



# De novo design of small beta barrel proteins

David E. Kim<sup>a,b,c</sup> , Davin R. Jensen<sup>d</sup>, David Feldman<sup>a,b</sup>, Doug Tischer<sup>a,b</sup>, Ayesha Saleem<sup>a,b</sup>, Cameron M. Chow<sup>a,b</sup> , Xinting Li<sup>a,b</sup>, Lauren Carter<sup>a,b</sup>, Lukas Milles<sup>a,b</sup> , Hannah Nguyen<sup>a,b</sup> , Alex Kang<sup>a,b</sup> , Asim K. Bera<sup>a,b</sup> , Francis C. Peterson<sup>d</sup>, Brian F. Volkman<sup>d</sup> , Sergey Ovchinnikov<sup>e,f</sup> , and David Baker<sup>a,b,c,1</sup>

Edited by William DeGrado, University of California San Francisco, San Francisco, CA; received May 10, 2022; accepted January 27, 2023

Small beta barrel proteins are attractive targets for computational design because of their considerable functional diversity despite their very small size (<70 amino acids). However, there are considerable challenges to designing such structures, and there has been little success thus far. Because of the small size, the hydrophobic core stabilizing the fold is necessarily very small, and the conformational strain of barrel closure can oppose folding; also intermolecular aggregation through free beta strand edges can compete with proper monomer folding. Here, we explore the de novo design of small beta barrel topologies using both Rosetta energy-based methods and deep learning approaches to design four small beta barrel folds: Src homology 3 (SH3) and oligonucleotide/oligosaccharide-binding (OB) topologies found in nature and five and six up-and-down-stranded barrels rarely if ever seen in nature. Both approaches yielded successful designs with high thermal stability and experimentally determined structures with less than 2.4 Å rmsd from the designed models. Using deep learning for backbone generation and Rosetta for sequence design yielded higher design success rates and increased structural diversity than Rosetta alone. The ability to design a large and structurally diverse set of small beta barrel proteins greatly increases the protein shape space available for designing binders to protein targets of interest.

protein design | small beta barrels | high-throughput screening | machine learning

While there has been considerable success in de novo design of all-alpha and alpha-beta folds (1–6), the design of globular all-beta proteins from scratch has only been achieved quite recently (7–9). The lag in progress was likely due to the tendency of all-beta designs to aggregate or be unstable, and also due to a focus on engineering “ideal” proteins with canonical secondary structure and short turn connections. All-beta folds often contain nonlocal backbone hydrogen-bonding interactions between residues that are distant in sequence that can be difficult to design since free strand edges can interact intermolecularly causing the formation of amyloid aggregates (10, 11). Furthermore, globular beta-folds often contain curved sheets and without incorporating nonideal features like beta-bulges and glycine kinks within the bent strands, considerable conformational strain can arise from steric clashes and unfavorable left-handed backbone twist. The incorporation of such nonideal features enabled the successful de novo design of an eight-stranded beta barrel (7), a single beta-sheet that twists to make nonlocal hydrogen bond interactions between the first and last strands closing the sheet into a barrel-like structure. However, the de novo design of smaller barrel-like proteins with six or fewer strands and less than 70 amino acids has not to our knowledge been achieved.

Small beta barrel proteins with six or fewer beta strands, such as SH3 domains and OB-folds, have an extremely diverse functional repertoire in nature involving interactions with proteins, peptides, oligonucleotides, oligosaccharides, structural scaffolding, shape shifting, and more. The features of these folds giving rise to this vast functional diversity include a large sequence space compatible with their simple structure, versatile binding sites involving concave sheet surfaces and pockets flanked by variably structured beta-turns and loops, and self-shape complementarity, which enables assembly into higher order oligomers. Learning how to systematically generate such proteins could thus considerably expand the range of activities and functions accessible to de novo protein design. For design of protein binding and other functions, small proteins also have considerable intrinsic advantages: They can be encoded on single long oligonucleotides (~230 nt) that can be synthesized in parallel in very large arrays greatly reducing the cost of gene synthesis, the density of binding (or other function) sites per unit mass is higher, and for therapeutic applications they can penetrate better into solid tumors.

We set out to design soluble beta barrels of less than 70 amino acids. We chose to focus on four different beta barrel topologies, two of which occur frequently in nature, and two of which occur rarely if at all (Fig. 1 *A–D*). The first two, the SH3 and OB folds, consist of

## Significance

De novo design of mini-proteins for engineering new functions has primarily focused on all-alpha and alpha-beta folds. Despite the enormous functional diversity observed in naturally occurring beta barrels less than 70 residues in length, the de novo design of these structures has not been achieved. Here, we describe the de novo design and characterization of four different classes of small beta barrels. The results provide insight into the determinants of folding and design of this important class of proteins, and provide routes to the design of high-affinity binding proteins.

Author affiliations: <sup>a</sup>Department of Biochemistry, University of Washington, Seattle, WA 98195; <sup>b</sup>Institute for Protein Design, University of Washington, Seattle, WA 98195; <sup>c</sup>HHMI, University of Washington, Seattle, WA 98195; <sup>d</sup>Department of Biochemistry, Medical College of Wisconsin, Milwaukee, WI 53226; <sup>e</sup>Division of Science, Faculty of Arts and Sciences, Harvard University, Cambridge, MA 02138; and <sup>f</sup>John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA 02138

Author contributions: D.E.K., D.F., D.T., and D.B. designed research; D.E.K., D.R.J., D.F., D.T., A.S., C.M.C., X.L., L.C., L.M., H.N., A.K., A.K.B., and S.O. performed research; D.E.K., D.R.J., D.F., A.K.B., F.C.P., and B.F.V. analyzed data; and D.E.K., D.F., D.T., and D.B. wrote the paper.

The authors declare no competing interest.

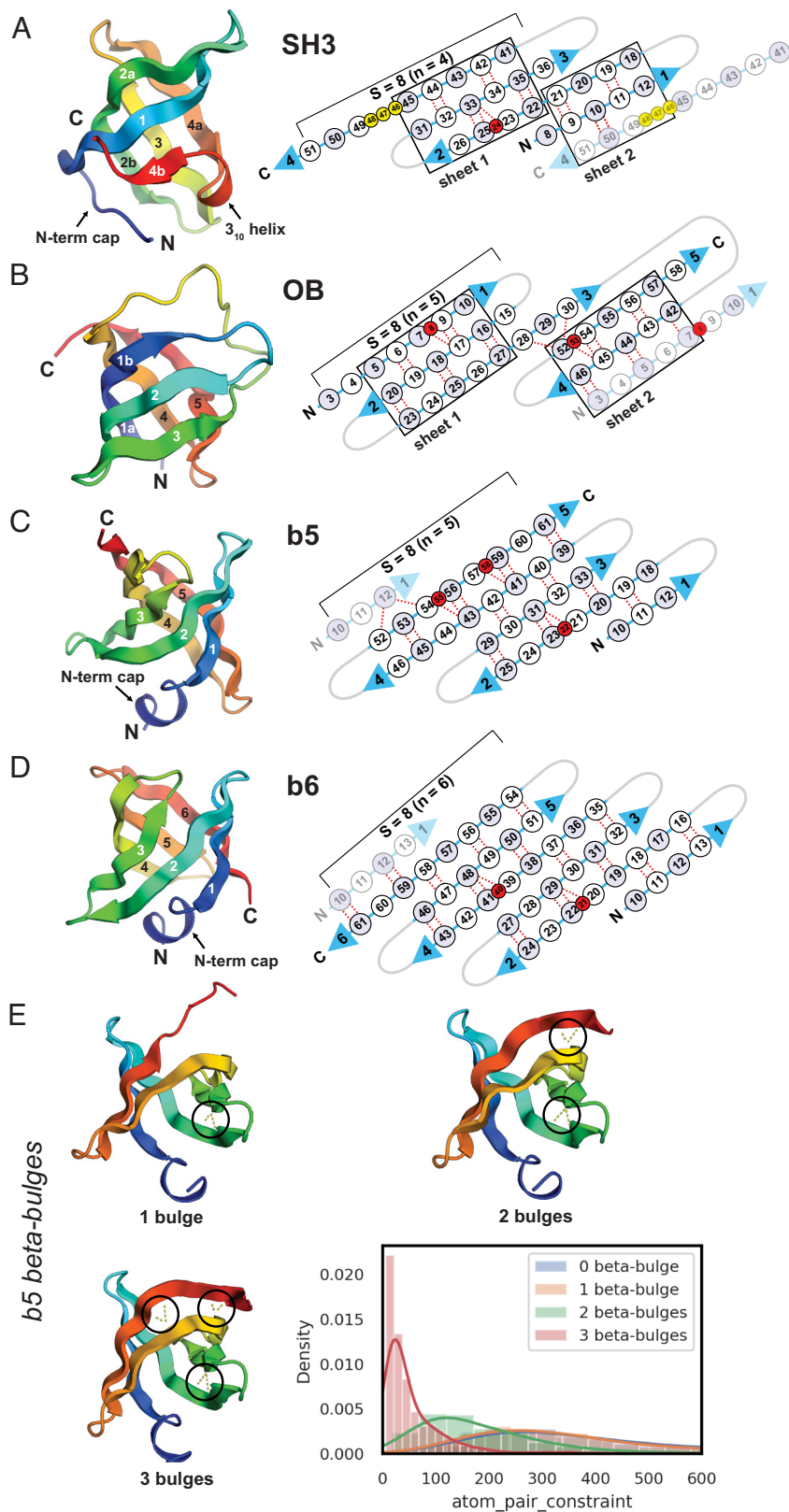
This article is a PNAS Direct Submission.

