Cite as: J. Dauparas *et al.*, *Science* 10.1126/science.add2187 (2022).

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]*

[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**

While 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, with outstanding performance in both in silico and experimental tests. On native protein backbones, ProteinMPNN has a sequence recovery of 52.4%, compared to 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, cryoEM and functional studies by rescuing previously failed designs, 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 like Rosetta treat sequence design as an energy optimization problem, searching for the combination of amino acid identities and conformations that have 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 need for compute intensive explicit consideration of sidechain 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 broadly applicable to design of monomers, cyclic oligomers, protein nanoparticles, and protein-protein interfaces. We began from a previously described message passing neural network (MPNN) with 3 encoder and 3 decoder layers and 128 hidden dimensions which predicts protein sequences in an autoregressive manner from 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 PDB were split into train, validation and test sets (80/10/10) based on the CATH (*8*) protein classification (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 below; interatomic distances evidently provide a better inductive bias to capture interactions between residues than dihedral angles or N-Cα-C frame orientations. We next introduced edge updates in addition to the node updates in the backbone encoder neural network (experiment 2). Combining 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 32-48 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 multi-chain 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 multi-chain design problems (see below), 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 indicating if 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 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 combine 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. Multi-state 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 Aug 02, 2021) determined by X-ray crystallography or cryoEM to better than 3.5Å resolution and with less 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 s versus 258.8 s on a single CPU for 100 residues), ProteinMPNN had a much higher overall native sequence recovery (52.4% vs 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 less 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 over interface residues, 53% for homomers and 51% for heteromers (Fig. 2B). In all three cases, sequence recovery correlated closely with residue burial ranging from 90-95% 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. For homomers, we found best results with averaging unnormalized probabilities (rather than normalized probabilities) between symmetry related positions (fig. S1C); because of the non-local context sequence recovery is no longer a monotonic function of the average Cβ neighbor distance (fig. S1B).

Training with backbone noise improves model performance for protein design

While 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 (std=0.02Å) had been added improved sequence recovery on confident protein structure models generated by AlphaFold (average pLDDT>80.0) from UniRef50, while the sequence recovery on unperturbed PDB structures significantly decreased (Table 1); crystallographic refinement may impart 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 where the protein backbone geometry is not known at atomic resolution.

AlphaFold (*10*) and RoseTTAfold (*13*) produce remarkably good structure predictions for native proteins given multiple sequence alignments which can contain substantial co-evolutionary and other information reflecting aspects of the 3D structure, but generally produce much poorer structures when provided only with a single sequence. We reasoned that ProteinMPNN might generate single sequences for native backbones more strongly encoding the structures than the original native sequence, as 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 design target structures, but following ProteinMPNN redesign 54.1% were confidently

predicted to fold to close to the target structures (Fig. 2F). This should substantially increase the utility of these scaffolds for design of small molecule binding and enzymatic functions.

We found further 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) as 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 2-3 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 IDDT cutoffs). In protein design calculations, the models trained with larger amounts of noise have the advantage of generating sequences which more strongly map to the target structures by prediction methods (this increases the frequency of designs passing 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 that the diversity of sequences generated by MPNN 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 the 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

