

Design and Structure of Protein Cages Based on Helical Fusion and Machine Learning

Pablo San Segundo-Acosta, Johanne Le Coq, Jasminka Boskovic, Robin A. Aglietti, Peter Bowers, Todd O. Yeates, and Roger Castells-Graells*



Cite This: <https://doi.org/10.1021/jacs.6c11491>



Read Online

ACCESS |



Metrics & More

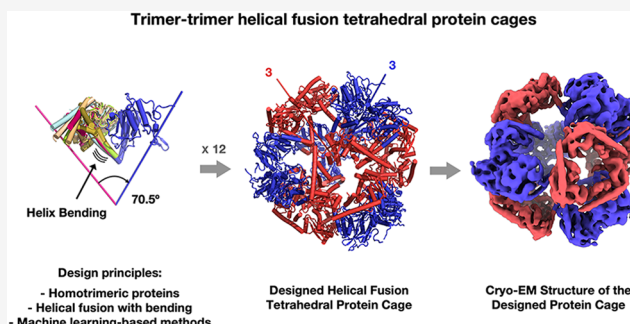


Article Recommendations



Supporting Information

ABSTRACT: Self-assembling protein cages are versatile nanoscale architectures with broad applications in drug delivery, vaccine development, and structural biology. Historically, two main strategies have been used to construct such cages: genetic fusion of oligomeric domains connected by helical linkers, and computational interface design using either physics-based or machine learning-based methods. Here, we extend the original fusion approach using modern AI algorithms and more sophisticated treatments of helix bending to create protein cages with novel architectures composed exclusively of trimeric building blocks arranged in tetrahedral symmetry. Of 15 designs tested experimentally, multiple sequence variants of two of these designs assembled predominantly into soluble, monodisperse particles of the expected size, with native molecular masses of 633 kDa (T33-Fus-1A, B) and 638 kDa (T33-Fus-2). Cryo-electron microscopy (cryo-EM) structures of three distinct sequence variants spanning from 3.0 to 3.9 Å in resolution confirmed the intended structures in atomic detail, with C- α RMSD values over the entire assemblies as low as 2 Å. The predicted modes of helix bending were similarly validated. The results highlight the impact of methodological improvements for achieving a level of regularity and design precision that has largely evaded prior applications of the fusion approach. These findings expand the prospects and accessible design space for self-assembling protein nanomaterials.



INTRODUCTION

Protein cages are symmetric macromolecular assemblies constructed from multiple copies of one or more protein subunits. Their unique architecture, typically combining a defined interior cavity with a programmable exterior surface, has motivated extensive efforts to design them for applications in biotechnology and medicine.^{1–13}

The design concepts and first successful strategy for creating novel protein cages were introduced by Padilla, Colovos, and Yeates,¹⁴ who recognized that fusing two oligomeric protein domains through a rigid α -helical linker could enforce specific geometric relationships between their symmetry axes. When two (or more) axes of symmetry are brought into orientations that intersect at certain angles, multiple copies of the resulting fusion protein are driven to self-assemble into a closed, cage-like particle with tetrahedral, octahedral, or icosahedral symmetry. The underlying geometric rules for diverse constructions have been laid out.^{14–16} Following the creation of a tetrahedral cage, the oligomeric fusion approach was subsequently extended to produce a cubic cage with octahedral symmetry.¹⁷ In both cases, and in related work by others,^{18–20} challenges related to linker flexibility and the formation of somewhat distorted geometries or alternative assembly states

highlighted the need for more precise control over the fusion geometry and the sequence of the connecting segments.

A complementary approach pioneered by King et al.,²¹ relies on computational design of de novo protein–protein interfaces to drive assembly. Physics-based methods using the Rosetta software suite enabled the creation of tetrahedral,^{21,22} octahedral,²¹ and icosahedral²³ cages from one- and two-component systems. Motivated by the need to improve experimental success rates for design, machine learning (ML) algorithms were subsequently incorporated into numerous interface design studies. The introduction of ProteinMPNN²⁴ for sequence design has markedly improved experimental outcomes for designed protein nanomaterials.^{25–27} More recently, RFdiffusion²⁸ and related generative methods have opened the scope of de novo backbone generation, enabling the creation of higher-order symmetric and pseudosymmetric assemblies.^{29–32}

Received: June 4, 2026

Revised: August 6, 2026

Accepted: August 7, 2026

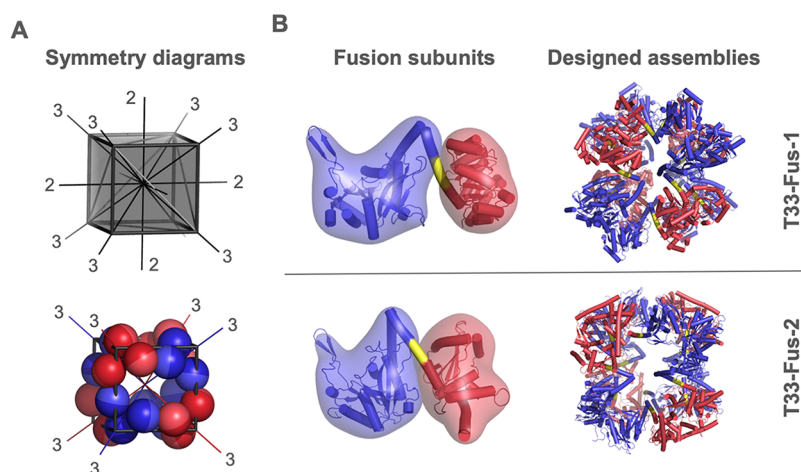


Figure 1. Design principles for trimer–trimer helical fusion cages and schematic models. (A) The symmetry axes for tetrahedral symmetry are shown on a cube (top) with two different types of trimeric components situated on alternating vertices (bottom). The 2-fold axes of symmetry, though they remain, are not drawn. (B) Elements of the design: two homotrimeric proteins (in red and blue) of known structure with terminal α -helices are identified and connected in tetrahedral symmetry such that their helices point toward one another. A continuous helical linker (in yellow) bridges the two trimers, and the junction sequence is optimized using ProteinMPNN.

While modern computational tools have helped advance many interface design-based studies, comparable advances for fusion-based assembly approaches have been more limited.^{33–35} Here, we expand and improve upon previous fusion-based protein cage design work through two main advances: the use of machine learning methods to optimize key regions of the fusion constructs generated, and an algebra-based consideration of natural α -helix bending. We describe a strategy combining classical mechanics and machine learning algorithms to create protein cages from two trimeric building blocks arranged in tetrahedral symmetry, with multiple successful designs confirmed by high-resolution cryo-EM structure determination.

