Rapid and automated design of two-component protein nanomaterials using ProteinMPNN

Robbert J. de Haas[a] (ID), Natalie Brunette[b,c] (ID), Alex Goodson[b,c] (ID), Justas Dauparas[b,c] (ID), Sue Y. Yi[b,c], Erin C. Yang[b,c], Quinton Dowling[b,c] (ID), Hannah Nguyen[b,c] (ID), Alex Kang[b,c] (ID), Asim K. Bera[b,c], Banumathi Sankaran[d], Renko de Vries[a], David Baker[b,c,e] (ID), and Neil P. King[b,c,1] (ID)

The design of protein–protein interfaces using physics-based design methods such as Rosetta requires substantial computational resources and manual refinement by expert structural biologists. Deep learning methods promise to simplify protein–protein interface design and enable its application to a wide variety of problems by researchers from various scientific disciplines. Here, we test the ability of a deep learning method for protein sequence design, ProteinMPNN, to design two-component tetrahedral protein nanomaterials and benchmark its performance against Rosetta. ProteinMPNN had a similar success rate to Rosetta, yielding 13 new experimentally confirmed assemblies, but required orders of magnitude less computation and no manual refinement. The interfaces designed by ProteinMPNN were substantially more polar than those designed by Rosetta, which facilitated in vitro assembly of the designed nanomaterials from independently purified components. Crystal structures of several of the assemblies confirmed the accuracy of the design method at high resolution. Our results showcase the potential of deep learning–based methods to unlock the widespread application of designed protein–protein interfaces and self-assembling protein nanomaterials in biotechnology.

protein design | ProteinMPNN | nanomaterials

Deep learning has revolutionized the field of protein design. Typical design paradigms require three fundamental steps: backbone generation, amino acid sequence design, and structure prediction to evaluate the quality of the designed sequences. Deep learning structure prediction methods such as trRosetta (1), RoseTTAFold (2), AlphaFold2 (3), and ESMfold (4) quickly and accurately generate models of proteins and protein complexes from amino acid sequences. Methods for de novo backbone generation such as hallucination, inpainting, and diffusion have significantly enhanced robustness and versatility compared to previous approaches and have been used to design de novo protein monomers, homooligomers, proteins bearing functional motifs, and protein- and DNA-binding proteins (5–11). Likewise, deep learning methods for sequence design such as ABACUS-R (12), proteinGAN (13), GVP-GNN (14, 15), and ProteinMPNN (16, 17) have demonstrated exceptional performance both in silico and in experimentally characterized proteins, especially monomers and homooligomers (18). In the only reported side-by-side comparisons to date, ProteinMPNN substantially outperformed Rosetta in the design of de novo protein homooligomers (16) and binders (19).

In nature, many protein assemblies with sophisticated functions are constructed from multiple distinct protein subunits or oligomers, which has motivated the development of methods for designing such assemblies for biotechnological applications (20–23). Rosetta has been a powerful tool for achieving this through protein–protein interface design, but the designed interfaces often rely primarily on hydrophobic packing and require significant manual intervention during the design process to eliminate unnecessary mutations (20, 24–27). While hydrophobic packing provides a strong driving force for assembly, it also tends to make the unassembled protein building blocks prone to aggregation, which can complicate their manufacture. By contrast, the interfaces in naturally occurring hierarchically structured protein complexes often include a higher fraction of polar residues, which maximizes assembly fidelity by minimizing off-target aggregation (28–30). Methods capable of designing custom multicomponent protein assemblies with native-like interfaces would promote the development of new protein-based technologies. For example, a licensed protein nanoparticle vaccine for SARS-CoV-2 (31) uses a variant of the computationally designed two-component icosahedral complex I53-50 that was engineered specifically to enable independent purification and in vitro assembly of the two building blocks (25, 28), a feature that was critical for commercial-scale manufacturing.

Here, we explore the design of multicomponent protein assemblies using ProteinMPNN and establish a fully automated design method that generates novel nanomaterials with high efficiency and accuracy. We benchmark its performance against

Significance

We address the current limitations in design of protein–protein interfaces by employing ProteinMPNN, a deep learning method, to create tetrahedral two-component protein nanomaterials that outperform the established physics-based Rosetta design method in terms of computational efficiency and manual refinement. Importantly, the interfaces designed by ProteinMPNN exhibit enhanced polarity, facilitating the seamless assembly of nanomaterials in vitro, which is crucial for efficient biotechnological manufacturing. Our findings demonstrate the potential of deep learning in democratizing protein interface design and showcases the potential of advanced AI methods in speeding up the development of the next generation of protein-based technologies.

Rosetta-based design and find that ProteinMPNN generates interfaces with a higher fraction of polar residues, which in several cases yields oligomeric building blocks with favorable solution properties.

Results

To directly compare the two design methods, we used ProteinMPNN to generate new amino acid sequences for 27 tetrahedral protein assemblies that were previously designed using Rosetta (20). These assemblies comprise four copies each of two distinct trimeric building blocks, arranged on opposing poles of the three-fold axes of tetrahedral point group symmetry (the "T33" architecture; Fig. 1). In the original publication (20), four of the 27 previously designed complexes successfully adopted the target architecture. Since then, negatively stained electron micrographs of one additional complex, T33-23, revealed monodisperse tetrahedral assemblies of the expected size and morphology following purification (*SI Appendix*, Fig. S1).

Nanoparticle Design with ProteinMPNN. Our method for designing nanoparticle interfaces using ProteinMPNN is depicted in Fig. 1. For each of the 27 original T33 designs, we first used the Rosetta SymDofMover to slightly vary the rigid body rotational (ω) and translational (r) degrees of freedom of each building block. This allowed us to generate 100 docked configurations that were close, but not identical, to the original design. Next, two contacting trimeric components were extracted and ProteinMPNN was used to select the optimal side chain identities for the same sets of interface residues originally considered for design using Rosetta. ProteinMPNN sequence design was rapid, requiring only ~1 s per sequence, compared to several minutes for Rosetta design. To evaluate the structural features of the designs and enable direct comparison to their Rosetta-designed counterparts, we threaded each ProteinMPNN-designed sequence onto its corresponding dock and evaluated several Rosetta-based interface metrics, including residue counts, clash check, predicted binding energy (ddG), interface surface area, shape complementarity, and the number of buried unsatisfied hydrogen bonding groups. We found that ProteinMPNN and Rosetta yielded designs with roughly similar scores according to these metrics, although the distributions of shape complementarity and predicted binding energy density were slightly better for the Rosetta designs (*SI Appendix*, Fig. S2). We selected a maximum of three variants of each original design for experimental characterization after ranking the designs by shape complementarity, resulting in a total of 76 designs that passed our filter cut-offs (*Materials and Methods*) without any manual intervention (*SI Appendix*, Table S2). We named these ProteinMPNN-designed nanoparticles by appending a period and a numeric identifier to the name of the original design from which each was derived (e.g., T33-01.1 or T33-25.3).

