HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2022 July 18.

Published in final edited form as:

Nature. 2021 December ; 600(7889): 547–552. doi:10.1038/s41586-021-04184-w.

De novo protein design by deep network hallucination

Ivan Anishchenko[1,2,#], **Samuel J. Pellock**[1,2,#], **Tamuka M. Chidyausiku**[1,2], **Theresa A. Ramelot**[3], **Sergey Ovchinnikov**[4], **Jingzhou Hao**[3], **Khushboo Bafna**[3], **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], **David Baker**[1,2,5,*]

[1]Department of Biochemistry, University of Washington, Seattle, WA 98105

[2]Institute for Protein Design, University of Washington, Seattle, WA 98105

[3]Department of Chemistry and Chemical Biology, and Center for Biotechnology and Interdisciplinary Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180

[4]John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA 02138

[5]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105

Abstract

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]. We investigated 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 (KL-divergence) between the distance distributions predicted by the network and the background distribution. Optimization from different random starting points results in novel proteins spanning a very wide range of sequences and predicted structures. We obtained synthetic genes encoding 129 network hallucinated sequences, expressed and purified the proteins in *E. coli*, and found that

*Corresponding author (dabaker@uw.edu).

#These authors contributed equally.

Contributions

I.A., S.J.P., T.M.C. and D.B. designed the study. I.A., S.O. developed the hallucination methodology. I.A. performed all in silico studies. T.M.C., S.J.P., J.H., L.C., and C.M.C. expressed and purified proteins for CD, NMR, and crystallography experiments. T.M.C. and S.J.P. performed SEC–MALS and CD experiments. S.J.P. and A.K. performed crystallization screening. S.J.P., A.K.B., and F.D. determined crystal structures. S.J.P., C.N., G.T.M, and T.A.R. performed structural comparisons of design models and experimental structures. T.A.R., J.H., K.B. and G.T.M. performed NMR experiments and analyzed the data. I.A., S.J.P., T.A.R., G.T.M. and D.B. wrote the manuscript with input from all authors.

Code availability

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 by the *trRosetta* structure modeling 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.

Competing interests

G.T.M is a co-founder of Nexomics Biosciences, Inc.

27 folded to monodisperse species with circular dichroism spectra consistent with the hallucinated structures. We determined the 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 physically-based models, to the de novo design of proteins with new functions.

Introduction

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 which 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 described to date address the classical de novo protein design problem of simultaneously generating both a new backbone structure and an amino acid sequence which encodes it.

Deep neural networks trained to predict distances between amino acid residues in protein 3D 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 on 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 structure quite accurately for de novo designed proteins from just a single sequence -- i.e., in the complete absence of co-evolution information[3]. The trRosetta model also predicted effects of amino acid substitutions on folding 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.

We wondered if the information stored in the many parameters of the trRosetta network could be used to generate physically plausible backbones and amino acid sequences which encode them. Methods such as Google's DeepDream[18] take networks trained to recognize faces and other patterns in images, and invert these 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 decided to take 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 100 amino acid sequences, 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-residue sequence separation and protein length dependent background distributions (Fig. 1b,c and Methods; the background distributions are obtained from a network trained on the entire PDB). 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 the distance and orientation maps of the mutated sequence using the network, and accepting the move based on the change in the KL-divergence to the background distribution according to the standard Metropolis criterion (Fig. 1c and methods). The increase in KL-divergence aggregated over all 2,000 simulation trajectories is shown in Fig. 1d; in almost all cases after ~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 into protein 3D structures by direct minimization with trRosetta[3].

We used this approach to generate two thousand 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 ~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 specific 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 random number seeds converge on different sequence-structure pairs (Fig. 2a and b). A 2D map of the space spanned by the structures (Fig. 2b) was generated by multidimensional scaling of their pairwise 3D structural similarity (TM-score, see Methods). The structures span all alpha, all beta and mixed alpha-beta 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 striking 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 the latter. Both de novo designed proteins and the hallucinated proteins generated here have regular alpha helices and beta 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 129, the lowest energy structures found in the simulations were close to the corresponding hallucinated structures with Cα-RMSDs <3.0 Å, and for all 129 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 physically-based energy model, the network generated sequences do indeed encode the corresponding structures.

We next sought to determine how the computer hallucinations behave in the real world by obtaining synthetic genes for the 129 proteins, and expressing and purifying them from *E. coli* (see Methods). Of these, 27 yielded size exclusion chromatography (SEC) peaks corresponding to monomeric or small oligomeric species (Figs. 3d, 4d, Extended Data Fig. 3d, and Suppl. 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 and Extended Data Fig. 3e,k), with the characteristic profiles of all alpha helical proteins for the all alpha helical designs (Fig. 3e and Extended Data Fig. 3e), and of alpha-beta proteins for the alpha-beta designs (representatives are shown Fig. 4e and Extended Data Fig. 3k). Twenty-one of the proteins were highly thermostable with apparent melting temperatures above 70 °C (Extended Data Figs. 2d,h and 3f,l); the alpha-beta designs in Fig. 4 are particularly stable as none undergo unfolding transitions up to 95 °C (Extended Data Fig. 2h). 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 funneled landscapes leading into the target structure (Figs. 3c and 4c). Taken together, these data indicate that the network hallucinated proteins can fold into a wide range of stable structures with the predicted secondary structures.

We next investigated the three dimensional structures of the hallucinated proteins. We determined the solution NMR structure for design 0515 to be a monomeric antiparallel four-helix bundle (1D estimated ^{15}N T_1 ~ 780 ms, ^{15}N T_2 ~ 77 ms, and τ_c ~ 9.6 ns at 25 °C); the ensemble of 20 structures had a Cα root mean square deviation (RMSD) to the hallucinated model of ~1.84 Å (Extended Data Table 1, Fig. 5a, b, and Extended Data Fig. 4). We also succeeded in determining 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 RMSD between model and crystal structure is 2.53 Å over all 100 residues (Fig. 5c, d, and Extended Data Fig. 5). The agreement observed between these two experimental and hallucinated structures suggests 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, 0217 forms a dimer in the crystal structure (Extended Data Fig. 5), consistent with SEC-MALS analysis and NMR rotational correlation time measurements (^{15}N T_1 ~2.0 s, ^{15}N T_2 ~32 ms, and τ_c ~25 ns at 25 °C, Suppl. Fig. 2)[22]. Sequences generated by the network were modeled as monomers, but the 0217 model displays clear amphipathic sequence patterning across the 3-helix topology, with numerous solvent-exposed hydrophobic residues (Suppl. Fig. 2) that mediate the dimer contacts in the crystal structure. NMR data on design 0417 are consistent with the alpha/beta fold of the hallucinated model, and both SEC and NMR relaxation measurements (^{15}N T_1 ~730 ms, ^{15}N T_2 ~77 ms, and τ_c ~10.4 ns at 25 °C, Extended Data Fig. 6) indicate that it is primarily monomeric but with some transient self-association in solution. The network appears to incorporate sequence features associated with the protein-protein interfaces of the native oligomeric proteins included in the PDB training set, likely explaining why many of the network hallucinated proteins self associate. Consistent with this, we found that substitution of surface hydrophobic residues with polar residues in a subset of the hallucinations that formed oligomers resulted in monomeric species by SEC (Suppl. Fig. 3; Supplementary discussion).

One of these surface-modified hallucinations, 0738_mod, yielded crystals and the 2.4 Å resolution structure was determined. Structural superposition of the 0738 model and 0738_mod crystal structure revealed a 3.68 Å Cα RMSD over 96 residues (Fig. 5f, g, and Extended Data Fig. 7). Despite register shifts upon superposition of the entire structure and model, the N- and C-terminal halves of the crystal structure align remarkably well to the corresponding regions in the hallucinated model with backbone RMSDs 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 0.021 (Suppl. Table 1). To determine if the lack of explicit 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 designs 0515 and 0217 (dimer) suggest 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; 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 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 10 years have sought to distill the key features of protein structures and protein sequence-structure relationships using physically-based models like Rosetta, and then have used these models to design idealized structures that embody these features based on the principle that proteins fold to their lowest free energy states[23,24]. The resemblance of the hallucinated structures to these idealized proteins -- in the regularity of the secondary structures, shortness of the loops, etc. -- is remarkable. Indeed, the most similar structure in the PDB to the 0738_mod structure is the de novo designed protein Top7 (Suppl. 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 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 and 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.

