Article

De novo protein design by deep network hallucination

https://doi.org/10.1038/s41586-021-04184-w

Received: 18 September 2020

Accepted: 21 October 2021

Published online: 1 December 2021

Check for updates

Ivan Anishchenko[1,2,7], Samuel J. Pellock[1,2,7], Tamuka M. Chidyausiku[1,2], Theresa A. Ramelot[3,4], Sergey Ovchinnikov[5], Jingzhou Hao[3,4], Khushboo Bafna[3,4], Christoffer Norn[1,2], Alex Kang[1,2], Asim K. Bera[1,2], Frank DiMaio[1,2], Lauren Carter[1,2], Cameron M. Chow[1,2], Gaetano T. Montelione[3,4] & David Baker[1,2,6]✉

There has been considerable recent progress in protein structure prediction using deep neural networks to predict inter-residue distances from amino acid sequences[1–3]. Here we investigate whether the information captured by such networks is sufficiently rich to generate new folded proteins with sequences unrelated to those of the naturally occurring proteins used in training the models. We generate random amino acid sequences, and input them into the trRosetta structure prediction network to predict starting residue–residue distance maps, which, as expected, are quite featureless. We then carry out Monte Carlo sampling in amino acid sequence space, optimizing the contrast (Kullback–Leibler divergence) between the inter-residue distance distributions predicted by the network and background distributions averaged over all proteins. Optimization from different random starting points resulted in novel proteins spanning a wide range of sequences and predicted structures. We obtained synthetic genes encoding 129 of the network-'hallucinated' sequences, and expressed and purified the proteins in *Escherichia coli*; 27 of the proteins yielded monodisperse species with circular dichroism spectra consistent with the hallucinated structures. We determined the three-dimensional structures of three of the hallucinated proteins, two by X-ray crystallography and one by NMR, and these closely matched the hallucinated models. Thus, deep networks trained to predict native protein structures from their sequences can be inverted to design new proteins, and such networks and methods should contribute alongside traditional physics-based models to the de novo design of proteins with new functions.

Deep learning methods have shown considerable promise in protein engineering. Networks with architectures borrowed from language models have been trained on amino acid sequences and used to generate new sequences without considering protein structure explicitly[4,5]. Other methods have been developed to generate protein backbones without consideration of sequence[6], and to identify amino acid sequences that either fit well onto specified backbone structures[7–10] or are conditioned on low-dimensional fold representations[11]; models tailored to generate sequences and/or structures for specific protein families have also been developed[12–16]. However, none of these methods address the classical de novo protein design problem of simultaneously generating both a new backbone structure and an amino acid sequence that encodes it.

Deep neural networks trained to predict distances between amino acid residues in 3D protein structures from amino acid sequence information have increased the accuracy of protein structure prediction[1–3]. These models take as input large sets of aligned sequences, and a major contributor to distance-prediction accuracy is the extent of co-evolution between the amino acid identities at pairs of positions. Following up an initial observation by AlphaFold in the 13th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction[17], we found that the trRosetta deep neural network trained using multiple sequence information could consistently predict three-dimensional structure accurately for de novo designed proteins from just a single sequence—that is, in the complete absence of co-evolution information[3]. The trRosetta model also predicted effects of amino acid substitutions on folding that were consistent with biophysical expectation[3]. These results suggested that during training, the trRosetta network was going beyond exploiting co-evolution information and learning fundamental relationships between protein sequence and structure.

Here we investigate whether the information stored in the many parameters of protein structure prediction networks can be used to generate physically plausible backbones and amino acid sequences that encode them. Methods such as Google's DeepDream[18] take networks trained to recognize faces and other patterns in images, and invert these

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Department of Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, Troy, NY, USA. [4]Center for Biotechnology and Interdisciplinary Sciences, Rensselaer Polytechnic Institute, Troy, NY, USA. [5]John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA, USA. [6]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [7]These authors contributed equally: Ivan Anishchenko, Samuel J. Pellock. ✉e-mail: dabaker@uw.edu

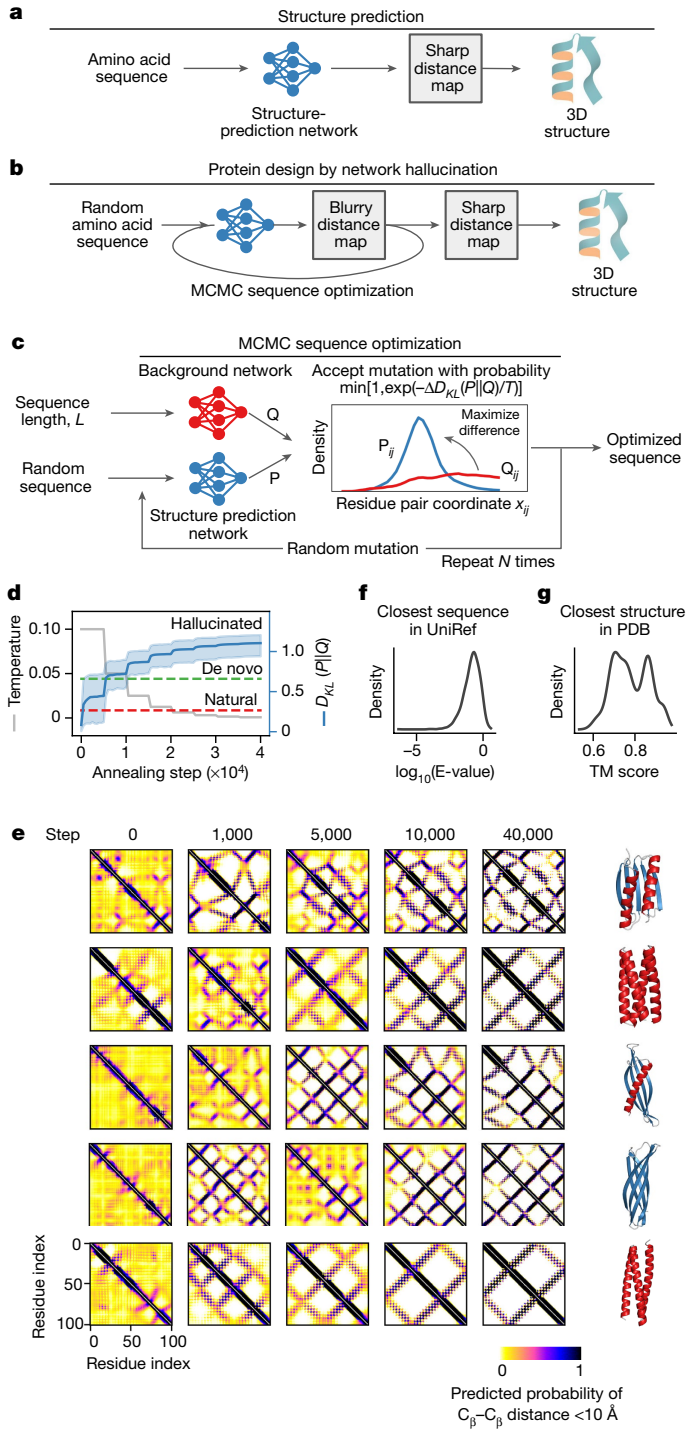


Fig. 1 | Overview of protein hallucination approach. a, Structure prediction methods such as trRosetta[3] and AlphaFold[2] employ a deep neural network to predict inter-residue geometries (reliable predictions have sharp 2D distance and orientation maps) from a single sequence or a multiple sequence alignment, and then the 3D structure is reconstructed by constrained minimization. **b**, Network predictions for a random sequence are not confident (blurry 2D maps); to transform a random sequence into one encoding a new folded protein, we introduce multiple single amino acid substitutions into the sequence using a Markov chain Monte Carlo (MCMC) algorithm, optimizing the sharpness of the 2D maps. **c**, Schematic of the MCMC procedure; P_{ij} and Q_{ij} denote trRosetta and background network predictions for a residue pair (i,j) and x is one of the residue-residue coordinates d,ω,θ,φ (Methods). **d**, Annealing trajectories averaged over 2,000 runs show a monotonic increase in the Kullback–Leibler divergence (contrast of the distance maps) with increasing Monte Carlo optimization. The mean and 0.01 and 0.99 quantiles are shown in blue; temperature profile (arbitrary units) is shown in grey. **e**, Distance maps become progressively sharper along the Monte Carlo trajectories as exemplified by five hallucinated sequences with different protein structure topologies. **f**, Hallucinated sequences are unrelated to the naturally occurring protein sequences in the UniRef90 database: median BLAST *e*-value of the closest hit is 0.17. **g**, Hallucinated structures range in similarity to protein structures in the PDB with average TM-scores to the closest match of 0.78.

by starting from arbitrary input images and adjusting them to be more strongly recognized as faces (or other patterns) by the network—the resulting images are often referred to as hallucinations because they do not represent any actual face, but what the neural network views as an ideal face. We used a similar approach to explore whether networks trained to predict structures from sequences could be inverted to generate brand-new 'ideal' protein sequences and structures.

The trRosetta network predicts distributions of distances and orientations between all pairs of residues in a set of aligned protein sequences for a protein family (Fig. 1a); in benchmark tests this network outperformed other methods[3]. Instead of inputting a naturally occurring sequence, we instead generated completely random sequences 100 amino acids in length, and fed these to the network (Fig. 1b). As expected for random sequences, which have a vanishingly small probability of folding to a defined structure, the distance distributions were diffuse and much less featured than those obtained with actual protein sequences. We then sought to optimize the sequences such that the network-predicted distance and orientation maps were as different as possible (had the highest Kullback–Leibler divergence) from residue-to-residue sequence separation and protein length-dependent generic protein background distributions (Fig. 1b, c and Methods). For each sequence, we carried out a Monte Carlo simulated annealing trajectory in sequence space: each step consists of substituting a randomly selected amino acid at a randomly selected position in the sequence, predicting distance and orientation distributions for all pairs of residues based on the mutated sequence using the network, and accepting the move on the basis of the change in the Kullback–Leibler divergence of the predicted distributions to the corresponding background distributions, summed over all residue pairs, according to the standard Metropolis criterion (Fig. 1c and Methods). The increase in Kullback–Leibler divergence aggregated over all 2,000 simulation trajectories is shown in Fig. 1d; in almost all cases, after around 20,000 Monte Carlo steps, the resulting distance maps were at least as featured (non-uniform) as those predicted for naturally occurring sequences and structurally confirmed de novo proteins designed using Rosetta. The predicted distance maps become progressively sharper during the course of the simulations, and trajectories started from different random sequences resulted in very different sequences and structures (Fig. 1e). We converted the final sharpened distance and orientation maps to protein 3D structures by direct minimization with trRosetta[3].

We used this approach to generate 2,000 new proteins with sequences predicted by the trRosetta network to fold into well defined structures, and compared their sequences and structures to native protein sequences and structures. The similarity of the hallucinated sequences to native protein sequences was very low, with best Blast[19] *e*-values to the Uniprot database of around 0.1 (Fig. 1f). Just as simulated images of cats generated by deep network hallucination are clearly recognizable as cats, but differ in detail from the cat images the network was trained on, the predicted structures resemble but are not identical to native structures in the PDB, with TM-align scores of 0.6–0.9 (Fig. 1g). The overall distributions of hallucinated sequences and structures are very different from those of naturally occurring proteins of the same (100-residue) length which were used during trRosetta training (Extended Data Fig. 1a–e).