While in silico native protein sequence recovery is a useful benchmark, the ultimate test of a protein design method is its ability to generate sequences which fold to the desired structure and have the desired function when tested experimentally. We evaluated ProteinMPNN on a representative set of design challenges encompassing protein monomer design, protein nanocage design, and protein function design. 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 expressed in *E. 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 (AF). Starting from a random sequence, a Monte Carlo trajectory is carried out optimizing the extent to which AF 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 AF were encoded in synthetic genes, and we attempted to express 150 proteins in *E. coli*. However, the AF generated sequences were mostly insoluble (median soluble yield: 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 as described above. The designed sequences were again encoded in synthetic genes and the proteins produced in *E. coli*. The success rate was far higher: of 96 designs we attempted to express in *E. coli*, 73 were expressed solubly (median soluble yield: 247 mg per liter of culture equivalent; Fig. 3A) and 50 had the target monomeric or oligomeric state as assessed by SEC (Fig. 3, A and C). Many of the proteins were highly thermo-stable, 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 (TM-score=0.56 against PDB) than most de novo designed proteins (Fig. 3D). The alpha-beta protein structure contains 5 beta strands and 4 alpha 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 sidechains predicted using AlphaFold from the ProteinMPNN sequence fitting nearly perfectly into the electron density (Fig. 3D). Crystal structures and cryo-EM structures of ten cyclic homo-oligomers with 130-1800 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 containing 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 C_5/C_6 cyclic oligomers built with Rosetta based on sequences designed with Rosetta, and a second set with sequences designed using ProteinMPNN. For the Rosetta designed set, 40% (out of total 10) were soluble and none had the correct oligomeric state confirmed by SEC-MALS. For the ProteinMPNN designed set, 88% (out of total 18) were soluble and 27.7% 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 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 MW (~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 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 to 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 scaffolding 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 scaffolding a proline rich SH3-binding motif (PPPRPPK) 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 as 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, 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 RosettaDesign 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 sidechain packing calculations, achieves much higher protein sequence recovery on native backbones (52.4% vs 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, achieved lower native sequence recoveries, and with the exception of a TIM barrel design study (*6*) have not been extensively validated using crystallography and cryoEM. 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 as 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 which fold; for example in Rosetta design calculations hydrophobic amino acids are often restricted on the protein surface as 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. While 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 judgement 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, "Generative models for graph-based protein design" in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett, Eds. (Neural Information Processing Systems Foundation, 2019), pp. 15741–15752.
2. Y. Zhang, Y. Chen, C. Wang, C. C. Lo, X. Liu, W. Wu, J. Zhang, ProDCoNN: Protein design using a convolutional neural network. *Proteins* **88**, 819–829 (2020). doi:10.1002/prot.25868 Medline
3. Y. Qi, J. Z. H. Zhang, DenseCPD: Improving the accuracy of neural-network-based computational protein sequence design with DenseNet. *J. Chem. Inf. Model.* **60**, 1245–1252 (2020). doi:10.1021/acs.jcim.0c00043 Medline
4. B. Jing, S. Eismann, P. Suriana, R. J. L. Townshend, R. Dror, "Learning from protein structure with geometric vector perceptrons," 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, Fast and flexible protein design using deep graph neural networks. *Cell Syst.* **11**, 402–411.e4 (2020). doi:10.1016/j.cels.2020.08.016 Medline
6. N. Anand, R. Eguchi, I. I. Mathews, C. P. Perez, A. Derry, R. B. Altman, P. S. Huang, Protein sequence design with a learned potential. *Nat. Commun.* **13**, 746 (2022). doi:10.1038/s41467-022-28313-9 Medline
7. C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, A. Rives, Learning inverse folding from millions of predicted structures. bioRxiv 2022.04.10.487779 [Preprint] (2022); https://doi.org/10.1101/2022.04.10.487779.
8. C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, J. M. Thornton, CATH—A hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108 (1997). doi:10.1016/S0969-2126(97)00260-8 Medline
9. B. Uria, I. Murray, H. Larochelle, "A deep and tractable density estimator" in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing, T. Jebara, Eds. (JMLR, 2014), pp. 467–475.
10. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). doi:10.1038/s41586-021-03819-2 Medline
11. A. Leaver-Fay, M. J. O'Meara, M. Tyka, R. Jacak, Y. Song, E. H. Kellogg, J. Thompson, I. W. Davis, R. A. Pache, S. Lyskov, J. J. Gray, T. Kortemme, J. S. Richardson, J. J. Havranek, J. Snoeyink, D. Baker, B. Kuhlman, Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol.* **523**, 109–143 (2013). doi:10.1016/B978-0-12-394292-0.00006-0 Medline
12. J. K. Leman, B. D. Weitzner, S. M. Lewis, J. Adolf-Bryfogle, N. Alam, R. F. Alford, M. Aprahamian, D. Baker, K. A. Barlow, P. Barth, B. Basanta, B. J. Bender, K. Blacklock, J. Bonet, S. E. Boyken, P. Bradley, C. Bystroff, P. Conway, S. Cooper, B. E. Correia, B. Coventry, R. Das, R. M. De Jong, F. DiMaio, L. Dsilva, R. Dunbrack, A. S. Ford, B. Frenz, D. Y. Fu, C. Geniesse, L. Goldschmidt, R. Gowthaman, J. J. Gray, D. Gront, S. Guffy, S. Horowitz, P.-S. Huang, T. Huber, T. M. Jacobs, J. R. Jeliazkov, D. K. Johnson, K. Kappel, J. Karanicolas, H. Khakzad, K. R. Khar, S. D. Khare, F. Khatib, A. Khramushin, I. C. King, R. Kleffner, B. Koepnick, T. Kortemme, G. Kuenze, B. Kuhlman, D. Kuroda, J. W. Labonte, J. K. Lai, G. Lapidoth, A. Leaver-Fay, S. Lindert, T. Linsky, N. London, J. H. Lubin, S. Lyskov, J. Maguire, L. Malmström, E. Marcos, O. Marcu, N. A. Marze, J. Meiler, R. Moretti, V. K. Mulligan, S. Nerli, C. Norn, S. Ó'Conchúir, N. Ollikainen, S. Ovchinnikov, M. S. Pacella, X. Pan, H. Park, R. E. Pavlovicz, M. Pethe, B. G. Pierce, K. B. Pilla, B. Raveh, P. D. Renfrew, S. S. R. Burman, A. Rubenstein, M. F. Sauer, A. Scheck, W. Schief, O. Schueler-Furman, Y. Sedan, A. M. Sevy, N. G. Sgourakis, L. Shi, J. B. Siegel, D.-A. Silva, S. Smith, Y. Song, A. Stein, M. Szegedy, F. D. Teets, S. B. Thyme, R. Y.-R. Wang, A. Watkins, L. Zimmerman, R. Bonneau, Macromolecular modeling and design in Rosetta: Recent methods and frameworks. *Nat. Methods* **17**, 665–680 (2020). doi:10.1038/s41592-020-0848-2 Medline
13. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021). doi:10.1126/science.abj8754 Medline
14. V. Mariani, M. Biasini, A. Barbato, T. Schwede, lDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013). doi:10.1093/bioinformatics/btt473 Medline
15. B. I. M. Wicky, L. F. Milles, A. Courbet, R. J. Ragotte, J. Dauparas, E. Kinfu, S. Tipps, R. D. Kibler, M. Baek, F. DiMaio, X. Li, L. Carter, A. Kang, H. Nguyen, A. K. Bera, D. Baker, Hallucinating symmetric protein assemblies. *Science* 10.1126/science.add1964 (2022).
16. N. P. King, J. B. Bale, W. Sheffler, D. E. McNamara, S. Gonen, T. Gonen, T. O. Yeates, D. Baker, Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510**, 103–108 (2014). doi:10.1038/nature13404 Medline
17. S. Boyoglu-Barnum, D. Ellis, R. A. Gillespie, G. B. Hutchinson, Y.-J. Park, S. M. Moin, O. J. Acton, R. Ravichandran, M. Murphy, D. Pettie, N. Matheson, L. Carter, A. Creanga, M. J. Watson, S. Kephart, S. Ataca, J. R. Vaile, G. Ueda, M. C. Crank, L. Stewart, K. K. Lee, M. Guttman, D. Baker, J. R. Mascola, D. Veesler, B. S. Graham, N. P. King, M. Kanekiyo, Quadrivalent influenza nanoparticle vaccines induce broad protection. *Nature* **592**, 623–628 (2021). doi:10.1038/s41586-021-03365-

