Novel backbone methods for de novo protein design

Isaac D. Lutz

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

David Baker, Chair

Hao Yuan Kueh

Drew Sellers

Georg Seelig

Program Authorized to Offer Degree:

Department of Bioengineering

Abstract

Novel backbone methods for de novo protein design

Isaac D. Lutz

Chair of the Supervisory Committee:

David Baker

Department of Biochemistry

Solving a protein design problem first requires sampling suitable backbones given the needs and constraints of the problem. The available structural space of backbones is vast, containing countless potential solution backbones for a given problem. Previous methods to explore this space include parametric sampling, fragment assembly, and more recently, generative deep learning methods. With more advanced methods and algorithms, we can more effectively sample this space to solve new protein design problems. In this work, I present three protein design projects focused on backbone sampling methods. First, I describe a traditional parametric approach to redesign heterodimers for synthetic protein logic. Second, I describe a generative reinforcement learning approach I developed to design protein architectures from the top–down. This method can fill arbitrary volumes and enables the design of capsids for vaccine antigen presentation. Third, I describe a collection of methods used to accomplish helical peptide recognition. The resulting high-affinity binders are useful as capture reagents for disease diagnosis, and can be engineered into biosensors.

Acknowledgments

I would like to recognize and thank the many scientists in the Institute for Protein Design and collaborators at other institutions who made this work possible and supported me through my graduate research. These include the authors listed in the manuscripts in this work, as well as countless others that helped me through valuable discussions, experimental assistance, and general guidance. It has truly been a pleasure working in this unique space with you all.

Thank you to Basile Wicky, for his fantastic mentorship and for providing a crucial foundation in protein design. Thank you to Shunzhi Wang and Chris Norn, for making reinforcement learning in protein design a reality. Thank you to Susana Vázquez Torres and Phil Leung, for taking helical peptide recognition to the highest affinity. Lastly, thank you to David Baker, for guiding me at every step along the way and supporting my exploration of this exciting field.

Table of Contents

Chapter 1

Parametric sampling for heterodimer-based protein logic

Introduction

De novo protein design is a sampling problem, requiring first the sampling of backbone structures followed by the sampling of amino acid sequences to fold to the desired structure. Backbone methods must generate structures suitable for solving a given protein design problem, by both satisfying the constraints of the problem as well as providing an optimal solution given desired structural attributes. The structural space of solution backbones is usually vast for a given problem, and so an effective method must be able to balance the exploration of diverse solutions with adequate sampling of those most likely to work. A variety of backbone sampling methods have been utilized, including parametric sampling, fragment assembly, and deep learning hallucination (*1*–*3*). Although these methods have solved many protein design problems, there are plenty of other problems with no effective ways to sample solutions, and backbone sampling methods in general are relatively underexplored. With improvements on existing methods and novel, advanced algorithms for backbone sampling, new structural spaces can be accessed to expand the field of de novo protein design.

Main Text

Recent work from the Baker lab describes the de novo design of protein heterodimers (*4*). These heterodimers consist of split four-helix bundles with hydrogen-bond networks installed at their interfaces to confer specificity (*4*). The authors successfully designed and characterized a large set of highly orthogonal pairs, and proposed that these programmable orthogonal interactions, analogous to Watson-Crick base pairing of nucleic acids, could lead to the design and application of protein logic (*4*). However, an obstacle still remains: these heterodimer pairs do not readily associate when expressed separately and combined, and require co-expression or denaturation and reannealing after combining to behave properly (*4*). This is likely due to monomer instability, as each heterodimer half consists of a two-helix hairpin with only a minimal protein core and a large exposed hydrophobic interface. As a result, these halves tend to irreversibly homodimerize when expressed on their own, severely limiting downstream applications.

We redesigned a set of these heterodimers by using parametric sampling to add a third buttressing helix to each of the hairpins (Fig. 1A). We used exhaustive grid sampling to generate many candidate buttressing helices per heterodimer, selecting parameter ranges and increments for helix distance, tilt, phase, and inversion. Experimental characterization through binding assays (Fig. 1B) and x-ray crystallography (Fig. 2) proves that the redesign successfully created a stable protein core for each half of the heterodimers and enabled homodimer exchange. These heterodimers can now be used for protein logic applications, including a new way to perform computation using networks of promiscuously interacting heterodimer halves. This project demonstrates that simple grid parametric sampling is an effective backbone sampling method and sufficient to solve many protein design problems and unlock new functions.

Conclusion

Parametric sampling is a simple yet powerful tool in protein design that can be used to solve a wide array of protein design problems. The approach demonstrates the efficacy of simple, ideal backbone elements like parametric alpha helices and short loops. However, parametric sampling is difficult to scale and apply to more complex protein design problems: the approach is very manual, requiring the choice of parameter ranges and increments as well as specification of an overall protein fold. This requires many choices from the protein designer given their understanding of the problem setup, and limits solutions to those that the designer chooses to parametrically describe. For some highly complex or constrained protein design problems, simply coming up with a potential solution protein fold can be nearly impossible. With a generative method, we could take the same proven ideal backbone elements and make the sampling process automatic, enabling discovery of solution backbones beyond human capabilities. The job of the protein designer then becomes problem formulation, in describing the constraints and desired properties of possible solutions (i.e. designing from the "top-down"). The job of the generative method is to find creative and optimal backbone solutions provided the problem formulation for a given protein design challenge, unlocking scale and complexity in protein backbone design. In the next chapter I will describe such a method, which uses reinforcement learning to accomplish automated and top-down protein design.

Figures

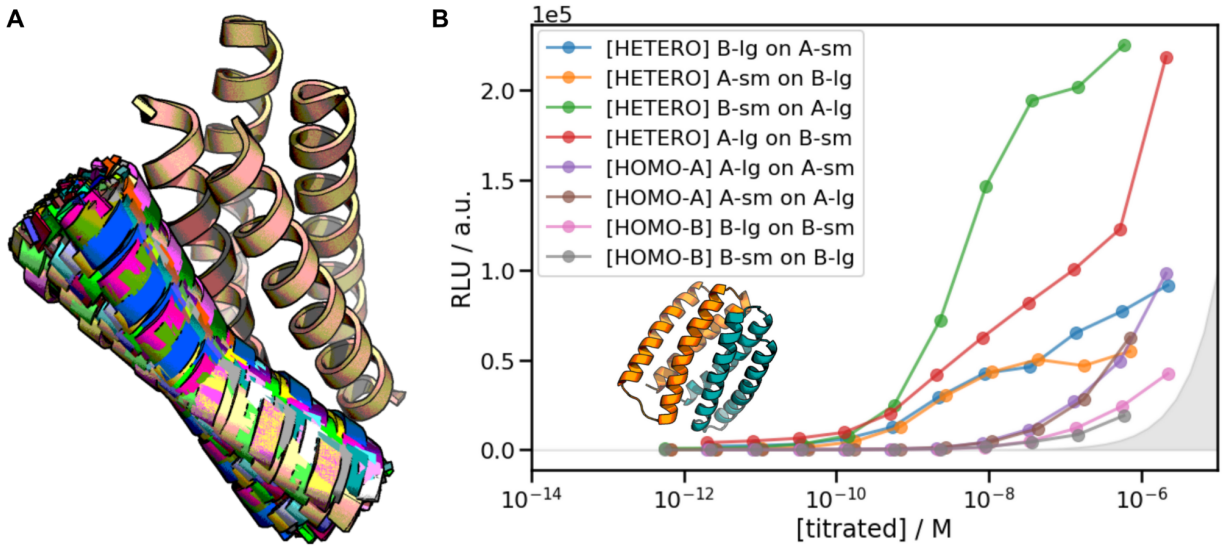

Fig. 1. Redesign of heterodimers to enable homodimer exchange.
(**A**) Grid sampling a third buttressing helix for each heterodimer hairpin along parameter ranges and increments for helix distance, tilt, phase, and inversion. (**B**) Example binding assay for redesigned heterodimer pairs. Heterodimer halves were separately expressed then mixed, demonstrating higher affinity heterodimeric binding and successful homodimer exchange.

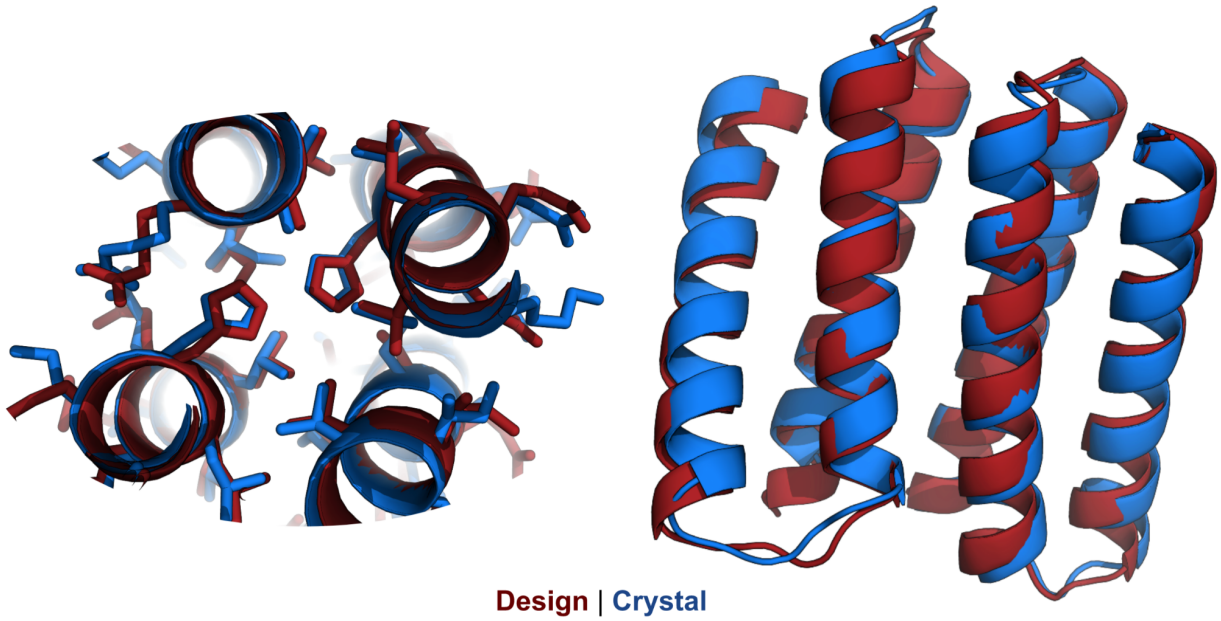

Fig. 2. Crystal structure of a redesigned heterodimer closely matches the design model. Interface atomic accuracy and an overall RMSD of 1.033 Å further demonstrates the success of the computational redesign strategy.

References

1. Grigoryan, G., & Degrado, W. F. (2011). Probing designability via a generalized model of helical bundle geometry. *Journal of molecular biology*, 405(4), 1079–1100. https://doi.org/10.1016/j.jmb.2010.08.058
2. Jacobs, T. M., Williams, B., Williams, T., Xu, X., Eletsky, A., Federizon, J. F., Szyperski, T., & Kuhlman, B. (2016). Design of structurally distinct proteins using strategies inspired by evolution. *Science* (New York, N.Y.), 352(6286), 687–690. https://doi.org/10.1126/science.aad8036
3. Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A. K., DiMaio, F., Carter, L., Chow, C. M., Montelione, G. T., & Baker, D. (2021). De novo protein design by deep network hallucination. *Nature*, 600(7889), 547–552. https://doi.org/10.1038/s41586-021-04184-w
4. Chen, Z., Boyken, S. E., Jia, M., Busch, F., Flores-Solis, D., Bick, M. J., Lu, P., VanAernum, Z. L., Sahasrabuddhe, A., Langan, R. A., Bermeo, S., Brunette, T. J., Mulligan, V. K., Carter, L. P., DiMaio, F., Sgourakis, N. G., Wysocki, V. H., & Baker, D. (2019). Programmable design of orthogonal protein heterodimers. *Nature*, 565(7737), 106–111. https://doi.org/10.1038/s41586-018-0802-y

Chapter 2

Top-down design of protein architectures with reinforcement learning

As originally published in Lutz, I. D., Wang, S., Norn, C., Courbet, A., Borst, A. J., Zhao, Y. T., Dosey, A., Cao, L., Xu, J., Leaf, E. M., Treichel, C., Litvicov, P., Li, Z., Goodson, A. D., Rivera-Sánchez, P., Bratovianu, A. M., Baek, M., King, N. P., Ruohola-Baker, H., & Baker, D. (2023). Top-down design of protein architectures with reinforcement learning. *Science* **380**, 266-273. doi:10.1126/science.adf6591

Abstract

As a result of evolutionary selection, the subunits of naturally occurring protein assemblies often fit together with substantial shape complementarity to generate architectures optimal for function in a manner not achievable by current design approaches. We describe a "top-down" reinforcement learning–based design approach that solves this problem using Monte Carlo tree search to sample protein conformers in the context of an overall architecture and specified functional constraints. Cryo–electron microscopy structures of the designed disk-shaped nanopores and ultracompact icosahedra are very close to the computational models. The icosohedra enable very-high-density display of immunogens and signaling molecules, which potentiates vaccine response and angiogenesis induction. Our approach enables the top-down design of complex protein nanomaterials with desired system properties and demonstrates the power of reinforcement learning in protein design.

Main Text

Multisubunit protein assemblies play critical roles in biology and are the result of evolutionary selection for function of the entire assembly. Therefore, the subunits in structures such as icosahedral viral capsids often fit together almost perfectly (*1*, *2*). In contrast to direct evolutionary selection on overall system properties, de novo protein design has generated protein architectures using a "bottom-up" hierarchical approach (Fig. 1A, left) in which monomeric structures are first docked into symmetric oligomers (*3–6*) and then assembled into closed assemblies with tetrahedral, octahedral, or icosahedral symmetry (*7–14*) or open assemblies such as two-dimensional (2D) layers and 3D crystals (*15–19*). An advantage of this hierarchical approach is that the multiple interfaces that stabilize the assembly can be validated independently (the first by characterization of the symmetric oligomer and the second by characterization of the nanomaterial assembly from the preformed oligomer), considerably increasing the robustness of the overall design process. Although such designed assemblies are already proving useful for biomedicine in immunobiology and other areas, as highlighted by the recent approval of a de novo–designed COVID vaccine (*20–23*), the bottom-up approach does have limitations. The properties of the assembly are limited to what can be generated from the available oligomeric building blocks, at least one of the subunit-subunit interfaces must be strong enough to stabilize a cyclic oligomeric substructure in isolation, and, more generally, there is no way to directly optimize the properties of the overall assembly.

We sought to overcome the limitations of bottom-up protein complex design by developing a top-down approach (Fig. 1A, right) that starts from a specification of the desired properties (overall symmetry, porosity, etc.) of the structure and systematically builds up subunits that pack together to optimize these properties. We reasoned that protein fragment assembly (*24–28*), which can generate a wide variety of monomeric protein structures, could provide a suitable mechanism for generating diversity. Previous design approaches such as SEWING have built up proteins from fragments, optimizing for monomer stability at each step (*29*), but we aimed instead to optimize for overall system properties, which could involve trading off monomer stability for increased subunit-subunit interaction strength and other properties. To enable such end state–based optimization, we turned to reinforcement learning (RL), which has achieved considerable success recently in different fields of artificial intelligence, such as self-driving cars (*30*), the AlphaGo program that defeats top human players in the game of Go (*31*, *32*), and algorithm development (*33*). Monte Carlo tree search (MCTS) (*34*, *35*) is an RL algorithm that finds optimal series of choices within a search tree. In MCTS, choices are selected randomly at each branch point to find a path down the tree, and after exploring a path, the state is evaluated, and probabilities at each branch point back-propagated up the tree are reweighted accordingly such that subsequent iterations are more likely to lead to optimal paths.

Backbone sampling by MCTS

We sought to develop a MCTS algorithm for generating protein complexes that builds up the monomeric subunits from protein fragments directly optimizing for prespecified global structural properties. We set up the tree search such that at each step in the tree, a short protein fragment is appended at either the N terminus or C terminus of the growing chain. The number of fragments to consider at each step is a trade-off between the rapidity of learning (with a smaller number, weights on each choice can be learned more quickly) and the total diversity of structures that can be generated (which increases with the number of choices at each step). We chose to balance these factors by using as building blocks parametrically generated straight helices, which are fully described by a single parameter (the length, which we allow to vary from nine to 22 residues), followed by short loops clustered into 316 bins (derived from clustering loops in a large helical protein database; see the materials and methods). The search begins with the selection of one of the helix possibilities and then alternates between the addition of a loop or a helix choice at either terminus. Once a loop bin is chosen, we select randomly from the closely related loop backbones within the cluster (Fig. 1B, left). Although this is a far narrower set of local structures than observed in native protein structures, we found in preliminary explorations that a wide variety of compact protein shapes could be readily generated from such building blocks. Building up a 100-residue protein backbone with this approach requires about five helix and four loop additions, yielding a total number of possibilities of $\sim 1 \times 10^{17}$, with additional structural diversity from the variation in loop backbones within a bin. The size of the search tree grows exponentially with the number of structural elements, so the space of possibilities is more effectively explored for monomers with fewer helices than for larger monomers.

The search is modulated based on the specific problem specification through geometric constraints that are applied at each step in the search tree and score functions that are evaluated only after full structures are completed. Potential moves consisting of helix or loop fragments are selected at each level of the search tree only if they pass geometric constraints that can be evaluated before the assembly of the entire structure; these include internal clashes and overall shape constraints (see the materials and methods for a full list of geometric constraints). Upon selection of a move passing the geometric constraints, its probability is upweighted, as are the probabilities of all prior moves leading to this point in the search tree. Completed backbones are evaluated using score functions that assess how well the overall generated structure satisfies the user specification of the problem to be solved (Fig. 1C and materials and methods), and the probabilities of selection of each move at each step along the search tree are reweighted accordingly. As individual move weights become increasingly biased after many traversals through the search tree, the generated complete backbones have higher and higher scores (fig. S1). Because each iteration takes on average only tens of milliseconds, high-scoring backbones can be sampled at scale by searching over tens of thousands of iterations. To address the classical RL problem of balancing exploration with exploitation (*30*–*32*), the search is initialized from

many independent trees, and the maximum probability of any one move is capped (see the materials and methods).

We first tested the MCTS approach in silico at the protein monomer level, choosing as a test problem the generation of protein backbones with arbitrarily prespecified overall shapes. To our knowledge, there are no current approaches for addressing this problem. A specified build volume is represented on a grid, and the MCTS is initialized randomly within the volume. At each move, only additions that stay within the specified volume are accepted. For a range of prescribed shapes, including regular polyhedra and letters from the alphabet, the ensembles of generated structures closely fill the specified volumes, and individual backbones have the prespecified shapes (fig. S2). The average sequence length of the solutions increases through the optimization as the choices of moves and combinations of moves that lead to satisfaction of the input constraints are learned, enabling traversal further down the search tree (fig. S1A).

We next sought to generalize the MCTS to the design of symmetric nanomaterials by applying symmetry operators to generate assemblies with the desired symmetry at each step in the search tree. Each move (helix or loop addition) is assessed by considering not only the growing monomer, but also its interactions with all nearby symmetry mates, computed using transformation matrices specifying each symmetry operator; moves that introduce steric clashes are discarded (Fig. 1, B and C). We tested these capabilities in silico by designing cyclic assemblies with symmetries C5 through C12, as well as tetrahedral, octahedral, icosahedral, and quasisymmetric icosahedral assemblies of up to 240 subunits (figs. S3 and S4). We found that by providing different geometric constraints and score functions to guide the search, we could control properties such as shape, size, porosity, and termini position from the top down (figs. S3 to S6).

Nanopore construction using constrained symmetric MCTS

As a first experimental test of the MCTS approach, we applied it to the highly constrained design challenge of filling the space between two previously designed cyclic protein rings (*6*, *36*) to generate disk-shaped structures with a central nanopore (Fig. 2A). Filling this substantial but irregularly shaped space such that there are no large voids between the two rings is not straightforward with previously described protein design methods. We approached this challenge with MCTS by geometrically constraining the search to the space between the two rings, requiring dense packing such that the only large void in the resulting assembly is the pore of the inner C6 ring. Both the inner and the outer ring have C6 symmetry, and the search tree was initialized to start at the N termini of the outer ring and simultaneously build six subunits that collectively fill the empty space. We performed the MCTS for each of 2000 placements of a set of different inner rings with a range of inner pore sizes inside a constant outer ring (for each

inner ring, we sampled rotations around and translations along the common cyclic symmetry axis). We selected backbones that fully filled the space between the two rings, designed sequences with ProteinMPNN (*37*), and selected for experimental characterization 32 designs predicted to assemble into the designed assemblies by AlphaFold (AF) (*38*). Of these, we found that 28 were soluble and could be purified and 11 formed particles with the expected size and shape by negative-stain electron microscopy (nsEM). nsEM 3D reconstructions for two designs had an overall shape closely consistent with that of the design models (Fig. 2B; some C7 2D class averages were also obtained; fig. S7). We obtained a cryo–electron microscopy (cryo-EM) map of a third design at 5.1-Å resolution and found it to be closely consistent with the design model: The alpha helices of the model are clearly within the contours of the density (Fig. 2C and fig. S8). The MCTS solution effectively satisfies the design criteria: The space between the two original rings is completely filled in, generating a disk-like structure with a narrow circular pore in the center. We are not aware of any previously designed or naturally occurring proteins that have this overall shape, which could be very useful for downstream nanopore-based sensing applications. More generally, these results demonstrate that the MCTS approach can solve highly constrained protein design problems.

Top-down design of mini-icosahedra

We next explored the use of MCTS to generate icosahedral assemblies by using 59 transformation matrices to compute symmetry mates for a growing monomer. We sought to design very small, closely packed capsids inaccessible by other design methods, and developed geometric constraints and score functions to specifically favor such structures (Fig. 1 and materials and methods). The end state–based score functions include measures of cage porosity and interface designability, as well as external placement of at least one terminus to enable fusion constructs (Fig. 1C). Given a specification of the length and number of helices in the monomer and the size of the overall assembly, we initialized millions of MCTS trajectories starting from a short helical fragment randomly placed within a specified upper distance bound of the origin in a random orientation and performed 10,000 iterations for each to generate a large set of diverse structures. The MCTS generated closely packed icosahedral assemblies in silico, which span a structural space distinct from that of native and previous de novo icosahedra, with shorter sequence lengths than any previously described protein icosahedra and porosities comparable to the densely packed capsids generated by evolution (Fig. 1D).

The MCTS method rapidly generates tens of thousands of candidate icosahedral assemblies, and we experimented with approaches for rapidly designing sequences that stabilize these assemblies in a manner compatible with our overall top-down approach. In previous bottom-up nanocage design studies, the sequences and backbones of the oligomeric building blocks are pre-optimized, so only the new interface formed between the building blocks in the cage is designed, and the

overall backbone is kept largely fixed (11). By contrast, with the top-down MCTS approach, the entire sequence must be designed, with backbone relaxation to optimize sequence-structure compatibility both within and between the monomers and to increase interface shape complementarity. A deep neural network trained to learn the sequence and structure relationships of native proteins was used to generate amino acid sequence profiles for each position in the newly generated backbones, which were used in turn to bias amino acid selection in the sequence design stage using Rosetta design (materials and methods and figs. S9 to S15). The resulting designs were filtered on the basis of interface contact molecular surface area (38), shape complementarity, predicted binding energy, exposed surface hydrophobicity, and AF (39) prediction similarity to the design model (see the materials and methods). The rigid body and internal degrees of freedom of the selected icosahedral assemblies were then optimized by Rosetta symmetric relaxation (40, 41), starting from both the Rosetta design model of the assembly and the AF-predicted structure of the monomer mapped back onto the assembly. To further increase sequence-structure compatibility, we repeated this design-predict-relax cycle three times, at each iteration performing sequence design on the full assemblies generated in the previous iteration, mapping back the predicted monomer structures into the assemblies, and relaxing the full structure in Rosetta. We applied this sequence design and backbone refinement procedure to 220,000 of the MCTS-generated backbones and selected 368 designs for experimental characterization (detailed filtering processes are described in the materials and methods and figs. S11 to S13).

Linear gene fragments encoding each design with hexahistidine purification tags were cloned into an Escherichia coli expression vector, and the proteins produced in E. coli in a 96-well format were purified by immobilized metal affinity chromatography (IMAC) pull-down. A total of 208 of the 368 designs were expressed and soluble as assessed by SDS–polyacrylamide gel electrophoresis. To evaluate particle formation, we performed nsEM on the IMAC elution fraction for each soluble sample. Two designs (RC_I_1 and RC_I_2, RL capsid with I symmetry, design 1 and 2) formed uniform particles with the expected size and shape (Fig. 3, A and B). Size-exclusion chromatography (SEC) of both designs yielded single peaks with an apparent molecular weight in the range expected for these assemblies (Fig. 3, C and D). The designed assemblies had the expected alpha-helical circular dichroism (CD) spectra and apparent melting temperatures above 65°C. nsEM analysis showed that assembly morphologies were retained after 1 hour of treatment at 95°C and subsequent cooling to 25°C (Fig. 3, F and H, and fig. S17).

To evaluate the accuracy of our design strategy, we determined the structures of SEC-purified RC_I_1 and RC_I_2 capsid particles using cryo-EM (Fig. 4 and fig. S18). For RC_I_1, 3D reconstruction yielded a 2.5-Å-resolution cryo-EM atomic model that closely matched the computational design (Fig. 4, A and B, and fig. S19). The N-terminal helices of two monomers pack in an antiparallel fashion to form the primarily hydrophobic C2 interface, whereas the two helices near the C terminus form the C5 interface with their neighbors (Fig. 4, B and C). Small

apertures (diameter ~13 Å) present at the C3 axes of the capsid make the N termini available for genetic fusion (Fig. 4C). Over the designed monomer, the root mean square deviation (RMSD) between the cryo-EM structure and the design model is 0.76 Å (Fig. 4D); a single rotamer flip (Phe63) and tilting of the C-terminal helix results in a slight expansion of the overall cage diameter, resulting in an RMSD over all 60 subunits of 3.72 Å (Fig. 4E). For RC_I_2, the 2.9-Å cryo-EM structure of design RC_I_2 was even closer to the design model (Fig. 4, F and G, and fig. S20), with RMSDs at the C2 and C5 interfaces of 0.66 and 0.27 Å, respectively (Fig. 4H). The RC_I_2 monomer adopts the designed three-helical bundle fold with a 0.59-Å RMSD to the design model (Fig. 4I), and the overall assembly is almost identical to the design model with a 1.39-Å RMSD over all 60 subunits (Fig. 4J). The C2 interface is situated near the extended C terminus of the monomer, allowing for potential monomeric or dimeric genetic fusions. The C5 pentameric interface is mediated by interactions between the N-terminal helices, which point inward and enable functionalization of the interior of the capsid. With diameters of 13 and 10 nm for RC_I_1 and RC_I_2, respectively, and associated monomer lengths of 67 and 54 residues, the designed mini-capsids are considerably smaller than most viral capsids.

Applications of top-down–designed capsids

The compact size and corresponding small exterior surface area of the designed particles enables the display of 60 or 120 copies of N- and/or C-terminal fused proteins with exceptionally high density: six or more times higher than previously designed icosahedral cages. We set out to explore whether this higher density could lead to greater biological efficacy in signaling and vaccine applications. We began by exploring the robustness of the designs to substantial sequence changes and to fusion of proteins to their outward-facing termini.

To evaluate robustness to sequence changes, we used ProteinMPNN (*37*) to generate diverse sequences for the RC_I_1 capsid backbone, and the designs were filtered using the AF and Rosetta metrics described above. Two of six experimentally tested ProteinMPNN designs, RC_I_1-H9 and RC_I_1-H11 (the former designed by ProteinMPNN using the working capsid backbone Cα coordinates as input, the latter the idealized polyA backbone without any backbone optimization and relaxation), assembled into the designed I1 symmetric capsid as evidenced by IMAC, SEC, and nsEM. A 3-Å cryo-EM structure of RC_I_1-H11 was almost identical to the design model, with a monomeric RMSD of 0.60 Å (Fig. 5A and fig. S21) and a very low full-cage RMSD over all 60 subunits of only 0.96 Å. RC_I_1-H9 and RC_I_1-H11 have on average 46% sequence divergence from the parent capsid and 30% sequence difference from each other, including highly diverse interface residue selections (fig. S22; for example, the errant Phe^{63} of the parent capsid was redesigned to Glu^{63} in RC_I_1-H11, likely accounting at least in part for the closer agreement of RC_I_1-11 with the design model). These results demonstrate

that the RL approach can generate directly designable protein backbone geometries with a high degree of accuracy.

We evaluated the robustness of the designs to genetic fusion by fusing SpyTag, SpyCatcher (*42*), and green fluorescent protein (GFP) proteins to the RC_I_1-H11 capsid with an N-terminal $(GGS)_n$ linker (Fig. 5B and figs. S23 to S25). In all cases, SEC elution profiles and nsEM micrographs showed monodisperse particles of the expected size and shape (see the materials and methods). The 2D class averages (inset) revealed spherical structures similar to that of the original icosahedral capsid, with additional density at the periphery of the particles, consistent with fused proteins connected to scaffolds through a flexible linker. Unlike a larger cage, nuclear localization sequence–tagged capsids fused to GFP are efficiently translocated into the nucleus, opening the door to nuclear delivery of high-valency protein and DNA-organizing constructs (fig. S26).

To assess the efficacy of the designed capsids in activating cellular signaling pathways by clustering cell surface receptors, we fused 60 copies of the angiopoietin 1 (Ang1) F domain (Fd), which binds the Tie2 receptor, to RC_I_1-H11 using SpyTag-SpyCatcher conjugation (*14*, *18*, *43*) (see the materials and methods and fig. S27). We found that the F domain–displaying capsids had very high potency in driving FOXO1 exclusion from the nucleus (Fig. 5, C and D), activating the AKT pathway (Fig. 5D and fig. S28, A to C) and stabilizing nascent blood vessels formed from human umbilical vein endothelial cells (HUVECs; Fig. 5E) (*43–49*). The Fd-displaying capsids (0.16 nM RC_I_1-H11-Fd) elicited stronger responses than a 10-fold greater concentration of a much larger F-domain–presenting icosahedral nanoparticle (I53-50) (*12*, *43*); the elevated potency likely results from the higher surface display density [to facilitate comparison, concentrations at the bottom of Fig. 5D are in terms of Fd monomer (0.16 nM capsid × 60 Fd copies per capsid = 10 nM Fd)]. The 0.16 nM (10 nM Fd) capsid also elicited stronger responses than 100 nM Ang1. The F domain–displaying capsid is thus an exceptionally potent Tie2-activating ligand. The designed capsid is also far easier to produce and much more stable than Ang1 and thus could be useful in stimulating differentiation and regeneration.

The high surface presentation density enabled by the designed scaffolds provides a route to investigating the effect of packing density on the elicitation of immune responses by nanoparticle-based immunogens. As a first step in this direction, we fused trimeric influenza hemagglutinin (HA) to the N terminus of I1-capsid RC_I_1 using a $(GS)_6$ linker. The fusion protein was expressed and secreted from mammalian cells and clearly forms HA-displaying particles according to SEC and nsEM (fig. S29 and Fig. 5F). Biolayer interferometry showed binding of both 5J8 [anti-HA head antibody (*50*)] and CR9114 [anti-HA stem antibody (*51*)] immunoglobulin G to HA capsids (Fig. 5G), indicating that the HA remains antigenically intact when displayed on the surface of the capsids. We immunized mice with HA-displaying RC_I_1, as well as a much larger icosahedral immunogen, HA-I53_dn5 (*52*), which has previously been

shown to elicit protective responses against influenza and is currently being evaluated in clinical trials (*53*). We found that HA-displaying RC_I_1 elicited a strong antibody response against vaccine-matched HA that was greater than that produced by the clinical vaccine candidate by a small but statistically significant amount (Fig. 5H). These results indicate that the high antigen presentation density enabled by top-down design can yield robust immune responses.

Conclusion

Our top-down RL approach enables the solution of design challenges inaccessible to previous bottom-up design methods. Cryo-EM structures confirm the design of 54- and 67-residue proteins that assemble into 60-subunit icosahedra with both internal monomer and overall assembly structure nearly identical to the computational models, and of disk-shaped nanopores generated by densely filling the space between cyclic protein rings with different diameters. Both the icosahedra and the disk designs are distinct from any previously designed or naturally occurring structures; the former have smaller subunits, smaller radii, and lower porosities, and the latter have narrow central pores within large, circular, otherwise nonporous structures. These structures could not have been built with previous bottom-up approaches. For the icosahedra, generating the shape complementarity of the interfaces requires the context of the full capsid structure, possible only through a top-down approach, and for the disks, densely filling a prescribed volume from preexisting building blocks is generally not possible. The density of protein chains and termini available for fusion to the icosahedra is considerably greater than the most compact previously designed assembly, enabling fusion to functional protein domains to generate bioactive nanoparticles. The Ang1 F domain–displaying capsids are potent activators of angiogenesis, and the influenza HA–displaying capsids elicit strong anti-HA antibody responses in mice. The capability of the MCTS approach to optimize any set of specified geometric criteria in a top-down fashion provides a route to potent, multivalent cellular receptor agonists and vaccines that are custom designed to rigidly scaffold immunogen or receptor-binding monomers and precisely position them relative to one another. More generally, our results demonstrate the power of RL for protein design, which we expect can be increased further by the incorporation of policy and value networks (*30–32*, *54*) to further guide the search.

Figures

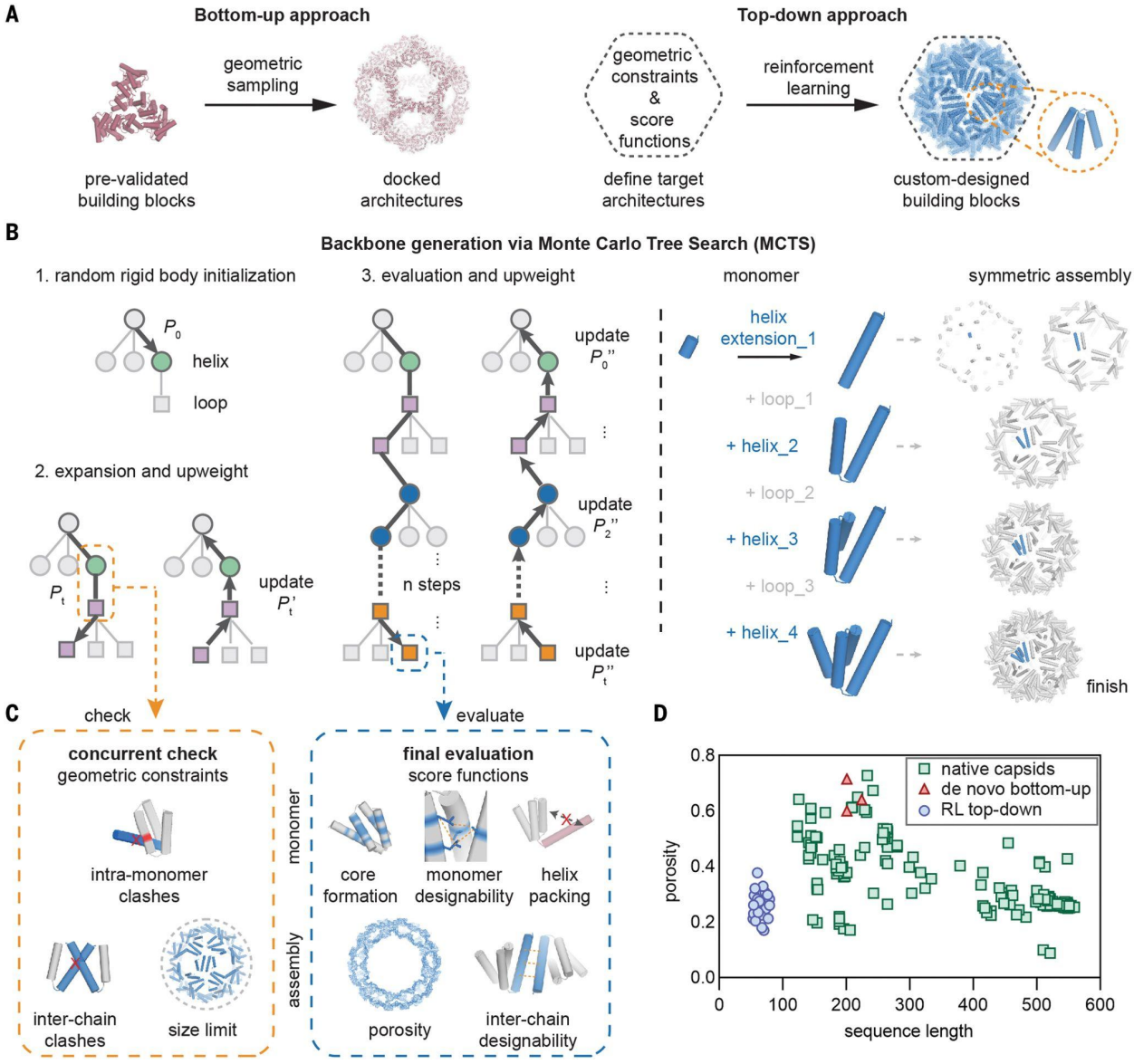

Fig. 1. Top-down design strategy and computational pipeline.
(**A**) Bottom-up (left) and top-down (right) strategies to protein assembly design. (**B**) (Left) MCTS architecture for monomer backbone generation. During each simulation, a helix stub is initialized at a random rigid body start position, and different configurations of helices and loops are sampled and constructed sequentially with probability P_t stored in each edge to build the search tree. Each move is checked against a set of predefined geometric constraints during the expansion stage and then updates probabilities P_t' afterward. Upon successful completion of a search tree, the monomer is evaluated by score functions and probabilities P_t'' are back-propagated to update all of the search tree edges. (Right) Symmetric transformations are applied to build an icosahedral capsid in parallel with monomers using the MCTS generative algorithm. (**C**) Concurrent geometric check (left) is performed at every step of the

expansion stage and the search tree is terminated if there are violations. Final evaluation (right) with a series of score functions is performed upon completion of a simulation for monomers and assemblies. (**D**) In silico RL-generated capsids (blue) occupy a distinct structural space compared with de novo–designed protein cages (red) and natural capsids (green).