RESULTS

Computational Design Pipeline for Trimer–Trimer Helical Fusion Cages

Our design strategy combines geometric connection of trimeric building blocks with ML-based sequence optimization at the helical junction and any other regions where incidental contacts might occur between the two trimeric components (Figure 1). A tetrahedral cage composed of two distinct trimeric components (termed a T33 architecture) comprises 4 copies of each trimer type, with different types occupying alternating vertices of a cube (Figure 1). If the two trimer types are situated at roughly similar distances from the center of the assembly, the overall architecture tends to resemble a cube. If one trimer type is more distal compared to the other, the overall shape tends to resemble a tetrahedron. When the T33 architecture is based on a genetic fusion of the two component types, the cage or nanoparticle is thus assembled from 12 identical copies of a subunit bearing the two component types connected by a linker, which is typically a continuous α -helix that spans between the two domains.

We began by mining the Protein Data Bank (PDB)³⁶ for homotrimeric proteins possessing terminal α -helices of at least 10 residues, reasoning that such helices provide the rigid structural element necessary for a predictable fusion geometry. Candidate trimers were required to have helices at either their N- or C-terminus. All pairwise combinations of a C-terminal-

helical component and an N-terminal-helical component were then evaluated for consistency with tetrahedral point group symmetry after joining their ends by an α helical linker. The added linker was allowed to be a maximum of 7 residues to a minimum of -2 ; the latter case means the native helical termini are overlapping.

The underlying geometric requirement for tetrahedral symmetry is that the 3-fold axes of symmetry carried by the two components must intersect in three-dimensional space, at an angle of precisely 70.5° (i.e., $\arccos(1/3)$). Thus, as a first step in design selection, combinatorial choices for the two components and the linker length were tested; the two termini were joined by an idealized α -helix and the angle between the two symmetry axes generated were calculated (based on vector dot products) along with the closest approach distance between the two symmetry axes. Of course, the likelihood that any construct would satisfy such constraints exactly is zero, leading to an important challenge in design selection. In our initial fusion-based design studies,^{14,37} we adopted a naive approach, essentially favoring design candidates where the two criteria (angle and separation distance between axes) were within some limits, with the expectation that a certain amount of helix bending would permit assembly. In the present study we employed a more structurally informed approach.

In our evaluation and selection of suitable candidates, we considered whether particular geometric perturbations—i.e., those needed to make a given design comply with the symmetry axis intersection requirements—could be accommodated by naturally allowable helix deformation. For illustration, consider a scenario where the geometric perturbation needed to meet the idealized angular design requirements—here, intersection at an angle of 70.5° —would require the connecting helix to be stretched along its axis. That mode of deformation is generally not observed in α helical segments, so even a minor deviation along that mode of deformation would not be a suitable starting point for design.

We developed an algebraic treatment of natural α -helix variability to aid in evaluating design candidates. Briefly, we examined the variation across natural α -helical structures using linear algebra and normal-mode analysis. By characterizing the degree to which different modes of deformation are sampled in

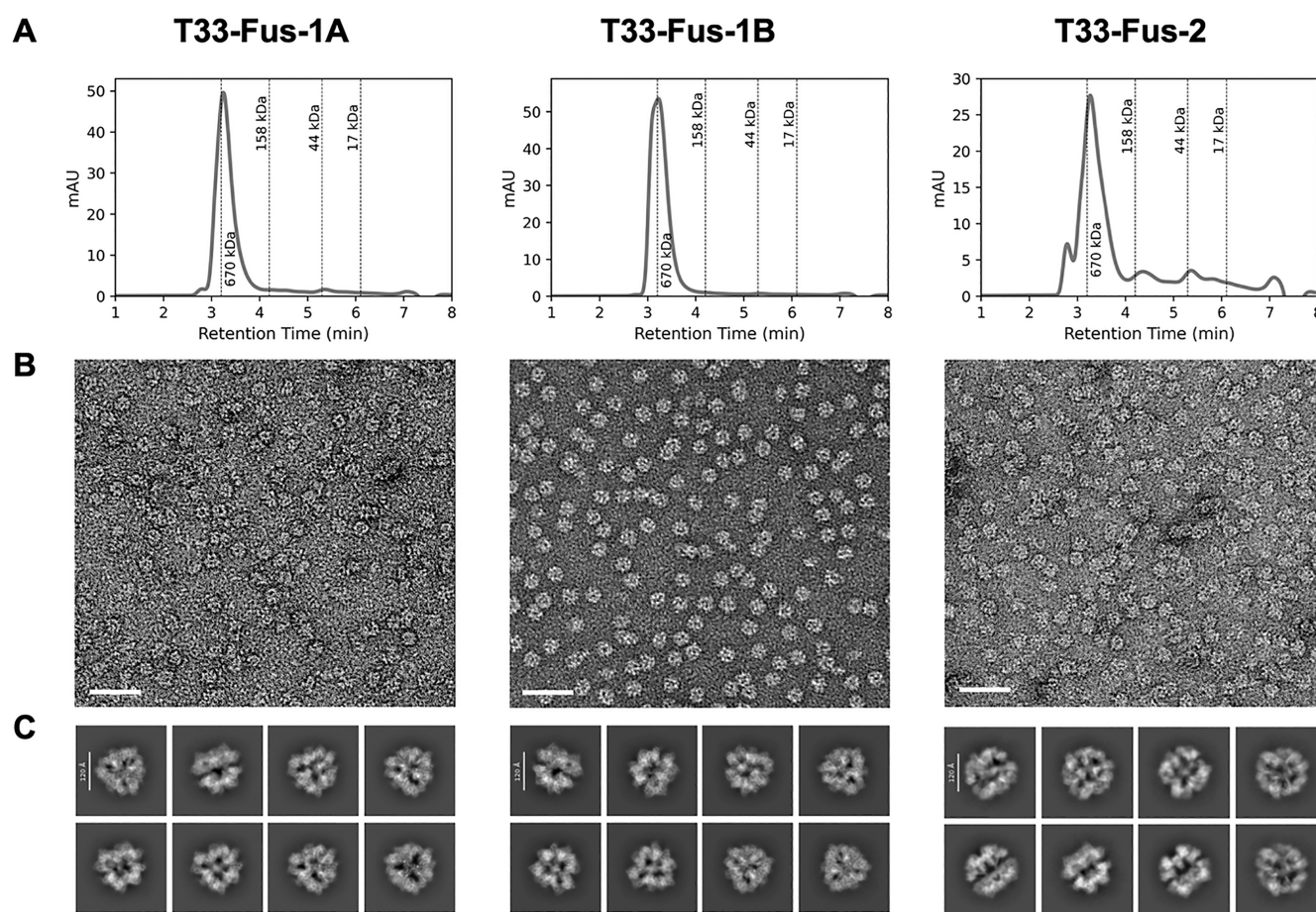


Figure 2. Experimental characterization of assembled protein cages. (A) SEC chromatograms of the three successful designs showing a single major peak at the expected elution volume for a 12-subunit assembly. (B) Representative negative-stain TEM micrographs revealing ~ 150 Å particles with apparent tetrahedral symmetry. Scale bar is 50 nm. (C) 2D classes showing different views of the particles from the cryo-EM data.

