doi:10.1038/nature13404

Accurate design of co-assembling multi-component protein nanomaterials

Neil P. King[1,2]*, Jacob B. Bale[1,3]*, William Sheffler[1]*, Dan E. McNamara[4], Shane Gonen[1,5], Tamir Gonen[5], Todd O. Yeates[4,6,7] & David Baker[1,2,8]

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 and variety of two-component materials that it makes accessible suggest a route to the construction of functional protein nanomaterials tailored to specific applications.

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[1,2], 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[3–6]. 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[6–16]. 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[17–20]. 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[3,4,8,11,21], 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 construction from combinations of building blocks, greater control over the timing of assembly, and enhanced modularity through independently addressable subunits. Although multi-component protein assemblies have recently been generated using disulphide bonds[14,22], flexible genetic linkers[11,15,22], or stereotyped coiled-coil interactions to drive assembly[14,15], 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[23,24]. Advances in computational protein modelling and design algorithms have recently made it possible to design such interfaces[25–29] and thereby direct the formation of novel self-assembling protein nanomaterials with atomic-level accuracy[7,9,10], 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 Fig. 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 (Fig. 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 (Fig. 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 (Fig. 1c). The score function used during docking favours interfaces with high densities of

[1]Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA. [2]Institute for Protein Design, University of Washington, Seattle, Washington 98195, USA. [3]Graduate Program in Molecular and Cellular Biology, University of Washington, Seattle, Washington 98195, USA. [4]UCLA Department of Chemistry and Biochemistry, Los Angeles, California 90095, USA. [5]Janelia Farm Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, Virginia 20147, USA. [6]UCLA-DOE Institute for Genomics and Proteomics, Los Angeles, California 90095, USA. [7]UCLA Molecular Biology Institute, Los Angeles, California 90095, USA. [8]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA.

*These authors contributed equally to this work.

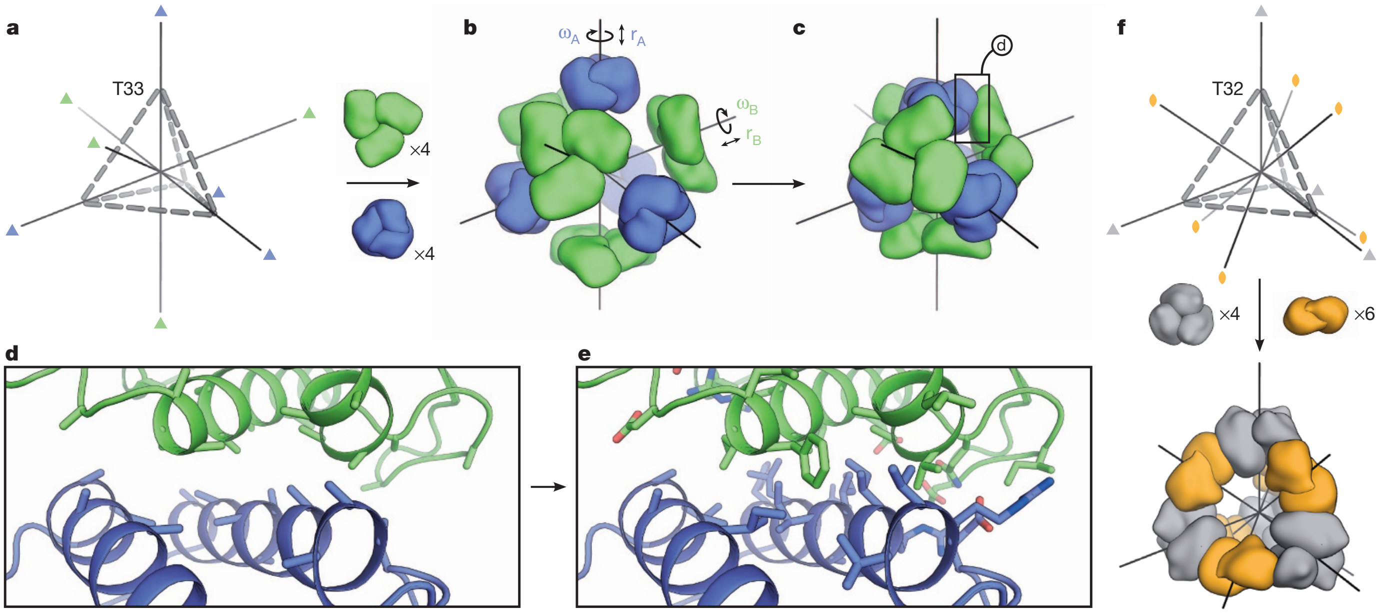


Figure 1 | Overview of the computational design method. a, The T33 architecture comprises four copies each of two distinct trimeric building blocks (green and blue) arranged with tetrahedral point group symmetry (24 total subunits; triangles indicate three-fold symmetry axes). **b**, Each building block 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.

contacting residues in well-anchored regions of the protein structure that are less likely to change conformation on mutation of surface side chains (Fig. 1d). RosettaDesign[30,31] 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[24] (Fig. 1e). The end result is a pair of new amino acid sequences, one for each building block, predicted to stabilize the modelled interface and drive assembly to the specific target configuration.

