Accurate prediction of protein structures and interactions using a 3-track neural network

Minkyung Baek[1,2], Frank DiMaio[1,2], Ivan Anishchenko[1,2], Justas Dauparas[1,2], Sergey Ovchinnikov[3,4], Gyu Rie Lee[1,2], Jue Wang[1,2], Qian Cong[5,6], Lisa N. Kinch[8], R. Dustin Schaeffer[6], Claudia Millán[9], Hahnbeom Park[1,2], Carson Adams[1,2], Caleb R. Glassman[10,11], Andy DeGiovanni[12], Jose H. Pereira[12], Andria V. Rodrigues[12], Alberdina A. van Dijk[13], Ana C. Ebrecht[13], Diederik J. Opperman[14], Theo Sagmeister[15], Christoph Buhlheller[15,16], Tea Pavkov-Keller[15,17], Manoj K Rathinaswamy[18], Udit Dalwadi[19], Calvin K Yip[19], John E Burke[18], K. Christopher Garcia[20], Nick V. Grishin[6,7,8], Paul D. Adams[12,21], Randy J. Read[9], David Baker[1,2,22*]

Affiliations:

[1]Department of Biochemistry, University of Washington; Seattle, WA98195, USA

[2]Institute for Protein Design, University of Washington; Seattle, WA98195, USA

[3]Faculty of Arts and Sciences, Division of Science, Harvard University; Cambridge, MA02138, USA

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

[5]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center; Dallas, TX, USA

[6]Department of Biophysics, University of Texas Southwestern Medical Center; Dallas, TX, USA

[7]Department of Biochemistry, University of Texas Southwestern Medical Center; Dallas, TX, USA

[8]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center; Dallas, TX, USA

[9]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge; Cambridge, U.K.

[10]Program in Immunology, Stanford University School of Medicine, Stanford, CA 94305, USA

[11]Departments of Molecular and Cellular Physiology and Structural Biology, Stanford University School of Medicine, Stanford, CA 94305, USA

[12]Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[13]Department of Biochemistry, Focus Area Human Metabolomics, North-West University; 2531 Potchefstroom, South Africa

[14]Department of Biotechnology, University of the Free State; 205 Nelson Mandela Drive, Bloemfontein, 9300, South Africa

[15]Institute of Molecular Biosciences, University of Graz; Humboldtstrasse 50, 8010, Graz, Austria

[16]Medical University of Graz; Graz, Austria

[17]BioTechMed-Graz; Graz, Austria

[18]Department of Biochemistry and Microbiology, University of Victoria; Victoria, British Columbia, Canada

[19]Life Sciences Institute, Department of Biochemistry and Molecular Biology, The University of British Columbia; Vancouver, British Columbia, Canada

[20]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305, USA

[21]Department of Bioengineering, University of California Berkeley, Berkeley, CA 94720, USA

[22]Howard Hughes Medical Institute, University of Washington; Seattle, WA98195, USA

*Corresponding author. Email: dabaker@uw.edu

Abstract: DeepMind presented remarkably accurate predictions at the recent CASP14 protein structure prediction assessment conference. We explored network architectures incorporating related ideas and obtained the best performance with a 3-track network in which information at the 1D sequence level, the 2D distance map level, and the 3D coordinate level is successively transformed and integrated. The 3-track network produces structure predictions with accuracies approaching those of DeepMind in CASP14, enables the rapid solution of challenging X-ray crystallography and cryo-EM structure modeling problems, and provides insights into the functions of proteins of currently unknown structure. The network also enables rapid generation of accurate protein-protein complex models from sequence information alone, short circuiting traditional approaches which require modeling of individual subunits followed by docking. We make the method available to the scientific community to speed biological research.

One-Sentence Summary: Accurate protein structure modeling enables the rapid solution of protein structures and provides insights into function.

The prediction of protein structure from amino acid sequence information alone has been a longstanding challenge. The bi-annual Critical Assessment of Structure (CASP) meetings have demonstrated that deep learning methods such as AlphaFold (*1*, *2*) and trRosetta (*3*), that extract information from the large database of known protein structures in the PDB, outperform more traditional approaches that explicitly model the folding process. The outstanding performance of DeepMind's AlphaFold2 in the recent CASP14 meeting (https://predictioncenter.org/casp14/zscores_final.cgi) left the scientific community eager to learn details beyond the overall framework presented and raised the question of whether such accuracy could be achieved outside of a world-leading deep learning company. As described at the CASP14 conference, the AlphaFold2 methodological advances included 1) starting from multiple sequence alignments (MSAs) rather than from more processed features such as inverse covariance matrices derived from MSAs, 2) replacement of 2D convolution with an attention mechanism that better represents interactions between residues distant along the sequence, 3) use of a two-track network architecture in which information at the 1D sequence level and the 2D distance map level is iteratively transformed and passed back and forth, 4) use of an SE(3)-equivariant Transformer network to directly refine atomic coordinates (rather than 2D distance maps as in previous approaches) generated from the two-track network, and 5) end-to-end learning in which all network parameters are optimized by backpropagation from the final generated 3D coordinates through all network layers back to the input sequence.

Network architecture development

Intrigued by the DeepMind results, and with the goal of increasing protein structure prediction accuracy for structural biology research and advancing protein design (*4*), we explored network architectures incorporating different combinations of these five properties. In the absence of a published method, we experimented with a wide variety of approaches for passing information between different parts of the networks, as summarized in the Methods and table S1. We succeeded in producing a "two-track" network with information flowing in parallel along a 1D sequence alignment track and a 2D distance matrix track with considerably better performance than trRosetta (BAKER-ROSETTASERVER and BAKER in Fig. 1B), the next best method after AlphaFold2 in CASP14 (https://predictioncenter.org/casp14/zscores_final.cgi).

We reasoned that better performance could be achieved by extending to a third track operating in 3D coordinate space to provide a tighter connection between sequence, residue-residue distances and orientations, and atomic coordinates. We constructed architectures with the two levels of the two-track model augmented with a third parallel structure track operating on 3D backbone coordinates as depicted in Fig. 1A (see Methods and fig. S1 for details). In this architecture, information flows back and forth between the 1D amino acid sequence information, the 2D distance map, and the 3D coordinates, allowing the network to collectively reason about relationships within and between sequences, distances, and coordinates. In contrast, reasoning about 3D atomic coordinates in the two-track AlphaFold2 architecture happens after processing of the 1D and 2D information is complete (although end-to-end training does link parameters to

some extent). Because of computer hardware memory limitations, we could not train models on large proteins directly as the 3-track models have many millions of parameters; instead, we presented to the network many discontinuous crops of the input sequence consisting of two discontinuous sequence segments spanning a total of 260 residues. To generate final models, we combined and averaged the 1D features and 2D distance and orientation predictions produced for each of the crops and then used two approaches to generate final 3D structures. In the first, the predicted residue-residue distance and orientation distributions are fed into pyRosetta (*5*) to generate all-atom models. In the second, the averaged 1D and 2D features are fed into a final SE(3)-equivariant layer (*6*), and following end-to-end training from amino acid sequence to 3D coordinates, backbone coordinates are generated directly by the network (see Methods). We refer to these networks, which also generate per residue accuracy predictions, as RoseTTAFold. The first has the advantage of requiring lower memory (for proteins over 400 residues, 8GB rather than 24GB) GPUs at inference time and producing full side chain models but requires CPU time for the pyRosetta structure modeling step.

