLLKHGADVNARDTDGWTPLHLAADNGHLEIVEVLLKYGADVNA
QDAYGLTPLHLAADRGHLEIVEVLLKHGADVNAQDKFGKTAFDISIDNGNEDLAEILQKLN

Subunit B:

MPHLVIEATANLRLETSPGELLEQANKALFASGQFGEADIKSRFVTLEAYRQGTAAVERAYLHACLSILDGRDIATRTLLGASLCA
VLAEAVAGGGEEGVQVSVEVREMERLSYAKRVVARQRLEHHHHHH

- DARP16

Subunit A:

MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIYKIMGEIGSKGKIEGISEERIAW
LLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKVLTVTREFGIGAEAAAYLLALSDLLFLLARVIEIEDLGKKLLEA
ARAGQDDEVRILMANGADVNAHDDQGSTPLHLAAWIGHPEIVEVLLKHGADVNARDTDGWTPLHLAADNGHLEIVEVLLKYGADVN
AQDAYGLTPLHLAADRGHLEIVEVLLKHGADVNAQDKFGKTAFDISIDNGNEDLAEILQKLN

Subunit B:
MPHLVIEATANLRLETSPGELLEQANKALFASGQFGEADIKSRFVTLEAYRQGTAAVERAYLHACLSILDGRDIATRTLLGASLCA
VLAEAVAGGGEEGVQVSVEVREMERLSYAKRVVARQRLEHHHHHH