These docking and design procedures were implemented by extending the Rosetta software[31,32] 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 (Extended Data Fig. 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[3,4,21]. Combining even two types of symmetry elements (as in the present study) can give rise to a large number of distinct symmetric architectures with

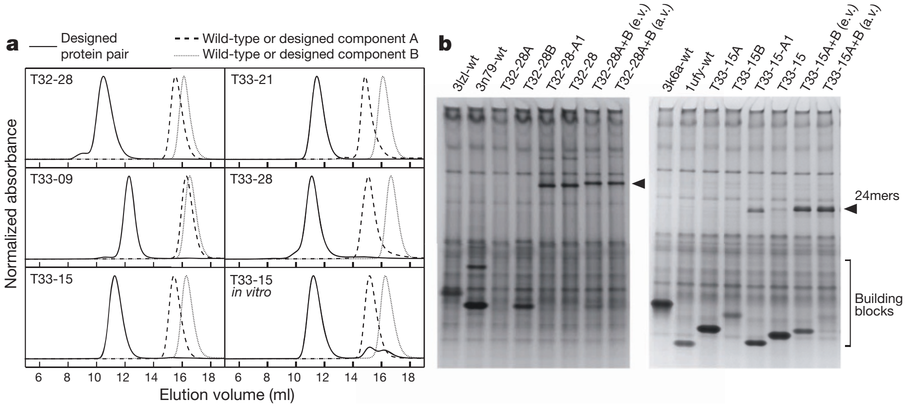


Figure 2 | Experimental characterization of co-assembly. a, SEC chromatograms of the designed pairs of proteins (solid lines) and the wild-type oligomeric proteins from which they were derived (dashed and dotted lines). The co-expressed designed proteins elute at the volumes expected for the target 24-subunit 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 slowly migrating species (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).

