PROTEIN DESIGN

# Robust deep learning–based protein sequence design using ProteinMPNN

J. Dauparas[1,2], I. Anishchenko[1,2], N. Bennett[1,2,3], H. Bai[1,2,4], R. J. Ragotte[1,2], L. F. Milles[1,2], B. I. M. Wicky[1,2], A. Courbet[1,2,4], R. J. de Haas[5], N. Bethel[1,2,4], P. J. Y. Leung[1,2,3], T. F. Huddy[1,2], S. Pellock[1,2], D. Tischer[1,2], F. Chan[1,2], B. Koepnick[1,2], H. Nguyen[1,2], A. Kang[1,2], B. Sankaran[6], A. K. Bera[1,2], N. P. King[1,2], D. Baker[1,2,4]*

Although deep learning has revolutionized protein structure prediction, almost all experimentally characterized de novo protein designs have been generated using physically based approaches such as Rosetta. Here, we describe a deep learning–based protein sequence design method, ProteinMPNN, that has outstanding performance in both in silico and experimental tests. On native protein backbones, ProteinMPNN has a sequence recovery of 52.4% compared with 32.9% for Rosetta. The amino acid sequence at different positions can be coupled between single or multiple chains, enabling application to a wide range of current protein design challenges. We demonstrate the broad utility and high accuracy of ProteinMPNN using x-ray crystallography, cryo–electron microscopy, and functional studies by rescuing previously failed designs, which were made using Rosetta or AlphaFold, of protein monomers, cyclic homo-oligomers, tetrahedral nanoparticles, and target-binding proteins.

The protein sequence design problem is to find, given a protein backbone structure of interest, an amino acid sequence that will fold to this structure. Physically based approaches such as Rosetta treat sequence design as an energy optimization problem, searching for the combination of amino acid identities and conformations that has the lowest energy for a given input structure. Recently, deep-learning approaches have shown promise in rapidly generating candidate amino acid sequences given monomeric protein backbones without the need for compute-intensive explicit consideration of side chain rotameric states (*1*–*7*). However, the methods described thus far do not apply to the full range of current protein design challenges and have not been extensively validated experimentally.

We sought to develop a deep learning–based protein sequence design method that is broadly applicable to the design of monomers, cyclic oligomers, protein nanoparticles, and protein-protein interfaces. We began from a previously described message-passing neural network (MPNN) with three encoder and three decoder layers and 128 hidden dimensions that predicts protein sequences in an autoregressive manner from the N to C terminus using protein backbone features—distances between Cα-Cα atoms, relative Cα-Cα-Cα frame orientations and rotations, and backbone dihedral angles—as input (*1*). We first sought to improve performance of the model on recovering the amino acid sequences of native single-chain proteins given their backbone structures. A set of 19,700 high-resolution single-chain structures from the Protein Data Bank (PDB) were split into train, validation, and test sets (80/10/10) based on the CATH (*8*) protein classification database (see methods). We found that including distances between N, Cα, C, O, and a virtual Cβ placed based on the other backbone atoms as additional input features resulted in a sequence recovery increase from 41.2% (baseline model) to 49.0% (experiment 1) (see Table 1); interatomic distances evidently provide a better inductive bias to capture interactions between residues than dihedral angles or N-Cα-C frame orientations. We also observed performance improvements with edge updates in addition to the node updates in the backbone encoder neural network (experiment 2). Combining the additional input features and edge updates leads to a sequence recovery of 50.5% (experiment 3). To determine the range over which backbone geometry influences amino acid identity, we tested 16, 24, 32, 48, and 64 nearest–Cα neighbor neural networks (fig. S1A) and found that performance was saturated at 32 to 48 neighbors. Unlike the protein structure prediction problem, locally connected graph neural networks can accurately model the structure-to-sequence mapping problem because the optimality of an amino acid at a particular position is largely determined by the immediate protein environment.

To enable application to a broad range of single- and multichain design problems, we replaced the fixed N to C terminal decoding order with an order-agnostic autoregressive model in which the decoding order is randomly sampled from the set of all possible permutations (*9*). This also resulted in a modest improvement in sequence recovery (Table 1; experiment 4). Order-agnostic decoding enables design in cases where, for example, the middle of the protein sequence is fixed and the rest needs to be designed, as in protein binder design where the target sequence is known; decoding skips the fixed regions but includes them in the sequence context for the remaining positions (Fig. 1B). For multichain design problems (see discussion later in the text), to make the model equivariant to the order of the protein chains, we kept the per chain relative positional encoding capped at ±32 residues (*10*) and added a binary feature that indicates whether the interacting pair of residues are from the same or different chains.

We used the flexible decoding order to fix residue identities in sets of corresponding positions (the residues at these positions are decoded at the same time). For example, for a homodimer backbone with two chains A

**Table 1. Improvements in model performance on native protein sequence recovery.** Test accuracy (percentage of correct amino acids recovered) and test perplexity (exponentiated categorical cross-entropy loss per residue) for models trained on the native backbone coordinates (value to the left of the slash) and models trained with Gaussian noise (SD = 0.02 Å) added to the backbone coordinates (value to the right of the slash). Noise was only added during training, and all test evaluations are with no added noise. The final column shows sequence recovery on 5000 AlphaFold protein backbone models, with average predicted lDDT > 80.0, randomly chosen from UniRef50 sequences.

| Noise level when training: 0.00 Å/0.02 Å | Modification | Number of parameters in millions | PDB test accuracy (%) | PDB test perplexity | AlphaFold model accuracy (%) |
|---|---|---|---|---|---|
| Baseline model | None | 1.381 | 41.2/40.1 | 6.51/6.77 | 41.4/41.4 |
| Experiment 1 | Add N, Cα, C, Cβ, O distances | 1.430 | 49.0/46.1 | 5.03/5.54 | 45.7/47.4 |
| Experiment 2 | Update encoder edges | 1.629 | 43.1/42.0 | 6.12/6.37 | 43.3/43.0 |
| Experiment 3 | Combine 1 and 2 | 1.678 | 50.5/47.3 | 4.82/5.36 | 46.3/47.9 |
| Experiment 4 | Experiment 3 with random decoding | 1.678 | 50.8/47.9 | 4.74/5.25 | 46.9/48.5 |