The 3-track models with attention operating at the 1D, 2D, and 3D levels and information flowing between the three levels were the best models we tested (Fig. 1B), clearly outperforming the top 2 server groups (Zhang-server and BAKER-ROSETTASERVER), BAKER human group (ranked second among all groups), and our 2-track attention models on CASP14 targets. As in the case of AlphaFold2, the correlation between multiple sequence alignment depth and model accuracy is lower for RoseTTAFold than for trRosetta and other methods tested at CASP14 (fig. S2). The performance of the 3-track model on the CASP14 targets was still not as good as AlphaFold2 (Fig. 1B). This could reflect hardware limitations that limited the size of the models we could explore, alternative architectures or loss formulations, or more intensive use of the network for inference. DeepMind reported using several GPUs for days to make individual predictions, whereas our predictions are made in a single pass through the network in the same manner that would be used for a server; following sequence and template search (~1.5 hours), the end-to-end version of RoseTTAFold requires ~10 minutes on an RTX2080 GPU to generate backbone coordinates for proteins with less than 400 residues, and the pyRosetta version requires 5 minutes for network calculations on a single RTX2080 GPU and an hour for all-atom structure generation with 15 CPU cores. Incomplete optimization due to computer memory limitations and neglect of side chain information likely explain the poorer performance of the end-to-end version compared to the pyRosetta version (Fig. 1B; the latter incorporates side chain information at the all-atom relaxation stage); since SE(3)-equivariant layers are used in the main body of the 3-track model, the added gain from the final SE(3) layer is likely less than in the AlphaFold2 case. We expect the end-to-end approach to ultimately be at least as accurate once the computer hardware limitations are overcome, and side chains are incorporated.

The improved performance of the 3-track models over the 2-track model with identical training sets, similar attention-based architectures for the 1D and 2D tracks, and similar operations in inference (prediction) mode suggests that simultaneously reasoning at the multiple sequence alignment, distance map, and three-dimensional coordinate representations can more effectively extract sequence-structure relationships than reasoning over only MSA and distance

map information. The relatively low compute cost makes it straightforward to incorporate the methods in a public server and predict structures for large sets of proteins, for example, all human GPCRs, as described below.

Blind structure prediction tests are needed to assess any new protein structure prediction method, but CASP is held only once every two years. Fortunately, the Continuous Automated Model Evaluation (CAMEO) experiment (*7*) tests structure prediction servers blindly on protein structures as they are submitted to the PDB. RoseTTAFold has been evaluated since May 15th, 2021 on CAMEO; over the 69 medium and hard targets released during this time (May 15th, 2021 ~ June 19th, 2021), it outperformed all other servers evaluated in the experiment including Robetta (*3*), IntFold6-TS (*8*), BestSingleTemplate (*9*), and SWISS-MODEL (*10*) (Fig. 1C).

We experimented with approaches for further improving accuracy by more intensive use of the network during sampling. Since the network can take as input templates of known structures, we experimented with a further coupling of 3D structural information and 1D sequence information by iteratively feeding the predicted structures back into the network as templates and random subsampling from the multiple sequence alignments to sample a broader range of models. These approaches generated ensembles containing higher accuracy models, but the accuracy predictor was not able to consistently identify models better than those generated by the rapid single pass method (fig. S3). Nevertheless, we suspect that these approaches can improve model performance and are carrying out further investigations along these lines.

In developing RoseTTAFold, we found that combining predictions from multiple discontinuous crops generated more accurate structures than predicting the entire structure at once (fig. S4A). We hypothesized that this arises from selecting the most relevant sequences for each region from the very large number of aligned sequences often available (fig. S4B). To enable the network to focus on the most relevant sequence information for each region while keeping access to the full multiple sequence alignment in a more memory efficient way, we experimented with the Perceiver architecture (*11*), updating smaller seed MSAs (up to 100 sequences) with extra sequences (thousands of sequences) through cross-attention (fig. S4C). Current RoseTTAFold only uses the top 1000 sequences due to memory limitations; with this addition, all available sequence information can be used (often over 10,000 sequences). Initial results are promising (fig. S4D), but more training will be required for rigorous comparison.

Enabling experimental protein structure determination

With the recent considerable progress in protein structure prediction, a key question is what accurate protein structure models can be used for. We investigated the utility of the RoseTTAFold to facilitate experimental structure determination by X-ray crystallography and cryo-electron microscopy and to build models providing biological insights for key proteins of currently unknown structures.

Solution of X-ray structures by molecular replacement (MR) often requires quite accurate models. The much higher accuracy of the RoseTTAFold method than currently available

methods prompted us to test whether it could help solve previously unsolved challenging MR problems and improve the solution of borderline cases. Four recent crystallographic datasets (summarized, including resolution limits, in table S2), which had eluded solution by MR using models available in the PDB, were reanalyzed using RoseTTAFold models: glycine N-acyltransferase (GLYAT) from *Bos taurus* (fig. S5A), a bacterial oxidoreductase (fig. S5B), a bacterial surface layer protein (SLP) (Fig. 2A) and the secreted protein Lrbp from the fungus *Phanerochaete chrysosporium* (Fig. 2B and fig. S5C). In all four cases, the predicted models had sufficient structural similarity to the true structures that led to successful MR solutions (see Methods for details; the per-residue error estimates by DeepAccNet (*12*) allowed the more accurate parts to be weighted more heavily). The increased prediction accuracy was critical for success in all cases, as models made with trRosetta did not yield MR solutions.

To determine why the RoseTTAFold models were successful, where PDB structures had previously failed, we compared the models to the crystal structures we obtained. The images in Fig. 2A and fig. S5 show that in each case, the closest homolog of the known structure was a much poorer model than the RoseTTAFold model; in the case of SLP, only a distant model covering part of the N-terminal domain (38% of the sequence) was available in the PDB, while no homologs of the C-terminal domain of SLP or any portion of Lrbp could be detected using HHsearch (*13*).

Building atomic models of protein assemblies from cryo-EM maps can be challenging in the absence of homologs with known structures. We used RoseTTAFold to predict the p101 $G_{\beta\gamma}$ binding domain (GBD) structure in a heterodimeric $PI3K_{\gamma}$ complex. The top HHsearch hit has a statistically insignificant E-value of 40 and only covers 14 residues out of 167 residues. The predicted structure could readily fit into the electron density map despite the low local resolution (Fig. 2C, top; trRosetta failed to predict the correct fold with the same MSA input (fig. S6)). The C_{α}-RMSD between the predicted and the final refined structure is 3.0 Å over the beta-sheets (Fig. 2C, bottom).

Providing insights into biological function

Experimental structure determination can provide considerable insight into biological function and mechanism. We investigated whether structures generated by RoseTTAFold could similarly provide new insights into function. We focused on two sets of proteins: first, G protein-coupled receptors of currently unknown structure, and second, a set of human proteins implicated in disease. Benchmark tests on GPCR sequences with determined structures showed that RoseTTAFold models for both active and inactive states can be quite accurate even in the absence of close homologs with known structures (and better than those in current GPCR model databases (*14*, *15*); fig. S7) and that the DeepAccNet model quality predictor (*12*) provides a good measure of actual model accuracy (fig. S7D). We provide RoseTTAFold models and accompanying accuracy predictions for closed and open states of all human GPCRs of currently unknown structure.