Screening and Characterization of Assembly State. The two components of the 76 designs were encoded as pairs in bicistronic expression plasmids that appended a hexahistidine tag to one of the components. Clarified lysates from 2 mL *Escherichia coli* expression cultures were screened for nanoparticle assembly using polyacrylamide gel electrophoresis (PAGE) under nondenaturing (native) conditions. Twenty-four designs yielded bands that migrated in the range expected of assemblies approaching ~1 MDa in molecular weight (*SI Appendix*, Fig. S3). These 24 potential hits were purified by immobilized metal affinity chromatography (IMAC), reevaluated by native PAGE, and also analyzed by

(sodium dodecyl sulfate–polyacrylamide gel electrophoresis) (SDS-PAGE) to determine which protein pairs co-eluted, a suggestion of successful assembly (*SI Appendix*, Fig. S4). Promising designs were then purified by size-exclusion chromatography (SEC), revealing 13 that eluted as single, symmetric peaks corresponding to ~1 MDa assemblies (Fig. 2 and *SI Appendix*, Fig. S5). Negatively stained electron micrographs of these 13 designs showed that all of them formed homogeneous assemblies of the expected size and shape (Fig. 2). Comparing 2D class averages of each nanoparticle to projections calculated from the computational design models confirmed that all 13 designs assemble to the intended architectures. These results establish that ProteinMPNN can accurately design self-assembling protein nanomaterials with a similar success rate to Rosetta—13/76 (17%) vs. 5/27 (18%)—but much more simply and efficiently.

The 13 successful designs were derived from 9 of the 27 distinct two-component protein complexes used as starting points (Fig. 2). Two successful designs were obtained from T33-08, two from T33-28, and all three designs based on T33-22 successfully assembled. In these cases, the related design models were highly similar, with backbone RMSD and amino acid sequence identities differing on average by only 0.52 Å over the asymmetric unit (ASU) and 16 distinct amino acid changes out of 45 positions considered for design, respectively. Notably, successful designs were obtained for 8 docked configurations that failed to yield experimentally confirmed assemblies in the original publication. Although ProteinMPNN did not generate successful variants for four of the five previously confirmed T33 nanoparticles (T33-09, T33-15, T33-21, and T33-23), verified assemblies were obtained for 13 of the 27 docked configurations between the two sequence design methods. This high success rate indicates that the simple docking method used—which is available as part of the RPXDock software package (32)—is effective at identifying "designable" docks. One docked configuration yielded successful designs using both Rosetta (T33-28) and ProteinMPNN (T33-28.2 and T33-28.3). Detailed comparisons of the interfaces of these designs are provided below.

Comparison of ProteinMPNN- and Rosetta-Designed Interfaces. As an initial comparison of deep learning- and Rosetta-designed interfaces, we visualized the interface residues of each component of the successful ProteinMPNN designs, using color to highlight polar side chains (oxygen atoms colored red, nitrogen atoms colored blue; Fig. 3A and *SI Appendix*, Fig. S6). We found that while both sets of interfaces formed well-packed and chemically complementary interactions, the ProteinMPNN-designed interfaces appeared to have more polar residues, especially near the boundary regions of the interface.

We then quantitatively compared the interfaces using several structural metrics. Many of these were similar between the two sets of designs, as noted previously (*SI Appendix*, Fig. S2). We found a major difference in the number of mutations to the input scaffolds ProteinMPNN made compared to Rosetta: Given the same interface, ProteinMPNN made approximately twice as many mutations (36 ± 5 vs. 17 ± 3; Fig. 3B). ProteinMPNN also tended to make more hydrophobic-to-polar mutations (Fig. 3C) but a similar number of polar-to-hydrophobic mutations (Fig. 3D). When normalized by the total number of mutations, it became clear that ProteinMPNN changed hydrophobes into polars at a similar rate as Rosetta (Fig. 3E), but the likelihood of ProteinMPNN converting a polar side chain into a hydrophobic one was much lower (Fig. 3F). Quantification of the overall fraction of polar residues at each interface showed that ProteinMPNN designs on average had a higher fraction of polar side chains (Fig. 3G). For instance, T33-18.2 has a predominantly polar (59%) interface with only a few well-packed
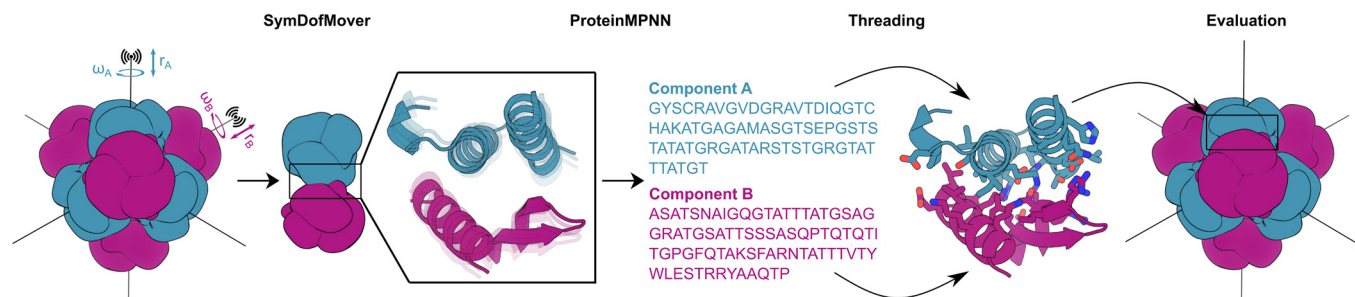
Fig. 1. Schematic of two-component tetrahedral nanoparticle design using ProteinMPNN. The rotational (ω) and translational (r) symmetrical degrees of freedom of a set of previously described T33 docked configurations were sampled using the Rosetta SymDofMover to generate a range of closely related backbones. The backbones of two trimers (one A, one B) were extracted and new amino acid sequences at the A–B interface were generated with ProteinMPNN. The sequences were threaded onto the component backbones using the Rosetta PackRotamerMover to enable evaluation in the context of the full symmetric assembly using Rosetta scoring metrics.

hydrophobes, compared to 43% in the original T33-18 (Fig. 3A). This difference between the two methods is even more remarkable considering that the Rosetta designs were manually refined to remove unnecessary polar-to-hydrophobic mutations: the unmodified Rosetta outputs contained even more hydrophobic side chains. This can be explained because the objective of the Rosetta score function is to minimize the energy of the system, and hydrophobic packing is strongly rewarded (33). Rosetta does not explicitly consider the higher likelihood of oligomeric component aggregation due to surface-exposed hydrophobes, whereas ProteinMPNN was trained on natural protein–protein interfaces that have evolved to balance binding and solubility.
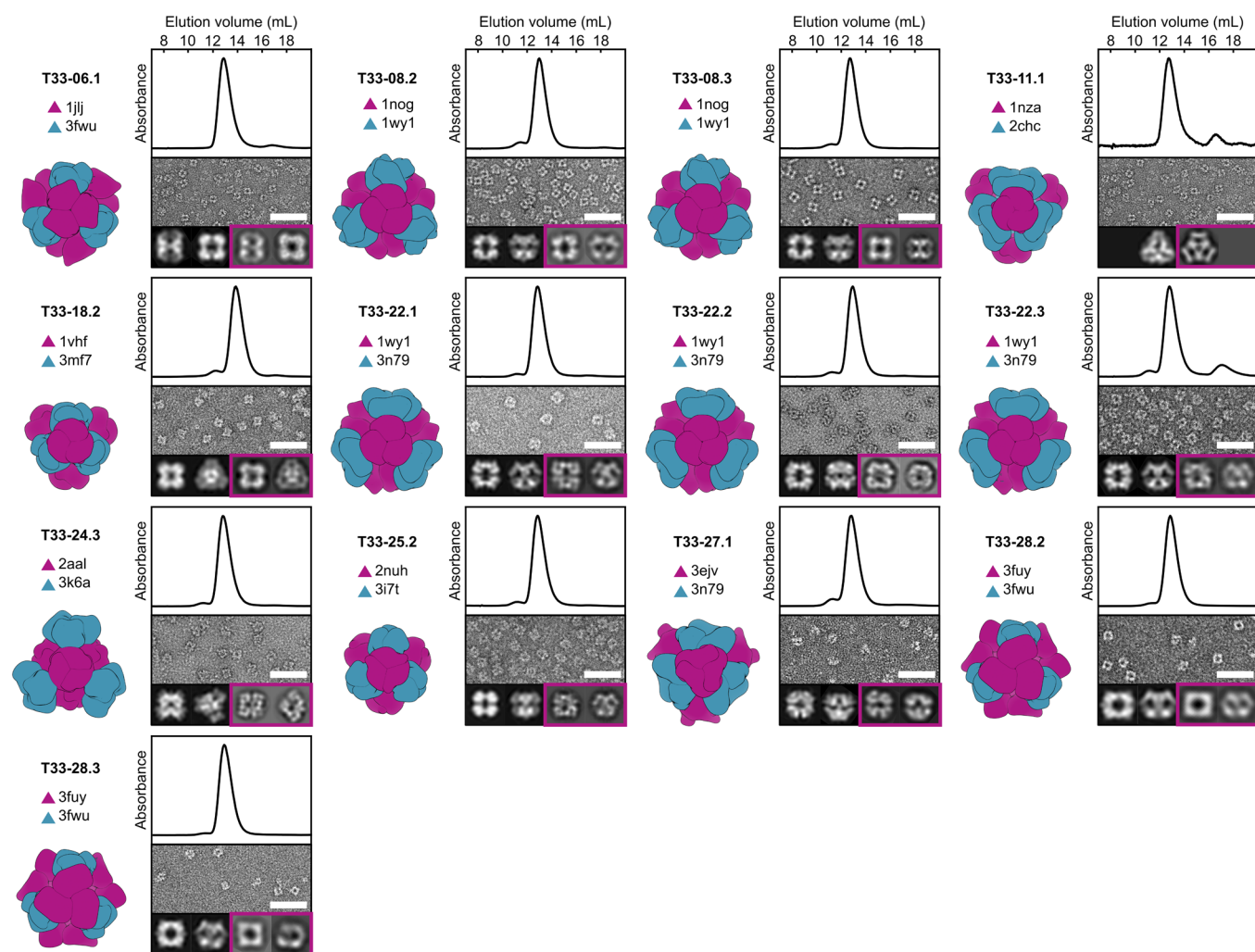

