Cite as: B. I. M. Wicky *et al.*, *Science* 10.1126/science.add1964 (2022).

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]*

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.

†These authors contributed equally to this work.

***Corresponding author. Email: dabaker@uw.edu**

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*–*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 (*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) (figs. S1 and S2). 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 (figs. S4 and S5). 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 (*33*). 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) (*34*), 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 (figs. S4, S6, and S7).

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 figs. S8 and S9), 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 and fig. 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 Cα RMSD of 0.6 Å across all designs, with resolutions ranging from 1.8 to 3.4 Å) (fig. S11 and 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. 3, A and B, and 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. 3C and 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. 3D and 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. 3E and 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 homolog to this design has a TM-score of only 0.59 (Fig. 3F and 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. 3G and 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, shorthanded 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) (*35*), 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. 4A and 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. 4B and figs. S13 to S16). 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 Cα 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. 3, H and I, and 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. 3J and 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 (*36*), 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 and table S3) and structure (median best TM-score between biounits from the PDB and HALS of 0.57) (Fig. 1C and 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 (*37*, *38*) which more directly sample in structure space, should enable the design of even more complex protein structures and assemblies.

REFERENCES AND NOTES

1. H. Garcia-Seisdedos, C. Empereur-Mot, N. Elad, E. D. Levy, Proteins evolve on the edge of supramolecular self-assembly. *Nature* **548**, 244–247 (2017). doi:10.1038/nature23320 Medline
2. I. G. Johnston, K. Dingle, S. F. Greenbury, C. Q. Camargo, J. P. K. Doye, S. E. Ahnert, A. A. Louis, Symmetry and simplicity spontaneously emerge from the algorithmic nature of evolution. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2113883119 (2022). doi:10.1073/pnas.2113883119 Medline
3. S. E. Ahnert, J. A. Marsh, H. Hernández, C. V. Robinson, S. A. Teichmann, Principles of assembly reveal a periodic table of protein complexes. *Science* **350**, aaa2245 (2015). doi:10.1126/science.aaa2245 Medline