The hallucinated sequences and their associated structures are quite diverse—different Monte Carlo trajectories starting from different

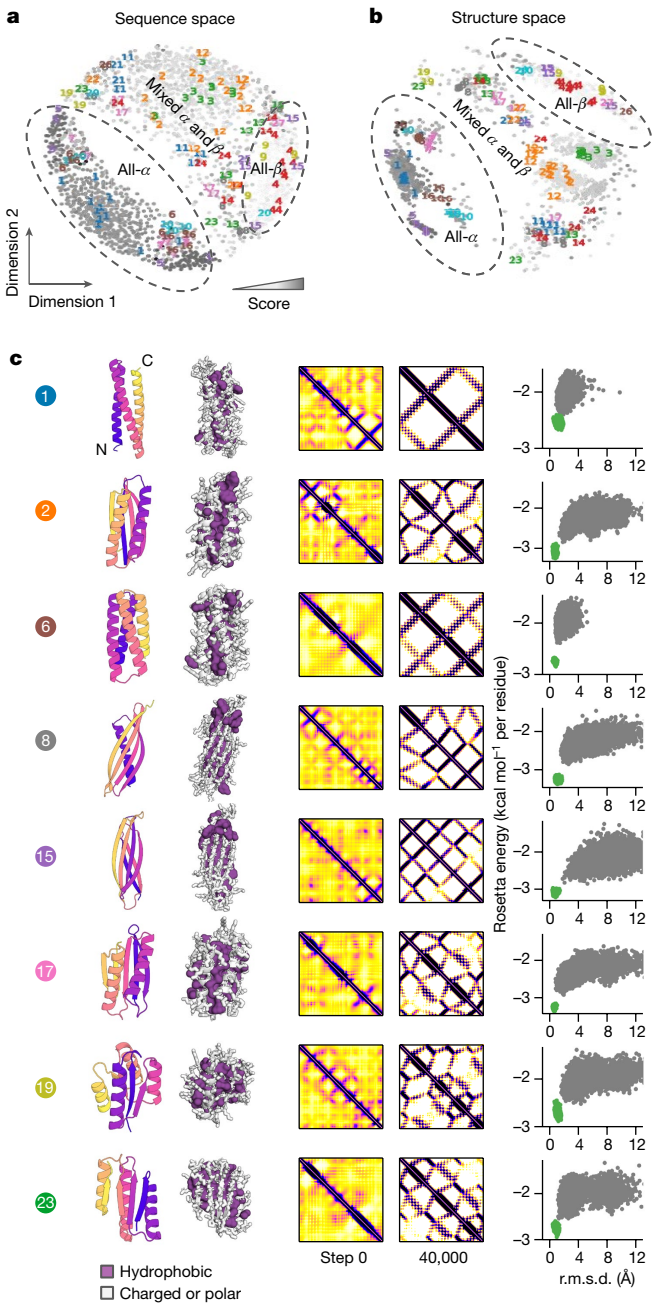


Fig. 2 | Overview of computational results. a, Multidimensional scaling generated representation of the sequence space covered by the 2,000 hallucinated proteins; BLAST bit score was used to measure the distance between proteins. Each grey dot represents one design colour-coded by the score from the network (a darker grey colour corresponds to a higher score). The 129 experimentally tested designs belong to 27 structural clusters, indicated by coloured numbers. **b**, Multidimensional scaling representation of the structural space covered by the 2,000 hallucinated proteins; (1 − TM-score) was used as the distance measure. Each grey dot represents one design; the greyscale value indicates the score from the network (darker grey corresponds to a higher score, equation (6)). The 129 experimentally tested designs fall into 27 structural clusters shown by coloured numbers. **c**, Examples of hallucinated designs of various topologies. First column, ribbon depiction of protein backbone coloured from blue (N terminus) to yellow (C terminus). Second column, hydrophobic core. Third column, distance maps at the beginning and end of the hallucination trajectory. Fourth column, folding energy landscapes from large-scale Rosetta ab initio structure prediction calculations; points represent lowest-energy structures sampled starting from an extended chain (grey points) and starting from the hallucinated design model (green points). The energy landscapes funnel into the energy minimum corresponding to the designed structure, providing independent, albeit in silico, evidence that the hallucinated sequences encode the hallucinated structures.

random number seeds converge on different sequence–structure pairs (Fig. 2a, b). We generated a 2D map of the space spanned by the structures (Fig. 2b) by multidimensional scaling of their pairwise 3D structural similarity (TM-score; Methods). The structures span all α-, all β- and mixed α–β-fold classes, with 95 different sub-folds at a TM-score clustering threshold of 0.75. Representative examples of structures from the 27 predominant clusters are shown in Fig. 2c. A prominent feature of these structures is that their backbone structures resemble the 'ideal' proteins generated by de novo protein design more than native proteins, despite the fact that the network was trained on native proteins. Both de novo designed proteins and the hallucinated proteins generated here have regular α-helices and β-sheets, and lack the long loops and other idiosyncrasies of native protein structures (Extended Data Fig. 1f, g).

We used Rosetta in silico protein folding simulations[20] to assess the extent to which the hallucinated sequences encode the hallucinated structures according to the Rosetta forcefield[21]. This is a completely orthogonal test as the network was trained exclusively on native protein structures, and has no access to the Rosetta energy function. We generated folding energy landscapes using large scale de novo folding simulations starting from an extended chain for 129 of the hallucinated proteins spanning a wide range of sequences and structures (Fig. 2c). For 82 out of the 129 proteins, the lowest-energy structures found in the simulations were close to the corresponding hallucinated structures with Cα root mean square deviation (r.m.s.d.) values below 3.0 Å, and for all 129 proteins, the lowest-energy structure sampled starting from the design model was lower in energy than any other structure obtained starting from an extended chain. Thus, according to the Rosetta physics-based energy model, the network-generated sequences do indeed encode the corresponding structures.

Next, we experimentally characterized the computer-generated hallucinations by obtaining synthetic genes for the 129 proteins, and expressing and purifying them from *E. coli* (Methods). Of these 129, 27 yielded size-exclusion chromatography (SEC) peaks corresponding to monomeric or small oligomeric species (Figs. 3d, 4d, Extended Data Fig. 3d,j, Supplementary Fig. 1) that were subsequently examined by circular dichroism (CD) spectroscopy. In all cases, the CD spectra were consistent with the target structures (Figs. 3e, 4e, Extended Data Fig, 3e,k), with the characteristic profiles of all α-helical proteins for the all α-helical designs (Fig. 3e, Extended Data Fig. 3e), and of α–β-proteins for the α–β-designs (representatives are shown in Fig 4e, Extended Data Fig. 3k). Twenty-one of the proteins were highly thermostable, with apparent melting temperatures above 70 °C (Extended Data Figs. 2d,2h,3f,3l); the α–β-designs in Fig. 4 were particularly stable, as none undergo unfolding transitions up to 95 °C. The experimentally validated proteins span a wide range of topologies, and all of the sequences are predicted by Rosetta large-scale energy calculations to have funnelled landscapes leading to the target structure (Figs. 3c, 4c). Together, these data indicate that the network-hallucinated proteins can fold into a wide range of stable structures with the predicted secondary structures.

We determined the solution NMR structure for design 0515 to be a monomeric antiparallel four-helix bundle (1D estimated ^{15}N $T_1 \approx 780$ ms, ^{15}N $T_2 \approx 77$ ms and $\tau_c \approx 9.6$ ns at 25 °C); structure quality assessment scores indicate a high-quality structure (Extended Data Table 1). The ensemble of 20 structures had a Cα r.m.s.d. relative the hallucinated model of approximately 1.82 Å (Fig. 5a, b, Extended Data Fig. 4). We also determined a 2.9 Å resolution crystal structure of design 0217, which revealed a 3-helix bundle with an overall fold similar to the hallucinated model; the backbone r.m.s.d. between model and crystal structure

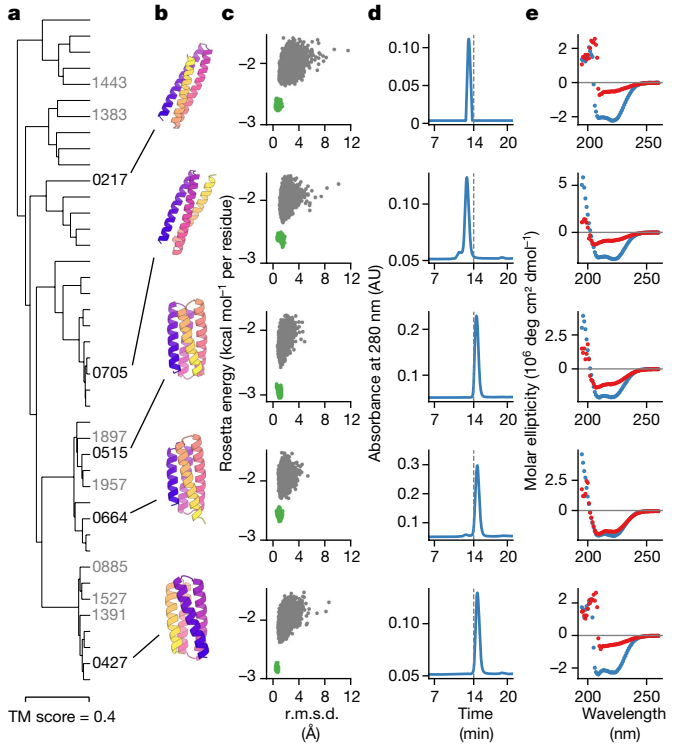


Fig. 3 | Experimental characterization of α-helical network-hallucinated proteins. a, Dendrogram showing 42 all-α designs clustered by structural similarity (TM-score); thermostable designs with CD spectra consistent with the target structure are labelled by their IDs. **b**, Three-dimensional models of the hallucinated designs. **c**, Ab initio folding funnels from Rosetta. **d**, SEC–MALS traces of purified protein. **e**, Circular dichroism spectra at 25 °C (blue) and 95 °C (red). Additional examples of stable α-helical designs marked in grey in **a** are shown in Extended Data Fig. 3.

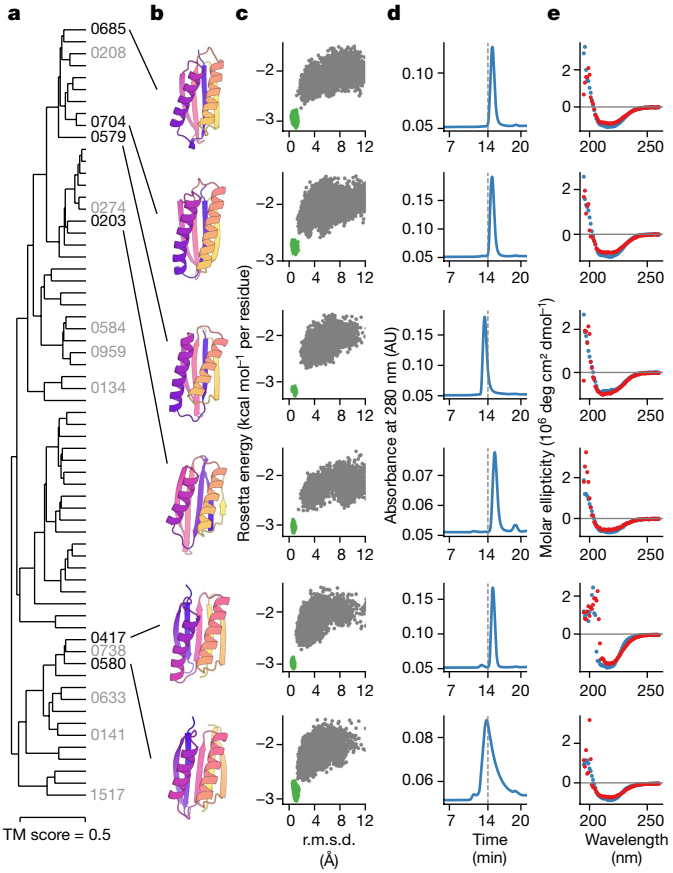


Fig. 4 | Experimental characterization of network-hallucinated proteins with mixed α–β structures. a, Dendrograms showing representative hallucinated protein designs with mixed α–β structures, clustered by TM-score; thermostable designs with CD spectra consistent with the target structure are labelled by their IDs. **b**, Three-dimensional models of the hallucinated designs. **c**, Ab initio folding funnels from Rosetta. **d**, SEC–MALS traces of purified protein. **e**, Circular dichroism spectra at 25 °C (blue) and 95 °C (red). Additional examples of stable mixed α–β - designs marked in grey in **a** are shown in Extended Data Fig. 3.

was 2.53 Å over all 100 residues (Fig. 5c, d, Extended Data Fig. 5). The agreement observed between these two experimental and hallucinated structures suggests that the network can accurately generate protein backbones and sequences that encode them.

As noted above, many of the hallucinated proteins form oligomers in solution. For example, design 0217 forms a dimer in the crystal structure (Extended Data Fig. 5), consistent with multi-angle light scattering coupled with SEC (SEC-MALS) analysis and NMR rotational correlation time measurements[22] (^{15}N $T_1 \approx 2.0$ s, ^{15}N $T_2 \approx 32$ ms and $\tau_c \approx 25$ ns at 25 °C, Supplementary Fig. 2). Sequences generated by the network were modelled as monomers, but the 0217 model displays clear amphipathic sequence patterning across the three-helix topology, with numerous solvent-exposed hydrophobic residues (Supplementary Fig. 2) that mediate the dimer contacts in the crystal structure. We also characterized hallucinated model 0417 by NMR, which revealed a spectrum consistent with the α–β-fold of the hallucinated model, and both SEC and NMR-relaxation measurements (^{15}N $T_1 \approx 730$ ms, ^{15}N $T_2 \approx 77$ ms and $\tau_c \approx$ 10.4 ns at 25 °C; Extended Data Fig. 6) indicate that it is primarily monomeric. However, temperature and buffer screening studies indicated transient self-association in solution, probably owing to the exposed hydrophobic residues (Extended Data Fig. 6), that ultimately precluded structure determination by NMR. The network appears to incorporate sequence features associated with the protein–protein interfaces of the native oligomeric proteins included in the PDB training set, probably explaining why many of the network-hallucinated proteins form dimers, higher order oligomeric species and soluble aggregates.