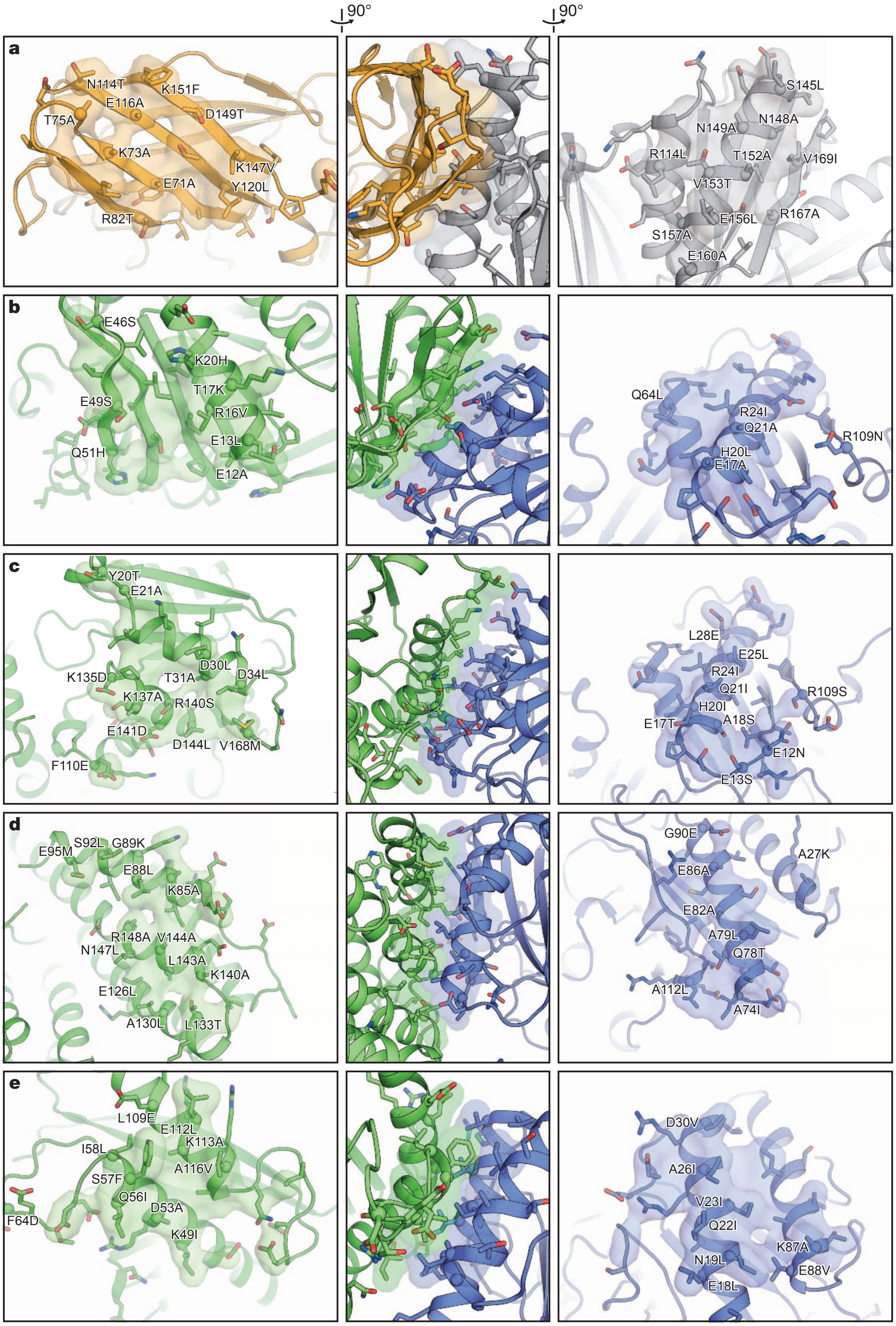


Figure 3 | Designed interfaces of two-component protein nanomaterials. The models of the designed interfaces in each component (left, 'A' component; right, 'B' component) of T32-28 (**a**), T33-09 (**b**), T33-15 (**c**), T33-21 (**d**) and T33-28 (**e**) are shown, and side views of each interface as 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.

a range of possible morphologies, including those with dihedral and cubic point-group symmetries, as well as helical, layer and space group symmetries (ref. 21 and T.O.Y., manuscript in preparation).

In this study we targeted two distinct tetrahedral architectures: the T33 architecture described above and the T32 architecture shown in Fig. 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[33] 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 (Extended Data Fig. 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 (Extended Data Fig. 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 (Fig. 2a and Supplementary Table 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 (Extended Data Fig. 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 (Fig. 2b, Extended Data Fig. 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 (Fig. 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, highlighting the shape and chemical complementarity generated by the many amino acid mutations introduced during design, are presented in Fig. 3. Qualitatively, the interfaces reflect the hypothesis underlying the design protocol: 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. The successful designs are quantitatively similar to the other designs 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; Extended Data Fig. 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.

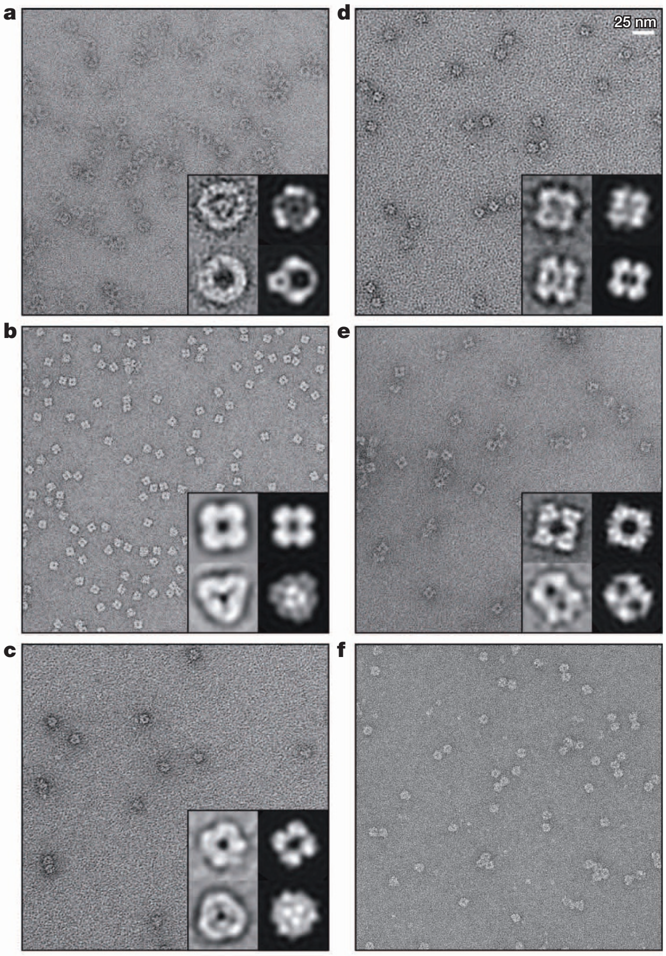


Figure 4 | Electron micrographs of designed two-component protein nanomaterials. Negative-stain electron micrographs for co-expressed and purified T32-28 (**a**), T33-09 (**b**), T33-15 (**c**), T33-21 (**d**) and T33-28 (**e**) are shown to scale (scale bar at top right, 25 nm). For each co-expressed 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 Fig. 2). Micrographs of unpurified, *in vitro*-assembled T33-15 as well as T33-15A and T33-15B in isolation are shown in Extended Data Fig. 5.