Copyright © 2023 the Author(s). Published by PNAS. This open access article is distributed under [Creative Commons Attribution License 4.0 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).

<sup>1</sup>To whom correspondence may be addressed. Email: dabaker@uw.edu.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2207974120/-/DCSupplemental>.

Published March 10, 2023.



**Fig. 1.** SH3, OB, b5, and b6 topologies and beta-strand 2d blueprint maps (A–D). SH3 and OB topologies consist of two orthogonally packed antiparallel three-stranded beta sheets. In the SH3 fold, sheet 2 consists of an N-terminal hairpin paired with the C-terminal strand, whereas in the OB fold, sheet 2 consists of a C-terminal hairpin paired with the N-terminal strand. The last strand in SH3 typically contains a 3<sub>10</sub> helix (yellow circles in the 2d map). In the 2D maps to the right of the structure cartoons, beta-strand residues are represented as circles labeled by residue number (white circles are residues that point toward the core and light blue the surface), backbone hydrogen bonds as red dashed lines, and beta-bulges as small red circles. (E) Addition of beta bulges to promote b5 beta-strand pairing and barrel closure. Examples of structure models with 1, 2 and 3 bulges are shown, together with a histogram of backbone H-bond (N–O) distance constraint scores (*atom\_pair\_constraint*) from backbone generation runs using blueprints with 0 to 3 beta-bulges. A score of 0 represents correctly modeled H-bond distances: the incorporation of beta bulges considerably improves the extent of hydrogen bonding, strand pairing, and barrel closure.

two orthogonally packed three-stranded antiparallel beta sheets where one sheet involves three consecutive beta strands and the other a beta hairpin paired with a single terminal strand. In the SH3 fold, the terminal strand is at the C terminus, and in the OB-fold, at the N terminus. The second two, which we refer to as b5 and b6, are up-and-down barrel-like folds consisting of a single antiparallel beta-sheet with five and six strands, respectively, that twists into a closed barrel by pairing the N and C termini. While larger barrels with eight or more strands are common, small up-and-down barrels are surprisingly rare in nature despite their simple topology; at the time of writing this manuscript, we were not aware of any structures with b6 topologies in the Protein Data Bank (PDB).

## Results

**Blueprint-Based Design.** We first explored the design of small beta barrels using a Rosetta “blueprint” approach similar to that used to design larger eight-stranded beta barrels (7). Blueprint maps specifying the secondary structure and their lengths, ABEGO backbone torsion angle bins (12, 13), and backbone hydrogen bonds were used to build sequence agnostic barrel-like backbones using Rosetta Monte Carlo fragment assembly starting from an extended chain. A key feature of the structure blueprints is the hydrogen bond pairing between the beta strands, which defines the overall topology and shear number ( $S$ ), the shift in strand registry between the first and last strands (14, 15).  $S$  determines the tilt of the strands with respect to the central axis of the barrel, the barrel diameter, and the core packing arrangement. Small beta barrel structures in nature typically have  $S \cong 2n$  where  $n$  is the number of strands. For the SH3 and OB designs, blueprints were constructed with  $S = 8$  or  $10$  based on representative PDB structures (*SI Appendix, Fig. S1*). For b5 and b6 designs, we experimented with a range of blueprints by generating structures through Rosetta fold assembly calculations, and blueprints which did not give rise to the targeted 3D structures were discarded. For the b5 blueprints, parallel pairing of strands 1 and 5 completed a barrel with  $S = 8$ . Initial b5 backbone generation attempts failed to complete hydrogen bond strand pairing of the last hairpin due to the bend required for the last strand. To address this issue, we incorporated additional beta-bulges in the blueprint which allowed the pairing to continue along the bend (*Fig. 1E*). Strand pairing was also improved with the addition of beta-bulges for b6 designs. DALI (16) and TM-align (17) searches of the PDB with the generated b5 and b6 backbone structures did not identify matches with similar topologies; the top hits were SH3 and larger beta barrels of lipocalin-like topologies with the exception of a match (TM-score of 0.57 and rmsd of  $C\alpha$  atoms of 2.58 Å) to chain I of the yeast mitochondrial large subunit (18), which has long loops connecting the strands. For all topologies, blueprint variations were made in the beta-turns using the following common ABEGO turn types observed in native small barrel-like folds: GG (canonical type I'  $\beta$ -turn), AA (type I  $\beta$ -turn), EA (type II'  $\beta$ -turn), and AAG ( $\beta$ -turn with intrinsic G1 bulge). Additional variations were made in loop regions and the lengths of N-terminal helix barrel capping structures. Distance constraints were also included to help maintain the relative position of N-terminal helix and loop capping structures if present (*SI Appendix, Fig. S2*).

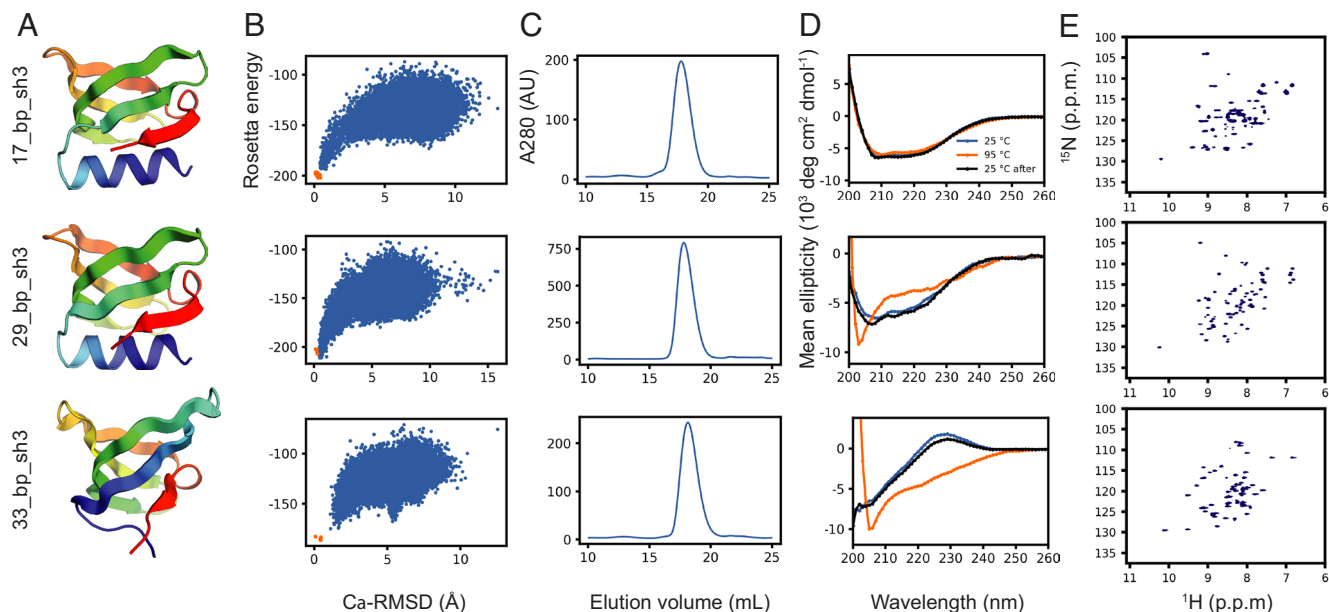
For each blueprint, we extensively sampled structure and sequence space by running up to 10,000 independent trajectories of backbone generation followed by Rosetta sequence design calculations (*Materials and Methods*) on Rosetta@home (<https://boinc.bakerlab.org/rosetta>). Designs were selected for Rosetta energy landscape calculations based on design quality metrics (*Materials and Methods*) and chosen for in vitro

biophysical characterization based on the extent to which the energy landscape funneled into the designed structure using the ff\_metric (19). Sequence similarity to native proteins in the NCBI nonredundant protein database was not detectable using BLAST with an E-value threshold of 0.1 for all designs selected for experimental characterization. Synthetic genes encoding the designs were ordered and the proteins were expressed in *Escherichia coli* and purified using nickel column affinity chromatography. 72 out of 80 designs expressed, 43 were soluble, and 19 and 9 appeared to be monomers and dimers by size exclusion chromatography (SEC), respectively. Eleven monomers and five dimers had far-UV circular dichroism (CD) spectra suggesting folded structures, and 11 designs had close to the expected number of well-dispersed sharp peaks from  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR spectroscopy [6 SH3, 3 OB, 1 b6, and 1 b5; 2 (1 OB and 1 b5) appeared to be dimers according to SEC].