Protein structures can provide insight into how mutations in key proteins lead to human disease. We identified human proteins without close homologs of known structure that contain multiple disease-causing mutations or have been the subject of intensive experimental investigation (see Methods). We used RoseTTAFold to generate models for 693 domains from such proteins. Over one-third of these models have a predicted lDDT > 0.8, which corresponded to an average C_α-RMSD -RMSD of 2.6 Å on CASP14 targets (fig. S8). Here, we focus on three examples that illustrate the different ways in which structure models can provide insight into the function or mechanisms of diseases.

Deficiencies in TANGO2 (transport and Golgi organization protein 2) lead to metabolic disorders, and the protein plays an unknown role in Golgi membrane redistribution into the ER (*16*, *17*). The RoseTTAFold model of TANGO2 adopts an N-terminal nucleophile aminohydrolase (Ntn) fold (Fig. 3A) with well-aligned active site residues that are conserved in TANGO2 orthologs (Fig. 3B). Ntn superfamily members with structures similar to the RoseTTAFold model suggest that TANGO2 functions as an enzyme that might hydrolyze a carbon-nitrogen bond in a membrane component (*18*). Based on the model, known mutations that cause disease (magenta spheres in Fig. 3A) could act by hindering catalysis (R26K, R32Q, and L50P, near active site) or produce steric clashes (G154R) (*19*) in the hydrophobic core. By comparison, a homology model based on very distant (<15% sequence identity) homologs had multiple alignment shifts that misplace key conserved residues (fig. S9 and table S3)

The ADAM (A Disintegrin And Metalloprotease) and ADAMTS families of metalloproteases are encoded by over 40 human genes, mediate cell-cell and cell-matrix interactions (*20*, *21*) and are involved in a range of human diseases, including cancer metastasis, inflammatory disorders, neurological diseases and asthma (*21*, *22*). The ADAMs contain prodomain and metalloprotease domains; the fold of the metalloprotease is known (*23*, *24*), but not that of the prodomain, which has no homologs of known structure. The RoseTTAFold predicted structure of the ADAM33 prodomain has a lipocalin-like beta-barrel fold (Fig. 3C) belonging to an extended superfamily that includes metalloprotease inhibitors (MPIs) (*25*). There is a cysteine in an extension following the predicted prodomain barrel; taken together, these data are consistent with experimental data suggesting that the ADAM prodomain inhibits metalloprotease activity using a cysteine switch (*26*). Conserved residues within ADAM33 orthologs line one side of the barrel and likely interact with the metalloprotease (Fig. 3D).

Transmembrane spanning Ceramide synthase (CERS1) is a key enzyme in sphingolipid metabolism which uses acyl-CoA to generate ceramides with various acyl chain lengths that regulate differentiation, proliferation, and apoptosis (*27*). Structure information is not available for any of the CerS enzymes or their homologs, and the number and orientation of transmembrane helices (TMH) are not known (*28*). The RoseTTAFold CERS1 model for residues 98 to 304 (Pfam TLC domain) (*29*) includes six TMH that traverse the membrane in an up and down arrangement (Fig. 3E). A central crevice extends into the membrane and is lined with residues required for activity (His182 and Asp213) (*30*) or conserved (W298), as well as a pathogenic mutation (H183Q) found in progressive myoclonus epilepsy and dementia that

decreases ceramide levels (*31*). This active site composition (His182, Asp 213, and potentially a neighboring Ser212) suggests testable reaction mechanisms for the enzyme (Fig. 3F).

Direct generation of protein-protein complex models

The final layer of the end-to-end version of our 3-track network generates 3D structure models by combining features from discontinuous crops of the protein sequence (two segments of the protein with a chain break between them). We reasoned that because the network can seamlessly handle chain breaks, it might be able to predict the structure of protein-protein complexes directly from sequence information. Rather than providing the network the sequence of a single protein, with or without possible template structures, two or more sequences (and possible templates for these) can be input, with the output the backbone coordinates of two or more protein chains. Thus, the network enables the direct building of structure models for protein-protein complexes from sequence information, short circuiting the standard procedure of building models for individual subunits and then carrying out rigid-body docking. In addition to the great reduction in compute time required (complex models are generated from sequence information in ~30 min on a 24G TITAN RTX GPU), this approach implements "flexible backbone" docking almost by construction as the structures of the chains are predicted in the context of each other. We tested the end-to-end 3-track network on paired sequence alignments for complexes of known structures (*32*) (see Methods and table S4 for details) containing two (Fig. 4A) or three (Fig. 4B) chains, and in many cases, the resulting models were very close to the actual structures (TM-score (*33*) > 0.8). Information on residue-residue co-evolution between the paired sequences likely contributes to the accuracy of the rigid body placement as more accurate complex structures were generated when more sequences were available (fig. S10). The network was trained on monomeric proteins, not complexes, so there may be some training set bias in the monomer structures, but there is none for the complexes.

To illustrate the application of RoseTTAFold to complexes of unknown structure with more than three chains, we used it to generate models of the complete four-chain human IL-12R/IL-12 complex (Fig. 4C and fig. S11). A previously published cryo-EM map of the IL-12 receptor complex indicated a similar topology to that of the IL-23 receptor; however, the resolution was not sufficient to observe the detailed interaction between IL-12Rβ2 and IL-12p35 (*34*). Such an understanding is important for dissecting the specific actions of IL-12 and IL-23 and generating inhibitors that block IL-12 without impacting IL-23 signaling. The RoseTTAFold model fits the experimental cryo-EM density well and identified a shared interaction between Y189 in IL-12p35 and G115 in IL-12Rβ2 analogous to the packing between W156 in IL-23p19 with G116 in IL-23R. In addition, the model suggests a role for the IL-12Rβ2 N-terminal peptide (residue 24-31) in IL-12 binding not observed in the IL-12 cryo-electron microscopy (IL-12Rβ2 D26 may interact with nearby K190 and K194 in IL-12p35), which may provide an avenue to target the interaction between IL-12 and IL-12Rβ2 specifically.

Conclusions

RoseTTAFold enables solutions of challenging X-ray crystallography and cryo-EM modeling problems, provides insight into protein function in the absence of experimentally determined structures, and rapidly generates accurate models of protein-protein complexes. Further training on protein-protein complex datasets will likely further improve the modeling of the structures of multiprotein assemblies. The approach can be readily coupled with existing small molecule and protein binder design methodology to improve computational discovery of new protein and small molecule ligands for targets of interest. The simultaneous processing of sequence, distance, and coordinate information by the three-track architecture opens the door to new approaches incorporating constraints and experimental information at all three levels for problems ranging from cryo-EM structure determination to protein design.

References and notes

1. A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, Improved protein structure prediction using potentials from deep learning. *Nature*. **577**, 706–710 (2020).

2. John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Kathryn Tunyasuvunakool, Olaf Ronneberger, Russ Bates, Augustin Žídek, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Anna Potapenko, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Martin Steinegger, Michalina Pacholska, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, Demis Hassabis., in *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction*.

3. J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, D. Baker, Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 1496–1503 (2020).

4. I. Anishchenko, T. M. Chidyausiku, S. Ovchinnikov, S. J. Pellock, D. Baker, De novo protein design by deep network hallucination. *bioRxiv* (2020), p. 2020.07.22.211482.

