UCLA

UCLA Electronic Theses and Dissertations

Title

Design of finite and infinite proteinaceous nanomaterials

Permalink

https://escholarship.org/uc/item/64v1g3dt

Author

Meador, Kyle

Publication Date

2023

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Design of finite and infinite proteinaceous nanomaterials

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Biochemistry, Molecular and Structural Biology

by

Kyle Meador

2023

© Copyright by

Kyle Meador

2023

ABSTRACT OF THE DISSERTATION

Design of finite and infinite proteinaceous nanomaterials

By

Kyle Meador

Doctor of Philosophy in Biochemistry, Molecular and Structural Biology

University of California, Los Angeles, 2023

Professor Todd O. Yeates, Chair

Over the early 21st century, structural biology has laid a robust foundation for our understanding of protein molecules. The atomic principles of their structure, how their sequence specifies such a structure, and conversely, how to enumerate an amino acid sequence to adopt a fold, are problems each closer to solved than not. Once the domain of material science and chemistry, materials of various shapes and sizes can now be made out of protein precursors. Designed protein building blocks have been synthesized that self-assemble into tetrahedral, octahedral, and icosahedral molecular cages, as well as infinitely ordered lattices such as two dimensional layers and three dimensional crystals. This thesis provides descriptions used to both create and apply protein nanomaterials towards the study of biochemical phenomenon. Particular interest is given to methods of searching for native like contacts and emulating their assembly into defined quaternary structures. By taking inspiration from nature and utilizing hypothesis driven symmetric materials engineering, I demonstrate the creation of proteins which form new materials in the laboratory and methods to engineer existing materials, including a high resolution imaging scaffold. These advances remove barriers to nanomaterials development and deepen the understanding of protein molecules.

The dissertation of Kyle George Meador is approved.

Gerard C. Wong

Heather Maynard

Jose A. Rodriguez

Todd O. Yeates, Committee Chair

University of California, Los Angeles

2023

Dedicated to my grandpa, George, whose passion, attention to detail, and convictions are foundations of this work. And my aunt, Jennifer, with whom the curiosity of these intellectual pursuits binds me, despite the energetic barrier of our family ties.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ACKNOWLEDGEMENTS

Chapter 3 of this dissertation is a version of a published manuscript: Laniado, J., Meador, K. & Yeates, T. O. A fragment-based protein interface design algorithm for symmetric assemblies. *Protein Eng., Des. Sel.* **34**, gzab008 (2021). Reproduced with permission. The research was conceived by T.O.Y. and J.L. The code was written by T.O.Y. and J.L. The fragment database was constructed by J.L. The example SCMs were constructed by J.L. and K.M. The post facto analysis of designed protein cages was performed by K.M. and J.L. The manuscript was prepared by T.O.Y., K.M. and J.L.

Chapter 4 of this dissertation contains a version of a manuscript in preparation: Meador, K. & Yeates, T. O. SymDesign align-helices: Automated pipeline for modeling protein fusions in hierarchical symmetric systems. (2023). The research was conceived by T.O.Y. The code was written by K.M and T.O.Y. The manuscript was prepared by K.M. and T.O.Y.

Chapter 4 also contains a version of a published manuscript: Castells-Graells, R., Meador, K., Arbing, M. A., Sawaya, M. R., Gee, M., Cascio, D., Gleave, E., Debreczeni, J. É., Breed, J., Leopold, K., Patel, A., Jahagirdar, D., Lyons, B., Subramaniam, S., Phillips, C. & Yeates, T. O. Cryo-EM structure determination of small therapeutic protein targets at 3 Å-resolution using a rigid imaging scaffold. *Proc. Natl. Acad. Sci. United States Am.* **120**, e2305494120 (2023). Reproduced with permission. R.C.-G., K.M., S.S., C.P., and T.O.Y. designed research; R.C.-G., K.M., M.A.A., M.R.S., M.G., D.C.,E.G., J.É.D., J.B., K.L., A.P., D.J., B.L., and S.S. performed research; R.C.-G., K.M., M.A.A., M.R.S., K.L., A.P., D.J., B.L.,S.S., C.P., and T.O.Y. analyzed data; and R.C.-G., K.M., and T.O.Y. wrote the paper.

Chapter 5 of this dissertation is a version of a submitted manuscript: Meador, K., Castells-Graells, R., Aguirre, R., Sawaya, M., Arbing, M., Sherman, T., Senarathne, C., & Yeates, T. O. A Suite of Designed Protein Cages Using Machine Learning Algorithms and Protein Fragment-Based Protocols. *Structure* (2023). Submitted. Conceptualization, KM and TOY; Methodology, KM, RCG, RA, MS, and TOY; Software, KM; Formal Analysis, KM; Investigation, KM, RCG, RA, MS, MA, TS, and CS; Resources, TOY; Data Curation, KM, RCG, and MS; Writing - Original Draft, KM and TOY; Writing - Review & Editing, KM and TOY; Visualization, KM and RCG; Supervision, KM and TOY; Funding Acquisition, TOY.

Finally, I would like to acknowledge the National Institutes of Health training grant (T32GM008496) for funding support.

VITA/BIOGRAPHICAL SKETCH

Education

University of Colorado Boulder, Boulder, Colorado, 2011 – 2014
B.A. in Molecular, Cellular, and Developmental Biology
B.A. in Neuroscience

Awards/Honors

Chemistry Biology Interface Training Grant: July 2018 – June 2021
Dorothy Radcliffe Dee Fellowship: September 2017
Graduation with Distinction: 2014
Arts & Sciences Dean's Scholarship: 2012 – 2014
Dean's List: 2012 – 2014

Publications

Gladkov, N., Scott, E. A., **Meador, K.,** Lee, E. J., Laganowsky, A. D., Yeates, T. O., Castells-Graells, R. Design of a symmetry-broken tetrahedral protein cage by a method of internal steric occlusion. *bioRxiv.* (2023) https://doi.org/10.1101/2023.11.08.566319

Meador, K., Castells-Graells, R., Aguirre, R., Sawaya, M., Arbing, M., Sherman, T., Senarathne, C., & Yeates, T. O. A Suite of Designed Protein Cages Using Machine Learning Algorithms and Protein Fragment-Based Protocols. *Structure.* (2023). Submitted. https://doi.org/10.1101/2023.10.09.561468

Castells-Graells, R., **Meador, K.**, Arbing, M. A., Sawaya, M. R., Gee, M., Cascio, D., Gleave, E., Debreczeni, J. É., Breed, J., Leopold, K., Patel, A., Jahagirdar, D., Lyons, B., Subramaniam, S., Phillips, C. & Yeates, T. O. Cryo-EM structure determination of small therapeutic protein targets at 3 Å-resolution using a rigid imaging scaffold. *Proc. Natl. Acad. Sci. United States Am.* 120, e2305494120 (2023). https://dx.doi.org/10.1073/pnas.2305494120

Laniado J., **Meador K.,** Yeates T. O., A fragment-based protein interface design algorithm for symmetric assemblies. *Protein Eng Des Sel 34, gzab008* (2021). https://doi.org/10.1093/protein/gzab008

Meador K., Wysoczynski C. L., Norris A. J., Aoto J., Bruchas M. R., Tucker C. L., Achieving tight control of a photoactivatable Cre recombinase gene switch: new design strategies and functional characterization in mammalian cells and rodent. *Nucleic Acids Res* (2019). https://doi.org/10.1093/nar/gkz585

Meador K. G., Killian H. R., Proctor D. M., Gressman K. M., Nuanes V. A., Dansereau O. J., Shattuck K. L., Gallagher C. K., Santos K. A., Thaden T. D., Espinoza J. M., Banks M. E., Dodier-Thurow E. R., Gillette C. M., Lata R. F., Miller J. L., Riquelme C. A., Leinwand L. A., Harvey P. A., Mitochondrial Remodeling During Physiological Cardiac Hypertrophy in the Burmese Python. *International Journal of Applied Science and Technology*. 5, 18–29 (2015)

Patents

Yeates, T. O., Castells-Graells, R., and Meador, K. DARPin Backbones and Rigidified Electron Microscopy Imaging Scaffolds. International Application No. PCT/US2023/074156, filed September 14, 2023

Yeates, T. O., Meador, K., Castells-Graells, R., and Arbing, M. A. Self assembling tetrahedral protein cages. U.S. Application No. 63/514,276, filed July 18, 2023

Research Experience

Yeates Lab, Department of Chemistry and Biochemistry, University of California, Los Angeles
January 2018 – September 2023

Tucker Lab, Department of Pharmacology, University of Colorado, Denver
April 2016 – June 2017

Falke Lab, Department of Biochemistry, University of Colorado, Boulder
August 2015 – February 2016

CU Change Lab, The Center for Health and Addiction: Neuroscience, Genes, and Environment, University of Colorado, Boulder
May 2013 - July 2015

Teaching and Mentoring

Mentored 10 students or technicians in lab techniques, experimental planning, interpretation, computational literacy, and data analysis.

Teaching assistant for 57 students in Biochemistry: Metabolism and Regulation seminar, 16 students in Biochemical Methods II lab course, 21 undergrad and graduate students in Structural Molecular Biology seminar, and 39 students in Physical Biochemistry.

Other Professional Experience

Organization for Cultural Diversity in Science (OCDS), UCLA, 2019 - 2021
Civic Engagement Co-Chair, 2020 - 2021

Graduate Biochemistry Student Association (gBSA), UCLA, 2017 - 2021
Treasurer, 2018 - 2019
Philanthropic Chair, 2019 - 2021

Chapter 1: Introduction

Protein design

Protein design can be viewed most logically as an inverse of protein folding. With protein folding, the question rests with how can a three dimensional structure be selected that most realistically places a sequence of amino acids in a configuration which minimizes the energy of the entire polymer. Here, Levinthal's paradox helps frame the enormity of the space that could be searched. The inverse question, can a protein sequence be specified which adopts a desired amino acid backbone configuration, is a question we are only capable of answering in the light of protein structure. As more protein structures have been revealed, it has become apparent that specifying a protein that adopts a particular shape provides the rational means to control molecular processes in a capacity similar to the diversity of natural proteins. In the pursuit of this goal, it is particularly important to specify the atomic configuration with enough confidence that the resulting sequence accomplishes the functional goal. Even the smallest of atomic deviations can detract from the desired functional outcome.

As the number of protein structures available has increased so too has the ability to design new ones. Still, it remains difficult to exactly predict the outcome of a design with atomic precision [1,2], especially in cases where dynamics of higher order interactions are involved, which constitute most of the functional roles proteins serve in nature. Most designs are quantified as successful based on root mean squared deviation (RMSD) [3], however, differences between crystal and solution state structures from which design templates are made [4], plus an incomplete grasp of the energetics and sampling trajectories to modeling full proteins from scratch, cause substantial deviations between design and reality. In the best-case scenario, these errors diminish total accuracy, however in most cases, designs fail completely. Given these difficulties, it is immensely more difficult to design two or more proteins which are

dependent on each other. Indeed, the success rates of multicomponent designs are far lower than independent monomeric designs [5–8].

Symmetry and self-assembly

To achieve success at large length scales, the most fundamental solution is to use symmetry. Using symmetric principles, large structures can be described using repetitive transformations of individual units. In this way, designing ultra-structured materials boils down to the design of a single unit and layering the unit upon a symmetric framework [9]. As the best practitioner of protein design to date, evolution routinely utilizes symmetric principles to generate many of the complex biological structures that organisms rely on for proper functioning. As far as finite structures, simple multimeric proteins rely on a few symmetric relationships. More complex biological assemblies—microtubules, nuclear pores, flagellar motors, and viral capsids—use multiple protein types often with individual symmetric properties to occupy highly specific functional roles.

Just as symmetry reduces the degrees of freedom in protein design, the driving reason for biological systems to utilize symmetry is the reduced genetic information to specify robust and intricate complexes. To compensate for the lost information, processes of self-assembly accompany symmetry which allow the same level of complexity to be realized from a single genetic origin. For example, at equilibrium, complexes are nearly completely dominated by the assembled species [10] which results in no sacrifice in the overall functionality, yet allows a single gene to accomplish more with less. Key then to harnessing symmetry in protein design is strict adherence to self-assembly within the design specifications. As the intended complex can only result from complete assembly of individual monomers, assembly pathways that are inefficient or inadequately specified result in off pathway outcomes [11]. The designs must cooperatively participate and follow self-assembly kinetics if they are to achieve the thermodynamic stability of the end state [12].

When it comes to creating assemblies of infinite bound, biology provides fewer examples as the nature of these extended materials create problems within the confines of a cell. A few biologic entities that contain symmetry operations consistent with 2D plane symmetry groups typically operate as surfaces or in closed topologies, including the bacterial S-layer [13], shell proteins of bacterial microcompartments [14], or the bacterial chemotaxis system [15]. As far as natural 3D lattices, the best examples result from applications where dense or stable storage is necessary such as metabolites [16], signaling molecules [17], or pathological agents [18]. Every instance follows a pattern using selective protein processing, either modifying protein shape, charge, and availability of ionic interactions to control ordered growth [19]. All of these steps require encapsulation to segregate the crystallization process from the dynamic cytoplasmic environment, except in cases where crystallization proves advantageous to an intracellular pathogen. In addition to the self-assembly processes which govern finite symmetric assemblies, lattice type assemblies involve a phase transition associated with the processes of crystallization.

Crystallization based assembly processes

In understanding crystallization assembly dynamics, the thermodynamic and kinetic properties of both nucleation and growth processes must be considered. First, a nucleation center, which constitutes a critical mass of materials with orientational similarity, must reach thermodynamic stability. Nucleation constitutes that major kinetic barrier and stands as the limiting step governing the phase transition [20]. Various snapshots of nucleation have been captured from model proteins giving insight into the mechanism [21–23]. All models of crystallization depend upon supersaturation, however models typically delineate between classical nucleation theory, where molecules in solution overcome a phase transition barrier to assume a solid crystal nucleus, or nonclassical theories, which contain intermediate stages between the liquid and solid state such as liquid-liquid phase separation, oligomer formation, or aggregation [24]. The

core of the process is similar regardless of the precise physical state where the nucleation mechanism occurs. Upon supersaturation, dense solution conditions force molecules into confined intermolecular orientations. During random fluctuations in this dense state, the energy of the system is raised due to the limited diffusional mobility. However, increased energy can be relieved upon accumulating energetically favored intermolecular contacts which decrease the systems enthalpy through molecular packing and increase entropy by liberating first shell waters into the bulk solvent [21]. If formation of a sufficient number of similar contacts results in an orientationally preferred association, additional molecules can subsequently pattern onto this solid phase with the same underlying orientation, sacrificing rigid body degrees of freedom for improved intermolecular contacts. Formation of a bonafide crystal nucleus requires that a substantial energy minima forms whereby the surface area to volume ratio of the nascent condensate surpasses a critical size and the kinetics of association and dissociation, as well as solution concentrations, result in the inevitable descent down the free energy potential to perpetuate growth of the nucleus [25].

During crystal growth, patterning of molecules off the nucleus grows the crystal into a size range amenable to structural characterization [20]. As the large entropic penalty of nucleus formation has been satisfied, this stage occurs rapidly as subsequently the molecules need only to adsorb to the solid in an orientationally specific manner. The decreased entropy in confining molecules to the growing crystal is more than offset by the gain in enthalpy from atomic interactions, and the increased entropy of the remaining molecules in the diluted solution, especially water [25]. The rate of growth largely depends on the rate at which the first molecule in a new lattice plane adsorbs to the crystal face. Subsequent molecules adsorb with fewer degrees of freedom due to supporting contacts provided by molecules on the current and internal layers. At this point, the process proceeds down the free energy landscape until an equilibrium with solution molecules is reached.

Orientational dependence in protein design

To realize orientationally dependent protein components, the choice of designed connection has a large impact on the design outcome. Current research has demonstrated a number of methods to connect macromolecules in distinct ways, which can be separated into covalent and non-covalent interactions [9]. Generally, covalent connections between proteins form a single continuous protein. These unary strategies are exemplified by fusion of an α-helical linker [26,27] or a simple flexible linkers if the linkers contain multivalent contacts to tether multiple sites in an orientationally dependent way [1]. Unary fusion techniques are generally easier to implement as they only require the design of the connecting termini. However, the reduced degrees of freedom in selecting interacting protein termini reduce the number of suitable geometric dispositions. Further, any deviation from perfect geometry can lead to unforeseen deviations along the degrees of freedom which ultimately affect the design outcome [27,28].

On the other hand, non-covalent interactions between two components do not readily assemble if one element is absent from the system. These types of binary design schemes are routinely used by biological systems to couple protein domains and typically form more stable designs as the number of contacting atoms in the interaction between components is increased. However, designing non-covalent interfaces involves greater uncertainty. As the hydrophobic effect is the primary means to drive interface energetics, protein surfaces must be endowed with significant hydrophobic substitutions. These alterations to the surface make this method prone to insolubility [5,6]. Either the protein fold cannot support the increase in exposed hydrophobics resulting in misfolding, or non-specific interactions and aggregation cause designed components to sequester from the soluble globule state [29].

The choice of connection also has important implications depending on the chosen symmetry and connection topology. For materials that are governed by self-assembly, unary designs offer a more fool proof method to satisfy kinetic constraints. These designs undergo self-assembly with the rates of the oligomeric units from which they are based and thus are

subject to rapid and complete assembly given accurate orientational specification. With binary assembly, the strength of the non-covalent interaction can be a deterrent to complete assembly [10]. In the case of extended materials, the benefit of binary design is preferred even with the downfall of decreased designability. As previously discussed, self-assembly of unbound materials naturally creates adverse effects during production in the bounded environment of the cell. Methods of cell free protein synthesis (CFPS) may permit unary strategies to be explored, however, they are not preferred, given the scale of protein production in CFPS [30] and the relative ease offered by cellular expression systems such as *E. coli*. As a binary paradigm allows separation between protein production and material assembly steps, and the biochemical determinants of crystallization are highly variable, working with a non-covalent design allows for finer control over the experimental variables.

Many considerations must be made into the nature of the non-covalent interaction. Ligand or metal mediated interactions are easily deployable, however, require introducing specific chemical linkers into the design or substitution of metal coordinating residues such as histidine or cysteine into interfaces. These methods require unique solutions for each symmetry combination [31] and are prone to nonspecific interactions or unspecified material expansion [32]. More robust design frameworks generate redesigned protein interfaces upon alignment of a pair of components with the underlying symmetry operation. This methodology has driven the success of most recently designed cages, filaments, 2D layers, and 3D crystals [2,5,6,33–39] specifically by using the Rosetta suite of computational tools [40,41]. Interface design is flexible for any shape or symmetry, however, generating strong non-covalent interfaces has typically leveraged large hydrophobic surfaces, which are prone to expression difficulties from insoluble protein. These interfaces also present issues specifically during assembly and expansion of infinite materials, as unsatisfied hydrophobic patches are present at each edge of the crystal. An alternative route for creating interfaces involves utilizing statistical sampling from the wealth of structures of protein complexes [42] and guiding design through emulation of natural interfaces

[43,44]. As nature consistently produces viable proteins, the odds of making biologically feasible designs increases the closer we get to emulating it. Indeed, in the successful crystal designs to date, the interfaces developed were free from extensive hydrophobic contacts. In the first implementation, a p6 layer was created by packing of α-helices. Polarity was then engineered into opposite ends of helical termini. Assembly of p6 layers on top of one another expanded the material in the third dimension, generating a P6 crystal [45]. In the second implementation, symmetrization of interfaces enables greater packing with less use of hydrophobics [38].

Principles of specific and favorable interfaces

To more accurately model protein interfaces, studying the patterns which make natural interfaces possible is a good place to start. Protein interfaces can be divided into three groups from the observed interchain contacts in the PDB: those which belong to obligate oligomers which are nearly completely assembled globules, those of specific protein-protein complexes, which can range from nearly permanently to transiently associated, and those arising solely from crystal contacts [43]. Key in all of these interfaces is the extent and type of buried surface area (BSA) between the two subunits. For non-obligate and obligate protein interfaces, the mean BSA is around 2,000 Å^2 with a typical minimum of around 1,200 Å^2, while only 900 Å^2 seems sufficient to encode sufficient specificity. Within interfaces, there are distinctive sets of atoms that together make a stable o-ring contact pattern. Though definitions differ, the o-ring consists of core residues, whose atoms combine to have > 25 % relative solvent accessible surface area (SASA) in the globule state, but are < 25 % relative SASA upon burial in a complex, rim residues, whose atoms have relative SASA > 25% in the complex state and support residues, those that are < 25 % relative SASA in the globule state, but contribute BSA to the interface [46]. Whereas core residues [also including support] make significant contributions (>2 kcal/mol) to binding energy, rim residues make minimal contributions to binding energy and exist primarily to seal bulk solvent from the core [42]. When compared to the rest of the protein

surface, there are noticeable residue propensities, evolutionary sequence conservation, and means of interaction that arise, with slight differences between the core and rim depending on whether the interface is a protein complex or an obligate oligomer. Charged residues are noticeably absent, with the exception of Arg, while there is an enrichment in aromatic residues and certain aliphatics, especially in the core. Generally, amino acid usage biases towards more entropically favored side chains. Crystal contacts depart significantly from the above observations. At their interfaces, they have decreased BSA, hydrogen bonds, and residue propensity, lacking specific “O-ring” morphology, and contain far more ordered waters. Additionally, many more contacts exist between small conformationally limited side chains and backbone atoms [43].

These properties are in stark contrast to the recent demonstrations of ultra-structured biologics using Rosetta de novo interface design which are significantly hydrophobic. Typically BSA hydrophobic propensity is not manifest in large patches on the protein surface. Instead, polar residues interspace most hydrophobic patches to thwart non-specific hydrophobic interactions and aggregation [47,48]. Comparison of protein complexes versus the corresponding atoms in the same monomeric structure, shows remarkably that 13% of interface atoms display SASA upon removal of the complexed protein which are completely buried in the monomeric state. This conformational diversity is unaccounted for in current design schemes and involves a number of backbone atoms, of which ~20% of the interface is typically composed. Additionally, hydrogen bonding is present in significant amounts (1 H-bond/75 $Å^2$ polar BSA, 1/190-210 $Å^2$ BSA) and typically involves backbone heteroatoms. Although there is noticeable bulk water exclusion from the core, structurally significant waters occur in wet interfaces providing coordination and bonding specificity [43]. These additional interface characteristics are noticeably absent from computational design models, which until only recently have taken hydrogen bonding into effect [8] while full prediction of water molecules in interfaces is not routinely utilized

or accurate. These shortcomings limit the ability to produce soluble, robust interfaces and need to be overcome to achieve higher success.

References

1. Sinclair, J. C., Davies, K. M., Vénien-Bryan, C. & Noble, M. E. Generation of protein lattices by fusing proteins with matching rotational symmetry. *Nature Nanotechnology* **6**, 558 (2011).

2. Shen, H. *et al.* De novo design of self-assembling helical protein filaments. *Science* **362**, 705–709 (2018).

3. Kufareva, I. & Abagyan, R. Homology Modeling. *Methods in molecular biology (Clifton, N.J.)* **857**, 231–57 (2012).

4. Sikic, K., Tomic, S. & Carugo, O. Systematic Comparison of Crystal and NMR Protein *Structure*s Deposited in the Protein Data Bank. *The Open Biochemistry Journal* **4**, 83–95 (2010).

5. King, N. P. et al. Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510**, nature13404 (2014).

6. Bale, J. B. et al. Accurate design of megadalton-scale two-component icosahedral protein complexes. Science **353**, 389–394 (2016).

7. Jacobs, T. et al. Design of structurally distinct proteins using strategies inspired by evolution. Science **352**, 687–690 (2016).

8. Chen, Z. et al. Programmable design of orthogonal protein heterodimers. Nature (2018) doi:10.1038/s41586-018-0802-y.

9. Yeates, T. O. Geometric Principles for Designing Highly Symmetric Self-Assembling Protein Nanomaterials. *Annual Review of Biophysics* **46**, 1–20 (2016).

10. Deeds, E. J., Bachman, J. A. & Fontana, W. Optimizing ring assembly reveals the strength of weak interactions. *Proc. Natl. Acad. Sci.* **109**, 2**348**–2353 (2012).

11. Bray, D. & Lay, S. Computer-based analysis of the binding steps in protein complex

formation. Proc. Natl. Acad. Sci. **94**, 13493–13498 (1997).

12. Wargacki, A. J. et al. Complete and cooperative in vitro assembly of computationally designed self-assembling protein nanomaterials. *Nature Communications* (2021) doi:10.1038/s41467-021-21251-y.

13. Sleytr, U. B. & Beveridge, T. J. Bacterial S-layers. *Trends in Microbiology* **7**, 253–260 (1999).

14. Kerfeld, C. A. et al. Protein Structures Forming the Shell of Primitive Bacterial Organelles. Science **309**, 936–938 (2005).

15. Erbse, A. H. & Falke, J. J. The Core Signaling Proteins of Bacterial Chemotaxis Assemble To Form an Ultrastable Complex. *Biochemistry-us* **48**, 6975–6987 (2009).

16. Banerjee, S. et al. Structure of a heterogeneous, glycosylated, lipid-bound, in vivo-grown protein crystal at atomic resolution from the viviparous cockroach Diploptera punctata. *IUCrJ* **3**, 282–293 (2016).

17. Dodson, G. & Steiner, D. The role of assembly in insulin's biosynthesis. *Curr Opin Struc Biol* **8**, 189–194 (1998).

18. Höfte, H. & Whiteley, H. R. Insecticidal crystal proteins of Bacillus thuringiensis. *Microbiol. Rev.* **53**, 242–255 (1989).

19. Doye, J. & Poon, W. Protein crystallization in vivo. *Current Opinion in Colloid & Interface Science* **11**, 40–46 (2006).

20. Protein Crystallography. *springer* **1607**, (2017).

21. Michinomae, M., Mochizuki, M. & Ataka, M. Electron microscopic studies on the initial process of lysozyme crystal growth. *Journal of Crystal Growth* **197**, (1999).

22. Liu, Y., Wang, X. & Ching, C. Toward Further Understanding of Lysozyme Crystallization: Phase Diagram, Protein−Protein Interaction, Nucleation Kinetics, and Growth Kinetics. *Crystal Growth & Design* **10**, **548**–558 (2010).

23. Sauter, A. et al. Real-Time Observation of Nonclassical Protein Crystallization Kinetics. *Journal of the American Chemical Society* (2014) doi:10.1021/ja510533x.

24. Nanev, C. N. Advancements (and challenges) in the study of protein crystal nucleation and growth; thermodynamic and kinetic explanations and comparison with small-molecule crystallization. *Prog Cryst Growth Ch* 100484 (2020) doi:10.1016/j.pcrysgrow.2020.100484.

25. Vekilov, P. G., Feeling‐Taylor, A. R., Yau, S. & Petsev, D. Solvent entropy contribution to the free energy of protein crystallization. *Acta Crystallographica Section D* **58**, 1611–1616 (2002).

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

27. Lai, Y.-T. et al. Structure of a designed protein cage that self-assembles into a highly porous cube. *Nature Chemistry* **6**, nchem.2107 (2014).

28. Padilla, J. E., Colovos, C. & Yeates, T. O. Nanohedra: Using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proceedings of the National Academy of Sciences* **98**, 2217–2221 (2001).

29. Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N. & Levy, E. D. Proteins evolve on the edge of supramolecular self-assembly. Nature 548, 244 (2017).

30. Krinsky, N. et al. A Simple and Rapid Method for Preparing a Cell-Free Bacterial Lysate for Protein Synthesis. *PLOS ONE* **11**, e0165137 (2016).

31. Yeates, T. O., Liu, Y. & Laniado, J. The design of symmetric protein nanomaterials comes of age in theory and practice. *Current Opinion in Structural Biology* (2016) doi:10.1016/j.sbi.2016.07.003.

32. Malay, A. D. et al. An ultra-stable gold-coordinated protein cage displaying reversible assembly. Nature (2019) doi:10.1038/s41586-019-1185-4.

33. King, N. P. et al. Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. Science 336, 1171–1174 (2012).

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

35. Bale, J. B. et al. Structure of a designed tetrahedral protein assembly variant engineered to

have improved soluble expression. *Protein Sci* **24**, 1695–1701 (2015).

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

37. Ben-Sasson, A. J. et al. Design of biologically active binary protein 2D materials. Nature **589**, 468–473 (2021).

38. Li, Z. et al. Accurate Computational Design of 3D Protein Crystals. *Biorxiv* 2022.11.18.517014 (2022) doi:10.1101/2022.11.18.517014.

39. Ueda, G. et al. Tailored design of protein nanoparticle scaffolds for multivalent presentation of viral glycoprotein antigens. *eLife* (2020) doi:10.7554/elife.57659.

40. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D. & André, I. Modeling symmetric macromolecular structures in Rosetta3. *Plos One* **6**, e20450 (2011).

41. Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. in *Methods in Enzymology* vol. 487 545–74 (2011).

42. Bogan, A. A. & Thorn, K. S. Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology* **280**, 1–9 (1998).

43. Janin, J., Bahadur, R. P. & Chakrabarti, P. Protein–protein interaction and quaternary structure. *Quarterly Reviews of Biophysics* **41**, 133–180 (2008).

44. Gao, M. & Skolnick, J. iAlign: a method for the structural comparison of protein–protein interfaces. *Bioinformatics* **26**, 2259–2265 (2010).

45. Lanci, C. J. et al. Computational design of a protein crystal. Proceedings of the National Academy of Sciences **109**, 7304–7309 (2012).

46. Levy, E. D. A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution. *J Mol Biol* **403**, 660–670 (2010).

47. Larsen, T. A., Olson, A. J. & Goodsell, D. S. Morphology of protein–protein interfaces. Structure 6, 421–427 (1998).

48. Panganiban, B. et al. Random heteropolymers preserve protein function in foreign

environments. Science **359**, 1239–1243 (2018).

Chapter 2: Design of protein crystals in two space groups demonstrates kinetic implications towards the realization of ordered protein lattices

Kyle Meador, Joshua Laniado, Todd O. Yeates

Introduction

Once the domain of material scientists and chemists, designed proteins can now be utilized as the building blocks to synthesize biocompatible nanomaterials. Ever since their utilization to deduce the atomic structures of biomolecules [1], crystalline materials have had a tremendous impact on the life sciences. To date, over 200,000 structures have been deposited in the Protein Data Bank (PDB), of which ~90% have been solved using crystallography followed by x-ray diffraction. Despite providing the foundation for structural biology, current methods in crystallography have only predicted crystal formation a priori on a handful of occasions [2,3]. Crystallographers integrate molecular biology with principles of physical chemistry to force homogeneous solutions of concentrated macromolecules into ordered crystal assemblies [4]. The fact that crystallization has proven so valuable, despite the lack of understanding to which factors influence every novel crystallization trial, represents a large gap in our knowledge.

To address the gaps in our knowledge of crystallography and advance materials science applications, we set out to design and assemble 3D crystals from combinations of lower symmetry protein building blocks. The design process samples the spatial orientations of two different symmetric protein complexes in order to position their individual symmetry axis preferentially along a global symmetric reference frame which combine to create the full specification for patterning an infinitely repeating 3D lattice. We pursued two different space groups for crystal design, utilizing three different methodologies to stabilize the orientation between oligomeric components with sufficient rigidity to facilitate ordered crystal growth. During

characterization, we show through various methodologies that by kinetically controlling the assembly of the pairs of molecules, the choice of design methodology directly affects how likely the designs can realize the intended lattice. These efforts uncover design principles that relate the outcome of the phase transitions observed to the protein design technique used and provide unique insight into the nature of this complex problem. These discoveries, along with concurrent developments in the field [3], should enable even more success on this frontier problem in the years to come.

Results

Space group considerations

Despite a general understanding of the thermodynamic processes at play, all crystal structures display varied nucleation and growth owing to the diversity of shapes and packing arrangements that chiral molecules sample. This aspect is what makes crystallization so difficult to predict. Given an unknown shape and orientation, there is no way to predict how effective packing will occur. However, as crystal nucleation is dependent on entropic factors, space groups which have more degrees of freedom (higher dimensionality) are favored during macromolecular crystallization [5]. Higher degrees of freedom provides more opportunities for a molecule to transition from a nucleation event, past the critical size, and towards crystal growth. As a sharp contrast to the random contacts formed during typical crystallization by unknown molecules, during design, the shapes and symmetries of the molecules are selected to favorably assume a lattice consistent with the symmetry operations of a predetermined space group. This knowledge allows the question of designed crystalline materials to largely remove the nucleation step from assembly and simplify to a model that only considers crystal growth via self-assembly.

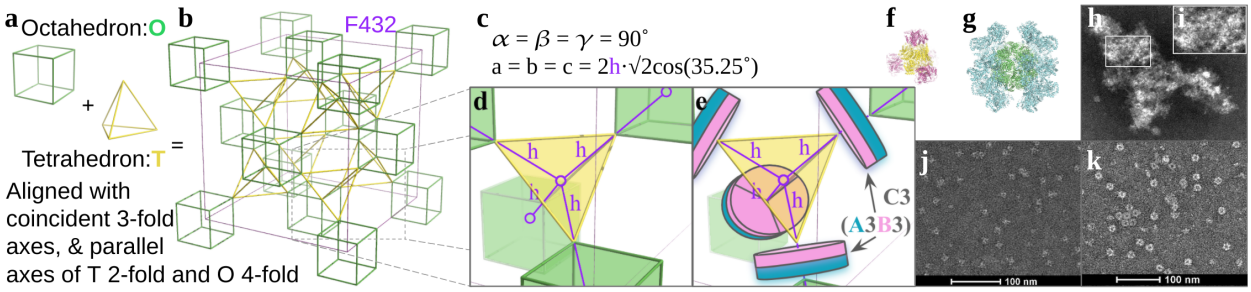
With the numerous possibilities for crystalline materials that can be formed from combinations of symmetry [6], careful consideration of the different space groups is necessary to

uncover the most favorable possibilities. Out of the 65 chiral space groups, there are 20 which can be recapitulated in SCMs. Of these, selection of the component symmetry operations is crucial in determining design potential. From the observation that nucleation favors higher dimensionality, D, where D = S + L - C, and S is the unit cell degrees of freedom, L is lattice degrees of freedom, and C is the minimal crystal contact order [5], we can apply the inverse logic to reduce dimensionality and favor the designed crystal over other alternatives. As all SCMs require only a single new contact, the contribution of C can be ignored during dimensionality considerations. Of the other two parameters, S and L, L is the only influential parameter as all 20 SCM space groups have S=6. Accordingly, minimal lattice degrees of freedom correspond to the cubic space groups where L=1. Finally, the ring size (R) establishes the minimum path to traverse from one molecule in the lattice, through a network of lattice contacts, back to the original molecule. Larger R have more capability to collapse upon unforeseen flexibility [7,8], leading to unintended design confirmations. Ring expansion is also possible, however entropically less favorable. Therefore, a minimum R can be prioritized to limit alternative conformations from forming.

Design and assembly of fused protein domains in topology F432:{T}{O}

Given geometric design considerations, the F432:{T}{O} topology offers one of the most promising routes for crystal design. The F432 space group has favorable dimensionality, minimal R value, and can be composed of highly symmetric point groups, T and O 2, 14 (Figure 2.1a). The designed interface lies between the vertices of the T and O point group when they are aligned upon a coincident three-fold axis of the space groups body diagonal and are oriented so the two-fold symmetry axis of T is parallel to the four-fold symmetry axis of O (Figure 2.1b). Using higher symmetry requires the least number of modifications given the large number of existing contacts. Additionally, the site where the point group symmetries interact occurs at a three-fold axis, allowing all design interactions to act in a multivalent nature (fig1d). For all

designs, we opted for a binary design strategy where protein sequences are engineered at the three-fold junction, utilizing an A and B group attached to both the O and T oligomeric units, respectively, allowing production of the O-A and T-B oligomers separately (Figure 2.1e). Our first strategy places an existing, structurally verified interface, on the three-fold axis between T and O oligomers. Cyclical three-fold (C3) symmetric assemblies made up of six chains total, three of type A, and three of type B (i.e A_3B_3 hetero-hexamers) are modeled into the interface using fusion of one monomer of O to one monomer of A, as well as fusion of a monomer of T to a monomer of B via flexible linkers. Representative designs have fusions of the affinity domain clustered around their three-fold vertices (Figure 2.1f,g). An additional strategy was to create affinity by splitting a portion of either the O and T assembly and fusing the smaller split peptide sequence onto the complementary oligomer. For example, a small isolated domain of O is removed and fused to the terminus of T. Upon mixing, the lacking residues of O will be satisfied



by the residues appended onto the T oligomer [9,10].

Figure 2.1. F432 Design scheme and preliminary mixing.

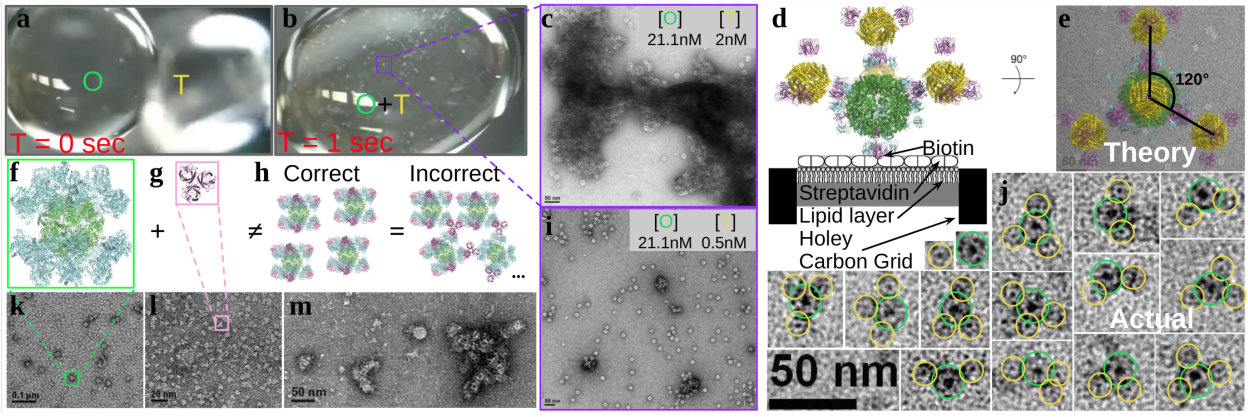
a) Combining O (green cube) and T (yellow pyramid) symmetries in the outlined orientation gives **b)** the F432 unit cell with **c)** cubic dimensions. **d)** Magnified contact geometries of O, T where the distance (h) from the center of mass (open circles) of T, to center of mass of O is variable for each F432 design. **e)** Fusion of natural C3 symmetric interfaces (A3B3, blue/pink) along h connects O and T, to A and B chains, respectively, using flexible linkers. **f)** Example design components displaying engineered interface domain on the outside of T and **g)** O oligomers. **j)** EM of T component, **k)** O component, and **h)** after crystallization attempt by overnight mixing. **i)** Magnification reveals a lack of apparent order indicating agglomerative assembly.

Twenty designs (comprising 20 individual T and O particles, where a pair of T and O represents one design) of the A_3B_3 interface fusion type along with 8 designs of split interfaces were computationally designed. We created nucleotide sequences which code the fusions and split proteins, and cloned these for expression in *E. coli*. Expression of separate O and T oligomers indicated limited solubility, especially as both components from a design pair must be soluble to test the design hypothesis. This was a surprising result given the limited modification involved to each protein in the multi-domain fusion. The result may indicate the assembly pathways of the component oligomers are disrupted upon placing an additional multimeric interaction into the assembly process. The result for split proteins was not as surprising given the interior regions exposed can cause significant hydrophobicity given the multivalency of the oligomer [11].

Crystallization dynamics of soluble designs

Separately soluble components for one A_3B_3 design and one pair of split protein designs were identified. The individual components from the A_3B_3 fusion can be visualized as homogeneous distributions of particles under negative stain electron microscopy (EM) (fig j,k) as the result of purification using size exclusion chromatography (SEC). Soluble designs were mixed to attempt crystallization of the F432 lattice. After mixing, no microscale order was apparent, however EM studies revealed numerous amorphous aggregates which have no significant crystallinity (Figure 2.1h,i). The interface in the characterized design displays less than 10 nM affinity [12]. Additionally, symmetry induced avidity lowers the effective K_d even further [13], likely into the picomolar range in our case. The effect of such strong interactions is in stark contrast to the interfaces that dominate crystals [14], and although designs may have satisfied geometric search and certainly displayed affinity criteria compatible with crystal nucleation, kinetic traps along the assembly route led to agglomeration.

To understand how such agglomeration was occurring, we tested the hypothesis that catastrophic crystal nucleation was occuring as a result of the strong interfaces in our designs. Given the designs were binary, we could employ biochemical techniques to control the assembly process by modulating both physical and chemical barriers. However, even mixing of dilute concentrations of individual components, we could visually observe a phase transition moments after mixing (Figure 2.2a,b). We first attempted to probe nucleation kinetics by controlling the concentration of each component during mixing, in effect perturbing the supersaturation of nucleation. We mixed particles in a number of concentration ratios, however we observed that a 40:1 ratio (21 nM: 0.5 nM) was required to limit solid phase formation to an extent that individual nuclei could be observed (Figure 2.2c,d). We additionally attempted a mixture regime where the O-A assembly (Figure 2.2f) was mixed with only the interfacial region (B_3) of the T oligomer (Figure 2.2g). This mixing scheme is unable to form the designed crystal as it lacks oligomeric T lattice contacts, however, the interface contacts (O-A:B) can still interact. If this occurs in an ideal manner, all O-A monomers should be occupied by a single B, ideally with an entire B_3 complex occupying the vertices of the O-A octahedron. If this is not the case a number of alternative topologies could result (Figure 2.2h). We observed individual distributions of homogenous precursor molecules (Figure 2.2k,l), however, upon mixing an equal



stoichiometry of monomers, we saw larger agglomerates, similar to the prior mixing of the full O-A and T-B particles.

Figure 2.2. Crystal nucleation and growth characterization.

a) Mixing dilute solutions of pure Tetrahedral and Octahedral components results in **b)** immediate precipitation. **c)** Precipitation is the macroscopic manifestation of agglomerates visible under EM and **i)** is concentration dependent. **f,k)** Mixing Octahedral component with **g,l)** Tetrahedral interface domain alone reveals similar agglomeration **m)** to full component mixing. **h)** Correct theoretical addition with no interface unbound compared to the incorrect reality. **d)** Solid phase crystal synthesis on EM grids facilitates imaging nucleation intermediates. Biotinylated sample is bound to streptavidin coated EM grids orienting the interface domain orthogonal to the grid surface, priming nucleation. Alternating application of symmetric components grows crystal nuclei, component by component. **e)** The result of one cycle of growth reveals numerous differences between theoretical geometry and **j)** observed geometries.

Solid phase crystal synthesis to image nucleation intermediates

The formation of higher order solid species upon addition of B_3 alone, indicated that the geometric constraints of the engineered contacts were not adequate to specify the intended lattice. To understand how the engineered contacts were improperly situated, we performed experiments to characterize the binding geometry. We devised a controlled nucleation approach similar in concept to solid phase synthesis where growth is controlled via compartmentalized reactions [15]. Briefly, one of the components is immobilized on a solid support via covalent interactions; this molecule serves as the nucleation site. Excess unbound protein is washed away leaving a clean support equilibrated in the binding buffer. Next, incubation of the support with a second solution, containing only the crystal counterpart, results in complete stoichiometric binding between bound nuclei and the second component. Subsequently, the support is again washed of excess unbound protein. This process is repeated, alternating the identity of the protein in each additional step and washing to remove excess unbound protein. Given the interaction dissociation rate of the design is measured in days [12], infinite dilution ideally leaves all molecules bound.

We performed this experiment using electron microscopy grids as a solid support medium whereby a single B_3 interfacial nuclei was attached to the grid through a biotinylated

handle that is site specifically attached to the non-binding surface of the B_3. This handle enables attachment to a 2D layer of streptavidin which was previously incubated with biotinylated lipids that coat the microscopy grid [16]. Given the three-fold nature of the B_3 and a single attachment to each monomer of B, this specifies a vertically oriented attachment with multiple points of contact which bestows the connection with more rigidity. The theoretical specification of one full round of solid phase crystal synthesis–application of each of the O and T oligomers once–leaves a single O-A particle bound to each support nuclei and subsequently bound T-B particles only present at sterically available sites. These positions include the A interaction domain at the vertice perpendicular, but opposite the support, and the vertices lying 70.5° off this axis, whereby three identical sites, 120° apart, create a characteristic geometric signature when viewed with the support perpendicular to the imaging plane (Figure 2.2d,e). We observed a range of geometric outcomes for the initial nuclei (Figure 2.2j). Given the nature of this experiment has multiple assembly processes, all prone to slight deviations from the ideal, the most important conclusion was the non-uniformity of the visualized nuclei sites. Although some have characteristic three-fold addition, most deviate in angle or in stoichiometry. If any of these off target assembly geometries occur during crystal growth, which we hypothesize is the case given the data presented, lattice defects ensure. Further, if even one incorrect interface occurs in the 96 copy number unit cell, order at the lattice length scale is affected and crystal formation poisoned.

F432 design conclusions

Given these results, it became apparent that our design was almost entirely decoupled from typical crystallization energetics which are dominated by the kinetics of nucleation and lattice growth [17]. We observe that under a regime of strong intermolecular contacts, nucleation was not dependent upon supersaturation. As the significant favorability of the interfaces overcomes the phase transition energetic barrier quite readily, we were confronted with quite the opposite problem. Rampant energetics caused lattice defects to become permanently

enmeshed in the growing solid phase, wherein dissociation rates of non-conforming molecules were completely overwhelmed by association rates and the entire design adopts significant disorder. Additional techniques to control this process include mutagenesis of interface “hot spot” residues to reduce affinity, which could provide valuable reductions in the overall rate of formation, however, general geometric flexibility in the O-A fusion may be too significant for any interfacial manipulation experiments to overcome.

Applying lessons to design of P432:{C3}{D4} crystals

We applied the principles uncovered from F432 to inform on additional design methodologies that are needed to generate successful crystal designs. As our next target, we considered the cubic space groups F23, I432, and P432 which use various combinations of two symmetry components and demonstrate promising dimensionality characteristics [5]. These space groups have cubic lattices, however, sacrifice an increasing ring size as the result of greater degrees of freedom (Table 2.1 and 2.2). Such a tradeoff is inherent when reducing the symmetry order of the two components as less symmetry present in each individual component requires more symmetric specification be supplied in their relative orientation. Although there is no change in the number of engineered contacts for any SCM (only one), there is increased design difficulty when higher order symmetry is an emergent property of the design. As we demonstrated, errors in angles of association result in defects during growth. As more of these angles are not inherent to the components, they must be specified through design.

Point Group	Number of oligomers
O	49
T	99
C3	1164
C4	217
D2	1906
D3	618

D4	234

Table 2.1. Number of oligomeric building block structures for common crystal symmetries

Oligomeric counts accessed May 2019 according to 90% sequence clustering, expression in E. Coli, and 3.0 Å resolution for x-ray crystallography PDB structures.

Space Group	Symmetry Combination	Combinatorial Design Space	Degrees of freedom (DOF)	Total Design Space
F23	C3 + D2	2218584	2	9E+06
F23	C3 + D3	719352	2	3E+06
P432	C3 + D4	272376	2	1E+06
I432	C4 + D3	134106	2	5E+05
I432	D3 + D4	144612	2	6E+05
F23	T + T	9801	1	1E+04
F432	T + O	4851	1	5E+03

Table 2.2. Degrees of freedom for two-component crystalline space groups

Symmetry operations, space groups, and total design space of favorable crystal designs using structural templates from the PDB. O - octahedron, T - tetrahedron, C - cyclic, D - dihedral.

Additional benefits arise when exploring crystals composed of lower order symmetries, primarily related to increased sampling of design candidates. First, the number of protein components in the PDB increases as the symmetry is decreased which means more pairs of designs can be sampled offering better solutions. Second, lower order symmetries offer greater designability from increased diversity of surface features at potential interface sites. Whereas high symmetry O and T oligomers have relatively smooth surface features, such as in F432 designs, using C3, C4, D2, D3, and D4 oligomers, there are greater surface features and thus generally improved shape complementary between molecules can be achieved. Third, these features allow interface design strategies to be employed which have demonstrated success in design of de novo interfaces [18,19]. Non-covalent interfaces are the primary way proteins

associate in biological systems and offer a flexible means to tune the ability of proteins to interact.

To access such surface features and interface design potential we required additional computational capabilities to specify orientations between oligomers, identify features, and specify sequences which create the desired interfaces. The result of these endeavors is outlined in chapter 3 of this work. Briefly, using a pool of oligomeric components, we use a three-step computational modeling scheme to find suitable design candidates. First, we sample pairs of protein components in configurations deemed feasible given the symmetric topology to create docked poses with no atomic clashes. Second, we examine the complementarity of their interfaces using natural protein complexes as interface templates and select for the most ideal properties. Third, we redesign the side chains present at residues in the new interface, bestowing them with atoms which are capable of producing viable interfaces. These methods rely on heuristics previously articulated with modifications that demonstrate promise in increasing success based on sampling statistically favorable interfaces from structurally characterized complexes rather than purely computational scoring.

We applied these docking and design principles to the production of designs in the P432:{C3}{D4} architecture. Designs with varying interface sizes and modeled interaction potential were chosen to span interface strength ranging from protein complexes, to crystal contacts [20]. In total, 48 designs were synthesized, each containing one C3 and one D4 component. We proceeded to characterize these designs by individual protein expression. Expression testing indicated that 15 designs demonstrated both components were soluble and therefore amenable to crystallization trials.

Structural validation of alternative assembly pathway

We tested crystallization of the most robust construct (4E/10E) using both commercial and custom crystal screens. After purification of both components separately, the concentrations

of each were measured and adjusted for equal stoichiometry. We prepared a mix of the samples at concentrations of 50, 5, and 0.5 μM immediately before crystallization trays were set up to prevent long incubation times outside of the crystallization experiment. After hours, many wells indicated formation of small pre-crystalline or microcrystalline species. Within three days, one well was identified to contain three crystal forms with cubic, rectangular, and needle-like morphologies. The condition contained crystals from both the 50 and 5 μM mixing conditions. The biggest cubic crystals were extracted, mounted and frozen. We collected data at the APS synchrotron, beamline E which produced diffraction to ~3 Å. Immediately, we could tell the crystal was not of the design given the deviating unit cell parameters and space group of I422. The structure was solved by molecular replacement and found to contain only one of the two designed components, indicating that it had preferentially crystallized even in a stoichiometrically equal mixture. The sequestration of one component from the other indicates that the interactions of this component with itself were preferred over the designed contacts with the complementary pair. Whether this was a consequence of the particular crystallization environment or an implementation detail of the design remains to be seen. However, it is hypothesized that the design contains a potential loop region that was unmodeled during design steps. We became aware upon examination and superposition of a homologous structure that this may be the case. If this is the case, potentially unfavorable clashes may prohibit the correct interface association, favoring the preferential self crystallization observed.

Screening for crystalline assembly via high throughput methodologies

There exists a number of hurdles to test for assembly into the crystalline state, especially in a manner consistent with the scale necessary to find successful candidates. Two methods were examined with the aim of miniaturizing this process and enabling discovery methods past routine purification of binary components and crystallization trials. First, we subjected designs to cell free protein synthesis (CFPS) [21,22] to enable characterization of two component systems in

amore unary fashion, where crystal growth occurs simultaneously with protein production, however, without the constraints of cellular inhibition and membrane encapsulation. Second, we conducted trials to investigate small angle x-ray scattering (SAXS) as a high throughput screening technique to identify designs which are capable of forming microcrystals [7,23].

Through CFPS trials, it was discovered that the quantities of protein synthesis are variable depending on the quality of the CFPS mixture, the particular DNA used, and the method of detection. Through all attempts to characterize production, we couldn't obtain large enough yields to reliably visualize protein expression. This was independent of whether the protein expressed in cells, resulting in a high probability of false negatives. Although detection was difficult with SDS-PAGE and western blotting, radiography could be pursued to improve detection using C-14, which is typically used as the preferred method for such dilute samples [24]. We didn't pursue radiography due to issues of laboratory safety and hurdles establishing such workflows. Given these difficulties imposed significant problems with detection, we pursued alternative methods.

As an exploration into the potential for SAXS to discover crystallinity, we set out to understand the quantities of materials necessary to detect crystals of various sizes from regular buffer conditions. If conditions can be located which are favorable for microcrystal data collection, more experimental protein solutions could be used to screen for the presence of designed microcrystals. We set up batch crystallization trials of microcrystalline samples with known crystalline sizes [25]. By performing crystallization with either 12, 10, 8, 6, 4, 3, 2, or 1 mg/ml lysozyme, the size distribution and number of crystals is reported to vary. We were only able to visually confirm crystals in the 12-6 mg/ml concentrations. Next we performed a dilution series of these crystalline assemblies either with crystallization buffer or fixative solution to cross link crystals so that upon dilution, they stay assembled. Finally, crystalline suspensions of 15 µl were subjected to capillary based SAXS experiments. We used incident x-rays of 11 keV and a

3.5 m sample-detector distance that enabled detection of characteristic Bragg peaks in the range of 0.0033 to 0.33 q, corresponding to the 0, 1, and 2 lattice indices.

Analysis of SAXS intensity versus radius demonstrated the expected Bragg peaks in only the most concentrated micro batch crystallization experiments (12, 10, 8 mg/ml) and only in the undiluted and two-fold diluted samples. Scattering of these samples was independent of preparation in fixative. Through additional investigation it was discovered that many experimental details contribute to the successful data collection. For instance, as the crystals are quite large, they are subject to gravitational effects and each of the wells with more crystals were visually inspected to have crystalline precipitation remaining after autosampler injection. Although such experiments confirmed that concentrated solutions of microcrystalline samples could be viably characterized by SAXS, more dilute and potentially semi-ordered microcrystalline formations remain a much less likely prospect.

These explorations demonstrate the continued difficulty of characterizing such assemblies via more rapid and high throughput means. Development of techniques in SAXS to minimize sample size, improve autosampler retention, and improve detection exposure areas may enable more robust detection of such solid assemblies. Additionally for CFPS, if more robust preparations are utilized, either through commercial kits, optimized nucleic acid inputs, or improved quantification methods, CFPS offers the unique capabilities to screen for such assemblies [22]. It is likely that either of these techniques could also be complemented with micro electron diffraction (microED). Our attempts at utilizing electron microscopy were complicated by the lack of ordered assemblies produced. It's also expected that the dehydration of lattices with large unit cell sizes may be a particular constraint and cryogenic microED is a prerequisite for such studies of designed protein lattices. Given these experiments, it is believed that current discovery and characterization methods are best suited to separate binary production of proteins and controlled mixing using typically protein crystallization methodologies.

P432 design conclusions

Though we have been unsuccessful in locating designed crystals in the space group P432, our experiments continue to reveal fundamental features of the design space and methods for uncovering successful candidates. The crystallization of a single component involved in the design revealed an unmodelled design consideration that may routinely result in the pursuit of such materials. As hydrophobic molecular interactions result in promiscuous, transient interactions, their explicit use in interface design of particularly crowded macromolecular structures, such as crystals, are a means by which such mutations could have unintended consequences in causing off-target interactions. For symmetric proteins, such as the oligomers which construct most crystalline morphologies, this appears to be even more so the case [11]. As there are multiple crystalline morphologies to choose from, as well as remaining uncharacterized P432 designs in the laboratory, there are certainly many more discoveries that await as we understand the complete boundaries of designed crystalline assemblies.

References

1. KENDREW, J. C. *et al.* Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 Å. Resolution. *Nature* **185**, 422–427 (1960).

2. Lanci, C. J. et al. Computational design of a protein crystal. *Proceedings of the National Academy of Sciences* **109**, 7304–7309 (2012).

3. Li, Z. et al. Accurate Computational Design of 3D Protein Crystals. *Biorxiv* 2022.11.18.517014 (2022) doi:10.**110**1/2022.11.18.517014.

4. Protein Crystallography. *springer* **1607**, (2017).

5. Wukovitz, S. W. & Yeates, T. O. Why protein crystals favour some space-groups over others. *Nat Struct Mol Biology* **2**, nsb1295-1062 (1995).

6. Laniado, J. & Yeates, T. O. A complete rule set for designing symmetry combination materials from protein molecules. *Proc National Acad Sci* 202015183 (2020)

doi:10.1073/pnas.2015183117.

7. Lai, Y.-T. et al. Structure of a designed protein cage that self-assembles into a highly porous cube. *Nature Chemistry* **6**, nchem.2107 (2014).

8. Laniado, J. et al. Geometric Lessons and Design Strategies for Nanoscale Protein Cages. *ACS Nano* (2021) doi:10.1021/acsnano.0c07167.

9. Jullien, N., Sampieri, F., Enjalbert, A. & Herman, J. Regulation of Cre recombinase by ligand‐induced complementation of inactive fragments. *Nucleic Acids Research* **31**, e131–e131 (2003).

10. Zeng, Y. et al. A split transcriptional repressor that links protein solubility to an orthogonal genetic circuit. *Acs Synth Biol* **7**, 2126–2138 (2018).

11. Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N. & Levy, E. D. Proteins evolve on the edge of supramolecular self-assembly. Nature **548**, 244 (2017).

12. Desmyter, A. et al. Viral infection modulation and neutralization by camelid nanobodies. Proceedings of the National Academy of Sciences 110, E1371–E1379 (2013).

13. Klein, D. E., Lee, A., Frank, D. W., Marks, M. S. & Lemmon, M. A. The Pleckstrin Homology Domains of Dynamin Isoforms Require Oligomerization for High Affinity Phosphoinositide Binding. *J Biol Chem* **273**, 27725–27733 (1998).

14. Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. A Dissection of Specific and Non-specific Protein–Protein Interfaces. *J Mol Biol* **336**, 943–955 (2004).

15. Merrifield, R. Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide. *Journal of the American Chemical Society* **85**, 2149–2154 (1963).

16. Han, B.-G. et al. Long shelf-life streptavidin support-films suitable for electron microscopy of biological macromolecules. *Journal of Structural Biology* **195**, 238–244 (2016).

17. Nanev, C. N. Advancements (and challenges) in the study of protein crystal nucleation and growth; thermodynamic and kinetic explanations and comparison with small-molecule crystallization. *Prog Cryst Growth Ch* 100484 (2020) doi:10.1016/j.pcrysgrow.2020.100484.

18. Fleishman, S. J. et al. Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin. Science **332**, 816–821 (2011).

19. King, N. P. et al. Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. Science 336, 1171–1174 (2012).

20. Janin, J., Bahadur, R. P. & Chakrabarti, P. Protein–protein interaction and quaternary structure. *Quarterly Reviews of Biophysics* **41**, 133–180 (2008).

21. Jewett, M. C. & Swartz, J. R. Mimicking the Escherichia coli cytoplasmic environment activates long-lived and efficient cell-free protein synthesis. *Biotechnol Bioeng* **86**, 19–26 (2004).

22. Abe, S. et al. Cell-free protein crystallization for nanocrystal structure determination. *Sci Rep-uk* **12**, 16031 (2022).

23. Li, T., Senesi, A. J. & Lee, B. Small Angle X-ray Scattering for Nanoparticle Research. *Chemical Reviews* **116**, (2016).

24. Silverman, A. D., Kelley-Loughnane, N., Lucks, J. B. & Jewett, M. C. Deconstructing Cell-Free Extract Preparation for in Vitro Activation of Transcriptional Genetic Circuitry. *ACS Synthetic Biology* (2019) doi:10.1021/acssynbio.8b00430.

25. Falkner, J. C. et al. Generation of Size-Controlled, Submicrometer Protein Crystals. *Chem Mater* **17**, 2679–2686 (2005).