We succeeded in solving solution NMR structures of two of the monomeric designs, which had the expected number of peaks in  $^1\text{H}$ - $^{15}\text{N}$  HSQC spectra (29\_bp\_sh3 and 33\_bp\_sh3; *Fig. 2*). Both solution structures matched closely to the computational designs with rmsds of 1.2 Å and 2.3 Å for 29\_bp\_sh3 and 33\_bp\_sh3, respectively (*Fig. 3*). Although the blueprints for these designs were based on native structures, there were substantial designed structural differences that were recapitulated in the solved structures. The helix in 29\_bp\_sh3 was shifted inward toward the rest of the structure with an average change in distance of 5.1 Å among the first 5  $C\alpha$  atoms compared with the native structure from which the blueprint was obtained (1KQ1). This part of the helix is involved in intermolecular interactions within the 1KQ1 homohexameric assembly, and shifting the helix likely improved monomer packing. The second turn in 33\_bp\_sh3 was shortened by one residue compared with the blueprint source (3P8D) through the replacement of a three residue turn with AAG backbone torsion angle bins to EA, which forms a type II'  $\beta$ -turn.

We were also able to obtain a high-resolution crystal structure of a third design, 17\_bp\_sh3, which closely matched the designed model with an rmsd of 1.6 Å (*Fig. 3*). This design was monomeric as confirmed by SEC-MALS (*SI Appendix, Fig. S3A*) and had high thermal stability with no change in CD spectra at 95 °C (*Fig. 2*). There were nearly double the expected number of NMR peaks at 10 °C and broadened peaks closer to the expected number at 25 °C suggesting exchange of states possibly between monomer and dimer. In the crystal structure, two three-stranded sheets from two monomers come together to form a six-stranded sheet, which may correspond to the transient state observed by NMR. While overall the crystal structure is very similar to the design model, in the third hairpin consisting of strands 3 and 4 (the “distal loop” in SH3 terminology), strand 4 is involved in the antiparallel inter-strand paired dimer interface causing the hairpin to shift outward (*SI Appendix, Fig. S4*).

As mentioned in the introduction and apparent in the dimeric crystal structure, all-beta structures have been difficult to design likely due to their tendency to oligomerize through intermolecular strand pairing. Indeed, 74 percent of the soluble experimentally characterized designs were oligomers, and 40 percent of the expressed designs aggregated. To investigate sequence and structure features enriched in soluble monomeric designs, we used array-based oligonucleotide synthesis to synthesize genes encoding tens of thousands of designs and screened for folding and oligomerization state using recently developed high-throughput assays. A total of 13,990 blueprint designs were made with varying combinations of the common ABEGO turn types described above and first screened for protease resistance using a previously described yeast surface proteolysis assay (3), followed by a recently developed



**Fig. 2.** Experimental characterization of blueprint designs 17\_bp\_sh3, 29\_bp\_sh3, and 33\_bp\_sh3. (A) Design models. (B) Rosetta energy landscapes. The blue points represent the lowest energy structures from independent Rosetta ab initio structure prediction trajectories starting from an extended chain. The red points represent refined models using the Rosetta FastRelax protocol starting from the design model. The X-axis is the  $C\alpha$  rmsd from the design model and the Y-axis is the Rosetta all-atom energy. (C) Size exclusion chromatograms of purified designs. (D) Far-UV circular dichroism spectra at 25 °C (blue), 95 °C (red), and back to 25 °C from 95 °C (black). (E) Two-dimensional  $^1H$ - $^{15}N$  HSQC NMR spectra at 25 °C.

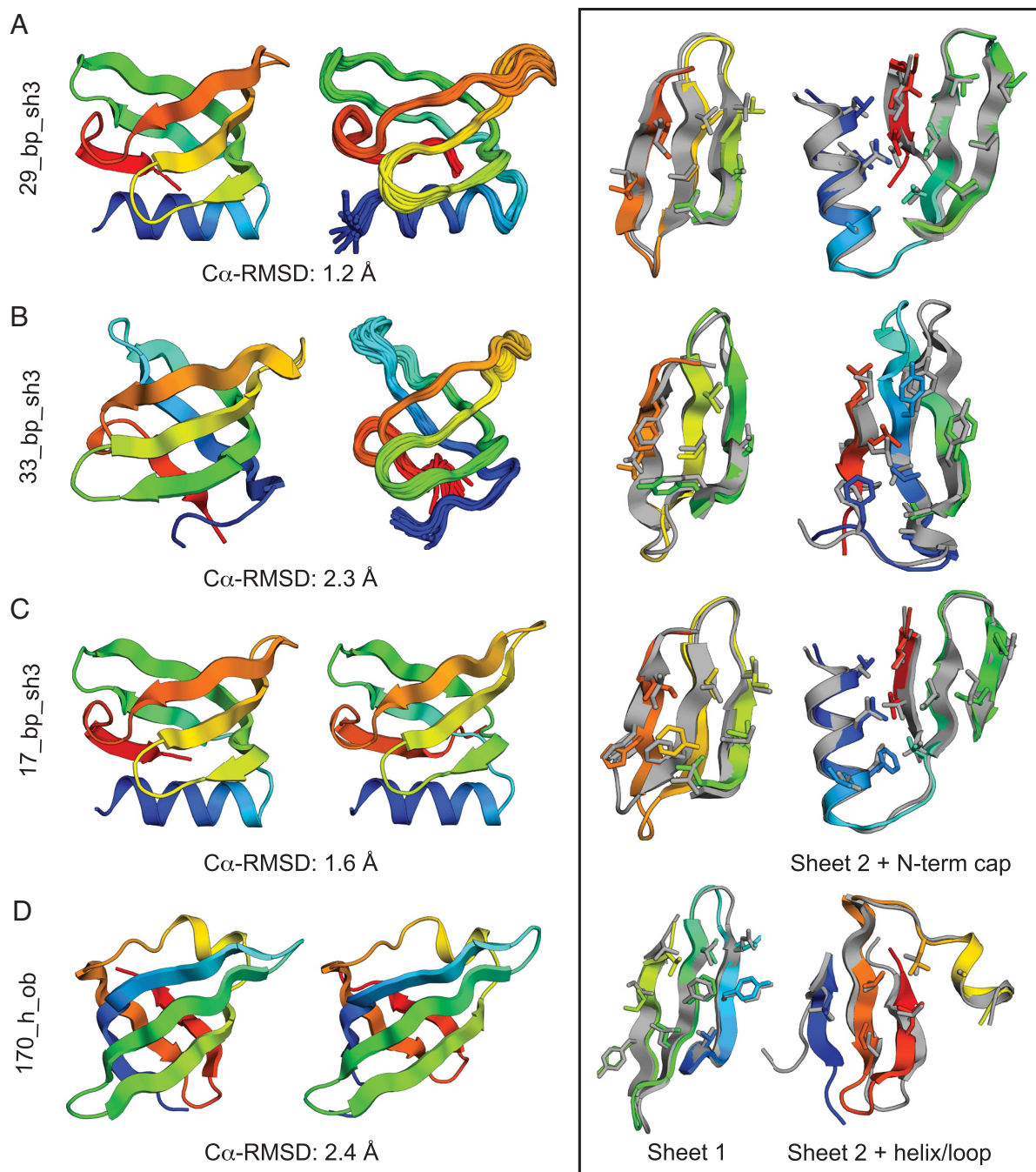
secondary screen that can identify soluble monomeric designs in a complex mixture using SEC and mass spectrometry (MS) barcode technology (20, 21); See *SI Appendix, Supplementary text and Materials and Methods* for detailed descriptions of both assays). In a previous study (3), iterative design-and-build test cycles using the protease assay resulted in systematic improvement in designs of ~40 residue all alpha and alpha/beta structures based on a linkage between protease resistance and proper folding and stability. However, only six percent of the small beta barrel designs were found to be protease resistant, and experimental characterization of the top protease-resistant designs in each fold class (see *SI Appendix, Supplementary text* for details) suggested the screen has a selection bias toward oligomerization for all beta proteins (77 percent of the top protease-resistant soluble designs were oligomers). Despite this selection bias, in the subsequent solution MS barcode screen, we were able to identify 99 protease stable designs (85 SH3, 12 b6, 1 b5 and 1 OB) in SEC fractions that were within the expected elution volume range for their monomeric states. From these designs, we selected 19 for in vitro biophysical characterization and all 19 expressed and were soluble, 16 appeared to be monomers (11 SH3, 4 b6, and 1 b5) as suggested by monodisperse SEC peaks with monomeric elution profiles (three were polydisperse, two with a likely dimer shoulder and another with a broad peak spanning both monomer and dimer fractions), and 17 had CD spectra suggesting structure. We selected six designs for  $^1H$ - $^{15}N$  HSQC NMR spectroscopy, and all six had spectra suggesting folded structure (4 b6, 1 b5, and 1 SH3); five had close to the expected number of well-dispersed sharp peaks at 25 °C and 1 (b6) had well-dispersed sharp peaks at 37 °C.