natural proteins, the presumptive helical deformation that would be required to idealize the symmetry axis requirements for any given design candidate could be evaluated against the allowable modes of α -helix variability (see [Methods](#)). The resulting measure of helical deformation, as a pseudo energy, was used as the primary score for selecting designs.

Regarding other design metrics, candidate designs where the fusion geometry would lead to undue collisions between subunits were discarded. Cases where incidental contacts (or near-contacts) occurred were retained; in those cases, limited amino acid optimization was also undertaken at those positions. We further evaluated the interior cavity radius of candidate designs, discarding outlying cases with very open structures, which were judged to be subject to potentially problematic collapse. Finally, we restricted our attention to cases where both the N- and C-termini of the fusion subunit were exposed on the exterior surface, anticipating applications involving further fusion to additional peptides, antigens, or effector domains. Fifteen distinct design candidates were selected for sequence optimization by machine learning.

The ProteinMPNN model²⁴ was used to generate a series of sequence variants for each candidate design. Sequence design was performed in the context of intact 12-subunit symmetric assemblies. In all cases, the residues in the linking helix and any residues in the component domains in direct contact with those residues were subject to design. Any points of incidental contact between the multiple chains were also included in

design. The remainder of the protein sequences were held fixed at their wild-type identities. For each design, 100 sequence designs were generated by ProteinMPNN, after which 2–4 constructs were chosen for experimental evaluation based on probability scores and similarity to the consensus sequence ([Table S3](#)).

Experimental Characterization of Designed Cages

Genes encoding the top 2–6 fusion protein designs across 15 sequence families were codon-optimized for expression in *Escherichia coli* and synthesized commercially. Proteins were expressed in BL21(DE3) LOBSTR cells and purified by immobilized metal affinity chromatography (IMAC) followed by size-exclusion chromatography (SEC). Of the 45 designs tested, five produced a prominent SEC peak eluting at a volume consistent with the expected molecular weight of the 12-subunit tetrahedral assembly, of which three were selected for follow-up characterization. Hereafter those designs are referred to as T33-Fus-1A and T33-Fus-1B (~ 633 kDa with A or B referring to sequence variants) and T33-Fus-2 (~ 638 kDa). The T33-Fus-1 variants are based on fusion of PDB 3E35 and 4E38, while T33-Fus-2 is based on fusion of PDB 2WAM and 3SY. The remaining designs either failed to express in soluble form, eluted as aggregates in the void volume, or showed prominent peaks corresponding to monomeric or lower-order oligomeric species.

An analysis of amino acid sequences showed that the two sequence variants of T33-Fus-1 differ from each other at 6 positions, and from the starting fusion model (prior to ProteinMPNN) at 12 positions for T33-Fus-1A and 10 positions for T33-Fus-1B (Figure S1). For the second construct, T33-Fus-2 differs from the starting fusion model at 12 positions (Figure S2).

SDS-PAGE analysis of the SEC peak fractions confirmed the presence of a single band at the expected subunit molecular weight for each of the three successful designs (Figures 2A and S3 and Table S1). Negative-stain transmission electron microscopy (NS-TEM) of the peak fractions revealed populations of well-dispersed, roughly spherical particles with diameters of approximately 150 Å (15 nm) for both designs, consistent with the dimensions predicted from the computational models (Figure 2B). The particle images showed apparent 2-fold and 3-fold symmetry views characteristic of tetrahedral objects, supporting the intended cage architecture (Figure 2C). No large aggregates or fibrillar structures were observed.

Cryo-EM Structure Determination and Identification of Assembly Intermediates

All three SEC-positive designs (T33-Fus-1A, T33-Fus-1B and T33-Fus-2) were taken forward to high-resolution cryo-EM analysis. Purified proteins were vitrified on Quantifoil grids and imaged on a 200 kV Glacios microscope equipped with a Falcon 4 detector. Image processing and 3D reconstruction yielded density maps at overall resolutions of 3.0 Å (T33-Fus-1A), 3.9 Å (T33-Fus-1B) and 3.8 Å (T33-Fus-2), determined by the gold-standard FSC criterion at the 0.143 threshold (Figures 3A,E, S4, and S5). The refined atomic models (Figure 3B,F) confirmed the intended tetrahedral T33 architecture. The maps showed clearly resolved secondary structure elements, side-chain densities for bulky residues, and the continuous α -helical linker connecting the two trimeric domains, enabling unambiguous model building (Figure 3C,G). Because the structures of T33-Fus-1A and T33-Fus-1B are nearly identical at the backbone level while the resolution for the former is significantly better, further detailed analysis on this construct was based on T33-Fus-1A.

For T33-Fus-1A and T33-Fus-2, the experimentally determined structures agree remarkably well with the computational models. Superpositions of the cryo-EM structures onto the computational design models yielded overall backbone RMSD values of 0.92 and 2.51 Å for T33-Fus-1A and T33-Fus-2 respectively when overlapping a single subunit. Slightly higher RMS deviations are observed when overlapping the entire 12-subunit assemblies: 2.0 Å for T33-Fus-1A and 4.9 Å for T33-Fus-2. Given their diameters of approximately 170 Å in both cases, those deviations correspond to fractional errors of only 1.2% and 1.9% for the two constructs.