Our work opens up a large set of exciting avenues to explore. On the sampling side, the Monte Carlo approach could perhaps 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[26] or catalytic sites, around which the network can hallucinate new protein inhibitors or enzyme catalysts. Unlike traditional protein design calculations, where 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. In our hallucination approach, trRosetta can be replaced by recently developed higher accuracy protein structure prediction methods[27,28], which should increase the precision with which new proteins can be hallucinated. 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.

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:

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

The first term on the right, *P(structure|sequence)*, is related to the protein structure prediction problem where one seeks for the most probable structure for a given protein sequence, while 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 that 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 which takes 1- and 2-site features derived from a multiple sequence alignment or a single sequence as an input and produces a 2D output 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 5 inter-residue angles (comprising the full set of 6 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, while 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 (Suppl Fig. 7; this can loosely be viewed as representing a generic “molten globule” state). 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 KL-divergence over all residue pairs (i, j) and distance and angle distributions

$$D_{KL}\left(P_{trRosetta} \middle|\middle| Q_{background}\right) = \sum_{x \in \{d,\omega,\theta,\varphi\}} \left[\frac{1}{L^2}\sum_{i,j=1}^{L}\sum_{k=1}^{N_x} p_{x,ijk} \log\left(\frac{p_{x,ijk}}{q_{x,ijk}}\right)\right] \quad (4)$$

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 KL-divergence of the amino acid composition of a sequence from that of the PDB as a whole

$$-D_{KL}\left(f_a\middle|\middle|f_a^{PDB}\right) = -\sum_{a=1}^{20} f_a log\left(\frac{f_a}{f_a^{PDB}}\right) \quad (5)$$

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

Protein hallucination.

We optimized the combined objective function

$$F = D_{KL}\left(P_{trRosetta}\middle|\middle|Q_{background}\right) - D_{KL}\left(f_a\middle|\middle|f_a^{PDB}\right) \quad (6)$$

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:

$$A_i = min[1, exp(-(F_i - F_{i-1})/T)] \quad (7)$$

(i.e., if A_i is smaller than a uniform random number $u \in [0, 1]$). 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 modeling score (TM-score)[29]. 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 (Eq. 4), and the cross entropy between the final hallucinated structure Y and the 6D coordinate distributions generated by trRosetta for the sequence:

$$score = D_{KL}(P||Q) + [CE(Q,Y) - CE(P,Y)] \quad (8)$$

$$CE(P,Y) = \sum_{x \in \{d,\omega,\theta,\varphi\}} \left[\frac{1}{L^2}\sum_{i,j=1}^{L}\sum_{k=1}^{N_x} y_{x,ijk} log\left(p_{x,ijk}\right)\right] \quad (9)$$

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 Eq. 8 [$CE(Q, Y) - CE(P, Y)$] 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. One-hundred twenty-nine 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 kanamycin). These overnight cultures were used to inoculate either 50 mL (for screening) or 500 mL (for crystallography) of Studier autoinduction media[30] 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 size-exclusion chromatography 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 analyzed by circular dichroism (CD) or flash frozen in liquid nitrogen for later analysis. The resulting samples were generally >95% homogeneous on SDS-PAGE gels.

Circular dichroism experiments.

To determine secondary structure and thermostability of the designs far-ultraviolet CD measurements were carried out with an JASCO 1500. 260 to 195 nm wavelength scans were measured at every 10 °C interval from 25 to 95 °C. Temperature melts monitored dichroism signal at 220 nm in steps of 2 °C/min with 30 s of equilibration time. Wavelength scans and temperature melts were performed using 0.35 mg/ml 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 (ThermoScientific) using predicted extinction coefficients[31].

NMR sample preparation.

Samples for NMR studies were prepared following standard protocols developed by the Northeast Structural Genomics Consortium[32,33]. Initial sample preparation was carried out on a fee-for-service basis by Nexomics Biosciences, Inc. Selected designs were expressed in *E. coli* BL21 (DE3) cells as U-^{15}N-enriched proteins, using MJ9 minimal media[34] with

antibiotic kanamycin (50 μg/ml), and $(^{15}NH_4)_2SO_4$ as the sole source of nitrogen. For midi-scale production[33], 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 hr. 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 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 >~98% homogeneous, based on SDS-PAGE. Samples were concentrated to ~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 L cultures were prepared using MJ9 minimal media[34] 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 harvested by centrifugation (2,270 × g for 1 hr), 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 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 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 [^1H-^{15}N]-HSQC or [^1H-^{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 rotational correlation times, τ_c[22]. RDC data collection on both isotropic and partially aligned samples was performed at 600 MHz using a 2D interleaved ^{15}N-^1H- HSQC IPAP experiment to measure couplings[35]. All NMR spectra were processed using NMRPipe and NMRDraw[36] and visualized in NMRFAM-SPARKY[37]. Backbone resonance assignments for 0515

were determined using a standard triple-resonance NMR strategy, with a suite of fast pulsing and BEST double and triple resonance experiments provided within NMRlib[38], 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[39] was used for NUS within TopSpin 3.2 (Bruker) and subsequently reconstructed using sparse multidimensional iterative lineshape-enhanced[40] (SMILE) reconstruction within NMRPipe[36]. Resonance assignments were determined by manual refinement of resonance assignments obtained from the I-PINE web server[41]. Assignment validation was done with cmap images generated using AutoAssign[42]. Peak intensities from 3D NOESY spectra, together with dihedral angle constraints determined from backbone chemical shift data using TALOS-N[43], were used as input for structure determination. NOESY peak assignments were made automatically using Cyana[44,45], together with the programs RPF and ASDP to guide manual correction of NOESY peak assignments[46,47]. The lowest energy 20, of 100 structures calculated, were then refined in explicit water using CNS[48] 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[49] (Suppl. Table 1). Resonance assignments and NMR data were deposited in the BioMagResDataBase (ID 30890), and coordinates and restraints in the PDB (ID 7M5T).

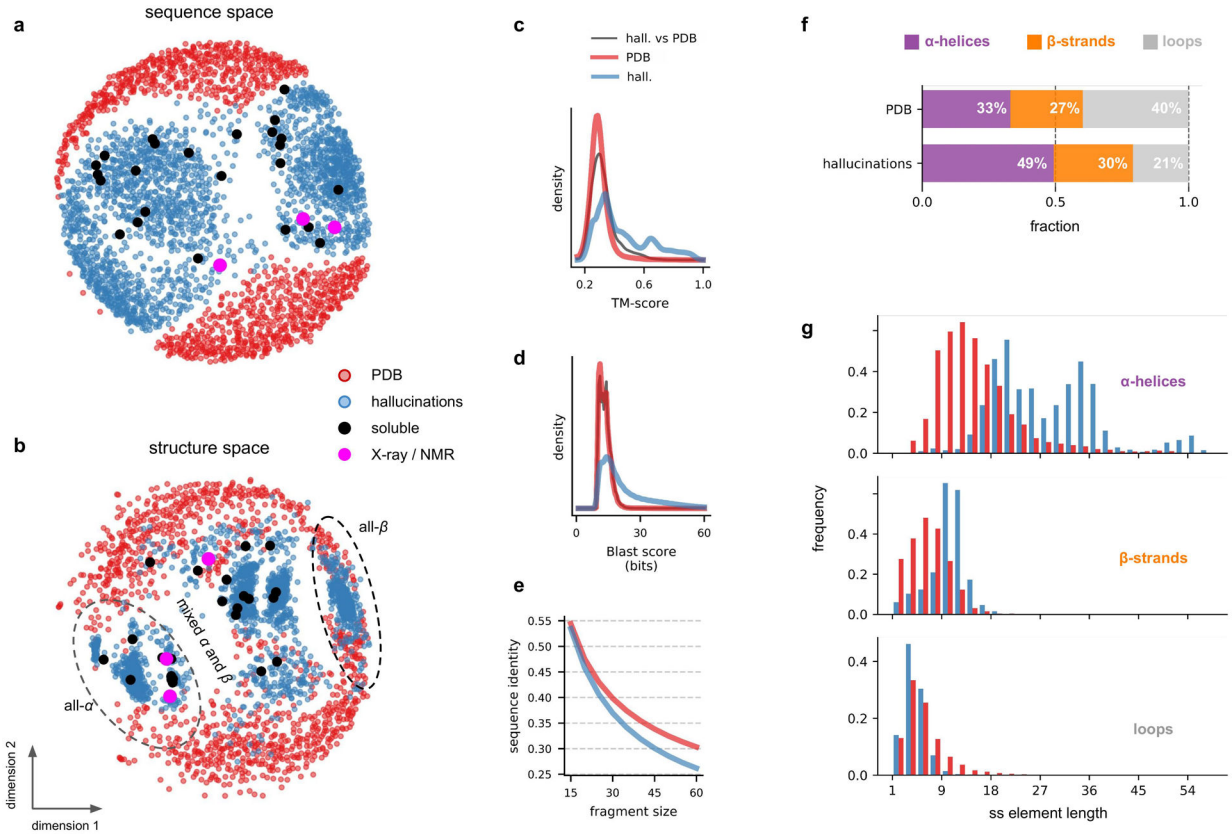
Crystallography sample preparation, data collection, and analysis.