With an 84 percent success rate for soluble monomer expression, and 74 percent confirmed by CD, the results highlight the power of the MS barcode approach to identify soluble monomeric designs from large libraries, which should be useful in future efforts to design all-beta folds and proteins that tend to oligomerize more generally. Comparison of the sequences and structures of the 99 soluble monomeric designs to the full set of experimentally characterized designs revealed that the soluble monomeric designs on

average have improved local sequence-structure mapping, more favorable Lennard–Jones interactions, and reduced spatial aggregation (SAP) scores (22) (*SI Appendix, Fig. S6*). Filtering designs based on these criteria should increase future design success rates.

**Design by Deep Network Hallucination.** While the Rosetta-based design work described above was in progress, there were concurrent developments in de novo design using deep network hallucination. In this approach, a deep neural network trained for structure prediction is used to guide optimization of an initial random sequence into a sequence predicted to fold into some structure. As originally described, the target structure was not specified in advance, and this “free hallucination” approach generated a wide range of folded structures (23). While folded monomeric proteins were obtained for all alpha and alpha/beta proteins, no folded proteins were obtained for all beta hallucinations. Since our interest here was in small beta barrel structures, we adapted the hallucination procedure to generate structures with this topology by incorporating into the loss function a second term, which favors the formation of specific structural elements (Fig. 4); in the beta barrel case a particular arrangement of the beta strands. For this “constrained hallucination” approach, we used trRosetta which predicts interresidue distances and orientations [the more accurate AlphaFold2 (24) and RoseTTAFold (25) had not yet been developed]. Sequence optimization was carried out by back propagating gradients assessed on the distances back to the input sequences (26).

Constrained hallucination has the advantages of not only requiring less computing to generate high-quality structures versus the Rosetta blueprint methods but also simplifying the design of diverse structures. The blueprint approach requires specifying the topology, secondary structure lengths, and other structural features for each design while having to follow rules consistent with known protein structures; however, this is not necessary with constrained hallucination since the design rules are largely encoded in the network, albeit in the form of millions of parameters that are not readily human interpretable. To investigate whether constrained

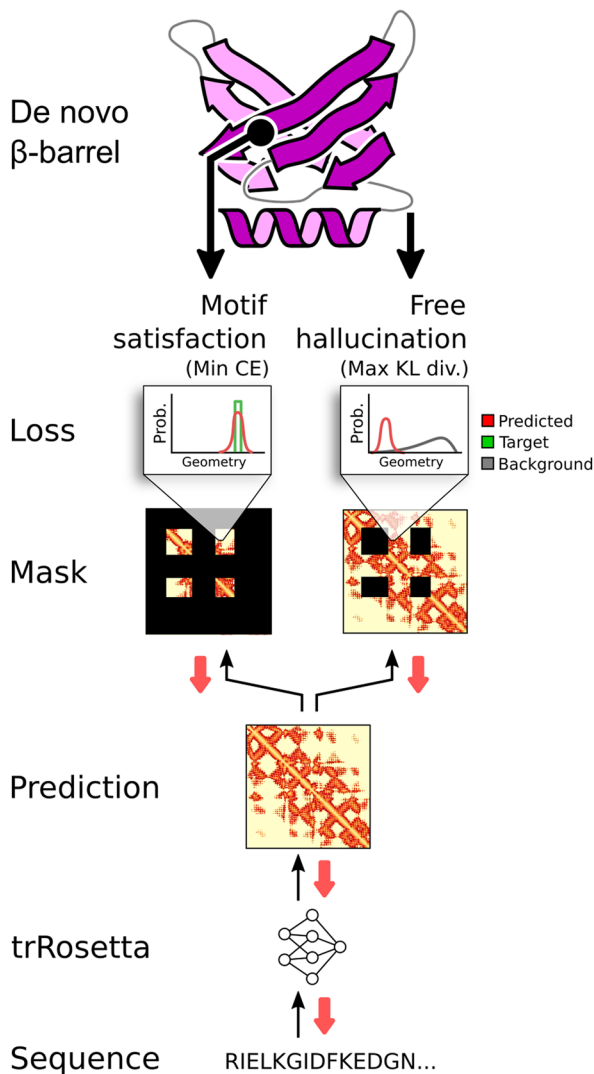


**Fig. 3.** Comparison of design models (far Left) to NMR structures (A and B) and crystal structures (C and D). The box shows a comparison of core side-chain packing from design models (rainbow) superimposed with determined structures (gray). The C $\alpha$  rmsds of the design models to the determined structures are noted.

hallucination could be used to design a structurally diverse set of high-quality barrel-like structures, we focused on hallucinating the turn and loop regions from blueprint designs and native structures with varying lengths (*Materials and Methods*). A total of 100 random-length hallucinations were sampled for each run for a total of 66,000 trRosetta distance and orientation outputs. The results were organized by groups, according to fold class and hallucinated length signature (e.g., 1KQ1\_L5L8L6L4L6L1 corresponding to the hallucinated lengths from N to C terminus for the SH3-like 1KQ1 fold), and for each group, the hallucination with lowest total loss and highest accuracy (how close the trRosetta distances were to the input distances) was selected for 3D structure energy minimization resulting in 13,984 hallucinated designs with unique length signatures within each fold class. Producing such

diversity in unique loop lengths using structure blueprints would be quite difficult in comparison.

Previous studies (26) have suggested that using trRosetta for structure generation followed by Rosetta for sequence design can generate more funneled energy landscapes than Rosetta alone and deeper free energy minima than trRosetta alone due to the strengths of Rosetta in modeling atomic-level detailed packing interactions in the folded structure. We thus redesigned the sequences of the hallucinations using Rosetta FastDesign (27). Since the success rates of the blueprint designs were low, we first focused on improving the sequence design protocol to optimize design quality features. One feature that particularly stood out as being suboptimal compared with native monomeric proteins of similar topology and size was the amount of buried nonpolar surface area (NPSA) from



**Fig. 4.** Protein fold generation by constrained hallucination. Initially, a random sequence is passed to trRosetta, which predicts the six degrees of freedom between the backbones of all pairs of residues. The predictions are masked into two regions. One contains all intramotif residue pairs over which the mean cross entropy between the predicted geometric distributions and the desired geometric distributions is calculated (*Left side*). For all other residue pairs, the mean KL divergence between background distributions and the predicted distributions is calculated (*Right side*). The final loss is the sum of these two values. The sequence is updated by backpropagating the gradient of the loss with respect to the sequence logits (red arrows). This process is repeated until the loss converges. In this work, the core beta strands were selected as the fixed structure motif, and the loops were freely hallucinated.

hydrophobic amino acids, which is an important determinant of folding stability (3). The average buried NPSA per residue for the monomeric native set (*SI Appendix, Selection of Native Structures*) was  $62 \text{ \AA}^2$  compared with 53 for the experimentally characterized designs. The lower values compared with native structures resulted from disallowing hydrophobic amino acids on the solvent exposed surface by the Rosetta LayerDesign protocol (28). To shift the buried NPSA distribution closer to the native distribution, we modified the layer settings to allow high beta-sheet propensity hydrophobic amino acids, Ile, Val, Tyr, and Trp, on the surface layer of strands, and biased the frequencies of these amino acids along with Thr, Met, and Phe to match the frequencies observed in the native set using the Rosetta residue type composition score term [*aa\_composition* (29)]; we did not include Met and Phe in the modified layer settings since their average frequencies on strand