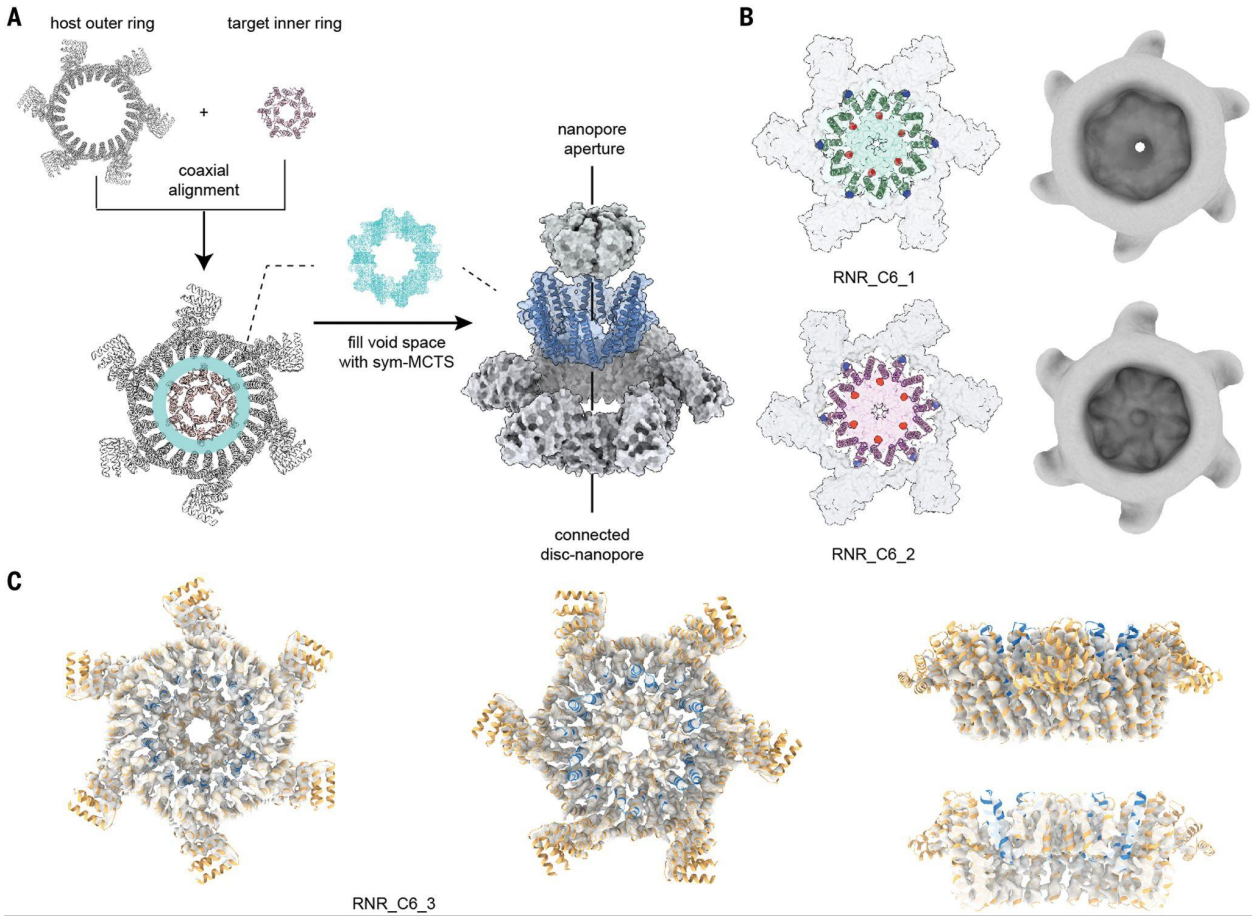

Fig. 2. Disk-nanopore design with symmetric MCTS.
(**A**) Schematic illustration of MCTS-based sampling to build space-filling connectors between two concentric rings to generate disk-like structures with different nanopore inner diameters. The inner ring was placed in the center of a host outer ring, varying the rotation and vertical offset, which generates different void volumes (teal; middle panel above arrows). MCTS was then performed to densely fill these void volumes (blue). (**B**) Design models (left column) and nsEM 3D ab initio reconstruction maps (right column) of two connected disk-nanopores (RNR_C6_1 and RNR_C6_2). The symmetric MCTS sampling built helices to connect the inner ring C terminus and outer ring N terminus (highlighted in red and blue, respectively, in the left column). (**C**) The cryo-EM map at 5.1-Å resolution for design RNR_C6_3 viewed from the top, bottom, and side is very close to the design model, with a narrow circular pore in the center of an otherwise nonporous disk-like structure.

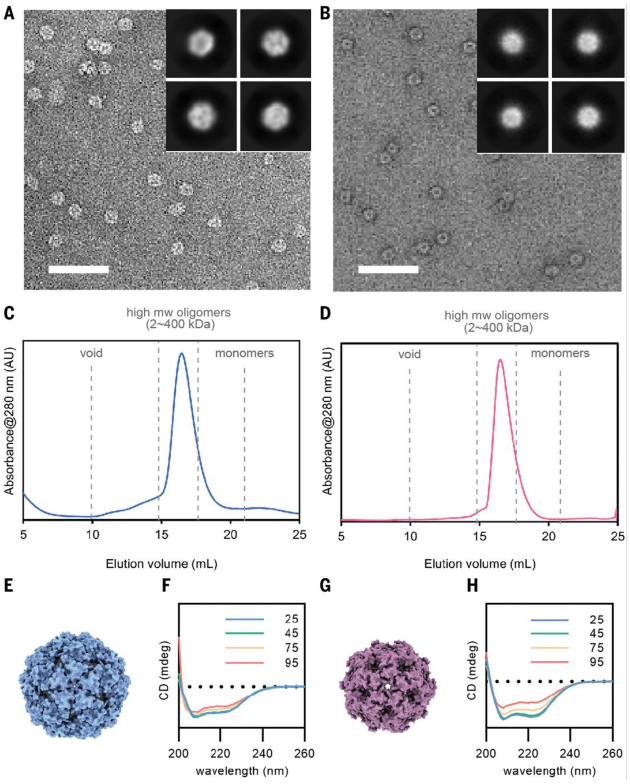

Fig. 3. Experimental characterization of designed capsids RC_I_1 and RC_I_2.
(**A** and **B**) Representative nsEM micrographs and reference-free 2D class averages (inset) for RC_I_1 (left) and RC_I_2 (right). Scale bar, 200 nm. (**C** and **D**) A single peak was observed for each SEC elution profile near the expected elution volumes for the target complexes. (**E** and **G**) Capsid computational design models. (**F** and **H**) Circular dichroism spectra measured at different temperatures (°C).

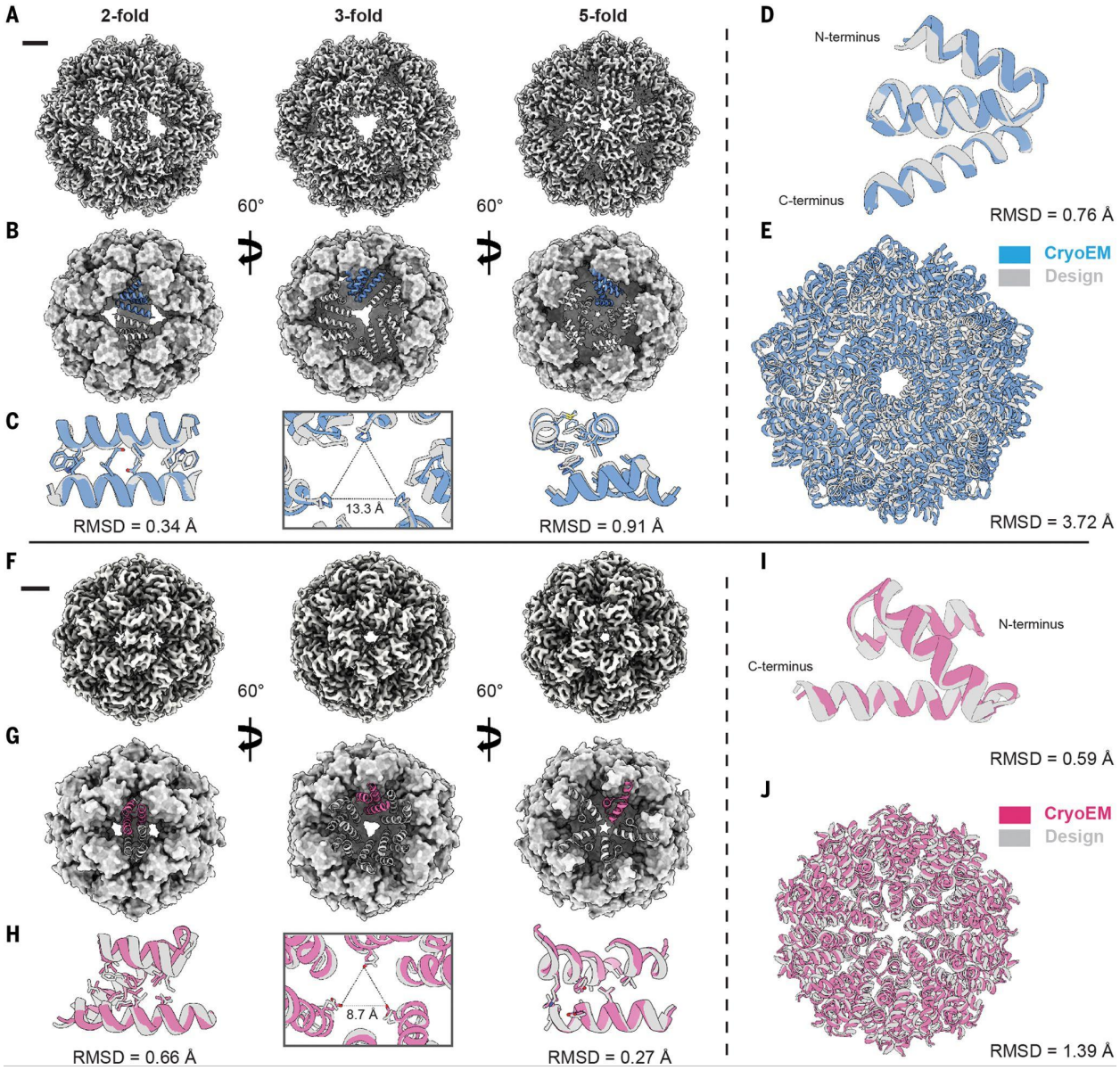

Fig. 4. Near-atomic resolution cryo-EM structures of designed capsids match design models.
(**A**) A 2.5-Å cryo-EM reconstruction of RC_I_1 viewed along the three symmetry axes. Scale bar, 20 Å. (**B**) Cryo-EM structure of RC_I_1 highlighting monomer packing and interfaces along each symmetry axis. (**C**) Overlay and RMSD calculations for RC_I_1 compared with the design model for each symmetry interface (cryo-EM is shown in blue; design is shown in gray). (**D**) Overlay and RMSD calculation for a single monomer of RC_I_1. (**E**) Overlay and RMSD calculation for the entire 60-mer RC_I_1 capsid. (**F**) A 2.9-Å cryo-EM reconstruction of RC_I_2 viewed along the three symmetry axes. Scale bar, 20 Å. (**G**) Cryo-EM structure of RC_I_2 highlighting monomer packing and interfaces along each symmetry axis. (**H**) Overlay and RMSD calculations for RC_I_2 compared with the design model for each symmetry interface (cryoEM is shown in pink; design is shown in gray). (**I**) Overlay and RMSD calculation for a single monomer of RC_I_2. (**J**) Overlay and RMSD calculation for the entire RC_I_2 capsid.

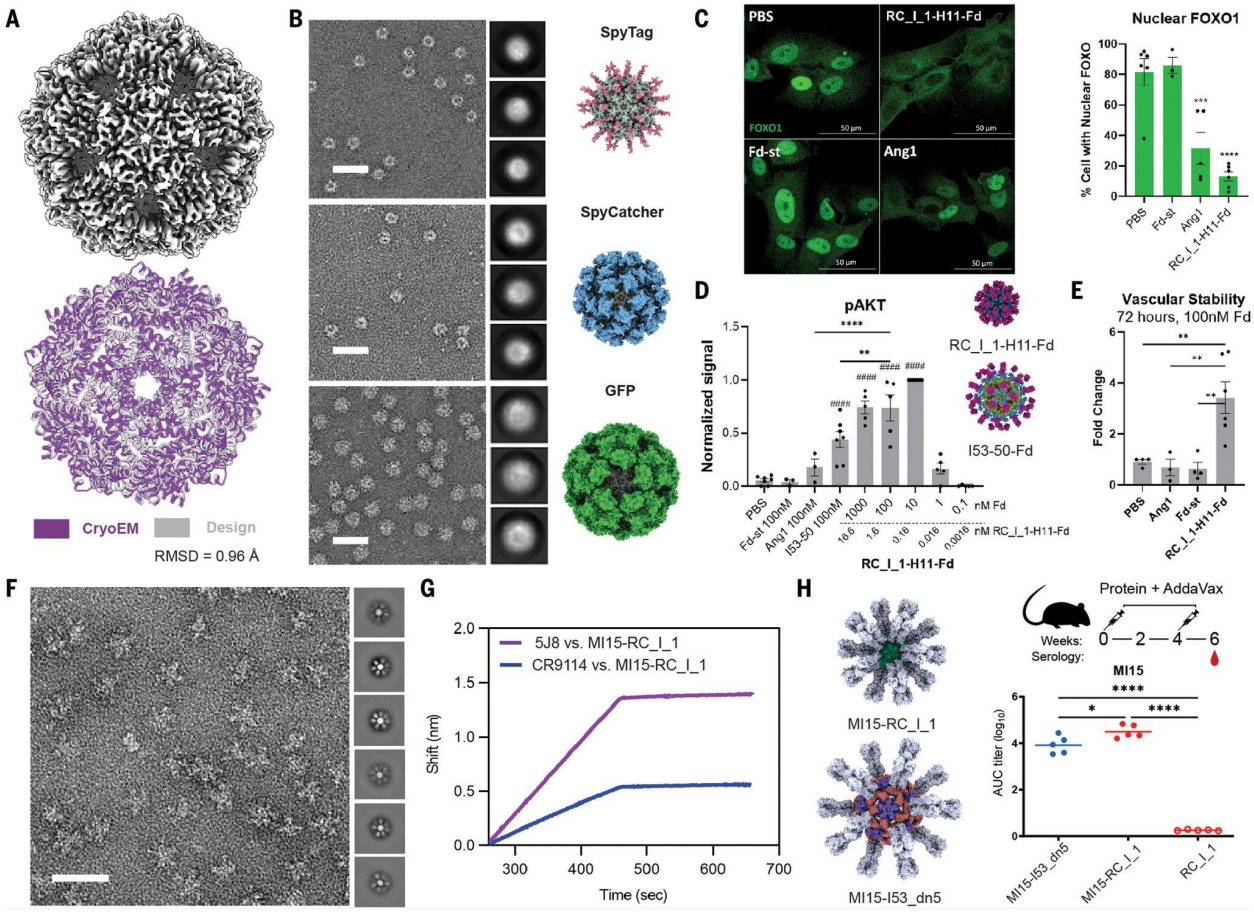

Fig. 5. Applications of designed capsids.

(**A**) Robustness of RC_I_1 capsid to sequence redesign using ProteinMPNN. The 3-Å-resolution cryo-EM reconstruction of RC_I_1-H11 (top) reveals a close agreement between the experimental structure (purple) and the design model (gray) with a RMSD of 0.96 Å. The RC_I_1-H11 structure is nearly identical to RC_I_1 despite considerable sequence differences [bottom; residue differences are highlighted in red (RC_I_1-H11) and teal (RC_I_1)]. (**B**) From top to bottom, models and representative nsEM images of spyTag-, spyCatcher-, and GFP-fused (to N terminus) RC_I_1-H11 with 2D class averages. Scale bar, 50 nm. (**C** and **D**) RC_I_1-H11-Fd activates Tie2 downstream Akt phosphorylation and FOXO1 translocation. Serum-starved HUVECs were treated with serially diluted RC_I_1-H11-Fd (1000-0.1 nM), Fd-st (100 nM), Ang1 (100 nM), I53-50 (100 nM), or phosphate-buffered saline (PBS) control for 15 min before protein lysate collection for Western blot analysis, or cells were fixed for FOXO1 antibody stain. (C) Left, representative confocal images of HUVECs immunofluorescence stained with FOXO1 antibody. Right, quantification showing the percentage of cells with nuclear FOXO1; 100 cells were counted in each biological replicate. Levels of significance were compared with PBS control in the FOXO1 graph. (D) Quantification of Western blot showing pAKT signal normalized to RC_I_1-H11-Fd at 10 nM. RC_I_1-H11-Fd induces a significantly higher signal than the previously characterized I53-50-Fd (inset) at 100 nM Fd equivalent. (**E**) Quantification of vascular stability by averaging the number of nodes, meshes, and tubes calculated at the 72-hour time point using the

Angiogenesis Analyzer plug-in in ImageJ (fig. S28D). In (C) to (E), *P* values were calculated using one-way ANOVA with Bonferroni's multiple-comparisons test in Prism for comparing groups of two or more; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; ****$P < 0.0001$; significance over PBS control is noted as # in (D). (**F**) Representative nsEM micrograph and 2D class averages (inset) of mammalian cell secreted RC_I_1 particle flexibly fused with M15 influenza HA (MI15-RC_I_1). Scale bar, 50 nm. (**G**) The RC_I_1 displayed HA is antigenically intact, reacting with both head (5J8) and stem (CR9114) anti-HA antibodies in biolayer interferometry experiments. (**H**) Models of RC_I_1 (top) and I53_dn5 (bottom) displaying MI15 influenza HA (left); the presentation is considerably denser in the former. Top right: Mouse immunization schedule. Bottom right: HA-specific antibody titers in immune sera. Statistical significance was determined using one-way ANOVA with Tukey's multiple-comparisons test; *$P < 0.05$; ****$P < 0.0001$. The RC_I_1 display format produces a higher antibody titer than the I53_dn5 nanoparticle currently in clinical trials.

References

1. R. Zandi, D. Reguera, R. F. Bruinsma, W. M. Gelbart, J. Rudnick, Origin of icosahedral symmetry in viruses. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 15556–15560 (2004). doi:10.1073/pnas.0405844101
2. T. Douglas, M. Young, Viruses: Making friends with old foes. *Science* **312**, 873–875 (2006). doi:10.1126/science.1123223
3. J. E. Padilla, C. Colovos, T. O. Yeates, Nanohedra: Using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 2217–2221 (2001). doi:10.1073/pnas.041614998
4. 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
5. J. Zhu, N. Avakyan, A. Kakkis, A. M. Hoffnagle, K. Han, Y. Li, Z. Zhang, T. S. Choi, Y. Na, C.-J. Yu, F. A. Tezcan, Protein assembly by design. *Chem. Rev.* **121**, 13701–13796 (2021). doi:10.1021/acs.chemrev.1c00308
6. 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, D. Baker, Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022). doi:10.1126/science.add1964
7. N. P. King, W. Sheffler, M. R. Sawaya, B. S. Vollmar, J. P. Sumida, I. André, T. Gonen, T. O. Yeates, D. Baker, Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **336**, 1171–1174 (2012). doi:10.1126/science.1219364
8. Y.-T. Lai, D. Cascio, T. O. Yeates, Structure of a 16-nm cage designed by using protein oligomers. *Science* **336**, 1129 (2012). doi:10.1126/science.1219351
9. Y.-T. Lai, E. Reading, G. L. Hura, K.-L. Tsai, A. Laganowsky, F. J. Asturias, J. A. Tainer, C. V. Robinson, T. O. Yeates, Structure of a designed protein cage that self-assembles into a highly porous cube. *Nat. Chem*. **6**, 1065–1071 (2014). doi:10.1038/nchem.2107
10. N. P. King, J. B. Bale, W. Sheffler, D. E. McNamara, S. Gonen, T. Gonen, T. O. Yeates, D. Baker, Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510**, 103–108 (2014). doi:10.1038/nature13404
11. Y. Hsia, J. B. Bale, S. Gonen, D. Shi, W. Sheffler, K. K. Fong, U. Nattermann, C. Xu, P.-S. Huang, R. Ravichandran, S. Yi, T. N. Davis, T. Gonen, N. P. King, D. Baker, Corrigendum: Design of a hyperstable 60-subunit protein icosahedron. *Nature* **540**, 150 (2016). doi:10.1038/nature18010
12. 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
13. 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

14. R. Divine, H. V. Dang, G. Ueda, J. A. Fallas, I. Vulovic, W. Sheffler, S. Saini, Y. T. Zhao, I. X. Raj, P. A. Morawski, M. F. Jennewein, L. J. Homad, Y.-H. Wan, M. R. Tooley, F. Seeger, A. Etemadi, M. L. Fahning, J. Lazarovits, A. Roederer, A. C. Walls, L. Stewart, M. Mazloomi, N. P. King, D. J. Campbell, A. T. McGuire, L. Stamatatos, H. Ruohola-Baker, J. Mathieu, D. Veesler, D. Baker, Designed proteins assemble antibodies into modular nanocages. *Science* **372**, eabd9994 (2021). doi:10.1126/science.abd9994
15. C. J. Lanci, C. M. MacDermaid, S. G. Kang, R. Acharya, B. North, X. Yang, X. J. Qiu, W. F. DeGrado, J. G. Saven, Computational design of a protein crystal. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 7304–7309 (2012). doi:10.1073/pnas.1112595109
16. J. C. Sinclair, K. M. Davies, C. Vénien-Bryan, M. E. M. Noble, Generation of protein lattices by fusing proteins with matching rotational symmetry. *Nat. Nanotechnol.* **6**, 558–562 (2011). doi:10.1038/nnano.2011.122
17. S. Gonen, F. DiMaio, T. Gonen, D. Baker, Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. *Science* **348**, 1365–1368 (2015). doi:10.1126/science.aaa9897
18. A. J. Ben-Sasson, J. L. Watson, W. Sheffler, M. C. Johnson, A. Bittleston, L. Somasundaram, J. Decarreau, F. Jiao, J. Chen, I. Mela, A. A. Drabek, S. M. Jarrett, S. C. Blacklow, C. F. Kaminski, G. L. Hura, J. J. De Yoreo, J. M. Kollman, H. Ruohola-Baker, E. Derivery, D. Baker, Design of biologically active binary protein 2D materials. *Nature* **589**, 468–473 (2021). doi:10.1038/s41586-020-03120-8
19. Z. Li, S. Wang, U. Nattermann, A. K. Bera, A. J. Borst, M. J. Bick, E. Yang, W. Sheffler, B. Lee, H. Nguyen, A. Kang, R. Dalal, J. Lubner, Y. Hsia, H. Haddox, A. Courbet, Q. Dowling, A. Favor, A. Etemadi, N. I. Edman, W. Yang, B. Sankaran, B. Negahdari, D. Baker, Computational design of de novo 3D protein crystals. bioRxiv 2022.11.18.517014 [Preprint] (2022); https://doi.org/10.1101/2022.11.18.517014.
20. A. C. Walls, B. Fiala, A. Schäfer, S. Wrenn, M. N. Pham, M. Murphy, L. V. Tse, L. Shehata, M. A. O'Connor, C. Chen, M. J. Navarro, M. C. Miranda, D. Pettie, R. Ravichandran, J. C. Kraft, C. Ogohara, A. Palser, S. Chalk, E.-C. Lee, K. Guerriero, E. Kepl, C. M. Chow, C. Sydeman, E. A. Hodge, B. Brown, J. T. Fuller, K. H. Dinnon 3rd, L. E. Gralinski, S. R. Leist, K. L. Gully, T. B. Lewis, M. Guttman, H. Y. Chu, K. K. Lee, D. H. Fuller, R. S. Baric, P. Kellam, L. Carter, M. Pepper, T. P. Sheahan, D. Veesler, N. P. King, Elicitation of potent neutralizing antibody responses by designed protein nanoparticle vaccines for SARS-CoV-2. *Cell* **183**, 1367–1382.e17 (2020). doi:10.1016/j.cell.2020.10.043
21. P. S. Arunachalam, A. C. Walls, N. Golden, C. Atyeo, S. Fischinger, C. Li, P. Aye, M. J. Navarro, L. Lai, V. V. Edara, K. Röltgen, K. Rogers, L. Shirreff, D. E. Ferrell, S. Wrenn,

D. Pettie, J. C. Kraft, M. C. Miranda, E. Kepl, C. Sydeman, N. Brunette, M. Murphy, B. Fiala, L. Carter, A. G. White, M. Trisal, C.-L. Hsieh, K. Russell-Lodrigue, C. Monjure, J. Dufour, S. Spencer, L. Doyle-Meyers, R. P. Bohm, N. J. Maness, C. Roy, J. A. Plante, K. S. Plante, A. Zhu, M. J. Gorman, S. Shin, X. Shen, J. Fontenot, S. Gupta, D. T. O'Hagan, R. Van Der Most, R. Rappuoli, R. L. Coffman, D. Novack, J. S. McLellan, S. Subramaniam, D. Montefiori, S. D. Boyd, J. L. Flynn, G. Alter, F. Villinger, H. Kleanthous, J. Rappaport, M. S. Suthar, N. P. King, D. Veesler, B. Pulendran, Adjuvanting a subunit COVID-19 vaccine to induce protective immunity. *Nature* **594**, 253–258 (2021). doi:10.1038/s41586-021-03530-2

22. P. S. Arunachalam, Y. Feng, U. Ashraf, M. Hu, V. V. Edara, V. I. Zarnitsyna, P. P. Aye, N. Golden, K. W. M. Green, B. M. Threeton, N. J. Maness, B. J. Beddingfield, R. P. Bohm, J. Dufour, K. Russell-Lodrigue, M. C. Miranda, A. C. Walls, K. Rogers, L. Shirreff, D. E. Ferrell, N. R. Deb Adhikary, J. Fontenot, A. Grifoni, A. Sette, D. T. O'Hagan, R. Van Der Most, R. Rappuoli, F. Villinger, H. Kleanthous, J. Rappaport, M. S. Suthar, D. Veesler, T. T. Wang, N. P. King, B. Pulendran, Durable protection against SARS-CoV-2 Omicron induced by an adjuvanted subunit vaccine. bioRxiv [Preprint] (2022); https://doi.org/10.1101/2022.03.18.484950.
23. J. Y. Song, W. S. Choi, J. Y. Heo, J. S. Lee, D. S. Jung, S.-W. Kim, K.-H. Park, J. S. Eom, S. J. Jeong, J. Lee, K. T. Kwon, H. J. Choi, J. W. Sohn, Y. K. Kim, J. Y. Noh, W. J. Kim, F. Roman, M. A. Ceregido, F. Solmi, A. Philippot, A. C. Walls, L. Carter, D. Veesler, N. P. King, H. Kim, J. H. Ryu, S. J. Lee, Y. W. Park, H. K. Park, H. J. Cheong, Safety and immunogenicity of a SARS-CoV-2 recombinant protein nanoparticle vaccine (GBP510) adjuvanted with AS03: A randomised, placebo-controlled, observer-blinded phase 1/2 trial. *EClinicalMedicine* **51**, 101569 (2022). doi:10.1016/j.eclinm.2022.101569
24. K. T. Simons, R. Bonneau, I. Ruczinski, D. Baker, Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **37**, 171–176 (1999). doi:10.1002/(SICI)1097-0134(1999)37:3+<171:AID-PROT21>3.0.CO;2-Z
25. D. Gront, D. W. Kulp, R. M. Vernon, C. E. M. Strauss, D. Baker, Generalized fragment picking in Rosetta: Design, protocols and applications. *PLOS ONE* **6**, e23294 (2011). doi:10.1371/journal.pone.0023294
26. E. Verschueren, P. Vanhee, A. M. van der Sloot, L. Serrano, F. Rousseau, J. Schymkowitz, Protein design with fragment databases. *Curr. Opin. Struct. Biol.* **21**, 452–459 (2011). doi:10.1016/j.sbi.2011.05.002
27. N. Koga, R. Tatsumi-Koga, G. Liu, R. Xiao, T. B. Acton, G. T. Montelione, D. Baker, Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012). doi:10.1038/nature11600
28. R. Pearce, X. Huang, G. S. Omenn, Y. Zhang, De novo protein fold design through sequence-independent fragment assembly simulations. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2208275120 (2023). doi:10.1073/pnas.2208275120
29. T. M. Jacobs, B. Williams, T. Williams, X. Xu, A. Eletsky, J. F. Federizon, T. Szyperski,

B. Kuhlman, Design of structurally distinct proteins using strategies inspired by evolution. *Science* **352**, 687–690 (2016). doi:10.1126/science.aad8036
30. B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. A. Sallab, S. Yogamani, P. Perez, Deep reinforcement learning for autonomous driving: A survey. *IEEE Trans. Intell. Transp. Syst.* **23**, 4909–4926 (2022). doi:10.1109/TITS.2021.3054625
31. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015). doi:10.1038/nature14236
32. D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016). doi:10.1038/nature16961
33. A. Fawzi, M. Balog, A. Huang, T. Hubert, B. Romera-Paredes, M. Barekatain, A. Novikov, F. J. R Ruiz, J. Schrittwieser, G. Swirszcz, D. Silver, D. Hassabis, P. Kohli, Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature* **610**, 47–53 (2022). doi:10.1038/s41586-022-05172-4
34. R. Coulom, "Efficient selectivity and backup operators in Monte-Carlo tree search," in *Computers and Games*, H. J. van den Herik, P. Ciancarini, H. H. L. M. Donkers, Eds. (Springer, 2007), vol. 4630 of *Lecture Notes in Computer Science*, pp. 72–83; http://link.springer.com/10.1007/978-3-540-75538-8_7.
35. L. Kocsis, C. Szepesvári, "Bandit based Monte-Carlo planning," in *Machine Learning: ECML* 2006, J. Fürnkranz, T. Scheffer, M. Spiliopoulou, Eds. (Springer, 2006), vol. 4212 of *Lecture Notes in Computer Science*, pp. 282–293; http://link.springer.com/10.1007/11871842_29.
36. J. P. Hallinan, L. A. Doyle, B. W. Shen, M. M. Gewe, B. Takushi, M. A. Kennedy, D. Friend, J. M. Roberts, P. Bradley, B. L. Stoddard, Design of functionalised circular tandem repeat proteins with longer repeat topologies and enhanced subunit contact surfaces. *Commun. Biol.* **4**, 1240 (2021). doi:10.1038/s42003-021-02766-y
37. 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* **378**, 49–56 (2022). doi:10.1126/science.add2187
38. L. Cao, B. Coventry, I. Goreshnik, B. Huang, W. Sheffler, J. S. Park, K. M. Jude, I. Marković, R. U. Kadam, K. H. G. Verschueren, K. Verstraete, S. T. R. Walsh, N. Bennett, A. Phal, A. Yang, L. Kozodoy, M. DeWitt, L. Picton, L. Miller, E.-M. Strauch, N. D. DeBouver, A. Pires, A. K. Bera, S. Halabiya, B. Hammerson, W. Yang, S. Bernard, L.

Stewart, I. A. Wilson, H. Ruohola-Baker, J. Schlessinger, S. Lee, S. N. Savvides, K. C. Garcia, D. Baker, Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022). doi:10.1038/s41586-022-04654-9

39. 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
40. L. G. Nivón, R. Moretti, D. Baker, A Pareto-optimal refinement method for protein design scaffolds. *PLOS ONE* **8**, e59004 (2013). doi:10.1371/journal.pone.0059004
41. P. Conway, M. D. Tyka, F. DiMaio, D. E. Konerding, D. Baker, Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23**, 47–55 (2014). doi:10.1002/pro.2389
42. B. Zakeri, J. O. Fierer, E. Celik, E. C. Chittock, U. Schwarz-Linek, V. T. Moy, M. Howarth, Peptide tag forming a rapid covalent bond to a protein, through engineering a bacterial adhesin. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E690–E697 (2012). doi:10.1073/pnas.1115485109
43. Y. T. Zhao, J. A. Fallas, S. Saini, G. Ueda, L. Somasundaram, Z. Zhou, I. Xavier Raj, C. Xu, L. Carter, S. Wrenn, J. Mathieu, D. L. Sellers, D. Baker, H. Ruohola-Baker, F-domain valency determines outcome of signaling through the angiopoietin pathway. *EMBO Rep.* **22**, e53471 (2021). doi:10.15252/embr.202153471
44. S. Han, S.-J. Lee, K. E. Kim, H. S. Lee, N. Oh, I. Park, E. Ko, S. J. Oh, Y.-S. Lee, D. Kim, S. Lee, D. H. Lee, K.-H. Lee, S. Y. Chae, J.-H. Lee, S.-J. Kim, H.-C. Kim, S. Kim, S. H. Kim, C. Kim, Y. Nakaoka, Y. He, H. G. Augustin, J. Hu, P. H. Song, Y.-I. Kim, P. Kim, I. Kim, G. Y. Koh, Amelioration of sepsis by TIE2 activation-induced vascular protection. *Sci. Transl. Med.* **8**, 335ra55 (2016). doi:10.1126/scitranslmed.aad9260
45. I. Kim, H. G. Kim, J.-N. So, J. H. Kim, H. J. Kwak, G. Y. Koh, Angiopoietin-1 regulates endothelial cell survival through the phosphatidylinositol 3′-Kinase/Akt signal transduction pathway. *Circ. Res.* **86**, 24–29 (2000). doi:10.1161/01.RES.86.1.24
46. M. Kim, B. Allen, E. A. Korhonen, M. Nitschké, H. W. Yang, P. Baluk, P. Saharinen, K. Alitalo, C. Daly, G. Thurston, D. M. McDonald, Opposing actions of angiopoietin-2 on Tie2 signaling and FOXO1 activation. *J. Clin. Invest.* **126**, 3511–3525 (2016). doi:10.1172/JCI84871
47. C. Daly, V. Wong, E. Burova, Y. Wei, S. Zabski, J. Griffiths, K.-M. Lai, H. C. Lin, E. Ioffe, G. D. Yancopoulos, J. S. Rudge, Angiopoietin-1 modulates endothelial cell function and gene expression via the transcription factor FKHR (FOXO1). *Genes Dev.* **18**, 1060–1071 (2004). doi:10.1101/gad.1189704
48. K. L. DeCicco-Skinner, G. H. Henry, C. Cataisson, T. Tabib, J. C. Gwilliam, N. J.

Watson, E. M. Bullwinkle, L. Falkenburg, R. C. O'Neill, A. Morin, J. S. Wiest, Endothelial cell tube formation assay for the in vitro study of angiogenesis. *J. Vis. Exp.* **91**, e51312 (2014).
49. N. P. J. Brindle, P. Saharinen, K. Alitalo, Signaling and functions of angiopoietin-1 in vascular protection. *Circ. Res.* **98**, 1014–1023 (2006). doi:10.1161/01.RES.0000218275.54089.12
50. J. C. Krause, T. Tsibane, T. M. Tumpey, C. J. Huffman, C. F. Basler, J. E. Crowe Jr., A broadly neutralizing human monoclonal antibody that recognizes a conserved, novel epitope on the globular head of the influenza H1N1 virus hemagglutinin. *J. Virol.* **85**, 10905–10908 (2011). doi:10.1128/JVI.00700-11
51. C. Dreyfus, N. S. Laursen, T. Kwaks, D. Zuijdgeest, R. Khayat, D. C. Ekiert, J. H. Lee, Z. Metlagel, M. V. Bujny, M. Jongeneelen, R. van der Vlugt, M. Lamrani, H. J. W. M. Korse, E. Geelen, Ö. Sahin, M. Sieuwerts, J. P. J. Brakenhoff, R. Vogels, O. T. W. Li, L. L. M. Poon, M. Peiris, W. Koudstaal, A. B. Ward, I. A. Wilson, J. Goudsmit, R. H. E. Friesen, Highly conserved protective epitopes on influenza B viruses. *Science* **337**, 1343–1348 (2012). doi:10.1126/science.1222908
52. G. Ueda, A. Antanasijevic, J. A. Fallas, W. Sheffler, J. Copps, D. Ellis, G. B. Hutchinson, A. Moyer, A. Yasmeen, Y. Tsybovsky, Y.-J. Park, M. J. Bick, B. Sankaran, R. A. Gillespie, P. J. Brouwer, P. H. Zwart, D. Veesler, M. Kanekiyo, B. S. Graham, R. W. Sanders, J. P. Moore, P. J. Klasse, A. B. Ward, N. P. King, D. Baker, Tailored design of protein nanoparticle scaffolds for multivalent presentation of viral glycoprotein antigens. *eLife* **9**, e57659 (2020). doi:10.7554/eLife.57659
53. S. Boyoglu-Barnum, D. Ellis, R. A. Gillespie, G. B. Hutchinson, Y.-J. Park, S. M. Moin, O. J. Acton, R. Ravichandran, M. Murphy, D. Pettie, N. Matheson, L. Carter, A. Creanga, M. J. Watson, S. Kephart, S. Ataca, J. R. Vaile, G. Ueda, M. C. Crank, L. Stewart, K. K. Lee, M. Guttman, D. Baker, J. R. Mascola, D. Veesler, B. S. Graham, N. P. King, M. Kanekiyo, Quadrivalent influenza nanoparticle vaccines induce broad protection. *Nature* **592**, 623–628 (2021). doi:10.1038/s41586-021-03365-x
54. D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, D. Hassabis, Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017). doi:10.1038/nature24270
55. I. Lutz, Protein backbone MCTS, Zenodo, (2023); https://doi.org/10.5281/zenodo.7709840.
56. S. Wang, C. Norn, I. Lutz, 2023 RL capsid design, Zenodo (2023); https://doi.org/10.5281/zenodo.7758067.
57. S. J. Fleishman, A. Leaver-Fay, J. E. Corn, E.-M. Strauch, S. D. Khare, N. Koga, J. Ashworth, P. Murphy, F. Richter, G. Lemmon, J. Meiler, D. Baker, RosettaScripts: A scripting language interface to the Rosetta macromolecular modeling suite. *PLOS ONE* **6**, e20161 (2011). doi:10.1371/journal.pone.0020161

58. T. J. Brunette, M. J. Bick, J. M. Hansen, C. M. Chow, J. M. Kollman, D. Baker, Modular repeat protein sculpting using rigid helical junctions. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 8870–8875 (2020). doi:10.1073/pnas.1908768117
59. B. Coventry, "npose v1.0," Github (2021); https://github.com/bcov77/npose.
60. Q. Zhou, "PyMesh: Geometry processing library for Python" (2018); https://pymesh.readthedocs.io/.
61. G. J. Rocklin, T. M. Chidyausiku, I. Goreshnik, A. Ford, S. Houliston, A. Lemak, L. Carter, R. Ravichandran, V. K. Mulligan, A. Chevalier, C. H. Arrowsmith, D. Baker, Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017). doi:10.1126/science.aan0693
62. J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, D. Baker, Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 1496–1503 (2020). doi:10.1073/pnas.1914677117
63. 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
64. Y. Song, F. DiMaio, R. Y.-R. Wang, D. Kim, C. Miles, T. Brunette, J. Thompson, D. Baker, High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735–1742 (2013). doi:10.1016/j.str.2013.08.005
65. 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
66. C. Suloway, J. Pulokas, D. Fellmann, A. Cheng, F. Guerra, J. Quispe, S. Stagg, C. S. Potter, B. Carragher, Automated molecular microscopy: The new Leginon system. *J. Struct. Biol.* **151**, 41–60 (2005). doi:10.1016/j.jsb.2005.03.010
67. M. Sun, C. M. Azumaya, E. Tse, D. P. Bulkley, M. B. Harrington, G. Gilbert, A. Frost, D. Southworth, K. A. Verba, Y. Cheng, D. A. Agard, Practical considerations for using K3 cameras in CDS mode for high-resolution and high-throughput single particle cryo-EM. *J. Struct. Biol.* **213**, 107745 (2021). doi:10.1016/j.jsb.2021.107745
68. P. Emsley, K. Cowtan, Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004). doi:10.1107/S0907444904019158
69. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010). doi:10.1107/S0907444910007493
70. R. Y.-R. Wang, Y. Song, B. A. Barad, Y. Cheng, J. S. Fraser, F. DiMaio, Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. *eLife* **5**, e17219 (2016). doi:10.7554/eLife.17219
71. V. B. Chen, W. B. Arendall 3rd, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, D. C. Richardson, MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010). doi:10.1107/S0907444909042073

72. B. A. Barad, N. Echols, R. Y.-R. Wang, Y. Cheng, F. DiMaio, P. D. Adams, J. S. Fraser, EMRinger: Side chain-directed model and map validation for 3D cryo-electron microscopy. *Nat. Methods* **12**, 943–946 (2015). doi:10.1038/nmeth.3541
73. E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004). doi:10.1002/jcc.20084
74. E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, T. E. Ferrin, UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021). doi:10.1002/pro.3943
75. J. Wang, S. Lisanza, D. Juergens, D. Tischer, J. L. Watson, K. M. Castro, R. Ragotte, A. Saragovi, L. F. Milles, M. Baek, I. Anishchenko, W. Yang, D. R. Hicks, M. Expòsit, T. Schlichthaerle, J. H. Chun, J. Dauparas, N. Bennett, B. I. M. Wicky, A. Muenks, F. DiMaio, B. Correia, S. Ovchinnikov, D. Baker, Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022). doi:10.1126/science.abn2100
76. J. R. Whittle, A. K. Wheatley, L. Wu, D. Lingwood, M. Kanekiyo, S. S. Ma, S. R. Narpala, H. M. Yassine, G. M. Frank, J. W. Yewdell, J. E. Ledgerwood, C. J. Wei, A. B. McDermott, B. S. Graham, R. A. Koup, G. J. Nabel, Flow cytometry reveals that H5N1 vaccination elicits cross-reactive stem-directed antibodies from multiple Ig heavy-chain lineages. *J. Virol.* **88**, 4047–4057 (2014). doi:10.1128/JVI.03422-13
77. J. Votteler, C. Ogohara, S. Yi, Y. Hsia, U. Nattermann, D. M. Belnap, N. P. King, W. I. Sundquist, Designed proteins induce the formation of nanocage-containing extracellular vesicles. *Nature* **540**, 292–295 (2016). doi:10.1038/nature20607
78. D. Feldman, A. Singh, J. L. Schmid-Burgk, R. J. Carlson, A. Mezger, A. J. Garrity, F. Zhang, P. C. Blainey, Optical pooled screens in human cells. *Cell* **179**, 787–799.e17 (2019). doi:10.1016/j.cell.2019.09.016

Methods

Monte Carlo tree search backbone sampling

The backbone generation method described here uses Monte Carlo tree search (MCTS), an RL algorithm, to choose secondary structure fragments of helices and loops in a decision tree to append to either terminus of a growing protein backbone. With reinforcement based on provided geometric constraints and score functions, iterations through the tree lead to more and more optimal backbone designs as the tree search is guided to better paths.

