PROTEIN DESIGN

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]*

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 seven designs are very similar to the computational models (median root mean square deviation: 0.6 angstroms), as are three cryo–electron microscopy structures of giant 10-nanometer rings with up to 1550 residues and C_{33} 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 since the inception of the protein design field (*6*, *7*). While there have been a number of successes (*8*–*10*), current approaches require specification of the structure of the monomers in advance and, with the exception of parametrically designed helical bundles (*11*, *12*), have involved rigid-body docking of previously characterized monomers into higher-order symmetric structures followed by interface optimization to confer low energy to the assembled state (*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 novel 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*).

[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.
*Corresponding author. Email: dabaker@uw.edu
†These authors contributed equally to this work.

Computational approach

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 [predicted local distance difference test (pLDDT); per-residue structural accuracy (*27*); and pTM, an estimate of the template modeling (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 (~1 to 7 CPU-days 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 call 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 of <0.5, the cutoff for fold assignment in CATH/SCOP (*29*)], indicating considerable generalization beyond the PDB training set.

Experimental biophysical characterization

We selected 150 designs with AF2 pLDDT > 0.7 and pTM > 0.7 for experimental testing. However, virtually none showed appreciable soluble expression when produced in *Escherichia 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 overfitting during MCMC optimization leading to the generation of adversarial sequences that the network confidently predicts are structured but in actuality have aberrant 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 was validated with AF2 (these dozens of evaluations, compared with 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 while RF2 did not confidently predict the structure of most of the original AF2 hallucinated sequences, it did successfully predict 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–multiangle light scattering (SEC-MALS) (Fig. 1G and fig. S10). CD analysis of the soluble samples indicated that 67/71 (94%) 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 but over-fitted sequences (which exhibit various pathologies; fig. S5) and highlighting the power of ProteinMPNN to

generate sequences that 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 heat-treated samples indicated that most designs retained their oligomeric state, suggesting that ProteinMPNN-designed HALs are thermostable (Fig. 1H and fig. S9).

Structure determination

To evaluate design accuracy, we attempted crystallization of 19 designs and succeeded in solving crystal structures for seven (three C_2, two C_3, and two C_4 designs) (Fig. 2). All crystal