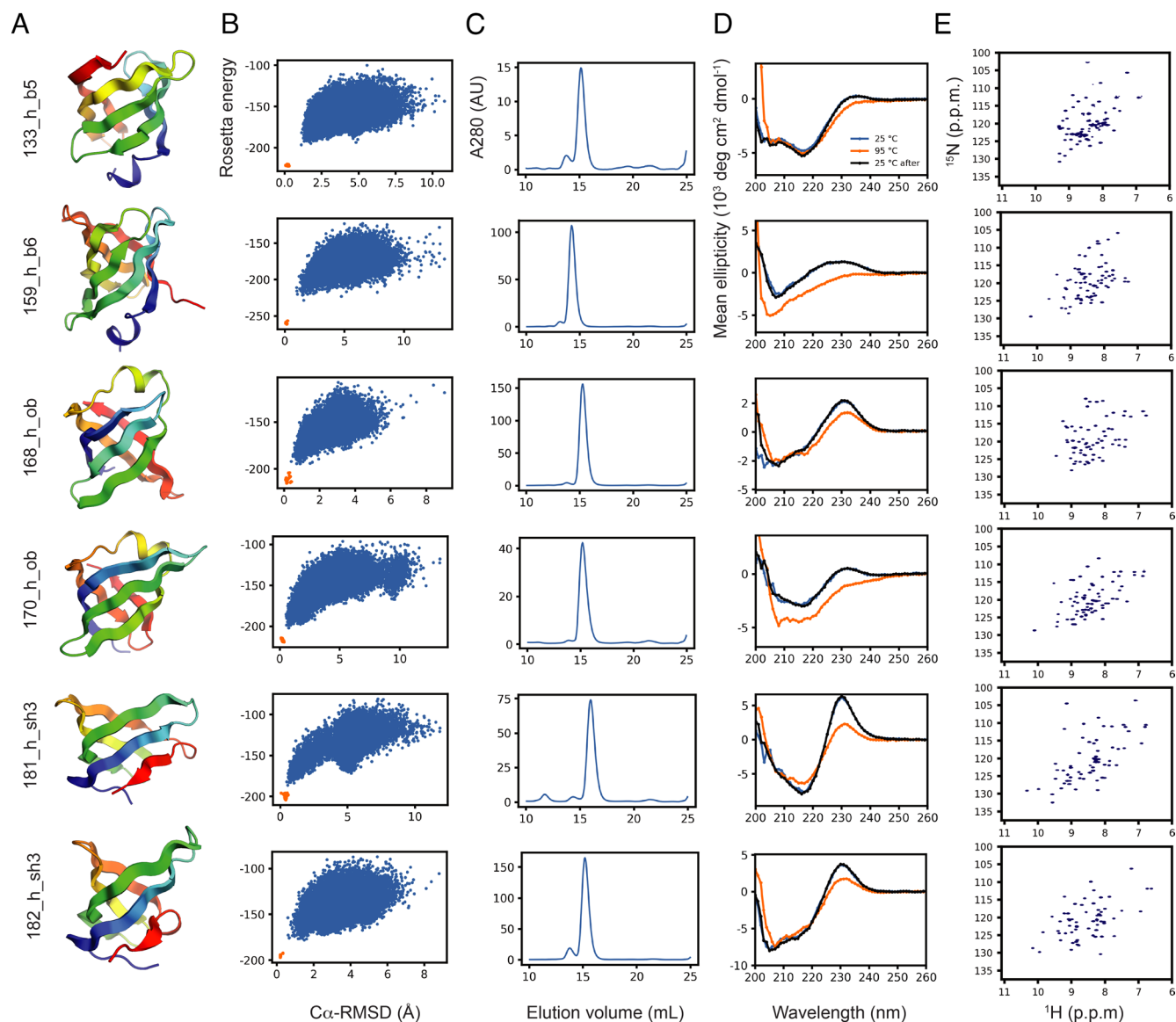
surfaces in the native set were significantly lower at 0.22 and 0.26, respectively, versus 0.60, 0.69, 0.52, and 0.72 for Ile, Val, Tyr, and Trp, respectively]. We also used fragment-based position-specific scoring matrices [PSSMs (19)] to improve local sequence to structure compatibility.

Multiple FastDesign variations were run for each hallucinated design and then filtered based on design quality metrics (*Materials and Methods*). The selected designs had a buried NPSA distribution closer to the native distribution (*SI Appendix, Fig. S8*) with an average value of  $59 \text{ \AA}^2$ , and the average Rosetta energy per residue improved to  $-3.6 \text{ kcal/mol}$  from  $-1.9$  compared with the original hallucinated designs. Rosetta energy landscape calculations were then run for each design, and 71 were selected for experimental characterization based on the quality of their energy landscape funnels using the *ff\_metric*. Seventy designs expressed, 68 were soluble, and 28 and 27 were likely monomers and dimers according to SEC with single monodisperse peaks, respectively. Of note, 22 out of the 27 designs characterized by CD spectroscopy had spectra suggesting structure, and 24 (9 OB, 7 SH3, 5 b6, and 3 b5) out of 26 designs characterized by  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR spectroscopy (*SI Appendix, Fig. S12*) had close to the expected number of well-dispersed sharp peaks. Biophysical characterization results of six hallucinated design examples are shown in Fig. 5. Most of the successful designs that were monomeric and folded based on SEC and HSQC were SH3 and OB folds (7 OB, 7 SH3, 3 b6, and 2 b5). Comparing sequence and structure features of the monomeric designs with well-resolved HSQC spectra to the full set of designs generated with FastDesign (45,761 designs) revealed enrichment of features similar to those in the MS barcode screen: increased hydrophobicity, more favorable Lennard Jones and solvation interactions, and improved local sequence–structure compatibility (*SI Appendix, Fig. S7*).

We were able to obtain a high-resolution crystal structure of a hallucinated OB fold, 170\_h\_ob, with a  $C\alpha$  rmsd of  $2.38 \text{ \AA}$  from the designed model (Fig. 3D); the hallucinated length signature for this design was L3L6L2L13L4L1). 29 out of the 59 backbone residues in the designed model were unconstrained during hallucination. Most of the model matched closely to the crystal structure with a  $C\alpha$  rmsd of  $0.82 \text{ \AA}$  spanning 45 residues; however, there were significant differences in the first four residues of the N terminus and ten residues spanning the last hairpin turn and strand in the C terminus. Since the N-terminal residues are polar and largely solvent exposed, the deviations in this region are not surprising. The differences in the C terminus involved a strand register shift caused by a designed beta-bulge missing in the crystal structure. These structural deviations may in part be due to the additional linker and histidine tag residues added to the expression construct.

## Discussion

The range of computational and experimental approaches used in this study is a testament to the recent blossoming of de novo protein design. On the computational front, Rosetta blueprint-based design and deep learning hallucination highlight the contrast between traditional physically (and human expert) based design and deep learning approaches in which the physical principles and expertise are replaced by the millions of parameters of the network—the tradeoff is loss of physical transparency for a gain in procedural simplicity and accuracy. On the experimental front, traditional biophysical characterization, while providing detailed structural and thermodynamic information, is intrinsically low throughput as the methods (NMR and X-ray structure determination in particular) are time and resource intensive, and recently



**Fig. 5.** Experimental characterization of constrained hallucination designs. (A) Design models. (B) Rosetta energy landscapes as described in Fig. 2 legend. (C) Size exclusion chromatograms of purified designs. These designs were run on an FPLC system with a relative elution offset of around  $-2.5$  mL compared with the system used for the blueprint designs in Fig. 2. (D) Far-UV circular dichroism spectra at 25 °C (blue), 95 °C (red), and back to 25 °C from 95 °C (black). (E) Two-dimensional  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR spectra at 25 °C.

developed yeast display proteolysis and MS barcoding approaches enable the evaluation of a much larger range of designs albeit at considerably lower resolution. Here, we summarize the major lessons from this combination of approaches, both about beta barrel design and about the strengths and weaknesses of the approaches themselves, and describe the opportunities opened up by the capability of designing small beta barrels with high accuracy and stability.

First, our results provide insight into the relative strengths and weaknesses of traditional protein design using physically based models such as Rosetta, and deep learning approaches. The Rosetta blueprint-based approach provides considerable control over backbone geometry and topology, but design rules need to be understood in advance for successful blueprints, which makes the overall design process more difficult. For this reason, blueprints have been primarily used to design relatively simple and idealized structures, and it can be tedious to sample considerable structural diversity as blueprints must be made for each case. Rosetta fragment-based assembly for blueprints with considerable nonlocal

interactions can be compute time intensive, as are the full-atom design and relaxation steps in the design end game. In contrast, the machine learning-based hallucination strategy explored here does not provide as precise control over backbone geometry, placement of atoms, and molecular interactions but readily provides more diverse solutions satisfying the overall design conception. The deep learning model used to guide the design process essentially has prior knowledge of the local structural features that promote structure formation, for example the connections/turns/hairpins between strand pairs and N-term capping motifs in the context of a beta barrel-like topology. The best results came from a combination of methods: using hallucination for backbones and Rosetta for sequence design and structure relaxation. Roughly 90 percent of these designs that were characterized with NMR were folded compared with around 30 percent of the blueprint designs (See Table 1 for the results' comparison). Recent work from our group suggests that ML methods may soon take over the sequence design step as well. Overall, while the deep learning approaches are simpler to use because of the huge amount of information

**Table 1. Experimental results' comparison of Rosetta blueprint designs, including those from protease stability and mass spectrometry (MS) barcode screens, and deep learning-based hallucinated designs.**

Designs	Expressed	Soluble	Monomer or dimer	Folded 25 °C (CD)	Folded 95 °C (CD)	Folded 25 °C (HSQC)
bp*	90% (72/80)	54% (43)	35% (28)	62% (16/26)	8% (2)	31% (11/35)
bp ps†	88% (21/24)	54% (13)	17% (4)	67% (2/3)	0% (0)	n/a
bp ps + MS barcode	100% (19/19)	100% (19)	95% (18)	89% (17/19)	11% (2)	83% (5/6)
Hallucinated	99% (70/71)	96% (68)	77% (55)	81% (22/27)	11% (3)	92% (24/26)

\*blueprint (bp) designs.

†protease stable (ps) designs.

about protein structure and protein sequence-structure relationships inherent in the models, they have the considerable disadvantage that the information guiding design is spread over the millions of parameters of the networks; in contrast the Rosetta blueprint approach and design methodology, while more cumbersome, makes all of the concepts feeding into the design explicit in the blueprint, and the details of the structure are determined by the physically transparent force field.