As expected, the agreement between observed and designed models is nearly exact for the trimeric cores, while relatively minor deviations occur in the helical linker region. The linker helix is continuous and well-ordered in the density maps for both constructs, indicating that the ProteinMPNN-designed junction sequences support stable helical geometry in the context of the assembled cage. Indeed, an examination of the helical fusion segments shows that the predicted helical bending, which was an important element in the design strategy, is recapitulated in the observed structures for both

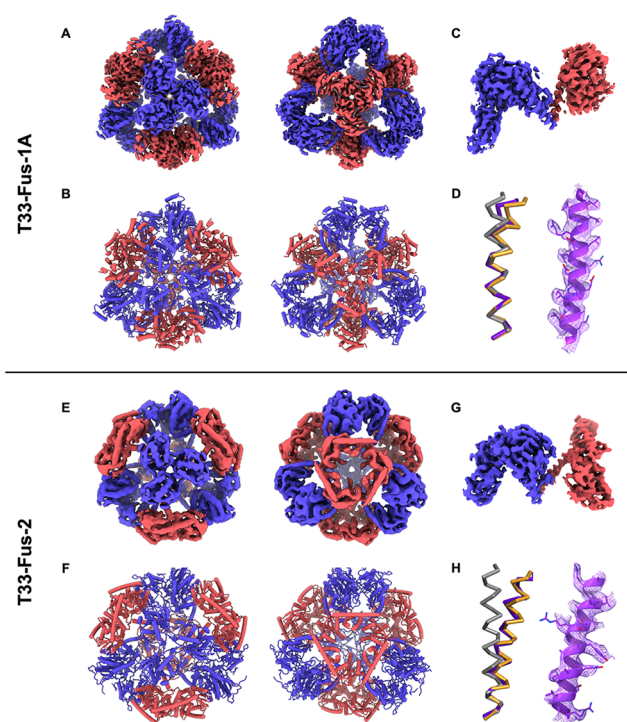


Figure 3. Cryo-EM structures of designs T33-Fus-1A and T33-Fus-2. (A, E) Cryo-EM density map viewed along the 3-fold axis of the N-terminal trimer domain (left) and 180° rotated cryo-EM map to show the 3-fold axis of the C-terminal trimer domain (right) for T33-Fus-1A and T33-Fus-2, respectively; (B, F) Refined atomic model showing the tetrahedral arrangement of two distinct trimeric domains (N-terminal trimer, blue; C-terminal trimer, red) connected by helical linkers, viewed along the 3-fold axis of the N-terminal trimer domain (left) and 180° rotated model to show the 3-fold axis of the C-terminal trimer domain (right) for T33-Fus-1A and T33-Fus-2, respectively; (C, G) Close-up view of the cryo-EM structure for a single subunit for T33-Fus-1A and T33-Fus-2, respectively; (D, H) An alignment of the modeled α helical linker (orange) in the T33-Fus-1A (D) and T33-Fus-2 cage (H), the observed structure of the α -helix (purple), and an idealized α -helix (gray) (left). The helices were aligned by optimal overlap at their N-terminal five amino acids. The deviation at the C-terminal end highlights the helix bending that was introduced in the design model based on an algebraic treatment of α -helix deformation (see Methods). The modeled bending was largely recapitulated in the observed cryo-EM structure. A fit of the refined structure to the density map is shown (right).

T33-Fus-1A (Figure 3D) and T33-Fus-2 (Figure 3H). The observed helical segments overlap much more closely with the slightly bent design models than they do with an idealized α helix. The T33-Fus-2 case is particularly striking; the deviation from ideal helix geometry is pronounced, while the match between model and observed geometry is remarkable (Figure 3H).

DISCUSSION

The present results demonstrate useful advances for applying the helical fusion strategy to protein cage design. The consideration of permissible modes of helix bending or deformation allows for more rational geometric scoring of candidate designs, while ML-based sequence design provides a rational approach for choosing amino acid sequences at positions where non-native residues are called for (e.g., at incidental contacts or in the linker itself where there is no

native sequence). Prior structural studies on fusion-based cages provide important points of comparison for both technical advances introduced in the present study.

Concerning the issue of helical deformation, a series of structural studies on prior fusion-based assemblies emphasized the presence of irregularly shaped assemblies or alternative assembly forms. The first designed protein cage by Padilla et al. in 2001 is notable.¹⁴ A series of structural studies^{38,39} showed that diverse forms were typically obtained in X-ray crystallographic experiments. These were generally the result of compression and asymmetric deformation, sometimes reaching 18 Å RMS deviations. Studies on other fusion cages showed deformations sufficient to cause alternative assembly: symmetry T and D3 instead of O (17) and symmetry D3 instead of T.^{40,41} In these prior cases of helical fusion (with the exception of Lai et al. 2015),⁴⁰ the natural oligomeric components appeared to be intact, with unexpected helical deformation being the cause of the design deficiencies. The structures obtained in the present study, where natural helical variation was analyzed during design, are much more robust to those deficiencies. The successful designs elucidated here are only a subset of the total designed sequences that were experimentally tested. Nonetheless, they represent the first fusion-based cages we are aware of where well-ordered, homogeneous and symmetric architectures could be obtained in abundance and characterized in atomic detail.

Similarly, the importance of choosing optimal sequences in and around the helical fusion is made clear by the history of fusion-based cages. Polymorphism that challenged the structural characterization of the original design¹⁴ was eventually overcome by judicious mutations in the linker region.^{37–39} Those studies predated robust software for sequence design. The present work, exploiting some of the latest machine learning tools for sequence design at key positions, led to highly ordered and symmetric assemblies.

Looking forward, the computational pipeline employed here is readily extensible. Higher-order symmetries (octahedral, icosahedral) and alternative oligomer combinations (e.g., trimer-tetramer or trimer-pentamer fusions) could be explored using the same framework. Integration with generative backbone design tools such as RFdiffusion²⁸ could further expand the accessible design space by enabling de novo generation of protein domains suitable for joining by helical junction backbones rather than relying solely on natural oligomeric components of known structure. Indeed, other recent studies have exploited helical fusions in pipelines using machine learning methods.^{33,35} In the context of possible extensions to the present work, ML methods could be further exploited to simultaneously consider α -helix bending and sequence design, and to include structure prediction metrics in scoring.

CONCLUSIONS

Designed protein cages serve as platforms for a growing range of applications, including multivalent antigen display for vaccine development,^{1,13} targeted drug delivery,³ protease-triggered disassembly for controlled release,⁴² and rigid scaffolds for cryo-EM imaging of small therapeutic proteins.^{43–45} Reliably engineering such novel architectures remains challenging. Improved strategies therefore offer expanded opportunities for successful development. Recent computational developments have played a critical role in expanding the set of available protein cages designed using

interface design approaches,^{25–27} or de novo fold generation.^{28,30,46} The computational developments in the present work provide similar advances for fusion-based design approaches, in particular by incorporating ProteinMPNN into the sequence design stages.²⁴ Notwithstanding the utility of ML methods demonstrated here and in other related work, classical mechanical principles remain important. As illustrated here, combining an algebraic or mechanical treatment of structural variability with modern ML methods diversifies the paths available for engineering robust protein cages and perhaps even more complex architectures.