x Medline
18. A. C. Walls, B. Fiala, A. Schäfer, S. Wrenn, M. N. Pham, M. Murphy, L. V. Tse, L. Shehata, M. A. O'Connor, C. Chen, M. J. Navarro, M. C. Miranda, D. Pettie, R. Ravichandran, J. C. Kraft, C. Ogohara, A. Palser, S. Chalk, E.-C. Lee, K. Guerriero, E. Kepl, C. M. Chow, C. Sydeman, E. A. Hodge, B. Brown, J. T. Fuller, K. H. Dinnon III, L. E. Gralinski, S. R. Leist, K. L. Gully, T. B. Lewis, M. Guttman, H. Y. Chu, K. K. Lee, D. H. Fuller, R. S. Baric, P. Kellam, L. Carter, M. Pepper, T. P. Sheahan, D. Veesler, N. P. King, Elicitation of potent neutralizing antibody responses by designed protein nanoparticle vaccines for SARS-CoV-2. *Cell* **183**, 1367–1382.e17 (2020). doi:10.1016/j.cell.2020.10.043 Medline
19. J. Marcandalli, B. Fiala, S. Ols, M. Perotti, W. de van der Schueren, J. Snijder, E. Hodge, M. Benhaim, R. Ravichandran, L. Carter, W. Sheffler, L. Brunner, M. Lawrenz, P. Dubois, A. Lanzavecchia, F. Sallusto, K. K. Lee, D. Veesler, C. E. Correnti, L. J. Stewart, D. Baker, K. Loré, L. Perez, N. P. King, Induction of potent neutralizing antibody responses by a designed protein nanoparticle vaccine for respiratory syncytial virus. *Cell* **176**, 1420–1431.e17 (2019). doi:10.1016/j.cell.2019.01.046 Medline
20. L. Cao, B. Coventry, I. Goreshnik, B. Huang, W. Sheffler, J. S. Park, K. M. Jude, I. Marković, R. U. Kadam, K. H. G. Verschueren, K. Verstraete, S. T. R. Walsh, N. Bennett, A. Phal, A. Yang, L. Kozodoy, M. DeWitt, L. Picton, L. Miller, E.-M. Strauch, N. D. DeBouver, A. Pires, A. K. Bera, S. Halabiya, B. Hammerson, W. Yang, S. Bernard, L. Stewart, I. A. Wilson, H. Ruohola-Baker, J. Schlessinger, S. Lee, S. N. Savvides, K. C. Garcia, D. Baker, Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022). doi:10.1038/s41586-022-04654-9 Medline
21. J. Dauparas, S. O, S. Duerr, dauparas/ProteinMPNN: ProteinMPNN (v1.0.0). Zenodo (2022); https://doi.org/10.5281/zenodo.6941302.
22. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you need" in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Eds. (Neural Information Processing Systems Foundation, 2017), pp. 5999–6009.
23. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
24. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the inception architecture for computer vision" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 2818–2826.
25. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017). doi:10.1038/nbt.3988 Medline
26. Y. Zhang, J. Skolnick, TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005). doi:10.1093/nar/gki524 Medline
27. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. Devito, M. Raison, A. Tejani, S. Chilamkurth, "Pytorch: An imperative style, high-performance deep learning library" in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett, Eds. (Neural Information Processing Systems Foundation, 2019), pp. 7994–8005.
28. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, "Neural message passing for quantum chemistry" in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup, Y. W. The, Eds. (JMLR, 2017), pp. 1263–1272.
29. B. Dang, M. Mravic, H. Hu, N. Schmidt, B. Mensa, W. F. DeGrado, SNAC-tag for sequence-specific chemical protein cleavage. *Nat. Methods* **16**, 319–322 (2019). doi:10.1038/s41592-019-0357-3 Medline
30. W. Kabsch, XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010). doi:10.1107/S0907444909047337 Medline
31. M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011). doi:10.1107/S0907444910045749 Medline
32. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007). doi:10.1107/S0021889807021206 Medline
33. P. Emsley, K. Cowtan, Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004). doi:10.1107/S0907444904019158 Medline
34. P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010). doi:10.1107/S0907444909052925 Medline
35. C. J. Williams, J. J. Headd, N. W. Moriarty, M. G. Prisant, L. L. Videau, L. N. Deis, V. Verma, D. A. Keedy, B. J. Hintze, V. B. Chen, S. Jain, S. M. Lewis, W. B. Arendall III, J. Snoeyink, P. D. Adams, S. C. Lovell, J. S. Richardson, D. C. Richardson, MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018). doi:10.1002/pro.3330 Medline

