PROTEIN DESIGN

Scalable protein design using optimization in a relaxed sequence space

Christopher Frank[1,2], Ali Khoshouei[1,2], Lara Fuβ[1,2], Dominik Schiwietz[1,2], Dominik Putz[1,2], Lara Weber[1,2], Zhixuan Zhao[3], Motoyuki Hattori[3], Shihao Feng[4], Yosta de Stigter[1,2], Sergey Ovchinnikov[5,6]*, Hendrik Dietz[1,2]*

Machine learning (ML)–based design approaches have advanced the field of de novo protein design, with diffusion-based generative methods increasingly dominating protein design pipelines. Here, we report a "hallucination"-based protein design approach that functions in relaxed sequence space, enabling the efficient design of high-quality protein backbones over multiple scales and with broad scope of application without the need for any form of retraining. We experimentally produced and characterized more than 100 proteins. Three high-resolution crystal structures and two cryo–electron microscopy density maps of designed single-chain proteins comprising up to 1000 amino acids validate the accuracy of the method. Our pipeline can also be used to design synthetic protein-protein interactions, as validated experimentally by a set of protein heterodimers. Relaxed sequence optimization offers attractive performance with respect to designability, scope of applicability for different design problems, and scalability across protein sizes.

The landscape of protein design has been fundamentally transformed by machine learning (ML) methods (*1*–*10*). Structural prediction networks such as AlphaFold2 (AF2) (*11*), ESMFold (*12*), and RoseTTAFold2 (*13*) enable a variety of protein design tasks by accurately predicting protein structures from input sequences (*11*–*13*), which enables filtering of candidate designs. Generative models based on diffusion and flow matching (*6*, *8*, *14*–*18*), such as RFDiffusion (*19*) and Chroma (*20*), have gained substantial popularity for their ability to create de novo protein designs for various design tasks (*19*, *20*).

An alternative approach considers leveraging structure prediction networks such as AF2 through iterative sequence evolution in a process that was previously coined "hallucination" (*3*, *4*, *7*). However, slowly converging random search algorithms (*3*, *4*, *7*) and challenges with implementing robust gradient descent–based optimization in a discrete sequence space (*21*) have hindered expanding this approach to more complex protein design tasks. Here, we hypothesized that gradient descent–based hallucination toward target objectives could be improved by operating beyond the confines of discrete (i.e., physically realistic) protein sequence space to facilitate smoother and more direct optimization transitions (Fig. 1A).

To implement this "relaxed hallucination" process, we expanded upon prior work that facilitated backpropagation through the AF2 network (*21*, *22*). In this framework, a sequence is input into the network and a loss based on a target objective is calculated using the resulting predicted structure. The loss is then backpropagated with respect to the input sequence and a gradient is obtained. This gradient is used to update the input sequence toward the target objective. Updating the sequence with the obtained gradient usually does not produce a one-hot–encoded sequence, but rather a logit-like or position-specific scoring matrix (PSSM). These "relaxed" representations are physically unrealistic because each residue position is populated seemingly by a superposition of all the 20 amino acids, each with a specific numeral weight. Previous methodologies commonly forced the updated, relaxed sequence back into a real-world one-hot–encoded sequence representation by applying argmax operations (*21*, *23*), causing substantial deviations away from the optimal gradient direction (Fig. 1A, top). In our approach, which we call relaxed sequence optimization (RSO), we directly return the updated relaxed sequence back into the structure prediction network (Fig. 1A, bottom, and movies S1 and 2) and iterated until convergence.

RSO exhibited rapid and stable convergence (fig. S1A and movies S1 to S4) and improved performance relative to previous protocols (*3*, *21*). By incorporating loss functions to numerically measure the differences between predicted designs and targets, RSO thus enables rapid prototyping for various target properties without the need for retraining. We found that RSO can converge along complex gradients, allowing the design of intricate design problems, including binder design and functional site scaffolding, and enabling the design of large single-chain proteins comprising up to 1000 amino acids (Fig. 1B).

Once RSO is converged, our pipeline discards the relaxed sequence and feeds the converged backbone geometry to the protein message-passing neural network (ProteinMPNN) module (*5*) to generate candidate protein sequences for the converged backbone geometry (Fig. 1C). ProteinMPNN is a key component because it was specifically trained to design protein sequences that will also fold experimentally into a given backbone structure. The ProteinMPNN-generated sequences are handed to structure prediction networks such as ESMFold or AF2 to repredict structures, which are then tested for agreement with the converged backbone geometry initially produced by RSO. This pipeline facilitates the swift engineering of de novo–designed proteins by simply adjusting the loss function to address user-defined design tasks.

Computational benchmarking

To evaluate the quality of RSO-designed backbones, we conducted designability tests (*9*, *19*) by generating sequence sets with ProteinMPNN for RSO-designed backbones and repredicting the encoded structure using ESMFold. We assessed similarity to the initial RSO backbones by calculating the root mean square deviation (RMSD) and the template-modeling (TM) score (*24*) and then selecting the best matching sequence.

We RSO-designed protein chains of increasing lengths, from 100 amino acids up to 1000 amino acids, with 100 candidate backbones per length. Our loss function optimized for confidence, a small radius of gyration, and a high number of intrachain contacts, and reduced helical content. For each backbone design, we created eight ProteinMPNN candidate sequences (using the soluble weights) (*25*), which were then (re)predicted with ESMFold (Fig. 2, A to C) and AF2 single sequencing (fig. S1B).