To determine whether self-association of hallucinated designs was mediated by surface hydrophobic patches, we substituted these with polar residues in a subset of the hallucinations that formed oligomers, and SEC revealed that several were converted to monomeric species (Supplementary Fig. 3). One of these surface-modified hallucinations, 0738_mod, yielded crystals, and we determined the crystal structure of this protein at a resolution of 2.4 Å. Structural superposition of the 0738 model and the 0738_mod crystal structure revealed a 3.68 Å C α r.m.s.d. over 96 residues (Fig. 5f, g, Extended Data Fig. 7). Despite register shifts upon superposition of the entire crystal structure and hallucinated model, the amino- and carboxy-terminal halves of the crystal structure align remarkably well to the corresponding regions in the hallucinated model with backbone r.m.s.d. values of 1.32 Å over 57 residues, and 2.17 Å over 43 residues for the N-terminal and C-terminal half, respectively (Fig. 5h), with many of the sidechain rotamers recovered. This is a notable result given that the network operates on the backbone level only in the structure-generation process. The accuracy does not reflect PDB memorization; the closest BLAST hits in the PDB for the N and C terminal halves have *e*-values of 0.29 and 0.63, respectively.

The high similarity of the NMR and crystal structures to the hallucinated structure models demonstrate that the hallucination process solves the classic de novo protein design problem, despite having no explicit knowledge of the physics of protein folding. The hallucinated sequences are unrelated to those of proteins of known structures; the sequences of the three hallucinated proteins whose structures we solved here all have *e*-values worse than 0.02 (Supplementary Table 1). To determine whether the lack of explicit

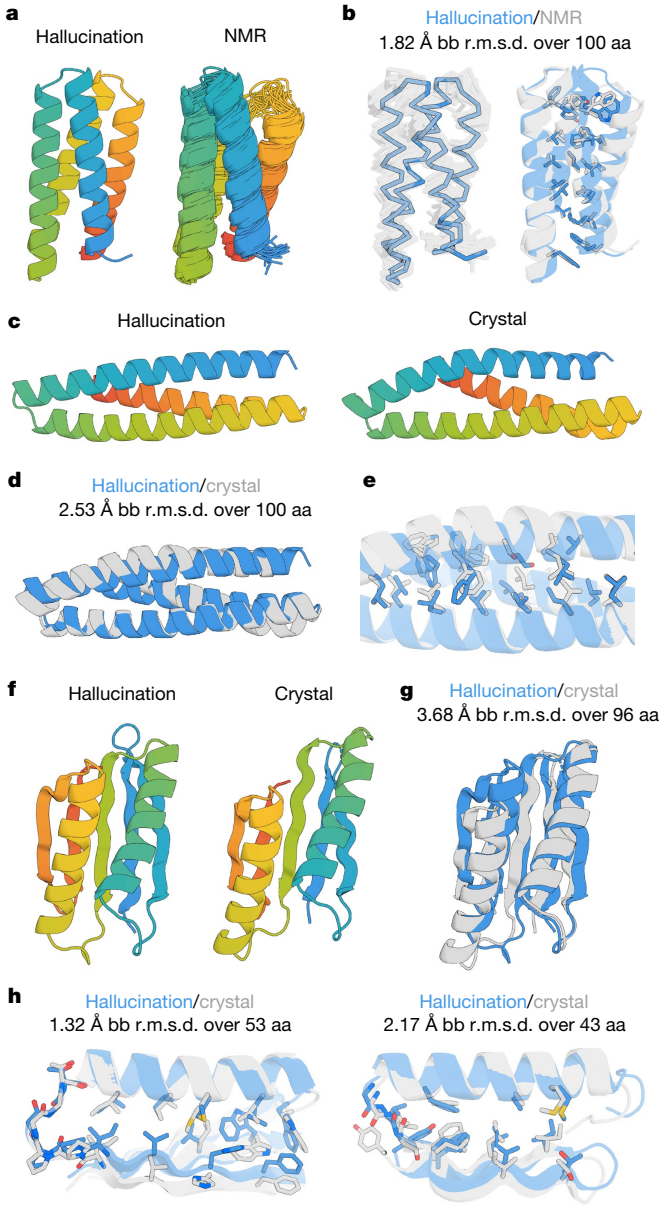


Fig. 5 | Structural analysis of network-hallucinated proteins. a, Hallucination model (left) and NMR ensemble structure (right) of design 0515. **b**, Superposition of NMR ensemble (transparent grey) and hallucinated model (outlined blue) of design 0515 and overlay of the medoid NMR structure and model with side chains shown. **c**, Structures of the hallucinated model 0217 (left) and the crystal structure (right). **d**, Superposition of the hallucination model (blue) and crystal structure (grey) of design 0217. **e**, Zoomed in overlay of the crystal structure (grey) and hallucination model (blue) of 0217 with side chains shown as sticks. **f**, Hallucination model of design 0738 (left) and the crystal structure of the surface-modified 0738_mod (right). **g**, Superposition of the 0738 hallucination model and the 0738_mod crystal structure. **h**, Superposition of the N-terminal section (left) and of the C-terminal section (right) of the 0738 hallucination model (blue) and the 0738_mod crystal structure (grey). Standalone structures in **a**, **c**, **f** are coloured from N terminus (blue) to C terminus (red).

treatment of side chains could lead to population of alternatively packed states, we investigated the dynamic properties in solution of design 0515 solved by NMR, as well as for 0217 and 0738_mod, for which structures were determined by X-ray crystallography (Extended Data Fig. 8). The solution data for design 0515 and the 0217 dimer indicate well-ordered structures in solution, with internal dynamics typical of small natural proteins. For design 0738_mod the solution data indicate multiple monomeric conformations in solution in slow conformational exchange. We anticipate that future incorporation of an explicit sidechain representation in the hallucination method could reduce such structural heterogeneity.

Conclusion

Our results demonstrate that a deep neural network trained exclusively on native sequences and structures can generalize to create new proteins with sequences unrelated to those of native proteins that fold into stable structures. Many of the hallucinated proteins that we found are monomeric, stable, have the expected secondary structure, and are strongly predicted to fold to the target structure by Rosetta in completely orthogonal calculations (we did not use Rosetta in any way for either sequence generation or selection for experimental characterization). The close agreement between experimental solution NMR and crystal structures with the corresponding hallucinated design models for the three proteins that we characterized in detail suggest that many of these proteins fold into the predicted hallucinated structures.

De novo protein design efforts over the past ten years have sought to distill the key features of protein structures and protein sequence–structure relationships using physics-based models such as Rosetta, and have then used these models to design idealized structures that embody these features on the basis of the principle that proteins fold to their lowest free-energy states[23,24]. The hallucinated structures show a remarkable resemblance to these idealized proteins—in the regularity of the secondary structures, shortness of the loops, and other characteristics. Indeed, the most similar structure in the PDB to the 0738_mod structure is the de novo designed protein Top7 (Supplementary Fig. 4). During training on large numbers of irregular native protein structures, the deep neural network evidently learned to encode ideal protein structure properties very similar to those encoded by expert protein designers using more traditional scientific approaches, albeit representing them in very different ways (in the millions of parameters in the network rather than the very much smaller number of parameters of the backbone-generation methods and the force field in Rosetta as well as other approaches). Current efforts in applying deep learning to a wide range of scientific problems will reveal whether this distilling of essential features occurs more generally.

Experimental analysis of the hallucinated designs by SEC and NMR indicate that a number of these proteins formed soluble aggregates or smaller homo-oligomers rather than monomers. There are several features of the approach that could account for this. First, trRosetta was trained on native protein structures, including many homo- and hetero-oligomers, and hence the model may not have learned to disfavour surface hydrophobic residues to the extent required for highly soluble monomers. This may have been a particular contributor to the low success rate for β-sandwich designs, which had multiple surface hydrophobic residues, perhaps reflecting antibody and other structures with extensive immunoglobulin-fold interdomain interactions (Supplementary Fig. 5). The homodimer interface observed in the crystal structure of 0217 may be representative of the interfaces formed in the many discrete homo-oligomers we observed. As illustrated by the conversion of several selected oligomers to monomers by substituting surface hydrophobic residues with polar residues, this shortcoming can be addressed relatively easily (Supplementary Fig. 3b). Second, the trRosetta model is inherently low in resolution, as sidechain atoms and packing interactions are not represented explicitly. This could limit the depth of the native free-energy minimum, and hence the occupancy of the designed states compared with alternative possibly aggregation-prone states. One common example of this is core overpacking, as most structural differences between the hallucinated models and experimental structures occur at locations where multiple

large hydrophobic residues were placed in the protein core (Supplementary Fig. 6); this could also account for the structural heterogeneity in solution observed for design 0738_mod (Extended Data Fig. 8). Complementing trRosetta design with an explicit all-atom design method like Rosetta could combine the strengths of both approaches[25].

Our work opens up a large set of avenues of research to explore. Our hallucination approach can be readily extended to design new proteins using the recently developed RoseTTAFold[26] and AlphaFold2[27] networks[28,29]. On the sampling side, the Monte Carlo approach can be made more efficient by direct gradient-based minimization by tracing the gradients back to the inputs[25]. The loss function can be generalized to include specific structural features—for example, binding motifs[30] or catalytic sites —around which the network can hallucinate new protein inhibitors or enzyme catalysts[28]. Unlike traditional protein design calculations, in which properties of the target scaffold such as the overall topology and/or the secondary structure element lengths and locations are specified in advance, through a structure 'blueprint' or other approach, the ability of the network to hallucinate plausible protein structures from scratch makes building a supporting scaffold around a desired functional site much more straightforward, since the structure need not be mapped out in advance. The network can come up with a wide range of different protein topology solutions for a given problem with no restrictions on sequence length[30]. More generally, our work demonstrates the power of generative deep learning approaches for molecular design, which will undoubtedly continue to grow over the coming years.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-04184-w.

1. Xu, J. Distance-based protein folding powered by deep learning. *Proc. Natl Acad. Sci. USA* **116**, 16856–16865 (2019).
2. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
3. Yang, J. et al. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl Acad. Sci. USA* **117**, 1496–1503 (2020).
4. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-*N* protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021).
5. Madani, A. et al. ProGen: language modeling for protein generation. Preprint at https://arxiv.org/abs/2004.03497 (2020).
6. Anand, N., Eguchi, R. & Huang, P. S. Fully differentiable full-atom protein backbone generation. In *ICLR 2019 Workshop* https://openreview.net/forum?id=SJxnVL8YOV (2019).
7. Wang, J., Cao, H., Zhang, J. Z. H. & Qi, Y. Computational protein design with deep learning neural networks. *Sci Rep.* **8**, 6349 (2018).
8. Ingraham, J., Garg, V. K., Barzilay, R. & Jaakkola, T. Generative models for graph-based protein design. in *ICLR 2019 Workshop* https://openreview.net/forum?id=SJgxrLLKOE (2019).
9. Anand, N., Eguchi, R. R., Derry, A., Altman, R. B. & Huang, P.-S. Protein sequence design with a learned potential. Preprint at https://doi.org/10.1101/2020.01.06.895466 (2020).
10. Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A. & Kim, P. M. Fast and flexible protein design using deep graph neural networks. *Cell Syst.* **11**, 402–411.e4 (2020).
11. Karimi, M., Zhu, S., Cao, Y. & Shen, Y. De novo protein design for novel folds using guided conditional Wasserstein generative adversarial networks. *J. Chem. Inf. Model.* **60**, 5667–5681 (2020).
12. Davidsen, K. et al. Deep generative models for T cell receptor protein sequences. *eLife* **8**, e46935 (2019).
13. Costello, Z. & Martin, H. G. How to hallucinate functional proteins. Preprint at https://arxiv.org/abs/1903.00458 (2019).
14. Eguchi, R. R., Anand, N., Choe, C. A. & Huang, P.-S. IG-VAE: generative modeling of immunoglobulin proteins by direct 3D coordinate generation. Preprint at https://doi.org/10.1101/2020.08.07.242347 (2020).
15. Repecka, D. et al. Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* **3**, 324–333 (2021).
16. Hawkins-Hooker, A. et al. Generating functional protein variants with variational autoencoders. *PLoS Comput. Biol.* **17**, e1008736 (2021).
17. Senior, A. W. et al. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins* **87**, 1141–1148 (2019).
18. Mordvintsev, A., Olah, C. & Tyka, M. Inceptionism: going deeper into neural networks. *Google AI Blog* https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html (2015).
19. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
20. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93 (2004).
21. Park, H. et al. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* **12**, 6201–6212 (2016).
22. Rossi, P. et al. A microscale protein NMR sample screening pipeline. *J. Biomol. NMR* **46**, 11–22 (2010).
23. Koga, N. et al. Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
24. Dou, J. et al. De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485–491 (2018).
25. Norn, C. et al. Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad Sci. USA* **118**, e2017228118 (2021).
26. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
27. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
28. Wang, J. et al. Deep learning methods for designing proteins scaffolding functional sites. Preprint at https://doi.org/10.1101/2021.11.10.468128 (2021).
29. Jendrusch, M., Korbel, J. O. & Sadiq, S. K. AlphaDesign: A de novo protein design framework based on AlphaFold. Preprint at https://doi.org/10.1101/2021.10.11.463937 (2021).
30. Tischer, D. et al. Design of proteins presenting discontinuous functional sites using deep learning. Preprint at https://doi.org/10.1101/2020.11.29.402743 (2020).