Chapter 3: A fragment-based protein interface design algorithm for symmetric assemblies

The following is a reprint of a research article from:

Protein Engineering, Design and Selection

34, 1-10 (2021)

DOI: 10.1093/protein/gzab008

Reprinted by permission from OUP Journals.

peds
protein engineering design & selection

OXFORD

Original Article

A fragment-based protein interface design algorithm for symmetric assemblies

Joshua Laniado[1], Kyle Meador[2], and Todd O. Yeates[1,2,3,*]

Abstract

Theoretical and experimental advances in protein engineering have led to the creation of precisely defined, novel protein assemblies of great size and complexity, with diverse applications. One powerful approach involves designing a new attachment or binding interface between two simpler symmetric oligomeric protein components. The required methods of design, which present both similarities and key differences compared to problems in protein docking, remain challenging and are not yet routine. With the aim of more fully enabling this emerging area of protein material engineering, we developed a computer program, nanohedra, to introduce two key advances. First, we encoded in the program the construction rules (i.e. the search space parameters) that underlie all possible symmetric material constructions. Second, we developed algorithms for rapidly identifying favorable docking/interface arrangements based on tabulations of empirical patterns of known protein fragment-pair associations. As a result, the candidate poses that nanohedra generates for subsequent amino acid interface design appear highly native-like (at the protein backbone level), while simultaneously conforming to the exacting requirements for symmetry-based assembly. A retrospective computational analysis of successful vs failed experimental studies supports the expectation that this should improve the success rate for this challenging area of protein engineering.

Key words: algorithms, docking, protein design, protein interfaces, secondary structure, self-assembly, symmetry

Introduction

A range of emerging bionanotechnology applications rely on designing protein molecules to bind and associate with each other in a geometrically specific fashion. Among such applications, those aimed at creating novel, self-assembling symmetric architectures, such as protein cages and extended protein arrays, place especially strict demands on achieving atomically precise associations (Yeates *et al.*, 2016). When such precision can be achieved by design, diverse protein-based materials with tailored spatial and biochemical properties can be produced. As examples, cubic and icosahedral protein cages (Padilla *et al.*, 2001; King *et al.*, 2012, 2014; Bale *et al.*, 2016; Cannon *et al.*, 2020a,b), as well as extended protein arrays (Ben-Sasson *et al.*; Sinclair *et al.*, 2011; Gonen *et al.*, 2015; Suzuki *et al.*, 2016), are finding wide ranging uses as biotherapeutics (e.g. for vaccines) (Brouwer *et al.*, 2019; Marcandalli *et al.*, 2019; Ueda *et al.*, 2020), as scaffolds for enzyme organization or atomic imaging (Ernst *et al.*, 2019; Liu *et al.*, 2019; Heater *et al.*, 2020; McConnell *et al.*, 2020), and as nanoscale containers for molecular encapsulation and delivery (Liang *et al.*, 2014; Edwardson *et al.*, 2020).

Owing to their complexity, as well as our incomplete understanding of their behavior, protein molecules present challenging subjects for design. In protein engineering studies, these challenges often manifest through unpredictable outcomes from mutagenesis, frequently leading to proteins that are prone to misfolding and aggregation. Improved computational methods are addressing those challenges, making it increasingly feasible to mutate the surface of two suitably chosen proteins to create a binding interface between them (Fleishman *et al.*, 2011; Fallas *et al.*, 2017; Pearce et al., 2019). Similar goals are being reached using *de novo* polypeptides as components (Chevalier *et al.*, 2017; Chen *et al.*, 2018; Adihou *et al.*, 2020; Cao *et al.*, 2020). Yet, despite exciting progress, relatively low success rates are still common in application areas where precision and predictability are essential, generally requiring many design trials to achieve a smaller number of correctly assembling protein designs. A

general view on the current challenges in designing novel protein–protein interfaces is that computational methods do not necessarily generate (prospective) interfaces that mimic native protein– protein interfaces (Stranges and Kuhlman, 2013). The difficulty of the task is heightened in design problems where additional spatial constraints must be met, beyond those required simply for binding. For the design of symmetric cages and regular arrays, for instance, the novel interface must bring the two component proteins together under exacting rules of symmetry; e.g. if each component is part of a naturally symmetric oligomer, then the interface must cause the symmetry axes of the separate components to intersect at a precisely prescribed angle. Such complex constraints confound the problem of designing optimal, native-like interfaces.

In addressing the problem of interface design in the context of symmetric assembly, the strategy introduced by King *et al.* (2012) prioritized the symmetric constraint part of the problem. There, oligomeric building blocks were docked by systematically sampling the rigid body degrees of freedom allowed by the point symmetry of the target assembly. As a result of the high dimensionality search space and the large number of different component oligomers con-sidered for docking, a rapid first-pass scoring was used to identify configurations that were potentially suitable for design: the number of Cβ contacts between the docked oligomeric building blocks. Naturally, only a minute fraction of candidate poses chosen under such coarse criteria present interfaces that are similar in atomic detail to those from natural protein–protein complexes. Subsequent amino acid sequence design and additional filtering steps were required to identify interfaces that might exhibit native-like properties. Newer protocols have shown the value of considering known residue pair interactions during docking (Fallas *et al.*, 2017) and prioritizing interfacial hydrogen bonding during sequence design (Boyken *et al.*, 2016; Chen *et al.*, 2018; Cannon *et al.*, 2020b).

The expansive database of known protein structures provides valuable empirical frameworks for evaluating proteins in terms of secondary structure motifs (Finkelstein and Ptitsyn, 1987; Guharoy and Chakrabarti, 2007; Gao and Skolnick, 2010; Xie *et al.*, 2015; Zhou and Grigoryan, 2015). Recent exercises in protein design have begun to prioritize the consideration of secondary structure motifs and the atomic details of how they tend to associate in native proteins (Tischer *et al.*; Silva *et al.*, 2019). For instance, threading helical fragments together produces novel fold topologies that retain features of observed tertiary motifs (Jacobs *et al.*, 2016; Brunette *et al.*, 2020). Further, sequence design using statistical models of tertiary structure segments has competed with or outperformed physicochemical energy functions in routine design tasks (Zhou *et al.*, 2020). The growing focus on secondary structure associations motivates an attempt to bring those principles to bear on the class of design problems related to symmetry-based assemblies.

Here, we describe algorithms and software that expand motif-based design methodologies to symmetric docking applications—e.g. cubic cages and extended protein arrays. Our new program is parameterized to exploit recent theoretical work articulating the geometric rules for designing wide ranging nanoscale materials built from combinations of oligomeric protein components—i.e. symmetry combination materials (SCMs) (Laniado and Yeates, 2020). Strategic choices are discussed for program optimization based on fragment-based lookup tables and separation of rotational vs translational subspace searches. Prospective novel designs are discussed, along with a retrospective analysis of successfully designed protein cages.

Results

Docking under symmetry constraints

The goal of the program developed here was to enable fragment-based docking for the design of self-assembling materials based on the principles of combined symmetries. The essential idea for building highly symmetric materials from simpler protein oligomers was described by Padilla *et al.* (2001), with diverse variations demonstrated in recent years (Bale *et al.*, 2016; King *et al.*, 2014; Lai *et al.*, 2014; Cannon *et al.*, 2020a). A complete articulation of all possible SCMs was recently completed (Laniado and Yeates, 2020); 124 different kinds of architectures can be created by introducing a new interface between two oligomeric components. In addition to various cage types based on the Platonic solids, 35 kinds of 2-D arrays and 76 kinds of 3-D arrays were identified as targets possible for design. Each of the distinct SCMs presents a different set of rigid body constraints, and complementary rigid body degrees of freedom, for sampling allowable arrangements of the two oligomers to be docked. For the present work, we have integrated the design rules for all possible SCMs within a new program, nanohedra.

We developed a general docking framework, applicable to all SCMs, that performs a search over multiple rigid body degrees of freedom relating two oligomeric building blocks (Fig. 1). The number of degrees of freedom depends on the symmetric system being constructed, ranging from a minimum of 1 to a maximum of 5 (Laniado and Yeates, 2020). Exploiting advantages of precalculation methods, we were able to factor the search problem for all scenarios into a search over rotational degrees of freedom (for cases where they exist), followed by direct calculation of optimal translational values by linear algebra methods, thereby avoiding the need to explicitly search translational degrees of freedom for each rotation. Identifying favorable docking arrangements within the allowable rigid body search space is made possible by precomputing common protein–protein fragment configurations from known structural data.

Fragment-based elements

Focusing on short segments of protein structure makes it possible to reduce computational complexity with lookup or 'hash' tables. To this end, we chose to categorize local protein structure using 5-residue fragments. Heuristically, a 5-residue segment is long enough to capture secondary structures types, including α-helical and β-strand conformations, as well as loop structures, while being short enough to model the allowable space of conformations with acceptable precision and coverage using a tractable number of representatives. Using a curated set of known protein–protein interfaces (see Methods), we

computed the most highly represented 5-residue fragment types found at interfaces using nearest neighbor clustering (on Cα RMSD) for a randomly sampled subset of fragments. We experimented with different similarity cutoff criteria and settled on a 0.75 Å cluster inclusion limit, which maximized fragment coverage, while ensuring stringent constraints on backbone geometry.

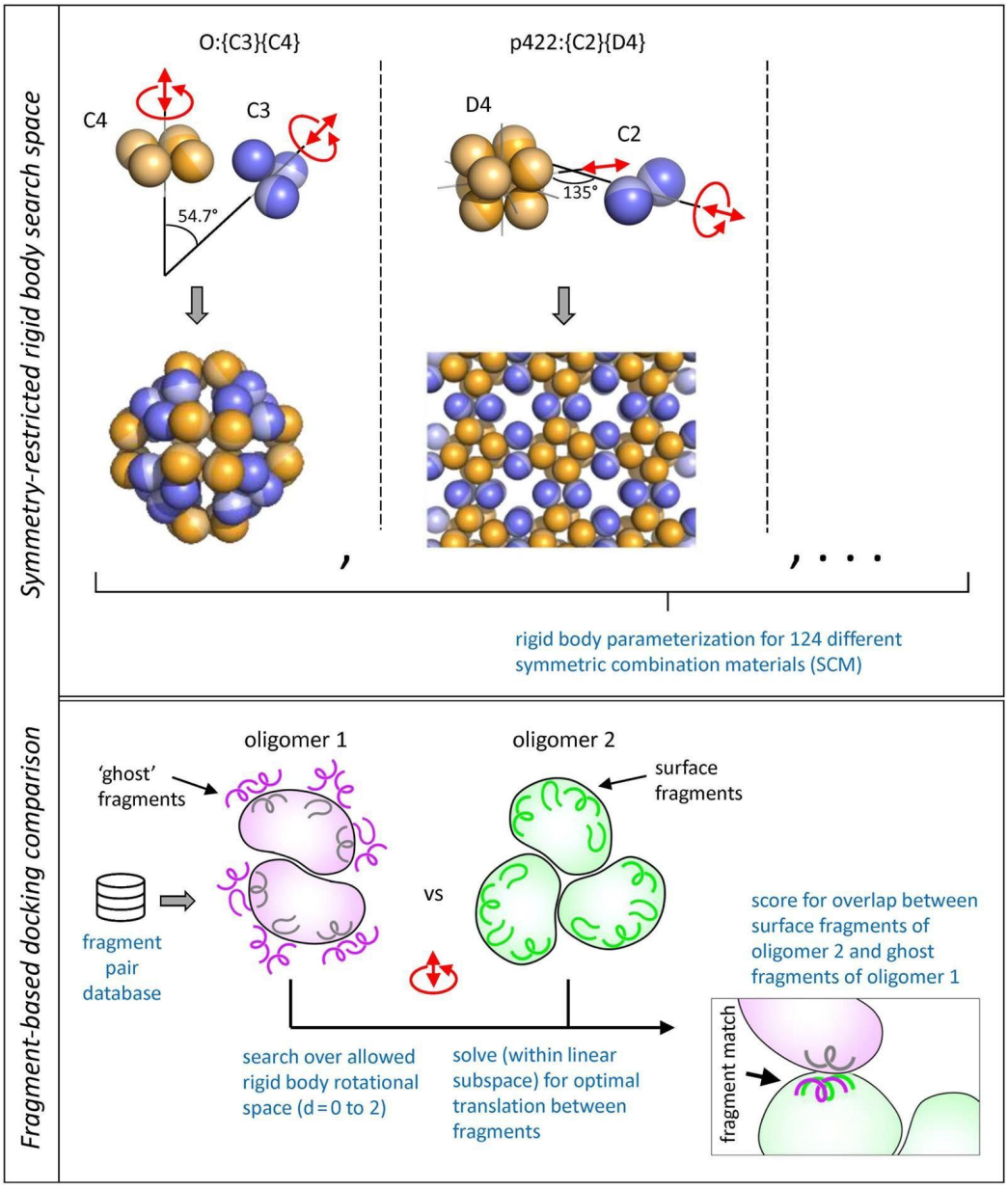


Fig. 1 Scheme illustrating two major aspects of the Nanohedra program for designing SCMs from two oligomeric protein components. The top panel shows examples of two SCM types (of 124 types possible), focusing on the geometric rules that must be satisfied when bringing the two different oligomeric components into specific contact. In each case, the red arrows indicate the rigid body degrees of freedom available, which must be explored computationally in a search for favorable docking configurations that would be amenable to amino acid sequence design at the emergent protein–protein interface. Nanohedra encodes the specific rigid body parameterization required for constructing all 124 SCM types (Laniado and Yeates 2020). The bottom panel highlights the use of protein fragment pair libraries as the essential feature for selecting favorable design poses for subsequent interface design. This allows nanohedra to generate native-like interfacial backbone arrangements for design. Program operation is made computationally tractable through various precalculation schemes. One of these involves the decoration of the first oligomer with 'ghost fragments' (based on a library of favorable fragment pair configurations), after which the search for suitable docking poses is reduced to a problem of identifying allowable oligomeric arrangements wherein surface fragments belonging to oligomer 2 overlap closely with ghost fragments covering oligomer 1.

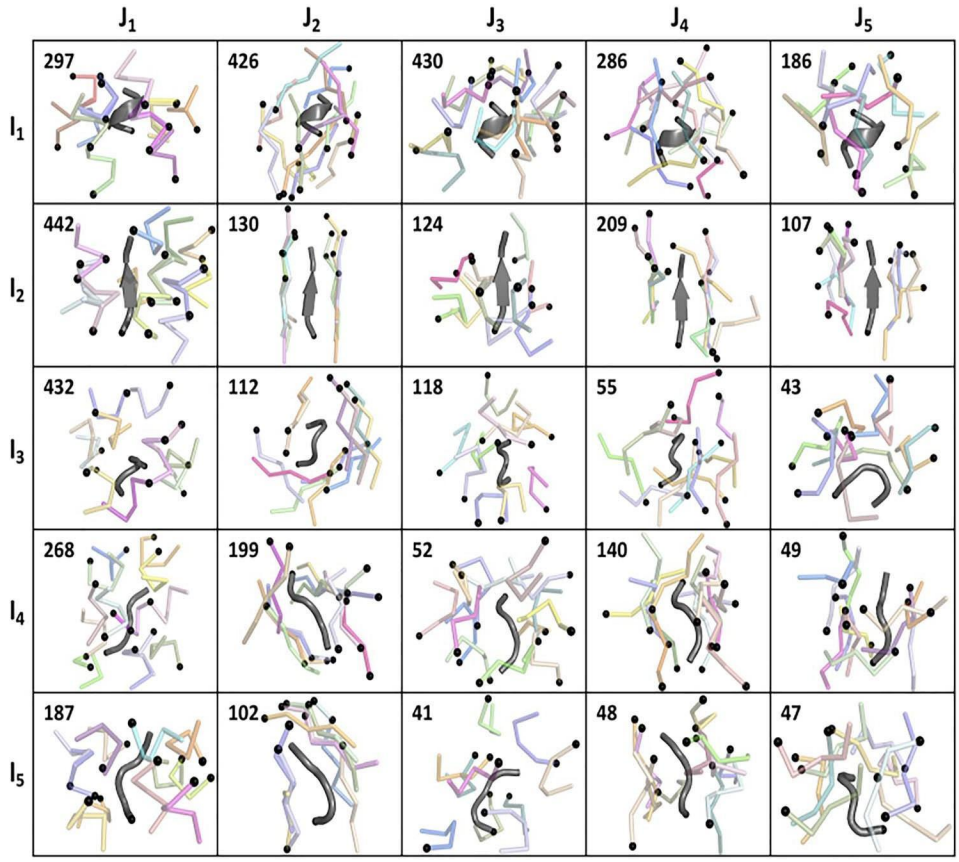


Fig. 2 Interface fragment database. For each of the 25 *i, j* fragment pair possibilities, a single representative *i* fragment is shown in gray and a subset of cluster representatives of the top 20 most populated clusters are shown in color for the spatially clustered *j* fragments. N-termini are marked with black spheres. The total number of unique *i, j* clusters is indicated in the top left corner of each frame.

As expected, different fragment clusters were populated to different degrees; those representing canonical α-helical and β-strand conformations were much more densely populated than those representing different loop conformations, or cases where regular secondary structure transitioned into loops. The top five clusters were sufficient to represent 61.4% of the candidate fragments, with the highest observed cluster corresponding to an α-helical conformation, followed by a β-strand conformation then three coiled conformations (Fig. 2). Rather than considering a larger number of clustered fragment conformations, we retained five cluster types in order to maximize statistical power in subsequent steps.

Following individual fragment clustering, a paired fragment–fragment clustering procedure was applied to interprotein contacts from the same protein–protein interface set. This problem was simplified by a form of coordinate reduction. A set of three 'guide coordinates', built on the C-alpha atom of the central residue of the 5-residue fragment (Fig. S1), was associated with the representative fragment from each individual fragment cluster (see Methods); note that three *x, y, z* coordinates (nine variables) are sufficient to specify 6D rigid body orientation and position in 3D space. This provided a generalized scheme for clustering the relative spatial arrangement between fragment–fragment pairs. Briefly, for every instance where a 5-residue fragment (type *i*) from one protein was found in spatial contact with a 5-residue fragment (type *j*) from another protein, each fragment in the *i, j* pair is assigned to one of the five individual fragment types. This allows placement of the representatives' guide coordinates onto the coordinate frame of the observed fragment pair. Next, the guide coordinate pair was transformed to put the *i* guide coordinate set in a canonical setting (i.e. at the origin with internal axes along principle directions). The resulting *j* guide coordinate set is then stored, providing a full representation of the relative spatial arrangement of the *i, j* fragment pair instance. The *j* guide coordinate sets were then used as the basis for a final nearest neighbors clustering step where the resulting cluster index, *k*, represents the different spatial modes that tend to be populated by specific *i, j* fragment pairs (Fig. 2). Clustering at this pairwise stage was based on a relatively strict similarity criterion (1 Å guide coordinate RMSD) to establish separate conformational and amino acid preferences for relatively finely discriminated fragment– fragment arrangements (Fig. S1). The resulting data structure is a triplet of (*i, j, k*) indices, each carrying a 9D coordinate point that captures a frequently observed spatial relationship *k,* between a specific *i, j* fragment pair type. In addition, owing to the cartesian nature of the embedding, a 9 x 9

covariance matrix (approximately rank 6) provides a quadratic description of the spatial variation for the given *i, j, k* fragment pair cluster; that information can be used to analyze permissible deviations. Observed central residue amino acid frequencies are also stored for each *i, j, k* cluster and can be used to guide subsequent sequence design steps. Ultimately, a total of 97 935 *i, j* fragment pairs from observed structures were grouped into 4530 *i, j, k* clusters specifying geometrically defined 3-D fragment associations.

Precoating by 'ghost fragments'

Having established common fragment pairing conformations in advance enables a precalculation protocol with important computational time savings. As a set-up to docking trials between two oligomers, one of the component oligomers (the first oligomer) is decorated with a large set of prospective 'ghost fragments' (Fig. S2 and Methods). These ghost fragments represent preferred interaction potentials based on the orientation of the fragments on the surface of the first oligomer. The precalculated database of representative *i, j, k* fragment pairs described previously serves as the source for constructing this set of ghost fragments, which is intended to be inclusive for the backbone configurations of the second oligomer that might comprise frequently observed interactions with the first. Depending on the size of the first protein, the ghost fragments may number in the thousands. Once the ghost fragment set is calculated, the subsequent fragment-based docking scheme is reduced to a problem of identifying orientations that might bring surface fragments of the second oligomer into near coincidence with the ghost fragments decorating the first. The exploitation of preferred positions for noncovalent interacting groups bears similarity to other recent computational design applications (Zhou and Grigoryan, 2015; Fallas *et al.*, 2017; Polizzi and DeGrado, 2020).

Identifying favorable rotations and translations

The outer loop of the docking calculations applies candidate rotation values to the two oligomers, if those degrees of freedom exist. The symmetric construction schemes for SCMs provide a maximum of one degree of rotational freedom for each component oligomer (e.g. about the unique symmetry axis of a cyclic oligomer). For each choice of rotational values for the two oligomers, a calculation is performed to test which of the possible pairs of fragments (chosen from the surface fragments of 2 and the ghost fragments of 1) are in nearly equivalent orientations, as would be required for near-overlap under any choice of translation. This step involves a large number of possible fragment pairs to be considered, as well as somewhat complex numerical calculations for orientation comparisons. We found it critical to shorten this calculation with further precalculation methods and hash tables. We assign each fragment (based on its guide coordinates) to a set of three Euler angles describing its orientation, with the Euler angles discretized into 10° bins. With a triplet of orientation indices assigned to each fragment, we are able to look up in a precalculated 6D Boolean (true/false) table whether or not the sets of Euler angle triplets assigned to the two fragments are within a prescribed angular discrepancy (with an accuracy of roughly 10°).

The steps described above rapidly identify pairs of fragments (a surface fragment of oligomer 2 and a ghost fragment surrounding oligomer 1) that could be nearly coincident under the chosen orientation values and *some* translational values between the oligomers. It is critical, however, that the translational relationships conform to those that are prescribed by the particular symmetry rules of the SCM being constructed. Some SCM types have three translational degrees of freedom while some have as few as one. Importantly, our program encodes those translational restrictions for all SCM types, based on tables provided in Laniado and Yeates (2020). For every pair of candidate fragments that have compatible orientations, our program calculates the optimal translation for overlap, *within the allowable space of rigid body translations for the given SCM type*. This is performed using a linear least-squares calculation, with the error value based on RMS deviation between the two sets of guide coordinates as a function of translational degrees of freedom. We then store the translational parameters for cases where the RMSD for the optimal overlap is within a prescribed cutoff (e.g. 1 Å).

Ultimately, a suitable docking arrangement between the oligomers is one where multiple candidate fragment pairs could be brought into near coincidence for the same (or highly similar) choices of the rotational and translational parameters. For fastest performance, we found it efficacious to perform the docking analysis in a rapid first pass over a reduced set of candidate fragment pairs (e.g. requiring at least one helix–helix association), followed by a second pass wherein the translational values established in the first pass serve to restrict consideration of additional fragment pairs in the second pass, with an attendant reduction in CPU time.

We found the procedures described above critical for reducing the CPU times to levels that were compatible with docking large sets of candidate oligomer pairs. Other approaches could also be considered, though we emphasize that procedures that might appear beneficial for certain kinds of symmetric construction choices are sometimes problematic for other types of constructions, e.g. depending on the types and numbers of the rigid body degrees of freedom. The system we developed applies universally to all 124 SCM types.

Heuristic scoring

For each satisfactory docking configuration, a nanohedra score is calculated based on the collection of favorable fragment pairs identified, with the goal of evaluating how well the docked interface is supported by the underlying fragment observations. To compute the nanohedra score, for each instance where a favorable surface fragment *vs* ghost fragment pair has been identified, a similarity score (z) is first calculated by dividing the RMSD obtained between the surface and ghost fragments by the mean RMSD for member fragments comprising the ghost fragment's *i, j, k* cluster (precalculated during fragment database creation), with a low value of z indicating a close similarity. If z is less than a prescribed threshold value (e.g. 2),

the inverse of 1 plus z squared is taken to give a match score, ranging from 0 to 1 with 1 indicating a perfect match. This match score for each fragment is propagated to each of the five residues comprising the fragments on oligomers 1 and 2. In this way, each residue in the protein–protein interface might inherit multiple component scores, since each residue might belong to overlapping fragments participating in favorable fragment–fragment pairs. For each such interfacial residue, its assigned match score(s) are first ranked in descending order and are then weighted by $1/2^{\text{rank}-1}$ (rank $>$ 0) for a final summation. This weighting scheme bounds the final score for each residue to a maximum of 2. The weighted match scores are then summed across interfacial residues to give the final nanohedra score for the identified docking configuration.

Program considerations

Nanohedra is a command line tool. It can be operated in one of three modes: query, dock or postprocessing. The docking mode executes the main procedures described in the present work. The user specifies the desired symmetry material outcome or SCM type, i.e. the specification of the two component symmetries and their resulting assembly type. Directory paths are input to specify the file locations for the oligomeric protein structures to be tested. The program output comprises pdb files with candidate docked poses in various forms (asymmetric unit within the final symmetry, docked oligomers and an expanded symmetry). Other information includes the final nanohedra score and the spatial transformation matrices mapping the canonically oriented coordinates onto the candidate pose. To guide subsequent design of the resulting interface, sequence information is output in the form of amino acid frequencies based on amino acid composition information tabulated from the fragment database (Fig. S3).

The computer time for execution depends critically on the size of the proteins (because larger proteins carry more surface fragments), the number of rotational degrees of freedom for sampling and the rotational sampling interval. Times on a single CPU core (2.5 GHz) can range from 2 to 24 h, with typical applications exploiting multicore clusters. Computer memory requirements also depend on the sizes of the proteins and the size of the symmetry group generated by the final assembly. Requirements range from roughly 8 to 25 GB. The user can override various default settings, e.g. angular sampling in rotational searching (−rot_step1/−rot_step_2) or the minimum number of fragment–fragment pairs needed for a well-docked pose (−min_matched).

Query mode is an informational mode that helps the user understand different options and certain symmetry aspects of the material to be designed: e.g. what kinds of resulting SCM materials can be constructed from a given combination of components, and conversely what component oligomer types would be needed to construct different SCMs according to various target criteria, such as the dimensionality of the resulting material (cage vs layer vs 3D crystal), the underlying rotational symmetry or specific geometric features (like network properties) of the material to be designed. Different material properties will be advantageous in different experimental contexts, and this mode captures the full space of design types recently articulated (Laniado and Yeates, 2020). A final postprocessing mode provides tools for ranking the output candidate poses, with options to sort by different criteria, e.g. according to the final nanohedra matching score or according to the numbers of fragment pairs identified in the match.

The program is implemented in Python with the exception of one routine that is written in Fortran (orient_oligomer). Python dependencies include biopython (Cock *et al.*, 2009), numpy (Harris *et al.*, 2020) and scikit-learn (the BallTree method is used to test for clashes) (Pedregosa *et al.*, 2011). Nanohedra also uses the freeSASA program to calculate solvent accessible surface areas (Mitternacht, 2016). The program code has been made available on GitHub.

Prospective SCMs

To demonstrate the universality of our fragment-based docking approach, we constructed prospective SCMs of six distinct types, with representatives from point (1), layer (2) and space group symmetries (3), based on component oligomer symmetry types ranging from C3 to tetrahedral (T). Nanohedra was run with default docking parameters, and for each SCM type, a search of the rigid body degrees of freedom inherent in each system produced numerous viable candidates with varied orientations and positions and different interfacial secondary structure compositions. Postprocessing revealed numerous poses with high nanohedra scores. One representative structure produced for each of the six SCM types tested is displayed in Figure 3. The results demonstrate the viability of the described method at producing assemblies conforming to a selected symmetric material.

In each example, the resulting interface exhibits native-like properties with respect to interfacial backbone–backbone associations. High structural complementarity between oligomers is apparent from the overlap between the ghost fragments of the first oligomer and the matched surface fragments of the second oligomer. The interfaces vary in the extent of regular secondary structure involvement, with each oligomer contributing at least one continuous

secondary structure element to the interface, and ranging from 8 (F23:{C3}{T}) to 20 (p222:{D2}{D2}) unique fragment matches (Fig. 3). Many of the docked configurations comprise extensive helical interactions, with interfaces containing anywhere from two to five helices (see F23:{C3}{T} and I432:{C4}{D3}, respectively). The contribution from β-strands is also apparent as both T:{C3}{C3} and p222:{D2}{D2} designs have mixed α/β interface motifs, despite prioritizing helix–helix pairs in first-pass searching. Additionally, matched interface fragments are sometimes enhanced by fortuitous contacts involving coiled segments surrounding regular secondary structures. These characteristics are reminiscent of patterns observed in nature (Guharoy and Chakrabarti, 2007).

Post facto analysis of designed protein cages

Prior work in designing protein assemblies has shown the challenges of generating computational designs that produce the desired exper-imental outcomes; success rates remain relatively low, as failures can manifest at many crucial

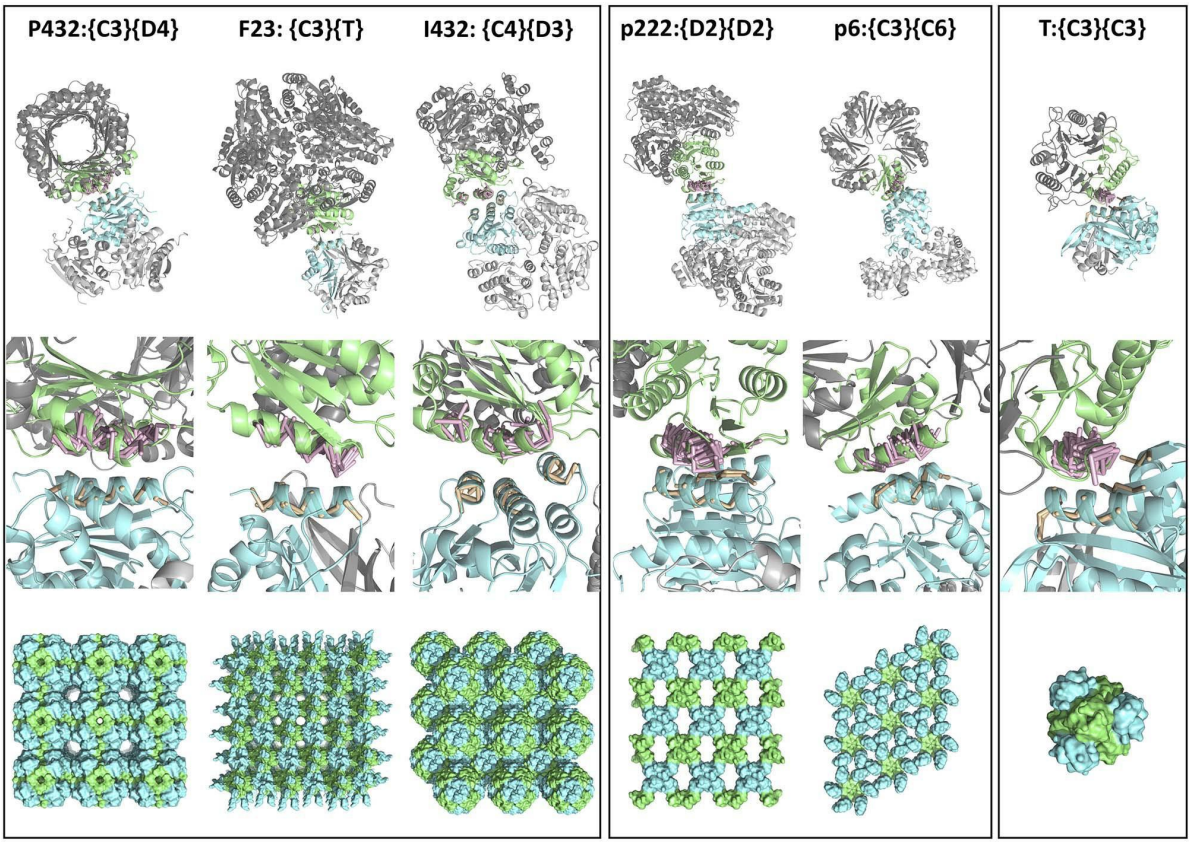


Fig. 3 Prospective SCMs. Six example SCMs generated by Nanohedra are shown: a finite tetrahedral cage (right column), two 2D layers (middle columns) and three 3D crystals (left columns). The top row illustrates the docked oligomeric building blocks that are required to construct the final material. Chains directly implicated in the docked interface are colored, while symmetrically related chains are in gray. Closeups of the interfaces are shown in the middle row. Oligomer 1 (blue) surface fragments (tan) and associated ghost fragments (pink) are shown. Ghost fragments are matched with secondary structure elements on the surface of oligomer 2 (green). The resulting symmetrically expanded materials are displayed in the bottom row. PDB accession codes used to generate the prospective materials shown are 1OSC and 1NQ3 for T: {C3}{C3}, 4O5O and 1UAY for p222: {D2}{D2}, 1OSC and 1GTZ for F23: {C3}{T}, 2B34 and 3BBC for I432: {C4}{D3}, 1VHC and 2A10 for p6: {C3}{C6}, 4XCW and 1DHN for P432: {C3}{D4}.

junctions. The favorable features of nanohedra—constructing designs based on native-like interfacial packing— will ultimately require experimental tests that are ongoing and not presented here. Nonetheless, the results of several recent design trials provide an opportunity to evaluate the prospective advantages of nanohedra ahead of new experimental trials.

For a retrospective analysis, we asked whether nanohedra could distinguish experimentally validated designed protein assemblies among a larger body of prospective computational designs that were unsuccessful. We focused on designed protein cages, for which there are more than a dozen successful cases validated in atomic detail, along with more than a hundred computational designs that led to experimental failure. We ran nanohedra on these designs to see if there was a difference in the generation of candidate poses that matched prior design targets between the two sets; this would argue that nanohedra has the capacity to generate computational designs that have improved experimental success rates. For each prior design (in both categories of experimental successes and failures), we took the two component oligomers in standard orientations (i.e. not corresponding to the previously designed configurations) and ran nanohedra to generate prospective designs for symmetric cages of the desired symmetry. While nanohedra was able to recapitulate the target in nearly all cases, there were differences in the extent to which the design target was ranked favorably compared to other potential designs comprising the two oligomers. We clustered all poses in the top 2000 output, then examined the ranked output to see where in the list of candidate poses (if at all) we could find configurations closely matching the target that was experimentally tested in earlier work. For the group of experimental successes (n = 14), we were able to recapitulate 70% of the design targets within the top 12 scoring pose clusters for each combination of building blocks and 100% of the targets within the top 88 ranked poses (Fig. 4). In contrast, for the design set derived from experimental failures (n = 138), we were unable to identify closely matching poses for 25% of designs and had to search until rank 112 in order to recapitulate 70% of the designed poses. These calculations clearly show that, among earlier computational designs, those that went on to experimental success are much more readily recapitulated using nanohedra compared to designs that failed. This indicates that using motifs present in native interfaces leads to improved search heuristics for biologically confirmed

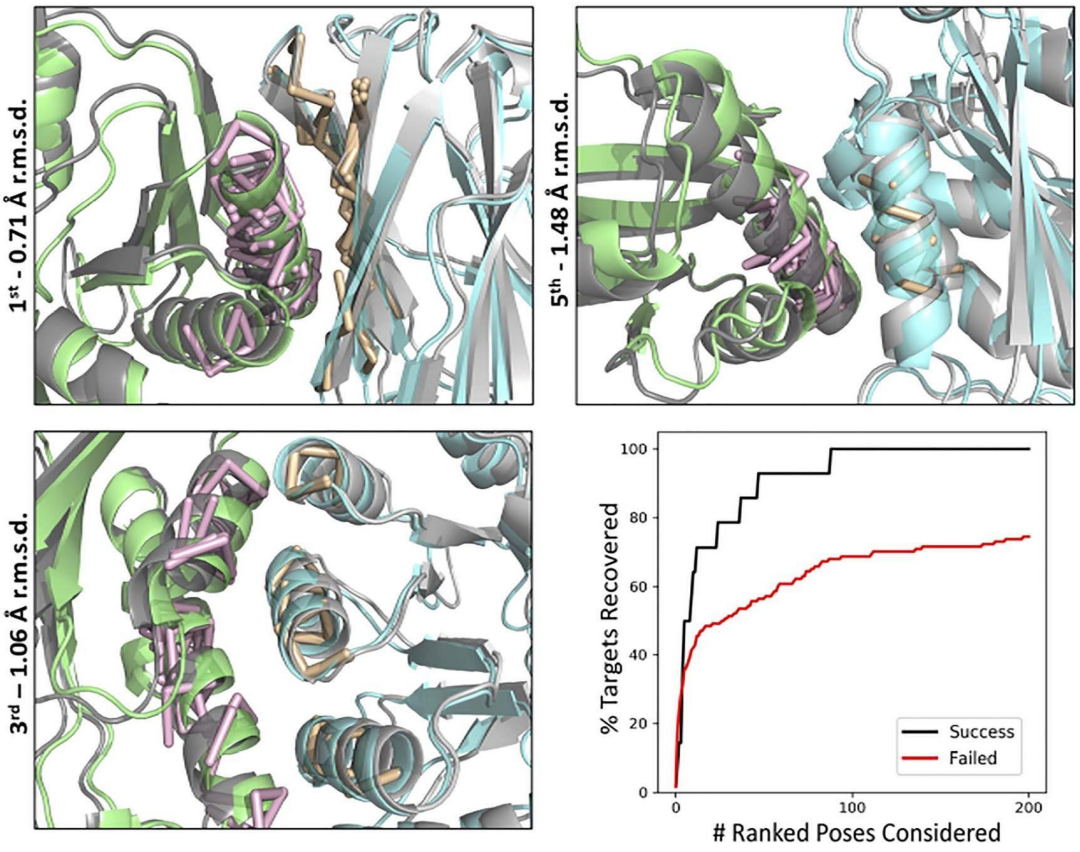


Fig. 4 Post facto analysis of designed protein cages. Three representative examples of successfully recapitulated poses (top and bottom left). The crystal structures of the target designs are shown in gray and Nanohedra predictions are shown in color. The iRMSD indicates the agreement, for atoms near the interface, between the docked pose and the crystal structure. The numerical value indicates the rank of the docked pose. The crystal structure of T32–28 (4NWN) displays a 0.71 Å iRMSD with the first ranked pose (top left). The crystal structure of T33–15 (4NWO) exhibits a 1.48 Å iRMSD with the fifth ranked pose (top right). The crystal structure of I53–40 (5IM5) shows a 1.06 Å iRMSD with the third ranked pose (bottom left). The ROC curve for all successful and failed designs shows the percentage of targets recovered according to the number of clustered nanohedra poses considered (bottom right). iRMSD of 3.0 Å or less was considered as a recovered pose. iRMSD—interface root mean squared deviation.

symmetric material designs. We further note that having to go to rank 12 to recapitulate most of the earlier experimental successes does not preclude that poses ranked higher by nanohedra could quite plausibly lead to successful experimental constructions, different in orientation from those validated earlier.

Discussion

Until now, designing symmetric protein assemblies has remained challenging to new entrants, as the process requires somewhat expert knowledge about symmetric construction and intertwined issues of how to sample allowable degrees of freedom in the context of docking software. Previously successful studies in executing two-component symmetric docking have used the Rosetta TCdock protocol, which requires the user to specify the symmetry rules through the use of symmetry definition files (King *et al.*, 2014). This is possible for symmetries already enumerated, but the majority of recently described SCMs (Laniado and Yeates, 2020) present a remaining challenge for specifying allowable degrees of freedom in the context of existing design software. By enumerating the allowable degrees of freedom for each SCM in a comprehensive, facile framework, nanohedra will empower a broader group of users to explore the large space of possible symmetric designed materials. This should accelerate the development of novel designed materials by protein and biomaterial engineers.

Nanohedra harnesses the power of a recently established theoretical framework (Laniado and Yeates, 2020) to enable the construction of a universe of possible protein-based nanomaterials. In addition, nanohedra's docking algorithm implements a novel fragment-based approach for assessing whether docked solutions resemble biological interfaces. Importantly, the nanohedra score does not depend on commonly used heuristic simplifications (such as number of Cβ–Cβ contacts) for rapid assessment of binding likelihood, as has been used in various docking studies (King *et al.*, 2014; Bale *et al.*, 2016; Fallas *et al.*, 2017). Instead, its statistical representation of clustered secondary structure elements exploits empirical

knowledge of typical packing motifs found in native protein interfaces. The implications of this choice, as demonstrated in our retrospective analysis of successful versus failed designs, are notable. In agreement with previous findings, geometric packing alone captures many essential elements of protein interaction (Jiang *et al.*, 2003). Furthermore, although nanohedra is geared for design of symmetric materials, our results point to potential opportunities for advances in macromolecular docking and interface design in other contexts.

This first version of nanohedra will admit future improvements along various lines, including GPU enhancements to increase speed. Additionally, Python3 compatibility and a seamless integration of all software dependencies would further lower the barrier to entry for non expert users. These developments are underway. Critical experimental studies will be needed to evaluate the most important assertions concerning the expected advantages of generating assemblies with native-like interfaces. We also emphasize that successful design requires judicious amino acid interface design as a final step. This subsequent design step is separate from the construction of native-like (backbone-level) poses provided by nanohedra, though as noted above, nanohedra provides valuable information about specific amino acid preferences favored in the fragment interfaces. This position-specific frequency information can be exploited by sequence design programs in the final step of design.

Methods

Fragment database generation

To generate the fragment library, all non redundant, biologically relevant interfaces from high-resolution structures were gathered from the PDB. For homomers, structure codes for biological assemblies referenced in the QSBio database (Dey *et al.*, 2018) were used to extract all assemblies that were verified with a confidence ranking of 'high' or 'very high'. For heteromers, biological assemblies were identified using PISA (Krissinel and Henrick, 2007). The homo- and heteromer sets were next filtered to include only representative structures clustered at 90% sequence identity with a reported resolution ≤2.0 Å, experimental expression in *Escherichia coli*, no nucleic acids and no membrane proteins. For each identified structure, all unique interfaces between two separate chains were extracted excluding chains with less than 10 residues or fewer than 5 Cβ atoms within 8 Å of a second chain. From the resulting chain pairs, interchain Cβ distances were computed and residues that were 8 Å or less apart were selected as residue pairs across the interface. For each residue in the interface residue pair, the preceding and following two residues (i.e. i – 2 through i + 2) were included in the observation and the resulting 5-residue segments were stored, first as an individual 5-residue segment (individual fragments), and second as a pair of 5-residue segments across the interface (paired fragments). For residues with multiple conformations, the A conformation was chosen. Selenomethionine residues were not considered.

From the pool of individual fragments, a subset was chosen to perform all-against-all RMSD measurements followed by nearest neighbor clustering. The top five neighbor clusters were selected as the clustering population significantly decreased after this point. From each of the top five clusters, the fragment with the most neighbors was selected as a cluster representative, centered on the origin and stored. Each of these five clustered fragments represents one unique type of individual fragment, and the instance with the most neighbors was chosen as the fragment representative. For the saved paired fragments, both fragments in the pair were queried for membership in one of the five individual fragment types according to a Cα RMSD threshold of 0.75 Å. If one of the fragments in the pair did not belong to an individual fragment type, the pair was discarded from further classification. Next, each fragment in the fragment pair was subjected to a structural superimposition on its corresponding matched individual fragment representative. This centered one fragment in the pair at the origin aligned to its structural representative, while maintaining the relative position of the partner fragment to this aligned fragment. Once in this orientation, a set of three guide coordinates was stored, one coordinate at the partner fragment's central Cα atom, the second displaced by a unit vector along the C-alpha to subsequent carbonyl carbon vector and the third displaced by a unit vector perpendicular to the previous vector and lying in the plane formed by the C-alpha atom, the subsequent carbonyl carbon and the preceding amide nitrogen. This guide coordinate set, stored for each fragment observation, describes the transformation of the partner fragment's central Cα atom, and its relative orientation with respect to the aligned individual fragment representative. In this way, each partner fragment provides a unique spatially encoded and secondary structure-dependent observation of the interaction potential surrounding each individual fragment type.

Finally, for each individual fragment representative, and for each set of secondary structure-dependent guide coordinates of that fragment representative, a subset of those guide coordinates was subjected to all against all RMSD calculations followed by nearest neighbor clustering. The resulting guide coordinate clusters were binned with a maximum of 1 Å deviation, requiring at least four members in the cluster to be considered. From this set of guide coordinate clusters, all possible guide coordinates were subjected to membership in the resulting clusters by testing for the minimal RMSD to an established cluster. If a cluster with RMSD less than 1 Å could not be located, the guide coordinates were disregarded as outliers. This procedure was applied for each partner secondary structure associated with each fragment representative.

For each i, j, k fragment pair cluster, the cluster representative fragment coordinates and guide coordinates were stored. Additionally, the cluster size, mean guide coordinate RMSD and observed amino acid pair frequencies for central fragment residues were stored. The top 75% most populated i, j, k clusters were then chosen for our final fragment database.

Docking prospective SCMs

From the set of 124 possible SCM types, we chose six as diverse representatives for presentation in this study; note that all 124 were tested for mathematical and computational correctness in our earlier study (Laniado and Yeates, 2020). For each of these SCM types, homo-oligomers matching the design criteria were

curated from the PDB by searching for the desired point group symmetry, X-ray resolution better than 2.5 Å, a helical content greater than 30% and *Escherichia coli* as the organism used for protein expression. Structures containing membrane proteins or nucleic acids were removed. Biological assemblies were identified using QSBio (Dey *et al.*, 2018), and representatives clustered at 70% sequence identity were then downloaded from the Protein Data Bank. A few candidate oligomeric building blocks were then selected for pair-wise docking with nanohedra using the default parameters.

Design recapitulation

The dataset for the design recapitulation experiments was generated by selecting all successfully designed two-component tetrahedral, octahedral and icosahedral designs from previously published work (King *et al.*, 2014; Bale *et al.*, 2015, 2016; Brouwer *et al.*, 2019; Cannon *et al.*, 2020b; Ueda *et al.*, 2020); these cases met the criteria of agreement between the model and an experimentally determined atomic model. Failed designs (e.g. described as insoluble or unknown oligomerization state) were also identified from earlier studies (King *et al.*, 2014; Bale *et al.*, 2016).

For each successful design, the two component oligomers used for docking were extracted from the deposited PDB structure of the protein cage. For the failed designs, the PDB structures of the native oligomeric building blocks were used. Default nanohedra docking parameters were used with the exception of a 2° rotational sampling step instead of 3° for each component oligomer. Docking proceeded until all rotational degrees of freedom had been sampled. For 4NWN, we had to modify the initial default helix–helix fragment search to strand-helix. Since the dimeric component is mainly composed of β-strands on its surface, suitable docked configurations could not be identified with the default initial helix–helix search. Only the default helix–helix fragment search was used for failed designs, and designs were not considered in rare cases where no helix–helix interaction was present.

The Cα interface RMSD (iRMSD) was computed between the target design and each nanohedra output pose. For successfully designed structures, the coordinates deposited in the PDB were used as a reference. Models of the failed designs noted in earlier studies (King *et al.*, 2014; Bale *et al.*, 2016) were obtained from Neil King and Jacob Bale. For each design target, the 2000 docked poses with the lowest iRMSD to the design target were selected and nearest neighbor clustering was performed using all to all iRMSD calculations. Interfaces within 1 Å iRMSD threshold were clustered, then each cluster was ranked according to the nanohedra score of the cluster representative.

Amino acid frequency plots

The nanohedra program outputs amino acid frequencies for the central residue in the docked pose for each surface–ghost fragment match that has been identified. To calculate this frequency distribution, frequencies are retrieved from the fragment database for the corresponding *i*, *j*, *k* cluster for each surface–ghost fragment pair. When multiple surface–ghost fragments are identified for the same residue, the frequency distribution is a sum of the individual amino acid frequencies, proportionally weighted by the corresponding surface–ghost fragment match score. In this instance, the final distribution reflects the separate constraints of all identified fragments. Weighting the frequency distribution in this way provides a quantitative output for how well the amino acid identities from the fragment library fit within the specified docked conformation. To visualize these distributions, at each residue the resulting frequencies were transformed into multiple sequence alignments and sequence logos were generated using the WebLogo server (Crooks *et al.*, 2004).

Code Availability

The nanohedra source code is freely available at https://github.com/ nanohedra/nanohedra.

Supplementary Data

Supplementary data are available at *PEDS* online.

Author Contributions

The research was conceived by T.O.Y. and J.L. The code was written by T.O.Y. and J.L. The fragment database was constructed by J.L. The example SCMs were constructed by J.L. and K.M. The post facto analysis of designed protein cages was performed by K.M. and J.L. The manuscript was prepared by T.O.Y., K.M. and J.L.

Funding

This work was supported by National Science Foundation (Grant CHE-1629214). K.M. was supported by National Institutes of Health training grant (T32GM008496).

Acknowledgements

We thank Duilio Cascio and Alex Lisker for computing support and Michael Sawaya for helpful discussions. We thank Neil King and Jacob Bale for providing us with the models of their two-component cage designs.

Conflict of Interest

The authors declare no conflicting interests. Paper edited by: Professor Christopher Snow.

References

Adihou, H., Gopalakrishnan, R., Förster, T. *et al.* (2020) *Nat. Commun.*, **11**, 5425.

Bale, J.B., Gonen, S., Liu, Y. *et al.* (2016) *Science*, **353**, 389–394.

Bale, J.B., Park, R.U., Liu, Y., Gonen, S., Gonen, T., Cascio, D., King, N.P., Yeates, T.O., Baker, D. (2015) *Protein Sci.*, **24**, 1695–1701.

Ben-Sasson, A.J., Watson, J., Sheffler, W., *et al.* (2021) *Nature*, **589**, 468–473. Boyken, S.E., Chen, Z., Groves, B. *et al.* (2016) *Science*, **352**, 680–687.

Brouwer, P.J.M., Antanasijevic, A., Berndsen, Z. *et al.* (2019) *Nat. Commun.*, **10**, 4272.

Brunette, T.J., Bick, M.J., Hansen, J.M., Chow, C.M., Kollman, J.M., Baker, D. (2020) *Proc. Natl. Acad. Sci. USA*, **117**:8870–8875.
Cannon, K.A., Nguyen, V.N., Morgan, C., Yeates, T.O. (2020a) *ACS Synth. Biol.*, **9**, 517–524.
Cannon, K.A., Park, R.U., Boyken, S.E., Nattermann, U., Yi, S., Baker, D., King, N.P., Yeates, T.O. (2020b) *Protein Sci.*, **29**, 919–929.
Cao, L., Goreshnik, I., Coventry, B. *et al.* (2020) *Science*, **370**, 426–431.
Chen, Z., Boyken, S.E., Jia, M. *et al.* (2018) *Nature*.**565**, 106–111.
Chevalier, A., Silva, D.A., Rocklin, G.J. *et al.* (2017) *Nature*, **550**, nature23912.
Cock, P.J.A., Antao, T., Chang, J.T. *et al.* (2009) *Bioinformatics*, **25**, 1422–1423.
Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E. (2004) *Genome Res.*, **14**, 1188–1190.
Dey, S., Ritchie, D.W., Levy, E.D. (2018) *Nat. Methods*, **15**, 67.
Edwardson, T.G.W., Tetter, S., Hilvert, D. (2020) *Nat. Commun.*, **11**,
5410. Ernst, P., Plückthun, A., Mittl, P.R.E. (2019) *Sci. Rep.*, **9**, 15199.
Fallas, J.A., Ueda, G., Sheffler, W. *et al.* (2017) *Nat. Chem.*, **9**, 353–360.
Finkelstein, A.V. and Ptitsyn, O.B. (1987) *Prog. Biophys. Mol. Biol.*, **50**, 171–190.
Fleishman, S.J., Whitehead, T.A., Ekiert, D.C., Dreyfus, C., Corn, J.E., Strauch, E.M., Wilson, I.A., Baker, D. (2011) *Science*, **332**, 816–821.
Gao, M. and Skolnick, J. (2010) *Proc. Natl. Acad. Sci.*, **107**, 22517–22522.
Gonen, S., DiMaio, F., Gonen, T., Baker, D. (2015) *Science*, **348**,
1365–1368. Guharoy, M. and Chakrabarti, P. (2007) *Bioinformatics*, **23**, 1909–1918.
Harris, C.R., Millman, K.J., Walt, S.J.v.d. *et al.* (2020) *Nature*, **585**,
357–362. Heater, B.S., Yang, Z., Lee, M.M., Chan, M.K. (2020) *J. Am. Chem. Soc.*, **142**, 9879–9883.
Jacobs, T., Williams, B., Williams, T., Xu, X., Eletsky, A., Federizon, J., Szyper- ski, T., Kuhlman, B. (2016) *Science*, **352**, 687–690.
Jiang, S., Tovchigrechko, A., Vakser, I.A. (2003) *Protein Sci.*, **12**,
1646–1651. King, N.P., Bale, J.B., Sheffler, W., McNamara, D.E., Gonen, S., Gonen, T., Yeates, T.O., Baker, D. (2014) *Nature*, **510**, nature 13404.
King, N.P., Sheffler, W., Sawaya, M.R., Vollmar, B.S., Sumida, J.P., André, I., Gonen, T., Yeates, T.O., Baker, D. (2012) *Science*, **336**, 1171–1174.
Krissinel, E. and Henrick, K. (2007) *J. Mol. Biol.*, **372**, 774–797.
Lai, Y.T., Reading, E., Hura, G.L., Tsai, K.L., Laganowsky, A., Asturias, F.J., Tainer, J.A., Robinson, C.V., Yeates, T.O. (2014) *Nat. Chem.*, **6**, nchem. 2107.
Laniado, J. and Yeates, T.O. (2020) *Proc. Natl. Acad. Sci.*, **117**:31817–31823.
Liang, M., Fan, K., Zhou, M., Duan, D., Zheng, J., Yang, D., Feng, J., Yan, X. (2014) *Proc. Natl. Acad. Sci.*, **111**, 14900–14905.
Liu, Y., Huynh, D.T., Yeates, T.O. (2019) *Nat. Commun.*, **10**, 1864.
Marcandalli, J., Fiala, B., Ols, S. *et al.* (2019) *Cell*, **176**, 1420–1431.e17.
McConnell, S.A., Cannon, K.A., Morgan, C., McAllister, R., Amer, B.R., Clubb, R.T., Yeates, T.O. (2020) *ACS Synth. Biol.*, **9**, 381–391.
Mitternacht, S. (2016) *F1000research*, **5**, 189.
Padilla, J.E., Colovos, C., Yeates, T.O. (2001) *Proc. Natl. Acad. Sci.*, **98**, 2217–2221.
Pearce, R., Huang, X., Setiawan, D., Zhang, Y. (2019) *J. Mol. Biol.*, **431**, 2467–2476.
Pedregosa, F., Varoquaux, G., Gramfort, A. *et al.* (2011) *J. Mach. Learn. Res.*, **12**, 2825–2830.
Polizzi, N.F. and DeGrado, W.F. (2020) *Science*, **369**,
1227–1233. Silva, D.A., Yu, S., Ulge, U.Y. *et al.* (2019)
Nature, **565**, 186–191.
Sinclair, J.C., Davies, K.M., Vénien-Bryan, C., Noble, M.E. (2011) *Nat. Nan- otechnol.*, **6**, 558.
Stranges, P.B. and Kuhlman, B. (2013) *Protein Sci.*, **22**, 74–82.
Suzuki, Y., Cardone, G., Restrepo, D., Zavattieri, P.D., Baker, T.S., Tezcan, F.A. (2016) *Nature*, **533**, 369.
Tischer, D., Lisanza, S., Wang, J., Dong, R., Anishchenko, I.K., Milles, L., Ovchinnikov, S. and Baker, D. *biorxiv.* doi: doi.org/10.1101/2020.11. 29.402743..
Ueda, G., Antanasijevic, A., Fallas, J.A. *et al.* (2020) *Elife.* **9**, e57659.
Xie, Z.R., Chen, J., Zhao, Y., Wu, Y. (2015) *BMC Bioinformatics*, **16**, 14.
Yeates, T.O., Liu, Y., Laniado, J. (2016) *Curr. Opin. Struct. Biol.*, **39**,
134–143. Zhou, J. and Grigoryan, G. (2015) *Protein Sci.*, **24**, 508–524.
Zhou, J., Panaitiu, A.E., Grigoryan, G. (2020) *Proc. Natl. Acad. Sci.*, **117**, 1059–1068.

Supplementary Information

A fragment-based protein interface design algorithm for symmetric assemblies

Joshua Laniado
Kyle Meador
Todd O. Yeates

This PDF file includes

Supplementary Text

Supplementary Figures S1 – S3

SUPPLEMENTARY FIGURES

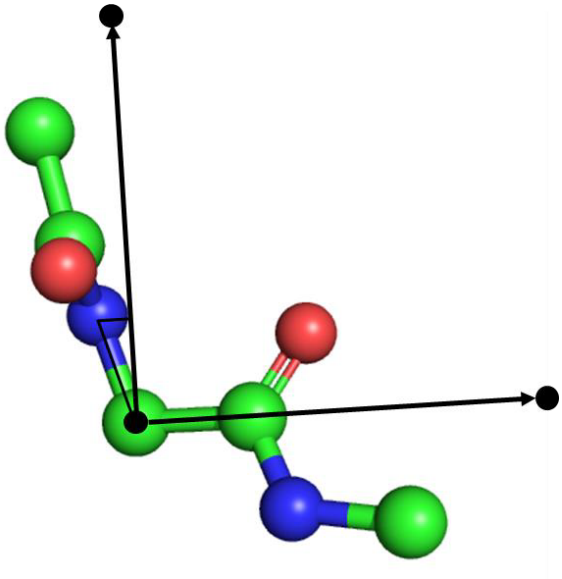


Fig. S1. Guide coordinates. An orthonormal three-atom system is constructed on the central C-alpha position of a 5-residue fragment to provide a reduced representation of its position and orientation. The unit vector length for typical calculations in Nanohedra is set to 3Å.

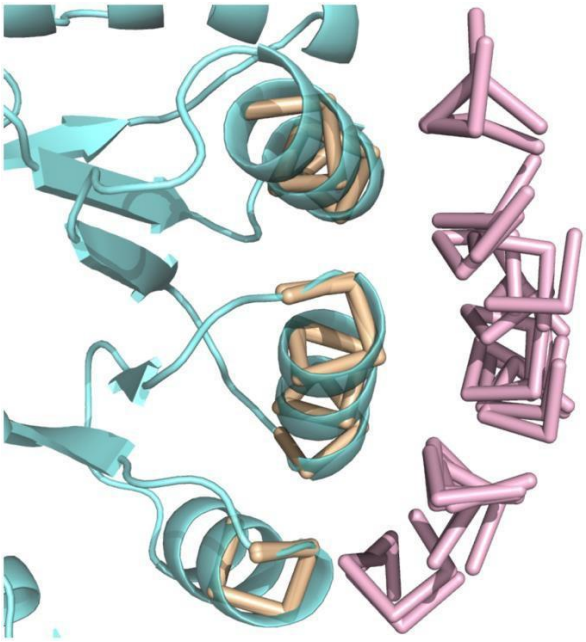


Fig. S2. Example illustration of ghost fragments. Oligomer 1 (cyan) is shown with its surface fragments (tan) and associated ghost fragments (pink), which represent candidate fragment associations.