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Molecular Engineering Graduate Program, University of Washington, Seattle, WA, USA. [4]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [5]Department of Physical Chemistry and Soft Matter, Wageningen University and Research, Wageningen, Netherlands. [6]Berkeley Center for Structural Biology, Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley Laboratory, Berkeley, CA, USA.
*Corresponding author. Email: dabaker@uw.edu

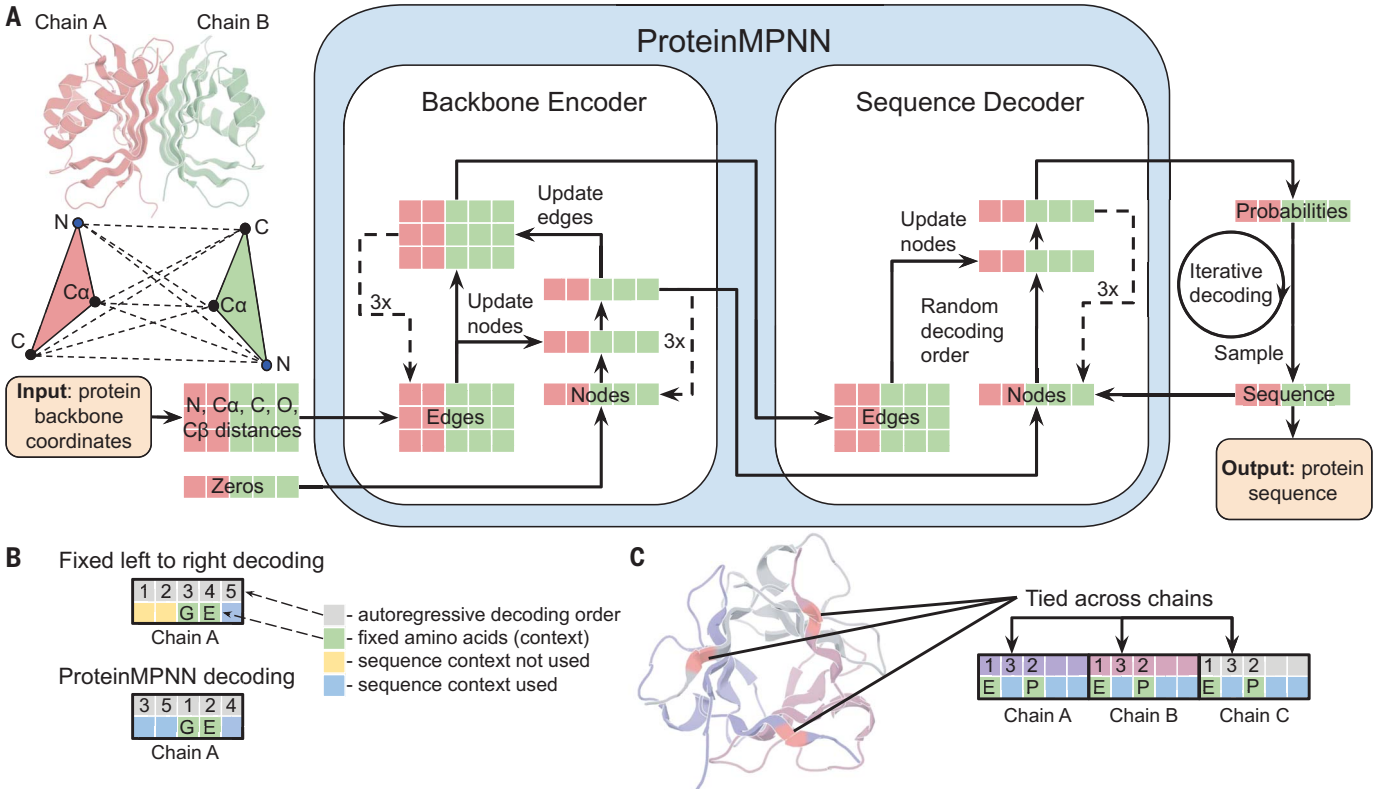


**Fig. 1. ProteinMPNN architecture.** (**A**) Distances between N, Cα, C, O, and virtual Cβ are encoded and processed using a message-passing neural network (Encoder) to obtain graph node and edge features. The encoded features, together with a partial sequence, are used to generate amino acids iteratively in a random decoding order. (**B**) A fixed left-to-right decoding cannot use sequence context (green) for preceding positions (yellow), whereas a model trained with random decoding orders can be used with an arbitrary decoding order during the inference. The decoding order can be chosen such that the fixed context is decoded first. (**C**) Residue positions within and between chains can be tied together, enabling symmetric, repeat protein, and multistate design. In this example, a homotrimer is designed with the coupling of positions in different chains. Predicted unnormalized probabilities for tied positions are averaged to get a single probability distribution from which amino acids are sampled.

and B with sequence $A_1$, $A_2$,... and $B_1$, $B_2$,..., the amino acids for chains A and B have to be the same for corresponding indices; we implement this by predicting unnormalized probabilities for $A_1$ and $B_1$ first and then combining these two predictions to construct a normalized probability distribution from which a joint amino acid is sampled (Fig. 1C). For pseudosymmetric sequence design, residues within or between chains can be similarly constrained; for example, for repeat protein design, the sequence in each repeat unit can be kept fixed. Multistate design of single sequences that encodes two or more desired states can be achieved by predicting unnormalized probabilities for each state and then averaging; more generally, a linear combination of predicted unnormalized probabilities with some positive and negative coefficients can be used to upweight or downweight specific backbone states to achieve explicit positive or negative sequence design. The architecture of this multichain and symmetry-aware (positionally coupled) model, which we call ProteinMPNN, is outlined schematically in Fig. 1A. We trained ProteinMPNN on protein assemblies in the PDB (as of 2 August 2021) determined by x-ray crystallography or cryo–electron microscopy (cryo-EM) to better than 3.5-Å resolution and with fewer than 10,000 residues (see methods).

For a test set of 402 monomer backbones, we redesigned sequences using Rosetta fixed backbone combinatorial sequence design [one round of the PackRotamersMover (*11*, *12*) with default options and the beta_nov16 score function] and ProteinMPNN. Although requiring only a small fraction of the compute time (1.2 versus 258.8 s on a single CPU for 100 residues), ProteinMPNN had a much higher overall native sequence recovery (52.4 versus 32.9%), with improvements across the full range of residue burial from protein core to surface (Fig. 2A). Differences between designed and native amino acid biases for the core, boundary, and surface regions for the two methods are shown in fig. S2.