Fig. 2. Structural characterization of 13 co-expressed ProteinMPNN-designed tetrahedral nanoparticles. *Left*: Computational design model and the PDB entry from which each of the two trimeric components (A, purple; B, blue) were derived. *Top:* Co-expressed nanoparticles were analyzed by SEC using a Superdex 200 10/300 column to determine their size and purity. Thirteen out of the 76 designed nanoparticles eluted at ~13 mL, consistent with an expected molecular weight (MW) of ~1 MDa. The absorbance was measured at 230 nm and normalized. *Middle:* Negatively stained electron micrographs of co-expressed tetrahedral nanoparticles. (Scale bar: 50 nm.) *Bottom Left:* Two representative 2D projections calculated from the design model. *Bottom Right* (boxed in purple): corresponding experimentally determined 2D class averages. In all cases, the 2D class averages closely resemble the 2D projections. In the case of T33-11.1, only a single preferred particle orientation was observed experimentally.
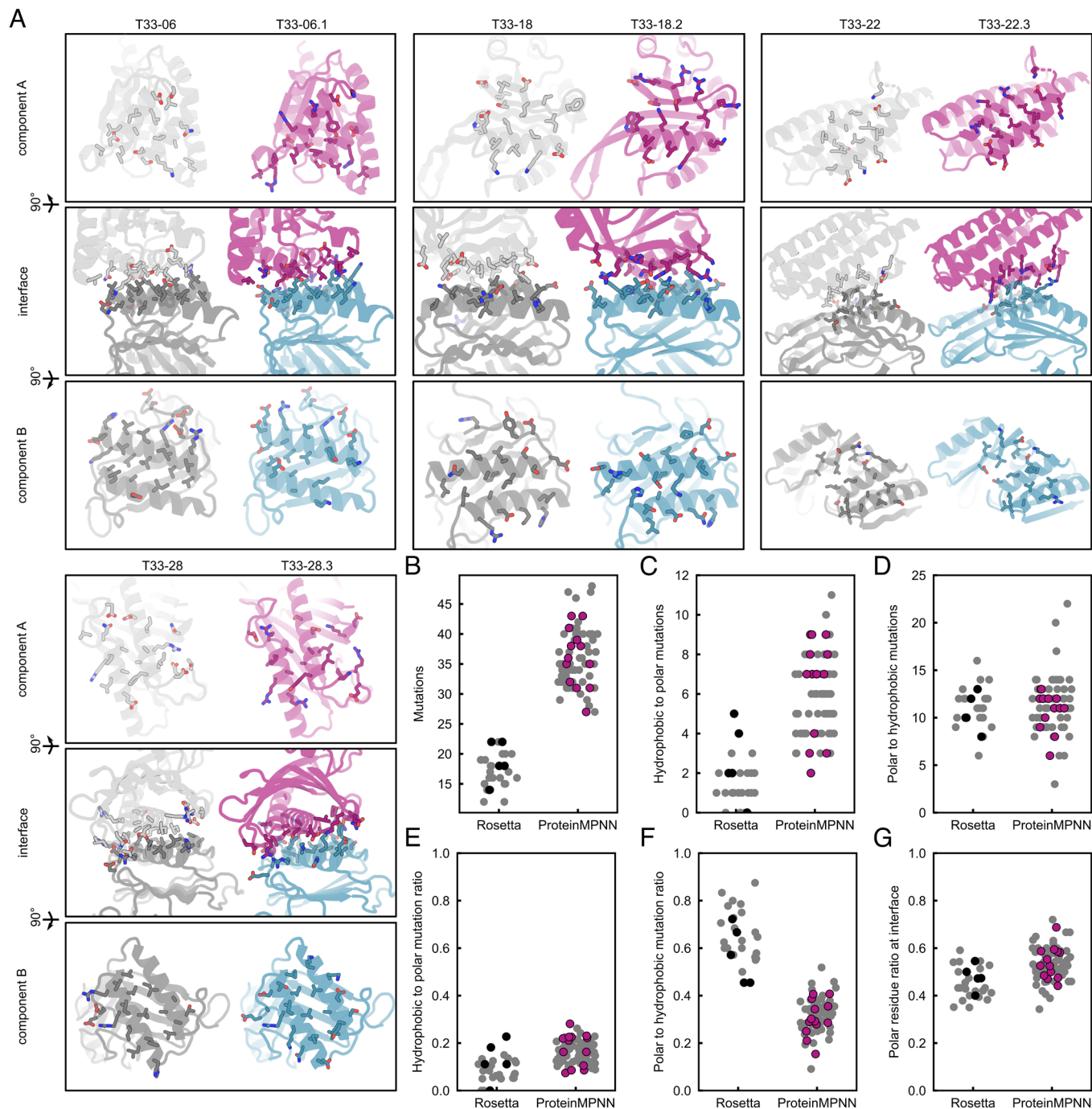
Fig. 3. Comparison of Rosetta and ProteinMPNN-designed interfaces. (*A*) Qualitative comparison of four representative Rosetta-designed interfaces (gray) and ProteinMPNN-designed interfaces (purple/blue). All interface residues are displayed as sticks, with oxygen and nitrogen atoms colored red and blue to highlight polar residues. (*B–G*) Quantitative analyses between the Rosetta (27) and ProteinMPNN design (76) sets, with (*B*) the number of mutations made, (*C*) the number of hydrophobic-to-polar mutations, (*D*) the number of polar-to-hydrophobic mutations, (*E*) the ratio of hydrophobic-to-polar mutations, (*F*) ratio of polar-to-hydrophobic mutations, and (*G*) the ratio of polar residues at the interface. Designs that failed to assemble into target assemblies are shown in gray, while successful designs are shown in black (five Rosetta designs) or purple (13 ProteinMPNN designs).