4. S. K. Burley, H. M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. D. Costanzo, C. Christie, J. M. Duarte, S. Dutta, Z. Feng, S. Ghosh, D. S. Goodsell, R. K. Green, V. Guranovic, D. Guzenko, B. P. Hudson, Y. Liang, R. Lowe, E. Peisach, I. Periskova, C. Randle, A. Rose, M. Sekharan, C. Shao, Y.-P. Tao, Y. Valasatava, M. Voigt, J. Westbrook, J. Young, C. Zardecki, M. Zhuravleva, G. Kurisu, H. Nakamura, Y. Kengaku, H. Cho, J. Sato, J. Y. Kim, Y. Ikegawa, A. Nakagawa, R. Yamashita, T. Kudou, G.-J. Bekker, H. Suzuki, T. Iwata, M. Yokochi, N. Kobayashi, T. Fujiwara, S. Velankar, G. J. Kleywegt, S. Anyango, D. R. Armstrong, J. M. Berrisford, M. J. Conroy, J. M. Dana, M. Deshpande, P. Gane, R. Gáborová, D. Gupta, A. Gutmanas, J. Koča, L. Mak, S. Mir, A. Mukhopadhyay, N. Nadzirin, S. Nair, A. Patwardhan, T. Paysan-Lafosse, L. Pravda, O. Salih, D. Sehnal, M. Varadi, R. Vařeková, J. L. Markley, J. C. Hoch, P. R. Romero, K. Baskaran, D. Maziuk, E. L. Ulrich, J. R. Wedell, H. Yao, M. Livny, Y. E. Ioannidis; wwPDB consortium, Protein Data Bank: The single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019). doi:10.1093/nar/gky949 Medline
5. D. S. Goodsell, A. J. Olson, Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153 (2000). doi:10.1146/annurev.biophys.29.1.105 Medline
6. T. Handel, W. F. DeGrado, De novo design of a Zn^{2+}-binding protein. *J. Am. Chem. Soc.* **112**, 6710–6711 (1990). doi:10.1021/ja00174a039
7. P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, P. S. Kim, High-resolution protein design with backbone freedom. *Science* **282**, 1462–1467 (1998). doi:10.1126/science.282.5393.1462 Medline
8. J. A. Fallas, G. Ueda, W. Sheffler, V. Nguyen, D. E. McNamara, B. Sankaran, J. H. Pereira, F. Parmeggiani, T. J. Brunette, D. Cascio, T. R. Yeates, P. Zwart, D. Baker, Computational design of self-assembling cyclic protein homo-oligomers. *Nat. Chem.* **9**, 353–360 (2017). doi:10.1038/nchem.2673 Medline
9. A. R. Thomson, C. W. Wood, A. J. Burton, G. J. Bartlett, R. B. Sessions, R. L. Brady, D. N. Woolfson, Computational design of water-soluble α-helical barrels. *Science* **346**, 485–488 (2014). doi:10.1126/science.1257452 Medline
10. P.-S. Huang, K. Feldmeier, F. Parmeggiani, D. A. F. Velasco, B. Höcker, D. Baker, *De novo* design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016). doi:10.1038/nchembio.1966 Medline
11. S. E. Boyken, Z. Chen, B. Groves, R. A. Langan, G. Oberdorfer, A. Ford, J. M. Gilmore, C. Xu, F. DiMaio, J. H. Pereira, B. Sankaran, G. Seelig, P. H. Zwart, D. Baker, De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science* **352**, 680–687 (2016). doi:10.1126/science.aad8865 Medline
12. L. Doyle, J. Hallinan, J. Bolduc, F. Parmeggiani, D. Baker, B. L. Stoddard, P. Bradley, Rational design of α-helical tandem repeat proteins with closed architectures. *Nature* **528**, 585–588 (2015). doi:10.1038/nature16191 Medline
13. J. B. Bale, S. Gonen, Y. Liu, W. Sheffler, D. Ellis, C. Thomas, D. Cascio, T. O. Yeates, T. Gonen, N. P. King, D. Baker, Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **353**, 389–394 (2016). doi:10.1126/science.aaf8818 Medline
14. I. Vulovic, Q. Yao, Y.-J. Park, A. Courbet, A. Norris, F. Busch, A. Sahasrabuddhe, H. Merten, D. D. Sahtoe, G. Ueda, J. A. Fallas, S. J. Weaver, Y. Hsia, R. A. Langan, A. Plückthun, V. H. Wysocki, D. Veesler, G. J. Jensen, D. Baker, Generation of ordered protein assemblies using rigid three-body fusion. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015037118 (2021). doi:10.1073/pnas.2015037118 Medline