In our tests, RSO produced designs with a lower RMSD (indicating better matching) relative to RFDiffusion, particularly for larger protein sizes, as evaluated by ESMFold (Fig. 2B). RSO successfully generated promising designs for proteins as large as 1000 amino acids with median TM scores of 0.89 for 1000 amino acids (Fig. 2C). Although RFDiffusion performed well for smaller proteins, it faced challenges with producing viable backbones beyond 600 amino acids. In these tests, we also observed a trend in which the RMSD for larger proteins increased when using AF2 single sequence for repredicting (fig. S1B), whereas this trend was absent in ESMFold repredictions. By supplementing information about the target backbone as an

[1]Laboratory for Biomolecular Nanotechnology, Department of Biosciences, School of Natural Sciences Technical University of Munich, 85748 Garching, Germany. [2]Munich Institute of Biomedical Engineering, Technical University of Munich, 85748 Garching, Germany. [3]State Key Laboratory of Genetic Engineering, Shanghai Key Laboratory of Bioactive Small Molecules, Collaborative Innovation Center of Genetics and Development, Department of Department of Physiology and Neurobiology, School of Life Sciences, Fudan University, Yangpu District, Shanghai 200438, China. [4]Changping Laboratory, Beijing 102200, China. [5]Faculty of Applied Sciences, Harvard University, Cambridge, MA, USA. [6]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA.
*Corresponding author. Email: so3@mit.edu (S.O.); dietz@tum.de (H.D.)

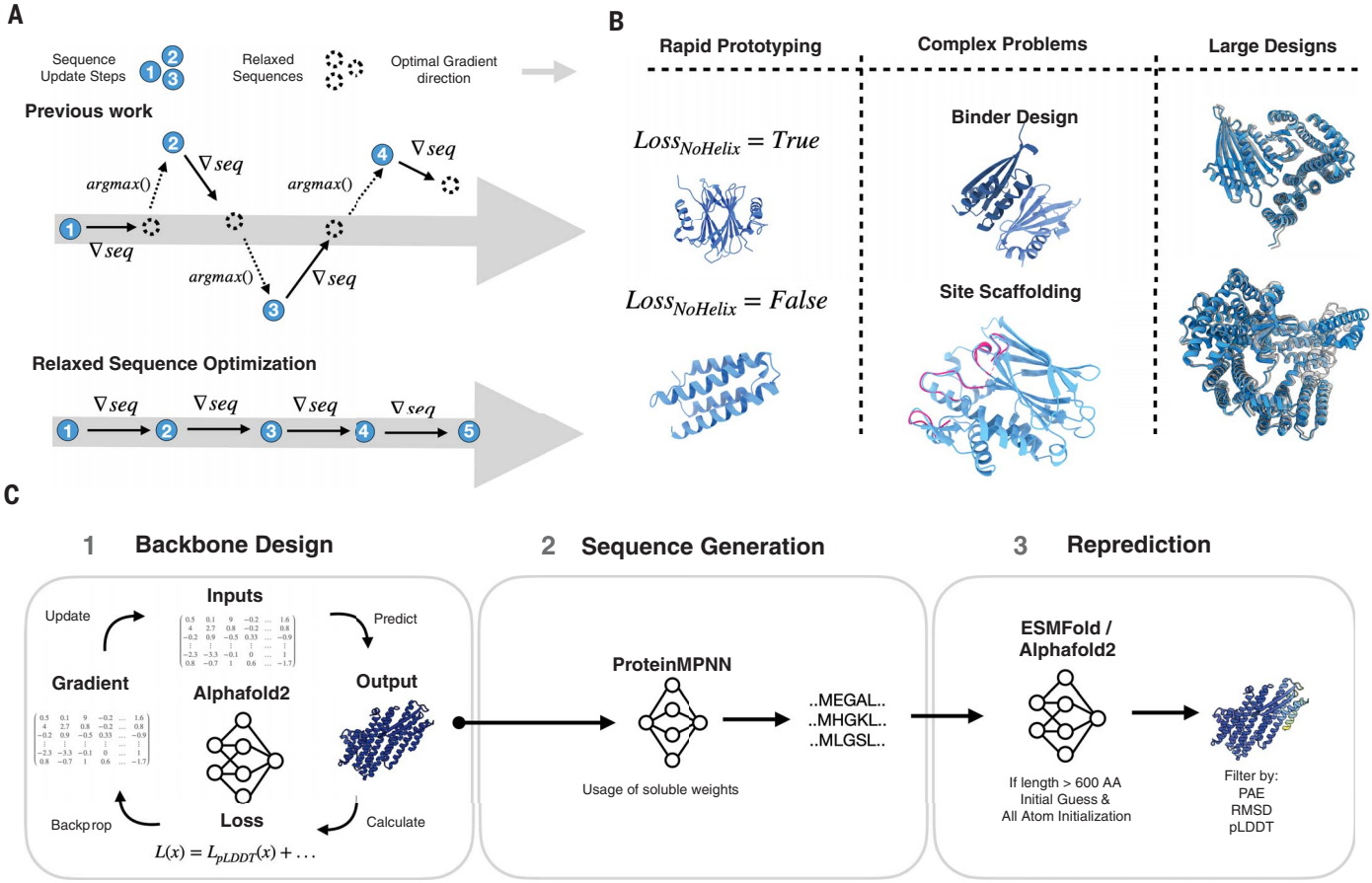

Fig. 1. Schematics of the protein design pipeline. (**A**) Schematic representation how the free gradient descent in RSO enables an efficient search for minima of the loss function. (**B**) Exemplary design tasks that can be accomplished using the RSO method. (**C**) Schematic view of the design process consisting of backbone design, sequence generation with ProteinMPNN, and candidate design filtering with ESM-Fold/AF2.

initial guess (*26*, *27*) to AF2's Evoformer module and by invoking "big bang initialization" (*28*), AF2 single sequencing could also (re)predict the larger proteins with improved quality, approaching the level of ESMFold (Fig. 2D).

We also tested whether the RSO relaxed sequences generated together with a converged backbone could be used directly for creating candidate sequences. Simply converting the relaxed sequences into one-hot–encoded sequences using argmax operations resulted in strong deviations of the repredicted structures from the target backbones (fig. S1C). Using more sophisticated approaches such as simulated annealing from the relaxed sequences improved the in silico success rates. This means that the in silico structures repredicted from the candidate sequences matched better to the initial backbone design, but the experimental testing of the sequences produced by annealing from relaxed sequences showed poor success rate (fig. S1D), which is consistent with previous work on AF2-only designed sequences (*3*, *4*). AF2 structure predictions exhibited a high tolerance toward mutations (fig. S1E), which could conversely limit their ability to distinguish between valid and adversarial sequences. The inclusion of a component such as ProteinMPNN, which is specifically constructed for creating valid sequences for a given input backbone, thus improves the overall experimental success.