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[50]. For 0217, molecular replacement (MR) 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 2000 models (1000 trRosetta and 1000 *ab initio*) had all side chains removed past the gamma carbon, and was run through Phaser[51]. 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[52], 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[53] and cycles of Phenix refinement[54] were used to build the final model. For 0217 final Ramachandran favored and outliers were 99% and 0%, respectively. For 0738_mod refinement, final Ramachandran favored 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[55]. 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 RMSDs are reported. Protein structure figures were made in PyMOL[56].

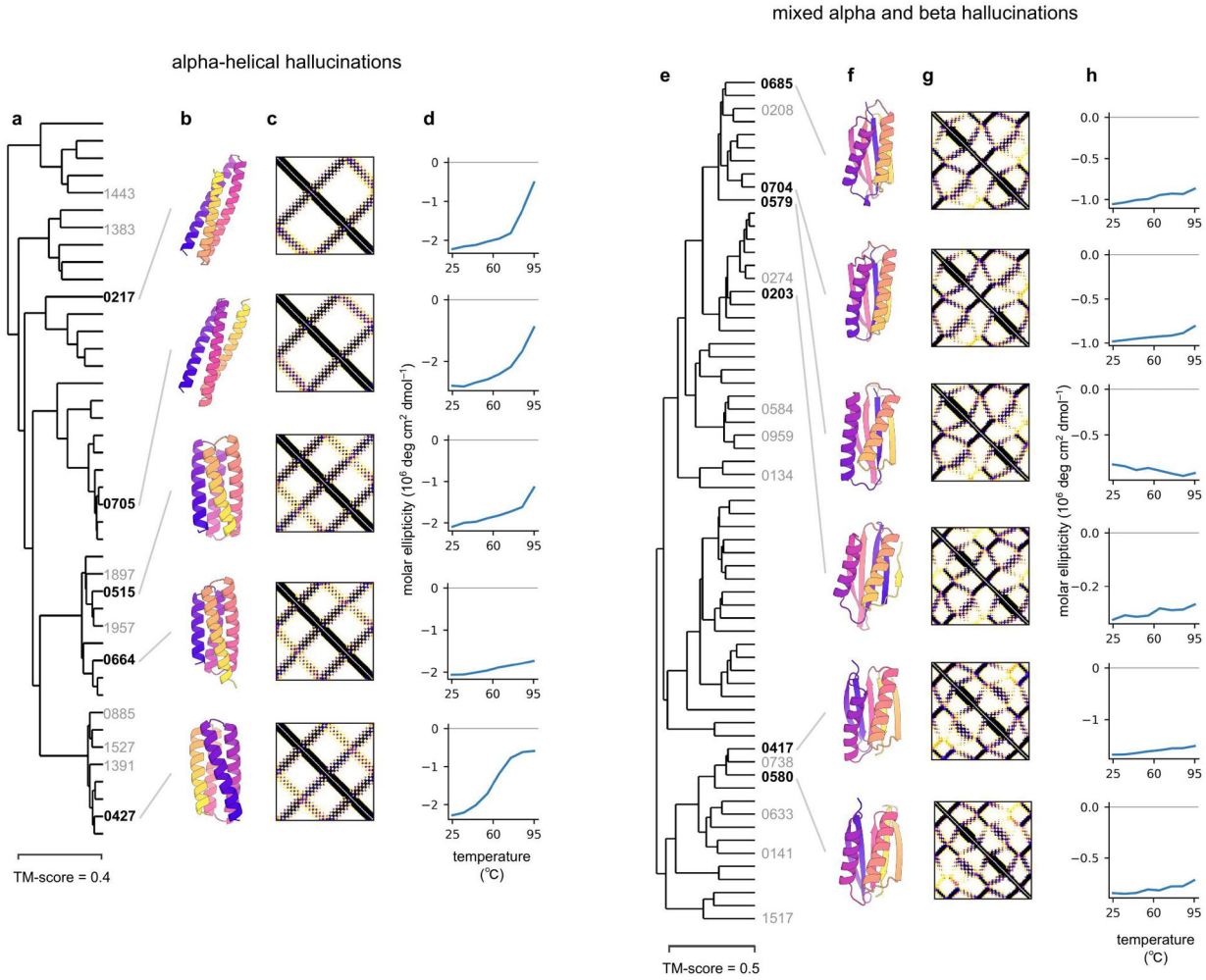
Extended Data



Extended Data Figure 1.

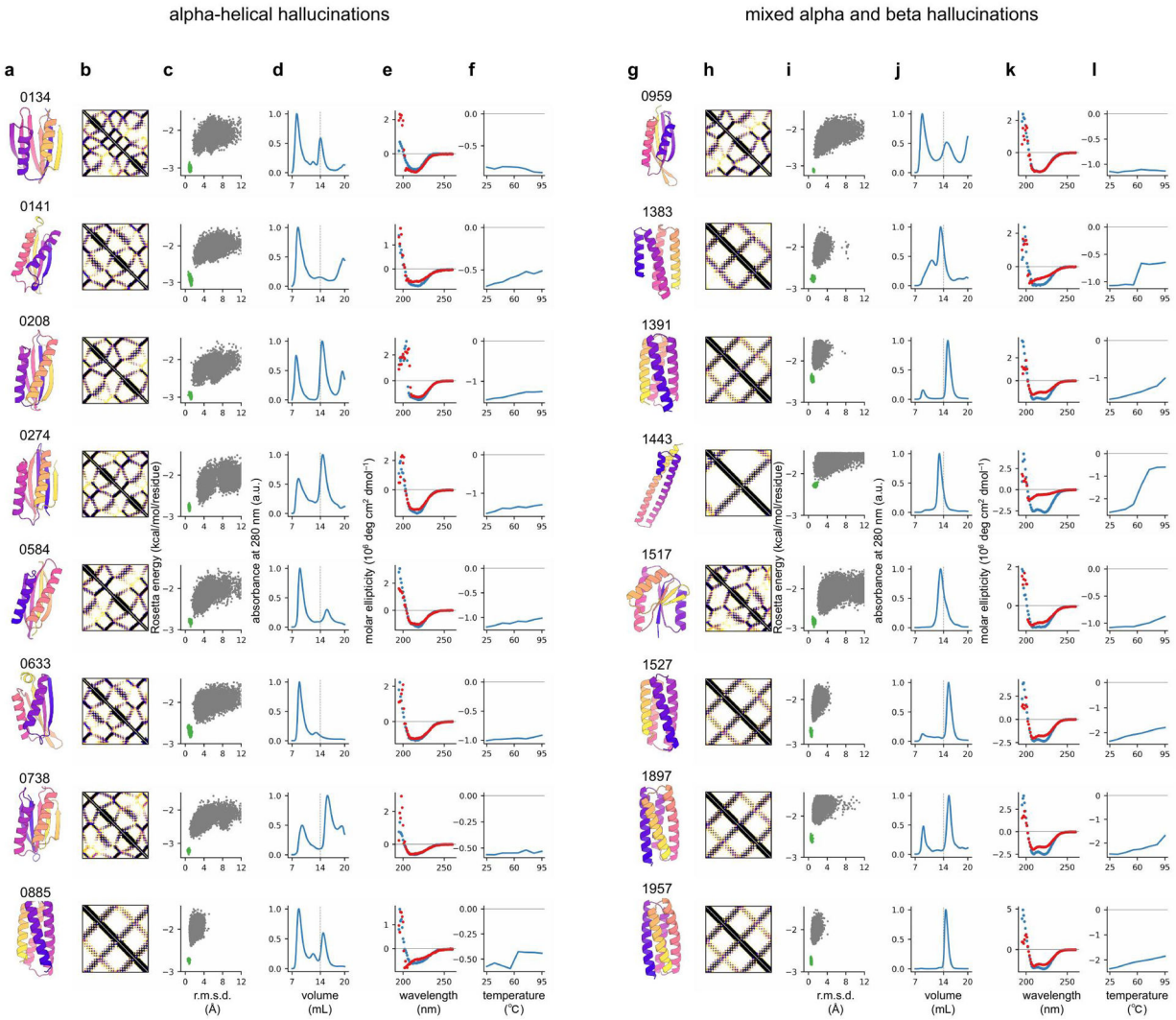
Comparison of the hallucinated designs to proteins with known structure and of similar length (100 ± 10 aa) from the trRosetta training set. **a,b**) Multidimensional scaling plots of the sequence (**a**) and structure (**b**) spaces covered by the 2,000 hallucinated proteins (blue dots) along with 1,110 proteins of similar length from the trRosetta training set (red dots). These scatter plots show that subspaces spanning 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.