Although none of the metrics evaluated were able to discriminate successful from unsuccessful designs, we note that the ProteinMPNN interfaces also had higher (i.e., worse) predicted binding energies on average according to the Rosetta score function (*SI Appendix*, Fig. S2). This result is unsurprising considering that Rosetta explicitly selected mutations that improved Rosetta energy during design. Although experimental determination of interface binding strength would be required to validate these predictions, it is intriguing to consider that studies from our group and others have indicated that relatively weak interfaces can be advantageous for the assembly of protein complexes constructed from oligomeric building blocks (28–30). This observation, combined with the more polar nature of the ProteinMPNN-designed components, suggested that they may outperform the Rosetta designs in their ability to assemble in vitro.

In Vitro Assembly of ProteinMPNN-Designed Tetrahedral Nanomaterials. The ability to control the assembly of two-component protein nanomaterials by mixing independently purified components in vitro simplifies their manufacture and
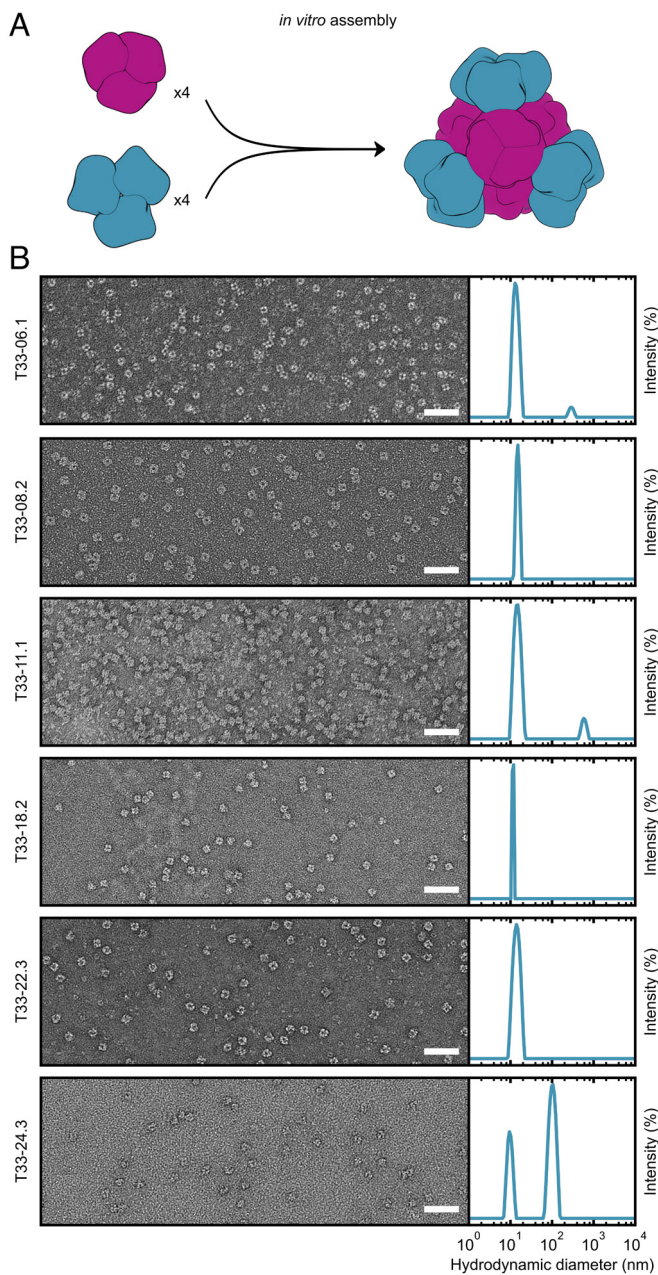
A

in vitro assembly

B

Fig. 4. In vitro assembly of ProteinMPNN-designed tetrahedral nanoparticles. (*A*) In vitro assembly of equimolar components A and B at 25 μM in TBS buffer. (*B*) *Left:* Negatively stained electron micrographs for 6 ProteinMPNN-designed in vitro assembled nanoparticles. (Scale bar: 50 nm.) *Right:* DLS of in vitro assembled tetrahedral nanoparticles. A scattering peak centered around the expected diameter (~15 nm) was present for all assembled nanoparticles. For T33-06.1, T33-11.1, and T33-24.3, a second aggregate peak of diameter >100 nm was observed. However, in all instances these peaks represented <1% of the total mass.

enables advanced functionalization, such as the generation of mosaic nanoparticle immunogens that elicit broadly protective immune responses (34–37). To evaluate in vitro assembly of our 13 ProteinMPNN-designed nanomaterials, we recloned each component individually with an appended hexahistidine tag. We were able to successfully purify both components of six complexes by IMAC and SEC (T33-06.1, T33-08.2, T33-11.1, T33-18.2, T33-22.2, and T33-24.3; Fig. 2). We mixed the components of each complex in a 1:1 molar ratio at approximately 25 μM in TRIS-buffered saline (TBS; 25 mM Tris pH 8.0, 150 mM NaCl, and 1 mM TCEP), as depicted in Fig. 4A. All six nanomaterials assembled efficiently and formed monodisperse particles that

in negatively stained micrographs were indistinguishable from those obtained by coexpression in *E. coli* (Fig. 4B). Dynamic light scattering (DLS) of the assembled materials also indicated efficient formation of the particles of the expected size, with minimal aggregation. By contrast, T33-15 was the only one of the Rosetta designs that could be assembled in vitro from purified components (20).

High-Resolution Structure Determination. We obtained crystal structures of three of our designs (T33-18.2, T33-27.1, and T33-28.3) to evaluate the accuracy of ProteinMPNN at high resolution (Fig. 5*A* and *SI Appendix*, Table S1). In each case, the crystal structure closely matched the computational design model, with backbone RMSD of 0.92, 1.14, and 0.63 Å over the two-chain ASUs of T33-18.2, T33-27.1, and T33-28.3, respectively (Fig. 5*B*). Like other structurally characterized computationally designed assemblies constructed from naturally occurring protein oligomers (20, 25, 38, 39), these minor deviations largely arise from small rigid body movements of the oligomeric building blocks rather than substantial backbone rearrangements. Although ProteinMPNN does not by default explicitly model side chain configurations during design (16), we threaded the ProteinMPNN-designed sequences onto each docked configuration using Rosetta to generate full-atom design models. Comparing the side chains of the highest-resolution structure (T33-18.2, with a resolution of 1.92 Å) to its design model revealed that many of the atomic interactions at the designed interface were recapitulated in the crystal structure, both hydrophobic packing interactions as well as hydrogen bonds and electrostatic interactions between polar side chains at the core of the interface (Fig. 5*B*). The configurations of polar side chains around the periphery of the interface deviated more frequently from those predicted in the design model. These data establish that ProteinMPNN can match the high accuracy of Rosetta in the design of self-assembling protein nanomaterials.