The RSO-generated proteins are structurally diverse (fig. S1F) and mostly globular (fig. S2A). The addition of a helical loss can reduce a bias toward generating helical secondary structures (Fig. 2E). To investigate the novelty of the proteins, we compared them with the entirety of currently known structures as provided by the Protein Data Bank (PDB) using foldseek (*29*). As in previous methods (*19*), for small proteins, there was significant similarity to existing proteins in the PDB, whereas for larger designs, there were fewer and fewer homologs in the PDB, suggesting that RSO may leverage AF2 generalization beyond the known protein space to create truly new folds (Fig. 2F and fig. S2B). Regarding computational efficiency, the use of back-propagation in RSO increases the time that it takes to complete one backbone design iteration for larger designs relative to RFDiffusion (fig. S2C) (however, note that RFDiffusion had difficulties with creating valid designs beyond 600 amino acids). In terms of the success rate for generating designs with <3 Å RMSD to their target, RSO and RFDiffusion were similarly graphics processing unit (GPU) efficient (fig. S2D).

We also tested RSO on more complex design tasks, including discontinuous site scaffolding using a previously developed scaffolding problem benchmark set (*19*) (Fig. 2G). We used the same settings for all designs and adapted the same success criterion of whole backbone RMSD <2.0 Å and motif RMSD <1.0 Å combined with having high confidence [predicted local-distance difference test (pLDDT) >85]. We chose the maximum length of each designed loop between motifs and kept the full sequence of the scaffolded structural elements constant (one-hot) for simultaneous design of discrete amino acids and relaxed sequences. We used two design approaches, called "fixed" or "free" on the basis of whether the motifs were fixed in three-dimensional space with or without a structural template, respectively. RSO successfully found solutions for all designs tasks