5. S. Chaudhury, S. Lyskov, J. J. Gray, PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*. **26**, 689–691 (2010).

6. F. B. Fuchs, D. E. Worrall, V. Fischer, M. Welling, SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. *arXiv [cs.LG]* (2020), (available at http://arxiv.org/abs/2006.10503).

7. J. Haas, A. Barbato, D. Behringer, G. Studer, S. Roth, M. Bertoni, K. Mostaguir, R. Gumienny, T. Schwede, Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins*. **86 Suppl 1**, 387–398 (2018).

8. L. J. McGuffin, R. Adiyaman, A. H. A. Maghrabi, A. N. Shuid, D. A. Brackenridge, J. O. Nealon, L. S. Philomina, IntFOLD: an integrated web resource for high performance protein structure and function prediction. *Nucleic Acids Research*. **47** (2019), pp. W408–W413.

9. J. Haas, R. Gumienny, A. Barbato, F. Ackermann, G. Tauriello, M. Bertoni, G. Studer, A. Smolinski, T. Schwede, Introducing "best single template" models as reference baseline for the Continuous Automated Model Evaluation (CAMEO). *Proteins*. **87**, 1378–1387 (2019).

10. A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, R. Lepore, T. Schwede, SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*. **46** (2018), pp. W296–W303.

11. A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, J. Carreira, Perceiver: General Perception with Iterative Attention. *arXiv [cs.CV]* (2021), (available at http://arxiv.org/abs/2103.03206).

12. N. Hiranuma, H. Park, M. Baek, I. Anishchenko, J. Dauparas, D. Baker, Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat. Commun.* **12**, 1340 (2021).

13. M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger, J. Söding, HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*. **20**, 473 (2019).

14. A. J. Kooistra, S. Mordalski, G. Pándy-Szekeres, M. Esguerra, A. Mamyrbekov, C. Munk, G. M. Keserű, D. E. Gloriam, GPCRdb in 2021: integrating GPCR sequence, structure and function. *Nucleic Acids Res.* **49**, D335–D343 (2021).

15. B. J. Bender, B. Marlow, J. Meiler, Improving homology modeling from low-sequence identity templates in Rosetta: A case study in GPCRs. *PLoS Comput. Biol.* **16**, e1007597 (2020).

16. L. S. Kremer, F. Distelmaier, B. Alhaddad, M. Hempel, A. Iuso, C. Küpper, C. Mühlhausen, R. Kovacs-Nagy, R. Satanovskij, E. Graf, R. Berutti, G. Eckstein, R. Durbin, S. Sauer, G. F. Hoffmann, T. M. Strom, R. Santer, T. Meitinger, T. Klopstock, H. Prokisch, T. B. Haack, Bi-allelic Truncating Mutations in TANGO2 Cause Infancy-Onset Recurrent

Metabolic Crises with Encephalocardiomyopathy. *Am. J. Hum. Genet.* **98**, 358–362 (2016).

17. C. Rabouille, V. Kondylis, TANGOing along the protein secretion pathway. *Genome Biol.* **7**, 213 (2006).

18. M. P. Milev, D. Saint-Dic, K. Zardoui, T. Klopstock, C. Law, F. Distelmaier, M. Sacher, The phenotype associated with variants in TANGO2 may be explained by a dual role of the protein in ER-to-Golgi transport and at the mitochondria. *J. Inherit. Metab. Dis.* **44**, 426–437 (2021).

19. S. R. Lalani, P. Liu, J. A. Rosenfeld, L. B. Watkin, T. Chiang, M. S. Leduc, W. Zhu, Y. Ding, S. Pan, F. Vetrini, C. Y. Miyake, M. Shinawi, T. Gambin, M. K. Eldomery, Z. H. C. Akdemir, L. Emrick, Y. Wilnai, S. Schelley, M. K. Koenig, N. Memon, L. S. Farach, B. P. Coe, M. Azamian, P. Hernandez, G. Zapata, S. N. Jhangiani, D. M. Muzny, T. Lotze, G. Clark, A. Wilfong, H. Northrup, A. Adesina, C. A. Bacino, F. Scaglia, P. E. Bonnen, J. Crosson, J. Duis, G. H. B. Maegawa, D. Coman, A. Inwood, J. McGill, E. Boerwinkle, B. Graham, A. Beaudet, C. M. Eng, N. A. Hanchard, F. Xia, J. S. Orange, R. A. Gibbs, J. R. Lupski, Y. Yang, Recurrent Muscle Weakness with Rhabdomyolysis, Metabolic Crises, and Cardiac Arrhythmia Due to Bi-allelic TANGO2 Mutations. *Am. J. Hum. Genet.* **98**, 347–357 (2016).

20. T. G. Wolfsberg, P. Primakoff, D. G. Myles, J. M. White, ADAM, a novel family of membrane proteins containing A Disintegrin And Metalloprotease domain: multipotential functions in cell-cell and cell-matrix interactions. *J. Cell Biol.* **131**, 275–278 (1995).

21. T. Klein, R. Bischoff, Active metalloproteases of the A Disintegrin and Metalloprotease (ADAM) family: biological function and structure. *J. Proteome Res.* **10**, 17–33 (2011).

22. S. Zhong, R. A. Khalil, A Disintegrin and Metalloproteinase (ADAM) and ADAM with thrombospondin motifs (ADAMTS) family in vascular biology and disease. *Biochem. Pharmacol.* **164**, 188–204 (2019).

23. P. Orth, P. Reichert, W. Wang, W. W. Prosise, T. Yarosh-Tomaine, G. Hammond, R. N. Ingram, L. Xiao, U. A. Mirza, J. Zou, C. Strickland, S. S. Taremi, H. V. Le, V. Madison, Crystal structure of the catalytic domain of human ADAM33. *J. Mol. Biol.* **335**, 129–137 (2004).

24. S. Takeda, T. Igarashi, H. Mori, S. Araki, Crystal structures of VAP1 reveal ADAMs' MDC domain architecture and its unique C-shaped scaffold. *EMBO J.* **25**, 2388–2396 (2006).

25. D. R. Flower, A. C. North, C. E. Sansom, The lipocalin protein family: structural

and sequence overview. *Biochim. Biophys. Acta*. **1482**, 9–24 (2000).

26. H. E. Van Wart, H. Birkedal-Hansen, The cysteine switch: a principle of regulation of metalloproteinase activity with potential applicability to the entire matrix metalloproteinase gene family. *Proceedings of the National Academy of Sciences*. **87** (1990), pp. 5578–5582.

27. M. Levy, A. H. Futerman, Mammalian ceramide synthases. *IUBMB Life* (2010), p. NA–NA, , doi:10.1002/iub.319.

28. J. L. Kim, B. Mestre, S.-H. Shin, A. H. Futerman, Ceramide synthases: Reflections on the impact of Dr. Lina M. Obeid. *Cellular Signalling*. **82** (2021), p. 109958.

29. E. Winter, C. P. Ponting, TRAM, LAG1 and CLN8: members of a novel family of lipid-sensing domains? *Trends in Biochemical Sciences*. **27** (2002), pp. 381–383.

30. S. Spassieva, J.-G. Seo, J. C. Jiang, J. Bielawski, F. Alvarez-Vasquez, S. Michal Jazwinski, Y. A. Hannun, L. M. Obeid, Necessary Role for the Lag1p Motif in (Dihydro)ceramide Synthase Activity. *Journal of Biological Chemistry*. **281** (2006), pp. 33931–33938.