We further evaluated ProteinMPNN on a test set of 690 monomers, 732 homomers (with fewer than 2000 residues), and 98 heteromers. The median sequence recoveries over all residues were 52% for monomers, 55% for homomers, and 51% for heteromers, and the median sequence recoveries over interface residues were 53% for homomers and 51% for heteromers (Fig. 2B). In all three cases, sequence recovery correlated closely with residue burial, ranging from 90 to 95% in the deep core to 35% on the surface (fig. S1B); the amount of local geometric context determines how well residues can be recovered at specific positions.

### Training with backbone noise improves model performance for protein design

Although protein sequence design approaches have often focused on maximizing sequence recovery for protein backbones from high-resolution crystal structures, this is not necessarily optimal for actual protein design applications. We found that training models on backbones to which Gaussian noise (SD = 0.02 Å) had been added improved sequence recovery on confident protein structure models generated by AlphaFold [average predicted local-distance difference test (IDDT) > 80.0] from UniRef50, whereas the sequence recovery on unperturbed PDB structures significantly decreased (Table 1); crystallographic refinement may impart

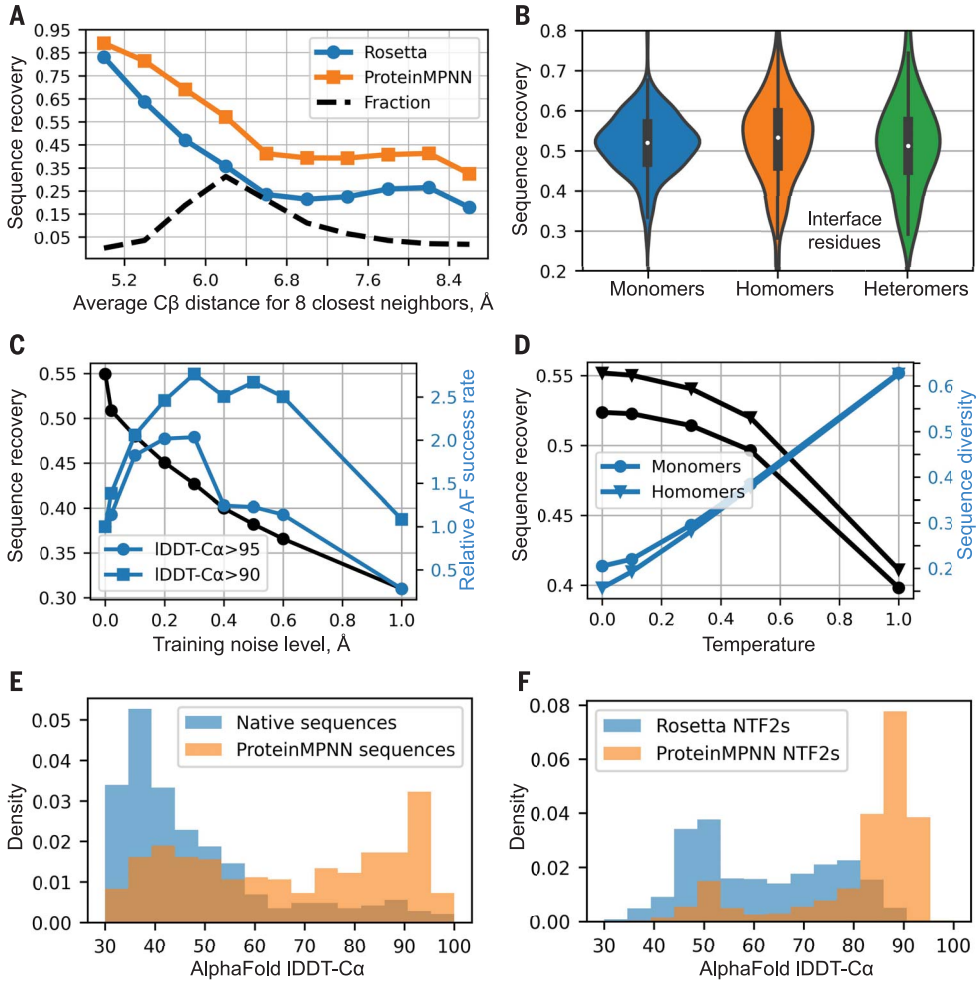


**Fig. 2. In silico evaluation of ProteinMPNN.** (**A**) ProteinMPNN has higher native sequence recovery than Rosetta. The average Cβ distance of the eight closest neighbors (*x* axis) reports on burial, with the most-buried positions on the left and the more-exposed positions on the right; ProteinMPNN outperforms Rosetta at all levels of burial. Average sequence recovery for ProteinMPNN was 52.4% compared with 32.9% for Rosetta. (**B**) ProteinMPNN has high sequence recovery for monomers and for both homo-oligomer and hetero-oligomer interfaces (Cβ-Cβ < 8 Å); violin plots are for 690 monomers, 732 homomers, and 98 heteromers. (**C**) Sequence recovery (black) and relative AlphaFold success rates (blue) as a function of training noise level. For higher accuracy predictions (circles), smaller amounts of noise are optimal (1.0 corresponds to a 1.8% success rate), whereas to maximize prediction success at a lower accuracy cutoff (squares), models trained with more noise are better (1.0 corresponds to a 6.7% success rate). (**D**) Sequence recovery and diversity as a function of sampling temperature. (**E**) Redesign of native protein backbones with ProteinMPNN considerably increases AphaFold prediction accuracy compared with the original native sequence using no multiple sequence information. Single sequences (designed or native) were input in both cases. Dark orange indicates overlap. (**F**) ProteinMPNN redesign of previous Rosetta-designed NTF2 fold proteins (3000 backbones in total) results in considerably improved AlphaFold single-sequence prediction accuracy. Dark orange indicates overlap.

some memory of amino acid identity in the backbone coordinates, which is captured by models trained on crystal structure backbones and reduced by the addition of noise. Robustness to small displacements in atomic coordinates is a desirable feature in real-world applications for which the protein backbone geometry is not known at atomic resolution.