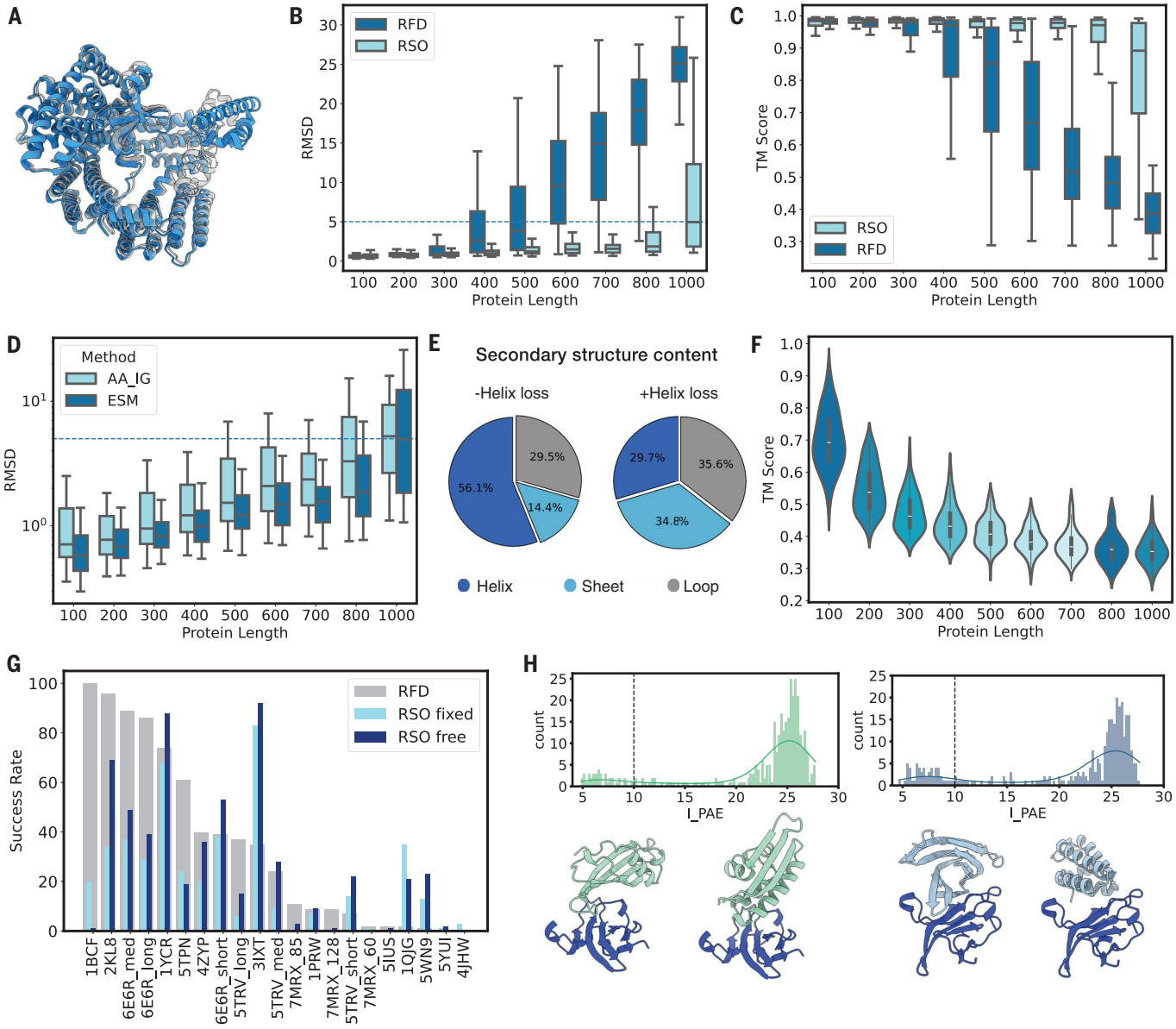

Fig. 2. In silico benchmarking of relaxed sequence optimization. (**A**) Exemplary ESM-Fold structure prediction overlaid to a RSO-designed backbone. (**B** and **C**) Comparison between RSO and RFDiffusion (RFD) for Ca-RMSD and TM score. RFDiffusion was run with 200 denoising steps and RSO with 100 steps. Eight MPNN sequences were generated using "soluble weights," and reprediction was done with ESMFold. (**D**) Boxplot showing how AF2 single sequencing supplemented with initial guess and all-atom initialization (AA_IG) can achieve RMSD values between design and prediction similar to ESMFold (ESM). (**E**) Pie charts showing secondary structure content for backbones generated without or with helix loss. (**F**) Violin plot showing the highest reported TM score of the repredicted proteins from (B) against the PDB. TM scores were calculated using foldseek (*24*). Higher TM values mean higher homology to known protein structures. (**G**) Conditional benchmarking results using the benchmarking set from (*14*). (**H**) Binder design campaigns against the human activin type II A and and B receptor. Histograms show the distribution of the interface-predicted align error (I_PAE) for designed binders. Rendering shows exemplar AF2 predictions of binder candidates bound to the receptor. PDBs used for design were 5NH3 (*31*) and 5NGV (*31*).

(Fig. 2G, fig. S2E, and table S3). The addition of templates ("fixed") reduced the average steps needed to converge toward low-RMSD designs, enabling the production of more candidate designs per unit time. The template-free method yielded designs of similar quality but converged more slowly. Combining a distogram-based loss with a RMSD loss yielded improved performance relative to using a pure frame-aligned point error–based loss (fig. S2F). Overall, RFDiffusion and RSO performed similarly well in these scaffolding problems (Fig. 2G), with success rates varying in a design task–specific fashion.

To test whether RSO can also generate protein-protein binders, we designed binders toward the activin type 2 A and B receptors following a previously reported strategy (*19*, *26*, *30*). After AF2 multimer filtering, we obtained between 10 and 16% binder candidates with an interaction-predicted aligned error (I_PAE) <10 (Fig. 2H), indicating the successful in silico design of promising binder candidates (*19*, *26*, *30*). The binder candidates were structurally diverse and included all beta to beta-alpha mixes or classical helical bundles (fig. S3, A and B). In addition, RSO yielded promising in silico