Methods

Approach

The general protein design problem can be formulated in probabilistic terms as the finding of mutually compatible sequence–structure pairs such that the joint probability P(sequence,structure) is maximized. Using the chain rule for probabilities:

$$P(\text{sequence, structure}) = P(\text{structure}|\text{sequence}) \times P(\text{sequence}) \quad (1)$$

The first term, P(structure|sequence), is related to the protein structure prediction problem where one seeks for the most probable structure for a given protein sequence, whereas the second term P(sequence) accounts for general constraints on amino acid sequences. As described in the following sections, we sought to develop a heuristic objective function that captures both terms and is a function purely of the amino acid sequence, that we could then optimize through simulated annealing in sequence space.

Networks and objective function

The trRosetta protein structure prediction network, described in detail elsewhere[3], is a 2D residual-convolutional neural network that takes one- and two-site features derived from a multiple sequence alignment or a single sequence as an input and produces a 2D output ($P_{\text{trRosetta}}$) describing distances and orientations for all residue pairs in a protein in a probabilistic manner: for every residue pair (i,j), these generated maps contain predicted probability distributions over the Cβ–Cβ distance and five inter-residue angles (comprising the full set of six rigid-body degrees of freedom). When accurate, such 2D predictions can be straightforwardly translated into a 3D structure by direct minimization[2,3]. Random sequences give diffuse predictions, whereas existing de novo designs produce peaked distributions with low variance[3].

To quantify the sharpness of predicted structure distributions for a given sequence, we trained a background network similar in architecture to trRosetta and on the same training set[3], but not providing amino acid sequence identity information (Supplementary Fig. 7; this can loosely be viewed as representing a generic 'molten globule' state, $Q_{\text{background}}$). Predictions from trRosetta and the background network ($p_{x,ijk}$ and $q_{x,ijk}$ respectively) have the same form: for every residue pair (i,j) the networks generate probability distributions over binned 6D residue–residue distances and orientations $x \in \{d, \omega, \theta, \varphi\}$ (see ref. [3]. for details) with $\sum_k p_{x,ijk} = \sum_k q_{x,ijk} = 1$. We can then quantify the extent of contrast between the structure predicted for a given sequence and the background distribution as the mean Kullback–Leibler divergence (D_{KL}) over all residue pairs (i,j) and distance and angle distributions

$$D_{\text{KL}}\left(P_{\text{trRosetta}} || Q_{\text{background}}\right) = \sum_{x \in \{d,\omega,\theta,\varphi\}} \left[\frac{1}{L^2} \sum_{i,j=1}^{L} \sum_{k=1}^{N_x} p_{x,ijk} \log\left(\frac{p_{x,ijk}}{q_{x,ijk}}\right) \right] \quad (2)$$

where L is the protein length, and N_x the number of bins which coordinate x is discretized into ($N_d = 37$, $N_{\omega,\theta} = 25$, $N_\varphi = 13$).

To capture general sequence constraints, we used the negative Kullback–Leibler divergence of the amino acid composition of a sequence from that of the PDB as a whole

$$-D_{\text{KL}}\left(f_a || f_a^{\text{PDB}}\right) = -\sum_{a=1}^{20} f_a \log\left(\frac{f_a}{f_a^{\text{PDB}}}\right) \quad (3)$$

where f_a is the frequency of the 20 amino acids in a given sequence and f_a^{PDB} are the frequencies in the Protein Data Bank; pseudocounts are added to avoid zeros in the numerator.

Protein hallucination

We optimized the combined objective function

$$F = D_{\text{KL}}\left(P_{\text{trRosetta}} || Q_{\text{background}}\right) - D_{\text{KL}}\left(f_a || f_a^{\text{PDB}}\right) \quad (4)$$

using simulated annealing starting from a random amino acid sequence of length L (L = 100 throughout this study). At each step i, a random single amino acid substitution is made at a randomly selected position, and the move is accepted based on the Metropolis criterion:

$$A_i = \min[1, \exp(-(F_i - F_{i-1})/T)] \quad (5)$$

(that is, if A_i is smaller than a uniform random number $u \in [0,1]$), F_i and F_{i-1} are objective function values (equation(4)) at steps i and i–1 respectively. Each trajectory consisted of 40,000 attempted moves; the temperature T is 0.1 at the beginning of the trajectory and reduced by half every 5,000 steps. Cysteines were excluded to avoid complications from oxidation since we planned to produce the proteins in the reducing environment of the *E. coli* cytoplasm.

Design selection

Two-thousand proteins were generated using the hallucination procedure described above, and structurally compared to each other using the template modelling score[31] (TM-score). Average-linkage hierarchical clustering yielded 95 clusters with an average inter-cluster similarity of TM-score = 0.75. We scored each of the designs within the 30 most populous clusters (which had 7 or more members) based on the sum of the KL divergence with the background distribution (equation (2)), and the cross-entropy between the final hallucinated structure Y and the 6D coordinate distributions generated by trRosetta for the sequence:

$$\text{score} = D_{\text{KL}}(P||Q) + [\text{CE}(Q, Y) - \text{CE}(P, Y)] \quad (6)$$

$$\text{CE}(P, Y) = \sum_{x \in \{d,\omega,\theta,\varphi\}} \frac{1}{L^2} \sum_{i,j=1}^{L} \sum_{k=1}^{N_x} y_{x,ijk} \log(p_{x,ijk}) \quad (7)$$

where Y is the 3D structure as represented by all distances and orientations between all pairs of residues ($y_{x,ijk} = 1$ for the bin k observed in the hallucinated structure, is zero otherwise); CE(Q, Y) is calculated similarly. The second term in equation (6) [CE(Q, Y) − CE(P, Y)]assesses how well the hallucinated structure fits the trRosetta predicted structure distributions. For each cluster, we picked the top 50% or top 20 (whichever was smaller) structures with the highest scores (297 designs in total), and inspected these structures manually to filter out those with internal cavities or voids, extended surface hydrophobic patches, and misformed secondary structure elements; three clusters were completely eliminated due to poor model quality. 129 hallucinated sequences from the remaining 27 structural clusters (no more than 10 designs per cluster) were selected for experimental testing.

Protein expression and purification

Genes coding for the selected 129 designs were synthesized and cloned into pET28b(+) expression vector with an additional 21-residue N-terminal sequence containing a His-tag and thrombin cleavage site to aid purification (full sequence: MGSSHHHHHHSSGLVPRGSHM). These plasmids were purchased from Genscript and expressed in *E. coli* BL21(DE3) cells. Starter cultures were grown overnight at 37 °C in lysogeny broth (LB) with added antibiotic (50 μg ml^{-1} kanamycin). These overnight cultures were used to inoculate either 50 ml (for screening) or 500 ml (for crystallography) of Studier autoinduction media[32] supplemented with antibiotic, and grown overnight. Cells were harvested by centrifugation and resuspended in 25 ml lysis buffer (20 mM imidazole in PBS containing protease inhibitors), and lysed by microfluidizer. PBS buffer contained 20 mM $NaPO_4$, 150 mM

NaCl, pH 7.4. After removal of insoluble pellets, the lysates were loaded onto nickel affinity gravity columns to purify the designed proteins by immobilized metal-affinity chromatography (IMAC).

Size-exclusion chromatography for screening

Following IMAC purification, designs were further purified by SEC on ÄKTAxpress (GE Healthcare) using a Superdex 75 10/300 GL column (GE Healthcare) in PBS buffer. The monomeric or smallest oligomeric fractions of each run (eluting at approximately 14 ml) were collected and immediately analysed by CD spectroscopy or flash frozen in liquid nitrogen for later analysis. The resulting samples were generally > 95% homogeneous on SDS–PAGE gels. For additional characterization by SEC-MALS, we analysed SEC-purified samples with elution buffer of 50 mM Tris-HCl, 150mM NaCl pH 8.0 at 1 ml min−1 over a Superdex 75 10/300 column in line with a Heleos multi-angle static light scattering and an Optilab T-rEX detector (Wyatt Technology Corporation). The data was then analysed using ASTRA (Wyatt Technologies) to calculate the weighted average molar mass (Mw) of the selected species and the number average molar mass (Mn) to determine monodispersity by polydispersity index (PDI) = Mw/Mn.

Circular dichroism experiments

To determine secondary structure and thermostability of the designs far-ultraviolet CD measurements were carried out with an JASCO 1500. The 260 to 195 nm wavelength scans were measured at every 10 °C intervals from 25 °C to 95 °C. Temperature melts monitored dichroism signal at 220 nm in steps of 2 °C min^{-1} with 30 s of equilibration time. Wavelength scans and temperature melts were performed using 0.35 mg ml^{-1} protein in PBS buffer with a 1-mm path length cuvette. Protein concentrations were determined by absorbance at 280 nm measured using a NanoDrop spectrophotometer (Thermo Scientific) using predicted extinction coefficients[33].

NMR sample preparation

Samples for NMR studies were prepared following standard protocols developed by the Northeast Structural Genomics Consortium[34,35]. Initial sample preparation was carried out on a fee-for-service basis by Nexomics Biosciences. Selected designs were expressed in *E. coli* BL21 (DE3) cells as U-^{15}N-enriched-enriched proteins, using MJ9 minimal media[36] with antibiotic kanamycin (50 µg ml^{-1}), and $(^{15}NH_4)_2SO_4$ as the sole source of nitrogen. For midi-scale production[35], 50-ml cultures were grown at 37 °C to OD_{600} 0.6 to 0.8 units, and protein production was induced with 1 mM IPTG at 25 °C over several hours. Cells were then harvested by centrifugation at 5,000*g*. Cell pellets were resuspended in lysis buffer (50 mM Tris-HCl, 0.5 M NaCl, 20 mM imidazole, pH 8.0, with protease inhibitor cocktail), cells were disrupted by sonication, and the resulting suspension centrifuged at 13,000*g* for 45 min. The supernatants from each fermentation were then purified in parallel using a set of 1-ml Ni-NTA HisTrap HP columns (GE Healthcare). For each column, the elution peak fraction was collected, and the purified protein was exchanged into NMR buffer 1 (20 mM Tris-HCl, pH 7.5, 100 mM NaCl). These samples were each more than about 98% homogeneous, based on SDS–PAGE. Samples were concentrated to around 0.5 mM protein concentration, and prepared in 3-mm Shigemi NMR tubes. Following initial screening, buffer conditions were further optimized by microscale NMR screening with various buffers and aggregation disrupting additives, using 1.7 mm NMR tubes, as described elsewhere[22].

U-^{15}N,^{13}C-enriched design 0515 protein samples for structure determination were prepared using a similar protocol. In this case, 1 litre cultures were prepared using MJ9 minimal media[36] with ^{13}C-glucose and $(^{15}NH_4)_2SO_4$ as the sole sources of carbon and nitrogen, respectively. Following initial growth at 37 °C, expression was induced with IPTG, and the cultures were shifted to 17 °C. Cells were collected by centrifugation (2,270*g* for 1 h), cell pellets were resuspended in 25 ml lysis buffer (PBS with 40 mM imidazole and protease inhibitor cocktail), and cells were disrupted by sonication. The insoluble pellet was sedimented by centrifugation (32,000*g* for 45 min), and the supernatant was applied to a 2.5 ml Ni-NTA column (Hispur Ni-NTA superflow agarose, ThermoFisher) equilibrated with the same lysis buffer. The protein was eluted from the column with steps of 75, 100, 150, 200 and 500 mM imidazole. The elution peak fraction was collected, dialyzed into NMR buffer 2 (25 mM HEPES, 50 mM NaCl, 0.02% NaN_3, pH 7.4), concentrated to ~0.9 mM protein concentration, and prepared in a 5-mm Shigemi NMR tube for data collection with addition of 5% D_2O (v/v). This sample was >98% homogeneous by SDS–PAGE analysis, and >95% isotope enriched based on MALDI-TOF mass spectrometry. Samples were prepared for residual dipolar coupling (RDC) data collection by dilution of a ^{15}N-labeled 0515 NMR sample with Pf1 phage (25 mg/ml) alignment medium.