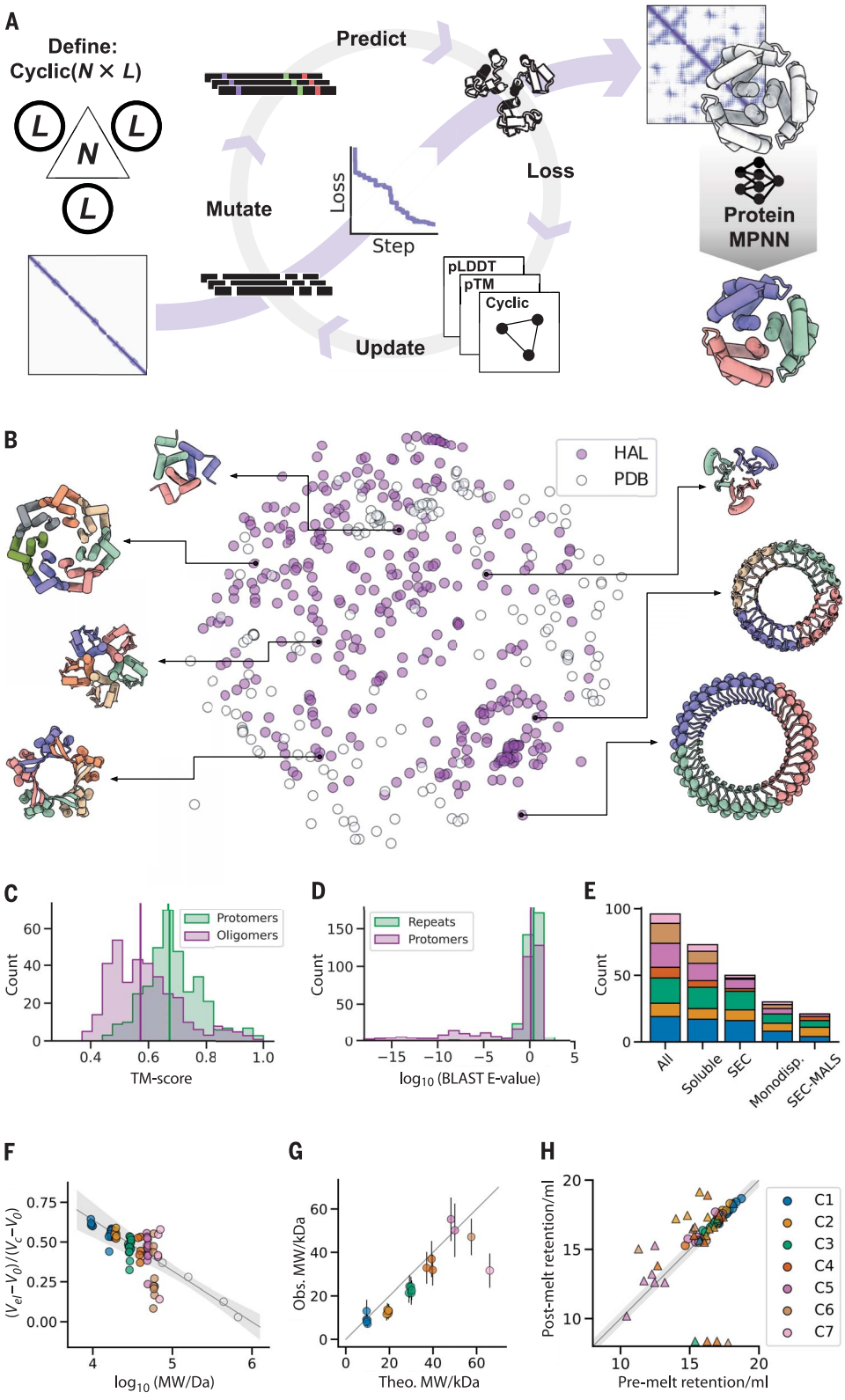

Fig. 1. Hallucinating symmetric protein assemblies. (**A**) Starting from choice of a cyclic symmetry and protein length, a random sequence is optimized by MCMC through the AF2 network until the resulting structure fits the design objective, followed by sequence redesign with ProteinMPNN. (**B**) The method generates structurally diverse outputs, quantified here by multidimensional 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**) Number of ProteinMPNN design successes at different levels of characterization. Monodisp., monodisperse. (**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 using the color scheme is shown in (H). In (G) and (H), the line indicates parity.