Helix and loop elements

For this method, fragment assembly through the decision tree uses two types of backbone secondary structure elements to select as choices: helices and loops. Helix choices are selected by number of residues of a parametrically-generated straight alpha-helix. This simplification severely limits the size of the structural search space and corresponding decision tree, allowing for more efficient exploration of the space. For example, allowing helix additions in the range of 9-22 residues in length results in only 14 possible choices of helices. Furthermore, simple parametrically-generated alpha helices have been used successfully to solve a wide variety of protein design problems, and so we hypothesized that this simplified sampling strategy would work adequately for many types of problems while avoiding complications arising from more complicated structural elements such as beta-sheets.

Similarly, we sought to use a library of short loop structures for loop choices to increase the likelihood of sampling successful designs. We used a library of approximately 26 thousand loops of 3-5 residues in length extracted from a set of de novo 4 helix bundles designed using the Rosetta BluePrintBDR (*56*) and filtered by loop RMSD to the PDB (*57*). We chose to bin the loops by structural similarity in order to reduce the number of choices in the decision tree and allow for alternate loop selection as a means to provide slight structural deviation and diversity. Loop choices in the decision tree are therefore the selection of a loop bin followed by the random selection of a loop from within that bin. This strategy required clustering based on parameters describing the helix following loop addition to enable structural similarity in subsequent choices using different loops within the same bin. We used *k*-means clustering on a set of 7 parameters describing the helix following each loop: 3 for helix translation, 3 for helix direction vector, and 1 for helix phase. We wanted a limited number of bins with a minimum number of loops per bin and a limited range in the distribution of loops per bin in order to avoid sampling biases resulting from over or under populated bins. The loop library was binned into 316 bins representing 316 possible loop choices in the decision tree, where each bin contains between 5 and 392 loops, with 81 loops in a bin on average.

Initialization and decision tree choices

To build backbones using this decision tree, a starting point is required for initialization. This starting point is fixed for a given tree, as backbones are built in place so a sequence of choices to sample a given backbone may only be valid in the context of a single starting point. This starting point can be an alpha-helical 'stub' placed in space with random translation and rotation, or a pre-existing structure off of which to build. Helix and loop choices from the decision tree are appended to the starting point to build the backbone using the npose package (*58*) to align the chosen element to the growing structure. The loops in the loop library have additional short 4-residue helices on either end to allow for accurate alignment to helices. Helix and loop choices may be appended to only a single terminus (N or C) or to both termini as specified by the user. For example, when building off of a randomly placed stub (as used in all cases in this work), the first set of choices in the decision tree will be 14 options for helix lengths between 9 and 22 residues extended off of the C terminus. The following choices in the next level of the tree consist of 732 options, or 316 loop bins each for loop additions off of either terminus. After a loop is chosen, the only allowed options are helix additions from the same terminus, as each loop must be followed by a helix.

MCTS expansion and evaluation

Random paths through this decision tree of helix and loop choices may lead to viable backbones, but it is computationally intractable to exhaustively explore all paths through the tree and most paths will result in either poor quality backbones that are suboptimal or invalid backbones that violate geometric constraints. With MCTS, paths are initially chosen with uniform random probability, but each iteration informs the choice of subsequent paths to guide the sampling towards better backbones. Upon initialization, all possible choices at each new branch point in the decision tree are assigned the same probability weight. Every choice is made at random, but weights may be adjusted with every iteration to increase their likelihood of selection. These probability weights are changed for each sampled backbone based on geometric constraints that are assessed at every choice as well as score functions that are only assessed for the final structure. Both geometric constraints and score functions must be rapidly calculated to allow for efficient sampling through thousands of iterations of the tree search. As score functions are typically slower to assess and often require a full structure for accuracy, they are only calculated on final structures that pass a specified minimum sequence length and minimum number of helices.

Geometric constraints

Geometric constraints are rapidly assessed for each helix and loop choice so that only valid choices are accepted. The first type of geometric constraint is the clash constraint, used for all backbone sampling with this method. Distances are calculated between all new backbone atoms for the potential helix or loop option and all other specified atoms. These specified atoms always include the growing backbone that is being sampled through the tree search, but may also include atoms of other prespecified structures to avoid in space. These other atoms can also include neighboring subunits when sampling symmetric assemblies, as was used for sampling the icosahedral backbones. With every new potential choice, transformation matrices are used to calculate the atoms of all nearby subunits with this additional choice (rather than every subunit to speed up computation), so that symmetric assemblies can be built in place while ensuring no clashes. For all clash checks, all atom distances must be above a specified clash threshold of 2.85 Å.

Simple space bounds may also be provided as geometric constraints, for example limits in Cartesian space where new atoms may not be placed. The diameter of the growing backbone can also be geometrically constrained by calculating distances between all atoms of the growing structure so that no new choice is allowed where any distance exceeds a specified diameter threshold. More complicated volume constraints can be provided through the use of *.obj* files specifying a mesh surface. A grid of points 1Å apart encompassing the entire volume is queried using the pymesh package (*59*) to find a winding number for each point. This number specifies whether the point falls within or outside of the volume. All points falling within the volume (winding number >= 1) are added to a hash table specifying all voxels of that volume. All atoms for every potential helix or loop choice are converted to their corresponding voxels, and these voxels are rapidly checked for their presence in the hash table such that only choices where all new atoms are within the specified volume are allowed.

As geometric constraints are assessed for every choice in the tree, only valid choices are allowed. With each valid choice selected, the probability of that choice is upweighted by a fixed value (set to 100), as are the probabilities of every prior choice leading back up the path in the tree. This way, a valid choice is rewarded not only for its initial selection, but also for subsequent valid choices that it enables.

Termination cases

Each iteration through the search tree will end at one of a number of termination cases. As only choices passing all geometric constraints are permitted, termination may occur after reaching a set limit for the number of invalid choice attempts at a given branch point. This trial limit serves to constrain the amount of time spent on an iteration, as many paths will lead to points in the tree

where few or no possible valid choices exist. The trial limit is 4 at the start of the tree, and increases by an additional value of 3 times the number of helices progressing down the tree, as later branch points will typically have fewer valid choice options and thus require additional attempts to find a valid choice if one exists. Furthermore, it is advantageous to spend compute time on iterations where part of a valid backbone is built already over iterations reaching early dead-end paths. Iterations may also terminate upon reaching preset limits on sequence length and number of helices, which are checked following each choice addition. Upon termination in instances where the last addition was a loop, the loop is removed to avoid lower quality backbones with dangling loops lacking a succeeding helix. Lastly, termination may occur upon picking an additional 'no choice' option. This option is introduced as an additional choice at loop branch points after passing a prespecified number of helices, and provides a means of terminating backbones that may already be of good quality that would be worsened with the addition of further choices. The probability for the 'no choice' termination is initialized at 0.1 when the minimum number of helices is first reached, then is initialized with an added probability of 0.05 for each helix added past the minimum, as it will typically become a more optimal option. Upon all cases of termination, backbones that exceed the minimum number of helices and an additional preset minimum sequence length are evaluated using specified score functions, and the tree path probability weights corresponding to the backbone are updated.

Score functions

Score functions are only evaluated at the end of an iteration on backbones that exceed a minimum number of helices and minimum sequence length, as score functions may be computationally costly and typically only make sense to evaluate on a completed structure that resembles desired backbones. We developed score functions with speed and generalization in mind, to enable rapid sampling and a diversity of solutions.

In this work, we describe three score functions assessing monomer quality. These functions attempt to reward backbones that could potentially be designed into stable monomers. The core formation score uses a cone in front of each residue's C alpha - C beta vector to calculate the percent of the backbone to classify as 'core' by sidechain neighbors (*60*). The helix packing score uses the same sidechain neighbors calculation to quantify a worst helix, which has the minimum average sidechain neighbors. This helix may be penalized, as if it dangles off of the structure it could lead to an unstable monomer. Lastly, the monomer designability score, repurposed from RPXdock, calculates transforms for all residue pairs and uses them to check a hash table for pair 'hits' suggesting sidechain designability (*4*). The score function returns the percent of 9mers with hash table hits in the core or core boundary region (classified by sidechain neighbors), indicating the potential designability of the monomer backbone.

We developed two additional score functions for assessing icosahedral assembly quality. These functions attempt to reward assemblies with closed surfaces and good interfaces between subunits. The porosity score rapidly approximates how porous the surface of the icosahedral assembly is by calculating the maximum volume filled across every spherical shell bounded by two atoms. The set of shells is defined by the radial distances of all pairs of atoms in a single subunit, plus an additional carbon Van der Waal (VDW) radius (1.7 Å) for the outer atom and minus the same 1.7 Å for the inner atom. For each of these shells, the approximate volume filled is the sum of all atom sphere volumes divided by the volume of the shell, using the carbon VDW radius once more to calculate all atom sphere volumes. The maximum fraction filled for the set of shells approximates the porosity of the icosahedral assembly. The porosity score also indirectly enriches for assemblies with better interfaces, as surface closure requires contact between neighboring subunits. The interface designability score further assesses interface quality by using the same atom pair transform hash table as the monomer designability score. The function checks atom pair transforms from one subunit to all neighboring subunits, and returns the total number of hash table hits only if there are hits present across two or more interfaces, thus evaluating the presence and designability of multiple subunit interfaces.

After evaluation, all scores are combined into a single value to reward backbones that are optimal across many or all provided score functions. This is done by passing each score through a sigmoid activation function, with parameters specifying the desired score values and weight relative to other scores. The product of these sigmoid-activated scores is passed through a final sigmoid activation provided it exceeds a minimum threshold, resulting in a final overall score between 0 and 1. This overall score is multiplied by a fixed scalar (set to 5,000) then used to upweight the probability weights of each choice down the decision tree path leading to the final backbone.

Upweighting

Probabilities for the path of decision tree choices are upweighted at every step for satisfying geometric constraints and upon termination after evaluating score functions. The relative values of the upweighting amounts and the initialized probability weights determine the extent to which upweighting will affect subsequent iterations through the tree, by increasing the likelihood that a given choice will be randomly selected again. Loop choices have an initialized probability weight of 10, while helix initial probability weights are variable depending on allowed lengths. As there are many more loop options than helix options at each branch point, the starting weights for helices are normalized such that the sum of helix choice weights equals the sum of loop choice weights. This adjustment ensures that a given upweighting amount will increase the probability of a helix or a loop choice equally, as otherwise helix choice diversity would be lost much more rapidly. As discussed earlier, the additional ‘no choice’ termination option is added separately with its own initialization weight, which begins at 10% of the sum of all other choices

at a loop branch point and increases by 5% for each additional helix past the minimum helix specification. Geometric constraints are upweighted for a path through the tree by adding a value of 100 to each probability weight, while score functions are upweighted by adding a value of 5,000 multiplied by the calculated overall score, which is between 0 and 1.

To avoid excessive upweighting of single choices leading to a loss of diversity in output structures in later iterations, we implemented a limit to the maximum probability that any single choice may have. At the start of the tree this limit is 0.4, and increases by a factor of 0.025 times the number of choices down the tree up to a maximum value of 0.6. In instances where upweighting would exceed this limit, the probability weight is instead set to a value to equal the probability limit.

Volume filling sampling

To sample backbones within volumes representing a variety of shapes (Supplementary Fig.S2), we first generated the desired shapes using CAD software. These shapes were converted to volume hash tables to use as geometric constraints. For each shape, we ran thousands of different MCTS runs, each initialized randomly within the volume. We first used 500 test iterations to check that backbone lengths and scores were improving indicating a reasonable initialization point, then ran an additional 50,000 iterations for each passing MCTS run. We permitted helices between 6 and 23 residues in length, and set a minimum of 6 helices and 100 residues of sequence length. For the purpose of filling the desired shapes, we provided only sequence length as a score function, so that the MCTS would prioritize finding paths that fit as many residues in the volume as possible. For Supplementary Fig.S2, we provide for each shape the 5 longest backbones found, as well as our favorite backbone from the same set of 5 most closely resembling the shape.

Sampling different symmetries, porosities, and sizes

To sample the range of symmetries in Supplementary Fig.S3, we ran thousands of different MCTS runs, each randomly initialized and sampled within geometric constraints. For the cyclic symmetries, building was constrained to a CAD-generated cylinder, while for the cage symmetries, building was constrained by outer and inner radius limits. We permitted helices between 9 and 22 residues in length, and varied the minimum and maximum numbers of helices and residues as well as the geometric size constraints to obtain designs of different sizes (Supplementary Fig.S3-6). We provided transformation matrices to compute all symmetry mates. For the cage symmetries, we used the same five score functions as for the icosahedral assembly sampling production runs. To sample cages with higher porosities (Supplementary Fig.S5), we changed the sigmoid activation of the porosity score function to instead reward lower scores. For the cyclic symmetries, we used the same score function set with the exception of the porosity

score. We ran each initialized search tree for 10,000 iterations. For Supplementary Fig.S4, we designed symmetries using ProteinMPNN and predicted monomer folding with AlphaFold. We selected characteristic designs for each symmetry or property to construct the figures.

Nanopore sampling production runs

To build nanopores via void filling between outer and inner C6 rings, we first generated docks of previously designed inner and outer rings, sampling inversion, translation, and rotation about the C6 symmetry axis (in this case the z-axis). The outer ring is a modified parametrically generated design (*36*), while the inner rings are designs that were generated using symmetric protein hallucination (*6*). We provided all backbone atoms of the inner and outer rings to restrict the x and y -dimensions of the build to the space between the rings. We geometrically constrained the z-dimension for each dock by constructing a cone-shaped cartesian limit based on the position of the inner ring relative to the outer ring. We permitted helices between 7 and 25 residues in length. We required at least 3 helices, but placed no limits on maximum number of helices, nor minimum or maximum number of residues, provided the geometric space to fill was severely constrained. Similarly, we provided sequence length as the only score function to upweight, rewarding longer builds that effectively filled the narrow space to close the surface of the nanopore. We initialized thousands of C6 symmetric MCTS runs, building from the six N-termini of the outer ring. For each run, we output at most the 5 longest sequence length builds with distances between build N-terminus and inner ring C-terminus of < 20 Å to allow for linking of chains into a single C6 construct using protein inpainting (*74*).

Icosahedral assembly sampling production runs

To sample small icosahedral backbones, we used random stub initialization within a sphere of radius 60 Å. For each MCTS run, we used 500 test iterations to check that the initialization point was reasonable, followed by 10,000 iterations. We permitted helices between 9 and 22 residues in length. We set a minimum of 3 helices, a maximum of 7 helices, a minimum of 50 residues, and a maximum of 80 residues. We set a maximum radial distance of 75 Å to constrain the size of the assemblies. We found that this search space was constrained enough to efficiently find a large number of diverse and high quality backbones of similar sizes. We used the three monomer score functions and two icosahedral assembly score functions described previously. Each score was passed through a sigmoid activation. The product of these five scores was multiplied by an additional reweight factor of 0.05, then passed through a final sigmoid activation. The sigmoid activations each have three relative weights and the following equation:

$$y = \frac{m}{1 + e^{-10a*(score-b)}}$$

These weights are listed for each score in Table S1.

For each run, we output at most the 5 highest scoring backbones, where all output backbones required a core formation score greater than or equal to 0.2, a helix packing score greater than 2.0, a monomer designability score greater than or equal to 0.9, a porosity score greater than or equal to 0.45, and an interface designability score greater than 17. Each run produced on average fewer than 1 backbone with these strict score thresholds. We repeated the randomly-initialized MCTS icosahedral assembly run millions of times, each with an independent tree explored through 10,000 iterations, to produce a library of approximately 1 million icosahedral backbones. Tree search iterations took on average 0.045 seconds, with each run taking on average 7.5 minutes. We estimate this library took approximately 1.5 million CPU hours to generate.

Sequence length and porosity analysis of native, bottom-up, and top-down icosahedra

To perform the sequence length and porosity analysis in Figure 1d, we first downloaded all icosahedral homo 60-mers from the PDB (as of August 1, 2022). All partial and non-native structures were removed, as well as the much longer sequence length PDB 3J3I for graph clarity, leaving a library of 240 icosahedra. A total of 3 icosahedral homo 60-mer "bottom-up" structures were obtained from databases at the IPD. A random selection of 25 MCTS-generated designs were selected for the "top-down" structures from the set of designs screened experimentally. Amino acid sequence length for a single subunit was determined from the structure files. A ray tracing method was utilized to find a measure of porosity for each structure, in which 10,000 rays with random direction were traced from the origin to the outside of the icosahedron. The fraction of rays that leave the capsid without coming within 0.5 Å of an amino acid atom is reported as the porosity. The choice of 10,000 random rays was benchmarked for 10 random native icosahedra against a larger more accurate run of 100,000 random rays, and found to be > 95% accurate in all 10 cases.

Protein BiRNN model for sequence profile prediction

The trRosetta (*61*) training set was used to train a deep neural network (ProteinBiRNN) for the prediction of each amino acid probability at each residue position given solely the backbone structure. The architecture of the deep neural network consists of several convolutional layers, a bidirectional recurrent neural network and a residue-level attention layer. All the input features are directly derived from the backbone structure and calculated on-the-fly. The 1D features include: 1) per-residue torsion angle phi, psi and omega; one hot encoded secondary structure type (H, L, E); and number of sidechain neighbors). The 2D features include: 1) inter-residue distance; relative orientation geometries (Supplementary Fig.S10). The network predicts the amino acid identity of each residue and we used categorical cross-entropy to measure the loss for the training. All the trainable parameters restrained by the L_2 penalty and dropout keeping probability 90% is used.

Sequence design, evaluation, and selection

The computational sequence design pipeline contains multiple modules (Supplementary Fig.S9). Position-specific scoring matrices (PSSM) were generated with the ProteinBiRNN model. Using tertiary structures as input, either a monomer or a multi-chain fragment or the entire capsid, the ProteinBiRNN predicts sequence probability distribution matrices for each residue position, based on its local backbone neighborhood environment. For each MCTS-generated capsid backbone, two ProteinBiRNN profile predictions were calculated for both monomer fold (pssm_chA) and asymmetric subunit that contain all interfaces (pssm_int), respectively.

To reduce unnecessary calculations, a fast pre-design step (< 2 min per CPU) was first performed to examine whether a monomer building block would preferentially fold into the designed backbone configuration. Residue type constraints (pssm_chA) were set by using the FavorSequenceProfile mover during Rosetta Fastdesign. Designed monomer scaffolds with unfavorable Rosetta score/residue (> -2.7) and AlphaFold metrics (pLDDT < 80, RMSD > 1.5 Å) were excluded from the pool for downstream studies. For example, in a test set of around 80,000 MCTS-generated backbones, about 30 % of designs passed the above filtering criteria (Supplementary Fig.S11).

Next, an interface screening stage of symmetric RosettaDesign calculation was performed for all capsids to design contacting assembly interfaces and protein cores in a single step, as guided by pssm_int. We applied a Rosetta FastDesign protocol, similar to the one described in a recent publication (*38*), to activate between side-chain rotamer optimization and gradient-descent-based energy minimization. Symmetric Rosetta backbone relaxation and small random perturbation to local backbone positions were observed to improve upon sequence encoding and interface shape complementarity, as compared to conventional cage design protocols with fixed input backbones (*11*). Computational metrics of the final design models were calculated using Rosetta, which includes ddG, shape complementarity and interface buried SASA, contact molecular surface, among others, for design selection (Supplementary Fig.S13). AF metrics suggest the sequence encoding obtained by ProteinBiRNN PSSM-guided Rosettadesign outperforms conventional layer design (Supplementary Fig.S14). All the script and flag files to run the programs are provided in the Supplementary Information. With backbone relaxation, a typical symmetric capsid design calculation takes about 3.5 hours on average to finish on a CPU.

To evaluate the interface quality of designed capsids, symmetric interfaces between two neighboring chains (C2,C3, and C5) were extracted and analyzed independently (Supplementary Fig.S12). To ensure capsid formation, we hypothesized that at least two relatively strong interfaces need to exist, where a first one is needed for homo-oligomeric formation, and a second one closes the capsid. Inspired by previous studies (*11*), we set the cutoff for relatively strong interface based on its calculated Rosetta metrics, including binding energy (ddG < -20) and

contact molecular surface (cms >180) with appreciable hydrophobic side chain packing (cms_apolar> 120).

During the final sequence optimization stage, Rosetta and ProteinMPNN (*37*) were used to design the selected capsid set with good backbone dockings (≥ 2 strong interfaces). An iterative backbone optimization process was cycled three times during the final sequence optimization step to improve interface shape complementarity and designed backbone robustness. For each design round, the Rosetta relaxed design model and its corresponding AF monomer prediction (from previous round) were TMaligned (*62*) back to the starting position and used as input for capsid design using Rosetta or ProteinMPNN (Supplementary Fig.S15). To design capsid sequences with the ProteinMPNN module, the tertiary structure of a full capsid was used as input while all 60 sequences were tied to be identical to enforce icosahedral symmetry. A typical proteinMPNN design calculation takes about 5 mins on a GPU for a homo-60 mer capsid with ~ 70 residues per monomer, which is significantly faster than conventional Rosetta Fastdesign. To generate pdb models for the MPNN output sequences, the output sequences were grafted onto the input backbone using Rosetta SimpleThreadingMover (*63*) followed by symmetric relaxation. Finally, all final capsid designs were evaluated by Rosetta (monomer and interface metrics) and Alphafold (monomer pLDDT and RMSD) for design selection.

Protein expression and purification

For small scale screening, synthetic genes encoding capsid sequences were optimized for E. coli expression and purchased from IDT (Integrated DNA Technologies) as linear eblock DNA in 96-well format. We performed Golden gate assemblies following the standard protocol (New England Biolab) to clone each synthetic gene into a pET29b vector with a hexahistidine affinity tag. Plasmids were then transformed into BL21* (DE3) (Invitrogen) E. coli competent cells and grew overnight in 1 mL of Studier autoinduction media. The cells were harvested by centrifugation, resuspended and lysed in 200 µL Bugbuster protein extraction reagent (Millipore). Following another centrifugation, the supernatant was purified by Ni^{2+} immobilized metal affinity chromatography (IMAC) with Ni-NTA charged magnetic beads (Qiagen). Magnetic beads with bound cell lysate were washed with 200 µL of washing buffer for three times (150 mM NaCl, 25 mM Tris pH 8.0, 30 mM imidazole) and eluted with 35 µL of elution buffer (150 mM NaCl, 25 mM Tris pH 8.0, 500 mM imidazole). Soluble fractions were analyzed by SDS-PAGE to check for a single intense band at expected molecular weights. The selected IMAC elution of soluble designs were diluted to appropriate concentration with a buffer (1:20, 150 mM NaCl, 25 mM Tris pH = 8.0) and analyzed by negative-stain electron microscopy (see below).

For large scale protein expression, plasmids were transformed into BL21* (DE3) E. coli competent cells. Single colonies from agar plate with 100 mg/L kanamycin were inoculated in 50

mL of studier autoinduction media, and the expression continued at 37 °C for over 24 hours. The cells were harvested by centrifugation, resuspended in 150 mM NaCl, 25 mM Tris pH 8.0, and lysed by 20 mL Bugbuster protein extraction buffer (Millipore, 50 mL Bugbuster lysis buffer contains - 49 mL Bugbuster reagent, 0.75 mL 2M imidazole, 1 protease inhibitor tablet, 50 mg lysozyme). Following another centrifugation, the supernatant was purified by 1 mL Ni^{2+} IMAC with Ni-NTA Superflow resins (Qiagen). Resins with bound cell lysate were washed with 10 mL (bed volume 1 mL) of washing buffer (same as above) for two times and eluted with 3 mL of elution buffer (same as above). To improve colloidal stability of the capsid protein, glycine was added to the IMAC elution to reach a final concentration of 100 mM, which was subsequently concentrated and purified via size exclusion chromatography (SEC) in 25 mM pH 8.0 Tris buffer (150 mM NaCl) on a Superose 6 Increase 10/300 gel filtration column (Cytiva). The resulting samples were generally > 95 % homogeneous on SDS-PAGE gels. SEC-purified designs were concentrated by 10K concentrators (Amicon) and quantified by UV absorbance at 280 nm and stocked at 4 ºC.

The HA mini capsid gene was composed of the A/Michigan/57/2015 (MI15) H1N1 HA ectodomain with the Y98F mutation (*75*) fused via a 6(GS) linker to the RC_I_1 capsid with an N-terminal hexahistidine affinity tag. This gene was codon optimized for human cell expression and made in the CMV/R mammalian expression vector by Genscript. Transient transfection into HEK293F cells was carried out using PEI MAX. After 4 days of expression, mammalian cell supernatants were clarified via centrifugation and filtration. IMAC was used to initially purify HA capsids from cell supernatant by adding in 1 ml of Ni2+ sepharose excel resin per 100 ml supernatant, along with 5 ml of 1 M Tris, pH 8.0 and 7 ml of 5 M NaCl. Batch binding was left to proceed for 30 min at room temperature with shaking. Resin was then collected in a gravity column, washed with 5 column volumes of 50 mM Tris, pH 8.0, 500 mM NaCl, 20 mM imidazole, and his-tagged HA capsids were eluted using 50 mM Tris, pH 8.0, 500 mM NaCl, 300 mM imidazole. Purification by SEC was then carried out on a Superose 6 Increase 10/300 gel filtration column equilibrated in 25 mM Tris, pH 8.0, 150 mM NaCl.

Bio-layer interferometry (BLI)

BLI was carried out using an Octet Red 96 system, at 25°C with 1000 rpm shaking. Antibodies were diluted to 10 ug/ml in kinetics buffer (PBS with 0.5% serum bovine albumin and 0.01% Tween) and then loaded onto protein A tips for 200 s. HA mini capsid was diluted to 500 nM in kinetics buffer and its association was measured for 200 s, followed by dissociation for 200 s in kinetics buffer alone.

Transmission negative-stain electron microscopy (nsEM)

Selected IMAC elutions and cage fractions from SEC traces were diluted to about 0.5 μM (monomeric component concentration) for negative-stain EM characterization. A drop of 5 μL sample was applied on negatively glow discharged, formvar/carbon supported 400-mesh copper grids (Ted Pella, Inc.) for 1 min. The grid was blotted and stained with 5 μL of uranyl formate, blotted again, and stained with another 5 μL of uranyl formate for 60 s before final blotting. 2% uranyl formate was used for all samples.

The screening and data collection was performed on a 120kV Talos L120C transmission electron microscope (Thermo Scientific) with a BM-Ceta camera using EPU 2.0. All nsEM datasets were processed by the CryoSparc software (*64*). Micrographs were imported into the CryoSparc web server and the contrast transfer function (CTF) was corrected. All the picked particles were 2D classified for 20 iterations into 50 classes. Particles from selected classes were used for building the ab-initio initial model. The initial 3D model was homogeneously refined using C1 and the corresponding Icosahedral symmetry.

Circular dichroism (CD) experiments

To study the secondary structure and thermodynamics of the designed capsid proteins, CD measurements were performed with an JASCO 1500. The 200 to 195 nm wavelength scans were measured at every 10 °C intervals from 25 to 95 °C. For each measurement, a 1-mm path length cuvette was loaded with protein concentrations at 0.2 mg/mL in TBS buffer, as measured by Nanodrop at 280 nm using predicted extinction coefficients.

CryoEM sample preparation

To prepare cryoEM sample grids for the capsids, 3 μL of 0.5 - 1.0 mg/mL of capsid proteins in 150 mM NaCl, 25 mM Tris (pH = 8.0), 100 mM Glycine was applied to glow-discharged Quantifoil R 2/2 300 mesh copper grids overlaid with a thin layer of carbon. Vitrification was performed on a Mark IV Vitrobot with a wait time of either 5 or 7.5 seconds, a blot time of 0.5 seconds, and a blot force of either 0 or -1 before being immediately plunged frozen into liquid ethane. The sample grids were clipped following standard protocols before being loaded into the microscope for imaging.

CryoEM data collection

Data collection was performed automatically using either Leginon (*65*) or SerialEM to control either a ThermoFisher Titan Krios 300 kV equipped with a standalone K3 Summit direct electron detector with an energy filter, or a ThermoFisher Glacios 200 kV equipped with a standalone K2

Summit direct electron detector (*66*). All three capsids were collected using counting mode with random defocus ranges spanned between -0.7 and -2.0 μm using image shift, with one-shot per hole on a Glacios for RC_1-H11, and 6 shots per hole on a Krios for RC_I_1 and RC_I_2. 3,888, 12,825, and 1,315 movies were collected with a pixel size of 0.84 Å, 0.84 Å, and 1.16 Å for RC_I_1, RC_I_2, and RC_I_1-H11 capsids, respectively. A total dose of ~60 $e^-/Å^2$ was applied to RC_I_1 and RC_I_2 on a Titan Krios, and ~50 $e^-/Å^2$ for RC_1_I-H11 on a Glacios.

For nanopore designs, CryoEM grids were prepared by diluting protein samples with TBS 1 to 10 times immediately before applying 3.5 μL to glow-discharged 400 mesh, C-flat, 2 micron holes, 2 micron spacing, CF-2/2-4C (CF-224C-100) (Electron Microscopy Sciences) cryoEM grids. Grids were blotted using a blot force of 0 and 5.5 second blot time at 100% humidity and 4 °C and plunge-frozen in liquid ethane using a Vitrobot Mark IV (FEI Thermo Scientific). cryoEM grids were screened on a Glacios transmission electron microscope (FEI Thermo Scientific) operated at 200 kV and equipped with a K3 Summit direct detector. Automated Glacios data collection was carried out using SerialEM software at a nominal 54 magnification of 36,000× (0.883 Å/pixel). Movies were acquired in counting mode fractionated in 50 frames of 200 ms at 8.5 e-/pixel/sec for a total dose of ~65 $e^-/Å^2$.

CryoEM data processing

All data processing was carried out in CryoSPARC and CryoSPARC Live (*64*). Alignment of movie frames was performed using Patch Motion with an estimated B-factor of 500 $Å^2$, with a maximum alignment resolution set to 3. Defocus and astigmatism values were estimated using Patch CTF with default parameters. RC_I_1 particles were initially picked in a reference-free manner using Blob Picker and extracted with a box size of 320 pixels. This was followed by a round of 2D classification and subsequent template-picking using the best 2D class averages low-pass filtered to 20 Å, for a total of 805,450 picked particles. For RC_I_1-H-11, 491,297 particles were picked with Template Picker using templates from the RC_I_1 dataset, and were extracted with a box size of 320 pixels. For RC_I_2, 769,399 particles were ultimately picked and extracted with a box size of 320 pixels after a round of reference-free picking using Blob Picker and a subsequent Template Picker job using templates derived from 2D class averages which were low-pass filtered to 20 Å. Rounds of reference-free 2D classification were next performed in CryoSPARC with a maximum alignment resolution of 6 Å for each dataset. The best classes for each sample that revealed clearly visible secondary-structural elements were used for 3D *ab initio* determination using the C1 symmetry operator. RC_I_1 and RC_I_1-H-11 datasets were next corrected for local particle motion followed by 3D *ab initio* using C1 and I symmetry, respectively. This was followed by a 3D refinement with I symmetry for RC_I_1 for a final global resolution estimate of 2.5 Å. For RC-I_1-H-11, *ab initio* was followed by Non-Uniform Refinement and Local Refinement for a final global resolution estimate of 3.0 Å. For RC_I_2, data was processed entirely in CryoSPARC Live during collection. After selection

of the best 2D class averages following template picking, an ab-initio job was run using C1 symmetry on 100,000 particles, and a final refinement performed with I symmetry applied using 283,552 of the best particles from 2D classification, yielding a global resolution estimate of 2.93 Å. Nanopore particles were initially picked in a reference-free manner using Blob Picker and extracted with a box size of 320 pixels. This was followed by multiple rounds of 2D classification and subsequent template-picking using the best 2D class averages. The best classes that revealed clearly visible secondary-structural elements, a total of 19,418 particles, were used for 3D *ab initio* determination using the C1 symmetry operator. This was followed by a 3D non uniform refinement with C6 symmetry, global CTF refinement, local refinement and homogeneous refinement for a final global resolution estimate of 5.13 Å (Fig. S8). Local resolution estimates were determined in CryoSPARC using an FSC threshold of 0.143. 3D maps for the half maps, final unsharpened maps, and the final sharpened maps for each capsids RC_I_1, RC_I_1-H11, and RC_I_2, and nanopore design RNR_C6_3 were deposited in the EMDB under accession number EMD-28860, EMD-28858, EMD-28859, and EMD-29939 respectively.

CryoEM model building and validation

The de novo predicted design models for each capsid (reported here) were used as initial references for building the final cryoEM structures. The models were manually edited and trimmed using Coot (*67*, *68*). We further refined each structure in Rosetta using density-guided protocols (*69*). EM density-guided molecular dynamics simulations were next performed using Interactive Structure Optimization by Local Direct Exploration (ISOLDE), with manual local inspection and guided correction of rotamers and clashes throughout simulated iterations. ISOLDE runs were performed at a simulated 25 Kelvin, with a round of Rosetta density-guided relaxation performed afterward. This process was repeated iteratively until convergence and high agreement with the map was achieved. Multiple rounds of relaxation and minimization were performed on the complete capsids, followed by human inspection for errors after each step. Throughout this process, we applied strict non-crystallographic symmetry constraints in Rosetta (*69*). Phenix real-space refinement was subsequently performed as a final step before the final model quality was analyzed using Molprobity (*70*) and EM ringer (*71*). Figures were generated using either UCSF Chimera (*72*) or UCSF ChimeraX (*73*). The final structures for RC_I_1, RC_I_2, and RC_I_1-H11 were deposited under PDB accession numbers 8F54, 8F53, and 8F4X, respectively.

Cell culture

Human Umbilical Vein Endothelial Cells (HUVECs) were acquired from Lonza (C2519AS). Cells were grown on 0.1% gelatin-coated (Sigma, G1890-100G) 35-mm cell culture dish in EGM2 media described previously (*43*). HUVECs were expanded and serially passaged to reach passage 7 before experiments.

RC_I_1-H11-Fd conjugation

F-domain fused with SpyTag were incubated with RC_I_1-H11 capsid fused with SpyCatcher for 4 hours at room temperature on nultation. Samples were analyzed on a SDS–PAGE gel to confirm at least 90% conjugation was reached (Supplementary Fig.S27).

pAKT titration

Passage 7 HUVECs at 80% confluency were rinsed with 1× PBS (Gibco, 10010023) twice. The cells were then starved in DMEM low glucose 1 g/l (Gibco, 11885-084) serum-free media for 16 hours. At 16-h timepoint, cells will be treated with RC_I_1-H11-Fd (1000-0.1 nM), Fd-st (100 nM), Ang1 (100 nM), I53-50 (100 nM), or PBS for 15 minutes at 37°C. After treatment, the media was aspirated, and cells were washed once with 1× PBS before harvesting protein for immunoblotting (Supplementary Fig.28).

Immunoblotting

Cells were lysed with 130 µl of previously described lysis buffer (*43*). Cell lysates were collected in a fresh Eppendorf tube with 4× Laemmli Sample buffer (Bio-Rad, 1610747) containing 10% beta-mercaptoethanol (Sigma-Aldrich, M7522-100) added and then heated at 95°C for 10 min. 30 µl of protein sample per well was loaded and separated on a 4– 10% SDS–PAGE gel for 30 min at 250 Volt using running buffer (7.2 g glycine, 1.5 g Tris-base, and 0.5g SDS diluted in 1L DI water). The proteins were then transferred on a nitrocellulose membrane for 12 min using transfer butter (7.3 g Tris-Base, 3.6 g glycine, and 0.46 g SDS in 1L DI water). Post-transfer, the membrane was blocked in 5% bovine serum albumin for 1 hour. Then the membrane was probed with primary antibodies overnight: pAkt-S473 (Cell Signaling, catalog# 9271S) at 1:2,000, b-Actin (Cell Signaling, catalog# 3700S) at 1:10,000, and S6 (Cell Signaling, catalog#2217S) at 1:1000. Membranes with primary antibodies were incubated at 4°C, overnight on a rocker. Then the membranes were washed with TBST buffer (2.4 g Tri-HCl, 8 g NaCl, and 1 mL Tween-20 diluted in 1L water with pH adjusted to 7.4) 3 times at 5-min intervals. Following washes, the pAKT membrane was blocked in 5% milk at room temperature for 1 hour and then incubated in the respective HRP-conjugated secondary antibody (BioRad, catalog#1721019) at 1:2,000 dilution in 5% milk for 1 hour. All other primary antibodies were removed and washed 3 times before adding secondary antibodies at 1:10,000 for 1-hour incubation. After 1 hour, membranes were washed with TBST (3 times, 5 min of interval) and developed using Chemiluminescence developer and imaged using Bio-Rad ChemiDoc Imager. Data were quantified using the ImageJ software to analyze band intensity. pAKT band intensity was divided by band intensity of actin or s6. All signaling levels are normalized to RC_I_1-H11-Fd at 10 nM signal levels as an internal positive control. EC_{50} is calculated using FindEC anything in Prism, GraphPad.

Immunofluorescence staining of FOXO1

FOXO1 analysis was done as described before (*43*). Briefly, passage 7 HUVECs were seeded on glass coverslips coated with 0.1% gelatin and cultured until confluency. Once confluent, cells were starved for 16 hours in low glucose DMEM (1 g/l D-glucose). Then, cells were stimulated RC_I_1-H11-Fd, Fd-st, or Ang1 at 100 nM of F-domain concentration for 15 min before fixing with 4% PFA (EMS, 15710) for 15 min. The fixed cells were washed three times at 5 min each before blocking for 1 hour with 3% BSA and 0.1% Triton diluted in PBS while on nutation. The cells were incubated with FOXO1 (Cell Signaling, catalog#2880) antibody diluted at 1:100 in blocking agent overnight. After the primary antibody, the cells were washed 3 times at 5 min each with PBS while on nutation. The cells were then incubated with secondary antibodies at 1:200 for 1 hour and 20 min at 37°C. After secondary antibodies, cells were washed for three times at 10 min each with PBS on nutation. Coverslips were sealed using VECTASHIELD plus DAPI (Vector laboratories, H-2000-2) upside-down on glass slides for analysis in confocal (Leica).