AlphaFold (*10*) and RoseTTAFold (*13*) make very good structure predictions for native proteins, given multiple sequence alignments that can contain substantial coevolutionary and other information that reflects aspects of the three-dimensional (3D) structure, but generally produce less-accurate structure models when provided with only a single sequence. We reasoned that ProteinMPNN might generate sequences for native backbones that more strongly encode the structures than the original native sequences, because evolution, in most cases, does not optimize for stability. Indeed, we found that ProteinMPNN sequences generated for native backbones were predicted to fold to these structures much more confidently and accurately by AlphaFold than the original native sequences (Fig. 2E). ProteinMPNN also strengthened the sequence-to-structure mapping for designed backbones: Over a set of de novo–designed ligand binding pocket–containing scaffolds generated using Rosetta, only 2.7% of the original designed sequences were predicted to fold to the target structures, but after ProteinMPNN redesign, 54.1% were confidently predicted to fold to the target structures (Fig. 2F). This should substantially increase the utility of these scaffolds for the design of small-molecule binding and enzymatic functions.

We further found that the strength of the single sequence–to-structure mapping, as assessed by AlphaFold, was higher for models trained with additional backbone noise. As noted above, the average sequence recovery for crystallographically refined backbones decreases with increasing amounts of noise added during training (Fig. 2C) because these models blur out local details of the backbone geometry. However, sequences generated by noised ProteinMPNN models are more robustly decoded into 3D coordinates by AlphaFold, likely because noised models focus more on overall topological features as encoded by, for example, the overall polar-nonpolar sequence pattern than local structural details. For example, a model trained with 0.3-Å noise generated two to three times more sequences with AlphaFold predictions within lDDT-Cα (*14*) of 95.0 and 90.0 of the true structures than unnoised or slightly noised models (Fig. 2C; training with higher levels of noise increases success rates for less-stringent lDDT cutoffs). In protein design calculations, the models trained with larger amounts of noise have the advantage of generating sequences that more strongly map to the target structures by prediction methods (this increases the frequency at which designs pass prediction-based filters and may, correspondingly, also increase the frequency of folding to the desired target structure).

Because the sequence determinants of protein expression, solubility, and function are not perfectly understood, in most protein design applications, it is desirable to test multiple designed sequences experimentally. We found