The design and experimental work here provide two new protein cages (and sequence variations thereon), with homogeneous and well-ordered three-dimensional structures not achievable before for fusion-based constructions. The protocol described, including with additional ML features, could be expanded to generate further sets of protein cages based on fusions. Besides not depending on the creation of novel protein–protein interfaces, fusion-based constructions offer additional counterpoints to interface-design methods. For example, the fusion approach typically involves single-component construction, whereas interface design approaches typically, though not always (see King et al.),²¹ involve multicomponent design. In some settings, single subunit compositions are expected to offer manufacturing advantages. Fusion-type cages have also been noted to be particularly suited for conversion to protease-sensitive forms with potential for triggered disruption.⁴² In addition, by avoiding the need for interface design, fusion-based approaches offer prospects for creating novel architectures where the complete sequence deviates from fully native sequences by only a small number of amino acids. That feature could be useful in future applications where limiting immunogenicity is a goal.

METHODS

Computational Design

Algebraic Analysis of Natural α -Helix Deformation. Possible strategies for incorporating minor helix deformation into designed proteins include dynamical simulation, machine learning, clustering, and algebraic approximations based on known structural data. A strategy was developed here based on the latter. A set of α helical segments of length 7 (roughly 2 turns) were extracted from high resolution structures ($d_{\text{min}} < 1.5$ Å) in the PDB.³⁶ That ensemble was then aligned spatially at the N-terminal three residues, and oriented in a canonical frame of reference based on the bond vectors from the first C- α to its amide nitrogen and its bonded carbonyl carbon. The pattern of observed deformation across the ensemble was then approximated as a problem of characterizing the rigid body space spanned by the set of $i + 6$ residues from the ensemble. While rigid body space (in three dimensions) depends on six variables, for convenience the problem was treated in an augmented 9-dimensional space, based on the three x, y, z coordinates of three reference atoms (C- α , N and C atoms) from the $i + 6$ residues. The symmetric 9×9 metric mass tensor from that ensemble was calculated, and the eigenvalues and eigenvectors extracted. A linear treatment of this type is suitable as long as rotational magnitudes are relatively small, as explained by Goldstein,⁴⁷ and that was judged to be the case here. As expected, the three smallest eigenvalues (of 9) were small enough to be ignored (accounting for less than 0.2% of the variation), leaving 6 that embody the modes and magnitudes of natural helix variation, to the degree that that variation can be captured by the atomic positions at the ends of a semiflexible segment. Interestingly, 90% of the total structural variation (i.e., the helix deformation) was captured by just the top 3 eigenvalues; 96% was captured by the top 4. The remaining two possible modes are practically unobserved. The finding that

certain mathematically permissible modes of structural variation of the helix (e.g., direct stretching) are essentially forbidden is consistent with intuition and an understanding of atomic bonding but has not been described before in this form as far as we are aware. Having extracted the permissible modes and magnitudes of bending (in the form of atomic shifts of the $i + 6$ residue in an established frame of reference), any helical conformation subject to some degree of nonideality can be evaluated based on standard equations for probability using a quadratic Gaussian form. Note that this approach to characterizing bending was naive to underlying amino acid sequence effects and to the influences of surrounding structural features. The important interplay between those effects is recaptured in the sequence design step.

Construction of Fusion Models. Homotrimeric protein structures were retrieved from the PDB.³⁶ Terminal α -helices were identified using DSSP.⁴⁸ Combinatorial fusions were generated computationally by joining the helical ends of the two components, inserting an ideal helical linker with lengths up to 7 residues, or allowing the terminal ends of the two components to overlap by 2 residues (an effective linker length of -2). For each such potential construct, the symmetry axes carried by the two components were analyzed algebraically to obtain the angle between the axes and their separation at closest approach. Cases where those two values were within limits that might be mitigated by helix bending were retained for optimization. Perturbations to helix geometry were introduced according to the principal modes of variation obtained above. Specifically, the most probable deformation consistent with the exact symmetry requirements was obtained by sampling small deformation steps with magnitudes dictated by the eigenvalues associated with the principal modes and identifying the closest match to the symmetry requirements in each step. The total deformation was scored as a pseudoenergy based on the sum of squares of deformation magnitudes divided by normal mode eigenvalues. Candidate fusion subunits were evaluated computationally for clashes between the two fused domains and between symmetry-related copies of the fusion subunit in the context of tetrahedral symmetry.

Sequences for the linker region and flanking positions were designed using ProteinMPNN.²⁴ A total of 100 sequences were generated per backbone. Designs for experimental consideration were chosen by first filtering for the top 50% based on probability scores and then by scoring based on divergence from the consensus amino acid at each position. Designs closest to the consensus sequence overall were prioritized.

Protein Expression and Purification

Genes encoding the designed fusion proteins were codon-optimized for *E. coli* expression and synthesized by GenScript. Constructs were cloned into the pET-30a vector with a C-terminal His₆ tag. Plasmids were transformed into *E. coli* LOBSTR-BL21(DE3) cells. Cultures were grown in TB medium supplemented with kanamycin at 37 °C to an OD₆₀₀ of 1.2, induced with 0.25 mM IPTG, and incubated at 15 °C for 16 h. Cells were harvested by centrifugation, resuspended in lysis buffer (30 mM Tris, 200 mM NaCl, 1 mM TCEP, 20 mM imidazole, 0.1% LDAO, 1 mM PMSF, pH 8.0) supplemented with benzonase nuclease, and lysed by sonication. Clarified lysate was applied to Ni-Charged MagBeads, washed with wash buffer (30 mM Tris, 200 mM NaCl, 1 mM TCEP, 20 mM imidazole, 50 mM sucrose, 0.01% Tween 20, pH 8.0), and eluted with elution buffer (30 mM Tris, 200 mM NaCl, 1 mM TCEP, 500 mM imidazole, 50 mM sucrose, 0.01% Tween 20, pH 8.0). The purified proteins were dialyzed into storage buffer (30 mM Tris, 200 mM NaCl, 5 mM EDTA, 1 mM TCEP, 50 mM sucrose, 0.01% Tween 20, pH 8.0) before being further purified by size-exclusion chromatography (SEC) on a Superose 6 Increase 10/300 GL column equilibrated in 30 mM Tris, 200 mM NaCl, 5 mM EDTA, pH 8.0 and running at a flow rate of 0.35 mL/min.