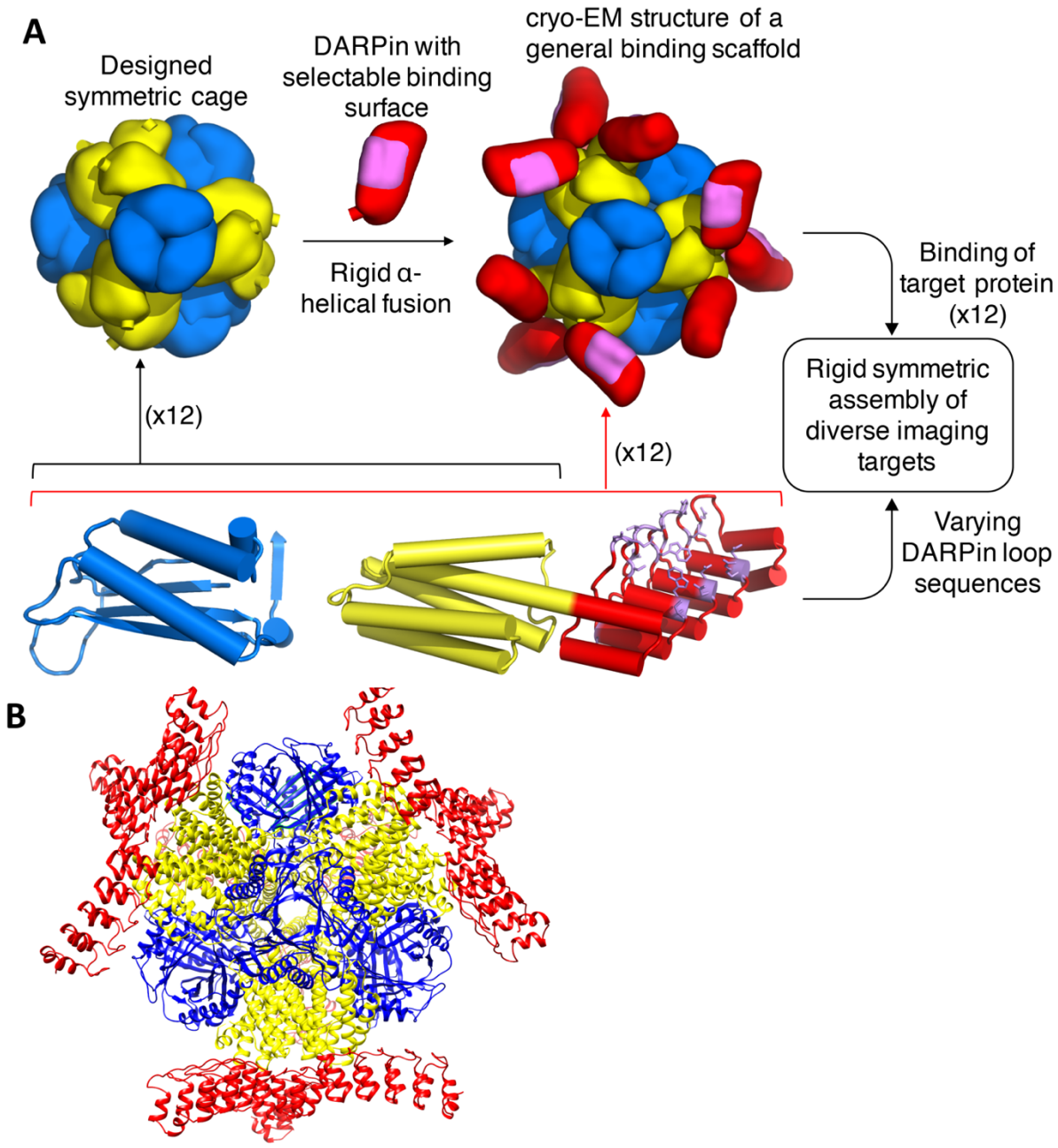


Figure 7.1. A molecular scaffolding system for modular display of macromolecules for cryo-EM imaging.

A. Schematic diagram for a scaffolding system built upon a designed symmetric protein cage; the example shown is a tetrahedrally symmetric cage with 24 subunits in $a_{12}b_{12}$ stoichiometry (A subunits in yellow and B subunits in blue). At least one of the subunits needs to have an α-helical

terminus (cylinder). The α-helical termini of the cage subunit (yellow) and the DARPin subunit (red) can be joined in a rigid fashion through genetic fusion, forming a general binding scaffold. The cryo-EM structure of a binding scaffold is solved in this study. The DARPin subunit contains variable loops (highlighted in pink) whose amino acid sequence can be selected to confer binding to a wide range of specific macromolecules of interest. In principle, binding a macromolecule of interest to the designed scaffold results in the symmetric display of 12 copies of the molecule. B. Detailed view of the specific scaffold, DARP14, that was designed and characterized in this study, with subunits colored as in panel A.

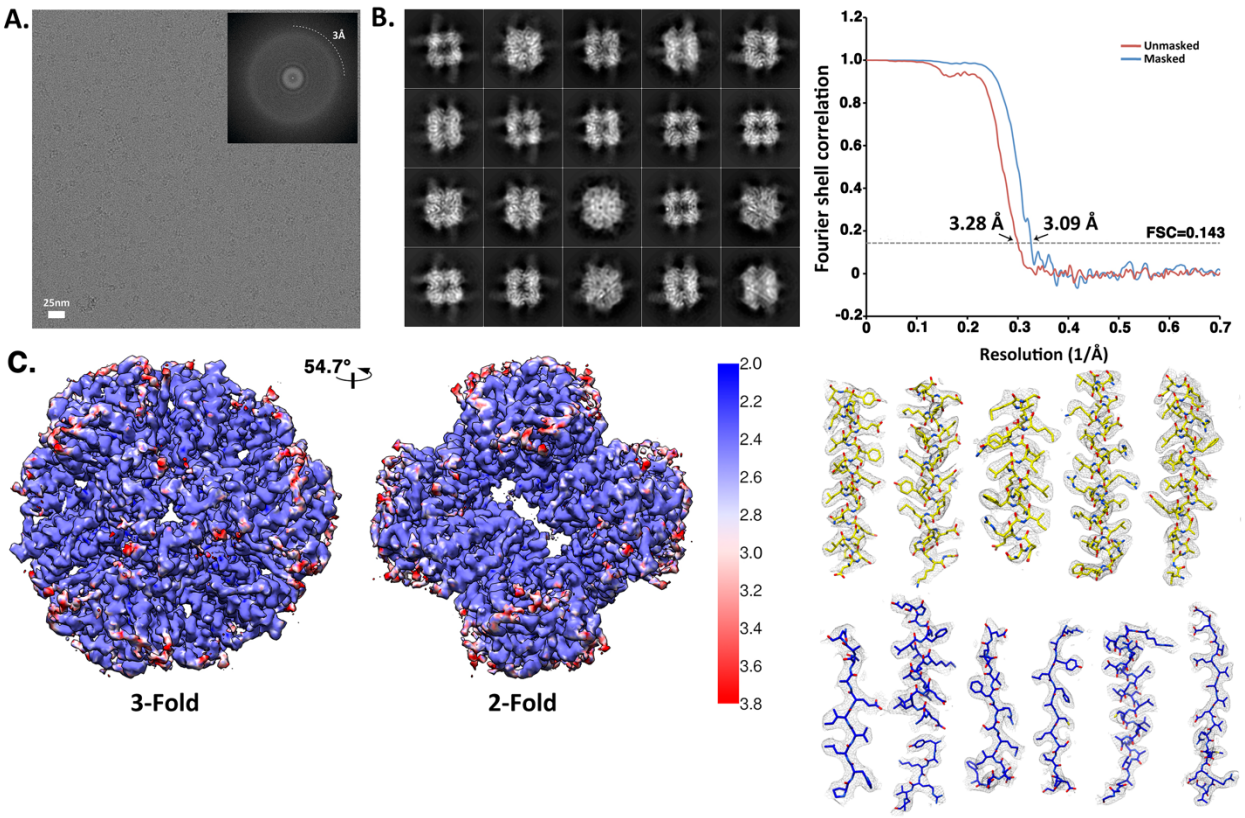


Figure 7.2. Cryo-EM structure of DARP14 symmetric cage core.

