Computational design of co-assembling multi-component protein nanomaterials

Jacob Barile Bale

A dissertation

Submitted in partial fulfillment of the

Requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

David Baker, Chair

Wilhelmus G.J. Hol

Barry L. Stoddard

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Computational design of co-assembling multi-component protein nanomaterials

Jacob Barile Bale

Chair of Supervisory Committee:

Professor David Baker

Biochemistry

Molecular self- and co-assembly of proteins into highly ordered symmetric complexes is an elegant and powerful means of patterning matter at the atomic scale and a hallmark of biological systems. Inspired by the exquisite forms and functions achieved by such protein-based molecular machines and materials in nature, my dissertation has focused on the development of methods for the atomically-accurate design of novel symmetric protein complexes. Specifically, I have focused on the design of materials formed through the co-assembly of multiple copies of two or more distinct protein subunits. The ability to design such multi-component materials with high accuracy has remained an outstanding challenge in the field of protein engineering, but offers great potential for a wide range of applications, including vaccine design, targeted delivery, and renewable energy. Here I present the results of my efforts, including the accurate design of five novel tetrahedral and ten novel icosahedral protein complexes formed through the co-assembly of two distinct types of protein subunits. These results represent a significant advance in protein

design and nanotechnology, opening the door to a new generation of genetically programmable materials tailored to specific applications.

Table of Contents

Acknowledgements

First, I would like to thank Neil King and William Sheffler, who have been my primary mentors and research partners throughout the majority of my graduate work. I consider myself very lucky to have been able to work so closely with such great individuals and I cannot thank them enough for the mentorship and support they have provided me over the course of the last five years. I would also like to thank Justin Siegel for the wonderful opportunities and mentorship he provided me during the initial portion of my graduate career.

This work would not have been possible without the immeasurable contributions of the many other developers of the Rosetta macromolecular modeling package, upon which the methods presented herein are based. For their assistance and/or advice in computational matters, I would specifically like to thank Darwin Alonso, Scott Boyken, TJ Brunette, Javier Castellanos, Aaron Chevalier, Frank DiMaio, Jorge Fallas, Sarel Fleishman, Alex Ford, Per Greisen, Yang Hsia, Possu Huang, Ken Jung, Chris King, Neil King, David La, Tom Linsky, Jeremy Mills, Rocco Moretti, Vikram Mulligan, Una Nattermann, Lucas Nivon, Gustav Oberdorfer, Fabio Parmeggiani, Gabriel Rocklin, Yifan Sang, William Sheffler, Lei Shi, Justin Siegel, Daniel-Adriano Silva, Eva-Maria Strauch, and George Ueda.

For experimental work, Lauren Carter, Daniel Ellis, Jorge Fallas, Jasmine Gallaher, Shane Gonen, Inna Goreshnik, Yang Hsia, Neil King, Jeremy Mills, Una Nattermann, Brooke Nickerson, Rachel Park, Fabio Parmeggiani, Rashmi Ravichandran, Justin Siegel, Chantz Thomas, and Clancey Wolf have all been of great help.

Our collaborators in the Yeates Lab at UCLA and Gonen Lab at Janelia Research Campus have been fantastic to work with and played critical roles in the characterization of the

designed materials. In particular I would like to thank Michael Collazo, Duilio Cascio, Shane Gonen, Tamir Gonen, Yuxi Liu, Dan McNamara, and Todd Yeates.

I would also like to thank the members of my Doctoral Advisory Committee, Jesse Bloom, Wim Hol, Christine Queitsch, and Barry Stoddard for taking time from their busy schedules to provide support and advice about various aspects of my research and career development. And I want to give special thanks to Barry Stoddard for the mentorship and opportunities he provided me during my first year rotation in his lab and to Christine Queitsch for providing me my first research opportunity as a undergraduate at the University of Washington, which helped inspire me to apply to graduate school and eventually led me to join the Baker Lab for my thesis work.

I of course also need to thank my thesis adviser, David Baker, for all of the incredible opportunities he has provided me. His passion for science and fearless approach to research is both infectious and inspiring, and it has been a privilege and an honor to work with him the past five years.

Lastly, and most of all, I would like to thank my wife, Megan, and my family. Words cannot express the depth of my gratitude for the love and support they have provided me throughout the years.

Introduction

Designing protein complexes with spatial order and functionality rivaling those of natural assemblies is challenging due to the complex forces governing protein structures and interactions. It was only within the last few years that methods were first developed for the design of novel symmetric assemblies with atomic-level accuracy[1-4]. While these results represented a significant advance in the field of protein nanomaterials, they were restricted to the design of relatively simple structures consisting of a single type of subunit and offered no means by which the assembly state of the materials could be controlled. We set out to overcome these limitations by developing a new method in which multiple copies of two or more distinct protein subunits are designed to co-assemble into a specific target architecture. This approach provides many potential advantages over the previous state of the art, including a larger search space for design, the ability to independently functionalize the different components, and the ability to control assembly by mixing independently produced components.

Our approach consists of two basic features, symmetric docking and *de novo* protein-protein interface design, intended to identify pairs of amino acid sequences that will fold and co-assemble into desired quaternary structures. Using oligomeric proteins of known structure as the building blocks for design, *de novo* protein-protein interfaces are created between pairs of building blocks to both provide the energetic driving force for assembly and rigidly define the relative orientation of the building blocks such that their symmetry axes align with those of the target symmetric architecture. This method extends and improves upon a previously developed method within the Rosetta macromolecular modeling suite for the design of one-component self-assembling protein nanomaterials[1]. A major update was carried out to Rosetta's symmetry

machinery[5] to allow the simultaneous modeling of multiple distinct subunit types in all of the symmetry groups relevant to protein structure. Using this updated symmetric framework, a new docking algorithm was developed to enable efficient sampling of the additional symmetric degrees of freedom associated with multi-component assemblies and additional functionality was introduced to restrict design to the appropriate interfaces between oligomeric building blocks and to ensure correct calculation of the various metrics used to evaluate the quality of the design models.

Two-component tetrahedra were chosen as the initial design targets used to demonstrate our approach. 57 designs, in two distinct tetrahedral architectures, were experimentally characterized. Five designs were confirmed by size exclusion chromatography and electron microscopy to form materials closely matching the size and shape of the design models, four of which were confirmed at high resolution by X-ray crystallography. In addition, it was shown that assembly of one of the designed materials could be controlled through mixing of independently purified components. These results were published in 2014 in Nature, titled *Accurate design of co-assembling multi-component protein nanomaterials*[6], and comprise the entirety of Section 1 and 2 here.

Low yield of soluble protein was a common issue with the unsuccessful tetrahedral designs and also prevented X-ray structure determination for one of the five successful designs, termed T33-09. This observation, combined with a lack of discernible differences in the calculated metrics of interface quality for successful and unsuccessful design models, suggested that developing methods to increase soluble expression of the designs may improve our design approach. Toward that end, new variants of T33-09 were designed and experimentally characterized in which a subset of the solvent-exposed side chains on each subunit were mutated

to either positively or negatively charged amino acids, a strategy which has previously been shown to be effective at increasing protein solubility[7,8] and is an enticing option for improving our designed nanomaterials as it avoids the need to mutate core or interface 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, which was once again found to match closely with the experimentally determined structure. These results were published in 2015 in Protein Science, titled *Structure of a designed tetrahedral protein assembly variant engineered to have improved soluble expression*[9], and comprise the entirety of Section 3 and 4 here.

Although the results with two-component tetrahedra served as a promising demonstration of our design methods, many applications will likely require the ability to design much larger and more complex protein assemblies. With that in mind, two-component icosahedra were chosen as the next design target; icosahedra possess the highest symmetry of all platonic solids and generate the maximum enclosed volume for symmetric assemblies formed from of a given size protein subunit[10,11]. 183 designs, spanning three distinct icosahedral architectures, were experimentally characterized. Ten were found to form materials closely matching the design models, three of which were validated by X-ray crystallography. Comprised of 120 subunits each, with molecular weights (1.8 to 2.8 MDa) and dimensions (240 to 400 Å diameter) rivaling those of small viral capsids, these are to our knowledge the first icosahedral protein complexes designed with atomic-level accuracy and the largest designed protein materials to date confirmed by X-ray crystallography. The architectures of the designs appear to be unique among known icosahedral protein structures, occupying new regions of the protein assembly universe, which have either not yet been explored by natural evolution or are otherwise undiscovered at present

in natural systems. And the large lumens of the designs, combined with their multi-component nature and potential to control assembly via mixing of in vitro purified components, makes them well suited for encapsulation of a broad range of materials including small molecules, nucleic acids, polymers, and other proteins. These results are currently being written up as an article titled *Accurate design of megadalton-scale, co-assembling icosahedral protein complexes* and comprise the entirety of Section 5 and 6 here.

Section 1: Computational design of two-component tetrahedral protein nanocages

Abstract

The self-assembly of proteins into highly ordered nanoscale architectures is a hallmark of biological systems. The sophisticated functions of these molecular machines have inspired the development of methods to engineer self-assembling protein nanostructures; however, the design of multi-component protein nanomaterials with high accuracy remains an outstanding challenge. Here we report a computational method for designing protein nanomaterials in which multiple copies of two distinct subunits co-assemble into a specific architecture. We use the method to design five 24-subunit cage-like protein nanomaterials in two distinct symmetric architectures and experimentally demonstrate that their structures are in close agreement with the computational design models. The accuracy of the method and the number of two-component materials that it makes accessible could provide a general route to the construction of functional protein nanomaterials tailored to specific applications.

Background and Motivation

The unique functional opportunities afforded by protein self-assembly range from the dynamic cellular scaffolding provided by cytoskeletal proteins to the encapsulation, protection and delivery of viral genomes to new host cells by virus capsids. Although natural assemblies can be repurposed to perform new functions[12,13], this strategy is limited to the structures of existing proteins, which may not be suited to a given application. To overcome this limitation, methods for designing novel self-assembling proteins are of considerable interest[14-17]. The

central challenge in designing self-assembling proteins is to encode the information necessary to direct assembly in the structures of the protein building blocks. Although the complexity and irregularity of protein structures resulted in slow initial progress in this area, advances in computational protein design algorithms and new approaches such as metal-mediated assembly have recently yielded exciting results[1-4,17-23]. Despite these advances, the self-assembling protein structures designed so far have been relatively simple, and continued improvements in design strategies are needed in order to enable the practical design of functional materials.

The level of structural complexity available to self-assembled nanomaterials generally increases with the number of unique molecular components used to construct the material. This is illustrated by DNA nanotechnology, in which specific and directional interactions between hundreds of distinct DNA strands allow the construction of nanoscale objects with essentially arbitrary structures[24-27]. In contrast, designing well-ordered multi-component protein nanomaterials has remained a significant challenge. Multiple distinct intermolecular contacts are necessary to drive the assembly of such materials[14,15,18,19,28], and programming new, geometrically precise interactions between proteins is difficult. Compared to homo-oligomers, multi-component protein nanomaterials offer several advantages: a wider range of possible structures due to their combinatorial nature, greater control over the timing of assembly, and enhanced modularity through independently addressable building blocks. Although multi-component protein assemblies have recently been generated using disulphide bonds[21,29], flexible genetic linkers[19,22,29], or stereotyped coiled-coil interactions to drive assembly[21,22], the flexibility of these relatively minimal linkages has generally resulted in materials that are somewhat polydisperse. Most natural protein assemblies, on the other hand, are constructed from protein–protein interfaces involving many contacts distributed over large interaction surfaces that serve

to precisely define the positions of the subunits relative to each other[30,31]. Advances in computational protein modelling and design algorithms have recently made it possible to design such interfaces[32-36] and thereby direct the formation of novel self-assembling protein nanomaterials with atomic-level accuracy[1-3], but the methods reported so far have been limited to the design of materials comprising only a single type of molecular building block. Here we expand the structural and functional range of designed protein materials with a general computational method for designing two-component co-assembling protein nanomaterials with high accuracy.

Computational Design Method

Our method centres on encoding the information necessary to direct assembly in designed protein–protein interfaces. In addition to providing the energetic driving force for assembly, the designed interfaces also precisely define the relative orientations of the building blocks. We illustrate the method in **Figure 1.1** using the dual tetrahedral architecture (designated here as T33) as an example. In this architecture, four copies each of two distinct, naturally trimeric building blocks are aligned at opposite poles of the three-fold symmetry axes of a tetrahedron (**Figure 1.1a**). This places one set of building blocks at the vertices of the tetrahedron and the other at the centres of the faces, totalling 12 subunits of each protein. Each trimeric building block is allowed to rotate around and translate along its three-fold symmetry axis (**Figure 1.1b**); other rigid body moves are disallowed because they would lead to asymmetry. These four degrees of freedom are systematically explored during docking to identify configurations with symmetrically repeated instances of a novel inter-building-block interface that is suitable for design (**Figure 1.1c**). The score function used during docking favours interfaces with high densities of contacting residues in well-anchored regions of the protein structure that are less

likely to change conformation on mutation of surface side chains (**Figure 1.1d**). RosettaDesign[37,38] is then used to sample the identities and configurations of the side chains near the inter-building-block interface, generating interfaces with features resembling those found in natural protein assemblies such as well-packed hydrophobic cores surrounded by polar rims[31] (**Figure 1.1e**). The end result is a pair of new amino acid sequences, one for each building block, predicted to stabilize the modelled interface and thereby spontaneously drive assembly to the specific target configuration.

These docking and design procedures were implemented by extending the Rosetta software[5,38] to enable the simultaneous modelling of multiple distinct symmetrically arranged protein components. The new protocol allows the different components to be arranged and moved independently according to distinct sets of symmetry operators (**Figure 2.1**). This enables the design strategy described above to be generalized to a wide variety of symmetric architectures in which multiple symmetric building blocks are combined in geometrically specific ways[14,15,28]. Combining even two types of symmetry elements (as in the present study) can give rise to a large number of distinct symmetric architectures with a range of possible morphologies, including those with dihedral and cubic point-group symmetries, as well as helical, layer and space group symmetries[28].

In this study we targeted two distinct tetrahedral architectures: the T33 architecture described above and the T32 architecture shown in **Figure 1.1f**, in which the materials are formed from four trimeric and six dimeric building blocks aligned along the three-fold and two-fold tetrahedral symmetry axes. We docked all pairwise combinations of a set of 1,161 dimeric and 200 trimeric protein building blocks of known structure in the T32 and T33 architectures (Supplementary Methods). This resulted in a large set of potential novel nanomaterials: 232,200

and 19,900 docked protein pairs, respectively, with a given pair often yielding several distinct promising docked configurations. Interface sequence design calculations were carried out on the 1,000 highest scoring docked configurations in each architecture, and the designs were evaluated on the basis of predicted binding energy, shape complementarity[39] and size of the designed interfaces, as well as the number of buried unsatisfied hydrogen-bonding groups (Supplementary Methods). After filtering on these criteria, 30 T32 and 27 T33 materials were selected for experimental characterization (**Figure 2.2**). The 57 designs were derived from 39 distinct trimeric and 19 dimeric proteins, and contained an average of 19 amino acid mutations per pair of subunits compared to the native sequences. The designed interfaces resided mostly on elements of secondary structure, both α-helices and β-strands, with nearby loops generally making minor contributions.

Screening and Characterization of Assembly State

Synthetic genes encoding each designed pair of proteins were cloned in tandem in a single expression vector to allow inducible co-expression in *Escherichia coli* (Supplementary Methods). Polyacrylamide gel electrophoresis (PAGE) under denaturing and non-denaturing (native) conditions was used to rapidly assess the level of soluble expression and assembly state of the designed proteins in clarified cell lysates. For most of the designs, either one or both of the designed proteins was not detectable in the soluble fraction, suggesting that insoluble expression is a common failure mode for the designed materials. Given that the majority of the mutations introduced by our method are changes from polar to hydrophobic residues at the designed interfaces, it is likely that the insolubility of these designs is due to either misfolding or non-specific aggregation of the designed protein subunits. Nevertheless, several designed protein pairs yielded single bands under non-denaturing conditions that migrated more slowly than the

wild-type proteins from which they were derived, suggesting assembly to higher-order species (**Figure 2.3**). These proteins were subcloned to introduce a hexahistidine tag at the carboxy terminus of one of the two subunits and purified by nickel affinity chromatography and size exclusion chromatography (SEC). Five pairs of designed proteins, one T32 design (T32-28) and four T33 designs (T33-09, T33-15, T33-21 and T33-28), eluted together during nickel affinity chromatography and yielded dominant peaks at the expected size of approximately 24 subunits when analysed by SEC (**Figure 1.2a** and **Table 2.1**).

We tested the ability of each of the five materials to assemble *in vitro* by expressing the two components in separate *E. coli* cultures and mixing them at various points after cell lysis (**Figure 2.3**). Native PAGE revealed that in two cases (T33-15 and T32-28) the two separately expressed components efficiently assembled to give the designed materials *in vitro* when equal volumes of cell lysates were mixed (**Figure 1.2b**, **Figure 2.3a,c**). Adjusting the volume of each lysate in the mixture to account for differences in the level of soluble expression of the two components allowed for more quantitative assembly. In the case of T33-15, the two components of the material could also be purified independently: T33-15A and T33-15B each eluted from the SEC column as trimers in isolation. After mixing the two purified components in a 1:1 molar ratio and allowing a two-hour incubation at room temperature, the mixture eluted from the SEC column predominantly at the volume expected for the 24-subunit assembly, with small amounts of residual trimeric building blocks remaining (**Figure 1.2a**). It is thus possible to control the assembly of our designed materials by simply mixing the two independently produced components.

The details of the designed interfaces for the five materials are presented in **Figure 1.3**. The interfaces reflect the hypothesis underlying the design protocol in that they feature well-

packed and highly complementary cores of hydrophobic side chains residing mostly in elements of secondary structure, surrounded by polar side chains lining the periphery of the hydrophobic cores. Quantitatively, the successful designs do not stand out from the others according to the interface metrics used to select designs for experimental characterization (predicted binding energy, shape complementarity, interface size and number of buried unsatisfied hydrogen-bonding groups; **Figure 2.4**). The similarity of the successful and unsuccessful designs according to these structural metrics, combined with the observed insolubility of many of the designs, suggests that focusing on improving the level of soluble expression of the designed proteins could substantially improve the success rate of our approach in the future.

Structural Characterization of the Designed Materials

Negative-stain electron microscopy of the five designed materials confirmed that they assemble specifically to the target architectures (**Figure 1.4**). For each material, fields of monodisperse particles of the expected size and symmetry were observed, confirming the homogeneity of the materials suggested by SEC. Particle averaging yielded images that recapitulate features of the computational design models at low resolution. For example, class averages of T33-09 revealed roughly square or triangle-shaped structures with well-defined internal cavities that closely resemble projections calculated from the computational design model along its two-fold and three-fold axes (**Figure 1.4, T33-09 inset**). Micrographs of T33-15 assembled *in vitro* as described above were indistinguishable from those of co-expressed T33-15 (**Figure 1.4** and **Figure 2.5**), demonstrating that the same material is obtained using both methods.

We solved X-ray crystal structures of four of the designed materials (T32-28, T33-15, T33-21 and T33-28) to resolutions ranging from 2.1 to 4.5 Å (**Figure 1.5** and **Tables 2.2 and**

2.3). In all cases, the structures reveal that the inter-building-block interfaces were designed with high accuracy: comparing a pair of chains from each structure to the computationally designed model yields backbone root mean square deviations (r.m.s.d.) between 0.5 and 1.2 Å (**Figure 1.5 right** and **Table 2.4**). In the structures with resolutions that permit detailed analysis of side-chain configurations (T33-15 and two independent crystal forms of T33-21), 87 of 113 side chains at the designed interfaces adopt the predicted conformations (**Tables 2.5 and 2.6**). As intended, the designed interfaces drive the assembly of cage-like nanomaterials that closely match the computational design models: the backbone r.m.s.d. over all 24 subunits in each material range from 1.0 to 2.6 Å (**Figure 1.5 left** and **Table 2.4**). The precise control over interface geometry offered by our method thus enables the design of two-component protein nanomaterials with diverse nanoscale features such as surfaces, pores and internal volumes with high accuracy.

Discussion

Owing to the unique functions accessible to self-assembling proteins, there is intense interest in engineering protein nanomaterials for applications in various fields. Most efforts so far have focused on repurposing naturally occurring protein assemblies, a strategy that is ultimately limited by the structures available and their tolerances for modification. Similarly, although directed evolution is a powerful method for protein engineering[40,41] and can be used to improve, for example, the packaging capability of existing protein nanocontainers[42,43], it is difficult to envision how it could accurately generate new protein nanomaterials with target structures defined at the atomic level. Our results demonstrate that computational protein design provides a general route for designing novel two-component self-assembling protein nanomaterials with high accuracy. The combinatorial nature of two-component materials greatly expands the number and variety of potential nanomaterials that can be designed. For example, in this study

we used 1,361 protein building blocks to dock over 250,000 distinct protein pairs in two target architectures with tetrahedral point group symmetry, resulting in a very large set of potential nanomaterials exhibiting a variety of sizes, shapes and arrangements of chemically and genetically addressable functional groups, loops and termini. With continued effort to increase the success rate of protein–protein interface design and reduce the rate of designed proteins that express insolubly, it should become possible to simultaneously design multiple novel interfaces in a single material, which would enable the construction of increasingly complex materials built from more than two components.

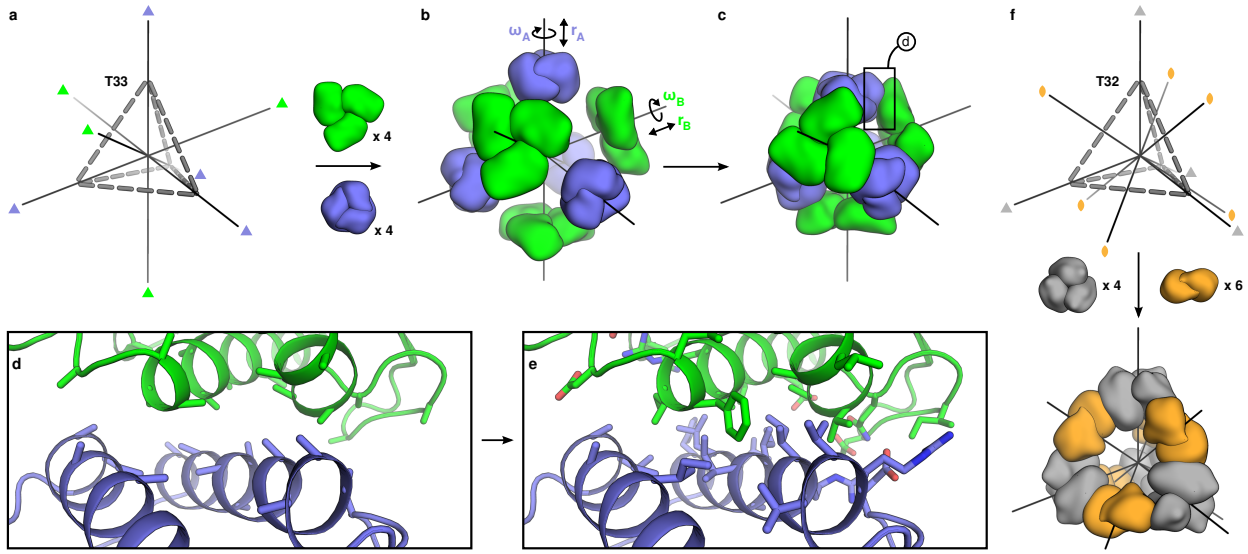
The conceptual framework that underlies our method—symmetric docking followed by protein–protein interface design—can be generally applied to a wide variety of symmetric architectures, including repetitive protein arrays that extend in one, two or three dimensions. Multi-component materials are advantageous in these extended architectures because the uncontrolled self-assembly of a single-component material inside the cell can complicate biological production[16,19,28]. We have shown that the two components of the designed materials T32-28 and T33-15 can be produced separately and mixed *in vitro* to initiate assembly of the designed structure. With new symmetric modelling algorithms capable of handling the additional degrees of freedom associated with these architectures, the accurate computational design and controllable assembly of complex, multi-component protein fibres, layers and crystals should also be possible.

The capability to design highly homogeneous protein nanostructures with atomic-level accuracy and controllable assembly should open up new opportunities in targeted drug delivery, vaccine design, plasmonics and other applications that can benefit from the precise patterning of matter on the subnanometre to 100-nanometre scale. Extending beyond static structure design,

methods for incorporating the kinds of dynamic and functional behaviours observed in natural protein assemblies should make possible the design of novel protein-based molecular machines with programmable structures, dynamics and functions.

Figures

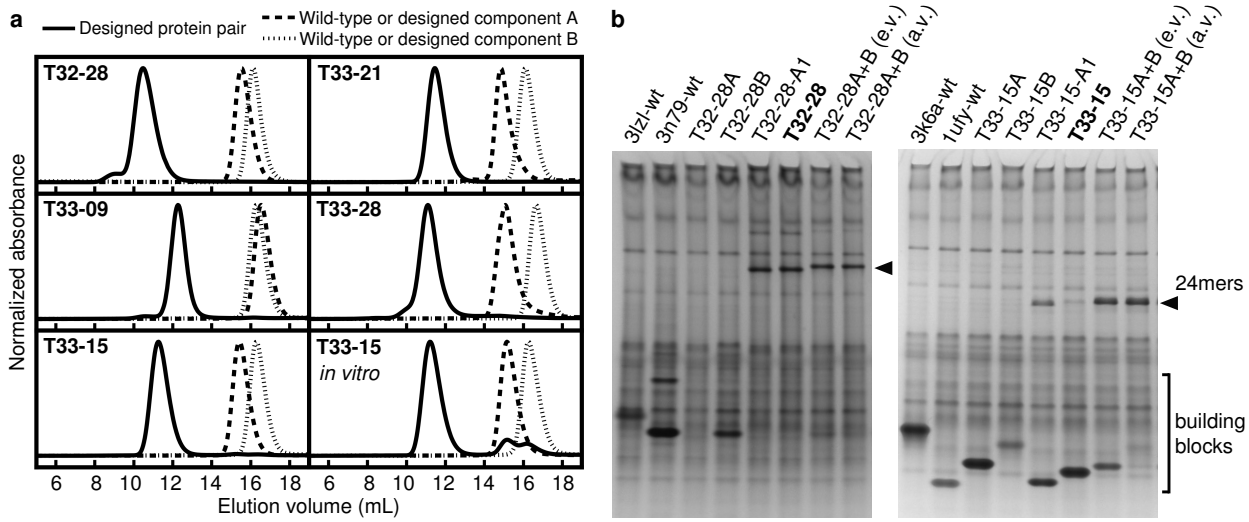
Figure 1.1



Overview of the computational design method. a, The T33 architecture comprises four copies each of two distinct trimeric building blocks (green and blue) arranged with tetrahedral point-group symmetry (24 total subunits; triangles indicate three-fold symmetry axes). **b**, Each building block has two rigid-body degrees of freedom, one translational (r) and one rotational (ω), that are systematically explored during docking. **c**, The docking procedure, which is independent of the amino acid sequence of the building blocks, identifies large interfaces with high densities of contacting residues formed by well-anchored regions of the protein structure. The details of such an interface, boxed here, are shown in **d**. **e**, Amino acid sequences are designed at the new interface to stabilize the modelled configuration and drive co-assembly of

the two components. **f**, In the T32 architecture, four trimeric (grey) and six dimeric (orange) building blocks are aligned along the three-fold and two-fold symmetry axes passing through the vertices and edges of a tetrahedron, respectively.