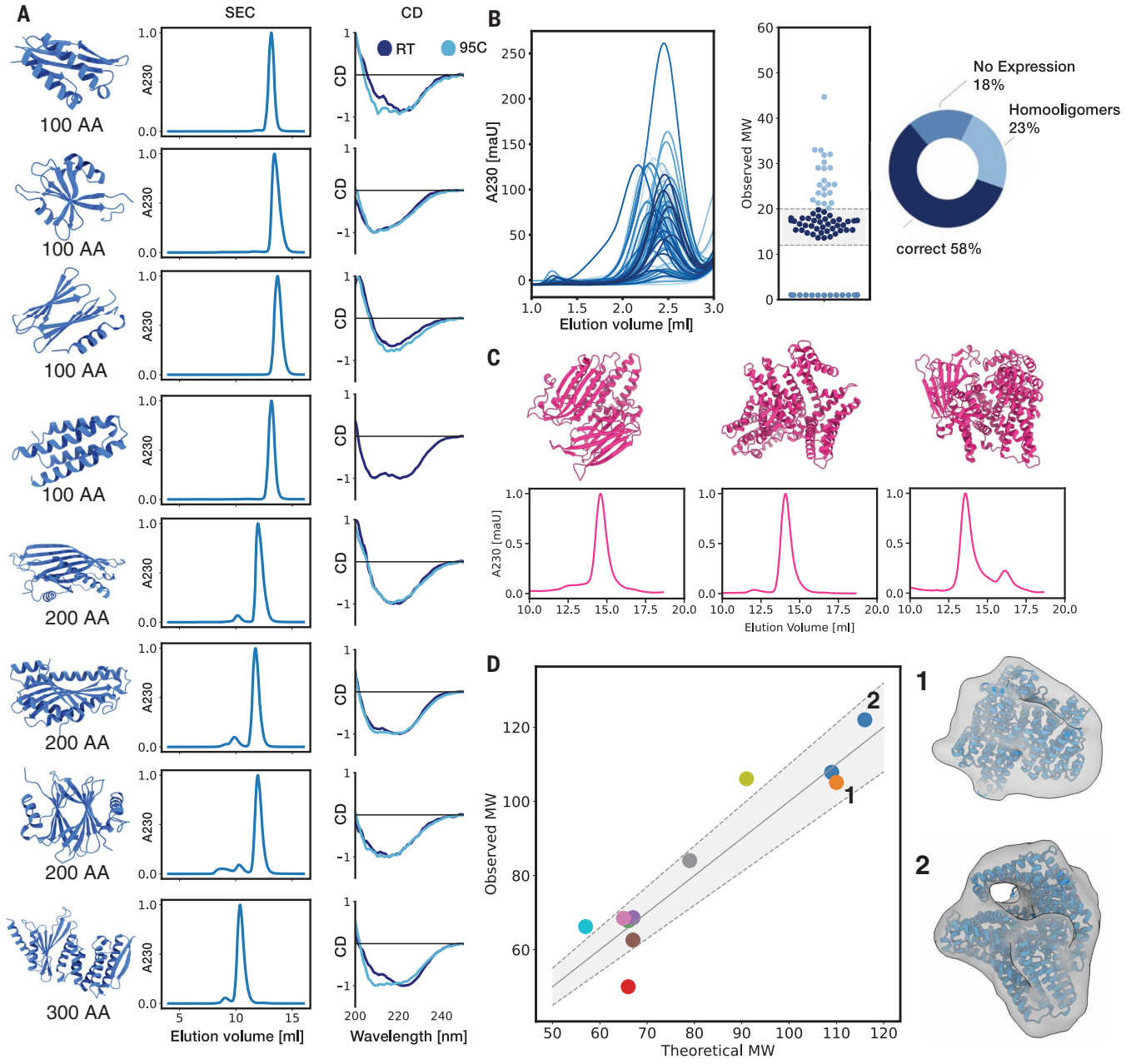

Fig. 3. Experimental characterization of designed monomers. (**A**) Biophysical analysis of designed monomers. Left: protein models representing AF2 predictions. Middle: SEC traces collected with Superdex 75 Increase Resin. Right: circular dichroism spectra collected at room temperature and at 95°C. (**B**) Left: overlay of SEC traces for 76 proteins collected with a Superdex 200 Increase 5/150 column. Middle: molecular weights determined based on the sec elution volumes peak positions. Right: pie chart showing the success rate of the binder design campaign. (**C**) Characterization of large proteins through SEC. Top: AF2-predicted models. Bottom: SEC traces acquired with Superdex 200 Increase 10/300. (**D**) Left: observed versus expected molecular weight (kDa) as obtained through SEC for designs ranging from 500 to 1000 amino acids. Right: nsTEM reconstructions of large monomers overlayed with AF2 model.

binder candidates for other challenging binder design problems, including designing a two-domain connector for the human growth hormone receptor (fig. S3, C and D) and the design of binders to viral surface receptors (fig. S3F).

Experimental validation

Using RSO, we designed 85 proteins with sizes ranging from 100 to 300 amino acids. Nine monomeric proteins were purified exemplarily through affinity chromatography and analyzed using native size-exclusion chromatography (SEC). Eight of those proteins expressed well, showed one predominant peak on SEC, and eluted at the expected fraction given their designed molecular weight (Fig. 3A). Circular dichroism spectroscopy gave characteristic spectra that agreed with the expectation derived from the designed secondary structure content of the proteins (Fig. 3A). The proteins were thermostable up to 95°C, which is consistent with previous reports on the high thermostability of de novo–designed proteins (*19*, *30*, *31*). The remaining 76 proteins were expressed and purified in a higher-throughput fashion adopting previously described strategies (*3*, *19*, *30*); 58% of them had molecular weights matching the

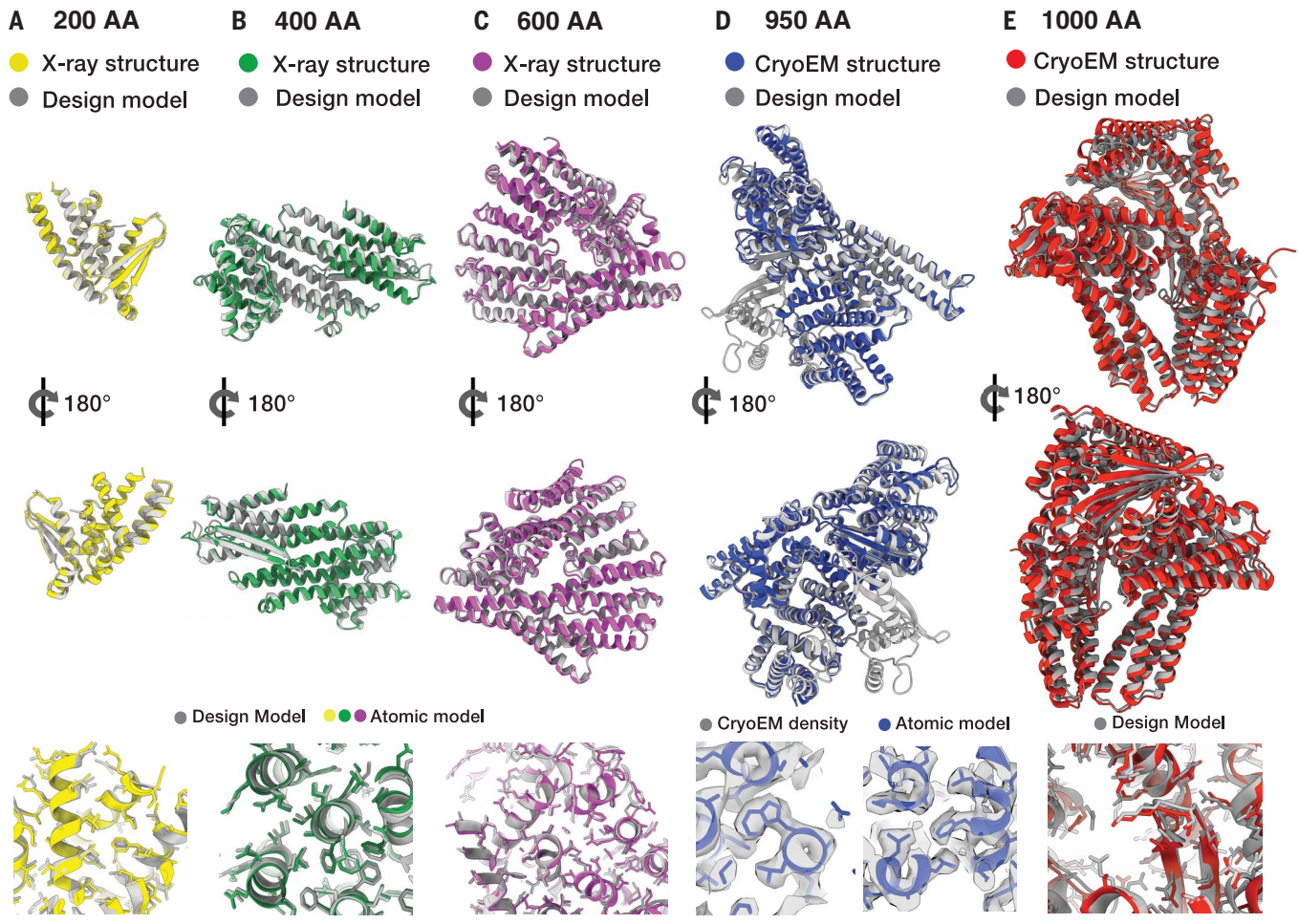

Fig. 4. Structural characterization of designed proteins. (**A** to **C**) Top and middle: overlay of an experimentally determined crystal structure (colored) of a 200–, 400–, and 600–amino acid designed protein with the AF2-predicted structure based on sequence (gray). Bottom: magnified views into the structures. (**D** and **E**) Top and middle: overlay of experimentally determined atomic models from the cryo-EM densities (colored) of a 950– and 1000–amino acid designed protein with the ESMFold-predicted structure based on sequence (gray). (D) Bottom: overlay of the atomic model (blue) into the cryo-EM density map. (E) Bottom: magnified view into the structure. In (A) to (E) for design models, we show the predicted structures from AF2 or ESMFold, whichever matches the initial hallucinated backbone the best (showing lowest RMSD) because this allows us to show the placement of side chains.

expectation from molecular weight–calibrated SEC (Fig. 3B).