ACKNOWLEDGMENTS

The authors wish to thank Sergey Ovchinnikov, Chris Norn, David Juergens, Jue Wang, Frank DiMaio, Ryan Kibler, Minkyung Baek, Sanaa Mansoor, Luki Goldschmidt, and Lance Stewart for helpful discussions. The authors would also like to 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 collected at The Advanced Light Source (ALS) and 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., D.B.), the Audacious Project at the Institute for Protein Design (A.B., A.K., B.K., F.C., T.F.H., R.J.dH., N.P.K., D.B.), a grant from the 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., D.B.), Howard Hughes Medical Institute Hanna Gray fellowship grant GT11817 (N.Beth.), 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., 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.), the National Science Foundation Graduate Research Fellowship (DGE-2140004, P.J.Y.L), the Howard Hughes Medical Institute (A.C., H.B., D.B.), and the National Institutes of Health, National Institute of General Medical Sciences, P30 GM124169-01(B.S.). We thank Microsoft and AWS for generous gifts of cloud computing credits. **Author contributions:** Conceptualization: JD, LFM, BIMW, AC, RJdH, HB, NBen; Methodology: JD, IA, PJYL; Software: JD, TFH, DT, BK, FC; Validation: JD, NBen, HB, AKB, BS, AK, HN, SP, PJYL, NBeth, RJdH, LFM, BIMW, AC, RJR; Formal analysis: JD, LFM, BIMW, RJR, NBen; Resources: JD, DB; Data curation: IA, JD, HB, ; Writing – original draft: JD, DB; Writing – review and editing: JD, DB; Visualization: JD, RJR, RJdH, HB, LFM, BIMW, PJYL, HB; Supervision: DB, NPK; Project administration: JD; Funding acquisition: JD, DB. **Competing interests:** 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*)