NMR data collection and structure determination

NMR data for initial NMR screening was collected at 298 K on a Bruker Avance III HD 700 MHz spectrometer at The City University of New York. Additional NMR screening and structure analysis data were collected at the indicated temperatures on a Bruker Avance II 600 MHz and 800 MHz spectrometer systems in the Center for Biotechnology and Interdisciplinary Studies at Rensselaer Polytechnic Institute. NMR screening was done by recording 2D [^{1}H-^{15}N]-HSQC or [^{1}H-^{15}N]-SOFAST-HMQC spectra, and by measurements of ^{15}N T_1 and T_2 relaxation times using 1D NMR spectra to provide estimated[22] rotational correlation times, τ_c. RDC data collection on both isotropic and partially aligned samples was performed at 600 MHz using a 2D interleaved ^{15}N-^{1}H- HSQC IPAP experiment to measure couplings[37]. All NMR spectra were processed using NMRPipe and NMRDraw[38] and visualized in NMRFAM-SPARKY[39]. Backbone resonance assignments for 0515 were determined using a standard triple-resonance NMR strategy, with a suite of fast pulsing BEST double and triple resonance experiments provided within NMRlib[40], including 2D [^{1}H-^{15}N]-SOFAST-HMQC, 2D [^{1}H-^{13}C]-BEST-HSQC, and 3D BEST-HNCO, BEST-HNCA, BEST-HNCACB, and BEST-HNcoCACB. Additionally, standard CcoNH TOCSY, ^{15}N TOCSY-HSQC, HBHAcoNH, and 3D NOESY (τ_{mix} = 100 ms) spectra implemented with nonuniform sampling (NUS) were collected to complete assignments. A 50% Poisson gap sampling schedule[41] was used for NUS within TopSpin 3.2 (Bruker) and subsequently reconstructed using sparse multidimensional iterative lineshape-enhanced[42] (SMILE) reconstruction within NMRPipe[38]. Resonance assignments were determined by manual refinement of resonance assignments obtained from the I-PINE web server[43]. Assignment validation was done with cmap images generated using AutoAssign[44]. Peak intensities from 3D NOESY spectra, together with dihedral angle constraints determined from backbone chemical shift data using TALOS-N[45], were used as input for structure determination. NOESY peak assignments were made automatically using Cyana[46,47], together with the programs RPF and ASDP to guide manual correction of NOESY peak assignments[48,49]. The lowest energy 20, of 100 structures calculated, were then refined in explicit water using CNS[50] with the addition of 70 backbone one-bond ^{1}H-^{15}N RDCs. Structure quality analyses were performed on the final ensemble of 20 models using RPF and PSVS software[51] (Table S1). Resonance assignments and NMR data were deposited in the BioMagResDataBase with ID 30890, and coordinates and restraints in the PDB with ID 7M5T.

Crystallography sample preparation, data collection, and analysis

Protein was expressed and purified as described for initial screening. Crystal screening was performed using Mosquito LCP by STP Labtech. Crystals were grown in 800 mM succinic acid pH 7.0 for 0217. For 0738_mod, crystals were grown in 15% (v/v) ethanol and 40% (v/v) pentaerythritol propoxylate (5/4 PO/OH). Resultant crystals were looped and flash cooled in liquid nitrogen. Data was collected on 24-ID-C at NECAT, APS, at the wavelength of 0.97918 Å at 100K temperature. Both datasets were subsequently processed with HKL2000 and Scalepack suite[52]. For 0217, molecular replacement was carried out using predicted models from two sources: trRosetta predictions[3], and classical Rosetta ab initio structure predictions[20]. While both sets of predictions yielded

converged ensembles on a single topology, the classical ab initio models had significant diversity within that ensemble. Each of the 2,000 models (1,000 trRosetta and 1,000 ab initio) had all side chains removed past the gamma carbon, and was run through Phaser[53]. A single solution was found in I 2 3 from one of the ab initio models with two copies in the asymmetric unit and a TFZ score of 13.3 (no other model yielded a TFZ score >8). Sidechains were rebuilt and the model was refined with Rosetta-Phenix[54], yielding a map with readily interpretable density. For 0738_mod molecular replacement was carried out using the trRosetta model with deleted loops. Manual rebuilding in Coot[55] and cycles of Phenix refinement[56] were used to build the final model. For 0217 final Ramachandran favoured and outliers were 99% and 0%, respectively. For 0738_mod refinement, final Ramachandran favoured and outliers were 96% and 0%, respectively. Coordinates and structure factors were deposited to the PDB for 0217 and 0738_mod with corresponding PDB IDs 7K3H and 7M0Q; crystallographic data collection and refinement statistics are provided in Extended Data Table 2.

Structural alignment generation and analysis

Structural alignments comparing NMR and crystal structures to hallucinated models were performed using the Theseus maximum likelihood superpositioning tool[57]. In cases where parts of the crystal structure were missing, corresponding regions in the hallucinated model were removed and subsequent superposition was performed. Alignments were performed in 'backbone' alignment mode and resulting classical pairwise r.m.s.d. values are reported. Protein structure figures were made in PyMOL[58].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The atomic coordinates of the crystal structures for designs 0217 and 0738_mod, as well as the NMR structure for design 0515 have been deposited in the RCSB Protein Data Bank with the accession numbers 7K3H, 7M0Q and 7M5T, respectively. NMR chemical shifts, NOESY peak lists, and spectral data have been deposited in the BioMagResDB, BMRB ID 30890. Amino acid sequences and structure models for all 2K designs described in the manuscript are freely available for download at https://files.ipd.uw.edu/pub/trRosetta/hallucinations2K.tar.gz. Amino acid sequences and 3D structures of the generated designs were compared to known protein sequences and structures in UniProt (https://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2017_12/uniref/) and the Protein Data Bank (11 March 2020), respectively.

Code availability

The computer code used to generate the hallucinated proteins described in the manuscript was made publicly available as a part of trDesign Github package (https://github.com/gjoni/trDesign); corresponding structural models were generated by the trRosetta structure modelling script available for free download at https://yanglab.nankai.edu.cn/trRosetta/download/. The Rosetta software suite was used to perform ab initio prediction calculations. Rosetta is freely available for academic users on Github, and can be licensed for commercial use by the University of Washington CoMotion Express License Program.

31. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
32. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).
33. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. How to measure and predict the molar absorption coefficient of a protein. *Protein Sci.* **4**, 2411–2423 (1995).
34. Acton, T. B. et al. Preparation of protein samples for NMR structure, function, and small-molecule screening studies. *Methods Enzymol.* **493**, 21–60 (2011).
35. Xiao, R. et al. The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. *J. Struct. Biol.* **172**, 21–33 (2010).
36. Jansson, M. et al. High-level production of uniformly 15N-and 13C-enriched fusion proteins in *Escherichia coli*. *J. Biomol. NMR* **7**, 131–141 (1996).
37. Ottiger, M., Delaglio, F. & Bax, A. Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra. *J. Magn. Reson.* **131**, 373–378 (1998).
38. Delaglio, F. et al. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
39. Lee, W., Tonelli, M. & Markley, J. L. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **31**, 1325–1327 (2015).
40. Favier, A. & Brutscher, B. NMRlib: user-friendly pulse sequence tools for Bruker NMR spectrometers. *J. Biomol. NMR* **73**, 199–211 (2019).
41. Hyberts, S. G., Milbradt, A. G., Wagner, A. B., Arthanari, H. & Wagner, G. Application of iterative soft thresholding for fast reconstruction of NMR data non-uniformly sampled with multidimensional Poisson gap scheduling. *J. Biomol. NMR* **52**, 315–327 (2012).
42. Ying, J., Delaglio, F., Torchia, D. A. & Bax, A. Sparse multidimensional iterative lineshape-enhanced (SMILE) reconstruction of both non-uniformly sampled and conventional NMR data. *J. Biomol. NMR* **68**, 101–118 (2017).
43. Lee, W. et al. I-PINE web server: an integrative probabilistic NMR assignment system for proteins. *J. Biomol. NMR* **73**, 213–222 (2019).
44. Moseley, H. N. B., Sahota, G. & Montelione, G. T. Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. *J. Biomol. NMR* **28**, 341–355 (2004).
45. Shen, Y. & Bax, A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J. Biomol. NMR* **56**, 227–241 (2013).
46. Güntert, P., Mumenthaler, C. & Wüthrich, K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283–298 (1997).
47. Herrmann, T., Güntert, P. & Wüthrich, K. Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J. Biomol. NMR* **24**, 171–189 (2002).
48. Huang, Y. J., Powers, R. & Montelione, G. T. Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. *J. Am. Chem. Soc.* **127**, 1665–1674 (2005).
49. Huang, Y. J., Tejero, R., Powers, R. & Montelione, G. T. A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins* **62**, 587–603 (2006).
50. Brünger, A. T. et al. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
51. Bhattacharya, A., Tejero, R. & Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. *Proteins* **66**, 778–795 (2007).
52. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
53. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
54. DiMaio, F. et al. Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nat. Methods* **10**, 1102–1104 (2013).
55. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
56. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D* **75**, 861–877 (2019).
57. Theobald, D. L. & Wuttke, D. S. Accurate structural correlations from maximum likelihood superpositions. *PLoS Comput. Biol.* **4**, e43 (2008).
58. *The PyMOL Molecular Graphics System* version 2.4 (Schrödinger, 2021).
59. Zweckstetter, M. NMR: prediction of molecular alignment from structure using the PALES software. *Nat. Protoc.* **3**, 679–690 (2008).
60. Montelione, G. T. & Wagner, G. 2D Chemical exchange NMR spectroscopy by proton-detected heteronuclear correlation. *J. Am. Chem. Soc.* **111**, 3096–3098 (1989).

Acknowledgements We thank R. Xiao, G. Liu and A. Wu (Nexomics Biosciences), for assistance in initial NMR protein production; J. Aramini for assistance with NMR data collection for initial HSQC screening; R. Ballard and X. Li for mass spectrometry assistance; and R. Divine and R. Kibler for AKTA scripting. This work was funded by grants from the NSF (DBI 1937533 to D.B. and I.A., and MCB 2032259 to S.O.), the NIH (DP5OD026389 to S.O.), Open Philanthropy (C.C. and A.B.), Eric and Wendy Schmidt by recommendation of the Schmidt Futures program (F.D. and L.C.), and the Audacious project (A.K.), the Washington Research Foundation (S.J.P.), Novo Nordisk Foundation Grant NNF17OC0030446 (C.N.). This work was also supported in part by NIH grants R01 GM120574 (G.T.M.) and R35GM141818 (G.T.M.), and the Howard Hughes Medical Institute (D.B. and T.M.C.). We also acknowledge computing resources provided by the Hyak supercomputer system funded by the STF at the University of Washington, and Rosetta@Home volunteers in ab initio structure prediction calculations, and thank staff at Northeastern Collaborative Access Team at Advanced Photon Source for the beamline, supported by NIH grants P30GM124165 and S10OD021527, and DOE contract DE-AC02-06CH11357. We acknowledge the NMR Core Facility resources at Rensselaer Polytechnic Institute and thank S. McCallum for providing valuable support.

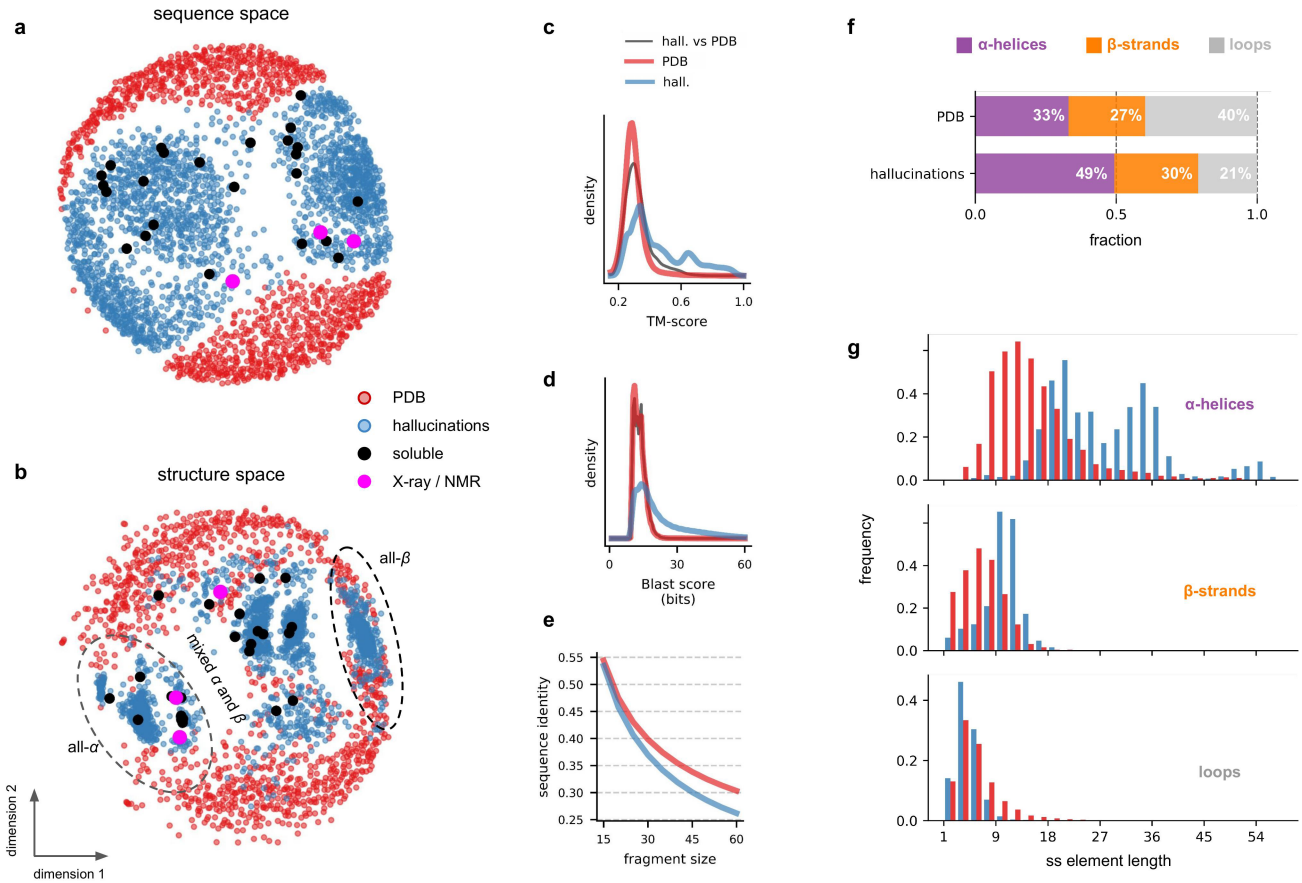
Competing interests G.T.M. is a founder of Nexomics Biosciences. The other authors declare no competing interests.