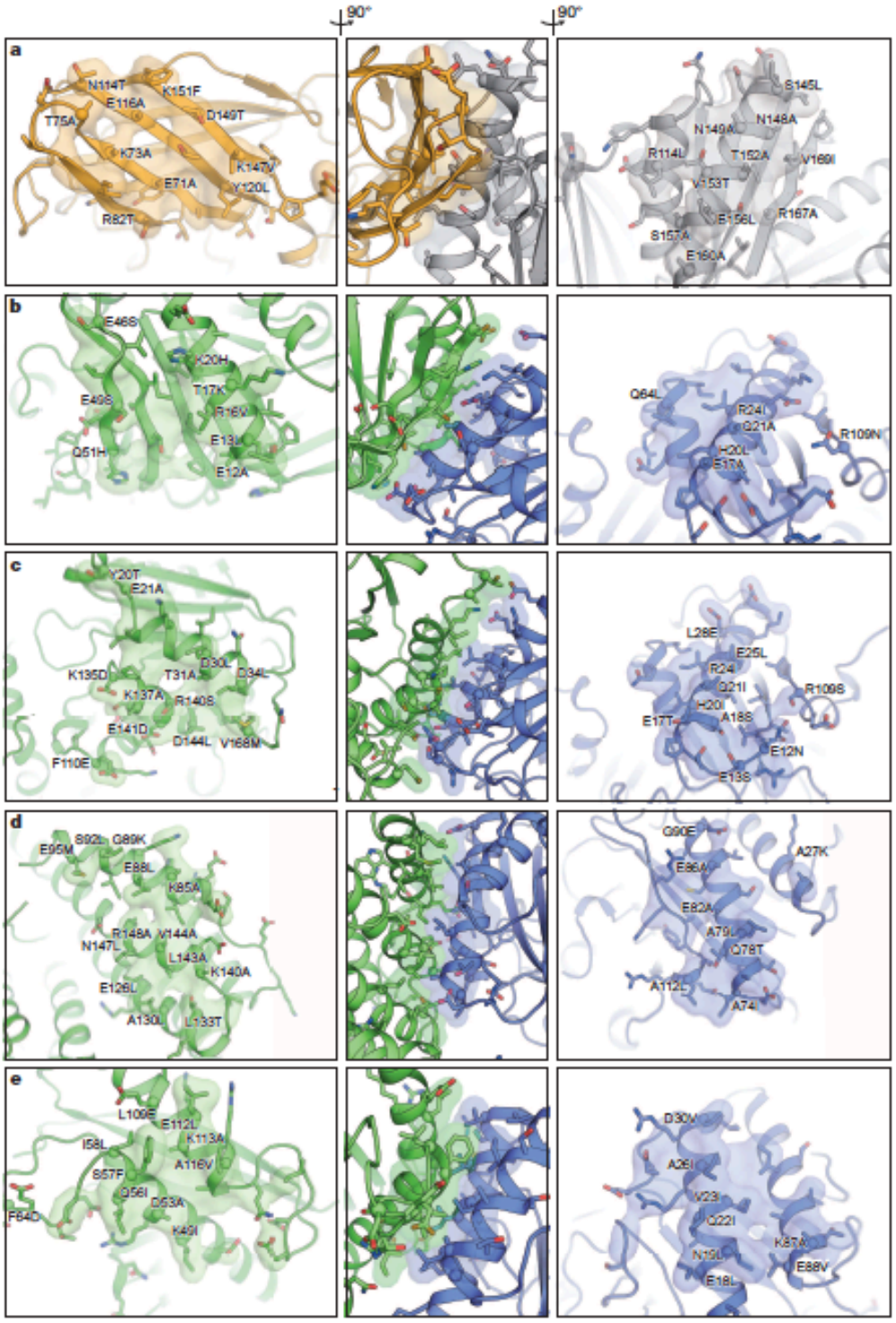
Figure 1.2



Experimental characterization of co-assembly. a, SEC chromatograms of the designed pairs of proteins (solid lines) and the wild-type oligomeric proteins from which they were derived (dashed and dotted lines). The co-expressed designed proteins elute at the volumes expected for the target 24-subunit nanomaterials, whereas the wild-type proteins elute as dimers or trimers. The T33-15 *in vitro* panel shows chromatograms for the individually produced and purified designed components (T33-15A and T33-15B) as well as a stoichiometric mixture of the two components. **b**, Native PAGE analysis of *in vitro*-assembled T32-28 (left panel) and T33-15 (right panel) in cell lysates. Lysates containing the co-expressed design components (A1-tagged, lane 5; hexahistidine-tagged, lane 6) reveal slowly migrating species ("24mers", arrows) not present in lysates containing the wild-type or individually expressed components (lanes 1–4). Mixing equal volumes (e.v.) of crude lysates containing the individual designed components

yields the same assemblies (lane 7), although some unassembled building blocks remain due to unequal levels of expression (particularly for T33-15). When the differences in expression levels are accounted for by mixing adjusted volumes of lysates (a.v.), more efficient assembly is observed (lane 8).

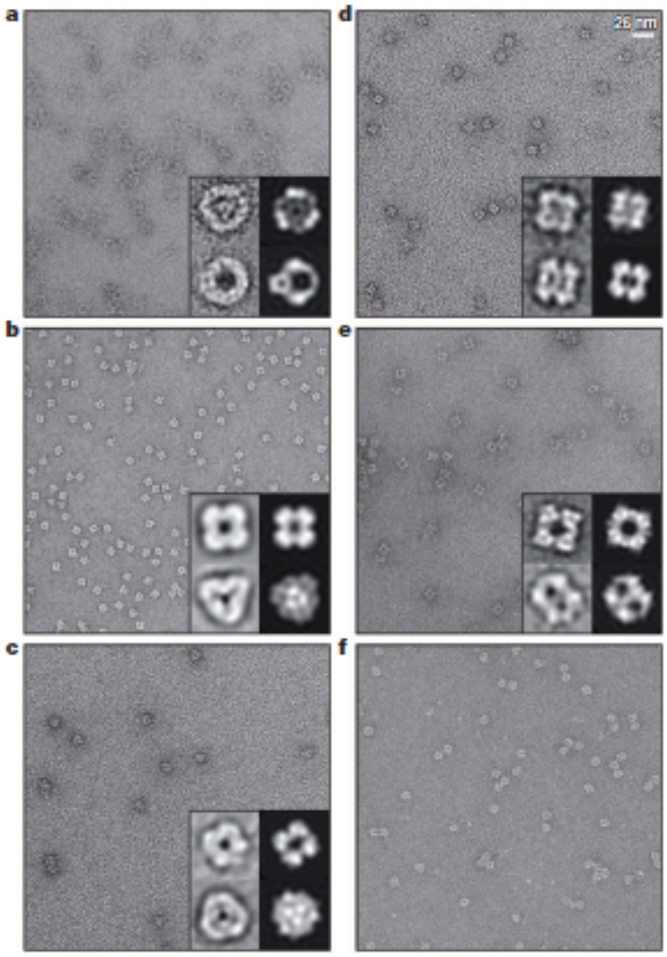
Figure 1.3



Modelled interfaces of designed two-component protein nanomaterials. The models of the designed interfaces in each component of T32-28 (**a**), T33-09 (**b**), T33-15 (**c**), T33-21 (**d**) and

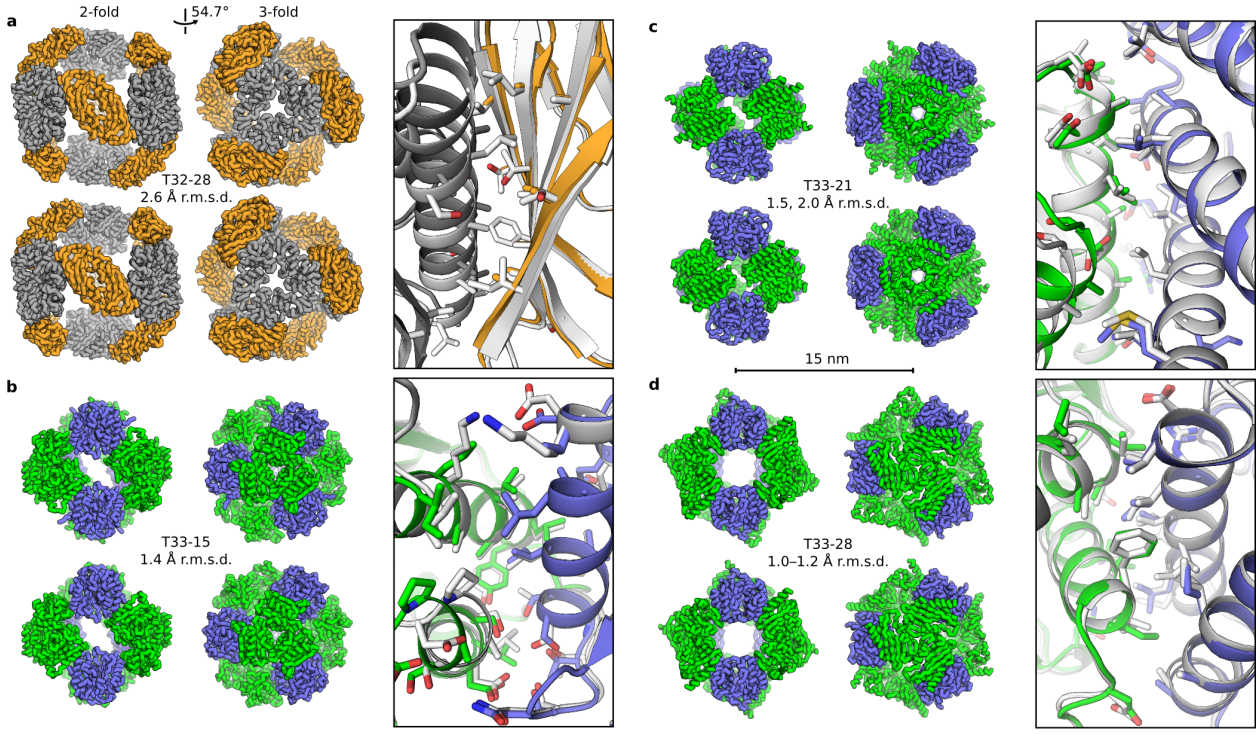
T33-28 (**e**) are shown at left or right, and side views of each interface as a whole are shown at centre (see arrows indicating rotations at top). Each image is oriented such that a vector originating at the centre of the tetrahedral material and passing through the centre of mass of the designed interface would pass vertically through the centre of the image. The side chains of all amino acids allowed to change identity or conformation during the interface design procedure are shown in stick representation. The alpha carbon atoms of positions that were mutated during design are shown as spheres, and the mutations are labelled. To highlight the morphologies of the contacting surfaces, atoms within 5 Å of the opposite building block are shown in semi-transparent surface representation. Oxygen atoms are red; nitrogen, blue; and sulphur, orange.

Figure 1.4



Electron micrographs of designed two-component protein nanomaterials. Negative-stain electron micrographs for co-expressed and purified T32-28 (**a**), T33-09 (**b**), T33-15 (**c**), T33-21 (**d**), and T33-28 (**e**) are shown to scale (scale bar at top right, 25 nm). For each co-expressed material, two different class averages of the particles (top and bottom) are shown in the insets (left) alongside back projections calculated from the computational design models (right). **f**, Micrograph of a T33-15 sample prepared by stoichiometrically mixing the independently purified components (T33-15A and T33-15B) *in vitro* and purifying the assembled material by SEC (see **Figure 1.2**). Micrographs of unpurified, *in vitro*-assembled T33-15 as well as T33-15A and T33-15B in isolation are shown in **Figure 2.5**.

Figure 1.5



Crystal structures of designed two-component protein nanomaterials. The computational design models (top) and X-ray crystal structures (bottom) are shown at left for T32-28 (**a**), T33-15 (**b**), T33-21 (**c**) and T33-28 (**d**). Views of each material are shown to scale along the two-fold and three-fold tetrahedral symmetry axes (scale bar at centre, 15 nm). The r.m.s.d. values between the backbone atoms in all 24 chains of the design models and crystal structures are indicated. For T33-21 (**c**), r.m.s.d. values are shown for both crystal forms (images are shown for the higher-resolution crystal form with backbone r.m.s.d. 2.0 Å), while the r.m.s.d. range for T33-28 (**d**) derives from the four copies of the fully assembled material in the crystallographic asymmetric unit. At right, overlays of the designed interfaces in the design models (white) and crystal structures (grey, orange, green and blue) are shown. Owing to the limited resolution of the T32-28 structure, the amino acid side chains were not modelled beyond the beta carbon. For

the interface overlays, the crystal structures were aligned to the design models using the backbone atoms of two subunits, one of each component.

Section 2: Supplementary information for computational design of two-component tetrahedral protein nanocages

Materials and Methods

Generating Dimeric and Trimeric Scaffold Proteins for Design

Sets of homodimeric and homotrimeric protein structures were curated for use as input to our design protocol. First, the PISA database[44] was searched for all homodimeric or homotrimeric proteins passing the default criteria for dissociation energy, accessible surface area, buried surface area, percent buried surface area, and average chain length. The IDs obtained from PISA were then provided as input for the advanced search tool in the Protein Data Bank (PDB, http://www.rcsb.org/pdb/) to select proteins clustered at 90% sequence identity with: 1) X-ray resolution less than 2 Å, 2) chain lengths of 75 to 200 amino acids, and 3) *Escherichia coli* as the host organism for protein expression.

Coordinates for each of the selected PDB IDs were downloaded from the biological assemblies in the PDB (ftp://ftp.wwpdb.org/pub/pdb/data/biounit/coordinates) and standardized for input to Rosetta. For biological assemblies containing multiple models with one chain per model, each model was treated as a separate chain. For assemblies containing multiple models with multiple chains per model, only the first model was considered. 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[45]) 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. A new biological assembly was created in PyMOL by superimposing copies of the design input chain

onto all other chains, and the assembly's symmetry axis was aligned along the vector [0,0,1] and its center of mass translated to the origin. Assemblies were discarded that were found to be too asymmetric, as assessed by the dispersion of symmetry axes implied by each tuple of symmetrically related atoms. For PDB IDs with multiple biological assemblies, the assembly with the lowest biological unit number found to match the expected C2 or C3 symmetry was chosen for design. The final set of 1,161 homodimeric and 200 homotrimeric proteins are listed in **Tables 2.7 and 2.8**, respectively.

Multi-component Symmetric Modeling

A major update to Rosetta's existing symmetric modeling framework[5] was carried out to enable modeling of multi-component systems; that is, systems consisting of multiple distinct protein subunits, each associated with a distinct symmetry group. Within this updated framework, each distinct subunit can be modified independently of one another, with the changes propagated to all symmetrically related copies. All of Rosetta's design and modeling functionality accessible to one-component symmetries is now accessible to multi-component symmetries as well, including efficient scoring calculations and sampling of symmetric degrees of freedom (DOFs). These changes to Rosetta's symmetry machinery are illustrated in Extended **Figure 2.1** and described briefly below. In both the one-component and multi-component case, the symmetry of a given target architecture (for visual examples of the T32 and T33 architectures, see **Figure 1.1**) is passed to Rosetta in the form of a symmetry definition file. The multi-component symmetry definition file syntax is largely the same as the one-component syntax, with the additional requirement that the jumps connecting the protein subunits to the fold tree must specify which component is connected to each symmetry element (the T32 and T33 symmetry definition files used in this study are provided as examples in the **Supplementary**

Files via the zipped archive T32_T33_design_models.zip). Previously, only a single connection was allowed from the symmetric fold tree into the asymmetric unit. Thus, when modeling a system with multiple distinct symmetric components, only one such component could have its internal DOFs preserved. For example, in the D32 system shown in **Figure 2.1**, if only one connection into the asymmetric unit is allowed, then one must choose to root the connection of the two subunits in the asymmetric unit to either the two-fold axis (**Figure 2.1a**) or the three-fold axis (**Figure 2.1b**). If both are connected to the three-fold axis, for example, rotations around this connection will correctly preserve the internal DOFs of the trimer, but disrupt the internal DOFs of the dimer (**Figure 2.1b**). By enabling multiple connections from the symmetric fold tree into the asymmetric unit, the multi-component extension of symmetric modeling in Rosetta allows the asymmetric unit to be broken down into substructures that are independently managed by the symmetric fold tree. Using a multi-component symmetric fold tree in our D32 example allows the trimer to connect directly to the three-fold axis and the dimer to connect directly to the two-fold axis, thus any motions allowed by the symmetric architecture preserve the internal DOFs of both building blocks (**Figure 2.1c**).

Two-component Symmetric Docking

An application was written within Rosetta to dock two distinct oligomeric building blocks in higher order symmetries in order to identify docked configurations suitable for interface design. The required inputs for the `tcdock` application are one PDB file containing a single subunit of the first scaffold component and a second PDB file containing a single subunit of the second scaffold. First, the subunits are arranged at the origin according to the symmetry specified by command-line options. Then the full space of connected symmetric configurations is sampled by systematically varying the translational and rotational degrees of freedom in the

system. In the cubic point group symmetries modeled in the present study, the two translational degrees of freedom are represented internally in a polar-coordinate form such that the separation between components can be determined by a slide operation, thus sampling only configurations in which the two components are in direct contact. This reduces the search space from four to three dimensions, making a search through all possible configurations formed from all possible combinations of a large set of building blocks computationally tractable. In order to test all four possible orientations of the two building blocks (inside/inside, inside/outside, outside/inside, outside/outside), two separate docking runs are performed in which the orientation of one of the building blocks is reversed by setting the command-line option `tcdock::reverse` to true. Configurations in which backbone or beta carbon atoms from different building blocks clash (distance between backbone amide nitrogen and carbonyl oxygen atoms <= 2.6 Å; distance between all other backbone/beta carbon atom pairs <= 3.0 Å) are discarded. In each non-clashing configuration, a measure of the suitability of the configuration for design ("designability") is calculated. While several command line options exist to modify the exact nature of the designability score, in this study it was calculated as the sum of the number of beta carbon contacts between building blocks (where a contact is defined as two beta carbon atoms within 12 Å), weighted by the type of secondary structures on which the contacting positions exist (by setting the `tcdock::cb_weight_secstruct` command line option to true) and the average degree of connectivity[46] of the contacting positions within the building block (by setting the `tcdock::cb_weight_average_degree` command line option to true). The designability score was thus calculated to favor the selection of docked configurations with large numbers of contacting residues on well-anchored regions of protein structure. In addition to contacts between building blocks of the two different components, two-component systems can also

possess contacts between symmetrically related building blocks of the same component. We refer to these as inter- and intra-component contacts, respectively. The command-line options `tcdock::intra`, `tcdock::intra1`, and `tcdock::intra2` control the contribution to the designability score of intra-component contacts between both components, between component 1, and between component 2, respectively.

Data and PDB files (optional) are output for a user-defined number of top scoring docked configurations (set by the `tcdock::topx` command-line option). The data, which can be saved by redirecting the output of the run to a log file, includes for each configuration the values for each of the rigid body DOFs that define that configuration, the designability score, the number of carbon beta contacts between building blocks, the number of contacting residues between building blocks, the average score per carbon beta contact, and the average score per contacting residue. A two-component docking example, including the specific command-line options used in this study, can be found in the **Supplementary Files** via the zipped archive T32_T33_docking.zip.

In this study, the 1,161 dimers and 200 trimers in our scaffold sets provided 232,200 unique pairwise combinations of trimers with dimers (T32), and 19,900 unique pairwise combinations of trimers (T33). Docking was carried out for each of these unique combinations with or without the `tcdock::reverse` option set to true, for a total of 504,200 independent docking trajectories. The `tcdock::intra` option was set to false such that intra-component contacts were not included in the calculated scores.

For each unique scaffold combination, the 3 top scoring T33 docks were selected. This set of 59,700 distinct configurations was ranked by the average designability score per contacting residue and the top 1,000 used as input for interface design. For T32, data was output for the 40

top scoring docked configurations per docking trajectory. This set of 18,576,000 distinct configurations was filtered to remove all configurations with less than 80 contacting residues between building blocks and ranked by the average designability score per contacting residue. This set was filtered to retain only the one top ranked configuration for each unique scaffold pair and the top 1,000 configurations were used as input for interface design.

Two-component Protein-protein Interface Design

A set of protein-protein interface design protocols was developed within Rosetta to identify mutations predicted to drive assembly of two distinct protein building blocks into higher order symmetric complexes. The design functionality was broken into modular components (Mover, Filter, and TaskOperation classes[38,47]) to facilitate future code development and reuse in other design applications. All components of the protocol can be accessed using the RosettaScripts user interface[47], which provides users the ability to rapidly rearrange and modify each step of the design process without having to change the underlying C++ code. We describe below the specific implementation that was used in this work.

The design process was split into four stages: I) interface design, II) shape complementarity optimization, III) automated reversion, and IV) resfile-based refinement. This process does not include any explicit "negative design" steps where alternative structural states are considered, but rather focuses on stabilizing a particular structural state and relying on the requirement for geometrical and chemical complementarity to produce stable protein-protein association to enforce interaction specificity. 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; the latter can be produced simply by concatenating the two scaffold protein PDB files used as input for docking (see **Generating dimeric and trimeric scaffold proteins for design**) and changing

the chain of the second subunit to "B". In addition, initial values for the translational and rotational symmetric rigid body DOFs, obtained from docking, must be provided in order to define the starting symmetric configuration. All design calculations are performed on the two independent subunits and propagated symmetrically[3] (see **Multi-component symmetric modeling**). Unless specified otherwise, all calculations were performed with the standard `score12` scorefunction.

Stage I: *Interface design.* During the initial stage, a combinatorial optimization of the side chains at the inter-building block interface is performed to replace the original amino acids from the input scaffold with residues predicted to form a low-energy protein-protein interface. Most of the mutations present in finalized designs selected for experimental characterization are introduced during Stage I.

Multiple design trajectories are carried out for each docked configuration. At the start of each trajectory, the symmetric rigid body DOFs are perturbed in order to sample nearby docked configurations. The behavior of these perturbations can be set by the user, including specifying whether to sample values from a user-defined grid of angles and displacements or randomly from user-defined uniform or Gaussian distributions of angles and displacements. Trajectories yielding docked configurations with clashing backbones (distance between backbone amide nitrogen and carbonyl oxygen atoms <= 2.6 Å; distance between all other backbone/beta carbon atom pairs <= 3.0 Å) are discarded prior to interface design based on user-defined cutoff values for the allowable number of clashing atoms. In each of the remaining trajectories, interface residues are selected according to the following three criteria: 1) the residue must have a beta carbon (alpha carbon in the case of glycine) within a user-defined cutoff distance to a beta carbon (alpha carbon in the case of glycine) in a different building block (in this study the default

10 Å cutoff was used), 2) the residue must have a nonzero solvent accessible surface area (calculated with a 2.2 Å radius probe) when the protein subunits are in the unbound state, and 3), with the exception of residues that have high Lennard-Jones repulsive scores (`fa_rep`), the residue must not be making contacts (any heavy atoms within 5 Å) with other subunits in the same oligomeric building block. Residues matching all three criteria are considered designable, with the exception of proline and glycine, which are restricted to repacking. An expanded set of residues, in which criteria 3 is not enforced, is also used in certain portions of the protocols outlined in stages I through IV. 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 are therefore also interface positions, but not all interface positions are design positions. These positions are updated at multiple points throughout stages I through IV; appending any positions that newly satisfy the selection criteria to the previously defined sets. All residues not in the selected sets remain fixed throughout the design process. In addition, mutations to proline, glycine, or cysteine are prohibited unless explicitly specified otherwise by the user via a resfile (see Stage IV). Optionally, a reduced amino acid set can be used during Stage I such that only the native amino acid and a subset of the 20 common amino acids are considered during design at each design position.

Once the design positions have been selected, an initial round of design is carried out using the standard RosettaDesign algorithm[37] and a version of the Rosetta scorefunction, `soft_rep`, in which the Lennard-Jones repulsive term (`fa_rep`) is down-weighted to favor tightly packed interfaces. The scorefunction is then set back to `score12` and the Rosetta energy is minimized through a series of small changes to the design position side chain configurations and the symmetric rigid body DOFs (i.e., the side chains and rigid body DOFs are symmetrically

"minimized"). Designs with contacting interface areas not meeting user-defined thresholds are optionally discarded. For those designs passing the interface area cutoffs, the design positions are updated and a second round of interface design is carried out using the standard RosettaDesign algorithm with the `score12` scorefunction. The design position side chains are then repacked and the interface position side chains and rigid body DOFs subjected to another round of minimization.

Several metrics are used to gauge the quality of the interfaces resulting from this first stage of design and to select designs to carry forward to shape complementarity optimization in Stage II. These metrics include: 1) the number of buried unsatisfied hydrogen bonds at the designed interface, 2) the shape complementarity[39] of the designed interface, and 3) the predicted binding energy of the interface, defined as the difference in energy between the bound and unbound (individual building blocks) state following repacking of the side chains at the design positions and minimization of the side chains at the interface positions in the unbound state. For each passing design: 1) the values of the final rigid body DOFs are output to a scorefile along with the metric values and the standard `score12` score terms and 2) a standard Rosetta resfile is generated containing each of the design positions and their amino acid identities.

In this study, 100 independent design trajectories were run for each of the top 1,000 docked T32 and T33 configurations (*vide supra*). At the start of each trajectory, the building blocks were displaced 2 Å away from the assembly's center of mass along their symmetry axes (to allow space for side chain packing since the structures were modeled without side chain atoms beyond the beta carbon during docking), and the translational and rotational rigid body DOFs were perturbed by sampling randomly from Gaussian distributions with standard deviations of 0.75 Å and 2 degrees, respectively. Perturbed configurations yielding more than 8

clashing backbone atoms were removed from further consideration. A reduced amino acid set was employed during this stage of the design process such that only retention of the wild-type amino acid or mutations to the following 8 amino acids were allowed: alanine, aspartate, isoleucine, leucine, asparagine, serine, threonine, and valine. This reduced amino acid set was used in order to strongly bias the residues at the designed interfaces towards amino acids with smaller, less flexible side chains. These residues, having fewer preferred side chain configurations than residues with more rotatable bonds (e.g., methionine), should tend to be more "preconfigured"[48] towards the designed configurations. Several recent design studies, including our own, have suggested preconfiguration as a common feature in successful designs[1,35,49,50]. Additionally, during all RosettaDesign steps in all stages, the chi2 angle for aromatic side chains being repacked or designed was restricted to between 70 and 110 degrees.

T32 design trajectories yielding contacting interface areas of less than 1,100 $Å^2$ or greater than 2,000 $Å^2$ following the first round of design were discarded. The passing T32 designs were further filtered at the end of Stage I to remove those that had more than 45 mutations or 8 buried unsatisfied hydrogen bonds at the designed interface, a predicted binding energy greater than -12 REU, or a shape complementarity score of less than 0.60. The T33 design trajectories were filtered based on contacting interface areas at the end of Stage I rather than after the first round of design, discarding those that yielded contacting interface areas of less than 600 $Å^2$. The passing T33 designs were further filtered to remove those with more than 100 mutations or 10 buried unsatisfied hydrogen bonds at the designed interface (both essentially placeholder values not expected to aggressively filter designs), a predicted binding energy greater than -12 REU, or a shape complementarity of less than 0.55. The resulting 1,292 T32 designs and 593 T33 designs were subjected to the protocol described in Stage II below.

Stage II: *Shape complementarity optimization.* Shape complementarity is a geometric measure of the quality of the puzzle-like packing in proteins[39]. Crystal structures of natural protein-protein interfaces show strong conservation of high shape complementarity relative to irrelevant protein-protein contacts such as those arising from crystal packing[31], suggesting the importance of this property in forming stable and well-defined interactions. In Stage II, mutations are identified that improve the shape complementarity of the designed interfaces without disrupting other score metrics. Shape complementarity is an orthogonal metric to the Rosetta energy function, and therefore optimizing this metric explicitly as described below can in some cases introduce mutations that would have scored marginally poorly during sequence optimization in Stage I. However, due to the large computational cost associated with calculating shape complementarity and the small number of mutations that resulted in this study from Stage II, in the future we suggest considering explicit shape complementarity optimization only for designs of special interest that have low shape complementarity scores.

The first step in Stage II is to regenerate the initial design from the two input scaffolds: 1) the new rigid body DOFs output from Stage I are used to reposition the subunits in the fully assembled state, 2) the interface positions are re-selected using the same criteria as before, with the exception that all positions specified in the input resfile are included regardless of whether or not they fulfill the criteria in the input state, 3) the resfile output from stage I is used as input to the RosettaDesign algorithm to reintroduce the initial design mutations, and 4) the interface position side chains are subjected to one or more rounds of minimization and/or repacking. Next, greedy optimization[51] is used to test individual reversions to the native amino acid at all mutated residues. This reversion step has since been deprecated in favor of the reversion protocol outlined in Step III, but is described here since it was used to produce the designs

reported in this study. A custom reversion score is used in which individual mutations are filtered to remove those that increase the number of buried unsatisfied hydrogen bonds at the designed interface and scored according to the sum of the predicted binding energy, the total Rosetta energy, and a residue type constraint energy favoring the native amino acid. The potential reversions are then combined one at a time proceeding from the individually best scoring to worst scoring reversions at each position, only accepting those that do not increase the number of buried unsatisfied hydrogen bonds at the designed interface and improve the reversion score in the context of all previously accepted mutations (the buried unsatisfied hydrogen bond criterion is optional; it was used for the T32 designs, but not T33). Following another one or more rounds of side chain minimization and/or repacking at the interface positions, greedy optimization is used to increase the shape complementarity of the designed interface as follows. First, mutations to all amino acids except cysteine, glycine, and proline are tested individually at each design position (defined by the input resfile). Each mutation is ranked by the shape complementarity of the design interface as long as the mutation does not: 1) increase the total Rosetta energy by more than 2.0 Rosetta energy units (REU), 2) decrease the predicted binding energy by 1.0 REU, 3) introduce any new unsatisfied hydrogen bonds, or 4) increase the Dunbrack energy (`fa_dun`) by more than 2.5 REU (the `fa_dun` criterion is optional; it was used for the T32 designs, but not T33). Next, mutations are combined one at a time proceeding from the best scoring to worst scoring individual mutations, only accepting those that still pass the same three or four criteria and improve the shape complementarity in the context of all previously accepted mutations. During both the reversion and shape complementarity optimization, all of the interface positions are subjected to one round of minimization, repacking, and minimization prior to evaluating the effects of each mutation.

In addition to the standard Rosetta scores, the following metrics are used to assess the quality of each design following one or more rounds of interface position side chain minimization and/or repacking: 1) the total number of mutations, 2) the number of buried unsatisfied hydrogen bonds at the interface, 3) the average degree of each design position 4) the RosettaHoles[52] packing score, 5) the average total Rosetta energy, `fa_rep`, `fa_dun`, and Lennard-Jones attractive energy (`fa_atr`) for each filter position, 6) the contacting interface area, 7) the predicted binding energy, 8) the shape complementarity, and 9) the change in predicted binding energy resulting from individual mutations of each interface side chain to alanine (i.e., a computational alanine scan of the designed interface). Those designs passing a set of user-defined thresholds for each metric are subsequently subjected to visual inspection to further filter the designs. A scorefile with the metric values and the standard `score12` score terms, and a resfile containing the design positions and their amino acid identities are generated for each design at the end of Stage II.

In this study, the designs resulting from Stage II were filtered to remove those with a shape complementarity score less than 0.65 (T32 only), predicted binding energies of greater than -25 REU, a positive RosettaHoles score for the designed interface (T32 only), an interface area less than 1,200 $Å^2$, or more than 1 buried unsatisfied hydrogen bond at the designed interface. The 283 passing T32 designs and 107 passing T33 designs were visually inspected and manually curated down to a list of 68 T32 designs and 38 T33 designs that were subjected to the reversion protocol outlined in Stage III.