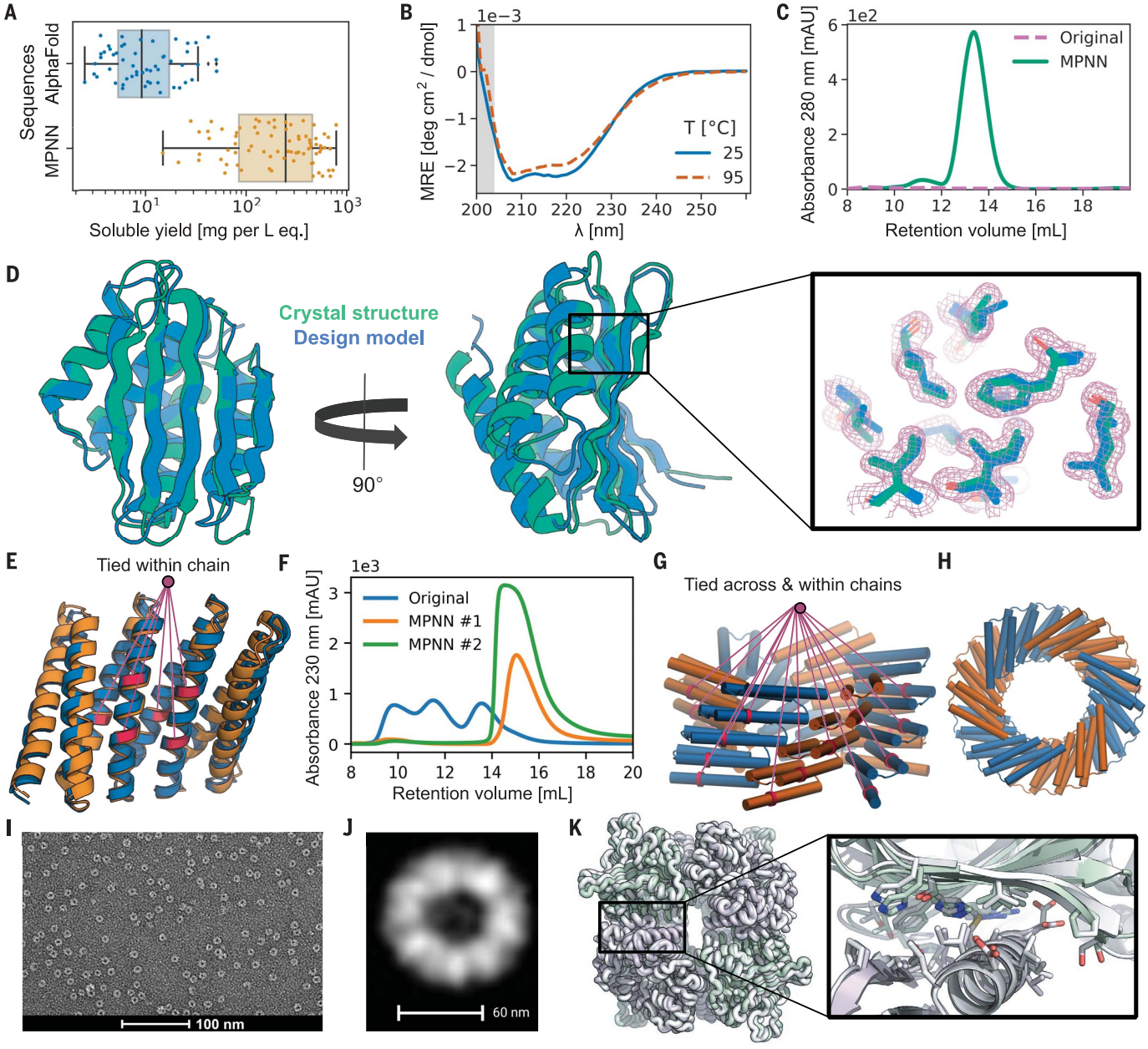


**Fig. 3. Structural characterization of ProteinMPNN designs.** (**A**) Comparison of soluble protein expression over a set of AlphaFold hallucinated monomers and homo-oligomers (blue) and the same set of backbones with sequences designed using ProteinMPNN (orange) ($N$ = 129). The total soluble protein yield after expression in *E. coli*, obtained from the integrated area under size exclusion traces of nickel-NTA–purified proteins, increases considerably from the barely soluble protein of the original sequences after ProteinMPNN rescue (median yields for 1 liter of culture equivalent are 9 and 247 mg, respectively). Boxes represent the quartiles of the soluble yield distribution and whiskers show the rest of it. (**B** to **D**) In-depth characterization of a monomer hallucination and corresponding ProteinMPNN rescue from the set in (A). Like almost all of the designs in (A), the sequence and structural similarities to the PDB of the design model are very low [expected value (E-value) = 2.8 against UniRef100 using HHblits; TM-score = 0.56 against PDB]. As shown in (B), the ProteinMPNN-rescued design has high thermostability, with a virtually unchanged circular dichroism profile at 95°C compared with 25°C. MRE, mean residue ellipticity. Shown in (C) is a SEC profile of the failed original design overlaid with the ProteinMPNN sequence design, which has a clear monodisperse peak at the expected retention volume. mAU, milli–absorbance units. As shown in (D), the crystal structure of the ProteinMPNN (PDB ID 8CYK) design is nearly identical to the design model (2.35-Å RMSD over 130 residues); see fig. S5 for additional information. The right panel shows model side chains in the electron density; crystal side chains are in green, and AlphaFold side chains are in blue. (**E** and **F**) ProteinMPNN rescue of the Rosetta design made from a perfectly repeating structural and sequence unit. Residues at corresponding positions in the repeat unit were tied during ProteinMPNN sequence inference. Shown in (E) are a backbone design model (orange) and MPNN redesigned sequence AlphaFold model (blue) with tied residues indicated by lines (~1.2-Å error over 232 residues). Shown in (F) is a SEC profile of the immobilized-metal affinity chromatograph (IMAC)–purified original Rosetta design and two ProteinMPNN redesigns. (**G** and **H**) Tying residues during ProteinMPNN sequence inference both within and between chains to enforce both repeat protein and cyclic symmetries. Shown in (G) is a side view of the design model. A set of tied residues are shown in red. Shown in (H) is a top-down view of the

design model. (**I**) Negative-stain electron micrograph of the purified design. (**J**) Class average of images from (I) closely match the top-down view in (H). (**K**) Rescue of the failed two-component Rosetta tetrahedral nanoparticle design T33-27 (*16*) by ProteinMPNN interface design. After ProteinMPNN rescue, the nanoparticle assembled readily with high yield, and the crystal structure (gray) is very nearly identical to the design model (green and purple) (backbone RMSD of 1.2 Å over two complete asymmetric units forming the ProteinMPNN-rescued interface).

that the diversity of sequences generated by ProteinMPNN could be considerably increased, with only a very small decrease in average sequence recovery, by carrying out inference at higher temperatures (Fig. 2D). We also found that a measure of sequence quality derived from ProteinMPNN, the averaged log probability of the sequence given the structure, correlated strongly with native sequence recovery over a range of temperatures (fig. S3A), enabling rapid ranking of sequences for selection for experimental characterization.

#### Experimental evaluation of ProteinMPNN

Although in silico native protein sequence recovery is a useful benchmark, the ultimate test of a protein design method is its ability to generate sequences that fold to the desired structure and have the desired function when tested experimentally. We evaluated ProteinMPNN on a representative set of protein monomer, assembly, and function design challenges. In each case, we attempted to rescue previous failed designs with sequences generated using Rosetta or AlphaFold; we kept the backbones of the original designs fixed but discarded the original sequences and generated new ones using ProteinMPNN. Synthetic genes encoding the designs were obtained, and the proteins were expressed in *Escherichia coli* and characterized biochemically and structurally.

We first tested the ability of ProteinMPNN to design amino acid sequences for protein backbones generated by deep network hallucination using AlphaFold. Starting from a random sequence, a Monte Carlo trajectory is carried out to optimize the extent to which AlphaFold predicts the sequence to fold to a well-defined structure (*15*). These calculations generated a wide range of protein sequences and backbones for both monomers and oligomers that differ considerably from those of native structures. In initial tests, the sequences generated by AlphaFold were encoded in synthetic genes, and we attempted to express 150 proteins in *E. coli*. However, the AlphaFold-generated sequences were mostly insoluble (median soluble yield of 9 mg per liter of culture equivalent; Fig. 3A). To determine whether ProteinMPNN could overcome this problem, we generated sequences for a subset of these backbones with ProteinMPNN; residue identities at symmetry-equivalent positions were tied by averaging unnormalized probabilities. The designed sequences were again encoded in synthetic genes, and the proteins were produced in *E. coli*. The success rate was far higher: Of the 96 designs that we attempted to express in *E. coli*, 73 were expressed solubly (median soluble yield of 247 mg per liter of culture equivalent; Fig. 3A) and 50 had the target monomeric or oligomeric state as assessed by size exclusion chromatography (SEC) (Fig. 3, A and C). Many of the proteins were highly thermostable, with secondary structure being maintained up to 95°C (Fig. 3B).

We solved the x-ray crystal structure of one of the ProteinMPNN monomer designs with a fold more complex [template modeling (TM)–score of 0.56 against PDB] than most de novo–designed proteins (Fig. 3D). The α-β protein structure contains five β strands and four α helices and is close to the design target backbone (2.35 Å over 130 residues), demonstrating that ProteinMPNN can accurately encode monomer backbone geometry in amino acid sequences. The accuracy was particularly high in the central core of the structure, with side chains predicted using AlphaFold from the ProteinMPNN sequence fitting nearly perfectly into the electron density (Fig. 3D). Crystal structures and cryo-EM structures of 10 cyclic homo-oligomers with 130 to 1800 amino acids were also very close to the design target backbones (*15*). Thus, ProteinMPNN can robustly and accurately design sequences for both monomers and cyclic oligomers.

We next took advantage of the flexible decoding order of ProteinMPNN to design sequences for proteins that contain internal repeats, tying the identities of proteins in equivalent positions. We focused on previously suboptimal Rosetta designs of repeat protein structures and found that many could be rescued by ProteinMPNN redesign; an example is shown in Fig. 3, E and F.