A. Representative motion-corrected cryo-electron micrograph of DARP14. (inset) Fourier transformation showing visible thon rings to ~3Å. B. Reference-free 2D class averages highlighting good alignment of the cage and clear density for fused 17kDa DARPins. C. Overview of a ~3.1 Å reconstruction of the cage core. Top: Representations of unfiltered local resolution viewing down the 3-fold and 2-fold symmetry axes highlighting extensive areas at atomic resolutions of ~2.5 Å. Middle: Fourier shell correlation (FSC) curves of unmasked and masked reconstructions. Bottom: refined model fit into density for subunit A (yellow) and subunit B (blue) of the cage core. All secondary structure elements are represented along with selected loop regions.

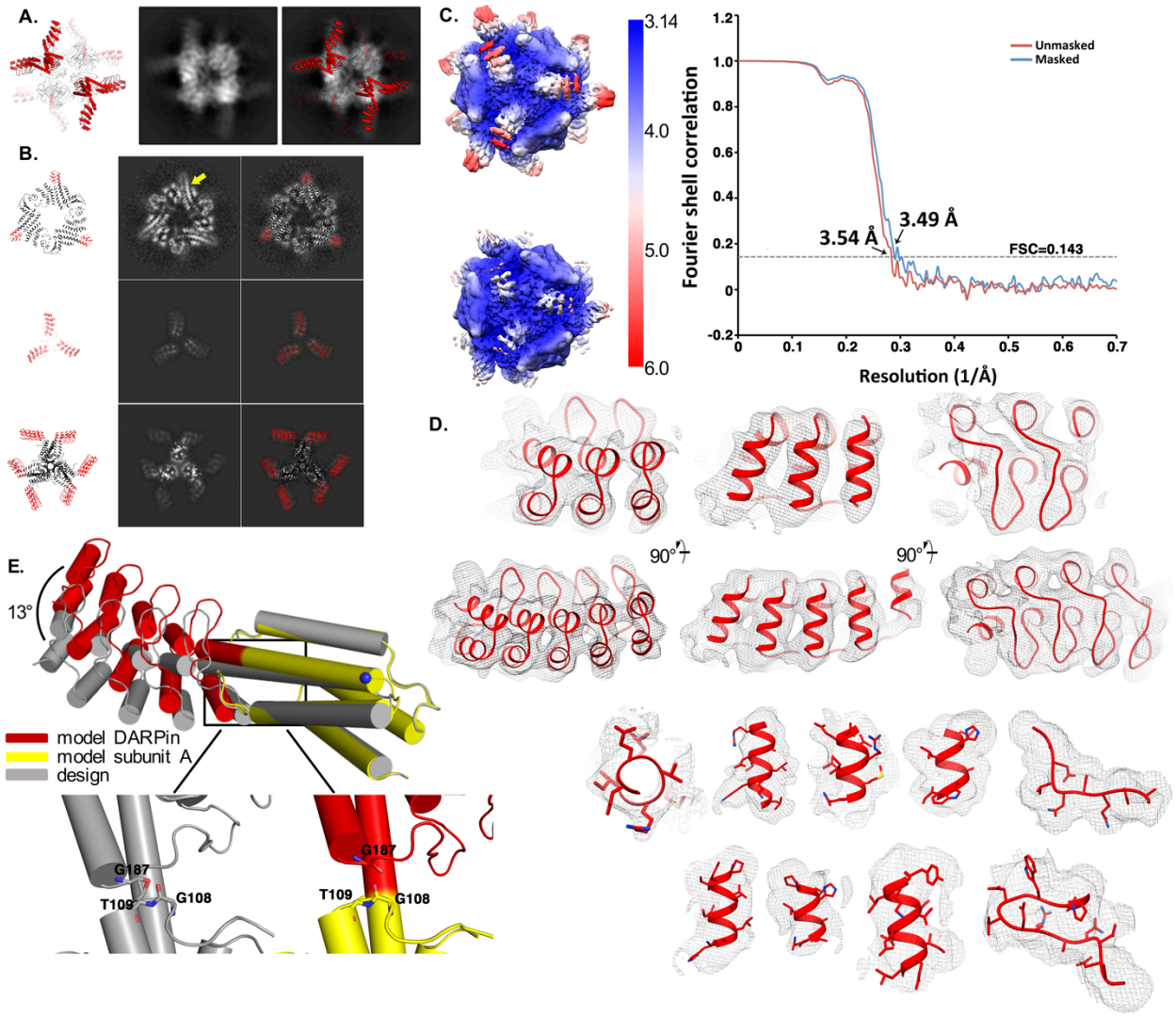


Figure 7.3. Cryo-EM reconstruction of DARPin displayed on the symmetric cage.