Stage III: *Automated reversion*. The purpose of this third stage in the design process is to identify, via an automated computational process, mutated residues predicted to not be critical for assembly and to revert them back to their native amino acid identities. This helps to

minimize the number of mutations being made to the scaffold proteins and reduces the amount of refinement required in Stage IV. Many of the amino acid changes made during this Stage constitute reversion of surface hydrophobic residues that were introduced during Stages I and II back to the wild-type, often polar, amino acid.

The first step is to regenerate the design from the two input scaffolds using the rigid body DOFs from stage I and the resfile output from stage II: 1) the rigid body DOFs are used to reposition the subunits in the fully assembled state, 2) the interface positions are re-selected in the same manner as in Stage II, 3) the resfile is used as input to the RosettaDesign algorithm to reintroduce the initial design mutations, and 4) one round of interface position side chain and rigid body DOF minimization, side chain repacking, and minimization is performed. Next, greedy optimization is used to revert mutations to the native amino identities as follows. During the first part of the optimization algorithm, each reversion is tested individually and ranked by the change in shape complementarity as long as the reversion does not: 1) decrease the predicted binding energy by more than 2.0 REU, 2) increase the number of buried unsatisfied hydrogen bonds at the interface, or 3) decrease the shape complementarity of the interface by more than 0.02. During the second stage, reversions that passed the first stage are combined one at a time proceeding from the best scoring to the worst scoring individual mutations, only accepting those that still pass the three criteria above in the context of all previously accepted mutations. Optimization is terminated if a mutation passes these criteria but causes the predicted binding energy to be greater than a user-defined threshold (in this study, -15 REU was used for the T32 designs and -17 REU for the T33 designs) or the shape complementarity to be less than 0.65. During both stages of optimization, all interface positions are subjected to one round of minimization, repacking, and minimization prior to evaluating the effects of each mutation.

Furthermore, during the combining stage, the reference structure for measuring the change in shape complementarity is reset after each accepted mutation.

Following one round of rigid body and side chain minimization, side chain repacking, and minimization, the full suite of additional metrics are once again evaluated (as outlined at the end of Stage II) with the additional calculation of a Boltzmann weighted estimation of the probability of each designed side chain configuration in the unbound state[48]. For each design, the values of the final rigid body DOFs are output to a scorefile along with the additional metrics and the standard `score12` score terms, and a resfile is generated containing the interface positions and their amino acid identities.

All 68 T32 designs and 38 T33 designs resulting from Stage III were run through the resfile-based refinement protocols outlined in Stage IV below.

Stage IV: *Resfile-based refinement*. The final stage of the design process involves 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 is used to generate modified resfiles for each design containing a small number of user-defined mutations relative to the resfiles output from Stage III. These mutations are often reversions to the wild-type amino acid that scored poorly during Stage III but are judged visually to be compatible with assembly. With recent improvements to the Rosetta energy function that were implemented after this work was performed[53] and improved design protocols (J.B.B, unpublished data), we believe that fully automated design without user-guided mutations will soon be practicable. Two different protocols, `resfile_optimize` and `resfile_design`, are used to test the user-defined mutations. In both protocols, the starting configuration is generated from the two input scaffolds using the rigid body DOFs from the previous round of design.

The `resfile_optimize` protocol uses greedy optimization to test the user-defined mutations. First the reverted design resulting from Stage III is regenerated using the unmodified resfile output from Stage III together with the standard RosettaDesign algorithm, and the side chains specified in the resfile are minimized, repacked, and minimized. Next, user-defined mutations are tested individually at each design position. Each mutation is ranked by the change in shape complementarity of the designed interface, if the mutation does not: 1) decrease the predicted binding energy by greater than 2.0 REU or 2) decrease the shape complementarity of the designed interface by more than 0.02. The passing mutations are then combined one at a time proceeding from the best ranked to the worst ranked individual mutations, only accepting those that still do not decrease the binding energy by more than 2.0 REU or the shape complementarity by more than 0.02 in the context of all previously accepted mutations. Optimization is terminated if a mutation passes these criteria, but causes the predicted binding energy to be greater than -15 REU or the shape complementarity to be less than 0.63. All positions specified in the input resfile are subjected to one round of minimization, repacking, and minimization prior to evaluating the effects of each mutation. Furthermore, during the combining stage, the reference structure for measuring the change in predicted binding energy and the change in the shape complementarity is reset after each accepted mutation.

The `resfile_design` protocol on the other hand, simply takes the starting design configuration generated using the rigid body DOFs from the previous round of design and applies the standard RosettaDesign algorithm with the user-defined resfile.

In both protocols, the symmetric rigid body DOFs and the side chains specified in the input resfile are minimized, side chains repacked, and minimized prior to calculating the full suite of design metrics. This process is iterated until designs are obtained which are deemed

suitable for experimental characterization or until the user decides the designs are no longer worth pursuing. In this study, 30 T32 and 27 T33 designs were selected for experimental characterization (**Figure 2.2**).

Source Code, Examples, and Design Models

Source code is freely available to academic users through the Rosetta Commons agreement (http://www.rosettacommons.org); SVN revision 52869 was used to perform the design calculations in this study. Examples of each stage of the two-component docking and interface design process can be found in the **Supplementary Files** via the zipped archives T32_T33_docking.zip and T32_T33_design.zip. The design models of T32-28, T33-09, T33-15, T33-21, and T33-28 in PDB format, along with corresponding symmetry definition files, are provided in the **Supplementary Files** via the zipped archive T32_T33_design_models.zip.

Protein Expression, Lysate Screening, and Purification

As described in the main text, each designed material comprises a pair of designed proteins. The two components are referred to here by the name of the designed material followed by the suffix "A" or "B". Enumerated in **Table 2.1** are the amino acid sequences for the five designs that were experimentally characterized in detail in this study (T32-28, T33-09, T33-15, T33-21, and T33-28) along with the wild-type proteins from which these designs were derived (referred to by their PDB ID followed by the suffix "-wt"). The amino acid sequences of the two C-terminal tags used in this study are also presented in **Table 2.1**.

Codon-optimized genes encoding the designed and corresponding wild-type proteins were either purchased (Gen9) or constructed from sets of purchased oligonucleotides (Integrated DNA Technologies) by recursive PCR[54]. All genes were cloned using the Gibson assembly

method[55] into a variant of the pET29b expression vector (Novagen) that had been digested by NdeI and XhoI restriction endonucleases (New England Biolabs). The genes encoding the wild-type proteins were each cloned into the vector individually, while the genes encoding the designed proteins were cloned in pairs along with the following intergenic region derived from the pETDuet-1 vector (Novagen):

5′-TAATGCTTAAGTCGAACAGAAAGTAATCGTATTGTACACGGCCGCATAATCGAAATTAATACGACTCACTA
TAGGGGAATTGTGAGCGGATAACAATTCCCCATCTTAGTATATTAGTTAAGTATAAGAAGGAGATATACAT-3′

The constructs for the designed protein pairs thus possessed the following set of elements from 5′ to 3′: NdeI restriction site, upstream gene, intergenic region, downstream gene, XhoI restriction site. The upstream genes encoded components denoted with the suffix "A"; the downstream genes encoded the "B" components (**Table 2.1**). This allowed for co-expression of the designed protein pairs in which both the upstream and downstream gene had their own T7 promoter/*lac* operator and ribosome binding site.

The pET29b variant used for the initial constructs appended the A1 peptide tag (**Table 2.1**) to the C terminus of each wild-type gene and to the downstream gene of each designed protein pair for fluorescent labeling via the AcpS system[56]. For purification purposes, vectors encoding C-terminally His-tagged versions of the designed protein pairs, the individual protein components, and the corresponding wild-types were subsequently constructed by subcloning (via Gibson assembly) into the standard pET29b vector between the NdeI and XhoI restriction sites. As with the A1 peptide tag, the 6xHis tag was only appended to the downstream component in the co-expression constructs.

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 OD_{600} 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 assembly by subjecting cleared lysates to native (non-denaturing) PAGE as described previously[1]. Single bands for each of the five successful materials were visible when stained with GelCode Blue (Thermo Scientific, **Figure 1.2b**, **Figure 2.3 lane 5**). In these initial screens, all constructs were tested under both the 22° C and the 37° C expression conditions. Based on these results, in all subsequent work T32-28, T33-28, and the corresponding wild-type proteins were expressed at 22° C, while T33-09, T33-15, T33-21, and the corresponding wild-type proteins were expressed at 37° C.

For purification, 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, and the lysates were cleared by centrifugation and filtered through 0.22 μM filters (Millipore). For T32-28A and 3LZL-wt containing samples, DTT was excluded from all buffers and 1mM $CuSO_4$ added to the lysis buffer in accordance with previous work on the 3LZL-wt protein that revealed copper binding sites at the dimeric interface and putative copper-dependent dimerization[57]. The proteins were purified from the filtered supernatants by nickel affinity chromatography on HisTrap HP columns (GE Life Sciences) and eluted using a linear gradient of imidazole (0.02–0.5 M). Fractions containing pure protein(s) of interest were pooled, concentrated using centrifugal filter devices (Sartorius Stedim Biotech), and further purified on a Superdex 200 30/100 gel filtration column (GE Life Sciences) 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, *in vitro* mixing, electron microscopy, and X-ray crystallography.

Analytical Size Exclusion Chromatography

Analytical SEC was performed on a Superdex 200 30/100 gel filtration column (GE Life Sciences) using 25 mM TRIS pH 8.0, 150 mM NaCl, 1 mM DTT as the running buffer. The designed materials were loaded onto the column with each component present at a subunit concentration of 50 μM. Individual designed components and wild-type proteins were loaded 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 a set of globular protein standards.

In vitro Mixing

Individual components of the five successful designs were expressed from pET29b vectors encoding C-terminally His-tagged versions of each component (under the same induction conditions outlined above). Lysates containing corresponding pairs of designed components were mixed either immediately following lysis (crude lysates) or after clearance by centrifugation (cleared lysates). Each was mixed in either a one-to-one volumetric ratio or with adjusted volumetric ratios intended to account for observed differences in expression levels of the two components in each designed pair. After incubating for two hours at room temperature, insoluble material was cleared by centrifugation and the samples subjected to native PAGE analysis. For comparison, these samples were analyzed together with cleared lysates of unmixed component A and B, and cleared lysates from co-expressed A1-tagged designs, co-expressed His-tagged designs, and corresponding His-tagged wild-types (**Figure 2.3**). Bands

corresponding to the assembled state were clearly visible in the crude lysate mixtures of T32-28 and T33-15 (**Figure 1.2b**, **Figure 2.3a, c lanes 7–8**). Corresponding bands for T32-28 and T33-15 were also visible in the cleared lysate mixtures, although noticeably less intense in the case of T32-28 (**Figure 1.2b**, **Figure 2.3a, c lanes 9–10**). It is also noteworthy that while the A1-tagged co-expression construct of T33-09 (**Figure 2.3b lane 5**) yielded a visible band for the assembled material, such a band was not clearly visible in the lysate of the His-tagged co-expression construct (**Figure 2.3b lane 6**). The low yield of purified His-tagged construct precluded crystal trials for this designed material. However, T33-09 clearly expresses solubly and assembles to the target architecture as shown by size exclusion chromatography (**Figure 1.2a**) and electron microscopy (**Figure 1.3**). Thus the concentration of the His-tagged assembly in lysates simply appears to be below the detection limit of our native PAGE analysis.

Based on the results from the mixed lysates experiments, T32-28 and T33-15 were additionally subjected to *in vitro* mixing experiments from purified components. Each of the C-terminally His-tagged components was purified by nickel affinity and gel filtration chromatography, and the purified components were mixed in a 1:1 molar ratio with each component present at a subunit concentration of 50 μM. Following incubation for two hours at room temperature, the mixtures were subjected to analytical size exclusion chromatography. The purifications and size exclusion chromatography were carried out as described above, with the exception that 5% (v/v) glycerol was added to all buffers. While T33-15 assembled efficiently from the independently purified components, T32-28 yielded only a small peak for the assembly product. The purified T32-28A component eluted significantly earlier than 3lzl-wt, indicating that lack of assembly in this case may be due to self-association of T32-28A in the absence of T32-28B.

Negative Stain Electron Microscopy

2–3 µl of purified T32-28, T33-09, T33-15, T33-21 and T33-28 samples at concentrations ranging from 0.01 mg mL^{-1} to 5 mg mL^{-1} were applied to glow discharged, carbon coated 200-mesh copper grids (Ted Pella, Inc.), washed with Milli-Q water and stained with 0.075% uranyl formate as described previously[58]. Grids were visualized for assembly validation and optimized for data collection. Screening and data collection was performed 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 either 49,000x (T33-21 and T33-28) or 60,000x (T32-28, T33-09 and T33-15) magnification at the specimen level. The contrast of all micrographs was enhanced in Fiji[59].

Coordinates for 3,910 (T32-28), 29,153 (T33-09), 18,197 (T33-15), 5,478 (T33-21) and 13,715 (T33-28) unique particles were obtained for averaging using either Ximdisp[60] or EMAN[61]. Extracted frames of these particles were used to obtain class averages by refinement in either SPIDER[62] or IMAGIC[63] using multiple rounds of MSA (multivariate statistical analysis) and MRA (multi-reference alignment). Low-resolution (17–30 Å) volumes from the design models were calculated using SPIDER and inspected in UCSF Chimera[64]. Back-projection images were computed in SPIDER on the low-resolution volumes and visualized using WEB[62].

In vitro-assembled, SEC-purified T33-15 (**Figure 1.2**) was also imaged as described above (**Figure 1.3**), along with SEC-purified T33-15A and T33-15B in isolation (**Figure 2.5**; see ***In vitro* mixing**). Images of T33-15 assembled *in vitro* without subsequent SEC purification were also obtained. For these images, SEC-purified T33-15A (0.64 mg mL^{-1}) and T33-15B

(0.56 mg mL^{-1}) were mixed in a 1:1 volumetric ratio, and grids prepared of the mixture after 2 hours at room temperature were imaged as described above (**Figure 2.5**).

Crystallization of T32-28

T32-28 was crystallized with hanging drop vapor diffusion at room temperature. Crystals were formed within four days by mixing 1 uL of 11.7 mg mL^{-1} protein and 1 uL of a 500 uL well solution containing only 1.675 M D,L-malic acid at pH 7.0. The crystals were cryo-protected in 2.0 M lithium sulfate and soaked for 20 seconds. The crystals diffracted to at least 4.5 Å and the asymmetric unit contained 12 molecules of T32-28A and 12 molecules of T32-28B in space group $P3_121$.

Crystallization of T33-15

As described above, crystals of T33-15 were grown within one week by mixing 1 uL of 7.6 mg mL^{-1} protein and 1 uL of a 500 uL well solution containing 100 mM sodium cacodylate at pH 6.5, 200 mM calcium acetate, and 28% (v/v) PEG 300. Crystals were cryo-protected by successive 30-second soaking in 10 uL solutions of mother liquor with glycerol added at final concentrations of 5%, 10%, 15%, and 20%. The crystals diffracted to at least 2.8 Å and the asymmetric unit contained one molecule each of T33-15A and T33-15B molecules in space group F432.

Crystallization of T33-21 in Space Group R32 and $F4_132$

T33-21 was crystallized similarly as described above. Crystals grew within three weeks following the mixing of 1 uL of 8.6 mg mL^{-1} protein and 0.5 uL of a 200 uL well solution containing 100 mM citric acid pH at 5.0 and 800 mM ammonium sulfate. Crystals were cryo-protected with 2.0 M lithium sulfate as described above. The crystals diffracted to at least 2.0 Å

and the asymmetric unit contained 4 molecules each of T33-21A and T33-21B in space group R32.

Alternatively, crystals also grew within one week by mixing 0.5 uL of 8.6 mg mL^{-1} protein and 1 uL of a 200 uL well solution containing 100 mM Bis-Tris at pH 5.5 and 2.12 M ammonium sulfate. Cryo-protection was performed with 2.0 M lithium sulfate as described above. These crystals diffracted to at least 2.6 Å and the asymmetric unit contained one molecule each of T33-21A and T33-21B in space group $F4_132$.

Crystallization of T33-28

T33-28 was crystallized as described above. Crystals grew within three days in hanging drops containing 0.5 uL of 15.8 mL^{-1} protein and 0.5 uL of a 200 uL well solution containing 100 mM sodium citrate tribasic dihydrate pH at 5.6, 200 mM ammonium acetate, and 24% (v/v) (+/-)-2-methyl-2,4-pentanediol. Cryo-protection involved passage of the crystal through drops of paratone-N oil until no more mother liquor appeared present around the crystal. The crystals diffracted to at least 3.5 Å and the asymmetric unit contained 48 molecules each of T33-28A and T33-28B in space group $P2_1$.

Crystallographic Data Collection and Structure Determination

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 for T32-28, T33-15, T33-21 (space group R32), T33-21 (space group $F4_132$), and T33-28 at detector distances of 650 mm, 450 mm, 300 mm, 300 mm, and 575 mm; with 0.5°, 0.5°, 0.2°, 0.5°, and 0.5° degree oscillations; and at wavelengths of 0.9793 Å, 0.9792 Å, 1.0393 Å, 0.9716 Å, and 0.9793 Å, respectively.

Data reduction, integration, and scaling were performed with XDS/XSCALE[65]. The program PHASER[66] was used to determine all crystal structures by molecular replacement (MR). For T33-15 and T33-21 structures, the MR search models were the original PDB scaffolds for each computationally designed component. The MR search models for the structures of T33-28 and T32-28 were models of the tetrahedral assemblies with and without side-chain atoms beyond β-carbons, respectively.

The X-ray diffraction data collected for T32-28 underwent additional processing in XSCALE to visualize anomalous scattering from copper ions anticipated in the T32-28A subunits. The data was scaled with unmerged Friedel mates and the resultant electron density map was used to calculate an anomalous difference Fourier map with the refined model in PHENIX[67]. The anomalous difference Fourier map revealed 12 roughly spherical difference peaks above 6σ in a tetrahedral arrangement that, when overlaid with the final, refined model of T32-28 and the structure of the wild-type scaffold protein (PDB ID 3LZL), clearly correspond to the copper binding sites in 3LZL/T32-28A. Although the anomalous difference peaks in the calculated map were not used to model copper ions in the final deposited structure, this analysis provided strong support for the accuracy of the moderate resolution T32-28 model. All deposited structure factors used for refinement were scaled with merged Friedel mates.

Crystallographic Refinement

All refinement steps were run using the phenix.refine module of PHENIX. Molecular replacement solutions were first refined with rigid body refinement, and then underwent individual coordinate refinement in addition to other strategies. Refinement strategies were tested comparing grouped and individual atomic displacement parameter (ADP) refinement, translation libration screw-motion (TLS) group definitions, and simulated annealing[68]. Each

refinement protocol was iteratively run while the quality of the model between runs was assessed in COOT using the $2mF_o$-DF_c with unfilled F_{obs} map and the mF_o-DF_c difference map[69]. Subsequent cycles of alternating refinement and model adjustment in COOT were performed to obtain the final refined models.

T32-28, T33-15, T33-21 (space group $F4_132$), and T33-28 were refined with individual isotropic ADP parameterization with 1 TLS group per polypeptide chain. T32-28 was refined as a model comprised of glycine, alanine, proline, and all other side chains truncated to the β-carbon due to poor electron density visibility in regions occupied by side chains. T33-15 was refined with reference model restraints assigned to T33-15B from chain A of PDB entry 1UFY. T33-21 (space group R32) was refined with individual isotropic ADP parameterization and 3-8 TLS group definitions per chain determined near residual minimization from the TLSMD server[70].

Model quality was assessed during and after refinement using geometric validation and MolProbity[71] tools as a part of the PHENIX suite. Structures of T33-15, T33-21, and T33-28 contain 97-100% of the residues within the most favored regions of the Ramachandran plot. Residues in the disallowed regions of the Ramachandran plot are found in T32-28 at positions where the phi and psi angles of the scaffold protein are also disallowed. T32-28, T33-15, and both T33-21 structures have ERRAT scores of 97.0%, 96.6%, 99.4%, and 98.2%, respectively. ERRAT[72] scores indicate the percentage of residues that fall below the 95% confidence limit for erroneous modeling. The large asymmetric unit of the T33-28 structure was inspected with VERIFY3D[73] due to incompatibility with ERRAT, and resulted in a passing score of greater than 80% of residues scored greater than or equal to 0.2 in the 3D/1D profile. Data collection and refinement statistics are reported in **Table 2.2** for T32-28, T33-15, and T33-28; and **Table 2.3**

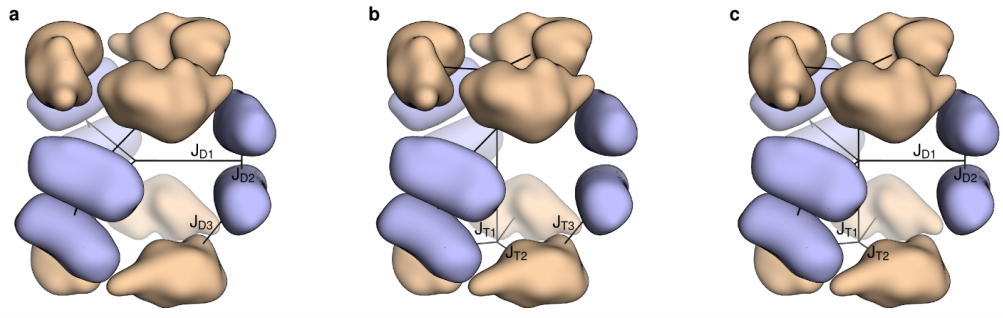
for the two structures of T33-21. The coordinates of the final models and the merged structure factors have been deposited in the Protein Data Bank with PDB codes 4NWN, 4NWO, 4NWR, 4NWP, and 4NWQ. All images of protein structures were made using PyMOL.

Quantitative Comparison of Crystal Structures and Design Models

Root mean square deviations (r.m.s.d.) over all backbone atoms (N, Cα, C, O) were calculated between each design model and corresponding crystal structure using the `pair_fit` command in PyMOL (**Table 2.4**). Additionally, for the three crystal structures with resolutions better than or equal to 2.8 Å (4NWO, 4NWP, and 4NWQ), differences in side chain dihedral angles (Δchi) of the design models (T33-15 and T33-21) and crystal structures were calculated for each interface position (**Tables 2.5 and 2.6**, see **Two-component protein-protein interface design** above for an explanation of how interface positions were determined). Positions at which the absolute value of Δchi was less than 25 degrees for all calculatable chi angles were considered to have adopted the predicted conformation; those positions without modeled side chains atoms beyond the beta carbon were not considered in this analysis. For positions with alternative side chain conformations, only conformation A was considered for these calculations.

Figures

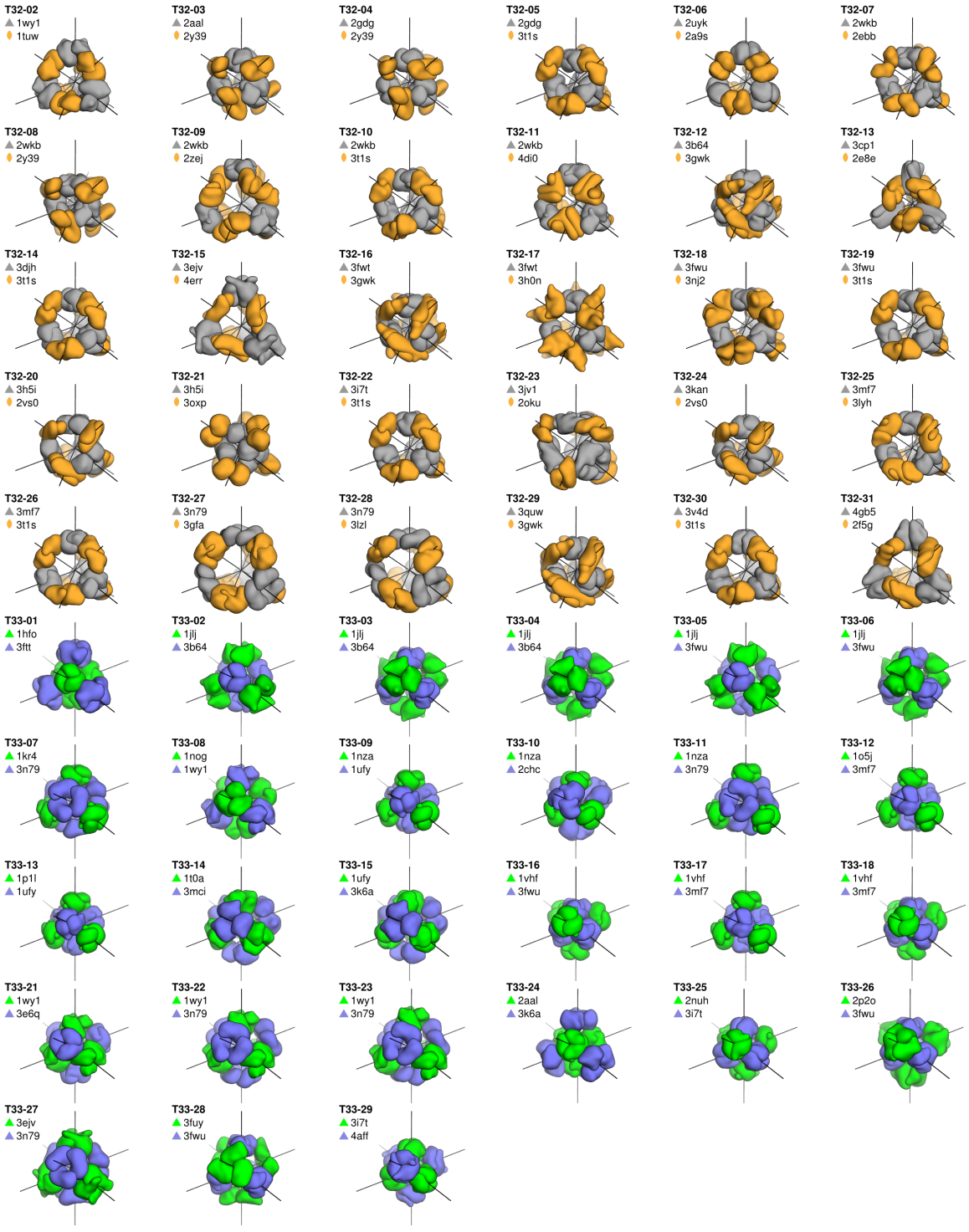
Figure 2.1



Comparison of one-component and multi-component symmetric fold trees. Within the Rosetta macromolecular modelling suite, the connections between residues in a protein structure are represented as a directed, acyclic, graph referred to as a "fold tree"[5,74]. When modelling multiple subunits in symmetric systems, the rigid body orientations of the subunits can be controlled by modifying the appropriate connections in the fold tree. In this work, we have extended Rosetta to allow multiple, independently managed connections from the fold tree to the subunits in the asymmetric unit (ASU) of the modelled structure. To demonstrate the new behaviour enabled by this change, three different symmetric fold tree representations of a D32 architecture are shown. In this architecture, which is used because of its relative simplicity, two trimeric building blocks (wheat) are aligned along the three-fold rotational axes of D3 point-group symmetry and three dimeric building blocks (light blue) are aligned along the two-fold rotational axes. **a**, The dimer-centric one-component symmetry case. Rigid-body degree of freedom (RB DOF, black lines) J_{D3} connecting the dimer subunit to the trimer subunit in the ASU is downstream of RB DOFs J_{D1} and J_{D2} controlling the dimer subunit; in this case the positions of the trimeric subunits depend on the positions of the dimeric subunits. **b**, The trimer-

centric one-component symmetry case. RB DOF J_{T3} connecting the trimer subunit to the dimer subunit in the ASU is downstream of RB DOFs J_{T1} and J_{T2} controlling the trimer subunit; in this case the positions of the dimeric subunits depend on the positions of the trimeric subunits. **c**, The multi-component symmetry case. With multi-component symmetric modelling, the RB DOFs controlling the trimer subunit (J_{T1} and J_{T2}) and the dimer subunit (J_{D1} and J_{D2}) in the ASU are independent. In this case the positions of the dimeric subunits do not depend on the positions of the trimeric subunits and vice versa, allowing the internal DOFs for each building block (J_{T2} and J_{D2}) to be maintained while moving the building blocks independently (J_{T1} and J_{D1}). See the Supplementary Methods for additional discussion.