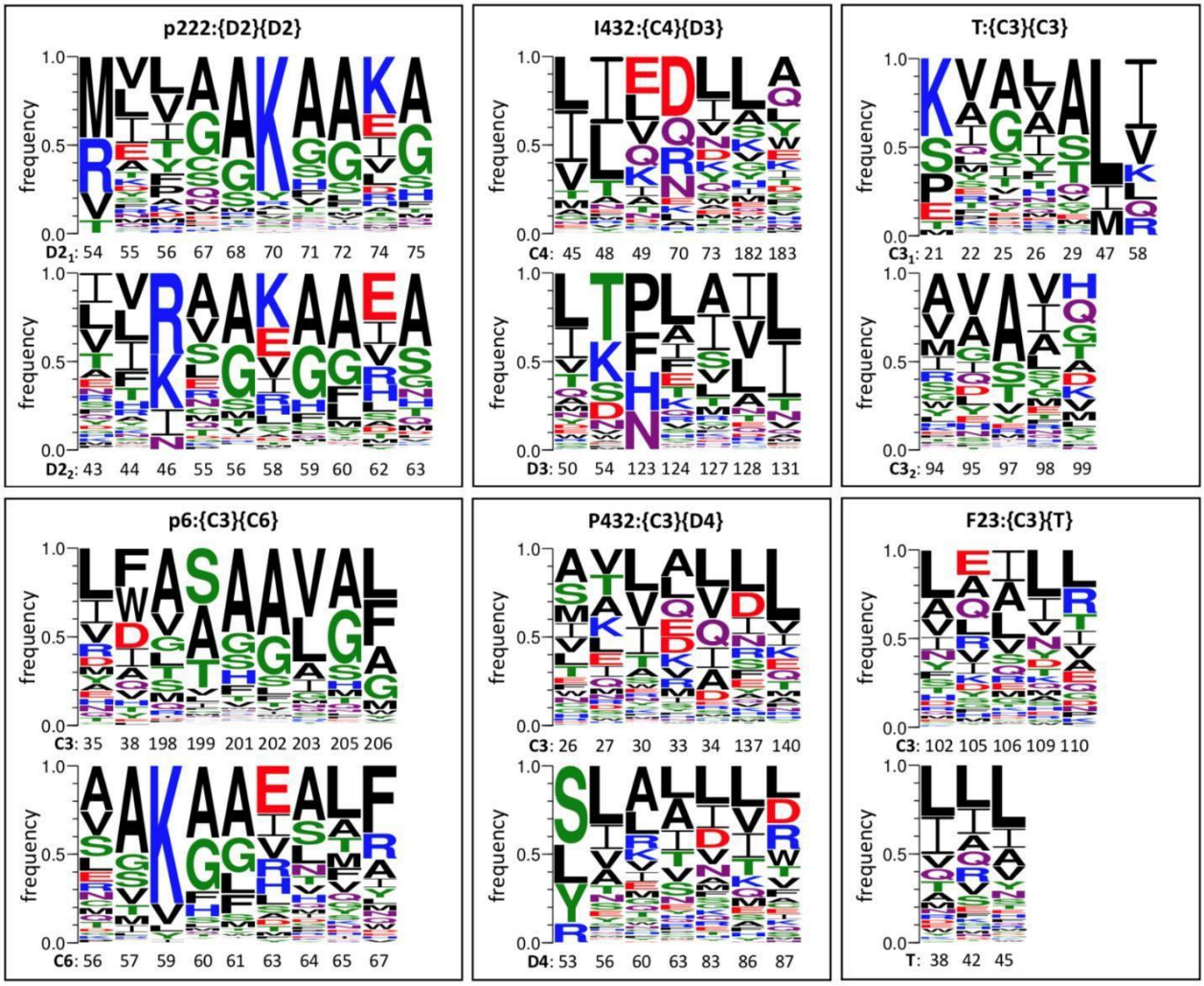


Fig. S3. Deduced amino acid preferences for prospective SCMs in Figure 3. For each case, the top and bottom diagrams indicate the preferences for the two oligomeric components, whose symmetry types are noted to the left of the participating residue numbers in the interface.

SUPPLEMENTARY TEXT

<u>PDB IDs and design names used for design recapitulation experiments</u>

PDB IDs of the experimentally validated designs:
4NWN, 4NWO, 4NWP, 4NWR, 4ZK7, 5CY5, 5IM4, 5IM5, 5IM6, 6P6F, 6VFH, 6VFI, 6VFJ, 6VL6

Names of the 'failed' designs:
I32-01, I32-03, I32-05, I32-07, I32-08, I32-12, I32-13, I32-14, I32-15, I32-16, I32-17, I32-20, I32-22, I32-23, I32-24, I32-25, I32-27, I32-31, I32-33, I32-34, I32-35, I32-36, I32-37, I32-38, I32-39, I32-40, I32-41, I32-45, I32-46, I32-49, I32-52, I32-53, I32-54, I32-55, I32-56, I32-60, I32-62, I32-

64, I32-65, I32-68, I32-70, I52-01, I52-07, I52-09, I52-10, I52-11, I52-12, I52-14, I52-17, I52-18, I52-20, I52-22, I52-23, I52-24, I52-26, I52-27, I52-28, I52-29, I52-31, I52-34, I52-35, I52-36, I52-38, I52-39, I52-40, I52-41, I52-42, I52-43, I52-44, I52-46, I53-06, I53-09, I53-12, I53-15, I53-16, I53-21, I53-23, I53-25, I53-27, I53-28, I53-29, I53-33, I53-35, I53-37, I53-43, I53-48, I53-49, I53-58, I53-61, I53-62, I53-63, I53-64, I53-67, I53-68, I53-70, I53-72, I53-74, I53-75, I53-79, I53-80, I53-82, I53-83, T32-02, T32-03, T32-06, T32-07, T32-08, T32-09, T32-11, T32-17, T32-18, T32-20, T32-24, T32-25, T32-26, T32-27, T33-01, T33-02, T33-03, T33-04, T33-05, T33-06, T33-07, T33-08, T33-11, T33-12, T33-13, T33-14, T33-16, T33-17, T33-18, T33-22, T33-23, T33-24, T33-25, T33-26, T33-27, T33-29

Fragment database

For every fragment pair cluster in the fragment database, a PDB file containing the extracted and canonically oriented representative is stored in the GitHub repository in the following directory:
/fragment_database/Top75percent_IJK_ClusterRepresentatives_1A. Additionally, all of the PDB entries from which these fragment representatives were derived are listed below:

4bku, 2cwq, 2o16, 3l77, 4dyv, 3eei, 3ek3, 2bfd, 1qbz, 3g1t, 2d4u, 4ep4, 4cw4, 5vip, 2d8d, 3h7f, 3ntl, 3l1n, 4p3m, 1g0o, 4wbd, 2fic, 1nkd, 2cxn, 2pbr, 4is2, 1vgm, 3mzv, 3frq, 1yxm, 3iso, 3b4w, 1hxh, 2gjl, 3b5n, 2c07, 2h8o, 4wec, 1g3k, 2yo3, 4od8, 3ucs, 5eck, 2x9g, 3osu, 4pz1, 3rr1, 3a5r, 4jo7, 1l5x, 3afn, 3svt, 1f4n, 3q1p, 3mms, 1d9c, 1vqu, 1ykd, 4f7u, 1pwb, 1pr9, 2ah6, 3hht, 3ged, 3t88, 3n3a, 3m3h, 5jge, 2pd6, 3a9z, 1t6o, 1t2a, 4wk5, 3is3, 3n9t, 1k3s, 1a7w, 5w83, 3shg, 4mqb, 2hq1, 3cp1, 2wsb, 3pko, 1xuq, 1jq0, 1eye, 5nps, 4fc7, 1yib, 3m91, 3kxc, 2bgk, 4dn2, 3lxy, 1n7s, 2dyo, 4ipi, 4m89, 1zv8, 4tkl, 2pr5, 5hvz, 3rwb, 2ieq, 1sum, 3wmi, 4kae, 3etn, 3w8e, 6cwp, 1orr, 3k31, 4qto, 2x02, 4is0, 3b09, 4dqx, 3wfv, 5jje, 1ybz, 3qk8, 1pn9, 4qfh, 4liw, 6ijf, 3f6d, 3v1t, 2gdz, 2f22, 3uce, 2qyo, 4i0b, 6akl, 5omb, 4ni5, 1aj8, 4m8s, 5yl9, 4o9a, 1k04, 2z4v, 2ibl, 4oy3, 4avm, 3oid, 2dul, 3obi, 4e3z, 3zv4, 2p58, 1g8e, 2p8u, 1x1t, 1xsv, 2hrz, 5h66, 3frc, 3h7a, 4gis, 3n27, 1gco, 4nbw, 3aha, 4x7y, 4atm, 2ag5, 1bkj, 2zsi, 3tl3, 3csx, 4oun, 2p8c, 4n5m, 4q04, 4h15, 4xr9, 2cvz, 3ado, 4q94, 2guk, 4jro, 3wp8, 4bvq, 4wzx, 4hrg, 1j20, 3o38, 3etq, 1vke, 1dug, 2hng, 4msp, 3eer, 4l8p, 2j9u, 4cyd, 3h6p, 4kqw, 1vmg, 2i7g, 4fp4, 4nbt, 4nbv, 4wba, 2ogi, 4hz2, 3f3x, 1gg1, 4g10, 3bg2, 6fpg, 4g41, 2qib, 2hx5, 2for, 3ljk, 1nff, 3n0l, 2yva, 4mh4, 1wt6, 4i6r, 2pnf, 2yf9, 4p1m, 2zhz, 3qxz, 3npk, 3f6f, 4mso, 3bem, 4lls, 4wxg, 4wj9, 2oyc, 3ux2, 2pqq, 2o0r, 5mvw, 2nt8, 1mdo, 2pa8, 3f13, 4ps2, 3egw, 4muz, 3box, 1nwh, 3vtc, 4laf, 3gr4, 3qhq, 4aj9, 3sf6, 3r9p, 3ls9, 2ybq, 4o47, 1bg7, 3cnu, 3t7c, 1r66, 3ttv, 3ho9, 6dey, 1iye, 2chp, 3l84, 1zmt, 1zav, 2cf7, 2e1z, 3sbf, 4kxv, 1s0a, 4ls6, 3e60, 2eo5, 3zwq, 3s46, 1gsu, 2c3f, 1vef, 3a2v, 3ijf, 2y2z, 4c1l, 3tl2, 2inc, 3m0g, 2ph3, 4csr, 3ond, 4duq, 4cv2, 4cy9, 1i0h, 1sff, 3q58, 3op4, 3wb9, 3eof, 4lsm, 2ae2, 3vmk, 4one, 4pxo, 4iy7, 3m0f, 3k4i, 3ozf, 1vhw, 3o04, 4mow, 3tjr, 4grd, 3mpz, 2pzm, 2dtx, 1zsx, 3lrt, 2qae, 4fha, 3tw9, 4iel, 2iwz, 3mt6, 3rcy, 5fcf, 1lo7, 2f5g, 4ahq, 3pgx, 3imf, 2z98, 1fmc, 2fzv,

2ehh, 1uay, 2c0r, 4k6o, 4go7, 3it4, 4mqq, 4k6f, 3qtp, 2wyu, 2cb1, 2v40, 3ucx, 3l7o, 4fs3, 4jqp, 2ix4, 3zrq, 2ox4, 1wy2, 3r4z, 4lui, 4phj, 4dgq, 4kwh, 4g81, 2e5f, 2jam, 4bmn, 4hp8, 3ors, 3uve, 3q0e, 3bhg, 3bl6, 1yhl, 5xoq, 2eg2, 1xw8, 3ge4, 4ag3, 1e93, 3u0g, 3wds, 1xql, 3o9z, 1w53, 3gy1, 2eq8, 1ii5, 4ny3, 4qol, 1prx, 3e04, 3zcd, 1f3a, 2i87, 4onc, 3a04, 1xw5, 2yxn, 1qwl, 3qpm, 1pl8, 1o1x, 3op7, 1s57, 3ro0, 4h9n, 1zk4, 4aov, 3icc, 1ks2, 2d7v, 3kkz, 3ech, 2b18, 4ham, 1ozh, 2q7v, 3dj6, 1fp1, 2cyb, 1njk, 3bm1, 3jva, 3khy, 4ijn, 2p2w, 3fpc, 2d5k, 4g91, 4h8e, 4aum, 2ibp, 2ffx, 2dc0, 1ujn, 3pvd, 1q08, 3exe, 1yqg, 1lk5, 1f5v, 4m20, 4rj2, 1tu7, 1bxk, 4hhp, 3o0h, 3elb, 6mpz, 3flu, 3sfw, 3bjk, 3oec, 3ts7, 3lvf, 4jzx, 2ob5, 3i1j, 3ndn, 3r1i, 3mdk, 1o5k, 1zxx, 1pjc, 1b4p, 1uzn, 1vf1, 3uw3, 1d7o, 2a72, 1n7k, 4e4f, 2ql8, 3p9c, 1rcq, 3ib5, 3tso, 3chv, 2pa6, 4itu, 1y9a, 3fpl, 3ooo, 4lmp, 4xeu, 2pzh, 4dmg, 3tji, 3o03, 2cjg, 3iav, 5o6t, 4ese, 3nvt, 4weo, 1xmp, 4adm, 3ay7, 3vp5, 1lqa, 6a7v, 1fw1, 1e6w, 3nx3, 2d1y, 4ma9, 4ued, 4hoj, 2eq7, 3gd6, 2e0c, 4f66, 1wwr, 3ak8, 4f4r, 4kct, 1rcd, 3g0s, 3hid, 3ai3, 4lfg, 2v2p, 3ce6, 4dbd, 4ddo, 3zqi, 1o54, 3o1c, 4lvu, 5ein, 2izz, 4e4y, 1n3l, 2f6g, 1j55, 1jqb, 4rhf, 2qfa, 3pu9, 3e5n, 3n6w, 1o5x, 4io1, 4r31, 4lcq, 1wve, 1o5j, 2ibd, 1b5p, 3dhz, 6h7b, 3n37, 2gcq, 2dko, 3ndo, 1f1u, 1nuy, 2wns, 1fec, 2gtd, 3ixq, 3gzy, 4ed9, 3awd, 1pyo, 1gwf, 1u11, 3tvj, 2x4d, 5xhz, 2zj3, 1zuk, 3qp6, 2zu2, 3e49, 3vcn, 4plg, 5fvk, 4okv, 2z3h, 3dms, 4p7t, 2pn8, 4hah, 1xng, 5vap, 3uh0, 1vpm, 3oj6, 3a8g, 4zgj, 2eq6, 4kbf, 4rit, 4mkn, 1r7a, 2x5f, 4qvt, 3k40, 4dbh, 1va0, 1wlu, 4e98, 2jah, 1mxi, 3ccg, 2xwl, 3mws, 1o4s, 6g5g, 3ewy, 4oqz, 2ptz, 3ke4, 5t46, 3p4u, 3kgw, 4eqs, 1zjr, 1j9j, 4o5o, 2p0u, 3r87, 3s4k, 3zho, 3odg, 3gem, 2rfv, 2fym, 3bqy, 3ei9, 1xg5, 3kqf, 4kms, 3oow, 1zjj, 1nog, 1s3z, 1iq6, 2g76, 4q0a, 3t32, 2hjp, 3ld3, 4wnd, 2pok, 6mjc, 3ekg, 4ip7, 2w5w, 2fuj, 1c3c, 1izc, 2ynz, 3qns, 3l9w, 3nwr, 4cok, 4izh, 1hqs, 1d3y, 4jdp, 1nu5, 3n4j, 3no4, 2yob, 2hvw, 1tu1, 4edh, 5gp7, 4l9p, 5mj3, 4air, 3kus, 3dc5, 3bn4, 1q98, 3dk9, 1ecf, 3r6h, 3ppi, 2i6d, 3l8u, 4xfj, 2o23, 4jak, 2h9b, 3kw2, 4dye, 4em8, 4ld8, 3f0h, 3ha2, 4h51, 4ox6, 3sdo, 2vpk, 1pym, 4lsb, 3ci3, 2yqu, 3iq1, 4wd2, 3q62, 4f32, 1zn8, 4u5w, 3k7p, 3sz7, 2o7c, 2p5v, 3jrz, 4jb8, 1o4v, 4jr2, 4mb6, 3d4p, 3wgq, 2b8t, 3zrj, 3b46, 4q34, 3r6k, 4fiv, 2d29, 4yhy, 4mn9, 4p61, 2akz, 4lvc, 1ox0, 4bvx, 1yxy, 1mzh, 1y2m, 2qve, 3hvv, 3rsi, 2e8e, 1m7s, 4jnq, 4ghg, 2geb, 4ei6, 1e5m, 4h0p, 1f8m, 5zjg, 5kp7, 1t9m, 4ln1, 3di4, 4ob0, 1j8b, 1jub, 4k7z, 1q7e, 3h4o, 3r77, 4glj, 2ptr, 3vub, 4c5s, 2j5s, 3ry8, 1f1x, 2q3b, 1lj9, 3ehu, 3ndd, 4mzw, 3evk, 3r3s, 2y4r, 4hmw, 1dlj, 3qy1, 1sov, 3wlv, 2p1f, 1pzg, 4ho1, 4eso, 4ms8, 4inc, 4l6w, 2ze3, 4ds3, 2nyn, 3dmo, 1o60, 1h05, 4gwg, 3kom, 1zb9, 3emf, 2bmo, 4ef8, 2cfu, 4uak, 2hxi, 3p8k, 2fe3, 5bpk, 3vpb, 4mch, 4hyr, 2cb0, 2yhw, 4jh2, 2ab0, 1xg4, 4d13, 2q7w, 1qxo, 4k8g, 4jj7, 2ogd, 1iho, 3m9y, 3rjs, 3va8, 4e08, 1gkm, 3u9r, 1w27, 1sg4, 2q0l, 1ig0, 3c8e, 3rf4, 2if5, 3urh, 3i3g, 3ztv, 4ofx, 1rew, 3e9a, 3ej9, 4axj, 4a0s, 4koq, 4fdx, 1duv, 4dh4, 2ou6, 1r5t, 1z4e, 1jq5, 2d1c, 3l07, 3tak, 4u8f, 3u4j, 1wwk, 4y99, 4fkx, 2o08, 2onf, 4ogd, 1vpb, 1h72, 3n3m, 1aie, 4rle, 1mqe, 2v6a, 4fay, 1cq4, 3qwa, 1uwz, 1ygt, 5xlu, 3mz0, 1t3i, 2xsu, 2o2p, 4ffk, 3sm1, 2y53, 4c8i, 1z0s, 1gk9, 1y0y, 1euh, 5by8, 2z5b, 1oc2, 2c1s, 3bl9, 4yfb, 1nvm, 4lrt, 4kna, 1p7k, 3ba1, 2xqq, 3typ, 3d9a, 3nua, 3s1t, 6gsv, 1udv, 3b33, 2wqi, 2iks, 1rp0, 1qb7, 3h12, 1uuf, 2v5j, 3m21, 4m1e, 4pua, 4egu, 4ytw, 3mb5, 2vu5, 1uwk, 4rhs, 2xcz, 4hrv, 4hst, 3ve9, 4usl, 4gkb, 2cwl, 4gci, 4i8p, 4xa8, 4d6q, 1mp9, 2nyi, 2kin, 3chb, 4tsh, 1lb3, 2dg5, 1vjl, 3dy0, 1g2o, 3hg7, 4ae7, 4hvc, 4npi, 3kzn, 3zxq, 2ekl, 1n13, 1qcz, 3lyx, 2qg8, 3ge6, 1m3s, 1gtz, 2hiq, 1rlk, 1wn2, 2vxt, 4xb6, 3kcc, 4l8e, 1iat, 1js1, 3bzq, 5wy2, 3a68, 4mum, 1tcv, 1kcb, 4bmu, 1msc, 3las, 4n45, 2qj8, 4b2h, 2qiw, 2vbf, 3o74, 4j2f, 3ndc, 3b02, 4kp7, 1yya, 4o6r, 1i07, 1sbk, 4xxv, 5n8b, 3vpc, 3awk, 1x54, 2pqm,

2c2u, 4fxi, 1x6v, 1b9m, 2npn, 1scj, 1mka, 3o4r, 3e97, 2c3b, 3aa0, 5ted, 3tjj, 4k3l, 4jbb, 4hnl, 5me5, 4aeq, 2z3z, 4jbg, 5om2, 6bw9, 1l2w, 2dc1, 3lp6, 4mf4, 3whr, 2fiu, 4q75, 4gel, 2bky, 3qag, 3zqu, 3c3y, 1t0a, 2qud, 1u0k, 4e5m, 1q7s, 2z1m, 4rao, 3ty4, 3ry0, 3fyn, 3v9o, 3r2r, 3h2d, 4nog, 1v84, 3lwz, 2fnu, 3gg2, 4njy, 1r9c, 2f9w, 2zsc, 6at6, 6cph, 2xgz, 1otf, 3kti, 1i9g, 3dod, 4weq, 1g2q, 3meq, 1cnz, 2o3j, 3nd1, 4wvi, 3uko, 3tdt, 3ju4, 2q3p, 4g5a, 2ypo, 4r2x, 1d5i, 2o28, 3lx3, 6aq7, 2d51, 1t4b, 2fwv, 4qam, 4e1p, 3vtn, 3o1k, 4evy, 2arz, 3mhy, 1kjq, 4otm, 3qjh, 1lk3, 2e6f, 4qq3, 1wly, 2yc3, 1f3u, 4cn0, 1osp, 3ep6, 1xbw, 4bl0, 3rdu, 6icc, 3h9e, 1m93, 3p4e, 2ih3, 4qyo, 5ow0, 2gxq, 2ib8, 4e45, 3wih, 2j32, 4o5l, 1cdc, 1yqd, 4i9b, 3kxq, 6a3w, 5ncw, 3g48, 3eot, 3i90, 1mtp, 1p1l, 1f9z, 3bcw, 2bc3, 4ajy, 3vw9, 1usm, 4pbc, 3ss6, 4usi, 2ltn, 2g5l, 2z3c, 3mff, 3dgp, 2fur, 1j05, 3lzl, 4oo7, 2nml, 3ers, 4hiz, 3hi2, 6eh4, 1ng2, 3njn, 1lgp, 1vj0, 5nhw, 6bfs, 2xt2, 1zgx, 3cls, 4pyj, 2vu1, 2nw2, 2ux9, 1su2, 2f0c, 3acz, 2ixd, 2a1h, 1he7, 2i2c, 5w5z, 3afo, 2h1t, 3oti, 2f01, 4miy, 4rsp, 2avt, 1ov3, 3asu, 4m2m, 4jj2, 4itx, 6fun, 2hhz, 4ku0, 6gny, 3prl, 1mvf, 1jyo, 5v1v, 3pjl, 2aps, 3ids, 3ht1, 3laa, 3wjp, 6fbk, 3nw4, 4k7x, 6mlc, 5n22, 1a4i, 3phc, 3l9y, 2eiy, 4rd7, 1je0, 2j27, 4yx7, 4lmb, 4je1, 2o66, 4lmy, 4q3n, 2wvg, 1wz3, 3cog, 3v1y, 5ctd, 4m1g, 4l19, 2an1, 1ve2, 4cay, 4r82, 2z30, 4kam, 6ba5, 2e0n, 3arn, 3q80, 1w70, 4jbn, 3d2y, 1x12, 2pv2, 4f4e, 1kr4, 3c9u, 5f67, 3qhx, 4ep8, 1or7, 1p9h, 1k9u, 3bvf, 1f18, 2coi, 5tvo, 2q2h, 3lv4, 2yo2, 3erp, 2zdp, 1vkn, 4aan, 1zzg, 4iwk, 4bf5, 3fwn, 4pdc, 2q2i, 1xvq, 2rbd, 1jvb, 4hem, 2cih, 5nwg, 1vlg, 5dhm, 1inl, 6bjz, 1cs1, 2gw8, 1vlr, 3qq6, 2zsl, 3q46, 2xdp, 1h6w, 3euo, 4dpl, 4udt, 1wl4, 2o4j, 3fv9, 2qap, 2r8o, 4ne3, 4r60, 3e8o, 2gff, 1vfs, 2vc6, 4eei, 3fkc, 2zyj, 3lc0, 1ns5, 2vre, 3hpw, 4f47, 1xa3, 2pn6, 3mbk, 4grn, 5ni9, 2y4d, 4ox8, 1ok7, 3r9t, 3i28, 4dza, 1ykw, 2nxw, 1umd, 2c31, 2pgw, 1f9r, 3kh8, 2xfx, 2bjf, 4w5k, 1uxj, 3v3w, 2dkj, 4dzh, 2xsx, 2nuh, 4lfy, 2ha8, 4lfl, 3g8y, 1moq, 2p3e, 1w85, 4jem, 3lot, 4of4, 3myb, 4bi6, 3nk6, 4kkm, 3epr, 3hqn, 1l6r, 5cq2, 4ez8, 1bi5, 4a57, 1qak, 3bio, 3jtm, 1wtj, 3fpz, 3mad, 2egv, 3oc7, 4jad, 2pfm, 3jtx, 2yg3, 3bq9, 1o4w, 4rg1, 1ooe, 3gr3, 3uj2, 2rkf, 4nbu, 4gmk, 3ta6, 3a2o, 1vim, 4lza, 4oox, 3t3w, 1vjo, 3nco, 3rpe, 3lqs, 2o9v, 4aty, 4yfv, 3jz6, 4c9b, 5ovo, 2vd8, 2vxn, 2wu9, 3vqj, 4wnn, 1vl2, 1u8f, 1gad, 4dd5, 1lq9, 1zod, 4oyk, 3ioy, 3ddh, 4beu, 4b98, 4fnq, 4imq, 2r37, 4u6d, 2w3q, 2epi, 5vq3, 2c4n, 4grc, 2f8a, 2fr5, 4ia6, 4bgl, 3ucj, 4efc, 3njd, 3mbd, 2j91, 1w6u, 3tj8, 1twd, 2j1p, 2qj2, 2nx8, 2z1n, 1iom, 3m0m, 4h7p, 3dr3, 2bd0, 4llo, 4o9k, 4ffc, 3g9k, 3r1w, 2xqh, 4k00, 3pss, 1um0, 1y0h, 4efi, 4bv1, 3vpd, 4cz5, 2b0a, 3nyt, 1z9p, 2z26, 2w8t, 4n6a, 4tm5, 2pij, 3rr2, 3ixc, 4hzd, 4w78, 4ir0, 1sr4, 2jf2, 3dwg, 6f45, 1g97, 4ea9, 3bri, 3i5t, 3bpk, 3cj8, 3tk8, 5nq0, 2h0u, 1tvz, 1y42, 3pmo, 4e6u, 2d5m, 4eqy, 4dq6, 4itb, 1k2x, 4qgr, 1n0w, 6hy2, 3g5o, 1h3f, 4k02, 2ps1, 2pwy, 3pc3, 2v3z, 2ex2, 1lwd, 3mmz, 2i8b, 1o20, 4aff, 3sz3, 3ovp, 3h6x, 2gpc, 3k9w, 1h2b, 1wvg, 2qde, 6da1, 3g3k, 2y7e, 3fdo, 3ivr, 2ord, 1zch, 3fcp, 3oig, 3fa5, 3b7h, 1k3y, 2pex, 3e39, 3g0t, 3r6f, 4l8l, 2vws, 3v1v, 3ge3, 2dfa, 1pb0, 4rk4, 4fgj, 3mcw, 4h2h, 3ngs, 2qq0, 4x54, 4jed, 4f1w, 3k2v, 4fn4, 5o9e, 3qy3, 2nuw, 3enk, 3ohe, 2vd4, 1xky, 3n01, 4dxk, 3p2n, 5y53, 2v1p, 3eus, 3s1l, 3glv, 3dgb, 4qec, 1sqs, 4o6v, 3o76, 3i47, 4kn5, 2znd, 2e2r, 1on3, 1gde, 1k92, 1ll2, 3f1l, 2xz9, 4jig, 4bgv, 2ebb, 3ono, 2fmm, 3l8a, 2uv4, 1m6j, 4nda, 4jpf, 1vra, 1psr, 4cr6, 3p8l, 2cdu, 2v9l, 1j5x, 6cxt, 1a05, 4ic3, 4cv4, 1w5q, 3lw3, 2rk9, 2vm8, 1ohl, 4iqg, 3sl7, 3jth, 4uuu, 3lcn, 3swo, 5d7f, 4ri6, 4nr0, 4eg0, 4a25, 4ath, 5k99, 2ef8, 1p9b, 4ecj, 1ne7, 3sgv, 3lyp, 3msu, 1f74, 3q98, 4uux, 1w6g, 3mf7, 1n7h, 1ixl, 1vkc, 3iwc, 3i44, 3keo, 3cu0, 4mf7, 3tfw, 2b1y, 4ouj, 2wqk, 1mo0, 2eh6, 3onr, 2b5v, 2gdg, 1np6, 1i0r, 3n5b, 4u3w, 4wkz, 4rot, 2adf, 1tgj, 2pn2, 3sdb, 4js0, 3cje, 3d40, 4o1k, 4hi7, 3lyh,

1y55, 1vdh, 4bmo, 4hfm, 2gw4, 1eo6, 1i7q, 4f3w, 4n18, 4pv2, 3rnr, 4f0b, 2bez, 2yzj, 3dn7, 4h31, 3f4f, 2e4n, 1rkx, 2vez, 3u55, 4b79, 1oth, 4j4r, 2dya, 4o5h, 4lgo, 4omn, 1x13, 3heq, 1l1q, 4n7i, 2h3g, 1yar, 4hh3, 2hpe, 4b4u, 4n0v, 3wnv, 4faz, 1q6z, 2rcz, 2y3q, 4c5e, 5mzw, 3n8b, 2y3n, 2gbw, 3zbg, 2cch, 2w8n, 2xwx, 4j1v, 1d6j, 1vhy, 2okg, 4utu, 6f6r, 4agh, 3lv0, 2j8m, 3pfe, 3oit, 2dlb, 2pbp, 3f3s, 1dwk, 4hws, 1ppv, 2ykf, 1y1x, 4edy, 5vmr, 2oek, 1kew, 3qhd, 1k51, 3n4i, 1y7l, 3n10, 2zej, 4n1d, 4wks, 2aef, 4ag7, 4wjq, 2pvq, 4jb7, 5jpo, 3erj, 4eez, 4ak5, 2q8g, 4fce, 1yhf, 1q8f, 4bk8, 3lqw, 2igq, 1qre, 1m5s, 3plx, 4e3x, 3bsw, 1nh8, 2p0a, 3h8u, 3q1x, 5t86, 3gwa, 3lru, 4dqn, 2rij, 4a0z, 4wk1, 2aqs, 2ewh, 1xhd, 4juu, 2axw, 2hlj, 3ho7, 4m98, 5m02, 1kjv, 3liy, 3cww, 3tbh, 3q6s, 4ear, 2zzd, 4lhr, 3nrq, 3i4q, 2dp9, 4ob5, 2qsi, 4nat, 4gk6, 6ddm, 1vfj, 3hmz, 2c2x, 5was, 4u1e, 2zyz, 4p3c, 3s2r, 3mdx, 2jg1, 3eg4, 2gx0, 3riq, 3d7j, 1ljo, 4lx3, 3lss, 2rk3, 2q4v, 2r4f, 3ajv, 4lx2, 2cdc, 4jks, 2vt1, 2fa8, 2hkn, 2pii, 1uku, 4chi, 1ocy, 4zox, 2rbb, 3e4v, 1k2e, 3fj5, 3ano, 1y9w, 1n12, 2ifx, 1w2w, 4le9, 1i52, 3hza, 1kq3, 3cjs, 2dxq, 4umx, 2ovg, 2o38, 4ecp, 3ulb, 3ewn, 3ft7, 3pxx, 2x8h, 3l7x, 2pjs, 1p1j, 1u6e, 1rkd, 4b0h, 2xn6, 1qtn, 2xty, 1guz, 3ncq, 3n1u, 5k8j, 3goa, 2bti, 5swk, 2xg5, 4p3w, 2d4p, 5ksa, 2rey, 3t9y, 1upi, 2y1e, 3kgz, 4mls, 3tbm, 1p5v, 5jjz, 2i8d, 2a50, 3lx1, 3lf4, 2ixk, 2w7r, 1go3, 2g4o, 4r85, 3e8m, 2wua, 2b8m, 2fom, 6mil, 2g2s, 3ssb, 3hm4, 5e5u, 1r6l, 4xo9, 3uuw, 2z1u, 4dn7, 1wkq, 3d5p, 3p48, 3bwu, 2cc6, 1v70, 4mj0, 4a0t, 2arr, 2c2i, 2air, 4boq, 3cqr, 1tvx, 3ozs, 4jde, 2q5w, 3nkd, 1fur, 1vdk, 4o3c, 4j5u, 3sxy, 3stp, 4ror, 1mki, 3ws7, 4m6u, 2f9i, 4hfg, 3vgl, 4fgs, 5adu, 4l57, 1wqa, 3moy, 1kol, 2x4k, 3vpz, 2qhf, 3t6c, 2p91, 3acd, 3gdm, 4gcm, 4e4u, 4m32, 1fe0, 1on2, 3ht5, 1hyu, 3uzo, 2wqf, 3g0o, 3da8, 2fre, 3oyt, 3emv, 3bwv, 5oxz, 3mti, 3tlo, 4c2e, 6grh, 3sw5, 4uma, 3exq, 3l6b, 3mqd, 2p2d, 1o22, 1hxp, 3lnl, 1j3q, 5wxl, 1vyo, 2bjk, 4c45, 4luk, 1o4t, 4ds1, 2ikk, 1v7p, 3a14, 2qme, 3lt0, 1vhf, 2a8j, 1tqx, 3es4, 3mwb, 3hna, 4v2k, 2egu, 3lio, 4jci, 2nqt, 1vj2, 4ozj, 2bnm, 4m1q, 4o4v, 4tq1, 1u7i, 1oaj, 4dxm, 2woz, 4g9p, 2wn3, 5c2u, 3ce1, 1r29.

Chapter 4: Design strategies for rigid display of proteins on symmetric scaffolds

The following chapter contains a reprint of a research article from

Proceedings of the National Academy of Sciences

Cryo-EM structure determination of small therapeutic protein targets at 3 Å resolution using a rigid imaging scaffold

120, 37 (2023)

DOI: 10.1073/pnas.2305494120

SymDesign align-helices: Automated pipeline for modeling protein fusions in hierarchical symmetric systems

Kyle Meador, Todd O. Yeates

Abstract

Modeling rigid protein fusion has become an important area in the domains of protein engineering as a means to create orientational dependence between two molecules. We present a computational tool to model helical protein fusions through an easy to use, yet customizable interface. We demonstrate how multiple protein conformations and local and global symmetry can be applied to this problem meeting biologically relevant systems with the necessary modeling capabilities. Importantly, new methods for modeling the inherent flexibility of the helix fusion allow us to sample important degrees of freedom to this design space. Together these features take into account all aspects of dynamics and assembly, allowing informed design of new molecules. This tool has already enabled development of new symmetric designs for imaging of small proteins and stands to immediately extend to other applications, including multienzyme materials and vaccine development.

Introduction

In protein engineering applications, it is often desirable to create orientational dependence between two or more molecules to impart novel functions. For instance, protein receptors can be specified to act on ligands in particular ways to create functional outcomes and colocalization can even be applied to create proteinaceous materials. In addition to non-covalent protein-protein interfaces, an established means to stably define the orientation of protein chains is through genetic fusion [1–6]. In applications where a predetermined orientation is necessary, a continuous alpha helix fusion can be a particularly attractive solution. As helices contain a predictable angular rise and pitch, these parameters can be used to vary the orientation of one protein to another, provided they contain helices with compatible directionality to create a single helical fusion between the respective pair.

Although methods of helical fusion have been employed for many years, access to novice protein designers is complicated by computational and logistical hurdles. Modeling two proteins is complicated by considerations such as whether each protein exists in a multicomponent system, local and global symmetry [7], and stitching together a chimeric sequence from the original sequences at the new junction. In addition, computational fluency has been necessary to use protein design tools designed for command line usage [8]. Recent trends to reduce hurdles and improve community engagement include open source code [9,10]and improving the access and usability of programming environments. A recent development has been the creation of sharable, interactive notebooks which pair documentation, visualization, and code simultaneously to improve user experience and engage users with the tools [11].

To extend methodologies of helical fusion to the larger community, we present a tool, provided both as a notebook for online usage (such as Google Colab) and a command-line application, to conformationally model helical fusions in conformationally complex systems. The user can control a range of specific parameters to explicitly search outcomes or simply explore the entire design space by enumerating all the degrees of freedom. Importantly, considerations

for real biological environments such as multicomponent systems and local and global symmetry are inherently handled, reducing barriers to modeling complex hierarchical interactions. In addition, we define new insights into how helices bend, i.e. their inherent spring-like flexibility, and utilize these observations to augment the conformational search. Each of these features is performed in an automated pipeline to make modeling complex biological systems simpler, allowing users to focus on the biological relevance of design choices.

Results

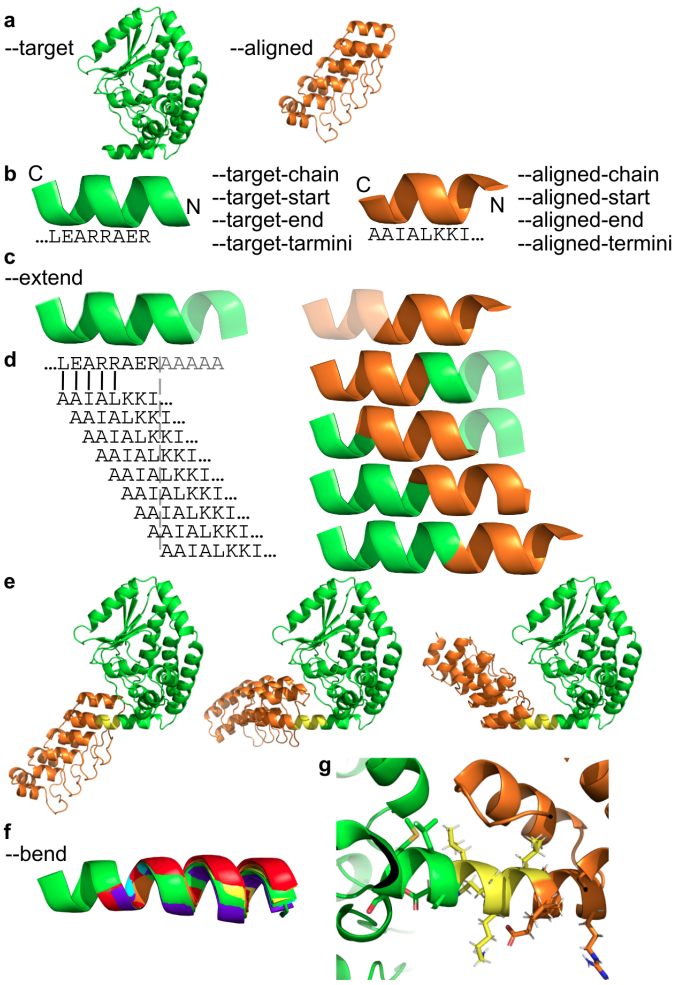
Enumerating the space of helical fusions

The essence of using a helix to connect two separate protein entities lies in four steps. First, two molecular systems and their secondary structure elements are defined and compatible helical regions are selected from each component (Figure 4.1a,b). Second, pairs of helical elements are iteratively overlaid, varying the alignment and thus the helical pitch (Figure 4.1d). Third, candidate overlaps are checked for clashes of the resulting extra-helical atomic positions and if deemed structurally viable are genetically fused, removing extra segments and linking the remaining sequence/structure (Figure 4.1e). Finally, the contiguous structural domain should be modified with a sequence that can support the resulting junction of the once separate entities (Figure 4.1g).

For each of these design steps, the default parameters perform a complete enumeration of the least invasive fusion of any pair of helical N- and C- termini from two separate protein systems (specified with `--target` and `--aligned`), or multiple flags can be specified to parameterize the task on satisfying design objectives. Definition of the helices to be aligned can be achieved by selecting chains of interest (`--target-chain`, `--aligned-chain`), residue numbering specification (`--target-start/ --target-end`, `--aligned-start/ --aligned-end`), and/or termini of interest (`--target-termini`, `--aligned-termini`)

(Figure 4.1b). This parameterization allows the same search procedure to occur at helices located even in the middle of molecules. Additionally, the length to `--extend` the overlapping helix past the identified segments generates longer alignments based on ideal helical parameters allowing the exploration of different fusion geometries (Figure 4.1c). Finally, at each aligned point, the number of times to sample the helical `--bend` from a distribution of helical parameters can identify solutions which deviate within the observed space of helical conformational flexibility (Figure 4.1f). All configurations resulting from the alignment procedure that additionally pass atomic clashing criteria are noted as satisfactory alignments and serve as outputs from fusion. The user can optionally provide specified parameters to determine what atomic violations constitute clashes including `--ignore-clashes`, `--ignore-pose-clashes`, `--ignore-symmetric-clashes`, `--clash-distance`, or `--clash-criteria` to modify how atoms are measured for clashes.

Figure 4.1. Helical alignment



a) Two protein systems are specified as input, one is the target (green) which will remain fixed in space and the other is the aligned molecule (orange) which will be moved to the reference frame of the target. **b)** Methods of identification of alpha helices to be utilized in the alignment procedure. **c)** The identified helices can be extended to a specified number of residues with ideal helical parameters. Extensions are represented by transparent segments. **d)** The helical alignment procedure proceeds by performing a structural superposition at each successive index of the two identified helices. The sequence of the alignment serves as a reference for the structural location of the aligned helix which can be seen to move along the target helix as the alignment is exhaustively scanned. **e)** The various fusions that result from alignment and helical superposition colored by the structure of origin, with the overlap in yellow. The target remains fixed in space while the aligned adopts the conformation specified by the alignment onto the target. Non participating structural features are discarded and converted into a single protein entity with one corresponding sequence. **f)** During alignment the bend of the alignment can be varied. A demonstration of various aligned bends with the ideal helix in green. Red shifted helices bend to greater extent in one direction, while blue shifted helices bend the other. **g)** After fusion, the resulting side chains make new contacts and an amino acid sequence that supports the chimera can be designed.

Symmetry aware molecular modeling

Creating an adequate computational representation of a biological entity requires that structural information be mapped to the genetic sequence which specifies its creation. In scenarios with a one to one correspondence, i.e. a monomer, sequence modifications are directly related to the structure and vice versa (Figure 4.2a). However, higher stoichiometries, as is typified by oligomerization as a result of molecular symmetry, requires mapping spatially distinct, but informationally identical copies to a single sequence (Figure 4.2b). Such a mapping allows propagation of sequence modifications, while spatial manipulations of the oligomeric ensemble are mapped to a single encoding sequence. Additionally, as molecules seldom exist in isolation, any conceivable multicomponent relationship must be addressable to allow accurate manipulation and description of the interactions between two or more entities, accounting for their respective symmetry states (Figure 4.2c). Such symmetric relationships account for an extraordinary amount of the observed protein complexity [12]. As such, explicit accounting of hierarchical symmetry in sequence-structure relationships enables modeling of biologically relevant molecules for engineering applications.

We apply symmetric principles to the task of protein fusion to enable the exploration of spaces with varied stoichiometry and multivalency. With or without symmetry, the number of distinct protein entities in either the target or aligned structure is flexible. Each respective entity from the target and aligned molecular systems will be fused creating one system with all members. If symmetric modeling is required, the target structure can span a wide range of `--symmetry` architectures which are automatically parameterized (see Symmetry input). The aligned structure can not be parameterized with symmetry for such a protocol, as a valid symmetric construction when two or more symmetric systems are combined requires strict constraints [13]. Deviation from these ideal geometries is anticipated to agglomerate biological assemblies into a dendritic solid phase [14]. As such, the target system's symmetry is accounted for when generating fusions and results in the aligned system being situated in the target

reference frame. This results in satisfactory alignments endowing the aligned system with the symmetric properties of the target entity to which fusion occurs.

To demonstrate the range of outcomes possible using symmetric environments, we aligned helices in each of five varying symmetric architectures (fig 3). As the aligned molecule, we selected the DARPin evolved to bind GFP (PDB entry 5ma8) as has been the subject of recent scaffolding works using alpha-helical fusion [15,16] and performed alignment with default parameters to the targets: 1) the biological assembly of the human chromodomain Y-like protein (PDB entry 2gtr), a C3 oligomer (fig 3a), 2) Ferritin (PDB entry 6b8f), a single component octahedron (fig 3b), 3) the designed protein cage, T33-51, (PDB entry 5yc5), a two component tetrahedron architecture (fig 3c), 4) the crystal lattice of endo-alpha-N-acetylgalactosaminidase from *B. longum* (PDB entry 2zxq), which is capable of supporting guest molecules [16] (fig 3d), and 5) a designed crystalline lattice (unpublished model), comprising a two component P432:{C3}{D4} architecture (fig 3e).

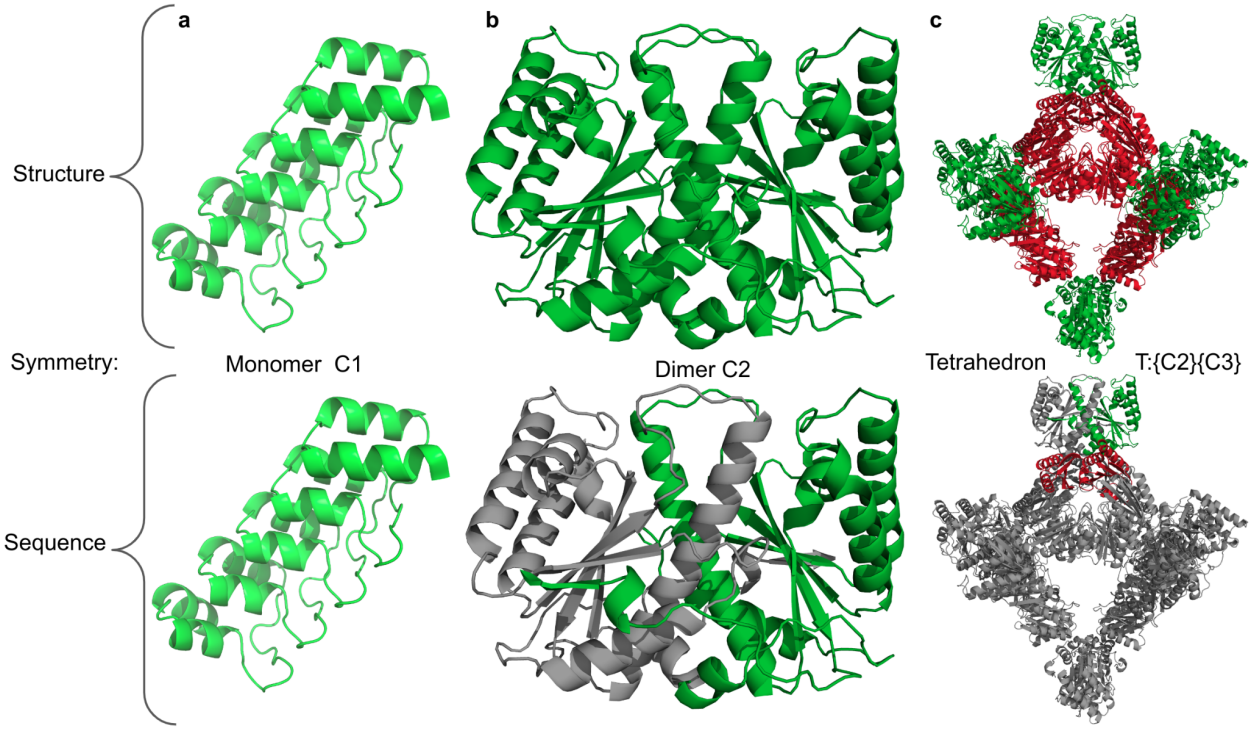


Figure 4.2. Symmetric sequence/structure relationships

a) A monomeric protein has one subunit and thus only one sequence copy. **b)** A dimeric protein has two subunits, therefore two sequences, however only one of which is unique. Modifications to the sequence or structure must be reflected through the sequence structure mapping. **c)** A T:{C2}{C3} Tetrahedron with 12 copies of each of two proteins (C2, green, C3, red). There are a total of 24 subunits and 24 sequences, only 2 of which are unique. The various protein-protein interfaces between each monomer of the C2 and C3 molecules, as well as between the C2 and C3 require accounting for stoichiometry during structural mapping and relation of structural information to their genetic precursors.

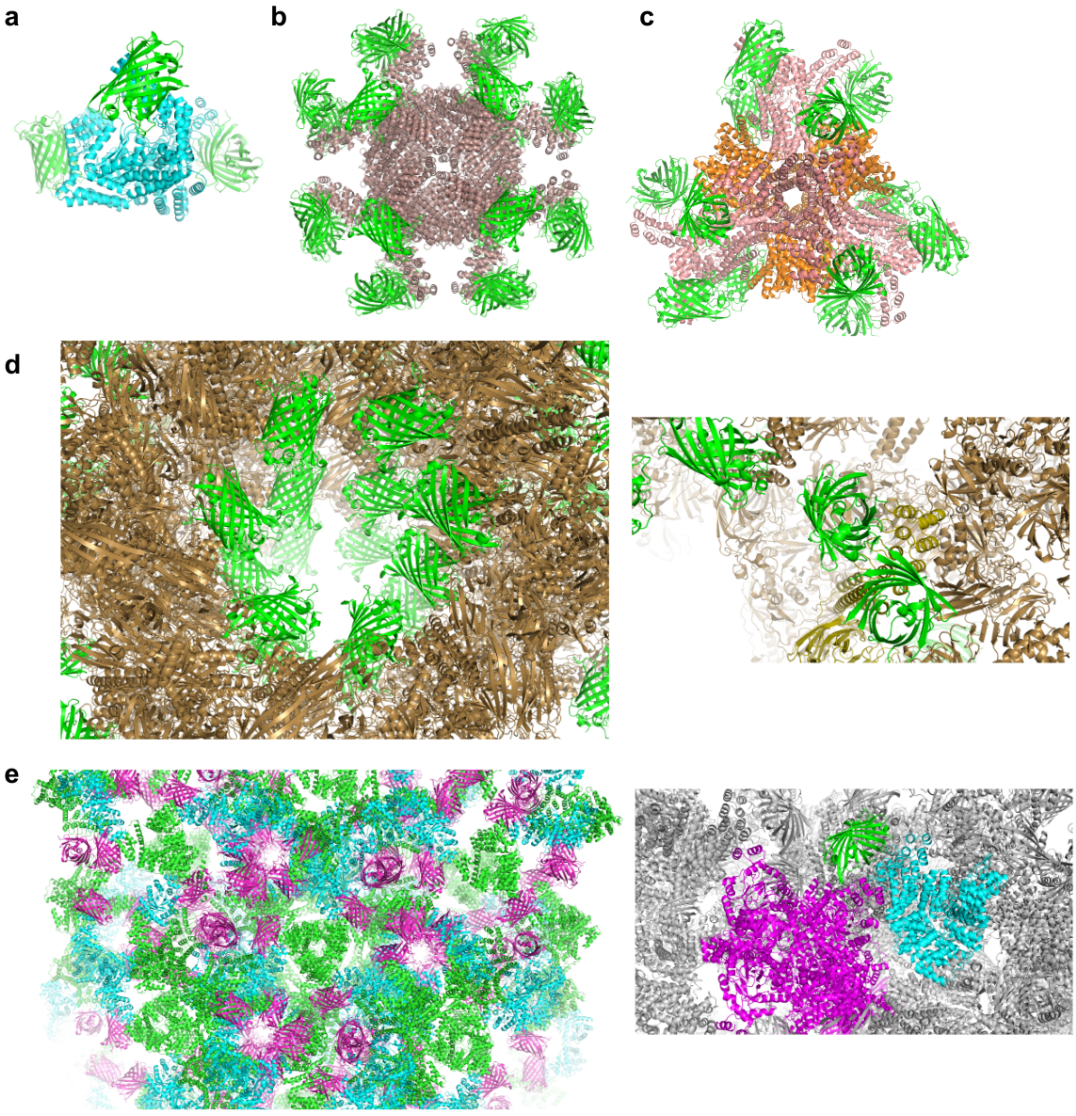


Figure 4.3. Helical fusion of various symmetric states

a) A C3 trimer fusion (cyan) with the second chain of the aligned molecule, in this case GFP, symmetrized. The fusion endows the C1 symmetry of the DARPin/GFP complex to assume the symmetry of the C3 molecule and display 3 copies of GFP. **b)** Octahedral fusion (ruby) with the aligned GFP bound at all 24 positions on the assembly. **c)** Tetrahedral T:{C3}{C3}{C1} fusion complex. The fused entity (pink) is C3 symmetric, where it participates in interactions to make a two component tetrahedral (T) assembly. The second entity in the tetrahedron (orange) is also C3 and assumes a single molecular system with the fused entity. Finally 12 copies of the GFP (C1, green) are bound to the pink fusion and displayed on the surface. **d)** Lattice symmetry $P6_5$ demonstrating the fused molecule (brown) in a crystalline environment and the bound GFP (green) occupying every monomeric binding site in a crystalline pore created along the crystallographic 6-fold screw axis (6 monomers/unit cell). The inset shows one monomer of the DARPin/GFP complex and the site of fusion in the lattice. **e)** Lattice symmetry P432:{C3}{D4}{C1} with two protein entities (C3 with fusion to DARPin, cyan) (D4, magenta) creating the lattice contacts. The GFP (C1 green) is observed at each of the DARPin fusions to the C3, however demonstrates an overall octahedral placement given the lattice symmetry. The

inset shows close up of the C3 fusion, D4, and GFP proteins (each monomer present in 24 copies/unit cell).

Sampling helical bend

By expanding the degrees of freedom of the rigid body alignment search procedure, the space for fusion becomes vastly more useful. Though helical parameters can be articulated [17], we sought a better understanding of the actual modes which deviations of helices are sampled in natural proteins (see Methods). Briefly, utilizing helical segments from high resolution x-ray crystallography structures from the PDB, we extracted seven residue alpha helical segments and for each, performed an alignment to the c-terminal residue of an ideal alpha helix. Next, we extracted the coordinates for the local reference frame of each residue in each aligned helical segment, where each local reference frame consists of three sets of three x, y, z coordinate positions, each situated along a basis vector in the local reference frame. After collecting, flattening, and stacking the coordinates at each residue into a 9-D vector, we stacked each residue in each helical segment into a 9-D metric tensor. Using an eigenvalue and eigenvector analysis of the covariation present in the 9-D tensor, six principle components were identified corresponding to the six main modes of deviation available within alpha helices. We are unaware of previous treatments on this point.

We parameterized these modes of variation in order to sample a rotation matrix from a random gaussian distribution that describes how helices naturally modulate which we refer to as helical bend. The usefulness of such bending becomes apparent when engineering helical fusions. First, they provide a useful source of variation in the outcome of sampling, one that is based on the physical principles of these secondary structures in nature (fig 4a). Second, helical fusions serve as a connection between two larger domains which can move non-cooperatively and cause the helix to become a hinge point. Utilizing information about how this hinge bends can allow secondary contacts to stabilize such a fusion into a more rigid outcome [18] or enable accurate modeling of such connection [19].

To demonstrate how the combination of helical fusion, symmetric modeling, and helical bending, can be applied toward the engineering of larger function molecules, we demonstrate various modeling outcomes as these parameters are manipulated. An important aspect of design of symmetric proteins is that in the context of the full assembly, subunit interactions can exist between different symmetric copies that allow engineering additional contact points. Such contact points can stabilize a fusion junction (fig 4). These can provide further atomic stability to the resulting fusion and require symmetry aware modeling to accurately capture and further incorporate into design.

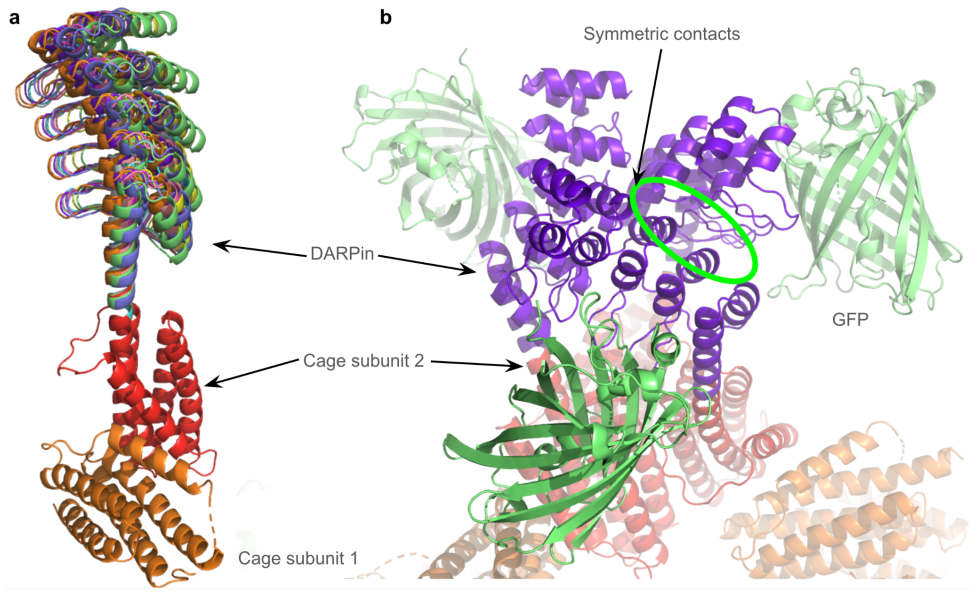


Figure 4.4. Helical bending and symmetric application of bending

a) The two-component asymmetric unit with subunit 1 and 2 acts as the target for helical fusion to a DARPin. Each of the different colored backbones signify one of 10 different bending models possible for the DARPin. The target remains fixed with respect to the global frame so that its orientation is consistent with any symmetry which is propagated to the aligned molecule. **b)** Full two-component tetrahedral assembly with focus on the 3-fold axis of symmetry. Cage subunit 1 (orange) is bound to cage subunit 2 (red), which is helically fused at the N-termini to the DARPin (purple), which is complementary to GFP (green). This particular orientation has additional symmetric contacts between the aligned DARPin as a result of symmetrization and modifications along the bending modes.

Programmatic interface

The tool can be operated from either the command line or in a .ipynb notebook on a personal computer, such as with Jupyter, or through the web using Google Colab. The `align-helices` module from the symdesign molecular modeling program contains all the implementation features available from notebook implementations and is compatible with higher throughput workflows. We provide an implementation as an .ipynb notebook as well as a link to a Google Colab notebook which automatically handles aspects of computational environmental set up and dependencies. An example of the GUI used to specify parameters for the Notebook application is depicted in Figure 4.5 alongside the corresponding command line input. As all

command line flags occupy the same namespace as those exemplified in the notebook GUI, this should make further engagement and programming particularly accessible.

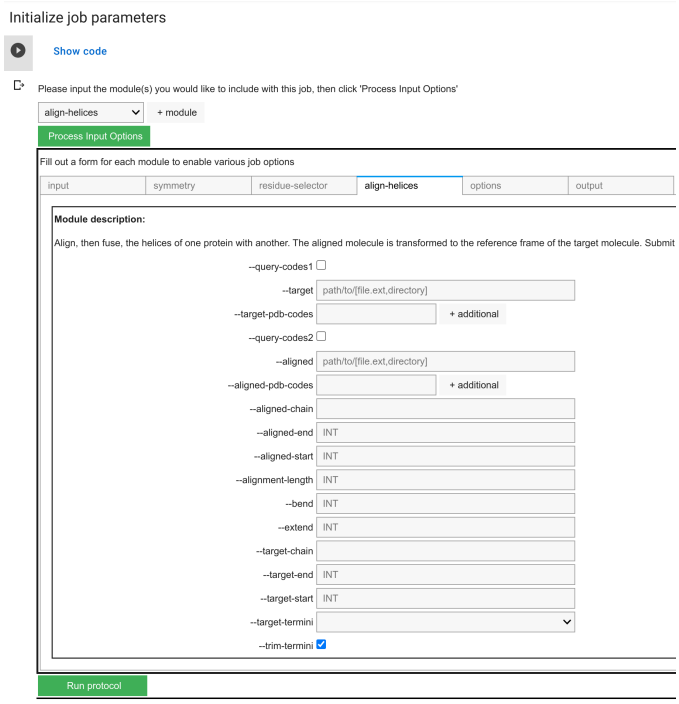


Figure 4.5. Programmatic implementation

a) Typical input tabs present in the Google Colab notebook. Protocol specific flags can be specified from the align-helices protocol. Additional options can be accessed in the tabs such as symmetry, residue-selector, and options.