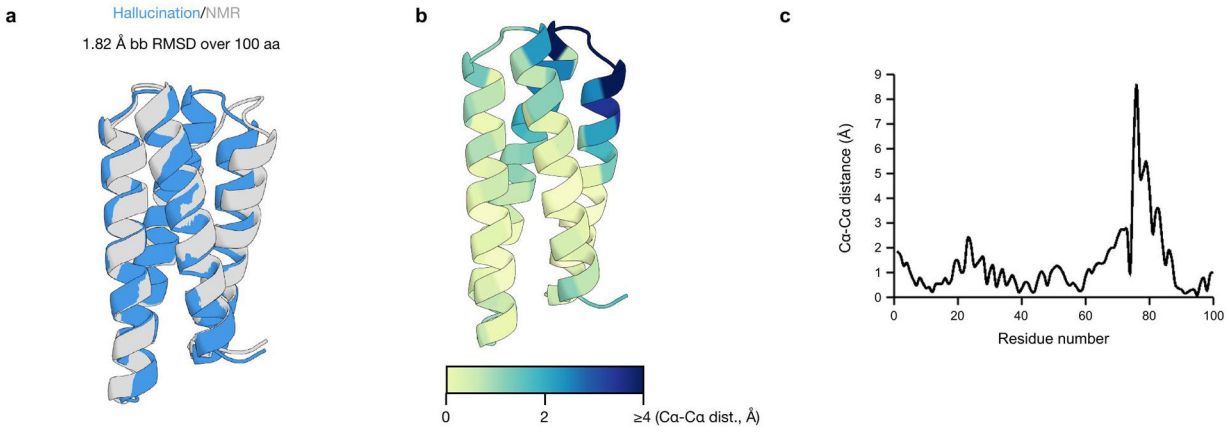
Extended Data Figure 2.

Structure similarity dendrograms (**a,e**), 3D structure models (**b,f**), predicted distance maps (**c,g**), and temperature dependence of circular dichroism signal at 220 nm in the 25–95 °C temperature range (**d,h**) for all-alpha and mixed alpha and beta hallucinations respectively.



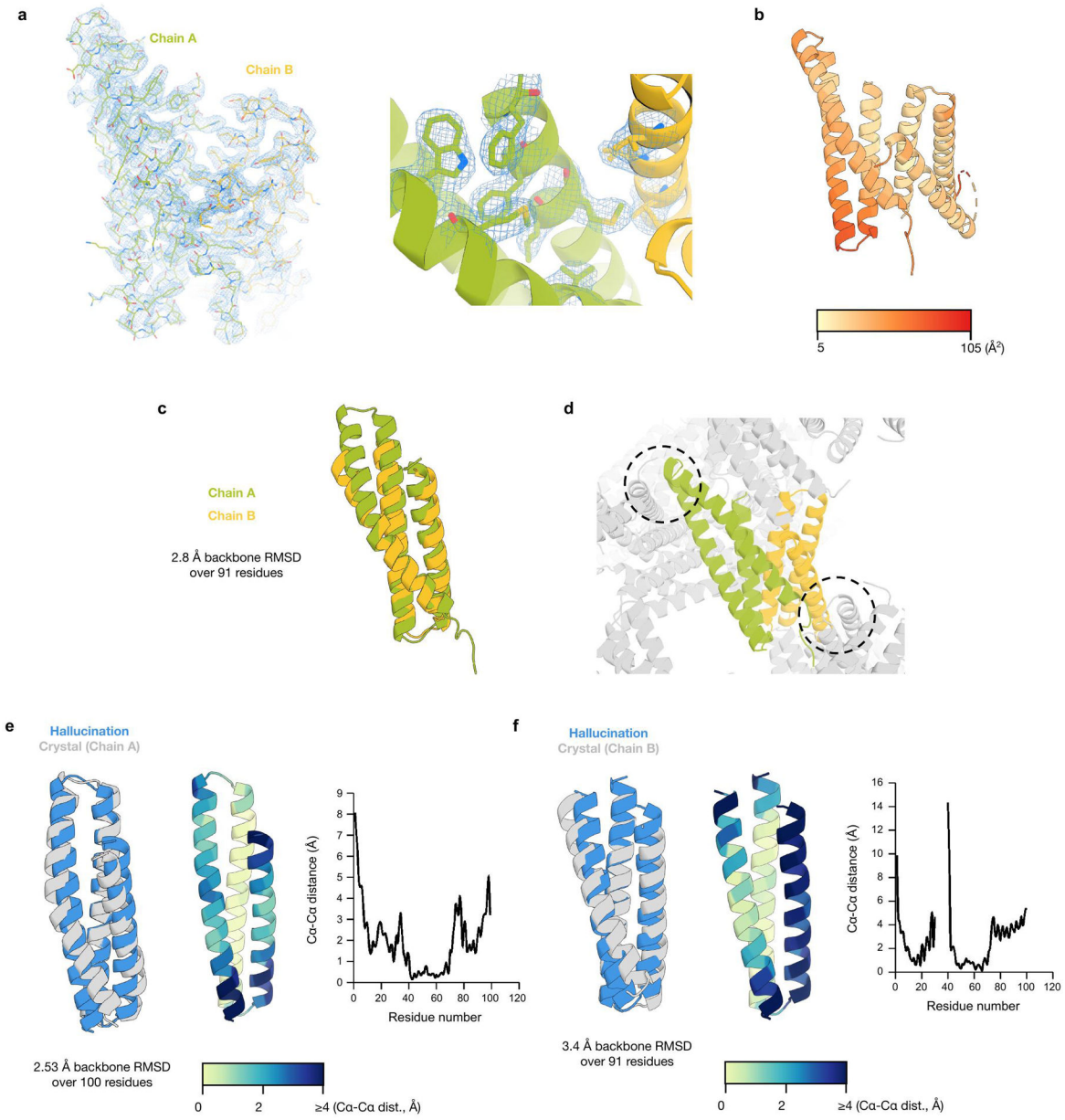
Extended Data Figure 3.

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



Extended Data Figure 4.

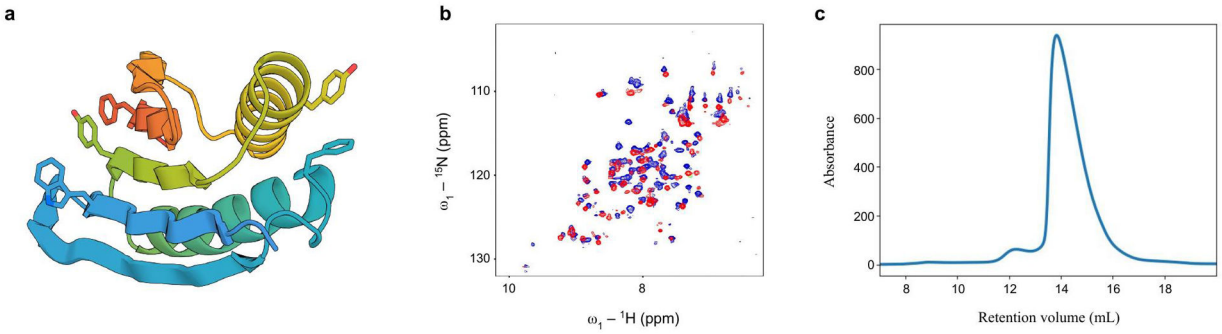
a) Superposition of hallucinated model (blue) and NMR medoid structure (gray) of 0515 reveal 1.85 Å backbone RMSD over 100 residues **b**) Hallucinated model of 0515 colored by distance between Cα-Cα pairs between model and NMR medoid structure after structural superposition and **c**) corresponding plot of per-residue Cα-Cα distance difference between model and NMR medoid structure.



Extended Data Figure 5.