A. Comparison of the DARP14 design (DARPins in red and cage subunits A and B in black and white) to one 2D class average with overlay (right) highlighting density for DARPin helices protruding from the cage. B. Three comparisons of the calculated model and slices of reconstructions. Top: Focus on the extended helix where DARPins are fused (yellow arrow). Middle: Top view showing the DARPin arms and clear density for each helical repeat. Bottom: Side-slice. C. Local resolution of unfiltered ~3.5Å reconstruction where the subunit A and fused DARPin were masked during refinement for higher resolution of those areas. Left top: Low contouring level to show entire reconstruction. Left bottom: Higher contouring level highlighting

the near-atomic detail of DARPin repeats. Right: FSC of unmasked and masked DARPin reconstructions. D. Highlight of DARPin density in different regions with the fitted model. Top two rows: High sigma level highlighting DARPin helical repeats 1-3 (top) and lower sigma level highlighting all 5 helical DARPin repeats as a top view (left), side view (middle) and bottom view (right). Views related by 90° rotations. Bottom two rows: Density fit of DARPin model from various helices (including one top view of helix 2) and two views of loop regions where the amino acid sequence for the DARPin would be varied for binding to cognate target molecules. E. Comparison between the computational design for DARP14 and the cryo-EM density-fitted model, showing a small displacement of the fitted model from the design. The designed DARP14 and the cryo-EM model were aligned on their A subunits (which is named chain B in PDB 4NWP). Top: there is a ~13° rotation around an axis going through the view of plane at the blue dot between the design and the EM model. Bottom: zoom in at Gly 187 in the first turn on the DARPin, which is in steric clash to Gly108 and Thr 109 from subunit A in the design (bottom left). This clash is relieved in observed model and probably contribute to the DARPin stability in currently observed orientation (bottom right).