Additional information

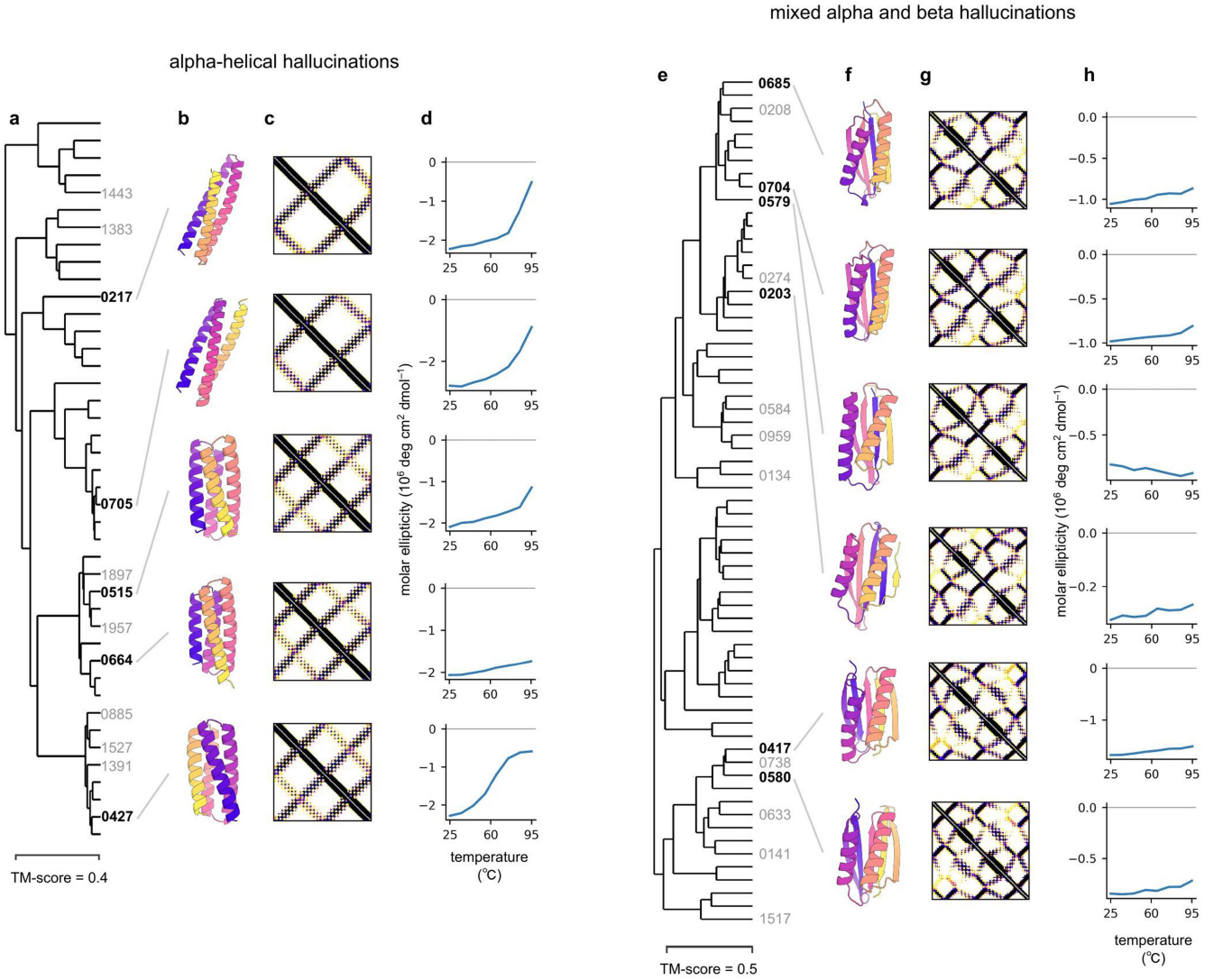
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-04184-w.

Correspondence and requests for materials should be addressed to David Baker.

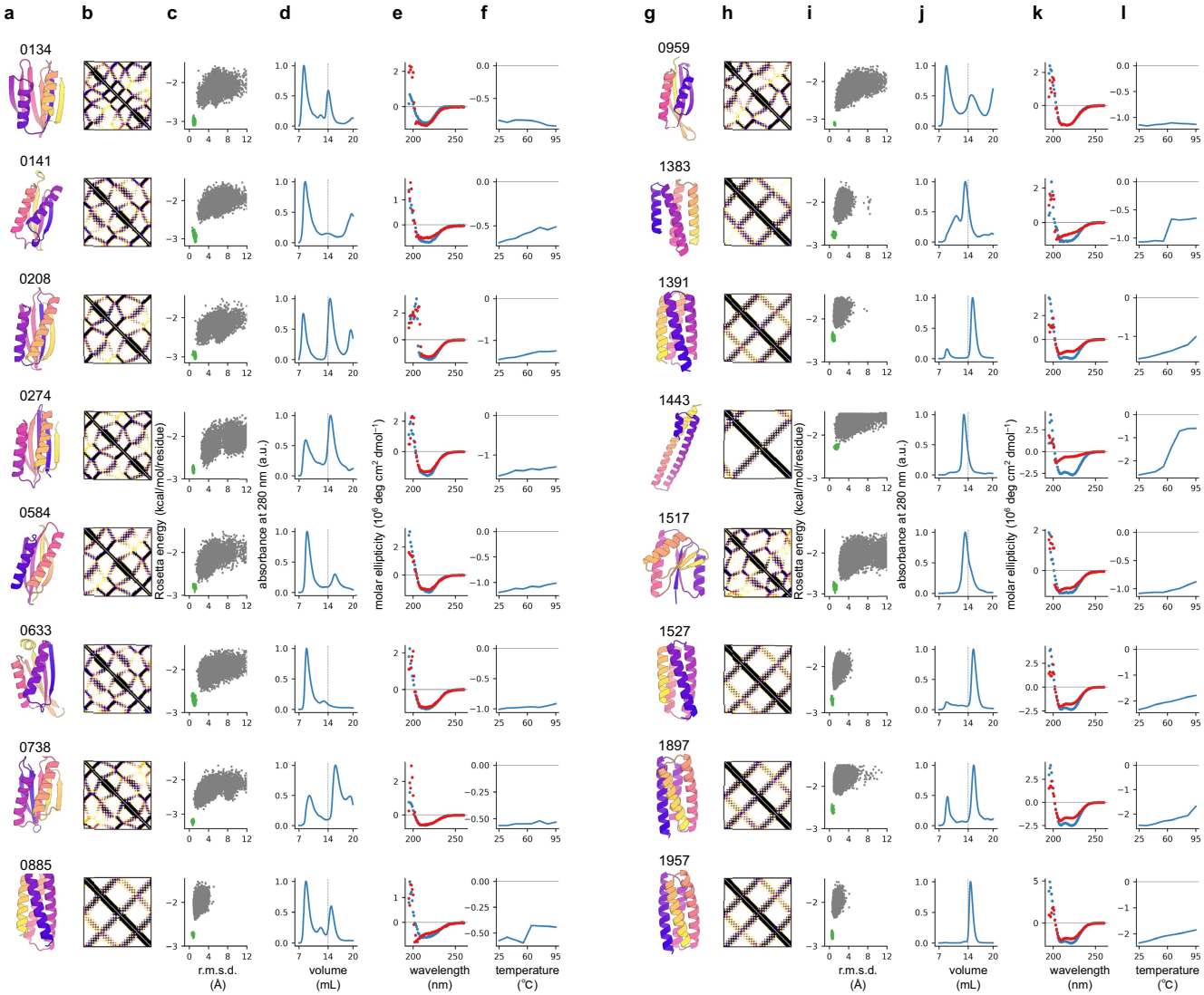
Peer review information *Nature* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



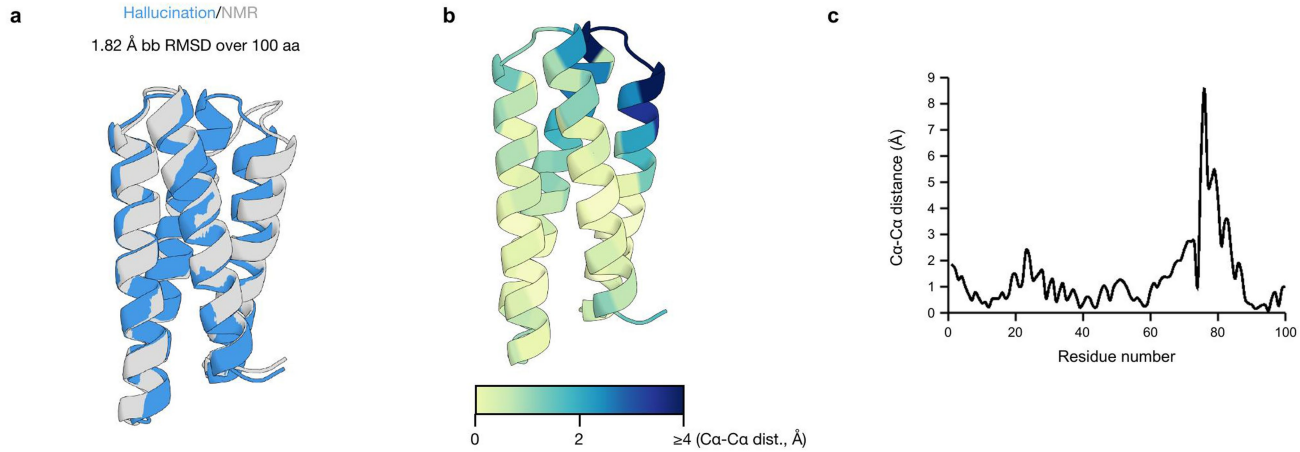
Extended Data Fig. 1 | Comparison of the hallucinated designs to proteins with known structure and of similar length (100 +/− 10 aa) from the trRosetta training set. a,**b**) Multidimensional scaling plots of the sequence (**a**) and structure (**b**) spaces covered by the 2,000 hallucinated proteins (blue dots) along with 1,110 proteins of similar length from the trRosetta training set (red dots). These scatter plots show that subspaces spanned by hallucinated proteins and natural proteins of similar size (100 +/− 10 aa) are quite distinct; the network is not simply recapitulating native proteins of the same length. Soluble and structurally characterized hallucinations are marked by black and magenta dots respectively. **c**,**d**) Distributions of pairwise structure (**c**) and sequence (**d**) similarities for hallucinated and natural proteins. The hallucinated proteins are more similar to each other (blue lines) than they are to natural proteins (grey lines). **e**) Sequence comparisons (gappless threading) of fragments of various size (15,20,...,60 aa) from the hallucinated designs (blue) and natural 100 (+/− 10) aa-long proteins (red) to other proteins from the trRosetta training set. There is no apparent tendency for the trRosetta-based design procedure to "copy over" sequence fragments from the proteins in the training set into the hallucinated designs. **f**,**g**) Secondary structure content of the hallucinated designs and natural 100 aa-long proteins from the training set. Hallucinations are more ideal than natural proteins in having less loops but longer secondary structure elements.