Negative-Stain Electron Microscopy

Purified protein at 0.05–0.2 mg/mL was applied to glow-discharged carbon-coated copper grids, blotted, and stained with 2% (w/v) uranyl acetate. Micrographs were collected on a Tecnai G2 Spirit

microscope equipped with a 4k × 4k TemCam-F416 CMOS camera (TVIPS) operating at 120 kV and on a 200 kV JEM-2200FS (JEOL) microscope equipped with a K3 direct electron detector (Gatan).

Cryo-EM Sample Preparation, Data Collection, and Processing

For cryo-EM, purified samples of designs T33-Fus-1A, T33-Fus-1B, and T33-Fus-2, were applied at a concentration of 0.3 mg/mL to glow-discharged Quantifoil R1.2/1.3 300 mesh copper grids, blotted for 3 s at 4 °C and 100% humidity using a Vitrobot Mark IV (Thermo Fisher Scientific), and plunge-frozen in liquid ethane. Data were collected on a Glacios microscope at 200 kV equipped with a Falcon 4. Movies were recorded at a calibrated pixel size of 0.93 Å/pixel with a total dose of $\sim 40 \text{ e}^-/\text{\AA}^2$.

All structures were determined using cryoSPARC.⁴⁹ For T33-Fus-1A, a preliminary analysis in cryoSPARC Live was used to generate an *ab initio* model and a low-resolution refined map, which served as input for 2D template generation. A total of 4200 movies were imported in EER format with 40 frames per movie. After patch motion correction and patch CTF estimation, micrographs were manually curated using thresholds of -0.5 to $-3.5 \text{ }\mu\text{m}$ average defocus and a maximum CTF-fit resolution of 4.5 Å, retaining 3237 micrographs. Particles were picked with the template picker (200 Å particle diameter) using the 2D templates generated during cryoSPARC Live preprocessing and curated through particle inspection using local power and normalized cross-correlation thresholds. A total of $\sim 413,000$ particles were extracted in a 416-pixel box binned to 120 pixels. 2D classification was used to select $\sim 109,000$ particles, followed by two nonuniform refinement steps: the first in C1 (using the *ab initio* model from preprocessing as the initial volume) and the second with tetrahedral (T) symmetry imposed. The refined particles were re-extracted at lower binning (416-pixel box binned to 320 pixels) and subjected to several rounds of homogeneous and nonuniform refinement with T symmetry. The resulting volume served as input for per-particle reference-based beam-induced motion correction, and the particles were re-extracted unbinned in a 416-pixel box. These $\sim 99,000$ motion-corrected particles were subjected to further nonuniform refinement. To obtain a less anisotropic consensus map of the whole protein cage, the orientation of the particles was then rebalanced (retaining $\sim 33,000$ particles), yielding a 3.70 Å map. For focused refinement, the full (prerebalancing) particle set was subjected to symmetry expansion under T symmetry (~ 1.2 million expanded particles), locally refined, and then rebalanced (retaining $\sim 435,000$ particles). A final local refinement with a mask covering the asymmetric unit yielded a reconstruction at 2.99 Å (Figure S6 and Table S2).

For T33-Fus-1B, a similar workflow was followed. A total of 4600 movies were imported into cryoSPARC in EER format with 40 frames per movie. A subset of 200 movies was preprocessed to generate an *ab initio* reconstruction and a preliminary low-resolution refined map for 2D template generation. After patch motion correction and patch CTF estimation of all movies, the micrographs were manually curated using thresholds of -0.5 to $-3.5 \text{ }\mu\text{m}$ average defocus and a maximum CTF-fit resolution of 4.5 Å, retaining 4000 micrographs. Particles were picked with the template picker (200 Å particle diameter) using the templates generated during preprocessing. After particle inspection, $\sim 451,000$ particles were extracted in a 416-pixel box binned to 120 pixels. One round of 2D classification retained $\sim 344,000$ particles, which were used to generate an *ab initio* reconstruction. The best resulting map served as input for two nonuniform refinement steps: first in C1 and then with T symmetry imposed. The refined particles were re-extracted in a 416-pixel box binned to 320 pixels, and the $\sim 295,000$ retained particles were further refined by nonuniform refinement, followed by per-particle reference-based beam-induced motion correction. The orientation of the $\sim 295,000$ corrected particles was rebalanced (retaining $\sim 184,000$ particles) to obtain a consensus refined map by homogeneous refinement (4.24 Å). The particles were then symmetry-expanded under T symmetry (~ 2.6 million particles), and a final local

refinement with a mask covering the asymmetric unit was performed, yielding a map at 3.91 Å resolution (Figure S7 and Table S2).

For T33-Fus-2, a subset of 1200 movies was preprocessed to generate an *ab initio* reconstruction and a preliminary low-resolution refined map for 2D template generation. After patch motion correction and patch CTF estimation of all movies (4086 movies), the micrographs were manually curated, retaining 2893 micrographs. Particles were picked with the template picker (250 Å particle diameter) using the templates generated during preprocessing. After particle inspection, ~202,000 particles were extracted in a 320-pixel box binned to 96 pixels. One round of 2D classification retained ~47,000 particles, which were refined by homogeneous refinement, first in C1 and then with T symmetry imposed. The particles were re-extracted in a 320-pixel box binned to 256 pixels and refined again by homogeneous refinement with T symmetry, yielding a map at 5.11 Å. The particles were then symmetry-expanded under T symmetry (~561,000 particles), and a final local refinement with a mask covering the asymmetric unit was performed, yielding a map at 3.76 Å resolution (Figure S8 and Table S2).

Atomic models were built into the density using Coot⁵⁰ and refined with Phenix.⁵¹

■ ASSOCIATED CONTENT

Data Availability Statement

The cryo-EM structures of the protein cages, and their associated atomic coordinates, have been deposited into the Electron Microscopy Data Bank (EMDB) and the Protein Data Bank (PDB) with EMDB accession codes EMD-59307, EMD-59308, EMD-59315, EMD-59316, EMD-59317, EMD-59318, and PDB accession codes 33AR, 33AS, 33AT, 33 AU, 33AV, 33AW, respectively. The datasets have been deposited in the Electron Microscopy Public Image Archive (EMPIAR) with EMPIAR accession codes EMPIAR-13888, EMPIAR-13889 and EMPIAR-13890. The sequences of the protein designs are included in the Supporting Information.