The crystal structures of T33-28 (ref. 20) and T33-28.3 yielded an opportunity to directly compare experimental structures of interfaces designed by Rosetta and ProteinMPNN. The designs had almost identical overall configurations, with a backbone RMSD of only 0.72 Å over the ASU. However, their interfaces differed by 33 mutations, of which 11 were at core positions. For example, a key core interaction in both assemblies features an aromatic side chain at position 56 in T33-28A that is packed into a complementary hydrophobic groove on the other subunit. ProteinMPNN placed a Tyr in this position, with its hydroxyl forming a hydrogen bond with Thr30 on T33-28B, instead of the more hydrophobic Phe selected by Rosetta (Fig. 5*C*). Two additional hydrophobic-to-polar mutations form hydrogen bonds at the core of the T33-28.3 interface: Ile22Lys in T33-28.3B and Ile48Asp in T33-28.3A (Fig. 5*D*). Several more polar residues at the boundary regions of T33-28.3 allowed the formation of more than 10 favorable hydrogen and ionic interactions, compared to zero polar interactions across the T33-28 interface (*SI Appendix*, Fig. S7). Most of these interactions were accurately recapitulated in the T33-28.3 crystal structure, although at the periphery some deviations of polar rotamer conformations were observed due to interaction with water molecules. Finally, a poorly packed region on the periphery of the T33-28 interface was redesigned by ProteinMPNN to feature a tyrosine residue from T33-28.3B (Tyr87) that is involved in several favorable packing interactions across the interface, including a cation–pi interaction with Arg61 from T33-28.3A that was observed in the crystal structure (Fig. 5*E*). The latter interaction was present in all three ProteinMPNN designs, while cation–pi interactions are rarely designed by Rosetta. In addition to highlighting the different types of interactions designed by each algorithm, these comparisons between the Rosetta- and ProteinMPNN-designed T33-28
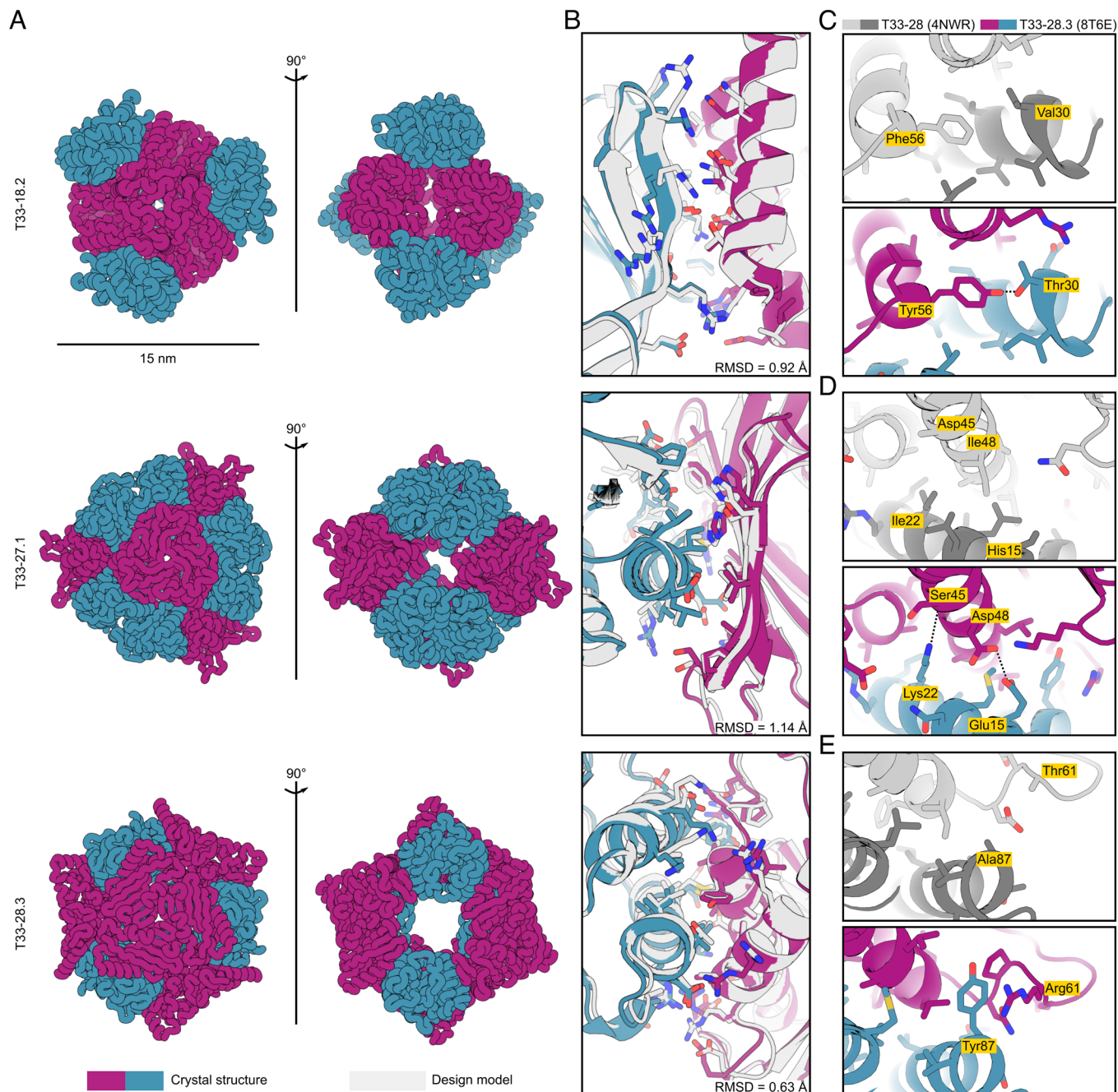
Fig. 5. Crystal structures of ProteinMPNN-designed tetrahedral nanoparticles. (*A*) Ribbon displays of crystal structures of three tetrahedral nanoparticles: T33-18.2 (PDB ID: 8T6C), T33-27.1 (PDB ID: 8T6N), and T33-28.3 (PDB ID: 8T6E). (*B*) Atomic interactions at the designed interfaces. Side chains of residues with atoms within 5.5 Å of the opposing subunit are shown as sticks. Light gray color indicates the ProteinMPNN design model and purple-blue colors indicate the crystal structure of components A and B, respectively. In all three instances, the crystal structures matched the design models with high accuracy, with backbone RMSD over the ASU of: 0.92 Å (T33-18.2), 1.14 Å (T33-27.1), and 0.63 Å (T33-28.3). (*C–E*) Comparison of selected interface interactions in the crystal structures for T33-28 (PDB ID: 4NWR; ref. 20) and T33-28.3 (PDB ID: 8T6E). (*C*) In T33-28.3A the hydrophobic Phe56 is replaced with a Tyr with its hydroxyl group hydrogen bonding with Thr33 on T33-28.3B. (*D*) T33-28.3 features two interface stabilizing polar interactions at the core. The Lys22 amino group forms a hydrogen bond with the backbone carbonyl of Ser45, and Asp48 forms hydrogen bonds with Glu15. In contrast, the core positions for T33-28 (Ile22 and Ile48) featured exclusively hydrophobic packing. (*E*) In a poorly packed region of T33-28, T33-28.3 features a cation-pi interaction between Tyr87 and Arg61. Note: in the crystal structure of T33-28, only the Cβ atom of the His15, Asp45, and Thr61 side chains were resolved.

assemblies establish how highly divergent designed interface sequences can drive the assembly of nearly geometrically identical self-assembling protein materials.

Discussion

Computationally designed self-assembling proteins are a promising technology platform that has begun to yield commercial products. Two-component nanoparticles similar to those reported here have been used to encapsulate and deliver molecular cargoes (40–42), to scaffold proteins for structural characterization by cryo-electron microscopy (43–45), and to display antigens in repetitive arrays that, when used as vaccines, elicit robust and in some cases broadly protective immune responses (22, 34, 35, 46–49). The recent licensure of a protein nanoparticle vaccine for SARS-CoV-2 established computationally designed self-assembling proteins as a commercial technology (31, 50). Nevertheless, the technology is still relatively new, and the assemblies designed to date are relatively simple. By contrast,

the remarkably sophisticated self-assembling proteins observed in nature—and the highly specialized functions they perform—hint at the technological potential of designed protein assemblies. Continued methods development will help realize this potential by making possible the design of synthetic protein assemblies that rival the structural and functional complexity of the molecular machines of the cell.