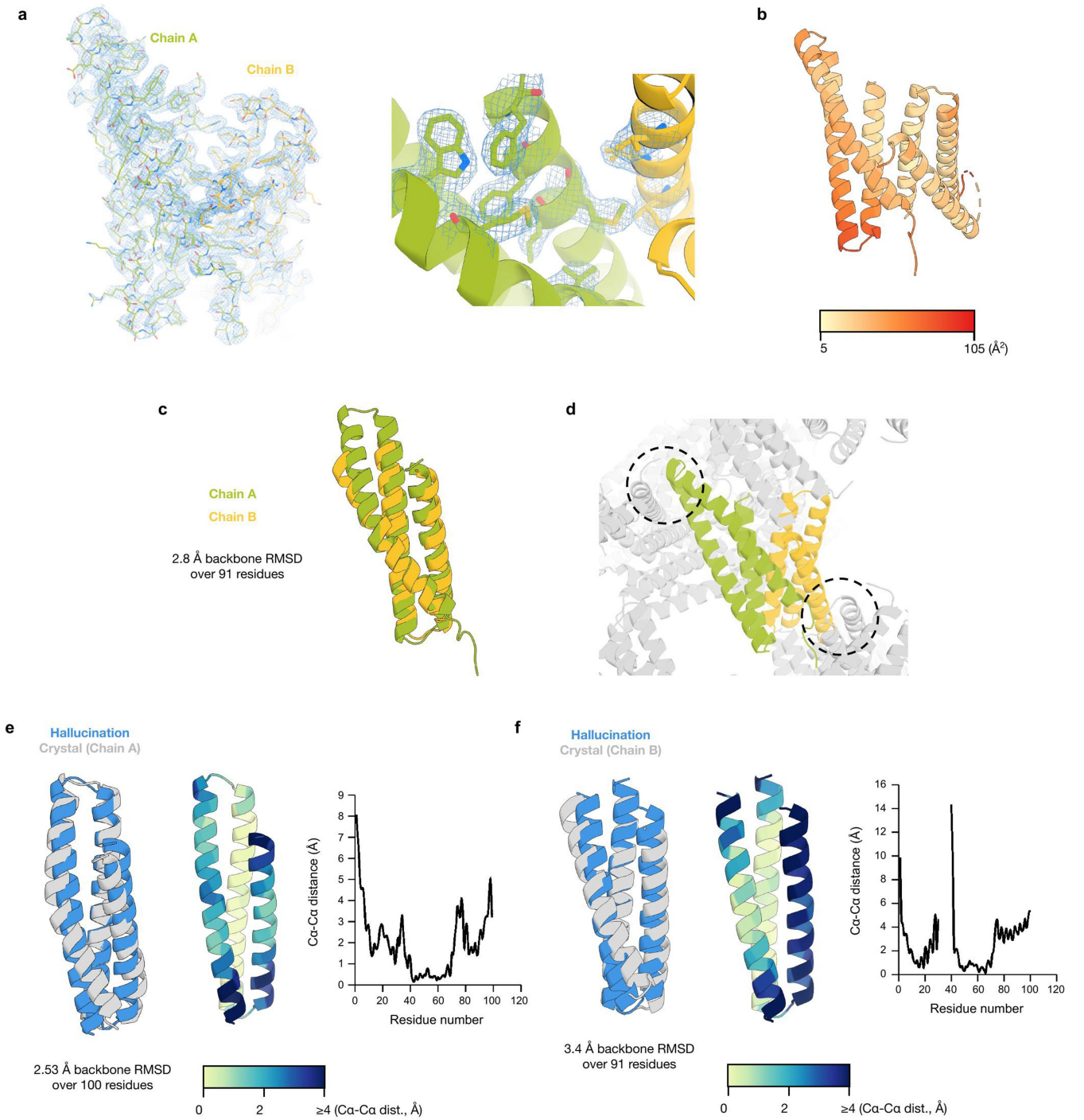
Extended Data Fig. 2 | Additional data on the experimentally characterized all-α and mixed α–β network-hallucinated proteins. a,**e**) Dendrograms showing representative hallucinated protein designs clustered by TM-score; thermostable designs with CD spectra consistent with the target structure are labelled by their IDs. **b**,**f**) Three-dimensional models of the hallucinated designs. **c**,**g**) Predicted distance maps at the end of the hallucination trajectory. **d**,**h**) Temperature dependence of CD signal at 220 nm in the 25-95 °C temperature range.



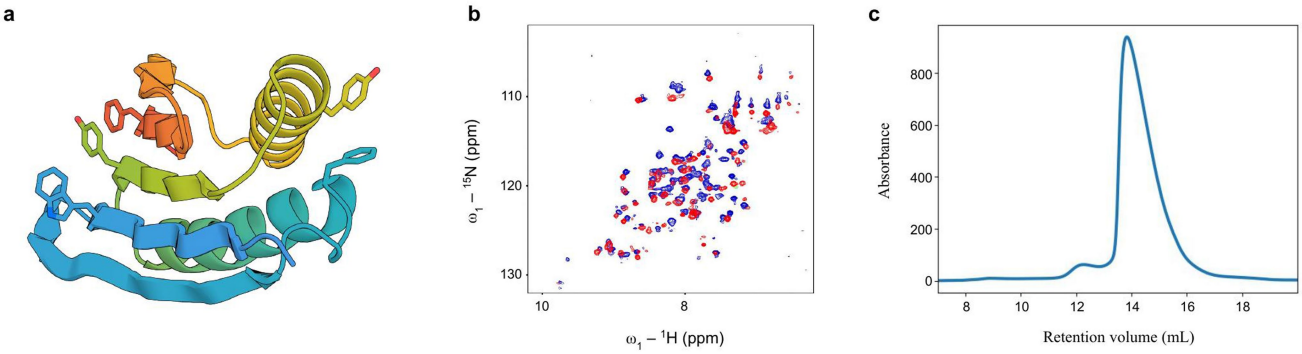
Extended Data Fig. 3 | Additional examples of thermostable hallucinations with CD spectra consistent with the target structure. a,**g**) 3D structure models of the hallucinated designs. **b**,**h**) Predicted distance maps at the end of the hallucination trajectory. **c**,**i**) ab initio folding funnels from Rosetta. **d**,**j**) Size-exclusion chromatography traces. **e**,**k**) Circular dichroism spectra at 25 °C (blue) and 95 °C (red). **f**,**l**) Temperature dependence of Circular Dichroism signal at 220 nm in the 25 to 95 °C temperature range.



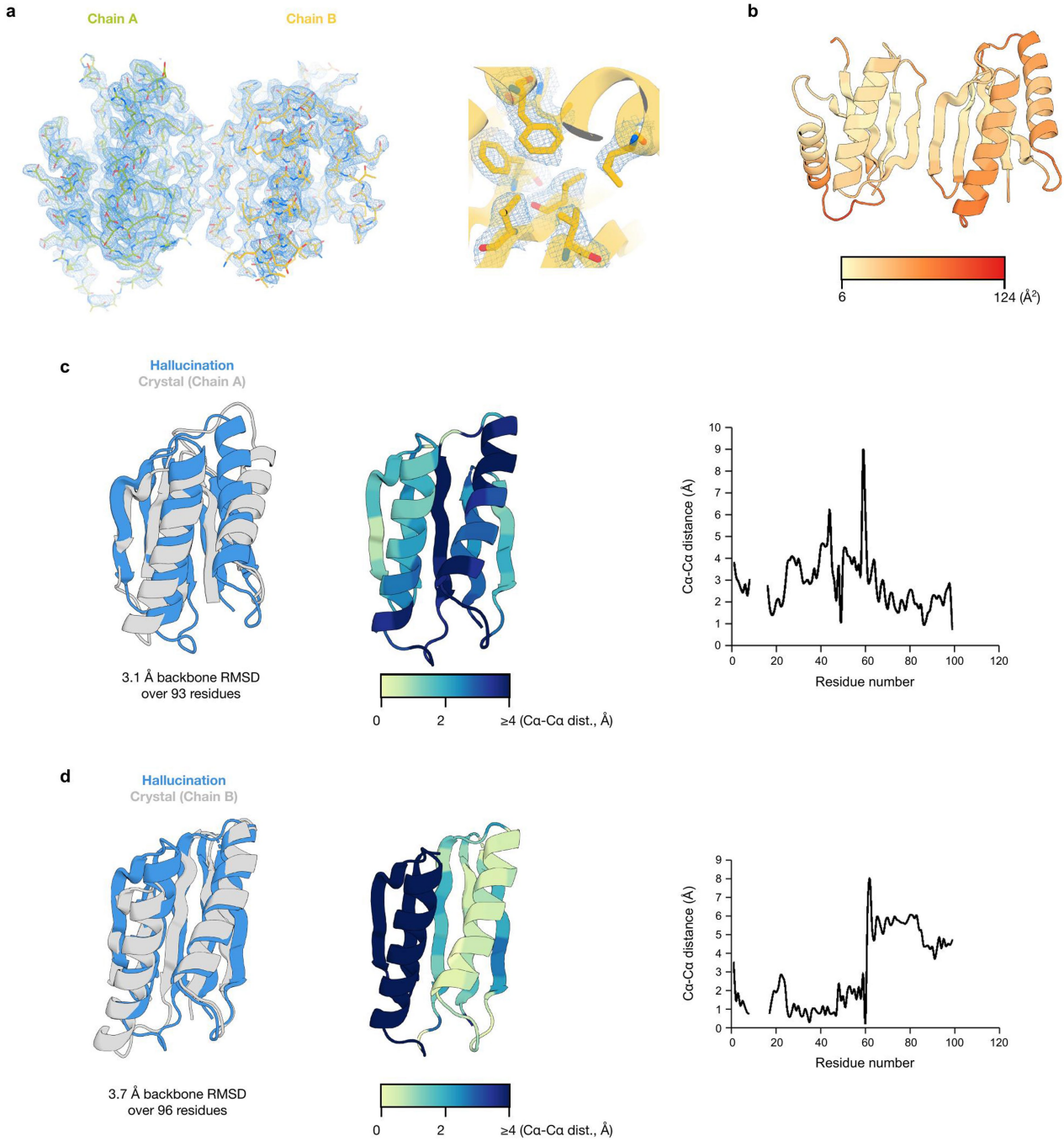
Extended Data Fig. 4 | Comparison of 0515 NMR structure to hallucinated model. a) Superposition of hallucinated model (blue) and NMR medoid structure (gray) of 0515 reveal 1.82 Å backbone r.m.s.d. over 100 residues **b**) Hallucinated model of 0515 colored by distance between Cα-Cα pairs between model and NMR medoid structure after structural superposition and **b**) corresponding plot of per-residue Cα-Cα distance difference between model and NMR medoid structure.



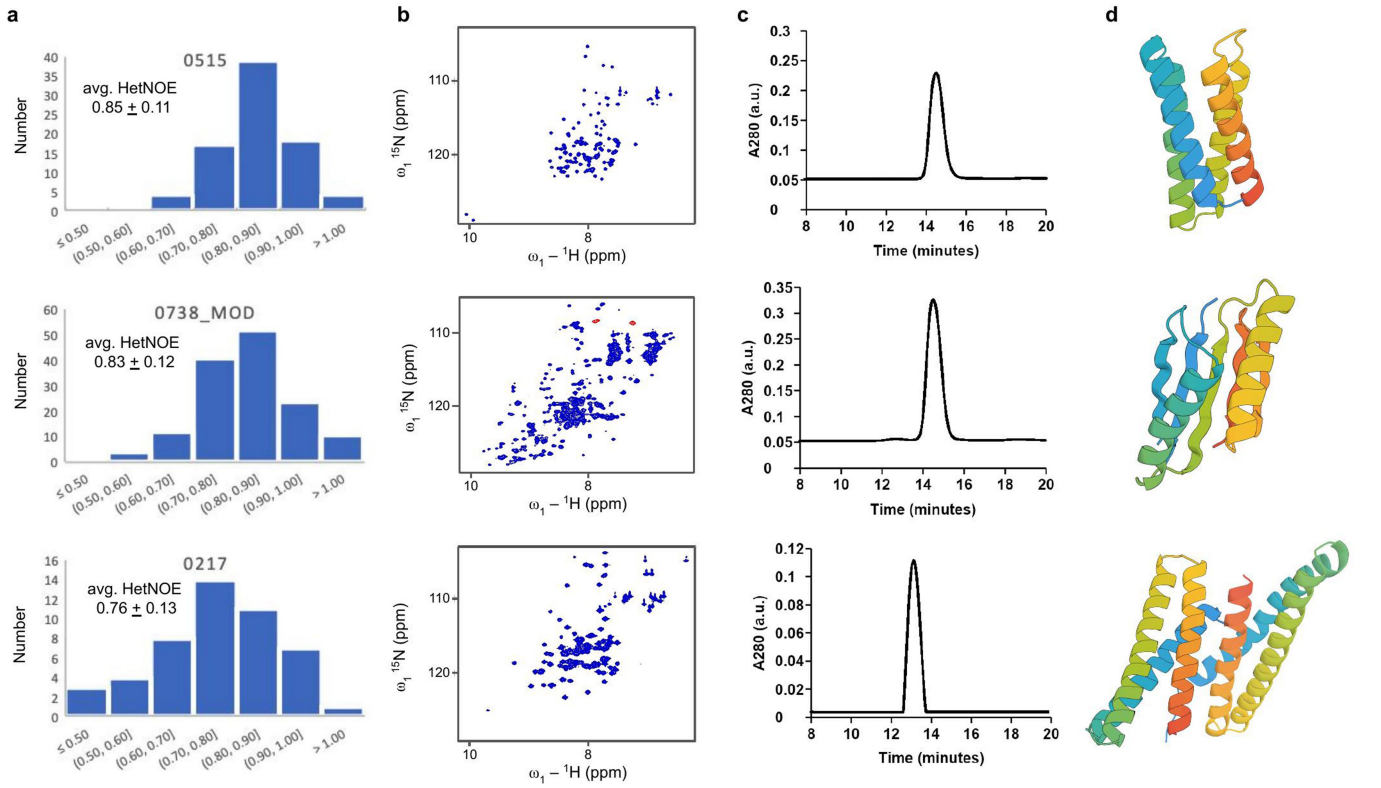
Extended Data Fig. 5 | Structural analysis of 0217 and comparison to hallucinated model. a) Representative electron density (2Fo-Fc, 1σ) over entire asymmetric unit (left) and core packing regions (right) of hallucination 0217. **b**) Both chains of the crystal structure colored by B-factor. **c**) Structural superposition of chains observed in the asymmetric unit reveal a 2.8 Å backbone r.m.s.d. over 91 residues. **d**) Crystal lattice contacts for chain A (green) and chain B (yellow) may explain structural differences observed between chains. Circled regions highlight where chain A is an ordered helix-loop-helix and chain B is disordered. **e**) Hallucinated model of 0217 colored by distance between Cα-Cα pairs between model and crystal structure after structural superposition and corresponding plot of per-residue Cα-Cα distance difference between model and crystal structure. **f**) Structural superposition of the hallucinated model and chain B of the 0217 crystal structure (left), 0217 model colored by Cα-Cα distance between hallucination and crystal structure (middle), and per residue Cα-Cα distance between hallucination and crystal structure per residue (right).