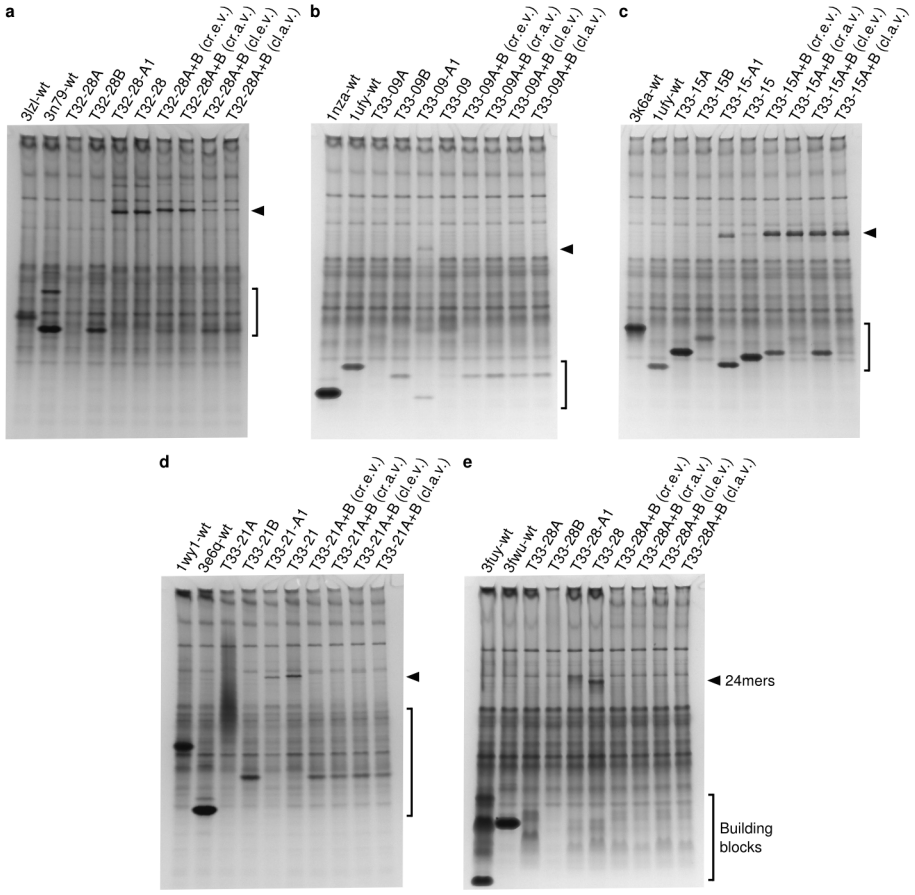
Figure 2.2



Models of the 57 designs selected for experimental characterization. Smoothed surface representations of each of the 30 T32 and 27 T33 designs are shown. The trimeric component of

each T32 design is shown in grey and the dimeric component in orange. The two different trimeric components of each T33 design are shown in blue and green. The tetrahedral two-fold and three-fold symmetry axes (black lines) are shown passing through the centre of each component. Each design is named according to its symmetric architecture (T32 or T33) followed by a unique identification number. The pairs of scaffold proteins from which the designs are derived are also indicated.

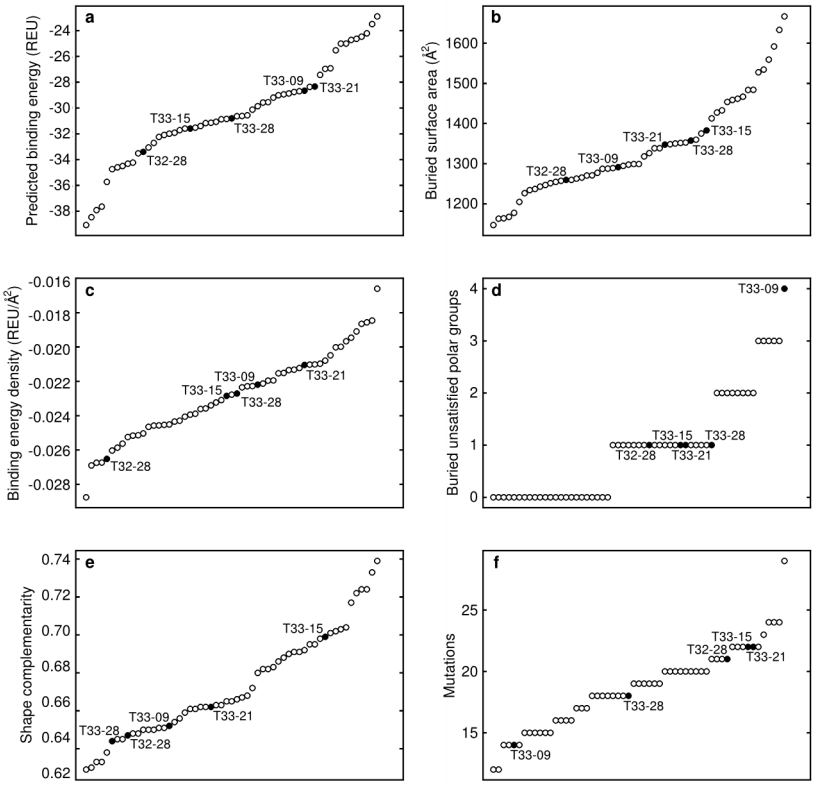
Figure 2.3



Native PAGE analysis of cleared cell lysates. Each gel contains cleared lysates pertaining to **a**, T32-28, **b**, T33-09, **c**, T33-15, **d**, T33-21, or **e**, T33-28. Lane 1 is from cells expressing the wild-

type scaffold for component A and lane 2 the wild-type scaffold for component B. Lanes 3 and 4 are from cells expressing the individual design components and lanes 5 and 6 the co-expressed components. Lanes 7 and 8 are from samples mixed as crude equal volume or crude adjusted volume (cr.e.v. or cr.a.v.) lysates, while lanes 9 and 10 are from samples mixed as cleared lysates (cl.e.v. or cl.a.v.). Lane 5 is from cells expressing the C-terminally A1-tagged constructs; all other lanes are from cells expressing the C-terminally His-tagged constructs. An arrow is positioned next to each gel indicating the migration of 24-subunit assemblies and the gel regions containing unassembled building blocks are bracketed. Each gel was stained with GelCode Blue. Portions of the gels in **a** and **c** are also shown in Fig. 2b.

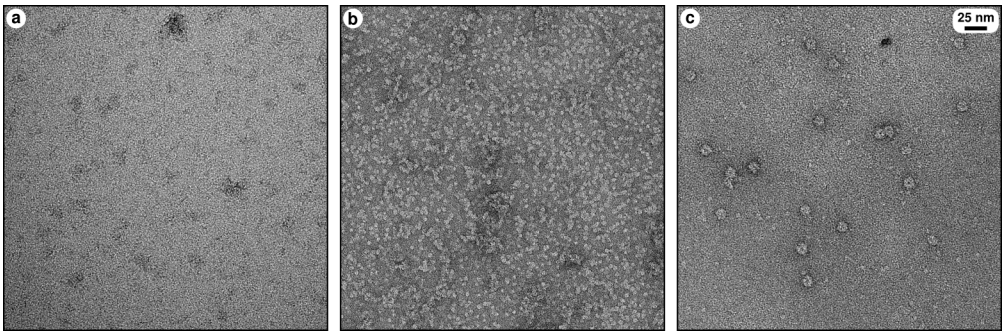
Figure 2.4



Structural metrics for the computational design models. Selected metrics related to the designed interfaces are plotted for the 57 designs that were experimentally characterized,

including **a**, the predicted binding energy measured in Rosetta energy units (REU), **b**, the surface area buried by each instance of the designed interface, **c**, the binding energy density (calculated as the predicted binding energy divided by the buried surface area), **d**, the number of buried unsatisfied polar groups at the designed interface, **e**, the shape complementarity of the designed interface, and **f**, the total number of mutations in each designed pair of proteins. Each circle represents a single design; the five successful materials are plotted as filled circles and labelled. In each plot, the designs are arranged on the *x* axis in order of increasing value of the metric analysed.

Figure 2.5



Electron micrographs of T33-15A and T33-15B in isolation, and *in vitro*-assembled T33-15 (unpurified). Negative stain electron micrographs of independently purified T33-15 components (left and middle) and unpurified, *in vitro*-assembled T33-15 (right) are shown to scale (scale bar at right, 25 nm).

Tables

Table 2.1 | Amino acid sequences of relevant designs, wild-type scaffolds, and tags

Name	Sequence	Comments
3LZL-wt	MGEVPIGDPKELNGMEIAAVYLQPIEMEPRGIDLAASLADIHLEA DIHALKNNPNGFPEGFWMPYLTIAYELKNTDTGAIKRGTLMPMV	Dimeric scaffold for T32-28A

	ADDGPHYGANIAMEKDKKGGFGVGNYELTFYISNPEKQGFGRH VDEETGVGKWFEPFKVDYKFKYTGTPK	
3N79-wt	MSQAIGILELTSIAKGMELGDAMLKSANVDLLVSKTICPGKFLLM LGGDIGAIQQAIETGTSQAGEMLVDSLVLANIHPSVLPAISGLNSV DKRQAVGIVETWSVAACISAADRAVKGSNVTLVRVHMAFGIGG KCYMVVAGDVSDVNNAVTVASESAGEKGLLVYRSVIPRPHEAM WRQMVEG	Trimeric scaffold for T32-28B
1NZA-wt	MEEVVLITVPSEEVARTIAKALVEERLAACVNIVPGLTSIYRWQG EVVEDQELLLLVKTTTHAFPKLKERVKALHPYTVPEIVALPIAEG NREYLDWLRENTG	Trimeric scaffold for T33-09A
1UFY-wt	MVRGIRGAITVEEDTPEAIHQATRELLLKMLEANGIQSYEELAAV IFTVTEDLTSAFPAEAARQIGMHRVPLLSAREVPVPGSLPRVIRVL ALWNTDTPQDRVRHVYLREAVRLRPDLESAQ	Trimeric scaffold for T33-09B and T33-15B
3K6A-wt	MSKAKIGIVTVSDRASAGIYEDISGKAIIDTLNDYLTSEWEPIYQVI PDEQDVIETTLIKMADEQDCCLIVTTGGTGPAKRDVTPEATEAVC DRMMPGFGELMRAESLKFVPTAILSRQTAGLRGDSLIVNLPGKPK SIRECLDAVFPAIPYCIDLMEGPYLECNEAVIKPFRPKAK	Trimeric scaffold for T33-15A
1WY1-wt	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAK HYVDEEMKGILEEIQNDIYKIMGEIGSKGKIEGISEERIKWLEGLIS RYEEMVNLKSFVLPGGTLESAKLDVCRTIARRAERKVATVLREF GIGKEALVYLNRLSDLLFLLARVIEIEKNKLKEVRS	Trimeric scaffold for T33-21A
3E6Q-wt	MPHLVIEATANLRLETSPGELLEQANAALFASGQFGEADIKSRFV TLEAYRQGTAAVERAYLHACLSILDGRDAATRQALGESLCEVLA GAVAGGGEEGVQVSVEVREMERASYAKRVVARQR	Trimeric scaffold for T33-21B
3FUY-wt	MESVNTSFLSPSLVTIRDFDNGQFAVLRIGRTGFPADKGDIDLCLD KMKGVRDAQQSIGDDTEFGFKGPHIRIRCVDIDDKHTYNAMVYV DLIVGTGASEVERETAEELAKEKLRAALQVDIADEHSCVTQFEM KLREELLSSDSFHPDKDEYYKDFL	Trimeric scaffold for T33-28A
3FWU-wt	MPVIQTFVSTPLDHHKRENLAQVYRAVTRDVLGKPEDLVMMTF HDSTPMHFFGSTDPVACVRVEALGGYGPSEPEKVTSIVTAAITKE CGIVADRIFVLYFSPLHCGWNGTNF	Trimeric scaffold for T33-28B
T32-28A	MGEVPIGDPKELNGMEIAAVYLQPIEMEPRGIDLAASLADIHLEA DIHALKNNPNGFPEGFWMPYLTIAYALANADTGAIKTGTLMPMV ADDGPHYGANIAMEKDKKGGFGVGTYALTFLISNPEKQGFGRH VDEETGVGKWFEPFVVTYFFKYTGTPK	
T32-28B	MSQAIGILELTSIAKGMELGDAMLKSANVDLLVSKTISPGKFLLM LGGDIGAIQQAIETGTSQAGEMLVDSLVLANIHPSVLPAISGLNSV DKRQAVGIVETWSVAACISAADLAVKGSNVTLVRVHMAFGIGG KCYMVVAGDVLDVAAAVATASLAAGAKGLLVYASIIPRPHEAM WRQMVEG	
T33-09A	MEEVVLITVPSALVAVKIAHALVEERLAACVNIVPGLTSIYRWQG SVVSDHELLLLVKTTTHAFPKLKERVKALHPYTVPEIVALPIAEG NREYLDWLRENTG	
T33-09B	MVRGIRGAITVEEDTPAAILAATIELLLKMLEANGIQSYEELAAVI FTVTEDLTSAFPAEAARLIGMHRVPLLSAREVPVPGSLPRVIRVLA LWNTDTPQDRVRHVYLNEAVRLRPDLESAQ	
T33-15A	MSKAKIGIVTVSDRASAGITADISGKAIILALNLYLTSEWEPIYQVI PDEQDVIETTLIKMADEQDCCLIVTTGGTGPAKRDVTPEATEAVC DRMMPGFGELMRAESLKEVPTAILSRQTAGLRGDSLIVNLPGDP ASISDCLLAVFPAIPYCIDLMEGPYLECNEAMIKPFRPKAK	

T33-15B	MVRGIRGAITVNSDTPTSIIIATILLLEKMLEANGIQSYEELAAVIFT VTEDLTSAFPAEAARQIGMHRVPLLSAREVPVPGSLPRVIRVLAL WNTDTPQDRVRHVYLSEAVRLRPDLESAQ	
T33-21A	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAK HYVDEEMKGILEEIQNDIYKIMGEIGSKGKIEGISEERIAWLLKLIL RYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKVLTVTREF GIGAEAAAYLLALSDLLFLLARVIEIEKNKLKEVRS	
T33-21B	MPHLVIEATANLRLETSPGELLEQANKALFASGQFGEADIKSRFV TLEAYRQGTAAVERAYLHACLSILDGRDIATRTLLGASLCAVLAE AVAGGGEEGVQVSVEVREMERLSYAKRVVARQR	
T33-28A	MESVNTSFLSPSLVTIRDFDNGQFAVLRIGRTGFPADKGDIDLCLD KMIGVRAAQIFLGDDTEDGFKGPHIRIRCVDIDDKHTYNAMVYV DLIVGTGASEVERETAEEEAKLALRVALQVDIADEHSCVTQFEM KLREELLSSDSFHPDKDEYYKDFL	
T33-28B	MPVIQTFVSTPLDHHKRLLLAIIYRIVTRVVLGKPEDLVMMTFHD STPMHFFGSTDPVACVRVEALGGYGPSEPEKVTSIVTAAITAVCGI VADRIFVLYFSPLHCGWNGTNF	
A1	LEGGDSLDMLEWSL	C-terminal tag used for fluorescent labeling and lysate screening
6xHis	LEHHHHHH	C-terminal tag used for nickel affinity chromatography

Table 2.2 | Crystallographic statistics for T32-28, T33-15, and T33-28 data collection and refinement. Statistics in parentheses refer to the highest resolution shell.

	T32-28 (PDB ID 4NWN)	**T33-15 (PDB ID 4NWO)**	**T33-28 (PDB ID 4NWR)**
Data collection			
Space group	$P3_121$	F432	$P2_1$
Cell dimensions			
a, b, c (Å)	246.01, 246.01, 290.94	213.52, 213.52, 213.52	124.91, 189.25, 376.83
α, β, γ (°)	90, 90, 120	90, 90, 90	90, 90.02, 90
Resolution (Å)	4.5	2.8	3.5
R_{merge} (%)	13.8 (94.6)	12.3 (177)	14.4 (60.1)
$CC_{1/2}$ (%)	99.7 (58.6)	99.9 (71.8)	99.4 (68.5)
CC* (%)	99.9 (85.9)	100 (91.4)	99.8 (90.2)
Mean I/σ(I)	13.20 (2.17)	19.80 (2.16)	8.95 (2.39)
Completeness (%)	98.31 (97.93)	99.91 (100)	98.80 (99.57)
Multiplicity	7.3 (7.5)	13.6 (14.3)	3.7 (3.9)
Wilson B-factor	184	79.3	90.5

Refinement			
Resolution range (Å)	93.93 - 4.5 (4.66 - 4.5)	75.49 - 2.8 (2.901 - 2.8)	94.21 - 3.5 (3.625 - 3.5)
No. reflections	59814 (5903)	10783 (1045)	217956 (21869)
R_{work}/R_{free} (%)*	29.7/34.3	20.2/25.2	26.1/29.9
No. atoms	20307	2011	88861
Protein	20307	2008	88861
Ligand/ion	0	1	0
Water	0	2	0
Average B-factors	216	72.6	91.7
Protein	216	72.6	91.7
Ligand/ion	N/A	111.5	N/A
Water	N/A	56.6	N/A
Protein residues	4075	285	12686
R.m.s. deviations			
Bond lengths (Å)	0.003	0.003	0.002
Bond angles (°)	0.55	0.77	0.49
Ramachandran favored (%)	97	98	97
Ramachandran outliers (%)	0.15	0	0
MolProbity clashscore†	0.89	2.26	4.61

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

† The number of unfavorable all-atom steric overlaps ≥ 0.4Å per 1000 atoms.

Table 2.3 | Crystallographic statistics for T33-21 data collection and refinement. Statistics in parentheses refer to the highest resolution shell.

	T33-21 R32 (PDB ID 4NWP)	**T33-21 $F4_132$ (PDB ID 4NWQ)**
Data collection		
Space group	R32	$F4_132$
Cell dimensions		
a, b, c (Å)	113.35, 113.35, 634.88	272.18, 272.18, 272.18
α, β, γ (°)	90, 90, 120	90, 90, 90
Resolution (Å)	2.1	2.8
R_{merge} (%)	11.2 (118)	12.2 (120)
$CC_{1/2}$ (%)	99.8 (74.9)	99.9 (87.8)
CC* (%)	100 (92.5)	100 (96.7)
Mean I/σ(I)	14.46 (2.48)	20.89 (3.14)
Completeness (%)	99.94 (99.97)	99.99 (99.95)

Multiplicity	9.7 (9.8)	19.8 (20.3)
Wilson B-factor	37.7	69.0
Refinement		
Resolution range (Å)	93.78 - 2.1 (2.175 - 2.1)	96.23 - 2.8 (2.9 - 2.8)
No. reflections	92425 (9127)	21830 (2129)
R_{work}/R_{free} (%)*	18.8/21.8	18.2/19.6
No. atoms	8248	2112
Protein	7882	2041
Ligand/ion	141	55
Water	225	16
Average B-factors	42.5	73.1
Protein	42.2	72.5
Ligand/ion	64.5	98.8
Water	40.6	64
Protein residues	1046	269
R.m.s. deviations		
Bond lengths (Å)	0.004	0.001
Bond angles (°)	0.67	0.41
Ramachandran favored (%)	100	99
Ramachandran outliers (%)	0	0
MolProbity clashscore†	1.87	1.2

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

† The number of unfavorable all-atom steric overlaps ≥ 0.4Å per 1000 atoms.

Table 2.4 | Root mean square deviations (r.m.s.d.) between crystal structures and design models

Design model	Crystal structure	Global r.m.s.d. (Å)*	Two-chain r.m.s.d. (Å)†	Contents of asymmetric unit	Structure used for superposition‡
T32-28	4NWN	2.586	1.246	One cage (24 subunits)	Asymmetric unit
T33-15	4NWO	1.433	0.876	One chain of each component (2 subunits)	One cage generated from crystallographic 2- and 3-folds
T33-21	4NWP	1.962	0.924	4 chains of each component (8 subunits)	One cage generated from one crystallographic 3-fold
T33-21	4NWQ	1.482	0.765	One chain of each component (2 subunits)	One cage generated from crystallographic 2- and 3-folds
T33-28	4NWR	0.965	0.503	Four complete cages (96 subunits)	One complete cage from the asymmetric unit
T33-28	4NWR	0.965	0.548	Four complete cages (96 subunits)	One complete cage from the asymmetric unit
T33-28	4NWR	1.195	0.567	Four complete cages (96 subunits)	One complete cage from the asymmetric unit
T33-28	4NWR	1.212	0.477	Four complete cages (96 subunits)	One complete cage from the asymmetric unit

*Global r.m.s.d. was calculated over all 24 subunits of each design model and corresponding subunits in each crystal structure.

†Two-chain r.m.s.d. was calculated over chains A and B of each design model and corresponding subunits in each crystal structure.

‡24 subunits composing one complete cage were derived from each crystal structure as indicated and the chains renamed to match the corresponding names in the design models. In the case of T33-28, four different sets of r.m.s.d. calculations were carried out, one for each of the four cages contained in the asymmetric unit of 4NWR.

Table 2.5 | Side chain chi value comparison of T33-15 crystal structure (PDB ID 4NWO) and design model.

Residue*	Δchi1	Δchi2	Δchi3	Δchi4	Δchi5
I9	1.6	4.2			
T10	4.8				
V11	1.1				
N12	-0.8	-3.4			

S13	-8.3				
T15	-0.5				
P16	2.0	-0.7	-0.9	2.4	
T17	135.3				
S18	116.3				
I20	1.3	-13.5			
I21	4.8	-1.4			
I24	-7.1	-4.4			
L25	-1.1	-6.8			
E28	-103.8	-14.8	79.8		
K29	-16.6	-	-	-	
E32	-	-	-		
Q64	90.3	145.9	-25.5		
I65	-1.8	1.7			
R86	-5.2	-8.2	-1.3	-3.6	7.5
L108	1.6	-0.6			
S109	3.9				
E110	-	-	-		
T140	-				
I143	2.2	-9.0			
K146	100.0	3.9	-	-	
I149	-98.7	-11.1			
L150	0.8	-9.1			
N153	13.3	4.6			
L154	3.5	3.4			
E226	-	-	-		
S227	-1.3				
K229	-	-	-	-	
E230	-	-	-		
V231	-				
D255	-9.8	4.6			
P256	-4.9	3.3	-0.2	-3.0	
S258	-10.3				
S260	0.5				
D261	2.4	18.8			
L264	103.8	108.6			
N285	-6.7	-7.6			
M288	-8.4	2.0	-10.2		
I289	-4.5	-3.3			
Pass[†]	**29**	**23**	**4**	**3**	**1**
Fail[‡]	**7**	**2**	**2**	**0**	**0**

***** Residue numbers refer to positions in the T33-15 design model.

[†] Number of residues where $|\Delta chi| < 25$.

‡ Number of residues where |Δchi| ≥ 25. Residues with missing atoms in the crystal structure, for which a Δchi value could not be determined, are indicated with a dash. All Δchi valuses are reported in degrees.

Table 2.6 | Side chain chi value comparison of T33-21 crystal structures (PDB IDs 4NWP and 4NWQ) and design model.

	T33-21 vs. 4NWP					T33-21 vs. 4NWQ				
Residue*	**Δchi1**	**Δchi2**	**Δchi3**	**Δchi4**	**Δchi5**	**Δchi1**	**Δchi2**	**Δchi3**	**Δchi4**	**Δchi5**
K52	-	-	-	-		108.5	-0.1	0.7	-1.5	
E54	1.4	-	-			4.4	-7.1	2.8		
S57	9.5					6.0				
E58	13.8	-	-			82.2	90.5	-38.0		
E59	-	-	-			-97.1	95.5	14.7		
R60	-0.2	4.1	-4.2	18.9	-0.2	2.1	5.4	-8.1	-3.2	0.0
I61	0.7	-8.0				-6.4	-13.6			
W63	3.0	-3.3				2.3	1.6			
L65	-5.1	12.7				-11.2	-3.6			
K66	110.0	-1.8	121.6	120.9		106.5	-25.5	-119.7	1.4	
I68	2.4	-10.0				-1.1	-12.1			
L69	0.1	3.0				-3.0	-1.3			
M72	33.0	-0.3	2.5			1.1	-3.8	11.4		
L102	-4.7	9.6				-4.7	8.0			
R103	-7.4	-5.5	-5.4	-3.4	0.3	-7.8	-5.2	-5.3	-1.5	0.7
L106	-2.0	-10.6				-4.8	-13.7			
T109	-7.4					-9.8				
R110	-102.3	14.3	-4.1	39.6	0.1	-106.9	19.8	-1.9	17.6	0.4
I114	-6.7	-20.3				-4.2	4.6			
E117	0.3	168.2	-55.5			-16.4	160.2	5.0		
L123	-2.2	1.4				-2.2	1.8			
D127	-9.4	2.5				-9.8	14.0			
D145	-	-				-	-			
K175	3.5	-	-	-		-1.8	-5.0	2.6	-0.6	
D221	13.3	-21.7				19.4	-17.7			
I222	109.5	110.2				100.6	-9.5			
T224	-4.6					-4.7				
R225	-15.7	-6.2	0.4	-9.4	1.0	-12.3	-1.4	-6.0	2.4	0.2
T226	-5.1					-1.8				
L227	4.5	0.5				5.3	0.3			
S231	111.7					-4.9				

C233	-7.9					-2.1				
V235	-1.6					-6.8				
E238	-	-	-			-11.4	0.6	16.9		
E258	14.0	-7.2	62.5			-1.1	-13.0	84.3		
R259	-4.5	3.5	-52.7	73.8	-1.2	2.6	-5.9	-120.7	-85.6	0.0
L260	20.9	-12.0				6.8	-3.3			
S261	0.9					4.1				
Y262	-11.7	2.4				-5.4	-6.2			
K264	12.2	-1.4	-9.0	7.4		6.6	-1.7	-129.1	7.8	
R265	90.0	-33.2	139.5	-16.2	0.1	3.3	2.0	11.5	-6.7	0.1
Pass[†]	**31**	**23**	**6**	**5**	**6**	**34**	**28**	**12**	**9**	**6**
Fail[‡]	**6**	**3**	**5**	**3**	**0**	**6**	**4**	**5**	**1**	**0**

***** Residue numbers refer to positions in the T33-21 design model.

[†] Number of residues where $|\Delta chi| < 25$.

[‡] Number of residues where $|\Delta chi| \geq 25$. Residues with missing atoms in the crystal structure, for which a Δchi value could not be determined, are indicated with a dash. All Δchi values are reported in degrees.