Tube formation assay

Tube formation was assessed using previously described protocols (*43*, *48*). Briefly, a 24-well plate was pre-coated with 150 uL of 100% Matrigel for 30 minutes at room temperature before cell seeding. Passage 6 HUVECs were seeded at 1.5×10^5 cells/350uL density suspended in media (DMEM low glucose + 0.5% FBS ± 100nM of RC_I_1-H11-Fd). 24 hours after cell seeding, the old media is aspirated and replaced with fresh media without any treatment. The cells continue to be incubated for up to 72 hours. Capillary-like structures were observed, and 20 randomly selected microscopic fields were photographed under Nikon Eclipse Ti scope. Images were analyzed to quantify the number of nodes, meshes, and tubes in each image using the Angiogenesis Analyzer plug-in in ImageJ. Vascular stability was calculated by the average number of nodes, meshes, and tubes. Data were normalized to PBS vehicle as fold change.

Animal Immunizations

Female BALB/c mice were purchased from Envigo (order code 047) at 7 weeks of age and were maintained in a specific pathogen-free facility within the Department of Comparative Medicine at the University of Washington, Seattle. Mice were randomized, only immunized if they appeared healthy, and no mice were excluded. Prior to each immunization, immunogens were mixed with 1:1 vol/vol AddaVax (InvivoGen vac-adx-10) to reach a final dose of 1.5µg per injection. At 8 weeks of age, 5 mice per group were injected subcutaneously in the inguinal region with 100µL of immunogen for the prime immunization followed by the boost immunization 4 weeks later. Mice were bled via the submental route at weeks 2 and 6. Blood was collected in serum separator tubes (BD # 365967) and rested for 30 min at room temperature for coagulation. Serum tubes were then centrifuged for 10 min at 2,000 × *g* and serum was

collected and stored at -80°C until use. Animal experiments were conducted in accordance with the University of Washington's Institutional Animal Care and Use Committee.

ELISA

MI15 HA-foldon trimer was added to 96-well Nunc MaxiSorp plates (Thermo Scientific) at 5.0 μg/mL with 50 μL per well and incubated for 1 hour. Blocking buffer composed of Tris Buffered Saline Tween (TBST: 25 mM Tris pH 8.0, 150 mM NaCl, 0.05% (v/v) Tween20) with 5% Nonfat milk was then added at 200 μl per well and incubated for 1 hour. Next plates were washed, with all washing steps consisting of 3× washing with TBST, using a robotic plate washer (Biotek). 5-fold serial dilutions of serum starting at 1:100 were made in blocking buffer, added to plates at 50 μl per well, and incubated for 1 hour. Plates were washed again before addition of 50 μl per well of anti-mouse HRP-conjugated goat secondary antibody (CellSignaling Technology) diluted 1:2,000 in blocking buffer and incubated for 30 minutes. All incubations were carried out with shaking at room temperature. Plates were washed a final time, and then 100 μl per well of TMB (3,3′,5′,5-tetramethylbenzidine, SeraCare) was added for 2 minutes, followed by quenching with 100 μl per well of 1 N HCl. Reading at 450 nm absorbance was done on an Epoch plate reader (BioTek).

Nuclear Localization Assay

Plasmids: Transfer plasmids for lentivirus production were prepared by replacing the U6 promoter of lentiguide-BC-plasmid (Addgene #127168) (*76*) with an EF1a promoter, followed by capsid designs tagged with eGFP and three repeats of nuclear localization sequence (3xNLS). Designs included three capsids (RC_I_1, RC_I_1-H11, and RC_I_2) and a previously characterized de novo designed icosahedral cage (I3_01) (*77*).

Lentivirus production: Plasmid encoding GFP and NLS-tagged designs,, psPAX2 (Addgene #12260), and pMD2.G (Addgene #12259) were used to transfect Lenti-X 293T cells (Takara #632180) using Lipofectamine 3000 (Thermo Fisher Scientific L3000015) in a 4:3:2 mass ratio. At 4 hours post-transfection, the media (DMEM supplemented with 10% FBS and 100 U/mL penicillin-streptomycin) was exchanged. At 48 hours post-transfection, the virus were harvested and filtered through a 0.45 μm polypropylene filter (ThermoFisher #44504-PV).

Lentiviral transduction: HeLa cells at 106 cells/mL were transduced by adding viral supernatant to the media (DMEM supplemented with 10% FBS and 100 U/mL penicillin-streptomycin) supplemented with 8 μg/mL polybrene (Sigma #TR-1003-G) and centrifuging at 1000 g for 1.5 hours at 33°C. At 3 hours post-infection the media was exchanged. At 24 hours post-infection, 3 x 104 cells per well were seeded onto poly-D-lysine coated 96-well glass-bottom plates (Greiner Bio-one 655892) and incubated for 24 hours before imaging.

Imaging: The cells were fixed with 4% paraformaldehyde w/v in phosphate buffered saline (PBS) for 10 min at room temperature and subsequently stained with 1 μM Hoechst 33342 (Thermo Scientific) for 30 min and washed with PBS. The images were acquired with ECLIPSE Ti2-E inverted microscope (Nikon) using a CFI Plan Apochromat Lambda D 40X/0.95 objective, X-Light V3 spinning disk (CrestOptics), Hamamatsu Fusion camera (2304x2304 pixels, connected by CoaXPress) and analyzed with ImageJ software (http://rsb.info.nih.gov/ij).

Supplementary Figures

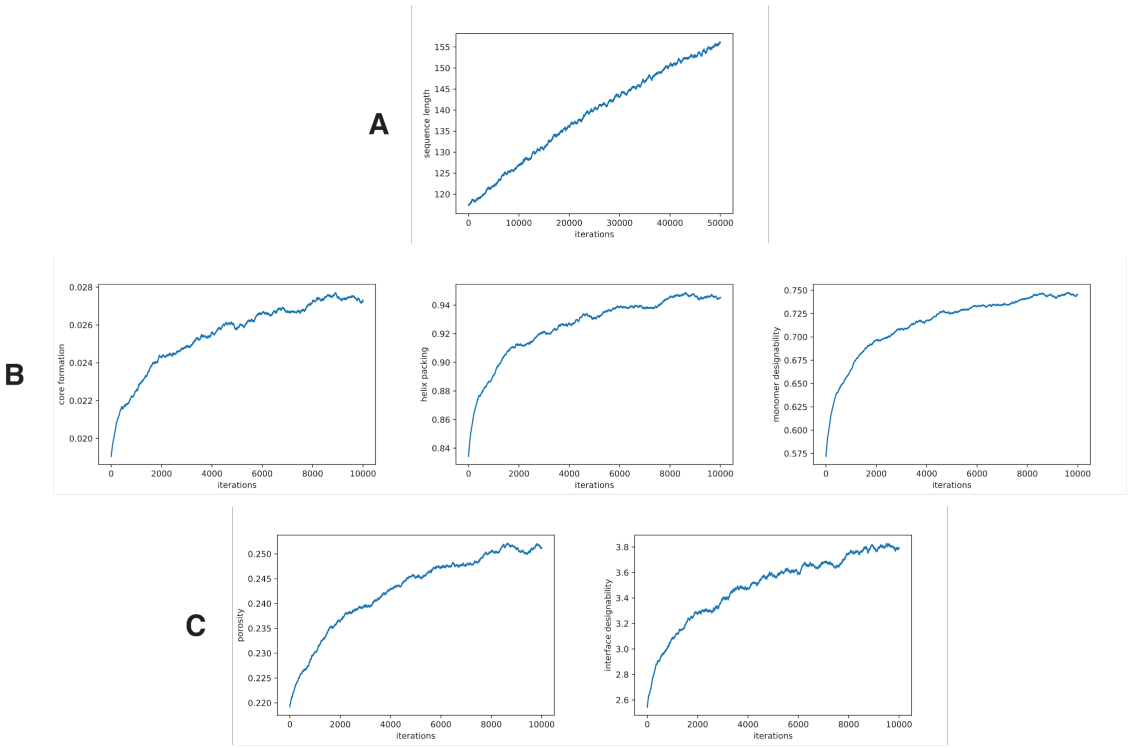

Fig. S1. Score function improvement across iterations. A rolling mean of 500 iterations was calculated across 100 independent MCTS runs to assess score function improvement over iterations. **A**. Sequence length improvement while sampling the L bracket shape from Supplementary Fig. 1. **B**. Monomer score function improvement while sampling icosahedra. **C**. Assembly score function improvement while sampling icosahedra.

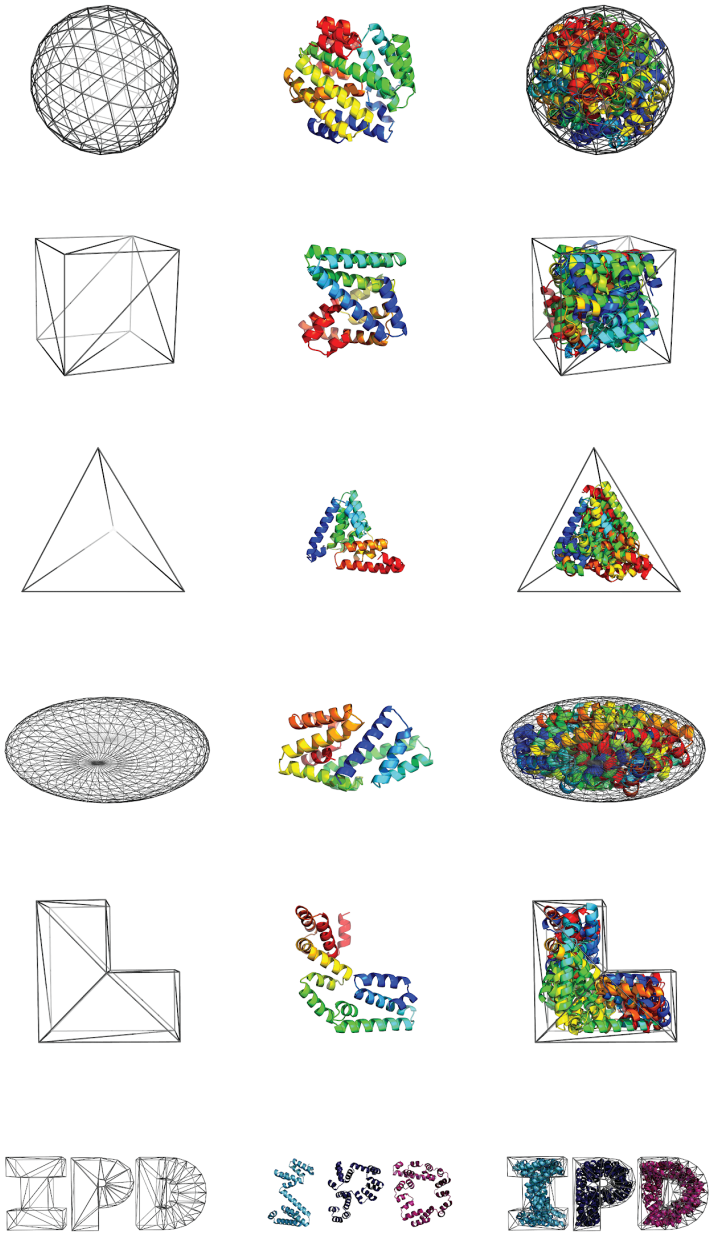

Fig. S2. Sampling protein backbones of arbitrarily prespecified shapes *in silico*. Left column, prespecified build volumes. Middle column, selected monomers matching desired shapes. Right column, overlayed 5 longest sequence length monomers filling each prespecified volume.

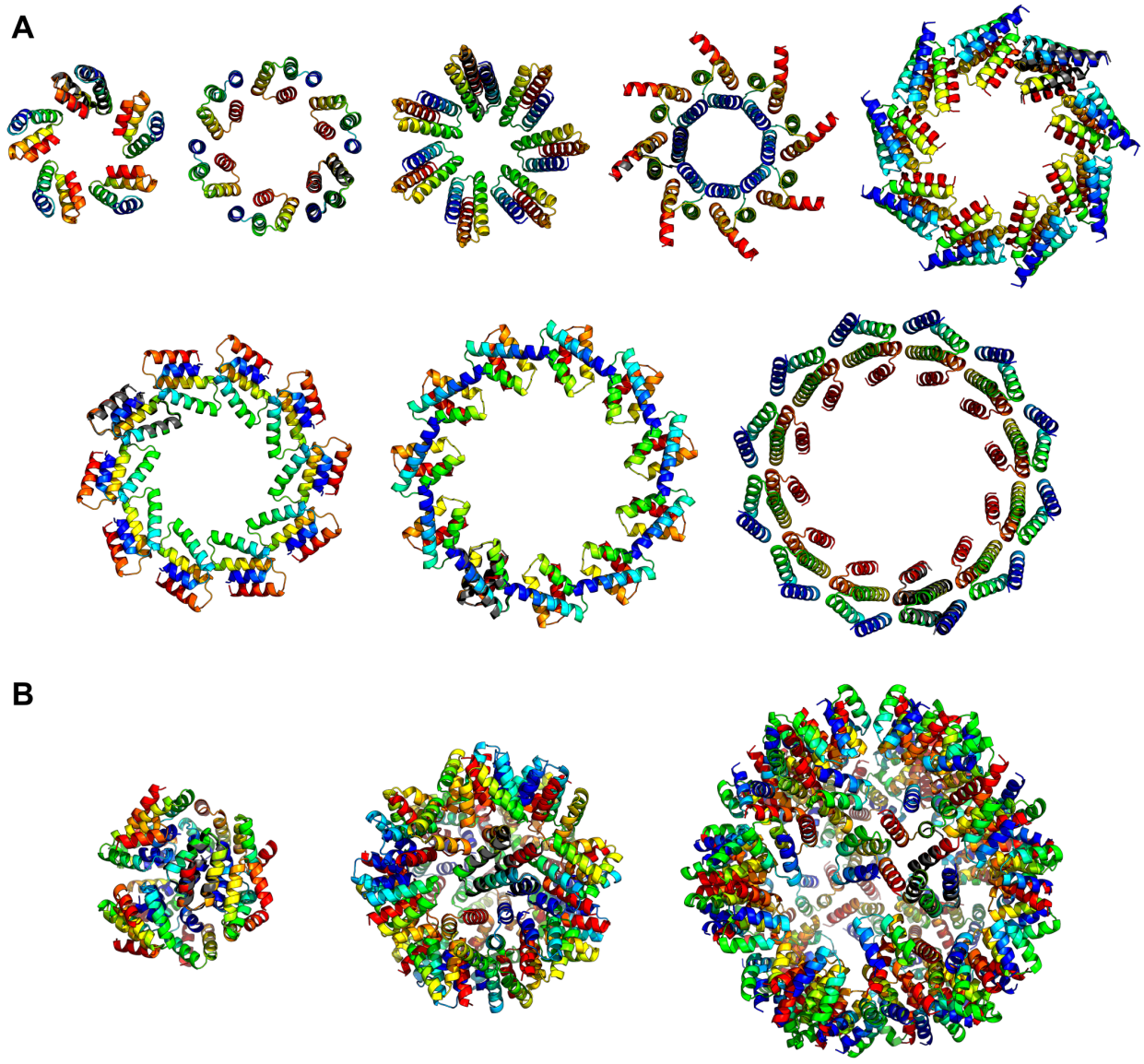

Fig. S3. *In silico* design of symmetric nanomaterials using MCTS. AlphaFold monomer predictions overlaid in dark gray. **A**, Cyclic symmetries C5, C6, C7, C8, C9, C10, C11, and C12 (left to right). **B**, Tetrahedral, octahedral, and icosahedral symmetries (left to right).

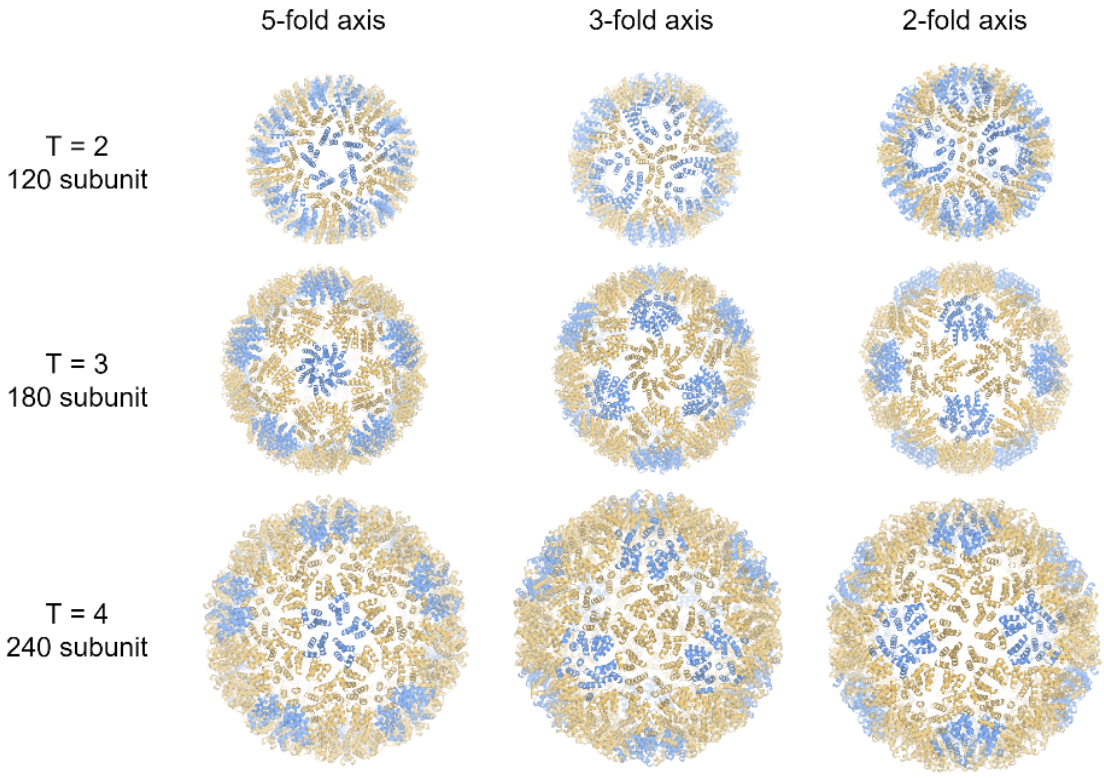

Fig. S4. Sampling quasi-symmetric icosahedral cages *in silico*.

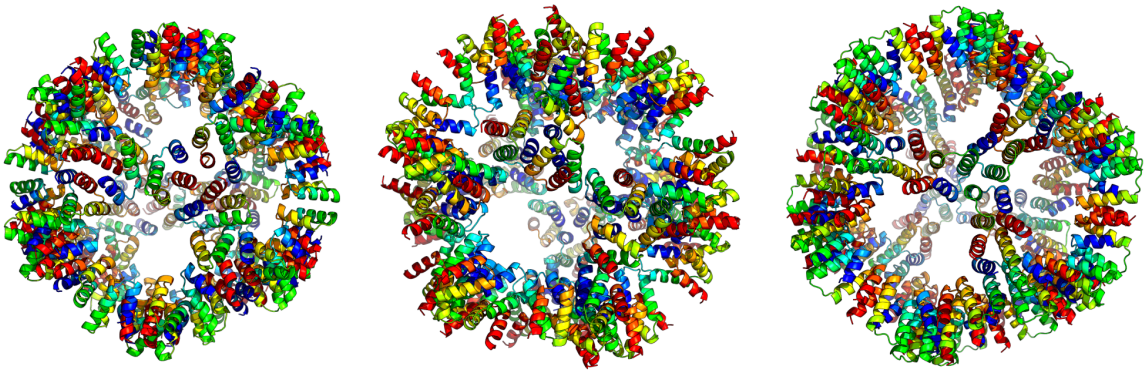

Fig. S5. Sampling high porosity icosahedral cages *in silico*.

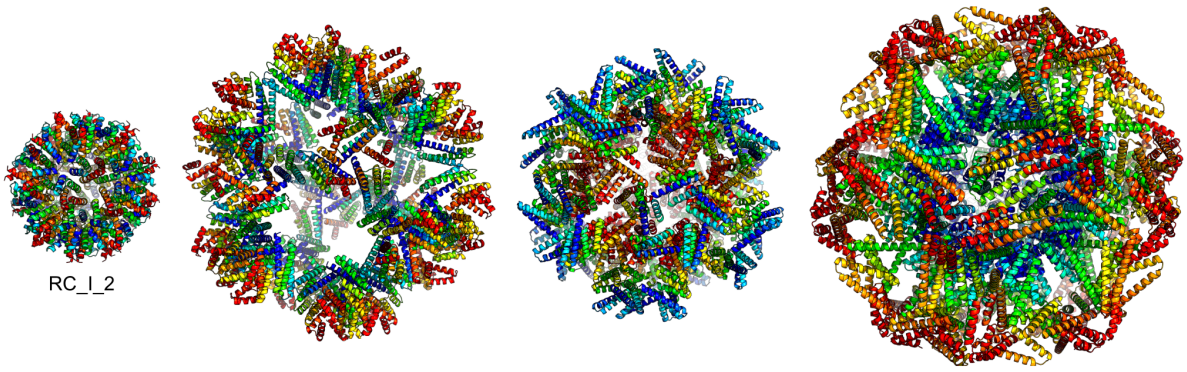

Fig. S6. Sampling larger icosahedral cages *in silico* as compared to capsid RC_I_2 (left).

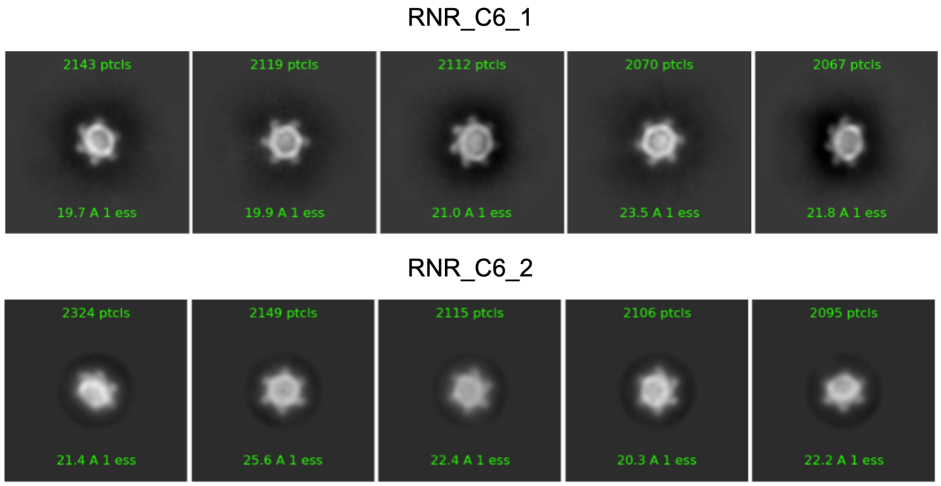

Fig. S7. nsEM 2D class average data for design RNR_C6_1 (top) and RNR_C6_2 (bottom).

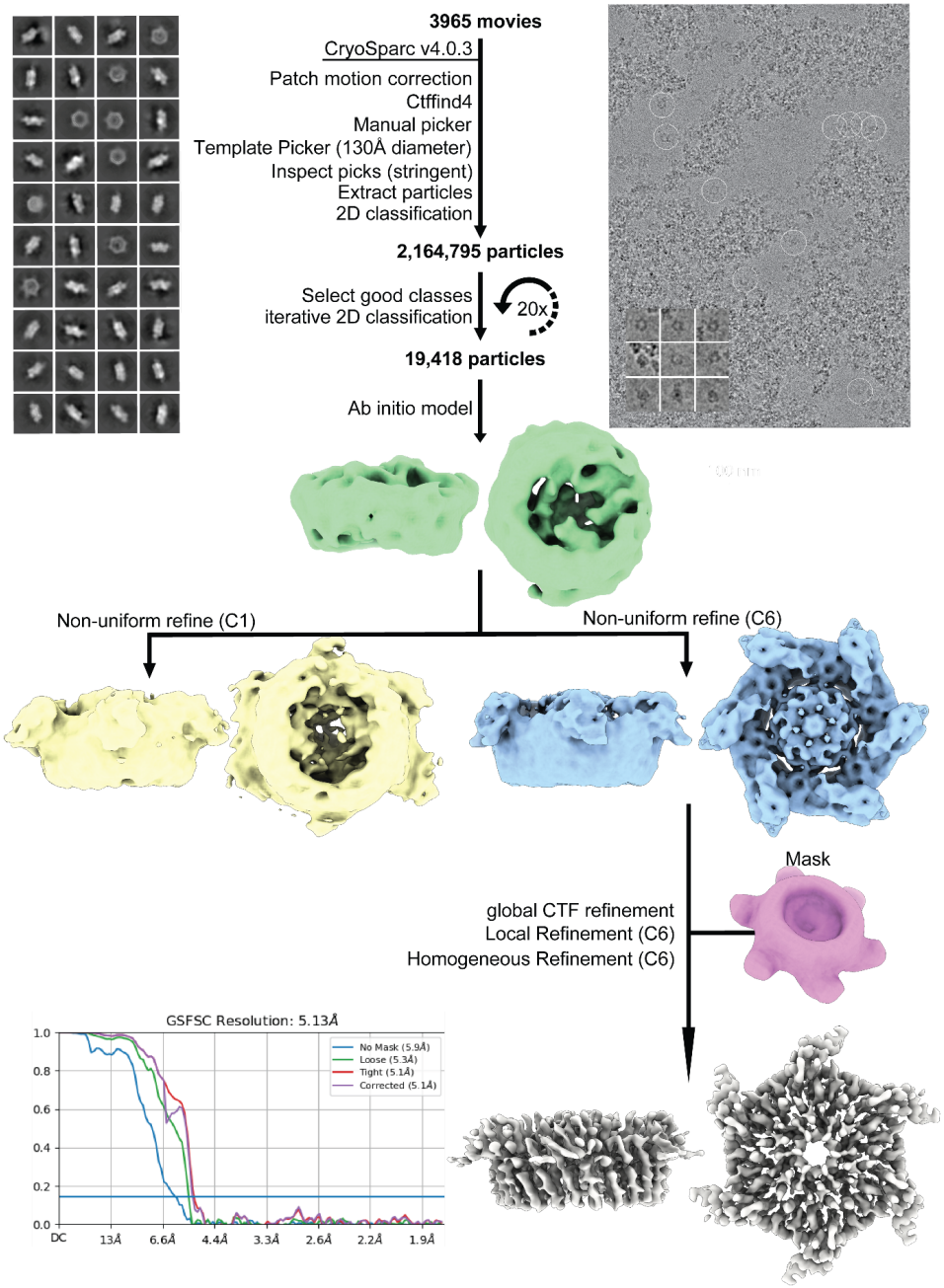

Fig. S8. Cryo-EM data processing flowchart and additional data for MCTS nanopore RNR_C6_3 density map reconstruction. Top panel: 2D Class averages, a typical cryoEM micrograph with example selected particles. Bottom panel: EM maps over the refinement process, and global resolution estimation plot.

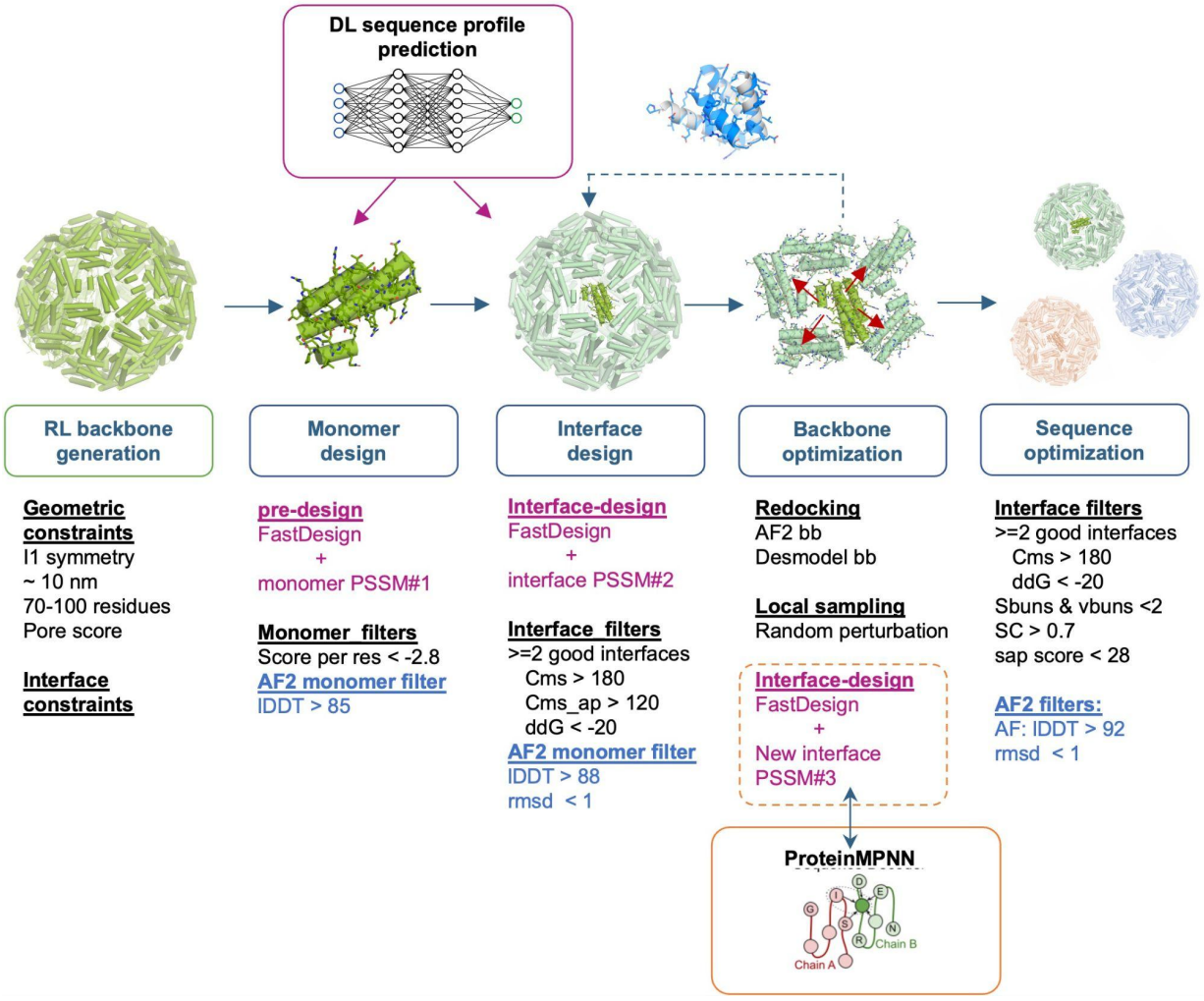

Fig. S9. Schematic illustration of capsid design pipeline.

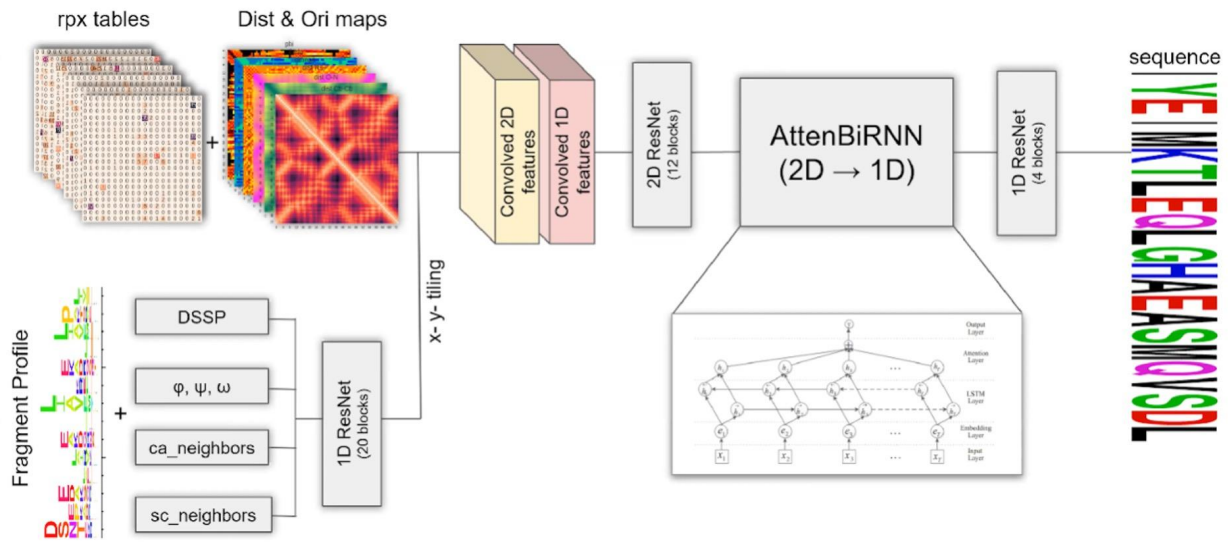

Fig. S10. Architecture of the Protein BiRNN model.

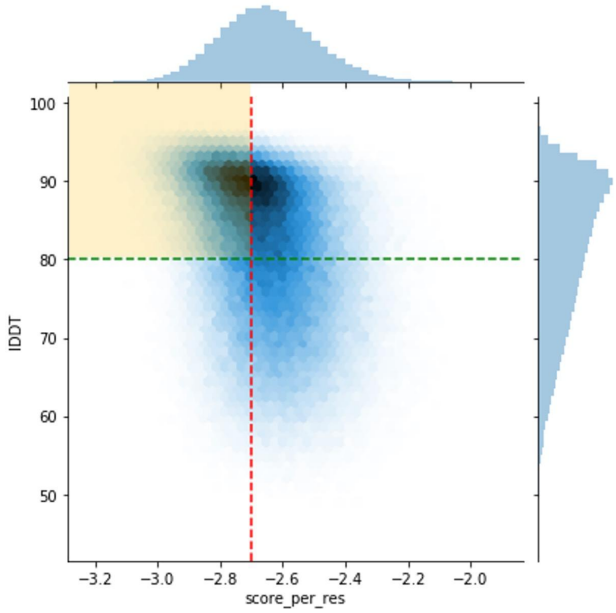

Fig. S11. Rosetta and AF2 filtering metrics for pre-designed monomers. Monomer backbones with Rosetta score_per_res < -2.7 and AF2 pLDDT > 80 were selected (the region highlighted in yellow) for downstream interface design.

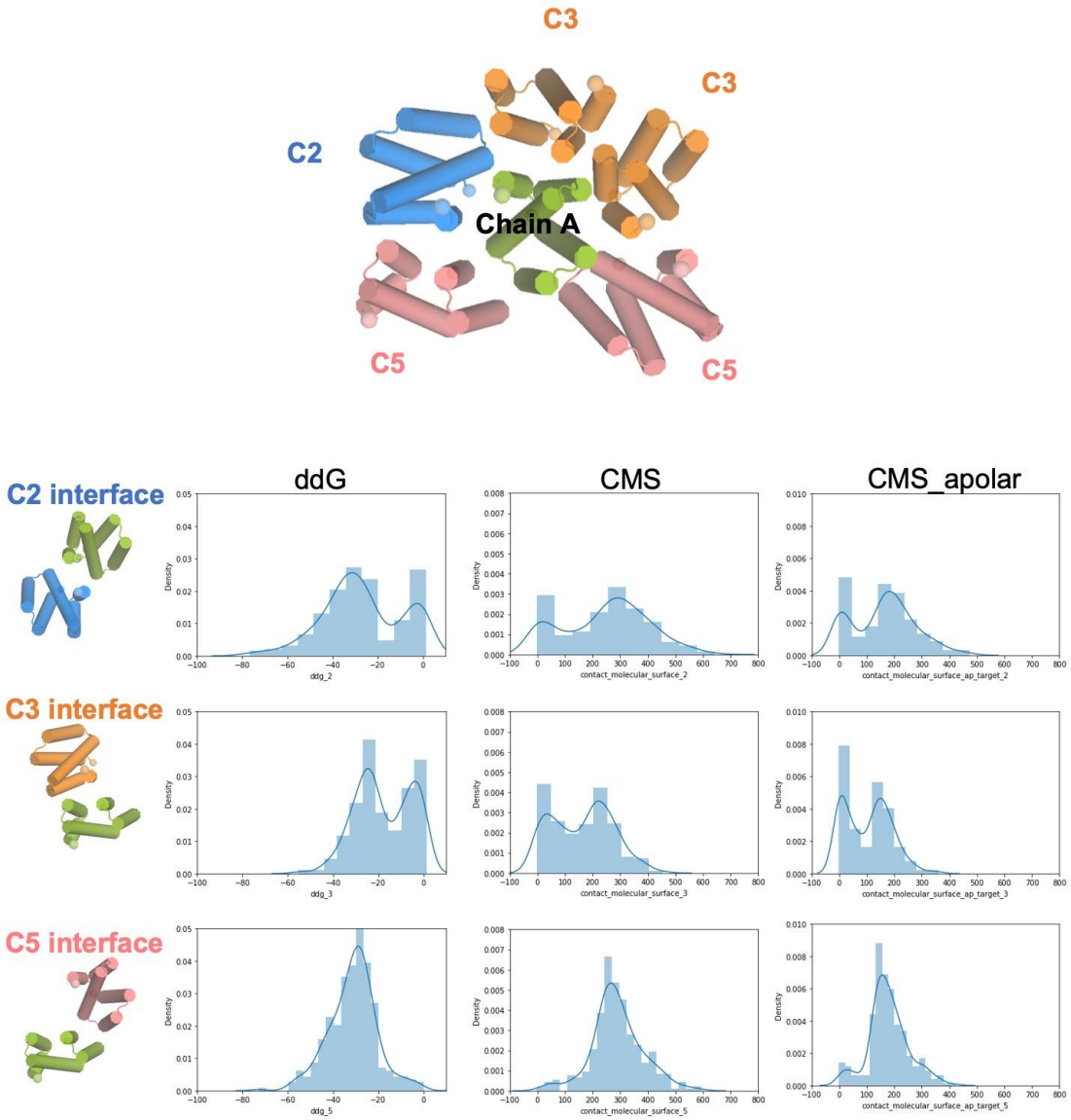

Fig. S12. Rosetta metrics for individual capsid interfaces by symmetry.