Structural characterization of the designed materials

Negative-stain electron microscopy of the five designed materials confirmed that they assemble specifically to the target architectures (Fig. 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 (Fig. 4b, inset). Micrographs of T33-15 assembled *in vitro* as described above were indistinguishable from those of co-expressed T33-15 (Fig. 4c, f and Extended Data Fig. 5c), 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 (Fig. 5 and Supplementary Tables 2 and 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

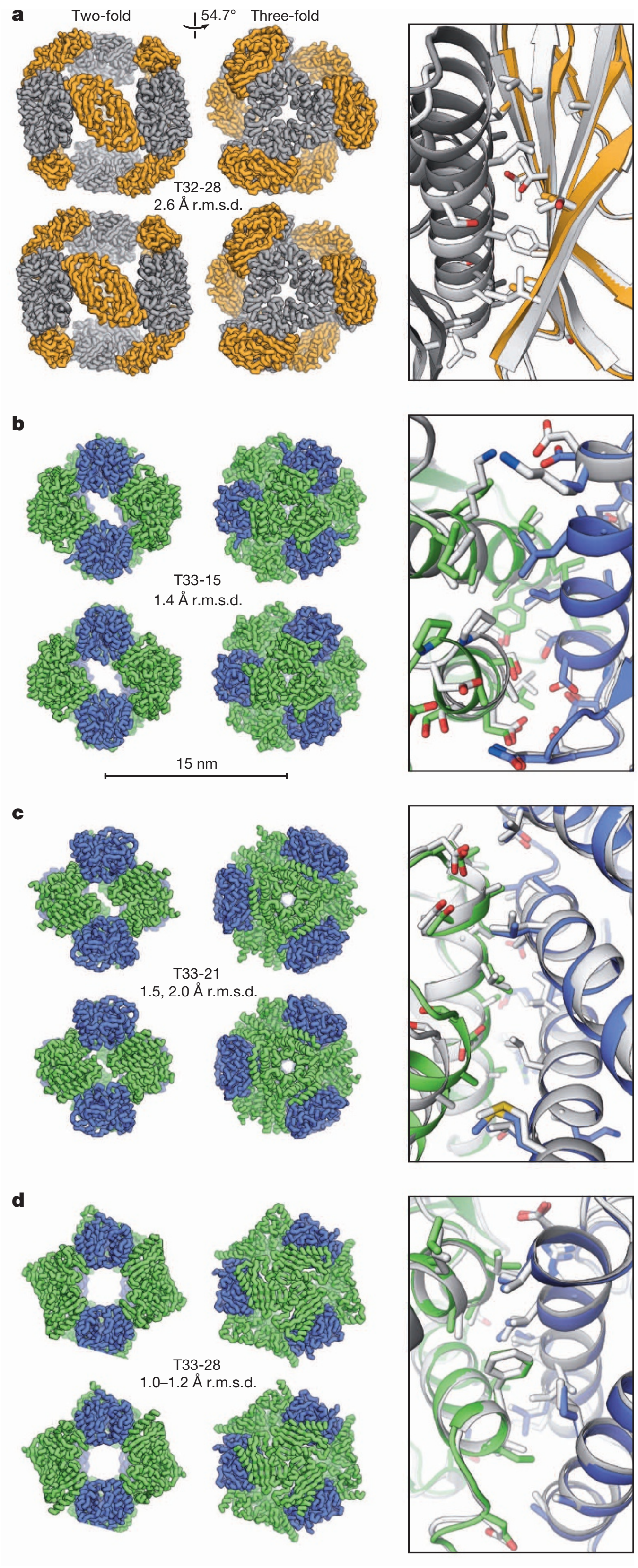


Figure 5 | Crystal structures of designed two-component protein nanomaterials. The computational design models (top) and X-ray crystal structures (bottom) are shown at left for T32-28 (**a**), T33-15 (**b**), T33-21 (**c**) and T33-28 (**d**). Views of each material are shown to scale along the two-fold 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.

deviations (r.m.s.d.) between 0.5 and 1.2 Å (Fig. 5 right and Extended Data Table 1). 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 (Supplementary Tables 5 and 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 (Fig. 5 left and Extended Data Table 1). 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[34,35] and can be used to improve, for example, the packaging capability of existing protein nanocontainers[36,37], 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. Multicomponent materials are advantageous in these extended architectures because the uncontrolled self-assembly of a single-component material inside the cell can complicate biological production[5,11,21]. 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.

METHODS SUMMARY

The symmetric modelling framework in Rosetta[31,32] was updated to enable the modelling of multi-component symmetrical structures. A new application, tcdock, docks pairs of protein scaffolds in higher-order symmetries, scoring each docked configuration according to its suitability for interface design. tcdock was used to dock all possible pairwise combinations of 200 trimeric scaffold proteins and all possible pairwise combinations of the same trimers and 1,161 dimeric proteins in the T33 and T32 symmetric architectures, respectively. New two-component protein–protein interface design protocols were used to design new amino acid sequences predicted to stabilize selected docked configurations. During the sequence design protocols, the symmetric rigid body degrees of freedom and the identities and conformations of the side chains at the inter-building-block interfaces were optimized to identify low-energy sequence-structure combinations. Thirty T32 and 27 T33 designs were selected for experimental characterization.