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

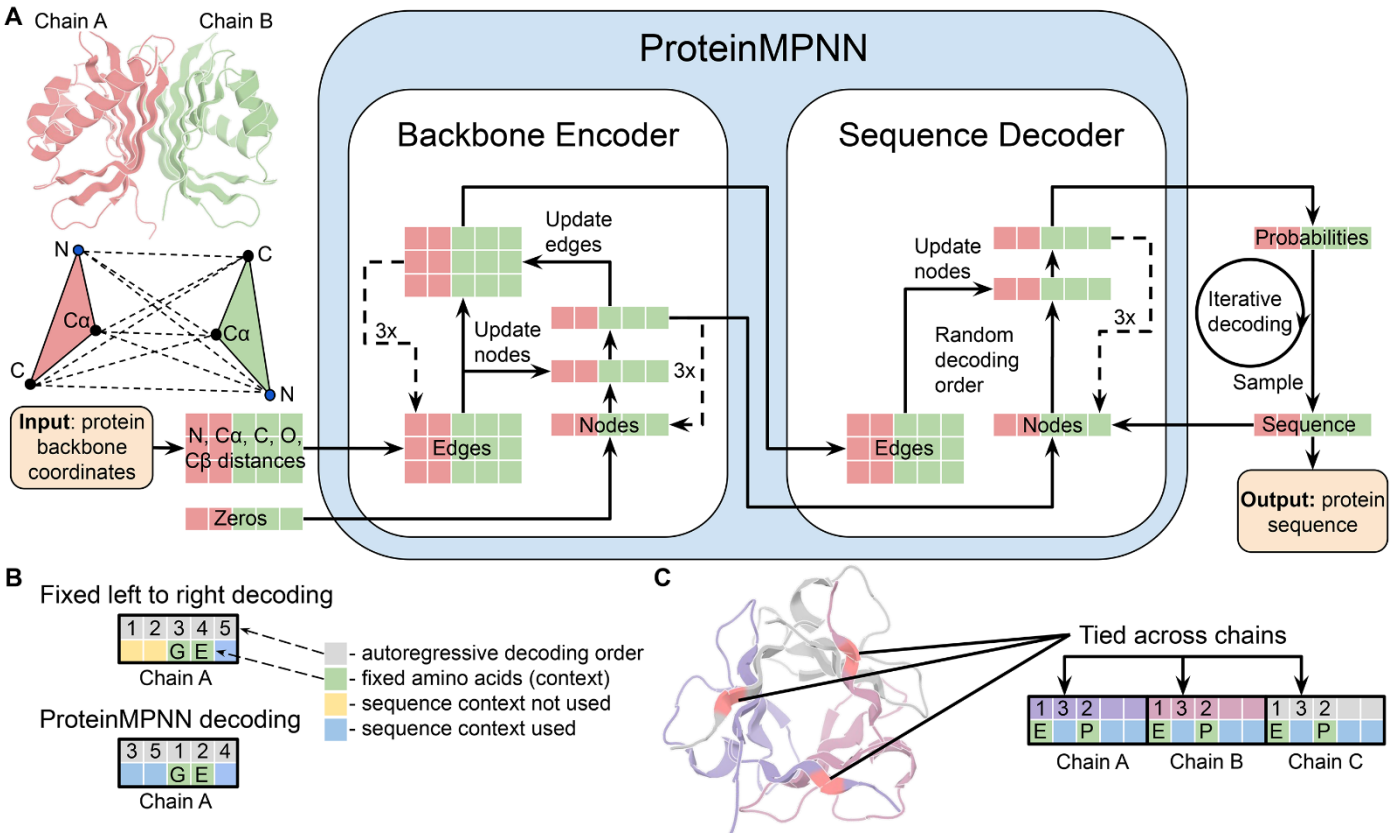

Fig. 1. ProteinMPNN architecture. (**A**) Distances between N, Cα, C, O, and virtual Cβ are encoded and processed using a message passing neural network (Encoder) to obtain graph node and edge features. The encoded features together with a partial sequence are used to generate amino acids iteratively 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 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 homo-trimer is designed with 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.