Second, our combination of experimental evaluation by high-throughput methods and more detailed individual biophysical characterization provides insights into the strengths and limitations of the different approaches. We explored two high-throughput methods here: a primary protease stability screen using yeast display followed by a secondary SEC-MS screen using peptide barcodes. The protease stability screen has the advantage of the direct link of genotype to phenotype on yeast, allowing next-generation sequencing to be used as a tremendously scalable readout. However, we observed a bias toward designs that oligomerize when expressed individually; protease resistance particularly for beta proteins may arise not only from folding but also from aggregation on the yeast surface. The SEC-MS approach has advantages over the protease stability screen in directly assessing solubility and oligomerization state in solution, but fewer variants can be tested in parallel due to the reduced sensitivity and multiplexing capacity of mass spectrometry compared with next-generation DNA sequencing. While we did not do a direct head-to-head comparison, we can compare the fractions of proteins tested using traditional more intensive biophysical methods resulting from the two screens that were folded. The largest differences were in the fraction of soluble designs and their oligomeric state; all 19 hits from the SEC-MS screen were soluble and 18 were monomers or dimers, whereas only 13 out of 24 hits from the protease stability screen were soluble and only four were monomers or dimers. Five out of the six SEC-MS screened designs that were characterized with <sup>1</sup>H-<sup>15</sup>N HSQC NMR were folded and 16 out of 19 were potentially folded monomers or dimers based on CD. With the improved success rate of deep learning-based designs, using SEC-MS as a primary screening tool may suffice.

Third, our results provide insight into the determinants of folding and design of small beta barrels. Inclusion of strategically placed beta bulges was necessary to achieve the strand bending required to form the b5 and b6 folds. Because of the small size of the core, incorporation of surface hydrophobic residues favored folding by increasing the overall hydrophobic burial through interactions between adjacent surface residues and resulted in lower Rosetta energies and likely increased stability. We observed more favorable Lennard-Jones interactions, more hydrophobic burial, and more optimal local sequence-structure relationships in both the soluble monomeric designs generated using Rosetta and in the hallucinated designs with well resolved NMR HSQC spectra compared with the overall population of experimentally characterized designs. A challenge in designing small beta barrels

is how to balance hydrophobic burial and packing while maintaining the backbone hydrogen bonds necessary for closing the barrel; our results suggest that increasing hydrophobic interactions contributes to stability and folding provided that they can be accommodated by the packing and hydrogen-bonding constraints of the barrel.

Fourth, our results provide insight into the extent to which the sampling of folds in nature reflects their inherent robustness or evolutionary history. We succeeded in the de novo design of all four of the target topologies, but overall had higher success rates on the SH3 and OB folds, which are widely sampled in nature, than the much rarer b5 and b6 folds. Successful designs of the b6 fold suggest that there is nothing in principle that disallows this topology. Overall, our lower success rates for b5 and b6 designs could reflect an intrinsically lower folding robustness or reduced rigidity of these folds, the difficulty of closing a barrel with a small number of strands, a greater tendency of all beta structures with short loops like b5 and b6 to oligomerize, or perhaps the fewer natural examples to learn from.

The ability to design small beta barrel-like folds should considerably enhance protein binder design. Their surfaces can tolerate significant sequence variation, and considerable backbone structural diversity can be sampled in the hairpins and turns (as for example in the long RT and n-src loops of SH3 domains, which play functional rather than structural roles). Together with the concave surface of the protein, well suited for peptide and protein binding, and the small size of the proteins which enables encoding on large scale oligo arrays, these designs complement the largely helical protein scaffold sets currently used for protein interface design. Toward this end, we provide here a scaffold set of 12,911 mini-barrels designed using constrained hallucination in the supplementary dataset archive (<https://doi.org/10.5281/zenodo.6529943>), which have a high probability of folding based on AlphaFold2 metrics (pLDDT > 88 and model rmsd < 2.0) shown to be accurate indicators for design success within our group.

## Materials and Methods

### Rosetta Blueprint Backbone Generation, Sequence Design, and Selection.

**Backbone generation.** 3D beta barrel-like protein backbones were constructed using a blueprint method described previously (1, 7, 30). Barrel-like topologies were generated using Rosetta Monte Carlo fragment assembly starting from an extended chain and guided by blueprints defining the secondary structure and ABEGO backbone torsion angle bins for each residue, and distance and angle constraints defining backbone hydrogen bonds. For SH3 and OB topologies, blueprints and constraints were initially extracted from native structures (*SI Appendix, Fig. S1*) and then updated to modify loops and turns for structural diversity. Design modifications included changing beta-hairpin turns to the common types observed in small native barrel-like folds GG (canonical type I' β-turn), AA (type I β-turn), EA (type II' β-turn), and AAG (β-turn with intrinsic G1 bulge) (7), not restricting ABEGO values in a five residue stretch of the long SH3 RT-loop in 1ZUY, replacing a loop crossing the top of the barrel between strands 3 and 4 in 4A21 with a 9 or 10 residue helix, and shortening a long hairpin loop between

strands 4 and 5 in 1C4Q to an AAG turn. b5 and b6 blueprints and constraints were manually constructed from scratch and improved for barrel assembly through trial and error by incorporating beta-bulges (Fig. 1E), backbone hydrogen bond constraints, and modifying strand and N-terminal helix lengths. Some b5 designs that were experimentally characterized did not have a closed barrel structure including hallucinated designs.

**Sequence design.** The Rosetta flexible backbone sequence design protocol, FastDesign (27), was used for each generated backbone, and the amino acids allowed at each position were restricted using the LayerDesign protocol, which classifies residue positions as either "core," "boundary," or "surface" layers according to their degree of burial within the design structure. Nonpolar amino acids were only allowed in the core layer and excluded from the surface layer. Rosetta resfiles were used to constrain beta-turn sequences to preferred profiles observed in native barrel structures as previously described (7). The amino acid identity was also constrained for key residues in N-terminal capping structures if present; for example, the third position in 3P8D-based designs was forced to be phenylalanine, which contributes to the protein core (SI Appendix, Fig. S2). A subset of designs were also generated without resfiles. For b5 and b6 designs, manual changes to blueprints, constraints, and resfiles were also made attempting to incorporate a core tyrosine or tryptophan in the first strand making a sidechain to backbone hydrogen bond to the N-terminal capping helix, similar to a tyrosine corner (31) commonly found in native beta barrels. Although such designs were selected for experimental characterization, none that were screened with HSQC NMR were folded.

The complete protocol including backbone generation followed by sequence design was coded in a Rosetta script file, and designs were generated using either Rosetta@home (<https://boinc.bakerlab.org/rosetta>) or an in-house computing cluster. The *beta\_nov16* version of the Rosetta full-atom energy function (32) and the *MonomerDesign2019* (33) relax script were used for all sequence designs with the exception of a subset of designs, which used the legacy relax script. The *MonomerDesign2019* relax script was slightly modified by removing the use of coordinate constraints (constraints that favor the starting input structure) for all Rosetta@home designs since the script feature was not available in the distributed Rosetta application. Some design variations were run on in-house clusters using the original *MonomerDesign2019* relax script and fragment-based sequence profiles (19) using *StructProfileMover* to improve local sequence to structure compatibility.

**Selection.** Designs were filtered using Rosetta scores divided by protein length *score\_per\_res*, *rama\_per\_res*, *omega\_per\_res*, *hbond\_sr\_bb*, and *hbond\_lr\_bb*, and filter metrics, which evaluate the agreement between atoms and local structure (*mismatch\_probability*) and the buried surface area of all atoms in hydrophobic residues (FAMILYVW) divided by protein length (*buried\_npsa\_per\_res*). Rosetta@home designs with Rosetta scores less than  $-50$  were initially filtered using the criteria: *score\_per\_res*, *rama\_per\_res*, and *omega\_per\_res* less than their average values, *mismatch\_probability*  $< 0.35$ , *buried\_npsa\_per\_res*  $> 50$ , and the sum of *hbond\_sr\_bb* and *hbond\_lr\_bb* divided by protein length  $< -0.75$ . This initial selection reduced the number of designs for compute intense "forward folding" energy landscape calculations, which involve large-scale sampling of Rosetta ab initio structure prediction trajectories starting from an extended chain. Final designs were selected based on the tendency for the energy landscapes to funnel toward the designed structures using the *ff\_metric* (19), an algorithm that evaluates the funnel by calculating the sum of rmsd in the lowest energy points. Designs were also selected through careful manual inspection of potential designs and their quality metrics.