31. N. Vanni, F. Fruscione, E. Ferlazzo, P. Striano, A. Robbiano, M. Traverso, T. Sander, A. Falace, E. Gazzerro, P. Bramanti, J. Bielawski, A. Fassio, C. Minetti, P. Genton, F. Zara, Impairment of ceramide synthesis causes a novel progressive myoclonus epilepsy. *Annals of Neurology*. **76** (2014), pp. 206–212.

32. Q. Cong, I. Anishchenko, S. Ovchinnikov, D. Baker, Protein interaction networks revealed by proteome coevolution. *Science*. **365**, 185–189 (2019).

33. Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality. *Proteins*. **57**, 702–710 (2004).

34. C. R. Glassman, Y. K. Mathiharan, K. M. Jude, L. Su, O. Panova, P. J. Lupardus, J. B. Spangler, L. K. Ely, C. Thomas, G. Skiniotis, K. C. Garcia, Structural basis for IL-12 and IL-23 receptor sharing reveals a gateway for shaping actions on T versus NK cells. *Cell*. **184**, 983–999.e24 (2021).

35. E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, T. E. Ferrin, UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

36. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, J. Wang, D. Baker, *RoseTTAFold: The first release of RoseTTAFold* (2021; https://zenodo.org/record/5068265).

37. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need. *arXiv [cs.CL]* (2017), (available at http://arxiv.org/abs/1706.03762).

38. J. Ho, N. Kalchbrenner, D. Weissenborn, T. Salimans, Axial Attention in Multidimensional Transformers. *arXiv [cs.CV]* (2019), (available at http://arxiv.org/abs/1912.12180).

39. K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell, A. Weller, Rethinking Attention with Performers. *arXiv [cs.LG]* (2020), (available at http://arxiv.org/abs/2009.14794).

40. R. Rao, J. Liu, R. Verkuil, J. Meier, J. F. Canny, P. Abbeel, T. Sercu, A. Rives, MSA Transformer. *bioRxiv* (2021), p. 2021.02.12.430858.

41. F. Ju, J. Zhu, B. Shao, L. Kong, T.-Y. Liu, W.-M. Zheng, D. Bu, CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction. *Nat. Commun.* **12**, 2535 (2021).

42. Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, Y. Sun, Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification. *arXiv [cs.LG]* (2020), (available at http://arxiv.org/abs/2009.03509).

43. 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).

44. M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding, M. Steinegger, Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45**, D170–D176 (2017).

45. M. Steinegger, M. Mirdita, J. Söding, Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods*. **16**, 603–606 (2019).

46. L. Zimmermann, A. Stephens, S.-Z. Nam, D. Rau, J. Kübler, M. Lozajic, F. Gabler, J. Söding, A. N. Lupas, V. Alva, A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* **430**, 2237–2243 (2018).

47. G. Bunkóczi, R. J. Read, Improvement of molecular-replacement models with Sculptor. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 303–312 (2011).

48. G. Bunkóczi, R. J. Read, phenix. ensembler: a tool for multiple superposition.

Comput Crystallogr Newsl. **2**, 8–9 (2011).

49. 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).

50. A. Vagin, A. Lebedev, in *ACTA CRYSTALLOGRAPHICA A-FOUNDATION AND ADVANCES* (INT UNION CRYSTALLOGRAPHY 2 ABBEY SQ, CHESTER, CH1 2HU, ENGLAND, 2015), vol. 71, pp. S19–S19.

51. Y. Wang, J. Virtanen, Z. Xue, Y. Zhang, I-TASSER-MR: automated molecular replacement for distant-homology proteins using iterative fragment assembly and progressive sequence truncation. *Nucleic Acids Res.* **45**, W429–W434 (2017).

52. A. J. McCoy, R. D. Oeffner, A. G. Wrobel, J. R. M. Ojala, K. Tryggvason, B. Lohkamp, R. J. Read, Ab initio solution of macromolecular crystal structures without direct methods. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 3637–3641 (2017).

53. G. Bunkóczi, B. Wallner, R. J. Read, Local error estimates dramatically improve the utility of homology models for solving crystal structures by molecular replacement. *Structure*. **23**, 397–406 (2015).

54. T. C. Terwilliger, Maximum-likelihood density modification. *Acta Crystallogr. D Biol. Crystallogr.* **56**, 965–972 (2000).

55. D. Liebschner, P. V. Afonine, M. L. Baker, G. Bunkóczi, V. B. Chen, T. I. Croll, B. Hintze, L. W. Hung, S. Jain, A. J. McCoy, N. W. Moriarty, R. D. Oeffner, B. K. Poon, M. G. Prisant, R. J. Read, J. S. Richardson, D. C. Richardson, M. D. Sammito, O. V. Sobolev, D. H. Stockwell, T. C. Terwilliger, A. G. Urzhumtsev, L. L. Videau, C. J. Williams, P. D. Adams, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol*. **75**, 861–877 (2019).

56. T. C. Terwilliger, R. W. Grosse-Kunstleve, P. V. Afonine, N. W. Moriarty, P. H. Zwart, L. W. Hung, R. J. Read, P. D. Adams, Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr. D Biol. Crystallogr.* **64**, 61–69 (2008).

57. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).

58. P. V. Afonine, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart, P. D. Adams, Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012).

59. 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 3rd, 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).

60. R. J. Read, A. J. McCoy, Using SAD data in Phaser. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 338–344 (2011).

61. J. Xu, M. McPartlon, J. Li, Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nature Machine Intelligence*, 1–9 (2021).

62. J. Yang, Y. Zhang, I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* **43**, W174–81 (2015).

63. D. Xu, Y. Zhang, Toward optimal fragment generations for ab initio protein structure assembly. *Proteins*. **81**, 229–239 (2013).

64. The UniProt Consortium, UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).

65. J. Pei, N. V. Grishin, The DBSAV Database: Predicting Deleteriousness of Single Amino Acid Variations in the Human Proteome. *J. Mol. Biol.* **433**, 166915 (2021).

66. L. S. Johnson, S. R. Eddy, E. Portugaly, Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*. **11**, 431 (2010).

67. S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. E. Tosatto, R. D. Finn, The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).

68. S. Bienert, A. Waterhouse, T. A. P. de Beer, G. Tauriello, G. Studer, L. Bordoli, T. Schwede, The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res.* **45**, D313–D319 (2017).

69. B. Mészáros, G. Erdos, Z. Dosztányi, IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337 (2018).

70. J. Hanson, K. K. Paliwal, T. Litfin, Y. Zhou, SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning. *Genomics, Proteomics & Bioinformatics*. **17** (2019), pp. 645–656.

71. F. Gabler, S.-Z. Nam, S. Till, M. Mirdita, M. Steinegger, J. Söding, A. N. Lupas, V. Alva, Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr. Protoc. Bioinformatics*. **72**, e108 (2020).

72. H. Cheng, R. D. Schaeffer, Y. Liao, L. N. Kinch, J. Pei, S. Shi, B.-H. Kim, N. V. Grishin, ECOD: an evolutionary classification of protein domains. *PLoS Comput. Biol.* **10**, e1003926 (2014).

73. R. Ayoub, Y. Lee, RUPEE: A fast and accurate purely geometric protein structure search. *PLoS One*. **14**, e0213712 (2019).