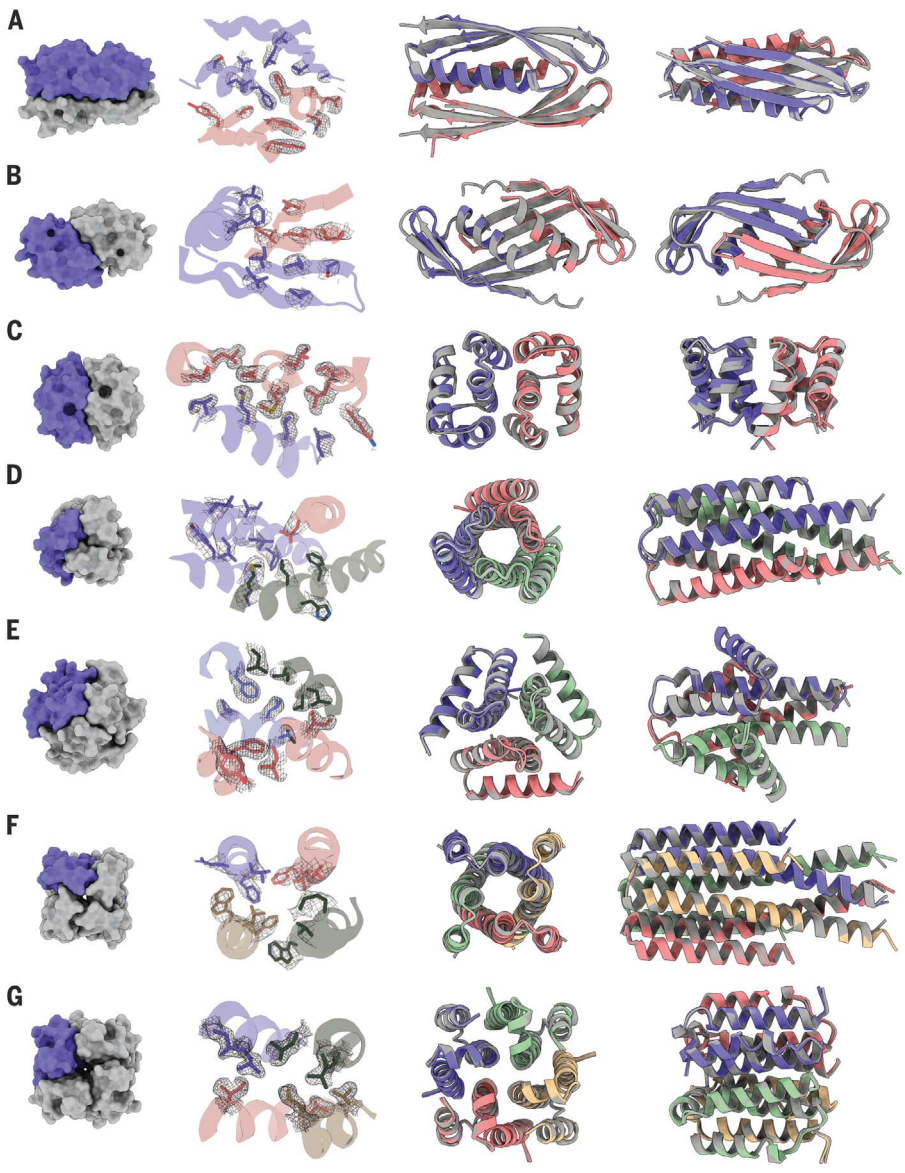

Fig. 2. Structures of HALs solved by x-ray crystallography are very close to their design models. (**A**) HALC2_ 062 (RMSD: 0.81 Å). (**B**) HALC2_065 (RMSD: 1.02 Å). (**C**) HALC2_068 (RMSD: 0.86 Å). (**D**) HALC3_104 (RMSD: 0.42 Å). (**E**) HALC3_109 (RMSD: 0.46 Å). (**F**) HALC4_135 (RMSD: 0.60 Å). (**G**) HALC4_136 (RMSD: 0.34 Å). For each row, the first panel (from the left) 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).

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 matched those of the design models (Fig. 2).

The solved structures exhibit notable diversity, with many intricate structural features. HALC2_062 (Fig. 2A) is a three-layer homodimer with a single helix from each protomer packed together between two outer β sheets (one from each protomer), whereas HALC2_065 (Fig. 2B) is also a mixed α/β homodimer 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 six interacting helices (three 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 considerably from its closest structural relative in the PDB (TM-score: 0.57) (Fig. 3C and table S2). HALC3_104 (Fig. 2D) is a homotrimeric 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 the 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 homotrimeric 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 C_4 symmetry instead of C_3 symmetry, 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 three-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 C_5 symmetry rather than the C_4 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 C_{42}) 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., C_{15} into a pentamer, shorthanded as $C_{15\text{-}5}$). 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 trajectory (<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 C_5 to C_{42} with their largest dimension ranging from 7 to 14 nm (median: 10 nm), which were then divided into homotrimers, -tetramers, -pentamers, -hexamers, -heptamers, -octamers, and a dodecamer, and the backbones were redesigned with ProteinMPNN (Fig. 1, A and B). Although the α/β topology of some of