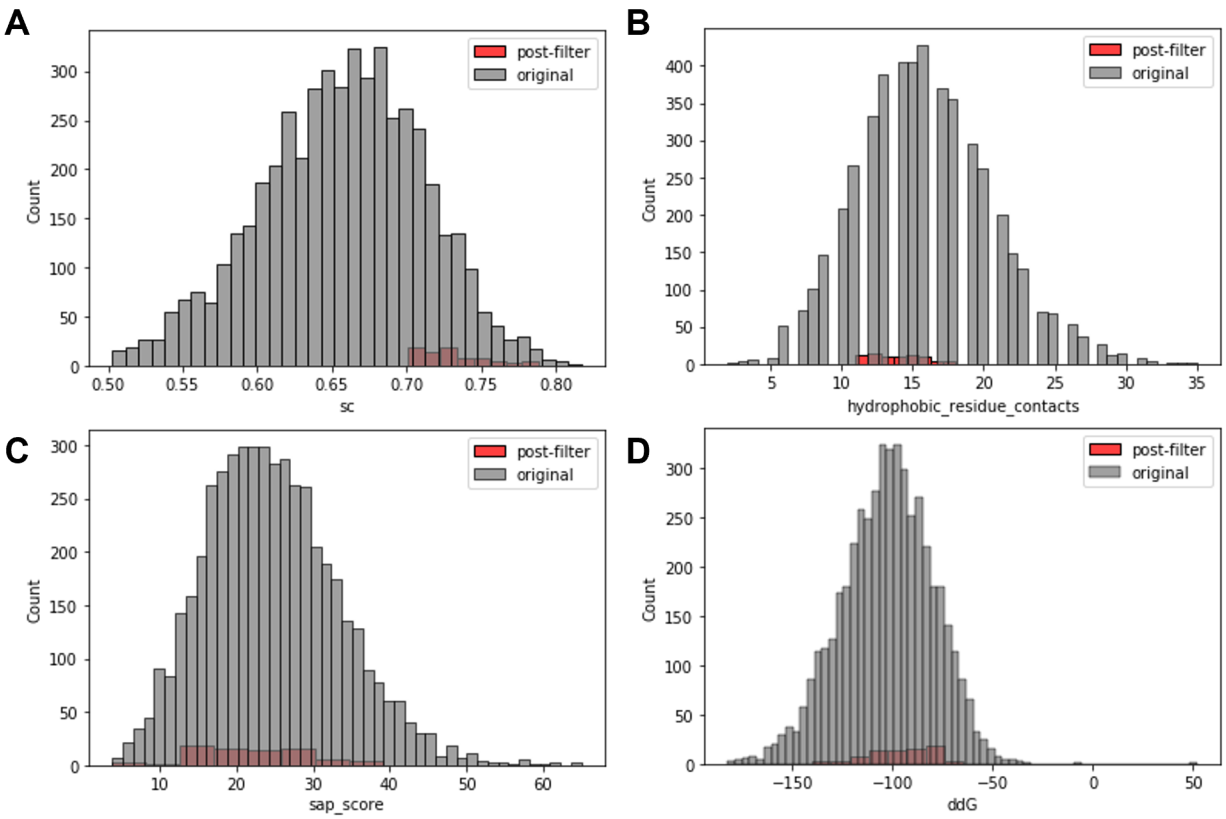

Fig. S13. Rosetta metrics for all designs (grey) as compared to selected designs for experimental characterization (red). **A**. Interface shape complementarity; **B**. Hydrophobic_residue_contacts at interface, **C**. Spatial aggregation propensity (SAP) score is a property of proteins that determines how aggregation prone they are; **D**. Interface ddG is a relative measure of Rosetta predicted binding energy for all interfaces.

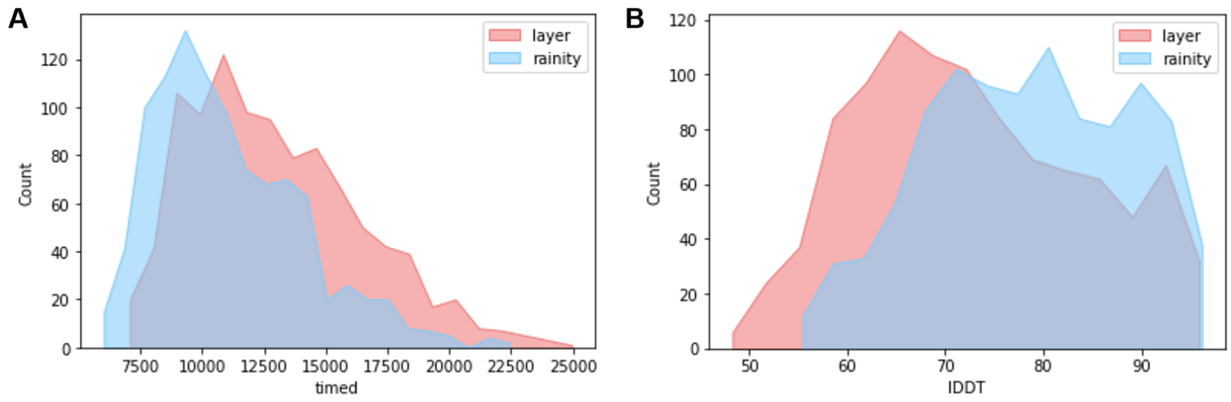

Fig. S14. A comparison of Rosetta based sequence design on 2000 scaffolds guided by conventional layer design (red) vs. pssm generated by protein BiRNN model (blue). **A**. The average computation time for a full cage design task is 13000 and 11000 cpu second for layer design and protein BiRNN respectively. **B**. The average designed monomer AF_pLDDT is 73 ± 12 and 79 ± 10 for layer design and protein BiRNN respectively. The percentile of monomers with AF_pLDDT > 80 is 30 % (layer design) vs. 45 % (protein BiRNN).

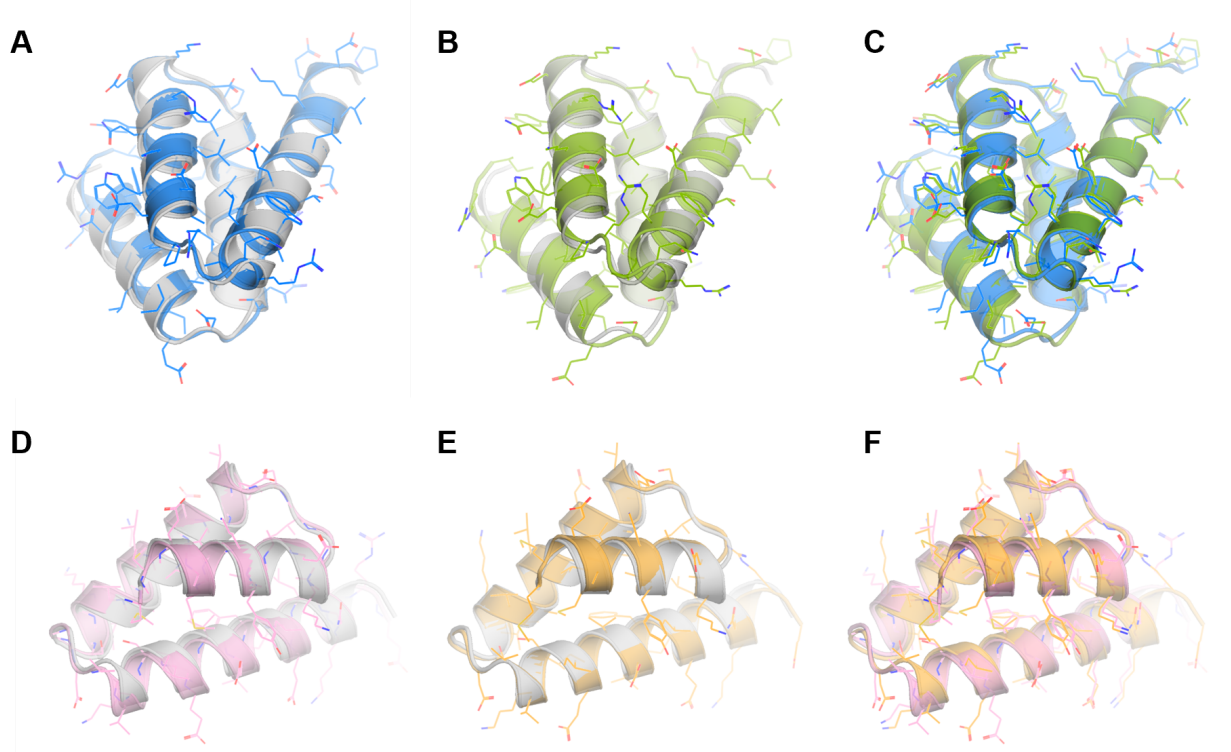

Fig. S15. Computational backbone optimization via Rosetta relaxation and AF prediction for RC_I_1(top row) and RC_I_2 (bottom row). **A,D**: Superposition of Rosetta relaxed models (RC_I_1: blue, RC_I_2: pink) and RL-polyA models (grey). **B,E**: Superposition of AF prediction models (RC_I_1: green, RC_I_2: orange) and RL-polyA models (grey). **C,F**: Superposition of AF prediction models (RC_I_1: green, RC_I_2: orange) and Rosetta relaxed models (RC_I_1: blue, RC_I_2: pink).

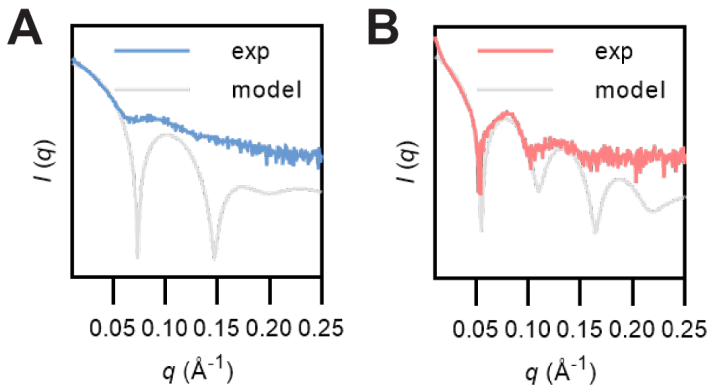

Fig. S16. 1D small-angle x-ray scattering patterns (SAXS, solid curves) of capsids RC_I_1 (**A**) and RC_I_2 (**B**) in buffer solution match well with profiles calculated from the design models (gray dashed curves). We note that there is a slight discrepancy between the model simulation and experimental SAXS patterns of the RC_I_1 sample, which we attribute to low overall signal intensity due to potential sample aggregation during shipment to the synchrotron beamline.

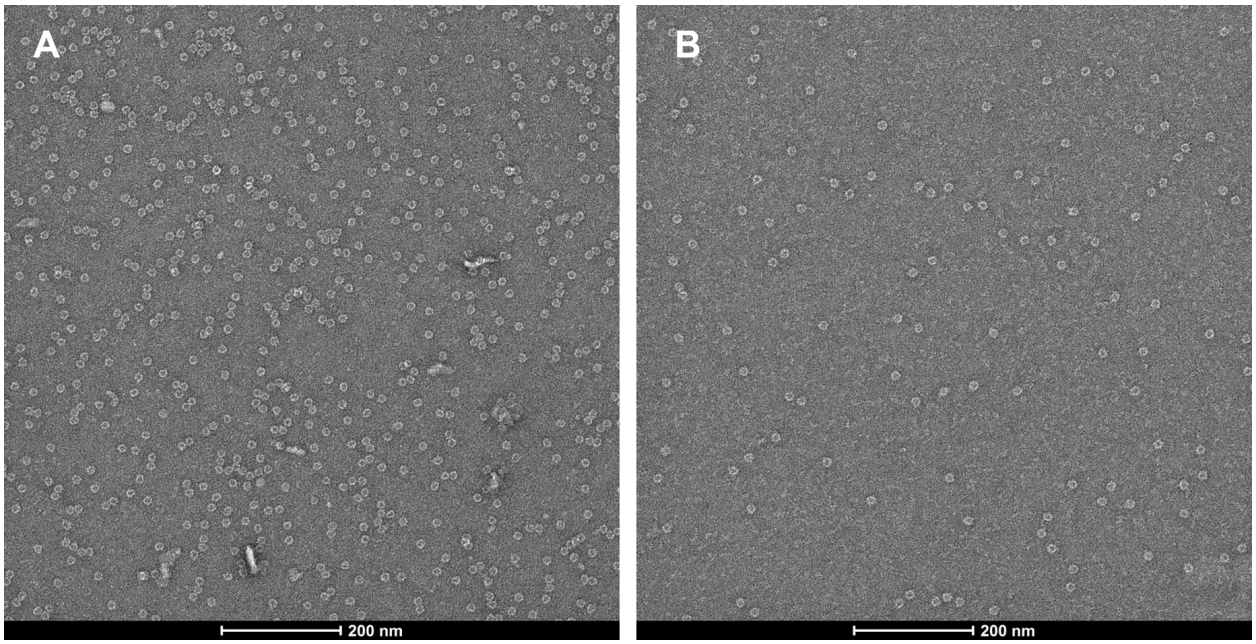

Fig. S17. Designed capsids show high thermal stability. Capsids RC_I_1-H11 (**A**) and its spycatcher fusion variants (**B**) were gradually heated up from room temperature to 95 ºC and incubated at 95 ºC for 1 hour before cooling back to room temperature. Capsids are observed to form mono-disperse assemblies post-thermal treatment as evidenced by nsEM.

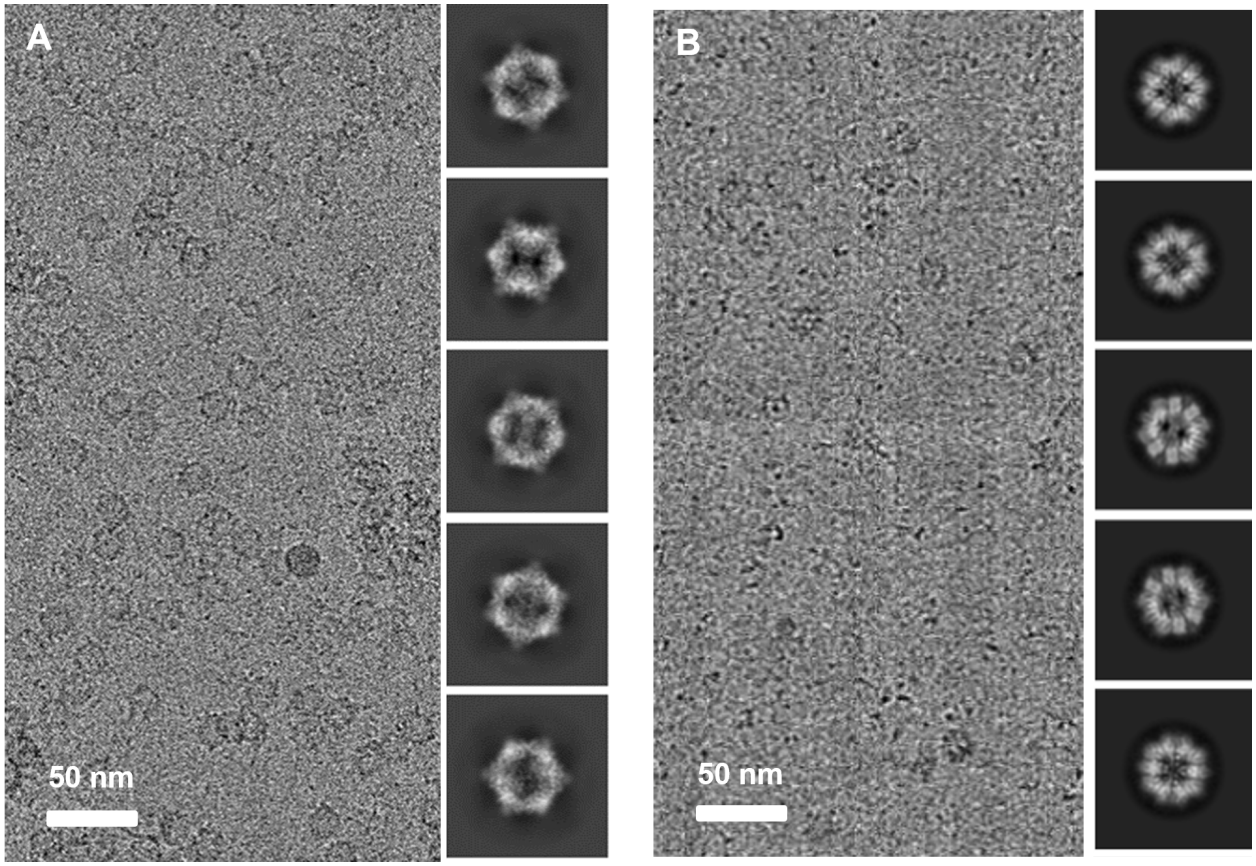

Fig. S18. Raw cryoEM micrographs and 2D class averages of capsids RC_I_1 (left) and RC_I_2 (right).

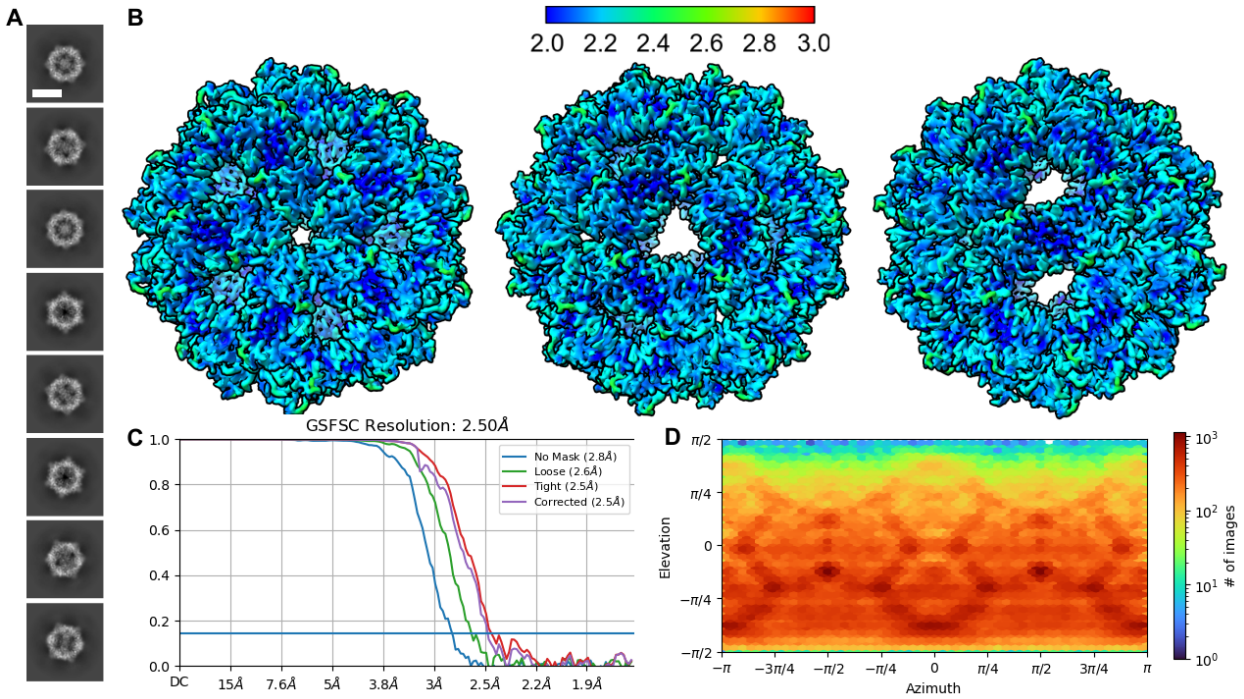

Fig. S19. CryoEM data and associated plots of RC_I_1. (**A**) 2D Class averages of RC_I_1 (Scale Bar = 10 nm). (**B**) CryoEM local resolution map of RC_I_1 viewed along three different angles. Local resolution estimates range from ~2.0 Å at the core to ~2.6 Å along the periphery. Regions of lower resolution (not shown here) correspond to attached 6× His Tag. (**C**) Global resolution estimation plot. (**D**) Orientational distribution plot demonstrating complete angular sampling.

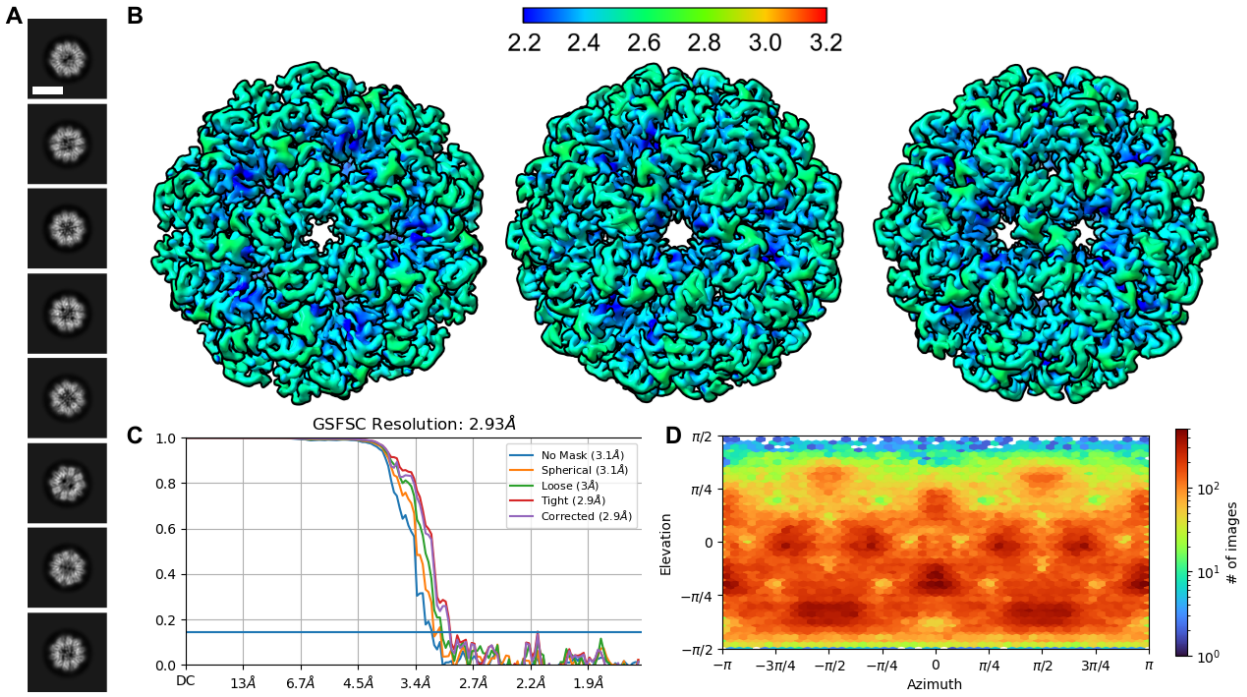

Fig. S20. CryoEM data and associated plots of RC_I_2. (**A**) 2D Class averages of RC_I_2 (Scale Bar = 10 nm). (**B**) CryoEM local resolution map of RC_I_2 calculated using an FSC value of 0.143 viewed along three different angles. Local resolution estimates range from ~2.2 Å at the core to ~2.6 Å along the periphery. Regions of lower resolution (not shown here) correspond to attached 6× His Tag. (**C**) Global resolution estimation plot. (**D**) Orientational distribution plot demonstrating complete angular sampling.

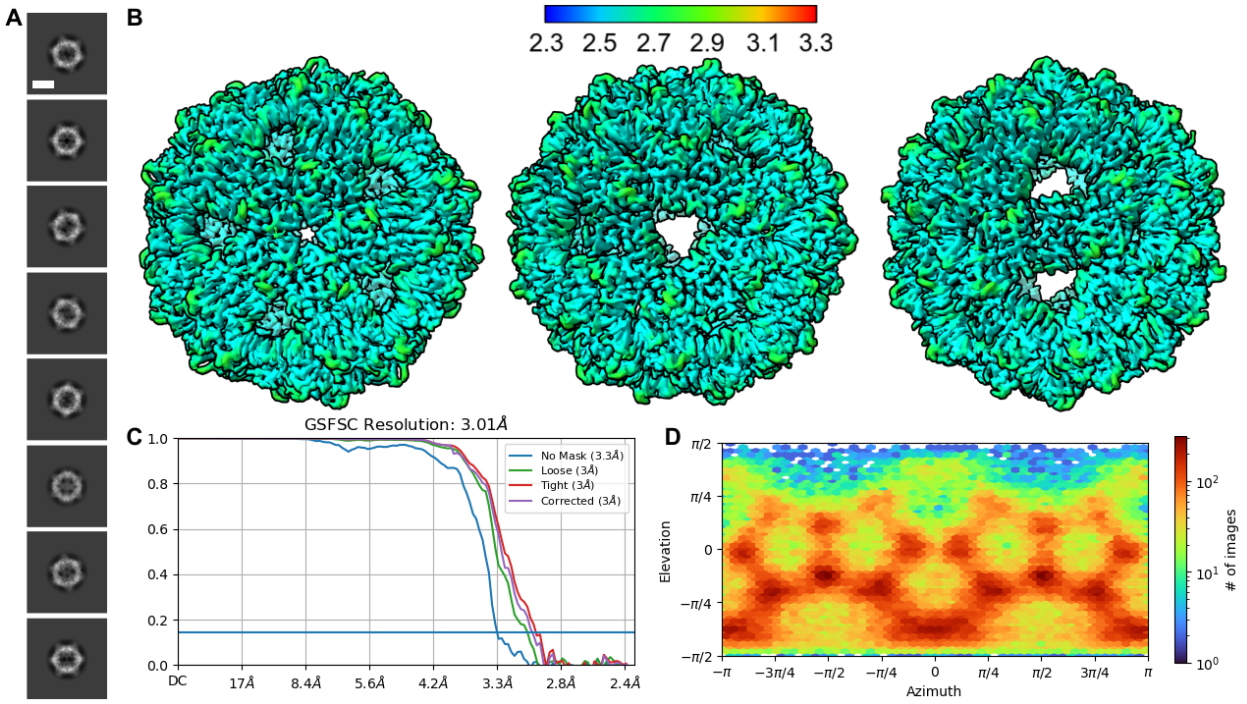

Fig. S21. CryoEM data and associated plots of RC_I_1-H11 (**A**) 2D Class averages of RC_I_1-H11 (Scale Bar = 10 nm). (**B**) CryoEM local resolution map of RC_I_1-H11 viewed along three different angles. Local resolution estimates range from ~2.3 Å at the core to ~2.6 Å along the periphery. Regions of lower resolution (not shown here) correspond to attached 6× His Tag. (**C**) Global resolution estimation plot. (**D**) Orientational distribution plot demonstrating near-complete angular sampling.

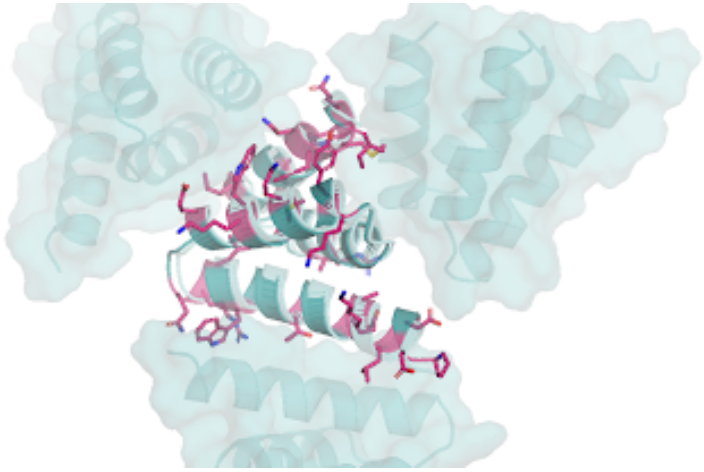

Fig.S22. The ProteinMPNN redesigned RC_I_1-H11 (purple) has a quite different sequence as compared to that of the parent design RC_I_1 (teal) made by Rosetta, including interface residues (side chains of RC_I_1-H11 are highlighted with stick representations).

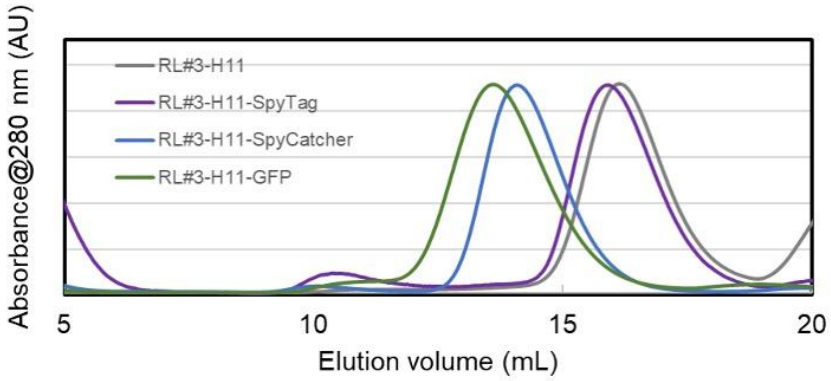

Fig. S23. SEC elution profiles for 4 different RC_I_1-H11 N-terminal fusions, from top to bottom: unmodified, spyTag, spyCatcher, and GFP, respectively.

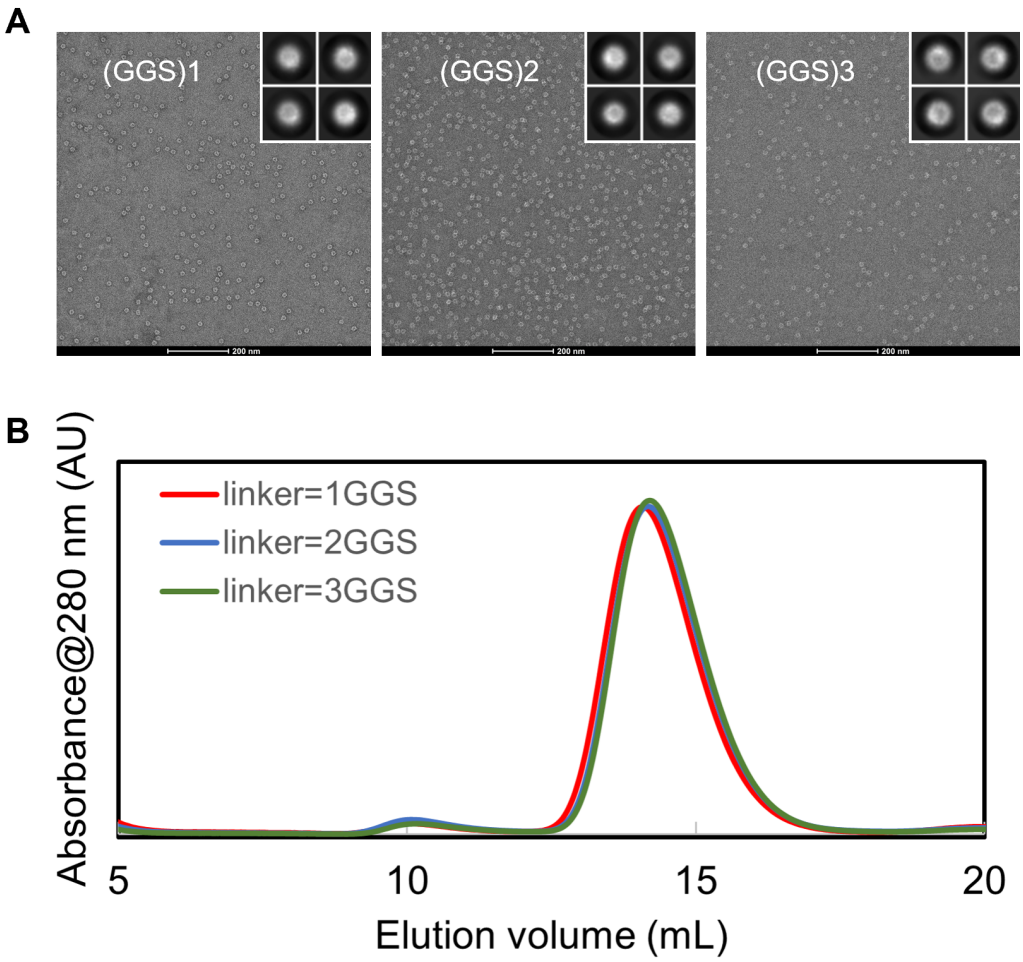

Fig. S24. nsEM characterization (inset: 2D class averages) and SEC elution profiles for RC_I_1-H11 spyCatcher fusions with different linker lengths.

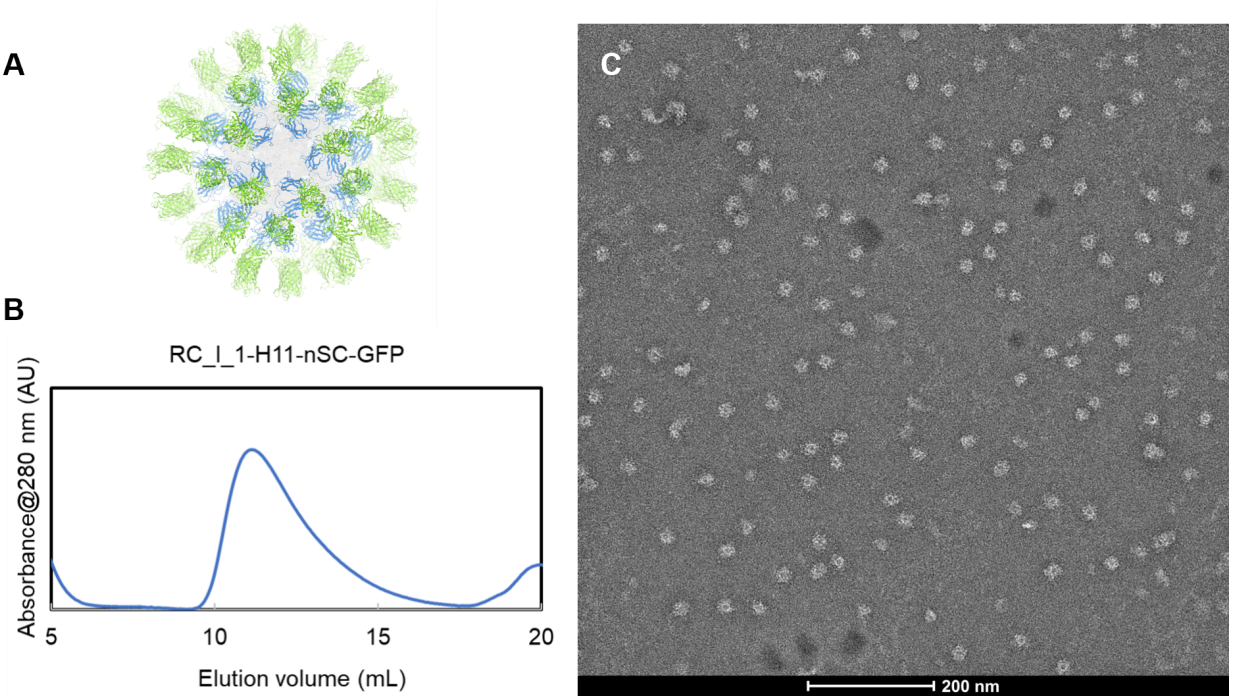

Fig. S25. A model of RC_I_1-H11 genetically fused with N-terminal GFP and SpyCatcher (**A**), which has been experimentally characterized by SEC elution profile (**B**) and nsEM (**C**).

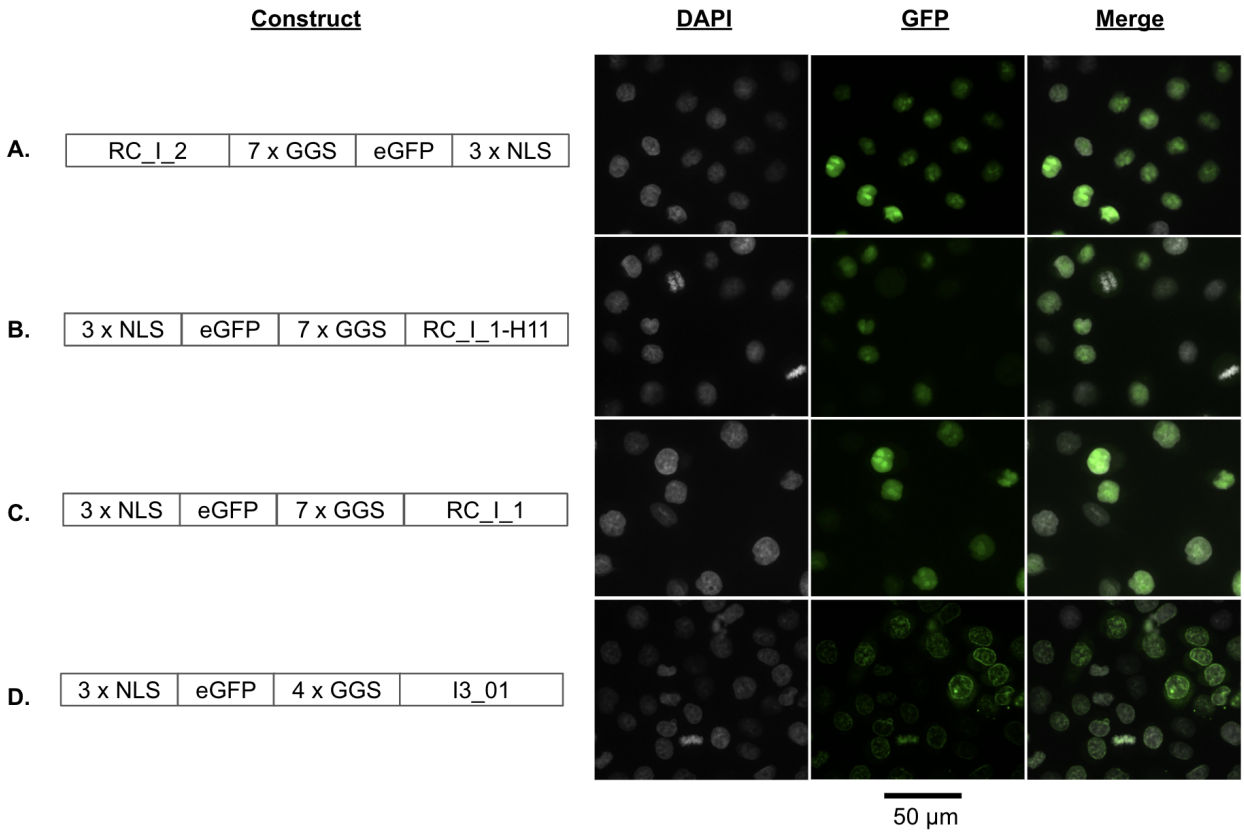

Fig. S26. Localization of NLS- and GFP-tagged constructs (green) within the nuclei of HeLa cells, 48h post-transfection. Nuclei are labeled in grey (DAPI). On the left, the components of each capsid construct (**A**, **B**, **C**, **D**) are depicted, while on the right, representative images from the DAPI and GFP channels of the corresponding constructs are shown. Images captured at 40× magnification. RC_I_1, RC_I_1-H11, and RC_I_2 have diameters of 10 to 13 nm, while I3_01 has a diameter of 28 nm.

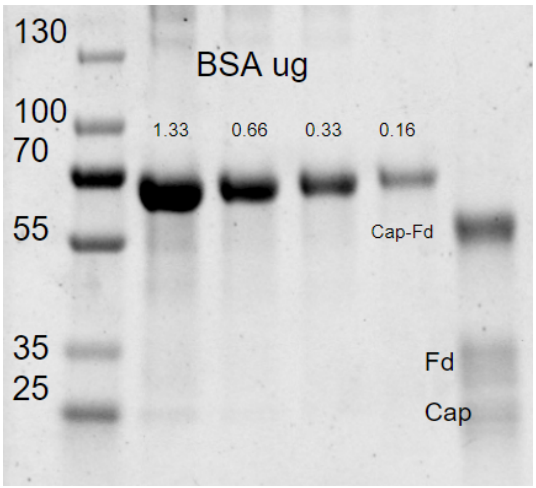

Fig. S27. SDS-Page gel confirmed successful RC_I_1-H11-Fd (Cap-Fd) conjugation with high efficiency.