Many recent developments in computational methods for modeling and designing proteins have focused on machine learning (4, 51–57). ProteinMPNN (16), based on the earlier graph-based message-passing neural network (MPNN) architecture of Ingraham et al. (17), has proven to be a robust and versatile tool for fixed-backbone sequence design: ProteinMPNN can be used as a sequence design module in a wide variety of protein redesign and de novo design tasks. As we have shown, this includes the design of self-assembling protein nanomaterials, where ProteinMPNN outperformed Rosetta, the previous state-of-the-art sequence design method, in a head-to-head comparison. Although the two methods generated experimentally confirmed assemblies with similar success rates, the speed and simplicity of ProteinMPNN are considerable advantages. In particular, its modest computational resource requirements and its elimination of the need for intensive manual review by structural biologists should make it accessible to a wider set of researchers than Rosetta, facilitating the application of protein design as a solution to a wider variety of challenges in biology and beyond.

We found an additional advantage of ProteinMPNN in its ability to design protein–protein interfaces with a higher proportion of polar residues than Rosetta. This directly translated to improved biophysical properties in the oligomeric components of our two-component protein nanomaterials. Specifically, the lower tendency of ProteinMPNN-designed components to aggregate resulted in a higher fraction of materials that could be assembled in vitro from individually purified components. In vitro assembly has been used in the manufacturing processes for several designed protein nanoparticle vaccines in clinical trials (35, 49) (NCT05664334 and NCT04750343), including a mosaic nanoparticle vaccine for influenza that codisplays four distinct hemagglutinin antigens on the same nanoparticle immunogen (34) (NCT04896086). Optimizing this property of designed protein nanomaterials is therefore of potential commercial relevance. Furthermore, as new computational methods enable the design of increasingly complex protein nanomaterials—such as those that break symmetry or comprise several different components (58–60)—hierarchical in vitro assembly will become even more preferred as a method of construction, mirroring the reticular synthesis of complex metal-organic frameworks and DNA nanotechnology objects (61–64). Our results suggest that the ability of ProteinMPNN to design components of protein nanomaterials with favorable solution properties will likely speed the development of the next generation of protein-based technologies.

Materials and Methods

Computational Design. The symmetrical degrees of freedom of the T33 nanoparticle docks were sampled with $\Delta 0.5$ Å translation (r) and $\Delta 1°$ rotation (ω) using the Rosetta SymDofMover, and 100 unique configurations per dock were generated. Because for some docks the interface spans multiple subunits of the homotrimers, the trimers of both components sharing a single nanoparticle interface were isolated, and the identical interface residues to King et al. (20) were identified. The interface sequences were optimized using ProteinMPNN, with all other residues kept fixed. A total of 16 sequences per sampled configuration (1,600 sequences per dock) were generated with ProteinMPNN using sampling temperatures of 0.2 and 0.3 and backbone noise of 0.05. Based on the loss score, the top 50% of the sequences were selected and threaded back onto the sampled ASUs using Rosetta resfiles, and nanoparticle design models

were generated using the SymDofMover. All residues were repacked using the Rosetta PackRotamerMover (the packer) with tetrahedral symmetry. Finally, a detailed evaluation of the designed interfaces was performed by filtering interfaces based on the following metrics: number of glycines = 0, number of methionines < 5, number of aromatics < 5, number of clashes < 3, shape complementarity > 0.5, predicted interface strength < 0 Rosetta Energy Units, and solvent accessible surface area buried at the designed interface >1,000 Å^2. For each of the 27 docks, the 3 highest shape complementarity designs were selected, with the exceptions of T33-15–where only 1 design passed the filter metrics–and T33-12, T33-16, and T33-29–where only 2 designs passed the filter metrics. This gave a total of 76 designs for the ProteinMPNN-designed set.

Bicistronic Expression, Lysate Screening, and Purification. Synthetic genes for designed proteins optimized for *E. coli* expression were purchased from Genscript ligated into the pET-29b(+) vector using the NdeI and XhoI restriction sites. A second ribosome binding site was inserted between the open reading frames of the two components of the bicistronic nanoparticle designs (AGAAGGAGATATCAT) such that the two proteins would be coexpressed. Only one of the components with the most accessible C-terminus carried a hexahistidine tag to facilitate copurification. Plasmids were cloned into BL21 (DE3) *E. coli* competent cells (New England Biolabs). Transformants were grown in 2 mL lysogeny broth (LB; 10 g Tryptone, 5 g Yeast Extract, 10 g NaCl) cultures in 96 deep-well plates at 37 °C for 2 h, induced with 1 mM IPTG, and then continued to shake at 37 °C for ~18 h. The cells were harvested and lysed by sonication using a plate sonicator (Qsonica) for 2 min in pulses of 20 s on, 40 s off in 25 mM Tris pH 8.0, 300 mM NaCl, 30 mM imidazole, 1 mM PMSF, 0.05% (w/v) DNAse I. Lysates were clarified by centrifugation for 15 min at 4,000 g in a swinging bucket rotor. Supernatants were applied to native PAGE gels and run at 100 V for ~4 h on ice. Potential nanoparticle hits had bands at ~1 MDa after staining with GelCode (Thermo Fisher Scientific).

The potential hits were expressed in 1 L LB cultures in 2 L baffled shake flasks. The cells were harvested and lysed by sonication using a microplate horn system (Qsonica) for 10 min with 10 s pulses at 80% amplitude in 25 mM Tris pH 8.0, 300 mM NaCl, 30 mM imidazole, 1 mM PMSF, 0.05% (w/v) DNAse I. Lysates were clarified by centrifugation at 24,000 g for 30 min and applied to a 2 mL column bed of Ni-NTA resin (Qiagen) for purification by IMAC. The resin was washed with 50 mL buffer 25 mM Tris pH 8.0, 300 mM NaCl, 30 mM imidazole. The protein of interest was eluted using 6 mL of 25 mM Tris pH 8.0, 300 mM NaCl, 300 mM imidazole. Directly after elution, 1 mM dithiothreitol (DTT; Sigma Aldrich) was added to eluates. Eluates were evaluated for coelution by SDS-PAGE and assembly by native PAGE. IMAC eluates that contained nanoparticle bands on native PAGE were concentrated in 100 kDa MWCO centrifugal filters (Amicon), sterile filtered (0.22 µm), and applied to a Superdex 200 Increase 10/300 (Cytiva) in 25 mM Tris pH 8.0, 150 mM NaCl, 1 mM DTT. Peaks eluting at ~13 mL strongly indicate nanoparticle formation and were fractionated and analyzed by SDS-PAGE to confirm the presence of both components.