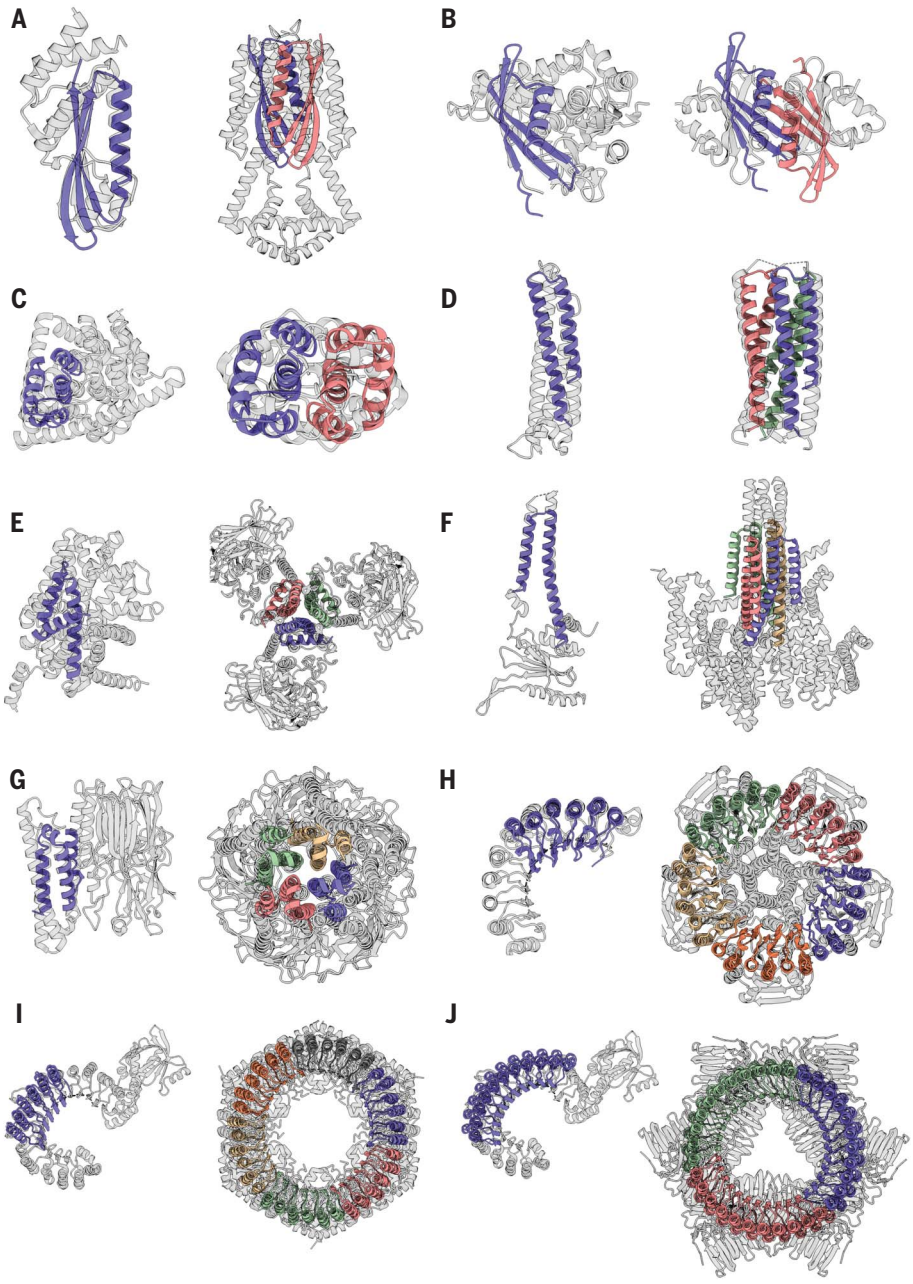

Fig. 3. Hallucinated structures differ considerably from their closest matches in the PDB. For each structure solved by crystallography (Fig. 2) or cryo-EM (Fig. 4B), the closest structural matches to the protomer and the oligomer are shown on the left and right, respectively. Designs are colored by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

these larger HALs is reminiscent of natural leucine-rich repeats (LRRs) (*35*), which is reflected by a median highest protomer TM-score 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 (basic local alignment search tool) 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 two-dimensional (2D) class averages and 3D ab initio reconstructed electron density maps for six designs with C_6 to C_{42} internal repeat symmetry (factorized as two C_5, three C_6, and one C_7) that clearly showed low-resolution structural features and diameters consistent with their designs (Fig. 4A and fig. S12). We selected three designs: one C_{15} homopentamer (HALC5-15_262), one C_{18} homohexamer (HALC6-18_265), and one C_{33} homotrimer (HALC3-33_343) for high-resolution single-particle cryo–electron microscopy (cryo-EM) 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-, 6.51-, and 6.32-Å resolution, respectively (Fig. 4B and figs. S13 to S16). HALC5-15_262 was originally designed as a homohexamer, but structure prediction calculations were more consistent with a pentameric structure of nearly identical protomer conformation and a very slightly shifted subunit interface (fig. S17); the cryo-EM structure is also a pentamer with a 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 cryo-EM, HALC5-15_262, HALC6-18_265, and HALC3-33_343, are 87, 99, and 100 Å in diameter, respectively, 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, whereas 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. Although both structures have matches to LRRs for their protomers (TM-score of 0.65 for both, but to different structures), the oligomeric assemblies are very 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 cryo-EM reconstruction; the protomer has a low TM-score