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/jacs.6c11491>.

Supporting figures and tables with sequence alignments, SDS-PAGE gel, cryo-EM maps data summary, cryo-EM data processing workflows, protein sequences, cryo-EM data collection, image analysis, modeling, refinement, and validation statistics (PDF)

Sequences for the designed trimer-trimer helical fusion cages (Table S3) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Roger Castells-Graells – Structural Biology Programme, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain; orcid.org/0000-0002-3985-6194; Email: rcastellsg@cnio.es

Authors

Pablo San Segundo-Acosta – Structural Biology Programme, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain

Johanne Le Coq – Structural Biology Programme, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain

Jasminka Boskovic – Structural Biology Programme, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain

Robin A. Aglietti – AvimerBio, San Diego, California 92121, United States

Peter Bowers – AvimerBio, San Diego, California 92121, United States

Todd O. Yeates – AvimerBio, San Diego, California 92121, United States

Complete contact information is available at:

<https://pubs.acs.org/doi/10.1021/jacs.6c11491>

Notes

The authors declare the following competing financial interest(s): T.O.Y., P.B., R.A.A., and R.C.-G. hold equity in AvimerBio. T.O.Y. and P.B., are inventors on a relevant patent application.

■ ACKNOWLEDGMENTS

We thank Duilio Cascio, Mark Arbing, Génesis Falcon, Alex Lisker, Michael Sawaya, Oscar Llorca, Rafael Fernández-Leiro, Lucas Tafur, and the members of the CNIO Structural Biology Programme for valuable feedback and for providing helpful support. We thank Pablo Guerra from the IBMB-CSIC CryoEM Platform for assistance during the sample preparation and microscope data acquisition. We acknowledge funding from Project IU16-014045 (CRYO-TEM) from Generalitat de Catalunya and by “ERDF A way of making Europe”, by the European Union. We acknowledge ALBA Synchrotron for access and support of the cryo-EM facilities under BAG Proposal No CNIO-BAG 20250370388. We thank the Fundación Occident Visiting Researchers Programme at the CNIO. We thank Fundació Catalunya La Pedrera for its continuous support. R.C.-G. and P.S.-A., Fellowships within the Generación D initiative, Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), are funded by the European Union NextGenerationEU funds, through the Plan de Recuperación, Transformación y Resiliencia (PRTR). The Castells-Graells laboratory is supported by the Agencia Estatal de Investigación (AEI/10.13039/501100011033), Ministerio de Ciencia, Innovación y Universidades and cofunded by the European Regional Development Fund (ERDF-EU) [PID2024-161096NA-I00 to R.C.-G.], and is also supported by the Carlos III Health Institute (ISCIII) via CNIO.

■ REFERENCES

- (1) Boyoglu-Barnum, S.; Ellis, D.; Gillespie, R. A.; et al. Quadrivalent influenza nanoparticle vaccines induce broad protection. *Nature* **2021**, *592*, 623–628.
- (2) Edwardson, T. G. W.; Levasseur, M. D.; Tetter, S.; et al. Protein Cages: From Fundamentals to Advanced Applications. *Chem. Rev.* **2022**, *122*, 9145–9197.
- (3) Edwardson, T. G. W.; Mori, T.; Hilvert, D. Rational engineering of a designed protein cage for siRNA delivery. *J. Am. Chem. Soc.* **2018**, *140*, 10439–10442.
- (4) Gladkov, N.; Scott, E. A.; Meador, K.; et al. Design of a symmetry-broken tetrahedral protein cage by a method of internal steric occlusion. *Protein Sci.* **2024**, *33*, No. e4973.
- (5) Gupta, A. K.; Esih, H.; Gradišar, H.; Jerala, R. Self-assembling protein cages: from coiled-coil module to machine learning-driven de novo design of next-generation biomaterials. *Mater. Adv.* **2026**, *7*, 1378–1396.
- (6) Hendricks, G. G.; Grigoryan, L.; Navarro, M. J.; et al. Computationally designed mRNA-launched protein nanoparticle immunogens elicit protective antibody and T cell responses in mice. *Sci. Transl. Med.* **2025**, *17*, No. eadu2085.