Individual Component Expression and Purification. Synthetic genes for individual components, each with a hexahistidine purification tag, were optimized for *E. coli* expression and purchased from Genscript ligated into the pET-29b(+) vector at the NdeI and XhoI restriction sites. The proteins were expressed in BL21(DE3) (New England Biolabs) in LB grown in 2 L baffled shake flasks. Cells were grown at 37 °C to an OD$_{600}$ ~ 0.6 and then induced with 1 mM IPTG. Expression temperature was reduced to 18 °C and the cells were shaken for ~18 h. The cells were harvested and lysed by sonication using a Qsonica Q500 for 10 min with 10 s pulses at 80% amplitude in 25 mM Tris pH 8.0, 300 mM NaCl, 30 mM imidazole, 1 mM PMSF, 0.05% (w/v) DNAse I. Lysates were clarified by centrifugation at 24,000 g for 30 min and applied to a 5 mL column bed of Ni-NTA resin (Qiagen) for purification by IMAC. Resin was prewashed with 25 mM Tris pH 8.0, 300 mM NaCl, 30 mM imidazole. The protein of interest was eluted using 25 mM Tris pH 8.0, 300 mM NaCl, 300 mM imidazole. DTT was added to eluates to a final concentration of 1 mM. Eluates were pooled, concentrated in 10 K MWCO centrifugal filters (Pall), sterile filtered (0.22 µm) and applied to either a Superdex 200 Increase 10/300 (Cytiva), or HiLoad 26/600 Superdex 200 pg SEC column (Cytiva) using 25 mM Tris pH 8.0, 300 mM NaCl, 1 mM TCEP buffer.

In Vitro Assembly. Total protein concentration of purified individual nanoparticle components was determined by measuring absorbance at 280 nm using a UV/vis spectrophotometer (Agilent Cary 3500 Multicell) and calculated extinction coefficients

(65). The assembly steps were performed at room temperature with addition in the following order: component A, followed by additional buffer as needed to achieve desired final concentration, and finally component B (in 25 mM Tris pH 8.0, 150 mM NaCl), with a molar ratio of A:B of 1:1. T33-08.2, T33-11.1, and T33-18.2 components were incubated at room temperature for at least 1 h in order to drive a more complete assembly. These assemblies were applied to a Superose 6 Increase 10/300 GL column (Cytiva) for purification by SEC using 25 mM Tris pH 8.0, 300 mM NaCl, 1 mM TCEP running buffer. Assembled nanoparticles were sterile filtered (0.22 μm) immediately prior to column application and following pooling of fractions. T33-06.1, T33-22.3, and T33-24.3 assembly reactions were incubated at room temperature for 18 h, then sterile filtered (0.22 μm) and analyzed directly, without subsequent SEC.

Negative Stain Electron Microscopy Collection and Processing. Tetrahedral nanoparticles were first diluted to 100 μg/mL in water prior to application of 3 μL of sample onto freshly glow-discharged 400-mesh copper grids (Ted Pella). Sample was incubated on the grid for 30 s before excess liquid blotted away with filter paper (Whatman). Then, 3 μL of 2% w/v uranyl formate (UF) stain was applied to the grid and immediately blotted away before an additional 3 μL of UF stain was applied. Stain was blotted off by filter paper, and a final 3 μL of UF stain was applied and allowed to incubate for ~30 s. Finally, the stain was blotted away and the grids were allowed to dry for 3 min. Prepared grids were imaged using EPU 2.0 on a 120 kV Talos L120C transmission electron microscope (Thermo Fisher Scientific) at 57,000× magnification with a BM-Ceta camera. Data processing was done in CryoSPARC (66), starting with CTF correction, particle picking, and extraction. Two or three rounds of 2D classification were done.

Crystallographic Data Collection and Structure Determination. All crystallization experiments were conducted using the sitting drop vapor diffusion method.

Crystallization trials were set up in 200 nL drops using the 96-well plate format at 20 °C. Crystallization plates were set up using a Mosquito LCP from SPT Labtech, then imaged using UVEX microscopes and UVEX PS-256 from JAN Scientific. Diffraction quality crystals formed in 0.9 M NPS 0.1 M Tris-BICINE pH 8.5, 30 % v/v of Glycerol and PEG 4000 for T33-18.2; 0.12 M ethylene glycol, 0.1 M Tris-BICINE pH 8.5, 30 % v/v of glycerol and PEG 4000 for T33-27.1; and 1.26 M sodium phosphate, 0.14 M potassium phosphate for T33-28.3.

Diffraction data were collected at the Advanced Light Source (ALS) HHMI beamline 8.2.1/ 8.2.2 and 5.0.1. at 1 Å wavelength. X-ray intensities and data reduction were evaluated and integrated using XDS (67) and merged/scaled using Pointless/Aimless in the CCP4 program suite (67, 68). Starting phases were obtained by molecular replacement using Phaser (69) using the designed model for the structures. Following molecular replacement, the models were improved using phenix.autobuild (70); efforts were made to reduce model bias by setting rebuild-in-place to false and using simulated annealing and prime-and-switch phasing. Structures were refined in Phenix (70). Model building was performed using COOT (71). The final model was evaluated using MolProbity (72). Data collection and refinement statistics are recorded in *SI Appendix*, Table S1. Data deposition, atomic coordinates, and structure factors reported in this paper have been deposited in the Protein Data Bank (PDB) (73), http://www.rcsb.org/, with accession codes 8T6C (T33-18.2), 8T6N (T33-27.1), and 8T6E (T33-28.3).

Data, Materials, and Software Availability. Design models (.pdb files) data have been deposited in T33 design models (10.5281/zenodo.8278877) (74).

Author affiliations: ^aDepartment of Physical Chemistry and Soft Matter, Wageningen University and Research, Wageningen 6078 WE, The Netherlands; ^bDepartment of Biochemistry, University of Washington, Seattle, WA 98195; ^cInstitute for Protein Design, University of Washington, Seattle, WA 98195; ^dMolecular Biophysics and Integrated Bioimaging, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; and ^eHHMI, Seattle, WA 98195