We also designed a set of larger proteins comprising 500 to 1000 amino acids. We filtered candidate designs using AF2 supplemented with initial guess (*26*) and big bang initialization (*28*) and/or ESMFold repredict-tion. Although the overall in silico structure of the proteins was mostly correct and predicted with high confidence, a fraction of repredicted candidate designs had minor deviations from the RSO-designed backbone, such as low-confidence loop regions or routing problems in which domains were connected with long, unstructured regions (fig. S4). We constructed genes for 14 of the large candidate proteins; 13 of them expressed and 11 had a dominant peak at the correct molecular weight as seen by SEC (Fig. 3, C and D, and figs. S5, A and B, and S6, A and B). The structures of three of the proteins large enough for analysis using negative-stain transmission electron microscopy (nsTEM) agreed with the designed structure (Fig. 3D and fig. S7).

Structural investigation

We designed and crystallized exemplarily three proteins comprising 200, 400, and 600 amino acids (figs. S5A, S8, and S9, A to G), and obtained crystal structures using x-ray diffraction at 2.2-Å (200 amino acids), 2.1-Å (400 amino acids), and 2.8-Å (600 amino acids) resolutions, respectively (Fig. 4, A to C). The atomic models constructed into the respective electron densities agreed with the predictions of the designed backbones with Ca-RMSD values (reported by TM score) of 0.90 Å (200 amino acids), 1.28 Å (400 amino acids), and 0.92 Å (600 amino acids), respectively. These low RMSD values reflect overall agreement with respect to the side-chain conformations and demonstrate the atomic design accuracy of RSO. Loops predicted with low confidence did show some deviations in the crystal structure relative to the designed geometry. However, loops predicted with high confidence were also reproduced well in the experimental structures (fig. S10, A and B).

We also performed single-particle cryo–electron microscopy (cryo-EM) with two larger candidate designs, having ~100 kDa molecular weight (comprising 950 and 1000 amino acids) (fig. S11, A and B). We determined and refined a high-resolution consensus cryo-EM map for the 950 amino acids design at a resolution of 2.7 Å (fig. S12, A and B), which allowed the out-of-the-box construction of an atomic model using Model Angelo (*32*) (Fig. 4D and fig. S13E). The final, refined atomic model constructed from the experimental cryo-EM data had a Ca-RMSD of 1.08 Å relative to the prediction of the designed structure. Regions in the design that