Sequence design

The last aspect of any fusion protocol is specifying the eventual sequence which should adopt the structural model. In the support of sequence design pursuits, all fusions can be explored further with the `design` module, accessible from the `+module` button from the protocol specification widget. For fusion, the default specification utilizes all residue locations of the entity fusion as well as any neighboring sites as designable positions using ProteinMPNN sequence inference. During `design`, the specified symmetry is respected and utilized towards sequence design calculations. Additionally, various programmatic flags exist to control which parts of a structural model are designable. This is particularly useful if the proteins of interest

have sensitive structural areas that are in proximity to the automatically identified fusion design positions using the default parameters. The flags include `--mask-chain`, `--design-chain`, and `--require-chain`, as well as the residue specific flags `--mask-residues`, `--design-residues`, and `--require-residues`.

Methods

Symmetry input

To properly define binary and higher entity number systems, symmetry must be accounted for locally, such as oligomeric relationships, and globally, such as inter-oligomeric relationships. A roadmap of possible binary relationships in symmetric assemblies and materials was laid out [13] and subsequently defined programmatically [20]. Here, these methodologies are extended to programmatically manipulate and measure such systems. The requirements for processing a symmetric input file are either 1) a symmetric assembly or 2) an asymmetric unit that sits in a specific reference frame. Such a reference frame is defined where the major axis of symmetry is coincident with the Z basis vector, and any minor symmetry axes are placed upon the X basis vector. For example, in an assembly with point group symmetry D_4, the 4-fold axis is coincident on Z, with one of the 2-folds coincident on X. For cubic symmetries such as octahedron (O), each equivalent 4-fold axis is placed upon each of the basis vectors, X, Y, and Z, situating a body diagonal 3-fold axis coincident with a line drawn from (1, 1, 1) to (-1, -1, -1). In the case of a tetrahedron (T) assembly, the 2-fold axis are equivalent to the 4-folds of O. An additional exception to the highest symmetry situated on the Z basis vector is for icosahedral symmetry, where the 2-fold axis is coincident with the Z axis as in T. In this case, the highest symmetry 5-fold is placed at a 31.72° angle from the Z axis on the XZ plane and the 3-fold axis is at a 20.91° angle from the Z axis on the YZ plane.

As every possible symmetric input is specified consistent with international tables of symmetry, the locations of sub-symmetric axes are defined and utilized to construct hierarchical relationships. To establish relationships between global, assembly symmetry, and local, oligomeric symmetry, global symmetry is searched for the presence of structural components at privileged positions which constitute member symmetric rotation axes of the global symmetry group. The presence of monomers in positions which simultaneously satisfy the global symmetry and local symmetry indicate which monomeric entities participate in an oligomeric entity, in which case the oligomeric entity meets the requirements to be a member of the global symmetry group. By only specifying the symmetry and ensuring the system frame of reference is available, we can flexibly and universally search for an such group member relationship regardless of the copy number and symmetric system; whether it be finite with only two monomers or infinitely many monomers from two or more biological entities.

Defining design positions

The starting aligned molecule is queried against all molecules in the resulting fusion reference frame for Cb contacts within the default 9 Å. Symmetrically related aligned molecules and any target molecule copies which approach the aligned asymmetric unit constitute important design positions. Similarly for the target molecule, any non-self interactions as a result of the fusion are included with a default Cb contact distance. These specifications cover the majority of the fusion site, however, all regions occuring between the molecules such as the result of an extension are also included in the design selection.

Description of alignment algorithm

Any input can have alignment performed given an adequate secondary structure of interest can be identified. By default, secondary structures are only identified at chain termini. Each of the termini are allowed to be trimmed back (default five residues) to find at the minimum, a five residue helix (parameterized with `--alignment-length`). Identified helices

are paired and for each overlapping five residue segment, helical alignment and positional clashing is performed. Alignment constitutes performing an RMSD superposition with the Kabsch algorithm and then applying the identified rotation and translation to the aligned molecule to put it in the frame of the target. Clash checking proceeds against the backbone and Cb atoms only, first the minimal asymmetric unit is checked then the asymmetric unit is checked in the context of the symmetric assembly. Any atomic approach of non-covalently bonded atoms less than 2.1 Å are discarded as clashing conformations. Alignment and clash checking proceeds iteratively at each five residue frame along the identified secondary structure element until all pairs of alignments have been exhausted. At this point, the next available termini pairing undergoes the search procedure, which proceeds for each entity pair from the target and aligned molecule. Depending on the number of alignments from each pairing of secondary structures, alignment results in a combinatorial space of $N_{target\text{-}entities}$ x $N_{align\text{-}entities}$ x $N_{alignments/pair}$ x $N_{termini\ avilable}$.

Linear algebra analysis of alpha-helix flexibility for domain fusions

We developed a linear algebra treatment to handle the modeling of alpha helix flexibility. Coordinates were extracted for a set of alpha helices from high quality protein structures in the PDB with X-ray crystallographic resolutions between 1.8 and 2 Å. Alpha helical protein segments were identified using the STRIDE program (ref). Segments of length 7 were analyzed; in order to avoid undesirable deviations at transitions from the ends of helices, we retained segments of length 7 after trimming the terminal residues from segments where 9 consecutive residues were classified as alpha helical. Then, all protein backbone segments were aligned by 3-D superposition of their N-terminal backbone atoms (N, C-alpha, C atoms).

After this alignment, we analyzed the variation of the C-terminal backbone positions across the collection of segments. To regularize the linear algebra analysis, we set up an orthonormal reference frame on each instance of the C-terminal residue. Specifically, the

vectors were: (1) vector from the C-alpha to the C atom, (2) vector in the plane of the C-alpha, C, and N atoms and perpendicular to vector 1, and (3) the cross product between 1 and 2. As a heuristic choice to balance the (otherwise affine or non-metric) consequences of rotational vs translational variations at the terminus, we chose the unit length of this reference frame to be 5 Å; the general behavior of the treatment was found to not depend sharply on that choice.

We then analyzed the collection of vectors of dimension 9 specified by the x, y, z, coordinates of the three positions defined by the orthonormal vectors above. The variation across these 9-D vectors captures the rigid body bending present in the set of backbone segments analyzed. We analyzed the variation in this 9-D space in the usual way, by constructing a metric tensor, ATA, where A is formed from the set of 9-D row vectors. A description of rigid body motion of a body in three dimensions requires 6 dimensions or degrees of freedom, while the metric tensor is 9 dimensional. Because rotational representations are approximately linear for sufficiently small rotations (Goldstein, Mechanics), the matrix A should span (mainly) 6 dimensions, as long as the rotational bending within the collection of segments is not large. We extracted the principal modes of variation using an eigenvector analysis of the metric tensor. As expected, only 6 of the resulting 9 eigenvalues are substantially greater than 0. These represent the 6 modes of flexible bending. The values of the 6 eigenvalues describe the magnitudes of the natural bending along 6 principal modes. Scaled to a sum of 1 in order to describe their relative magnitudes, those eigenvalues were: 0.424, 0.318, 0.161, 0.055, 0.025, 0.016. It is remarkable that the variability is nearly fully captured by just 3 modes of flexing; the first 3 eigenvalues account for more than 90% of the variation. This reflects important natural properties of the alpha-helical conformation. For completeness, and to enable protein engineering calculations where the reference protein domain is fused to an additional domain in the reverse C-to-N sequential order, parallel calculations were also performed with helical segments aligned at their C-terminus, with variation at the N-terminus analyzed; the numerical values were similar, as expected.

With the modes of flexing identified, hypothetical instances of helical bending can be generated readily, in a fashion that conforms to the natural variation observed. For illustration, 6 random Gaussian values (centered on 0 with standard deviation of 1) can be chosen for the magnitudes to be applied along the six principal modes; those random values are first multiplied by the respective eigenvalues (unnormalized) and then by the eigenvectors, and then summed. The resulting 9-D vector produces an instance of the reference frame describing the end of an alpha helix with random bending introduced. From there, fusion of the subsequent domain and its alpha helix is straightforward.

Extending and bending alignments with ideal helical parameters

By default, any identified helix is extended five additional residues to allow five additional frames of overlap (`--extend 10` is used to extend 10 residues). This ideal helix is derived from ideal Watson-Crick helicity parameters. An important point to increase the sampling degrees of freedom comes from sampling from ideal helical bending modes. Using the `--bend` flag, say with an argument `10`, each identified alignment can also have 10 randomly sampled transformations from a distribution of helical transformation modes applied to the aligned molecule. Each additional sample creates a slightly varied twist or curve that varies the positioning of residues downstream from the bend.

Discussion

Herein, we describe the computational procedure used for engineering of helical protein fusions. We enable its exploration through a feature oriented experience to enable the design of fused helical systems of various types and locations. This tool is accessible both as a command line tool and as a distributable notebook. Additionally, the source code is maintained at github.com/kylemeador/symdesign and can be applied to accomplish various protein modeling tasks.

To ensure the tool is up to the task of modeling relevant biological molecules, we extended programmatic modeling of symmetric systems to account for higher order accommodations, asymmetric participants, and the varied relationships that result from association of multiple entities. Importantly all that is required for investigating such complex models is prior knowledge of the local symmetry present for each entity involved as well as the global symmetry of the entire system. We demonstrate that using these symmetric representations, the task of modeling helical fusions can be augmented.

By incorporating parameters that describe helical bend we show that additional confirmations can be sampled which can realize important design outcomes from additional contacts. The application of helical fusion, bending, and symmetry comes together in modeling helical fusions which establish multiple connection points as a result of the sampled fusion, the bend, and the symmetric system. Not only do we specify the generation of new protein backbones, but we incorporate ProteinMPNN sequence design directly after sampling steps [21]. In this way, both helical fusions and extra non-covalent contacts can be designed to favor the sampled state. The conformations, and their designed sequences can further be investigated using structure prediction with AlphaFold [9] as well as symmetric outcomes with AlphaFold multimer [22]. We foresee that dissemination of this tool will have immediate benefits for protein design projects. Continued development and collaboration of these methods should find wide impact in expanding the capabilities of the protein design community.

References

1. Padilla, J. E., Colovos, C. & Yeates, T. O. Nanohedra: Using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proceedings of the National Academy of Sciences* **98**, 2217–2221 (2001).

2. Lai, Y.-T. *et al.* Structure of a designed protein cage that self-assembles into a highly porous cube. *Nature Chemistry* **6**, nchem.2107 (2014).

3. Cannon, K. A., Nguyen, V. N., Morgan, C. & Yeates, T. O. Design and Characterization of an Icosahedral Protein Cage Formed by a Double-Fusion Protein Containing Three Distinct Symmetry Elements. *Acs Synth Biol* (2020) doi:10.1021/acssynbio.9b00392.

4. Brunette, T. J. et al. Modular repeat protein sculpting using rigid helical junctions. *P Natl Acad Sci Usa* 201908768 (2020) doi:10.1073/pnas.1908768117.

5. Vulovic, I. et al. Generation of ordered protein assemblies using rigid three-body fusion. *Proc National Acad Sci* **118**, e2015037118 (2021).

6. Hsia, Y. et al. Design of multi-scale protein complexes by hierarchical building block fusion. *Nature Communications* (2021) doi:10.1038/s41467-021-22276-z.

7. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D. & André, I. Modeling symmetric macromolecular structures in Rosetta3. *Plos One* **6**, e20450 (2011).

8. Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. in *Methods in Enzymology* vol. 487 545–74 (2011).

9. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 1–11 (2021) doi:10.1038/s41586-021-03819-2.

10. Ahdritz, G. et al. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Biorxiv* 2022.11.20.517210 (2022) doi:10.1101/2022.11.20.517210.

11. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

12. Goodsell, D. S. & Olson, A. J. STRUCTURAL SYMMETRY AND PROTEIN FUNCTION. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153 (2000).

13. Laniado, J. & Yeates, T. O. A complete rule set for designing symmetry combination materials from protein molecules. Proc National Acad Sci 202015183 (2020) doi:10.1073/pnas.2015183117.

14. Hernández, N. E. et al. Stimulus-responsive self-assembly of protein-based fractals by

computational design. Nature Chemistry 605–614 (2019) doi:10.1038/s41557-019-0277-y.

15. Liu, Y., Gonen, S., Gonen, T. & Yeates, T. O. Near-atomic cryo-EM imaging of a small protein displayed on a designed scaffolding system. Proceedings of the National Academy of *Science*s (2018) doi:10.1073/pnas.1718825115.

16. Ernst, P., Plückthun, A. & Mittl, P. R. E. Structural analysis of biological targets by host:guest crystal lattice engineering. *Sci Rep-uk* **9**, 15199 (2019).

17. Crick, F. H. C. The Fourier transform of a coiled‐coil. *Acta Crystallogr.* **6**, 685–689 (1953).

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

19. 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* **10**, 1864 (2019).

20. Laniado, J., Meador, K. & Yeates, T. O. A fragment-based protein interface design algorithm for symmetric assemblies. *Protein Eng., Des. Sel.* **34**, gzab008 (2021).

21. Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. Science **378**, 49–56 (2022).

22. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. Biorxiv 2021.10.04.463034 (2021) doi:10.1101/2021.10.04.463034.

Cryo-EM structure determination of small therapeutic protein targets at 3 Å-resolution using a rigid imaging scaffold

The following is a reprint of a research article from

Proceedings of the National Academy of Sciences

120, 37 (2023)

DOI: 10.1073/pnas.2305494120

Reprinted with permission from PNAS.

PNAS RESEARCH ARTICLE | BIOCHEMISTRY
BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

Cryo-EM structure determination of small therapeutic protein targets at 3 Å-resolution using a rigid imaging scaffold

Roger Castells-Graells[a], Kyle Meador[b], Mark A. Arbing[a], Michael R. Sawaya[a], Morgan Gee[b], Duilio Cascio[a], Emma Gleave[c], Judit É. Debreczeni[c], Jason Breed[c], Karoline Leopold[d], Ankoor Patel[d], Dushyant Jahagirdar[d], Bronwyn Lyons[d], Sriram Subramaniam[d,e], Chris Phillips[c], and Todd O. Yeates[a,b,1]

Edited by Edward Egelman, University of Virginia, Charlottesville, VA; received April 5, 2023; accepted July 14, 2023

Cryoelectron microscopy (Cryo-EM) has enabled structural determination of proteins larger than about 50 kDa, including many intractable by any other method, but it has largely failed for smaller proteins. Here, we obtain structures of small proteins by binding them to a rigid molecular scaffold based on a designed protein cage, revealing atomic details at resolutions reaching 2.9 Å. We apply this system to the key cancer signaling protein KRAS (19 kDa in size), obtaining four structures of oncogenic mutational variants by cryo-EM. Importantly, a structure for the key G12C mutant bound to an inhibitor drug (AMG510) reveals significant conformational differences compared to prior data in the crystalline state. The findings highlight the promise of cryo-EM scaffolds for advancing the design of drug molecules against small therapeutic protein targets in cancer and other human diseases.

cryo-EM | small proteins | imaging scaffolds | protein design | cancer drugs

Significance

Cryoelectron microscopy (cryo-EM) is emerging as a major method for elucidating the structures of proteins in atomic detail. A key limitation, however, is that cryo-EM is applicable only to sufficiently large macromolecular complexes. This places a great many important proteins of smaller size, especially those of interest for therapeutic drug development, outside the reach of cryo-EM. We describe a protein engineering effort that overcomes the lower mass limit through the development of a modular imaging scaffold able to rigidly bind and display practically any small protein of interest, greatly increasing its effective mass. We show this technology can be used to visualize molecules, such as a key cancer protein, with important implications for drug design and biomedical research.

Cryoelectron microscopy (cryo-EM) is a rapidly expanding method for determining the atomic structures of large molecular assemblies. It is, however, problematic for determining the structures of small-to-medium-sized protein molecules. A size of about 38 kDa represents a likely theoretical lower limit (1), while about 50 kDa is a practical limit from current work (2). Accordingly, vast numbers of cellular proteins, including many of key therapeutic interest, remain beyond the reach of cryo-EM methods (3).

A potential workaround to the size limitation in cryo-EM is to bind a small protein of interest (the "cargo") to a much larger carrier (the "scaffold") in order to make it large enough to visualize readily. Ideas for scaffolding approaches go back several years (4–6). A key challenge is how to make the binding attachment between the scaffold and the cargo protein sufficiently rigid, as even minor flexibility in the attachment severely compromises the ability to reconstruct a high-resolution image of the bound cargo component. In addition, a general solution to the scaffolding problem calls for modular design, i.e., through the use of a scaffolding component that can be readily diversified to bind any given cargo protein of interest (7–10). Earlier work has explored the use of DARPins as the modular binding domain, genetically fused by way of a continuous alpha helical connection to self-assembling protein cages, to create large symmetric scaffolds for imaging (11–14). Diverse studies have made progress (2, 15–20) (*SI Appendix*, *Supplementary Text*), but further improvements are needed to develop a facile system for high-resolution cryo-EM of small proteins.

In the present study, we demonstrate a protein design advance that substantially rigidifies a cryo-EM scaffold based on fusion of a DARPin as the modular binding domain to a designed protein cage. Analogous to antibodies, sequence variations in the nonconserved loop regions of a DARPin protein can be selected in the laboratory in order to obtain a variant that binds nearly any protein of interest (21). To demonstrate utility in a critically important area of medicine, we have applied this rigidified cryo-EM scaffolding system to study mutant and drug-bound structures of the key oncogenic protein KRAS, which represents a major target for designing anticancer drugs.

Results and Discussion

Rigidification and Testing of an Imaging Scaffold. A previous cage-scaffold design reached a resolution of about 3.8 Å for the attached cargo protein (11, 12), but residual flexibility made it impossible to reach the higher resolution needed for reliable atomic interpretation (generally about 3 Å or better). In the earlier design, the individual DARPin arms—12 in total emanating from the tetrahedrally symmetric cage—protruded separately from each other, thus suffering from residual flexibility. To make further stabilizing contacts possible, we investigated alternative design choices for a scaffold. A different tetrahedral protein cage

Author contributions: R.C.-G., K.M., S.S., C.P., and T.O.Y. designed research; R.C.-G., K.M., M.A.A., M.R.S., M.G., D.C., E.G., J.É.D., J.B., K.L., A.P., D.J., B.L., and S.S. performed research; R.C.-G., K.M., M.A.A., M.R.S., K.L., A.P., D.J., B.L., S.S., C.P., and T.O.Y. analyzed data; and R.C.-G., K.M., and T.O.Y. wrote the paper.

Competing interest statement: S.S. is CEO of Gandeeva Therapeutics. T.O.Y. is CEO of AvimerBio. S.S. holds equity in Gandeeva Therapeutics. T.O.Y. holds equity in AvimerBio. R.C.-G., K.M., and T.O.Y. are inventors on a relevant patent application.

This article is a PNAS Direct Submission.

Copyright © 2023 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution License 4.0 (CC BY).

[1]To whom correspondence may be addressed. Email: yeates@mbi.ucla.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2305494120/-/DCSupplemental.

Published September 5, 2023.

known as T33-51 (22), when modeled with alpha helical linkers to DARPins, oriented the protruding arms to be in near-contact with each other; three DARPins come together at each of the four vertices of the tetrahedron (Fig. 1). Then, computational interface design methods were used to generate new amino acid sequences at the interfaces formed between three symmetry-related copies of the DARPin (*SI Appendix*, Fig. S1 and *Materials and Methods*). The designed interfaces between protruding DARPins were proposed to confer additional stability to these key binding components of the scaffold (Fig. 1). From 12 candidate sequence designs, five were validated by experimental tests to self-assemble into cage-like structures as intended (*Materials and Methods*).

Before employing the candidate cryo-EM scaffolds to image a protein target of major biological importance, we compared their performance in a test system, using the well-studied superfolder version of the green fluorescent protein (GFP) (23), 26 kDa in size, as the cargo protein. When bound to the imaging scaffold, the overall molecular weight of this complex is 972 kDa. As expected, experimental tests showed that all five scaffold candidates bound to GFP when the DARPin (genetically fused to the cage) was one previously established to bind GFP (*SI Appendix*). Initial cryo-EM datasets were collected on the five candidate scaffolds with GFP bound. Based on data processing of similar numbers of particle images from the five candidates, one design

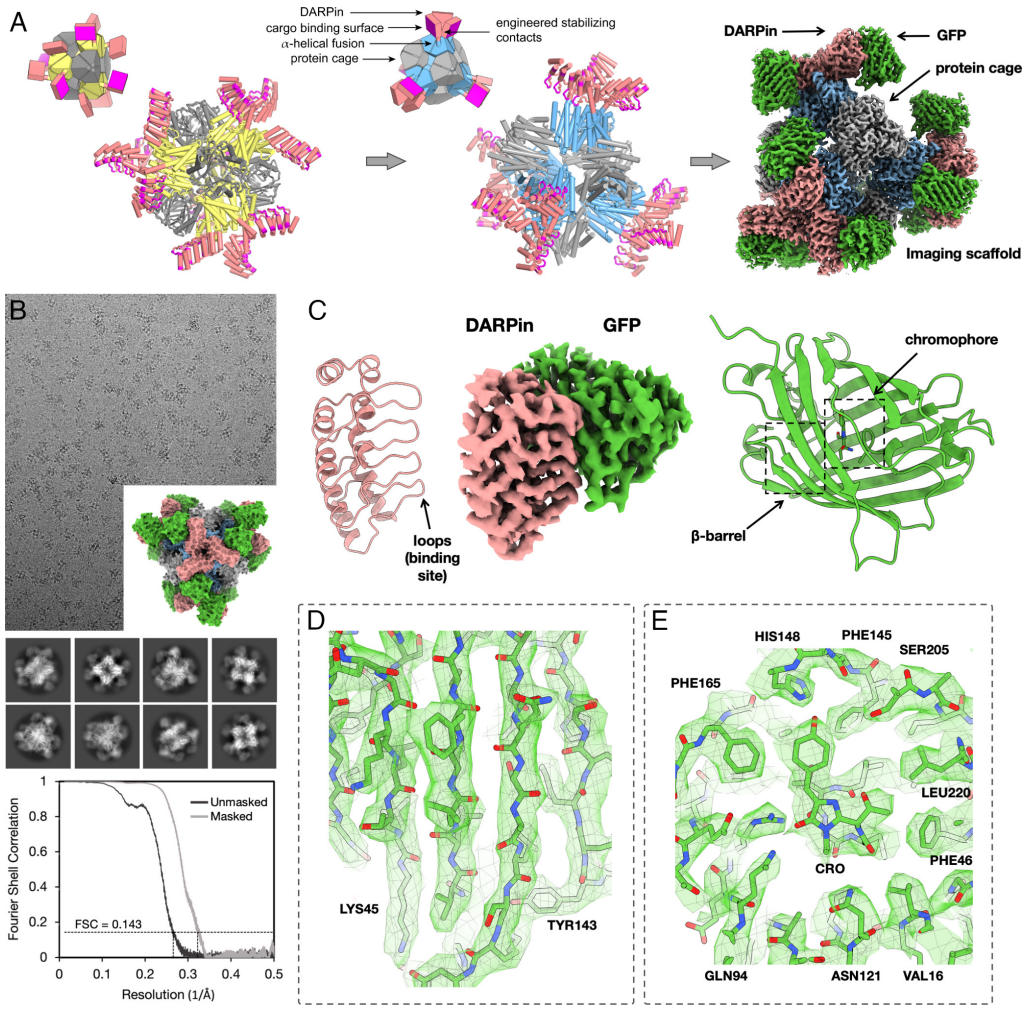


Fig. 1. Rigidified modular cryo-EM imaging scaffolds. (*A*, *Left*) A scheme for a previously described scaffold (11, 12), based on a self-assembling protein cage, displayed protruding DARPin domains as modular binders via continuous alpha helical fusions. The cage subunits bearing the continuous alpha helical fusion are shown in yellow. The other subunit type in this two-component cage is shown in gray. DARPin domains are colored in salmon with their hypervariable binding regions highlighted in magenta. (*A*, *Middle*) A redesigned scaffold based on similar principles, but with protruding DARPin arms disposed to make additional protein–protein contacts with symmetric copies of each other. Designed surface mutations at the newly created interface away from the hypervariable region stabilize the DARPin domain, allowing high-resolution cryo-EM imaging of bound cargo. The *Insets* provide simplified geometric diagrams of the scaffold constructions. (*A*, *Right*) Composite cryo-EM map after focused refinements of GFP bound to a rigidified imaging scaffold. (*B*) Cryo-EM micrograph of the rigidified imaging scaffold bound to GFP (model shown in *Inset*) and 2D classes from selected particles. An FSC plot illustrates agreement between independent half-maps obtained after focused classification and 3D reconstruction, masked around the GFP protein (resolution = 3.1 Å based on a correlation threshold of 0.143). (*C*, *Middle*) A view of the final density map covering the DARPin and its bound GFP protein. Ribbon models of the two components are shown on the sides. (*D* and *E*) Focused views of the density map covering several GFP beta-strands and the GFP chromophore with its surrounding amino acid side chains.

(designated RCG-10; *SI Appendix*) appeared to offer the most rigid presentation of the bound GFP cargo protein. This scaffold was therefore selected for further analysis and cryo-EM data processing. Following data processing from ~877,000 particles obtained from 3,575 cryo-EM movies, a 3-D density map was obtained in which the resolution of the central core of the scaffold was 2.7 Å, with a resolution of 3.1 Å for just the GFP component (Fig. 1 and *SI Appendix*, Figs. S4 and S5). The level of atomic detail is illustrated by the density for the GFP chromophore and side chains from the neighboring amino acid residues (Fig. 1).

In order to assess issues related to coordinate precision and potential perturbances caused by binding to the scaffold, we compared the bound protein structure to crystal structures of GFP in an unbound form. The binding of GFP to the DARPin did not lead to meaningful differences in the backbone, though a different rotamer is seen for a tyrosine residue (Tyr39). The rms deviation for the GFP displayed by the imaging scaffold compared to a crystal structure is 0.59 Å. For data quality and model refinement statistics, see *SI Appendix*, Table S1.

While the significant improvement in resolution of the cargo (compared to the previous, unrigidified scaffold) also reflects various advances in cryo-EM instrumentation and software, analysis of the data shows that the scaffold redesign did lead to a dramatic reduction in the flexibility of the cargo attachment, as anticipated (*SI Appendix*, Fig. S12). The success of the rigidification plan is evident in the pattern of agreement between the atomic model and the cryo-EM density map; the agreement Q-scores decrease steeply with distance from the core-DARPin hinge in the old design but remain nearly uniform in the new design (*SI Appendix*, Fig. S12). Importantly, this supports the hinge as a principal cause of reduced resolution of the cargo in the old design and the reduction in hinge flexibility as a major cause of improvement in the new design.

Additionally, we compared the ability of the deep-learning program ModelAngelo (24) to build de novo atomic models into the cryo-EM density maps. For the earlier 3.8-Å cryo-EM map, the program correctly built only 93 residues (including sidechain atoms) of 156 DARPin residues, a roughly 60% completion for the DARPin. Only 65 of 231 residues could be built for the GFP cargo, corresponding to only 28% completeness. For the new 3.1-Å cryo-EM map, ModelAngelo built all 156 residues of the DARPin domain correctly (100% success), including sidechains. For the GFP cargo, the program built 220 of 231 residues correctly (95% success), including sidechains. The missing residues are in loops (*SI Appendix*, Fig. S13).

Cryo-EM Structures of the Oncogenic KRAS Protein Bound to GDP. For biomedically relevant structural studies, we chose the KRAS protein as a target of high clinical importance. KRAS is a 19-kDa GTPase involved in signal transduction in cell proliferation pathways. KRAS is among the most prevalent human oncogenes, with mutations in KRAS occurring in about 25% of all cancers (25). Some of the most clinically relevant mutations occur at amino acid residues Gly12 and Gly13. Drugs bound to a minor cleft region of the protein near that location are of key pharmaceutical interest, including covalent inhibitors targeting cysteine mutants (i.e., G12C or G13C) (26–29). We therefore undertook a series of structural studies on known KRAS mutants, focusing on the degree of atomic interpretability in 3D density maps obtained using the cryo-EM scaffold described above; a DARPin with loop sequences that bind the GDP-bound form of KRAS was already known from prior work (30, 31), enabling the scaffold to be readily repurposed to image GDP-bound KRAS structures (*Materials and Methods*).

For imaging experiments, we investigated three different sequence variants of KRAS—single site mutants G12V, G12C, and G13C—in their GDP-bound forms. All three KRAS variants were found to bind with good occupancy to our cryo-EM scaffold (presenting the KRAS-specific DARPin). For mutant G13C, ~665,000 particles were obtained from 2,000 cryo-EM movies. Following similar data processing as before, we obtained a 3-D density map showing a resolution of 2.5 Å for the entire particle and 2.9 Å for the KRAS protein (Fig. 2 and *SI Appendix*, Figs. S7 and S8). Among

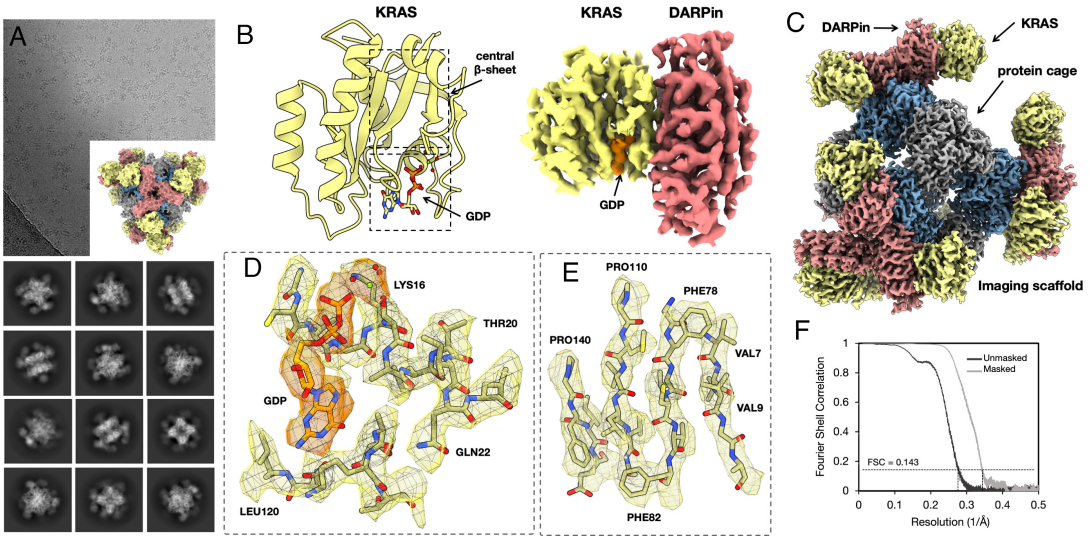


Fig. 2. Cryo-EM structure of KRAS on a rigidified imaging scaffold. (*A*) Cryo-EM micrograph of the rigidified imaging scaffold bound to KRAS (model shown in *Inset*) and 2D classes from the selected particles. (*B*) 3D reconstruction of a density map covering the DARPin and its bound KRAS protein. The GDP ligand is shown in orange. A ribbon model of the KRAS is shown on the left side. (*C*) Composite cryo-EM map after focused refinements of KRAS bound to a rigidified imaging scaffold. (*D* and *E*) Focused views of the density map covering the bound GDP ligand (orange density) and select regions of the KRAS structure. The Mg^{2+} ion is represented by a green sphere. (*F*) An FSC plot illustrates agreement between independent half-maps, obtained after focused classification and 3D reconstruction, masked around the KRAS protein (resolution = 2.9 Å based on a correlation threshold of 0.143).

other metrics of map quality, we assessed the ability of automatic protein model-building software to generate an atomic model for the protein without human intervention. Given the cryo-EM density map and the amino acid sequences for the DARPin and KRAS proteins, ModelAngelo (24) was able to build, de novo, a correct and nearly complete atomic model using default parameters (164 out of 166 residues for KRAS and 150 out of 157 for the DARPin). The amino acid sequence was correctly assigned throughout both KRAS G13C and DARPin molecules. Limited manual fitting was sufficient to join breaks in the chain where the density was weak for mobile loops in the proteins. The success of the modeling exercise shows the utility of the cryo-EM scaffolding approach for an automated structure determination pipeline.

As imaged here by cryo-EM, the KRAS protein matches closely to known structures of KRAS-GDP reported in previous X-ray crystallography studies (30, 31). Our refined structure of the G13C mutant overlaps with a previous X-ray crystal structure with an rms deviation of only 0.5 Å over protein backbone atoms.

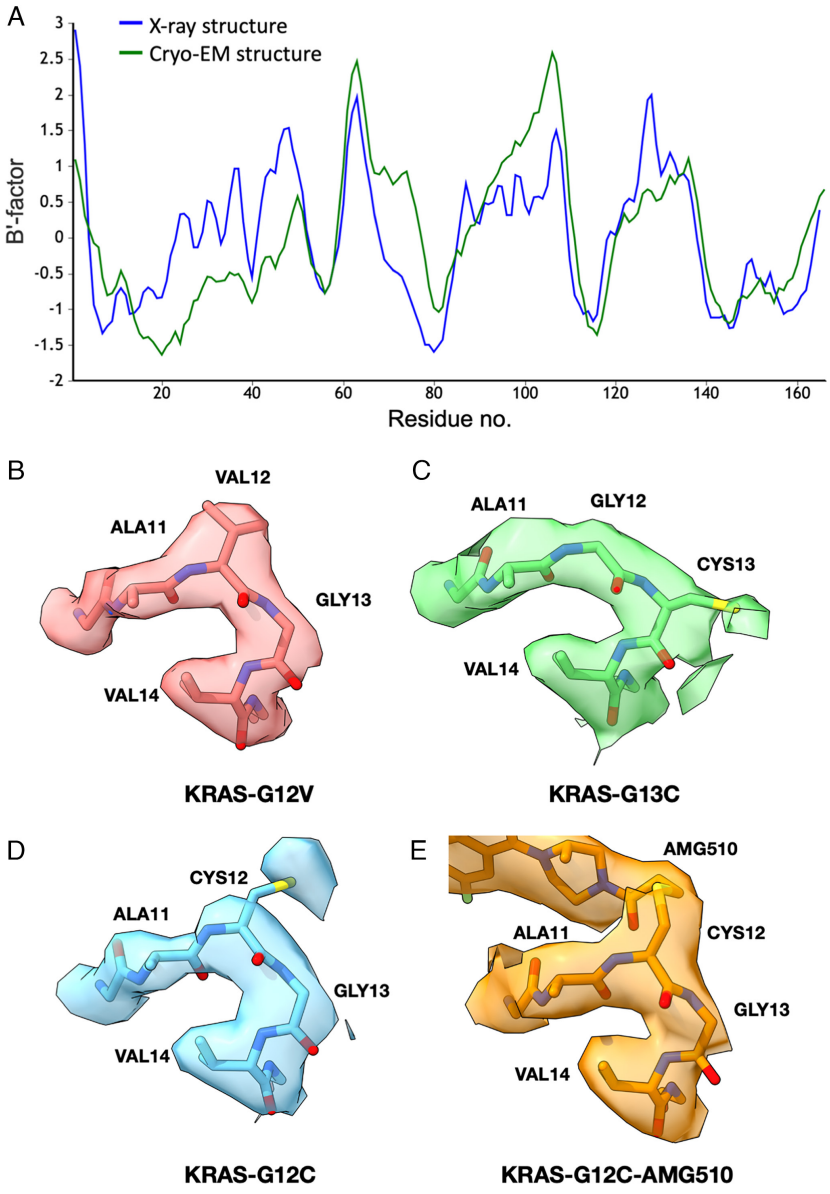


Fig. 3. Structural and dynamical interpretability of cryo-EM maps of KRAS and single-site mutants. (*A*) A plot of refined B-factors—a measure of flexibility or dynamic mobility—for the KRAS structure. Agreement is evident between the X-ray crystal structure (pdb 5o2s) and the cryo-EM structure, which was built and refined de novo (after setting B-factors to a uniform starting value of 20 Å^2). The B-factors are averaged over individual amino acid residues and smoothed over a three-residue window, then normalized for direct comparison using the BANΔIT toolkit (32). The calculated correlation coefficient is 0.65. (*B–E*) Cryo-EM density maps around the single site mutations for KRAS G12V, G13C, G12C, and G12C bound to AMG510. A higher-than-average mobility of Cys12 is also reported by X-ray crystallography (pdb6oim).

The region around the bound GDP cofactor further emphasizes the atomic interpretability (Fig. 2). A Mg^{2+} ion bound near the terminal GDP phosphate group is also clearly visible. An interpretation of protein flexibility and dynamics from the cryo-EM map also agreed well with prior data, as revealed by an analysis of B-factors (or atomic displacement parameters). When examined across the length of the KRAS protein sequence, the correlation coefficient was 0.65 for the atomic structure obtained by cryo-EM compared to an earlier structure reported by X-ray crystallography (Fig. 3*A*). This highlights that the resolution and map quality obtained by cryo-EM are high enough to provide detailed atomic interpretation as well as potentially important information about conformational flexibility.

Structures of additional KRAS mutants provided further opportunities to evaluate atomic interpretability. Following similar protocols as for the G13C mutant, for the G12V mutant, we obtained a final map reconstruction with a resolution of 2.4 Å for the entire particle and 3.1 Å around the KRAS protein (*Materials and Methods*). For the G12C mutant, the resolution was 2.2 Å for the entire particle and 3.0 Å around the KRAS protein (*Materials and Methods*). The maps and refined KRAS structures were all closely comparable, with significant differences in the maps occurring only at the mutated amino acid side chains, as anticipated (Fig. 3). As an assessment of coordinate precision, the rms deviation between the two most closely related cryo-EM structures (the G12V and G12C mutants) was 0.58 Å; this is slightly less than the differences when compared to previously reported X-ray crystal structure, which are between 0.73 and 1.1 Å (*SI Appendix*, Table S2).

Conformational Variations and Drug Binding to KRAS G12C. A minor or "cryptic" cleft in the KRAS protein around residues 12 and 13 has been a site of intense focus for drug design efforts (27–29). Substantial protein conformational changes occur in that region upon drug binding; energetic and structural differences caused by drug binding stabilize the KRAS protein in its inactive form, which binds preferentially to GDP. Understanding the conformational and energetic landscape of the KRAS protein in this binding cleft region is expected to advance the discovery of new cancer drugs. Among drugs targeting clinically important KRAS mutations are a subset that form covalent bonds to cysteine mutants in that site.

As a test of our cryo-EM scaffold for analyzing KRAS drug binding, we determined the structure of the KRAS G12C mutant bound to the covalent inhibitor drug AMG510 [also known as sotorasib; (33)]. Following similar data processing protocols as before, from a set of 69,949 particle images obtained from 2072 cryo-EM movies, we obtained a density map with a resolution of 2.2 Å for the entire particle and 3.2 Å around the KRAS protein bound to AMG510. The map revealed significant conformation changes in the KRAS G12C mutant protein upon binding the AMG510 inhibitor compared to the G12C structure without drug bound. This was anticipated based on prior X-ray crystal structures showing conformational changes in this key region upon drug binding (28, 34–37). Most notable, however, is that the AMG510-bound structure we obtained by cryo-EM differs in the drug-binding region from the structure of the same complex reported earlier by X-ray crystallography protein structure database (PDB 6oim). The nominal resolution in the cryo-EM map is lower than that reported for the X-ray crystal structure (1.65 Å) (33), but the density is sufficiently well resolved to derive a conformation for bound AMG510 that is different from that observed in the crystallographic structure (*SI Appendix*, Fig. S9), especially at the covalent attachment point (residue 12) and the loop residues 60-GQEEYSAM-67 (Fig. 4). The torsion angle at the covalent bond between Cys12 and the drug molecule AMG510 differs by about 100° in the cryo-EM model from the conformation reported in a crystallographic model of the same drug complex (Fig. 4). A movement of ~ 2.7 Å is evident in regions of the drug molecule around the isopropyl pyridyl group, distal from the point of covalent attachment to Cys 12. We assessed the confidence in our modeling of the AMG510 drug molecule in a test in which we refined atomic models separately into density maps produced using two independent half-datasets. For the drug molecule, the differences between the independent models were only 0.1 to 0.3 Å. This is considerably smaller than the coordinate differences observed in comparison to the reported X-ray structure, which exceeded 2 Å, supporting the conclusion that meaningful differences are being revealed between the reported X-ray and cryo-EM conformations for drug binding (*SI Appendix*, Fig. S14).

Motivated by differences observed in the drug-binding pocket of the KRAS G12C mutant, we surveyed the PDB for examples of KRAS G12C bound to other inhibitors or drug molecules. An analysis of a set of 12 such structures (pdb 7a47, 6pgp, 6pgo, 8dnj, 8dnk, 8dni, 7a1y, 5v9o, 5v9l, 4lv6, 4luc, and 4lyh), all elucidated by X-ray crystallography, highlights a substantial degree of conformational variability for the KRAS protein in the binding region. Some of this variation is clearly the result of differences in the chemical structures of the various bound drugs. But there are unexpected patterns. Interestingly, whereas the cryo-EM structure reported here for the AMG510 drug

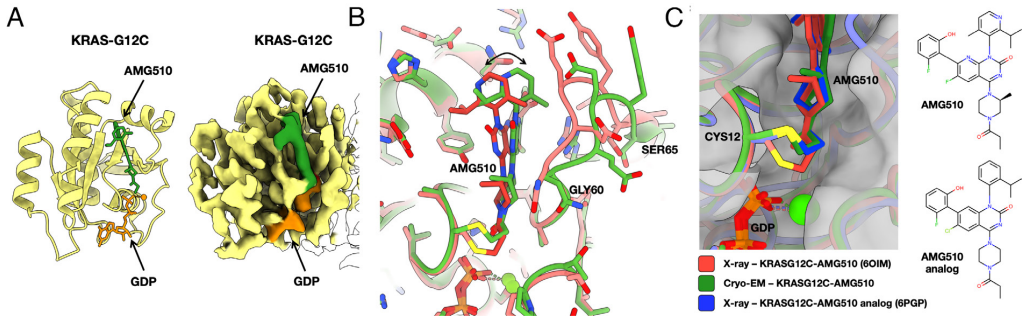


Fig. 4. Cryo-EM structure of KRAS G12C bound to AMG510. (*A*) A refined atomic model (*A*, *Left*) and a cryo-EM density map (*A*, *Right*) covering the KRAS protein, with the AMG510 drug molecule bound. The GDP ligand is shown in orange, and the AMG510 drug is in green. (*B*) Comparison between the cryo-EM structure and a prior X-ray crystal structure of KRAS G12C bound to AMG510. (*C*) Conformational variation at the covalent bond between Cys12 and the AMG510 and an AMG510 analog in X-ray and cryo-EM structures. At the thioether attachment, the cryo-EM model resembles an X-ray crystal structure of a complex with an AMG510 analog.

complex differs from a prior X-ray crystal structure of the identical complex (as discussed above), it matches more closely to an alternative X-ray crystal structure of a complex with a slightly different AMG510 analog (Fig. 4*C*). In particular, we note that the covalent attachment geometry for AMG 510 derived by cryo-EM occurs as well in the context of different drug bound complexes of KRAS G12C.

The findings on AMG510 binding suggest a substantial range of apparently low-energy conformations for the drug molecules and surrounding segments of the protein. The particular conformation observed appears to be affected at least in part by other molecular interactions. In the X-ray crystal structure, the drug-binding region (residues 62 to 73) is at a crystal packing interface (*SI Appendix*, Fig. S15*A*); conformational changes imposed by crystallographic molecular packing have long been studied and proven useful in uncovering conformational states involved in molecular function such as catalysis (38). Likewise, it is notable that in the cryo-EM structure, residue Met 67 is in contact with one of the DARPin domains protruding from the scaffold (*SI Appendix*, Fig. S15*B*). The observed variation across structures provides potentially useful insight into the conformational landscape for drug binding.

Conclusions

These initial structural findings serve as a starting point for deeper explorations of KRAS, and other small therapeutic protein targets, by cryo-EM scaffolding methods. Two immediate messages emerge. The first concerns feasibility. The rigidified scaffold described here provides a number of advantageous properties for cryo-EM structure determination—size, symmetry, and modular binding—making it suitable for future applications to many important systems. Second, the observation of conformational variability in drug binding emphasizes that cryo-EM approaches are likely to offer alternative structural views and distinct atomic frameworks for drug design efforts across broad areas of medicine.

Materials and Methods

Conformational Sampling of Rigidified Scaffolds. The N-terminal helix of DARP14-3G124Mut5 (12) was spatially aligned to the C-terminal helix of each subunit from the T33-51 cage (22). Using local programs, superpositions were performed between the first five helical residues of the DARPin to five residue windows from the terminal helical region of the protein cage, with different choices for the alignment segment from the protein cage. Following superposition, each conformation was evaluated for detrimental, overlapping collisions, and potentially favorable contacts in the fully assembled symmetric environment using local programs as well as visual inspection. Promising conformations–those where multiple protruding DARPin arms came into close proximity–were subjected to further conformational exploration by allowing for minor helix flexing. Modeling of allowable deviations from ideal alpha helix geometry was based on natural deviations observed in a large set of alpha helices extracted from high-resolution crystal structures.

Interface Design Calculation. All calculations were performed in the context of tetrahedral symmetry. For each sampled alignment and helical bend conformation, the resulting pose was relaxed into the REF2015 score function (39) using the FastRelax mover (40). Then, residues in the aligned helical fusion as well as any residues located in cage subunits or other DARPins (excluding variable loop regions) within 8 Å of the aligned DARPin were marked as designable. Further, all residues within 8 Å of designable residues were designated as packable. Sequence design trajectories were performed with a coordinate constraint applied to backbone atoms using Rosetta FastDesign with the InterfaceDesign2019 protocol (41) and REF2015 score function. We collected interface design metrics to quantify the resulting design success as compared to native interfaces (42). After analysis of the global design pool, we removed entire poses from consideration where the average design trajectory had a measured shape complementarity below 0.6, leaving eight viable poses for sampling sequence variations. Next, we ranked the design trajectories from each passing pose by applying a linear weighting scheme to the normalized metrics from each pose. These consisted of favoring fewer buried unsatisfied hydrogen bonds, lower interface energy (between complexed and unbound forms), higher interface shape complementarity, and lower interface solvation energy. Each normalized metric was equally weighted and summed to rank each trajectory. Finally, by examining the sequence diversity of the top candidates from each pose, we removed redundant sequence mutation patterns and selected 12 individual designs for characterization.

Protein Production. The sequences of the imaging scaffolds used in this paper are listed below. DNA fragments carrying the designed imaging scaffold sequences were synthesized (Integrated DNA Technologies and Twist Bioscience) and separately cloned into the vectors pET-22b (subunitB-DARPin) or pSAM (subunitA) (gifted from Jumi Shin, Addgene plasmid #45174; http://n2t.net/addgene:45174; RRID:Addgene_45174). The superfolder GFP V206A (sfGFP V206A) vector was previously described (12). DNA manipulations were carried out in *Escherichia coli* XL2 cells (Agilent). The proteins were expressed in *E. coli* BL21(DE3) cells (New England Biolabs) in Terrific Broth at 18 °C overnight with 0.5 mM IPTG induction at an OD_{600} of 1.0.

Upon collection of the cells, pellets were resuspended in buffer (50 mM Tris, 300 mM NaCl, 20 mM imidazole, pH 8.0) supplemented with benzonase nuclease, 1 mM PMSF, EDTA-free protease inhibitor cocktail (Thermo Scientific) and 0.1% LDAO and lysed using an EmulsiFlex C3 homogenizer (Avestin). The cell lysate was cleared by centrifugation at 20,000 × g for 20 min at 4 °C; the resulting supernatant was recovered and centrifuged at 10,000 × g for 10 min at 4 °C and then loaded onto a HisTrap column (GE Healthcare) pre-equilibrated with the same resuspension buffer. The imaging scaffold was eluted with a linear gradient to 300 mM imidazole. Upon elution, 5 mM EDTA and 5 mM BME were added immediately for designs 5, 8, 10, 13, and 14. The eluted proteins were concentrated using Amicon Ultra-15 100-kDa molecular weight cutoff for the imaging scaffold and 3-kDa molecular weight cutoff for the GFP protein. The concentrated proteins were further purified by size exclusion chromatography using a Superose six Increase column, eluted with 20 mM Tris pH 8.0, 100 mM NaCl, 5 mM BME, 5 mM EDTA for designs 5, 8, 10, 13, and 14 and 20 mM Tris, pH 8.0, and 100 mM NaCl for design 33. Chromatography fractions were analyzed by SDS–PAGE and negative stain EM for the presence of the imaging scaffold. KRAS G12V and KRAS G13C proteins were prepared as previously described by Kettle et al. (43).

The DNA sequence encoding wild-type KRAS (1 to 169) was synthesized (Genscript) and cloned into a pET28 vector with an N-terminal 6xHis tag followed by a TEV site. The G12C mutation was introduced using site-directed mutagenesis and confirmed by sequencing. Protein was expressed in BL21(DE3) cells in LB at 16 °C overnight, following induction at OD_{600} of 0.7 with 0.5 mM IPTG. After harvesting, cell pellets were resuspended in purification buffer (20 mM HEPES, pH 7.4, 300 mM NaCl, 0.5 mM TCEP, and 5 mM $MgCl_2$) supplemented with 1x EDTA-free protease inhibitor cocktail and 400 units benzonase and lysed by sonication. Cleared lysate was loaded onto a 1-mL HisTrap column (Cytiva), washed with 20 CV purification buffer +25 mM Imidazole, and eluted using an imidazole gradient to 500 mM Imidazole. Peak fractions were pooled, concentrated, and loaded onto a Superdex 75 Increase size-exclusion column in SEC buffer (purification buffer excluding $MgCl_2$). For AMG510-bound protein, KRAS G12C was incubated with AMG510 at a 2:1 molar ratio for 30 min and subjected to size-exclusion chromatography (Superose 6 Increase). Peak fractions yielded a mixture of AMG510-bound and free KRAS G12C (see *SI Appendix*, Fig. S10, first lane).

Either KRAS G12C or KRAS G12C-AMG510 was mixed with the imaging scaffold at a 2:1 molar ratio, incubated on ice for 5 min, and complex formation was confirmed through size-exclusion chromatography (Superose 6 Increase).

Negative Stain EM. The concentration of a 3.5-μL sample of fresh Superose six Increase eluent was adjusted to ~100 μg/mL, applied to glow-discharged Formvar/Carbon 400 mesh Cu grids (Ted Pella Inc) for 1 min and blotted to remove excess liquid. After a wash with filtered MilliQ water, the grid was stained with 2% uranyl acetate for 1 min. Images were taken on a Tecnai T12, a T20, a TF20, and a Talos F200C.

Cryo-EM Data Collection. Concentrated imaging scaffolds (1 to 10 mg/mL) were mixed with the GFP cargo or KRAS G13C/KRAS G12V/ KRAS G12C/KRAS

G12C–AMG510 to a molar ratio of 1:2 and diluted to a final concentration of 0.5 to 0.7 mg/mL. The final buffer composition was 20 mM Tris, pH 8.0, and 100 mM NaCl.

Quantifoil 300 mesh R2/2 copper grids were glow discharged for 30 s at 15 mA using a PELCO easiGLow (Ted Pella). A 1.8- to 3.5-μL volume of sample was applied to the grid at a temperature of 10 or 18 °C at ~100% relative humidity, followed by blotting and vitrification into liquid ethane using a Vitrobot Mark IV Thermo Fisher Scientific. Cryo-EM data were collected on an FEI Titan Krios cryoelectron microscope equipped with a Gatan K3 Summit direct electron detector and on a Titan Krios G4 cryoelectron microscope (Thermo Fisher Scientific) equipped with a Falcon4 direct electron detector in electron event registration mode. With the Gatan K3 Summit detector, movies were recorded with Leginon (44) and SerialEM (45) at a nominal magnification of 81,000× (calibrated pixel size of 1.1 Å per pixel) for designs 5, 8, 10, 13, 14, 33 (G13C) datasets and at a nominal magnification of 105,000× (calibrated pixel size of 0.856 Å per pixel) for design 33 (G12V) dataset, over a defocus range of −1.0 to −2.2 μm. With the Falcon4 detector, movies were recorded with the EPU automated acquisition software at a nominal magnification of 155,000× (calibrated pixel size of 0.5 Å per pixel), for design 33 (G12C and G12C-AMG510) datasets, over a target defocus range of −1.00 μm to −2.25 μm with increment steps of 0.25 μm and a total dose of 40 $e^{-}/Å^{2}$.

Fourier shell correlation (FSC) calculations are summarized in *SI Appendix*, Fig. S11. Plots showing dependence of resolution on the number of particles are shown in *SI Appendix*, Fig. S16.

Cryo-EM Data Processing and Model Building. Motion correction, CTF estimation, particle picking, 2D classification, and further data processing were performed with cryoSPARC v.3.2 (46). An initial set of particles was automatically picked using a blob-picker protocol. The extracted particles were 2D classified after which an ab initio reconstruction was generated. This reconstruction was then used for the 3D refinements enforcing T symmetry. The 3D structure was used to generate 2D projections of the particles and then used to repick the particles from the images using a template picker. The picked particles were extracted from the micrographs and went through 3D refinements enforcing T symmetry. The symmetry was then expanded, followed by further focused 3D classification without alignments and focused refinements using a mask encompassing the density for one DARPin and one cargo protein, GFP or KRAS, respectively. The best-resolved classes from the focused 3D classification were focused refined (C1 symmetry) performing local angular searches with the fulcrum at the center of mass of the mask. For the GFP imaging scaffold, we obtained an overall resolution of 2.7 Å for the entire particle and a resolution of 3.1 Å over the GFP protein, based on an FSC threshold of 0.143. For the KRAS G13C imaging scaffold, we obtained an overall resolution of 2.5 Å for the entire particle, and the resolution over the KRAS protein was 2.9 Å. We performed automatic de novo atomic model building into our KRAS G13C cryo-EM density using the program ModelAngelo (24) in the $COSMIC^{2}$ platform (47). The structure of GFP was built de novo using the automated chain tracing program, Buccaneer (48). The other three structures reported here were built starting from atomic models of close homologs, as noted in *SI Appendix*, Table S1. Manual adjustments to the models were performed using Coot (49), and automated refinement was performed using Phenix (50). Figures were prepared using ChimeraX (51, 52) and PyMOL (Version 2.0 Schrödinger, LLC)

Refinement into Half-Maps. We used refinement against independent half-maps (reconstructed from independent half-datasets) as an assessment of coordinate precision for the bound AMG510 drug molecule. Prior to independent real-space refinement, the molecules were subjected to computational simulated annealing—heating to 1,000 K and slow cooling to 300 K—in the program Phenix.

FSC Calculation. FSC plots were generated using the *mtriage* tool of Phenix (53). Each refined model and final map were submitted to *mtriage* along with two half-maps. Masked curves correspond to the use of a smoothed mask to perform FSC calculation only around the model (54).

Retrospective Test of Scaffold Structure Predictability by AI Methods. Given the important interplay between protein sequence design and protein structure prediction, we considered whether a leading machine learning algorithm, AlphaFold2 (55), would correctly predict the structure of our designed scaffold based on amino acid sequence. Such a success would argue that an unguided algorithm might have reached the same (or a similar) design result. A key element of the present scaffold design is the association of a homomeric protein trimer–based on a protein chain comprising a cage subunit fused to a DARPin—in such a fashion that stabilizing interactions occur between three copies of the DARPin; the trimer is mainly held together by association of the cage subunit component. When applied to our designed protein sequence, and specifying three chains to be associated, the AlphaFold2 program did not faithfully recapitulate the key stabilizing features between DARPins that were critical in rigidifying the scaffold to enable high-resolution imaging, and which were validated by cryo-EM. For example, residue ARG 254 was engineered to make a stabilizing interaction with residue ASP 181 from an adjacent DARPin. In our cryo-EM structure, those two residues come into atomic contact, as intended. In contrast, prediction by AlphaFold2 leaves those two residues ~15 Å apart, which is well beyond interaction distance. We furthermore attempted to use AlphaFold2 to computationally assemble the entire 24 subunit ($a_{12}b_{12}$) scaffold architecture given just the amino acid sequence information. That computational exercise did not assemble the cage subunits into a correct tetrahedral assembly. These results emphasize the importance in the present work of expert human input in the overall design strategy.

Data, Materials, and Software Availability. The structures of the imaging scaffolds and the protein targets, 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-29700 (56), EMD-29713 (57), EMD-29715 (58), EMD-29718 (59), EMD-29719 (60), and EMD-29720 (61) and PDB accession codes 8G3K (62), 8G42 (63), 8G47 (64), 8G4E (65), 8G4F (66), and 8G4H (67), respectively. The sequences of the protein designs are included in *SI Appendix*.

ACKNOWLEDGMENTS. We thank David Strugatsky and Peng Ge for assistance in cryo-EM data collection acquired at the Electron Imaging Center for Nanomachines at the University of California, Los Angeles California for NanoSystems Institute, and Alison Berezuk for assistance with cryo-EM data collection carried out at University of British Columbia, Vancouver. We also thank Kevin Cannon, Ivo Atanasov, and Wong Hoi Hui for training in cryo-EM. We thank Yi Xiao Jiang, Tom Dendooven, Jack Bravo, and Yuval Mazor for helpful discussions about cryo-EM data processing, and Tom Ceska, Matt Lucas, and Lee Freiburger for helpful KRAS-related discussions. We thank Chris Garcia and Nathanael Caveney for discussions regarding cryo-EM scaffolding, and Alex Lisker for computing support. This work was supported by NIH grant R01GM129854 (T.O.Y.). Additional resources for sample preparation and electron microscopy screening were supported by DOE grant DE-FC02-02ER63421.

Author affiliations: [a]Department of Energy, Institute for Genomics and Proteomics, University of California, Los Angeles, CA 90095; [b]Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095; [c]Discovery Sciences, R&D, AstraZeneca, Cambridge CB2 0AA, United Kingdom; [d]Gandeeva Therapeutics, Inc., Burnaby, British Columbia V5C 6N5, Canada; and [e]Department of Biochemistry and Molecular Biology, The University of British Columbia, Vancouver, BC V6T 1Z3, Canada

1. R. Henderson, The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28**, 171–193 (1995).
2. M. A. Herzik, M. Wu, G. C. Lander, High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM. *Nat. Commun.* **10**, 1032 (2019).
3. T. O. Yeates, M. P. Agdanowski, Y. Liu, Development of imaging scaffolds for cryo-electron microscopy. *Curr. Opin. Struct. Biol.* **60**, 142–149 (2020).
4. F. Coscia *et al.*, Fusion to a homo-oligomeric scaffold allows cryo-EM analysis of a small protein. *Sci. Rep.* **6**, 30909 (2016).
5. P. A. Kratz, B. Böttcher, M. Nassal, Native display of complete foreign protein domains on the surface of hepatitis B virus capsids. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 1915–1920 (1999).
6. T. G. Martin *et al.*, Design of a molecular support for cryo-EM structure determination. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E7456–E7463 (2016).
7. C. McMahon *et al.*, Yeast surface display platform for rapid discovery of conformationally selective nanobodies. *Nat. Struct. Mol. Biol.* **25**, 289–296 (2018).
8. M. S. Morrison, T. Wang, A. Raguram, C. Hemez, D. R. Liu, Disulfide-compatible phage-assisted continuous evolution in the periplasmic space. *Nat. Commun.* **12**, 5959 (2021).
9. H. K. Binz *et al.*, High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat. Biotechnol.* **22**, 575–582 (2004).
10. S. Rothenberger *et al.*, The trispecific DARPin ensovibep inhibits diverse SARS-CoV-2 variants. *Nat. Biotechnol.* **40**, 1845–1854 (2022).