1. J. Yang *et al.*, Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 1496–1503 (2020).
2. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
3. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
4. Z. Lin *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
5. I. Anishchenko *et al.*, De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
6. J. Wang *et al.*, Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
7. A. H.-W. Yeh *et al.*, De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).
8. J. Ingraham *et al.*, Illuminating protein space with a programmable generative model. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.12.01.518682 (Accessed 5 March 2024).
9. C. Frank *et al.*, Efficient and scalable de novo protein design using a relaxed sequence space. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.02.24.529906 (Accessed 5 March 2024).
10. I. D. Lutz *et al.*, Top-down design of protein architectures with reinforcement learning. *Science* **380**, 266–273 (2023).
11. B. I. M. Wicky *et al.*, Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).
12. Y. Liu *et al.*, Publisher correction: Rotamer-free protein sequence design based on deep learning and self-consistency. *Nat. Comput. Sci.* **2**, 526–526 (2022).
13. D. Repecka *et al.*, Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* **3**, 324–333 (2021).
14. B. Jing, S. Eismann, P. Suriana, R. J. L. Townshend, R. Dror Learning from protein structure with geometric vector perceptrons. arXiv [Preprint] (2020). https://doi.org/10.48550/arXiv.2009.01411 (Accessed 5 March 2024).
15. C. Hsu *et al.*, Learning inverse folding from millions of predicted structures. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.04.10.487779 (Accessed 5 March 2024).
16. J. Dauparas *et al.*, Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
17. J. Ingraham, V. Garg, R. Barzilay, T. Jaakkola, "Generative models for graph-based protein design" in *Advances in Neural Information Processing Systems* (Vancouver Convention Center, Vancouver, Canada, 2019), vol. 32, pp. 15820–15831.
18. N. Anand *et al.*, Protein sequence design with a learned potential. *Nat. Commun.* **13**, 1–11 (2022).
19. N. R. Bennett *et al.*, Improving de novo protein binder design with deep learning. *Nat. Commun.* **14**, 2625 (2023).
20. N. P. King *et al.*, Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510**, 103–108 (2014).
21. J. M. Fletcher *et al.*, Self-assembling cages from coiled-coil peptide modules. *Science* **340**, 595–599 (2013).
22. P. J. M. Brouwer *et al.*, Enhancing and shaping the immunogenicity of native-like HIV-1 envelope trimers with a two-component protein nanoparticle. *Nat. Commun.* **10**, 4272 (2019).
23. A. Khmelinskaia, A. Wargacki, N. P. King, Structure-based design of novel polyhedral protein nanomaterials. *Curr. Opin. Microbiol.* **61**, 51–57 (2021).
24. P. B. Stranges, B. Kuhlman, A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Sci.* **22**, 74–82 (2013).
25. J. B. Bale *et al.*, Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **353**, 389–394 (2016).
26. Y. Hsia *et al.*, Design of multi-scale protein complexes by hierarchical building block fusion. *Nat. Commun.* **12**, 1–10 (2021).
27. I. Vulovic *et al.*, Generation of ordered protein assemblies using rigid three-body fusion. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015037118 (2021).
28. A. J. Wargacki *et al.*, Complete and cooperative in vitro assembly of computationally designed self-assembling protein nanomaterials. *Nat. Commun.* **12**, 883 (2021).
29. E. D. Levy, E. Boeri Erba, C. V. Robinson, S. A. Teichmann, Assembly reflects evolution of protein complexes. *Nature* **453**, 1262–1265 (2008).
30. P. Ceres, A. Zlotnick, Weak protein−protein interactions are sufficient to drive assembly of hepatitis B virus capsids. *Biochemistry* **41**, 11525–11531 (2002).
31. A. C. Walls *et al.*, Elicitation of potent neutralizing antibody responses by designed protein nanoparticle vaccines for SARS-CoV-2. *Cell* **183**, 1367–1382.e17 (2020).
32. W. Sheffler *et al.*, Fast and versatile sequence-independent protein docking for nanomaterials design using RPXDock. *PLoS Comput. Biol.* **19**, e1010680 (2023).
33. R. F. Alford *et al.*, The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
34. S. Boyoglu-Barnum *et al.*, Quadrivalent influenza nanoparticle vaccines induce broad protection. *Nature* **592**, 623–628 (2021).
35. A. C. Walls *et al.*, Elicitation of broadly protective sarbecovirus immunity by receptor-binding domain nanoparticle vaccines. *Cell* **184**, 5432–5447.e16 (2021).
36. P. J. M. Brouwer *et al.*, Two-component spike nanoparticle vaccine protects macaques from SARS-CoV-2 infection. *Cell* **184**, 1188–1200.e19 (2021).
37. K. Sliepen *et al.*, Induction of cross-neutralizing antibodies by a permuted hepatitis C virus glycoprotein nanoparticle vaccine candidate. *Nat. Commun.* **13**, 7271 (2022).

38. Y. Hsia *et al.*, Corrigendum: Design of a hyperstable 60-subunit protein icosahedron. *Nature* **540**, 150 (2016).
39. N. P. King *et al.*, Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **336**, 1171–1174 (2012).
40. T. G. W. Edwardson, T. Mori, D. Hilvert, Rational engineering of a designed protein cage for siRNA delivery. *J. Am. Chem. Soc.* **140**, 10439–10442 (2018).
41. G. L. Butterfield *et al.*, Evolution of a designed protein assembly encapsulating its own RNA genome. *Nature* **552**, 415–420 (2017).
42. T. G. W. Edwardson, S. Tetter, D. Hilvert, Two-tier supramolecular encapsulation of small molecules in a protein cage. *Nat. Commun.* **11**, 5410 (2020).
43. Y. Liu, D. T. Huynh, T. O. Yeates, A 3.8 Å resolution cryo-EM structure of a small protein bound to an imaging scaffold. *Nat. Commun.* **10**, 1864 (2019).
44. Y. Liu, S. Gonen, T. Gonen, T. O. Yeates, Near-atomic cryo-EM imaging of a small protein displayed on a designed scaffolding system. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3362–3367 (2018).
45. R. Castells-Graells *et al.*, Rigidified scaffolds for 3 angstrom resolution cryo-EM of small therapeutic protein targets. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.09.18.508009 (Accessed 5 March 2024).
46. G. Ueda *et al.*, Tailored design of protein nanoparticle scaffolds for multivalent presentation of viral glycoprotein antigens. *Elife* 9, e57659 (2020).
47. T. U. J. Bruun, A.-M. C. Andersson, S. J. Draper, M. Howarth, Engineering a rugged nanoscaffold to enhance plug-and-display vaccination. *ACS Nano* **12**, 8855–8866 (2018).
48. R. Rahikainen *et al.*, Overcoming symmetry mismatch in vaccine nanoassembly through spontaneous amidation. *Angew. Chem. Int. Ed. Engl.* **60**, 321–330 (2021).
49. J. Marcandalli *et al.*, Induction of potent neutralizing antibody responses by a designed protein nanoparticle vaccine for respiratory syncytial virus. *Cell* **176**, 1420–1431.e17 (2019).
50. J. Y. Song *et al.*, Safety and immunogenicity of a SARS-CoV-2 recombinant protein nanoparticle vaccine (GBP510) adjuvanted with AS03: A randomised, placebo-controlled, observer-blinded phase 1/2 trial. *EClinicalMedicine* **51**, 101569 (2022).
51. M. Baek, D. Baker, Deep learning and protein structure modeling. *Nat. Methods* **19**, 13–14 (2022).
52. N. Ferruz, B. Höcker, Controllable protein design with language models. *Nat. Mach. Intell.* **4**, 521–532 (2022).
53. J. Wang, Protein sequence design by deep learning. *Nat. Comput. Sci.* **2**, 416–417 (2022).
54. A. Elnaggar *et al.*, ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
55. E. Nijkamp, J. Ruffolo, E. N. Weinstein, N. Naik, A. Madani, ProGen2: Exploring the boundaries of protein language models. *Cell Syst.* **14**, 968–978.e3 (2022).
56. B. Chen *et al.*, xTrimoPGLM: Unified 100B-scale pre-trained transformer for deciphering the language of protein. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.07.05.547496 (Accessed 5 March 2024).
57. A. Rives *et al.*, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016239118 (2021).
58. S. Lee *et al.*, Design of four component T=4 tetrahedral, octahedral, and icosahedral protein nanocages through programmed symmetry breaking. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.06.16.545341 (Accessed 5 March 2024).
59. Q. M. Dowling *et al.*, Hierarchical design of pseudosymmetric protein nanoparticles. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.06.16.545393 (Accessed 5 March 2024).
60. R. D. Kibler *et al.*, Stepwise design of pseudosymmetric protein hetero-oligomers. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.04.07.535760 (Accessed 5 March 2024).
61. H. Ramezani, H. Dietz, Building machines with DNA molecules. *Nat. Rev. Genet.* 21, 5–26 (2019).
62. B. Zhang *et al.*, Reticular synthesis of multinary covalent organic frameworks. *J. Am. Chem. Soc.* **141**, 11420–11424 (2019), 10.1021/jacs.9b05626.
63. N. C. Seeman, H. F. Sleiman, DNA nanotechnology. *Nat. Rev. Mater.* **3**, 1–23 (2017).
64. W. Xu *et al.*, Anisotropic reticular chemistry. *Nat. Rev. Mater.* **5**, 764–779 (2020).
65. M. R. Wilkins *et al.*, Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.* **112**, 531–552 (1999).
66. A. Punjani, J. L. Rubinstein, D. J. Fleet, M. A. Brubaker, cryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
67. W. Kabsch, XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
68. M. D. Winn *et al.*, Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).
69. A. J. McCoy *et al.*, Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
70. P. D. Adams *et al.*, PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
71. P. Emsley, K. Cowtan, Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
72. C. J. Williams *et al.*, MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).
73. H. M. Berman *et al.*, The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
74. R. J. de Haas, T33-design-models. Zenodo. https://doi.org/10.5281/zenodo.8278877. Accessed 6 March 2024.