Table 2.7 | List of homodimeric PDB entries used as scaffolds (PDB ID and biological unit number, separated by an underscore)

1a1x_1	1a3c_1	1ag9_1	1alu_1	1alv_1	1az5_1	1b0x_1	1b8z_1	1bgf_1	1bm9_1	1btk_1
1buo_1	1byf_1	1byr_1	1c02_1	1cdc_1	1ci4_1	1ciz_2	1coz_1	1cxq_1	1d7j_3	1d9c_1
1dnl_1	1dz3_1	1e7l_1	1ecs_1	1eeq_1	1ek3_1	1ep0_1	1esr_1	1etx_1	1evx_1	1ex2_1
1ext_1	1eyv_1	1f1e_1	1f1g_1	1f1m_1	1f9f_1	1f9z_1	1fit_1	1fmb_1	1fux_1	1fzv_1
1g2i_1	1g2q_1	1g8q_1	1gd7_1	1gvj_1	1gvp_1	1gy6_1	1gy7_1	1gyx_1	1h8x_1	1he7_1
1hgx_1	1ht9_1	1hur_1	1i0r_1	1i12_1	1i3c_1	1i4y_1	1i9d_2	1ic2_2	1ihr_1	1ilk_1
1iq6_1	1is6_1	1ixl_2	1izm_1	1j24_1	1j27_2	1j2r_2	1j3m_1	1j3q_1	1j55_1	1j7g_1
1j8b_1	1j98_1	1jc4_1	1jhc_1	1jhg_1	1jk3_2	1jr8_1	1jrl_2	1jya_1	1k04_1	1k2e_1
1k3s_1	1k66_1	1k8u_1	1k9u_1	1kcq_2	1kl9_2	1kll_1	1kso_1	1l1q_1	1l3p_1	1l8s_1
1lgp_1	1lj9_1	1lq9_1	1ly1_1	1m0d_1	1m1f_1	1m2d_1	1m4i_1	1mi8_1	1mjh_1	1mk4_1
1mka_1	1mkk_1	1mp9_1	1msc_1	1mxi_1	1my6_1	1my7_1	1n99_1	1n9l_2	1ng2_2	1njh_1
1nki_1	1np6_1	1nqd_1	1nrz_5	1ns5_1	1nu3_1	1nxm_1	1nzn_2	1o22_1	1o3u_1	1o4t_1

1o4w_1	1o50_1	1o6a_1	1o6d_2	1oh0_1	1ohp_1	1oiv_1	1on2_1	1oqc_1	1oru_1	1ovs_2
1p6o_1	1pbj_2	1pdo_1	1psr_1	1puc_1	1pvm_1	1py9_1	1pzw_1	1q08_1	1q7s_3	1q8b_2
1q98_1	1q9u_1	1qip_1	1qou_1	1qto_1	1qwi_1	1r1t_1	1r1u_1	1r29_1	1r5q_2	1r7j_1
1r7l_1	1r9c_1	1rdo_1	1rfy_1	1rlk_2	1rxq_2	1s4k_1	1s67_1	1s7i_1	1s7z_2	1s99_1
1sd4_1	1sei_1	1sgm_1	1sh8_1	1sjw_2	1sjy_1	1sk4_2	1sl8_2	1snd_1	1t82_1	1t92_1
1tc5_1	1tfe_1	1tgj_1	1to4_1	1tu1_1	1tuh_1	1tuv_2	1tuw_1	1tvd_1	1twu_1	1u2w_1
1u3y_1	1u5f_2	1u69_2	1u7i_1	1uat_2	1udv_1	1ues_1	1ukk_1	1usc_1	1usm_1	1usp_1
1ut7_1	1uww_1	1uz3_1	1v05_1	1v2z_1	1v70_1	1v8y_1	1v96_1	1v9y_1	1vc1_1	1vh5_1
1via_1	1vj2_1	1vje_1	1vjl_1	1vkc_1	1vki_1	1vl7_1	1vq3_2	1vr7_1	1vzg_1	1w53_1
1wc9_1	1wkq_1	1wlt_1	1wn2_1	1woc_1	1wpn_1	1wu9_1	1wwc_2	1wwi_2	1wz3_1	1x0j_1
1x2i_1	1x6i_1	1x82_1	1x8d_1	1xe1_2	1xfs_1	1xhn_1	1xqa_1	1xrk_1	1xs0_1	1xso_1
1xsq_1	1xty_1	1xvq_2	1y0b_1	1y0h_1	1y0u_1	1y5h_1	1y7r_1	1y9q_1	1y9w_1	1yb3_2
1ybx_1	1ybz_2	1yfu_1	1ygt_1	1yhf_2	1yib_1	1ylk_1	1ylm_1	1yo3_1	1yoa_1	1yr0_1
1ysp_2	1z0p_1	1z2w_3	1z4e_1	1z9n_1	1z9p_1	1zb9_1	1zdn_3	1zhq_1	1zhv_2	1zj6_2
1zlj_1	1zn8_1	1zo2_1	1zop_1	1zps_1	1zpv_1	1zpw_2	1ztd_1	1zva_1	1zwy_1	1zxk_1
2a15_1	2a67_1	2a6c_1	2a72_1	2a8n_1	2a9s_1	2aan_2	2aao_3	2akp_3	2aps_1	2aqs_1
2asf_1	2auw_2	2b06_1	2b0a_1	2b0v_2	2b18_1	2b1y_1	2b3s_3	2b5a_1	2b5g_1	2b6h_2
2b8m_1	2b9a_1	2bbe_2	2bdr_1	2bnl_1	2bsj_1	2bz1_1	2c2i_1	2c9q_1	2car_1	2cvd_1
2cvi_1	2cwz_1	2cyy_1	2d37_1	2d4p_2	2d4u_1	2d5m_1	2d7v_1	2d8d_1	2dc3_1	2dc4_1
2dlb_1	2dm9_1	2dob_2	2dp9_1	2dpf_1	2dql_1	2duy_2	2dvk_1	2dxq_1	2e1f_2	2e1n_1
2e6u_2	2e8e_1	2eb1_1	2ebb_1	2ecu_1	2een_1	2ef8_1	2efv_1	2egd_1	2eh3_1	2ehp_1
2ei5_1	2eiq_3	2ejn_1	2eo4_1	2erb_3	2esu_2	2f22_1	2f4p_1	2f5g_1	2f62_1	2f99_2
2f9h_1	2fa1_1	2fa5_1	2fbh_1	2fbn_1	2fck_1	2fd5_1	2fe3_1	2fex_1	2fhq_1	2fip_1
2fiu_1	2fj9_2	2fjt_1	2fl4_1	2fpr_1	2fq4_1	2fr2_2	2fre_1	2ftr_1	2fu4_1	2fyq_1
2fyx_1	2g0c_2	2g0i_1	2g1u_1	2g3a_1	2g3r_2	2g7s_1	2g84_1	2gax_3	2gbt_1	2ge7_1
2gen_1	2gff_1	2glz_1	2goj_1	2gpc_1	2gu9_1	2gux_1	2gxg_1	2gyq_1	2gzv_2	2h1t_1
2h2b_1	2h8e_1	2h9u_1	2ha8_1	2hbo_1	2hcm_1	2hhg_2	2hhz_1	2hiq_1	2hkv_1	2hl0_1
2hlj_1	2hng_1	2hq9_1	2hql_1	2hs1_1	2hsb_1	2htd_1	2huh_1	2hur_2	2hyt_1	2hzc_2
2hzt_2	2i02_1	2i51_1	2i7a_1	2i7d_1	2i8b_1	2i8d_1	2i8t_1	2ia1_1	2ict_2	2idl_1
2iek_1	2if5_1	2ifx_1	2ig6_1	2igi_1	2ikk_1	2imj_1	2iml_1	2ims_1	2imz_1	2inb_1
2isy_1	2iu5_1	2ivy_1	2iwq_1	2ixk_1	2j6b_1	2j6y_1	2j7j_1	2j8m_1	2jar_1	2jba_1
2jdj_1	2je3_1	2jlj_1	2lig_1	2nlv_1	2nrk_1	2nsa_2	2nwv_1	2nx4_1	2nx8_1	2nyb_1
2nyc_1	2nyi_1	2nz7_1	2nzo_1	2o08_1	2o28_1	2o38_1	2o4t_1	2o6f_1	2o70_1	2o7m_1

2o95_1	2o99_1	2oa2_1	2oai_1	2ob5_1	2od4_1	2od6_1	2oda_1	2oee_1	2ogi_1	2oik_1
2okf_1	2oku_1	2olm_1	2omo_1	2onf_1	2oo2_1	2ooj_1	2ook_1	2opo_1	2oqk_2	2oqm_1
2oso_1	2ou3_1	2ou5_1	2ou6_1	2ouf_1	2ovs_1	2owp_1	2oyn_1	2oyz_1	2ozh_1	2ozj_1
2p08_1	2p09_1	2p12_1	2p25_1	2p3w_1	2p5q_1	2p7o_1	2p84_1	2p8g_1	2p8i_1	2p92_1
2pa7_1	2pey_1	2pfb_1	2pfi_1	2pfw_1	2pjs_1	2pk8_1	2pkh_1	2pmr_1	2pn0_1	2pn2_1
2pq3_2	2pqv_1	2prx_1	2pwo_1	2pyt_1	2q03_1	2q0y_1	2q20_1	2q24_1	2q2f_1	2q2h_1
2q2i_1	2q30_1	2q3p_1	2q3t_1	2q3x_1	2q4n_1	2q5c_3	2q79_1	2q82_1	2q8o_1	2q9k_1
2q9r_1	2qe9_1	2qhk_1	2qjw_1	2qkp_1	2ql8_1	2qml_2	2qmm_1	2qnd_3	2qnl_1	2qnt_1
2qqz_1	2qrr_1	2qsi_1	2qsw_1	2qtr_1	2qud_1	2qvm_2	2qx0_1	2r0x_1	2r1i_1	2r47_1
2r4i_1	2r6u_1	2r6v_1	2r78_1	2rbb_1	2rc3_1	2rcz_1	2rey_1	2rh0_1	2rhm_1	2ril_1
2riq_1	2rk3_1	2rk9_1	2rkf_1	2rkh_1	2uv4_1	2v57_1	2v90_1	2vez_1	2vkl_1	2voc_1
2vpk_1	2vs0_1	2vsv_1	2vvp_1	2vvw_1	2w1r_1	2w2a_1	2w31_1	2w4e_1	2w7w_1	2wb6_1
2wce_1	2wcr_1	2wcu_1	2wcw_1	2wfc_1	2wnx_1	2wp7_1	2wra_1	2wtg_1	2wzo_1	2x3g_1
2x5c_1	2x5h_1	2x5r_1	2x7z_1	2xbq_1	2xdp_1	2xf1_1	2xhf_1	2xr4_1	2xrh_1	2xxc_1
2y0o_1	2y39_1	2y6w_1	2y78_1	2yfd_1	2ykz_1	2yqy_3	2ysk_1	2yvo_1	2ywl_1	2yxh_1
2yz1_3	2yzk_1	2z10_1	2z6d_1	2z8u_1	2z98_1	2zcm_1	2zdo_1	2zdp_1	2zej_1	2zgl_1
2znd_1	2zpm_2	2zvy_1	2zw2_1	2zxy_1	3a2y_2	3a5p_1	3a6r_1	3a6s_3	3acd_1	3agx_1
3ah7_1	3aly_3	3b02_1	3b09_1	3b33_1	3b47_1	3b5g_1	3b5t_1	3b76_1	3b7c_1	3b7h_1
3b9c_1	3bb9_1	3bcw_1	3bde_1	3bln_1	3bm1_1	3bm7_1	3bmz_1	3bn7_1	3bpj_1	3bpv_1
3bqx_1	3bri_1	3bs3_1	3but_1	3by8_2	3byr_1	3bzh_1	3bzt_2	3c0f_1	3c1d_3	3c1q_1
3c3m_1	3c97_1	3can_1	3cb0_1	3cby_1	3ce1_1	3cex_1	3cjd_1	3cje_1	3cjn_1	3cm3_1
3cng_1	3cnk_1	3cnu_1	3cp3_1	3ct6_1	3cu3_1	3czt_1	3d00_1	3d0f_1	3d0j_1	3d0w_1
3d5p_1	3d7a_1	3db7_1	3dcm_1	3df8_1	3dib_2	3dlo_1	3dm8_1	3dmc_1	3dn7_1	3dnx_1
3do8_1	3dpj_1	3dr6_1	3dsb_1	3dz8_1	3e10_1	3e17_1	3e2c_1	3e39_1	3e4v_1	3e5h_1
3e8o_1	3ebt_1	3ec6_1	3ec9_1	3ecf_1	3f3x_1	3f43_1	3f7e_1	3f7l_1	3f8h_1	3f8x_1
3f9s_1	3fcd_1	3fcn_1	3fd7_3	3ff0_1	3ffy_1	3fg9_1	3fgv_1	3fgy_1	3fh1_1	3fjs_1
3fkc_2	3flj_1	3fm2_1	3fm5_1	3fmb_1	3fn7_1	3fov_1	3fqm_1	3frq_1	3fu1_1	3fv6_1
3fwz_1	3fx7_1	3fxh_1	3fyb_1	3fyn_1	3g0k_1	3g13_1	3g14_1	3g16_1	3g26_1	3g2b_1
3g46_1	3g7p_1	3g8g_1	3g8k_1	3g8z_1	3gby_1	3gdw_1	3gfa_1	3ggq_1	3ggu_1	3ghj_1
3gla_1	3glv_1	3gm5_1	3gpv_1	3grd_1	3guz_1	3gwk_1	3gwn_1	3gxh_3	3gya_2	3gyd_1
3gzr_1	3h05_1	3h0n_1	3h0x_2	3h1s_1	3h2d_1	3h36_1	3h3h_1	3h4o_1	3h4y_1	3h51_1
3h6q_2	3h8h_2	3h8u_1	3h95_1	3ha2_1	3ha9_2	3hcz_2	3hdc_1	3hf5_1	3hhv_1	3hiu_1
3hix_5	3hk4_2	3hm4_1	3hmf_2	3hmz_1	3hoi_1	3hqx_1	3hr7_1	3ht1_1	3huh_1	3hup_1

3hvv_1	3hx9_1	3hyq_2	3hzb_1	3hzp_1	3i24_1	3i3g_1	3ia1_1	3ia8_3	3ibm_1	3ifj_3
3ift_2	3igr_2	3iis_1	3ijm_1	3ilx_1	3in8_1	3inq_1	3ip0_2	3ir3_1	3itf_1	3ix3_1
3jrz_1	3jtf_1	3jtw_1	3jtz_1	3jum_1	3jx9_1	3k0z_1	3k1e_1	3k21_1	3k2v_1	3k3v_2
3k67_1	3k69_1	3k86_1	3kb5_2	3kbe_1	3kbq_1	3kby_1	3kg0_2	3kgz_1	3kk4_1	3kkg_1
3kl1_1	3kol_1	3kor_1	3kpc_1	3ksh_1	3ksv_1	3kuv_1	3kwk_1	3kyz_1	3l18_1	3l1e_1
3l1n_2	3l34_1	3l3u_1	3l46_1	3l7h_1	3l7x_1	3l8u_1	3l9y_1	3lag_1	3las_1	3lb5_1
3lby_1	3le4_1	3le5_1	3leq_2	3lf6_1	3lfh_1	3lfp_1	3lfr_1	3lhc_1	3lhr_1	3lin_1
3lio_1	3llv_2	3lmo_1	3lqn_1	3lqy_2	3lr0_1	3lte_1	3lw3_1	3lwc_1	3lx7_1	3lyd_1
3lyg_1	3lyh_1	3lyx_1	3lza_1	3lzl_1	3m1e_1	3m5b_1	3m6j_1	3m8e_1	3m9z_1	3mcw_1
3mdp_1	3mgd_1	3mgm_1	3mhx_1	3mmh_1	3mng_1	3msh_1	3mti_1	3mtq_1	3mws_1	3myf_1
3n1s_1	3n4j_1	3n4w_1	3n6y_3	3n8b_1	3nad_1	3nbc_1	3neu_1	3nfc_1	3nj2_1	3njc_1
3nl9_1	3nqn_1	3nr1_1	3nrh_3	3nrp_1	3ny5_1	3nym_1	3o0m_1	3o10_1	3o1c_1	3o2e_2
3o2r_1	3o79_3	3oa4_1	3obh_1	3oga_1	3ogh_1	3ohe_1	3oj7_1	3oji_1	3okx_1	3oms_1
3on4_1	3oni_2	3oop_1	3ose_2	3ov8_1	3oxp_1	3p0t_1	3p2t_2	3pc6_1	3pg6_1	3pmd_1
3pn3_3	3pp9_1	3pr6_2	3pu7_1	3q20_1	3q34_1	3q3y_3	3q62_1	3q63_1	3q64_1	3q6a_1
3q7r_1	3q8i_2	3q90_1	3qbm_1	3qdo_1	3qfl_1	3qh6_2	3qmq_1	3qoo_1	3qp4_1	3qp8_1
3qs2_1	3qu1_1	3qzx_1	3r0n_1	3r5g_1	3r68_2	3r6a_1	3r6f_1	3rcp_2	3rd1_1	3rem_1
3rfi_2	3rkc_1	3rmh_1	3rmu_1	3rob_1	3rqi_1	3rt2_2	3s2r_1	3s45_1	3s6f_1	3s8i_1
3s9f_1	3sb1_1	3sd2_1	3sk2_1	3sl2_2	3sl7_1	3slz_1	3smd_2	3smj_3	3son_1	3soy_1
3svi_2	3sxm_1	3sz7_2	3szj_1	3t1s_1	3t43_1	3t46_2	3t8r_1	3t90_1	3t9y_1	3td4_4
3teq_1	3tgn_1	3tgv_1	3tj8_1	3tk0_1	3tnj_1	3tol_1	3trc_1	3typ_1	3tys_1	3u04_1
3u15_1	3u1d_1	3u2a_1	3u5v_1	3u6g_1	3u80_1	3ub6_1	3ucb_1	3ucg_1	3ufe_1	3uh9_1
3uie_1	3ulb_1	3ups_1	3urr_1	3uv0_1	3ux2_1	3vjz_1	3vk6_1	3vp5_1	3vql_1	3vub_1
3zrd_1	3zve_1	3zw5_1	3zxc_1	3zxq_1	3zy7_1	4a1i_1	4a5k_1	4a5n_1	4ae4_1	4aeq_1
4ag7_1	4agh_1	4alg_1	4avp_1	4ax2_1	4b4p_1	4b6i_1	4di0_1	4duq_1	4e08_1	4e0h_1
4e2g_1	4e74_1	4e7p_1	4eae_1	4egu_1	4em8_1	4err_1	4es1_1	4eun_1	4ew5_1	4ew7_1
4exo_1	4exr_1	4ezg_1	4f82_1	4f8y_1	4fak_1	4fiv_1	4flb_1	4fld_1	4g5a_1	4g6x_1
4gdh_1	4ghj_1	4giw_1	4go7_1	4gs3_2	4gwb_1					

Table 2.8 | List of homotrimeric PDB entries used as scaffolds (PDB ID and biological unit number, separated by an underscore)

1buu_1	1dbf_1	1dg6_1	1di6_1	1f7l_1	1fth_1	1gr3_1	1gu9_1	1gx1_1	1h7z_1	1h9m_1

1hfo_1	1idp_1	1iv2_1	1jd1_1	1jlj_1	1jq0_1	1jw8_2	1knb_1	1kr4_1	1lr0_2	1n2m_1
1nog_1	1nza_1	1o5j_1	1o91_1	1ocy_1	1oni_1	1ox3_1	1p1l_2	1pwb_1	1q5h_1	1q5x_1
1qu1_1	1rlh_2	1s55_1	1seh_1	1sjn_1	1t0a_1	1tcz_1	1td4_1	1u5x_1	1u9d_2	1ufy_1
1uku_1	1usn_2	1uuy_1	1uxa_1	1v3w_1	1ve0_1	1vfj_1	1vhf_2	1vmf_1	1vmh_1	1vph_1
1woz_1	1wy1_1	1x25_1	1xhd_2	1yq5_1	2aal_1	2bcm_1	2brj_1	2bt9_1	2bzv_1	2chc_1
2cu5_1	2cvl_1	2dt4_1	2e7a_1	2ed6_1	2eg2_1	2f0c_1	2fb6_2	2fvh_1	2g2d_1	2gdg_1
2gr8_1	2gw8_1	2h6l_1	2hx0_1	2ibl_1	2ieq_1	2ig8_1	2is8_1	2j2j_1	2j9c_1	2jb7_1
2nuh_2	2oj6_1	2ol1_1	2otm_1	2p2o_1	2p6c_1	2p6h_1	2p6y_1	2p9o_1	2pii_1	2qg8_1
2qih_1	2r32_1	2r6q_1	2rfr_1	2rie_1	2tnf_1	2uyk_1	2vnl_1	2w5p_1	2wds_1	2wh7_1
2wkb_1	2wpq_1	2wq4_1	2x4j_1	2xcz_1	2xdh_1	2xdj_1	2xx6_1	2y75_2	2yzj_1	2zhz_1
3aqe_7	3b64_1	3b8l_1	3bsw_1	3bzq_1	3c6v_1	3cc0_1	3ci3_1	3cp1_1	3d01_1	3d9x_1
3da0_1	3djh_1	3e6q_1	3eby_1	3efg_1	3ehw_1	3ejc_1	3ejv_1	3emf_1	3f09_1	3f0d_1
3f4f_1	3fq3_3	3ftt_1	3fuy_1	3fwt_1	3fwu_1	3gqh_1	3h5i_1	3h6x_1	3htn_1	3hwu_1
3hza_1	3i3f_1	3i7t_1	3ixc_1	3jv1_1	3k6a_1	3kan_1	3kjj_1	3laa_1	3lqw_1	3m1x_1
3mc3_1	3mci_1	3mdx_1	3mf7_1	3mhy_1	3mko_1	3mqh_1	3mxu_2	3n79_1	3ne3_2	3nfd_1
3o46_1	3oiu_2	3opk_1	3p48_1	3pzy_1	3qc7_1	3qr7_1	3quw_1	3r1w_1	3r3r_2	3rwn_1
3so2_1	3ta2_1	3tio_1	3tq5_1	3tqz_1	3uv9_2	3v4d_1	3vi6_2	3zw0_1	4aff_1	4g2k_1
4gb5_1	4gdz_1									

Section 3: Structure of a designed tetrahedral assembly variant engineered to have improved soluble expression

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.

Background and Motivation

Symmetric homomeric and heteromeric protein complexes perform a broad range of functions in biological systems[30,31]. Inspired by these natural protein-based molecular machines and materials, many efforts have been undertaken to design novel supramolecular protein structures[1-4,6,14-16,18-23,75-77]. 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[1,6]. 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[6].

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[7,8] 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.

Computational Design Strategy

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[6]. 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 as follows[5,38]. Using the original T33-09 design model as the starting point, with the same treatment of the backbone and rigid body DOFs as published previously[6], 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[45], yielding four new variants with 4 to 8 mutations per subunit compared to the original design (**Table 4.1**).

Screening for Improved Soluble Yield

Synthetic genes encoding the four designed variants (**Table 4.1**) 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 (**Figure 4.1**). We named this new design variant, which contains 5 mutations in the A component relative to the original design, T33-31 (**Figure 3.1a**).

Characterization by SDS-PAGE, Gel Filtration, and Electron Microscopy

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 (**Figure 3.1b**). In addition, gel filtration of individually expressed subunits purified by nickel affinity chromatography showed a substantial reduction of apparent soluble aggregate in the negatively charged subunit A sample compared to the original design (**Figure 4.2b**), suggesting that the negatively charged variant has less of a tendency to self-associate. 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 (**Figure 3.1c**). 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 (**Figure 3.1d**). 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 (**Figure 3.1e**)[6]. 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.

Structure Validation

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 (**Figure 3.2** and **Table 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 (**Figure 3.2**). 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.

Discussion

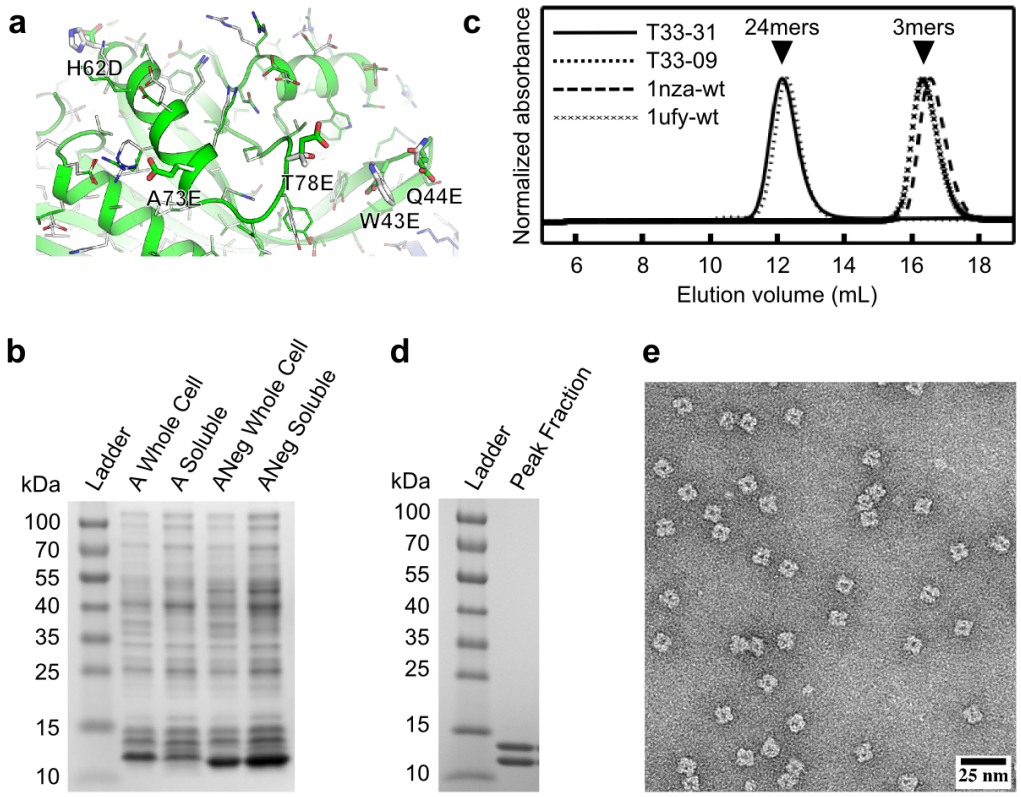
There are many possible reasons for the success of the T33-31 variant compared to the other combinations of subunits experimentally tested. SDS-PAGE analysis and gel filtration of the original A and B subunits showed that, when individually expressed, subunit A appeared to self-associate to form soluble aggregates whereas subunit B appeared to exist primarily as free trimer (**Figure 4.2**). Given this lower yield of soluble, non-aggregated subunit A compared to subunit B in the original design, it is perhaps not surprising that the best redesigned variant involved changes to subunit A, which decreased the tendency of the subunit to self-associate. It is at present less clear why the negatively charged version of subunit A worked better than the positively charged version or why the original version of subunit B worked better in combination with the negatively charged A subunit than the other versions of subunit B. It is possible that the greater total number of mutations in the positively charged variants (8 and 7 mutations for APos and BPos, respectively, compared to 5 and 4 for ANeg and BNeg, respectively), including a greater number of positions with switched charged states (e.g., a mutation from a glutamate to lysine), disrupted native interactions that stabilize the structures of the subunits. It is also possible that the behavior of the positively charged variants is complicated by interactions with cellular polyanions such as nucleic acids.

The results presented here 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. In this sense, the present work demonstrates how experimental characterization of computationally designed proteins generates

valuable feedback that can be used to improve the computational design methods. 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[78,79], 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[41,80,81] of protein nanostructures[43,82,83] should provide particularly powerful tools to help tailor these new nanomaterials for a wide variety of features and target applications.

Figures

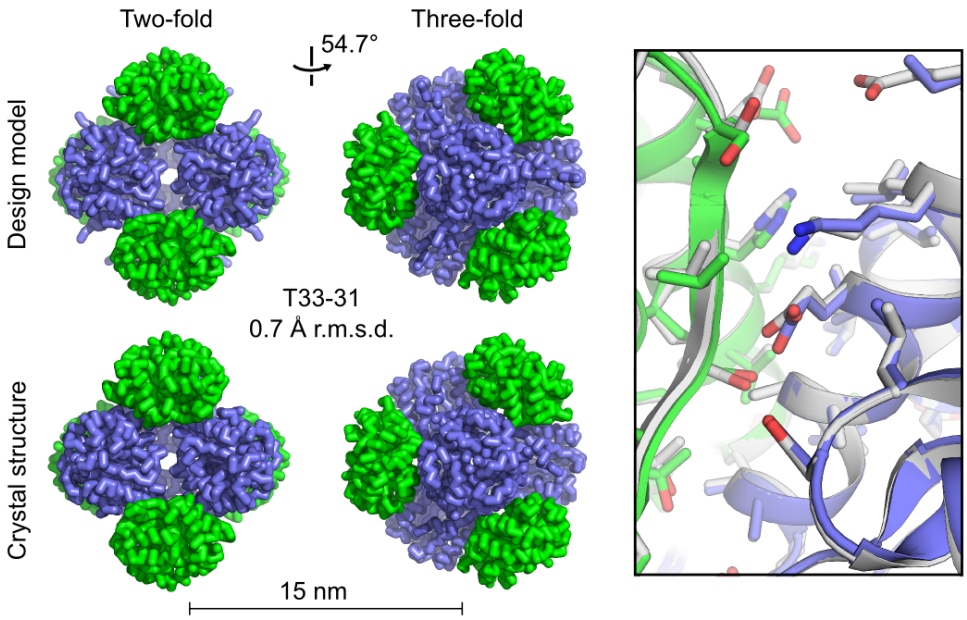
Figure 3.1



Experimental characterization of designed protein assembly T33-31 by SDS-PAGE, analytical SEC, and electron microscopy. a, Close-up of the original subunit A and negatively charged subunit A (Aneg) from the T33-09 (white) and T33-31 (green and blue) design models. The five surface residues mutated in T33-31 compared to T33-09 are labeled and shown as sticks. **b**, 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. **c**, 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. **d**, 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. **e**, Negative stain electron micrograph of *in vitro*-mixed, SEC-purified T33-31.

Figure 3.2



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.

Section 4: Supplementary information for structure of a designed tetrahedral assembly variant engineered to have improved soluble expression

Materials and Methods

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[6]. 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[6].

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[58]. 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[59].

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

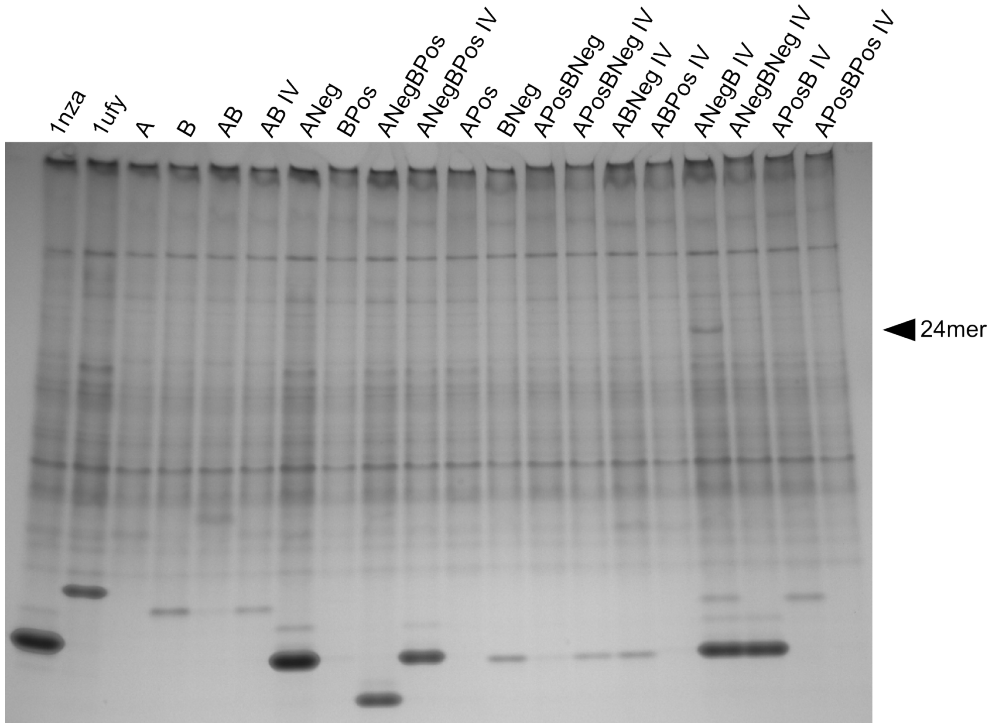
Diffraction data sets were collected at the Advanced Photon Source (APS) beamline 24-ID-C equipped with a Pilatus-6M detector. All data were collected at 100 K. Data were collected at a detector distance of 602 mm, with 0.5° oscillations, and at 0.979100 Å wavelength. The

crystals showed diffraction to 3.25 Å. The XDS/XSCALE package[65] 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[66] 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[84-87] 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[69], 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_ENREF_33[67]. 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.