We next experimented with enforcing both cyclic and internal repeat symmetry by tying positions both within and between subunits, as illustrated in Fig. 3G. We experimentally characterized a set of such $C_5$ and $C_6$ cyclic oligomers with backbones generated using Rosetta and with sequences designed either with Rosetta or with ProteinMPNN. For the Rosetta-designed set, only 4 of 10 designs tested were soluble and none had the correct oligomeric state confirmed by SEC-multiangle light scattering (SEC-MALS). For the ProteinMPNN-designed set, 16 out of 18 were soluble and 5 had the correct oligomeric state. We characterized the structure of one of the designs that was large enough for resolution of structural features by negative-stain EM (Fig. 3I), and image averages were closely consistent with the design model (Fig. 3J).

We next evaluated the ability of ProteinMPNN to design sequences that assemble into target protein nanoparticle assemblies. We started with a set of previously described protein backbones for two-component tetrahedral designs that were generated using a compute- and effort-intensive procedure that involved Rosetta sequence design followed by more than a week of manual intervention to decrease surface hydrophobicity and improve interface packing (*16*). We used ProteinMPNN to design 76 sequences spanning 27 of these tetrahedral nanoparticle backbones, tying identities at equivalent positions in the 12 copies of each subunit in the assemblies, and tested these sequences without further intervention. Upon expression in *E. coli* and purification by SEC, 13 designs formed assemblies with the expected molecular weight (~1 MDa) (fig. S4), including several new tetrahedral assemblies that had failed using Rosetta. We solved the crystal structure of one of these and found that it was very close to the design model [1.2-Å Cα root mean square deviation (RMSD) over two subunits; Fig. 3K]. Thus, ProteinMPNN can robustly design sequences that assemble into designed nanoparticle structures, which have proven useful for structure-based vaccine design (*17*–*19*). Sequence generation with ProteinMPNN is fully automated and requires only about 1 s per backbone, vastly streamlining the design process compared with the earlier Rosetta-based procedure.

As a final test, we evaluated the ability of ProteinMPNN to rescue previously failed designs of new protein functions using Rosetta. We chose as a challenging example the design of proteins that scaffold polyproline II helix motifs recognized by SH3 domains, where portions of the protein scaffold outside of the core SH3-binding motif make additional interactions with the target (the goal is to generate protein reagents with high affinity and specificity for individual SH3 family members). Backbones that scaffold a proline-rich SH3-binding motif (PPPRPPK; where P is proline, R is arginine, and K is lysine) recognized by the Grb2 SH3 domain were generated using Rosetta remodel (see legend of Fig. 4; the SH3-binding motif is colored in green in Fig. 4A), but sequences designed for these backbones and expressed in *E. coli* did not fold to structures that bind Grb2 (Fig. 4B; the design problem is challenging because very few native proteins have proline-rich secondary structure elements that closely interact with the core of the protein). To test whether ProteinMPNN could overcome this problem,