11. 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).
12. 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).
13. Q. Yao, S. J. Weaver, J.-Y. Mock, G. J. Jensen, Fusion of DARPin to aldolase enables visualization of small protein by Cryo-EM. *Structure* **27**, 1148–1155.e3 (2019).
14. I. Vulovic *et al.*, Generation of ordered protein assemblies using rigid three-body fusion. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015037118 (2021).
15. T. Uchański *et al.*, Megabodies expand the nanobody toolkit for protein structure determination by single-particle cryo-EM. *Nat. Methods* **18**, 60–68 (2021).
16. R. J. Cater *et al.*, Structural basis of omega-3 fatty acid transport across the blood–brain barrier. *Nature* **595**, 315–319 (2021).
17. X. Fan *et al.*, Single particle cryo-EM reconstruction of 52 kDa streptavidin at 3.2 angstrom resolution. *Nat. Commun.* **10**, 2386 (2019).
18. J. S. Bloch *et al.*, Development of a universal nanobody-binding Fab module for fiducial-assisted cryo-EM studies of membrane proteins. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2115435118 (2021).
19. X. Wu, T. A. Rapoport, Cryo-EM structure determination of small proteins by nanobody-binding scaffolds (Legobodies). *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2115001118 (2021).
20. K. Zhang *et al.*, Cryo-EM, protein engineering, and simulation enable the development of peptide therapeutics against acute myeloid leukemia. *ACS Cent. Sci.* **8**, 214–222 (2022).
21. Y. L. Boersma, A. Plückthun, DARPins and other repeat protein scaffolds: Advances in engineering and applications. *Curr. Opin. Biotechnol.* **22**, 849–857 (2011).
22. K. A. Cannon *et al.*, Design and structure of two new protein cages illustrate successes and ongoing challenges in protein engineering. *Protein Sci.* **29**, 919–929 (2020).
23. J.-D. Pédelacq, S. Cabantous, T. Tran, T. C. Terwilliger, G. S. Waldo, Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* **24**, 79–88 (2006).
24. K. Jamali, D. Kimanius, S. H. W. Scheres, A graph neural network approach to automated model building in cryo-EM maps. arXiv [Preprint] (2022). https://doi.org/10.48550/arXiv.2210.00006 (Accessed 7 February 2023).
25. S. Li, A. Balmain, C. M. Counter, A model for RAS mutation patterns in cancers: Finding the sweet spot. *Nat. Rev. Cancer* **18**, 767–777 (2018).
26. L. Huang, Z. Guo, F. Wang, L. Fu, KRAS mutation: From undruggable to druggable in cancer. *Signal Transduct. Target. Ther.* **6**, 1–20 (2021).
27. A. Mullard, Cracking KRAS. *Nat. Rev. Drug Discov.* **18**, 887–891 (2019).
28. J. M. Ostrem, U. Peters, M. L. Sos, J. A. Wells, K. M. Shokat, K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature* **503**, 548–551 (2013).
29. J. M. L. Ostrem, K. M. Shokat, Targeting KRAS G12C with covalent inhibitors. *Annu. Rev. Cancer Biol.* **6**, 49–64 (2022).
30. S. Guillard *et al.*, Structural and functional characterization of a DARPin which inhibits Ras nucleotide exchange. *Nat. Commun.* **8**, 16111 (2017).
31. N. Bery *et al.*, KRAS-specific inhibition using a DARPin binding to a site in the allosteric lobe. *Nat. Commun.* **10**, 2607 (2019).
32. F. Barthels, T. Schirmeister, C. Kersten, BANΔIT: B′-Factor analysis for drug design and structural biology. *Mol. Inform.* **40**, 2000144 (2021).
33. J. Canon *et al.*, The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity. *Nature* **575**, 217–223 (2019).
34. M. Mathieu *et al.*, KRAS G12C fragment screening renders new binding pockets. *Small GTPases* **13**, 225–238 (2022).
35. B. A. Lanman *et al.*, Discovery of a covalent inhibitor of KRASG12C (AMG 510) for the treatment of solid tumors. *J. Med. Chem.* **63**, 52–65 (2020).
36. M. Zeng *et al.*, Potent and selective covalent quinazoline inhibitors of KRAS G12C. *Cell Chem. Biol.* **24**, 1005–1016.e3 (2017).
37. K. Zhu *et al.*, Modeling receptor flexibility in the structure-based design of KRASG12C inhibitors. *J. Comput. Aided Mol. Des.* **36**, 591–604 (2022).
38. M. Sawaya, J. Kraut, Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: Crystallographic evidence. *Biochemistry* **36**, 586–603 (1997).
39. R. F. Alford *et al.*, The rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
40. L. G. Nivón, R. Moretti, D. Baker, A pareto-optimal refinement method for protein design scaffolds. *PLoS One* **8**, e59004 (2013).
41. J. B. Maguire *et al.*, Perturbing the energy landscape for improved packing during computational protein design. *Proteins* **89**, 436–449 (2021).
42. J. Janin, R. P. Bahadur, P. Chakrabarti, Protein–protein interaction and quaternary structure. *Q. Rev. Biophys.* **41**, 133–180 (2008).
43. J. G. Kettle *et al.*, Structure-based design and pharmacokinetic optimization of covalent allosteric inhibitors of the mutant GTPase KRASG12C. *J. Med. Chem.* **63**, 4468–4483 (2020).
44. C. Suloway *et al.*, Automated molecular microscopy: The new Leginon system. *J. Struct. Biol.* **151**, 41–60 (2005).
45. D. N. Mastronarde, Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
46. 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).
47. M. A. Cianfrocco, M. Wong-Barnum, C. Youn, R. Wagner, A. Leschziner, "COSMIC2: A science gateway for cryo-electron microscopy structure determination" in *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact, PEARC17* (Association for Computing Machinery, 2017), pp. 1–5.
48. K. Cowtan, The buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 1002–1011 (2006).
49. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
50. D. Liebschner *et al.*, Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix. *Acta Crystallogr. D Struct. Biol.* **75**, 861–877 (2019).
51. T. D. Goddard *et al.*, UCSF ChimeraX: Meeting modern challenges in visualization and analysis: UCSF ChimeraX visualization system. *Protein Sci.* **27**, 14–25 (2018).
52. E. F. Pettersen *et al.*, UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
53. P. V. Afonine *et al.*, New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Crystallogr. D Struct. Biol.* **74**, 814–840 (2018).
54. G. Pintilie, D.-H. Chen, C. A. Haase-Pettingell, J. A. King, W. Chiu, Resolution and probabilistic models of components in CryoEM maps of mature P22 bacteriophage. *Biophys. J.* **110**, 827–839 (2016).
55. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
56. R. Castells-Graells, M. R. Sawaya, T. O. Yeates, EMD-29700, Cryo-EM imaging scaffold subunits A and B used to display KRAS G12C complex with GDP. Electron Microscopy Database. https://www.ebi.ac.uk/emdb/EMD-29700. Deposited 8 February 2023.
57. R. Castells-Graells, M. R. Sawaya, T. O. Yeates, EMD-29713, KRAS G12C complex with GDP imaged on a cryo-EM imaging scaffold. Electron Microscopy Database. https://www.ebi.ac.uk/emdb/EMD-29713. Deposited 8 February 2023.
58. R. Castells-Graells, M. R. Sawaya, T. O. Yeates, EMD-29715, KRAS G12C complex with GDP and AMG 510 imaged on a cryo-EM imaging scaffold. Electron Microscopy Database. https://www.ebi.ac.uk/emdb/EMD-29715. Deposited 8 February 2023.
59. R. Castells-Graells, M. R. Sawaya, T. O. Yeates, EMD-29718, Green Fluorescence Protein imaged on a cryo-EM imaging scaffold. Electron Microscopy Database. https://www.ebi.ac.uk/emdb/EMD-29718. Deposited 9 February 2023.
60. R. Castells-Graells, M. R. Sawaya, T. O. Yeates, EMD-29719, KRAS G12V complex with GDP imaged on a cryo-EM imaging scaffold. Electron Microscopy Database. https://www.ebi.ac.uk/emdb/EMD-29719. Deposited 9 February 2023.
61. R. Castells-Graells, M. R. Sawaya, T. O. Yeates, EMD-29720, KRAS G13C complex with GDP imaged on a cryo-EM imaging scaffold. Electron Microscopy Database. https://www.ebi.ac.uk/emdb/EMD-29720. Deposited 9 February 2023.
62. R. Castells-Graells, M. R. Sawaya, T. O. Yeates, 8G3K, Cryo-EM imaging scaffold subunits A and B used to display KRAS G12C complex with GDP. Protein Data Bank. https://www.rcsb.org/structure/8G3K. Deposited 8 February 2023.
63. R. Castells-Graells, M. R. Sawaya, T. O. Yeates, 8G42, KRAS G12C complex with GDP imaged on a cryo-EM imaging scaffold. Protein Data Bank. https://www.rcsb.org/structure/8G42. Deposited 8 February 2023.
64. R. Castells-Graells, M. R. Sawaya, T. O. Yeates, 8G47, KRAS G12C complex with GDP and AMG 510 imaged on a cryo-EM imaging scaffold. Protein Data Bank. https://www.rcsb.org/structure/8G47. Deposited 8 February 2023.
65. R. Castells-Graells, M. R. Sawaya, T. O. Yeates, 8G4E, Green Fluorescence Protein imaged on a cryo-EM imaging scaffold. Protein Data Bank. https://www.rcsb.org/structure/8G4E. Deposited 9 February 2023.
66. R. Castells-Graells, M. R. Sawaya, T. O. Yeates, 8G4F, KRAS G12V complex with GDP imaged on a cryo-EM imaging scaffold. Protein Data Bank. https://www.rcsb.org/structure/8G4F. Deposited 9 February 2023.
67. R. Castells-Graells, M. R. Sawaya, T. O. Yeates, 8G4H, KRAS G13C complex with GDP imaged on a cryo-EM imaging scaffold. Protein Data Bank. https://www.rcsb.org/structure/8G4H. Deposited 9 February 2023.

Chapter 5: A Suite of Designed Protein Cages Using Machine Learning Algorithms and Protein Fragment-Based Protocols

Kyle Meador, Roger Castells-Graells, Roman Aguirre, Micheal Sawaya, Mark Arbing

Trent Sherman, Chethaka Senarathne, Todd O. Yeates

The following is a version of a manuscript in submission to the journal

Structure (2023)

Abstract

Designed protein cages and related materials provide unique opportunities for applications in biotechnology and medicine. In the present study, we apply new computational approaches to design tetrahedrally symmetric, self-assembling protein cages. For the generation of docked poses, we emphasize a protein fragment-based approach, while for *de novo* interface design, a comparison of computational protocols highlights the power and increased experimental success achieved using ProteinMPNN. In relating information from docking and design, we observe that agreement between sequence preferences for fragments and ProteinMPNN inference correlate with experimental success. Additional insights for designing polar interactions are highlighted by experimentally testing larger and more polar interfaces. In all, we report five structures for seven protein cages, along with two structures of intermediate assemblies, with atomic resolution in the best case reaching 2.0 Å. The new cages add substantially to the body of available protein nanoparticles, and to methodologies for their creation.

Introduction

Recent advances in protein design are making it possible to engineer self-assembling protein architectures of high complexity [1–3]. The seminal work by Padilla et al. in 2001 [4] laid the foundation for creating novel protein cages and other extended materials by exploiting principles of symmetry. The key idea was that bringing two different symmetry elements together (exemplified by protein oligomers) in a precisely defined arrangement is sufficient to dictate formation of surprisingly diverse and complex assembly outcomes, broadly referred to as symmetry combination materials (SCM). Considering just two-component types, 124 different architectural forms have been articulated mathematically [5]. For each such symmetry form, the combinatorial space for pairs of oligomeric protein building blocks, leads to an extraordinarily deep space for design. Importantly, only a sliver of this design space has been explored experimentally to date. Large machine learning models are enabling far more efficient search of protein conformational space [3,6–9] and effectively applying these tools to model protein materials will be key in the ultimate realization of their potential.

Recent achievements have demonstrated the potential of engineered protein cages and related types of SCMs towards applications in nanotechnology and medicine. Their biocompatibility as well as their size, topology, and multivalency have enabled the localization of target substrates [10], molecular delivery [11] or sequestration of payloads [12], and scaffolding of antigens [13–15], enzymes [16,17], or binders for high resolution imaging [18,19]. Notwithstanding these promising demonstrations, new methods are needed to fully harness the rich functionality and dynamics that are possible with protein assemblies, as exemplified by naturally evolved systems [20–25]. Indeed, emerging efforts to mimic the complex behaviors of natural systems are leading to exciting new design prospects [26,27].

Crucially, the predictable design of novel SCMs remains a challenge [28]. While the first approach for designing protein cages used genetic fusion [4,29], subsequent work (introduced by

King et al. 2012) has increasingly relied on designed protein-protein interfaces. A critical factor for success in such cases is the specification of a geometrically precise *de novo* interface that is accessible and cooperative with the surrounding energy landscape [30]. The amino acids at the interface must encode sufficient information to drive the emergence of quaternary structure [31], while preserving the energetics governing tertiary structure [32]. Importantly, diverse SCM's including cubic and icosahedral protein cages, can be created through the installation of a single interface type between candidate building blocks. Of course, only a minute fraction of possible protein pairs and orientations constitute suitable starting points, so identifying plausible docked positions (*i.e.* poses) is a critically important algorithmic consideration.

In the present study, we apply two computational developments towards the design of two-component tetrahedral protein cages. Addressing the challenge of generating protein interfaces resembling native complexes, we evaluate a protein fragment-based method that identifies poses with *de novo* interfaces exhibiting modes of association commonly observed in nature [33]. With regard to sequence design at the new interface, our first set of designs employed knowledge-based scoring functions. Spurred by those results, a second approach pursued a recently developed graph neural network to achieve greater success. Through biochemical and multi-state structural characterization, we validate a suite of novel protein cages. A comparison of design parameters across protocols (including prior efforts) with patterns of success or failure, offers insights into overcoming challenges in the design of protein interfaces. The new symmetric materials, design techniques, and accompanying software provide a foundation to explore a growing universe of protein materials.

Results and Discussion

Defining the target assembly

As a design target, we chose a tetrahedral architecture denoted T:{C3}{C3} (Figure 5.1a). In materials of this form, two trimeric protein oligomers (A_3 and B_3), each obeying C3

symmetry, are oriented so their 3-fold symmetry axis are coincident with a body diagonal of a cube. Key to the accurate design of bonafide T:{C3}{C3} assemblies is the specification of a fixed connection between the trimeric building blocks to enforce a strict intersecting angle of the two C3 axes. Herein, a *de novo* protein-protein interface is used to specify the relative orientation of the pair of trimers. If the interface is accurately modeled — the trimers associate strongly and with the correct geometry — four copies of each trimer will assemble in an overall stoichiometry of $A_{12}B_{12}$ to form a mid-nanometer scale, protein cage.

In generating candidate poses consistent with tetrahedral symmetry we adopted the fragment-based docking approach implemented in the program Nanohedra [33]. This approach analyzes the surface-exposed segments of two proteins and systematically identifies poses that allow the two components to interact using arrangements of fragments commonly observed in known protein-protein interfaces (Figure 5.1c). We performed all-to-all pairwise docking between 84 different trimers from the PDB, generating ~377,000 candidate poses from 3,584 pairwise trimer combinations. To prioritize interfaces composed of more extensive tertiary motifs, we filtered for poses that utilize fragment observations from three or more discrete secondary structure elements in the interface, resulting in ~45,000 candidates. Finally, we gathered coarse energy and area metrics for the asymmetric unit (ASU) and retained poses with greater than 1,200 Å buried surface area (BSA) in the interface, a negative calculated interface free energy, overall shape-based features (spike ratio), and less than 3 Å root mean squared deviation (RMSD) between backbone atoms from the initially docked pose and an energy minimized model to arrive at 590 candidate poses (see scouting protocol).

Interface design using fragments and knowledge-based hydrogen bond networks

We frame our interface design ideas around a three state model of protein folding and association [34] (Figure 5.1a,b). The first state consists of protein monomers. In the second state, multiple protein chains associate to form one oligomer. Importantly, designs that reach the

discrete oligomeric state avoid multiple off-pathway states that together constitute *solubility errors*. Well-formed oligomers serve as the precursors for the third state, the protein-protein complex, where sets of oligomers are further associated in a specific geometry to form a higher-level architecture. Herein, the complex state constitutes a tetrahedral protein cage. When a complex is not formed successfully from viable oligomers, the result constitutes some type of *design error*, reflecting a failure to establish the intended interface. To ensure the experiment captures feedback from the interface design task, which is critical for improving *de novo* interface design techniques [35,36], we aimed to limit *solubility errors*, *e.g.* by seeking to avoid undue hydrophobicity, and measure *design errors*.

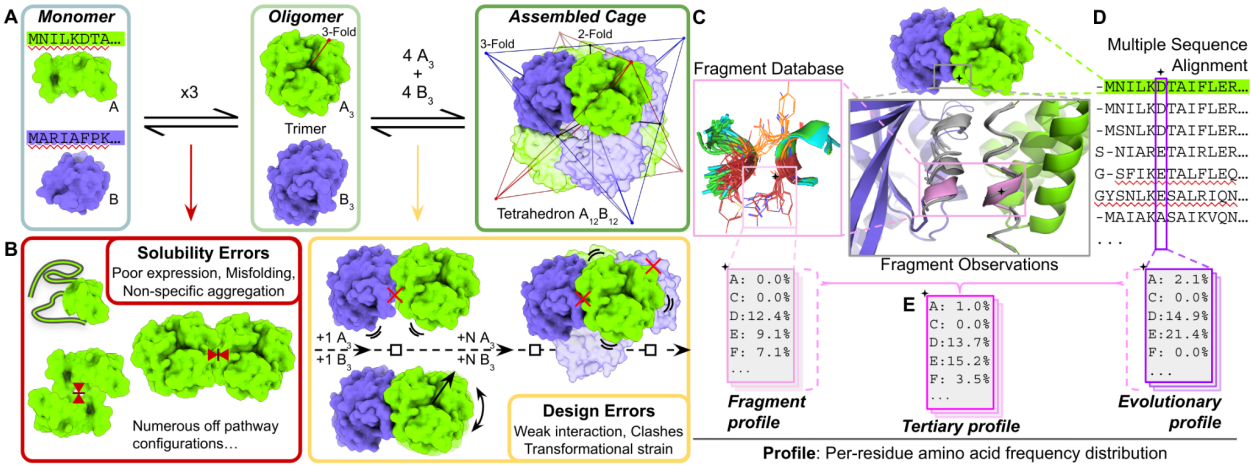


Figure 5.1. A design framework for creating protein cages.

a) A diagram of assembly states for two protein components culminating with the tetrahedral cage outlined by a multi-colored stellar octahedron, where each color constitutes one tetrahedron. **b)** Off pathway solubility and design errors lead to interface design failure. **c)** Fragment Observations are depicted in gray segments and span between protein components with a single observation highlighted in pink. A Fragment Database stores the amino acid frequencies for every residue position, with an example residue indicated with a star. Every fragment observation at one residue contributes proportionally to the frequency distribution. The stack of all fragment residues constitute a Fragment Profile. **d)** The initial protein sequence is used to search for homologous sequences. At the starred position, a frequency distribution of amino acids is calculated and constitutes one residue of the Evolutionary Profile. **e)** The fragment observation superposition modulates how much the Tertiary Profile reflects fragment or evolutionary information at each residue.

Our fragment-based docking (Nanohedra program) includes statistical information on amino acid preferences for tertiary structure (fragment-fragment) motifs [37]. We therefore sought to utilize this information to bias amino acid selections during sequence design. We calculated position-specific amino acid frequencies for every fragment residue in the nascent interface, collectively referred to as the fragment profile (Figure 5.1c). Similarly, an evolutionary profile was calculated from position-specific amino acid frequencies observed from multiple sequence alignments (MSA) of homologous proteins (Figure 5.1d). These two profiles were combined into the tertiary profile, which represents an amino acid distribution conditioned on the tertiary structure of the nascent interface and the underlying protein folds.

Early in method development, we observed that hydrophobic amino acids tended to be favored during design owing to Lennard-Jones score terms [38] and the default FastDesign ramping protocol [39]. As the fragment database utilized for pose identification contains motifs found in interfaces from native complexes, and those complexes typically demonstrate greater polar characteristics [40], we hypothesized that explicitly sampling polar amino acids could reduce the prevalence of hydrophobic atomic interactions while retaining well-packed and complementarity tertiary motifs [41]. To utilize fragment information for design and sampling, we developed a modified HBNet protocol [42], using fragment information to guide the HBNet search. We refer to this protocol as FragmentHBNet. Briefly, interface residues with fragment observations initiate the HBNet search for low energy hydrogen bond networks. In a subsequent step, the fragment profile is used to constrain amino acid sampling during packing and minimization at fragment residues around the hydrogen bond network residues. These dual searches effectively prioritize the most well packed cores that support low energy hydrogen bonding networks. Finally, the tertiary profile is used to guide sampling of the entire interface. Interface design trajectories with FragmentHBNet affected several key interface design metrics, including lowering contributions from hydrophobic BSA while increasing the number of hydrogen bonds.

The 590 candidate poses noted above were subjected to the FragmentHBNet protocol, producing a design model for the top 20 trajectories. All models were then filtered for shape complementarity > 0.68, measured BSA > 1,000 Å, fraction of hydrophobic BSA < 65%, buried unsatisfied hydrogen bonds < 3/1,000 $Å^2$ of BSA, and no residues with deviating Errat scores [43]. One design from each pose was prioritized and a mutation reversion protocol was performed (see Methods). From all reverted and original designs, we again prioritized one design per pose and finally manually inspected each design for missing segments near interfaces; poorly suited termini or unmodeled regions in the proximity of the interface were discarded. Lastly, poses

were clustered for similarity according to iAlign and 41 designs were selected for testing. We refer to this design set using the FragmentHBNet protocol as T33-fn.

Characterization of T33-fn Designs

For biochemical characterization, we added a polyhistidine tag ('His-tag') onto an exposed terminus of one of the two components and synthesized codon-optimized genes for simultaneous expression of both components in *Escherichia coli*. With the His-tag attached to only one component, we expected to purify both components if the interface design was successful. SDS-PAGE analysis of the immobilized metal affinity chromatography (IMAC) lysates revealed 35 cases wherein both genes were well-expressed. In 10 cases, both components were in the soluble fraction after clarifying the lysate by centrifugation. Eight of the soluble designs demonstrated co-elution of both components, however with variation in relative abundance, typically with higher quantities of the His-tagged trimer (Figure S5.1a). For all eight, we concentrated fractions with both components present and isolated the assemblies using size exclusion chromatography (SEC). The chromatograms indicated significant assembly for one design, while the other seven revealed species expected for component trimers, some larger sized fractions containing both components, and limited amounts of assembly-sized species (Figure S5.1b).

For the most well-behaved design, T33-fn10, biochemical characterization revealed properties expected from the design model (Figure 5.2). SEC coupled to small angle X-ray scattering (SAXS) data provided an experimental radius of gyration (Rg) of 5.55 nm which demonstrated excellent fit to the designed value of 5.53 nm (Figure 5.2a,b). Additionally, we obtained crystals that diffracted to modest resolution, with the best dataset reaching 6 Å (structure summary in Table 5.1). A crystal structure was solved by molecular replacement using the design model revealing eight copies of each monomer in the asymmetric unit (ASU). Upon application of crystal symmetry operators, two separate tetrahedral assemblies are

recapitulated, confirming the designed interface (Figure 5.2c,d). Both instances of the assembly show excellent agreement to the design model with mean values of 0.8 Å RMSD over all C-alpha atoms and a value of 0.99 for the assembly local distance difference test (LDDT) (Table 5.2).

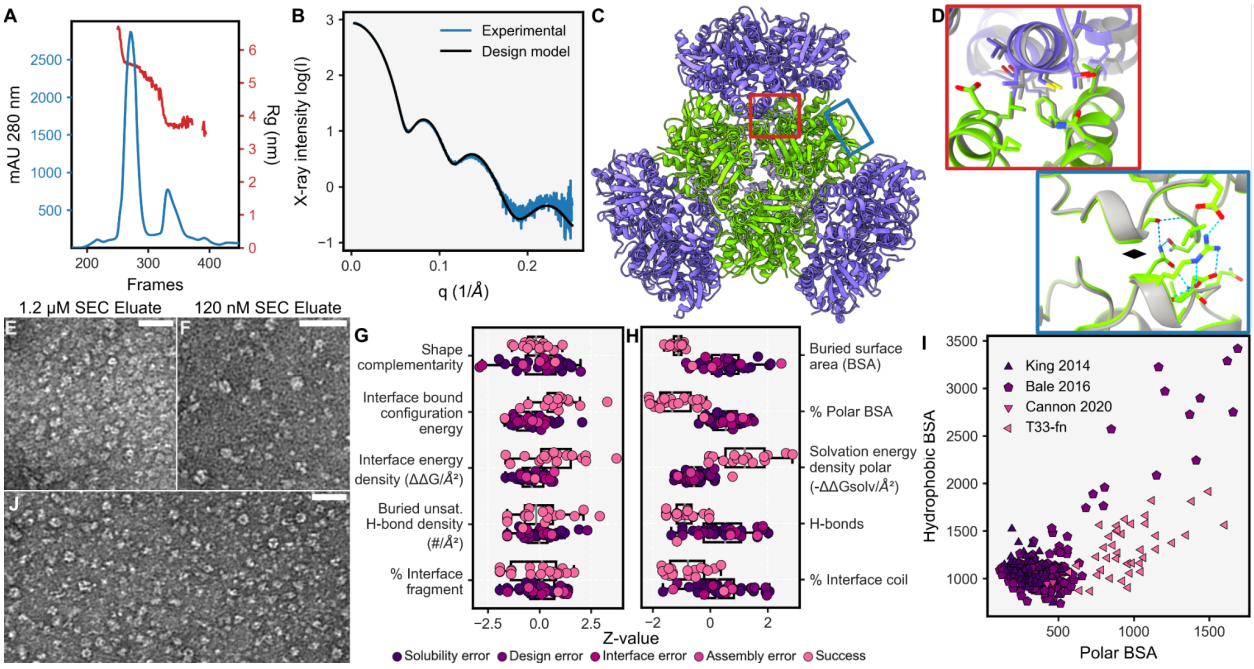


Figure 5.2. Biophysical and structural analysis of T33.

a) SEC-SAXS of T33-fn10 giving a measured Rg of 5.6 nm for the assembled cage. **b)** A SAXS X-ray scattering plot averaged over the fractions corresponding to cage species (blue) and a theoretical plot from the designed model (black). **c)** A crystal structure of T33-fn10 colored according to component trimer type (PDB: 8UJA). **d)** A close-up view of the structure (green/purple) superimposed on the designed interface (gray), split into two regions where the diamond represents a two-fold symmetry axis between separate trimers. **e-f)** Micrographs of T33-fn10 at 1.2 μM and 120 nM, showing disassembly upon dilution of particles. **g)** A comparison of design metrics for successful prior two-component SCM designs (top) and T33-fn designs (bottom) for design filters and **h)** protocol specific differences. Individual designs colored according to design outcome. **i)** Distribution of buried surface area (BSA) according to atomic polarity for two-component designs colored by publication. Markers indicate the design symmetry (triangle - Tetrahedral, pentagon - Icosahedral). **j)** A micrograph of T33-fn40 demonstrating cages after purification by SEC. All scale bars are 50 nm.

Table 5.1. T33-fn10 crystallographic data and refinement statistics

Dataset	T33-fn10
Data collection statistics	
Diffraction source	APS 24-ID-C
Wavelength (Å)	0.97918
Temperature (K)	100
Detector	DECTRIS PILATUS 6M-F
Crystal to Detector distance (mm)	600
Total Rotation Range (°)	360
Rotation per image (°)	0.5
Exposure time per image (s)	0.25
Space Group	$P2_13$
Cell dimensions	
a, b, c (Å)	225.63, 225.63, 225.63
α, β, γ (°)	90.0, 90.0, 90.0
Mosaicity (°)	0.283
Resolution Range (Å)	92.1-6.0 (6.15-6.00)
Rmerge	0.17
⟨I/σ(I)⟩	9.2 (1.1)
$CC_{1/2}$	99.7 (38.3)
Completeness (%)	99.9 (100.0)
No. reflections unique	18590
Redundancy	10.6 (10.6)
$R_{r.i.m.}$	0.174 (2.174)
Overall B factor from Wilson plot (Å^2)	382
Atomic Refinement Statistics	
Resolution range (Å)	92.1-6.0 (6.31-6.00)
No. of reflections, working set	9842 (1245)
No. reflections, test set	986 (138)
R_{work}	21.5 (33.1)
R_{free}	24.8 (37.6)
Molecules/asymmetric unit	16
No. atoms	21720
Average B-factor, all atoms (Å^2)	401.0
RMSD Bond Lengths (Å)	0.010
RMSD Bond Angles (°)	1.02
Ramachandran statistics (%)	
Favored	97.31
Allowed	2.37
Outliers	0.32
Accession code	
PDB (model)	8UJA

Validation by negative stain transmission electron microscopy (nsEM) revealed that specimens prepared at 1.2 μM of assembled complex appeared crowded but mostly homogenous, while specimens prepared at 10-fold dilution lack clear evidence of cage-like assemblies (Figure 5.2e,f). The designed interface results in apparently intact assemblies at concentrations in excess of 1 μM, but with limited assembly at lower concentrations. Designs with unequal stoichiometry were more difficult to characterize. For one design, T33-fn40, nsEM captured assemblies resembling a projection of the intended model (Figure 5.2j), though heterogeneity was noted in SDS-PAGE of SEC fractions (Figure S5.1b).

T33-fn interface design analysis

Notwithstanding the successful creation of T33-fn10, the low success rate, despite promising indications of soluble oligomers and evidence for their association (*i.e.* co-elution), indicated that most designs failed to assemble faithfully. To understand how design indications contribute to biochemical outcomes, we grouped the designs according to the classifications of Figure 5.1a, – *solubility errors*, *design errors*, and *success* – and compared these outcomes to metrics calculated for the design models after Rosetta refinement. We further separated *design errors* into three types: *design errors*, where both components are soluble yet don't interact, *interface errors*, where the components co-elute without higher order assembly, and *assembly errors*, where higher order species are observed, but no tetrahedral complex was observable. Finally, we included the successfully characterized two-component cage design models from King *et al.* 2014, Bale *et al.* 2016, and Cannon *et al.* 2020, to provide further understanding of design requirements.

Important interface metrics, including buried unsatisfied hydrogen bond density [36], shape complementarity [44], interface bound configuration energy [45], and calculated free interface energy density [46] indicated all designs were as favorable, if not more favorable than prior successful designs (Figure 5.2g). Areas of deviation highlight particular choices made in our design

protocol, such as the number of hydrogen bonds, fraction of polar BSA, and calculated solvation free energy density (Figure 5.2h). A notable observation concerns the total BSA in the T33-fn design set. Our initial solvent accessible surface area (SASA) measurement underestimated BSA by ~2-fold (see Supplement, BSA measurements). As a result, T33-fn had relatively large interfaces on average (2,287 $Å^2$), which were substantially polar in character, exceeding total BSA and polar BSA values of previously designed protein cages (Figure 5.2i). Interestingly, T33-fn interfaces utilized extensive loop/coil secondary structure elements, for which it appeared difficult to design adequate atomic interactions (Figure 5.2h).

Our design choices in this experiment were guided in substantial measure by a desire to minimize outcomes where unduly hydrophobic interfaces might lead to solubility errors. While designs were generally soluble, retrospective analysis suggests that T33-fn10 was the single design to possess a solvation free energy well within the range of prior successful designs. Though prior successes also had negative solvation energy values, where negative indicates that aqueous solvation (*i.e*. disassociation) is favored, T33-fn design choices manifested in negative energies overall (Figure S5.2a,b). Though numerous other factors are undoubtedly important, many designs likely lacked crucial hydrophobic contributions to drive interface formation. The challenges in optimizing parameter choices and protocols, clearly highlighted by this exercise, led us to incorporate machine learning techniques into a new protocol for designing protein cages.

Fragment-guided protein cage design using machine learning

In a second set of designs, we revisited both docking and design methodologies, augmenting each step with recent developments in machine learning, where available. For docking, we improved upon techniques for building block input and search of the docking space. For sequence design, we implemented the message passing graph neural network (MPNN) algorithm [47]. Finally, in filtering, we prioritized agreement between fragment observations and

neural network outputs to prioritize sequences that agreed with native interface constraints and where AlphaFold [6] predicted the intended oligomeric components.

As a first improvement, AlphaFold structure prediction was integrated into the design pipeline. In the early stages, we performed predictions to fill in unmodeled regions of the input oligomers; their absence in our earlier protocol led to uncertainty and tedious assessment of potential collisions. In a second step, we assessed whether designed sequences are predicted to fold into the oligomeric state. Both assessments improved confidence that selected designs would meet prerequisites for oligomeric and complex formation.

To improve identification of high-quality docked poses, we implemented three search heuristics in Nanohedra (see Methods). First, we searched for poses where the resulting interfaces had continuous fragment overlaps across multiple residues. This option only searches fragment pairs that participate in higher order relationships*, i.e*. a network of coordinating residues. Next, we clustered coarsely sampled poses into a smaller number of transformational groups, which reduced the search space while maintaining top docking candidates. Finally, each cluster was finely sampled along the rigid body degrees of freedom consistent within the symmetric architecture in order to identify poses with optimal docking metrics. With these improvements, we again docked poses in the T:{C3}{C3} architecture. From 55,611 pairwise combinations of 334 AlphaFold-curated trimers, a total of 72,419 poses, representing plausible backbone models for a novel two-component protein cage, were identified that contained more than four unique secondary structure elements (two from each interface).

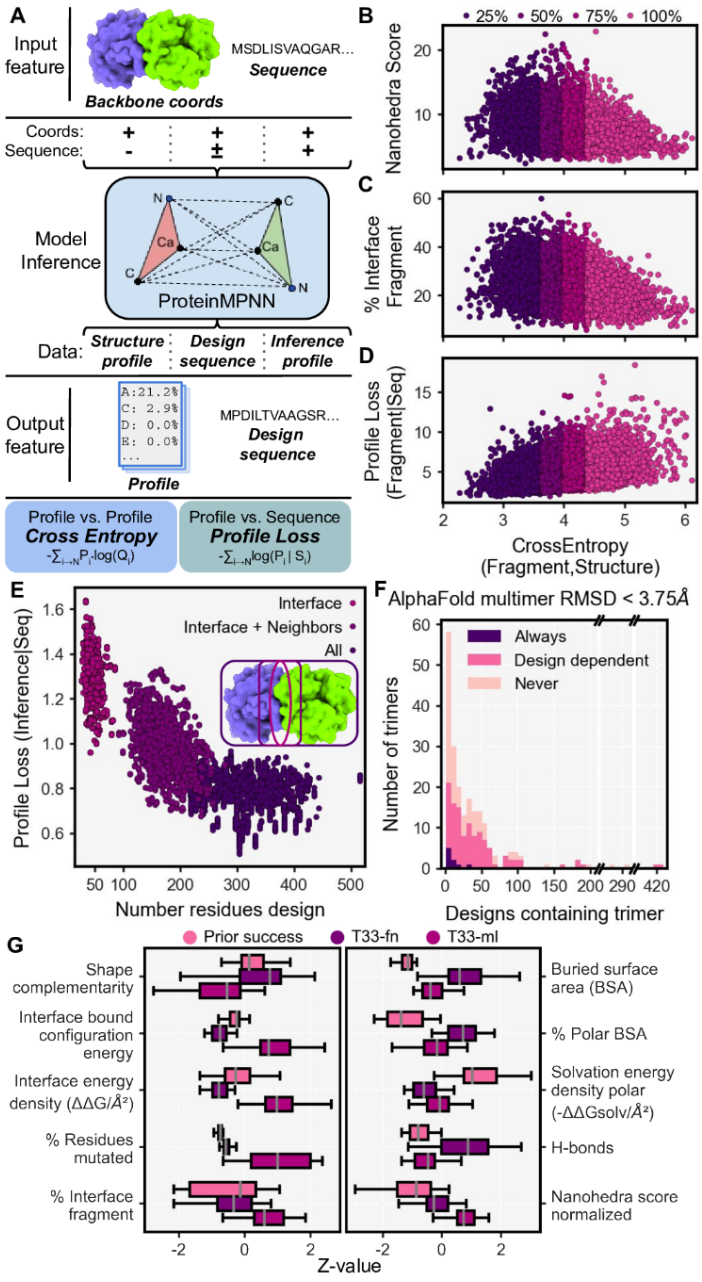


Figure 5.3. Characterization of sequences and poses for T33 machine learning (T33-ml) design models.

a) ProteinMPNN model inference informs on different features of the sequence-structure relationship. Coordinates and sequence serve as input features to vary inference methods and therefore output features. Output features are used to measure cross entropy and profile loss to inform on the fit of inference to other distribution profiles. **b-d)** Comparison of the ProteinMPNN structure profile and fragment profile cross entropy vs Nanohedra score (panel b); percent of interface with fragment observations (panel c); and fragment profile loss given the designed sequence (panel d). **e)** The distribution of ProteinMPNN scores for designed sequences compared with the number of residues chosen for design based on three selection protocols. **f)**

A stacked histogram of ensemble AlphaFoldInitialGuess folding outcome for designed sequences grouped according to each input trimer PDB identifier. **g)** Box plots showing the distribution of T33-ml design metrics scaled according to z-value compared to T33-fn and prior successful protein cages.

For the task of sequence design, we utilized ProteinMPNN [7]. Exploiting ProteinMPNN's computational efficiency, we deeply explored the generative sequence space conditioned on each of the candidate poses, producing ~3.7 million sequences. We performed inference on the backbone coordinates of each pose in the fully symmetric assembly form, using three separate protocols for residue selection comprising all residues, interface residues, or interface residues and their neighbors (see Methods). Across the protocols, we observed the ProteinMPNN score, *i.e* the ProteinMPNN inference profile negative log likelihood (NLL, Profile Loss, Figure 5.3a), was dependent on the number of designed residues (Figure 5.3e), while increasing the number of interface residues had no observable effect (Figure S5.2c).

To explore how ProteinMPNN inference could be used to prioritize docked poses for experimental testing, we investigated the ProteinMPNN structure profile, *i.e.* the amino acid probability distribution inferred only from the pose backbone coordinates (Figure 5.3a). We compared this amino acid profile to the fragment amino acid profile for the pose (described above) calculating the cross entropy (CE, Figure 5.3a) between the two, with lower CE values equating to closer agreement between probability distributions. Lower CE correlated with a larger Nanohedra score (Figure 5.3b) and a higher percentage of the total interface represented by fragments (Figure 5.3c). Importantly, lower CE values correlated with lower fragment profile loss given the inferred sequence (Figure 5.3d). A lower value for the profile loss maximizes the likelihood that, for a single designed sequence, the sequence will encode the sequence-structure relationship captured by fragment-based data. The agreement between fragment information and ProteinMPNN inference identifies poses for which sequences are more likely to encode structures that are compatible with the *de novo* interface.

Given these trends, we filtered for poses with high quality fragment observations whose individual sequences exhibited a fragment profile loss < 5. Additionally, we filtered for sequences where *all* residues were designed, ProteinMPNN score < 1, and the retention of native sequence was >=45%. We also implemented filters to ensure sequences designed in the complex state remain compatible with the oligomeric, unbound state. We filtered for total ProteinMPNN score in the unbound state < 880 and an evolution profile loss (analogous to prior profile loss, but utilizing the evolutionary profile) < 2.5 to avoid aberrant placement of hydrophobic residues.

For each of the 4,241 poses that passed these filters, we chose the single best designed sequence according to ProteinMPNN score and subjected it to predictive (computational) structural validation. On one hand, we took the designed sequence and threaded it into the docked pose, mutating every residue to the designed amino acid and then evaluating structural features. On the other hand, using AlphaFold multimer, we performed folding predictions in the oligomeric state to assess how strongly each sequence specified the proposed trimer. We performed folding without MSA features and instead used the AlphaFoldInitialGuess variant to bias folding on the pose coordinates [48]. We found many of the AlphaFold predictions recapitulated the intended trimeric forms in a sequence dependent manner. When grouped by building block identity, the majority of trimers (n=107) gave predicted oligomeric structures within 3.75 Å (C-alpha RMSD) of the known structure (Figure 5.3f). Further, we found many trimers failed to satisfy this threshold for any designed sequence (n=93), while few trimers folded correctly for every sequence analyzed (n=9).

Following predictive structural validation, we selected designs with shape complementarity >= 0.65, BSA >= 1,500 $Å^2$, buried unsatisfied hydrogen bond density <= 2/1,000 $Å^2$, and interface solvation free energy density >= -0.01/$Å^2$. Finally, we applied two different filters using folding calculations. In the most selective case, we included designs where both trimers successfully folded to within 3.75 Å RMSD (n=248), resulting in 17 designs. In a

more permissive case, we selected designs where only one component folded satisfactorily (n=1,438), yielding 16 designs. To increase design diversity, we added three designs where both trimers folded while we relaxed shape complementarity to >= 0.6 and tightened BSA to >= 1,800 $Å^2$. We additionally selected an equal number of sequences that were created either using the *interface* or *interface + neighbors* residue selection protocols. For the final 65 high quality designs, we subjected each structurally threaded model to thorough refinement in Rosetta and removed 26 designs based on deviating shape complementarity and solvation free energy density.

The resulting 39 T33-ml (machine learning) designs constitute a completely different set of sequences compared to prior work, including those from the T33-fn designs above (Figure 5.3g). Despite less stringent sequence design filters, in most cases the T33-ml metrics fell in ranges between those calculated for prior success and those from the T33-fn set (which were more polar). In most cases, the fraction (or percentage) of residues mutated was an order of magnitude higher than in past designs . Surprisingly, for calculated interface free energies, most designs had positive values, while negative energies indicate more favorable binding or subunit association (Figure S5.2d).

Experimental characterization of T33-ml designs

As before, we appended a single His-tag to a surface accessible terminus and synthesized 38 bicistronic genes for expression in *E. coli* (one design failed during gene synthesis). After expression and IMAC purification, 35 showed both components expressed, while both components were present in the soluble fraction in 24 cases. From these, both components co-eluted from IMAC (Figure S5.3) in 17 cases. These were subjected to SEC wherein four designs demonstrated an unambiguous peak at the expected elution volume for the full assembly (Figure 5.4). For the other 13, more complex elution patterns showed intermediate assemblies and individual trimeric species (or monomers), among minor

populations of intact assemblies (Figure S5.3b,S5.4). Further, SEC-SAXS validated these observations for the best behaving designs (Figure S5.5a). Investigation by nsEM showed particles possessing the predicted size and features of complete assemblies for 6 of the designs, with 6 out of 38 representing an experimental success rate of 16%.

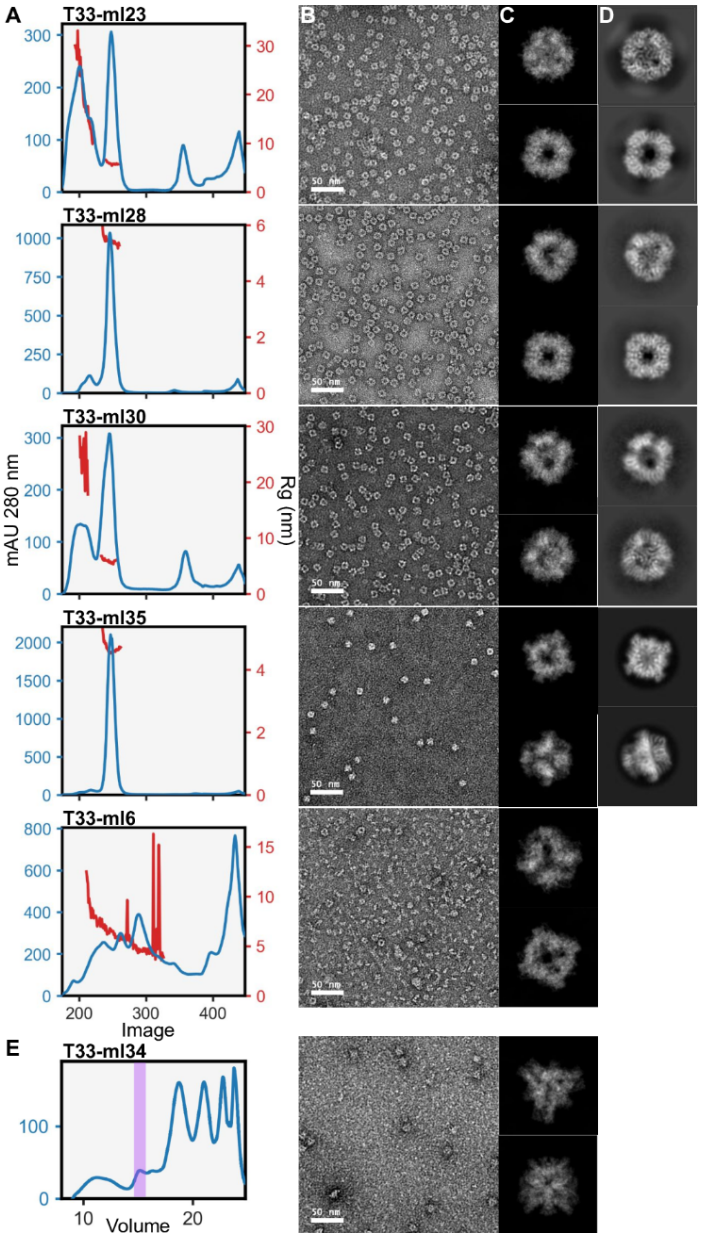


Figure 5.4. Biochemical characterization of six tetrahedral cages produced using machine learning protocols (T33-ml).

a) SEC curves in blue, with SAXS measured radius of gyration (where available), in red. **b)** nsEM micrographs. **c)** 2D projections of the design model and **d)** matching 2D cryo-EM class where possible. **e)** T33-ml34 demonstrates incomplete assembly by SEC, however, the IMAC elution visibly assembles under nsEM. The proposed assembly elution volume is shaded in purple.

We undertook atomic structure determination by cryo-electron microscopy (cryo-EM) for four designs that behaved robustly in solution to assess their accuracy with respect to the intended models. We acquired datasets where images demonstrated 2D classes matching model projections (Figure 5.4), and with sufficient orientational diversity to perform high-resolution 3D reconstructions (structure summary in Table 5.3). In all cases, the reconstructions were consistent with the expected positioning of trimers in the tetrahedral designs. Beginning with computationally designed models, we performed refinement into the resulting density maps, enforcing T symmetry. In Figure 5.5, we present the various cryo-EM structures alongside their respective design models to emphasize the close agreement between design and experiment.

Table 5.2. Structurally validated assembly statistics.

Design Name	Resolution (Å)	Assembly RMSD (Å)	ASU RMSD (Å)	Assembly LDDT
T33-fn10	6.0	0.75	0.45	0.99
T33-ml23	2.0	2.37	1.26	0.78
T33-ml23 $A_{12}B_9$	3.9	2.60	1.45	0.73
T33-ml28	2.7	1.73	1.08	0.86
T33-ml30	4.2	3.62	1.88	0.69
T33-ml35	2.9	1.55	1.11	0.85
T33-ml35 A_9B_{12}	4.4	1.55	1.42	0.85

Agreement between experimental structures and their respective design models. All structures are from cryo-EM except T33-fn10, which was from X-ray crystallography. RMSD was calculated between all corresponding C-alpha atoms in the structure model, which for the ASU, constitutes a single A-B pair, while for the assembly, it constitutes all symmetrically related copies of A and B. ASU - Asymmetric unit, RMSD - root mean squared deviation, LDDT - local distance difference test.

Table 5.3. Cryo-EM data collection, image analysis, modeling, refinement, and validation statistics.

Dataset	T33-ml23	T33-ml23 $A_{12}B_9$	T33-ml28	T33-ml30	T33-ml35	T33-ml35 A_9B_{12}
Data collection						
Microscope	Titan Krios	Titan Krios	Titan Krios	Titan Krios	Titan Krios	Titan Krios
Voltage (keV)	300	300	300	300	300	300
Detector	K3 Summit	K3 Summit	K3 Summit	K3 Summit	K3 Summit	K3 Summit
Nominal magnification	105,000	105,000	81,000	81,000	130,000	81,000
Acquisition software	SerialEM	SerialEM	SerialEM	SerialEM	SerialEM	SerialEM
Electron ose ($e^-/Å^2$)	40	40	40	47.5	40	47.5
Pixel size (Å)	0.86	0.86	1.1	1.1	0.65	1.1
Defocus range (μm)	[-0.50,-2.00]	[-0.50,-2.00]	[-0.50,-2.50]	[-0.50,-2.50]	[-0.50,-2.50]	[-0.50, -2.50]
Data processing						
Number of particles	3,900,288	47,115	857,483	7,114	54,116	66,561
Symmetry imposed	T	C1	T	T	T	C1
Resolution (Å)	2.0	3.9	2.7	4.2	2.9	4.4
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143
Refinement						
Map refinement	Global	Global	Global	Global	Global	Global
Initial model (PDB codes)	*in silico* (4iyq, 5dii)	*in silico* (4iyq, 5dii)	*in silico* (5dii, 6t76)	*in silico* (5dii, 6vvr)	*in silico* (5hrz,6gdx)	*in silico* (5hrz,6gdx)
Protein residues	3456	2904	3492	3528	2724	2364
Non-hydrogen atoms	27000	22100	26568	26196	21432	18615
R.M.S. deviations						
Bond lengths (Å)	0.007	0.014	0.004	0.004	0.007	0.004
Bond angles (°)	0.757	1.168	0.515	0.751	0.735	0.882
Validation						
MolProbity score	1.22	2.17	1.33	2.07	1.35	2.06
Clashscore	4.48	21.07	3.09	18.38	6.39	27.72
Poor rotamers (%)	0.43	0.47	0.90	0.0	0.55	0.0
Ramachandran (%)						
Favored	98.24	94.86	96.52	95.52	99.55	97.29
Allowed	1.76	5.14	3.48	4.48	0.45	2.71
Disallowed	0.0	0.0	0.0	0.0	0.0	0.0
Fit to map (CC_{mask})	0.79	0.65	0.80	0.80	0.87	0.77
Accession codes						
EMDB (maps)	42181	42390	42286	42355	42381	42382
PDB (model)	8UF0	8UN1	8UI2	8UKM	8UMP	8UMR

We evaluated the LDDT of the resulting structures in the context of the design models as well as the RMSD over the entire assembly and over one A-B heteromer from the cage (Table 5.2). The best resolved design, T33-ml23, reached 2 Å resolution, revealing atomic details such as holes in aromatic side chains (Figure S5.5b,c). Although the full assembly deviates from the design by 2.4 Å RMSD, the cage demonstrated high stability and rigidity, which enabled outstanding imaging resolution. In contrast, for the worst-resolved design, T33-ml30, refinement reached a resolution of 4.2 Å with a 3.6 Å RMSD for the whole assembly. Interestingly, in all cases minor rigid body adjustments occurred along the allowed degrees of freedom. Overall, this resulted in slightly larger structures than designed, with an average radius of gyration increase of 1.5 Å (3%).

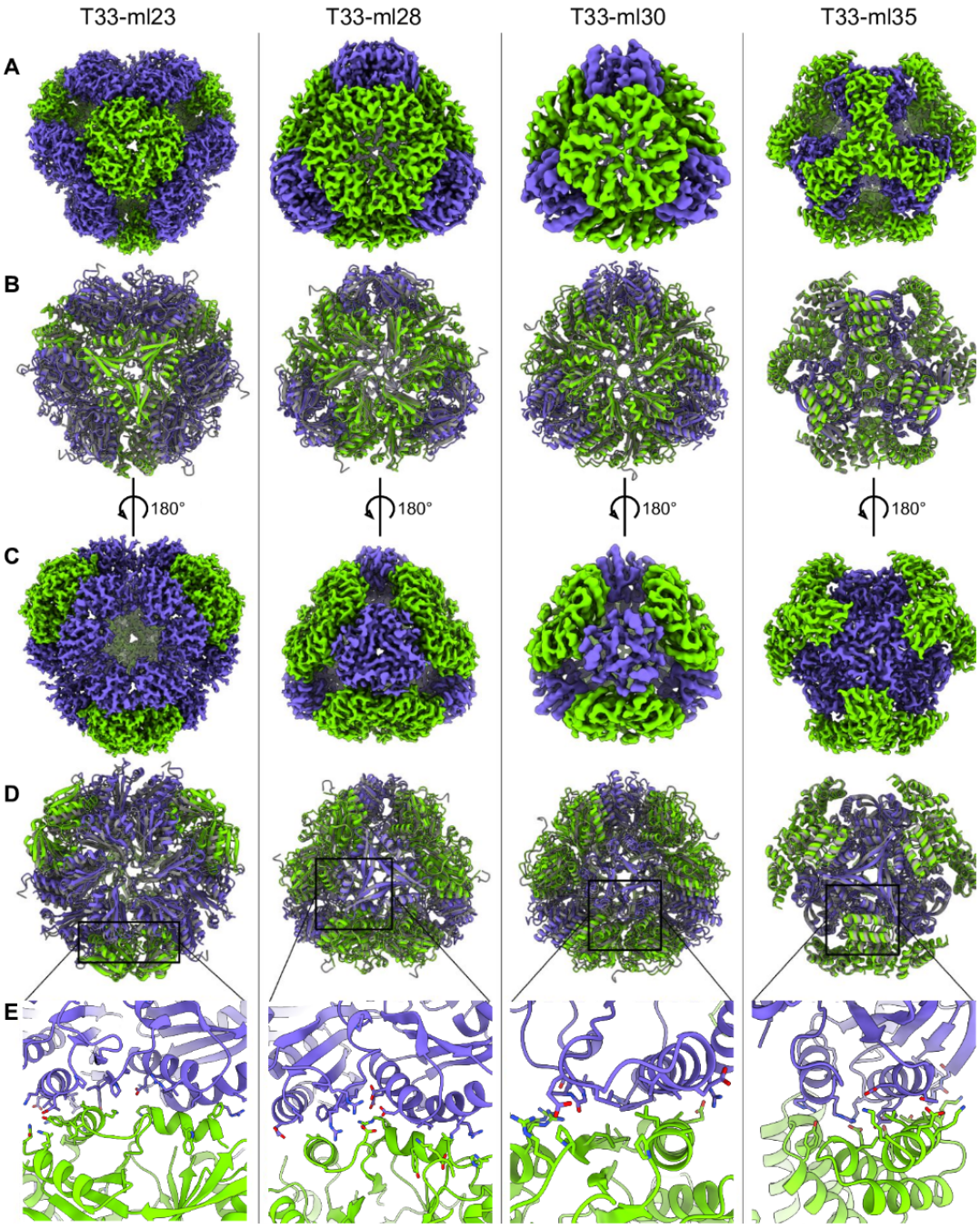


Figure 5.5. Structural characterization of two-component cages by cryo-EM.

a) A view down the 3-fold axis of the reconstructed cryo-EM map for the designs T33-ml23, T33-ml28, T33-ml30, and T33-ml35 (PDB: 8UFO, 8UI2, 8UKM, and 8UMP respectively) with the first component colored green and second component purple. **b)** The structure superimposed on the design model (gray).. **c)** The reference frame is rotated 180° to center the cryo-EM map and **d)** the design model, on the 3-fold axis of the second component. **e)** A close-up of the side-chains which participate in the *de novo* interfaces.

T33-ml interface design analysis

Cryo-EM structures of our T33-ml cages showed overall deviations greater than 1.5 Å from the designed models (Table 5.2), motivating a comparison of calculated interface metrics for the experimental structures vs the designed models. We found that the calculated interface energies were more favorable for the experimental structures, likely indicative of the atomic inaccuracies resulting from sequence based design models. We also noted that the values for calculated interface energies were mostly positive (tending to indicate dissociation) in contrast to energies typically prioritized by knowledge-based methods. Among other limitations of calculated energy values, protein concentrations, especially in situations where many components associate in a complex, are not adequately captured.

We classified the T33-ml designs according to biochemical outcome (discussed in *T33-fn interface design analysis*) to understand predictors of interface design success. We found that the ProteinMPNN score was not a strong predictor of design outcome (Figure 5.6b). When accounting only for soluble designs, the strongest predictor of biochemical outcome was a minimal fragment profile loss given the designed sequence (r^2=-0.59, Figure 5.6c). That is, successful designs tended to be those for which the sequence inferred by ProteinMPNN maximized the likelihood of utilizing amino acids favored by fragment observations. This finding highlights the benefit of using fragment-based sequence-structure information to guide pose selection and improve experimental outcomes for *de novo* designed interfaces.

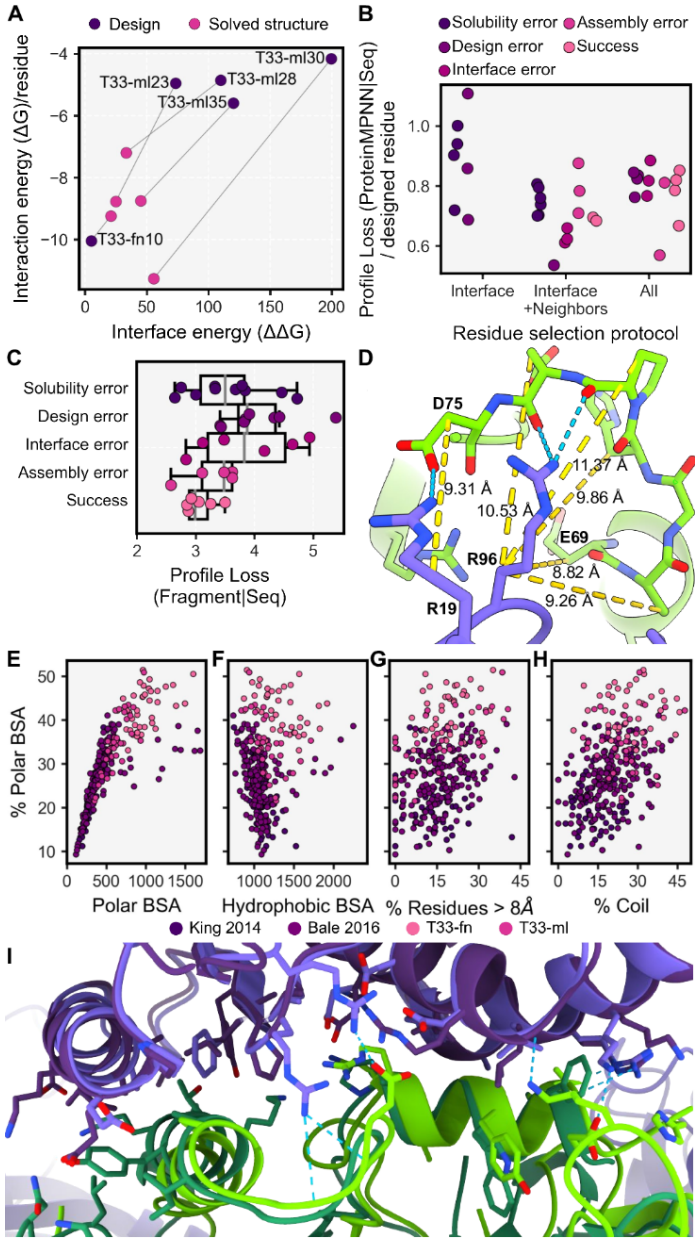


Figure 5.6. Understanding design outcomes.

a) Rosetta interaction energy vs interface energy for the structures and their design models presented in this work **b)** ProteinMPNN score per designed residue for biochemically characterized T33-ml designs grouped by residue selection protocol and colored according to design outcome. **c)** T33-ml designs grouped according to solubility outcome correlates with fragment profile loss.. **d)** Example of a hydrogen bonding interaction between residues (A:19 and A:96 and B:69:75 in T33-ml28) with C-beta distances greater than 8 Å .**e-h)** For all two-component cages from King *et al.* 2012, Bale *et al.* 2016, and this work, the distribution of polar buried surface area fraction (BSA) as a function of **e)** polar BSA, **f)** hydrophobic BSA, **g)** percentage of interface residues greater than 8 Å (C-beta to C-beta) from opposing residues and **h)** percentage of loop/coil secondary structure (Coloring according to publication or

T33-fn/T33-ml). **i)** Interface comparison of T33-ml23 (dark purple/dark green) to T33-ml28 (purple/green). Cross-interface hydrogen bonds and salt bridges are indicated in blue for T33-ml28.