15. Y. Hsia, R. Mout, W. Sheffler, N. I. Edman, I. Vulovic, Y.-J. Park, R. L. Redler, M. J. Bick, A. K. Bera, A. Courbet, A. Kang, T. J. Brunette, U. Nattermann, E. Tsai, A. Saleem, C. M. Chow, D. Ekiert, G. Bhabha, D. Veesler, D. Baker, Design of multi-scale protein complexes by hierarchical building block fusion. *Nat. Commun.* **12**, 2294 (2021). doi:10.1038/s41467-021-22276-z Medline
16. C. E. Correnti, J. P. Hallinan, L. A. Doyle, R. O. Ruff, C. A. Jaeger-Ruckstuhl, Y. Xu, B. W. Shen, A. Qu, C. Polkinghorn, D. J. Friend, A. D. Bandaranayake, S. R. Riddell, B. K. Kaiser, B. L. Stoddard, P. Bradley, Engineering and functionalization of large circular tandem repeat protein nanoparticles. *Nat. Struct. Mol. Biol.* **27**, 342–350 (2020). doi:10.1038/s41594-020-0397-5 Medline
17. D. D. Sahtoe, F. Praetorius, A. Courbet, Y. Hsia, B. I. M. Wicky, N. I. Edman, L. M. Miller, B. J. R. Timmermans, J. Decarreau, H. M. Morris, A. Kang, A. K. Bera, D. Baker, Reconfigurable asymmetric protein assemblies through implicit negative design. *Science* **375**, eabj7662 (2022). doi:10.1126/science.abj7662 Medline
18. I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, T. A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A. K. Bera, F. DiMaio, L. Carter, C. M. Chow, G. T. Montelione, D. Baker, De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021). doi:10.1038/s41586-021-04184-w Medline
19. M. Jendrusch, J. O. Korbel, S. K. Sadiq, AlphaDesign: A *de novo* protein design framework based on AlphaFold. bioRxiv 2021.10.11.463937 [Preprint] (2021). https://doi.org/10.1101/2021.10.11.463937.
20. L. Moffat, J. G. Greener, D. T. Jones, Using AlphaFold for Rapid and Accurate Fixed Backbone Protein Design. bioRxiv 2021.08.24.457549 [Preprint] (2021). https://doi.org/10.1101/2021.08.24.457549.
21. J. Wang, S. Lisanza, D. Juergens, D. Tischer, I. Anishchenko, M. Baek, J. L. Watson, J. H. Chun, L. F. Milles, J. Dauparas, M. Expòsit, W. Yang, A. Saragovi, S. Ovchinnikov, D. Baker, Deep learning methods for designing proteins scaffolding functional sites. bioRxiv 2021.11.10.468128 [Preprint] (2021). https://doi.org/10.1101/2021.11.10.468128.
22. S. Ovchinnikov, P.-S. Huang, Structure-based protein design with deep learning. *Curr. Opin. Chem. Biol.* **65**, 136–144 (2021). doi:10.1016/j.cbpa.2021.08.004 Medline
23. C. Norn, B. I. M. Wicky, D. Juergens, S. Liu, D. Kim, D. Tischer, B. Koepnick, I. Anishchenko, D. Baker, S. Ovchinnikov; Foldit Players, Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2017228118 (2021). doi:10.1073/pnas.2017228118 Medline
24. N. Anand, R. Eguchi, I. I. Mathews, C. P. Perez, A. Derry, R. B. Altman, P.-S. Huang, Protein sequence design with a learned potential. *Nat. Commun.* **13**, 746 (2022). doi:10.1038/s41467-022-28313-9 Medline
25. C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, A. Rives, Learning inverse folding from millions of predicted structures. bioRxiv 2022.04.10.487779 [Preprint] (2022). https://doi.org/10.1101/2022.04.10.487779.
26. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold.