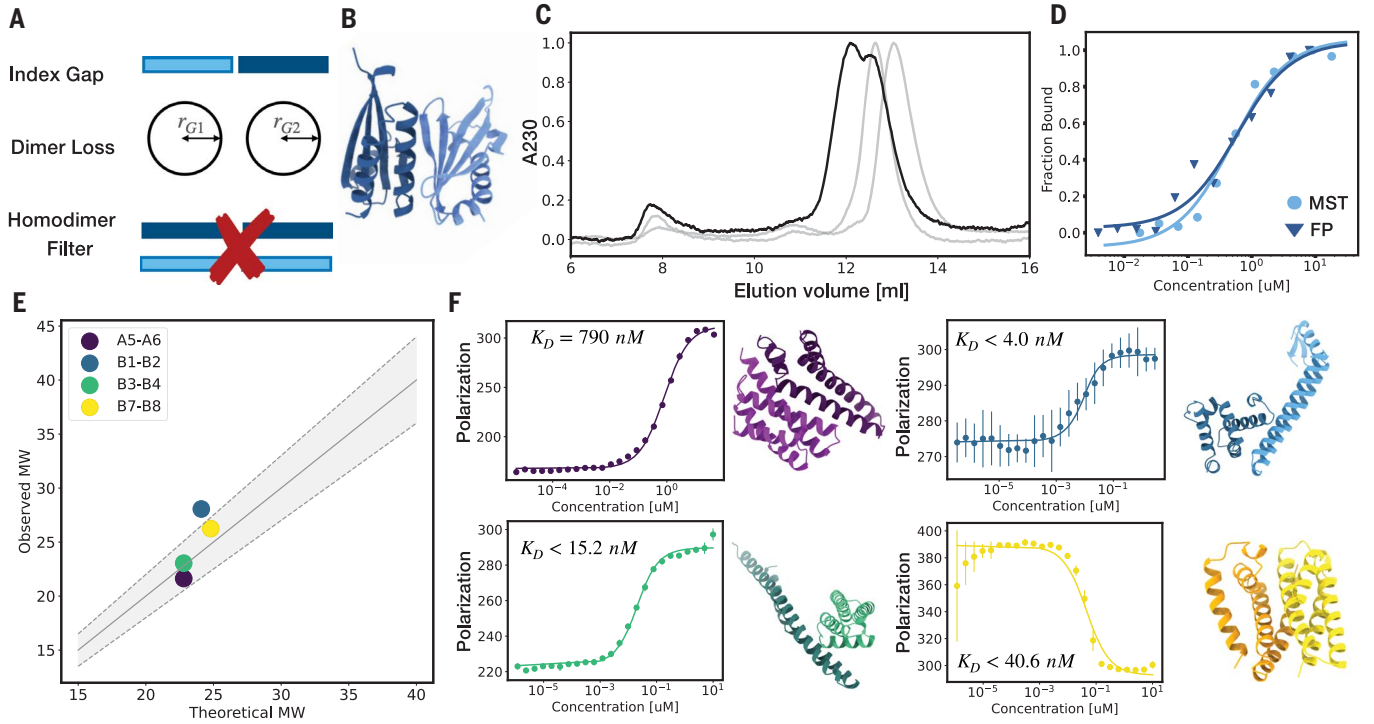

Fig. 5. Experimental characterization of protein interaction design. (**A**) Schematic overview of heterodimer design using RSO. (**B**) Rendering of the C5C6 heterodimer pair. (**C**) Characterization of heterodimeric pair C5C6 using SEC. (**D**) Binding isotherms of the C5C6 heterodimer pair acquired using the fluorescent polarization assay (triangles) and microscale thermophoresis (circles). (**E**) Expected versus observed molecular weight of a second set of heterodimers as determined through molecular weight–calibrated SEC. (**F**) Top: predicted structures of heterodimers analyzed in (E) and binding isotherms as obtained through fluorescent polarization. Error bars show SD of the mean computed from three replicates.

had low AF2 pLDDT also were modeled with low confidence by Model Angelo in the cryo-EM map, indicating flexibility (fig. S13F). Initially, we could not resolve the N-terminal region comprising 130 amino acids in this 950–amino acid design (fig. S14A), presumably due to flexibility. Single-particle cryo-EM analysis of samples stored at low temperatures revealed homodimeric species that were linked in a domain-swapping configuration through the flexible N-terminal part (fig. S14, B and C).

Single-particle cryo-EM analysis with the 1000–amino acid design candidate (Fig. 4E and fig. S11B) yielded a 3.3-Å resolution cryo-EM reconstruction (fig. S15). Using Model Angelo and PHENIX (*33*), we constructed an atomic model from the cryo-EM data. The model had a Ca-RMSD of 1.91 Å relative to the repredicted structure of the designed backbone (Fig. 4E and fig. S16), indicating good agreement between design and experimental structure.

These five experimentally determined protein structures demonstrate that RSO can be used for accurate protein structure design tasks.

Design of protein-protein interactions

Many current and emerging applications for de novo protein design involve designing protein-protein interactions. By including residue index gaps (*34*), AF2 can also be used for designing protein complexes comprising multiple separate chains (Fig. 5A). We used this feature to design homo-oligomers (fig. S17A) and heterodimers using RSO. To this end, we designed a loss function including two partial radius of gyration losses and an additional homodimer filter to favor heterodimers (Fig. 5A). This approach successfully created heterodimer designs in which individual monomers stayed monomeric when expressed separately, but formed a dimeric complex when mixed, as observed with SEC analysis (Fig. 5, B and C). Microscale thermophoresis and fluorescent polarization analysis yielded dissociation constants of 560 and 480 nM, respectively, for the heterodimeric design (Fig. 5D). We designed a second set of heterodimers but omitting in silico homo-oligomer filtering. These proteins independently formed homo-oligomers at the elution concentrations (fig. S17B), but at low concentrations, three out of the four proteins were monomeric. However, when we mixed the two distinct monomers designed for a heterodimer, all designs transitioned to the desired heterodimeric complex (Fig. 5E and fig. S17C). Binding affinity analysis using fluorescence polarization assays yielded dissociation constants for the heterodimeric interactions ranging from <4.0 to 790 nM (Fig. 5F). These findings demonstrate that RSO can also be used for interface design tasks in conjunction with ProteinMPNN sequence optimization.

Conclusions

The emerging de novo protein design pipeline comprises backbone design, sequence generation, and design filtering. ProteinMPNN and ESMFold provide fast and reliable methods for generating sequences for given backbones and for repredicting structures from candidate sequences for filtering, respectively. However, the quality of the input backbone remains a critical factor driving the overall success of protein design. Our work with RSO shows that innovations in backbone design methods can continue to push the boundaries of protein design. Specifically, RSO achieves high designability and efficiently generates promising in silico candidates for large proteins, including tasks such as site scaffolding and binder generation. It also allows design objectives to be encoded in custom loss functions, enabling rapid adaptation to individual research questions without retraining entire networks. Furthermore, RSO operates in sequence space, allowing backbone structure design with sequence constraints. High-quality backbones such as those produced by RSO can provide the foundation for ProteinMPNN to excel, whereas ProteinMPNN's capabilities allow us to fully

realize the potential of advanced backbone design.

The flexibility of this approach may expand toward building synthetic proteins with a variety of user-defined conformations by combining loss functions for multiple predicted states into one common gradient. RSO pushes the size range of designable protein monomers beyond the 100-kDa molecular weight barrier while retaining excellent accuracy, thereby approaching the size of therapeutically relevant protein scaffolds such as antibodies. The concept presented herein may likely extend to other structure prediction methods such as AlphaFold-3 (*35*) and RoseTTAFold All-Atom (*10*) to accomplish tasks such as small-molecule binding and protein-DNA hybrid structures with marginal modifications to the prediction networks.