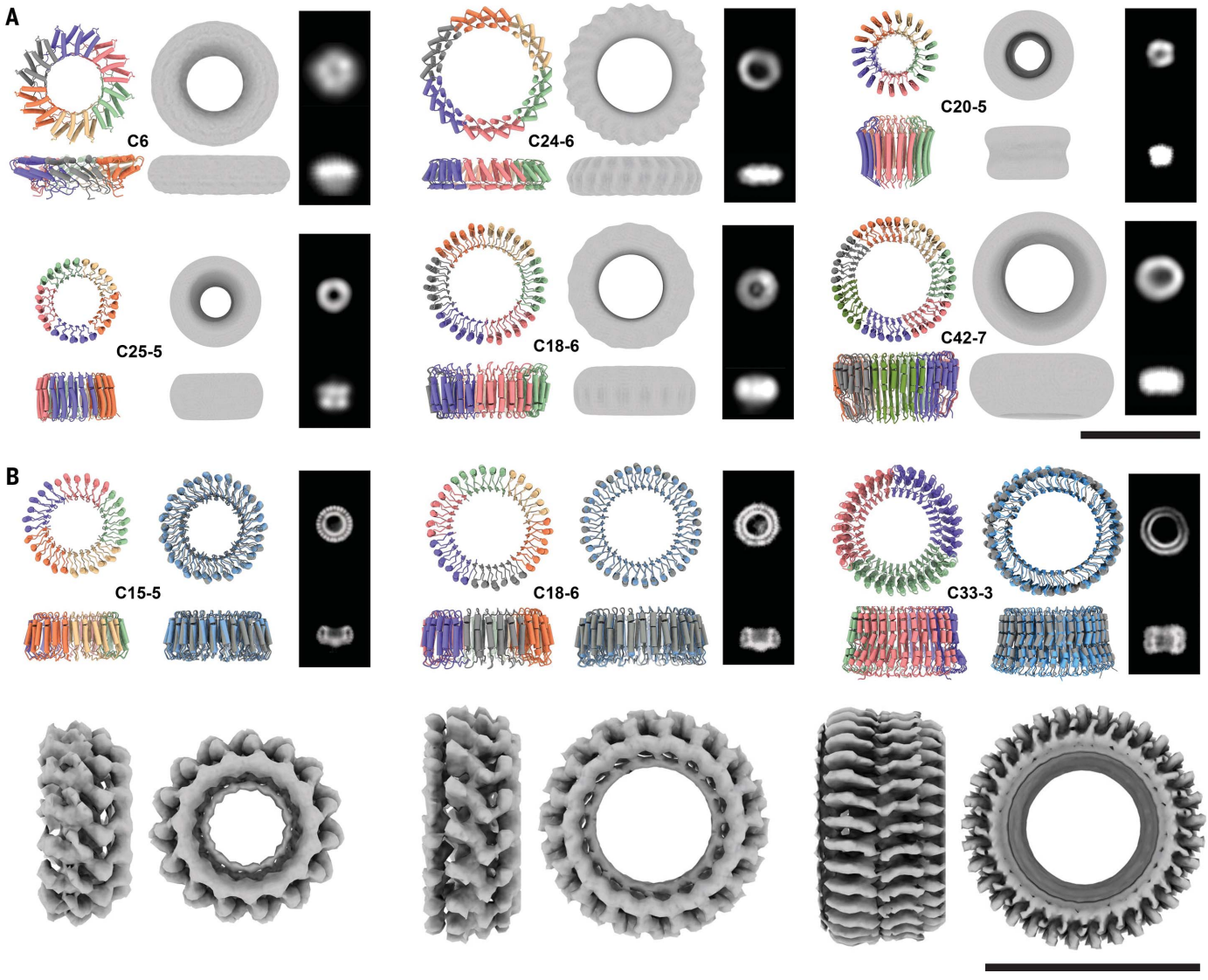

Fig. 4. Cryo–electron microscopy and negative stain electron microscopy validation of large HALs. For each design, the model is shown colored by chain and the corresponding internal symmetry (X) and oligomerization state (Y) are indicated (CX-Y). The electron density map is shown next to the model alongside characteristic 2D class averages. (**A**) Negative stain characterization of HALs. Ring diameters are 92, 110, 75, 80, 100, and 107 Å for HALC6_220, HALC24-6_316, HALC20-5_308, HALC25-5_341, HALC18-6_278, and HALC42-7_351, respectively. (**B**) Cryo-EM characterization of three large HALs. The ring diameters are 87, 99, 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 cryo-EM model (gray) and design model (blue). The computed backbone atom RMSD between the designed and experimental structure is 0.81, 1.69, and 2.30 Å, respectively (fig. S16). Bottom row: 4.38, 6.51, and 6.32 Å cryo-EM electron density maps. Scale bars, 10 nm.