The assembly states of the designed pairs of proteins were assessed by native PAGE, and those that migrated more slowly than the wild-type scaffolds were subjected to affinity purification and SEC. The ability of the materials to assemble *in vitro* was investigated by independently producing the two components, mixing them at various points after cell lysis, and analysing the mixtures by native PAGE and SEC. The materials were structurally characterized by negative-stain electron microscopy and particle averaging, and at high resolution by X-ray crystallography.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 January; accepted 25 April 2014.

Published online 25 May 2014.

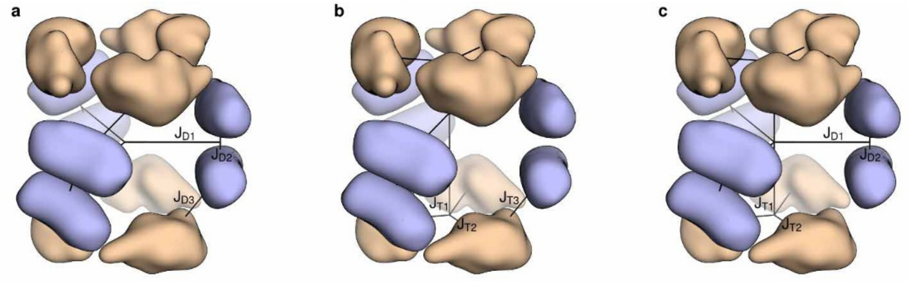
1. Howorka, S. Rationally engineering natural protein assemblies in nanobiotechnology. *Curr. Opin. Biotechnol.* **22,** 485–491 (2011).
2. Douglas, T. & Young, M. Viruses: making friends with old foes. *Science* **312,** 873–875 (2006).
3. Lai, Y. T., King, N. P. & Yeates, T. O. Principles for designing ordered protein assemblies. *Trends Cell Biol.* **22,** 653–661 (2012).
4. King, N. P. & Lai, Y. T. Practical approaches to designing novel protein assemblies. *Curr. Opin. Struct. Biol.* **23,** 632–638 (2013).
5. Sinclair, J. C. Constructing arrays of proteins. *Curr. Opin. Chem. Biol.* **17,** 946–951 (2013).
6. Salgado, E. N., Radford, R. J. & Tezcan, F. A. Metal-directed protein self-assembly. *Acc. Chem. Res.* **43,** 661–672 (2010).
7. King, N. P. *et al.* Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **336,** 1171–1174 (2012).
8. Brodin, J. D. *et al.* Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays. *Nature Chem.* **4,** 375–382 (2012).
9. Lanci, C. J. *et al.* Computational design of a protein crystal. *Proc. Natl Acad. Sci. USA* **109,** 7304–7309 (2012).
10. Stranges, P. B., Machius, M., Miley, M. J., Tripathy, A. & Kuhlman, B. Computational design of a symmetric homodimer using beta-strand assembly. *Proc. Natl Acad. Sci. USA* **108,** 20562–20567 (2011).
11. Sinclair, J. C., Davies, K. M., Venien-Bryan, C. & Noble, M. E. Generation of protein lattices by fusing proteins with matching rotational symmetry. *Nature Nanotechnol.* **6,** 558–562 (2011).
12. Lai, Y. T., Cascio, D. & Yeates, T. O. Structure of a 16-nm cage designed by using protein oligomers. *Science* **336,** 1129 (2012).
13. Der, B. S. *et al.* Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. *J. Am. Chem. Soc.* **134,** 375–385 (2012).
14. Fletcher, J. M. *et al.* Self-assembling cages from coiled-coil peptide modules. *Science* **340,** 595–599 (2013).
15. Boyle, A. L. *et al.* Squaring the circle in peptide assembly: from fibers to discrete nanostructures by de novo design. *J. Am. Chem. Soc.* **134,** 15457–15467 (2012).
16. Grigoryan, G. *et al.* Computational design of virus-like protein assemblies on carbon nanotube surfaces. *Science* **332,** 1071–1076 (2011).
17. Seeman, N. C. Nanomaterials based on DNA. *Annu. Rev. Biochem.* **79,** 65–87 (2010).
18. Rothemund, P. W. Folding DNA to create nanoscale shapes and patterns. *Nature* **440,** 297–302 (2006).
19. Ke, Y., Ong, L. L., Shih, W. M. & Yin, P. Three-dimensional structures self-assembled from DNA bricks. *Science* **338,** 1177–1183 (2012).
20. Han, D. *et al.* DNA gridiron nanostructures based on four-arm junctions. *Science* **339,** 1412–1415 (2013).
21. Padilla, J. E., Colovos, C. & Yeates, T. O. Nanohedra: using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proc. Natl Acad. Sci. USA* **98,** 2217–2221 (2001).
22. Usui, K. *et al.* Nanoscale elongating control of the self-assembled protein filament with the cysteine-introduced building blocks. *Protein Sci.* **18,** 960–969 (2009).
23. Goodsell, D. S. & Olson, A. J. Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 105–153 (2000).
24. Janin, J., Bahadur, R. P. & Chakrabarti, P. Protein-protein interaction and quaternary structure. *Q. Rev. Biophys.* **41,** 133–180 (2008).
25. Huang, P. S., Love, J. J. & Mayo, S. L. A de novo designed protein protein interface. *Protein Sci.* **16,** 2770–2774 (2007).
26. Jha, R. K. *et al.* Computational design of a PAK1 binding protein. *J. Mol. Biol.* **400,** 257–270 (2010).
27. Karanicolas, J. *et al.* A de novo protein binding pair by computational design and directed evolution. *Mol. Cell* **42,** 250–260 (2011).
28. Fleishman, S. J. *et al.* Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332,** 816–821 (2011).
29. Khare, S. D. & Fleishman, S. J. Emerging themes in the computational design of novel enzymes and protein-protein interfaces. *FEBS Lett.* **587,** 1147–1154 (2013).
30. Kuhlman, B. & Baker, D. Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA* **97,** 10383–10388 (2000).
31. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487,** 545–574 (2011).
32. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D. & Andre, I. Modeling symmetric macromolecular structures in Rosetta3. *PLoS ONE* **6,** e20450 (2011).
33. Lawrence, M. C. & Colman, P. M. Shape complementarity at protein/protein interfaces. *J. Mol. Biol.* **234,** 946–950 (1993).
34. Arnold, F. H. & Volkov, A. A. Directed evolution of biocatalysts. *Curr. Opin. Chem. Biol.* **3,** 54–59 (1999).
35. Jäckel, C., Kast, P. & Hilvert, D. Protein design by directed evolution. *Annu. Rev. Biophys.* **37,** 153–173 (2008).
36. Wörsdörfer, B., Pianowski, Z. & Hilvert, D. Efficient *in vitro* encapsulation of protein cargo by an engineered protein container. *J. Am. Chem. Soc.* **134,** 909–911 (2012).
37. Wörsdörfer, B., Woycechowsky, K. J. & Hilvert, D. Directed evolution of a protein container. *Science* **331,** 589–592 (2011).
38. Bradley, P. & Baker, D. Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. *Proteins* **65,** 922–929 (2006).