Nature **596**, 583–589 (2021). doi:10.1038/s41586-021-03819-2 Medline
27. V. Mariani, M. Biasini, A. Barbato, T. Schwede, lDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013). doi:10.1093/bioinformatics/btt473 Medline
28. Y. Zhang, J. Skolnick, TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005). doi:10.1093/nar/gki524 Medline
29. J. Xu, Y. Zhang, How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010). doi:10.1093/bioinformatics/btq066 Medline
30. A. Mordvintsev, C. Olah, M. Tyka, "Inceptionism: Going Deeper into Neural Networks," Google AI Blog, 17 June 2015; https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.
31. A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. arXiv:1412.1897 [cs.CV] (2015).
32. K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:1312.6034 [cs.CV] (2014).
33. J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, Robust deep learning–based protein sequence design using ProteinMPNN. *Science* 10.1126/science.add2187 (2022). doi:10.1126/science.add2187
34. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021). doi:10.1126/science.abj8754 Medline
35. B. Kobe, J. Deisenhofer, The leucine-rich repeat: A versatile binding motif. *Trends Biochem. Sci.* **19**, 415–421 (1994). doi:10.1016/0968-0004(94)90090-6 Medline
36. P. Guerra, M. González-Alamos, A. Llauró, A. Casañas, J. Querol-Audí, P. J. de Pablo, N. Verdaguer, Symmetry disruption commits vault particles to disassembly. *Sci. Adv.* **8**, eabj7795 (2022). doi:10.1126/sciadv.abj7795 Medline
37. N. Anand, T. Achim, Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. arXiv:2205.15019 [q-bio.QM] (2022).
38. B. L. Trippe, J. Yim, D. Tischer, D. Baker, T. Broderick, R. Barzilay, T. Jaakkola, Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. arXiv:2206.04119 [q-bio.BM] (2022).
39. S. Mukherjee, Y. Zhang, MM-align: A quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res.* **37**, e83 (2009). doi:10.1093/nar/gkp318 Medline
40. R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack Jr., R. Das, D. Baker, B. Kuhlman, T. Kortemme, J. J. Gray, The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017). doi:10.1021/acs.jctc.7b00125 Medline
41. M. C. Lawrence, P. M. Colman, Shape complementarity at protein/protein interfaces. *J. Mol. Biol.* **234**, 946–950 (1993). doi:10.1006/jmbi.1993.1648 Medline
42. D. T. Jones, Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999). doi:10.1006/jmbi.1999.3091 Medline
43. W. Kabsch, C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983). doi:10.1002/bip.360221211 Medline
44. N. Chennamsetty, V. Voynov, V. Kayser, B. Helk, B. L. Trout, Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 11937–11942 (2009). doi:10.1073/pnas.0904191106 Medline
45. B. Dang, M. Mravic, H. Hu, N. Schmidt, B. Mensa, W. F. DeGrado, SNAC-tag for sequence-specific chemical protein cleavage. *Nat. Methods* **16**, 319–322 (2019). doi:10.1038/s41592-019-0357-3 Medline
46. W. Kabsch, *XDS*. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010). doi:10.1107/S0907444909047337 Medline
47. M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, Overview of the *CCP*4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011). doi:10.1107/S0907444910045749 Medline
48. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, *Phaser* crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007). doi:10.1107/S0021889807021206 Medline
49. P. Emsley, K. Cowtan, *Coot*: Model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004). doi:10.1107/S0907444904019158 Medline
50. P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, *PHENIX*: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010). doi:10.1107/S0907444909052925 Medline
51. G. N. Murshudov, A. A. Vagin, E. J. Dodson, Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.* **53**, 240–255 (1997). doi:10.1107/S0907444996012255 Medline
52. C. J. Williams, J. J. Headd, N. W. Moriarty, M. G. Prisant, L. L. Videau, L. N. Deis, V. Verma, D. A. Keedy, B. J. Hintze, V. B. Chen, S. Jain, S. M. Lewis, W. B. Arendall 3rd, J. Snoeyink, P. D. Adams, S. C. Lovell, J. S. Richardson, D. C. Richardson, MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018). doi:10.1002/pro.3330 Medline
53. B. L. Nannenga, M. G. Iadanza, B. S. Vollmar, T. Gonen, Overview of electron crystallography of membrane proteins: Crystallization and screening strategies using negative stain electron microscopy. *Curr. Protoc. Protein Sci.* **17**, 17.15.1–17.15.11 (2013). doi:10.1002/0471140864.ps1715s72 Medline
54. T. Grant, A. Rohou, N. Grigorieff, *cis*TEM, user-friendly software for single-particle image processing. *eLife* **7**, e35383 (2018). doi:10.7554/eLife.35383 Medline
55. 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). doi:10.1038/nmeth.4169 Medline
56. A. Punjani, D. J. Fleet, 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol.* **213**, 107702 (2021). doi:10.1016/j.jsb.2021.107702 Medline
57. B. Carragher, N. Kisseberth, D. Kriegman, R. A. Milligan, C. S. Potter, J. Pulokas, A. Reilein, Leginon: An automated system for acquisition of images from vitreous ice specimens. *J. Struct. Biol.* **132**, 33–45 (2000). doi:10.1006/jsbi.2000.4314 Medline
58. S. Q. Zheng, E. Palovcak, J.-P. Armache, K. A. Verba, Y. Cheng, D. A. Agard, MotionCor2: Anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017). doi:10.1038/nmeth.4193 Medline
59. A. Rohou, N. Grigorieff, CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015). doi:10.1016/j.jsb.2015.08.008 Medline

ACKNOWLEDGMENTS

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." **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. **Data and materials availability:** All data are available in the main text or as supplementary materials. Data frame containing all protein information and experimental data, design models, scripts and computational methods are available on GitHub at https://github.com/bwicky/oligomer_hallucination and Zenodo. 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).

SUPPLEMENTARY MATERIALS
science.org/doi/10.1126/science.add1964
Materials and Methods
Figs. S1 to S17
Tables S1 to S4
References (*39*–*59*)

Submitted 27 May 2022; accepted 8 September 2022
Published online 15 September 2022
10.1126/science.add1964

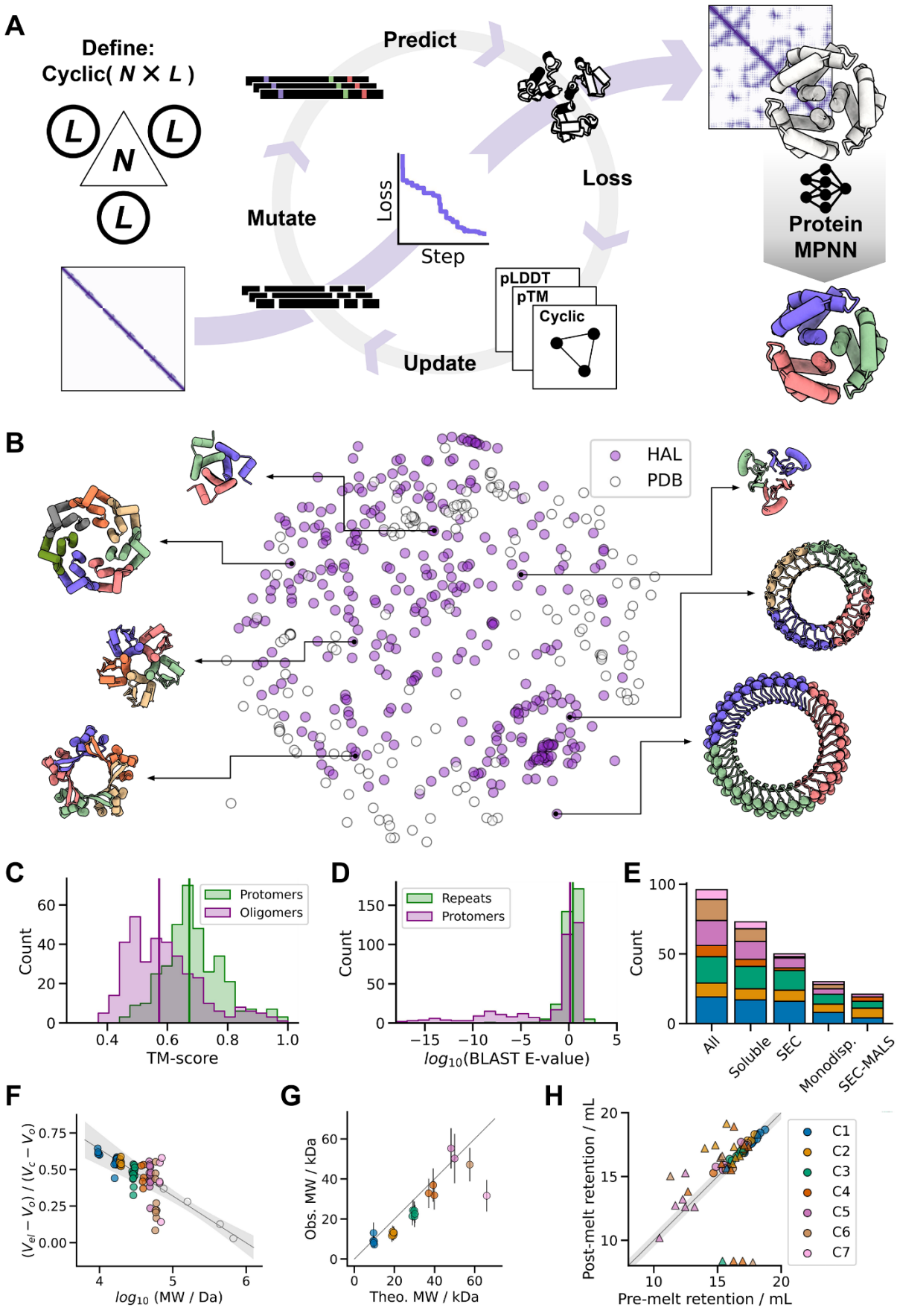

Fig. 1. Hallucinating symmetric protein assemblies. (**A**) Starting from choice of a cyclic symmetry and protein length, a random sequence is optimized by MCMC through the AF2 network until the resulting structure fits the design objective, followed by sequence re-design with ProteinMPNN. (**B**) The 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 (E) to (H), the data are categorized by cyclic symmetry classes. The legend is shown in (H).