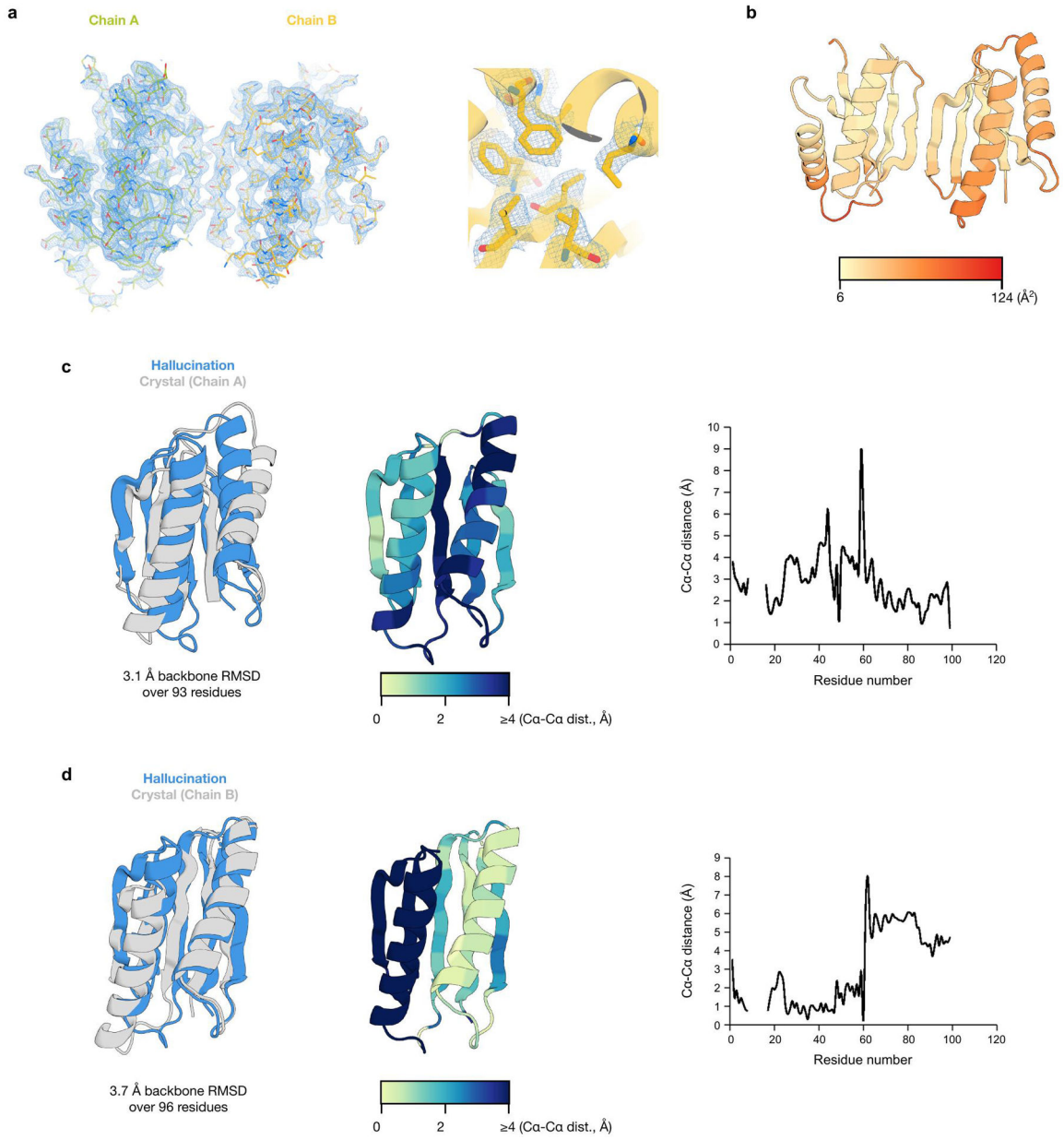
a) Representative electron density (2Fo-Fc, 1σ) over entire asymmetric unit (left) and core packing regions (right) of hallucination 0217. **b**) Both chains of the crystal structure colored by B-factor. **c**) Structural superposition of chains observed in the asymmetric unit reveal a 2.8 Å backbone RMSD 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).



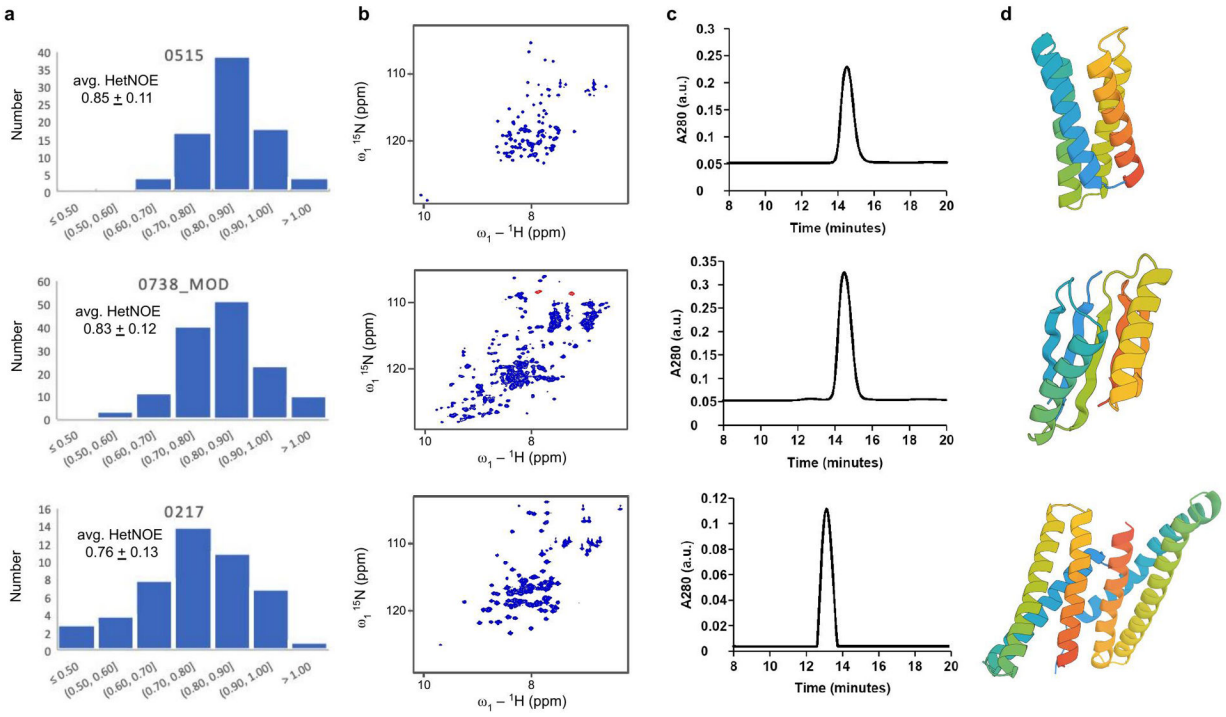
Extended Data Figure 6.

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^{e}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 Figure 7.

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 Figure 8.

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 sec 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 (288K), and different peaks at higher temperatures (308 and 318K), and confirmed by

detection of ^{15}N ZZ-exchange cross peaks at 318K with 600 and 750 ms mixing times (Bruker pulse sequence hsqcetexf3gp, data not shown)[58].

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 (°)	0.12
Max. dihedral angle violation (°)	3.50
Max. distance restraint violation (Å)	0.32
Deviations from idealized geometry	
Bond lengths (Å)	0.018
Bond angles (°)	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[46]. 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[55].

[f] RPF scores reflect the goodness-of-fit of the final ensemble of structures (including disordered residues) to the NOESY data and resonance assignment[43].

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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank R. Xiao, G. Liu and A. Wu, Nexomics Biosciences, Inc, for assistance in initial NMR protein production, and J. Aramini, City College of New York, for assistance in NMR data collection for initial HSQC screening. We thank R. Ballard and X. Li for Mass Spectrometry assistance as well as R. Divine and R. Kibler for AKTA scripting. This work was funded by grants from the NSF # DBI 1937533 (D.B. and I.A.), the NIH # DP5OD026389 (S.O.), Open Philanthropy (C.C., A.B.), Eric and Wendy Schmidt by recommendation of the Schmidt Futures program (F.D., 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 grant R01 GM120574 (to 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.

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 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.