Extended Data Fig. 6 | Structural analysis, NMR characterization, and SEC analysis of hallucinated sequence 0417. a) Hallucinated model with surface hydrophobics shown as sticks and **b**) [^{1}H-^{15}N]-SOFAST-HMQC spectra of hallucinated sequence 0417 before (red) and after (blue) buffer optimization. Spectrum before optimization (red) was obtained using a protein concentration of ~0.3 mM at 298K in 20 mM Tris-HCl, pH 7.2, 100 mM NaCl and spectrum acquired after optimization (blue) was obtained using a protein concentration of ~0.3 mM, at temperature of 323 K in a buffer of 20 mM sodium phosphate at pH 6.5, 50 mM NaCl, and 20% glycerol. The NMR data are consistent with a folded structure containing a mix of alpha and beta secondary structure. Even under optimized conditions, there is still evidence of exchange broadening (e.g. Trp side chain N$^{\varepsilon}$Hs are weak), resonances that appear only at high temperature and high glycerol concentrations, and some resonances that are doubled; all indications of transient self-association. **c**) Size-exclusion chromatography trace of 0417 displays a small additional peak corresponding to a larger oligomeric species which corroborates the NMR analysis.



Extended Data Fig. 7 | Structural analysis of 0738_mod and comparison to hallucinated model 0738. a) Representative electron density (2Fo-Fc, 1σ) over entire asymmetric unit (left) and core packing regions (right) of hallucination 0738_mod. **b**) Both chains of the crystal structure colored by B-factor. **c**) Structural superposition of the hallucinated model and chain A of the 0738_mod crystal structure (left), 0738_mod model colored by Cα-Cα distance between hallucination and crystal structure (middle), and per residue Cα-Cα distance between hallucination and crystal structure per residue (right). **d**) Hallucinated model of 0738_mod colored by distance between Cα-Cα pairs between model and crystal structure after structural superposition and corresponding plot of per-residue Cα-Cα distance difference between model and crystal structure.



Extended Data Fig. 8 | NMR and biochemical analysis of hallucinated sequences 0515, 0738_mod, and 0217. a) ^{1}H-^{15}N heteronuclear NOE (hetNOE) histograms for 0515 (82 non-overlapped peaks), 0738_mod (144 peaks), and 0217 (47 peaks), together with their average values. ^{1}H-^{15}N steady state heteronuclear NOEs were obtained from the ratio of cross peak intensities ($I_{saturated}/I_{equilibrium}$) with ($I_{saturated}$) and without ($I_{equilibrium}$) 3 s of proton saturation during the presat delay and recorded in an interleaved manner, split in TopSpin, processed identically using NMRPipe, and peak picked in SPARKY to obtain peak intensities. **b**) ^{1}H-^{15}N HSQC spectra of corresponding proteins collected at 800 MHz at 298 K in 25 mM HEPES, pH 7.4, 50 mM NaCl buffer and prepared in a 5-mm Shigemi NMR tubes for data collection with addition of 5% D_2O (v/v). These ^{15}N-enriched protein samples were prepared at concentrations of 0.4 mM, 0.15 mM, and 0.2 mM, respectively. **c**) SEC data demonstrating monodispersity of these proteins in solution, with predominantly monomer for 0515 and 0738_mod and predominantly dimer for 0217. SDS-PAGE data (not shown) show that each is > 95% homogeneous, which together with MALDI-TOF mass spectrometry indicate that the spectral heterogeneity observed is not due to chemical heterogeneity. **d**) Ribbon diagrams of the corresponding monomeric or dimeric protein structures. These results show that the three designs have characteristic dynamics in solution. The average hetNOE for the homodimer 0217 is lower than for 0515 and 0738_mod, and it has fewer peaks than expected due to exchange broadening. Although 0738_mod has a similar hetNOE distribution as monomeric 0515, it has more than double the expected number of peaks, indicating at least two folded conformations (for all or parts of the protein) in solution that are in slow conformational exchange on the NMR time-scale. This was further validated by the appearance of new peaks in spectra at lower temperature (288 K), and different peaks at higher temperatures (308 and 318 K), and confirmed by detection of ^{15}N ZZ-exchange cross peaks at 318 K with 600 and 750 ms mixing times (Bruker pulse sequence hsqcetexf3gp, data not shown)[60].

Extended Data Table 1 | NMR refinement statistics and quality scores for 0515

Secondary Structure	α−helices: 3-23, 27-48, 52-70, 79-99
NMR conformationally-restricting restraints	
Distance restraints	
Total NOE	2092
Intra-residue	470
Inter-residue	
Sequential ($\|i-j\|=1$)	505
Medium-range ($\|i-j\|<4$)	675
Long-range ($\|i-j\|>5$)	398
Hydrogen bond	140
Total dihedral angle restraints	175
φ	89
ψ	86
No. of restraints per residue	24.1
No. of long-range restraints per residue	4.1
No. of HN RDC restraints	70
Violations (mean)	
Distance RMS violation/restraint (Å)	0.01
Dihedral angle RMS violation/restraint (º)	0.12
Max. dihedral angle violation (º)	3.50
Max. distance restraint violation (Å)	0.32
Deviations from idealized geometry	
Bond lengths (Å)	0.018
Bond angles (º)	1.1
Average pairwise r.m.s. deviation** (Å)	
Heavy (all / ordered[d])	1.2 / 1.0
Backbone (all / ordered[d])	0.7 / 0.5
Model quality statistics [b]	
Molprobity Ramachandran statistics	
Most favored, allowed, disallowed regions (%)	99.6, 0.4, 0
Global quality scores (Raw / Z-score)[c]	
Procheck (φ-ψ)[d]	0.61 / 2.71
Procheck (all)[d]	0.31 / 1.89
Molprobity Clashscore	7.99 / 0.15
Verify3D	0.25 / -3.37
Prosall	1.23 / 2.40
RDC Q RMSD scores[e]	0.20 ± 0.01
RPF scores[f]	
Recall/Precision	0.97 / 0.95
F-measure/DP-score	0.96 / 0.82

**Pairwise r.m.s. deviation was calculated for the 20 lowest energy refined structures out of 100 calculated.

[b]Calculated using PSVS1.5[52]. Average distance violations were calculated using the sum over r^{-6}.

[c]Structure-quality Z-scores are computed relative to mean and standard deviations for a set of 252 X-ray structures < 500 residues, of resolution ≤ 1.80 Å, R-factor ≤ 0.25 and R-free ≤ 0.28; a positive value indicates a 'better' score.

[d]Based on ordered residue ranges [S(phi) + S(psi) > 1.8], 3-72, 79-99.

[e]Calculated with PALES[59] .

[f]RPF scores reflect the goodness-of-fit of the final ensemble of structures (including disordered residues) to the NOESY data and resonance assignment[49].

Extended Data Table 2 | Crystallographic data collection and refinement statistics

	0217 (7K3H)	0738_mod (7M0Q)
Data collection		
Space group	I23	$P3_1$
Cell dimensions		
a, *b*, *c* (Å)	135.1, 135.1, 135.1	46.3, 46.3, 82.5
α, β, γ (°)	90, 90, 90	90, 90, 120
Resolution (Å)	47.8-3.0 (3.1-3.0)[b]	50.0-2.4 (2.49-2.40)
R_{merge}	0.09 (2.5)	0.08 (1.1)
$I / \sigma I$	57.5 (3)	18.0 (2.6)
Wilson B-factor	45.0	29.0
Completeness (%)	100 (100)	99.8 (100)
Redundancy	39.6 (40.6)	10.7 (8.2)
Refinement		
Resolution (Å)		
No. reflections	8049 (509)	7132 (405)
R_{work} / R_{free}	0.24/0.27 (0.32/0.40)	0.21/0.25 (0.31/0.31)
No. atoms		
Protein	1530	1495
Ligand/ion	0	0
Water	19	20
B-factors		
Protein	36.6	41.3
Water	26.9	35.1
R.m.s. deviations		
Bond lengths (Å)	0.001	0.002
Bond angles (°)	0.38	0.43

[a]Data were collected from a single crystal.
[b]Values in parentheses are for the highest-resolution shell.

nature research

Corresponding author(s): David Baker

Last updated by author(s): Jul 22, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The computer code which was used to generate the hallucinated proteins described in the manuscript was made publicly available as a part of trDesign Github package (https://github.com/gjoni/trDesign); corresponding structural models were generated with the trRosetta structure modeling tool available for free download at https://yanglab.nankai.edu.cn/trRosetta/download/. The Rosetta software suite was used to perform ab initio prediction calculations; Rosetta is freely available for academic users on Github, and can be licensed for commercial use by the University of Washington CoMotion Express License Program.
Data analysis	Generated designs were compared to known protein sequences and structures using BLAST (version 2.6.0) and TMalign (released on 2017/07/08) respectively. Crystallographic data were analyzed with PHENIX (release 1.18rc2-3793) and Coot (v0.8.9). NMR data were collected using Bruker TopSpin (v3.2pl7) and processed and analyzed with NMRPipe (v10.9), NMRFAM-Sparky (v1.370), iPine, Talos-N, PDBStat (5.20.4), Cyana (3.98.13), CNS (v1.3), ASDP (v2.3), PSVS (v1.5).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The atomic coordinates of the crystal structures for designs 0217 and 0738_mod, as well as the NMR structure for design 0515 have been deposited in the RCSB

Protein Data Bank with the accession numbers 7K3H, 7M0Q and 7M5T respectively. NMR chemical shifts, NOESY peak lists, and spectral data have been deposited in the BioMagResDB, BMRB ID 30890. Amino acid sequences and structure models for all 2K designs described in the manuscript are freely available for download at https://files.ipd.uw.edu/pub/trRosetta/hallucinations2K.tar.gz. Amino acid sequences and 3D structures of the generated designs were compared to known protein sequences and structures in UniProt (Uniref90 v2017_12 https://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2017_12/uniref/) and the Protein Data Bank (2020/03/11) respectively.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The number of designs to generate (two thousand) was determined based on the structural diversity of the resulting hallucinations such that the pool covers the three major fold classes: all alpha, all beta and mixed alpha-beta. By selecting hallucinations from 27 predominant clusters with 7 and more members we ensured to focus our efforts on analyzing the most representative folds produced by trRosetta which have a high chance to re-occur should another 2k sample be generated. The total sample size for experimental testing (n = 129) was determined based on estimated workload as well as to maximize coverage of the hallucinated structure space by various folds. Selected pool of structures covers all alpha, mixed alpha and beta, and all beta topologies in about equal amounts.
Data exclusions	No data were excluded from analysis.
Replication	Only one round of design generation was performed. Protein expression and solubility was tested once or twice. All attempts to replicate expression and solubility screening experiments for further experimental characterization were successful. Structural characterization was performed once or twice with internal statistical validation.
Randomization	There was no randomized sample allocation in this work. All tested protein designs received identical treatment
Blinding	Blinding was not relevant to this work, since all tested proteins received identical treatment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging