

Hallucinating symmetric protein assemblies

B. I. M. Wicky^{1,2,†}, **L. F. Milles**^{1,2,†}, **A. Courbet**^{1,2,3,†}, **R. J. Ragotte**^{1,2}, **J. Dauparas**^{1,2}, **E. Kinfu**^{1,2}, **S. Tipps**^{1,2}, **R. D. Kibler**^{1,2}, **M. Baek**^{1,2}, **F. DiMaio**^{1,2}, **X. Li**^{1,2}, **L. Carter**^{1,2}, **A. Kang**^{1,2}, **H. Nguyen**^{1,2}, **A. K. Bera**^{1,2}, **D. Baker**^{1,2,3,*}

¹Department of Biochemistry, University of Washington, Seattle, WA, USA.

²Institute for Protein Design, University of Washington, Seattle, WA, USA.

³Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.

Abstract

Deep learning generative approaches provide an opportunity to broadly explore protein structure space beyond the sequences and structures of natural proteins. Here we use deep network hallucination to generate a wide range of symmetric protein homo-oligomers given only a specification of the number of protomers and the protomer length. Crystal structures of 7 designs are very close to the computational models (median RMSD: 0.6 Å), as are 3 cryoEM structures of giant 10 nanometer rings with up to 1550 residues and C33 symmetry; all differ considerably from previously solved structures. Our results highlight the rich diversity of new protein structures that can be generated using deep learning, and pave the way for the design of increasingly complex components for nanomachines and biomaterials.

Cyclic protein oligomers play key roles in almost all biological processes and constitute nearly 30% of all deposited structures in the Protein Data Bank (PDB, (1) (2–4)). Because of the many applications of cyclic protein oligomers, ranging from small molecule binding and catalysis to building blocks for nanocage assemblies (5), *de novo* design of such structures has been of considerable interest from the beginning of the protein design field

This work is licensed under a Creative Commons Attribution 4.0 International License, which allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use.

*Corresponding author. dabaker@uw.edu.

†These authors contributed equally to this work.

Author contributions: Conceptualization: AC, BIMW, LFM, DB

Methodology: AC, BIMW, LFM, DB

Software: AC, BIMW, LFM, JD, MB, FD, RDK

Validation: AC, BIMW, LFM, ST, EK, RJR, AKB

Formal analysis: AC, BIMW, LFM, DB, RJR

Investigation: AC, BIMW, LFM, ST, EK, XL, LC, AKB, AK, HN

Resources: AC, BIMW, LFM, MB, FD, DB

Data curation: AC, BIMW, LFM, DB, AKB, RJR, XL, LC

Writing – original draft: AC, BIMW, LFM, DB

Writing – review & editing: AC, BIMW, LFM, DB

Visualization: AC, BIMW, LFM, RJR

Supervision: DB

Project administration: AC, BIMW, LFM, DB

Funding acquisition: AC, BIMW, LFM, DB

Competing interests: BIMW, LFM, AC, RJR, JD, EK, ST, RDK, and DB are inventors on a provisional patent application submitted by the University of Washington for the design, composition and function of the proteins created in this study.

(6, 7). While there have been a number of successes (8–10), current approaches typically require specification of the structure of the protomers in advance. With the exception of parametrically designed structures (11, 12), design strategies involve rigid body docking of characterized monomers into higher order symmetric structures, followed by interface optimization to generate low-energy assembled states (13–17). The requirement that the protomer structure be specified in advance has limited the exploration of the full space of oligomeric structures, such as assemblies with more intertwined chains. For monomeric protein design, broad exploration of the space of possible structures has become possible by deep network hallucination: starting from a random amino acid sequence, Markov chain Monte Carlo (MCMC) optimization favoring folding to a well-defined state converges on new sequences that fold to new structures (18–21). By extension, we reasoned that deep network hallucination could enable the design of higher-order protein assemblies in one step, without prespecification or experimental confirmation of the structures of the protomers, provided that a suitable loss function specifying both protomer folding and assembly could be formulated (18–20, 22–25).

We set out to broadly explore the space of cyclic protein homo-oligomers by developing a method for hallucinating such structures that places no constraints on the structures of either the protomers or the overall assemblies. Starting from only a choice of chain length L and oligomer valency N (2 for a dimer, 3 for a trimer, etc.), the method carries out a Monte Carlo search in sequence space starting from a random sequence (Fig. 1A). The loss function guiding the search is computed by inputting N copies of the sequence into the AlphaFold2 (AF2) network (26), and combining structure prediction confidence metrics (pLDDT; per-residue structural accuracy (27), and pTM; an estimate of the TM-score (28)) with a measure of cyclic symmetry (the standard deviation of the distances between the center of mass of adjacent protomers within the predicted structure).

We found that monomers and dimeric to heptameric assemblies could readily be generated by this procedure for chains of 65 to 130 amino acids, with converging trajectories typically coalescing to cyclic homo-oligomeric structures within a few hundred steps (approximately 1 to 7 days of CPU-time for monomers to heptamers respectively, Fig. S1–2). The resulting structures are topologically diverse, spanning all- α , mixed α/β and all- β structures, and differ from the structures of cyclic *de novo* designs present in the PDB (Fig. 1B). These assemblies, which we term HALs, also differ from natural proteins in both structure (Fig. 1C) and sequence (Fig. 1D), with the median closest relatives in the PDB having TM-scores of 0.67 and 0.57 for the protomers and oligomers respectively (29% of the structures have TM-scores < 0.5 , the cutoff for fold assignment in CATH/SCOP (29)), indicating considerable generalization beyond the PDB training set.

We selected 150 designs with AF2 pLDDT > 0.7 and pTM > 0.7 for experimental testing. However, virtually none showed significant soluble expression when produced in *E. coli* (median soluble yield: 9 mg per liter of culture-equivalent, Fig. S3), and of the few that were marginally soluble none had both the expected oligomerization state by size-exclusion chromatography (SEC), and a circular dichroism (CD) profile consistent with the hallucinated structure. We speculated that this failure could be a consequence of over-fitting during MCMC optimization leading to the generation of adversarial sequences,

i.e. confidently-predicted sequences with unrealistic biophysical properties (Fig. S4–5). Adversarial samples have been generated by activation maximization in the context of image classification neural networks, which similarly leads to unrealistic outputs (30–32). To eliminate such over-fitting, we generated new sequences for the HAL backbones using the recently developed ProteinMPNN sequence design neural network (*accompanying manuscript: Dauparas et al.*). For each original backbone, 24 to 48 sequences were generated with ProteinMPNN, and assembly to the target oligomeric structure validated with AF2 (these dozens of evaluations compared to the hundreds performed during hallucination make overfitting much less likely). In addition, we independently evaluated the sequences using an updated version of RoseTTAFold (RF2) (33), and found that RF2 did not confidently predict the structure of most of the original AF2 hallucinated sequences, but successfully predicted almost all ProteinMPNN sequences (Fig. S4, S6–7).

We tested 96 ProteinMPNN-designed HALs with pLDDT > 0.75 and root-mean-square deviation (RMSD) to original backbone < 1.5 Å and found that 71/96 (74%) were expressed to high levels (median yield: 247 mg per liter of culture-equivalent), 50/96 (52%) had a SEC retention volume consistent with the size of the oligomer (of which 30 (60%) were monodisperse) (Fig. 1F and Fig. S8–9), and at least 21/96 (22%) had the correct oligomeric state when assessed by SEC-Multi Angle Light Scattering (SEC-MALS) (Fig. 1G and Fig. S10). CD analysis of the soluble samples indicated that 67/71 (96%) had secondary structure contents consistent with the designs (Fig. S9). These success rates are in stark contrast to those of the original AF2 hallucinated sequences, indicating that the MCMC procedure generates viable backbones with over-fitted sequences exhibiting various pathologies (Fig. S5), and highlights the power of ProteinMPNN to generate sequences which fold to a given backbone structure (Fig. 1E). We assessed the thermal stability of the 71 soluble HALs by CD spectroscopy, and found that 54 maintained their secondary structure up to 95 °C (Fig. S9). SEC characterization of the heated-treated samples indicated that most designs retained their oligomeric state, suggesting that ProteinMPNN-designed HALs are thermostable (Fig. 1H, S9).

To evaluate design accuracy we attempted crystallization of 19 designs and succeeded in solving crystal structures for seven (three C2s, two C3s and two C4s, Fig. 2). All crystal structures had the correct oligomerization state and closely matched the design models (median C α RMSD of 0.6 Å across all designs, with resolutions ranging from 1.8 to 3.4 Å, Fig. S11, Table S1). The side chain conformations in the crystal structures also closely match those of the design models (Fig. 2).

The solved structures exhibit striking diversity with many intricate structural features. HALC2_062 (Fig. 2A) is a three-layer homo-dimer with a single helix from each protomer packed together between two outer β -sheets (one from each protomer), while HALC2_065 (Fig. 2B) is also a mixed α/β homo-dimer, but has a single, continuous β -sheet shared between both chains, which wraps around two perpendicular paired helices. These two hallucinated structures are distinct from any structure in the PDB, with TM-scores to their best matches of 0.59 and 0.54 respectively (Fig. 4A–B, Table S2). HALC2_068 (Fig. 2C) is a fully helical dimer with an extensive interface formed by 6 interacting helices (3 from each protomer), with a single perpendicular helix buttressing the interfacial helices. Despite

the low secondary structure complexity and absence of long-range contacts, this design also differs significantly from its closest structural relative in the PDB (TM-score: 0.57, Fig. 4C, Table S2). HALC3_104 (Fig. 2D) is a homo-trimeric coiled-coil, with a central bundle of three helices, augmented by an outer-ring of three shorter helices that lie in the groove formed by adjacent protomer (the closest matching structure in the PDB has a TM-score of 0.88, Fig. 4D, Table S2). HALC3_109 (Fig. 2E) is a homo-trimeric three-layer all-helical structure, with three inner helices splaying outwards to contact two additional helices from the same protomers at angles of roughly 25° and 90°; the closest assembly in the PDB has a TM-score of 0.69 (Fig. 4E, Table S2). HALC4_135 (Fig. 2F) is a coiled-coil composed of helical hairpins reminiscent of HALC3_104, but with C4 symmetry instead of C3, and a discontinuous superhelical twist. Despite its simple topology, the closest structural homologue to this design has a TM-score of only 0.59 (Fig. 4F, Table S2). HALC4_136 (Fig. 2G) is composed of 3-helix protomers with eight outer helices encasing four almost fully hydrophobic inner helices, where two of the helices are rigidly linked through a 90° helical kink. The closest match in the PDB has a TM-score of 0.71, but the matched structure has C5 symmetry rather than the C4 symmetry of the design and crystal structure (Fig. 4G, Table S2).

Next, we sought to generate HALs of greater complexities across longer length-scales by extending the design specifications to structures of higher symmetry (up to C42) and longer oligomeric assembly sequence lengths (up to 1800 residues). To generate multiple possible oligomers from a single structure, we specified the MCMC trajectories as single-chains with internal sequence symmetry; the resulting structure-symmetric repeat proteins can be split into any desired oligomeric assembly compatible with factorization (e.g. C15 into a pentamer, shorthand as C15-5). To maximize the exploration of the design space while minimizing the use of computational resources, we devised an evolution-based computational strategy: many short MCMC trajectories (< 50 steps) outputs were clustered by structure prediction confidence metrics (pLDDT and pTM), and then used to seed new trajectories (see Supplementary Materials). Using this approach, we hallucinated cyclic homo-oligomers from C5 to C42 with their largest dimension ranging from 7 to 14 nm (median: 10 nm), which were then divided into homo-trimers, tetramers, pentamers, hexamers, heptamers, octamers, and dodecamer, and the backbones were re-designed with ProteinMPNN (Fig 1C). While the α/β topology of some of these larger HALs is reminiscent of natural Leucine Rich Repeats (LRRs, (34)), which is reflected by a median highest protomer TM-scores of 0.64, these ring-shaped structures differ considerably from the horseshoe folds of LRRs that do not close into cyclic structures. The closest oligomer structures in the PDB have a median TM-score of 0.47, and BLAST sequence similarity searches for the repetitive sequence motif do not return any significant hits (Fig. 1D); the hallucination process as in the earlier cases generalizes beyond the training set.

These larger HALs have overall molecular weights greater than 100 kDa, and thus were well-suited for structural characterization by electron microscopy (EM). We screened soluble large HALs with a SEC retention volume consistent with the size of their oligomeric state by negative stain EM (nsEM), and in most cases observed monodisperse particles of the expected size and circular shape. We obtained 2D class averages and 3D *ab initio* reconstructed electron density maps for six designs with C6 to C42 internal repeat symmetry

(factorized as: two C5s, three C6s, and one C7) that clearly showed low-resolution structural features and diameters consistent with their designs (Fig. 3A, Fig. S12). We selected three designs: one C15 homo-pentamer (HALC5–15_262), one C18 homo-hexamer (HALC6–18_265) and one C33 homo-trimer (HALC3–33_343) for high-resolution single particle cryoEM characterization. We collected datasets that produced 2D class averages with clear secondary structure feature placements, and 3D *ab initio* reconstruction and refinement yielded 3D electron density maps at 4.38 Å, 6.51 Å and 6.32 Å resolution respectively (Fig. 3B, Fig. S13–16). HALC5–15_262 was originally designed as a homo-hexamer, but structure prediction calculations were more consistent with a pentameric structure of nearly identical protomer conformation and only a very slightly shifted subunit interface (Fig. S17); the cryoEM structure is also a pentamer with an C α RMSD of 1.69 Å to this predicted structure (Fig. S16).

These hallucinated rings are giant structures quite unlike anything in the PDB. The three rings solved by cryoEM, HALC5–15_262, HALC6–18_265 and HALC3–33_343, are 87 Å, 99 Å and 100 Å in diameter and 40 to 50 Å high, with a continuous parallel β -sheet in the lumen of the pore, and outer helices that enforce the curvature and closure of the ring. HALC3–33_343 has a simple helix-loop-sheet structural motif as its repeating unit, while in HALC5–15_262 and HALC6–18_265, the repeating unit contains two distinct helix-loop-sheet elements, which produces an alternating helical outer pattern clearly observable in the 2D class averages. While both structures have matches to LRRs for their protomers (TM-score of 0.65 for both, but to different structures), the oligomeric assemblies are strikingly different from any natural protein (TM-scores of 0.48 and 0.49 respectively, Fig. 4H–I, Table S2). HALC3–33_343 has an unusual internal loop region breaking the outer helices midway in the repeat, producing a widening of the ring on one side, which is clearly visible in the cryoEM reconstruction; the protomer has a low TM-score (0.48) despite having an LRR-like topology, and the oligomer is even further from anything currently known (TM-score: 0.41, Fig. 4J, Table S2) The high structural symmetry of these designed complexes rivals that of natural proteins: the highest cyclic symmetry recorded in the PDB for naturally occurring proteins is C39 (Vault proteins (35), PDB 4HL8 and 7PKY), and there are no closed symmetric α/β ring-like structures.

Conclusion

Our deep learning-based approach to designing cyclic homo-oligomers jointly generates protomers and their oligomeric assemblies without the need for a hierarchical docking approach. We report a rich assortment of *de novo* protein homo-oligomers across the nanoscopic scale, with broad topological diversity while maintaining design constraints such as symmetry and oligomeric state. These hallucinated oligomers differ substantially from natural oligomers in both sequence (median lowest BLAST E-value against UniRef100 of 1.3 for the repeated sequence motifs, Fig. 1D, Table S3)) and structure (median best TM-score between biounits from the PDB and HALS of 0.57, Fig. 1C, Table S2); our computational pipeline interpolates and extends native fold-space rather than simply recapitulating memorized protein structures, demonstrating the power of deep learning to explore previously uncharted regions of the design landscape (Fig. 1B). Our results also highlight the power of the ProteinMPNN method for protein sequence design; of the 30

out of the 192 designs evaluated experimentally by either SEC-MALS, nsEM, cryoEM, or X-ray crystallography, 27 had the intended oligomeric state, and 7 out of 19 for which crystallization was attempted formed diffracting crystals (this is a considerably higher crystallization success rate than typical for Rosetta *de novo* designs, and suggests that ProteinMPNN may generate protein surfaces more likely to form crystal contacts). More generally, our results show that a rich diversity of protein structures and assemblies beyond what exists in the PDB can now be accessed by deep learning-based generative models.

The formalism described here can be extended to other types of complex design tasks, including the design of higher order point group symmetries, arbitrary symmetric or asymmetric hetero-oligomeric assemblies, oligomeric scaffolding of existing functional domains, and design of multiple states, provided a loss function describing the solution can be formalized and computed. Computational requirements and hardware memory limitations become bottlenecks for hallucination of increasingly large structures; the development of computationally less expensive structure prediction methods with fewer parameters, as well as generative approaches such as diffusion models ((36, 37)) which more directly sample in structure space, should enable the design of even more complex protein structures and assemblies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Ivan Anishchenko, Sergey Ovchinnikov, William Sheffler, Jesse Hansen, Christoffer Norn, Dmitri Zorine, Luki Goldschmidt, and Timothy Huddy for helpful discussions.

Funding:

This work was supported with funds provided by the Audacious Project at the Institute for Protein Design (AK, LC, XL, EK, ST, DB), a grant from the National Institute of General Medical Sciences (P41 GM 103533-24, RDK), an EMBO long-term fellowship (ALTF 139-2018, BIMW), a grant from the National Science Foundation (CHE-1629214, DB), the Open Philanthropy Project Improving Protein Design Fund (HN, AB, RJR, JD, DB), an Alfred P. Sloan Foundation Matter-to-Life Program Grant (G-2021-16899, AC, DB), a Human Frontier Science Program Cross Disciplinary Fellowship (LT000395/2020-C, LFM), an EMBO Non-Stipendiary Fellowship (ALTF 1047-2019, LFM), and the Howard Hughes Medical Institute (AC, DB). CryoEM was performed on a Glacios microscope purchased via the University of Washington Arnold and Mabel Beckman cryoEM center (DB) with a S10 award (S10OD032290), and at the Fred Hutchinson Cancer Center Electron Microscopy Shared Resource (supported by Cancer Center Support Grant P30 CA015704-40). X-ray crystallography utilized the Northeastern Collaborative Access Team beamlines, funded by the National Institute of General Medical Sciences from the National Institutes of Health (P30 GM124165), and the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. Molecular graphics and analyses were performed with UCSF ChimeraX developed with support from NIH P41-GM103311. We thank Microsoft and AWS for generous gifts of cloud computing credits. We thank the IPD Breakthrough Fund for support for the “*Design of selective pores and channels for sensing, filtration, and sequencing*”.