74. J. Pei, N. V. Grishin, AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*. **17**, 700–712 (2001).

75. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

76. A. M. Altenhoff, C.-M. Train, K. J. Gilbert, I. Mediratta, T. Mendes de Farias, D. Moi, Y. Nevers, H.-S. Radoykova, V. Rossier, A. Warwick Vesztrocy, N. M. Glover, C. Dessimoz, OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res.* **49**, D373–D379 (2021).

77. P. Benkert, M. Biasini, T. Schwede, Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*. **27**, 343–350 (2011).

78. L. Holm, Using Dali for Protein Structure Comparison. *Methods Mol. Biol.* **2112**, 29–42 (2020).

79. S. J. Hubbard, J. M. Thornton, naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*. **2** (1993).

80. A. Lafita, S. Bliven, A. Kryshtafovych, M. Bertoni, B. Monastyrskyy, J. M. Duarte, T. Schwede, G. Capitani, Assessment of protein assembly prediction in CASP12. *Proteins: Structure, Function, and Bioinformatics*. **86** (2018), pp. 247–256.

81. P. Conway, M. D. Tyka, F. DiMaio, D. E. Konerding, D. Baker, Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23**, 47–55 (2014).

82. M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, D. G. Higgins, Clustal W and Clustal X version 2.0. *Bioinformatics*. **23** (2007), pp. 2947–2948.

Acknowledgments: We thank Eric Horvitz, Naozumi Hiranuma, David Juergens, Sanaa Mansoor, and Doug Tischer for helpful discussions, David E. Kim for web-server construction, and Luki Goldschmidt for computing resource management. TPK thanks Bernd Nidetzky and Mareike Monschein from Graz University of Technology for providing protein samples for crystallization. DJO gratefully acknowledges assistance with data collection from scientists of Diamond Light Source beamline I04 under proposal mx20303. TS, CB, and TPK acknowledge the ESRF (ID30-3, Grenoble, France) and DESY (P11, PETRAIII, Hamburg, Germany) for provision of synchrotron-radiation facilities and support during data collection.

Funding: This work was supported by Microsoft (MB, DB, and generous gifts of Azure compute time and expertise), Open Philanthropy (DB, GRL), Eric and Wendy Schmidt by recommendation of the Schmidt Futures program (FD, HP), The Washington Research Foundation (MB, GRL, JW), National Science Foundation Cyberinfrastructure for Biological Research, Award # DBI 1937533 (IA), Wellcome Trust, grant number 209407/Z/17/Z (RJR), National Institute of Health, grant numbers P01GM063210 (PDA, RJR), DP5OD026389 (SO), RO1-AI51321 (KCG) and GM127390 (NVG), Mathers Foundation (KCG), Canadian Institute of Health Research (CIHR) Project Grant, grant numbers 168998 (JEB) and 168907 (CKY), the Welch Foundation I-1505 (NVG), Global Challenges Research Fund (GCRF) through Science & Technology Facilities Council (STFC), grant number ST/R002754/1: Synchrotron Techniques for African Research and Technology (START) (DJO, AAvD, ACE), Austrian Science Fund (FWF) projects P29432 and DOC50 (doc.fund Molecular Metabolism) (TS, CB, TP).

Author contributions: MB, FD, and DB designed the research; MB, FD, IA, JD, SO, JW developed deep learning network; GRL and HP analyzed GPCR modeling results; QC, LNK, RDS, NVG analyzed modeling results for proteins related to the human diseases; CRG KCG analyzed modeling results for the IL-12R/IL-12 complex; PDA, RJR, CA, FD, CM worked on structure determination; AAvD, ACE, DJO, TS, CB, TPK, MKR, UD, CKY, JEB, AD, JHP, AVR provided experimental data; MB, FD, GRL, QC, LNK, HP, CRG, PDA, RJR, DB wrote the manuscript; all authors discussed the results and commented on the manuscript.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: The GPCR models of unknown structures have been deposited to http://files.ipd.uw.edu/pub/RoseTTAFold/all_human_GPCR_unknown_models.tar.gz and http://files.ipd.uw.edu/pub/RoseTTAFold/GPCR_benchmark_one_state_unknown_models.tar.gz. The model structures for structurally uncharacterized human proteins have been deposited to http://files.ipd.uw.edu/pub/RoseTTAFold/human_prot.tar.gz. The atomic models have been deposited at the Protein Data Bank (PDB) with accession codes PDB: 7MEZ (full PI3K complex structure). The structures for GLYAT, oxidoreductase, SLP, and Lrbp proteins will be deposited in the PDB when final processing is completed. The method

is available as a server at https://robetta.bakerlab.org (RoseTTAFold option), and the source code and model parameters are available at https://github.com/RosettaCommons/RoseTTAFold or Zenodo (*36*).

Supplementary Materials

Materials and Methods

Figs. S1 to S17

Tables S1 to S4

References (*37*–*82*)