**Protein Expression and Purification.** Genes encoding blueprint designs were synthesized and cloned into the pET29b+ vector by Integrated DNA Technologies (IDT). Hallucinated designs were synthesized as eBlock gene fragments (IDT), cloned into the pET29b+ vector using Golden Gate assembly, and verified with Sanger sequencing (GENEWIZ). Plasmids were transformed into chemically competent Lemo21(DE3) *E. coli* (NEB) and protein expression was induced overnight in Studier autoinduction medium with 30  $\mu\text{g}/\text{mL}$  kanamycin at 37 °C. Cells were lysed by sonication in the presence of DNase and protease inhibitors, and designs were purified using immobilized metal affinity chromatography (IMAC) with Ni-NTA affinity resin (Qiagen). Hallucinated designs were initially expressed for small-scale size exclusion chromatography (SEC) screens in a 2 mL 96-deep well round-bottom plate, and cells were lysed using BugBuster (Millipore). Purification proceeded in a 96-well (800  $\mu\text{L}/\text{well}$ ) plate with a 25- $\mu\text{m}$  polyethylene frit (Agilent

Technologies) loaded with 75  $\mu\text{L}/\text{well}$  Ni-NTA affinity resin. For 50 and 500 mL scale expressions, the eluted purified protein was exchanged and concentrated into 50 mM NaPi (50 mM NaCl, pH 6.5), 20 mM NaPi (50 mM NaCl, pH 7.4), or 20 mM Tris buffer (100 mM NaCl, pH 8.0) using Amicon Ultra 3 kDa concentrators. Proteins were further purified by Akta Pure FPLC (GE Healthcare) SEC using a Superdex 75 increase 10/300 GL column (SI Appendix, Fig. S9). Based on calibration data, 15- to 17-mL fractions were presumed to be monomeric and 13- to 15-mL fractions dimeric. Monodispersed peaks that were likely monomeric or dimeric were collected and kept at 4 °C for further characterization experiments within a week or frozen for later analysis. Protein concentrations were determined by absorbance at 280 nm using a NanoDrop spectrophotometer (ThermoScientific) with predicted extinction coefficients (34).

See SI Appendix, Supplementary Materials and Methods for biophysical characterization, and high-throughput screens and analysis.

#### Constrained Hallucination and Rosetta Sequence Redesign.

**Input motifs and hallucination strategy.** Structural motifs used as input for constrained hallucination were taken from 188 structures consisting of the 104 experimentally characterized blueprint designs, 61 protease stable designs consisting of up to the top 10 stable designs for each fold group, 13 additional b5 designs, and 10 native structures. To maintain a barrel-like topology, we kept the central four residues of the beta strands and, if present, the N-terminal helix as input backbone segments while the rest of the structure was hallucinated with random lengths within plus or minus two residues from the original lengths. We also tried shorter input lengths of two to three central residues, but the resulting hallucinations were lower in quality with less regular secondary structure and in some cases different topologies. For folds with an N-terminal helix, we also ran hallucinations with the region between the helix and first strand included as an input backbone segment to preserve the helix capping structure. To design longer insertions such as extended beta-hairpins, we hallucinated the loops and turns with random lengths up to six residues or for loops with six or more residues, the original length plus one. The residue spans specifying segments for hallucination and their corresponding length spans to randomly sample were defined in a single command line argument.

**Sequence representation during optimization.** We represented the sequence being optimized as a one-hot-encoded multiple-sequence alignment (MSA)  $X \in \mathbb{R}^{N \times L \times A}$ , where  $L$  is the length of the first sequence in the alignment and the protein whose structure is being predicted,  $N$  is the number of aligned sequences, and  $A = 21$  is the number of amino acids plus gap character (although we do not use gaps during design), as previously described (26, 35). trRosetta accurately predicts the structure of de novo proteins even when only one sequence is used as input ( $N = 1$ ), so previous work on unconstrained protein design optimized a single sequence (23). When designing proteins to recapitulate structural constraints derived from natural backbones, however, better accuracy was obtained when we also provided a PSSM  $Y \in \mathbb{R}^{L \times A}$ . Normally when trRosetta is passed a full MSA, it calculates a PSSM feature that is the single-site amino acid frequency at each position. Since the PSSM potentially contains more information that could aid in structure prediction but we only pass a single sequence, we sought to mimic the PSSM by reinterpreting the logits of the sequence being optimized. The PSSM was the softmax of the sequence logits.

**Loss function.** The loss function was based on our previous work (35) and was composed of two parts,

$$\mathcal{L} = w_M \mathcal{L}_M + w_H \mathcal{L}_H,$$

where  $\mathcal{L}_M$  is the "motif" loss, which captures how well specific structural elements are recapitulated in the design, and  $\mathcal{L}_H$  is the "free hallucination" loss, which measures how well the sequence encodes the backbone shape.  $w_M$  and  $w_H$  are weights for each loss term.

For a protein of length  $L$ ,  $\mathcal{L}_M$  was calculated as the average categorical cross-entropy (CCE) of the desired interresidue geometry (one-hot encoded) to the network's predicted geometry  $p(y)$ .

$$\mathcal{L}_M = - \sum_{y \in \{d, \omega, \theta, \phi\}} \left[ \left( \sum_{i=1}^L \sum_{j \neq i}^L m_{ij} \log p(y_{ij} = y_i^0) \right) / \mathcal{A} \left( \sum_{i=1}^L \sum_{j \neq i}^L m_{ij} \right) \right],$$

$$\text{where } m_{ij} = \begin{cases} 1, & ||C\beta_i - C\beta_j|| \leq cce\_cutoff \text{ and } i, j \in \text{motif} \\ 0, & \end{cases}$$

$y \in \{d, \omega, \theta, \phi\}$  represents residue-residue distances and orientation angles and  $y^0$  is the value of the distance or angle in the reference motif. The features  $\theta$  and  $\phi$  are asymmetric and so all six interresidue degrees of freedom are encoded over  $y_{ij}$  and  $y_{ji}$ , which are both always present in the symmetric mask  $m$ . The cross entropy is only calculated over specific regions, denoted by the mask  $m$ , corresponding to the placement of the motifs in the primary sequence (defined at the beginning of a design run). The mask only included residues whose  $C\beta$  were within a cutoff distance of each other, typically 10 Å. The maximum cutoff distance ever used was 20 Å, as that is the maximum distance of trRosetta predictions.

The hallucination loss was identical to the previously published “free hallucination” loss (23), except that it was averaged only over residues not covered by the motif loss.

$$\mathcal{L}_M = - \sum_{y \in \{d, \omega, \theta, \phi\}} \left[ \sum_{i=1}^L \sum_{j \neq i}^L (1 - m_{ij}) KL(p(y_{ij}) || b(y_{ij})) \right] / 4 \left( \sum_{i=1}^L \sum_{j \neq i}^L (1 - m_{ij}) \right)$$

where the KL divergence over  $k$  geometric features is defined by:

$$KL(p || q) = \sum_k p_k \log(p_k / q_k),$$

and  $b(y_{ij})$  is a background distribution calculated by a separate background network and specific to the protein length. Essentially, it is the geometric distributions trRosetta predicts when given a scrambled sequence, which contains no real structural information, thus any predicted geometries that diverge from this background do so because of extra information provided by the input sequence. Minimizing this term maximizes how strongly the sequence encodes the predicted structure. See previous work on free hallucination approaches for more details (23).

**Optimization method.** We used gradient descent with a constant learning rate to optimize the MSA tensor, as previously described (26). To backpropagate gradients through the discrete one-hot sequence representation to the continuous logits, we used a reparameterization trick as previously described (26, 35–38).

**Rosetta sequence redesign and selection.** As noted in the main text, we sought to improve the hallucinated designs by redesigning the sequences using Rosetta FastDesign with LayerDesign settings allowing high beta-sheet propensity hydrophobic amino acids on the surface layer of strands, using PSSMs derived from Rosetta fragments, and biasing the composition of Ile, Val, Tyr, Trp, Thr, Met, and Phe to more closely match frequencies and counts observed in natives. Instead of using PSSMs directly generated from the structure profile mover (*StructProfileMover*) in Rosetta, the positional amino acid counts were used to generate PSSMs that were adjusted with BLOSUM62 (39) pseudocounts and background frequencies. In addition, 26 FastDesign variations (*SI Appendix, Table S8*) with *aa\_composition* score weights of 1.0 and 3.0 were run on Rosetta@home for each of the 13,984 hallucinated designs described in the main text. Some of the variations were run with standard LayerDesign settings, and an additional variation did not use the *aa\_composition* score. The variations consist of different amino acid composition biases and either allowing or disallowing Trp in the surface and boundary layers for LayerDesign.

Our strategy to select designs for experimental characterization consisted of five steps. First, we selected Rosetta@home designs using Rosetta filter metrics with the criteria: *pack*  $\geq 0.6$ , *cavity\_volume* = 0, *sbuns* = 0, *buns\_all*  $\leq 10$ ,

*buried\_npsa\_per\_res*  $\geq 50$ , and *score\_per\_res*  $\leq -3.0$ , and grouped the designs based on their fold class and hallucinated length signature. Second, for each group, we selected up to the top 100 scoring designs based on *score\_per\_res*, and then in the third step, we filtered the designs using: *pack*  $\geq 0.65$ , *mismatch\_probability*  $\leq 0.5$ , *worst9mer*  $\leq 0.5$ , *sap\_score*  $\leq 30$ , and *score\_per\_res*  $\leq$  the average *score\_per\_res*. In the fourth step, we used a logistic regression model described in *SI Appendix, Supplementary Materials and Methods* to select designs that were predicted to be protease stable with greater than 0.5 probability. Rosetta energy landscape calculations were run for these designs and then in the last step, we filtered the designs using: *ff\_metric*  $\leq 20$ , RoseTTAFold predicted IDDT  $\geq 0.8$ , and the SD of rmsds to the design after sampling refinement trajectories with FastRelax  $< 0.1$ . Up to the top 10 ranked designs based on *ff\_metric* for each fold were selected for experimental characterization.