Supplementary Information is available in the online version of the paper.

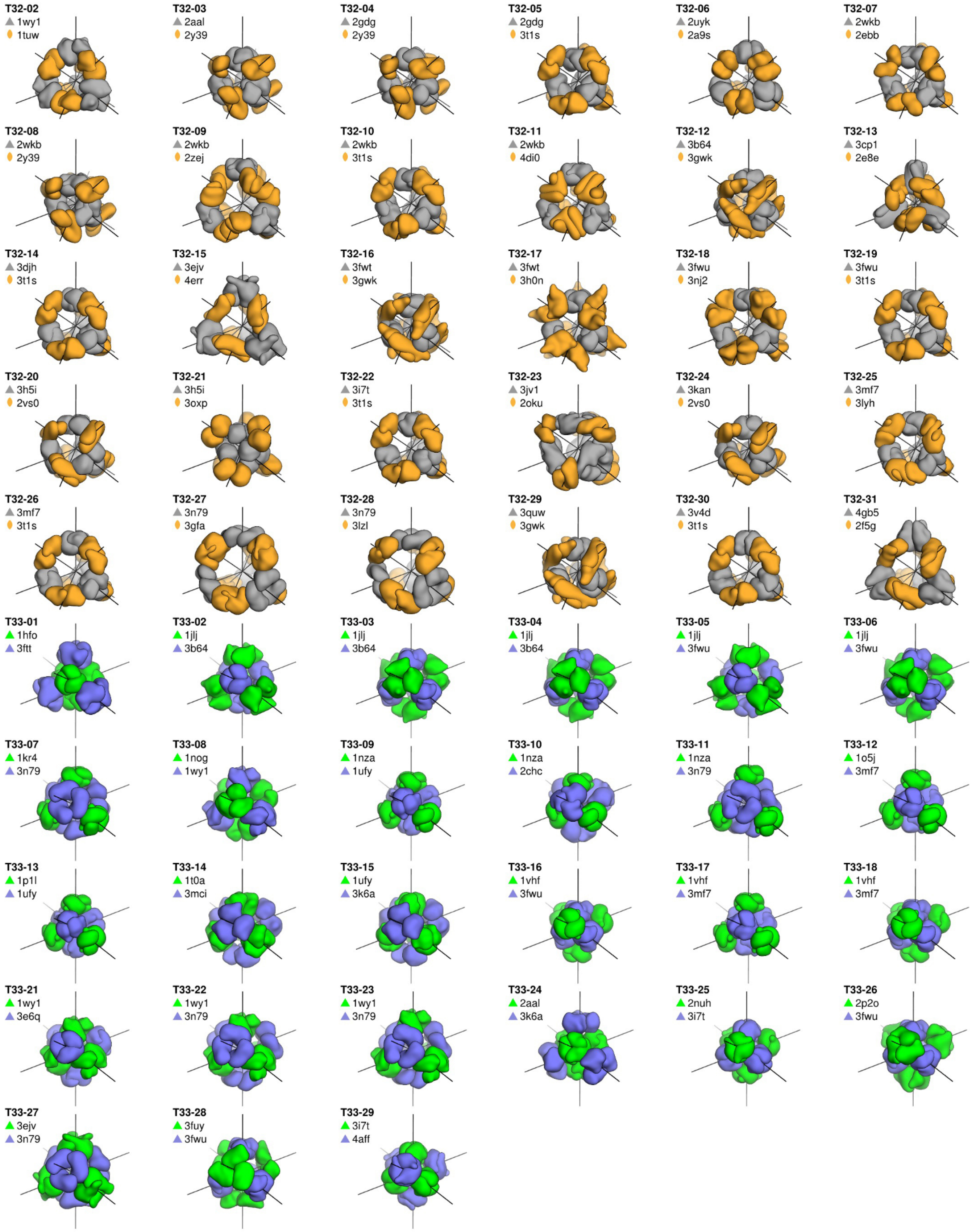
Acknowledgements We thank D. Shi and B. Nannenga (JFRC) for help with electron microscopy, F. DiMaio and R. Moretti for assistance with software development, P. Greisen for scripts used to compare side-chain conformations, J. Gallaher for technical assistance, M. Collazo for help with preliminary crystallization screening, D. Cascio and M. Sawaya for help with crystallographic experiments, and M. Capel, J. Schuermann and I. Kourinov at NE-CAT beamline 24-ID-C for help with data collection. This work was supported by the Howard Hughes Medical Institute (T.G. and D.B.) and the JFRC visitor program (S.G.), the National Science Foundation under CHE-1332907 (D.B. and T.O.Y.), grants from the International AIDS Vaccine Initiative, DTRA (N00024-10-D-6318/0024), AFOSR (FA950-12-10112) and DOE (DE-SC0005155) to D.B., an NIH Biotechnology Training Program award to D.E.M. (T32GM067555) and an NSF graduate research fellowship to J.B.B. (DGE-0718124). T.O.Y. and D.E.M. also acknowledge support from the BER programme of the DOE Office of Science. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding bodies.