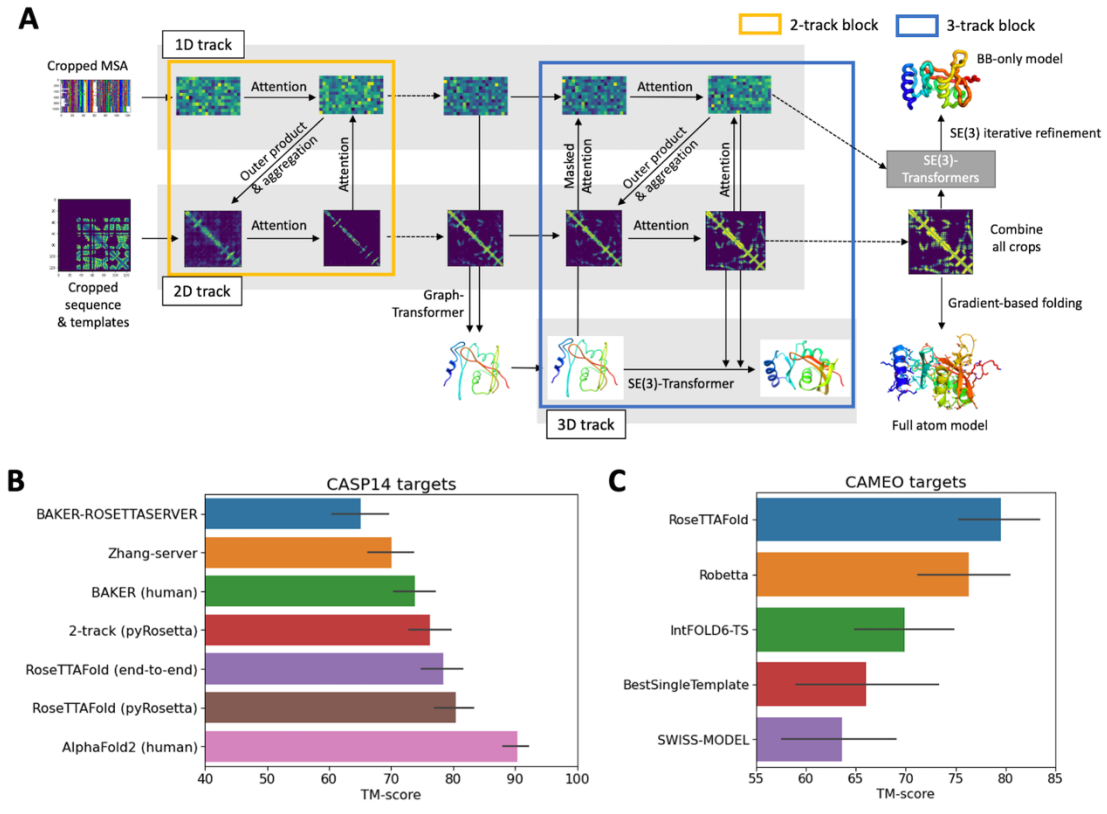

Fig. 1. Network architecture and performance. (A) RoseTTAFold architecture with 1D, 2D, and 3D attention tracks. Multiple connections between tracks allow the network to simultaneously learn relationships within and between sequences, distances, and coordinates (see Methods and fig. S1 for details). **(B)** Average TM-score of prediction methods on the CASP14 targets. Zhang-server and BAKER-ROSETTASERVER were the top 2 server groups while AlphaFold2 and BAKER were the top 2 human groups in CASP14; BAKER-ROSETTASERVER and BAKER predictions were based on trRosetta. Predictions with the 2-track model and RoseTTAFold (both end-to-end and pyRosetta version) were completely automated. **(C)** Blind benchmark results on CAMEO medium and hard targets; model accuracies are TM-score values from the CAMEO website (https://cameo3d.org/).

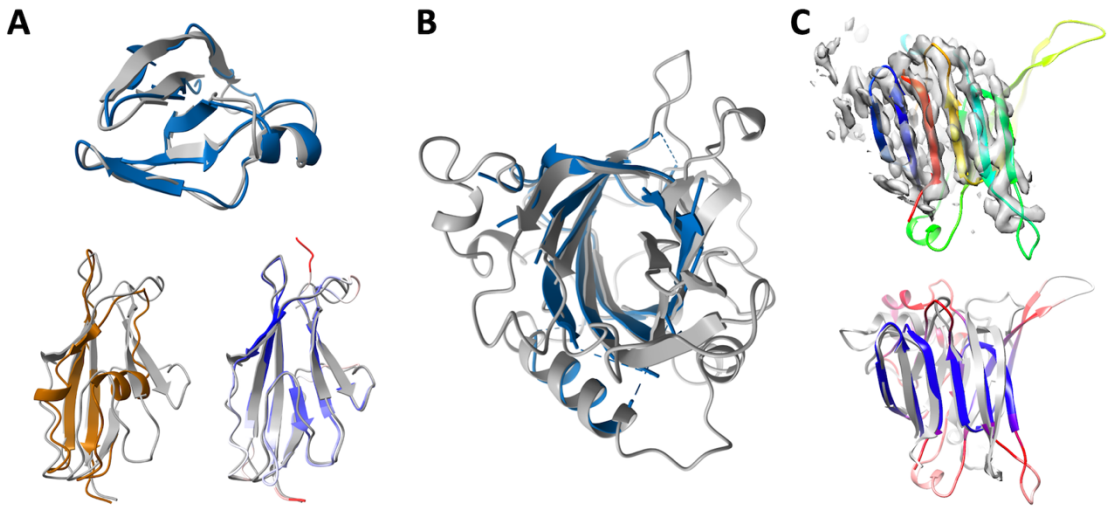

Fig. 2. Enabling experimental structure determination with RoseTTAFold. (A-B) Successful molecular replacement with RoseTTAFold models. **(A)** SLP. (top) C-terminal domain: comparison of final refined structure (gray) to RoseTTAFold model (blue); there are no homologs with known structure. (bottom) N-terminal domain: refined structure is in gray, and RoseTTAFold model is colored by the estimated RMS error (ranging from blue for 0.67 Å to red for 2 Å or greater). 95 C_α atoms of the RoseTTAFold model can be superimposed within 3 Å of C_α atoms in the final structure, yielding a C_α-RMSD of 0.98 Å. In contrast, only 54 C_α atoms of the closest template (4l3a, brown) can be superimposed (with a C_α-RMSD of 1.69 Å). **(B)** Refined structure of Lrbp (gray) with the closest RoseTTAFold model (blue) superimposed; residues having estimated RMS error greater than 1.3 Å are omitted (full model is in fig. S5C). **(C)** Cryo-EM structure determination of p101 $G_{\beta\gamma}$ binding domain (GBD) in a heterodimeric PI3K$_\gamma$ complex using RoseTTAFold. (top) RoseTTAFold models colored in a rainbow from the N-terminus (blue) to the C-terminus (red) have a consistent all-beta topology with a clear correspondence to the density map. (bottom) Comparison of the final refined structure to the RoseTTAFold model colored by predicted RMS error ranging from blue for 1.5 Å or less to red 3 Å or greater. The actual C_α-RMSD between the predicted structure and final refined structure is 3.0 Å over the beta-sheets. Figure prepared with ChimeraX (*35*).