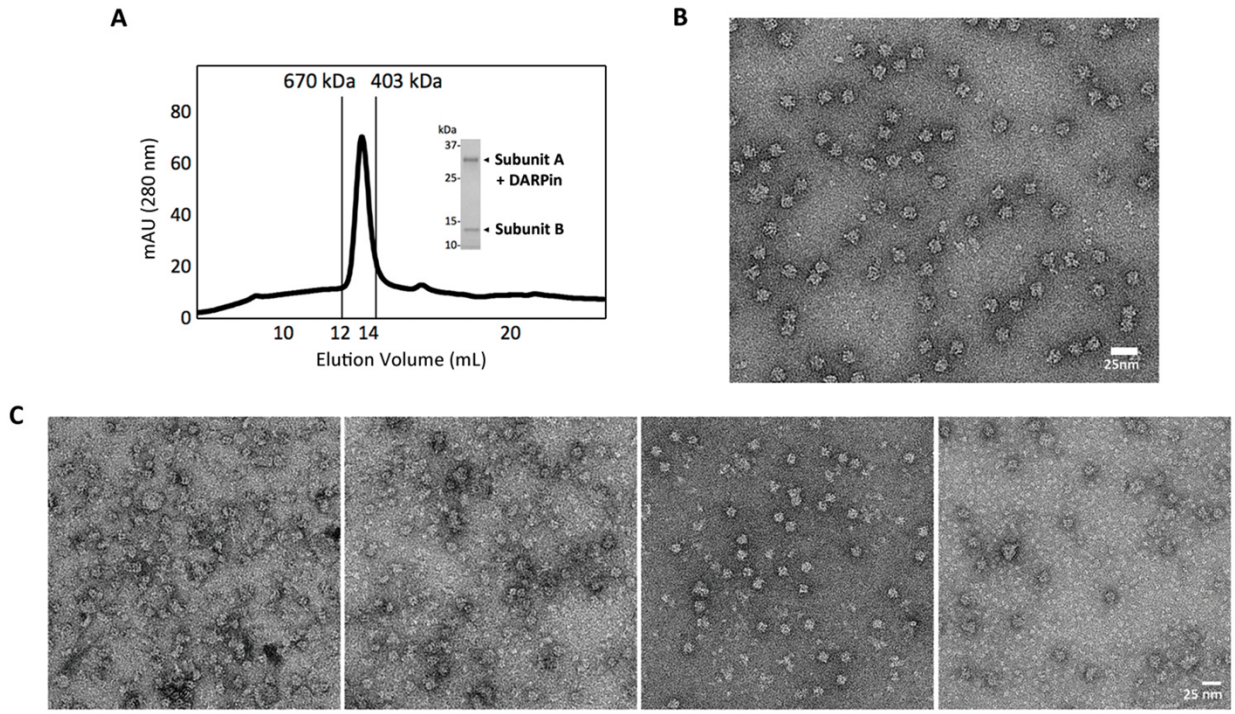
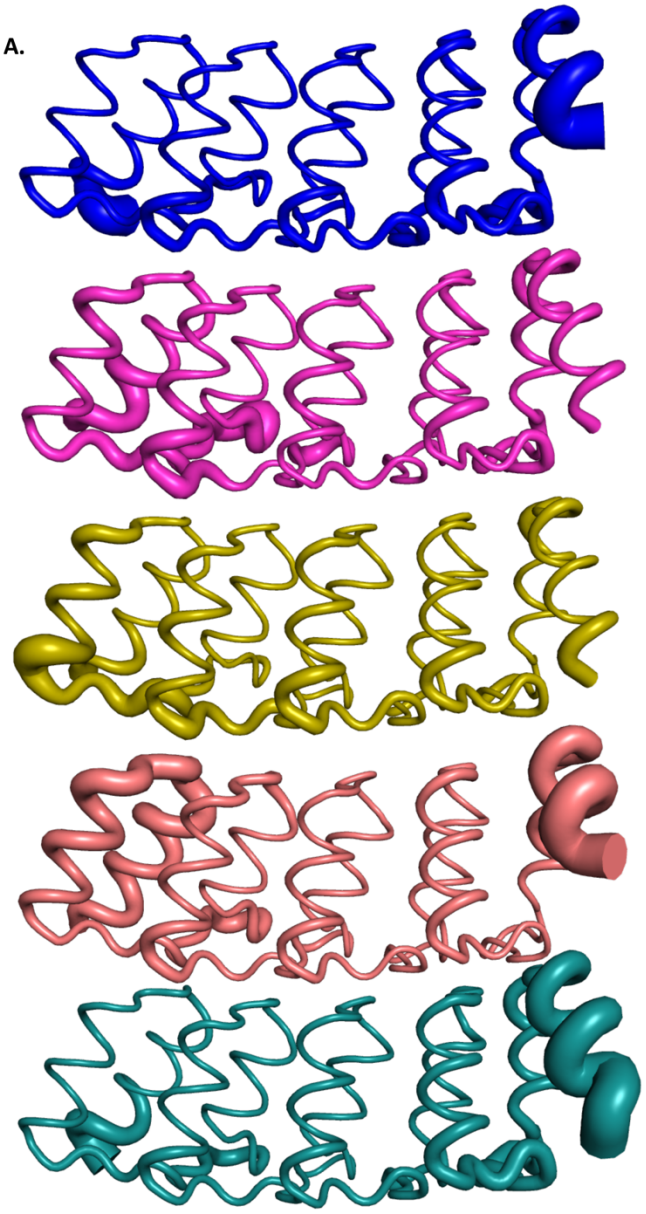


Fig. 7.S1 Designed DARPin-displaying cages form particles of expected size and shape.

A. Purification of the DARPin14 cage scaffold indicates a homogeneous preparation by size exclusion chromatography, SDS PAGE analysis and negative stain electron microscopy. The two components in DARP14 co-elutes and migrate as single peak with the correct retention volume from size-exclusion chromatography (Superose 6 Increase, GE Healthcare). B. Negative stain EM of DARP14. C. Negative stain EM of DARP10, DARP11, DARP12, and DARP16, from left to right. Particles of ~15nm diameter are clearly visible.

A.



B.

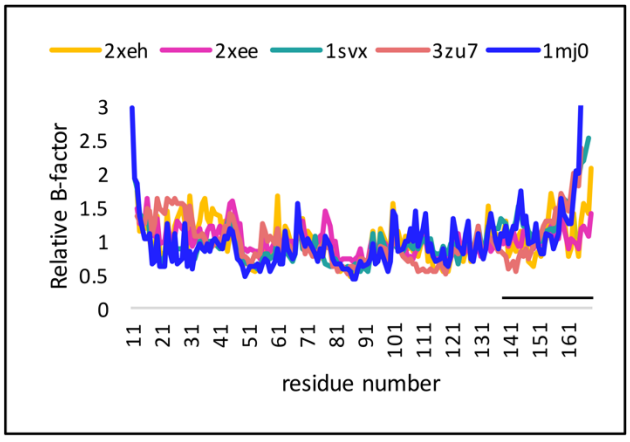

2xeh
2xee
1svx
3zu7
1mj0
Relative B-factor
3
2.5
2
1.5
1
0.5
0
11
21
31
41
51
61
71
81
91
101
111
121
131
141
151
161
residue number

Fig. 7.S2. Comparison of thermal atomic displacement parameters (B-factors) from previous DARPin crystal structures.