Figures

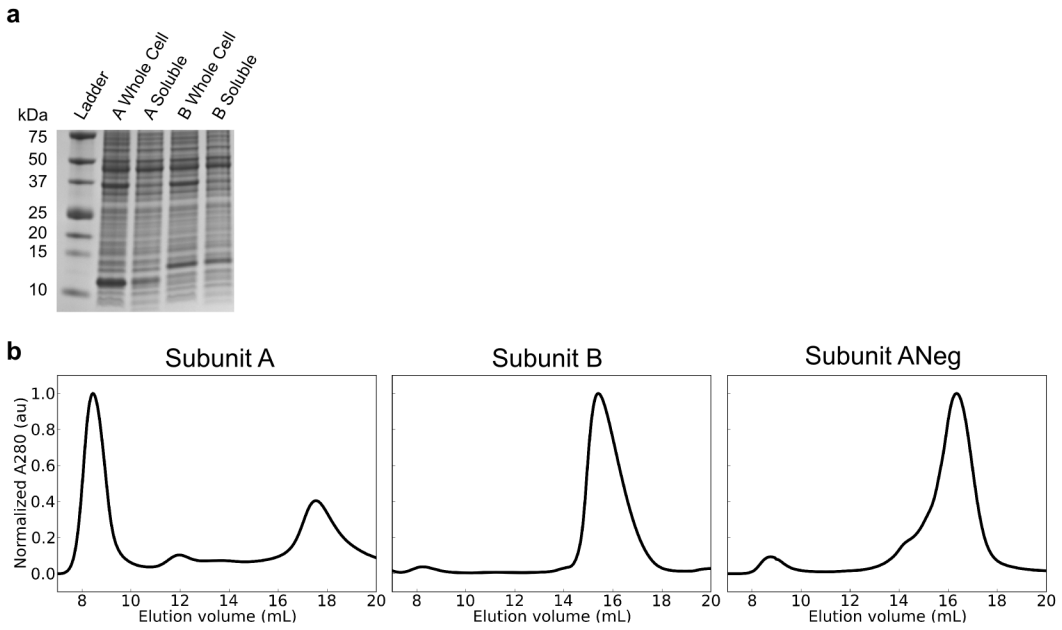
Figure 4.1



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 tag1. 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.

Figure 4.2



SDS-PAGE and gel filtration of the original subunits and negatively charged subunit A. a, SDS-PAGE analysis of whole cell and clarified lysates from cells expressing the original subunit A or subunit B. Bands are visible near the expected molecular weight of 12.5 kDa for subunit A in lanes 2 and 3 and near the expected molecular weight of 14.5 kDa for subunit B in lanes 4 and 5. Although a more intense band is observed for subunit A in the whole cell lysate than subunit **b**, a similar amount of subunit A and B are observed in the soluble fractions. (B) SEC chromatograms of nickel purified subunit A, B, and ANeg. Each of the proteins were individually expressed and purified via nickel affinity chromatography and the pooled and concentrated samples subjected to gel filtration with a Superdex 200 10/300 GL gel filtration

column as described in the Materials and Methods section. The majority of the nickel purified protein from the subunit A sample is observed to elute near the void volume (8 mL), indicating a high propensity to form large, soluble aggregates. In constrast, the primary peaks for the original subunit B and redesigned, negatively charged subunit A are both observed near the expected elution volume of for free timers in solution (~16 mL), with only very minor peaks observed near the void volume.

Tables

Table 4.1 | 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]	MEEVVLITVPSEEVARTIAKALVEERLAACVNIVPGLTSIYRWQGEVVEDQELLL LVKTTTHAFPKLKERVKALHPYTVPEIVALPIAEGNREYLDWLRENTG
T33-09A	MEEVVLITVPSALVAVKIAHALVEERLAACVNIVPGLTSIYRWQGSVVSDHELLL LVKTTTHAFPKLKERVKALHPYTVPEIVALPIAEGNREYLDWLRENTG
T33-09ANeg	MEEVVLITVPSALVAVKIAHALVEERLAACVNIVPGLTSIYREEGSVVSDHELLL LVKTTTDAFPKLKERVKELHPYEVPEIVALPIAEGNREYLDWLRENTG
T33-09APos	MEEVVLITVPSAKVAVKIAHALVKERLAACVNIVPGLTSIYRKKGSVVSDHELLL LVKTTTKAFPKLKERVKRLHPYKVPEIVALPIAEGNREYLRWLRENTG
1UFY[2]	MVRGIRGAITVEEDTPEAIHQATRELLLKMLEANGIQSYEELAAVIFTVTEDLTS AFPAEAARQIGMHRVPLLSAREVPVPGSLPRVIRVLALWNTDTPQDRVRHVYLRE AVRLRPDLESAQ
T33-09B	MVRGIRGAITVEEDTPAAILAATIELLLKMLEANGIQSYEELAAVIFTVTEDLTS AFPAEAARLIGMHRVPLLSAREVPVPGSLPRVIRVLALWNTDTPQDRVRHVYLNE AVRLRPDLESAQ
T33-09BNeg	MVRGIRGAITVEEDTPAAILAATIELLLKMLEANGIESYEELAAVIFTVTEDLTS AFPAEAARLIGMHRVPLLSAREVPVPGSLPRVIRVLALWNTDTPQDEVRHVYLNE AVELRPDLESDQ
T33-09BPos	MVRGIRGAITVEEDTPAAILAATIELLLKMLKANGIQSYKELAAVIFTVTEDLTS AFPAEAARLIGMHRVPLLSAREVPVPGSLPRVIRVLALWNTKTPQDRVRHVYLNK AKRLRPDLKSKQ

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 4.2 | 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.

Section 5: Accurate design of megadalton-scale, co-assembling icosahedral protein complexes

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 as well as mediate binding and entry into host cells. Inspired by these natural materials, here we report the computational design and experimental characterization of megadalton-scale, two-component icosahedral protein nanostructures. Ten designs spanning three distinct icosahedral architectures were found to form materials closely matching the design models. Comprising 120 subunits each, with molecular weights (1.8 to 2.8 MDa) and dimensions (240 to 400 Å diameter) comparable to those of small viral capsids, these results represent a new milestone in protein and nanomaterial engineering. The ability to design such large and complex structures with high accuracy presents exciting new opportunities for a broad range of applications, including vaccines, targeted delivery, and bioenergy.

Background and Motivation

The remarkable forms and functions of natural protein assemblies have inspired many efforts to engineer novel self- and co-assembling protein complexes[1-4,6,17-23,28,29,49,76,77,88,89]. A common feature of each of these approaches, as well as the natural 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, with many accompanying benefits[15,30]. Some of the most impressive and prevalent examples of symmetry in nature are icosahedral virus capsids. Icosahedra possess the highest symmetry of all possible regular convex polyhedra (tetrahedra, hexahedra, octahedra, dodecahedra, and icosahedra) and, importantly for the purpose of packaging viral genomes, generate the maximum enclosed volume for symmetric assemblies formed from of a given size protein subunit[10,11].

Despite exciting advances in protein nanomaterial engineering in the last few years, accurate design of icosahedral protein assemblies has yet to be demonstrated. Here we address this elusive goal by means of an approach in which multiple copies of two or more distinct types of oligomeric protein building blocks are docked in a target symmetric architecture and *de novo* protein-protein interfaces are designed between the oligomers to both rigidly define their relative orientation and provide the energetic driving force for assembly to the specific target configuration[6]. The co-assembling, multi-component nature of this approach provides a much larger search space for design and many potential advantages for downstream applications compared to approaches restricted to the design of self-assembling, homo-oligomeric complexes. We focus here on three different icosahedral architectures formed through the co-assembly of two distinct protein components.

Computational design

The two-fold, three-fold, and five-fold symmetry axes present within icosahedral symmetry provide three possible ways in which to construct icosahedra from pairwise combinations of two distinct types of oligomeric building blocks; we refer to these architectures as I53, I52 and I32 (**Figure 5.1**). 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 (I53 = Icosahedron constructed from 5mers and 3mers). Similarly, the I52 architecture is formed from twelve pentameric and thirty dimeric building blocks, and the I32 architecture is formed from twenty trimeric and thirty dimeric building blocks, wherein each building block is aligned along its corresponding five-fold, three-fold, or two-fold icosahedral symmetry axis.

The building blocks used in the present study were derived from pentameric, trimeric, and dimeric crystal structures in the Protein Data Bank (PDB), along with a small number of crystal structures of de novo designed oligomers not yet deposited in the PDB (Supplementary Materials and Methods, **Tables 6.1-6.3**). During design, pairs of building blocks were arranged in the I53, I52, or I32 symmetric architectures 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 during docking to identify configurations suitable for interface design (**Figure 5.1**, Supplementary Materials and Methods). In total, 14,400 pairs of pentamers and trimers, 50,400 pairs of pentamers and dimers, and 276,150 pairs of trimers and dimers were docked in the I53, I52, and I32 architectures, respectively.

Up to fifty top scoring configurations were output for each docking trajectory and filtered based on several metrics, including the total docking score and the number of contacting residues at the docked interface. Protein-protein interface design calculations were carried out on the resulting 66,115 I53, 35,468 I52, and 161,007 I32 configurations and the designs filtered based on interface area, predicted binding energy, and shape complementarity[39]. Following an additional stage of design aimed at minimizing the number of mutations from the native

sequences, 71 I53, 44 I52, and 68 I32 designs were selected for experimental characterization (Supplementary Materials and Methods, **Figures 6.1-6.3**). The 183 designs were derived from 23 distinct pentameric, 57 distinct trimeric, and 91 distinct dimeric protein scaffolds and contained an average of 20 mutations per design compared to the native sequences.

Experimental Characterization

Codon optimized genes encoding each pair of designed amino acid sequences were synthesized (Gen9 Inc.) and cloned into a vector for inducible co-expression in E. coli, with a single hexahistidine tag appended to the N- or C-terminus of one subunit in each pair (Supplementary Materials and Methods). The designed proteins were expressed at small scale in 96 well culture plates and purified via immobilized metal-affinity chromatography (IMAC) using nickel-coated filter plates (His MultiTrap FF, GE Healthcare). Clarified cell lysates and purification products were then subjected to polyacrylamide gel electrophoresis under denaturing conditions (SDS PAGE) to screen for soluble expression and co-purification (**Figure 6.4a**); because only one subunit in each designed pair possessed a hexahistidine tag, co-purification was used as an indication that the two subunits were interacting. Designs appearing to co-purify were subsequently analyzed by non-denaturing polyacrylamide gel electrophoresis to screen for slowly migrating species as an additional indication of assembly to higher order materials (**Figure 6.4b**). Designs appearing to co-purify and assemble were subsequently expressed at larger scale and purified by IMAC followed by size exclusion chromatography (SEC, **Figure 6.5**). 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 at elution volumes between 8.5 and 12 mL, corresponding well with expected

elution volumes based on the diameters of the computational design models (**Figure 5.2** and **Table 6.4**).

To further investigate the structures of these ten materials, small-angle X-ray scattering (SAXS) and negative stain electron microscopy were performed on the SEC-purified samples (Supplementary Materials and Methods). The large-scale features of the design models were all found to match well with the SAXS data, supporting the conclusion that these designs assemble to the intended three-dimensional configurations in solution. Plots of the log of the scattering intensity, I, as a function of scattering vector, q, show multiple large dips in the intensity in the region between 0.015 A-1 and 0.15 A-1, each of which is closely recapitulated in the profiles calculated from the design models (**Figure 5.2**). Negative stain electron microscopy provides additional low-resolution confirmation that I53-34, I53-40 I53-47, I53-50, I52-03, I52-33, I32-06, and I32-28 each assemble specifically to the target architectures (**Figure 5.3**). Micrographs of each designed material show fields of particles with the approximate size and shape of the design models. Particle averaging yields distinct staining patterns clearly matching the models at low resolution. The large trimeric and pentameric voids observed in the I52 and I32 averages, for instance, closely resemble the pores in the projections calculated 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 I52-32 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 particles are visible in electron micrographs of I52-32, and appear to match the design model at low resolution, many unassembled pentameric and dimeric components are also

visible (**Figure 6.6a-b**). Due to the heterogeneity of assembly states observed in this sample, particle averaging was not attempted. In addition, attempts to obtain electron micrographs of the I32-19 assembly were also unsuccessful, yielding what appear primarily to be partially assembled cages and aggregates (**Figure 6.6c-d**). Despite these findings, the results from SEC and SAXS strongly indicate both I52-32 and I32-19 form assemblies closely matching the design models in solution.

To further evaluate the accuracy of our designs, X-ray crystal structures were determined for one material from each of the three different architectures: I53-40, I52-32, and I32-28 (**Figure 5.4**). Each crystal structure was found to contain only a portion of the expected icosahedral assembly in the asymmetric unit, with the full icosahedra formed through application of crystal lattice symmetry. Although the resolution of the structures (3.5 to 5.0 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 1.0 Å, while the r.m.s.d. over all 120 subunits in each material ranges from 0.6 to 2.6 Å (**Figure 5.4**). 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 protein nanostructures to date to be verified by X-ray crystallography (**Figure 6.7**).

In addition to the ten designs discussed above, 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 (**Figure 6.8**). 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 (~35-40 nm vs. 29 nm, **Figure 6.9**). Retrospective analysis of the computational design models reveals I53-51, I32-10, and the ten successful designs are similar to the other 171 tested designs according to a wide range of computational metrics related to the designed interfaces, including the predicted binding energy, predicted binding energy density, shape complementarity[39], and the number of buried unsatisfied hydrogen bonding groups (**Figure 6.10**). I53-51 and I32-10 do however possess both a greater total number of mutations and a higher percentage of mutated residues than any of the other ten successful designs. We also found that the majority of the successful designs fall in the lower half of the range of tested interface sizes and number of mutated residues, and possess sequences with improved secondary structure propensities compared to the native scaffolds. Taken together, these results suggest that employing a more stringent cutoff on the number of mutated residues and exploring additional computational methods of assessing scaffold stability and foldability are likely to lead to increased success rates in future design efforts.

Discussion

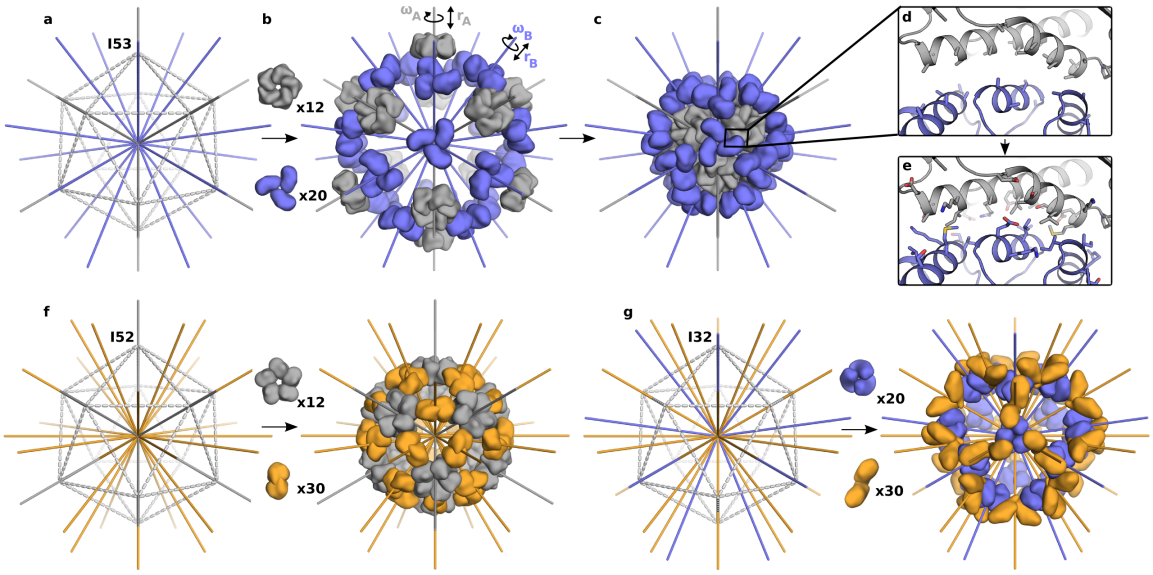
Given the prevalence of icosahedral symmetry in natural protein structures such as viral capsids, it is interesting to consider how our target architectures compare. Adhering to the triangulation number rules set out by Caspar and Klug[10], our design targets can be considered forms of T=1 assemblies in which the asymmetric unit is a heterodimer comprised of one subunit from each of our two different components. With regard to the overall architecture of the assemblies, the most similar naturally occurring structures of which we are aware are of the

Cowpea Mosaic Virus (CPMV) and related viral capsids with so called "pseudo T=3" symmetry. CPMV comprises 60 copies each of two distinct protein subunits, with the small (S) subunit arranged around the icosahedral 5-folds and the large (L) subunit around the 3-folds, and thus fits within the parameters of the I53 architecture (**Figure 6.11**). However, the S and L subunits possess three domains, each occupying spatially equivalent positions to those found in "T=3" capsids formed from 180 copies of a single type of subunit[90]. Our I53 designs display no such underlying features and therefore cannot be considered "pseudo T=3" capsids like CPMV. Furthermore, we are not aware of any natural assemblies 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 large lumens of our designed materials, combined with their multi-component nature and potential to control assembly in vitro via mixing of purified components[6], 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 chemical or genetic modifications, make them attractive candidates for a wide range of applications, including targeted drug delivery, vaccine design, and bioenergy.

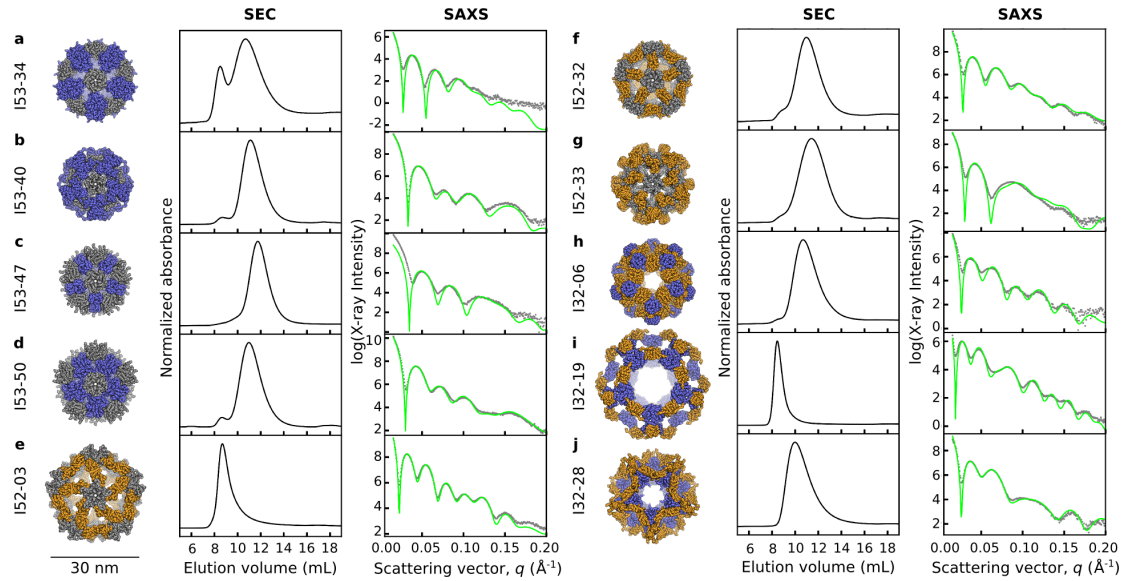
Figures

Figure 5.1



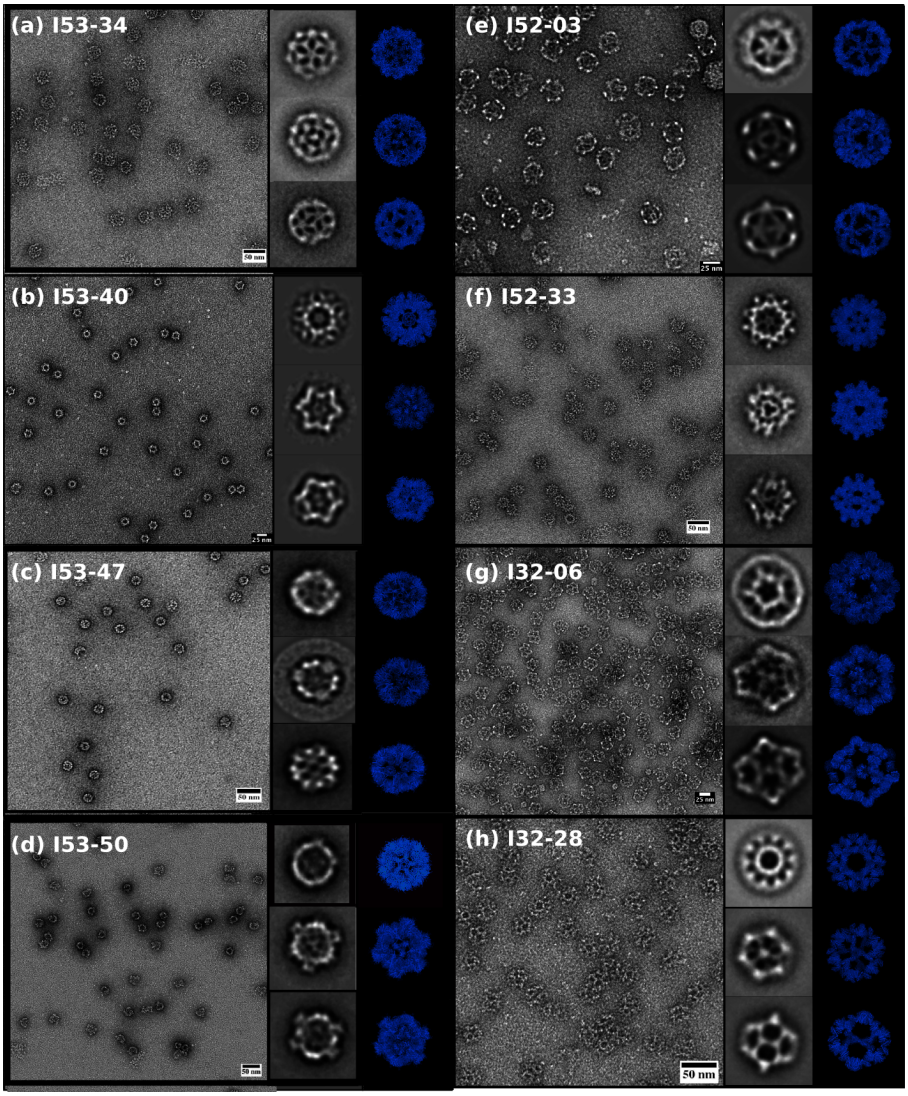
Overview of the design method and target architectures. **a,** A schematic illustration of icosahedral symmetry outlined with dashed lines, with the five-fold symmetry axes (grey) going through each vertex and three-fold symmetry axes (blue) going through each face of the icosahedron. **b-c,** 12 pentamers (grey) and 20 trimers (blue) are aligned along the 5-fold and 3-fold symmetry axes, respectively. Each oligomer possesses two rigid body degrees of freedom, one translational (r) and one rotational (ω) that are systematically sampled to identify configurations with large interfaces and high densities of contacting residues suitable for protein-protein interface design. **d-e,** Amino acid sequences are designed at the new interface to stabilize the modeled configuration. **f,** The I52 architecture is comprised of 12 pentamers (grey) and 30 dimers (orange) aligned along the five-fold and two-fold icosahedral symmetry axes. **g,** And the I32 architecture is comprised of 20 trimers (blue) and 30 dimers (orange) aligned along the three-fold and two-fold icosahedral symmetry axes.

Figure 5.2



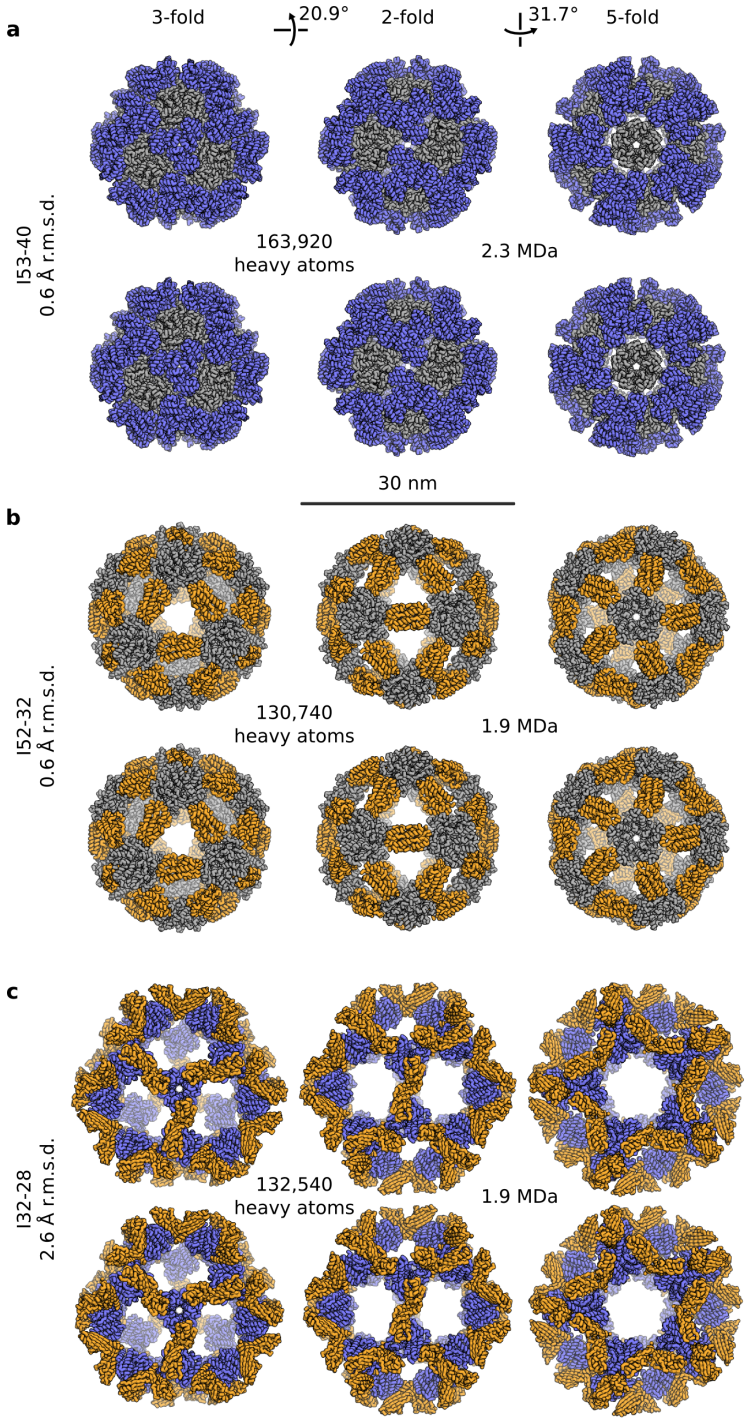
Experimental characterization by size exclusion chromatography and small-angle X-ray scattering. Computational design models (left), SEC chromatograms (middle), and SAXS profiles (right) are shown for **a,** I53-34 **b,** I53-40 **c,** I53-47 **d,** I53-50 **e,** I52-03 **f,** I52-32 **g,** I52-33 **h,** 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 assemblies and X-ray scattering intensities (grey dots) that match well with profiles calculated from the design models (green).

Figure 5.3



Characterization of the designed materials by electron microscopy. 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 are shown on the left side of each panel, with three different class averages (roughly corresponding to the five-fold, three-fold, and 2-fold symmetry axes) shown directly to the right of the raw micrographs along with back projections from the computational design models.

Figure 5.4



Crystal structures of the I53-40, I52-32, and I32-28 designs. Computational design models (top) and X-ray crystal structures (bottom) are shown for **a,** I53-40 **b,** I52-32 and **c,** I32-28.

Views of each material are shown to scale along the 3-fold, 2-fold, and 5-fold icosahedral symmetry axes (scale bar: 30 nm). The r.m.s.d. values provided are those between the backbone atoms in all 120 subunits of the design models and crystal structures. The number of heavy atoms and approximate molecular weight of each 120-subunit assembly are also provided.

Section 6: Supplementary information for accurate design of megadalton-scale, co-assembling icosahedral protcin complexes

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[45]) 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[44] 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 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 6.1**, homotrimeric scaffold set in **Table 6.2**, and homodimeric scaffold set in **Table 6.3**.

Symmetric docking

Symmetric docking was carried out as described previously[6], 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 structure 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 that had a score less than 180 and fewer than 35 or greater than 70 contacting residues at the interface. 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[6]. 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.

Protein-protein interface design

Protein-protein interface design was carried out as described previously[6], with the following exceptions. The design process was split into three stages: I) interface design, II) automated reversion, and III) resfile-based refinement, rather than the four stages used previously, which included an additional stage for shape complementarity optimization. During Stage I (interface design) the side chain conformations of the interaction motifs identified during docking were added to the side chain rotamer library used for design and only 10 design trajectories were carried out per docked configuration, compared to the 100 trajectories carried out previously. In addition, the criteria used to select designs at each stage were modified and extra features were added to the automated reversion protocol in order to identify and revert mutations that resulted in substantial losses to core packing or hydrogen bonding between backbone polar atoms and side chain atoms compared to the native scaffold.

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'-
TAATGCTTAAGTCGAACAGAAAGTAATCGTATTGTACACGGCCGCATAATCGAAATT
AATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCATCTTAGTATAT
TAGTTAAGTATAAGAAGGAGATATACTT-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 gene, intergenic region, downstream gene, 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 gene, intergenic region, downstream gene, XhoI restriction site. In each case, the upstream genes encode components denoted with the suffix "A"; the downstream genes encode the "B" components (**Table 6.4**). This allows for co-

expression of the designed protein pairs in which both the upstream and downstream genes have their own ribosome binding site, and in the case of the I53 designs, both genes also have their own T7 promoter/lac operator.