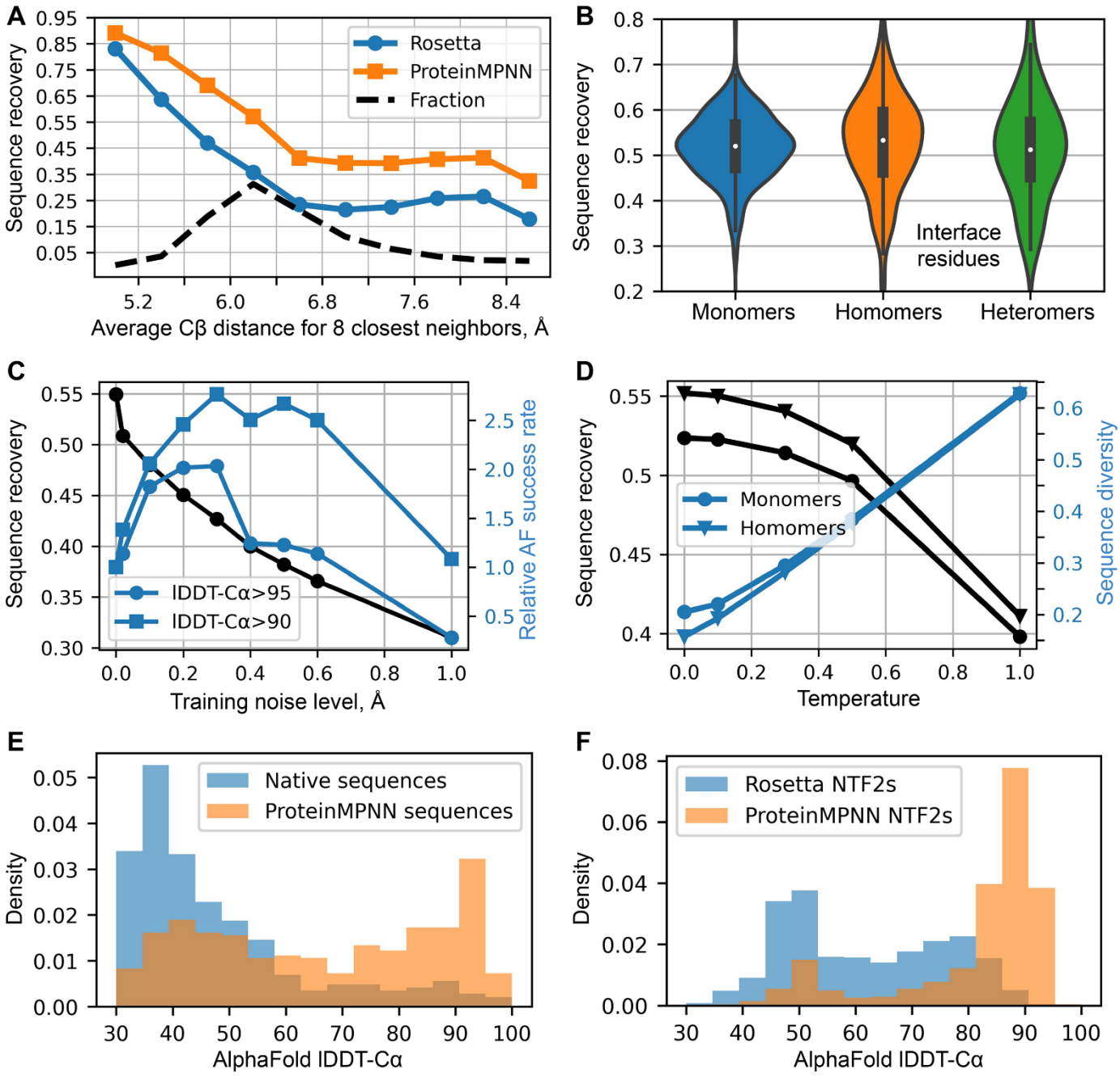

Fig. 2. In silico evaluation of ProteinMPNN. (**A**) ProteinMPNN has higher native sequence recovery than Rosetta. The average Cβ distance of the 8 closest neighbors (x axis) reports on burial, with most buried positions on the left and more exposed on the right; ProteinMPNN outperforms Rosetta at all levels of burial. Average sequence recovery for ProteinMPNN was 52.4%, compared to 32.9% for Rosetta. (**B**) ProteinMPNN has high sequence recovery for monomers and for both homo-oligomer and hetero-oligomer interfaces (Cβ-Cβ<8Å); violin plots are for 690 monomers, 732 homomers, 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 1.8% success rate), while to maximize prediction success at a lower accuracy cutoff (squares), models trained with more noise are better (1.0 corresponds to 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 to the original native sequence using no multiple sequence information. Single sequences (designed or native) were input in both cases. (**F**) ProteinMPNN redesign of previous Rosetta designed NTF2 fold proteins (3,000 backbones in total) results in considerably improved AlphaFold single sequence prediction accuracy.