A. Structures of different DARPins shown in sausage representation according to their B-factors. Structures are positioned in similar orientation with N-termini on the left and C-termini on the right. Top to bottom: consensus DARPin (1MJ0), consensus DARPins with stabilizing mutations on the C-terminal repeat (2XEE chain A & 2XEH chain A), DARPin specific for ERK2 binding as used here in DARP14 (3ZU7 chain B), DARPin specific for maltose binding protein (1SVX chain A). For clarity, ERK2 and maltose binding protein are omitted from 3ZU7 & 1SVX, respectively. The C-terminal repeat tends to have higher B-factor than other repeats, unless stabilized by additional mutations. B. Plot of relative B-factors across the primary sequences. B-factor for each residue is normalized against the average B factor of protein residues in the same chain. Black line indicates the C-terminal repeat region.

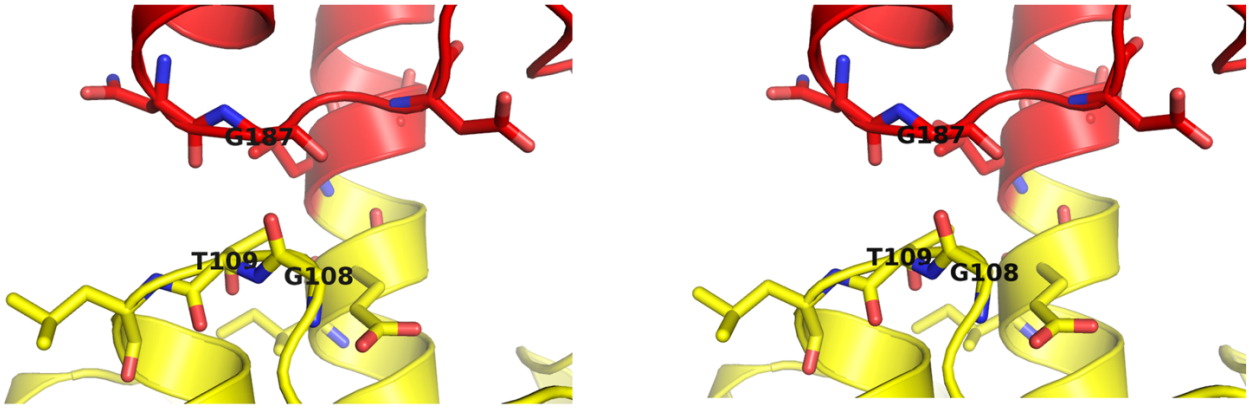


Fig. 7.S3. Details of the additional atomic contacts between the DARPin and the cage subunits in stereo view.

Coloring is the same as in Fig. 3E.

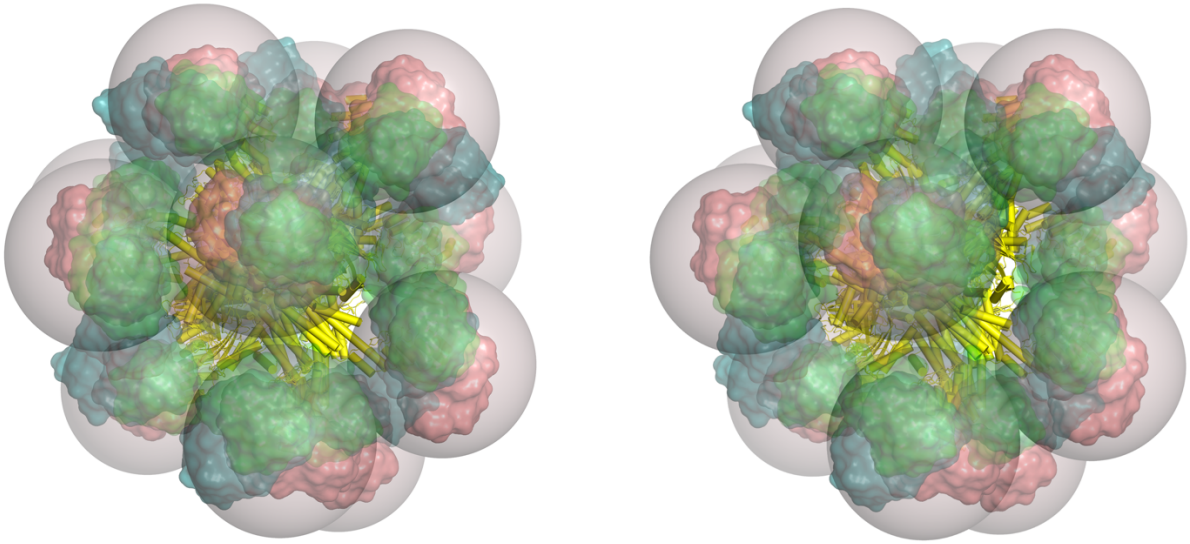


Fig. 7.S4. Predicted DARP14 target binding mode.