Data and materials availability:

All data is available in the main text or as supplementary materials. Scripts and computational methods are available on GitHub (https://github.com/bwicky/oligomer_hallucination), Crystallographic datasets have been deposited in the PDB

(accession codes: 8D03, 8D04, 8D05, 8D06, 8D07, 8D08 8D09). EM maps have been deposited in the EMDB (accession codes: EMD-27658, EMD-27659, EMD-27660).

References

- Garcia-Seisdedos H, Empereur-Mot C, Elad N, Levy ED, Proteins evolve on the edge of supramolecular self-assembly. *Nature*. 548, 244–247 (2017). [PubMed: 28783726]
- Johnston IG, Dingle K, Greenbury SF, Camargo CQ, Doye JPK, Ahnert SE, Louis AA, Symmetry and simplicity spontaneously emerge from the algorithmic nature of evolution. *Proc. Natl. Acad. Sci* 119, e2113883119 (2022).
- Ahnert SE, Marsh JA, Hernández H, Robinson CV, Teichmann SA, Principles of assembly reveal a periodic table of protein complexes. *Science*. 350, aaa2245 (2015).
- wwPDB consortium, Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res*. 47, D520–D528 (2019). [PubMed: 30357364]
- Goodsell DS, Olson AJ, Structural Symmetry and Protein Function. *Annu. Rev. Biophys. Biomol. Struct* 29, 105–153 (2000). [PubMed: 10940245]
- Handel T, DeGrado WF, De novo design of a Zn²⁺-binding protein. *J. Am. Chem. Soc* 112, 6710–6711 (1990).
- Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS, High-Resolution Protein Design with Backbone Freedom. *Science*. 282, 1462–1467 (1998). [PubMed: 9822371]
- Fallas JA, Ueda G, Sheffler W, Nguyen V, McNamara DE, Sankaran B, Pereira JH, Parmeggiani F, Brunette TJ, Cascio D, Yeates TR, Zwart P, Baker D, Computational design of self-assembling cyclic protein homo-oligomers. *Nat. Chem* 9, 353–360 (2017). [PubMed: 28338692]
- Thomson AR, Wood CW, Burton AJ, Bartlett GJ, Sessions RB, Brady RL, Woolfson DN, Computational design of water-soluble α -helical barrels. *Science*. 346, 485–488 (2014). [PubMed: 25342807]
- Huang P-S, Feldmeier K, Parmeggiani F, Fernandez Velasco DA, Höcker B, Baker D, De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol* 12, 29–34 (2016). [PubMed: 26595462]
- Boyken SE, Chen Z, Groves B, Langan RA, Oberdorfer G, Ford A, Gilmore JM, Xu C, DiMaio F, Pereira JH, Sankaran B, Seelig G, Zwart PH, Baker D, De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science*. 352, 680–687 (2016). [PubMed: 27151862]
- Doyle L, Hallinan J, Bolduc J, Parmeggiani F, Baker D, Stoddard BL, Bradley P, Rational design of α -helical tandem repeat proteins with closed architectures. *Nature*. 528, 585–588 (2015). [PubMed: 26675735]
- Bale JB, Gonen S, Liu Y, Sheffler W, Ellis D, Thomas C, Cascio D, Yeates TO, Gonen T, King NP, Baker D, Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science*. 353, 389–394 (2016). [PubMed: 27463675]
- Vulovic I, Yao Q, Park Y-J, Courbet A, Norris A, Busch F, Sahasrabudhe A, Merten H, Sahtoe DD, Ueda G, Fallas JA, Weaver SJ, Hsia Y, Langan RA, Plückthun A, Wysocki VH, Veesler D, Jensen GJ, Baker D, Generation of ordered protein assemblies using rigid three-body fusion. *Proc. Natl. Acad. Sci* 118, e2015037118 (2021).
- Hsia Y, Mout R, Sheffler W, Edman NI, Vulovic I, Park Y-J, Redler RL, Bick MJ, Bera AK, Courbet A, Kang A, Brunette TJ, Nattermann U, Tsai E, Saleem A, Chow CM, Ekiert D, Bhabha G, Veesler D, Baker D, Design of multi-scale protein complexes by hierarchical building block fusion. *Nat. Commun* 12, 2294 (2021). [PubMed: 33863889]
- Correnti CE, Hallinan JP, Doyle LA, Ruff RO, Jaeger-Ruckstuhl CA, Xu Y, Shen BW, Qu A, Polkinghorn C, Friend DJ, Bandaranayake AD, Riddell SR, Kaiser BK, Stoddard BL, Bradley P, Engineering and functionalization of large circular tandem repeat protein nanoparticles. *Nat. Struct. Mol. Biol* 27, 342–350 (2020). [PubMed: 32203491]
- Sahtoe DD, Praetorius F, Courbet A, Hsia Y, Wicky BIM, Edman NI, Miller LM, Timmermans BJR, Decarreau J, Morris HM, Kang A, Bera AK, Baker D, Reconfigurable asymmetric protein assemblies through implicit negative design. *Science*. 375, eabj7662 (2022).

18. Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, Hao J, Bafna K, Norn C, Kang A, Bera AK, DiMaio F, Carter L, Chow CM, Montelione GT, Baker D, De novo protein design by deep network hallucination. *Nature*. 600, 547–552 (2021). [PubMed: 34853475]
19. Jendrusch M, Korbel JO, Sadiq SK, AlphaDesign: A de novo protein design framework based on AlphaFold (2021), p. 2021.10.11.463937, doi:10.1101/2021.10.11.463937.
20. Moffat L, Greener JG, Jones DT, Using AlphaFold for Rapid and Accurate Fixed Backbone Protein Design (2021), p. 2021.08.24.457549, doi:10.1101/2021.08.24.457549.
21. Wang J, Lisanza S, Juergens D, Tischer D, Anishchenko I, Baek M, Watson JL, Chun JH, Milles LF, Dauparas J, Expòsit M, Yang W, Saragovi A, Ovchinnikov S, Baker D, Deep learning methods for designing proteins scaffolding functional sites (2021), p. 2021.11.10.468128, doi:10.1101/2021.11.10.468128.
22. Ovchinnikov S, Huang P-S, Structure-based protein design with deep learning. *Curr. Opin. Chem. Biol* 65, 136–144 (2021). [PubMed: 34547592]
23. Norn C, Wicky BIM, Juergens D, Liu S, Kim D, Tischer D, Koepnick B, Anishchenko I, Players Foldit, Baker D, Ovchinnikov S, Coral A, Bubar AJ, Boykov A, Valle Pérez AU, MacMillan A, Lubow A, Mussini A, Cai A, Ardill AJ, Seal A, Kalantarian A, Failer B, Lackersteen B, Chagot B, Haight BR, Ta tan B, Uitham B, Roy BG, de Melo Cruz BR, Echols B, Lorenz BE, Blair B, Kestemont B, Eastlake CD, Bragdon CJ, Vardeman C, Salerno C, Comisky C, Hayman CL, Landers CR, Zimov C, Coleman CD, Painter CR, Ince C, Lynagh C, Malania D, Wheeler DC, Robertson D, Simon V, Chisari E, Kai ELJ, Rezae F, Lengyel F, Tabotta F, Padelletti F, Boström F, Gross GO, McIlvaine G, Beecher G, Hansen GT, de Jong G, Feldmann H, Borman JL, Quinn J, Norrgard J, Truong J, Diderich JA, Canfield JM, Photakis J, Slone JD, Madzio J, Mitchell J, Stomieroski JC, Mitch JH, Altenbeck JR, Schinkler J, Weinberg JB, Burbach JD, Sequeira da Costa JC, Bada Juarez JF, Gunnarsson JP, Harper KD, Joo K, Clayton KT, DeFord KE, Scully KF, Gildea KM, Abbey KJ, Kohli KL, Stenner K, Takács K, Poussaint LL, Manalo LC, Withers LC, Carlson L, Wei L, Fisher LR, Carpenter L, Ji-hwan M, Ricci M, Belcastro MA, Leniec M, Hohmann M, Thompson M, Thayer MA, Gaebel M, Cassidy MD, Fagiola M, Lewis M, Pfütznerreuter M, Simon M, Elmassy MM, Benevides N, Kerr NK, Verma N, Shannon O, Yin O, Wolfteich P, Gummarsall P, Tlu cik P, Gajar P, Triggiani PJ, Guha R, Mathew Innes RB, Buchanan R, Gamble R, Leduc R, Spearing R, dos Santos Gomes RLC, Estep RD, DeWitt R, Moore R, Shneider SG, Zaccanelli SJ, Kuznetsov S, Burillo-Sanz S, Mooney S, Vasiliy S, Butkovich SS, Hudson SB, Pote SL, Denne SP, Schwegmann SA, Ratna S, Kleinfelter SC, Bausewein T, George TJ, de Almeida TS, Yeginer U, Barmettler W, Pulley WR, Wright WS, Willyanto, Lansford W, Hochart X, Gaiji YAS, Lagodich Y, Christian V, Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad. Sci* 118, e2017228118 (2021).
24. Anand N, Eguchi R, Mathews II, Perez CP, Derry A, Altman RB, Huang P-S, Protein sequence design with a learned potential. *Nat. Commun* 13, 746 (2022). [PubMed: 35136054]
25. Hsu C, Verkuil R, Liu J, Lin Z, Hie B, Sercu T, Lerer A, Rives A, Learning inverse folding from millions of predicted structures (2022), p. 2022.04.10.487779, doi:10.1101/2022.04.10.487779.
26. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D, Highly accurate protein structure prediction with AlphaFold. *Nature*. 596, 583–589 (2021). [PubMed: 34265844]
27. Mariani V, Biasini M, Barbato A, Schwede T, IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*. 29, 2722–2728 (2013). [PubMed: 23986568]
28. Zhang Y, Skolnick J, TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* 33, 2302–2309 (2005). [PubMed: 15849316]
29. Xu J, Zhang Y, How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*. 26, 889–895 (2010). [PubMed: 20164152]
30. Inceptionism: Going Deeper into Neural Networks. Google AI Blog, (available at <http://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>).

31. Nguyen A, Yosinski J, Clune J, Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images (2015), (available at <http://arxiv.org/abs/1412.1897>).
32. Simonyan K, Vedaldi A, Zisserman A, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps (2014), (available at <http://arxiv.org/abs/1312.6034>).
33. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D, Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. 373, 871–876 (2021). [PubMed: 34282049]
34. Kobe B, Deisenhofer J, The leucine-rich repeat: a versatile binding motif. *Trends Biochem. Sci* 19, 415–421 (1994). [PubMed: 7817399]
35. Guerra P, González-Alamos M, Llauro A, Casañas A, Querol-Audí J, de Pablo PJ, Verdaguer N, Symmetry disruption commits vault particles to disassembly. *Sci. Adv* 8, eabj7795 (2022).
36. Anand N, Achim T, Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models (2022), doi:10.48550/arXiv.2205.15019.
37. Trippe BL, Yim J, Tischer D, Baker D, Broderick T, Barzilay R, Jaakkola T, Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem (2022), doi:10.48550/arXiv.2206.04119.
38. Mukherjee S, Zhang Y, MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res.* 37, e83 (2009). [PubMed: 19443443]
39. Alford RF, Leaver-Fay A, Jeliazkov JR, O’Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, Labonte JW, Pacella MS, Bonneau R, Bradley P, Dunbrack RL, Das R, Baker D, Kuhlman B, Kortemme T, Gray JJ, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput* 13, 3031–3048 (2017). [PubMed: 28430426]
40. Lawrence MC, Colman PM, Shape Complementarity at Protein/Protein Interfaces. *J. Mol. Biol* 234, 946–950 (1993). [PubMed: 8263940]
41. Jones DT, Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol* 292, 195–202 (1999). [PubMed: 10493868]
42. Kabsch W, Sander C, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 22, 2577–2637 (1983). [PubMed: 6667333]
43. Chennamsetty N, Vovnov V, Kayser V, Helk B, Trout BL, Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci* 106, 11937–11942 (2009). [PubMed: 19571001]
44. Dang B, Mravic M, Hu H, Schmidt N, Mensa B, DeGrado WF, SNAC-tag for sequence-specific chemical protein cleavage. *Nat. Methods* 16, 319–322 (2019). [PubMed: 30923372]
45. Kabsch W, XDS. *Acta Crystallogr. D Biol. Crystallogr* 66, 125–132 (2010). [PubMed: 20124692]
46. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A, Wilson KS, Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr* 67, 235–242 (2011). [PubMed: 21460441]
47. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ, Phaser crystallographic software. *J. Appl. Crystallogr* 40, 658–674 (2007). [PubMed: 19461840]
48. Emsley P, Cowtan K, Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr* 60, 2126–2132 (2004). [PubMed: 15572765]
49. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung L-W, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr* 66, 213–221 (2010). [PubMed: 20124702]
50. Murshudov GN, Vagin AA, Dodson EJ, Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallogr. D Biol. Crystallogr* 53, 240–255 (1997). [PubMed: 15299926]

51. Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, Verma V, Keedy DA, Hintze BJ, Chen VB, Jain S, Lewis SM, Arendall III WB, Snoeyink J, Adams PD, Lovell SC, Richardson JS, Richardson DC, MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* 27, 293–315 (2018). [PubMed: 29067766]
52. Nannenga BL, Iadanza MG, Vollmar BS, Gonen T, *Curr. Protoc. Protein Sci*, in press, doi:10.1002/0471140864.ps1715s72.
53. Grant T, Rohou A, Grigorieff N, cisTEM, user-friendly software for single-particle image processing. *eLife.* 7, e35383 (2018). [PubMed: 29513216]
54. Punjani A, Rubinstein JL, Fleet DJ, Brubaker MA, cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* 14, 290–296 (2017). [PubMed: 28165473]
55. Punjani A, Fleet DJ, 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol* 213, 107702 (2021). [PubMed: 33582281]
56. Carragher B, Kisseberth N, Kriegman D, Milligan RA, Potter CS, Pulokas J, Reilein A, Leginon: An Automated System for Acquisition of Images from Vitreous Ice Specimens. *J. Struct. Biol* 132, 33–45 (2000). [PubMed: 11121305]
57. Zheng SQ, Palovcak E, Armache J-P, Verba KA, Cheng Y, Agard DA, MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* 14, 331–332 (2017). [PubMed: 28250466]
58. Rohou A, Grigorieff N, CTFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol* 192, 216–221 (2015). [PubMed: 26278980]

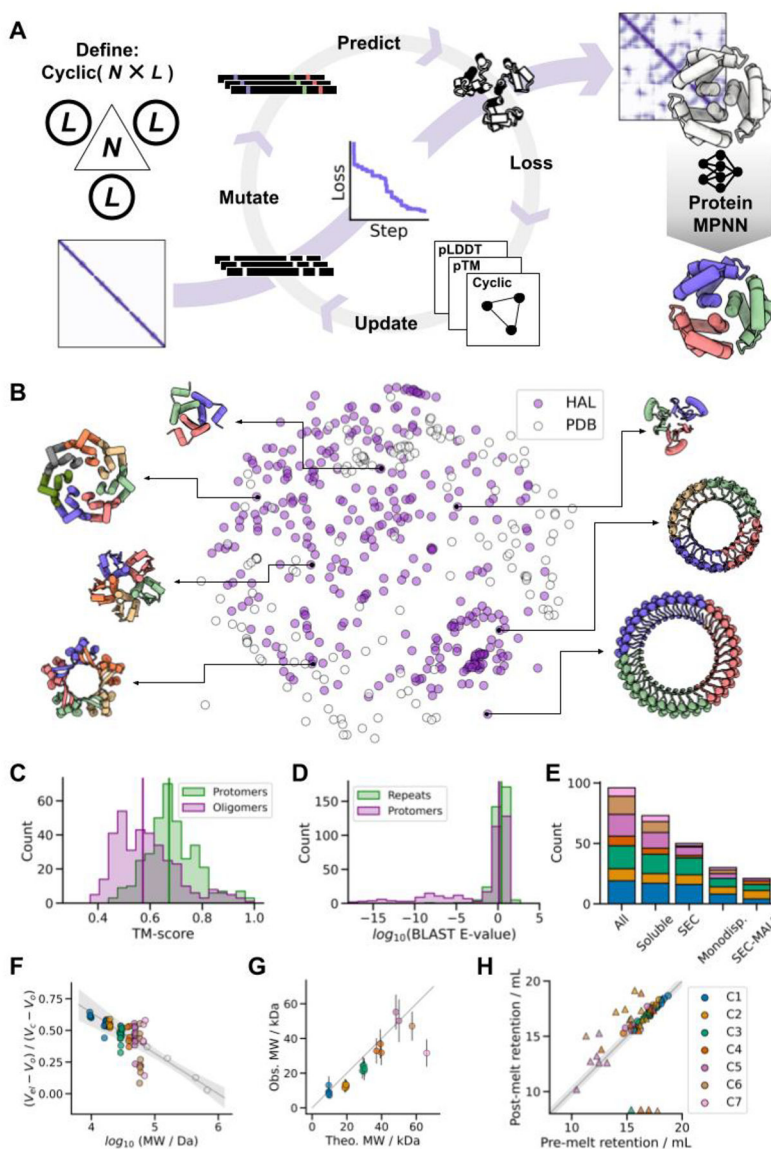


Fig. 1. Hallucinating symmetric protein assemblies

(A) Starting from choice of a cyclic symmetry and protein length, a random sequence is optimized by MCMC through the AF2 network until the resulting structure fits the design objective, followed by sequence re-design with ProteinMPNN. (B) The method generates structurally diverse outputs, quantified here by multi-dimensional scaling of protomer pairwise structural similarities between experimentally tested HALs ($N=351$) and all *de novo* cyclic oligomers present in the PDB ($N=162$). (C) Generated structures differ from those in the PDB. Median TM-scores to the closest match: 0.67 and 0.57 for the protomers and oligomers respectively (vertical lines). (D) Generated sequences are unrelated to naturally-occurring proteins. Median BLAST E-values from the closet hit in UniRef100: 2.6 and 1.3 for the repeat motifs and protomers respectively (vertical lines). (E) Success counts of ProteinMPNN-designed HALs at different levels of characterization. (F) Most soluble HALs have SEC retention volumes consistent with their oligomeric state.

The gray line shows the fit to calibration standards (open circles), and the shaded area represents the 95% confidence interval of the calibration. **(G)** The observed molecular weights of HALs from SEC-MALS are close to those computed from the design models. **(H)** ProteinMPNN-designed HALs are thermostable. Pre-melting and post-melting retention volumes are closely correlated; circles represent designs that remained monodisperse, while triangles indicate polydispersity after heat-treatment. In plots **E-H**, the data is categorized by cyclic symmetry classes. The legend is shown in **H**.