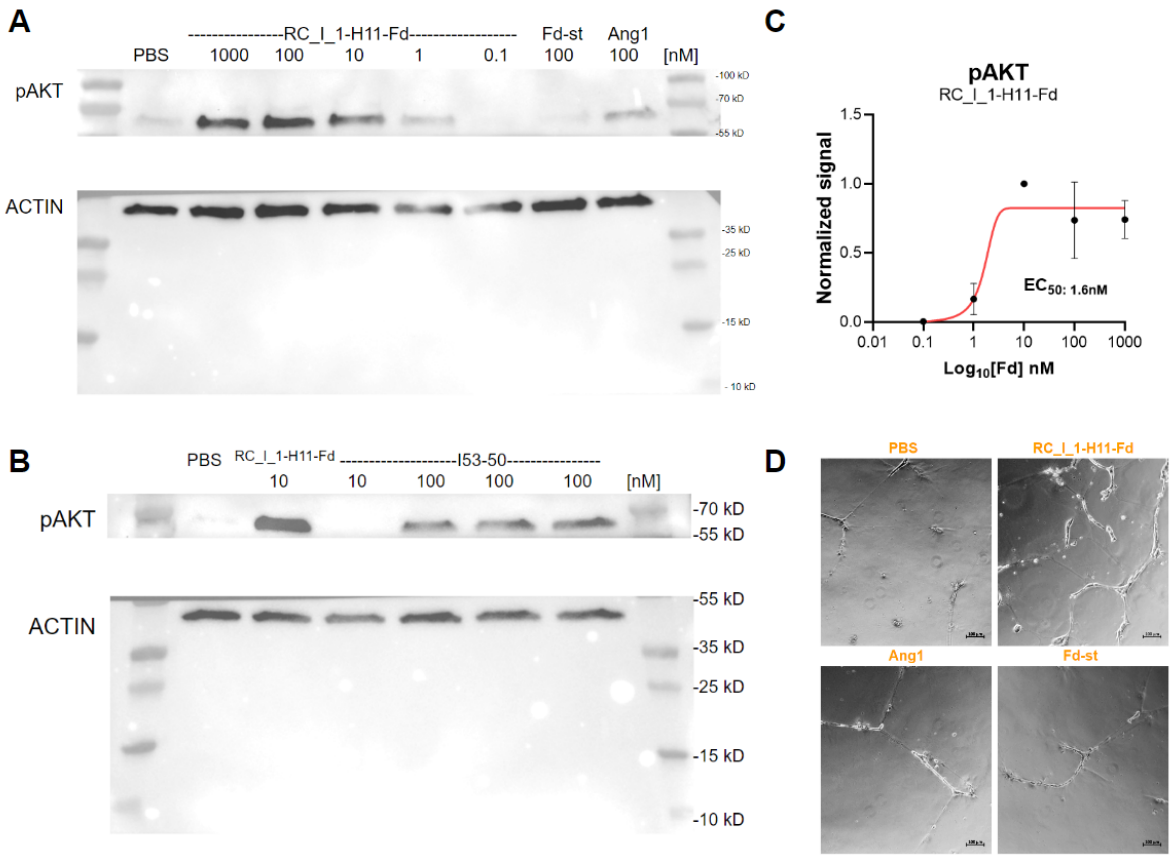

Fig. S28. Capisd-Fd constructs activate the Tie2 pathway. **A-B**) Representative western blot image for pAKT activation. Overlay of visible-light image of the membrane onto chemiluminescence signal from the bands. **C**) Quantifications of pAKT were fitted in titration curves using Prism to estimate EC50 : 1.6 nM. **D**) Representative images of the tube assay at 72-hour time point.

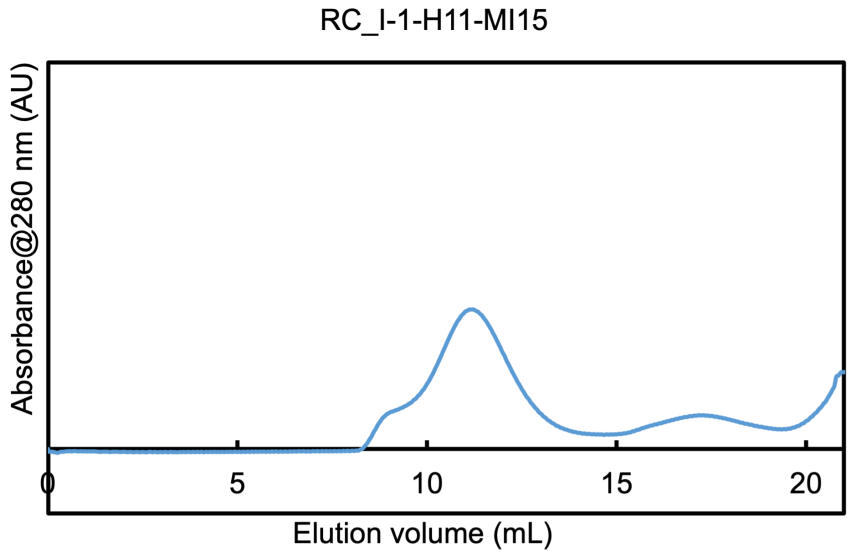

Fig. S29. SEC elution profile for RC_I_1-H11 hemagglutinin fusion with flexible linker.

Supplementary Tables

Table S1. Score function sigmoid activation weights.

score function	***m***	***a***	***b***
core formation	100	3	0.25
monomer designability	10	3	0.9
helix packing	1.1	0.5	2.1
porosity	100	1.8	0.61
interface designability	2	0.03	20
final	1	0.003	200

Table S2. Protein sequence information of nanopores.

Name	Sequence*
RN_C6_1	MGLGRETGDINAPFDPLLFAALRAMASSEDKANFEMIMKMFITSYATPETRPNVDTIL DYMALMTRLEGAEDTRAAFELDRLAGRALEEPDPERRERIREEREARLFRLLLRRAVR LQDEATLAAVLELLRTRDPAVAMLLEALALLAEGLKAALKAGDKELAERLEKEIERLA ELLLEFVEDLVRRGRLDTAAFALLLALALLAQLLKEALKAGDKELAERVEKEIERLAE ELKRVVEAAVEAGRLDLAADVLLEALELLAQLLKEALKAGDKELAKRVEKEIKRLTKE LLELAEKSLQEGLRLMAKYVESGGKEKELYEEAERLVDLAGDVLLETLELLNQLLEAA AKAGDKELAERIEKEIERLLEKLRVVALAKRAIEIAKEITEKNLDKEFAKLVAELLRA AAENPDLEAVKVARLALEIALQQPNSELSKEALKLAIRAINSDDELAKKVAALALEIA VEQPGSELSKEALKLAEEAIETDDEKAKKKALEALELALEQPGSEESKEALKKAKEEV EKAL
RN_C6_2	MGLNRETGDPNAPWDPLLFAALSIMRRSHDRRDFELVLQMFINNYGTPETPLNVDAIL DWMELLAADADTPAEDTRFAARLNRLSGAALRERDPARREALRRRRDELLLLRMLRHA VERDDDGALAAVLELLATRDPAVTAVLEALKQLAKILKEALKAGDKEIAKEILKEIEE LAKLLLELVRSLMRRGDLHTAALALLQAQAQLAKLAKEALKAGDKELAKEILKEIEEL AKELLEVVRAAVEAGDLELAADALLQAQAQLAKLAKEALKAGDKELAEEILKEIEELV KELLELARESRRRGLELMAEYVRSGGRDEDLLRRARDLLDLAADALLQTLKQLGKLLK AA AKAGDKELAKEILKEIKELVKELRGVVVAQAAVALAEEITRSGLDPEFAELVAELLEA AARNPRLLALEVARAALEVALQRPNTEEARRALRLAIRAIASPDELAQEVALAALRIA IERPGTEEARRALRLAERAIETDDEEAQREALEALRLALERPGTEEAREALERAREEV ERAL
RN_C6_3	MGLNRDTGDPNARWDPLLRALLRIMKQSYKQEDFNAIMDFIRSNYGTPETPLNVDAIM MYMMLMMQENPAEDTRRAFRLDELAGAALDETDPAKRAELRRRMEEFIFFHMLEHAFR KDDEGSLAAVLELLRTESPAIATLLELIRILVKAAKEALKAGNEEIAKKILKEIKENN KLLEEFVKGEIRRGKLFTAARALLITIRLLVKLAKEALKAGNEEIAKEVLKEIKENNK ELEKVVKAAIAAGDLDLAARALLQTLRLLVKLAKEALKAGNEEIAKEVLKEIKENTKE LLDIARRSLQEGLRLGAAYVRAGGREEELWRRAERLVRLAAEALLQTLKLLAKLLGEA AKAGNEEIAKEVQKEIKKIVKELRVVALAQRALEIARAITAQGLDPEFARLVAELLEA AAANPDELAVRVALRALEIALQQPNSQQAKRALELAIRAIRSPDELAQRVALRALEIA IQQPNSQQAKRALELAERAIRTPDAAAQEAALAALELALQQPGSPEAQAALAAAEAAV EAAL

Table S3. Protein sequence information of capsids and their variants.

Name	Sequence*	Notes
RC_I_1	MPDEDLKAELAATEAIWLLRQGRPEEVWKLMQRLYEKGDP ALWAVLRALLRSGDEIAILIAWNFMQRILEHHHHHH	First experimentally verified capsid from RosettaDesign
RC_I_1-H9	MMEEELRARVAATRALFLLEQGRPDEVVRLLEELLERGDP AIWDVLRALLESGDPVGKLIAEYFSRRLLEHHHHHH	ProteinMPNN redesign #1 of RC_I_1 backbone
RC_I_1-H11	MMEEERRRHLAAAEARFLLELGRPDEVLRLLERLLEEGDP ALFAALRELLESGDPLARLIAETVFRRLGSWGSLEHHHHH H	ProteinMPNN redesign #2 of RC_I_1 backbone
RC_I_2	MMMEAMVKYLAEKAGISEVEAAEIVLKAVKISGGDVVKSI ELVDLFIEILNKGREGSWGSLEHHHHHH	ProteinMPNN redesign of RC_I_2 backbone
RC_I_2-original	MVEESMVRYLSKHAGVSEDEAAKLVLKAVRISGGDVVKSI ELVDLFIEVINRGREGSWSGLEHHHHHH	Rosetta design, low expression

*His-tag, Trp for A280 absorbance, and N terminal Met are included in all ordered sequences

Table S4. Protein sequence information of capsid-fusions.

Name	Sequence*	Notes
RC_I_1-nHA	MKAILVVLLYTFTTANADTLCIGYHANNSTDTVDTVLEK NVTVTHSVNLLEDKHNGKLCKLRGVAPLHLGKCNIAGWI LGNPECESLSTASSWSYIVETSNSDNGTCFPGDFINYEE LREQLSSVSSFERFEIFPKTSSWPNHDSNKGVTAACPHA GAKSFYKNLIWLVKKGNSYPKLNQSYINDKGKEVLVLWG IHHPSTTADQQSLYQNADAYVFVGTSRYSKKFKPEIATR PKVRDQEGRMNYYWTLVEPGDKITFEATGNLVVPRYAFT MERNAGSGIIISDTPVHDCNTTCQTPEGAINTSLPFQNI HPITIGKCPKYVKSTKLRLATGLRNVPSIQSRGLFGAIA GFIEGGWTGMVDGWYGYHWQNEQGSGYAADLKSTQNAID KITNIVNSVIEKMNTQFTAVGKEFNHLEKRIENLNKKVD DGFLDIWTYNAELLVLLINERTLDYHDSNVKNLYEKVRN QLKNNAKEIGNGCFEFYHKCDNTCMESVKNGTYDYPKYS EEAKLNREKIDGVGSGSGSGSGSGSGSPDEDLKAELAAT EAIWLLRQGRPEEVWKLMQRLYEKGDPALWAVLRALLRS GDEIAILIAWNFMQRI	Hemagglutinin fused to n-terminus of the RC_I-1 (Three linker lengths were ordered: GS=10, GS=12, and GS =14) The GS = 12 construct was used in the mouse study.
RC_I_1-H11-nGFP	MRGHHHHHHGSSMRKGEELFTGVVPILVELDGDVNGHKF SVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLVTTLTY GVQCFARYPDHMKQHDFFKSAMPEGYVQERTISFKDDGT YKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYN FNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADHYQ QNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLE FVTAAGITHGMDELYKGGSGGSGGSGGSGGSGGSGGSME EERRRHLAAAEARFLLELGRPDEVLRLLERLLEEGDPAL FAALRELLESGDPLARLIAETVFRRLGSWSG	N-terminal fusion of GFP
RC_I_1-H11-nSpyCatcher	MGAMVDTLSGLSSEQGQSGDMTIEEDSATHIKFSKRDED GKELAGATMELRDSSGKTISTWISDGQVKDFYLYPGKYT FVETAAPDGYEVATAITFTVNEQGQVTVNGKATKGDAHI GGSGGSGGSMEEERRRHLAAAEARFLLELGRPDEVLRLL ERLLEEGDPALFAALRELLESGDPLARLIAETVFRRLGS WSGLEHHHHHH	N-terminal fusion of spycatcher
RC_I_1-H11-nSpyTag	MAHIVMVDAYKPTKGGSGGSMEEERRRHLAAAEARFLLE LGRPDEVLRLLERLLEEGDPALFAALRELLESGDPLARL	N-terminal fusion of spytag

	IAETVFRRLGSWSGLEHHHHHH	

Table S5. Cryo-EM data collection, refinement, and validation statistics

	#1 RNR_C6_3 (n/a) (EMD-29939)	#2 RC_I_1 (8F54) (EMD-28860)	#3 RC_I_2 (8F53) (EMD-28859)	#4 RC_I_1-H11 (8F4X) (EMD-28858)
Data collection and Processing (for each dataset):				
Microscope	FEI Glacios	FEI Titan Krios	FEI Titan Krios	FEI Glacios
Voltage (keV)	200	300	300	200
Camera	Gatan K3	Gatan K3	Gatan K3	Gatan K2
Magnification	36000	105000	105000	36000
Pixel size at detector (Å/pixel)	0.88	0.84	0.84	1.16
Total electron exposure ($e^-/Å^2$)	65	61	62	59
Exposure rate (e^-/pixel/sec)	8.5	15	15	8
Number of frames collected during exposure	50	100	100	50
Defocus range (μm)	-0.8 to -2.0	-0.7 to -2.0	-0.7 to -2.0	-0.7 to -2.0
Automation software (EPU, SerialEM or manual)	SerialEM	Leginon	Leginon	SerialEM
Micrographs collected (no.)	3,965	3,888	12,825	1,315
Micrographs used (no.)	1,909			
Total extracted particles (no.)	109,254	805,450	769,399	491,297

Table S5. Cryo-EM data collection, refinement, and validation statistics (continued)

	#1 RNR_C6_3 (n/a) (EMD-29939)	#2 RC_I_1 (8F54) (EMD-28860)	#3 RC_I_2 (8F53) (EMD-28859)	#4 RC_I_1-H11 (8F4X) (EMD-28858)
For each reconstruction:				
Final particles (no.)	19,418	678,727	325,728	153,765
Point-group or helical symmetry parameters	C6	I	I	I
Estimated error of translations/ rotations (if available)				
Resolution (global, Å)	5.13 Å	2.50 Å	2.93 Å	3.01 Å
- FSC 0.5 (unmasked/ masked)	7.32/5.40	3.10/2.71	3.38/3.13	3.47/3.29
- FSC 0.143 (unmasked/ masked)	5.85/5.13	2.82/2.50	3.12/2.93	3.33/3.01
Resolution range (local, Å)	3.6-5.9	2.0-3.0	2.2-3.2	2.3-3.3
Resolution range due to anisotropy (Å)				
Map sharpening *B* factor ($Å^2$) / (*B* factor range)	402.5	-124.6	-187.5	-131.1
Map sharpening methods	NU-refine Homo-refine Local refine	Auto	Auto	Auto

Table S5. Cryo-EM data collection, refinement, and validation statistics (continued)

	#1 RNR_C6_3 (n/a) (EMD-29939)	#2 RC_I_1 (8F54) (EMD-28860)	#3 RC_I_2 (8F53) (EMD-28859)	#4 RC_I_1-H11 (8F4X) (EMD-28858)
Model composition (for each model):				
Protein		66,960	51,060	32,760
Model Refinement (for each model):				
Refinement package - real or reciprocal space - resolution cutoff Model-Map scores -CC		Coot/ISOLDE/ Rosetta/ Namdinator/ PHENIX (real space)	Coot/ISOLDE/ Rosetta/ Namdinator/ PHENIX (real space)	Coot/ISOLDE/ Rosetta/ Namdinator/ PHENIX (real space)
R.m.s. deviations from ideal values				
- Bond lengths (Å)		0.23	0.26	0.26
- Bond angles (°)		0.48	0.45	0.37
Validation (for each model):				
Clashscore		0	1	1
Poor rotamers (%)		0	0	0
C-beta deviations		0	0	0
Ramachandran plot				
- Favored (%)		99	99	99
- Outliers (%)		0	0	0

Acknowledgements

Authors: Isaac D. Lutz[1,2,3]†, Shunzhi Wang[1,2]†*, Christoffer Norn[1,2,4]†, Alexis Courbet[1,2,5], Andrew J. Borst[1,2] , Yan Ting Zhao[1,6,7], Annie Dosey[1,2], Longxing Cao[1,2,8], Jinwei Xu[1,2], Elizabeth M. Leaf[1,2], Catherine Treichel[1,2], Patrisia Litvicov[1,6], Zhe Li[1,2], Alexander D. Goodson[1,2], Paula Rivera-Sánchez[4], Ana-Maria Bratovianu[4], Minkyung Baek[1,2,9], Neil P. King[1,2], Hannele Ruohola-Baker[1,3,6,7], David Baker[1,2,3]*

Affiliations:
[1]Department of Biochemistry, University of Washington, Seattle, WA, USA.
[2]Institute for Protein Design, University of Washington, Seattle, WA, USA.
[3]Department of Bioengineering, University of Washington, Seattle, WA, USA.
[4]BioInnovation Institute, DK2200 Copenhagen N, Denmark.
[5]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.
[6]Institute for Stem Cell and Regenerative Medicine, University of Washington, Seattle, WA, USA.
[7]Oral Health Sciences, University of Washington, Seattle, WA, USA.
[8]Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou, Zhejiang, China.
[9]School of Biological Sciences, Seoul National University, Seoul, Republic of Korea.

*Corresponding author. Email: dabaker@uw.edu (D.B.); swang523@uw.edu (S.W.)
†These authors contributed equally to this work.

We thank F. Dimaio, J. Dauparas, B. Coventry, T. Huddy, N. Woodall, R. Kibler, J. Watson, and I. Haydon for help with computational design and discussion; R. Kibler, D. Feldman, L. Milles, and N. Ennist for help with the experiments; and S. Dickinson and J. Quipse for help in maintaining and operating the electron microscopes used.

Funding: This work was conducted at the Advanced Light Source (ALS), a national user facility operated by Lawrence Berkeley National Laboratory on behalf of the Department of Energy, Office of Basic Energy Sciences, through the Integrated Diffraction Analysis Technologies (IDAT) program, supported by DOE Office of Biological and Environmental Research. Additional support comes from the National Institutes of Health project ALS-ENABLE (NIH grant P30 GM124169) and a High-End Instrumentation Grant S10OD018483. This work was also supported by the National Institute on Aging grant (grant 1U19AG065156-01 to I.D.L. and D.B.); Amgen (S.W.), the Audacious Project at the Institute for Protein Design (A.J.B., Z.L., L.C., H.R.-B., Y.T.Z., D.B., and N.P.K.); the NIH/National Institute of Dental and Craniofacial Research (NIDCR) (grant T90 DE021984 to Y.T.Z); the National Institute of Allergy and Infectious Diseases (NIAID grant 1P01AI167966 to N.P.K.); Novo Nordisk (foundation grant

NNF170C0030446 to C.N.); the Open Philanthropy Project Universal Flu Vaccine and Improving Rosetta Design (A.D., N.P.K., and D.B.); a Microsoft gift (M.B.); and the Department of Defense Peer Reviewed Medical Research Program (award W81XWH-21-1-0006 to H.R.-B. and D.B.).

Author contributions: I.D.L. and S.W. contributed equally and the author order was chosen arbitrarily; citations on CVs, etc., will be adjusted accordingly. S.W., I.D.L., C.N., and D.B. conceptualized the research. I.D.L. developed the RL backbone generation method. S.W. and C.N. developed the sequence design pipeline. L.C. and M.B. developed the DL sequence profile prediction method. S.W. designed the original capsids and performed the screening, expression, and characterization experiments. A.C. and J.X. designed the nanopore ring connector and performed cryo-EM characterization. A.J.B. designed cryo-EM experiments and optimization of sample purification conditions, performed the initial cryo-EM screening experiments, and optimized the cryo-EM freezing conditions. A.J.B., S.W., and Z.L. prepared additional cryo-EM grids and collected cryo-EM data. A.J.B. processed the cryo-EM data and built and solved the structures for each designed capsid. S.W. and A.D. designed and characterized the fusion capsids. A.D., C.T., E.M.L., and N.P.K. designed and performed the immunization studies for the HA-capsid fusions. A.D.G. produced and characterized HA-capsid proteins. Y.T.Z., P.L., and H.R.-B. designed and performed the cell signaling assays. C.N., P.R.S., and A.M.B. designed and performed the nuclear localization experiments. All authors analyzed the data. D.B. supervised the research. S.W., I.D.L., and D.B. wrote the manuscript with the input from the other authors. All authors revised the manuscript.
Competing interests: D.B., S.W., I.D.L., C.N., A.D., N.P.K., and A.J.B. are inventors on a provisional patent application (63/383,700) submitted by the University of Washington for the design, composition, and applications of the protein assemblies described in this work. The remaining authors declare no competing interests.

Data and materials availability: All data are available in the main text or the supplementary materials. MCTS backbone sampling code and structural coordinates of the in silico–generated examples are available at https://github.com/idlutz/protein-backbone-MCTS and archived at Zenodo (*55*). Code for the sequence design and filtering of assemblies is available at https://files.ipd.uw.edu/pub/2023_RL_capsid_design/sequence_design_pipeline.tar and archived at Zenodo (*56*). For RC_I_1, RC_I_2, and RC_I_1-H11 capsid structures, coordinates are deposited in the Protein Data Bank with the accession codes 8F54, 8F53, and 8F4X; cryo-EM density maps are deposited in the Electron Microscopy Data Bank (EMDB) with the accession codes EMD-28860, EMD-28859, and EMD-28858.

Chapter 3

Helical peptide recognition using the full suite of protein design tools

Adapted from Vázquez Torres, S., Leung, P. J. Y., Lutz, I. D., Venkatesh, P., Watson, J. L., Hink, F., Huynh, H.-H., Yeh, A. H.-W., Juergens, D., Bennett, N. R., Hoofnagle, A. N., Huang, E., MacCoss, M. J., Expòsit, M., Lee, G. R., Levine, P. M., Li, X., Lamb, M., Korkmaz, E. N., Nivala, J., Stewart, L., Rogers, J. M., & Baker, D. (2022). De novo design of high-affinity protein binders to bioactive helical peptides. bioRxiv 2022.12.10.519862. doi:10.1101/2022.12.10.519862

Introduction

In the following work, we approached and accomplished helical peptide binding, a previously unsolved protein design problem, using a wide array of different protein design tools. Helical peptides present a new challenge as compared to previous de novo protein binder targets, as they lack rigidity and require concave binding surfaces. Initially, we used parametric design to sample custom groove scaffolds, and found a notable improvement over traditional minibinder approaches. We found that the MCTS approach described in the previous chapter was able to automatically sample groove-shaped scaffolds as well. However, initial hits from both methods did not have sufficient affinities for downstream applications. Further modification of the parametric groove scaffolds with deep learning tools led to higher affinity. Additionally, we found that generative deep learning methods sampled high-affinity binders closely resembling the human-designed parametric backbones. The following work describes the process of achieving helical peptide recognition, which ultimately requires methods that sample custom backbones designed specifically to bind to helical peptide shapes.

Abstract

Many peptide hormones form an alpha-helix upon binding their receptors (*1–4*), and sensitive detection methods for them could contribute to better clinical management. De novo protein design can now generate binders with high affinity and specificity to structured proteins (*5*, *6*). However, the design of interactions between proteins and short helical peptides is an unmet challenge. Here, we describe parametric generation and deep learning-based methods for designing proteins to address this challenge. We show that with the RFdiffusion generative model, picomolar affinity binders can be generated to helical peptide targets either by noising and then denoising lower affinity designs generated with other methods, or completely de novo starting from random noise distributions; to our knowledge these are the highest affinity designed binding proteins against any protein or small molecule target generated directly by computation without any experimental optimization. The RFdiffusion designs enable the enrichment of parathyroid hormone or other bioactive peptides in human plasma and subsequent detection by mass spectrometry, and bioluminescence-based protein biosensors. Capture reagents for bioactive helical peptides generated using the methods described here could aid in the improved diagnosis and therapeutic management of human diseases (*7*, *8*).

Main Text

Peptide hormones, such as parathyroid hormone (PTH), neuropeptide Y (NPY), glucagon (GCG), and secretin (SCT), which adopt alpha helical structures upon binding their receptors (*1–4*), play key roles in human biology and are well established biomarkers in clinical care and biomedical research (Fig. 1a). There is considerable interest in their sensitive and specific quantification, which currently relies on antibodies that require substantial resources to generate, can be difficult to produce with high affinity, and often have less-than-desirable stability and reproducibility (*5*). Furthermore, the loop-mediated interaction surfaces of antibodies are not particularly well suited to high specificity binding of extended helical peptides. Designed proteins can be readily produced with high yield and low cost in E. coli and have very high stability, but while there have been considerable advances in de novo protein design to generate binders for folded proteins (*5*, *6*), the design of proteins that bind helical peptides with high affinity and specificity remains an outstanding challenge. Design of peptide-binding proteins is challenging for two reasons. First, proteins designed to bind folded proteins, such as picomolar affinity hyper-stable 50-65 residue minibinders (*5*), have shapes suitable for binding rigid concave targets, but not for cradling extended peptides. Second, peptides have fewer residues to interact with, and are often partially or entirely unstructured in isolation (*9*); as a result, there can be an entropic cost of structuring the peptide into a specific conformation (*10*), which compromises the favorable free energy of association. Progress has been made in designing peptides that bind to extended beta strand structures (*11*) and polyproline II conformations conformations (*12*) using protein side chains to interact with the peptide backbone, but such interactions cannot be made with alpha helical peptides due to the extensive internal backbone - backbone hydrogen bonding.

Design of helical peptide binding scaffolds

We set out to develop general methods for designing proteins that bind peptides in helical conformations. To fully leverage recent advances in protein design, we explored both parametric and deep learning-based approaches. For parametric generation, we reasoned that helical bundle scaffolds with an open groove for a helical peptide could provide a general solution to the helical peptide binding problem: the extended interaction surface between the full length of the helical peptide target and the contacting helices on the designed scaffold could enable the design of high affinity and specificity binding (Fig. 1b). In parallel, we reasoned that deep learning methods, which do not pre-specify scaffold geometries, could permit the exploration of different potential solutions to helical peptide binding.

Parametric design of groove scaffolds

We began by exploring parametric methods for generating backbones with overall "groove" shapes. Using the Crick parameterization of alpha-helical coiled coils (*13*), we devised a method to sample scaffolds consisting of a three-helix groove supported by two buttressing helices (Fig. 1c, see Supplementary Materials). We assembled a library of these scaffolds sampling a range of supercoiling and helix-helix spacings to accommodate a variety of helical peptide targets (Supplementary Fig. S1). We then used this library to design binders to PTH, GCG, and NPY, and screened 12 designs for each target using a nanoBiT split luciferase binding assay. Many of the designs bound their targets (3/12, 4/12, and 8/12 to PTH, GCG, and NPY) but with only micromolar affinities (see Supplementary Materials). These results suggest that groove-shaped scaffolds can be designed to bind helical peptides, but also that design method improvement was necessary to achieve high-affinity binding.

While powerful for generating and sampling a large number of potential scaffolds, the parametric generation approach has the limitation of building only from ideal building blocks, in this case parametric alpha helices. Deep learning methods do not have these limitations, and we explored whether RoseTTAFold inpainting (RF*joint*) (*14*), a model that can jointly design protein sequences and structures, could be used to improve the modest affinities of our parametrically-designed PTH binders (Fig. 2a). We used RF inpainting to extend the binders (non-parametrically) to incorporate additional interactions with the target peptide to take advantage of the full potential binding interface of the peptide. Out of 192 designs tested, 44 showed binding against PTH in initial yeast display screening. Following SEC purification, the best binder was found to bind at 6.1 nM affinity to PTH. Binding was quite specific: very little binding was observed to PTH related peptide (PTHrp), a related peptide sequence with 34% sequence identity (Fig. 2A). Overall, the affinity of the starting PTH binders was improved by approximately three orders of magnitude, and the highest-affinity binder had 19% greater surface area contacting the target peptide. We used the same design strategy to generate higher affinity binders for NPY and GCG. Using weak parametric binders as a starting point, we extended their binding interfaces and generated a ~231 nM affinity binder for GCG and a 3.5 μM binder for NPY after screening 96 designs (Supplementary Fig. S2).

As an alternative to de novo parametric design of scaffolds that contain grooves, we explored the threading of helical peptides of interest onto already existing designed scaffolds with interfaces that make extensive interactions with helical peptides (Fig. 2b). We started from a library of scaffolds that contained single helices bound by pseudorepetitive helical scaffolds. We then threaded sequences of peptides of interest onto the bound single helix and filtered to maximize interfacial hydrophobic interactions of the target sequence to the binder scaffold. The binders were then redesigned in the presence of the threaded target sequence with ProteinMPNN (*15*) and the complex was predicted with AF2 (*16*) (with initial guess (*6*)) and filtered on AF2 and

Rosetta metrics. Initial screening using yeast surface display identified 4/66 binders, which were expressed in E. coli. Following size exclusion chromatography (SEC) purification of the monomer fraction, all 4 of the designs were found to bind with sub-micromolar affinity using fluorescence polarization (FP), with the highest-affinity design binding with an affinity of 2.7 nM for SCT. Binding specificity was assessed with FP by measuring affinity for GCG, a related hormone to which SCT shares a significant degree of sequence identity (44%) and conformational homology (*1*, *2*). We found that the tightest SCT binder was only 4 fold selective for SCT over GCG, which suggested additional design strategies might be necessary to increase the quality of the binding interface and to achieve high-specificity binding (Fig. 2b).

Designing peptide binders by hallucination

We next explored the use of deep learning hallucination methods to generate helical peptide binders completely de novo, with no pre-specification of the desired binder geometry (from peptide sequence alone) (Fig. 2c). Hallucination or "activation maximization" approaches start from a network that predicts protein structure from sequence and carry out an optimization in sequence space for sequences which fold to structures with desired properties. This approach has been used to generate novel monomers (*17*), functional-site scaffolds (*14*) and cyclic oligomers (*18*). Hallucination using AlphaFold2 (AF2) or RosettaFold has a number of attractive features for peptide binder design. First, neither the binder nor the peptide structure needs to be specified during the design process, enabling the design of binders to peptides in different conformations (this is useful given the unstructured nature of many peptides in solution; disordered peptides have been observed to bind in different conformations to different binding partners (*9*)). Second, metrics such as the predicted alignment error (pAE) have been demonstrated to correlate well with protein binding (*6*), permitting the direct optimization of the desired objective, albeit with the possible hazard of generating adversarial examples (*18*).

We began by designing binders to the apoptosis-related BH3 domain of Bid (Fig. 1a). The Bid peptide is unstructured in isolation, but adopts an alpha-helix upon binding to Bcl-2 family members (*19*, *20*); it is therefore a model candidate for the design of helix-binding proteins. Starting from only the Bid primary sequence, and a random seed binder sequence (of lengths 60, 70, 80, 90 or 100 residues), we iteratively optimized the sequence of the binder through a Monte Carlo search in sequence space, guided by a composite loss function including the AF2 confidence (pLDDT, pTM) in the complex structure, and in the interaction between peptide and target (pAE). The trajectories typically converged in 5000 steps (sequence substitutions; Supplementary Fig. S3), and the output binder sequence was subsequently redesigned with ProteinMPNN, as previously described (*18*). All designed binders were predicted to bind to Bid in a helical conformation; the exact conformations differ between designs because only the amino acid sequence of the target is specified in advance. This protocol effectively carries out

flexible backbone protein design, which can be a challenge for traditional Rosetta based design approaches for which deep conformational sampling can be very compute intensive. Interestingly, in line with our prediction that "groove" scaffolds would offer an ideal topology for helical peptide binding, many of the binders from this approach contained a well-defined "groove" by eye, with the peptide predicted to make extensive interactions with the binder, typically helix-helix interactions.

47 of the hallucinated designs were tested experimentally (Supplementary Fig. S4a). Initial screening was performed with co-expression of a GFP-tagged Bid peptide and the HIS-tagged binders, with coelution of GFP and binder used as a readout for binding. 4 of these designs were further characterized, and showed soluble, monomeric expression even in the absence of peptide co-expression (Supplementary Fig. S4b). All designed proteins could be pulled-down using Bid BH3 peptide immobilized on beads (Supplementary Fig. S4c). Circular dichroism experiments indicated that the Bid peptide was unstructured in solution, and that helicity increased upon interaction with the hallucinated proteins, in line with the design prediction (Supplementary Fig. S4d). The binders were highly thermostable, and, unlike the native Bcl-2 protein Mcl-1, readily refolded after (partial) thermal denaturation at 95 °C (Supplementary Fig. S4e). Isothermal titration calorimetry revealed that all four bound Bid peptide, with the highest-affinity design binding having an affinity of 25 nM (Fig. 2c), a higher affinity interaction than with the native partner Mcl-1 (Supplementary Fig. S4f).

Peptide binder design with RFdiffusion

We next explored the design of binders using the RoseTTAFold-based denoising diffusion model RFdiffusion described in the accompanying paper (Watson et al.). RFdiffusion is much more compute efficient than hallucination, and is trained to directly generate a diversity of solutions to specific design challenges starting from random 3D distributions of residues that are progressively denoised. We reasoned that RFdiffusion could be used both for binder optimization (by sampling related conformations around a specific binder structure) and for fully de novo design starting from a completely random noise distribution.

A long-standing challenge in protein design is to increase the activity of an input native protein or designed protein by exploring the space of plausible closely related conformations for those with predicted higher activity. This is difficult for traditional design methods as extensive full atom calculations are needed for each sample around a starting structure (using molecular dynamics simulation or Rosetta full atom relaxation methods), and it is not straightforward to optimize for higher binding affinity without detailed modeling of the binder-target sidechain interactions. We reasoned that, in contrast, RFdiffusion might be able to rapidly generate plausible backbones in the vicinity of a target structure, increasing the extent and quality of

interaction with the target guided by the extensive knowledge of protein structure inherent in RoseTTAfold. During the reverse diffusion (generative) process, RFdiffusion takes random Gaussian noise as input, and iteratively refines this to a novel protein structure over many ("T") steps (typically 200). Partly through this denoising process, the evolving structure no longer resembles "pure noise", instead resembling a "noisy" version of the final structure. We reasoned that ensembles of structure with varying extents of deviation from an input structure could be generated by partially noising to different extents (for example, timestep 70), and then denoising to a similar, but not identical final structure (Fig. 3a, b).

We experimented with this approach starting from our parametrically-designed inpainted binders to GCG (with 231 nM affinity) and NPY (with 3.5 μM affinity) (Supplementary Fig. S2). Following partial noising and denoising, we identified designs that in silico, had significantly improved AF2 metrics compared to the starting design. The diversity compared to the starting design could be readily tuned by varying the time point to which the starting design was noised (Fig. 3a). Initial screening on yeast display revealed quite high binding success rates, with 25/96 designs binding GCG, and 20/96 binding NPY at 10 nM peptide concentration. The highest affinity designs were expressed in E. coli, purified, and their binding affinities were determined using FP. The highest-affinity binders were found to bind at subnanomolar affinities to GCG, and 5.6 nM to NPY (Fig. 3c). The designed proteins are quite specific: the GCG binders bound 10 times less tightly to SCT, which was chosen due to its high similarity to GCG. Impressively, the NPY binder did not show any cross-reactivity to peptide YY (PYY), which is a member of the NPY/pancreatic polypeptide family (*21*) and shares a high percentage of sequence similarity (63.5% for the sequences used in the assay).

Inspired by this success at optimizing binders with RFdiffusion, we next tested its ability to design binders to a different BH3 peptide, Bim and PTH completely de novo through unconditional binder design - providing RFdiffusion only with the sequence and structures of the two peptides in helical conformations, and leaving the topology of the binding protein and the binding mode completely unspecified (Fig. 4a). From this minimal starting information, RFdiffusion generated designs predicted by AF2 to fold and bind to the targets with high in silico success rates. A representative design trajectory is shown for PTH in Fig. 4b and Supplemental Video 1; starting from a random distribution of residues surrounding the PTH peptide in a helical conformation, in sequential denoising steps the residue shifts to surround the peptide and progressively organize itself into a folded structure which cradles the peptide along its entire surface.

We obtained synthetic genes encoding 96 designs for each target. Using yeast surface display, we found that 25 of the 96 designs bound to Bim at 10 nM peptide concentration. The highest affinity design, which purified as a soluble monomer, bound too tightly for steady state estimates of the dissociation constant (Kd); global fitting of the association and dissociation kinetics

suggest a Kd of ~100pM (Fig 4C). For PTH, we found that 56/96 of the designs bound by yeast surface display with sub-micromolar affinities. The highest affinity design again bound too tightly for accurate Kd estimation; instead FP data provides an approximate upper bound for the Kd<500 pM (Fig. 4c). Binding was also highly specific; no binding was observed to the related PTHrp (Fig. 4c). Circular dichroism temperature melts indicate that both binders are stable at 95°C (Fig 4C). The diffused from scratch binders again had considerable structural similarity to our starting groove binding concept.

Origins of higher affinity binding

The RFdiffusion scaffolds bind the peptides with extended helices in a manner not entirely different from our starting groove structures and the other designs described above. What is the origin of their higher affinity? Reasoning that de novo building of the designs in the presence of the target, rather than starting from pre-generated scaffolds, could increase the extent of shape matching between binder and target, we computed the contact molecular surface (*5*) for all of our designs in complex with the peptides. The average contact molecular surface for the partially diffused GCG binders and NPY increased by 33% and 29% respectively compared to the starting models, and the Rosetta ddG improved by 29% and 21% (Fig. S5a, S5b).

Comparison of solutions to the binding problem

Our results provide an interesting side by side comparison of human and machine based problem solving. Despite the differences in affinity, the deep learning methods typically came up with the same overall solution to the helical peptide binding design problem – groove shaped scaffolds with helices lining the binding site – as the human designers did in the first Rosetta parametric approaches. The increased affinity likely derives at least in part from higher shape complementarity resulting from direct building of the scaffold to match the peptide shape; the ability of RFdiffusion to “build to fit" provides a general route to creating high shape complementary binders to a wide range of target structures.

Design of protein biosensors for PTH detection

Given our success in generating de novo binders to clinically-relevant helical peptides, we next sought to test their use as detection tools for use in diagnostic assays. Compared to immunosensors, which often exhibit antibody denaturation, loss of conformational stability, and

wrong positioning of the antigen-binding site during sensor immobilization, de novo protein-based biosensors offer a more robust platform with high stability and tunability for diagnostics (*22*, *23*). To design PTH biosensors, we grafted the 6.1 nM PTH binder into the lucCage system (*24*), screened 8 designs for their luminescence response in the presence of PTH, and identified a sensitive lucCagePTH biosensor (LOD = 10 nM) with ~21-fold luminescence activation in the presence of PTH (Fig. 5a).

Enriching peptide targets from a complex mixture

We explored the use of our picomolar affinity RFdiffusion generated binder to PTH as a capture reagent in immunoaffinity enrichment coupled with liquid chromatography-tandem mass spectrometry (LC-MS/MS), a powerful platform for detecting low-abundance protein biomarkers in human serum (*25*). We evaluated the RFdiffusion binder in an LC-MS/MS assay for PTH in serum. PTH enrichment was quantified based on the analysis of the N-terminal peptide of a tryptic digestion of PTH in human plasma (*26–28*). (see Supplemental Materials). We found that the designed binder enabled capture of PTH from spiked buffer and spiked human plasma with recoveries of 53% and 43%, respectively (Fig. 5b). The very high thermal stability of the designed binders (Fig. 4c,d) suggests that bioactive peptide capture reagents could have much longer shelf lives than antibodies, and be amenable to harsher washing conditions enabling re-use of binder conjugated beads.