References

1. Xu J Distance-based protein folding powered by deep learning. Proc. Natl. Acad. Sci. U. S. A 116, 16856–16865 (2019). [PubMed: 31399549]
2. Senior AW et al. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710 (2020). [PubMed: 31942072]
3. Yang J et al. Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. U. S. A 117, 1496–1503 (2020). [PubMed: 31896580]
4. Biswas S, Khimulya G, Alley EC, Esvelt KM & Church GM Low-N protein engineering with data-efficient deep learning. Nat. Methods 18, 389–396 (2021). [PubMed: 33828272]
5. Madani A et al. ProGen: Language Modeling for Protein Generation. bioRxiv (2020) doi:10.1101/2020.03.07.982272.
6. Anand N, Eguchi R & Huang PS Fully differentiable full-atom protein backbone generation. In ICLR 2019 Workshop (2019).
7. Wang J, Cao H, Zhang JZH & Qi Y Computational Protein Design with Deep Learning Neural Networks. Sci. Rep 8, 6349 (2018). [PubMed: 29679026]
8. Ingraham J, Garg VK, Barzilay R & Jaakkola T Generative Models for Graph-Based Protein Design. In ICLR 2019 Workshop (2019).
9. Anand N, Eguchi RR, Derry A, Altman RB & Huang P-S Protein Sequence Design with a Learned Potential. bioRxiv (2020) doi:10.1101/2020.01.06.895466.

10. Strokach A, Becerra D, Corbi-Verge C, Perez-Riba A & Kim PM Fast and Flexible Protein Design Using Deep Graph Neural Networks. Cell Syst 11, 402–411.e4 (2020). [PubMed: 32971019]
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). [PubMed: 32945673]
12. Davidsen K et al. Deep generative models for T cell receptor protein sequences. eLife 8, e46935 (2019). [PubMed: 31487240]
13. Costello Z & Martin HG How to Hallucinate Functional Proteins. arXiv:1903.00458 (2019).
14. Eguchi RR, Anand N, Choe CA & Huang P-S IG-VAE: Generative Modeling of Immunoglobulin Proteins by Direct 3D Coordinate Generation. bioRxiv (2020) doi:10.1101/2020.08.07.242347.
15. Repecka D et al. Expanding functional protein sequence spaces using generative adversarial networks. Nat. Mach. Intell 1–10 (2021).
16. Hawkins-Hooker A et al. Generating functional protein variants with variational autoencoders. PLoS Comput. Biol 17, e1008736 (2021). [PubMed: 33635868]
17. Senior AW 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). [PubMed: 31602685]
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 SF, Gish W, Miller W, Myers EW & Lipman DJ Basic local alignment search tool. J. Mol. Biol 215, 403–410 (1990). [PubMed: 2231712]
20. Rohl CA, Strauss CEM, Misura KMS & Baker D Protein Structure Prediction Using Rosetta. Methods Enzymol. 383, 66–93 (2004). [PubMed: 15063647]
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). [PubMed: 27766851]
22. Rossi P et al. A microscale protein NMR sample screening pipeline. J. Biomol. NMR 46, 11–22 (2010). [PubMed: 19915800]
23. Koga N et al. Principles for designing ideal protein structures. Nature 491, 222–227 (2012). [PubMed: 23135467]
24. Dou J et al. De novo design of a fluorescence-activating β-barrel. Nature 561, 485–491 (2018). [PubMed: 30209393]
25. Norn C et al. Protein sequence design by conformational landscape optimization. Proc. Natl. Acad. Sci. U. S. A 118, (2021).
26. Tischer D et al. Design of proteins presenting discontinuous functional sites using deep learning. bioRxiv (2020) doi:10.1101/2020.11.29.402743.
27. Baek M et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science (2021) doi:10.1126/science.abj8754.
28. Jumper J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021) doi:10.1038/s41586-021-03819-2.
29. Zhang Y & Skolnick J TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302–2309 (2005). [PubMed: 15849316]
30. Studier FW Protein production by auto-induction in high density shaking cultures. Protein Expr. Purif 41, 207–234 (2005). [PubMed: 15915565]
31. Pace CN, 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). [PubMed: 8563639]
32. Acton TB et al. Preparation of protein samples for NMR structure, function, and small-molecule screening studies. Methods Enzymol. 493, 21–60 (2011). [PubMed: 21371586]
33. Xiao R et al. The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. J. Struct. Biol 172, 21–33 (2010). [PubMed: 20688167]
34. 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). [PubMed: 8616269]

35. Ottiger M, Delaglio F & Bax A Measurement ofJand Dipolar Couplings from Simplified Two-Dimensional NMR Spectra. Journal of Magnetic Resonance vol. 131 373–378 (1998). [PubMed: 9571116]
36. Delaglio F et al. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J. Biomol. NMR 6, 277–293 (1995). [PubMed: 8520220]
37. Lee W, Tonelli M & Markley JL NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics 31, 1325–1327 (2015). [PubMed: 25505092]
38. Favier A & Brutscher B NMRlib: user-friendly pulse sequence tools for Bruker NMR spectrometers. J. Biomol. NMR 73, 199–211 (2019). [PubMed: 31076970]
39. Hyberts SG, Milbradt AG, Wagner AB, 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). [PubMed: 22331404]
40. Ying J, Delaglio F, Torchia DA & 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). [PubMed: 27866371]
41. Lee W et al. I-PINE web server: an integrative probabilistic NMR assignment system for proteins. J. Biomol. NMR 73, 213–222 (2019). [PubMed: 31165321]
42. Moseley HNB, Sahota G & Montelione GT Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. J. Biomol. NMR 28, 341–355 (2004). [PubMed: 14872126]
43. 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). [PubMed: 23728592]
44. 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). [PubMed: 9367762]
45. 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). [PubMed: 12522306]
46. Huang YJ, Powers R & Montelione GT 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). [PubMed: 15701001]
47. Huang YJ, Tejero R, Powers R & Montelione GT A topology-constrained distance network algorithm for protein structure determination from NOESY data. Proteins 62, 587–603 (2006). [PubMed: 16374783]
48. Brünger AT et al. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr 54, 905–921 (1998). [PubMed: 9757107]
49. Bhattacharya A, Tejero R & Montelione GT Evaluating protein structures determined by structural genomics consortia. Proteins 66, 778–795 (2007). [PubMed: 17186527]
50. Otwinowski Z & Minor W Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol. 276, 307–326 (1997).
51. McCoy AJ et al. Phaser crystallographic software. J. Appl. Crystallogr 40, 658–674 (2007). [PubMed: 19461840]
52. DiMaio F et al. Improved low-resolution crystallographic refinement with Phenix and Rosetta. Nature Methods 10, 1102–1104 (2013). [PubMed: 24076763]
53. Emsley P, Lohkamp B, Scott WG & Cowtan K Features and development of Coot. Acta Crystallogr. D Biol. Crystallogr 66, 486–501 (2010). [PubMed: 20383002]
54. Liebschner D et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr. D Biol. Crystallogr 75, 861–877 (2019).
55. Theobald DL & Wuttke DS Accurate structural correlations from maximum likelihood superpositions. PLoS Comput. Biol 4, e43 (2008). [PubMed: 18282091]
56. Schrödinger LLC. The PyMOL Molecular Graphics System, Version 2.4. (2021).
57. Zweckstetter M NMR: prediction of molecular alignment from structure using the PALES software. Nat. Protoc 3, 679–690 (2008). [PubMed: 18388951]

58. Montelione GT & Wagner G 2D Chemical exchange NMR spectroscopy by proton-detected heteronuclear correlation. Journal of the American Chemical Society vol. 111 3096–3098 (1989).