- (7) Kanekiyo, M.; Wei, C. J.; Yassine, H. M.; et al. Self-assembling influenza nanoparticle vaccines elicit broadly neutralizing H1N1 antibodies. *Nature* **2013**, *499*, 102–106.
- (8) Lee, E. J.; Gladkov, N.; Miller, J. E.; Yeates, T. O. Design of Ligand-Operable Protein-Cages That Open Upon Specific Protein Binding. *ACS Synth. Biol.* **2024**, *13*, 157–167.
- (9) Miller, J. E.; Castells-Graells, R.; Arbing, M. A.; et al. Design of Beta-2 Microglobulin Adsorbent Protein Nanoparticles. *Biomolecules* **2023**, *13*, No. 1122.
- (10) Steinmetz, N. F.; Lim, S.; Sainsbury, F. Protein cages and virus-like particles: from fundamental insight to biomimetic therapeutics. *Biomater. Sci.* **2020**, *8*, 2771–2777.
- (11) Stupka, I.; Azuma, Y.; Biela, A. P.; et al. Chemically induced protein cage assembly with programmable opening and cargo release. *Sci. Adv.* **2022**, *8*, No. eabj9424.
- (12) Szyszka, T. N.; Siddique, R.; Loustau, A.; et al. High-Fidelity In Vitro Packaging of Diverse Synthetic Cargo into Encapsulin Protein Cages. *Angew. Chem., Int. Ed.* **2025**, *64*, No. e202422459.
- (13) Walls, A. C.; Fiala, B.; Schäfer, A.; et al. Elicitation of Potent Neutralizing Antibody Responses by Designed Protein Nanoparticle Vaccines for SARS-CoV-2. *Cell* **2020**, *183*, 1367–1382.
- (14) 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. U.S.A.* **2001**, *98*, 2217–2221.
- (15) Laniado, J.; Yeates, T. O. A complete rule set for designing symmetry combination materials from protein molecules. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 31817–31823.
- (16) Yeates, T. O. Geometric Principles for Designing Highly Symmetric Self-Assembling Protein Nanomaterials. *Annu. Rev. Biophys.* **2017**, *46*, 23–42.
- (17) Lai, Y.-T.; Reading, E.; Hura, G. L.; et al. Structure of a designed protein cage that self-assembles into a highly porous cube. *Nat. Chem.* **2014**, *6*, 1065–1071.
- (18) Badieyan, S.; Sciore, A.; Eschweiler, J. D.; et al. Symmetry-Directed Self-Assembly of a Tetrahedral Protein Cage Mediated by de Novo-Designed Coiled Coils. *ChemBioChem* **2017**, *18*, 1888–1892.
- (19) Cristie-David, A. S.; Koldewey, P.; Meinen, B. A.; Bardwell, J. C. A.; Marsh, E. N. G. Elaborating a coiled-coil-assembled octahedral protein cage with additional protein domains. *Protein Sci.* **2018**, *27*, 1893–1900.
- (20) Sciore, A.; Su, M.; Koldewey, P.; et al. Flexible, symmetry-directed approach to assembling protein cages. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 8681–8686.
- (21) King, N. P.; Sheffler, W.; Sawaya, M. R.; et al. Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **2012**, *336*, 1171–1174.
- (22) King, N. P.; Bale, J. B.; Sheffler, W.; et al. Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **2014**, *510*, 103–108.
- (23) Bale, J. B.; Gonen, S.; Liu, Y.; et al. Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **2016**, *353*, 389–394.
- (24) Dauparas, J.; Anishchenko, I.; Bennett, N.; et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **2022**, *378*, 49–56.
- (25) de Haas, R. J.; Brunette, N.; Goodson, A.; et al. Rapid and automated design of two-component protein nanomaterials using ProteinMPNN. *Proc. Natl. Acad. Sci. U.S.A.* **2024**, *121*, No. e2314646121.
- (26) Haas, C. M.; Jasti, N.; Dosey, A.; et al. From sequence to scaffold: Computational design of protein nanoparticle vaccines from AlphaFold2-predicted building blocks. *Proc. Natl. Acad. Sci. U.S.A.* **2025**, *122*, No. e2409566122.
- (27) Meador, K.; Castells-Graells, R.; Aguirre, R.; et al. A suite of designed protein cages using machine learning and protein fragment-based protocols. *Structure* **2024**, *32* (6), 751–765.
- (28) Watson, J. L.; Juergens, D.; Bennett, N. R.; et al. De novo design of protein structure and function with RFdiffusion. *Nature* **2023**, *620*, 1089–1100.
- (29) Dowling, Q. M.; Park, Y. J.; Fries, C. N.; et al. Hierarchical design of pseudosymmetric protein nanocages. *Nature* **2025**, *638*, 553–561.
- (30) Haas, C. M. et al. De novo design of protein nanoparticles with integrated functional motifs *bioRxiv* 2026 DOI: 10.64898/2025.12.19.695620.
- (31) Lee, S.; Kibler, R. D.; Ahn, G.; et al. Four-component protein nanocages designed by programmed symmetry breaking. *Nature* **2025**, *638*, 546–552.
- (32) Rankovic, S.; Carr, K. D.; Decarreau, J.; et al. Computational design of bifaceted protein nanomaterials. *Nat. Mater.* **2025**, *24*, 1635–1643.
- (33) Hsia, Y.; Mout, R.; Sheffler, W.; et al. Design of multi-scale protein complexes by hierarchical building block fusion. *Nat. Commun.* **2021**, *12*, No. 2294.
- (34) Jacobs, T. M.; Williams, B.; Williams, T.; et al. Design of structurally distinct proteins using strategies inspired by evolution. *Science* **2016**, *352*, 687–690.
- (35) Vulovic, I.; Yao, Q.; Park, Y. J.; et al. Generation of ordered protein assemblies using rigid three-body fusion. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, No. e2015037118.
- (36) Berman, H. M.; Westbrook, J.; Feng, Z.; et al. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (37) Lai, Y.-T.; Cascio, D.; Yeates, T. O. Structure of a 16-nm cage designed by using protein oligomers. *Science* **2012**, *336*, 1129.
- (38) Lai, Y.-T.; Tsai, K.-L.; Sawaya, M. R.; Asturias, F. J.; Yeates, T. O. Structure and Flexibility of Nanoscale Protein Cages Designed by Symmetric Self-Assembly. *J. Am. Chem. Soc.* **2013**, *135*, 7738–7743.
- (39) Lai, Y.-T.; Hura, G. L.; Dyer, K. N.; et al. Designing and defining dynamic protein cage nanoassemblies in solution. *Sci. Adv.* **2016**, *2*, No. e1501855.
- (40) Lai, Y.-T.; Jiang, L.; Chen, W.; Yeates, T. O. On the predictability of the orientation of protein domains joined by a spanning alpha-helical linker. *Protein Eng., Des. Sel.* **2015**, *28*, 491–500.
- (41) Laniado, J.; Cannon, K. A.; Miller, J. E.; et al. Geometric Lessons and Design Strategies for Nanoscale Protein Cages. *ACS Nano* **2021**, *15*, 4277–4286.
- (42) Miller, J. E.; Srinivasan, Y.; Dharmaraj, N. P.; et al. Designing Protease-Triggered Protein Cages. *J. Am. Chem. Soc.* **2022**, *144*, 12681–12689.
- (43) Agdanowski, M. P.; Castells-Graells, R.; Sawaya, M. R.; et al. X-ray crystal structure of a designed rigidified imaging scaffold in the ligand-free conformation. *Acta Crystallogr., Sect. F: Struct. Biol. Commun.* **2024**, *80*, 107–115.
- (44) Castells-Graells, R.; Meador, K.; Arbing, M. A.; et al. Cryo-EM structure determination of small therapeutic protein targets at 3 Å-resolution using a rigid imaging scaffold. *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120*, No. e2305494120.
- (45) Liu, Y.; Huynh, D. T.; Yeates, T. O. A 3.8 Å resolution cryo-EM structure of a small protein bound to an imaging scaffold. *Nat. Commun.* **2019**, *10*, No. 1864.
- (46) Ingraham, J. B.; Baranov, M.; Costello, Z.; et al. Illuminating protein space with a programmable generative model. *Nature* **2023**, *623*, 1070–1078.
- (47) Goldstein, H. *Classical Mechanics*; Addison-Wesley Press, 1950.
- (48) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.
- (49) Punjani, A.; Rubinstein, J. L.; Fleet, D. J.; Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **2017**, *14*, 290–296.
- (50) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and development of Coot. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66*, 486–501.
- (51) Liebschner, D.; Afonine, P. V.; Baker, M. L.; et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr., Sect. D: Struct. Biol.* **2019**, *75*, 861–877.