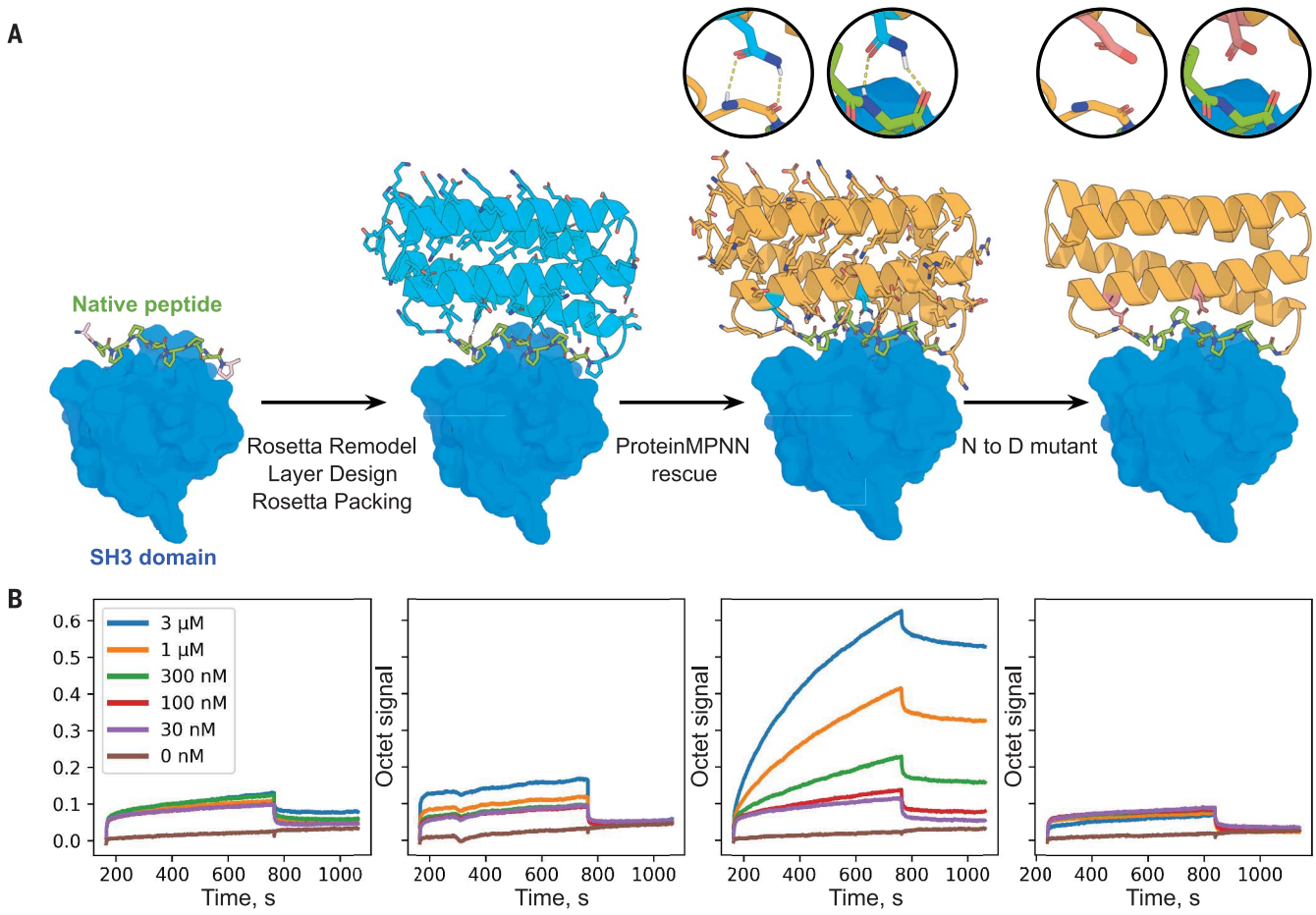


**Fig. 4. Design of protein function with ProteinMPNN.** (**A**) Design scheme. The first panel shows the structure (PDB ID 2W0Z) of a fragment of Gab2 peptide bound to the human Grb2 C-term SH3 domain (core SH3-binding motif PPPRPPK is in green; the target is rendered with surface and colored blue). In the second panel, helical bundle scaffolds were docked to the exposed face of the peptide using RIFDOCK (*20*), and Rosetta remodel was used to build loops connecting the peptide to the scaffolds. Rosetta sequence design with layer design task operations was used to optimize the sequence of the fusion (cyan) for stability, rigidity of the peptide-helical bundle interface, and binding affinity for the Grb2 SH3 domain. The third panel shows the ProteinMPNN redesign (orange) of the designed binder sequence; hydrogen bonds involving asparagine side chains between the peptide and base scaffold are shown in green and in the inset. In the fourth panel, mutation of the two asparagines (N) to aspartates (D) disrupts the scaffolding of the target peptide. (**B**) Experimental characterization of binding using biolayer interferometry. Biotinylated C-terminal SH3 domain from human Grb2 was loaded onto Streptavidin (SA) Biosensors, which were then immersed in solutions containing varying concentrations of SH3-binding peptide AIAPPPRPPKPSQ (first panel; A, alanine; I, isoleucine; S, serine; Q, glutamine) or of the designs (second to fourth panels) and then transferred to buffer lacking added protein for dissociation measurements. The ProteinMPNN design (third panel) has much greater binding signal than the original Rosetta design (second panel); this is greatly reduced by the asparagine-to-aspartate mutations (fourth panel).

we generated sequences for the same backbones while keeping the core SH3-binding motif sequence (PPPRPPK) fixed and expressed the proteins in *E. coli*. Biolayer interferometry experiments showed strong binding to the Grb2 SH3 domain (Fig. 4B), with considerably higher signal than the free proline-rich peptide; point mutations predicted to disrupt the design completely eliminated the binding signal. Thus, ProteinMPNN can generate sequences for challenging protein design problems even when traditional Rosetta design fails.

## Conclusion

ProteinMPNN solves sequence design problems in a fraction of the time required for physically based approaches such as Rosetta, which carry out large-scale side chain packing calculations; achieves much higher protein sequence recovery on native backbones (52.4 versus 32.9%); and rescues previously failed designs made using Rosetta or AlphaFold for protein monomers, assemblies, and protein-protein interfaces. Machine-learning sequence design approaches have been developed previously (*1*–*7*), including the message-passing method on which ProteinMPNN is based, but have focused on the monomer design problem, have achieved lower native sequence recoveries, and, with the exception of a triosephosphate isomerase (TIM) barrel design study (*6*), have not been extensively validated using crystallography and cryo-EM. Whereas structure prediction methods can be evaluated purely in silico, this is not the case for protein design methods: In silico metrics such as native sequence recovery are very sensitive to crystallographic resolution (fig. S3, B and C) and may not correlate with proper folding (even a single residue substitution, while causing little change in overall sequence recovery, can block folding); in the same way that language translation accuracy must ultimately be evaluated by human users, the ultimate test of sequence design methods is experimental characterization.

Unlike Rosetta and other physically based methods, ProteinMPNN requires no expert

customization for specific design challenges, and it should thus make protein design more broadly accessible. This robustness reflects fundamental differences in how the sequence design problem is framed. In traditional physically based approaches, sequence design maps to the problem of identifying an amino acid sequence whose lowest-energy state is the desired structure. This is, however, computationally intractable because it requires computing energies over all possible structures, including unwanted oligomeric and aggregated states; instead, as a proxy, Rosetta and other approaches carry out a search for the lowest-energy sequence for a given backbone structure, and structure prediction calculations are required in a second step to confirm that there are no other structures in which the sequence has still lower energy. Because of the lack of concordance between the design objective and what is being explicitly optimized, considerable customization can be required to generate sequences that fold; for example, in Rosetta design calculations, hydrophobic amino acids are often restricted on the protein surface because they can stabilize undesired multimeric states and, at the boundary region between the protein surface and core, there can be considerable ambiguity about the extent to which such restrictions should be applied. Although deep-learning methods lack the physical transparency of methods like Rosetta, they are trained directly to find the most probable amino acid for a protein backbone given all the examples in the PDB, and hence such ambiguities do not arise, making sequence design more robust and less dependent on the judgment of a human expert.

The high rate of experimental design success of ProteinMPNN, together with the compute efficiency, applicability to almost any protein sequence design problem, and lack of requirement for customization, should make it very broadly useful for protein design. ProteinMPNN-generated sequences also have a much higher propensity to crystallize, greatly facilitating structure determination of designed proteins (*15*). The observation that ProteinMPNN-generated sequences are predicted to fold to native protein backbones more confidently and accurately than the original native sequences (using single-sequence information in both cases) suggests that ProteinMPNN may also be widely useful in improving expression and stability of recombinantly expressed native proteins (with residues required for function kept fixed).