**Data, Materials, and Software Availability.** The structures for 29\_bp\_sh3 and 33\_bp\_sh3 have been deposited in the Protein Data Bank ([www.wwpdb.org](http://www.wwpdb.org)) with PDB IDs 7UWY and 7UWZ, and the Biological Magnetic Resonance Bank (BMRB) with BMRB IDs 31017 and 31018, and the structures for 17\_bp\_sh3 and 170\_h\_ob have been deposited with PDB IDs 7UR7 and 7UR8. The code to run constrained hallucinations is publicly available on the *sokrypton/TrDesign\_partialhal* GitHub repository ([https://github.com/sokrypton/TrDesign\\_partialhal](https://github.com/sokrypton/TrDesign_partialhal)). PDB models of the blueprint and hallucinated designs, including the scaffold set of 12,911 mini-barrel designs are available in supplementary dataset archive (<https://doi.org/10.5281/zenodo.6529943>). Commands, input files, and selection scripts for the designs, protease stability screen results, LASSO regression model and feature data, MS barcode library results, and native PDB sets are also available in the archive. All study data are included in the article and/or *SI Appendix*.

**ACKNOWLEDGMENTS.** We thank all Baker lab members, past and present, who have helped with this study particularly Anastassia A. Vorobieva, Ivan Anishchenko, Gyu Rie Lee, T. J. Brunette, Hugh Haddock, Brian Koepnick, Brian Coventry, and Longxing Cao for insightful discussions and suggestions, IA for help with model quality predictions, HH for help with the protease stability assay analysis, Chris Norn for developing the method to adjust Rosetta fragment PSSMs using BLOSUM62 pseudocounts and background frequencies, and Luki Goldschmidt and Patrick Vecchiato for computer and network infrastructure support. We also thank staff at the UW Institute for Protein Design (IPD) for initial help with protein production and the UW Biofabrication Center for running protease stability assay experiments. We acknowledge Rosetta@home volunteers and the Texas Advanced Computing Center at The University of Texas at Austin for providing computing resources, and staff at the Advanced Photon Source Northeastern Collaborative Access Team beamlines, funded by the National Institute of General Medical Sciences from the NIH grant P30 GM124165 and U.S. Department of Energy (DOE) contract DE-AC02-06CH11357. This work was supported with funds provided by the HHMI (D.E.K. and D.B.), a Microsoft gift (D.T. and D.B.), Audacious Project at the IPD (X.L., A.K.B., and D.B.), the Defense Advanced Research Projects Agency Synergistic Discovery and Design project HR001117S0003 contract FA8750-17-C-0219 (C.M.C., A.S., and D.B.), the Open Philanthropy Project Improving Protein Design Fund (L.C., H.N., and D.B.), the Human Frontier Science Program Cross Disciplinary Fellowship (LT000395/2020-C, L.M.) and European Molecular Biology Organization (EMBO) Non-Stipendary Fellowship (EMBO ALTF 1047-2019, L.M.), the DOE, Office of Science, grant DE-SC0018940 (A.K. and D.B.), the National Cancer Institute grant R01CA240339, the NIH grant DP5OD026389 (S.O.), the National Institute on Aging grant 5U19AG065156 (D.F. and D.B.), and the NIH grants R37 AI058072 and S10 OD020000 (B.F.V., F.C.P., and D.R.J.).

- N. Koga *et al.*, Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
- P.-S. Huang *et al.*, De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016).
- G. J. Rocklin *et al.*, Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
- B. Basanta *et al.*, An enumerative algorithm for de novo design of proteins with diverse pocket structures. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 22135–22145 (2020).
- T. J. Brunette *et al.*, Exploring the repeat protein universe through computational protein design. *Nature* **528**, 580–584 (2015).
- X. Pan, T. Kortemme, Recent advances in de novo protein design: Principles, methods, and applications. *J. Biol. Chem.* **296**, 100558 (2021).
- J. Dou *et al.*, De novo design of a fluorescence-activating  $\beta$ -barrel. *Nature* **561**, 485–491 (2018).
- E. Marcos *et al.*, De novo design of a non-local  $\beta$ -sheet protein with high stability and accuracy. *Nat. Struct. Mol. Biol.* **25**, 1028–1034 (2018).
- A. A. Vorobieva *et al.*, De novo design of transmembrane  $\beta$  barrels. *Science* **371**, eabc8182 (2021).
- W. Wang, M. H. Hecht, Rationally designed mutations convert de novo amyloid-like fibrils into monomeric beta-sheet proteins. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 2760–2765 (2002).
- J. S. Richardson, D. C. Richardson, Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 2754–2759 (2002).
- R. T. Wintjens, M. J. Rooman, S. J. Wodak, Automatic classification and analysis of alpha alpha-turn motifs in proteins. *J. Mol. Biol.* **255**, 235–253 (1996).
- Y.-R. Lin *et al.*, Control over overall shape and size in de novo designed proteins. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E5478–E5485 (2015).
- A. D. McLachlan, Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* **128**, 49–79 (1979).
- A. G. Murzin, A. M. Lesk, C. Chothia, Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis. *J. Mol. Biol.* **236**, 1369–1381 (1994).

16. L. Holm, Using dali for protein structure comparison. *Methods Mol. Biol.* **2112**, 29–42 (2020).
17. Y. Zhang, J. Skolnick, TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
18. A. Amunts *et al.*, Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343**, 1485–1489 (2014).
19. T. J. Brunette *et al.*, Modular repeat protein sculpting using rigid helical junctions. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 8870–8875 (2020).
20. P. Egloff *et al.*, Engineered peptide barcodes for in-depth analyses of binding protein libraries. *Nat. Methods* **16**, 421–428 (2019).
21. P. C. Havugimana *et al.*, A census of human soluble protein complexes. *Cell* **150**, 1068–1081 (2012).
22. N. Chennamsetty, V. Voynov, V. Kayser, B. Helk, B. L. Trout, Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 11937–11942 (2009).
23. I. Anishchenko *et al.*, De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
24. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
25. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
26. C. Norm *et al.*, Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2017228118 (2021).
27. G. Bhardwaj *et al.*, Accurate de novo design of hyperstable constrained peptides. *Nature* **538**, 329–335 (2016).
28. S. J. Fleishman *et al.*, RosettaScripts: A scripting language interface to the Rosetta macromolecular modeling suite. *PLoS One* **6**, e20161 (2011).
29. P. Hosseinzadeh *et al.*, Comprehensive computational design of ordered peptide macrocycles. *Science* **358**, 1461–1466 (2017).
30. P.-S. Huang *et al.*, RosettaRemodel: A generalized framework for flexible backbone protein design. *PLoS One* **6**, e24109 (2011).
31. J. M. Hemmingsen, K. M. Gernert, J. S. Richardson, D. C. Richardson, The tyrosine corner: A feature of most greek key  $\beta$ -barrel proteins. *Protein Sci.* **3**, 1927–1937 (1994).
32. R. E. Pavlovicz, H. Park, F. DiMaio, Efficient consideration of coordinated water molecules improves computational protein-protein and protein-ligand docking discrimination. *PLoS Comput. Biol.* **16**, e1008103 (2020).
33. J. B. Maguire *et al.*, Perturbing the energy landscape for improved packing during computational protein design. *Proteins* **89**, 436–449 (2021).
34. C. N. Pace, F. Vajdos, L. Fee, G. Grimsley, T. Gray, How to measure and predict the molar absorption coefficient of a protein. *Protein Sci.* **4**, 2411–2423 (1995).
35. D. Tischer *et al.*, Design of proteins presenting discontinuous functional sites using deep learning. bioRxiv [Preprint] (2020). <https://doi.org/10.1101/2020.11.29.402743> (Accessed 19 January 2022).
36. J. Linder, G. Seelig, Fast activation maximization for molecular sequence design. *BMC Bioinform.* **22**, 510 (2021).
37. N. Bogard, J. Linder, A. B. Rosenberg, G. Seelig, A deep neural network for predicting and engineering alternative polyadenylation. *Cell* **178**, 91–106.e23 (2019).
38. J. Linder, G. Seelig, Fast differentiable DNA and protein sequence optimization for molecular design. arXiv [cs.LG] [Preprint] (2020). <https://doi.org/10.48550/arXiv.2005.11275> (Accessed 18 January 2022).
39. S. Henikoff, J. G. Henikoff, Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10915–10919 (1992).