Variations and plasticity at the interfaces of T33-ml cages

We noted several cases where the observed structures revealed conformational adjustments that could be interpreted, *post-facto*, in atomic terms. For T33-ml30, a pair of helices, with high fragment representation, interact across the interface, while peripheral strand and loop/coil sections participate in polar interactions. Though ProteinMPNN selected arginine residues to satisfy hydrogen bond donor/acceptor pairs at the periphery, the structure reveals the polar groups were ultimately unsatisfied by interactions with other protein atoms. Instead, their burial was avoided through a rigid body rearrangement around the helical fragments (Figure 5.5e) which maintain the highest LDDT values for the pose (Figure S5.6). The T33-ml35 cage provides another example of this phenomenon. There, a rigid body shift compared to the design allows space for W90 and R94 of component B to fit in the interface, while adjacent fragment residues maintain high agreement with the design and participate in tight hydrophobic packing and extensive hydrogen bonding (Figure 5.5e).

Two cages, T33-ml23 and T33-ml28, offer a particularly unique situation for comparison. The trimeric components that formed the basis for these two cages happen to be structurally homologous and docked similarly (Figure 5.6i). Both cages share one trimer, a tandem-BMC microcompartment protein (PDB: 5dii) while the other trimers are CutA protein family homologs (PDB: 4iyq, 6t76). Despite overall structural similarity (2.9 Å RMSD overall assembly and 2.0 Å for the A-B subunit pair), the two cages have rather dissimilar sequences. As both cages had all residues selected for redesign there is only 58% overall sequence identity between the two cages (70% for 5dii, 35% for 4iyq/6t76), and only 44% over residues shared in the interfacial region. The interfaces also differ in overall character. For T33-ml23, the interface is more hydrophobic (75%), while the T33-ml28 interface (60% hydrophobic) presents over 10 polar interactions between the two trimeric components. This comparison highlights a high degree of

sequence/structure degeneracy within the designed assembly space, an observation consistent with patterns of natural evolution in protein-protein interfaces [49,50].

Increasing polar interactions in designed interfaces

We analyzed interface metrics for T33-fn, T33-ml, and all 2-component SCM designs docked by King *et al.* 2014 and Bale *et al.* 2016. Unsurprisingly, we found that the amount of polar BSA positively correlates with the fraction of polar BSA (r^2=0.783), however hydrophobic BSA was independent of the fraction of polar BSA (r^2=0.095) (Figure 5.6e,f). For the surveyed designs, these relationships indicate that increased polar contributions are primarily obtained by increasing the total BSA rather than replacing hydrophobic atoms below some necessary minimum.

We next examined conformational features to understand how designs managed to bury polar atoms at their interfaces. The fraction of polar BSA positively correlated with the percent of interface residues whose C-beta atoms are separated by greater than 8 Å (Figure 5.6g). In those cases, polar interactions come from side-chain to side-chain contacts (enriched at distant side-chain positions, Figure S5.7a), or through side-chain to backbone interactions, which can include C-beta distances up to 12 Å (Figure 5.6d). We next analyzed the extent of loop/coil segments in designs, as they make up a significant proportion of interface secondary structure in natural complexes [51] and inherently have more backbone polar atoms available for hydrogen bonding. The percent of interface loop/coil segments also positively correlated with the fraction of polar BSA (Figure 5.6h), which was independant fromthe percent of interface residues greater than 8 Å away (Figure S5.7b). It is notable that utilizing backbone noise within ProteinMPNN design enabled backbone variation or uncertainty to be sampled. This approach offers an advantage for generating designs with increased utilization of polar interactions available to loop/coil regions.

Structural analysis of partial assembly states

During processing of cryo-EM data for the T33-ml23 and T33-ml35 cages, ~20% of particles were classified in conformations representing partial assemblies along the route to a full $A_{12}B_{12}$ cage. We performed 3D reconstruction using C1 symmetry and built models into 3.9 Å and 4.4 Å density maps for T33-ml23 and T33-ml35 respectively (Figure 5.7). For each, the partial assembly structures were very similar to the intact cage, but with a single trimer absent from the reconstruction (Table S5.1).

A notable difference in the partial assemblies was the distance between equivalent backbone atoms across the missing trimer. For both intermediate structures, this distance increased by 1.0 Å compared to the full assembly. We take these displacements to reflect the magnitudes of the atomic rearrangements that occur during the stepwise or hierarchical assembly processes. We note that in T33-ml23, the missing component is a BMC trimer. Several studies have highlighted the flexibility of BMC proteins [52,53], and the observation of structural variation in partially assembled forms could reflect underlying flexibility of the components.

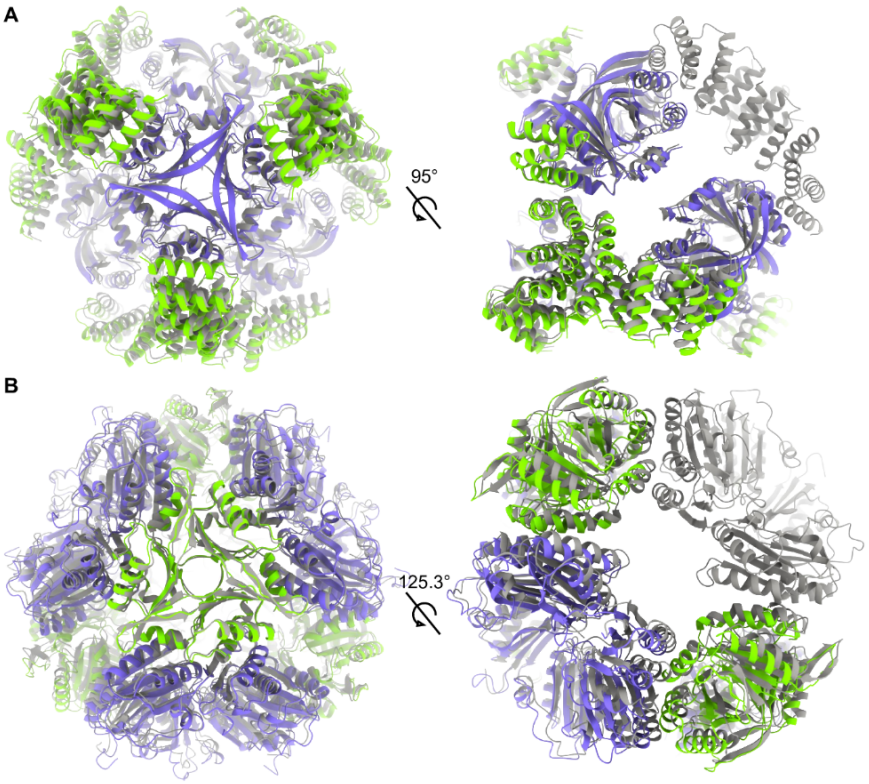


Figure 5.7. Structural comparison of intermediate and full assemblies.

a) T33-ml35 A_9B_{12} (green/purple, (PDB: 8UMR) superimposed on the gray $A_{12}B_{12}$ assembly which upon rotation shows B components from A_9B_{12} protrude into the site of the missing A_3 trimer. **b)** T33-ml23 $A_{12}B_9$ (green/purple, PDB: 8UN1) superimposed on the gray $A_{12}B_{12}$ assembly which upon rotation shows A components from the $A_{12}B_9$ intermediate relaxing away from the missing B_3 trimer.

At present, ascribing an underlying cause to the formation of partial structures is difficult. Such cases could arise when minor design defects (*e.g.* slight deviations from ideal angles between components) propagate, making addition of the last component dependent on substantial atomic movements (Figure 5.1a). Alternatively, partial assemblies could reflect kinetic traps, as has been discussed in the context of viral capsids [54]. Finally, for cages built from two (or more) distinct subunit types,one component could be depleted before the other during assembly. To this point, it is notable that many viral capsids that assemble from multiple subunit types employ a proteolytic mechanism to produce equally abundant subunit types from a longer polypeptide [55].

Conclusions

The present study demonstrates the successful design and structural validation of a new suite of nanoscale symmetric protein cages using protocols that prioritize fragment-based

sequence-structure relationships at *de novo* interfaces. These results elaborate on prior demonstrations of symmetric cages [4,28,29,41,56,57] and fragment-based design [58,59] by utilizing higher order positional relationships present in both fragments and graph-based neural networks to capture native protein properties.

For five of the new cages, we were able to determine atomic structures by X-ray crystallography or cryo-EM. Analysis of these cases generally confirmed the formation of intended tertiary motifs. Though non-helical motifs were employed in some cases, the designed interfaces were rich in alpha helical interactions overall. Differences between designed models and observed structures were generally small, but sometimes consequential in leading to unexpected, favorable atomic interactions.

From our studies, a comparison of different computational protocols – knowledge-based vs machine-learning – is complicated by the introduction of multiple different heuristic choices. Nonetheless, the high success rate (approximately 16%) that we achieved using the latest graph neural network approaches is notable. Introducing complex, native-like, structural features by intentional design is a major computational challenge, and this is particularly true with regard to polar interactions such as hydrogen bonds, which rely on greater atomic precision than hydrophobic interactions. Our successful results, as well as contemporary reports [60], indicate that machine learning methods are favorably suited for such complex tasks. Also notable for our design work was the allowance in machine learning for considerable backbone variation (*e.g.* by networks trained with backbone noise). In our study, this manifested in cases where important interfacial interactions were made between side chains in loop regions, which, in the absence of backbone variations in modeling, such as fixed backbone calculations, would not have been possible. The consideration of backbone variation – including in insertions and deletions – is an area of key importance in ongoing work on antibody design [61,62]. The machine learning approach also allowed a great deal of sequence novelty, i.e. design outside of the interface regions, while simultaneously improving success rates.

As a last point of interest, our studies also illuminated the structures of partially assembled cages, which are likely representative of intermediates in the multi-step assembly process for cage formation. Indeed, partial structures appeared to dominate in a number of our failed designs. Understanding partial structures and the events leading to them could be impactful in improving design successes in future work.

Acknowledgements

Conceptualization, KM and TOY; Methodology, KM, RCG, RA, MS, and TOY; Software, KM; Formal Analysis, KM; Investigation, KM, RCG, RA, MS, MA, TS, and CS; Resources, TOY; Data Curation, KM, RCG, and MS; Writing - Original Draft, KM and TOY; Writing - Review & Editing, KM and TOY; Visualization, KM and RCG; Supervision, KM and TOY; Funding Acquisition, TOY.

Code availability

All code is freely available and hosted at https://github.com/kylemeador/symdesign. To perform design as described in this publication, a Google colab document is available at https://bit.ly/symdesign-colab.

Methods

All descriptions of computational processes are implemented in the symdesign repository https://github.com/kylemeador/symdesign. To perform design as described in this publication, a Google colab document is available at https://bit.ly/symdesign-colab. Where applicable, functions or scripts that are indicated to perform the described procedures are indicated in italic and relative to the repository's main directory, i.e. *path/to/symdesign*. Flags and arguments for program operation are indicated in courier font.

Structural preprocessing and docking

Docking round 1 inputs

Each of the round 1 trimers (see Supplement) was retrieved from the Protein Data Bank (PDB) in July 2020 using the following filters: global symmetry symbol equal to C3, X-ray diffraction dataset with resolution <= 2.5 Å, helical content > 30%, and *E. coli* as the protein expression organism. Entries with more than one protein entity or that contained nucleic acid entities were removed. Entries with a rcsb_polymer_entity_annotation.type containing the words OPM, mpstruc, MemProtMD, PDBTM, or MEMBRANE PROTEIN were removed. All structures were clustered using the 70% sequence identity threshold available from the PDB. The clustered representatives were then cross referenced against QSBio to select only high or very high confidence biological assembly predictions [63], and the corresponding biological assembly was downloaded. 84 candidate oligomeric building blocks were selected for docking trials. Before docking, a monomer from the assembly was chosen and subjected to symmetric refinement into the REF2015 score function [38] using the FastRelax mover and the suggested flags for pareto optimal refinement [64].

Docking round 2 inputs

Each of the round 2 trimers (see Supplement) was retrieved from the PDB in March 2023 using the following filters: global symmetry symbol equal to C3, but not symmetry type

dihedral, X-ray diffraction dataset with resolution <= 3 Å, *E. coli* as the protein expression organism, and 80 <= number of residues <= 300. Entries with more than one protein entity or that contained nucleic acid entities were removed. Entries with a rcsb_polymer_entity_annotation.type containing the words OPM, mpstruc, MemProtMD, PDBTM, or MEMBRANE PROTEIN were removed. Additionally, if the entry title contained the keywords tail, fibre, shaft, head, spike, glycoprotein, ectodomain, or “receptor binding protein”, it was removed. All entries that satisfied these selection criteria were input into the thermophilic prioritization, assembly confirmation, and sequence clustering protocol as described subsequently.

First, the list of all organism taxonomy IDs were collected from ThermoBase_ver_1.0_2022 [65]. These taxonomic ids were used to filter the selection criteria described above for only entries in which the source organism is one of the thermophilic ids. Each matching thermophilic entry ID was then subjected to the 30% sequence clustering service provided at the PDB. The sequence cluster groups returned entry ID’s sorted according to resolution. Each subsequent ID was iteratively tested for a high or very high confidence biological assembly annotation from QSbio [63]. If a matching biological assembly entry wasn’t located within the cluster group in QSBio, the PDB was iteratively queried until an assembly with both “author_defined_assembly” and “author_and_software_defined_assembly” annotations was identified. This procedure was repeated in order of decreasing resolution until a match was found or all cluster group members were exhausted. For each cluster group where a confirmed assembly was located, the group was removed from further selection and the entry identifier was saved.

After selection of proteins from thermophilic taxonomic ids proteins, all remaining entry IDs from the initial selection criteria that weren’t found via thermophilic prioritization were again clustered according to 30% sequence identity from the PDB. Again, these non-thermophilic sequence clusters were subjected to the resolution sorting and assembly confirmation

procedure in the prior paragraph until all 30% clustered groups were either represented by a confirmed assembly or discarded when all members were exhausted with no confirmed assembly representative. Every entry that passed thermophilic prioritization, assembly confirmation, and sequence clustering had the corresponding biological assembly downloaded for docking.

Building blocks may be retrieved using this procedure from the PDB API using the script present at: *symdesign/tools/retrieve_oligomers.py*

Modeling missing density

Each trimeric PDB entry was preprocessed to model missing density in internal loops and termini using AlphaFold. The reference sequence associated with the PDB entry was queried using hhblits (see Running hhblits) and the resulting multiple sequence alignment was parsed into a format corresponding to AlphaFold msa feature arrays. All subsequent inference occurred as described in AlphaFoldInitialGuess inference using the reference sequence as the sequence input and the biological assembly coordinates as coordinate input. The only exception was that the AlphaFold msa features processed here were provided to the AlphaFold feature dictionary instead of using a blank feature dictionary.

Of the AlphaFold predicted structures, the structure with the lowest RMSD to the input asymmetric unit (measured over all originally present C-beta coords from one protomer of the trimer) was structurally aligned to the asymmetric unit and selected as the disorder modeled trimeric representative. Each trimeric assembly was examined by eye for successful recapitulation of the input biological assemblies by AlphaFold. If the extent of modeled disorder was minimal, i.e. only few residues at the termini were added, the original trimeric biological assembly was used in place of the AlphaFold model. In rare cases, there were larger deviations present when comparing the biological assembly to the AlphaFold model by eye which may have indicated erroneous alignment of the prediction to the biological assembly given

alternatively modeled residues or a lack of evolutionary evidence for the trimer resulting in weak predictions around the oligomeric interface. In all these cases the biological assembly was chosen if there was no unmodeled density, otherwise the trimer was discarded from consideration. Finally, all trimers were visually examined for large protruding features which caused prolonged instead of compact structures. These were additionally removed.

Whether a building block was provided in a file or retrieved from the PDB, the flag `--loop-model-input` will specify loop modeling should occur while performing preprocessing for the `nanohedra` module.

Generation of fragment observations

Fragment observations are identified as in Laniado et al. 2021 [33]. Observation proceeds by identifying interface residues for the pose, examining each residue for surface accessibility, and then matching surface exposed residues to a particular fragment type, where the fragment type is determined by structural match between neighboring residues of a specified interval (here +-2 residues) from the residue of interest (for a total of 5 residues) and corresponding fragments of protein structure from the fragment database (totalling 5 residues in length). The fragment type is similar to a secondary structure classification, however utilizes neighboring atoms for classification. Next, those surface accessible interface residues which match a fragment type are examined for potential fragment interaction by querying the identified fragment type against all unique fragment clusters from all fragment types. Fragment clusters constitute a spatially distinct observation of one fragment type with another and are clustered into dense, orientationally dependent groups by RMSD measurements. Fragments which do not clash with backbone or C-beta atoms constitute “ghost fragments” and represent each of the possible orientations between two fragment types which can potentially interact for the identified residues. Finally, ghost fragments are compared against another group of residues to match the other residue’s fragment type to the ghost fragment, fragment type. A match is identified if the

RMSD between the other fragment and the ghost fragment is below a threshold, in this case 1 Å. For the case of interface residue fragment observations, the grouping of other residues belong to the surface of another oligomer, which may be of the same or a different entity type.

Nanohedra docking

All docking procedures can be performed using the `nanohedra` module from the provided git repository.

Nanohedra fragment based docking routines were used as implemented in Laniado et al. 2021 [33]. For T33-fn designs, docking was performed with the following parameters `--minimum-matched` 3 `--initial-z-value` 1 `--match-value` 0.5. In lieu of providing a desired fragment type for docking search, a modification was made to automatically calculate the majority secondary structure type of all surface exposed residues. This secondary structure type was used as the source of initial fragment overlap searching.

For T33-ml docking, the following modifications were used. Input building blocks had their termini trimmed back to remove extended loop/coil segments that mainly arise as the result of preprocessing with AlphaFold. This trimming occurs through the flag `--trim-termini`. Fragment potentials, i.e. ghost fragments, were subjected to an additional search constraint that they must also initially match with another ghost fragment from the same protein component. We refer to this search technique and implement its usage through the flag `--continuous-ghosts` due to the use of multiple overlapping ghost fragments which overlap for a continuous region of the identified fragment potential. For instance, the residue i and i+4 of component 1 have continuous ghosts if their ghost fragments occupy the same location. Such a scenario is more likely if they are members to the same secondary structure and that secondary structure's potential contacts consist of regularly spaced contacts.

Transformational clustering, fine grained search of docked space

As our ideally docked protein-protein interfaces contain multiple fragment observations, the pose created by ideal overlap of any single fragment observation has the effect of creating multiple redundant docked observations with subtle variation in transformational parameters. In docking, typically an advantage of redundant observations can be realized through clustering. By grouping each similar pose into an ensemble of potential positions, a global analysis of ensemble density revealed positions with the highest potential. To both prioritize fine grained search within clustered ensembles and reduce the roughness produced by discrete sampling of the available degrees of freedom, we implemented a grid optimization procedure into the final stage of docking. Importantly, optimization proceeds within the available degrees of freedom (DOF), so each available transformational DOF is placed on a grid and combinatorially sampled. After any single round of optimization, those positions with the highest scores are selected and a new grid is sampled from the DOF until the optimization target function is achieved. In this work, we chose Nanohedra score as our optimization target function and sampled until the change in score from one round of optimization to the next fell below %5 improvement or sampling new grid positions fell below the resolution of adequately describing a unique transformation. Unique transformations were binned into a six dimensional transformation hash, consisting of three translations and three euler angles, to describe the unique parameters that specify a transformation of the two rigid bodies with regards to one another.

Metric calculation

Residue types

Given a docked pose will have new amino acids specified at interacting residue positions, *interface residues* were defined as residues based on C-beta C-beta atom distances. For docking measurements, the distance to identify residues was 9 Å, while for design, 8 Å was used. For all interfacial C-beta contact search procedures, residues are queried from a

heterotypic globular component. This criteria also includes residues from a symmetrically related, i.e. non-self oligomer, of the same protein entity. For measurements on designed structures, *interface residues* were identified by their contribution of BSA to the interface. In all cases where the residue is modeled with the amino acid glycine, the C-alpha atom is used to measure C-beta distances. *Neighbor residues* are defined as residues with a C-beta atom within 8 Å of another residue's C-beta atom. *Interface fragment residues* are defined as residue positions containing a fragment observation as identified by a fragment potential search. The interface *core*, *rim*, and *support* residues are defined according to [31].

Position specific profile calculation

Each profile constitutes a per-residue amino acid frequency distribution which is tabulated over all positions of the pose in question. All distributions are limited to the 20 canonical amino acids and positions where no information is present are discarded.

Fragment profile

Every fragment observation is an association between a tertiary motif observed during modeling and a database of structurally clustered fragment observations. To represent the sequence preferences associated with a single fragment cluster, the amino acids observed from each member of the cluster are tallied to create a cumulative amino acid frequency distribution for each residue position in the fragment cluster. The resulting set of distributions represents the amino acid probabilities associating observed sequence preferences with the structural motif.

When fragment observations are identified for a pose, multiple fragments may be present at each residue. To capture the full sequence-structure probability from multiple fragment observations, each observation references the respective amino acid distribution at the corresponding structural position from the cluster. After collecting all participating distributions, they are combined to reflect the contribution of each fragment observation to the total fragment potential. This is accomplished through scaling every fragment observation

distribution by both a match score and an interaction weight to enable summation to a single distribution.

For the match score,m, the structural match of the pose fragment observation to the representative fragment from the respective cluster is calculated according to equation 1.

$$m = \frac{1}{1 + (RMSD_{frag}/RMSD_{cluster})^2}$$

Eq. 1

Where the $RMSD_{frag}$ is the RMSD between the fragment observation in the pose and the RMSD of the representative fragment from the cluster, and the $RMSD_{cluster}$ is the average RMSD measured for all member fragments which belong to the cluster.

For the interaction weight, t – the interaction importance – each structurally aligned residue in a fragment cluster is measured for atomic contacts with the opposite fragment pair. In this measurement, the number of times that side-chain atoms from the residue of interest interact with atoms from the opposite fragment is normalized by the number of side-chain atoms in the residue. This normalization creates a relative interaction weight for this residue in this member fragment from the cluster. Finally, for every member of the cluster, the mean interaction weight of the same residue position is taken as the residue specific fragment interaction weight, t_r. Calculation of t_r is defined by equation 2.

$$t_r = \frac{1}{M} \sum_{F \in C} \frac{\sum_{a_r \in r_{sc}^{atoms}} \sum_{p \in F_{paired}^{residues}} \sum_{a_p \in p^{atoms}} 1 \; if \; a_r - a_p \; else \; 0}{\#\; r_{sc}^{atoms}}$$

Eq. 2.

Where f is the fragment observation from the set of all fragment observations from the fragment cluster, C. Where p is a residue from the set of residues belonging to the paired fragment

structure, $F^{residues}_{paired}$. Where a_r is an atom from the side-chain (sc) atoms of residue r and a_p is an atom from residue p and M is the number of members in the fragment cluster, C.

Equation 3 defines the process for creating a resulting amino acid distribution for a single residue.

$$P_{aa} = \sum_{aa \in AA} \sum_{F \in C} f^F_{aa} \times m_F t_F / mt_{total}$$

Eq. 3

Where AA is the set of all amino acids and aa is a single amino acid. Where F is a single observation from the set of all fragment observations, C. f^F_{aa} is the frequency of amino acid aa from fragment observation F. Where $m_F t_F$ is the match to the fragment cluster from observation i multiplied by the interaction potential for the particular fragment index associated with the fragment cluster from observation F. Finally, mt_{total} is the sum of all $m_F t_F$ observations for the residue and creates the denominator for the scaling factor for each F observation by its overall contribution importance.

The total fragment profile for a pose reflects all amino acid frequency distributions, P_{aa}, calculated for each residue in the pose with fragment observations.

Evolutionary profile

The hidden markov model written to the .hhm file extension was parsed and the model values for each amino acid at each residue were converted to amino acid frequencies. These values constituted the per-residue amino acid frequencies for evolutionary profile based calculations. All positions of an evolutionary profile that were generated based on a reference sequence, but were used in subsequent structural analysis were removed from the structure representation to keep the index of the reference residue aligned with the structure. This is important in adjusting disordered internal or terminal residues to have the proper profile alignment.

Tertiary profile

The tertiary profile is the per-residue amino acid distribution calculated from a combination of the fragment profile and the evolutionary profile. The tertiary profile assumes the amino acid frequencies of the evolutionary profile for residue positions where there are no fragment observations, and therefore, no fragment profile. For positions where there is fragment profile information, the resulting distribution is weighted by the fragment profile distribution with a maximum weight of alpha (here 0.5) with the inverse, (1 - alpha), contributed by the evolutionary profile. If the fragment observations are deemed weak, according to the fragment match, m and the fragment interaction, t, then alpha is reduced by a modifier which reflects the weakness. Such a modifier reduces alpha, and thus the fragment profile's contribution to the tertiary profile. The m and t value modifiers are calculated by comparison to the fragment database and cluster references, respectively. If the values fail to meet quality thresholds, alpha is modified by the proportion of the discrepancy.

During Rosetta design protocols, the tertiary profile specified which amino acids were available for sampling at each position. Those amino acids that had a frequency greater than 0, through the flag –use_occurance_data, were available for design. For the HBNet protocol, those amino acids that had a frequency greater than 0 for the fragment profile were utilized for the design of fragment residues.

Cross Entropy

The information required to represent one profile as a different profile was calculated by measuring the cross entropy between them. For all measurements, the following formula was used:

$$CE = -\sum_{i=1}^{N} P_i \times log(Q_i)$$

Eq. 4

Where for all residues i in a profile with length N, P is the “true” profile and Q is the profile of interest. When cross entropy is mentioned in the text or figures, the “true” distribution P, is listed first while the measured distribution Q, is listed second.

Negative log likelihood (Profile Loss)

The ability of information encoded by a sequence to capture the total information in a profile was calculated by measuring the negative log likelihood. For all measurements, the following formula was used:

$$NLL = -\sum_{i=1}^{N} log(P_i^{aa}) \mid S_i \rightarrow aa$$

Eq. 5

Where for all residues i in a profile with length N, P is the profile and S is the sequence of characters at each measurement from that profile. The amino acid, aa, at each S_i is the only value considered from the profile.

Nanohedra score

Nanohedra score was calculated according to the procedure in Laniado et al. 2021 [33] with the modification as the score was only tabulated for central fragment residues, not every fragment residue. For the Nanohedra score normalized, the Nanohedra score was divided by the number of fragment residues. This gives a maximum normalized value of 2.

Interface energy

All default energy terms from the REF2015 score function were calculated for each residue and residue neighbor in the interface and summed to yield various solvation energy measurement states. The interface energy complex is calculated on the full complex, while the bound version is calculated on the bound confirmation in the oligomeric state, and the unbound version is calculated on the minimized confirmation in the oligomeric state. To calculate the interface energy, the unbound is subtracted from the complex (complex - unbound), while to

calculate the bound configuration energy, the bound is subtracted from the complex (complex - bound).

Interface solvation energy

The energy terms, `lk_ball_wtd` and `fa_solv` were calculated for each residue and residue neighbor in the interface and summed to yield various solvation energy measurement states. The solvation energy complex is calculated on the full complex, while the bound version is calculated on the bound confirmation in the oligomeric state, and the unbound version is calculated on the minimized confirmation in the oligomeric state. To calculate the Interface solvation energy, the complex is subtracted from the unbound (unbound - complex), while to calculate the bound configuration energy, the bound is subtracted from the unbound (unbound - bound).

Interface bound configuration energy

To calculate the energy required to attain the atomic conformations utilized during interface complexation from the uncomplexed state (not including the process of complexation with the interface partner), the energy at each residue was captured for both states, the complexed conformation form of the oligomer and the oligomer after four rounds of atomic packing and minimization. Next, the energy difference between the individual residues in the bound conformations and the unbound conformations were taken with positive values indicating energy is needed to assume the bound conformation.

BSA calculation

The atoms participating in the interface are measured for solvent accessible surface area (SASA) in both the complexed state and the uncomplexed state (no repacking). Next, the total SASA in the uncomplex state was subtracted from the SASA in the complex state to find the difference. For segregation of SASA by atomic polarity, only atoms that were deemed polar

had SASA summed, while for hydrophobic BSA only non-polar atoms were summed. All calculations were performed using the program FreeSASA[66].

SS calculation

The program Stride [67] was used to calculate secondary structure for each residue. When secondary structure percentages were assigned to interface portions, the total number of interface residues was used as the denominator and the number of residues of a particular secondary structure type were the numerator

ProteinMPNN scores

ProteinMPNN score was calculated as in Dauparas et al. 2022 [7], where the score was the average over all residues in either the complexed state, i.e. bound interface, or the uncomplexed state, i.e. the interface was separated/. For ProteinMPNN score for a particular subset of residues, only those residues were averaged. The total ProteinMPNN score reflects a total summation of the individual scores for all residues.

Shape complementarity

Calculations were performed using the ShapeComplementarityFilter in Rosetta. During design calculations residues were selected based on 8 Å C-beta C-beta interface residue membership. For calculations performed during retrospective analysis, residues were included if they were interface residues according to BSA involvement.

Buried unsatisfied hydrogen bonds

The number of hydrogen bonds participating in the complexed interface state and the uncomplexed interface state (no repacking) were summed. The number of unsatisfied hydrogen bonds in the uncomplexed state was subtracted from the number of unsatisfied hydrogen bonds in the complexed state to find the difference.

Local distance difference test (LDDT)

The LDDT score [68] was utilized according to the original implementation to compare structures regardless of superposition technique. For cage assemblies, all reported values are the result of calculation of one chain that was perfectly symmetric.

Root mean squared deviation (RMSD)

Calculation was performed using the Kabsch algorithm for finding an optimal overlap and calculating an RMSD.

New hydrophobic collapse sites

The hydrophobic collapse index (HCI)[69] was calculated with the following modifications. Instead of the amino acid types FILV being classified as collapsable residues, the amino acid types were expanded to include FMILYVW. Additionally, the HCI threshold was modified to 0.48 from the reported 0.43 to maintain consistent overlap with observations of collapse [69]. The HCI was taken for both the designed sequence and the reference sequence, i.e. the sequence before sequence design occurred, and regions in the designed sequence which resulted in HCI larger than the HCI threshold, however, that were not larger than the HCI threshold for the reference sequence were determined to be new hydrophobic collapse sites.

Interface composition similarity

The residue burial types in the designed interfaces were calculated according to classification of either core, rim, and support residues. To measure the similarity between designed interfaces and natural interfaces, the number of interface residues of each type were compared to the number of residues expected based on the size of the interface. To find the expected values of each residue burial type, lines of best fit reported in Levy 2010[31], were used to calculate expectation, corresponding to: core = 0.01*BSA + 0.6, rim = 0.01*BSA - 2.5, and support = 0.006*BSA + 5. For the calculated BSA, the difference between the expected number of residues and actual number of residues in each classification, R, was found. Next the

percentage of this difference from the expected value was subtracted from 1. After the expected difference is summed for each residue, the mean value was taken as the interface composition similarity as in equation 4. Values of 1 indicate exact similarity to the expected interface composition values and 0 indicates no similarity.

$$similarity = 1/3 \sum_{R \in core,rim,support} 1 - \frac{abs(R_{expected} - R_{measured})}{R_{expected}}$$

Eq. 6

Spike ratio

The spike ratio is defined here as 1 minus the ratio of two distances, where one distance constitutes the center of mass of one protein component from the center of mass of the cage compared to the same distance in another protein component. In this comparison, the larger distance takes the denominator of the ratio so 1 indicates the center of mass of each component is equidistant from the cage center of mass, and a ratio of 0.5 indicates one component is two times the distance from the cage center of mass as the other.

Errat deviations

Each residue in a pose was subjected to measurement by Errat [43], which measures atomic contact patterns for highly unusual distributions. Any position was considered deviating as judged by an Errat score at the residue greater than 2 standard deviations over defined thresholds.

Pose selection

A weighting scheme was utilized to prioritize poses and designs based on evaluation of metrics calculated for each considered structure/sequence. Those poses or designs which demonstrated the largest weighted sum were selected, where the individual weighting terms are according to a provided weight and the normalized value of that pose/designs metric compared to all others considered.

590 candidate poses and prioritization of their designs for T33-fn characterized sequences

For selection of poses, the metrics, direction of prioritization, and weighting coefficient are as follows: shape complementarity of fragment residues, prioritize higher values, 0.3 weight; interface residue composition similarity, prioritize higher values, 0.4 weight; the percent of interface with fragment observations, prioritize higher values, 0.1 weight; and the percent of hydrophobic BSA, prioritize lower values, 0.2 weight.

For each set of designs from each filtered pose, the weighting scheme was applied similarly to select the best design, with the metrics, direction of prioritization, and weighting coefficient as follows: interface energy, prioritize lower values, 0.25 weight; interface bound configuration energy, prioritize lower values, 0.15 weight; atomic density of non-hydrogen interface atoms (following the local density metric [40]), prioritize higher values, 0.2 weight; the density of buried unsatisfied hydrogen bonds, prioritize lower values, 0.15 weight; shape complementarity of the interface, prioritize higher values, 0.1 weight; and the BSA to SASA ratio of the pose, prioritize lower values, 0.15 weight.

4,241 candidate poses for T33-ml

For T33-ml designs, in depth investigation occurred for 4,241 poses. These were selected according to the following filters. The type of residue selection protocol was all residues, there were no new hydrophobic collapse sites, the fraction of residues that were different amino acids from the original sequence, i.e. mutated positions, were less than 55 %, the ratio of fragment observations at each residue site (multiple fragment ratio) was greater than 2.5, the number of fragment observations at the interface was greater than 20, the total ProteinMPNN score in the unbound state was less than 880, the ProteinMPNN score in the complex state over all designed residues was less than 1, the profile loss of the sequence given the evolutionary profile was less than 2.5 (average of all residue positions), the profile loss of the sequence given the fragment profile was less than 5 (average of all fragment residues),

Finally, shape based features such as the spike ratio less than 0.5, and the distance between the cage center of mass and the minimal atomic distance for each component which is less than three times the distance for the component which is displaced further.

Finally, from the selected poses, one design was selected according to the lowest ProteinMPNN score and characterized through computational structure prediction and threading.

Design protocols

Symmetry

All design methodologies are untaken in the presence of the entire symmetric system in question. For oligomeric state measurements, this was in the trimeric system, while for complex state measurements, either the entire Tetrahedron complex was created, or where applicable, the minimal contacting group of protein chains was constructed which contain the minimal information necessary to fully model all possible interactions present in the complex. All modifications made to the sequence of oligomeric components was maintained at a single protomer representative (the captain) and propagated symmetrically to all identical, but spatially separate, symmetry mates.

Running hhblits

Per-residue hidden markov models were created using amino acid sequences submitted to hhblits [70] against the UniRef30_2020_02 database (retrieved from https://gwdu111.gwdg.de/~compbiol/uniclust/2020_02)/ and run with the following parameters `-ohhm` ***FILENAME.hhm*** `-oa3m` ***FILENAME.a3m*** `-hide_cons -hide_pred -hide_dssp -E 1E-06`.

iAlign clustering

Clustering according to interface alignment in iAlign [71] was performed. Where it was found that there were overlapping poses, the best pose was selected according to the weighting procedure described in 590 candidate poses and prioritization of their designs for T33-fn characterized sequences.

Tertiary constrained FastDesign

Metropolis criteria Monte Carlo optimization was performed using the FastDesign mover in Rosetta. FastDesign is set up with a SequenceProfile TaskOperation which limits the residues available for packing to those residues specified by the provided position specific scoring matrix file [72]. The profile of interest for this protocol was the tertiary profile. Additionally, design (defined as amino acid sampling) is only allowed at interface residue positions, while interface residue neighbors are allowed to pack (defined as sampling of residue conformations such as backbone and rotameric states). Five rounds of FastDesign sampling is performed in the REF2015 score function and the resulting asymmetric unit is written to file.

The implementation of this protocol can be found at *dependencies/rosetta/interface_design/design_profile.xml*.

Scouting FastDesign

Scouting constitutes a quick round of design where minimal residues are selected, few amino acid types are allowed for design, and all FastDesign is subjected to one cycle of design. The first set of residues designed constitutes the interface fragment residues and uses the FastDesign with protocol InterfaceDesign2019. The second set of residues set for design constitutes the remaining interface residues and uses the FastDesign protocol PolarDesign2019.

The implementation of this protocol can be found at *dependencies/rosetta/scout.xml*.

FragHBNet

The combination of HBNet with fragment residue packing prioritizes extensive hydrogen bond networks for their ability to simultaneously support well packed, hydrophobic interactions. The method is a modification to the search procedure of the MCMC HBNet protocol (Maguire 2021). Importantly, the `start_selector` keyword in the HBNet Mover uses a `residue_selector` which includes all observed fragment residues identified for the pose. The residue positions which are available for HBNet inclusion include interface residues (of which fragment residues belong), and interface neighbors. During HBNet search for interface residues utilize the amino acid types available from the union of the fragment and evolutionary profiles, where a frequency above 0 (via the flag `--use_occurence_data true`) is included, while for interface neighbors, only the wild-type amino acid type is available. HBNet proceeds for 50,000 Monte Carlo runs and accepts all found networks less than -0.65 REU, containing three or more residues. After HBNet search, the top 250 networks are input to the MultiplePoseMover which performs one cycle of FastDesign sampling (with REF2015 score function and PolarDesign2019 weights) at FragHBNet residues, defined as an interface residue, an identified hydrogen bond network residue, or a neighbor of a network residue. MCMC sampling at interface residues utilizes the intersection of amino acids available from the fragment and evolutionary profile (`--use_occurence_data true`), while those outside the interface are only allowed to utilize amino acids available from the evolutionary profile (`--use_occurence_data true`). After each candidate network is briefly designed, the entire group of candidate networks is ranked and the top N network candidates are selected. In this work, we used the top 20 network candidates.

Ranking proceeds for the highest N*3 (60 in this work) candidates measured according to FragHBNet residue shape complementarity and the lowest N*3 (60 in this work) measured heterotypic interface interaction energy per-FragHBNet residue. Finally, candidate poses that fall

within both of these rankings criteria are ranked in descending order according to the number of residues included in the FragHBNet. The top N (20 in this work) candidates are subjected to deeper sampling using 5 rounds of constrained FastDesign with HBNet participating residues restricted to repacking with an additional AtomPairConstraintGenerator utilized with `sd`=0.4 to minimize undesired loss of the identified hydrogen bond networks. This constraint was set up to mimic the energetic contributions of hydrogen bonds in REF2015 to allow HBNet residues to flexibly readjust as the remainder of the interface was designed.

The scripts for these design protocols are found at: *dependencies/rosetta/hbnet_scout.xml*, *dependencies/bin/sort_hbnet_silent_file_results.sh*, and *dependencies/rosetta/hbnet_design_profile.xml*.

Reversion criteria

For each residue in a design, reversion utilized a fuzzy prioritization mechanism wherein mutations are made to the original amino acid and measured for their effects [73]. The prioritization filter accepted mutations if they resulted in a decrease in shape complementarity less than 0.02 units and resulting in a negative value for calculated hydrogen bonding energy. Subsequently all passing mutations were ranked according to highest shape complementarity, lowest hydrogen bonding energy, and lowest unsatisfied hydrogen bonds. The best scoring reversions were accepted followed by iterative testing of remaining reversions. Those remaining were combined with prior accepted reversions and again tested for their cumulative impact, accepting if the above filters passed.

The implementation of this protocol can be found at *dependencies/rosetta/optimize.xml*.

ProteinMPNN

For sequence design, residues identified as designable by one of three protocols, all - every residue, interface - interface residues, or interface+neighbors - interface residues and

their neighbors, were specified as positions available for ProteinMPNN sequence inference. All positions and coordinates were symmetrized including symmetrically tying positions and using the ProteinMPNN.tied_sample() method. In protocols where not all residues were used for designs, the remaining positions were input as their wild-type identities into the model and contributed to the encoding and decoding steps. The final sequence and probabilities were trimmed to only the minimal sequence representing the asymmetric unit. Resulting probabilities were transformed to represent an inference profile, representing the per-residue amino acid frequency distribution predicted as a result of inference.

For structure profile creation, symmetrized coordinates were used as inputs and the ProteinMPNN class was called. The resulting log probabilities were transformed to normal log probabilities to create the ProteinMPNN structure profile, i.e. the per-residue amino acid frequency distribution predicted for the protein coordinates alone.

Rosetta refinement

To ensure that Rosetta energy calculations were carried out as accurately as possible, all designs were relaxed into Rosetta before metrics were acquired. Designs were input as an asymmetric unit, with symmetry set using symmetry definition files [74]. Symmetric refinement was performed into the REF2015 score function [38] using the FastRelax mover. Additionally, the suggested flags for pareto-optimal refinement [64] were included for five rounds of refinement including the flags:

```
-relax:ramp_constraints false
-no_optH false
-relax:coord_cst_stdev 0.5
-nblist_autoupdate true
-relax:bb_move false
-constrain_relax_to_start_coords
```

```
-use_input_sc
-relax:coord_constrain_sidechains
-flip_HNQ
-no_his_his_pairE
```

The implementation of this protocol can be found at *dependencies/rosetta/refine.xml*.

Threading of ProteinMPNN sequences to the designed structure

For structural metric measurement of ProteinMPNN designed sequences, the amino acid identities at each residue position were mutated to the designed ProteinMPNN sequence. Threaded designs were then refined in Rosetta following the Rosetta refinement procedure. For threading however, refinement only used one cycle of refinement to sample the structural state.

The implementation of this protocol can be found at *dependencies/rosetta/refine.xml*.

Refinement for structural analysis

For retrospective analysis of cage designs, all design models were subjected to an additional iteration of Rosetta refinement. In most cases, designs should be minimized after one round of Rosetta refinement. In T33-ml designs however, a higher fraction of amino acids were mutated and thus backbone rearrangements were expected. As a consequence, this extra refinement ensured further convergence to minimum energies.

Structure Prediction

AlphaFoldInitialGuess inference

For all inference performed using the AlphaFoldInitialGuess model, the amino acid sequence was processed into arrays corresponding to AlphaFold sequence features, and structure coordinates were processed into arrays to be input as coordinate positions to the AlphaFoldInitialGuess class which was derived from descriptions provided in Bennett et al. 2023 [48]. AlphaFold msa features were provided empty to the AlphaFold feature dictionary as described in [75]. All predictions were performed using AlphaFold multimer with the model

parameters `multimer_v3`. Residues that were missing in their entirety or had side chain atoms missing had atomic coordinates initialized at the origin (i.e. 0,0,0). The structure coordinates were then input into the `prev_pos` feature to emulate a prior round of model inference, at which point AlphaFold was run having assumed these starting coordinates were a result of a prior round of prediction. Predictions preceded using for all five of the provided AlphaFold multimer model parameters unless satisfactory confidence metrics were achieved (pLDDT ≥ 85), at which point prediction was terminated. After model inference, the most confident model was refined into the Amber scorefunction.

Biochemical characterization

Protein expression

Plasmids containing two genes of interest, but constituting a single design, were acquired from Twist Biosciences. Each plasmid contains a Kanamycin resistance gene for selection as well as a pET-Duet type interspatial expression cassette cloned between the genes of interest to enable bicistronic expression of the two proteins under the control of a LacO inducible T7 promoter. Plasmids were transformed into LOBSTR BL21(DE3)-RIL competent cells (Kerafast - EC1002) and always supplemented with 50 µg/ml of Kanamycin during growth. First, cells were grown to saturation overnight in 20 ml of PG media [76]. The following day, 10 ml of saturated culture was used to inoculate 1 L ZYM-5052 cultures [76] which was left to grow for either ~66 hours at 18°C, ~42 hours at 25°C, or 16 hours at 37°C. Cells were harvested by centrifugation at 4,000 xg for 10-15 minutes and quickly moved to freezing conditions at approximately -80°C. Pellets were transferred to -20°C after 24 hours at -80°C.

Immobilized metal affinity chromatography (IMAC)

Purification of proteins from cell pellets was performed with a 1:4 ratio of grams of pellet to milliliters of purification buffer (i.e. 10 g : 40 ml). Purification buffer contained 50 mM Tris-HCl, pH 7.5 at 4°C, 300 mM KCl, and 30 mM imidazole, while lysis buffer (utilized during cell lysis)

was supplemented with 0.5 mM EDTA, 300 µg/ml lysozyme, 100 µg/ml benzonase nuclease, and one tablet of protease inhibitor cocktail EDTA-free (Thermo Scientific)/50 ml of cellular suspension. Pellets were resuspended in lysis buffer with end over end rotation for ~30 minutes at 4°C until no visible chunks of cell pellet remained. Lysis of cellular resuspension proceeded with three passes of high pressure homogenization in a EmulsiFlex C3 homogenizer (Avestin) pressurized to > 15,000 psi. Lysate was then clarified using centrifugation at 20,000 xg for 30 minutes or 10,000 xg for 45 minutes at 4°C. The soluble supernatant fraction was collected and either incubated with 2 ml Ni-NTA resin/50 ml of soluble lysate (batch binding) or applied to a 5 ml HisTrap pre-equilibrated with purification buffer.

For batch binding, side over side rotation of the soluble lysate and Ni-NTA resin proceeded for 30 minutes at 4°C, after which the solution was allowed to stand for five minutes at 4°C, enough to separate the Ni-NTA resin and the supernatant by gravitational forces. Then, the supernatant was removed and the remaining Ni-NTA resin was collected and loaded into a filtered column where remaining soluble lysate was cleared from the Ni-NTA resin by gravity. The Ni-NTA column was washed by gravity flow with 10 column volumes (CV) of purification buffer supplemented to 60 mM imidazole (wash buffer). Finally, protein of interest was eluted from the resin with gravity flow using 9 CV of purification buffer supplemented with 250 mM imidazole (elution buffer).

For HisTrap binding, the soluble lysate was applied to the 5 ml column at 1 ml/min until all soluble lysate had passed through the column. Next, 10 CV of purification buffer were applied, followed by 2 CV of wash buffer and then a gradient from 60 mM imidazole to 300 mM imidazole (elution buffer concentration) was carried out for 3 CV. Finally, 10 CV of elution buffer was passed over the column.

Fractions corresponding to the lysate, the clarified lysate, the insoluble lysate, the unbound Ni-NTA supernatant (aka the flow through), the wash, and the elution were analyzed

by SDS-PAGE for bands at the hypothesized size for the proteins of interest and classified according to one of insoluble, soluble, or co-eluting based on the presence of both bands.

Size Exclusion Chromatography (SEC)

After IMAC purification, fractions containing proteins of interest were concentrated using Amicon Ultra 10 kDa MWCO concentrators until the sample reached approximately 500 µl volume or precipitation was visibly observed at which point concentration was immediately stopped. After concentration, the sample was clarified by centrifugation for 10 minutes at 16,900 xg to remove any precipitation from soluble protein. All chromatography was performed using the BioRad NGC HPLC system at 4°C. Analytical chromatography was performed using a 10/300 gl size column with either Superose6 Increase or Superdex200 Increase resin with flow at 0.3 ml/minute. For preparative scale chromatography, 16/60 scale columns with either Superose6 or Superdex200 residue was used while flowing at 0.9 ml/minute. Fractions corresponding to peaks in the chromatogram, observed by absorbance at 230/280 nm, were analyzed by SDS-PAGE for protein species of interest.

CryoEM specimen preparation

Purified samples of the designed assemblies were removed directly from SEC fractions if the fractions were deemed sufficiently concentrated (>0.5 mg/ml) otherwise were concentrated to ~1 mg/ml and stored at 4°C until specimen preparation. Upon freezing, QUANTIFOIL R 2/1 Cu 300 mesh grids (Cat # Q3100-CR1) were glow discharged for 30 seconds with a Pelco easiGlow glow discharger using 15 mA and a negative polarity. Immediately after glow discharge, grids were loaded into a Vitrobot Mark IV Thermo Fisher Scientific set to 100% humidity and blotted with 595 Filter Paper 55/20mm (Ted Pella Cat# 47000-100) using 1 total blot with a 0 second wait time, -4 blot force, 4 second blot time, and 0 second drain time and vitrified into liquid ethane. All specimens were prepared using a protective plastic facemask and a surgical mask to avoid contamination.

The sample for T33-ml30/T33-ml35 mixed were measured with absorbance of 0.583 as measured directly from SEC were frozen in SEC buffer consisting of 20 mM Tris-HCl pH 7.5 at 4°C, 150 mM KCl. The design T33-ml28 was diluted to 0.31 mg/ml using 20 mM Tris-HCl pH 8 at 214°C and 100 mM NaCl. The design T33-ml23 was diluted to 0.6 mg/ml using 20 mM Tris-HCl pH 8 at 21°C and 100 mM NaCl.

Cryo-EM data acquisition and processing

Cryo-EM data were collected on Titan Krios cryo-electron microscopes equipped with a Gatan K3 Summit direct electron detector. Movies were recorded with SerialEM [77] at a nominal magnification of 81,000× (calibrated pixel size of 1.1 Å per pixel), 105,000× (calibrated pixel size of 0.86 Å per pixel) and 130,000x (calibrated pixel size of 0.65 Å per pixel), over a defocus range of −0.5 to −2.5 µm and a total dose of 40 $e^{-}/Å^{2}$.

Motion correction, CTF estimation, particle picking, particle extraction, 2D classification, and additional data processing were performed with cryoSPARC [78]. An initial set of particles was automatically picked through the blob-picker method, extracted and 2D classified. Particles selected from 2D classes were used for ab initio reconstruction. This reconstruction was then used for the 3D refinements enforcing T symmetry. The 3D structure was used to generate 2D projections of the particles and then used to repick the particles from the images using a template picker. The re-picked particles were extracted from the micrographs, 2D classified and went through 3D refinements enforcing T symmetry. Particles were further classified using heterogeneous refinement and the best classes were used for 3D refinements enforcing T symmetry. For the T33-ml23, we obtained an overall resolution of 2.0 Å, based on an FSC threshold of 0.143. For the T33-ml28, we obtained an overall resolution of 2.7 Å, for T33-ml30 the resolution was 4.2 Å and for T33-ml35, 2.9 Å. Subsets of particles missing one trimer were isolated from the heterogeneous refinement step. 3D refinement with C1 symmetry of these particles missing a trimer resulted in cryo-EM maps for T33-ml23 at 3.9 Å and for T33-ml35, at

4.4 Å. Model building was performed using Coot [79], and automated refinement was performed using Phenix [80]. Figures were prepared using ChimeraX [81,82].

Data processing

Molecular replacement and refinement of T330fn10 in phaser

Molecular replacement was needed to solve for the phases of the T33-fn10 crystal. Search was carried out in phaser [83] using the computationally designed A:B, asymmetric unit model of the cage. Initial hits suggested the A:B model was present, however, low agreement was apparent. Models with increasing copy numbers of A:B yielded improved results with the model containing eight copies of each of A and B. Upon further inspections, pores in the lattice and omit maps indicated that additional molecules were present in the lattice. Another copy of the tetrahedral model was manually placed in the lattice and oriented in a non-clashing orientation. A final round of molecular replacement with the new model indicated as mobile and the old model fixed improved the fit even further as judged by omit maps without significant features.

Refinement of the initial molecular replacement model was carried out using phenix [80]. The refinement parameters that yielded the best fit with the smallest deviation between R_{work} and R_{free} included one round of rigid body refinement with grouped ADP. Next, all B-factors were manually set to 200 and another round of rigid body refinement was performed using a reference model as a restraint and subsequently refining by tls and grouped ADP. This resulted in an R_{work} of 0.22 and R_{free} of 0.25.

Cryo-EM map refinement

All maps were refined utilizing their corresponding design assembly model as the search model and the highest resolution map achieved from CryoEM data processing in phenix.real_space_refine [84]. First, the refine method `run=rigid_body` was used with `resolution`=4.2, `weight`=4, `ncs_constraints`=True,

ramachandran_restraints=True, c_beta_restraints=True, target_bonds_rmsd=0.02, and target_angles_rmsd=2.0. Rigid body refinement was carried out for five macrocycles, and in some cases longer until the measured CC converged when the design and model had large deviations. Next, run=minimization_global+local_grid_search+morphing+adp was used to refine individual atomic positions with the same flags as for rigid_body. After the first round of individual atomic parameters were refined, the model and map were opened for manual inspection in coot [85]. Deviating side chains and backbone segments identified by visual inspection and validation tools were refined into the density with the Real Space Refine Zone, and model termini were built or removed depending on additional sequence added during formatting sequence for expression. Finally edits in the captain NCS chain were propagated to all mate chains symmetrically and the file was saved for additional rounds of individual atomic refinement.

In most cases, after rigid body phenix refinement, each chain demonstrated asymmetry when superimposed on symmetrically related copies and thus asymmetrically fit into the density. Extra measures were taken to find the best fitting individual chains, one for each protein entity, and then the representative chains were manually symmetrized to fit in the map. This procedure improved the overall fit and allowed individual atomic site modifications to reach the highest obtainable CC.

SEC-SAXS

Initial processing of raw images proceeding according to the automated processing suite from SasTool which includes creating images with background subtraction, the calculation of 1D scattering intensity profiles, and radius of gyration measurements. Subsequent processing was performed to analyze the experimental results in the context of structural models. First, UV absorbance from SEC was used to identify peaks of interest for cage species. Next, SAXS

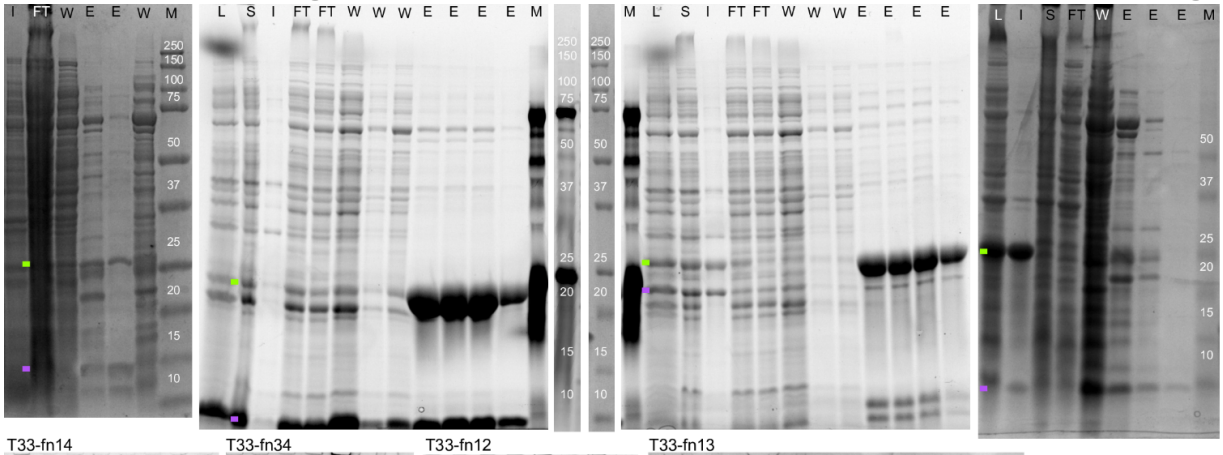
images representing radial scattering intensities from the central fractions of the corresponding SEC peaks were analyzed using PRIMUS from the ATSAS Data analysis software (version 3.2.1) [86]. The frames of interest were averaged together using the average tool to create average scattering intensities over the identified peak area. To calculate theoretical scattering profiles, the CRYOSOL model evaluation tool was used to convert .pdb files to scattering curves. Theoretical curves were superimposed on the scattering profile of the experimental frames and the scattering values were exported.

Supplementary Materials

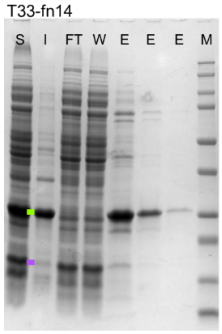
A Suite of Designed Protein Cages Using Machine Learning Algorithms and Protein Fragment-Based Protocols

Kyle Meador, Roger Castells-Graells, Roman Aguirre, Michael R. Sawaya, Mark A. Arbing, Trent Sherman, Chethaka Senarathne, Todd O. Yeates

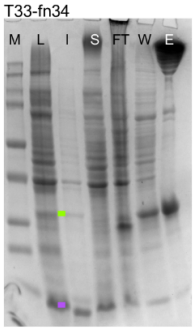
Supplemental Figure 5.1. Purification of T33-fn designs.