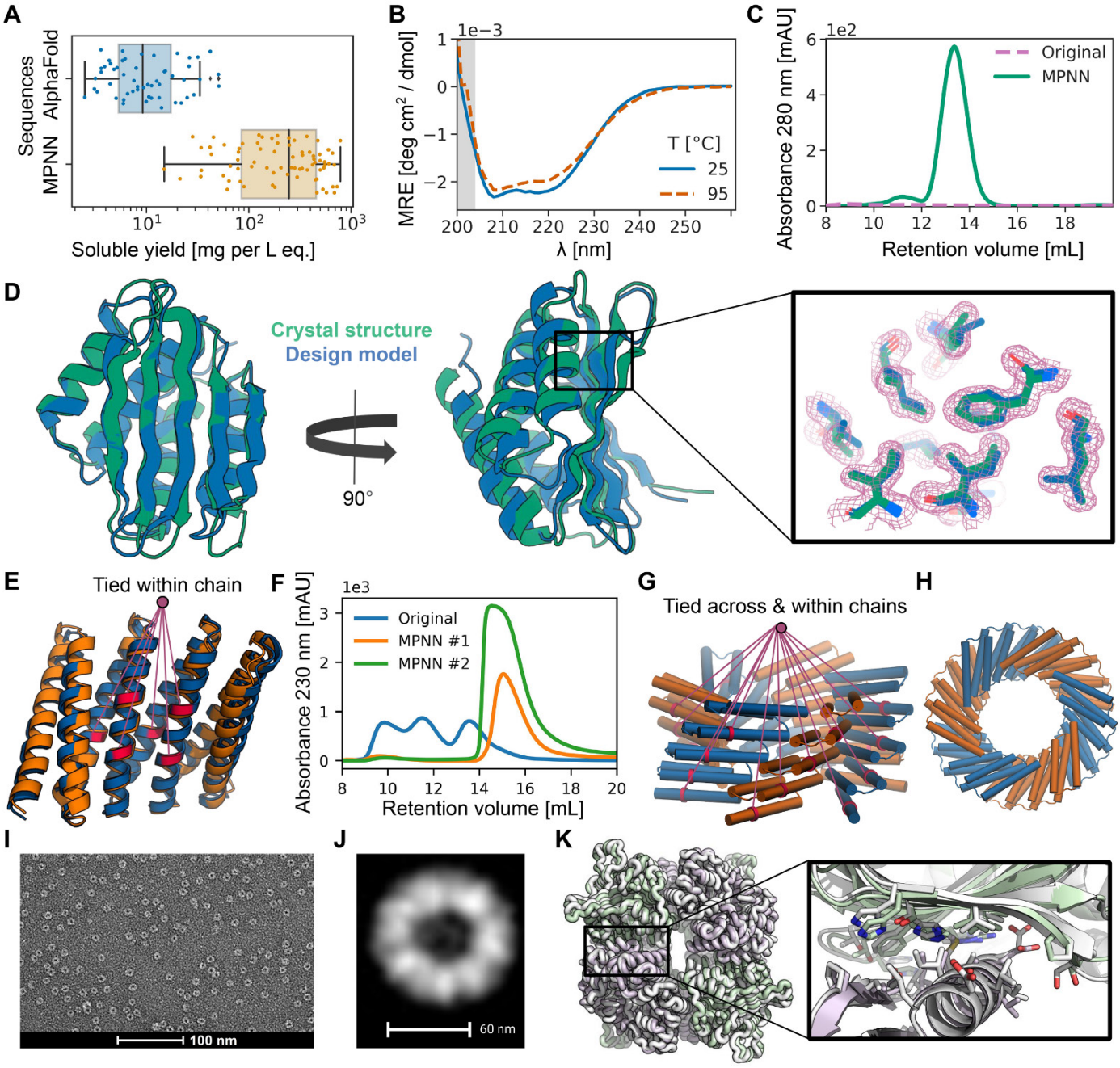

Fig. 3. Structural characterization of ProteinMPNN designs. (**A**) Comparison of soluble protein expression over a set of AlphaFold hallucinated monomers and homo-oligomers (blue) and the same set of backbones with sequences designed using ProteinMPNN (orange), N=129. The total soluble protein yield following expression in *E. coli*, obtained from the integrated area unders size exclusion traces of nickel-NTA purified proteins, increases considerably from the barely soluble protein of the original sequences following ProteinMPNN rescue (median yields for 1 L of culture equivalent: 9 and 247 mg respectively). (**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 similarity to the PDB of the design model are very low (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 to 25°C. Shown in (C) is a size exclusion (SEC) profile of the failed original design overlaid with the ProteinMPNN sequence design, which has a clear monodisperse peak at the expected retention volume. As shown in (D), the crystal structure of the ProteinMPNN (8CYK) design is nearly identical to the design model (2.35 Å RMSD over 130 residues); see fig. S5 for additional information. Right panel shows model sidechains in the electron density, in green crystal side chains, in blue AlphaFold side chains. (**E** and **F**) ProteinMPNN rescue of Rosetta design made from a perfectly repeating structural and sequence unit (DHR82). 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 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 purified design. (**J**) Class average of images from I closely match top down view in (H). (**K**) Rescue of the failed two-component Rosetta tetrahedral nanoparticle design T33-27 (*16*) by ProteinMPNN interface design. Following ProteinMPNN rescue, the nanoparticle assembled readily with high yield, and the crystal structure (grey) is very nearly identical to the design model (green/purple) (backbone RMSD of 1.2 Å over two complete asymmetric units forming the ProteinMPNN rescued interface).