REFERENCES AND NOTES

1. N. Ferruz, S. Schmidt, B. Höcker, *Nat. Commun.* **13**, 4348 (2022).
2. J. M. Singer *et al*., *PLOS ONE* **17**, e0265020 (2022).
3. B. I. M. Wicky *et al*., *Science* **378**, 56–61 (2022).
4. R. Verkuil *et al*., bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.12.21.521521.
5. J. Dauparas *et al*., *Science* **378**, 49–56 (2022).
6. B. L. Trippe *et al*., arXiv:2206.04119 [Preprint] (2023).
7. I. Anishchenko *et al*., *Nature* **600**, 547–552 (2021).
8. N. Anand, T. Achim, arXiv:2205.15019 [Preprint] (2022).
9. A. E. Chu, T. Lu, P.-S. Huang, *Nat. Biotechnol.* **42**, 203–215 (2024).
10. R. Krishna *et al*., *Science* **384**, eadl2528 (2024).
11. J. Jumper *et al*., *Nature* **596**, 583–589 (2021).
12. Z. Lin *et al*., bioRxiv 500902 [Preprint] (2022); https://doi.org/10.1101/2022.07.20.500902.
13. M. Baek *et al*., bioRxiv 542179 [Preprint] (2023); https://doi.org/10.1101/2023.05.24.542179.
14. A. E. Chu, L. Cheng, G. E. Nesr, M. Xu, P.-S. Huang, bioRxiv 542194 [Preprint] (2023); https://doi.org/10.1101/2023.05.24.542194.
15. Y. Lin, M. Lee, Z. Zhang, M. AlQuraishi, arXiv:2405.15489 [Preprint] (2024).
16. Y. Lin, M. AlQuraishi, arXiv:2301.12485 [Preprint] (2023); https://doi.org/10.48550/arXiv.2301.12485.
17. K. E. Wu *et al*., arXiv:2209.15611 [Preprint] (2022); https://doi.org/10.48550/arXiv.2209.15611.
18. S. Alamdari *et al*., bioRxiv 556673 [Preprint] (2023); https://doi.org/10.1101/2023.09.11.556673.
19. J. L. Watson *et al*., *Nature* **620**, 1089–1100 (2023).
20. J. B. Ingraham *et al*., *Nature* **623**, 1070–1078 (2023).
21. C. Goverde, B. Wolf, H. Khakzad, S. Rosset, B. E. Correia, *Prot. Sci.* **32**, e4653 (2023).
22. J. Wang *et al*., *Science* **377**, 387–394 (2022).
23. C. Norn *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2017228118 (2021).
24. Y. Zhang, J. Skolnick, *Proteins* **57**, 702–710 (2004).
25. C. A. Goverde *et al*., bioRxiv 540044 [Preprint] (2023); https://doi.org/10.1101/2023.05.09.540044.
26. N. R. Bennett *et al*., *Nat. Commun.* **14**, 2625 (2023).
27. J. P. Roney, S. Ovchinnikov, *Phys. Rev. Lett.* **129**, 238101 (2022).
28. H. Schweke *et al*., *Cell* **187**, 999–1010.e15 (2024).
29. M. van Kempen *et al*., bioRxiv 479398 [Preprint] (2022); https://doi.org/10.1101/2022.02.07.479398.
30. S. Vázquez Torres *et al*., *Nature* **626**, 435–442 (2024).
31. K. Wu *et al*., *Nature* **616**, 581–589 (2023).
32. K. Jamali *et al*., *Nature* **628**, 450–457 (2024).
33. P. V. Afonine, J. J. Headd, T. C. Terwilliger, P. D. Adams, "New tool: phenix.real_space_refine" (Computational Crystallography Newsletter, 2013); file:///C:/Users/swhite/Desktop/Phenix.real_space_refine_CCN_2013_07.pdf.
34. Data for: C. Frank *et al*., Figshare (2024); https://doi.org/10.6084/m9.figshare.27009724.
35. ColabDesign code for: C. Frank, Zenodo (2024); https://doi.org/10.5281/zenodo.13309081.

ACKNOWLEDGMENTS

We thank the DCI from EPFL, UNIL, and UNIGE for their support and for structural investigations of the K10 and K12 proteins; the team at NovoPro discoveries for structural studies; Google Cloud Services for providing computational resources; F. Praetorius, M. Pacesa, and L. Milles for useful discussions; J. Dauparas for help debugging the ColabDesign code; and the staff from BL17U1 at the Shanghai Synchrotron Radiation Facility (SSRF) for assistance during data collection; the Dubochet Center at EPFL Switzerland for technical support with cryo-EM; and the Shanghai Synchrotron for x-ray diffraction. The diffraction experiments were performed at SSRF BL02U1 (proposal no. 2020-SSRF-PT-014702-2). **Funding:** This work was supported by the European Research Council (Advanced Grant 101018465 to H.D.); the Deutsche Forschungsgemeinschaft (DFG Gottfried Wilhelm Leibniz Program Grant DI1500/3-1 to H.D.); TUM Innovation Network Projekt RISE (H.D. and C.F.); the National Institutes of Health (grant DP5OD026389 to S.O.); the National Science Foundation (grant MCB2032259 to S.O.); Amgen (S.O.); and the Innovative Research Team of High-level Local Universities in Shanghai, a key laboratory program of the Education Commission of Shanghai Municipality (ZDSYS14005). **Author contributions:** C.F., S.O., and H.D. conceived the study, with H.D. and S.O. providing supervision. C.F., S.O., and D.S. performed computational studies; C.F., L.F., L.W., D.S., and D.P. conducted wet lab experiments. C.F. was responsible for nsTEM. A.K. acquired cryo EM data at the Technical University of Munich (TUM). Z.Z. and M.H. conducted crystallization studies on the 200-amino acids design candidate and refined the model for K10 and K12. H.D. and S.O. acquired funding. CF wrote the original draft, and H.D., S.O., and M.Y. edited the manuscript. **Competing interests:** The authors declare no competing interests. H.D. is founder and CEO of CPTx GmbH. **Data and materials availability:** All sequences, designs, and AF2 predictions, as well as the negative stain TEM micrographs and an archived version of the exemplar notebook, are available on Figshare (*34*). The atomic models of crystal structures and of the cryo-EM densities are deposited to the PDB under 8S89 (400 amino acids Protein), 8YL8 (200 Form 1), 8YL4 (200 Form2), and 9EXZ (600 amino acids). The cryoEM dataset and models are deposited to the EMDB and PDB databases under EMD-50040 (K12) and 9F0L (K10) and 9EXK (K12) and EMD-50113 (K10). All notebooks and code are available through Zenodo (*35*) or Github: https://github.com/sokrypton/ColabDesign.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adq1741
Materials and Methods
Figs. S1 to S17
Tables S1 to S11
References (*36*–*51*)
Movies S1 to S4
MDAR Reproducibility Checklist

Submitted 2 May 2024; accepted 13 September 2024
10.1126/science.adq1741