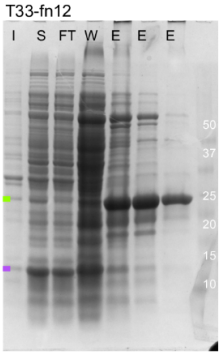
T33-fn14



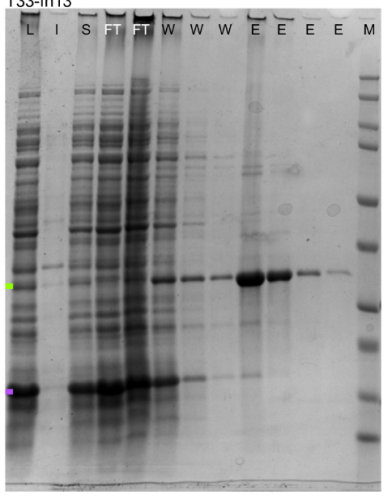
T33-fn34



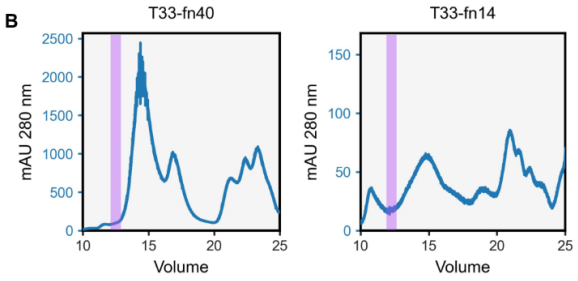
T33-fn12



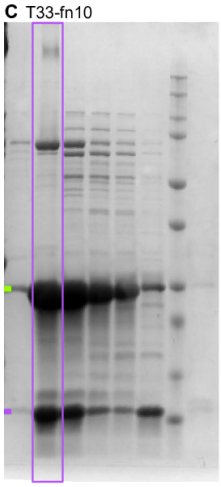
T33-fn13

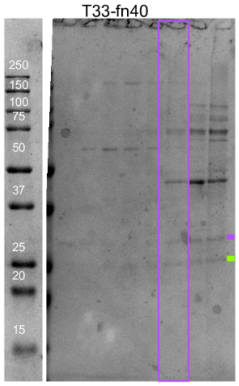


B



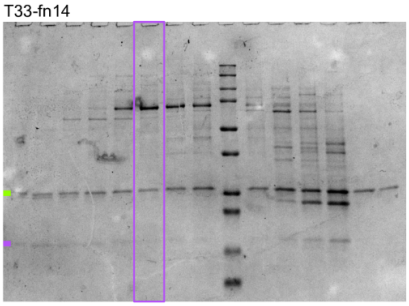
C T33-fn10





T33-fn40

T33-fn14



a) IMAC purification gels for selected designs. Samples are presented to indicate the extent of solubility after centrifugation where clarification leaves insoluble material as precipitation. Further, purification reveals the extent to which species co-elute from IMAC. Each gel is labeled according to the following possible labels to indicate the time point in which it was taken. M - molecular weight marker, L - lysate, I - insoluble lysate fraction, S - soluble lysate fraction, FT - soluble lysate after flowing through a Ni-NTA column, W - wash of the Ni-NTA column with 60 mM imidazole, E - elution of the Ni-NTA column with 250 mM imidazole. Multiple fractions indicate time resolved fractionation of the indicated separation type. For each gel, the green bar indicates the experiment size of the component with a His Tag, while the purple bar indicates the experimental size of the second component. **b)** Chromatograms from SEC for selected T33-fn designs. The expected assembly size is overlaid in purple on top of the chromatogram. **c)** SDS-PAGE gels corresponding to SEC runs presented in Figure 5.2. The design T33-fn40 indicates that multiple different species are present at fractions which are larger than trimers. For the designs T33-fn40 and T33-fn14, chromatograms and gels indicate species of assembled cages, intermediates, and trimeric species are present

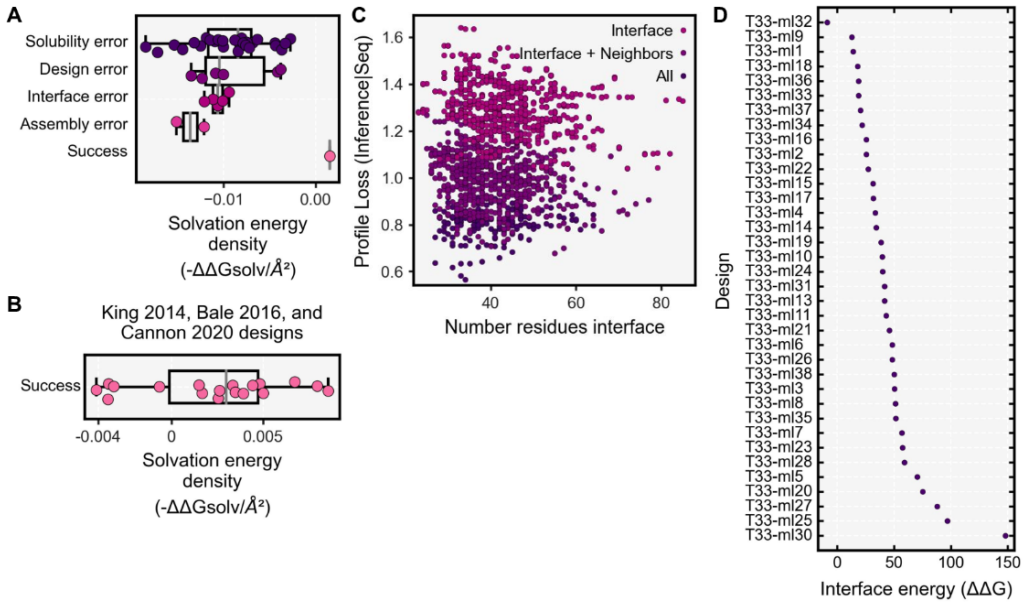
Alternative buried surface area calculations for T33-fn design filtering and selection

During T33-fn design calculations, programmatic assessment of buried surface area (BSA) was performed with the Rosetta SimpleMetrics based `PerResidueSasaMetric` which measures the solvent accessible surface area (SASA) on a per-residue basis. To define which residues which should be measured, `ResidueSelectors` including all interface residues (specified using a pair of `Index ResidueSelectors`, one for each side of the interface) were used. The measurement was performed on all residues in the unbound state as well as the bound state. To calculate BSA, for each of the two states the SASA values were summed across all residues in the state and next by subtracting the bound state from the unbound state. This procedure is analogous to the procedure described in the methods, with the only modification being the programmatic source used to calculate SASA values.

During retrospective analysis of T33-fn designs, it became apparent that the `PerResidueSasaMetric` based method was systematically lower than analogous measurements made using PyMol, FreeSASA[66], and in comparison with values reported previously in the literature for interfaces measured using this technique. We estimate the different calculations resulted in approximately 2x lower BSA using the

`PerResidueSasaMetric` than actually exists. In the text, when BSA is used to describe a filtering or selection step, the BSA values for T33-fn utilized the `PerResidueSasaMetric`, while for T33-ml designs, filtering utilized the FreeSASA based calculation as described in the methods. On the other hand, for all values reported in retrospective calculations such as Figure 5.2, 5.3 and 5.6, and Supplemental Figure 5.2 and 5.7, we utilized the FreeSasa based method to perform analysis and report results.

Supplemental Figure 5.2. Design calculations of interface free energy and ProteinMPNN score for design models.



a) The calculated solvation free energy density values for T33-fn designs are mostly negative while the only design that demonstrates positive solvation energy, *i.e.* solvation is not favored, was T33-fn10, the best characterized design. **b)** For prior successful designs [41,56,57], calculated solvation free energy density tends to be positive, while designs successfully form with negative calculated values. **c)** The ProteinMPNN score is plotted versus the number of interface residues within 8 Å for each pose. Each datapoint for the ProteinMPNN score is indicated with the color corresponding to the design protocol that was used for the sequence inference. **d)** The interface energy was calculated for each T33-ml design and sorted according to increasing values. All designs have a positive calculated interface energy.

PDB EntityID codes used for trimeric building block docking

Each EntityID was parsed for the EntryID. Either the QSBio high or very high assembly, or the PDB biological assembly 1 was used. All codes listed below were vetted for author defined assembly ensuring human annotation was utilized if QSBio annotations were not available.

T33-fn designs:

1df4_1,1ekq_1,1gu9_1,1hfo_1,1hwu_1,1ihc_1,1j2v_1,1lu9_1,1mr7_1,1nq3_1,
1o5j_1,1ode_1,1osc_1,1pd5_1,1v4n_1,1vfj_1,1vhc_1,1wdg_1,1wvt_1,1x25_1,
1zoi_1,1zvb_1,2cvl_1,2dj6_1,2ej5_1,2gdg_1,2gtr_1,2i9d_1,2pbq_1,2pd2_1,
2q0t_1,2rfr_1,2v81_1,2vky_1,2vx2_1,2yo2_1,2zhy_1,3cp1_1,3fwt_1,3fwu_1,
3hrx_1,3i82_1,3jv1_1,3l3s_1,3l7q_1,3l8r_1,3m6n_1,3mf7_1,3mjz_1,3n4h_1,
3oi9_1,3pac_1,3q98_1,3qc7_1,3qv0_1,3r0p_1,3tcr_1,3tf3_1,3wfv_1,3zmf_1,
3zo8_1,4d8m_1,4f47_1,4g9q_1,4glf_1,4gvr_1,4i61_1,4jcu_1,4jm7_1,4k2n_1,
4k3w_1,4kd6_1,4knp_1,4lk5_1,4m1a_1,4mi2_1,4mod_1,4myl_1,4n72_1,4nkj_1,
4o8u_1,4ogg_1,4wcz_1,4xcw_1

T33-ml designs:

1avq_1,1ca4_1,1dun_1,1gr3_1,1h9m_1,1hg4_1,1hl7_1,1j3l_1,1jlj_1,1jxz_1,
1k4m_1,1khx_1,1kr4_1,1lw1_1,1m68_1,1mvn_1,1mww_1,1n2m_1,1nkv_1,1nxj_1,
1og6_1,1otg_1,1p32_1,1pf5_1,1qre_1,1qwg_1,1rhy_1,1rj8_1,1s55_1,1sed_1,
1sg4_1,1td4_1,1u5z_1,1ui9_1,1v4n_1,1v6h_1,1v9o_1,1viy_1,1vl0_1,1w15_1,
1wck_1,1wp8_1,1wy1_1,1xho_1,1xx4_1,1ygs_1,1yox_1,1yq6_1,1yqf_1,1yx1_1,
1zcl_1,2a5z_1,2a7k_1,2ar3_1,2arh_1,2bcm_1,2c5q_1,2chc_1,2cu5_1,2d4n_1,
2dtt_1,2dyy_1,2ekm_1,2fbm_1,2flz_1,2fvh_1,2gvh_1,2h6l_1,2hqu_1,2i9d_1,
2ig8_1,2is8_1,2nmu_1,2ose_1,2p6h_1,2p6y_1,2pc5_1,2pd2_1,2q0t_1,2q35_1,
2q6o_1,2qg8_1,2qlp_1,2qs7_1,2r3u_1,2re9_1,2v2d_1,2ves_1,2vhe_1,2vx2_1,
2w5p_1,2wq4_1,2wr8_1,2x4j_1,2y77_1,2yw4_1,2yzj_1,2z5w_1,2zfh_1,3a76_1,
3ahp_1,3b64_1,3b8l_1,3c19_1,3c6v_1,3cyo_1,3dli_1,3dzv_1,3e99_1,3eat_1,
3eby_1,3ef8_1,3ejv_1,3fbq_1,3fd9_1,3fsc_1,3fuy_1,3fwu_1,3gdc_1,3gkb_1,
3gmj_1,3h5i_1,3hpd_1,3hrx_1,3htn_1,3hwu_1,3hyt_1,3i2b_1,3i3f_1,3i3u_1,
3i7t_1,3ifv_1,3io0_1,3irs_1,3jqy_1,3jv1_1,3k1s_1,3k93_1,3k9a_1,3ke5_1,
3kjk_1,3kwe_1,3l39_1,3l88_1,3l8r_1,3lao_1,3lke_1,3mae_1,3mc3_1,3mc4_1,
3mqh_1,3n79_1,3nhv_1,3nke_1,3o3w_1,3quw_1,3qv0_1,3r8y_1,3rf5_1,3soz_1,
3sti_1,3t5s_1,3ta4_1,3tcr_1,3tdt_1,3tjo_1,3tqf_1,3tyj_1,3ub1_1,3vbp_1,
3vnp_1,3wfv_1,3wia_1,3wv7_1,3x2y_1,3ziw_1,3zjb_1,4ad9_1,4b6r_1,4bfc_1,
4c82_1,4di1_1,4e38_1,4emh_1,4fur_1,4g9q_1,4gdz_1,4h6c_1,4hc8_1,4i61_1,
4isx_1,4iyq_1,4jcu_1,4jdn_1,4jf3_1,4jgs_1,4jj9_1,4jm7_1,4jpr_1,4jqs_1,
4kg8_1,4ki3_1,4kw2_1,4lho_1,4liy_1,4m17_1,4mej_1,4myo_1,4nrd_1,4nsm_1,
4ous_1,4r7t_1,4rfu_1,4tzu_1,4u5r_1,4uof_1,4usi_1,4wia_1,4wk3_1,4wrb_1,
4x3n_1,4xc5_1,4xcw_1,4xl8_1,4xqa_1,4y2l_1,4y6i_1,4you_1,5b2f_1,5bmo_1,

5cxd_1,5dii_1,5ds7_1,5ect_1,5eur_1,5fus_1,5h5p_1,5ha6_1,5hlj_1,5hpq_1,
5hrz_1,5ht7_1,5izs_1,5joq_1,5jru_1,5k21_1,5ka5_1,5kvb_1,5m62_1,5mvo_1,
5nz2_1,5o34_1,5ucq_1,5uif_1,5un0_1,5v13_1,5vjy_1,5wfg_1,5xum_1,5y5q_1,
5ycq_1,5yhu_1,5z1q_1,5z81_1,6ard_1,6as5_1,6bj7_1,6cuq_1,6cv6_1,6gdx_1,
6its_1,6ive_1,6j3m_1,6l8p_1,6ln3_1,6lnl_1,6lr3_1,6mhh_1,6mmq_1,6ny9_1,
6ood_1,6p7l_1,6p7o_1,6pnz_1,6qbw_1,6r5z_1,6t76_1,6tj2_1,6tjc_1,6ty6_1,
6u66_1,6u9c_1,6veh_1,6vvr_1,6vvw_1,6vw4_1,6we5_1,6wmg_1,6x7q_1,6xi6_1,
6xt4_1,6zmg_1,6zzm_1,7alg_1,7c1i_1,7cp2_1,7dda_1,7dph_1,7dsz_1,7fe6_1,
7kd9_1,7l7w_1,7m58_1,7ms9_1,7o1e_1,7o45_1,7o4z_1,7obj_1,7okc_1,7p4v_1,
7pkw_1,7qrr_1,7r1m_1,7rgv_1,7s45_1,7std_1,7tbp_1,7te3_1,7uuo_1,7ywf_1,
8abw_1,8bro_1,8del_1,8e62_1

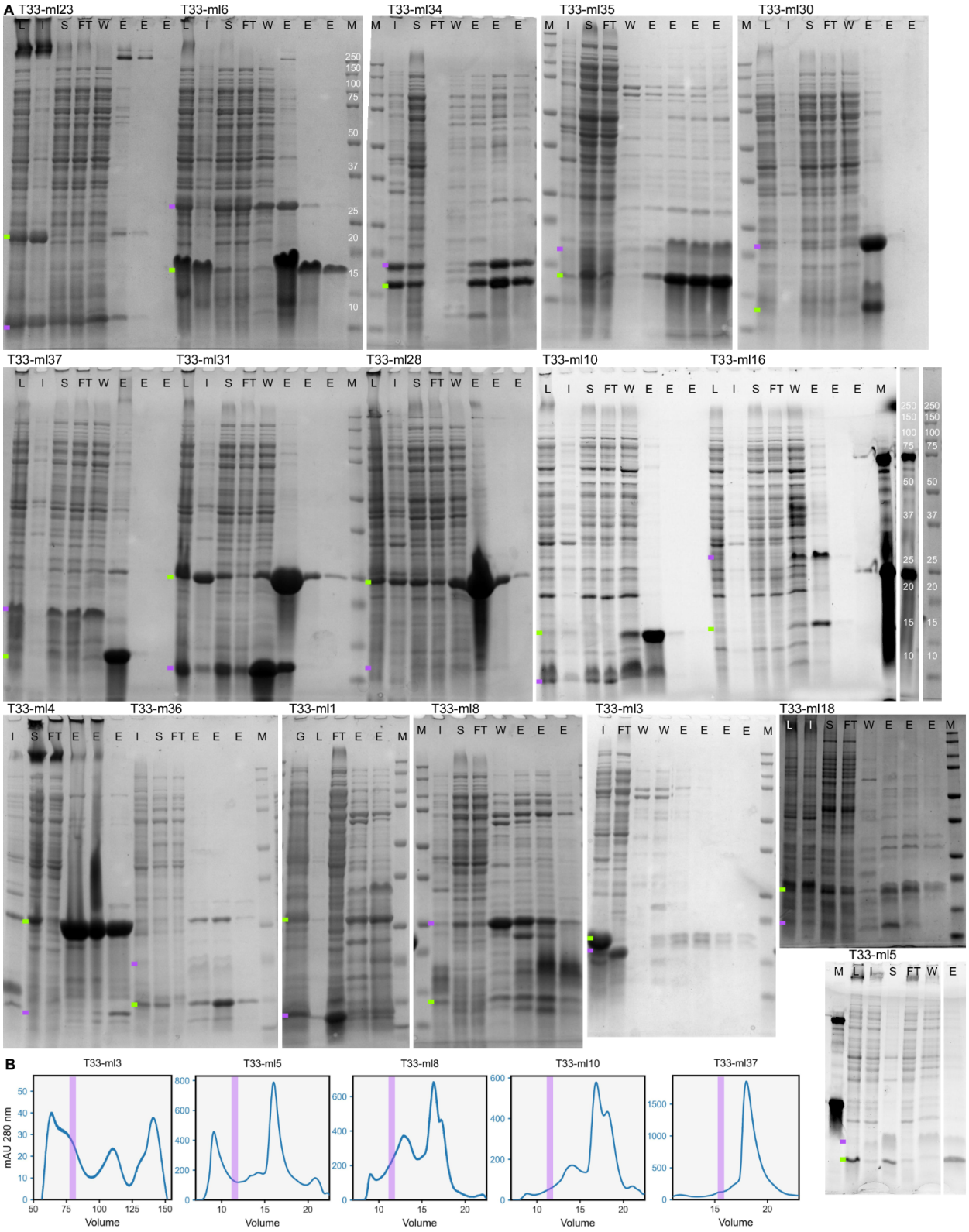
AlphaFoldInitialGuess trimeric predictions

Passing:
2zfh_1,3fwu_1,3k9a_1,3l39_1,5izs_1,6pnz_1,6vvw_1,7obj_1,8bro_1

Mixed:
1lw1_1,1m68_1,1mvn_1,1mww_1,1n2m_1,1nkv_1,1otg_1,1p32_1,1qwg_1,1sg4_1,
1v4n_1,1v6h_1,1v9o_1,1wp8_1,1xx4_1,1yqf_1,2dtt_1,2dyy_1,2ekm_1,2flz_1,
2fvh_1,2h6l_1,2i9d_1,2nmu_1,2p6h_1,2pd2_1,2q35_1,2qs7_1,2vx2_1,2y77_1,
2yw4_1,3ahp_1,3b64_1,3c6v_1,3cyo_1,3dli_1,3dzv_1,3ef8_1,3ejv_1,3h5i_1,
3hpd_1,3i2b_1,3i3f_1,3i3u_1,3i7t_1,3io0_1,3jv1_1,3ke5_1,3kjk_1,3mc3_1,
3n79_1,3nhv_1,3nke_1,3quw_1,3qv0_1,3rf5_1,3tcr_1,3wfv_1,3x2y_1,4bfc_1,
4c82_1,4g9q_1,4i61_1,4iyq_1,4jf3_1,4jgs_1,4jpr_1,4jqs_1,4ki3_1,4lho_1,
4nsm_1,4usi_1,4wrb_1,4xcw_1,4y6i_1,5bmo_1,5cxd_1,5dii_1,5ds7_1,5eur_1,
5ha6_1,5hrz_1,5joq_1,5ka5_1,5kvb_1,5uif_1,5z1q_1,6as5_1,6bj7_1,6cuq_1,
6gdx_1,6j3m_1,6l8p_1,6ln3_1,6lr3_1,6mhh_1,6mmq_1,6qbw_1,6t76_1,6veh_1,
6vvr_1,6vw4_1,6x7q_1,7m58_1,7ms9_1,7te3_1,8del_1

Failing:
1avq_1,1hl7_1,1j3l_1,1jlj_1,1jxz_1,1khx_1,1og6_1,1pf5_1,1rhy_1,1ui9_1,
1viy_1,1vl0_1,1yox_1,1yx1_1,1zcl_1,2a7k_1,2ar3_1,2c5q_1,2ig8_1,2is8_1,
2qlp_1,2ves_1,2vhe_1,3e99_1,3eby_1,3fsc_1,3gkb_1,3gmj_1,3hrx_1,3hyt_1,
3irs_1,3k93_1,3kwe_1,3lao_1,3lke_1,3mae_1,3mc4_1,3o3w_1,3r8y_1,3soz_1,
3tqf_1,3ub1_1,3vbp_1,3vnp_1,3wv7_1,3zjb_1,4b6r_1,4di1_1,4gdz_1,4isx_1,
4kw2_1,4m17_1,4mej_1,4myo_1,4nrd_1,4r7t_1,4rfu_1,4uof_1,4wia_1,4wk3_1,
5b2f_1,5fus_1,5jru_1,5m62_1,5o34_1,5ucq_1,5un0_1,5v13_1,5vjy_1,5wfg_1,
5xum_1,5z81_1,6cv6_1,6its_1,6ive_1,6lnl_1,6ny9_1,6p7l_1,6p7o_1,6tj2_1,
6ty6_1,6we5_1,6wmg_1,6zzm_1,7c1i_1,7cp2_1,7dsz_1,7l7w_1,7o45_1,7okc_1,
7rgv_1,7std_1,7tbp_1

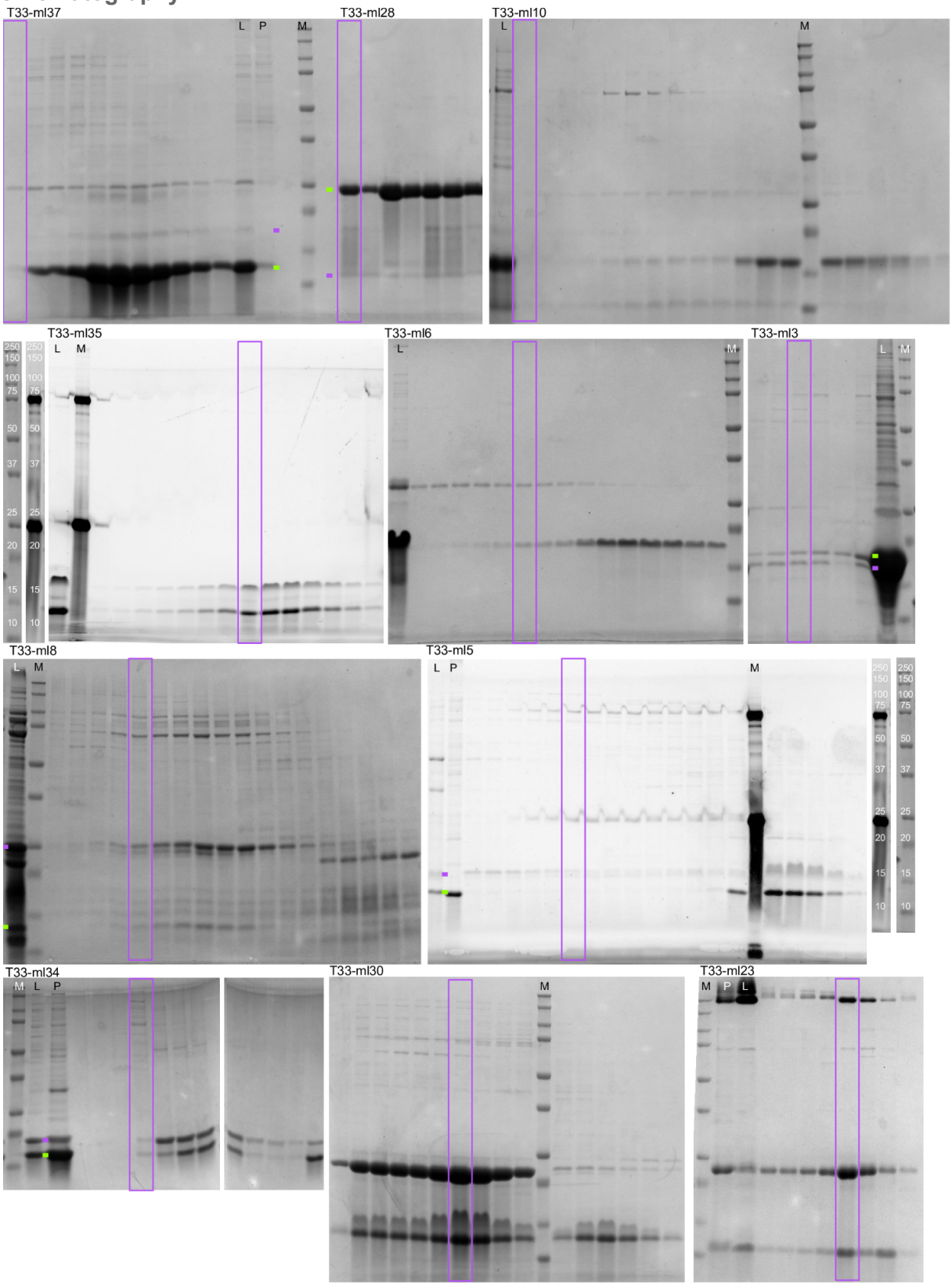
Supplemental Figure 5.3. Immobilized metal affinity chromatography of T33-ml designs.



a) IMAC purification gels for selected designs. Samples are presented to indicate the extent of solubility after centrifugation where clarification leaves insoluble material as precipitation.

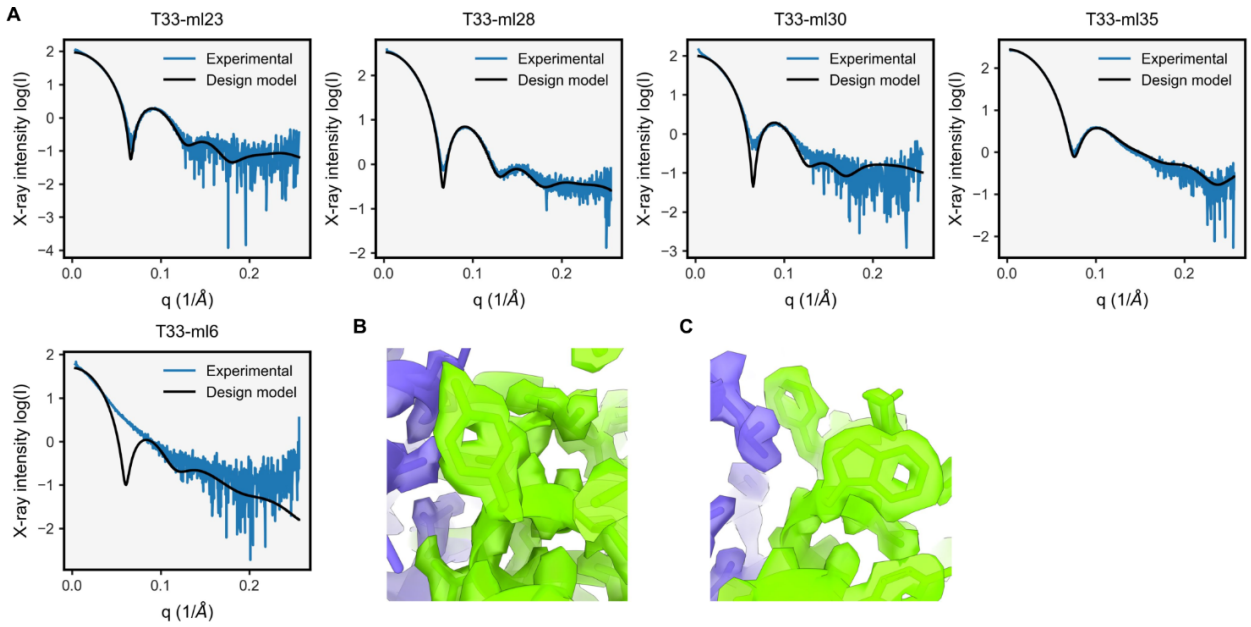
Further, purification reveals the extent to which species co-elute from IMAC. Each gel is labeled according to the following possible labels to indicate the time point in which it was taken. M - molecular weight marker, L - lysate, I - insoluble lysate fraction, S - soluble lysate fraction, FT - soluble lysate after flowing through a Ni-NTA column, W - wash of the Ni-NTA column with 60 mM imidazole, E - elution of the Ni-NTA column with 250 mM imidazole. Multiple fractions indicate time resolved fractionation of the indicated separation type. For each gel, the green bar indicates the experiment size of the component with a His Tag, while the purple bar indicates the experimental size of the second component. **b)** Chromatograms from SEC for selected designs classified as an assembly error. The expected assembly size (~11-12 ml) is overlaid in purple on top of the chromatogram.

Supplemental Figure 5.4. Separation of T33-ml assemblies using size exclusion chromatography.



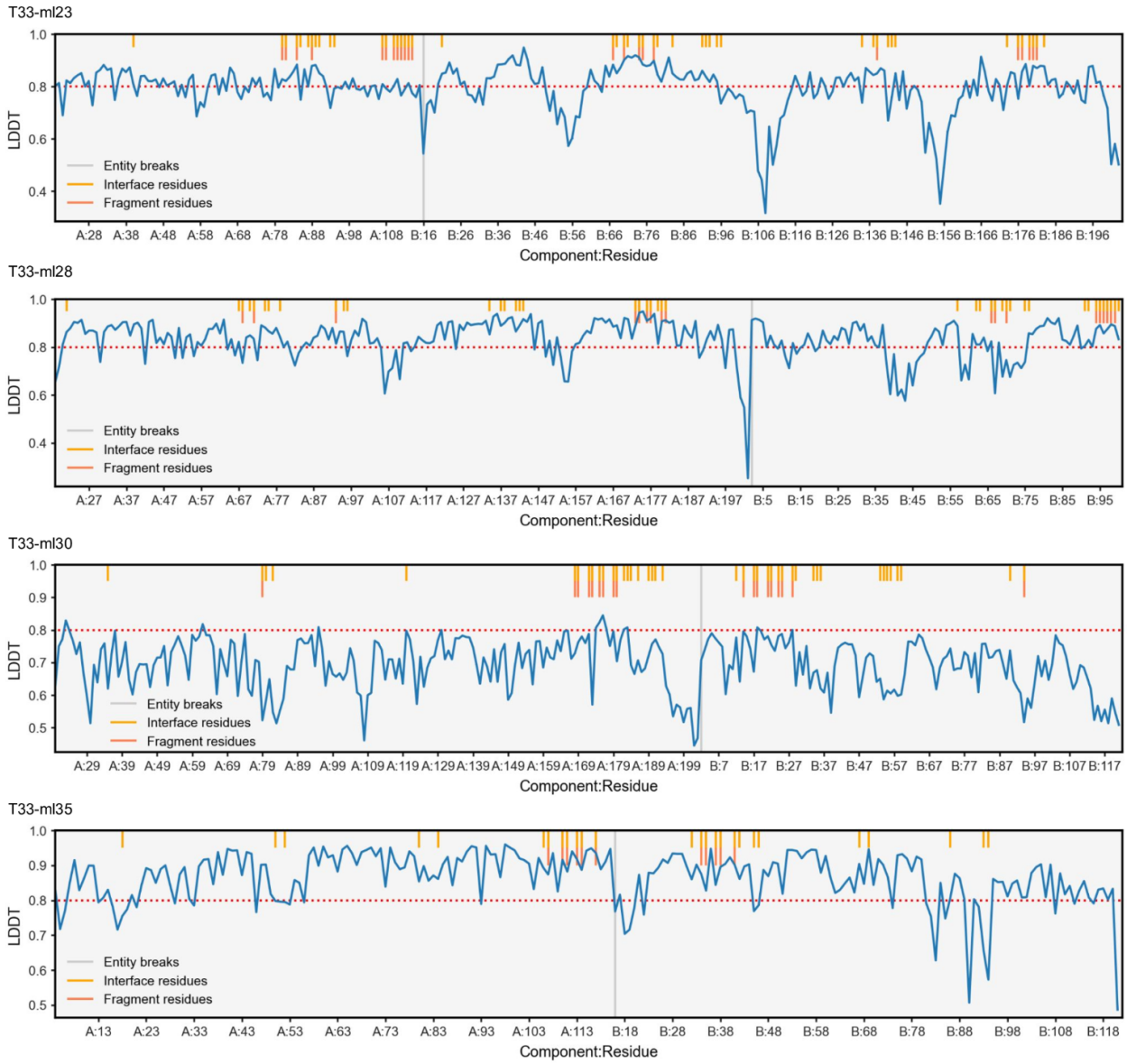
a) SDS-PAGE gels corresponding to SEC runs presented in Figure 5.4a and Figure S5.3b. The expected assembly fraction (~11-12 ml) is overlaid in purple on top of each gel and lanes are indicated as: M - with a molecular weight marker, L - protein loaded to SEC, P - precipitation during protein concentration. The gel of T33-ml34 has been stitched together to reflect the ordering of the lanes in the SEC experiment. Despite robust co-elution, SEC results in many species, from assembled cages, trimeric species bound together and even monomers which are unbound. **b)** The designs T33-ml3, T33-ml5, T33-ml8, T33-ml10, and T33-ml37 mostly elute from the size exclusion at the incorrect fraction, which is indicative of *assembly errors* that prevent full assembled cages from forming, while resulting in larger assemblies as a result of protein complexation.

Supplemental Figure 5.5. Validation of design models using small angle x-ray scattering and cryo-EM density.



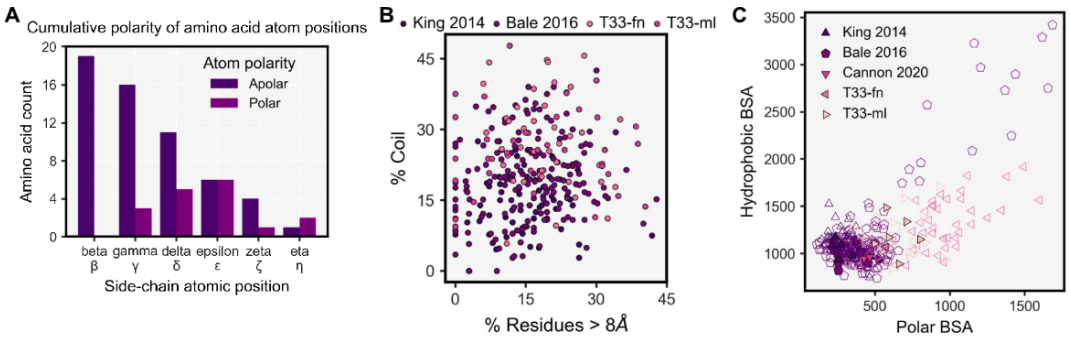
a) Averaged experimental small angle x-ray scattering (SAXS) profiles for images acquired from SEC peak fractions corresponding to assembled cages. The SAXS profiles (blue) correspond to images 238-258, 239-255, 229-249, 239-256, and 235-241 for T33-ml23, T33-ml28, T33-ml30, T33-ml35, and T33-ml6, respectively. Each scattering profile is plotted alongside the theoretical scattering calculated from design models (black). **b-c)** For the design T33-ml23, the 2.0 Å resolution allows aromatic side chains to be resolved. Example residues include component A Y85 (panel b) and W112 and F115 (panel c).

Supplemental Figure 5.6. Per-residue local density difference test for interface and interface fragment residues.



For the indicated designs, the local distance difference test (LDDT) was measured between the full cryo-EM assembly structure and the predicted design model. Values with greater than 0.8 (horizontal red line) indicate strong agreement. The location of interface residues (orange bar, upper segment; identified by 8 Å C-beta C-beta distances) and fragment residues (red bar, lower segment) are highlighted to indicate the extent of agreement in the modeled *de novo* interface. Entity breaks separate the A and B chains in each assembly.

Supplemental Figure 5.7. Retrospective analysis of polar interactions and interface areas for two-component designed protein cages.



a) Histogram of atomic polarity by side-chain position. For each canonical amino acid (n=20), the polarity of each atom in the side-chain (polar: N, O, S; apolar: C, H) is plotted at the side-chain atomic position it occupies. Polar atoms more frequently occupy atomic positions distal to the alpha carbon (Ca). **b)** The distribution of polar conformational features for design models. The fraction of coiled residues is independent of the fraction of residues greater than 8 Å apart as measured by Cb-Cb distance. **c)** Comparison of the interface area contributions from hydrophobic and polar buried surface area (BSA) interactions for 2-component designed cages considered in this analysis. The markers denote the symmetry of the assembly (triangle - tetrahedral, pentagon - icosahedral) and are filled with color if the assembly is reported as successful.

Supplemental Table 5.1. Comparison of intermediate assemblies to complete assembly structures.

Design Name	Assembly RMSD (Å)	ASU RMSD (Å)	Assembly LDDT
T33-ml23 $A_{12}B_9$	1.27	0.93	0.76
T33-ml35 A_9B_{12}	1.92	1.08	0.64

RMSD was calculated between all corresponding C-alpha atoms in the structure model, which for the ASU, constitutes a single A-B pair, while for the assembly, it constitutes all symmetrically related copies of A and B. To facilitate accurate comparison to the intermediate assemblies, for both designs, one trimer was removed from the full, symmetric assembly structure before measurements were performed. ASU - Asymmetric unit, RMSD - root mean squared deviation, LDDT - local distance difference test.

Designed sequences

T33-fn1-A

MKLIVAIVRPEKLYDVLRFLFHAGVRGLTLSRVQGHGGETERVETYRGTTVKMEFAEKVRLEIGVSEPFV
EATVIAILIAARTGEVGDGKIFVLPVEKVYRIRTGEEDEAAVTPVQ

T33-fn1-B

MDAQSAAKCLTAVRRHSPLVHSITNNVVTNFTANGLLALGASPVMAYAKSEVADMAKIAGALVLNIGTLS
KESLLAMTAAGLSANEHGVPVILDPVGAGATPTRTLAARFIIHAVRLAAIRGNAAEIAHTVGVTDWLIKG
VDAGEGGGDIIRLAQQAAQKLNTVIAITGEVDVIADTSHVYTLHNGHKLLTKVTGAGCLLTSVVGAFCAV
EENPLFAAIAAISSYGVAAQLAAQQTADKGPGSFQIELLNKLSTVTEQDVQEWATIERVTVSGHHHHHH

T33-fn2-A

MSLDYTTQQIIEKLRELKIVPVIALDNADDIIPLALTLAAAGLSVAEITFRSEAAADAIRLLRKISPDFL
IAAGTVLTAEQVHRAKRSGADFVVTPGLNPKIVKLCQDLNFPITPGVNNPMAIEIALEMGISAVKFFPAE
ASGGVKMIKALLGPYAQLQIMPTGGIGLHNIRDYLAIPNIVACGGSWFVEKKLIQSNNWEEIATLVKEVI
DIIK

T33-fn2-B

MHHHHHHSGAYRYRDIVVRKQDGFTHILLSTKSSENNSLNPEVMREVQSALSTAAADDSKLVLLSAVGSV
FCCGLDFIYFIRRLTDDRYRESTLMALAIAYFVNTFIQFKKPIIVAVNGPAIGLGASILPLCDVVWANEK
AWFQTPYTTFGQSPDGCSTVMFPKIMGGASANEMLLSGRKLTAQEACGKGLVSQVFWPGTFTQEVMVRIK
ELASCNPVVLEESKALVRCNMKDELIQANVREALVLTKIWGSAQGMDSMLKYLQRKIDEF

T33-fn3-A

MSLSYTTQQIIEKLKRLGIVPVIALDNADDILPLADTLAKNGLSVAEITFRSEAAADAIRLLRANRPDFL
IAAGTVLTAEQVVLAKSSGADFVVTPGLNPKIVKLCQDLNFPITPGVNNPMAIEIALEMGISAVKFFPAE
ASGGVKMIKALLGPYAQLQIMPTGGIGLHNIRDYLAIPNIVACGGSWFVEKKLIQSNNWKEIAYLVLEVL
IIIGE

T33-fn3-B

MHHHHHHSGMMTTSNAGAQQPNVEGRRFSPDQVRSVAPALEQYTQQRLYGDVWQRPGLNRRDRSLVTIAA
LIARGEAPALTYYADQALENGVKPSEISETITHLAYYSGWGKAMATVGPVSEAFAKRGIGQDQLAAVEST
PLPLDEEAEQQREDRVTRQFGSVAPGLVQYTTDYLFRDLWLRPDLAPRDRSLVTIAALISVGQVEQITFH
LNKALDNGLSLDQAAEVITHLAFYAGWPNAMSALPVALAVKFKRHS

T33-fn4-A

MSLSYTTQQIIEKLRELKIVPVIALDDYRDILGLAMVLAANGLSVAEITFRSEAAADAIRALRIFHPDFL
IAAGTVLTAEQVVLAKSSGADFVVTPGLNPKIVKLCQDLNFPITPGVNNPMAIEIALEMGISAVKFFPAE
ASGGVKMIKALLGPYAQLQIMPTGGIGLHNIRDYLAIPNIVACGGSWFVEKKLIKRNNGDEIARLVREVI
DIIKE

T33-fn4-B

MHHHHHHSGDALVDYAGPAATGGNVARLTLNSPHNRNALSSALVSQLHQGLRDASSDPAVRVVVLAHTGG
TFCAGADLSEAGSGGSPSSAYDMAVERAREMAALMRAIVESRLPVIAAIDGHVRAGGFGLVGACDIAVAG
PRSSFALNEAAIGVAPAIISLTLLPKLSARAAARYYLTGSQFDAREAEEIGLITAAALIVLAQVLMLAGA
VISGSPQGLAASKALTTAAVLERFDRDAERLAEESARLFVSDEAREGMLAFLENRAPNWFS

T33-fn5-A

METVETSAAPKPDGPYSQAIKVGNTLYVSGQIPIDEQTNTIVDGDIATQTAQVLLNIMAIVLAAGFSLSD
VAMAFVFLKDMNMFEDFNQTYALAFTDKPPARVTVEVSRLPKDALIEIAVICSKGSHHHHHH

T33-fn5-B
MAYRYRDIVVRKQDGFTHILLSTKSSENNSLNPEVMREVQSALSTAAKDSSKLVLLSAVGSVFCCGLDFI
YFIRRLTDDRLTESRKMAEAIRNFVNTFIKFQKPIIVAVNGPAIGLGASILPLCDVVWANEKAWFQTPYT
TFGQTPDGCSTVMFPKIMGGASANEMLLSGRKLTAQEACGKGLVSQVFWPGTFTQEVMVRIKELASCNPI
VLKQSKILVRSNMERELEKANDLEAYVLSKIWASAQGMDSMLKYLQRKIDEF
T33-fn6-A
MAYRYRDIVVRKQDGFTHILLSTKSSENNSLNPEVMREVQSALRKAQADESKLVLLSAVGSVFCCGLDFI
YFIRRLTDDRLDEAKKMAEAIRNFVNTFIQFTKPIIVAVNGPAIGLGASILPLCDVVWANEKAWFQTPYT
TFGQSPDGCSTVMFPKIMGGASANEMLLSGRKLTAQEACGKGLVSQVFWPGTFTQEVMVRIKELASCNPG
VLIVSKALVRSNMEMELEKANKLEAAVLLAIWALDDGMDSMLKYLQRKIDEF
T33-fn6-B
MHHHHHHGSLVRRIIFTDKAPDAIGAYSQAVLVDRTIYISGQLGMDPASGKLVPGGVIAETAQALLNILE
ILRAAGCRMTNVVKATVLLADINDFLDVAIVIAGFHTRSFPARAAYQVAALPKGGRVEIEAIAVQGPLTT
ASL
T33-fn7-A
MHHHHHHSGAYRYRDIVVRKQDGFTHILLSTKSSENNSLNPEVMREVQSALSTAAHDDSKLVLLSAVGSV
FCCGLDFIYFIRRLTDDRSRESLKMAEAIRNFVNTFIQFDKPIIVAVNGPAIGLGASILPLCDVVWANEK
AWFQTPYTTFGQSPDGCSTVMFPKIMGGASANEMLLSGRKLTAQEACGKGLVSQVFWPGTFTQEVMVRIK
ELASCNPVVLEISKALVRTNMEEELEQANADECSSLAYIWGLAQGMDSMLKYLQRKIDEF
T33-fn7-B
MPMFIVNTNVPRASVREGFLRRLTNALAAYTGKHRQYIAVHVVPDQLMTFSGTNDPCALCSLHSIGKIGG
EQNAALSAYLCILLSDDLKISPDRVYINYYDMNAANVGWNGDTFA
T33-fn8-A
MHHHHHHSGAYRYRDIVVRKQDGFTHILLSTKSSENNSLNPEVMREVQSALRTALYDDSKLVLLSAVGSV
FCCGLDFIYFIRRLTDDRKRESTKMAEAIKKFVITFIFFAKPIIVAVNGPAIGLGASILPLCDVVWANEK
AWFQTPYTTFGQSPDGCSTVMFPKIMGGASANEMLLSGRKLTAQEACGKGLVSQVFWPGTFTQEVMVRIK
ELASCNPVVLKESKHLVRLNMLEELYKANERECEVLKKIWGSAQGMDSMLKYLQRKIDEF
T33-fn8-B
MQWQTKLPLIAILRGITPDEAALHVAAVIKAGFDAVEIPLNSPQWEQSIPMIVAIYGEVALIGAGTVLKP
EQVDALARMGCQLIVTPNIHSEVIRRAVGYGMTVCPGCATATEAFTALEAGAQALKIFPSSAFGPQYIKA
LKAVLPSDIAVFAVGGVTPENLAQWIDAGCAGAGLGSDLYHQSLSLDATAKQALEFVEAYKRAVAL
T33-fn9-A
MHHHHHHSGMKVVVQIKDFDKVPQALRSVANLYADIKDAEIEVVLHQSAIKALLQDSPTRDITSLLIKAN
ILIVGCENSIRSQNLSHDQLQPGIKIVTSGVGEIVRKQSEGWIYLAL
T33-fn9-B
MPGMLPWTEQQFQLLGEIEEVELGRIQKRSGANLSRNWVMIPHVTHFDKTDITELEAFRKQQNEEAAKRK
LDVKITPVVFIMKAVAAALEQMPRFNSSLSEDGQRLTLKKYINIGVAVDTPNGLVVPVFDDVNKKGIIEL
SRELMTISKKARDGKLDTSEMVGGCFTISSIGGLGTTHFAPIVNAPEVAILGVSKSAMEPVWNGKEFVPR
LMLPISLSFDHRVIDGADGARFITIINNTLSDIRRLVM
T33-fn10-A
MKVVVQIKDFDKVPQALQSVLNLFLDLGNAEIEVVLHQSAIKALLLNSPTRSIIEELIKLNILIVGCEHS
IRSQNLDHRQLIDGIKIVRSGVGEIVRKQSEGWIYLAL

T33-fn10-B

MHHHHHHSGMSLRLERDGAVARLLIDRADRRNAFSLDMWQRLPELLAEASGDDALRVLVVKSANGGAFCA
GADIAELLANKDDAAFHLANQQAINRAQYELARFRLPTVAMVEGDCIGGGCGIALACDMRIAAPAARFGI
TPAKLGLVYPLHDVKLLVDLVGPGQARRLMFTGGLIDANEAHRIGLVELLGESEDALVGQLATVSSFSTQ
AIKSFVRRVLDGQVADDTLSLCVFASATLGADFREGTGAFLEKRPPVF

T33-fn11-A

MHHHHHHSGTDITANVVVSNPRPIFTESRSFKAVANGKIYIGQIDTDPVNPANQIPVYIENEDGSHVQIT
QPLIINAAGKIVYNGQLVKIVTVQGHSMAIYDANGSQVDYIANVLKYDPDQYSIEADKKFKLIKQIEDKI
QLILAAIASILRDLARIWKLIGE

T33-fn11-B

MSIDKLKHKLDDYAKDIKLNLSSITRSSVLDQEQLWGTLLASAAATRNKQVLADIKLDSTLYLDQREQHA
ALGAAAIMGMNNVFYRGRGFLEGRYDDLRPGLRMNIIANPGIPKANFELWSFAVSAINGCSHCLVAHEHT
LRTVGVDREAIFEALKAAAIVSGVAQALATIE

T33-fn12-A

MTDITANVVVSNPRPIFTESRSFKAVANGKIYIGQIDTDPVNPANQIPVYIENEDGSHVQITQPLIINAA
GKIVYNGQLVKIVTVQGHSMAIYDANGSQVDYIANVLKYDPDQYSIEADKKFKLIKQIEDKIEKILAAIA
HIEIDIALIKALIGE

T33-fn12-B

MHHHHHHSGDDPRLLSLFSAQREEDADIVIIGFPYDEGCVRNGGRAGAKKGPAAFRFFLQRLGSVENREL
NVNASHLKLYDAGDITASTLEEAHEKLESKVFTVLARGAFPFVIGGGNDQSAPNGRAMLRAFPGDVGVIN
VDSHLDVRPPLSDGRVHSGTPFRQLLEESSFDGSRFVEFACQGSQCGALHAAYVQANQGHLMWLSEVRKK
GAVRALAEAFKITGKNTFFSFDVDSLKSSDMPGVSCPAAVGLSAQEAFDMCFLAGSISTVMMMDMSELNP
LVEEYRSPRVAVYMFYHFVLGFATRP

T33-fn13-A

MTDITANVVVSNPRPIFTESRSFKAVANGKIYIGQIDTDPVNPANQIPVYIENEDGSHVQITQPLIINAA
GKIVYNGQLVKIVTVQGHSMAIYDANGSQVDYIANVLKYDPDQYSIEADKKFKLIAQIEGHLEAIASVLQ
SIINEIARIKKLIGE

T33-fn13-B

MHHHHHHSGDDPRLLSLFSAQREEDADIVIIGFPYDEGCVRNGGRAGAKKGPAAFRFFLQRLGSVNNLEL
NVDASHLKLYDAGDITASTLEEAHEKLESKVFTVLARGAFPFVIGGGNDQSAPNGRAMLRAFPGAVGVIN
VDSHLDVRPPLSDGRVHSGTPFRQLLEESSFDGRFFVEFACQGSQCGALHAQYVRDHQGVLMWLSEVRAL
GAVKALRLAFTLTGANTFFSFDVDSLKSSDMPGVSCPAAVGLSAQEAFDMCFLAGKTPEVMMMDMSELNP
LVEEYRSPRVAVYMFYHFVLGFATRSKPKAEN

T33-fn14-A

MTDITANVVVSNPRPIFTESRSFKAVANGKIYIGQIDTDPVNPANQIPVYIENEDGSHVQITQPLIINAA
GKIVYNGQLVKIVTVQGHSMAIYDANGSQVDYIANVLKYDPDQYSIEADKKFKLIKQIEDKIQQILEKIA
IIFRQLARIAKYIGE

T33-fn14-B

MHHHHHHSGMSLRLERDGAVARLLIDRADRRNAFSLDMWQDLPKLLAEARSDSALRVLVVKSANGGAFCA
GADIAELLANKDDALFHEENQEAINRAQYELARFRLPTVAMVEGDCIGGGCGIALACDMRIAAPAARFGI
TPAKLGLVYPLHDVKLLVDLVGPGQARRLMFTGGLIDANEAHRIGLVELLGESEDALVGQLATVSSFSTQ
AIKSFVRRVLDGQVADDAHSLNVFHLAFMTDDFREGTGAFLEKRPPVF

T33-fn15-A
MPFLQTIVSVSLDDQKRARLSLFYGMLCRKTLGIPGDQVMTAFSDKTPISFNGSTAPAAYVRVESWGEYA
PSKPKEMTAAIAAAIYAECGIPPERIYVFYYSTKHCGWNGHNF
T33-fn15-B
MHHHHHHSGAYRYRDIVVRKQDGFTHILLSTKSSENNSLNPEVMREVQSALSTAAADSASLVLLSAVGSV
FCCGLDFIYFIRRLTDDTKRESFKMAEAIRNFVNTFIQFQKPIIVAVNGPAIGLGASILPLCDVVWANEK
AWFQTPYTTFGQSPDGCSTVMFPKIMGGASANEMLLSGRKLTAQEACGKGLVSQVFWPGTFTQEVMVRIK
ELASCNPAVLRESKFLVRCNMKMELEQANEREAAHLKFIHAHAQGMDSMLKYLQRKIDEF
T33-fn16-A
MPFLQTIVSVSLDDRKRALLSTAYLYICREELGLALDSVMTAFSDKTPISFDGSTEPAAYVRVESWGEYA
PSKPKMMTPRIAAAITKECGIPKARIYVFYYSTKHCGWNGTNF
T33-fn16-B
MHHHHHHSGDALVDYAGPAATGGPVARLTLNSPHNRNALSSALVSQLHQGLRDASSDPAVRVVVLAHTGG
TFCAGADLSEAGSGGSPSSAYDMAVERAREMAALMRAIVESRLPVIAAIDGHVRAGGFGLVGACDIAVAG
LESSFALYEARIGVAPAIISLTLLPKLSARAAARYYLTHEKFDARRAEEIGLITMAAEDVDLLVALLVLA
VGSGSPQGLAASKALTTAAVLERFDRDAERLAEESARLFVSDEAREGMLAFLEKRLPNWFS
T33-fn17-A
MHHHHHHSGMVLKERQDGVLVLTLNRPEKLNAITGELLDALYAALKEGEEDREVRALLLTGAGRAFSAGQ
DLTEFGDHIPQYEDHLRRYNRVVEALSGLNKPLVVAVNGVAAGAGMSLALWGDLRLAAVGASFTTAFVRI
GLVPDSGLSFLLPRLVGLAKAQELLLLSPRLSAEEALALGLVHRVVPAEKLMEEALSLAKELAQGPTRAY
ALTKILLLETYRLSLTEALALEAILQGFAGLTKDHEEGVRAFREKRPPRFQGS
T33-fn17-B
MIVQQQNNLLRAIEAQEALLQLTVIGIKRLQARSGGRGGWETLERLIKKYTSTIASLIAESQNQQEK
T33-fn18-A
MHHHHHHSGMVLKERQDGVLVLTLNRPEKLNAITGELLDALYAALKEGEEDREVRALLLTGAGRAFSAGQ
DLTEFGDHHPRYGSHLQRYNRVVEALSGLEKPLVVAVNGVAAGAGMSLALWGDLRLAAVGASFTTAFVRI
GLVPDSGLSFLLPRLVGLAKAQELLLLSPRLSAEEALALGLVHRVVPAEKLMEEALSLAKELAQGPTRAY
ALTKKLLLETYRLSLTEALALERLAQAIAGMSQDHEEGVRAFREKRPPRFQGR
T33-fn18-B
MPLIRIDLTSDRSREQRRAIADAVHDALVEVLAIPARDRFQILTAHDPSDIIAEDAGLGSDKSPSVVIIH
VFTQAGRTIETKQRVFKAITLSLLPIGVMDADVFIAITENAPHDWSFAGGQNQYVQGELAIPATGAA
T33-fn19-A
MVLKERQDGVLVLTLNRPEKLNAITGELLDALYAALKEGEEDREVRALLLTGAGRAFSAGQDLTEFGDRK
PDYEAHLRRYNRVVEALSGLEKPLVVAVNGVAAGAGMSLALWGDLRLAAKGASFTTAFVRIGLVPDSGLS
FLLPRLVGLAKAQKLLLDSLKLSAEQALKLGLVHGVVKAHALMLEALLIARRLAQGPTRAYALTKKLLLE
TYRLSLTEALALEAVLQGQAGQTQDHEEGVRAFREKREPRFQGR
T33-fn19-B
MHHHHHHSGMTQTAPAAVAYSVNHAGVAAIVLDRPDASNALDEHMKTELLQALLAAGGDPAVRAVVMSAA
GKNFCVGQDLEEHVERLDDDPAHAMDTVREHYNPVLEALDAIKVPVVVAINGACVGAGLGLALGADIRIA
GQRAKFGTAFTGIGLAADSALSASLPRLIGASRATAMFLLGDTIDAPTAHTWGLVHEVVDEGSPADVANS
VAGRLAGGPTAAFSEVKELLRRNAVAPLGDVLEREASAQQRLGASVDHSAAVLAFEAKDKPIFYGD
T33-fn20-A

MAVSDQRLSEATKRELQDELQRAGHPQAPVIPDGWRMDFELGVTHFTMRKSHGDEEIILQLTGEDRSNEE
ITRTLDVLVVNGGKALVFGMSVEDGEFVINNVCFRHDGKLALDTSAEAQFQKSQLYMGPDLADLEDHLVD
SFTSYLSARGVNDTLANFIDQASLVFEQQNYLAWLLAINLFVS

T33-fn20-B

MLSVNEIAAEIVEDMLDYEEELRIESKKLSTGAIVVDCGVNVPGSYDAGIMYTQVCMGGLADVDIVVDTI
NDVPFAFVTEYTDHPAIACLGSQKAGWQIKVGKYFAMGSGPARALALKPLETMARIEYMDDARVAVIALE
ANQLPDDRHMTYMAIECFVRLENVYALVAPTASIVGSVQISGRIVQTAIFKMNEIGYDPKLIVSGAGRCP
ISPILENDLKAMGSTNDSMMYYGSVFLTVKKYDEILKNVPSCTSRDYGKPFYEIFKAANYDFYKIDPNLF
APAQIAVNDLETGKTYVHGKLNAEVLFQSYQIVLEEGSHHHHHH

T33-fn21-A

MADADQVLSAATQLELAIERTRAGLPEKPEIPPGWEIDRKPGVTHFTMRKSHGSETIILQLTGEDRSNEE
ITRTLDVLVVNGGKALVFGMSVEDGEFVINNVCFRKDGKLALDTSAEAQFQKSQLYMGPDLADLEEYLVD
SFTSYLSARGVNDTLANFIDQFSLWSEQADYEEWLESINQFMS

T33-fn21-B

MHHHHHHSGMSLRLERDGAVARLLIDRADRRNAFSLDMWQRLPELLKEASGDDALRVLVVKSANGGAFCA
GADIAELLANKDDAAFHAANTAAILLAMAELASFRLPTVAMVEGDCIGGGCGIALACDMRIAAPAARFGI
TPAKLGLVYPLHDVKLLVDLVGPGQARRLMFTGGLIDANEAHRIGLVELLGESEDALVGQLATVSSFSTQ
AIKSFVRAVLDGQIIDTAKSFAVFASAFEGADFREGTGAFLEKRPPVF

T33-fn22-A

MSLSQDGTLEQRQGDGVLTLTLGRAPAHPLSKETLARLKAALEWAMGDDSVHVLVIHGPGRIFCAGHDLK
EIGRHRADPDEGREEVTILFEECSALMLDLAHCPKPTIALVEGIATAAGLQLMAACDLAYASPAARFCLP
GVQNGGFCTTPAVAVSRVIGRRAVTEMALTGATYDADWALAAGLINRILPEAALATHVADLAGALAARNQ
APLRRGLETLNRHLELPLEQAYALATPVMVEHFMDPGRRHLDWID

T33-fn22-B

MHHHHHHSGSAVQPSFIRTNIGSTLRIIEEPQSDVYWIHMHADLAINPGRACFSTRLVDDITGYQTNLGQ
RLNTAGVLAPHVVLASDSDVFNLGGDLALFCQLIREGDRARLLDYAQRCVRGVHAFHVGLGARAHSIALV
QGNALGGGFEAALSCHTIIAEEGVMMGLPEVRFDLFPGMGAYSFMCQRISAHLAQKIMLEGNLYSAEQLL
GMGLVDRVVPPGFGELEIAKHIQKSQLTPHAWAAMQQVREMTTAVPLEEMMRITEIWVDTAMQLGEESLR
TMDQIVRKQSRRSGLDAG

T33-fn23-A

MKKIHTKNAPAARGPYVQGKIVGNLLFASGQVPLSPLSGKVIGTTIEEQTRQVLANIAAILGAAGTDFDH
VVKTTCFLSDIADFLPFNEVYADQFKSDFPARSAVEVARLPKNVKIEIEVIAELI

T33-fn23-B

MHHHHHHSGMMTTSNAGAQQPNVEGRRFSPDQVRSVAPALEQYTQQRLYGDVWQRPGLNRRDRSLVTIAA
LIARGEAPALTYYADQALENGVKPSEISETITHLAYYSGWGKAMATVGPVSEAFAKRGIGQDQLAAVEST
PLPLDEEDERAFATRVRNMFGDVAPGLVQYTTDYLFRDLWLRPDLAPRDRSLVTIAALISVGQVEQIYFH
LNKALDNGLSEEQAAEVITHLAFYAGWPNAMSALPVAKDVFKARRK

T33-fn24-A

MTTTVQNLIADINSLTSHLHEKDFLLTWEQTPDELKQVLDVAAALKALRADNISTKVFNSGLGISVFRDN
STRTRFSYASALNLLGLAQQDLDEGKSQIAHGETVRETANMISFCADAIGIRDDMYLGAGNAYMREVGAA
LDDGYKQGVLPQRPALVNLQCDIDHPTQSMADLAWLREHFGSLENLKGKKIAMTWAYSPSYGKPLSVPQG
IIGLMTRFGMDVTLAHPEGYDLIEEVVLIAQAAAGHSDGHYKQVTSMEEAFKDADIVYPKSWAPYKVMEE