Discussion

Antibodies have served as the industry standard for affinity reagents for many years, but their use is often hampered by variable specificity and stability (*29*, *30*). For binding helical peptides, the computationally designed helical scaffolds described in this paper have a number of structural and biochemical advantages. First, the extensive burial of the full length of an extended helix is difficult to accomplish with antibody loops, but very natural with matching extended alpha helices in groove shape scaffolds. Second, designed scaffolds are more amenable to incorporation into sensors as illustrated by the LucCage PTH sensor. Third, they are more stable, can be produced much less expensively, and could be more easily incorporated into affinity matrices for enrichment of peptide hormones from human serum. Fourth, peptide binders can achieve high affinity and specificity purely through computational methods, eliminating the need to use animals, which often mount weak responses to highly conserved bioactive molecules. Our MS based detection of peptides present at very low abundance in sera following enrichment using the designed binders could provide a general route forward for serological detection of a wide range of disease associated peptide biomarkers.

Our results highlight the emergence of powerful new deep learning methods for protein design. The inpainting and RFdiffusion methods were both able to improve on initial Rosetta designs, and the hallucination approach generated high affinity binders without requiring prespecification of the bound structures. Most impressively, the RFdiffusion method rapidly generated very high (picomolar) affinity and specific binders to multiple helical peptides. As described in the accompanying manuscript (Watson et al.), RFdiffusion is able to design binders to folded targets; here we demonstrate further that RFdiffusion can be used to improve starting designs by partial noising and denoising, and can generate binders to peptides starting from no information other than the target. To our knowledge, the Bim and PTH binding proteins diffused starting from random noise are the highest affinity binders to any target (protein, peptide, or small molecule) achieved directly by computational design with no experimental optimization. We expect both the de novo peptide binder design capability and the ability to resample around initial designs (before or after experimental characterization) to be broadly applicable.

Figures

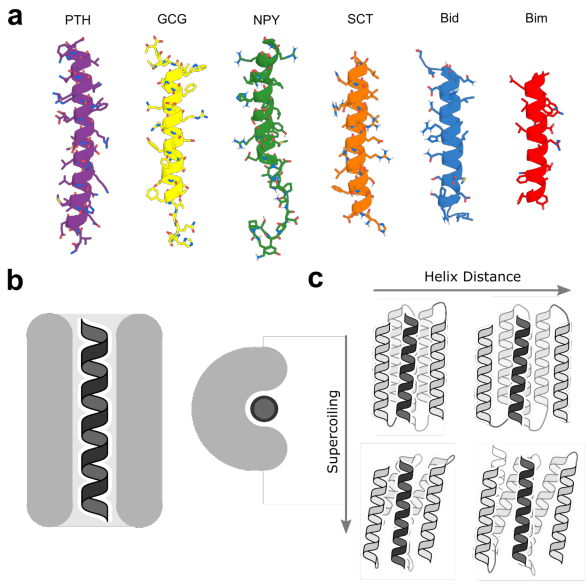

Figure 1. Binding helical peptides in groove scaffolds. (a) Helical peptide targets: parathyroid hormone (PTH), glucagon (GCG), neuropeptide Y (NPY), secretin (SCT), and the apoptosis-related BH3 domains of Bid and Bim. **(b)** "Open groove" structural solution to the helix binding problem. **(c)** Parametric approach to sampling of groove scaffolds varying supercoiling and helix distance to fit different targets.

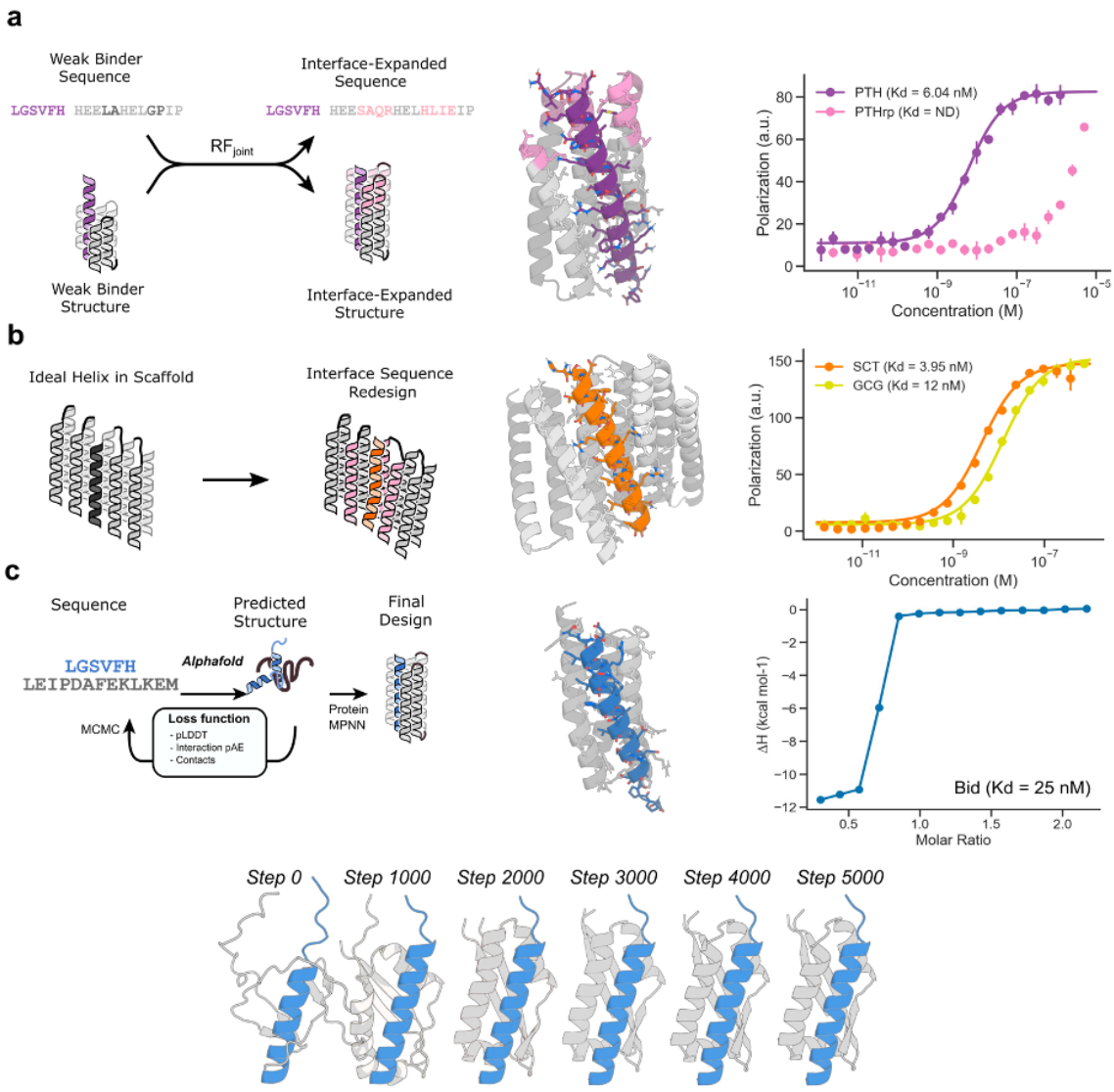

Figure 2. Design strategies for binding helical peptides. (a) Inpainting binder optimization: redesign of parametrically generated binder designs using RF*joint* inpainting to expand the binding interface. Left: schematic illustration of approach. Middle: original parametric scaffold (gray), inpainted design with extended interface (pink), and PTH target (purple). Right: Fluorescence polarization measurements with TAMRA-labeled targets indicate 6.1 nM binding to PTH and only weak binding to off-target PTH related peptide (PTHrp). **(b)** Thread target sequence and redesign: threading peptides onto pseudorepetitive protein scaffolds. Left: schematic illustration. Right: Design model of SCT based on repeat protein scaffold (grey) and SCT target (orange). Fluorescence polarization measurements with TAMRA-labeled targets indicate 3.95 nM binding to SCT and 12 nM binding to GCG. **(c)** Binder design with deep network hallucination. Top left: schematic illustration. Right, designed binder resulting from Monte Carlo optimization of binder sequence using AlphaFold over 5000 steps, with only target sequence (not structure) provided. Hallucinated binder (gray); target Bid peptide (blue). Isothermal titration calorimetry measurements (far right) indicate 25 nM binding to Bid. Bottom: hallucination trajectory starting from

random sequence (left) to final sequence (right); the protein folds around the peptide, which increases in helical content from step 0 to step 1000.

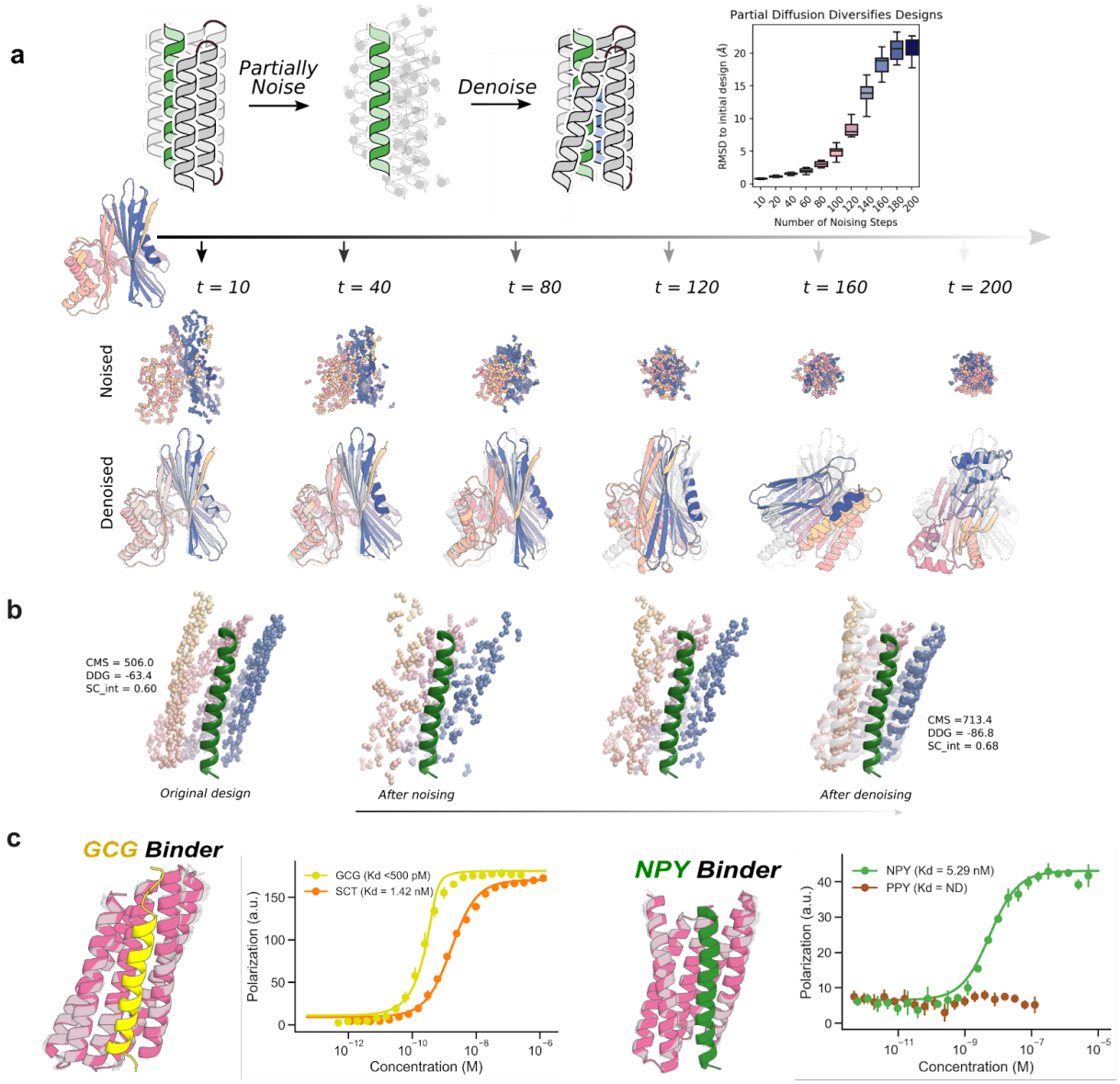

Figure 3. Peptide binder optimization with RF*diffusion*. (a) Top: Schematic showing partial noising and denoising using RF*diffusion*. A starting monomer (left) is partially noised for an increasing number of steps and then denoised resulting in designs (color) increasingly different from the original design (gray). Varying the noising stage from which denoising trajectories are initiated enables control over the extent of introduced structural variation. Bottom left: The distribution of RMSD to initial design vs number of partial noising steps. Bottom right: Starting from initial helix binder designs, we use partial diffusion to design optimized binders with improved shape complementarity. **(b)** Partial denoising trajectory starting from an initial NPY binder shown on the left. The final design (color) is shown on the right overlaid over the original design (gray). Contact molecular surface (CMS), Rosetta DDG (DDG) and interface shape complementarity (sc_int) values are reported for the original and optimized binder. **(c)** Diffused binders to GCG and NPY. Top left: Design models (gray) and AF2 predictions (pink, metrics in Supplementary Table 1), of diffused binders to GCG (yellow). Top right: FP measurements with FAM-labeled GCG indicate a sub-nanomolar binding affinity and selectivity over SCT. Bottom left: Design models (gray)

and AF2 predictions (pink, metrics in Supplementary Table 1), of diffused binders to NPY (green). Bottom right: FP measurements with FAM-labeled NPY indicate a binding affinity of 5.29 nM and no binding to PYY, demonstrating selectivity.

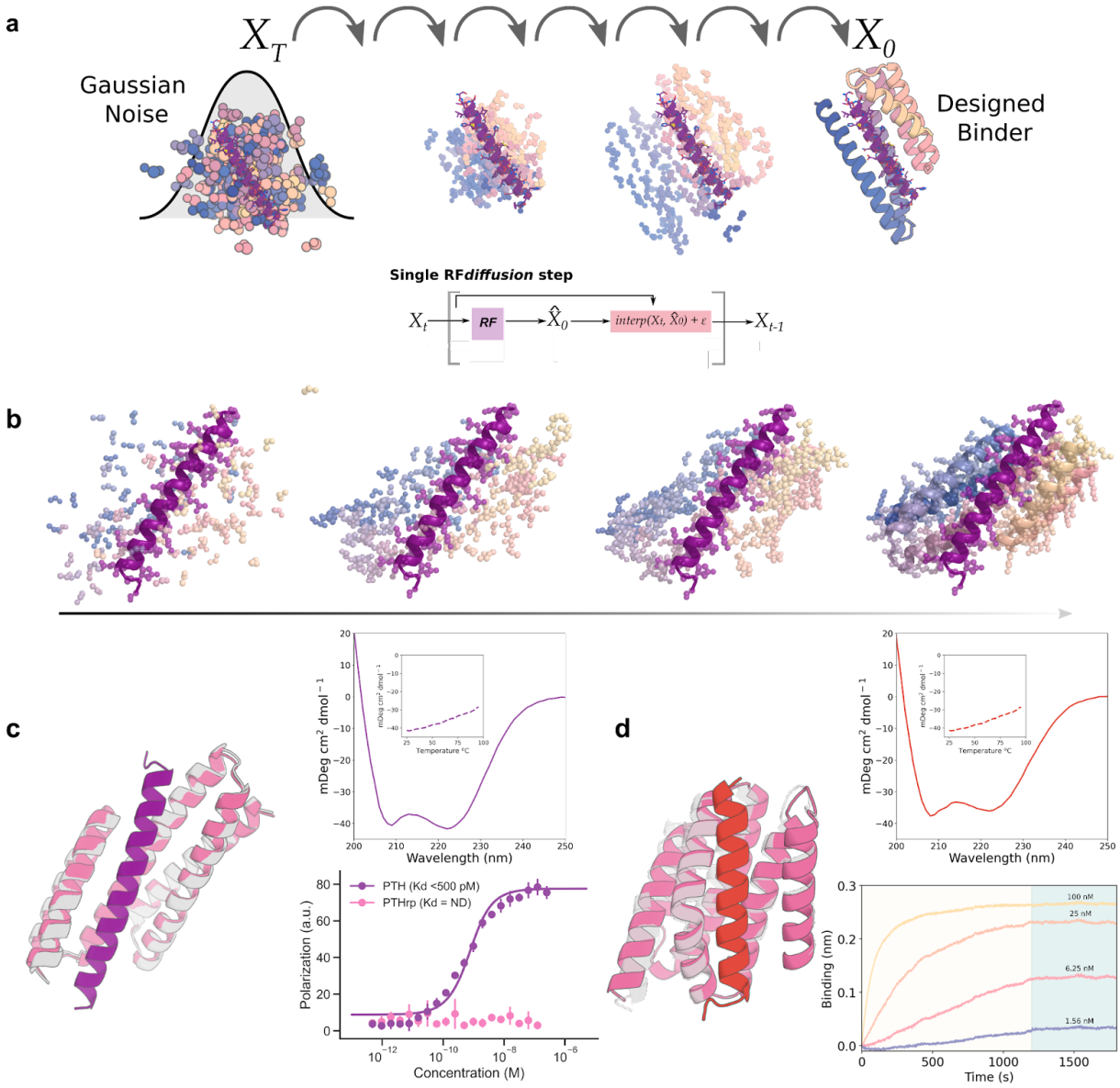

Figure 4. Peptide binder design with RF*diffusion*. (a) Schematic showing binder design using RFdiffusion. Starting from a random distribution of residues around the target peptide (X_T), successive RFdiffusion denoising steps progressively remove the noise leading by the end of the trajectory X_0 to a folded structure cradling the peptide. At each step t, RF*diffusion* predicts the final structure pX_0 given the current noise sample Xt, and a step that interpolates in this direction is taken to generate the input for the next denoising step Xt-1. **(b)** Denoising trajectory in the presence of PTH (purple, Supplementary Video 1). Starting from random noise (left), a folded structure starts to emerge, leading to the final designed binder (right). **(c)** Design of picomolar affinity PTH binders. Left: Design model (gray) and AF2 prediction (pink, metrics in Supplementary Table 1), of designed PTH binder (purple). Bottom right: Fluorescence polarization measurements with TAMRA-labeled PTH indicate a sub-nanomolar binding affinity and no binding for PTH related peptide, indicating high specificity (PTHrp). Top right: Circular dichroism data indicating that the binder has the designed helical secondary structure and does not undergo cooperative unfolding below 95°C (inset). **(d)**: Design of picomolar affinity Bim binders. Left:

Design model (gray) and AF2 prediction (pink, metrics in Supplementary Table 1), of designed Bim binder (red). Right bottom: Biolayer interferometry measurement of Bim binding indicates a sub-nanomolar affinity, with very slow dissociation kinetics. Biotinylated Bim was coupled to an Octet sensor, and incubated with the indicated concentrations of binder. The off rate is too slow to be accurately measured. Right top: CD data shows that the binder has helical secondary structure and is stable at 95°C (inset).

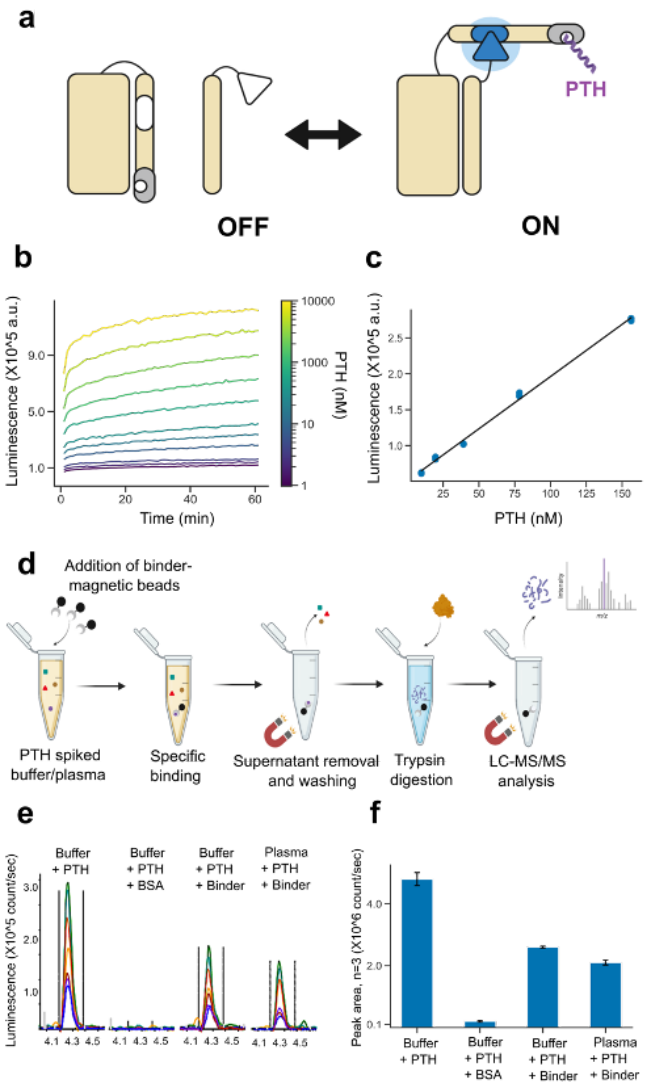

Figure 5. Application of designed binders to sensing and detection. (a) Protein biosensors for PTH detection. Left: Schematic of the grafted PTH lucCage biosensor, depicting the cage and latch (left, beige), key (right, beige), luciferase halves (inactive in white, active in blue), the PTH binder (red), and PTH peptide target (purple). Right: design model shown in the same color scheme. **(b)** Titration of PTH results in linear increases in luciferase luminescence. **(c)** Evaluation of the PTH biosensor at limiting concentrations of PTH indicates a 10 nM limit of detection (see methods). **(d-f)** The designed PTH binder enables robust recovery of PTH from complex mixtures. **(d)** Enrichment experiment schematic. **(e)** LC-MS/MS chromatograms for SVSEIQLMHNLGK, the N-terminal tryptic peptide of PTH; different peptide fragments detected by the LC-MS/MS assay are in different colors. **(f)** Mean chromatographic peak areas for triplicate measurements of each sample type. Error bars represent standard deviation.

References

1. Fukuhara, S. *et al.* Structure of the human secretin receptor coupled to an engineered heterotrimeric G protein. *Biochem. Biophys. Res. Commun.* **533**, 861–866 (2020).
2. Boesch, C., Bundi, A., Oppliger, M. & Wüthrich, K. 1H nuclear-magnetic-resonance studies of the molecular conformation of monomeric glucagon in aqueous solution. *Eur. J. Biochem.* **91**, 209–214 (1978).
3. Park, C. *et al.* Structural basis of neuropeptide Y signaling through Y1 receptor. *Nat. Commun.* **13**, 853 (2022).
4. Shimizu, N., Guo, J. & Gardella, T. J. Parathyroid hormone (PTH)-(1-14) and -(1-11) analogs conformationally constrained by alpha-aminoisobutyric acid mediate full agonist responses via the juxtamembrane region of the PTH-1 receptor. *J. Biol. Chem.* **276**, 49003–49012 (2001).
5. Cao, L. *et al.* Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).
6. Bennett, N. *et al.* Improving de novo Protein Binder Design with Deep Learning. 2022.06.15.495993 Preprint at https://doi.org/10.1101/2022.06.15.495993 (2022).
7. Kamani, F., Najafi, A., Mohammadi, S. S., Tavassoli, S. & Shojaei, S. P. Correlation of Biochemical Markers of Primary Hyperparathyroidism with Single Adenoma Weight and Volume. *Indian J. Surg.* **75**, 102–105 (2013).
8. Hu, L. & Xie, X. Parathyroid carcinoma with sarcomatoid differentiation: a case report and literature review. *Diagn. Pathol.* **15**, 142 (2020).
9. Wright, P. E. & Dyson, H. J. Linking folding and binding. *Curr. Opin. Struct. Biol.* **19**, 31–38 (2009).
10. Lazar, T., Tantos, A., Tompa, P. & Schad, E. Intrinsic protein disorder uncouples affinity from binding specificity. *Protein Sci. Publ. Protein Soc.* **31**, e4455 (2022).
11. Gisdon, F. J. *et al.* Modular peptide binders - development of a predictive technology as alternative for reagent antibodies. *Biol. Chem.* **403**, 535–543 (2022).
12. Wu, K. *et al.* De novo design of modular peptide binding proteins by superhelical matching. 2022.11.14.514089 Preprint at https://doi.org/10.1101/2022.11.14.514089 (2022).
13. Grigoryan, G. & DeGrado, W. F. Probing Designability via a Generalized Model of Helical Bundle Geometry. *J. Mol. Biol.* **405**, 1079–1100 (2011).
14. Wang, J. *et al.* Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
15. Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science* eadd2187 (2022) doi:10.1126/science.add2187.
16. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
17. Anishchenko, I. *et al.* De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).

18. Wicky, B. I. M. *et al.* Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).
19. Liu, Q. *et al.* Apoptotic regulation by MCL-1 through heterodimerization. *J. Biol. Chem.* **285**, 19615–19624 (2010).
20. Crabtree, M. D., Mendonça, C. A. T. F., Bubb, Q. R. & Clarke, J. Folding and binding pathways of BH3-only proteins are encoded within their intrinsically disordered sequence, not templated by partner proteins. *J. Biol. Chem.* **293**, 9718–9723 (2018).
21. Larhammar, D. Evolution of neuropeptide Y, peptide YY and pancreatic polypeptide. *Regul. Pept.* **62**, 1–11 (1996).
22. Säll, A. *et al.* Advancing the immunoaffinity platform AFFIRM to targeted measurements of proteins in serum in the pg/ml range. *PLOS ONE* **13**, e0189116 (2018).
23. Makaraviciute, A. & Ramanaviciene, A. Site-directed antibody immobilization techniques for immunosensors. *Biosens. Bioelectron.* **50**, 460–471 (2013).
24. Quijano-Rubio, A. *et al.* De novo design of modular and tunable protein biosensors. *Nature* **591**, 482–487 (2021).
25. Shi, J. *et al.* A distributable LC-MS/MS method for the measurement of serum thyroglobulin. *J. Mass Spectrom. Adv. Clin. Lab* **26**, 28–33 (2022).
26. Huynh, D. Q. Metrics for 3D Rotations: Comparison and Analysis. *J. Math. Imaging Vis.* **35**, 155–164 (2009).
27. Hoofnagle, A. N., Becker, J. O., Wener, M. H. & Heinecke, J. W. Quantification of thyroglobulin, a low-abundance serum protein, by immunoaffinity peptide enrichment and tandem mass spectrometry. *Clin. Chem.* **54**, 1796–1804 (2008).
28. Lopez, M. F. *et al.* Selected Reaction Monitoring–Mass Spectrometric Immunoassay Responsive to Parathyroid Hormone and Related Variants. *Clin. Chem.* **56**, 281–290 (2010).
29. Baker, M. Reproducibility crisis: Blame it on the antibodies. *Nature* **521**, 274–276 (2015).
30. Bradbury, A. & Plückthun, A. Reproducibility: Standardize antibodies used in research. *Nature* **518**, 27–29 (2015).

Methods

Identification of weak binder hits from parametric designs in pilot experiment

The first helical peptide binder hits were identified in pilot experiments screening for binding using the nanoBiT split luciferase assay (methods). These kinetic binding experiments were performed in cell lysate with no control over protein concentration, so candidate binders were selected qualitatively for showing some increase in luminescence signal over time above background noise, indicating likely binding activity. Additional pilot experiments indicated that this binding activity was all at very weak affinities, likely >100 nM. Therefore, these initial candidates were not further characterized, but rather selected for additional design to yield higher affinity binders.

Identification of weak binders for NPY and GCG using extended parametric designs

We used the RF inpainting approach to extend the binding interfaces of NPY and GCG weak binders hits from parametric design. However, the characterized proteins displayed low affinity binding to their targets, which was not enough for diagnostic applications.

Parametric design of groove-shaped scaffold library and use for binder design

The parametric groove-shaped scaffold library was sampled using a random sampling approach, where key parameters were selected randomly from distributions. An even distribution of bundle "lengths" was sampled, where each parametric helix was 15-19 residues long. A supercoiling value was randomly selected from a biased distribution favoring more supercoiled scaffolds, given these scaffolds were more likely to fail in the subsequent looping step. An average helix neighbor distance value was randomly selected from a normal distribution informed by native helical bundle geometries. The distance of each helix from its neighbors was independently randomly selected from a much tighter normal distribution centered at the preselected average helix neighbor distance value, to provide some noise within a given scaffold to helix distances and allow for heterogeneous amino acid selections. Values for helix phase and Z displacement were randomly sampled for each helix. The "groove" consisting of 3 helices was first sampled as a helical bundle using the Crick parameterization of alpha-helical coiled coils, around an imaginary central helix where the target was to later be docked. Next, the two buttressing helices were sampled with the same parameterization, but moved radially outward with randomly sampled helix neighbor distances as well as an additional randomly sampled tilt. This process was used to sample a set of 200k arrangements of 5 helices. Next, the Rosetta ConnectChainsMover was used to loop this set into approximately 135k successful scaffold backbones. These backbones were designed and filtered using Rosetta to yield a final library of 18 thousand scaffolds. This library was used to design binders to different helical peptide targets

using an adapted version of the miniprotein binder design computational pipeline used by Cao *et al.*[5].

Design of BIM peptide binders

We also experimented with unconditional binder design for the apoptosis-related peptide Bim (DMRPEIWIAQELRRIGDEFNAYYARR; PDB: 6X8O)- providing RF*diffusion* only with the sequence and structures of the two peptides in helical conformations, and leaving the topology of the binding protein and the binding mode completely unspecified. From this minimal starting information, RF*diffusion* generated designs predicted by AF2 to fold and bind to the targets with high *in silico* success rates. We obtained synthetic genes encoding 96 designs for each target. Using yeast surface display, we found that 25 of the 96 designs bound to Bim (10nM, no avidity). The highest affinity design, which purified as a soluble monomer, bound too tightly for steady state estimates of the dissociation constant (Kd); global fitting of the association and dissociation kinetics suggest a Kd of ~100pM. External potentials were used to promote interactions between the binder and target - specifically, the radius of gyration of the complex was minimized.

Gene construction of peptide hormone binders

The designed protein sequences were optimized to be both expressed in *S.cerevisiae* and *E. coli*. Linear DNA fragments (eBlocks, Integrated DNA Technologies) encoding design sequences included overhangs suitable for cloning into pETcon3 vector for yeast display[5] and Golden Gate cloning into LM627 vector for protein expression[18]. For initial testing hallucinated binders to Bid, binders were cloned into a modified LM627 vector. Specifically, Golden Gate cloning was used to generate sfGFP-Bid-STOP-[Binder]-SNAC-HISx6 assemblies.

Yeast display screening

For the yeast transformation, 50-60 ng of digested pETcon3 and 100 ng of insert (eBlocks, Integrated DNA Technologies) were transformed into *S. cerevisiae* EBY100 strain using the protocol described in ref[5] . EBY100 cultures were grown in C-Trp-Ura medium supplemented with 2% (w/v) glucose (CTUG). For induction of expression, yeast cells initially grown in CTUG were transferred to SGCAA medium supplemented with 0.2% (w/v) glucose and induced at 30 °C for 16–24 h. Cells were washed with PBSF (PBS with 1% (w/v) BSA) and labeled for 40 minutes with biotinylated peptide targets at room temperature using without-avidity labeling condition[5]. After incubation time, cells were washed and resuspended in PBSF for cell sorting (Attune NxT Flow Cytometer, Thermo Fisher Scientific).

NanoBiT screening

Linear gene fragments encoding binder design sequences and target peptide sequences were cloned into *E. coli* expression vectors using Golden Gate assembly; these vectors were pET28b(+) derivatives genetically fusing the smBiT and lgBiT halves of the NanoLuc® Luciferase (Promega) to the binders and peptides respectively. Resulting plasmids were transformed into BL21* (DE3) (Invitrogen) *E. coli* competent cells, then grown in 1mL TBII in 96-deepwell plates at 37C and 600 rpm. After 2 hours, expression was induced with IPTG (0.1 mM) and cells were incubated for an additional 4 hours. Cells were harvested by centrifugation (15 min at 4 kg), then resuspended in 100 uL lysis buffer (10 mM NaP pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$, 1 mg/mL lysozyme, 10 ug/mL DNAse I, 1 tablet Complete Protease inhibitor / 50 mL). Cells were incubated for 1 hour at room temperature and 600 rpm, then frozen (-80C for 30min) and thawed (37C at 600 rpm for 30min) twice. Lysate was cleared by centrifugation (20 min at 4 kg), and the soluble fraction was then transferred to a 96-well plate for use as stock protein/peptide for conducting the nanoBiT screen. Screens were assembled in 96-well Half Area Black Flat Bottom Polystyrene NBS Microplates (Corning 3686). Binder design smBiT lysate was diluted 12 uL into 1400 uL assay buffer (10 mM NaP pH 7.4, 150 mM NaCl), while target peptide lgBiT lysate was diluted 6 uL into 1400 uL assay buffer. Stock rows in the assay plate were prepared by mixing 40 uL substrate (499.2 uL assay buffer, 20.8 uL Nano-Glo® Luciferase Assay Substrate (Promega)) with 40 uL diluted binder design smBiT lysate, while experimental rows were prepared by adding 50 uL diluted target peptide lgBiT lysate. At read time, 50 uL of the stock row was added to the 50 uL experimental row and mixed quickly and carefully, then luminescence was read immediately for 5 min using a plate reader (Biotek Synergy Neo2).

Bicistronic protein expression

Hallucinated binders to Bid were screened by bicistronic expression with the Bid peptide. Plasmids encoding sfGFP-Bid-STOP-[Binder]-SNAC-HISx6 were cloned into *E. coli,* and 2 mL cultures of each of the 47 designs were grown overnight in LB. Cultures were diluted into TB medium, and grown to approximately OD280 0.6, before induction with 1 mM IPTG for 4 hours at 37ºC. Bacteria were lysed for 15 minutes in 300 B-PER (Thermo) + 1 mM PMSF, 0.1 mg/mL Lysozyme (Sigma), 0.01 mg/mL DNAse I. Lysates were clarified by centrifugation at 4000 *g* for 10 minutes, before purification on Ni-NTA resin (wash buffer: 20 mM Tris pH 8.0, 150 mM NaCl, 20 mM Imidazole; elution buffer: 20 mM Tris pH 8.0, 150 mM NaCl, 250 mM Imidazole). Eluates were assessed for GFP fluorescence on a fluorescence plate reader.

Peptide synthesis and purification

The PTH-TAMRA peptide was synthesized in-house on a CEM Liberty Blue microwave synthesizer. All L- and D-amino acids were purchased from P3 Biosystems. Oxyma Pure was purchased from CEM, DIC was purchased from Oakwood Chemical, diisopropyl ethylamine (DIEA) and piperidine were purchased from Sigma- Aldrich. Dimethylformamide (DMF) was

purchased from Fisher Scientific and treated with an Aldraamine trapping pack prior to use. Synthesis was done on a 0.1 mmol scale on CEM Cl-TCP(Cl) resin. Five equivalents of each amino acid were activated using 0.1 M Oxyma with 2% (v/v) DIEA in DMF, 15.4% (v/v) DIC, and coupled on resin for 4 min with double coupling if needed. This was followed by deprotection using 5 mL of 20% piperidine in DMF for 2 min at 95 °C. Global deprotection was accomplished TFA/Water/TIPS (95:2.5:2.5) for 3 hours. This deprotection mixture was precipitated in 30 mL of ice-cold ethyl ether, centrifuged and decanted, then washed twice more with fresh ether and dried under nitrogen to yield crude peptide for high pressure liquid chromatography (HPLC) purification.

The crude peptide was dried and dissolved in a mixture of ACN and water where the entire crude is soluble. This solution was purified on a C18 column in an Agilent HPLC instrument. A linear gradient of increasing ACN with 0.1% TFA was used to purify the samples. UV signal was monitored at 214 nm and all peaks were collected. Peaks were checked using ESI mass spectroscopy for the correct peptide mass. The purified peptide was then lyophilized for further use.

Protein expression and purification in *E. coli* for peptide hormone binders

Protein expression was performed using 50 mL of the Studier autoinduction media supplemented with kanamycin, and grown overnight at 37°C. The cells were harvested by spinning at 4,000 x g for 10 min and then resuspended in lysis buffer (100 mM Tris-HCl, 200 mM NaCl, 50 mM imidazole) supplemented with protease inhibitor tablets (Pierce™ Protease Inhibitor Tablets, EDTA-free). Then, the cells were lysed by sonication in a Qsonica, Q500 with a: 4-pronged horn for 2:30 min ON total, with an amplitude of 80%. Soluble fractions were clarified by centrifugation at 14,000 x g for 40 minutes, and were subsequently purified by affinity chromatography using bed Ni-NTA resin (Qiagen or Thermo Fisher) on a vacuum manifold. A series of washes using Low-salt buffer (20 mM Tris-HCl, 200 mM NaCl, 50 mM imidazole) and High- salt buffer (20 mM Tris-HCl, 1000 mM NaCl, 50 mM imidazole) were performed prior to elution with Elution buffer (20 mM Tris-HCl, 200 mM NaCl, 500 mM imidazole). After elution, protein samples were filtered and injected into an autosampler-equipped Akta pure system on a Superdex S75 Increase 10/300 GL column at room temperature. The SEC running buffer was 20mM Tris-HCl, 100mM NaCl pH 8. Selected fractions were pooled and concentrated using Spin filters (3 kDa molecular weight cutoff, Amicon, Millipore Sigma) and stored at 4 °C before downstream characterizations. Protein concentrations were determined by absorbance at 280 nm using a NanoDrop spectrophotometer (Thermo Scientific) using their extinction coefficients and molecular weights obtained from their amino acid sequences using the ProtParam tool.

Fluorescence polarization

Fluorescence polarization binding assays were carried out in 96-well plates (Corning 3686), with two-fold serial dilution of designed peptide binders in the presence of 0.5 nM fluorescently labeled peptide targets. Protein and peptide were diluted from their stock concentration into 20mM Tris-HCl pH 8, 100mM NaCl, 0.1% v/v Tween 20, and the protein was titrated in 2-fold serial dilutions onto constant peptide. After incubating the peptide and binder for one hour at room temperature, the fluorescence polarization was measured at the excitation and emission wavelengths of the FAM dye (485/530 nm) or the TAMRA dye (530/590 nm), in a Synergy Neo2 multi-mode plate reader. Titrations were conducted in replicate, and the Kd was fitted with SciPy. Specifically, curves were fit to N observations of an observed signal, $Signal_i$, at titrated concentrations $[A_{tot}]_i$ according to the following equation:

$$Signal_i = Baseline + Amplitude \frac{AB_{conc}([A_{tot}]_i, [B_{tot}], K_d)}{[B_{tot}]},$$

Where $[B_{tot}]$ is the known total concentration of the binder, $Baseline$ and $Amplitude$ are free parameters, and the concentration of the bound state $[AB]$ is computed as

$$AB_{conc}([A_{tot}]_i, [B_{tot}], K_d) = (([A_{tot}] + [B_{tot}] + K_d) \pm \sqrt{([A_{tot}] + [B_{tot}] + K_d)^2 - 4[A_{tot}][B_{tot}]})/2$$

The unknown parameters (K_D, $Baseline$ and $Amplitude$) were fit using scipy.optimize.curve_fit, $[B_{tot}]$ was additionally fit in the optimization, but only allowed to within 0.5 nM $\pm$ 0.1%.