Author Contributions N.P.K., J.B.B., W.S. and D.B. designed the research. N.P.K., J.B.B. and W.S. wrote program code and performed the docking and design calculations. N.P.K. and J.B.B. biophysically characterized the designed materials and prepared samples for structural analysis. S.G. characterized the designed materials by electron microscopy; S.G. and T.G. analysed electron microscopy data. D.E.M. crystallized the designed protein materials; D.E.M. and T.O.Y. analysed crystallographic data. N.P.K., J.B.B. and D.B. analysed data and wrote the manuscript. All authors discussed the results and commented on the manuscript.

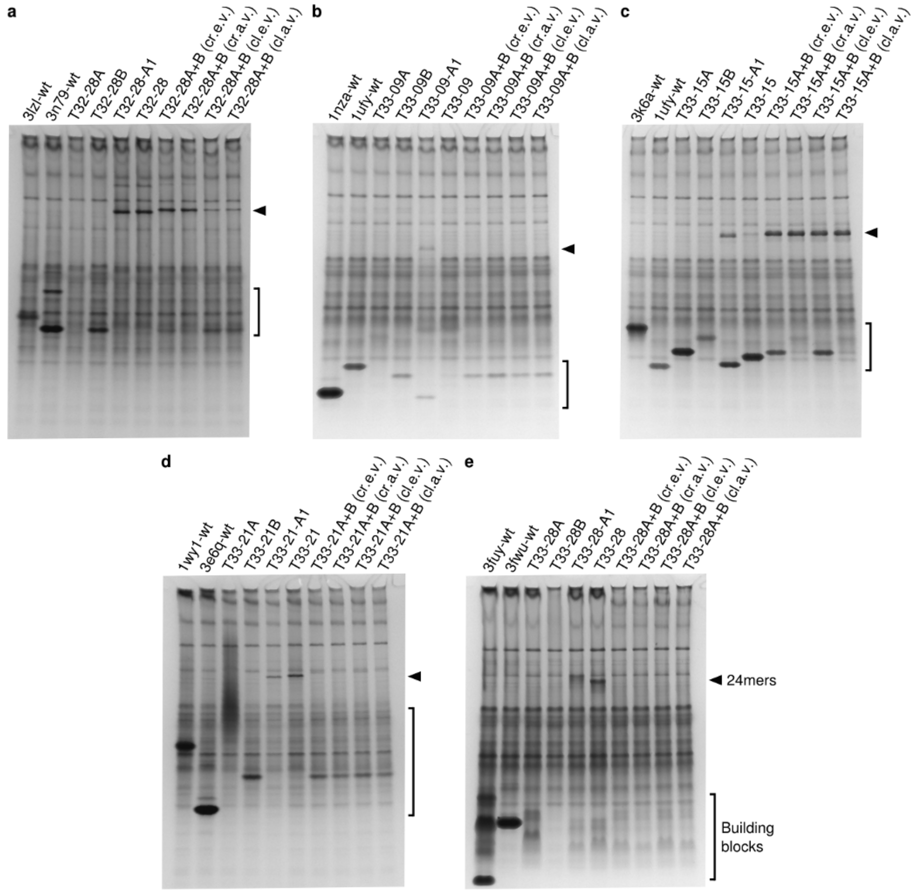
Author Information The crystal structures and structure factors for the designed materials have been deposited in the RCSB Protein Data Bank (http://www.rcsb.org/) under the accession codes 4NWN (T32-28), 4NWO (T33-15), 4NWP (T33-21, *R*32 crystal form), 4NWQ (T33-21, *F*$4_1$32 crystal form) and 4NWR (T33-28). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.B. (dabaker@u.washington.edu).