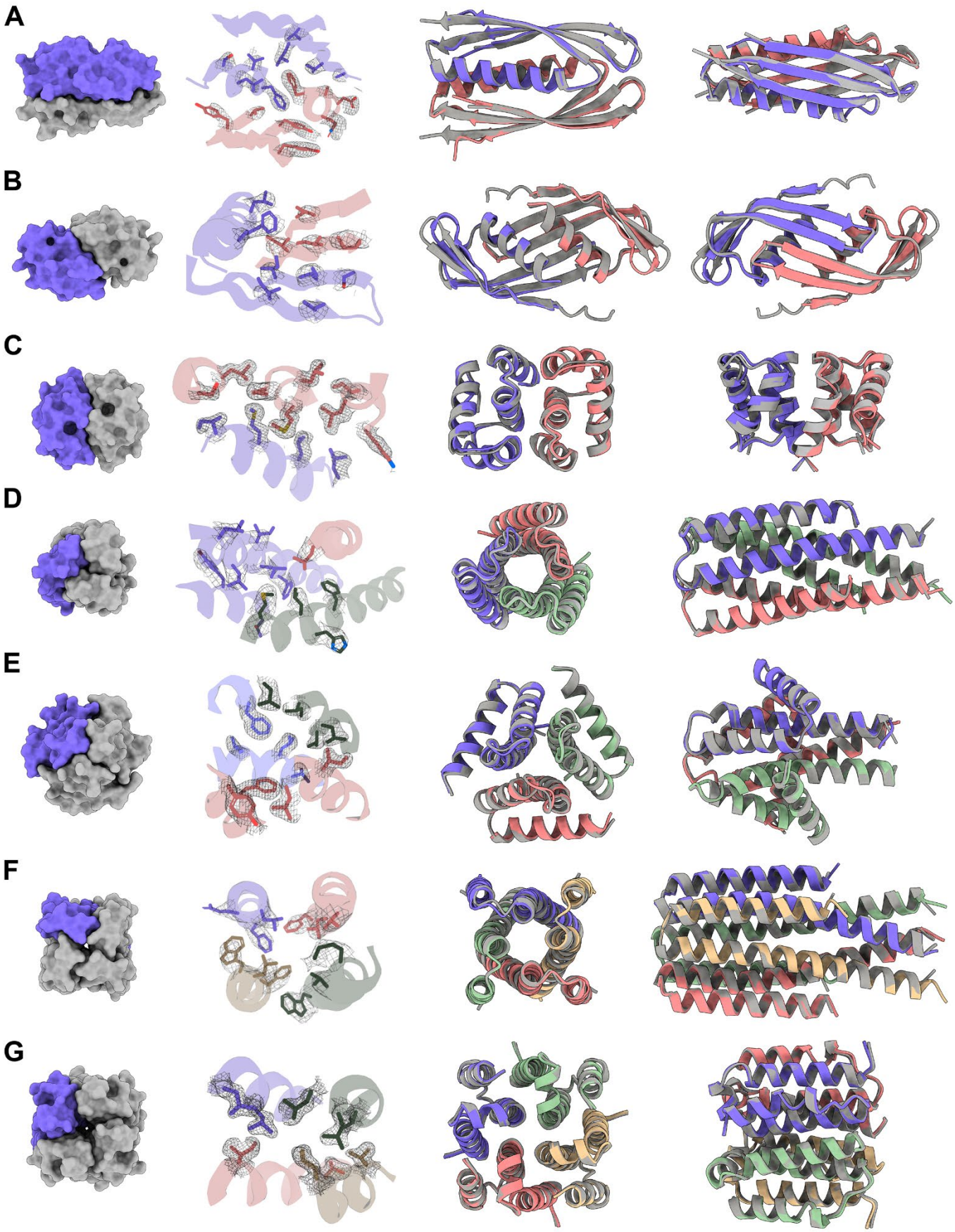

Fig. 2. Structures of HALs solved by X-ray crystallography compared to their design models. (**A**) HALC2_062 (RMSD: 0.81 Å). (**B**) HALC2_065 (RMSD: 1.02 Å). (**C**) HALC2_068 (RMSD: 0.86 Å). (**D**) HALC3_104 (RMSD: 0.42 Å). (**E**) HALC3_109 (RMSD: 0.46 Å). (**F**) HALC4_135 (RMSD: 0.60 Å). (**G**) 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).