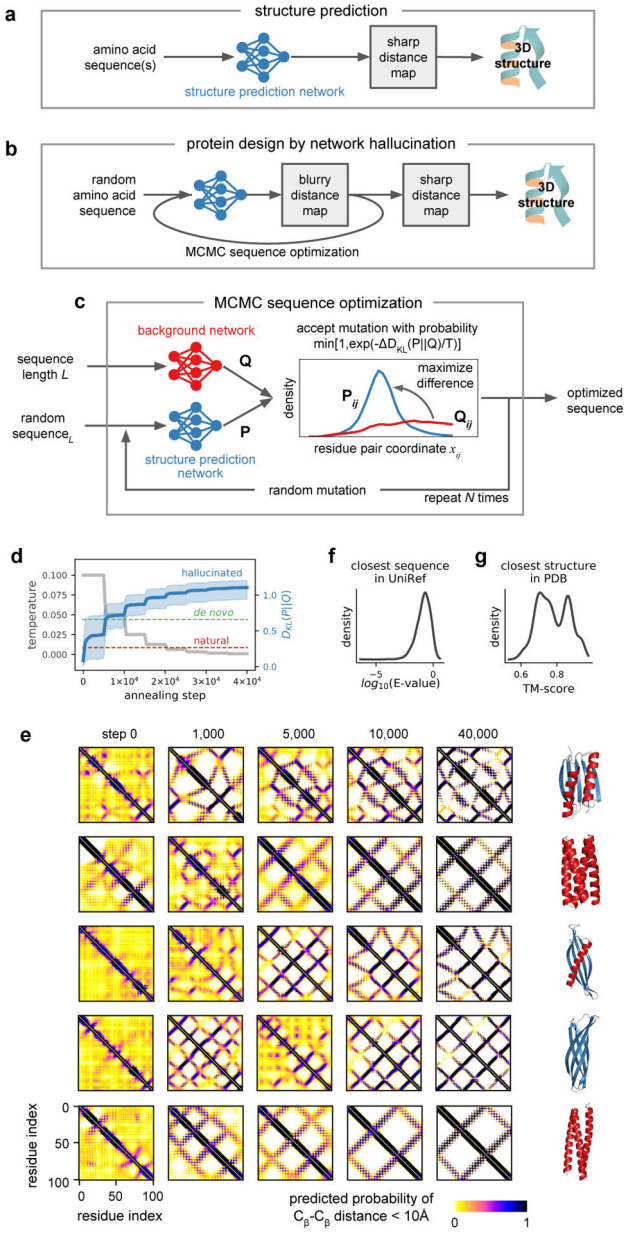


Figure 1:

Overview of protein hallucination approach. **a)** In structure prediction using trRosetta and other recent methods, a deep neural network is used 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 Markov chain Monte Carlo algorithm, optimizing the sharpness of the 2D maps. **c)** Schematic of the MCMC procedure. **d)** Annealing trajectories averaged over 2,000 runs show a monotonic increase in the KL-divergence (contrast of the distance maps) with increasing Monte Carlo

optimization. The mean and 0.01,0.99 quantiles are shown in blue; temperature profile 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 the protein structures in the PDB with average TM-scores to the closest match of 0.78.

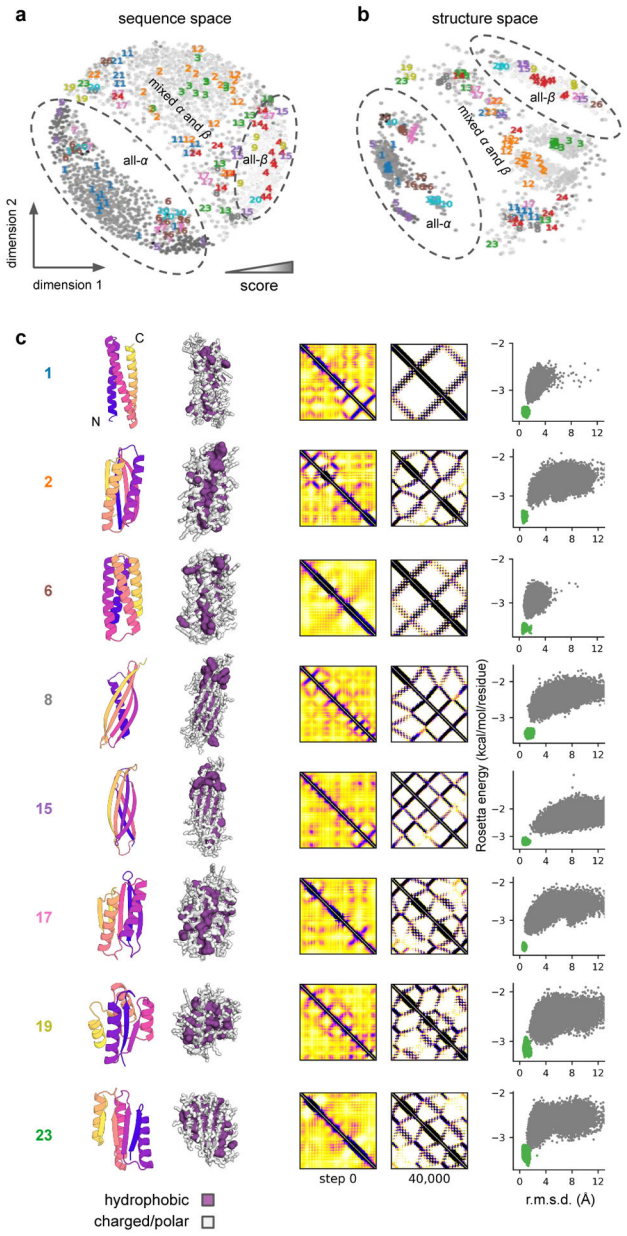


Figure 2:

Overview of computational results. **a)** Multidimensional scaling plot 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 color-coded by the score from the network (darker grey color corresponds to higher score). 129 experimentally tested designs belong to 27 structural clusters shown by colored numbers. **b)** Multidimensional scaling plot of the structural space covered by the 2,000 hallucinated proteins; (1 - TM-score) was used to measure the distance between proteins. Each grey dot represents one design; the gray-scale indicates the score from the network (darker grey corresponds to higher score, Eq. 8). The 129 experimentally tested designs fall into 27 structural clusters shown by colored numbers. **c)** Examples of hallucinated designs of various topologies.

First column, ribbon depiction of protein backbone colored from blue (N-terminus) to red (C-terminus); second column, hydrophobic core; third column, distance maps at the beginning and end of hallucination trajectory, and 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.

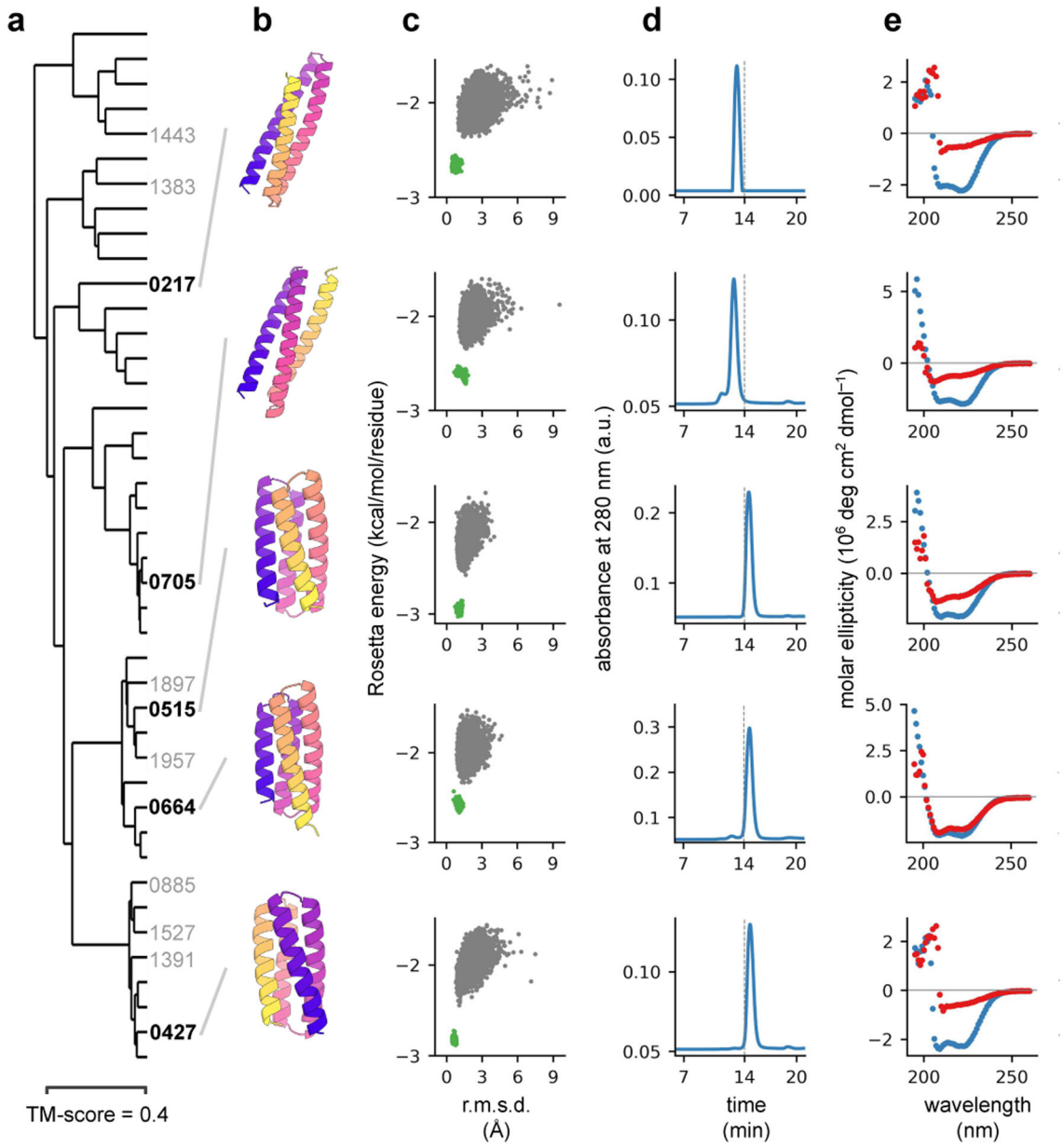


Figure 3.