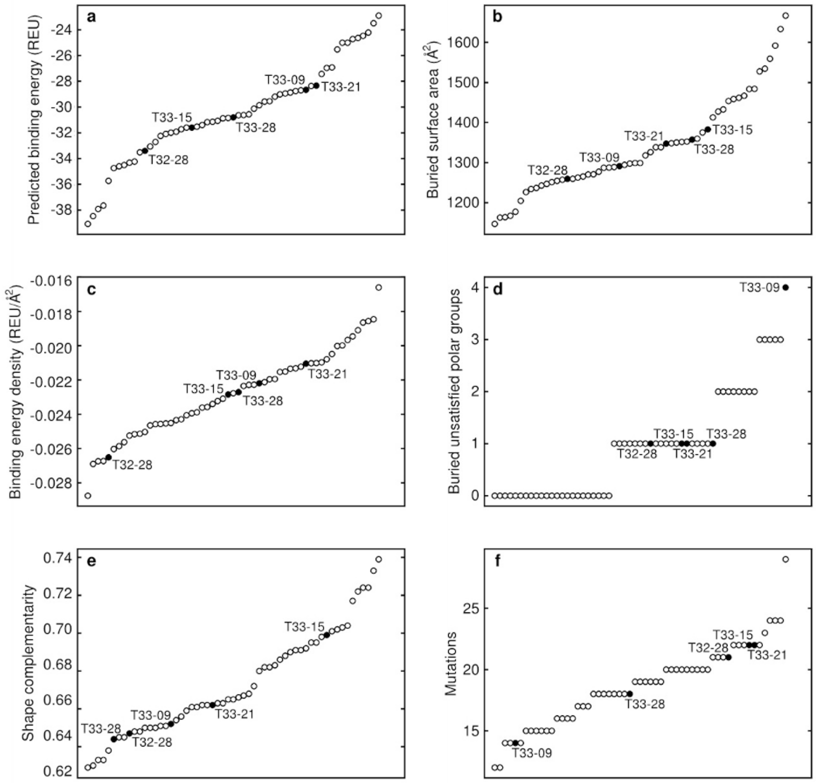
Extended Data Figure 1 | Comparison of one-component and multi-component symmetric fold trees. Within the Rosetta macromolecular modelling suite, the connections between residues in a protein structure are represented as a directed, acyclic, graph referred to as a 'fold tree'[32,38]. When modelling multiple 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 Supplementary Methods for additional discussion.



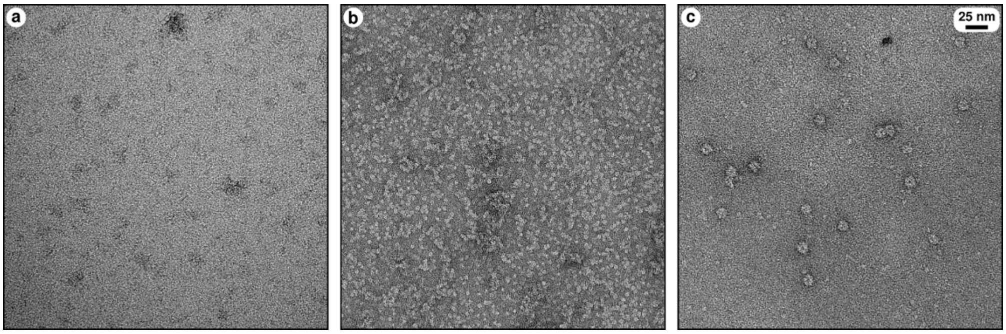
Extended Data Figure 2 | Models of the 57 designs selected for experimental characterization. Smoothed surface representations of each of the 30 T32 and 27 T33 designs are shown. The trimeric component of each T32 design is shown in grey and the dimeric component in orange. The two different trimeric 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.



Extended Data Figure 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.



Extended Data Figure 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.



Extended Data Figure 5 | Electron micrographs of *in vitro*-assembled T33-15. Negative stain micrographs of independently purified T33-15A (**a**) and T33-15B (**b**), as well as unpurified, *in vitro*-assembled T33-15 (**c**) are shown to scale (scale bar at right, 25 nm).

Extended Data Table 1 | 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 backbone r.m.s.d. was calculated over all 24 subunits of each design model and corresponding subunits in each crystal structure.

[†] Two-chain backbone 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.