(0.48) despite having an LRR-like topology, and the oligomer is even further from any currently known structure (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 IDs 4HL8 and 7PKY].

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, cryo-EM, 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 is typical for Rosetta de novo designs,

suggesting 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*) that 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, *Nature* **548**, 244–247 (2017).
2. I. G. Johnston *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2113883119 (2022).
3. S. E. Ahnert, J. A. Marsh, H. Hernández, C. V. Robinson, S. A. Teichmann, *Science* **350**, aaa2245 (2015).
4. S. K. Burley *et al*., *Nucleic Acids Res.* **47**, D520–D528 (2019).
5. D. S. Goodsell, A. J. Olson, *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153 (2000).
6. T. Handel, W. F. DeGrado, *J. Am. Chem. Soc.* **112**, 6710–6711 (1990).
7. P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, P. S. Kim, *Science* **282**, 1462–1467 (1998).
8. J. A. Fallas *et al*., *Nat. Chem.* **9**, 353–360 (2017).
9. A. R. Thomson *et al*., *Science* **346**, 485–488 (2014).
10. P.-S. Huang *et al*., *Nat. Chem. Biol.* **12**, 29–34 (2016).
11. S. E. Boyken *et al*., *Science* **352**, 680–687 (2016).
12. L. Doyle *et al*., *Nature* **528**, 585–588 (2015).
13. J. B. Bale *et al*., *Science* **353**, 389–394 (2016).
14. I. Vulovic *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015037118 (2021).
15. Y. Hsia *et al*., *Nat. Commun.* **12**, 2294 (2021).
16. C. E. Correnti *et al*., *Nat. Struct. Mol. Biol.* **27**, 342–350 (2020).
17. D. D. Sahtoe *et al*., *Science* **375**, eabj7662 (2022).
18. I. Anishchenko *et al*., *Nature* **600**, 547–552 (2021).
19. M. Jendrusch, J. O. Korbel, S. K. Sadiq, 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, bioRxiv 2021.08.24.457549 [Preprint] (2021). https://doi.org/10.1101/2021.08.24.457549.
21. J. Wang *et al*., bioRxiv 2021.11.10.468128 [Preprint] (2021). https://doi.org/10.1101/2021.11.10.468128.
22. S. Ovchinnikov, P.-S. Huang, *Curr. Opin. Chem. Biol.* **65**, 136–144 (2021).
23. C. Norn *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2017228118 (2021).
24. N. Anand *et al*., *Nat. Commun.* **13**, 746 (2022).
25. C. Hsu *et al*., bioRxiv 2022.04.10.487779 [Preprint] (2022). https://doi.org/10.1101/2022.04.10.487779.
26. J. Jumper *et al*., *Nature* **596**, 583–589 (2021).
27. V. Mariani, M. Biasini, A. Barbato, T. Schwede, *Bioinformatics* **29**, 2722–2728 (2013).
28. Y. Zhang, J. Skolnick, *Nucleic Acids Res.* **33**, 2302–2309 (2005).
29. J. Xu, Y. Zhang, *Bioinformatics* **26**, 889–895 (2010).
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 imagesarXiv:1412.1897 [cs.CV] (2015).
32. K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: visualising image classification models and saliency mapsarXiv:1312.6034 [cs.CV] (2014).
33. J. Dauparas *et al*., *Science* **378**, 49–56 (2022).
34. M. Baek *et al*., *Science* **373**, 871–876 (2021).
35. B. Kobe, J. Deisenhofer, *Trends Biochem. Sci.* **19**, 415–421 (1994).
36. P. Guerra *et al*., *Sci. Adv.* **8**, eabj7795 (2022).
37. N. Anand, T. Achim, Protein structure and sequence generation with equivariant denoising diffusion probabilistic modelsarXiv: 2205.15019 [q-bio.QM] (2022).
38. B. L. Trippe *et al*., Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problemarXiv:2206.04119 [q-bio.BM] (2022).
39. B. I. M. Wicky, L. F. Milles, A. Courbet, bwicky/oligomer_hallucination: Initial release, version 0.1, Zenodo (2022); https://doi.org/10.5281/zenodo.7112141.