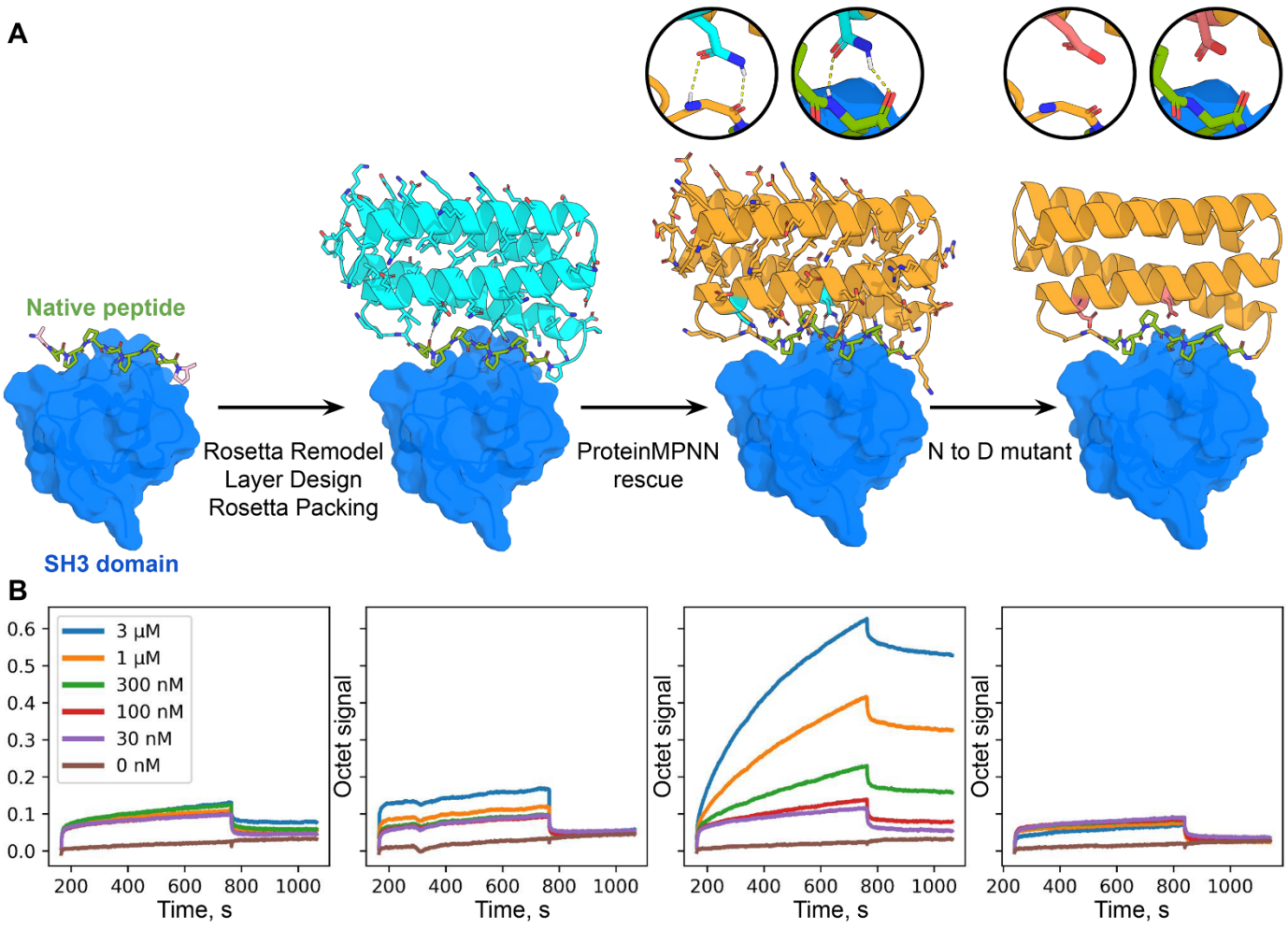

Fig. 4. Design of protein function with ProteinMPNN. (**A**) Design scheme. First panel; structure (PDB 2W0Z) of a fragment of Gab2 peptide bound to the human Grb2 C-term SH3 domain (core SH3-binding motif PPPRPPK is in green, target rendered with surface and colored blue). Second panel: helical bundle 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. Third panel; ProteinMPNN redesign (orange) of the designed binder sequence; hydrogen bonds involving asparagine sidechains between the peptide and base scaffold are shown in green and in the inset. Fourth panel; Mutation of the two asparagines to aspartates to disrupt the scaffolding of the target peptide. (**B**) Experimental characterization of binding using biolayer interferometry. Biotinylated C-term 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 (left), or of the designs (right panels), and then transferred to buffer lacking added protein for dissociation measurements. The ProteinMPNN design (3rd panel from the left) has much greater binding signal than the original Rosetta design (2nd panel from the left); this is greatly reduced by the asparagine to aspartate mutations (last panel). Note that all designs as well as the native peptide are fused with sfGFP at the C terminus.

Table 1. Single chain sequence design performance on CATH held out test split. Test accuracy (percentage of correct amino amino acids recovered) and test perplexity (exponentiated categorical cross entropy loss per residue) for models trained on the native backbone coordinates (left, normal font) and models trained with Gaussian noise (std=0.02Å) added to the backbone coordinates (right, bold font). Noise was only added during training and all test evaluations are with no added noise. The final column shows sequence recovery on 5,000 AlphaFold protein backbone models with average pLDDT > 80.0 randomly chosen from UniRef50 sequences.

Noise level when training: 0.00Å/0.02Å	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 instead of forward decoding	1.678	50.8/**47.9**	4.74/**5.25**	46.9/**48.5**

Robust deep learning–based protein sequence design using ProteinMPNN

J. DauparasI. AnishchenkoN. BennettH. BaiR. J. RagotteL. F. MillesB. I. M. WickyA. CourbetR. J. de HaasN. BethelP. J. Y. LeungT. F. HuddyS. PellockD. TischerF. ChanB. KoepnickH. NguyenA. KangB. SankaranA. K. BeraN. P. KingD. Baker

Science, **Ahead of Print** • DOI: 10.1126/science.add2187