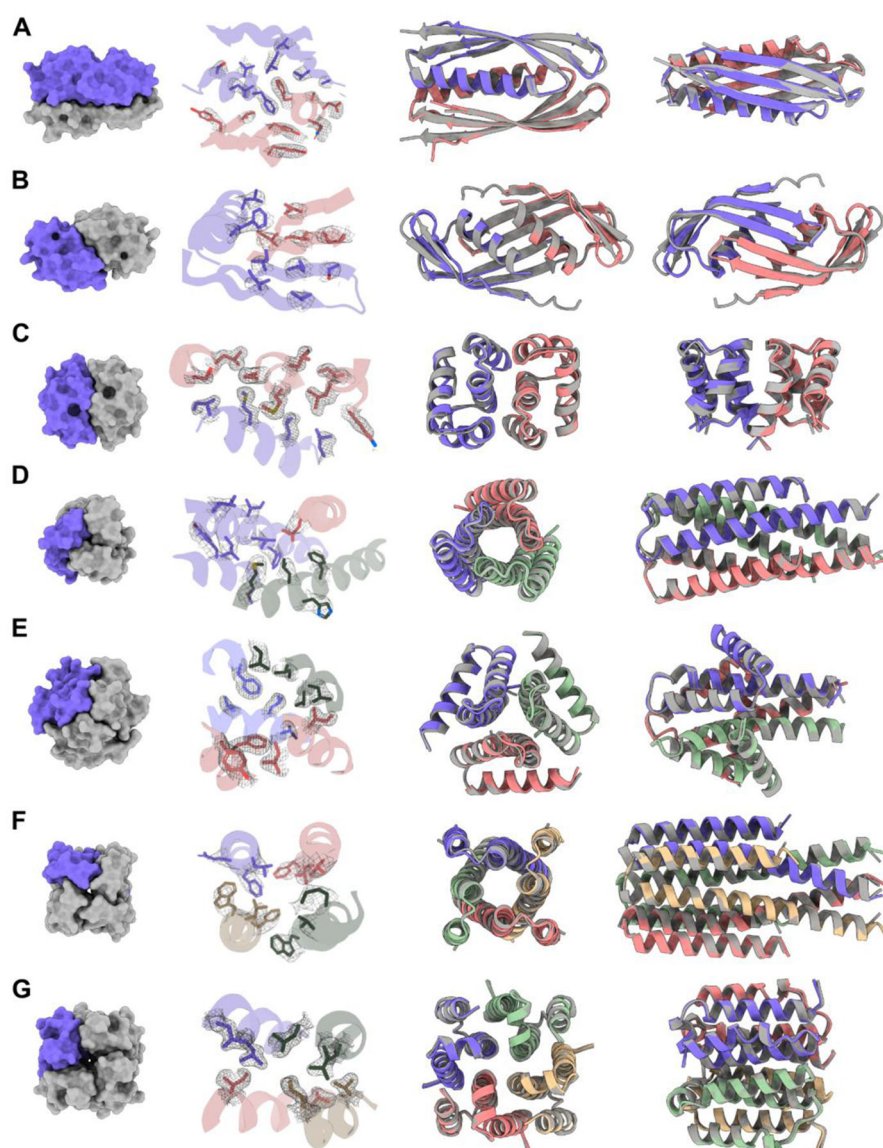


Fig. 2. Structures of HALs solved by X-ray crystallography compared to their design models. (A) HALC2_062 (RMSD: 0.81 Å). (B) HALC2_065 (RMSD: 1.02 Å). (C) HALC2_068 (RMSD: 0.86 Å). (D) HALC3_104 (RMSD: 0.42 Å). (E) HALC3_109 (RMSD: 0.46 Å). (F) HALC4_135 (RMSD: 0.60 Å). (G) HALC4_136 (RMSD: 0.34 Å). For each row, the first panel shows a surface rendering of the oligomer with one protomer highlighted in purple, the second highlights the side-chain rotamers of the design model to the 2mFo-DFc map (in gray), and the last two panels show two different orientations of the structural overlays between the model (gray) and the solved structure (colored by chains).