ACKNOWLEDGMENTS

We thank I. Anishchenko, S. Ovchinnikov, W. Sheffler, J. Hansen, C. Norn, D. Zorine, L. Goldschmidt, and T. Huddy for helpful discussions. **Funding:** This work was supported with funds provided by the Audacious Project at the Institute for Protein Design (A.K., L.C., X.L., E.K., S.T., and D.B.), a grant from the National Institute of General Medical Sciences (P41 GM 103533-24, to R.D.K.), an EMBO long-term fellowship (ALTF 139-2018, to B.I.M.W.), a grant from the National Science Foundation (CHE-1629214, to D.B.), the Open Philanthropy Project Improving Protein Design Fund (H.N., A.K.B., R.J.R., J.D., and D.B.), an Alfred P. Sloan Foundation Matter-to-Life Program Grant (G-2021-16899, to A.C. and D.B.), a Human Frontier Science Program Cross Disciplinary Fellowship (LT000395/2020-C, to L.F.M.), an EMBO Non-Stipendiary Fellowship (ALTF 1047-2019, to L.F.M.), and the Howard Hughes Medical Institute (A.C. and D.B.). Cryo-EM was performed on a Glacios microscope purchased by the University of Washington Arnold and Mabel Beckman Center for Cryo-EM (D.B.) with an 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 was performed using 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: A.C., B.I.M.W., L.F.M., and D.B. Methodology: A.C., B.I.M.W., L.F.M., and D.B. Software: A.C., B.I.M.W., L.F.M., J.D., M.B., F.D., and R.D.K. Validation: A.C., B.I.M.W., L.F.M., S.T., E.K., R.J.R., and A.K.B. Formal analysis: A.C., B.I.M.W., L.F.M., D.B., and R.J.R. Investigation: A.C., B.I.M.W., L.F.M., S.T., E.K., X.L., L.C., A.K.B., A.K., and H.N. Resources: A.C., B.I.M.W., L.F.M., M.B., F.D., and D.B. Data curation: A.C., B.I.M.W., L.F.M., D.B., A.K.B., R.J.R., X.L., and L.C. Writing – original draft: A.C., B.I.M.W., L.F.M., and D.B. Writing – review & editing: A.C., B.I.M.W., L.F.M., and D.B. Visualization: A.C., B.I.M.W., L.F.M., and R.J.R. Supervision: D.B. Project administration: A.C., B.I.M.W., L.F.M., and D.B. Funding acquisition: A.C., B.I.M.W., L.F.M., and D.B. **Competing interests:** B.I.M.W., L.F.M., A.C., R.J.R., J.D., E.K., S.T., R.D.K., and D.B. are inventors on a provisional patent application (63/368,093) 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 (*39*). Crystallographic datasets have been deposited in the PDB (IDs: 8D03, 8D04, 8D05, 8D06, 8D07, 8D08, and 8D09). EM maps have been deposited in the EMDB (accession codes: EMD-27658, EMD-27659, and EMD-27660). **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.add1964
Materials and Methods
Figs. S1 to S17
Tables S1 to S4
References (*40*–*60*)

View/request a protocol for this paper from *Bio-protocol*.

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

Hallucinating symmetric protein assemblies

B. I. M. Wicky, L. F. Milles, A. Courbet, R. J. Ragotte, J. Dauparas, E. Kinfu, S. Tipps, R. D. Kibler, M. Baek, F. DiMaio, X. Li, L. Carter, A. Kang, H. Nguyen, A. K. Bera, and D. Baker

Science, **378** (6615), .
DOI: 10.1126/science.add1964

Deep learning takes on protein design

Deep learning approaches such as Alphafold and Rosettafold have made reliable protein structure prediction broadly accessible. For the inverse problem, finding a sequence that folds to a desired structure, most approaches remain based on energy optimization. In two papers, a range of protein design problems were addressed through deep learning methods. Dauparas *et al*. built on recent deep learning protein design approaches to develop a method called ProteinMPNN. They validated designs experimentally and showed that ProteinMPNN can rescue previously failed designs made using Rosetta or Alphafold. Wicky *et al*. started from a random sequence and used Monte Carlo sequence search coupled with structure prediction by Alphafold to design cyclic homo-oligomers. Although the designs were generated to achieve stable expression, the sequences had to be regenerated using ProteinMPNN. This approach allowed for the design of a range of experimentally validated cyclic oligomers and paves the way for the design of increasingly complex assemblies. —VV