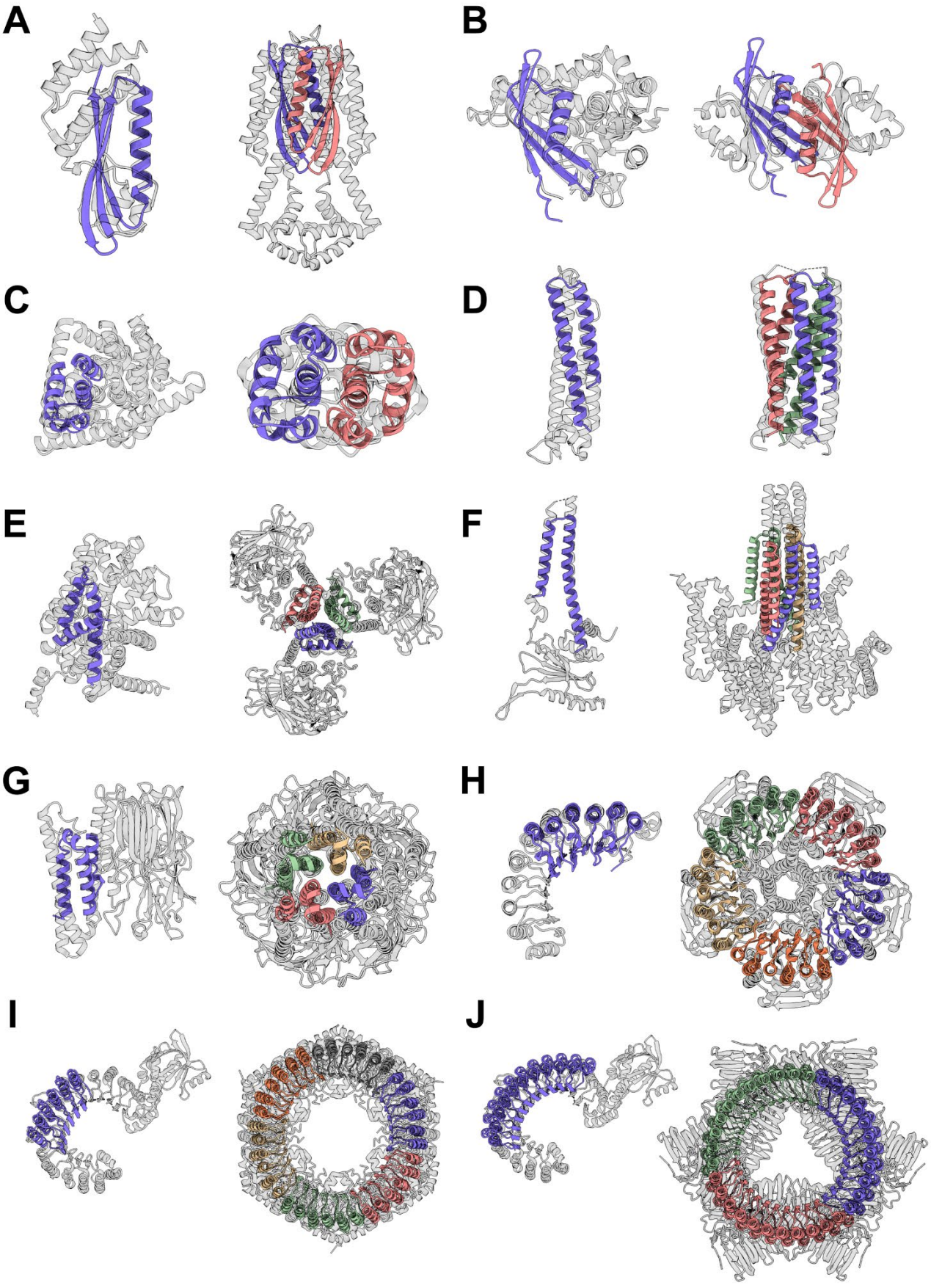

Fig. 3. Hallucinated structures differ significantly from their closest matches in the PDB. For each structure solved by crystallography (Fig. 2) or cryoEM (Fig. 4B), the closest structural match to the protomer and to the oligomer are shown on the left and right respectively. Designs are colored by 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) and (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).