RTELLRANDHEGLKALEKQCLAQNAQHKDWHCTEEMMELTRDGEALYMHCLPADISGVSCKEGEVTEGVF
EKYRIATYKEASWKPYIIAAMILSRKYADPGALLEQLLKENQPRVK
T33-fn24-B
MHHHHHHSGAYRYRDIVVRKQDGFTHILLSTKSSENNSLNAEVMKEVLDALHTADKDDSKLVLLSAVGSV
FCCGLDFIYFIRRLTDDRKRESSLMALAIRATAAFYAAFSKPIIVAVNGPAIGLGASILPLCDVVWANEK
AWFQTPYTTFGQSPDGCSTVMFPKIMGGASANEMLLSGRKLTAQEACGKGLVSQVFWPGTFTQEVMVRIK
ELASCNPVVLEESKALVRCNMKMELNQAIERECEVLKKIWGSAQGMDSMLKYLQRKIDEF
T33-fn25-A
MTGAGRSNDTGIVLQQAMLLMAIEAQQHLLQLTVWGIKQLQTRVLGGGGRWMQWDKEISNYTNTVYQILK
GAQSQQRKNKFDTLSL
T33-fn25-B
MHHHHHHSGSIEVLKAALSEYAKDIKLNLSSITRSSVLDQEQLWGTLLASAAATRNEQVLAMIAYQALDH
LSGGQFAAALGAAAIMGMNNVFYRGRGFLEGRYDDLRPGLRMNIIANPSIPKANFELWSFAVSAINGCSH
CLVAHEHTLRTVGVDREAIFEALKAAAIVSGVAQALATIEALSPS
T33-fn26-A
MAGAGQSNDSGIVQQQSNLLRAIEAQQHLLQLTVWGIKQLQTRVLGSGGAWGLWRGFIYKYTATVKRLLE
ESQNQQERNEKDLLALA
T33-fn26-B
MHHHHHHSGAYRTRDIVVRKQDGFTHILLSTKSSENNSLNPEVMRLVISSLSKAATGDSKLVLLSAVGSV
FCCGLDFIYFIRRLTHSRQLTSEQMARAIRIFVNTFIQFSKPIIVAVNGPAIGLGASILPLCDVVWANEK
AWFQTPYTTFGQTPDGCSAVMFPKIMGGASANEMLLSGRKLTAQEACGKGLVSQVFWPGTFTQEVMVRIK
ELASCEPAALMATKHAVWANMKMELEQANEIECEALKVRWGSAQGMDSMLKYLQRKIDEF
T33-fn27-A
MTDAGRKLDSIIVQQQANLLRAIEAQQHLLQLTVWGIKQLQTRVLGGGGRWMQWDKEISNYTNTVYRLLE
DSRYKQAALKLALKALA
T33-fn27-B
MHHHHHHSGAYRYRDIVVRKQDGFTHILLSTKSSENNSLNAEVQLEVQSALSTAAQDDSKLVLLSAVGSV
FCCGLDFIFFIRLLTRDREETSTTQAGADAGFVATFIQFKKPIIVAVNGPAIGLGASILPLCDVVWANEK
AWFQTPYTTFGQSPDGCSTVMFPKIMGGASANEMLLSGRKLTAQEACGKGLVSQVFWPGTFTQEVMVRIK
ELASCNPVVLEESKALVRCNMKMELEQANIREAQVLEKIWGKAQGMDSMLKYLQRKIDEF
T33-fn28-A
MTGAGFLNDWGIVQQQSNLLRAIEAQQHLLQLTVWGIKQLQTRVLGGGGRWMQWDKEISNYTNTVYRLLE
ESQNQQEINEHLLRHLA
T33-fn28-B
MHHHHHHSGAYRYRDIVVRKQDGFTHILLSTKSSENNSLNEQVMLYVRSALKKAATDDSKLVLLSAVGSV
FCCGLDFIYFIRRLTDDRDRESRFMAIAIREFVNTFIQFKKPIIVAVNGPAIGLGASILPLCDVVWANEK
AWFQTPYTTFGQSPDGCSTVMFPKIMGGASANEMLLSGRKLTAQEACGKGLVSQVFWPGTFTQEVMVRIK
ELASCNPVVLEESKALVRCNMKMELEQANERECEVLKKIWRSKQGMDSMLKYLQRKIDEF
T33-fn29-A
MPHLTLEYTDNLPEPQIRHLLFLLNGALLSRPEIFPVGGIRARAYRLSEYALADGSEPSDAFVHLRLQIG
AGRSDEQKKKTGDILFLILVAHFRAEFSQRGLMLSAEISEFSDKGTWKKNNIHARYRK
T33-fn29-B

MHHHHHHSGDDPRLLSLFSAQREEDADIVIIGFPYDEGCVRNGGRAGAKKGPAAFRFFLQRLGSVNNLEL
NVDASHLKLYDAGDITASTLEEAHEKLESKVFTVLARGAFPFVIGGGNDQSAPNGRAMLRAFPGDVGVIN
VDSHLDVRPPLSDGRVHSGTPFRQLLEESSFDGQRFVEFACQGSQCGALHAQYVRDHQGILMWLSEVRKL
GAYQALLIAFALTGSNTFFSFDVDSLKSSDMPGVSCPAAVGLSAQEAFDMCFLAGSDEQVMMMDMSELNP
LVEEYRSPRVAVYMFYHFVLGFALRSKPKAEN
T33-fn30-A
MHHHHHHSGEQLPQCETLILEKQGPTLVITINRPDVRNAMSLQMVAELSTIFSEIENDISIRAAVLRGAG
GHFCAGGDIEDMLEARAQKAGEGRDDPFYKLNRAFGQMIQQVNESSKVVIAITEGAVMGGGFGLACVSDL
AIAGPTAKFGMPETTLGVIPAQIAPFVVERIGLTQARRLALLGLRIDATEACKLGIVHQVAESEEQLSDM
LNQALERVRLCAPDATAETKALLHRVGHEAMAGLLDDAAEKFAAAIRGPEGAEGRMASLQDREPKWAELP
NQ
T33-fn30-B
MTQTAPAAVAYSVNHAGVAAIVLDRPEASNALDRTMKTELLQALLAAGGDPAVRAVVMSAAGKNFCVGQD
LAEHVEALRDDPANAMKTVEEHYNKVLEALDAIKVPVVVAINGACVGAGLGLALGADIRIAGQRAKFGTA
FTGIGLAADSALSASLPRLIGASRATAMFLLGDTIDAPTAHTWGLVHEVVDEGSPADVANSVAGRLAGGP
TAAFSEVKELLRRNAVAPLGTVLLKETIAQLRLGSSRDHSAAVEAFLAKDKPVFVGR
T33-fn31-A
MHHHHHHGSNDVLFSNHGRVAVITLNRGDRLNAWTTPMRETIIDALERFNRDPEVAAIIMTGKGREAFSA
GQDLSEAHDFDGERAVAWVKEWQRYYTALRSLSKPLVMALNGTAAGSAFQVALLGDIRVGHPGVRMGQPE
INAGIASTTGPWIMNAMLGMSRTIELTLTGRLMPADHCHRIGLIHVLTSEDLVFDEALLIATELAAKPPV
AMRLDKQRFREMTEPGFIDCIEAGERIQREAYDSGEPARMMEEFFSKRAK
T33-fn31-B
MSGIDTKQQNNLLSAIIAQQHLLQLTVWGIKQLQARSGGRGGWMAWDRHINNLTSIIHSHIKESLDQQEK
T33-fn32-A
MHHHHHHGSNDVLFSNHGRVAVITLNRPDRGNAWTTPMRETIIDALERFNRDPEVAAIIMTGAGNDIFSK
GQDLSEAHDFDGERAVAWVKEWQRYYTALRSLSKPLVMALNGTAAGSAFQVALLGDIRVGHEATFMGQPE
INAGIASTTGPWIMNAMLGMSRTIELTLTGRLMEAEECHRIGLIHLLVHESQVFDMALIIATNLAAKPPV
AMRLDKQRFREMTEPGFIDCIEAGERIQREAYDSGEPARMMEEFFSKRAK
T33-fn32-B
MILVYSTFPNEEKALEIGRKLLEMRLIACFNAFEIRSGYWKDGRIVQDKEWAAIFKTTEEKEHDLYEALR
LLHPYEFPAIFTLKVENVLEEYMALLRASVS
T33-fn33-A
MHHHHHHSGFQSMSNDVLFSNHGRVAVITLNRPERLNAWTTPMRETIIDALERFNRDPEVAAIIMTGAGQ
DAFSAGQDLSEAHDFDGERAVAWVKEWQRYYTALRSLSKPLVMALNGTAAGSAFQVALLGDIRVGHKHVR
MGQPEINAGIASTTGPWIMNAMLGMSRTIELTLTGRIMPAKECHRIGLIHYLTHESTVFDVALLIAEILA
RKPPVAMRLDKQRFREMTEPGFIDCIEAGERIQREAYDSGEPARMMEEFFSKRAK
T33-fn33-B
MPHIRVRGAEKEKVRDFTAGLADILGRAASDTASAFTFEYVETTFFFDGKEDDGLVFIEVLWFDRDSETR
ATIALLFTLKWKRITDKIVTIVFNPLIENMYYVDGKRF
T33-fn34-A
MHHHHHHSGPSSAIATLAPVAGLDVTLSDGVFSVTINRPDSLNSLTVPVITGIADAMEYASTDPEVKVVR
IGGAGRGFSSGAGISADDVSDGGGVPPDTIILEIERLVRAIAALPHPVVAVVQGPAAGVGVSIALACDVV

LASENAFFMLAFTKIGLMPDGGASALVAAAVGRIRAMQMALLPERLPAAEALAWGLVTAVYPADEFEAEV
DKVIARLLSGPAVAFAKTKLAINAATLTELSPALQRESLGQSVLLKSPDFVEGATAFQQRRTPNFTDR
T33-fn34-B
MIIVYTTFPDWESAEFVVKQLLLARMIACANLREHRAFYWWSNSIEEDKEVGAILKTRESLWRDLKEAIK
QLHPYDVPAIIRIDVDDVNNGYEEWLIEETQK
T33-fn35-A
MHHHHHHSGTQTAPAAVAYSVNHAGVAAIVLDRPEASNALDRTMKTELAAALKKAAGDESVRAVVMSAAG
KNFCVGQDFTEHAVALARDPRHAMDTVREHYNPVLEALDAIKVPVVVAINGACVGAGLGLALGADIRIAG
QRAKFGTAFTGIGLAADSALSASLPRLIGASRATAMFLLGDTIDAPTAHTWGLVHEVVDEGSPADVANSV
AGRLAGGPTAAFSEVKELLRRNAVAPLGDVLEREASAQQRLGASRDHSAAVKAFLAKDKPVFVGR
T33-fn35-B
MQWQTKLPLIAILRGITPDEALAHVGAVIDAGFDAVEIPLNSPQWEQSIPAIVDAYGDKALIGAGTVLKP
EQVKALADMGCQLIVTPNIHKEVIVAAVAFFMTVCPGCATATEAFTALEAGAQALKIFPSSAFGPQYIKA
LKAVLPSDIAVFAVGGVTPENLAQWIDAGCAGAGLGSDLYRAGQSVERTAQQAAAFVKAYREAVQL
T33-fn36-A
MIVDYSVNHAGVAAIVLRDAKNSNALDDGAKTELLHALLKAGGDPAVRAVVMSAAGKNFCVGQDLREHWI
ATAKDPAHAMDTVREHYNPVLEALDAIKVPVVVAINGACVGAGLGLALGADIRIAGQRAKFGTAFTGIGL
AADSALSASLPRLIGASRATAMFLLGDTIDAPTAHTWGLVHEVVDEGSPADVANSVAGRLAGGPTAAFSE
VKELLRRNAVAPLGDVLEREASAQQRLGASKDHRAALLAFMNKDKPVFVGR
T33-fn36-B
MHHHHHHSGSAVQPFIRTNIGSTLRIIEEPQRDVYWIHMHADLAINPGRACFSTRLVDDITGYQTNLGQR
LNTAGVLAPHVVLASDSEVFNLGGDLALFCQLIREGDRARLLDYAQRCVRGVHAFHVGLGARAHSIALVQ
GAALGGGFEAALSCHTIIAELAGSYGLPEVQNDLFPGMGAYSFMCQRISAHLAQKIMLWGNLFSALQLLG
MGLVDAVVSEGSGVDMVEFVIFISKRTPHAWAAMQQVREMTTAVPLEEMMRITEIWVDTAMQLGEKSLRR
MDELVKADSRRSGLDAG
T33-fn37-A
MHHHHHHSGMTQTAPAAVAYSVNHAGVAAIVLDRPEASNALDRTMKTELLQALLAAGGDPAVRAVVMSAA
GKNFCVGQDLAEHVEALREDPSNAMATVREHYNPVLEALDAISVPVVVAINGACVGAGLGLALGADIRIA
GQRAKFGTAFTGIGLAADSALSASLPRLIGASRATAMFLLGDTIDAPTAHTWGLVHEVVDEGSPADVANS
VAGRLAGGPTAAFTFVKYALRMNAVAPLGVVLDIEATFQQFLGASRDHSAAVEAFLAKDKPVFVGR
T33-fn37-B
MAGAGQSNDSGIVQQQSNLLQAIQRQLHLLELTVKGIKQLQTRVLGGGGLWTAIDLQISFMTEAVKRLLR
EAQEQQDRNEKDLLALA
T33-fn38-A
MHHHHHHSGTQTAPAAVAYSVNHAGVAAIVLDRPEASNALDRTMKYELLKALLLAAANLDVRAVVMSAAG
KNFCVGQDRDEHIEALRDDPKNAMDTVREHYNPVLEALDAIKVPVVVAINGACVGAGLGLALGADIRIAG
QRAKFGTAFTGIGLAADSALSASLPRLIGASRATAMFLLGDTIDAPTAHTWGLVHEVVDEGSPADVANSV
AGRLAGGPTAAFSEVKELLRRNAVAPLGDVLEREASAQQRLGASRDHSLAVKAFMADAKPIFVGR
T33-fn38-B
MTSTAVEITVKNAADIAIIGGSGLYQMQALTNKRSVRIATPYALPSDDIVLGELNGVTVAFLTRHGQGHR
LTPSEVPYRANIYALKSLGVRYIVSVSAVGSLQETLKPLDMVIPDQMIDMTKQRVSTFFGDGAVAHVSMA
DPLCPEVADILIRAYDNADIADGQCHAKATYVCIEGPQFSTRAESHWYRQMQADIIGMTNMPEAKLAREA

SIAYATLALVTDFDCWHPNEQAVSADYAIQNLMKNADNAQQVIKQAVALIASEQPKSIAHTALTQALVTP
VEAMSEETKLRLFALLP

T33-fn39-A

MTQTAPAAVAYSVNHAGVAAIVLDSPRNSNALDDEMKTELLQALLAAGGDPAVRAVVMSAAGKNFCVGQD
LFAHFAELRRDPAHAMDTVREHYNPVLEALDAIKVPVVVAINGACVGAGLGLALGADIRIAGQRAKFGTA
FTGIGLAADSALSASLPRLIGASRATAMFLLGDTIDAPTAHTWGLVHEVVDEGSPADVANSVAGRLAGGP
TAAFSEVKELLRRNAVAPLGDVLEREASAQQRLGASRDHSAAVEAFLAKDKPVFVGR

T33-fn39-B

MHHHHHHSGMSLRLERDGAVARLLIDRADRRNAFSLDMWLRLPELLAEASGDDALRVLVVKSANGGAFCA
GADIAELLANKDDGAFHDANQMAILRAQLELARFRLPTVAMVEGDCIGGGCGIALACDMRIAAPAARFGI
TPAKLGLVYPLHDVKLLVDLVGPGQARRLMFTGGLIDANEAHRIGLVELLGESEDALVGQLATVSSFSTQ
AIKSFVRRVLDGQVAMDADAAHVLASAYEGADFREGTGAFLEKRPPVF

T33-fn40-A

MSDLRLERDGAVARLLIDRPQNNNAFDTQMWQNLPVLLADASGDDALRVLVVKSANGGAFCAGADEHLLL
TRMLDDDWHAENQQAINRAQYELARFRLPTVAMVEGDCIGGGCGIALACDMRIAAPAARFGITPAKLGLV
YPLHDVKLLVDLVGPGQARRLMFTGGLIDANEAHRIGLVELLGESEDALVGQLATVSSFSTQAIKSFVRR
VLDGQVADDADSLRVFASAFKQKDFMEGQLAFAQNRPPVF

T33-fn40-B

MHHHHHHSGMYETIRYEVKGQVAWLTLNRPDQLNAFTEQMNAEVTKALKQAGADPNVRCVVITGAGEAFC
AGEDLSGVTEEMDHGDVLRSRYAPMMKALHHLEKPVVAAVNGKAAGAGMSLALACDFRLLSETASFAPAF
ISVGLVPDAGHLYYLPRLVGRAKALELAVLGVRVTALQAAKLGLATAVIPKHLWELAVKAYASALSNMPT
KAIGLIKRLLRESEETTFDRYLEREAECQRIAGLTSDHREGVKARNESRKPLFQGN

T33-fn41-A

MHHHHHHSGMSLRLERDGAVARLLIDRADRRNAFSLDMWQRLPELLKEASGDDALRVLVVKSANGGAFCA
GADIAELLANKDDAAFHAANQDAINYAQYELARFRLPTVAMVEGDCIGGGCGIALACDMRIAAPAARFGI
TPAKLGLVYPLHDVKLLVDLVGPGQARRLMFTGGLIDANEAHRIGLVELLGESEDALVGQLATVSSFSTQ
AIKSFVRRVLDGQVADDKQSLLYFAAAYHHADFREGTGAFLEKRPPVF

T33-fn41-B

MNDFLNSTSTVPEFVGASKIGDTIGMVIPNVDQQLLDKLHVTKQYETLGILSDRTGAGPQIMAMDEGIKA
TNMECIDVEWPRDTKGGGGHGCLIIIGGDDEKDARQAIRVALDNAARTFGDVYNAKAGHLELQFTARAAG
AAHLGLGAVEGKAFGLICGCPSGIGVVMGDKALKTDGVEPLNFTSPSHGTSFSNEGCLTITSAAAAVLTA
VLAGRRVGLKLLSQFGEEPKNDFQSYAK

T33-ml1-A

MYPVDLHMHTIKNDASFSTLSDYIARAKEKGIKLIAITDHGPSHPIAPHPEYYVRMKELPDVVDGVGVLR
GVEANILDTNGNIDVTPEMEKSLDLILAGLYESVYPPQSRAENTKALINAIASGKVHVISHPADPRYPVD
YRALAGAAAVAGVALEITEHAFGEEFPGAEPRARELARAVKEAGGYVALGSDAHHAWHLGRFEHAERVLR
EVGFPEERVLNRSPEKLLAFLESRGVPKKPAFADLGGSHHWGGHHHHHH

T33-ml1-B

MSISYRKLDIALSADGEEVLVEGFVLPTKFFENVIVTTMLNAAGTDEENINALLADVHAAGLDVSNYGKA
SEIYAKGDPEKRAEAEARRAEAEARRAELAAELSTPEAQAEAKKEKVLEAAELAARFGPAGVKAGL

T33-ml2-A

MHHHHHHGGSHHWGGKLLLAATGSPAARFFGELAKQFVPHFEVRAVLTEGALEFVDLSSLPAEVPVYTDE
DMRAAFKKEGDVILHIELADWADVLLIAPASINTIAEIASGLAPNLLLRIFAGWDLSKPVFIAPAMSQRE
YDNPATKENLKKLEERGVIIIPPVKGRGADGSVGNGVMAPPEKIADAVLGYLEARKALKKVVTS
T33-ml2-B
MNMAETYYEIGKKFEKTGAYDAAIHAYLAALAEDPNNAEAWYNLGKAYEKLGKYKEAIEAYKKALELDPK
NAEAWYNLGKAYEKLGDYKKALEEYLKSLELDPFNEEAKKNAKEAGKKGVLE
T33-ml3-A
MKLPNKVSLVAGTAEGATPENALHGARLNAGIGDVNLVPVSGIAPAGAEIVPLPELPPGALLPTAEASIV
SDVPGKTIAAAVAVGIPKDPSLPGIIATYAGEMSAEEARRRVELIVIEQFLQRGWELESIHSVAVEHTVK
RLGAALAAAVLWYKGGSHHWGGHHHHHH
T33-ml3-B
MTAEGETAVAIALLGDEDPHVRAEAAKKLGKIGDKEAVKPLIEALGDEDPAVRAAAALALGKIGDKEAVP
PLIGALLDEDPAVRVAAALALGKIGDKEAVPALILALLDEDEAVRVAAAVALGKIGDKEAVEPLVKALEK
EEGLVRKAAAIALEKIGGEEVKKAAEELAKKGEGEARKAAEEYLKKHKLE
T33-ml4-A
MLDEIAFADARILTPFTEADIERLLDALELEPGTRVVDLGCGTGYFLVLGAERKGITATGIDISELAIEK
ARELARERGVEDRVEFIHGDVSTYVAEEKVDVAACIGAEAYFGGIEGALKALEKSLKEGGIILLGVPYWR
TRPATEAEARACGFDSIDDLDTLAETVAKLEALGYRVIQIVLADEHGMEELYYGRLVQLDRWLREHPDHP
FAPALEAELATLAARYEKYVRRHLGYGVFALRKRLEGGSHHWGGHHHHHH
T33-ml4-B
MIVVLITVPSEEVARKIARAAVEGGLAAEVLIIPALTLYYRENGKVVERPVYLLLVLSTESKFPALLALV
KALHPEKVPLIVALPVVDGNPEYLRWVKLNTG
T33-ml5-A
MPMVIFECSDNIREEAKFEELFARLNPALASTGLFPLEEIVGRVHWVDTWQFADGQHDYAFVHVTIDVPA
GLSEEDRLLVLNGVFALLLGHLEPLMKEHLLYLSLELRVLPATLSRRWNNAIELFKGGSHHWGGHHHHHH
T33-ml5-B
MSVKIDVIRVEIPEGTWVIIGQSHSRSIVFDLSQTLQSASGRLRFGIAYCEASGKRLILHDGNDPALVEL
AKETALKIGAGHTFVIYIRNGRPEDILNRIKNIESVVRIFAATANPLQVLVAETDQGRGVIGVVDGYTPL
GVATAEDRAALAAALRAEGYKR
T33-ml6-A
MTAVFAAIVGFLDKKIEELKKIQKHKTLPKMSGGWELELNGTEAKLVREVDGYKVTVTFNINNSIPPTFD
GEEEPSQGQPVEEQPPELTSTPNFVVEIVKASQPDTALVFDCYYPEDEVGQEEEEEKPLFEIKEVSFQST
GESEWKDTNYTLNTDNLPEDLLLAFYALLAALGVDNDFAKELIELSTALEHQERITFFEQLRDFIA
T33-ml6-B
MNLAEKMYEAGKYFAAQGNYELAIIAYTLALLKDPNNAEAWYNLGKAYAALGKYEEAIEAYKKALALDPN
NAEAWYNLGGAYAGLGKYEEAIEYLEKALALDPNNELAKLMLKFAKLQLELEGGSHHWGGHHHHHH
T33-ml7-A
MHHHHHHGGSHHWGGMLFSFLHEEKKLGKIIVVDKGSGPEHVRSQLKTCGDYIDYVRFAGKTAAQMPAEV
VKEKIAIYHEKGIKVFPGASLFEEAVKKGKEDEFLAECKAAGFDAVEIGNLNINLSDEELEALIKKAKAA
GFEVFTVVGRADPKVDKLLSVSDIVRRIRRFLEAGADYVIIYGGSTGKGKGLYDENGNLKEEDLDYIKEN
VPMEKIIFEAPLEKQQKQLIEKFGPSVNIAEISFSDVITVANLRSGLRGDTFGKV
T33-ml7-B

MKERQISLLETLLSLYIDLLEVMADMAGKSGKYVLLDVREDLKFVEKNKIPGAIWLPVSLLEERIDELDP
SKTYVVYDYKGHSTNSYRALLILLKAGFEAYILSGDPKLLLG
T33-ml8-A
MLAFEFLHEEKELGKIIVVDTGTSPHHLKGQLETVGDFIDYVKFAGMTAAVMPKKVVCEKILIYHKHGIK
VMPGGTLFEKAVSKGKEKEFLLECRELGFDAIEISDLNIDLSDEELKKLIKMAKEEGFEVFTKVGRADKE
RDAKLTVEDIIAKINFYLEAGADYVIIYGGASGKGIGLYDENGKLKKEWLDEIKKNVDMSKIIFEAPLPE
QQKELLDKFGPSANLAEISLHDVARLAEMRYGLRGDTFGKV
T33-ml8-B
MHHHHHHGGSHHWGGGMAQAQTQGQEEEQKKKIVIIIRHGPEEPIYCVTPLRLAVVAAEQGYETTIVFTE
LGPELLNKLYWIEEMAKGGNPVTKYLLKAREKGVKIYVCEWSLEEICKLKKEDIIPGVEIIDDKDIIKLM
LEADVVIFF
T33-ml9-A
MHHHHHHGGSHHWGGMKAFEFLYEDFQRGLTVVLDKGLPPKFVEDYLKVCGDYIDFVKFGWGTSAVIDRD
VVKEKINYYKDWGIKVYPGGTLTEYAYSKGKFDEFLNECEKLGFEAVEISDASIDFSMKEMIDMIRKAKA
NGFMVLTLVGRKDPAKDAELTIVERVIRITAYLDAGADYVIIYGRESGKGKGLFDKEGKLKLDDLDILAS
SVDMSKVIFEAPQKSQQVALILKFGSSVNLANIAFDEVISLETLRRGLRGDTFGKV
T33-ml9-B
MSKTTVIYPGSFDPIHKGHVDLIERASKMFPRVVVAVVKGHHKKHTFSIVERLLLVEAAVGHLPNVEVRA
VDGLLVNVFKELKATAVLRGLRAVSDFEYEFQLANMNRQLDPHFEAVFLTPSEQYSFISSTLIQKLAANG
GDISQFVPPVVVAAFKALKGKGW
T33-ml10-A
MPLVVLVTVPSEEEARRIARALVERRLAAQVNIVPGLTSIYRRDGEVVEDQELLLLVFTTELRFPLLREL
VRSLHPEATPMIVALPVVDGNTDYLLWLLENTG
T33-ml10-B
MPFEKALYFLTYLSYTIDIAELSILIKKGDKSIIIVDVRDAEAYKECHIPTAISIPGNKINEESTKDLPK
DKTIITYCWGPACNGATRASRKFAELGFDVKRLIGGIEYWRKENGEVEGTLGAKADLFWNMKKESLEGGS
HHWGGHHHHHH
T33-ml11-A
MFSNKRVLVEKEGEAGIAVMKFKNPPVNSLSLEFLTEFVISLEKLENDKSIRGVILTSERPGIFSAGLDL
MEMYGRNPAHYAEYWKAVQELWLRLYLSNLTLISAINGASPAGGCLMALTCDYRIMADDDGYTIGLNESL
LGIVAPFWLKDNYVNTIGHRAAERALQLGTLFPPAEALKVGLVDAVVPPEAVLAAAKGTLAEWFQIPDHS
RQLTKSMMRKATADNLIKQREADIQNFTSFISRDSIQKSLHVYLEKLKQKKG
T33-ml11-B
MHHHHHHGGSHHWGGSSGLVPAGSHMRLTPHEQERLLLSYAAELARRRRARGLRLNHPEAIAVIADHILE
GARDGRSPGELAAAGQTVLGRDDVVGGVPEMLAEVQVEATFPRGTMTVTVERPIA
T33-ml12-A
MMSGWFPVKTTEELEVIDITPLVEAALKGAGLKNGLVLVYVPDVDAAIIVNTADPELLEDIVRHLRTLCD
PEGDWAYNKVEPNAHAYLGTALVGNSVVIPVRNGKLDLGKEQKVLFIDMDGPDTYTVKLMALEE
T33-ml12-B
MIKKPEFGLMQPPKKRVRQELSSVAEETIEAAFDFFDVDGDGKINKEELKKALHALGFAVNDMQIEALMA
AYDKDGDGYINKEEFKEIVELLRKNGGSHHWGGHHHHHH
T33-ml13-A

MNAVVQRTELGNGVVQITMKDESSRNGFSPSIVAGLKAALDAVIDDSSVKVVILTGYGNYFSSGASKEYL
LALTKGEVGVLNLVPLILDCPVPVIAAMQGHSFGGGLLLGLACDFVVFSQESVYATNESKYGFTPYAAAR
LILRRKLGSELAQEMLYTGENYRGKELAERGIPFPVVSRQDVLNYAQQLGQKIAKAPRLTLILLNIDARA
DLRAAYPAALRRELGLFSLTFSQPEIPERIQQEF
T33-ml13-B
MRRGLLPNDVWQADICEYKYKKYKYCLHIVVDTFSGAMSVSCKKKKTPLETIEALLQAISLLGRPKKIIS
DHDPAFRHGLTKAFCLSSGIELESYTPGDPSSSALVDAACKELKALLDRYLTENPELPLDNAINLALWEH
NQLKVVEPYGKTPWQLHHSGGSHHWGGHHHHHH
T33-ml14-A
MHHHHHHGGSHHWGGSELTVNVINGPNLRMLGFREPAVYGGTTFSELVELIEREAAELGLKAVVRQSDSE
AQLLKWIHLAALMAEPVILNAGGLTHTSVALRDACAELSAPLIEVHISNVHAREEFRRHSYLSPIATGVI
VGLGIQGYLLALRYLAEHVGT
T33-ml14-B
MPMFIVNTNVPRASVPDGFLSLLTRLLALLTGKPEKYIAVHVVPDQLMAFGGSSEPCALCSLHSIGKIGH
EQNRSYSKLLCTVLAQRLRISPDRVYINYYDMNPENVGWNNSTF
T33-ml15-A
MSLKDKKILIVEDSLEQAITIGLILVKYGYEVIIAGTGEQAVEYVSGGEYPDLILMDIELGEGMDGVQTA
LAIQQISELPVVFLTAHTEPAVVEKIRSVTAYGYVMKSATEQVLITIVEMALRLYEANVHANEG
T33-ml15-B
MHHHHHHGGSHHWGGGRSLVVIVNDRTAHGDQDKSGPLVVGLLRAAGFVVDGVVVVENDLSEIQNAVNTA
VIGGVDLVVTVGGTGVTPRDVAPEATQPLLDRELLGIAEAIRSSGLAAGVTEAGLSRGVAGISGSTLVVN
IAGSQEAVLVGLKTLLPMAIQIIEQLSSLEI
T33-ml16-A
MAEVSIKTKIKAKHRLYSKNLSEEENKMLFGSAAKKGGHEHNYTITVTVKGEIDPTTGLVINGTDLRIWI
EKAIILPLDNKNLNEDVPYFKTNVPTTENIAKYIKENLEKVLPKGLLSKVVVEETEEHKVTIKGEGGSHH
WGGHHHHHH
T33-ml16-B
MPAIILTTTPTEADARALAEGLLEKRLIAEAIITPNVTRIYLENGEIKSEKVVRMELYTVEEKVEAAMTY
IEAHHPDPIPPIIVIKPDKVSPKYKKWILEQTAL
T33-ml17-A
MAPTMTEFVGTAGGDTVGLVIANVDSLLHKHLGLDNTCRSIGIISARVGAPAQMMAADVAVQTTNTEVAT
IELPRDTKGGAGHGIFIVLKAADVSDARRAVEIALAMTDEYLGDVYLCDAGHLEVQFTARASLIFEKAFG
APSGQAFGIMHAAPAGVGMIVADTALKTADVKLITYGSPTNGVLSYTNEILITISGDERAVLKSLDAARK
AGLSILKDMGEKPVSMSEPTF
T33-ml17-B
MPLIRIDLTSTRSRLQRQLIAQAVHDALVEVLAIPARDRFQILTAHPISDIIAEDAGLGFTRSPDVVIIH
VFTQAGRTIETKQRVFAAITESLAPIGVAGSDVFIAITENAPHDWSFGFGSAQYVTGELAIPATGAAGGS
HHWGGHHHHHH
T33-ml18-A
MKILIVVTHGPEDLDRTYAPLFLAVVAAERGYKTSVFFMIKGPLLLNRDYIAKVALEGGNPYLEYLYKAK
QLGVEIYVCVQSLRDMCHLKEEDIIGGVKLVGGSTLIDLTLEADRTLFF
T33-ml18-B

MNLAEKMYKAGNAMYRKGQYTIAIIAYSLALLLDPKNAEAWYNLGNAYYAKGEYDDAIKAYEKALMLDPN
NAEAWYNLGNAYYAKGDYESAILAYQLALKLDPNNAEAKQNLANAKQKLALEGGSHHWGGHHHHHH
T33-ml19-A
MPMVTIRTNLPASEVPADFAAELTALLSKTLGVPADRIAVEVLPGVDLTFGGSREPVALITVESIGNLTP
EQTNLLTLQLTLLLQLRLGLPEDRVLILFHDLPASQVGRDGRTEAAA
T33-ml19-B
MHHHHHHGGSHHWGGSPSDPPRPALLMLELRSYALGLAVADAALRAAPVRLLLARPVEPGKALILLTGEE
EACRAALEAALRVAREGSGNLLDSVFIPAIHPQLLPFLLEEVAAPPLADPDEAVLVAEVRTPAAAIRAAN
AALEAAPVRLTRMRLAEHIGGKAYFTLTGRREDVLRAAQVIAEVAGEDLIDLRLIPRPHAALRGREFF
T33-ml20-A
MHHHHHHGGSHHWGGGMKVTFLGAAVVLIEGKKNIIIDPFISGNPVCPVKLEGLPKIDYILVTHGHGDHL
GDAVEIAKKNDATVISNYEICHYLGKKGVKTHAMHIGGSYLFDFGRVKMTPAVHGSGILDGDSMIYGGNP
SGFLIEIDGKKIYHAGDTGLTREMELLAEENVDVAFLPIGGNFVMDVKDAIKAAKMIKPKKVVPMHYGTW
ELIFADVEAFKAGVEAIGVECVILEPGESLEL
T33-ml20-B
MPVITVNTNVAEKSIPVFFQAALTNMMSKLLDVGKERMFVDLRSGANIMMGGDRNPCVFATVECIGRLNP
GSCALMAQEMEKMFIEHLNVRRERIVIRFIPVPAEFCSFNGKLHDVKEERDEYLE
T33-ml21-A
MTVPEFVGASEIGDTIGMVIPRVDQQLLDKLHVTKQYKTLGILSDRTGAGPQIMAMDEGIKATNMECIDV
EWPRDTKGGGGHGCLIIIGGDDPADARQAIRVALENLPRTFAGVFNAKAGHLEFQWTPRAAGAAHLGLGA
VEGKAFGLICGCPSGIGVVMGDKALKVAGVEPLNFTSPSHGTSFSNEGCLTITGDPRAVLAAVMAGAEVG
LKLLSQFGEEPVD
T33-ml21-B
MPLIRIDLTSDRSRFQRLAIAEAVHLALVEVLAIPERDRFQILTAHDPLDIIAEDAGLGFTRSPSVVIIH
VFTQAGRTIETKQRVFAAITEALAGIGVAGSDVFIAITENAPHDWSFGFGSAQYVTGELAIPATGAAGGS
HHWGGHHHHHH
T33-ml22-A
MHHHHHHGGSHHWGTVPEFVGASEIGDTIGLVIPRVDQQLLDKLHVTKQYKTLGIISDRTGAGPQIMAAD
EGIKATNVECIDVEWPRDTKGGGGHGCLIILGGEDPEDMRHAVRVALAELPRTFARVFYNKAGYAVFQYT
DAAAGAAHLGLGAVEGKPFGLIAGCPSGIGVVAADEALKVEGVEPLNFTSPSHGTSFSNEGCLTITGDPE
AVRLAVERGEAVAIRLLKTFGEEPKNDFPSYIK
T33-ml22-B
MTDPMKVILYIAMLELEKYIMRAAAAYALGKLGDLRAVPPLIKALKDEDAIVRAAAADALGKIGDLKAVP
PLIKALKDEDGAVRVSAAVALGKIGDLKAVEPLIKALKDEDAVVRVAAAIALGKIGDLRAVEPLIKALID
EKGKVQEAAALALGAIGGERVREAMEKLAEEGKGRARLYAVKYLGEHDAE
T33-ml23-A
MAIIETTTPTEEEAKAIAKKLLENRLIAEAIITPALTKIYRENGEIKSETVTRVTLYTEEENVPKAVTYI
KAIHPDPIPPIIVITPTDANPAYKGWVAFET
T33-ml23-B
MHHHHHHGGSHHWGGDPERPALGILELSSYARGVKVADAALKAAPVKLLKCEPVEPGRALIMLLGEPEDV
AKAMIAALDVAGLGSGNLIDYALIPEIHPQLLPFLKEYKKSEPIKDPNKAIIVAEVSTVAAAIEAADVAL
RLANVELTSMRLAEHIGGRASFTLIGDKEDVEKAARAIRGVAGERLLDLEIIEKPVEALIGNEFF

T33-ml24-A
MALIYTTTPTYEDAMNIAKKLLENKLIAYALIFSNITSVYVEEDEIHNNTECAVIMATVEEKVLKATLYI
EAIHPKDMPPIIVIVPADVSPRFQGWVYAKT
T33-ml24-B
MHHHHHHGGSHHWGGPARPALGVLLLTSIARGITVTDAALKAAPVRLLMSRPVCSGKHLNIFTGRPEEVL
TAMFAALETAGLGSGKLLDYAFIPALHPQLLRFLDAPVVADAWEEDTEAVAVVETTTPCAAIESADVALK
LAPVRLRDLRLAIGIAGKAYFTLAGREEDVRRAAKAVKGTAGDKLIELEFIARPVDELRGRLFF
T33-ml25-A
MDGEDALAAATAAEVAALLAILEAGLAALKALGFPLPDETGLDNRFLKALADWLRTEKALLTLREEALLR
LLRLLVERSA
T33-ml25-B
MHHHHHHGGSHHWGGEPDRPALGVLLLASIALGRAVADAALKAAPSLLLMSRPVCPGKHLIMMRGQVAEV
ETAMAAALATAGAGSGNLLDSAELPYAHEQLWRFLDAPVVADAWEEDTLAVLVVETATPCAAIRAADAAL
KTAPVTLRDMRLAIGIAGKAWFTLAGDPLAVLRAAVTVVAVAGDRLLRLEFIERPVDELRGRLFF
T33-ml26-A
MPVLTIATNVPSEAVPEGALLGLTLMLSELLGVPPEEIAVQITPDQRMVFGGSSEPCAICELKSIGKINA
EKNKELSAALTEFLERALGIPPERVLILFHNVKKENWGRNGGVFAGGSHHWGGHHHHHH
T33-ml26-B
MRMKYKVVIVTGVPGVGKSTVLKELEKIAKEKGIKIAVFDFFDYMLELAKKDGLVTKKDDIPFLPLDVLK
KLMKEAAKKIVEEAEKLLDEDGILLIDTQAVIKTNHGYVPGLPKFVMDVLKPDIIAVVEASPLDITRRML
ADTSRRLAYMGGGPGVAELMETERAAAIAAAIHTGAAVLFVRNAPGMERRAAERLLKAILNL
T33-ml27-A
MHHHHHHGGSHHWGPAGEPDRPALGVLELASIALGVAVADAALKAAPVLLLMARPVCSGKFLLVLRGEPE
AVRAAMEAALRTAGVGSGNLLDFLFLPAVHEQLLRFLDAPVVADALEDPDLALLVAETATPCAAIAAADA
ALKTAPVRLVDLRLAIGIAGKAWFVLAGAEEDVLRAALVVARVAGDRLLDLRFLPAPHDELRGRLFF
T33-ml27-B
MIKLSADKETVLVHGQELSTKFFLEVVVQTQLLAAGTNTALATQILALVLAAGLPVDDYGAYSRAFATGD
PALRAAAERVRAKAEAEREAMAAIHATPEEIAKAVAERKAREEALIKRFGNKGAAFGL
T33-ml28-A
MHHHHHHGGSHHWGGDPARPALGVLELKSYALGVAVADAALRAAPVELLKCEPVEPGKALIMIRGEPEAV
ARAMAAALETAKAGSGNLIDHAFIGRIHPALLPFLLEETAAPPIEDPDEAVLVVETKTVAAAIEAADAAL
DVAPVRLLRMRLSEHIGGKAYFVLAGDEEAVRKAARAVRAVAGEKLIDLRIIPRPHEALRGRLFF
T33-ml28-B
MPIALTVVPPEEAEPLARELVEAGLAAEVLLVPVRRIYREKGKVREEEVTLLLILVSREGVPALRAWIEA
RHPDDIPLFIVLAVDEEASNKRYLGYIAAETHLYSA
T33-ml29-A
MAPARPALLVLELSSYALGVEVADAALKAAPVELLLCRPIEPGKALIMLTGEPEAVEAAMKAALETAQEG
SGNLIASLFIPAIHPALLPFLLEPVRAPPLADPDEALLVAETSTVPAAIRAADEALRAAPVTLVRMDLAE
HIGGKASFVLTGELEDVVRAARVVVEVAGEDLIDLRIIPRPVAALRGRLFF
T33-ml29-B
MPMLIVYVPEGFSKAQKRQLLLLLHLAVVEALGVPLENVSIILTTVEPEDVLLGGKIGRPLAVVLVYILE
GLSPEQKAALIKALTEAVAKALGMDPENVSVIIVEVKPENFGVGNGKSAKEAGGSHHWGGHHHHHH

T33-ml30-A
MGPDEPERPALLILELKSYARGVRVADAALKAAPVRLLKCKIVEPGKALIMLTGRPEDVEKAYKAALTVA
NKGSGNLIDSVFIPAIHPALLPFLLEETPAPPLEDPDRALLFVEVKTVAAAIRAADAALRAAPVELVRMR
LSEHIGGKAVFALVGDPADVLRAAAVVAEVAGDQLLDIAIIPRPHPALLGREFF
T33-ml30-B
MPMLVVYVPEGYSEAQKRALLFRLAAAVVEATGTPLENVRIILTTYAPADVLLGGAIGVPLVVILVYLLE
GLSPEQKAALVKALTAAAAEALGVDPENIRVILVPVPPENFGVGNGKTAAEAGGSHHWGGHHHHHH
T33-ml31-A
MHHHHHHGGSHHWGGPAGEPDRPALGVLLLKSYARGVAVADAALKAAPSLLLMNRPVCPGKHLLMMRGQV
AEVEEAMRAALEEAGEGSGQLLASAFIPYAHEQLWRFLDAPVVADAIEEPDLAVAVVETKTPCAAIRAAD
AALKAAPVVLRDMRLAIGINGKARFTLEGKLVDVLEAAAVVIEVAGDDLISLSIIPRPHDELRGRWFF
T33-ml31-B
MADFHEQMATMFKNLAKILKAKNAAEVKDALKEMRKAALAAHKEVPPSLKDKPLNSQEMIEFHDEMLELA
WAIHDAAHLAKEGKIEEAKKKAEEILKMVSRLVSLY
T33-ml32-A
MHHHHHHGGSHHWGGMSISYRKLDIALSADKKTVLVFGQELSTKYFTEIVVTTMLNSTGSDMANSNRILN
DIHAAGLDAGDYGKYSRWWAQSNAQERQEAERRRKEAKAHQERLRAEKATVAAQLAAAAARLAEMRRLRE
RFGEAGIAAGL
T33-ml32-B
MNLAEKMYNAGQAMYRKGQYTIAIIAYTLSLLKDPKNAEAWYNLGQAYYKKGQYLDAIESYLKALTLDSS
NAEAWYNLGQAYYKLGHYEEAIEAYEKALALDPNNAEAKQNLGNAKQKLGLE
T33-ml33-A
MHHHHHHGGSHHWGGMSISYRKLDIALSADGREVLVFGQVLKTTFFKNIVVTTMLNSTGSDMANSNRILN
DIHAAGLDAGDYGKYSRWWAQSNAQERQEAERRRKEAKAHRAARRAALSTPEALAAATAEIEAERAALGA
RFGPAGLDAGL
T33-ml33-B
MRMFFKVVVVTGVPGVGKTTVIKELQGLAEKEGIKLYVVDFEDVMLEEAVARGLVEDRDKIRTLPLDILR
ELQKLAALRIRREALLALGASGILVVDTHALVKTVAGYYPGLPKFVMDILKPDMIAVVEASPEEVAARQA
RDTTRYRVDIGGVEGVKRLMENARAASIASAIQYASTVAIVENREGEAAKAAEELLRLIKNL
T33-ml34-A
MHHHHHHGGSHHWGGMSISYRKLDIALSADGEEVLVDGQVLPTRFFLDTVVTTMLNACGTDEENINEILA
DVHAAGLDVSNYGWASEVYKKGDPEKRAEAEARRAEAEARRAERRRRLASPEARRERRREEAERRLRLYE
RFGEAGLEAGL
T33-ml34-B
MTTEEEVVLAIAELFLPDPHARAEAAKKLGKIGDPEAVPALIRALFDPDPAVRAAAAKALGKIGDKEAVP
ALIVALFDPDPAVRVAAAKALGKIGDKEAVPALIEALFDPDPAVRVAAAIALGKIGDKEAVPALVRALKY
EEGLVREAAAIALKKIGGEEVKKAMEELAKFGEGEAKEFAEEYLKEN
T33-ml35-A
MNLAEKMYKAGQIEFAKGNYETAIIAYTLALLKDPNNAEAWYNLGEAYLALGNYEEAIEAYQKALELDPN
NAEAWYNLGEAYLALGDYDNAIEAFTKALELDPNNKTAKAGLKLAKEKKALE
T33-ml35-B

MTDLSSLIETADLRLLLTTVPTETEALYLALAAVEKGLAAEVLITPVTRVRRENGKLVVEDVYRLSFKTT
RERLDALVAWLQRRHPLALPECLVLTPIASSVAYRDWLRSSLQGGSHHWGGHHHHHH
T33-ml36-A
MNMARDFYRAGLIAYAKGEYETAIVAFQLALLLDPNNAEAWYNLGKAYYALGLYREAIEAYKKALELDPN
NAEAWYNLGKAYYALGDYESAIEAYKKALELDPNNVEAHANLHKAKKKLALE
T33-ml36-B
MPSYAVSSRAGLIDQERRAAVADLITALHSEILKIPRYLVQVIFNDLDAGALFLAGREAPEGHVWIHADI
ISGRTKEQKKAFLQALTVEVARVLGLPEEQVWVYVNEIPGENMTLFGQILPAPGEEEAWFATLPEELQKR
LADLRGGSHHWGGHHHHHH
T33-ml37-A
MNLANDFYEAGKEEFAKGRYNLAIVCFSLALLKDPNNAEAWYNLGKAYFALGKYDKAIEAYQKALELDPN
NAEAWYNLGLAYFALGNYKEAIEYYKKALELDPNNELAKLALKLAKEKLELE
T33-ml37-B
MAMPAVKLVIVTEKILLKDITRIILESGAKGFTVMNTGGIGSRERAGEGEPDIDKIRANIKFEVLCESRE
LAELIAEAIASKFFDKYAGIIYTCSAEVLYGHDFCGPEGSGSHHWGGHHHHHH
T33-ml38-A
MHHHHHHGGSHHWGGMNLRAAGPGWLFCPAHRPELFAKAAAAADVVILDLEDGVAESMKPGARENLRAHP
LDPERTVVRINAGGTADQARDLEALAGTAYTTVMLPKAESAAQVIELAPRDVIALVETARGAVCAAEIAA
ADPTVGMMWGAEDLIATLGGSSSRRADGAYRDVARHVRSTILLAASAFGRLALDAVHLDILDVEGLQEEA
RDAAAVGFDVTVCIHPSQIPVVRKAYRPSHEKLEWARLVLLNAQGKAGAFVFEGQMVDSPVLTHAETMLR
RAGEATSE
T33-ml38-B
MPSYAVSSRAGLIDRLRRLEVARLLTTLHRDIAVAPRYLVQVIFNDLDAGALFVAGAEAPEGHVWIHADI
RSGRTAQQKTDLLEQITSKVADVLELPPEHVWVYVNEIPGENMTEYGKLLPEPGKEEEWFATLPPGLQTV
LSA

References

1. Sahtoe, D. D. *et al.* Reconfigurable asymmetric protein assemblies through implicit negative design. *Science* **375**, eabj7662 (2022).

2. Wicky, B. I. M. et al. Hallucinating symmetric protein assemblies. Science eadd1964 (2022) doi:10.1126/science.add1964.

3. Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* 1–3 (2023) doi:10.1038/s41586-023-06415-8.

4. Padilla, J. E., Colovos, C. & Yeates, T. O. Nanohedra: Using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proceedings of the National Academy of Sciences* **98**, 2217–2221 (2001).

5. Laniado, J. & Yeates, T. O. A complete rule set for designing symmetry combination materials from protein molecules. *Proc National Acad Sci* 202015183 (2020) doi:10.1073/pnas.2015183**117**.

6. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 1–11 (2021) doi:10.1038/s41586-021-03819-2.

7. Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. Science **378**, 49–56 (2022).

8. Ingraham, J. et al. Illuminating protein space with a programmable generative model. *bioRxiv* 2022.12.01.518682 (2022) doi:10.1101/2022.12.01.518682.

9. Madani, A. et al. Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* 1–8 (2023) doi:10.1038/s41587-022-01618-2.

10. Ben-Sasson, A. J. et al. Design of biologically active binary protein 2D materials. Nature **589**, 468–473 (2021).

11. Marcandalli, J. et al. Induction of Potent Neutralizing Antibody Responses by a Designed Protein Nanoparticle Vaccine for Respiratory Syncytial Virus. *Cell* **176**, 1420-1431.e17 (2019).

12. Miller, J. E. et al. Design of Beta-2 Microglobulin Adsorbent Protein Nanoparticles.

Biomolecules **13**, 1122 (2023).

13. Martin, J. T. et al. Targeting HIV Env immunogens to B cell follicles in nonhuman primates through immune complex or protein nanoparticle formulations. *Npj Vaccines* **5**, 72 (2020).

14. Arunachalam, P. S. et al. Adjuvanting a subunit COVID-19 vaccine to induce protective immunity. Nature **594**, 253–258 (2021).

15. Boyoglu-Barnum, S. et al. Quadrivalent influenza nanoparticle vaccines induce broad protection. Nature 1–6 (2021) doi:10.1038/s41586-021-03365-x.

16. Heater, B. S., Yang, Z., Lee, M. M. & Chan, M. K. In Vivo Enzyme Entrapment in a Protein Crystal. *J Am Chem Soc* (2020) doi:10.1021/jacs.9b13462.

17. McConnell, S. A. et al. Designed Protein Cages as Scaffolds for Building Multienzyme Materials. *Acs Synth Biol* (2020) doi:10.1021/acssynbio.9b00407.

18. Ernst, P., Plückthun, A. & Mittl, P. R. E. Structural analysis of biological targets by host:guest crystal lattice engineering. *Sci Rep-uk* **9**, 15199 (2019).

19. Castells-Graells, R. et al. Cryo-EM structure determination of small therapeutic protein targets at 3 Å-resolution using a rigid imaging scaffold. *Proc. Natl. Acad. Sci. United States Am.* **120**, e2305494120 (2023).

20. Höfte, H. & Whiteley, H. R. Insecticidal crystal proteins of Bacillus thuringiensis. *Microbiol. Rev.* **53**, 242–255 (1989).

21. Kerfeld, C. A. et al. Protein *Structure*s Forming the Shell of Primitive Bacterial Organelles. Science **309**, 936–938 (2005).

22. Erbse, A. H. & Falke, J. J. The Core Signaling Proteins of Bacterial Chemotaxis Assemble To Form an Ultrasfigure Complex. *Biochemistry-us* **48**, 6975–6987 (2009).

23. Shin, S.-H. et al. Direct observation of kinetic traps associated with structural transformations leading to multiple pathways of S-layer assembly. Proc National Acad Sci **109**, 12968–12973 (2012).

24. Cameron, J. C., Wilson, S. C., Bernstein, S. L. & Kerfeld, C. A. Biogenesis of a Bacterial

Organelle: The Carboxysome Assembly Pathway. Cell **155**, 1131–**114**0 (2013).

25. Pastuzyn, E. D. et al. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. Cell (2018) doi:10.1016/j.cell.20**17**.12.024.

26. Wargacki, A. J. et al. Complete and cooperative in vitro assembly of computationally designed self-assembling protein nanomaterials. *Nature Communications* (2021) doi:10.1038/s41467-021-21251-y.

27. Courbet, A. et al. Computational design of mechanically coupled axle-rotor protein assemblies. Science **376**, 383–390 (2022).

28. Ueda, G. et al. Tailored design of protein nanoparticle scaffolds for multivalent presentation of viral glycoprotein antigens. *eLife* (2020) doi:10.7554/elife.57659.

29. Lai, Y.-T. et al. Structure of a designed protein cage that self-assembles into a highly porous cube. *Nature Chemistry* **6**, nchem.2107 (2014).

30. Gainza, P. et al. De novo design of protein interactions with learned surface fingerprints. Nature **617**, 176–184 (2023).

31. Levy, E. D. A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution. *J Mol Biol* **403**, 660–670 (2010).

32. Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L. & Tawfik, D. S. The Stability Effects of Protein Mutations Appear to be Universally Distributed. *J. Mol. Biol.* **369**, 1318–1332 (2007).

33. Laniado, J., Meador, K. & Yeates, T. O. A fragment-based protein interface design algorithm for symmetric assemblies. *Protein Eng., Des. Sel.* **34**, gzab008 (2021).

34. Tsai, C.-J., Xu, D. & Nussinov, R. Structural motifs at protein‐protein interfaces: Protein cores versus two‐state and three‐state model complexes. *Protein Sci* **6**, 1793–1805 (1997).

35. Fleishman, S. J. et al. Community-Wide Assessment of Protein-Interface Modeling Suggests Improvements to Design Methodology. J Mol Biol **414**, **2**89–302 (**20**11).

36. Stranges, P. B. & Kuhlman, B. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. Protein Sci **22**, 74–82

(2013).

37. Zhou, J. & Grigoryan, G. Rapid search for tertiary fragments reveals protein sequence–structure relationships. Protein Sci **24**, 508–524 (2015).

38. Alford, R. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **13**, 3031–3048 (2017).

39. Maguire, J. B. et al. Perturbing the energy landscape for improved packing during computational protein design. *Proteins Struct Funct Bioinform* (2021) doi:10.1002/prot.26030.

40. Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. A Dissection of Specific and Non-specific Protein–Protein Interfaces. J Mol Biol **336**, 943–955 (2004).

41. Cannon, K. A. et al. Design and structure of two new protein cages illustrate successes and ongoing challenges in protein engineering. Protein Sci **29**, 919–929 (2020).

42. Maguire, J. B., Boyken, S. E., Baker, D. & Kuhlman, B. Rapid Sampling of Hydrogen Bond Networks for Computational Protein Design. *J Chem Theory Comput* **14**, 2751–2760 (2018).

43. Colovos, C. & Yeates, T. O. Verification of protein structures: Patterns of nonbonded atomic interactions. Protein Sci 2, 1511–1519 (1993).

44. Lawrence, M. C. & Colman, P. M. Shape Complementarity at Protein/Protein Interfaces. J Mol Biol **234**, 946–950 (1993).

45. Fleishman, S. J., Khare, S. D., Koga, N. & Baker, D. Restricted sidechain plasticity in the structures of native proteins and complexes. Protein Sci 20, 753–757 (2011).

46. Jha, R. K. et al. Computational Design of a PAK1 Binding Protein. J. Mol. Biol. **400**, 257–270 (2010).

47. Ingraham, J., Garg, V. K., Barzilay, R. & Jaakkola, T. Generative models for graph-based protein design. *Advances in neural information processing systems* (2019).

48. Bennett, N. R. et al. Improving de novo protein binder design with deep learning. *Nat. Commun.* **14**, 2625 (2023).

49. Sudha, G., Singh, P., Swapna, L. S. & Srinivasan, N. Weak conservation of structural

features in the interfaces of homologous transient protein–protein complexes. Protein Sci 24, 1856–1873 (2015).

50. Dey, S. & Levy, E. D. PDB-wide identification of physiological hetero-oligomeric assemblies based on conserved quaternary structure geometry. Structure (2021) doi:10.1016/j.str.2021.07.012.

51. Guharoy, M. & Chakrabarti, P. Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein–protein interactions. *Bioinformatics* **23**, 1909–1918 (2007).

52. Tanaka, S., Sawaya, M. R. & Yeates, T. O. Structure and Mechanisms of a Protein-Based Organelle in Escherichia coli. Science **327**, 81–84 (2010).

53. Ochoa, J. M. et al. Symmetry Breaking and Structural Polymorphism in a Bacterial Microcompartment Shell Protein for Choline Utilization. Protein Sci (2020) doi:10.1002/pro.3941.

54. Endres, D. & Zlotnick, A. Model-Based Analysis of Assembly Kinetics for Virus Capsids or Other Spherical Polymers. *Biophys. J.* **83**, 1217–1230 (2002).

55. Spall, V. E., Shanks, M. & Lomonossoff, G. P. Polyprotein Processing as a Strategy for Gene Expression in RNA Viruses. *Semin. Virol.* **8**, 15–23 (1997).

56. King, N. P. et al. Accurate design of co-assembling multi-component protein nanomaterials. Nature **510**, nature13404 (2014).

57. Bale, J. B. et al. Accurate design of megadalton-scale two-component icosahedral protein complexes. Science **353**, 389–394 (2016).

58. Vanhee, P. et al. Protein-Peptide Interactions Adopt the Same Structural Motifs as Monomeric Protein Folds. Structure 17, 1128–1136 (2009).

59. Zhou, J., Panaitiu, A. E. & Grigoryan, G. A general-purpose protein design framework based on mining sequence–structure relationships in known protein structures. Proc National Acad Sci 117, 1059–1068 (2020).

60. Haas, R. J. de et al. Rapid and automated design of two-component protein nanomaterials

using ProteinMPNN. bioRxiv 2023.08.04.551935 (2023) doi:10.1101/2023.08.04.551935.

61. Adolf-Bryfogle, J. et al. RosettaAntibodyDesign (RAbD): A general framework for computational antibody design. *PLoS Comput. Biol.* **14**, e1006112 (2018).

62. Foley, G. et al. Engineering indel and substitution variants of diverse and ancient enzymes using Graphical Representation of Ancestral Sequence Predictions (GRASP). *PLOS Comput. Biol.* **18**, e1010633 (2022).

63. Dey, S., Ritchie, D. W. & Levy, E. D. PDB-wide identification of biological assemblies from conserved quaternary structure geometry. *Nature Methods* **15**, 67 (2018).

64. Nivón, L. G., Moretti, R. & Baker, D. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *Plos One* **8**, e59004 (2013).

65. DiGiacomo, J., McKay, C. & Davila, A. ThermoBase: A database of the phylogeny and physiology of thermophilic and hyperthermophilic organisms. *PLoS ONE* **17**, e0**26**8253 (2022).

66. Mitternacht, S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000research* **5**, 189 (2016).

67. Frishman, D. & Argos, P. Knowledge‐based protein secondary structure assignment. Proteins Struct Funct Bioinform 23, 566–579 (1995).

68. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics **29**, 2722–2728 (2013).

69. Wruck, F., Katranidis, A., Nierhaus, K. H., Büldt, G. & Hegner, M. Translation and folding of single proteins in real time. Proceedings of the National Academy of Sciences 114, E4399–E4407 (2017).

70. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**, 173–175 (2012).

71. Gao, M. & Skolnick, J. iAlign: a method for the structural comparison of protein–protein interfaces. Bioinformatics 26, 2259–2265 (2010).

72. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).

73. King, C. et al. Removing T-cell epitopes with computational protein design. Proc. Natl. Acad. Sci. **111**, 8577–8582 (2014).

74. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D. & André, I. Modeling symmetric macromolecular structures in Rosetta3. Plos One 6, e20450 (2011).

75. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

76. Studier, F. W. Protein production by auto-induction in high-density shaking cultures. *Protein Expression and Purification* **41**, 207–234 (2005).

77. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).

78. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nat. Methods **14**, 290–296 (2017).

79. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. Sect. D: Biol. Crystallogr.* **66**, 486–501 (2010).

80. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr. Sect. D **75**, 861–877 (2019).

81. Pettersen, E. F. et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

82. Goddard, T. D. et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Sci. **27**, 14–25 (2018).

83. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

84. Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. Sect. D: Struct. Biol.* **74**, 531–**54**4 (2018).

85. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. Acta Crystallogr. Sect. D: Biol. Crystallogr. **60**, 2126–2132 (2004).

86. Manalastas-Cantos, K. et al. ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. J. Appl. Crystallogr. 54, 343–355 (2021).