## REFERENCES AND NOTES


1. J. Ingraham, V. K. Garg, R. Barzilay, T. Jaakkola, in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, H. Wallach *et al*., Eds. (Neural Information Processing Systems Foundation, 2019), pp. 15741–15752.
2. Y. Zhang *et al*., *Proteins* **88**, 819–829 (2020).
3. Y. Qi, J. Z. H. Zhang, *J. Chem. Inf. Model.* **60**, 1245–1252 (2020).
4. B. Jing, S. Eismann, P. Suriana, R. J. L. Townshend, R. Dror, paper presented at the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
5. A. Strokach, D. Becerra, C. Corbi-Verge, A. Perez-Riba, P. M. Kim, *Cell Syst.* **11**, 402–411.e4 (2020).
6. N. Anand *et al*., *Nat. Commun.* **13**, 746 (2022).
7. C. Hsu *et al*., bioRxiv 2022.04.10.487779 [Preprint] (2022).
8. C. A. Orengo *et al*., *Structure* **5**, 1093–1108 (1997).
9. B. Uria, I. Murray, H. Larochelle, in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing, T. Jebara, Eds.(JMLR, 2014), pp. 467–475.
10. J. Jumper *et al*., *Nature* **596**, 583–589 (2021).
11. A. Leaver-Fay *et al*., *Methods Enzymol.* **523**, 109–143 (2013).
12. J. K. Leman *et al*., *Nat. Methods* **17**, 665–680 (2020).
13. M. Baek *et al*., *Science* **373**, 871–876 (2021).
14. V. Mariani, M. Biasini, A. Barbato, T. Schwede, *Bioinformatics* **29**, 2722–2728 (2013).
15. B. I. M. Wicky *et al*., *Science* **378**, 56–61 (2022).
16. N. P. King *et al*., *Nature* **510**, 103–108 (2014).
17. S. Boyoglu-Barnum *et al*., *Nature* **592**, 623–628 (2021).
18. A. C. Walls *et al*., *Cell* **183**, 1367–1382.e17 (2020).
19. J. Marcandalli *et al*., *Cell* **176**, 1420–1431.e17 (2019).
20. L. Cao *et al*., *Nature* **605**, 551–560 (2022).
21. J. Dauparas, S. O, S. Duerr, dauparas/ProteinMPNN: ProteinMPNN (v1.0.0). Zenodo (2022).


## ACKNOWLEDGMENTS


We thank S. Ovchinnikov, C. Norn, D. Juergens, J. Wang, F. DiMaio, R. Kibler, M. Baek, S. Mansoor, L. Goldschmidt, and L. Stewart for helpful discussions. We also thank the Meta AI protein team for sharing AlphaFold models generated for UniRef50 sequences. The Berkeley Center for Structural Biology is supported in part by the National Institutes of Health (NIH), National Institute of General Medical Sciences. Crystallographic data were collected at The Advanced Light Source (ALS), which is supported by the director, Office of Science, Office of 20 Basic Energy Sciences, and US Department of Energy under contract number DE-AC02- 05CH11231. **Funding:** This work was supported with funds provided by a gift from Microsoft (J.D., D.T., and D.B.), the Audacious Project at the Institute for Protein Design (A.K.B., A.K., B.K., F.C., T.F.H., R.J.d.H., N.P.K., and D.B.), a grant from the National Science Foundation (NSF) (DBI 1937533 to D.B. and I.A.), an EMBO long-term fellowship ALTF 139-2018 (B.I.M.W.), the Open Philanthropy Project Improving Protein Design Fund (R.J.R. and D.B.), Howard Hughes Medical Institute Hanna Gray fellowship grant GT11817 (N.Bet.), The Donald and Jo Anne Petersen Endowment for Accelerating Advancements in Alzheimer's Disease Research (N.Ben.), a Washington Research Foundation Fellowship (S.P.), an Alfred P. Sloan Foundation Matter-to-Life Program Grant (G-2021-16899; A.C. and D.B.), a Human Frontier Science Program Cross Disciplinary Fellowship (LT000395/2020-C; L.F.M.), an EMBO Non-Stipendiary Fellowship (ALTF 1047-2019; L.F.M.), an NSF Graduate Research Fellowship (DGE-2140004; P.J.Y.L), the Howard Hughes Medical Institute (A.C., H.B., and D.B.), and NIH, National Institute of General Medical Sciences, grant P30 GM124169-01 (B.S.). We thank Microsoft and Amazon Web Services (AWS) for generous gifts of cloud computing credits. **Author contributions:** Conceptualization: J.D., L.F.M., B.I.M.W., A.C., R.J.d.H., H.B., N.Ben.; Methodology: J.D., I.A., P.J.Y.L.; Software: J.D., T.F.H., D.T., B.K., F.C.; Validation: J.D., N.Ben., H.B., A.K.B., B.S., A.K., H.N., S.P., P.J.Y.L, N.Bet., R.J.d.H., L.F.M., B.I.M.W., A.C., R.J.R.; Formal analysis: J.D., L.F.M., B.I.M.W., R.J.R., N.Ben.; Resources: J.D., D.B.; Data curation: I.A., J.D., H.B., ; Writing – original draft: J.D., D.B.; Writing – review and editing: J.D., D.B.; Visualization: J.D., R.J.R., R.J.d.H., H.B., L.F.M., B.I.M.W., P.J.Y.L., H.B.; Supervision: D.B., N.P.K.; Project administration: J.D.; Funding acquisition: J.D., D.B. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data are available in the main text or as supplementary materials. ProteinMPNN code (*21*) is available at https://github.com/dauparas/ProteinMPNN.



## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.add2187
Materials and Methods
Figs. S1 to S12
Table S1
References (*22*–*35*)

View/request a protocol for this paper from *Bio-protocol*.


Submitted 27 May 2022; accepted 7 September 2022
Published online 15 September 2022
10.1126/science.add2187

# Robust deep learning–based protein sequence design using ProteinMPNN

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker

*Science* **378** (6615), . DOI: 10.1126/science.add2187

**Deep learning takes on protein design**

Deep learning approaches such as Alphafold and Rosettafold have made reliable protein structure prediction broadly accessible. For the inverse problem, finding a sequence that folds to a desired structure, most approaches remain based on energy optimization. In two papers, a range of protein design problems were addressed through deep learning methods. Dauparas *et al*. built on recent deep learning protein design approaches to develop a method called ProteinMPNN. They validated designs experimentally and showed that ProteinMPNN can rescue previously failed designs made using Rosetta or Alphafold. Wicky *et al*. started from a random sequence and used Monte Carlo sequence search coupled with structure prediction by Alphafold to design cyclic homo-oligomers. Although the designs were generated to achieve stable expression, the sequences had to be regenerated using ProteinMPNN. This approach allowed for the design of a range of experimentally validated cyclic oligomers and paves the way for the design of increasingly complex assemblies. —VV