Peptides used for the assay are shown in the following table:

Table 1. Fluorophore-labeled peptides used in Fluorescence polarization assays

Peptide name	Sequence	Supplier	Cat #	Fluorophore
PTH-TAMRA	SVSEIQLMHNLGKHLNSME RVEWLRKKLQDVHNF	In-house	NA	5-TAMRA
PTHrp-FAM	AVSEHQLLHDKGKSIQDLR RRFFLHHLIAEIHTAEIA	Phoenix Pharmaceuticals, Inc.	FG-056-08A	FAM
SCT-FAM	HSDGTFTSELSRLREGARLQ RLLQGLV	Phoenix Pharmaceuticals, Inc.	FG-067-03A	FAM
GCG-FAM	HSQGTFTSDYSKYLDSRRA QDFVQWLMNT	Addex Bio	ABBFO 2033	FAM
NPY-FAM	SKPDNPGEDAPAEDMARYY SALRHYINLITRQR	Phoenix Pharmaceuticals, Inc.	FG-049-04A	FAM
PPY-FAM	IKPEAAGEDASPEELNRYYA SLRHYLNLVTRQRY	Phoenix Pharmaceuticals, Inc.	FG-059-02A	FAM

Cloning, expression and purification of Bid-binding hallucinations, Avi-tagged Bid peptide and MCL-1

Bid-binding hallucinations were cloned into a pET28 vector, containing an N-terminal His_{10} and a PreScission cleavage site, using TEDA cloning[32] and transformed into XL-1-Blue chemically competent cells, single clones isolated and amplified and sequences confirmed by Sanger sequencing. Plasmids transformed into chemically competent BL21 DE3 *E. coli*, and plated onto LB agar plates supplemented with 100 ug/mL kanamycin. Single colonies were used to make starter cultures of LB with 100 ug/mL kanamycin and incubated overnight at 37 °C. 1:100 volume starter culture was added to autoinduction media Overnight Express Instant TB Medium (Novagen) in Ultra-Yield flasks (Thomson), with 100 µg/mL kanamycin, incubated at 37 °C for 5 hours, then 18 °C for 18 hrs. Cells harvested by centrifugation 6,000 rpm, 20 mins, 4 °C, and pellets were frozen at -80°C.

Defrosted cell pellets were resuspended in approx. 10 mL/g Lysis Buffer (50 mM potassium phosphate pH 7.0, 300 mM NaCl, 5 mM imidazole, 2 mM b-mercaptoethanol, 10% glycerol), supplemented with 60 µg/mL lysozyme, 1.4 µg/mL DNaseI, 0.05 mM PMSF. Cells were lysed by passing through French press twice, 18 kpsi. Lysate was clarified by centrifugation 18,000 *g*, 45 mins, 4 °C, and loaded onto HIS-Select Nickel affinity resin (Sigma) by gravity, resin washed with Wasg Buffer (50 mM potassium phosphate pH 7.0, 100 mM NaCl, 5 mM imidazole, 2 mM b-mercaptoethanol, 10% glycerol) and eluted with Wash Buffer containing 350 mM imidazole. Protein containing fractions (assessed by A_{280}) were combined, and further purified by size exclusion chromatography (SEC) using HiLoad 16/600 200 pg Superdex column (Cytiva) using ÅKTA FPLC system (Cytiva) equilibrated in 50 mM sodium phosphate pH 7.0, 1 mM DTT. Fractions were concentrated, concentration measured using A_{280} and predicted extinction coefficients[33], then flash frozen $N_{2(l)}$ for storage at -80 °C.

DNA corresponding to BH3 motif of human Bid Q79-G144 (Uniprot: P55957) was assembled by complementary oligos (IDT) and primer extension using Klenow fragment (NEB), and cloned using TEDA into pET28 with an N-terminal His_{10}, SUMO and C-terminal Avi. Expression and purification was carried out as for the hallucinations, except for co-transformation with a chloramphenicol-resistant BirA expressing plasmid, the addition of chloramphenicol 25 ug/mL in all cultures, with the addition of 40 µM BTN to the media before temperature was reduced to 18 °C. After SEC, His_{10}-SUMO was cleaved using ULP-1 protease, and His_{10}-SUMO removed using Ni resin, Bid-Avi peptide concentration was measured using A_{280}, and stored at -80°C. To express human Mcl-1 P166-G327 (Uniprot: Q07820) a pEQ80L vector with N-terminal His_6 and Avi-tag, for co-expression with BirA. Expression and purification was carried out as for the Bid-binding hallucinations, with the addition of 40 µM BTN to the media before temperature was reduced to 18 °C.

ITC

Isothermal titration calorimetry was carried out with an ITC200 (Micocal). Bid peptide was in the syringe, at ~300 μM, and binder (hallucination of Mcl-1) was kept in the cell (~25 μM), with both peptide and binder in matched buffer (sodium phosphate pH 7.0, 1 mM DTT). Temperature was held at 25 °C or 10 °C, as indicated. Fitting of titrations was carried out using 1-site binding, using manufacturers software (OriginLab).

Circular dichroism

Spectra were recorded for Bid peptide alone, Bid in complex with binders (hallucination or Mcl-1) and binders alone. All concentrations were 10 μM, in a 2 cm pathlength quartz cuvette. Spectra recorded on J-1500 Circular Dichroism Spectrophotometer, with temperature held at 25 °C, or ramped at 1 °C/min.

Pull-down

10 μL bead slurry Dynabeads M-280 Streptavidin (Thermo Fisher Scientific) were washed with Pull-Down Buffer (sodium phosphate pH 7.0, 1 mM DTT, 0.05% Tween20), incubated with saturating amounts of (Avi-tagged) Bid peptide 15 mins, 4 °C with rotation, beads were then incubated with free biotin 25 μM, and washed three times with ice cold Pull-Down Buffer. 10 μL of 2 μM binder (hallucination or Mcl-1) was incubated with pelleted beads for 30 mins, 4 °C, with rotation. Supernatant was recovered and the beads washed three times before resuspension in 10 μL Pull-Down Buffer. Both supernatant and washed beads were loaded onto denaturing SDS-PAGE, with protein detection by InstantBlue Coomassie staining.

Bio-layer Inferometry (BLI) Binding Experiments

BLI experiments were performed on an Octet Red96 (ForteBio) instrument, with streptavidin coated tips (Sartorius Item no. 18-5019). Buffer comprised 1X HBS-EP+ buffer (Cytiva BR100669) supplemented with 0.1% w/v bovine serum albumin. Tips were pre-incubated in the buffer for at least 10 minutes before use. Tips were then sequentially incubated in 50nM biotinylated Bim peptide (loading, 500s), buffer (baseline, 150s), designed binder (association, 1200s) and buffer (dissociation, 600s). Due to the extremely slow dissociation of Bim from the designed binders, it was not possible to calculate a precise K_D, but estimates suggest significantly sub-nanomolar affinity.

Design and characterization of lucCagePTH biosensor for parathyroid hormone detection

The detailed design protocol for the lucCage and lucKey sensor system was described previously (Nature, 2021, 482). In brief, the amino acid sequence (FELLDKLIELLRELIELTREYI) at the N-terminal end of the 6.1 nM PTH binder was grafted onto the latch region (residues 323 to 353) of lucCage. The Rosetta models were visually inspected and eight of them were selected for experimental validation. We produced, purified, and screened for the luminescence signal emitted from each biosensor in the presence of 5 μM PTH. From this process, we identified several hits showing increased luminescence upon adding PTH, of which we assigned the best one with a 21-fold activation as lucCagePTH. We then set up assays to evaluate the response of lucCagePTH with a range of PTH concentrations. 10 μl of 10 nM lucCagePTH, 10 μl of 10 nM lucKey, 10 μl of serial diluted PTH, and 40 μl of buffer (50% HBS-EP/50% Nano-Glo luciferase assay buffer) were pre-mixed and 30 μl of 100× diluted furimazine was injected immediately before luminescence kinetic acquisition. The luminescence measurements were taken every 1 min (0.1 s integration and 10 s shaking during intervals) for a total of 60 mins by Neo2 microplate reader. The linear region of luminescence responses to the corresponding PTH concentrations was fitted to a linear regression curve and the LOD was calculated as 3 × standard deviation of the response / the slope of the calibration curve.

Affinity enrichment of PTH analyzed by LC-MS/MS

Sample description

Recombinant human PTH protein was purchased from Sigma (#SAE 0192_100 ug, MA, USA) and reconstituted at 100 μg/mL in a 10 % acetonitrile, 0.1 % formic acid,1 mg/mL bovine serum albumin solution and stored in 40 μL aliquots at -20 °C. Dilutions at 1000 ng/mL and 62.5 ng/mL were prepared freshly as needed by dilution in the same acetonitrile, formic acid, albumin solution.

The plasma samples used were de-identified leftover clinical samples obtained from the clinical laboratories at the University of Washington Medical Center. The use of de-identified leftover clinical samples was reviewed by the University of Washington Human Subjects Division (STUDY00013706).

The evaluation of PTH immunoaffinity enrichment in buffer and plasma was performed in three process replicates using 8 different types of samples:

- Series A: Reconstitution buffer (10 % acetonitrile, 0.1 % formic acid,1 mg/mL bovine serum albumin in water) served as the blank.
- Series B: Reconstitution buffer spiked with PTH at 7.2 ng/mL was directly digested without the addition of beads and served as the Control sample (representing 100% recovery of PTH).
- Series C: Reconstitution buffer spiked with PTH at 7.2 ng/mL was incubated with beads blocked by bovine serum albumin before washing and digestion, which served as the negative control, to quantify non-specific binding in buffer.
- Series D: Reconstitution buffer spiked with PTH at 7.2 ng/mL was incubated with designed binder-conjugated beads before washing and digestion, which was used to quantify the affinity precipitation of PTH from buffer.
- Series E: Plasma was incubated with beads blocked by bovine serum albumin before washing and digestion, which was used to quantify non-specific binding in unspiked plasma.
- Series F: Plasma was incubated with designed binder-conjugated beads before washing and digestion, which was used to quantify affinity precipitation of PTH in plasma.
- Series G: Plasma spiked with PTH at 7.2 ng/mL was incubated with beads blocked by bovine serum albumin before washing and digestion, which was used to quantify non-specific binding in spiked plasma.
- Series H: Plasma spiked with PTH at 7.2 ng/mL was incubated with designed binder-conjugated beads before washing and digestion, which was used to quantify the affinity precipitation of PTH in spiked plasma.

Sample preparation and LC-MS/MS conditions

Affinity enrichment was performed in buffer or plasma at the protein level. Designed binders were conjugated to tosyl-activated Dynabeads M-280 according to the manufacturer's instructions and subsequently blocked using bovine serum albumin and Tris. The amino terminal peptide was analyzed after tryptic digestion of either pure protein in buffer, or after trypsin digestion of PTH that had been affinity precipitated by the designed binder-conjugated beads (or by the control/blocked magnetic beads). Briefly, PTH proteins in buffer/plasma were purified using PTH mini-binder conjugated-paramagnetic beads at room temperature, for 1 h. The beads were then washed 4 times with phosphate-buffered saline supplemented with CHAPS (0.1% 3-((3cholomidopropyl) dimethylammonio)-1-propanesulfate to reduce nonspecific interactions). The proteins that were affinity precipitated by the designed binder-conjugated-paramagnetic beads were suspended in 10 µL of a solution containing 10 % acetonitrile, 0.1 % formic acid, 1

mg/mL bovine serum albumin. The washed beads were then suspended with 30 µL of 30% isopropanol,100 mM ammonium bicarbonate, and digested at 37 °C for 30 min after adding 100 µL of 0.01 mg/mL trypsin in 10 mM hydrochloride acid. The liberated peptides were then removed from the beads using a magnet and analyzed using LC-MS/MS.

Peptides were analyzed by liquid chromatography-tandem mass spectrometry in the multiple reaction monitoring acquisition mode using an UHPLC I-Class Chromatography system coupled to a Xevo TQ-S triple quadrupole tandem mass spectrometer (Waters, MA, USA). Peptides were eluted from an Acquity UPLC HSS T3 1.8µm (C18, 2.1x50 mm, pore size 100 Å) analytical column (Waters) at 45 °C using 0.1 % formic acid, 2 % dimethylsulfoxide in LC-MS grade water as mobile phase A and 0.1 % formic acid, 2 % dimethylsulfoxide in LC-MS grade methanol as mobile phase B.

The liquid chromatography and mass spectrometry conditions are detailed in Tables 2, 3 and 4.

Table 2. Liquid chromatography conditions

Mobile phase	Phase A: 0.1 % formic acid, 2 % dimethylsulfoxide in water 0.1 % formic acid, 2 % dimethylsulfoxide in methanol
Column	Acquity UPLC HSS T3 1.8µm (C18, 2.1x50 mm, pore size 100 Å)
Temperature	45±5 °C
Flow rate	0.3 mL/min
Injection volume	20 µL
Gradient	0-0.5 min: 2% B at 0.3 mL/min 7.5: 98% B at 0.3 mL/min 7.6: 98% B at 0.6 mL/min 8.6: 2% B at 0.6 mL/min 9.9: 2% at 0.3 mL/min

Table 3. Mass spectrometry conditions	
Source polarity	ESI+
Capillary voltage	3.25 kV
Source Offset voltage	50 V
Desolvation Temp	600 °C
Desolvation Gas Flow	1000 L/h
Cone Gas Flow	150 L/h

Table 4. Multiple reaction monitoring conditions					
Peptide sequences	Q1 (m/z)	Q3 (m/z)	Cone (V)	Collision Energy (eV)	Ion type
HLNSMER.2	443.7136	218.1047	35	15	y3
HLNSMER.2	443.7136	261.6207	35	15	y4
HLNSMER.2	443.7136	318.6421	35	15	y5
HLNSMER.3	296.1448	218.1047	35	9	y3
HLNSMER.3	296.1448	261.6207	35	9	y4
HLNSMER.3	296.1448	318.6421	35	9	y5
HLNSMER.3	296.1448	435.202	35	9	y3
HLNSMER.3	296.1448	522.2341	35	9	y4
HLNSMER.3	296.1448	636.277	35	9	y5
HLNSM(+15.994915)ER.2	451.7111	226.1021	35	16	y3
HLNSM(+15.994915)ER.2	451.7111	269.6181	35	16	y4
HLNSM(+15.994915)ER.2	451.7111	326.6396	35	16	y5
HLNSM(+15.994915)ER.2	451.7111	451.1969	35	16	y3
HLNSM(+15.994915)ER.2	451.7111	538.229	35	16	y4
HLNSM(+15.994915)ER.2	451.7111	652.2719	35	16	y5
HLNSM(+15.994915)ER.3	301.4765	226.1021	35	10	y3
HLNSM(+15.994915)ER.3	301.4765	269.6181	35	10	y4
HLNSM(+15.994915)ER.3	301.4765	326.6396	35	10	y5
HLNSM(+15.994915)ER.3	301.4765	451.1969	35	10	y3
HLNSM(+15.994915)ER.3	301.4765	538.229	35	10	y4

ADVNVLTK.2	430.2478	574.3559	35	15	y5
ADVNVLTK.2	430.2478	673.4243	35	15	y6
ADVNVLTK.3	287.1676	181.1259	35	9	y3
ADVNVLTK.3	287.1676	230.6601	35	9	y4
ADVNVLTK.3	287.1676	361.2445	35	9	y3
ADVNVLTK.3	287.1676	460.313	35	9	y4
SLGEADK.2	360.1821	167.0921	35	12	y3
SLGEADK.2	360.1821	231.6134	35	12	y4
SLGEADK.2	360.1821	260.1241	35	12	y5
SLGEADK.2	360.1821	333.1769	35	12	y3
SLGEADK.2	360.1821	462.2195	35	12	y4
SLGEADK.2	360.1821	519.2409	35	12	y5
SLGEADK.3	240.4572	167.0921	35	7	y3
SLGEADK.3	240.4572	260.1241	35	7	y5
SLGEADK.3	240.4572	333.1769	35	7	y3
SLGEADK.3	240.4572	462.2195	35	7	y4
VEWLR.2	351.7003	229.1183	35	12	b2
VEWLR.2	351.7003	474.2823	35	12	y3
EDNVLVESHEK.2	649.8148	629.2889	35	23	y5
EDNVLVESHEK.2	649.8148	728.3573	35	23	y6
EDNVLVESHEK.2	649.8148	841.4414	35	23	y7
EDNVLVESHEK.3	433.5456	315.1481	35	14	y5
EDNVLVESHEK.3	433.5456	364.6823	35	14	y6

EDNVLVESHEK.3	433.5456	421.2243	35	14	y7
EDNVLVESHEK.3	433.5456	629.2889	35	14	y5
DAGSQRPR.2	443.7281	322.1854	35	15	y5
DAGSQRPR.2	443.7281	643.3634	35	15	y5
DAGSQRPR.2	443.7281	700.3849	35	15	y6
DAGSQRPR.3	296.1545	214.6401	35	9	y3
DAGSQRPR.3	296.1545	278.6693	35	9	y4
DAGSQRPR.3	296.1545	322.1854	35	9	y5
DAGSQRPR.3	296.1545	350.6961	35	9	y6
DAGSQRPR.3	296.1545	428.2728	35	9	y3
DAGSQRPR.3	296.1545	556.3314	35	9	y4
DAGSQRPR.3	296.1545	643.3634	35	9	y5
DAGSQRPR.3	296.1545	700.3849	35	9	y6
SVSEIQLMHNLGK.2	728.3849	527.2973	35	26	y9
SVSEIQLMHNLGK.2	728.3849	568.3202	35	26	y5
SVSEIQLMHNLGK.2	728.3849	635.3346	35	26	y11
SVSEIQLMHNLGK.2	728.3849	699.3607	35	26	y6
SVSEIQLMHNLGK.2	728.3849	812.4447	35	26	y7
SVSEIQLMHNLGK.2	728.3849	940.5033	35	26	y8
SVSEIQLMHNLGK.2	728.3849	1053.587	35	26	y9
SVSEIQLMHNLGK.2	728.3849	1269.662	35	26	y11
SVSEIQLMHNLGK.3	485.9257	159.1128	35	16	y3
SVSEIQLMHNLGK.3	485.9257	431.2613	35	16	y4

SVSEIQLMHNLGK.3	485.9257	470.7553	35	16	y8
SVSEIQLMHNLGK.3	485.9257	527.2973	35	16	y9
SVSEIQLMHNLGK.3	485.9257	568.3202	35	16	y5
SVSEIQLMHNLGK.3	485.9257	591.8186	35	16	y10
SVSEIQLMHNLGK.3	485.9257	635.3346	35	16	y11
SVSEIQLMHNLGK.3	485.9257	699.3607	35	16	y6
SVSEIQLMHNLGK.3	485.9257	812.4447	35	16	y7
SVSEIQLMHNLGK.3	485.9257	940.5033	35	16	y8
SVSEIQLM(+15.994915)HNLGK.2	736.3823	159.1128	35	26	y3
SVSEIQLM(+15.994915)HNLGK.2	736.3823	317.2183	35	26	y3
SVSEIQLM(+15.994915)HNLGK.2	736.3823	431.2613	35	26	y4
SVSEIQLM(+15.994915)HNLGK.2	736.3823	535.2948	35	26	y9
SVSEIQLM(+15.994915)HNLGK.2	736.3823	568.3202	35	26	y5
SVSEIQLM(+15.994915)HNLGK.2	736.3823	643.3321	35	26	y11
SVSEIQLM(+15.994915)HNLGK.2	736.3823	715.3556	35	26	y6
SVSEIQLM(+15.994915)HNLGK.2	736.3823	828.4396	35	26	y7
SVSEIQLM(+15.994915)HNLGK.2	736.3823	956.4982	35	26	y8
SVSEIQLM(+15.994915)HNLGK.2	736.3823	1069.582	35	26	y9
SVSEIQLM(+15.994915)HNLGK.3	491.2573	159.1128	35	16	y3
SVSEIQLM(+15.994915)HNLGK.3	491.2573	643.3321	35	16	y11

Data treatment

Data processing was performed with Skyline Daily version 21.1.1.223. The peak area for each peptide was determined as the sum of the peak areas of all selected transitions. The recovery over blocked-beads (RE) in spiked buffer and in spiked plasma was estimated using Equations 1, and 2, respectively.

$$RE_{buffer} = \frac{Peak\ area\ Series\ D}{Peak\ area\ Series\ B} \quad (1)$$

$$RE_{plasma} = \frac{Peak\ area\ Series\ H}{Peak\ area\ Series\ B} \quad (2)$$

Supplementary Figures

Figure S1. Parametric groove scaffold library. 45 scaffolds from the library of 18 thousand parametric groove scaffolds, demonstrating a range of supercoiling and helix distances to accommodate a range of helical peptide targets.

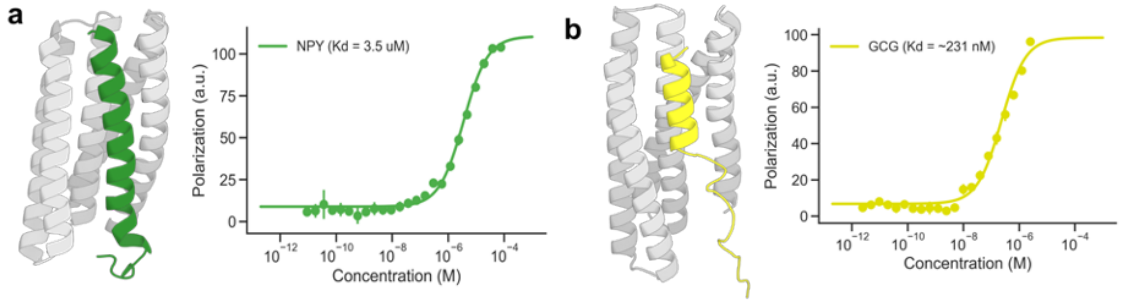

Figure S2. Inpainted peptide binders bound their targets with low affinity. (a) NPY binder. **(b)** Glucagon binder. AF2 predictions of the proteins and peptides are shown on the left. FP binding data is shown on the right.

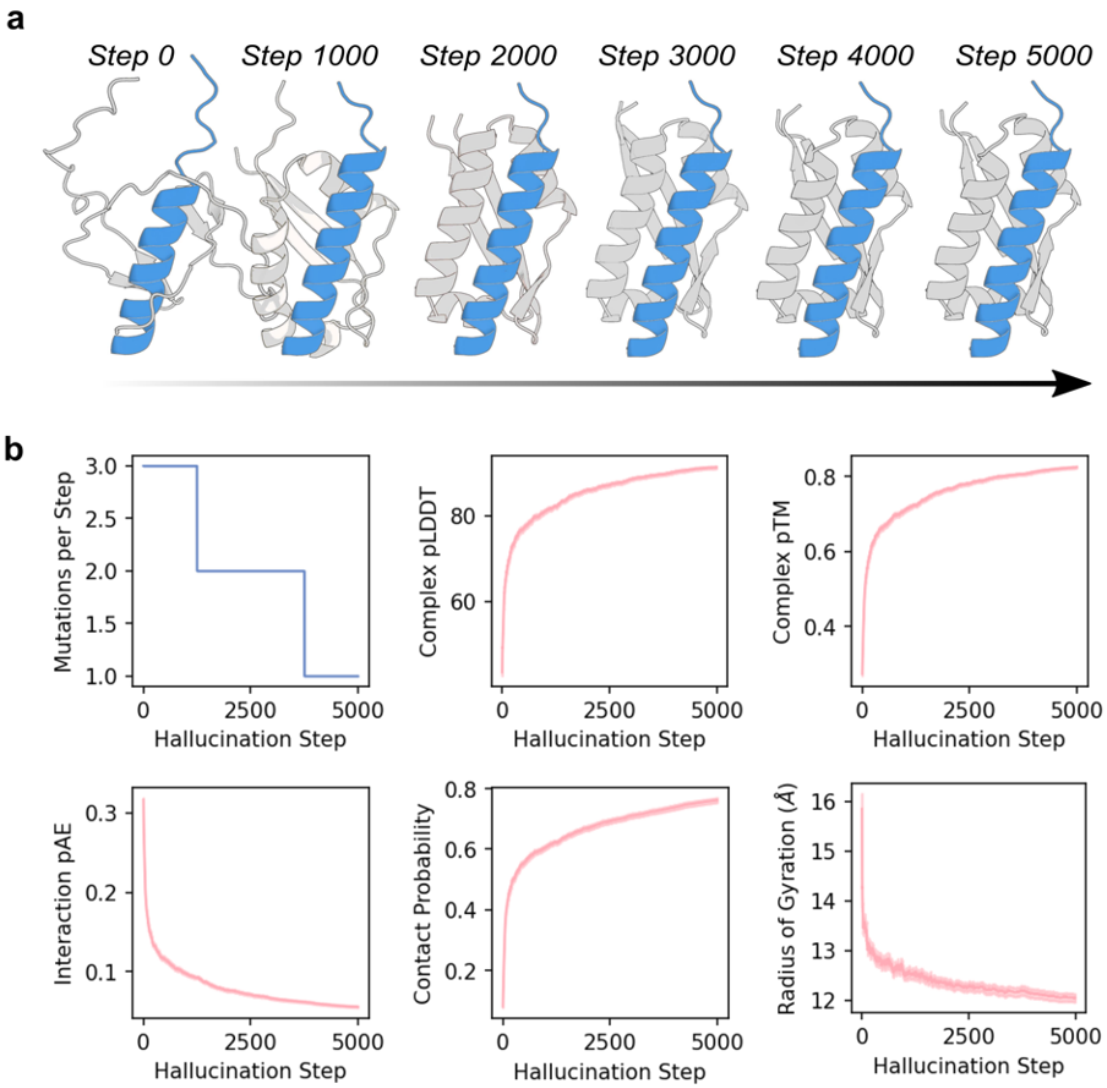

Figure S3. Hallucinating Bid Binders with AlphaFold2. (a) Example hallucination trajectory generating 70 amino acid binders to the peptide Bid (blue). Initially, AlphaFold2 predicts an unstructured "binder", but over 5000 steps, a binder is built up around the peptide. Crucially, no template structure is provided for the Bid peptide, allowing AF2 to predict its structure throughout. Note the predicted elongation of the helical structure in the peptide (blue, top) over the hallucination trajectory. **(b)** Hallucination trajectories approximately converge after 5000 steps. Left to right, top to bottom: The mutation rate at each step is decayed throughout the trajectory (1250 x 3 steps, 2500 x 2 steps, 1250 x 1 step). More mutations initially helps speed up hallucination, while a lower rate later on allows more gradual refinement. The AF2 confidence (pLDDT, pTM) in the bound structure increases throughout trajectories, while the pAE between peptide and binder (known to be a good correlator of binding) decreases. The contact probability also trends to convergence over the trajectories, while the proteins typically become more compact (radius of gyration). N=96 trajectories.

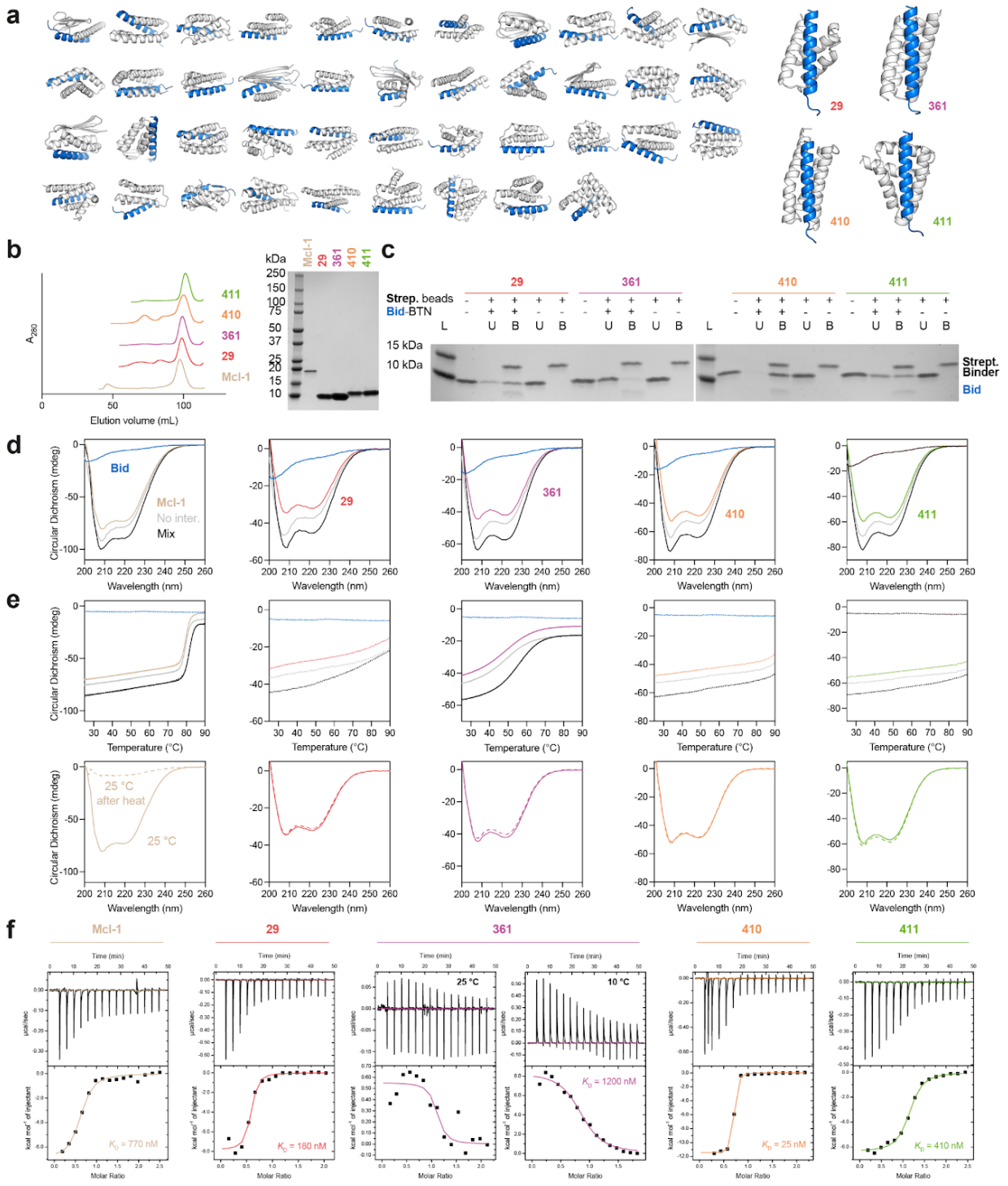

Figure S4. Hallucinated Bid binders were stable and bound Bid peptide with high affinity. (a) 47 hallucinated designs tested for initial experimental screening. **(b)** 4 designs were chosen for expression without Bid peptide. All expressed as monomeric proteins (assessed by preparative SEC) and were pure by SDS-PAGE. **(c)** All hallucinations could be pulled-down by biotinylated Bid immobilized on streptavidin magnetic beads. B = bound to bead, U = unbound, in supernatant. L = ladder. **(d)** Bid is unstructured in isolation by circular dichroism (CD), whereas all hallucinations were helical in isolation,

as predicted from the hallucinated structure. A 1:1 molar ratio of binder:Bid (Mix) produced greater helical signal than that predicted by the isolated spectra (No inter.) suggesting binding is inducing helix formation. **(e)** Melting with CD showed that hallucinations were thermostable, and binding to Bid increased thermostability (where measurable). All hallucinations would remain folded, or refold after heating and cooling, in contract to the natural binder Mcl-1 which precipitated in the process. **(f)** ITC showed that hallucinations bound to Bid, with μM to nM K_ds.

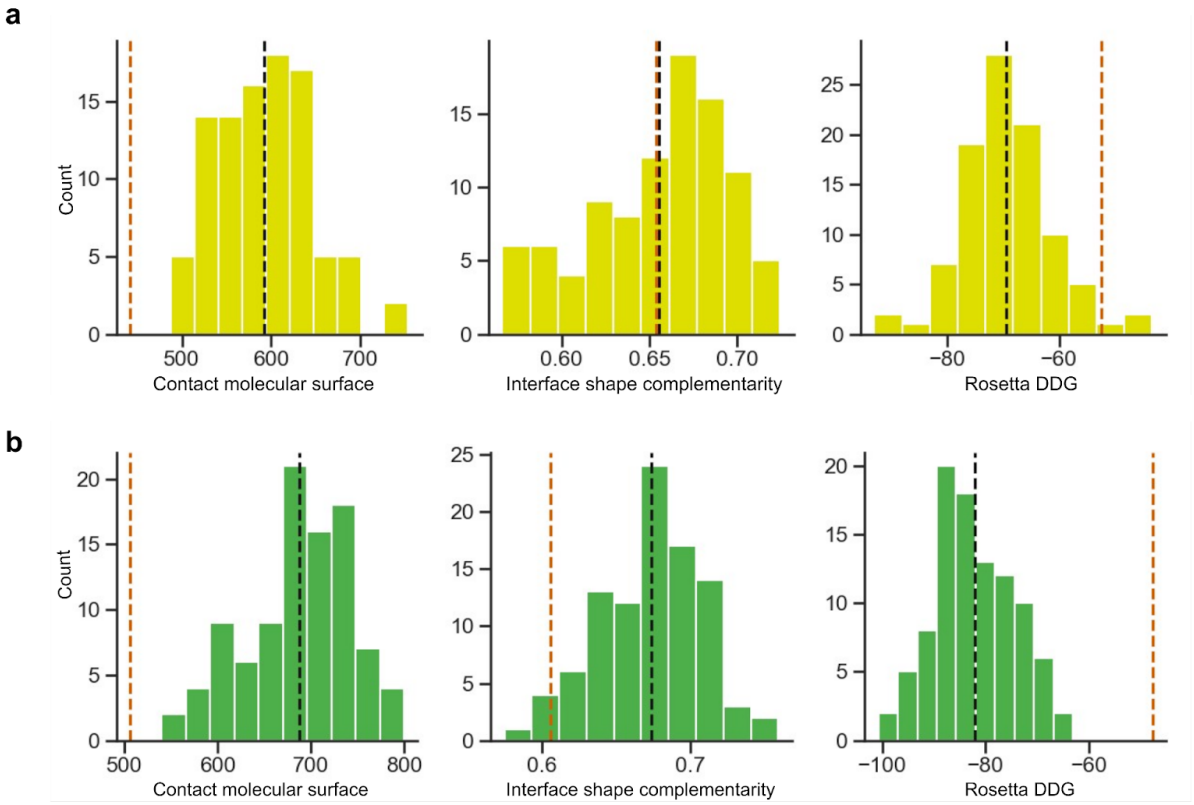

Figure S5. Binding metrics for partially diffused binders. (a) Computational metrics for 96 ordered partially diffused glucagon binders showed significant improvement in contact molecular surface (a measure of interface size and quality) and Rosetta ddG (a measure of interface predicted energy) over the starting design (vertical red lines). Distribution means are shown in black. **(b)** Computational metrics for 96 ordered partially diffused NPY binders showed significant improvement in contact molecular surface, Rosetta ddG, and interface shape complementarity (a measure of interface quality) over the starting design (vertical red lines). Means are shown in black.

Supplementary Video 1. A video of the diffusion trajectory for the fully diffused PTH binder can be seen at: https://www.bakerlab.org/wp-content/uploads/2022/11/diffusion_animation_PTHbinder_v6.mp4

Supplementary Tables

Table S1. AlphaFold metrics for partially and fully diffused binders.

	GCG Binder	NPY Binder	PTH Binder	Bim Binder
RMSD AF2 vs Design	0.62 Å	0.61 Å	0.78 Å	0.80 Å
AF2 interaction PAE	9.25	8.29	4.40	4.50
AF2 pLDDT for binder	95.52	93.41	94.3	96.6

Acknowledgements

Authors: Susana Vázquez Torres[‡1,2,3], Philip J. Y. Leung[‡1,2,4], Isaac D. Lutz[‡1,2,5], Preetham Venkatesh[‡1,2,3], Joseph L. Watson[1,2], Fabian Hink[6], Huu-Hien Huynh[7], Andy Hsien-Wei Yeh[1,2], David Juergens[1,2,4], Nathaniel R. Bennett[1,2,4], Andrew N. Hoofnagle[7], Eric Huang[8], Michael J MacCoss[8], Marc Expòsit[1,2,4], Gyu Rie Lee[1,2], Paul M. Levine[1,2], Xinting Li[1,2], Mila Lamb[1,2], Elif Nihal Korkmaz[1,2], Jeff Nivala[10,11], Lance Stewart[1,2], Joseph M. Rogers*[6], David Baker*[1,2,9].

Affiliations:
[1]Department of Biochemistry, University of Washington, Seattle, WA, USA.
[2]Institute for Protein Design, University of Washington, Seattle, WA, USA.
[3]Graduate Program in Biological Physics, Structure and Design, University of Washington, Seattle, WA, USA.
[4]Graduate Program in Molecular Engineering, University of Washington, Seattle, WA, USA.
[5]Department of Bioengineering, University of Washington, Seattle, WA, USA.
[6]Department of Drug Design and Pharmacology, University of Copenhagen, Jagtvej 160, 2100, Copenhagen, Denmark.
[7]Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA, USA.
[8]Department of Genome Sciences, University of Washington, Seattle, WA, USA.
[9]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.
[10]School of Computer Science and Engineering, University of Washington, Seattle, WA, USA.
[11]Molecular Engineering and Sciences Institute, University of Washington, Seattle, WA, USA.

‡Equal contribution
*To whom correspondence should be addressed

Funding: This work was supported with funds provided by a grant U19 AG065156 from the National Institute for Aging (S.V.T., M.M., E.H., A.H., H.H.H., I.L., D.B.), a gift from Amgen (J.W.), the Audacious Project at the Institute for Protein Design (A.H.-W.Y., D.B.), a gift from Microsoft Gift supporting Computational Protein Structure Prediction and Design at the Institute for Protein Design (D.J., D.B.), the Washington State General Operating Fund supporting the Institute for Protein Design (P.V.), a grant INV-010680 from the Bill and Melinda Gates Foundation Grant (D.J., J.W., D.B.), a NIH NIBIB Pathway to Independence Award (A.H.-W.Y., K99EB031913), a National Science Foundation Training Grant number EF-2021552 (P.L.), NERSC award BER-ERCAP0022018 (P.L.), the Open Philanthropy Project Improving Protein Design Fund (P.L., G.R.L.,D.B.), The Donald and Jo Anne Petersen Endowment for Accelerating Advancements in Alzheimer's Disease Research (N.Ben.), and the Howard Hughes Medical Institute (D.B.). J.M.R. and F.H. were supported by the Novo Nordisk Foundation (NNF19OC0054441 to J.M.R.). H.H.H is supported by a postdoctoral fellowship provided by

the Partnership for Clean Competition. We thank Microsoft and AWS for generous gifts of cloud computing resources.

Author contributions: D.B. directed the work. I.L. and S.V.T. designed, screened, and experimentally characterized the parametrically designed groove scaffold peptide binders. P.J.Y.L. and S.V.T. designed, screened and experimentally characterized the threaded peptide binders. J.L.W., developed the hallucination method for peptide binding. J.L.W., F.H., and J.M.R designed and experimentally characterized the hallucinated peptide binders. J.L.W. and S.V.T. designed and characterized the inpainted binders. S.V.T. and P.V. designed, screened, and experimentally characterized all the different classes of diffused peptide binders shown in this manuscript. J.L.W., D.J., and N.R.B. developed the RF*diffusion* algorithm used for peptide binder design. H.H.H, E.H., M.J.M., and A.N.H performed the LC-MS/MS peptide detection. A.H.-W.Y. designed and characterized the lucCagePTH biosensors and analyzed the sensing experiments. M.E. and G.R.L supported during yeast display binding screening. All authors reviewed and accepted the manuscript.