For purification purposes, each co-expression construct includes a 6x-histidine tag (HHHHHH) appended to the N- or C-terminus of one of the two protein coding regions.

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 from 2 to 4 mL of 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. Although adding 5 % (v/v) glycerol to all buffers used for purification of I32-19 appeared to moderately improve the results from electron microscopy, it did not appear to make much difference with results obtained from 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 6.6b**.

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, 2) 5 % (v/v) glycerol was added and the NaCl concentration increased to 300 mM in the I53-51 buffer, and 3) the NaCl concentration was increased to 500 mM in the I52-33 buffers. The designed materials were loaded onto the column with each component present at a subunit concentration of 20-50 μM.

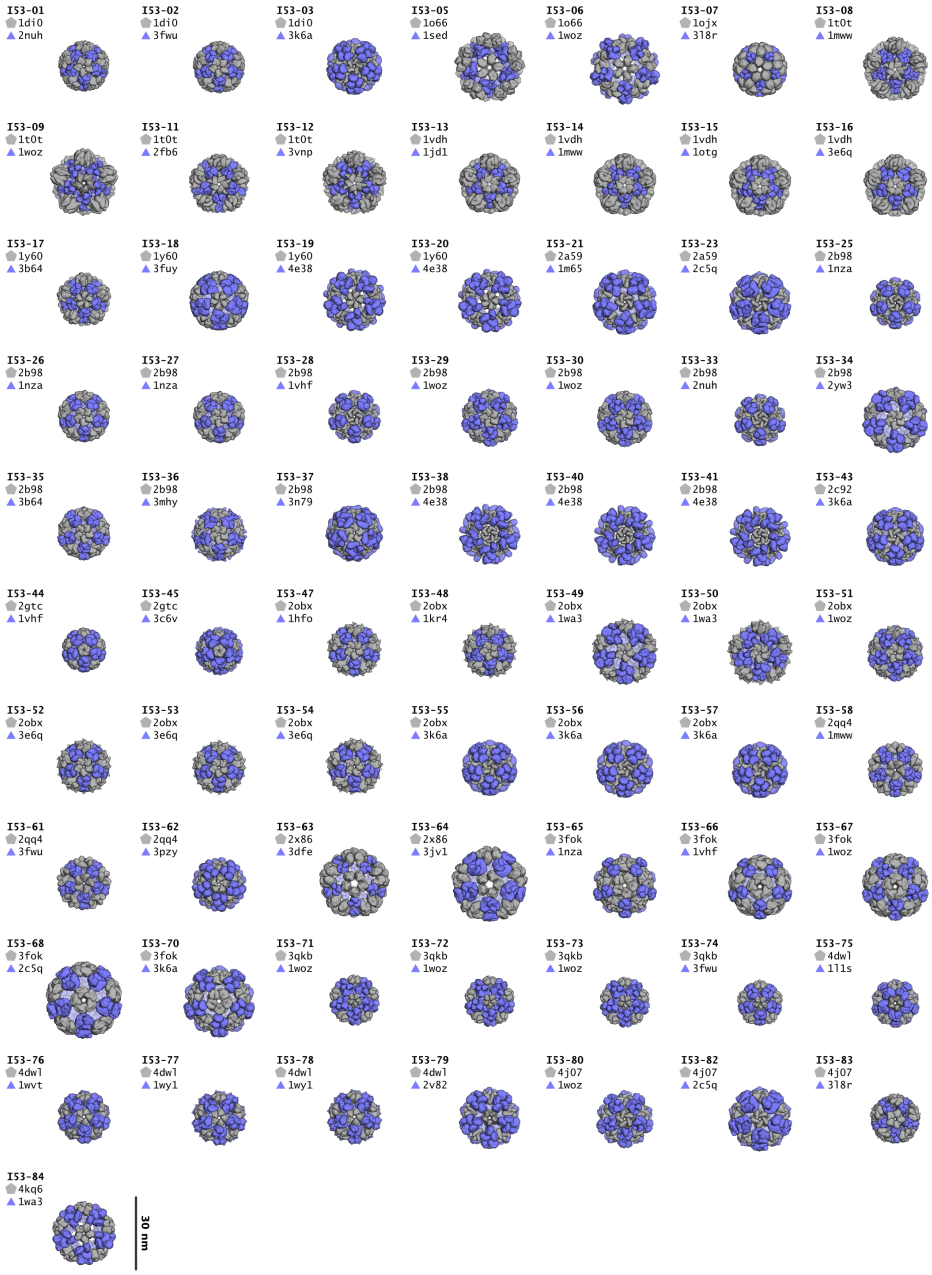
Small-angle X-ray scattering

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[91]. Purified samples were rerun over gel filtration in 25 mM TRIS pH 8.0, 150 mM NaCl, 2 % (v/v) glycerol, 1 mM DTT running buffer (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 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 samples concentrated ~2x-10x from the gel filtration fractions, with the gel filtration buffer and concentrator eluates used for buffer subtraction. Sequential exposures ranging from 0.5 to 5 seconds were taken at 12 keV to maximize signal to noise, with visual checks for radiation-induced damage to the protein. The FOXS algorithm[92,93] was then used to calculate scattering profiles from our design models and fit them to the experimental data.

Figures

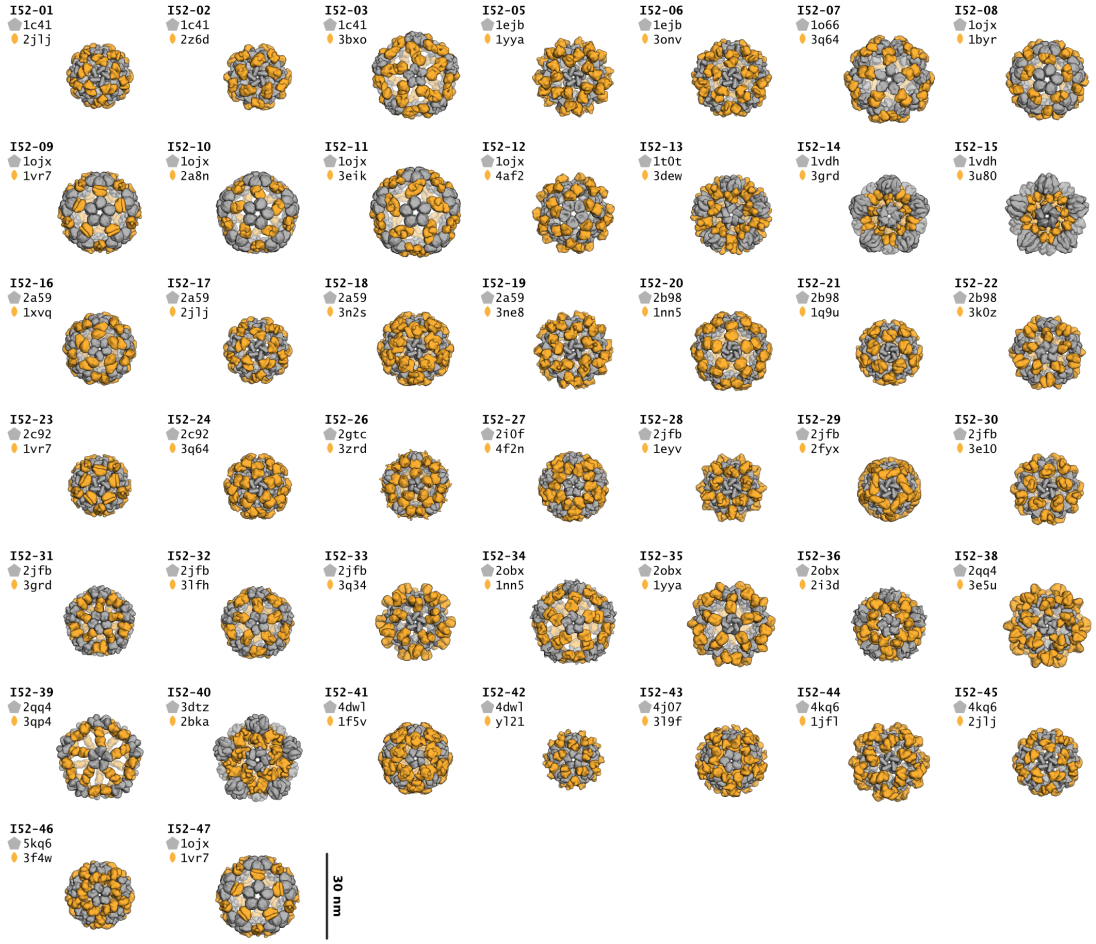
Figure 6.1



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.

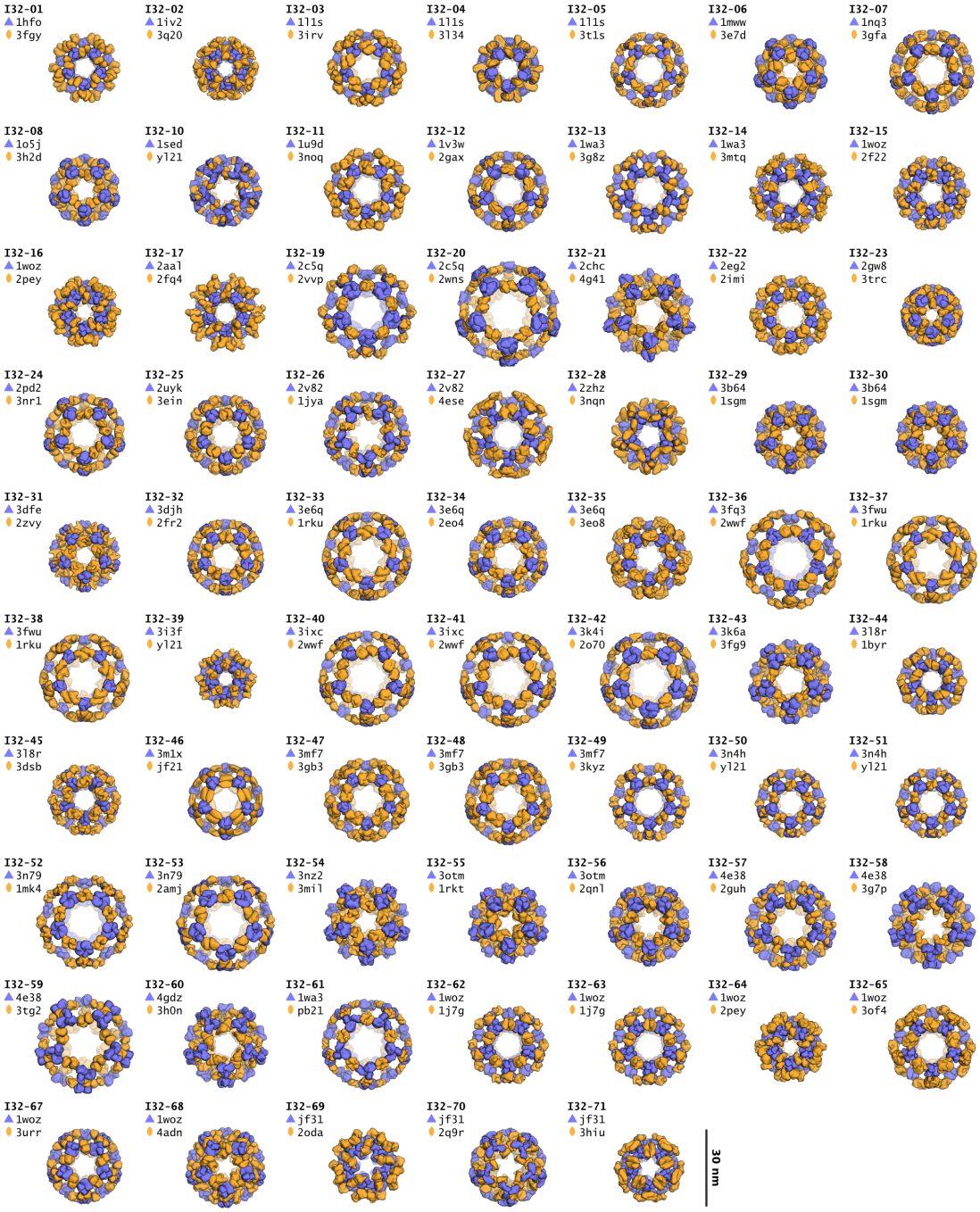
Figure 6.2



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.

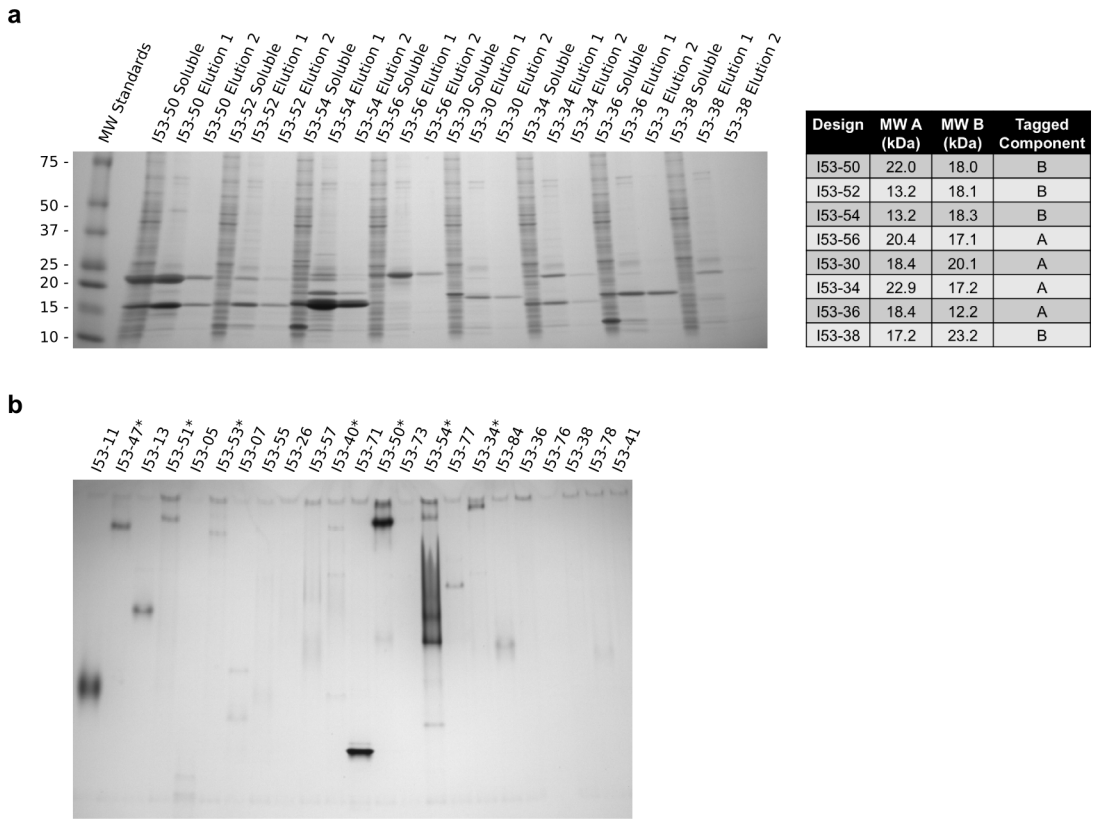
Figure 6.3



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.

Figure 6.4

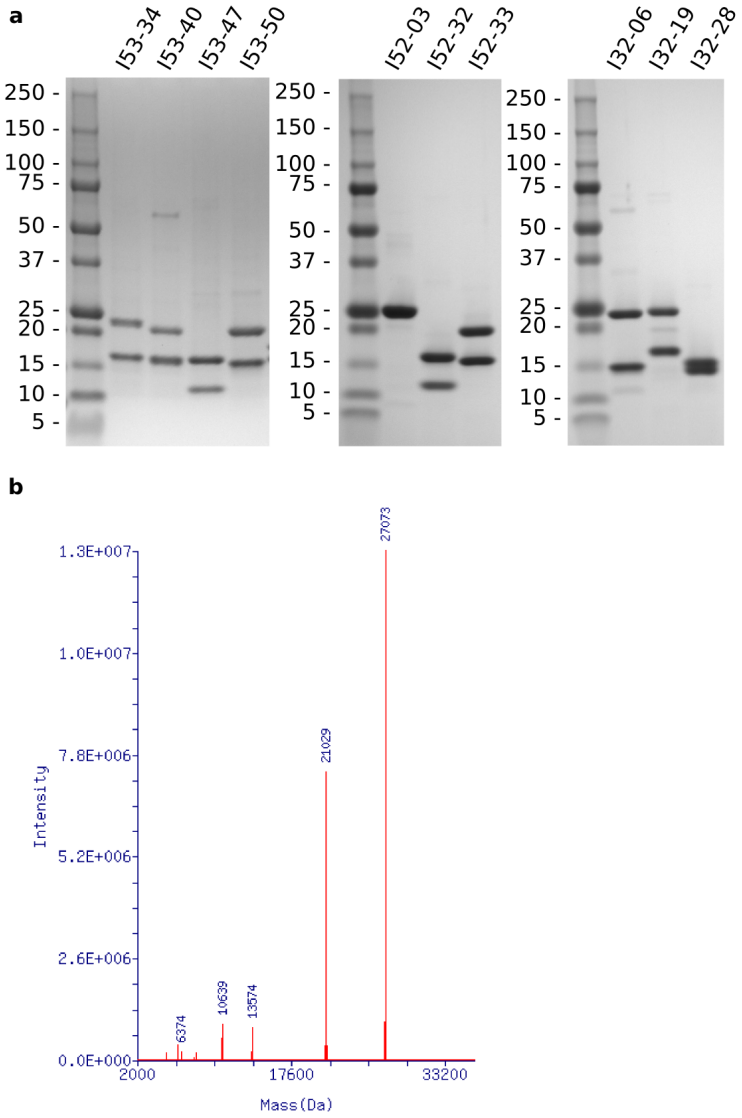


Design	MW A (kDa)	MW B (kDa)	Tagged Component
I53-50	22.0	18.0	B
I53-52	13.2	18.1	B
I53-54	13.2	18.3	B
I53-56	20.4	17.1	A
I53-30	18.4	20.1	A
I53-34	22.9	17.2	A
I53-36	18.4	12.2	A
I53-38	17.2	23.2	B

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).

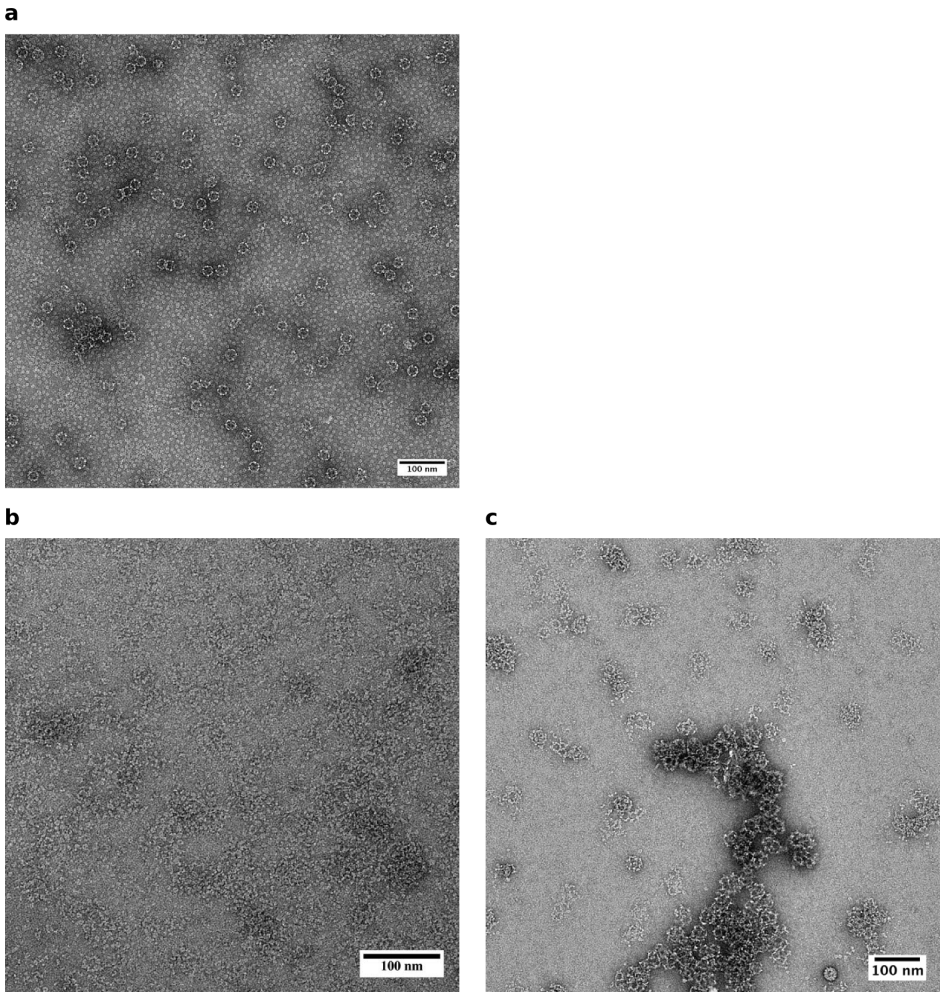
Figure 6.5



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.

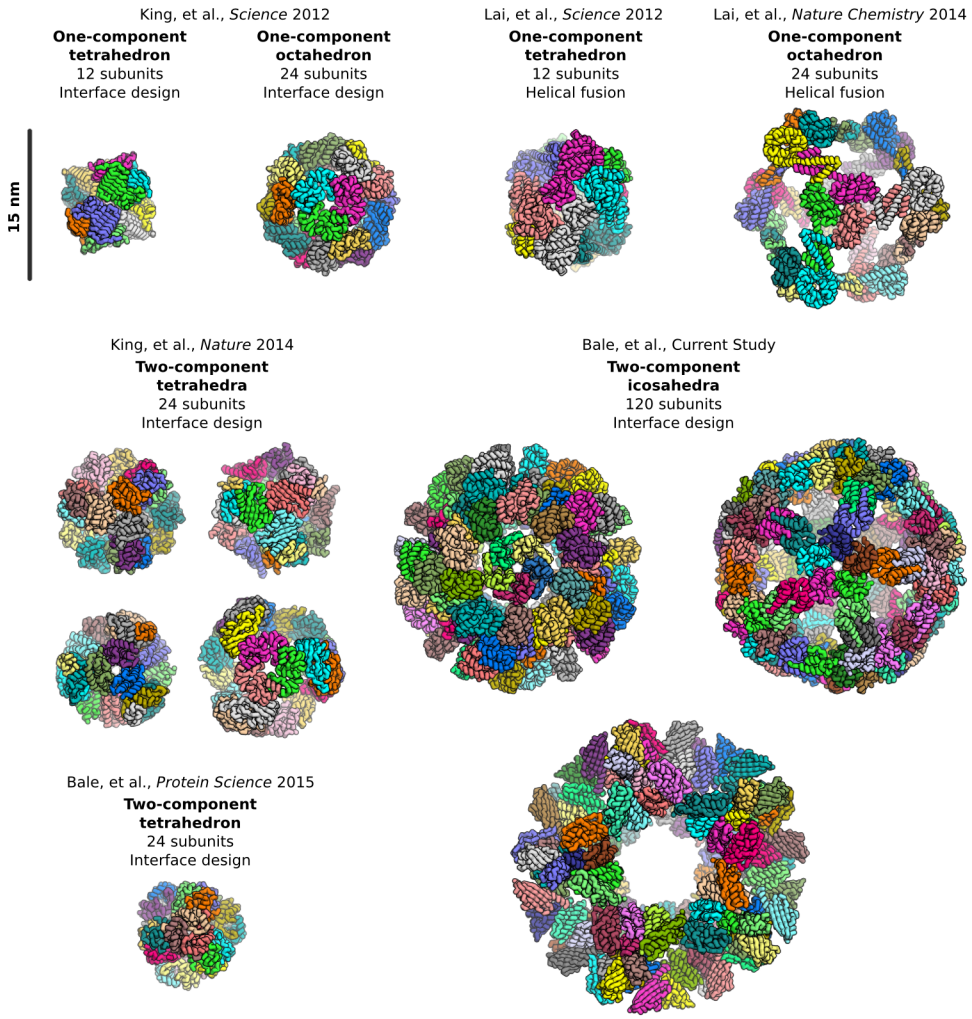
Figure 6.6



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.

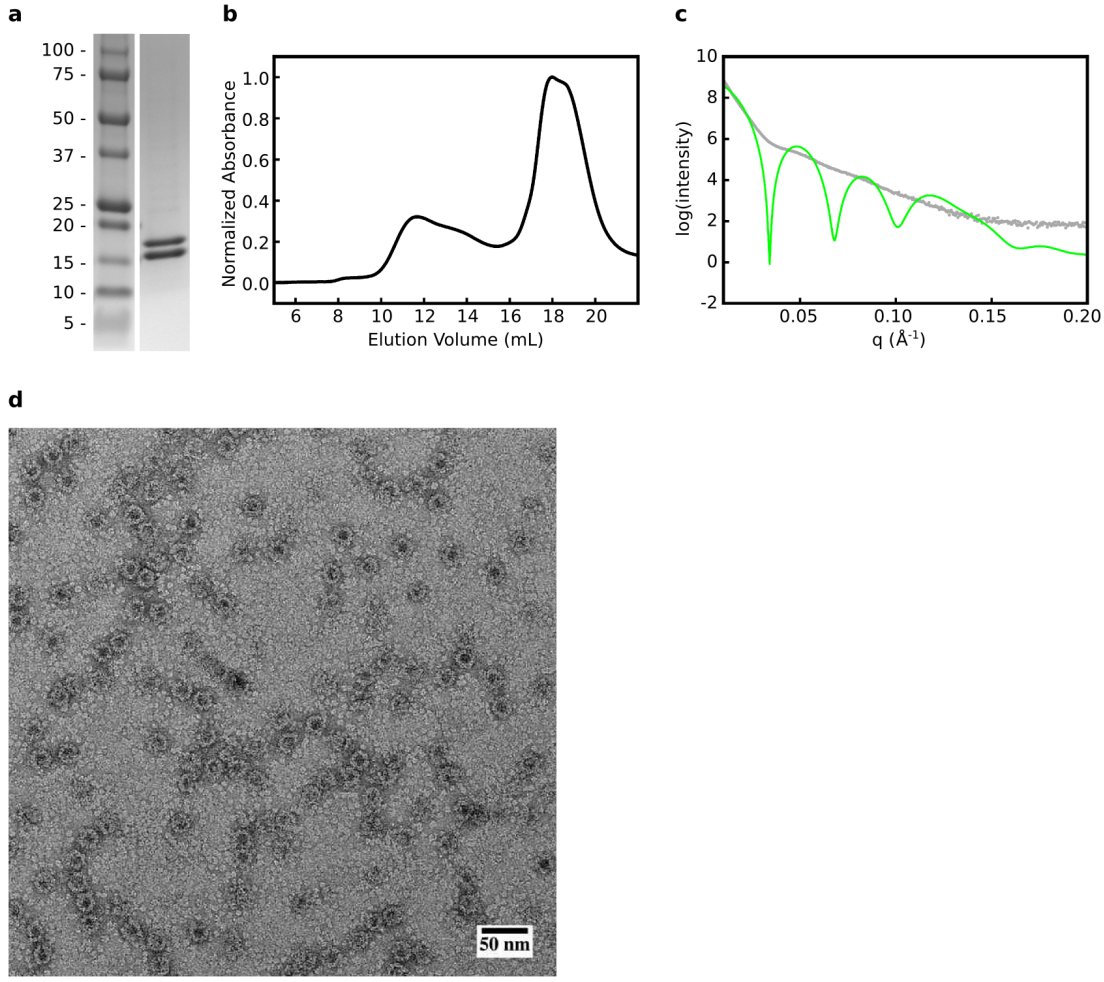
Figure 6.7



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.

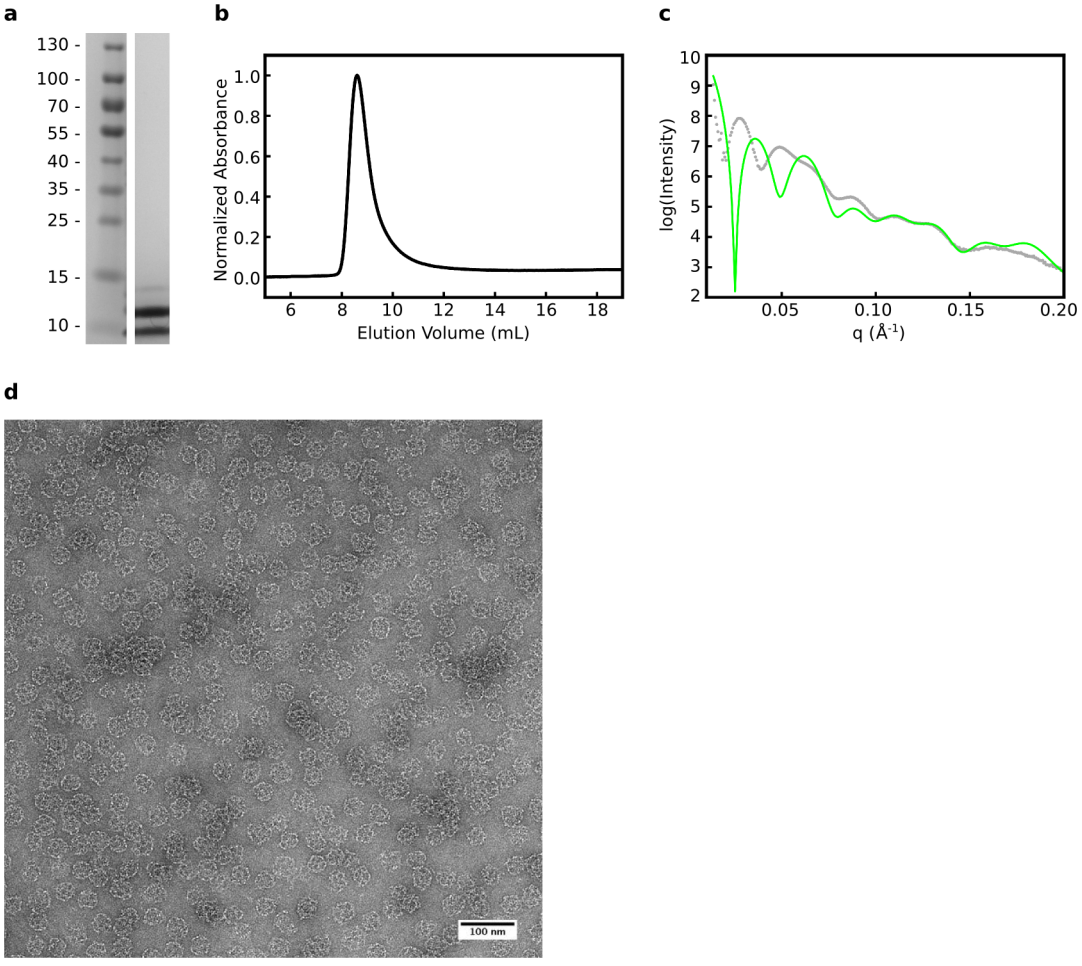
Figure 6.8



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.

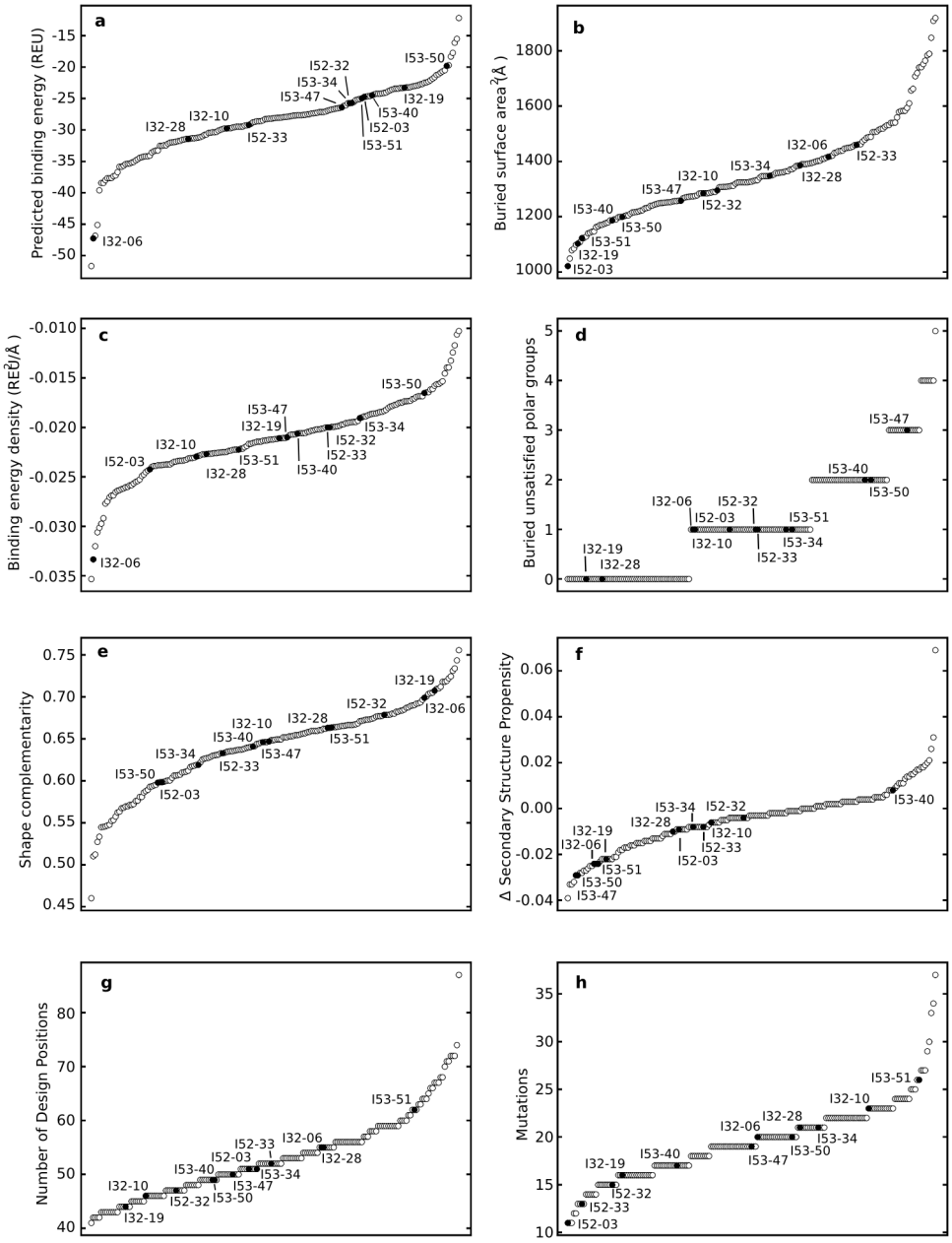
Figure 6.9



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.

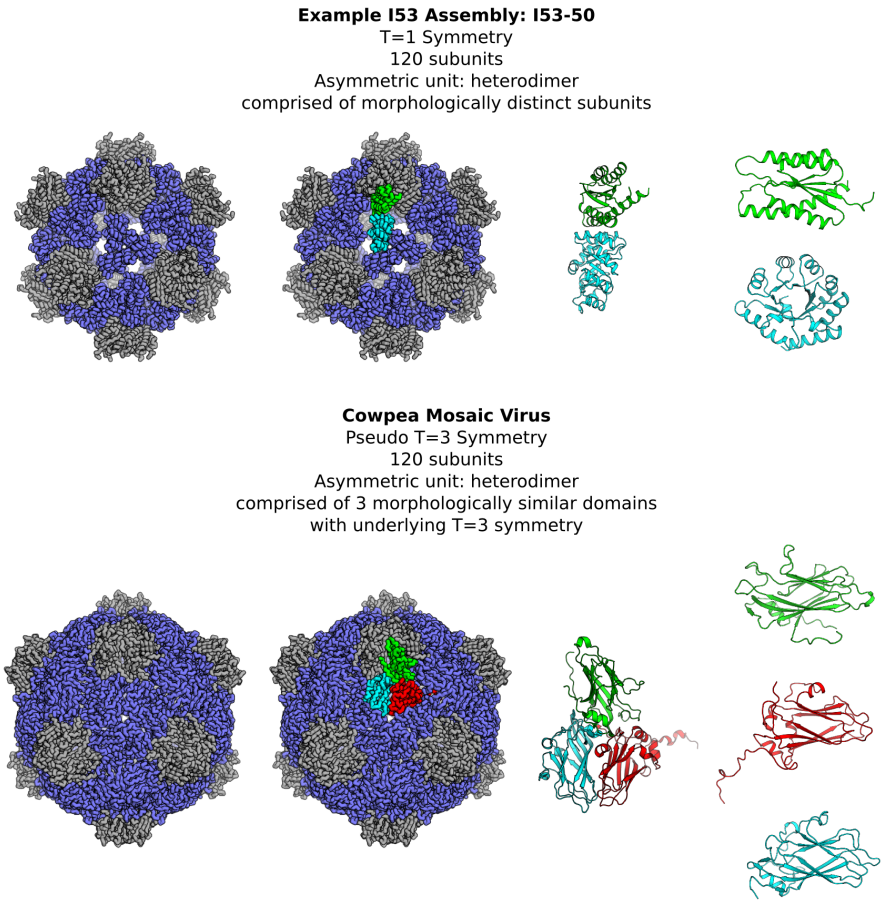
Figure 6.10



Metrics of the computational design models. Selected metrics related to the designed interfaces are plotted for the 183 designs that were experimentally characterized, including **a,** the predicted binding energy measured in Rosetta energy units (REU), **b,** the surface area buried by each instance of the designed interface, **c,** the binding energy density (calculated as the predicted

binding energy divided by the buried surface area), **d,** the number of buried unsatisfied polar groups at the designed interface, **e,** the shape complementarity of the designed interface, **f,** the change in secondary structure propensity of the designed sequences compared to the native sequences (negative scores indicate secondary structure predictions generated from the designed sequences match the backbone structure better than the predictions generated from the native sequences), **g,** the total number of residues allowed to mutate or change side chain conformations during design and **h,** the total number of mutations in each designed pair of proteins. Each circle represents a single design; the ten confirmed designs, as well as I53-51 and I32-10, are plotted as filled circles and labelled. In each plot, the designs are arranged on the x axis in order of increasing value of the metric analysed.

Figure 6.11



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.

Tables

Table 6.1 | 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 6.2 | 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 6.3 | 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 6.4 | Amino acid sequences of the 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 designs.

Name	Sequence	Scaffold ID
I53-34A	MRGSHHHHHHGMEGMDPLAVLAESRLLPLLTVRGGEDLAGLATVLELMGVGALEITLRTEKGLEALKALRKSGLLLGAGTVRSPKEAEAALEAGAAFLVSPGLLEEVAALAQARGVPYLPGVLTPTEVERALALGLSALKFFPAEPFQGVRVLRAYAEVFPEVRFLPTGGIKEEHLPHYAALPNLLAVGGSWLLQGDLAAVMKKVKAAKALLSPQAPG	2yw3 trimer
I53-34B	MTKKVGIVDTTFARVDMAEAAIRTLKALSPNIKIIRKTVPGIKDLPVACKKLLEEEGCDIVMALGMPGKAEKDKVCAHEASLGLMLAQLMTNKHIIEVFVHEDEAKDDDELDILALVRAIEHAANVYYLLFKPEYLTRMAGKGLRQGREDAGPARE	2b98 pentamer
I53-40A	MTKKVGIVDTTFARVDMASAAILTLKMESPNIKIIRKTVPGIKDLPVACKKLLEEEGCDIVMALGMPGKAEKDKVCAHEASLGLMLAQLMTNKHIIEVFVHEDEAKDDAELKILAARRAIEHALNVYYLLFKPEYLTRMAGKGLRQGFEDAGPARE	2b98 pentamer
I53-40B	MSTINNQLKALKVIPVIAIDNAEDIIPLGKVLAENGLPAAEITFRSSAAVKAIMLLRSAQPEMLIGAGTILNGVQALAAKEAGATFVVSPGFNPNTVRACQIIGIDIVPGVNNPSTVEAALEMGLTTLKFFPAEASGGISMVKSLVGPYGDIRLMPTGGITPSNIDNYLAIPQVLACGGTWMVDKKLVTNGEWDEIARLTREIVEQVNPLEHHHHHH	4e38 trimer
I53-47A	MPIFTLNTNIKATDVPSDFLSLTSRLVGLILSKPGSYVAVHINTDQQLSFGGSTNPAAFGTLMSIGGIEPSKNRDHSAVLFDHLNAMLGIPKNRMYIHFVNLNGDDVGWNGTTF	1hfo trimer
I53-47B	MNQHSHKDYETVRIAVVRARWHADIVDACVEAFEIAMAAIGGDRFAVDVFDVPGAYEIPLHARTLAETGRYGAVLGTAFVVNGGIYRHEFVASAVIDGMMNVQLSTGVPVLSAVLTPHRYRDSAEHHRFFAAHFAVKGVEAARACIEILAAREKIAALEHHHHHH	2obx pentamer

I53-50A	MKMEELFKKHKIVAVLRANSVEEAIEKAVAVFAGGVHLIEITFTVPDADTVI KALSVLKEKGAIIGAGTVTSVEQCRKAVESGAEFIVSPHLDEEISQFCKEKG VFYMPGVMTPTELVKAMKLGHTILKLFPGEVVGPQFVKAMKGPFPNVKFVPT GGVNLDNVCEWFKAGVLAVGVGSALVKGTPDEVREKAKAFVEKIRGCTE	1wa3 trimer
I53-50B	MNQHSHKDYETVRIAVVRARWHAEIVDACVSAFEAAMADIGGDRFAVDVFDV PGAYEIPLHARTLAETGRYGAVLGTAFVVNGGIYRHEFVASAVIDGMMNVQL STGVPVLSAVLTPHRYRDSDAHTLLFLALFAVKGMEAARACVEILAAREKIA ALEHHHHHH	2obx pentamer
I53-51A	MFTKSGDDGNTNVINKRVGKDSPLVNFLGDLDELNSFIGFAISKIPWEDMKK DLERVQVELFEIGEDLSTQSSKKKIDESYVLWLLAATAIYRIESGPVKLFVI PGGSEEASVLHVTRSVARRVERNAVKYTKELPEINRMIIVYLNRLSSLLFAM ALVANKRRNQSEKIYEIGKSW	1woz trimer
I53-51B	MNQHSHKDYETVRIAVVRARWHADIVDQCVRAFEEAMADAGGDRFAVDVFDV PGAYEIPLHARTLAETGRYGAVLGTAFVVNGGIYRHEFVASAVIDGMMNVQL STGVPVLSAVLTPHRYRSSREHHEFFREHFMVKGVEAAAACITILAAREKIA ALEHHHHHH	2obx pentamer
I52-03A	MGHTKGPTPQQHDGSALRIGIVHARWNKTIIMPLLIGTIAKLLECGVKASNI VVQSVPGSWELPIAVQRLYSASQLQTPSSGPSLSAGDLLGSSTTDLTALPTT TASSTGPFDALIAIGVLIKGETMHFEYIADSVSHGLMRVQLDTGVPVIFGVL TVLTDDQAKARAGVIEGSHNHGEDWGLAAVEMGVRRRDWAAGKTE	1c41 pentamer
I52-03B	MYEVDHADVYDLFYLGRGKDYAAEASDIADLVRSRTPEASSLLDVACGTGTH LEHFTKEFGDTAGLELSEDMLTHARKRLPDATLHQGDMRDFQLGRKFSAVVS MFSSVGYLKTVAELGAAVASFAEHLEPGGVVVVEPWWFPETFADGWVSADVV RRDGRTVARVSHSVREGNATRMEVHFTVADPGKGVRHFSDVHLITLFHQREY EAAFMAAGLRVEYLEGGPSGRGLFVGVPALEHHHHHH	3bxo dimer
I52-32A	MGMKEKFVLIITHGDFGKGLLSGAEVIIGKQENVHTVGLNLGDNIEKVAKEV MRIIIAKLAEDKEIIIVVDLFGGSPFNIALEMMKTFDVKVITGINMPMLVEL LTSINVYDTTELLENISKIGKDGIKVIEKSSLKM	3lfh dimer

I52-32B	MKYDGSKLRIGILHARWNLEIIAALVAGAIKRLQEFGVKAENIIIETVPGSF ELPYGSKLFVEKQKRLGKPLDAIIPIGVLIKGSTMHFEYICDSTTHQLMKLN FELGIPVIFGVLTCLTDEQAEARAGLIEGKMHNHGEDWGAAAVEMATKFNLE HHHHHH	2jfb pentamer
I52-33A	MAVKGLGEVDQKYDGSKLRIGILHARWNRKIILALVAGAVLRLLEFGVKAEN IIIETVPGSFELPYGSKLFVEKQKRLGKPLDAIIPIGVLIKGSTMHFEYICD STTHQLMKLNFELGIPVIFGVLTCLTDEQAEARAGLIEGKMHNHGEDWGAAA VEMATKFN	2jfb pentamer
I52-33B	MGANWYLDNESSRLSFTSTKNADIAEVHRFLVLHGKVDPKGLAEVEVETESI STGIPLRDMLLRVLVFQVSKFPVAQINAQLDMRPINNLAPGAQLELRLPLTV SLRGKSHSYNAELLATRLDERRFQVVTLEPLVIHAQDFDMVRAFNALRLVAG LSAVSLSVPVGAVLIFTARLEHHHHHH	3q34 dimer
I32-06A	MGSHHHHHHGMTDYIRDGSAIKALSFAIILAEADLRHIPQDLQRLAVRVIHA CGMVDVANDLAFSEGAGKAGRNALLAGAPILCDARMVAEGITRSRLPADNRV IYTLSDPSVPELAKKIGNTRSAAALDLWLPHIEGSIVAIGNAPTALFRLFEL LDAGAPKPALIIGMPVGFVGAAESKDELAANSRGVPYVIVRGRRGGSAMTAA AVNALASERE	3e7d dimer
I32-06B	MITVFGLKSKLAPRREKLAEVIYSSLHLGLDIPKGKHAIRFLCLEKEDFYYP FDRSDDYTVIEINLMAGRSEETKMLLIFLLFIALERKLGIRAHDVEITIKEQ PAHCWGFRGRTGDSARDLDYDIYV	1mww trimer
I32-10A	MEMDIRFRGDDLEALLKAAIMMIKAALKMGATITLSLDGNDLEIRITGVPEA ARKALATIAEVLAKTFGITVTRTIR	yl21 dimer
I32-10B	MDSMDHRIERLEYYIQLLVKTVDMDRYPFYALLIDKGLSKEEGESVMRICQA LSVALETLKALGQVTFDELLKIFAGALNEKLDVHETIFALYEQGLYQELMEV FIDIMKHFDLEHHHHHH	1sed trimer
I32-19A	MGSDLQKLQRFSTCDISDGLLNVYNIPTGGYFPNLTAISPPQNSSIVGTAYT VLFAPIDDPRPAVNYIDSVPPNSILVLALEPHLQSQFHPFIKITQAMYGGLM STRAQYLKSNGTVVFGRIRDVDEHRTLNHPVFAYGVGSCAPKAVVKAVGTNV	2c5q trimer

	QLKILTSDGVTQTICPGDYIAGDNNGIVRIPVQETDISKLVTYIEKSIEVDR LVSEAIKNGLPAKAAQTARRMVLKDYI	
I32-19B	MSGMRVYLGADHAGYELKQAIIAFLKMTGHEPIDCGALRYDADDDYPAFCIA AATRTVADPGSLGIVLGGSGNGEQIAANKVPGARCALAWSVQTAALAREHNN AQLIGIGGRMHTLEEALRIVKAFVTTPWSKAQRHQRRIDILAEYERTHEAPP VPGAPALEHHHHHH	2vvp dimer
I32-28A	MGDDARIAAIGDVDELNSQIGVLLAEPLPDDVRAALSAIQHDLFDLGGELCI PGHAAITEDHLLRLALWLVHYNGQLPPLEEFILPGGARGAALAHVCRTVCRR AERSIKALGASEPLNIAPAAYVNLLSDLLFVLARVLNRAAGGADVLWDRTRA H	2zhz trimer
I32-28B	MILSAEQSFTLRHPHGQAAALAFVREPAAALAGVQRLRGLDSDGEQVWGELL VRVPLLGEVDLPFRSEIVRTPQGAELRPLTLTGERAWVAVSGQATAAEGGEM AFAFQFQAHLATPEAEGEGGAAFEVMVQAAAGVTLLLVAMALPQGLAAGLPP ALEHHHHHH	3nqn dimer

Supplementary Files

These files contain:

T32-28, T33-09, T33-15, T33-21, and T33-28 design models:

T32_T33_design_models.zip

Example files from the docking protocol utilized in Section 1 and 2:

T32_T33_docking.zip

Example files from the design protocol utilized in Section 1 and 2:

T32_T33_design.zip

References

1. 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.

2. 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.

3. 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.

4. 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.

5. DiMaio F, Leaver-Fay A, Bradley P, Baker D, Andre I (2011) Modeling symmetric macromolecular structures in Rosetta3. PLoS One 6:e20450.

6. 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.

7. Lawrence MS, Phillips KJ, Liu DR (2007) Supercharging proteins can impart unusual resilience. J Am Chem Soc 129:10110-2.

8. 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.

9. Bale JB, Park RU, Liu Y, Gonen S, Gonen T, Cascio D, King NP, Yeates TO, Baker D (2015) Structure of a designed tetrahedral protein assembly variant engineered to have improved soluble expression. Protein Sci 24:1695-701.

10. Caspar DL, Klug A (1962) Physical principles in the construction of regular viruses. Cold Spring Harb Symp Quant Biol 27:1-24.

11. Zandi R, Reguera D, Bruinsma RF, Gelbart WM, Rudnick J (2004) Origin of icosahedral symmetry in viruses. Proc Natl Acad Sci U S A 101:15556-60.

12. Howorka S (2011) Rationally engineering natural protein assemblies in nanobiotechnology. Curr Opin Biotechnol 22:485-91.

13. Douglas T, Young M (2006) Viruses: making friends with old foes. Science 312:873-5.

14. Lai YT, King NP, Yeates TO (2012) Principles for designing ordered protein assemblies. Trends Cell Biol 22:653-61.

15. King NP, Lai YT (2013) Practical approaches to designing novel protein assemblies. Curr Opin Struct Biol 23:632-8.

16. Sinclair JC (2013) Constructing arrays of proteins. Curr Opin Chem Biol 17:946-51.

17. Salgado EN, Radford RJ, Tezcan FA (2010) Metal-directed protein self-assembly. Acc Chem Res 43:661-72.

18. 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.

19. 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.

20. Lai YT, Cascio D, Yeates TO (2012) Structure of a 16-nm cage designed by using protein oligomers. Science 336:1129.

21. 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.

22. 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.

23. 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.

24. Seeman NC (2010) Nanomaterials based on DNA. Annu Rev Biochem 79:65-87.

25. Rothemund PW (2006) Folding DNA to create nanoscale shapes and patterns. Nature 440:297-302.

26. Ke Y, Ong LL, Shih WM, Yin P (2012) Three-dimensional structures self-assembled from DNA bricks. Science 338:1177-83.

27. Han D, Pal S, Yang Y, Jiang S, Nangreave J, Liu Y, Yan H (2013) DNA gridiron nanostructures based on four-arm junctions. Science 339:1412-5.

28. Padilla JE, Colovos C, Yeates TO (2001) Nanohedra: using symmetry to design self assembling protein cages, layers, crystals, and filaments. Proc Natl Acad Sci U S A 98:2217-21.

29. Usui K, Maki T, Ito F, Suenaga A, Kidoaki S, Itoh M, Taiji M, Matsuda T, Hayashizaki Y, Suzuki H (2009) Nanoscale elongating control of the self-assembled protein filament with the cysteine-introduced building blocks. Protein Sci 18:960-9.

30. Goodsell DS, Olson AJ (2000) Structural symmetry and protein function. Annu Rev Biophys Biomol Struct 29:105-53.

31. Janin J, Bahadur RP, Chakrabarti P (2008) Protein-protein interaction and quaternary structure. Q Rev Biophys 41:133-80.

32. Huang PS, Love JJ, Mayo SL (2007) A de novo designed protein protein interface. Protein Sci 16:2770-4.

33. Jha RK, Leaver-Fay A, Yin S, Wu Y, Butterfoss GL, Szyperski T, Dokholyan NV, Kuhlman B (2010) Computational design of a PAK1 binding protein. J Mol Biol 400:257-70.

34. Karanicolas J, Corn JE, Chen I, Joachimiak LA, Dym O, Peck SH, Albeck S, Unger T, Hu W, Liu G and others (2011) A de novo protein binding pair by computational design and directed evolution. Mol Cell 42:250-60.

35. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Wilson IA, Baker D (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 332:816-21.

36. Khare SD, Fleishman SJ (2013) Emerging themes in the computational design of novel enzymes and protein-protein interfaces. FEBS Lett 587:1147-54.

37. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 97:10383-8.

38. 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.

39. Lawrence MC, Colman PM (1993) Shape complementarity at protein/protein interfaces. J Mol Biol 234:946-50.

40. Arnold FH, Volkov AA (1999) Directed evolution of biocatalysts. Curr Opin Chem Biol 3:54-9.

41. Jackel C, Kast P, Hilvert D (2008) Protein design by directed evolution. Annu Rev Biophys 37:153-73.

42. Worsdorfer B, Pianowski Z, Hilvert D (2012) Efficient in vitro encapsulation of protein cargo by an engineered protein container. J Am Chem Soc 134:909-11.

43. Worsdorfer B, Woycechowsky KJ, Hilvert D (2011) Directed evolution of a protein container. Science 331:589-92.

44. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372:774-97.

45. The PyMOL Molecular Graphics System. 1.5: Schrödinger, LLC 2012.

46. Fleishman SJ, Whitehead TA, Strauch EM, Corn JE, Qin S, Zhou HX, Mitchell JC, Demerdash ON, Takeda-Shitaka M, Terashi G and others (2011) Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol 414:289-302.

47. Fleishman SJ, Leaver-Fay A, Corn JE, Strauch EM, Khare SD, Koga N, Ashworth J, Murphy P, Richter F, Lemmon G and others (2011) RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. PLoS One 6:e20161.

48. Fleishman SJ, Khare SD, Koga N, Baker D (2011) Restricted sidechain plasticity in the structures of native proteins and complexes. Protein Sci 20:753-7.

49. Grueninger D, Treiber N, Ziegler MO, Koetter JW, Schulze MS, Schulz GE (2008) Designed protein-protein association. Science 319:206-9.

50. Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL and others (2013) Computational design of ligand-binding proteins with high affinity and selectivity. Nature 501:212-6.

51. Nivon LG, Bjelic S, King C, Baker D (2013) Automating human intuition for protein design. Proteins

52. Sheffler W, Baker D (2010) RosettaHoles2: a volumetric packing measure for protein structure refinement and validation. Protein Sci 19:1991-5.

53. Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R, Song Y, Kellogg EH, Thompson J, Davis IW, Pache RA, Lyskov S and others (2013) Scientific benchmarks for guiding macromolecular energy function improvement. Methods Enzymol 523:109-43.

54. Prodromou C, Pearl LH (1992) Recursive PCR: a novel technique for total gene synthesis. Protein Eng 5:827-9.

55. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, 3rd, Smith HO (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods 6:343-5.

56. Zhou Z, Cironi P, Lin AJ, Xu Y, Hrvatin S, Golan DE, Silver PA, Walsh CT, Yin J (2007) Genetically encoded short peptide tags for orthogonal protein labeling by Sfp and AcpS phosphopantetheinyl transferases. ACS Chem Biol 2:337-46.

57. Chan AC, Doukov TI, Scofield M, Tom-Yew SA, Ramin AB, Mackichan JK, Gaynor EC, Murphy ME (2010) Structure and function of P19, a high-affinity iron transporter of the human pathogen Campylobacter jejuni. J Mol Biol 401:590-604.

58. 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.

59. 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.

60. Smith JM (1999) Ximdisp--A visualization tool to aid structure determination from electron microscope images. J Struct Biol 125:223-8.

61. Ludtke SJ, Baldwin PR, Chiu W (1999) EMAN: semiautomated software for high-resolution single-particle reconstructions. J Struct Biol 128:82-97.

62. Frank J, Radermacher M, Penczek P, Zhu J, Li Y, Ladjadj M, Leith A (1996) SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. J Struct Biol 116:190-9.

63. van Heel M, Harauz G, Orlova EV, Schmidt R, Schatz M (1996) A new generation of the IMAGIC image processing system. J Struct Biol 116:17-24.

64. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25:1605-12.

65. Kabsch W (2010) Xds. Acta Crystallogr D Biol Crystallogr 66:125-32.

66. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. J Appl Crystallogr 40:658-674.

67. 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.

68. Painter J, Merritt EA (2006) Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. Acta Crystallogr D Biol Crystallogr 62:439-50.

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

70. Painter J, Merritt EA (2006) TLSMD web server for the generation of multi-group TLS models. Journal of Applied Crystallography 39:109-111.

71. Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 66:12-21.

72. Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 2:1511-9.

73. Luthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356:83-5.

74. Bradley P, Baker D (2006) Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. Proteins 65:922-9.

75. 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.

76. 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.

77. 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.

78. Stephanopoulos N, Francis MB (2011) Choosing an effective protein bioconjugation strategy. Nat Chem Biol 7:876-84.

79. Spicer CD, Davis BG (2014) Selective chemical protein modification. Nat Commun 5:4740.

80. Badran AH, Liu DR (2015) In vivo continuous directed evolution. Curr Opin Chem Biol 24:1-10.

81. 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.

82. Song WJ, Tezcan FA (2014) A designed supramolecular protein assembly with in vivo enzymatic activity. Science 346:1525-8.

83. 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.

84. Bricogne G (1997) [23] Bayesian statistical viewpoint on structure determination: Basic concepts and examples. 276:361-423.

85. 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.

86. 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.

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

88. Ringler P, Schulz GE (2003) Self-assembly of proteins into designed networks. Science 302:106-9.

89. Gonen S, DiMaio F, Gonen T, Baker D (2015) Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. Science 348:1365-8.

90. Lin T, Chen Z, Usha R, Stauffacher CV, Dai JB, Schmidt T, Johnson JE (1999) The refined crystal structure of cowpea mosaic virus at 2.8 A resolution. Virology 265:20-34.

91. Hura GL, Menon AL, Hammel M, Rambo RP, Poole FL, 2nd, Tsutakawa SE, Jenney FE, Jr., Classen S, Frankel KA, Hopkins RC and others (2009) Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nat Methods 6:606-12.

92. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A (2013) Accurate SAXS profile computation and its assessment by contrast variation experiments. Biophys J 105:962-74.

93. Schneidman-Duhovny D, Hammel M, Sali A (2010) FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic Acids Res 38:W540-4.