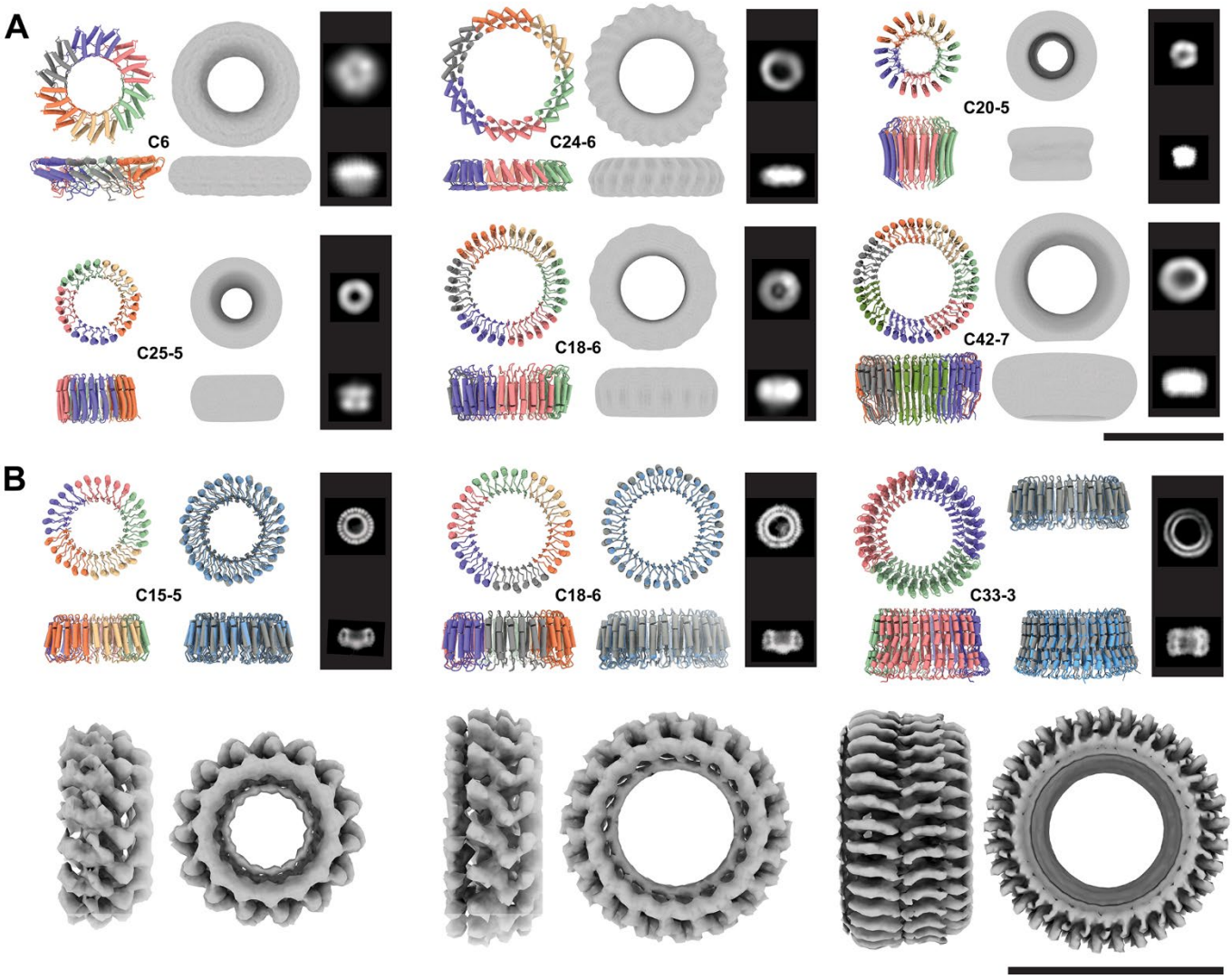

Fig. 4. Cryo-electron and negative stain electron microscopy validation of large HALs. For each design, the model is shown colored by chain and the corresponding internal symmetry (X) and oligomerization state (Y) are indicated (CX-Y). The electron density map is shown next to the model alongside characteristic 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 characterization 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.

Hallucinating symmetric protein assemblies

B. I. M. WickyL. F. MillesA. CourbetR. J. RagotteJ. DauparasE. KinfuS. TippsR. D. KiblerM. BaekF. DiMaioX. LiL. CarterA. KangH. NguyenA. K. BeraD. Baker

Science, **Ahead of Print** • DOI: 10.1126/science.add1964