Crystal structures between DARPins and different targets are aligned onto the DARP14 EM model (yellow helical tubes) in the DARPin portion. Different target proteins bind to their cognate DARPins with different orientations: GFP (PDB 5LEM) in green, maltose binding protein (PDB 1SVX) in teal, ERK2 (PDB 3ZU7) in salmon. A sphere (transparent brown) with 90 Å diameter can docked onto the DARP14 EM model without clashes.

Table 7.S1. CryoEM data table

Data collection/processing	
Microscope	Titan Krios
Voltage	300 kV
Camera	Gatan K2 direct electron detector
Camera mode	Super-resolution
Defocus range	-0.7 to -2.5 μm
Exposure time per video	5 sec
Dosage per frame	1.2 $e^-/Å^2$
Reconstructions	
Software	RELION 2.1-b1
Pixel Size	0.655 Å/Pixel
Symmetry	T
Cage core	
Particles	34,650
Overall resolution (unmasked)	3.29 Å
Overall resolution (masked)	3.09 Å
Cage with DARPin	
Particles	183,753
Overall resolution (unmasked)	3.54Å
Overall resolution (masked)	3.49Å
Model	
Protein residues	5400
Ramachandran	
Outliers	
Allowed	
Favored	

REFERENCES

1. Nogales E (2016) The development of cryo-EM into a mainstream structural biology technique. *Nat Methods* 13(1):24–27.

2. Glaeser RM (2016) How good can cryo-EM become? *Nat Methods* 13(1):28–32.

3. Merk A, et al. (2016) Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. *Cell* 165(7):1698–1707.

4. Campbell MG, Veesler D, Cheng A, Potter CS, Carragher B (2015) 2.8 Å resolution reconstruction of the Thermoplasma acidophilum 20S proteasome using cryo-electron microscopy. *eLife* 4:e06380.

5. Gao Y, Cao E, Julius D, Cheng Y (2016) TRPV1 structures in nanodiscs reveal mechanisms of ligand and lipid action. *Nature* 534(7607):347–351.

6. Banerjee S, et al. (2016) 2.3 Å resolution cryo-EM structure of human p97 and mechanism of allosteric inhibition. *Science* 351(6275):871–875.

7. Zhang X, et al. (2017) An Atomic Structure of the Human Spliceosome. *Cell* 169(5):918–929.e14.

8. Greber BJ, et al. (2017) The cryo-electron microscopy structure of human transcription factor IIH. *Nature* 549(7672):414–417.

9. Hirschi M, et al. (2017) Cryo-electron microscopy structure of the lysosomal calcium-permeable channel TRPML3. *Nature* advance online publication. doi:10.1038/nature24055.

10. Padilla JE, Colovos C, Yeates TO (2001) Nanohedra: Using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proc Natl Acad Sci* 98(5):2217–2221.

11. Lai Y-T, Cascio D, Yeates TO (2012) Structure of a 16-nm Cage Designed by Using Protein Oligomers. *Science* 336(6085):1129–1129.

12. Fletcher JM, et al. (2013) Self-Assembling Cages from Coiled-Coil Peptide Modules. *Science* 340(6132):595–599.

13. Huard DJE, Kane KM, Tezcan FA (2013) Re-engineering protein interfaces yields copper-inducible ferritin cage assembly. *Nat Chem Biol* 9(3):169–176.

14. King NP, et al. (2014) Accurate design of co-assembling multi-component protein nanomaterials. *Nature* 510(7503):103–108.

15. Bale JB, et al. (2016) Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* 353(6297):389–394.

16. Sciore A, et al. (2016) Flexible, symmetry-directed approach to assembling protein cages. *Proc Natl Acad Sci* 113(31):8681–8686.

17. Binz HK, Stumpp MT, Forrer P, Amstutz P, Plückthun A (2003) Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins. *J Mol Biol* 332(2):489–503.

18. Strickland D, Moffat K, Sosnick TR (2008) Light-activated DNA binding in a designed allosteric protein. *Proc Natl Acad Sci* 105(31):10709–10714.