Experimental characterization of alpha-helical hallucinations. **a)** Dendrogram showing 42 all-alpha designs clustered by structural similarity (TM-score); thermostable designs with CD spectra consistent with the target structure are labeled by their IDs. **b)** 3D structure 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 (blue) and 95 (red) °C. Contact maps and full temperature melts are in Extended Data Fig. 2, and additional examples of stable alpha-helical designs marked in grey in panel **a** are shown in Extended Data Fig. 3, Size-exclusion and CD plots are representative plots of duplicate experiments.

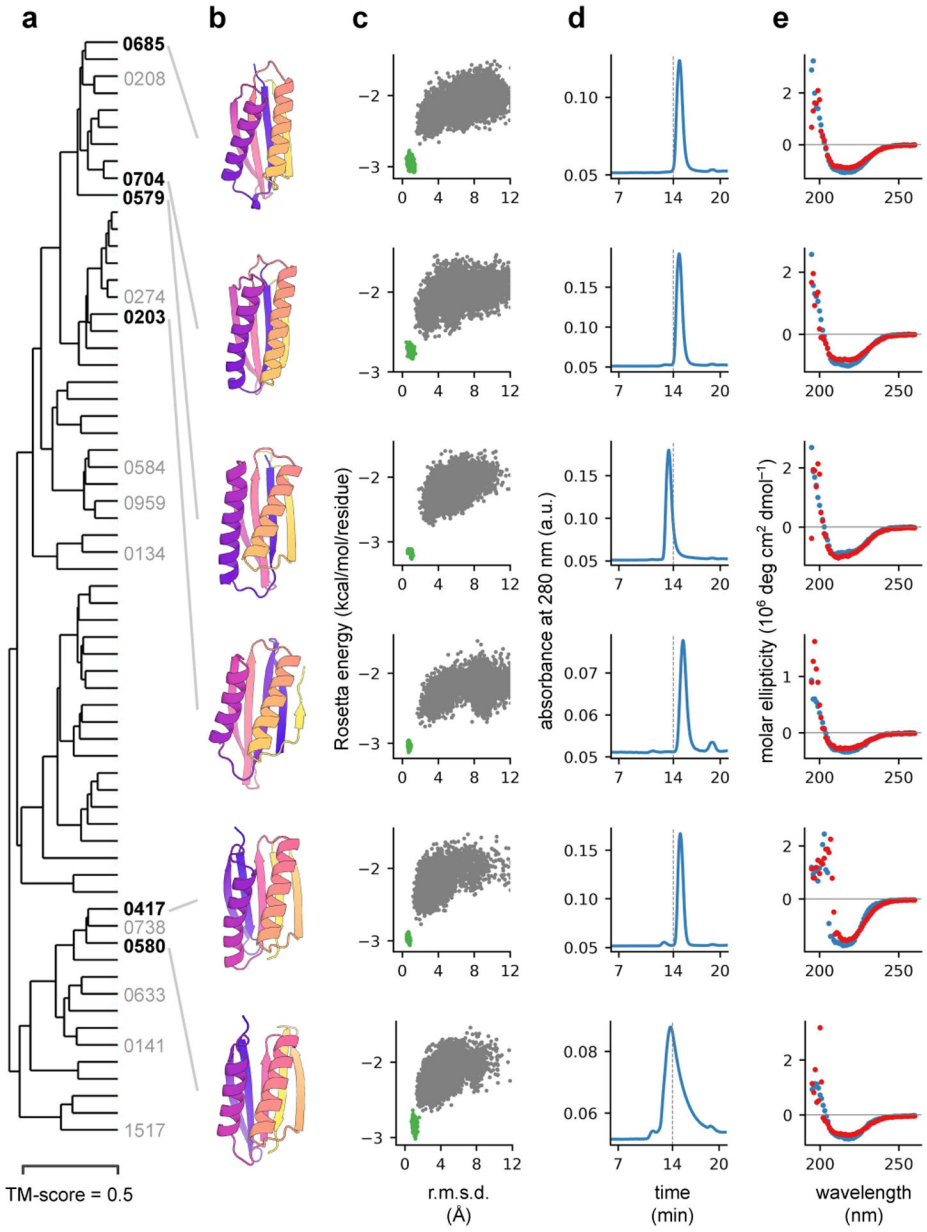


Figure 4.

Experimental characterization of mixed alpha and beta hallucinations. Columns are as in Fig. 3. Additional examples are shown in Extended Data Fig. 3.

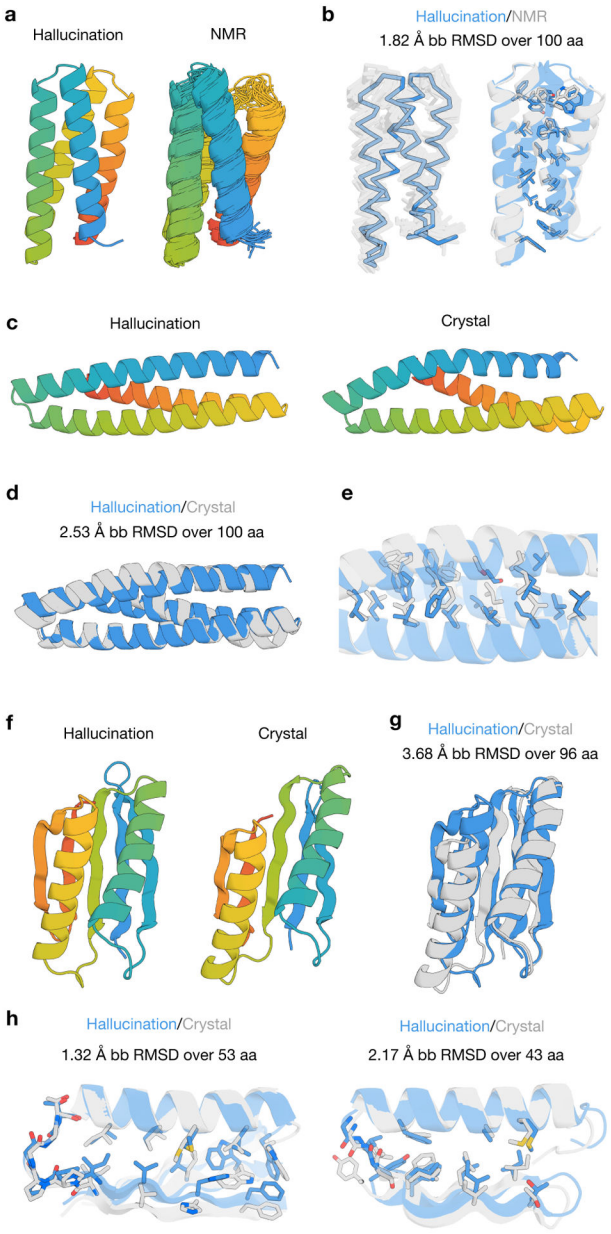


Figure 5.

Structural analysis of network hallucinated proteins. **a)** Hallucination model (left) and NMR ensemble structure of 0515 (right). **b**) Superposition of NMR ensemble (gray, transparent) and hallucinated model (blue, outlined) of 0515 and overlay of medoid NMR structure and model with side chains shown. **c)** Structures of the 0217 hallucination model (left) and crystal structure (right). **d)** Superposition of 0217 hallucination model (blue) and crystal structure (gray). **e)** Zoomed in overlay of 0217 crystal structure (gray) and hallucination model (blue) with side chains shown as sticks. **f)** Structures of 0738 model (left) and 0738_mod crystal structure (right). **g)** Superposition of 0738 hallucination model and 0738_mod crystal structure. **h)** Superposition of only the N-terminal section (left) and only the C-terminal section (right) of the 0738 hallucination model (blue) and 0738_mod crystal

structure (gray). Standalone structures are colored from N-terminus (blue) to C-terminus (red).