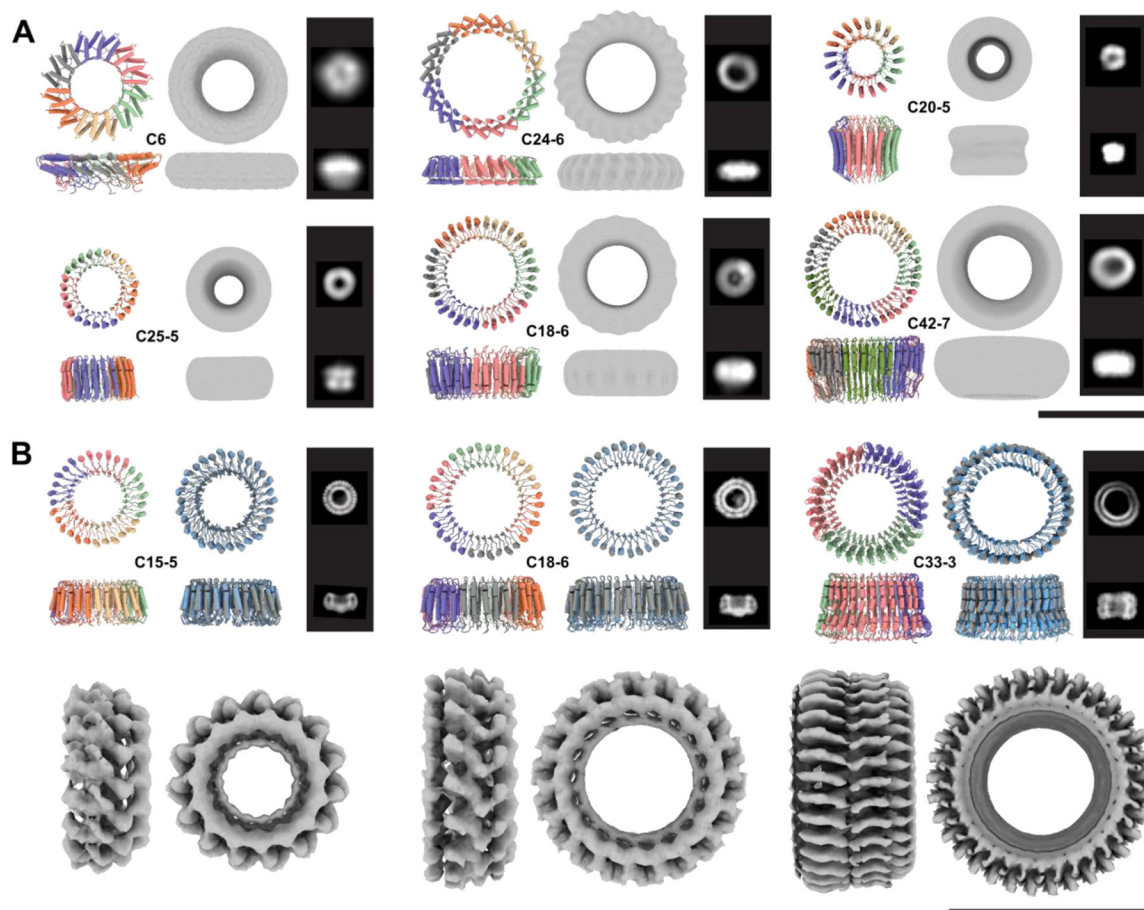


Fig. 3. Cryo-electron and negative stain electron microscopy validation of large HALs.

For each design, the model is shown colored by chain and the corresponding internal symmetry (X) and oligomerization state (Y) are indicated ($CX-Y$). The electron density map is shown next to the model alongside characteristic 2D class averages. **(A)** Negative stain characterization of HALs. Ring diameters are 92 Å, 110 Å, 75 Å, 80 Å, 100 Å, 107 Å, for HALC6_220, HALC24-6_316, HALC20-5_308, HALC25-5_341, HALC18-6_278 and HALC42-7_351, respectively. **(B)** CryoEM characterisation of three large HALs. The ring diameters are 87 Å, 99 Å, and 100 Å for HALC15-5_262, HALC18-6_265, and HALC33-3_343, respectively. Top row left panels: design model colored by chain; Top row, right panels: superpositions of the CryoEM model (gray) and design model (blue). The computed backbone atom RMSD between the designed and experimental structure are 0.81 Å, 1.69 Å, and 2.30 Å respectively (Fig. S16). Bottom row: 4.38 Å, 6.51 Å, and 6.32 Å cryoEM electron density maps. Scale bars = 10 nm.

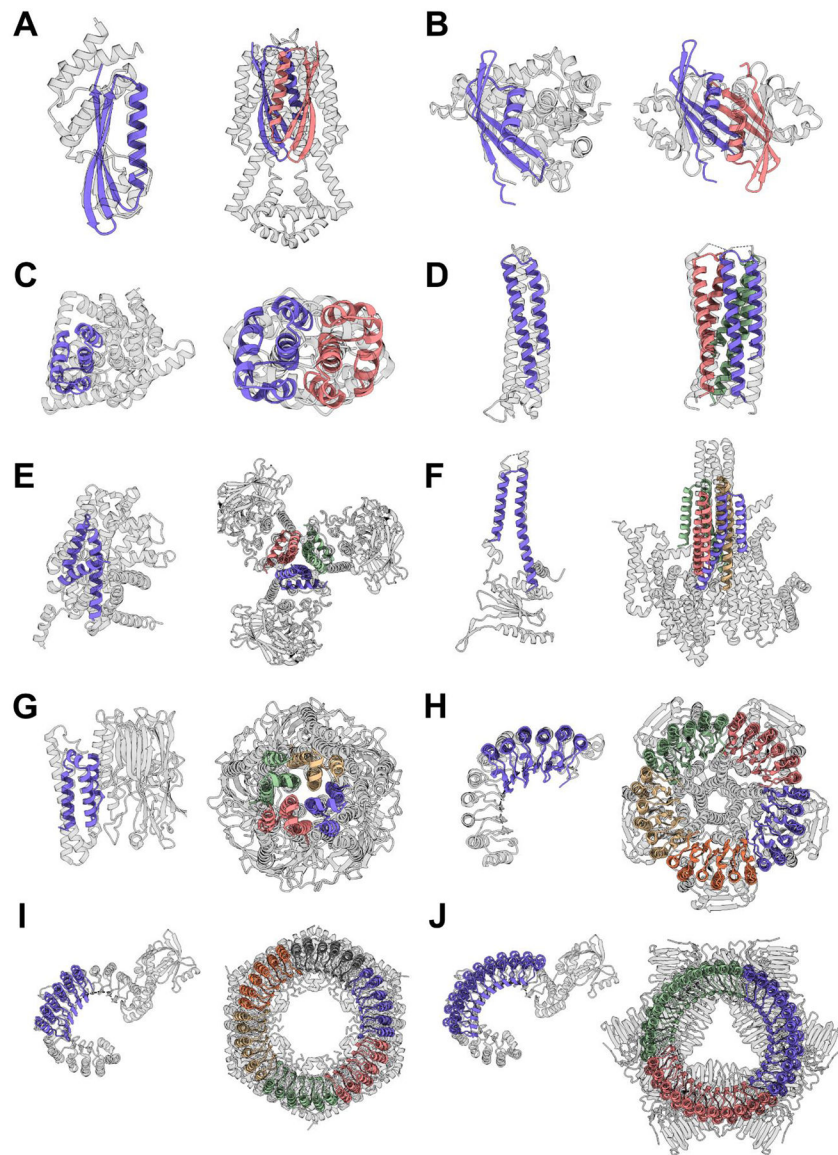


Fig. 4. Hallucinated structures differ significantly from their closest matches in the PDB. For each structure solved by crystallography (Fig. 2) or cryoEM (Fig. 3B), the closest structural match to the protomer and to the oligomer are shown on the left and right respectively. Designs are colored by chain and the closest matching PDB is shown in gray. In most cases the closest oligomer has an entirely different structure; this is particularly evident for the larger designs in **G-H**. TM-scores (protomer | oligomer) are indicated in parentheses, and the PDB IDs are reported in Table S2. **(A)** HALC2_062 (0.69 | 0.59). **(B)** HALC2_065 (0.67 | 0.54). **(C)** HALC2_068 (0.67 | 0.57). **(D)** HALC3_104 (0.87 | 0.88). **(E)** HALC3_109 (0.78 | 0.69). **(F)** HALC4_135 (0.80 | 0.59). **(G)** HALC4_136 (0.80 | 0.71). **(H)** HALC15-5_262 (0.65 | 0.46). **(I)** HALC18-6_265 (0.65 | 0.49). **(J)** HALC33-3_343 (0.49 | 0.41).