19. Sivaramakrishnan S, Spudich JA (2011) Systematic control of protein interaction using a modular ER/K α-helix linker. *Proc Natl Acad Sci* 108(51):20467–20472.

20. Batyuk A, Wu Y, Honegger A, Heberling MM, Plückthun A (2016) DARPin-Based Crystallization Chaperones Exploit Molecular Geometry as a Screening Dimension in Protein Crystallography. *J Mol Biol* 428(8):1574–1588.

21. Schütz M, et al. (2016) Generation of Fluorogen-Activating Designed Ankyrin Repeat Proteins (FADAs) as Versatile Sensor Tools. *J Mol Biol* 428(6):1272–1289.

22. Wu Y, et al. (2017) Rigidly connected multispecific artificial binders with adjustable

geometries. *Sci Rep* 7(1):11217.

23. Amstutz P, et al. (2005) Intracellular Kinase Inhibitors Selected from Combinatorial Libraries of Designed Ankyrin Repeat Proteins. *J Biol Chem* 280(26):24715–24722.

24. Veesler D, et al. (2009) Crystal Structure and Function of a DARPin Neutralizing Inhibitor of Lactococcal Phage TP901-1 COMPARISON OF DARPin AND CAMELID VHH BINDING MODE. *J Biol Chem* 284(44):30718–30726.

25. Plückthun A (2015) Designed Ankyrin Repeat Proteins (DARPins): Binding Proteins for Research, Diagnostics, and Therapy. *Annu Rev Pharmacol Toxicol* 55(1):489–511.

26. Sennhauser G, Grütter MG (2008) Chaperone-Assisted Crystallography with DARPins. *Structure* 16(10):1443–1453.

27. Jin P, et al. (2017) Electron cryo-microscopy structure of the mechanotransduction channel NOMPC. *Nature* 547(7661):118–122.

28. Wetzel SK, et al. (2010) Residue-Resolved Stability of Full-Consensus Ankyrin Repeat Proteins Probed by NMR. *J Mol Biol* 402(1):241–258.

29. Kramer MA, Wetzel SK, Plückthun A, Mittl PRE, Grütter MG (2010) Structural Determinants for Improved Stability of Designed Ankyrin Repeat Proteins with a Redesigned C-Capping Module. *J Mol Biol* 404(3):381–391.

30. Kummer L, et al. (2012) Structural and functional analysis of phosphorylation-specific binders of the kinase ERK from designed ankyrin repeat protein libraries. *Proc Natl Acad Sci* 109(34):E2248–E2257.

31. Pecqueur L, et al. (2012) A designed ankyrin repeat protein selected to bind to tubulin caps the microtubule plus end. *Proc Natl Acad Sci* 109(30):12011–12016.

32. Seeger MA, et al. (2013) Design, construction, and characterization of a second-

generation DARPin library with reduced hydrophobicity. *Protein Sci* 22(9):1239–1257.

33. Coscia F, et al. (2016) Fusion to a homo-oligomeric scaffold allows cryo-EM analysis of a small protein. *Sci Rep* 6:30909.

34. Lai Y-T, et al. (2014) Structure of a designed protein cage that self-assembles into a highly porous cube. *Nat Chem* 6(12):1065–1071.

35. Bale JB, et al. (2015) Structure of a designed tetrahedral protein assembly variant engineered to have improved soluble expression. *Protein Sci* 24(10):1695–1701.

36. Sivaramakrishnan S, Spink BJ, Sim AYL, Doniach S, Spudich JA (2008) Dynamic charge interactions create surprising rigidity in the ER/K α-helical protein motif. *Proc Natl Acad Sci* 105(36):13356–13361.

37. Zheng SQ, et al. (2017) MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods* 14(4):331–332.

38. XMIPP: a new generation of an open-source image processing package for electron microscopy (2004) *J Struct Biol* 148(2):194–204.

39. Scheres SHW (2016) Processing of Structurally Heterogeneous Cryo-EM Data in RELION. *Methods Enzymol* 579:125–157.

40. Kucukelbir A, Sigworth FJ, Tagare HD (2014) Quantifying the local resolution of cryo-EM density maps. *Nat Methods* 11(1):63–65.

41. Pettersen EF, et al. (2004) UCSF Chimera—A visualization system for exploratory research and analysis. *J Comput Chem* 25(13):1605–1612.

42. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66(Pt 4):486–501.

43. Conway P, Tyka MD, DiMaio F, Konerding DE, Baker D (2014) Relaxation of backbone

bond geometry improves protein energy landscape modeling. *Protein Sci* 23(1):47–55.

44. Wang RY-R, et al. (2016) Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. *eLife* 5:e17219.

45. Schrödinger, LLC (2015) The PyMOL Molecular Graphics System, Version 1.8.