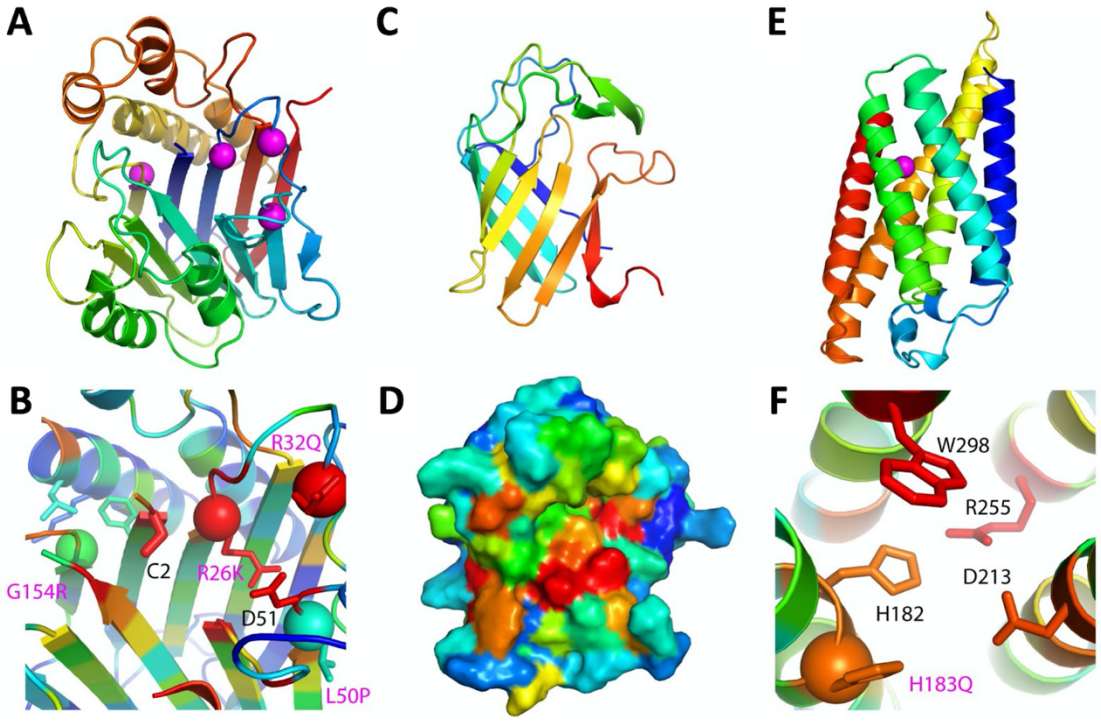

Fig. 3. RoseTTAFold models provide insights into function. (A) TANGO2 model, colored in a rainbow from the N-terminus (blue) to the C-terminus (red), adopts an Ntn hydrolase fold. Pathogenic mutation sites are in magenta spheres. **(B)** Predicted TANGO2 active site colored by ortholog conservation in rainbow scale from variable (blue) to conserved (red) with conserved residues in stick and labeled. Pathogenic mutations (spheres with wild-type side chains in the sticks) are labeled in magenta; select neighboring residues are depicted in the sticks. **(C)** ADAM33 prodomain adopts a lipocalin-like barrel shown in a rainbow from N-terminus (blue) to C-terminus (red). **(D)** ADAM33 model surface rendering colored by ortholog conservation from blue (variable) to red (conserved), highlighting a conserved surface patch. **(E)** CERS1 transmembrane structure prediction is colored from N-terminus (blue) to C-terminus (red), with a pathogenic mutation in TMH2 near a central cavity in magenta. **(F)** Zoom of CERS1 active site with residues colored by ortholog conservation from variable (blue) to conserved (red). Residues that contribute to catalysis (H182 and D213) or are conserved (W298 and D213) line the cavity. The conserved pathogenic mutation is adjacent to the active site.

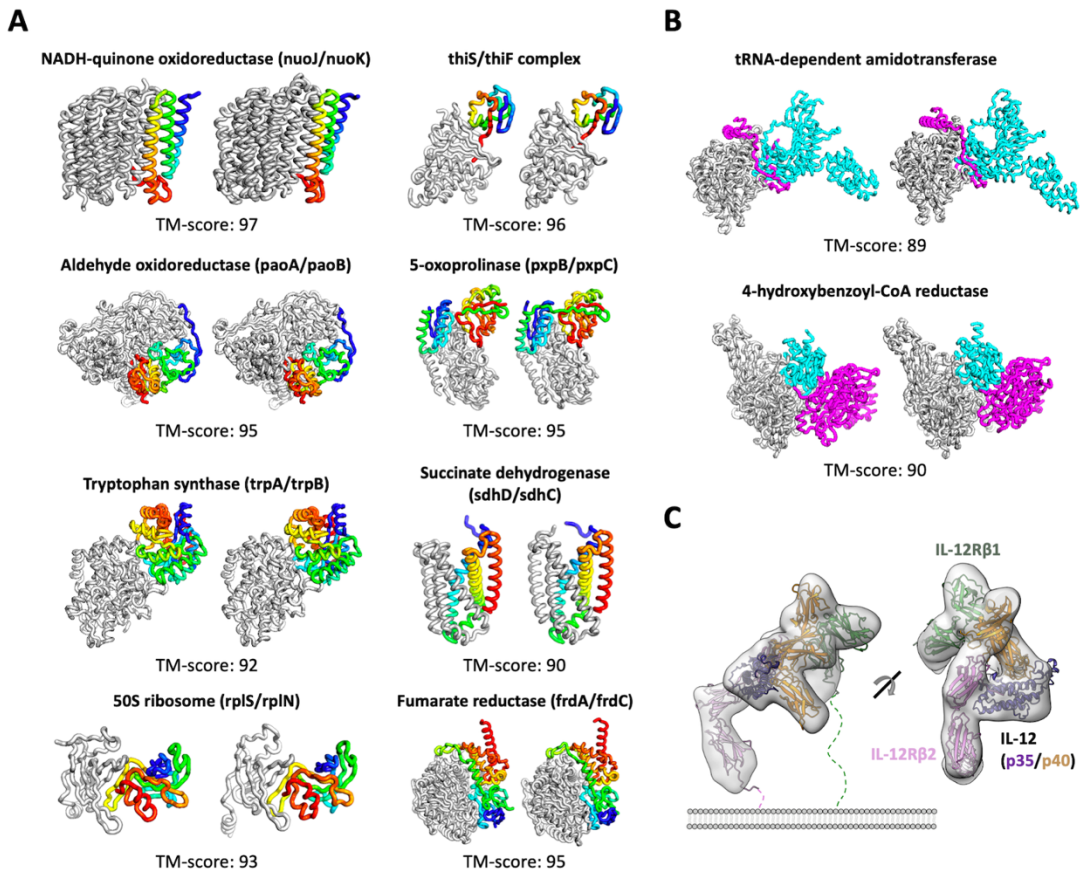

Fig. 4. Complex structure prediction using RoseTTAFold. (A, B) Prediction of structures of *E.coli* protein complexes from sequence information. Experimentally determined structures are on the left, RoseTTAFold models, on the right; the TM-scores below indicate the extent of structural similarity. **(A)** Two chain complexes. The first subunit is colored in gray, and the second subunit is colored in a rainbow from blue (N-terminal) to red (C-terminal). **(B)** Three chain complexes. Subunits are colored in gray, cyan, and magenta. **(C)** IL-12R/IL-12 complex structure generated by RoseTTAFold fits the previously published cryo-EM density (EMD-21645).

Science
AAAS

Supplementary Materials for

Accurate prediction of protein structures and interactions using a 3-track network

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K Rathinaswamy, Udit Dalwadi, Calvin K Yip, John E Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, David Baker*

Correspondence to: dabaker@uw.edu

This PDF file includes:

Materials and Methods
Figs. S1 to S17
Tables S1 to S4

Materials and Methods

Details of deep learning model

Initial Embedding

We describe the input multiple sequence alignments (MSA) as a matrix $x \in \mathbb{R}^{N \times L}$, where rows correspond to N sequences in the MSA, and columns are L positions in the aligned sequence. The input MSA is first tokenized to get the initial MSA features for further processing. Individual amino acids and gaps are regarded as character-level tokens (21 in total), and those are mapped to vectors having d_{msa} size through an embedding layer. The sinusoidal positional encoding (*37*) is added for residues in each sequence to let the network know the positional relationship. For the sequence dimension, an indicator for the query sequence instead of positional encoding is added because MSAs are unordered sets of sequences except the query sequence.

Template information is used to generate initial pair features by extracting pairwise distances and orientations from template structures for the aligned positions, along with 1D (positional similarity and alignment confidence scores) and scalar features (HHsearch probability, sequence similarity, and sequence identity) provided by HHsearch (*13*). Both features are concatenated to 2D inputs by tiling them along both axes of 2D inputs. Templates are first processed independently by one round of axial attention (row-wise attention followed by column-wise attention) (*38*) and then merged into a single 2D feature matrix using a pixel-wise attention mechanism. This processed feature matrix is then concatenated with the 2D-tiled query sequence embedding and projected to hidden dimension (d_{pair}) for pair features. The 2D sinusoidal positional encoding is also added.

Processing MSA features via self-attention

After embedding the input MSA as described in the previous section, each MSA update step has $\mathbb{R}^{N \times L \times d}$ features as input and output. The MSA features are processed by the axial attention approach (*38*) which alternates attention over rows and columns of the 2D features. To reduce memory usage, we used Performer architecture (*39*) for the column attention (attention over sequence dimension) that reduces the memory requirements from $O(LN^2)$ to $O(LN)$. We first compared this MSA encoder with a coevolution extractor (described in the next section) to the architecture with hand-crafted features (sequence profiles and inverse covariance matrices). As shown in Table S1 (architecture 1 vs 2), we found that having a learnable MSA encoder slightly improves distance and orientation prediction (Δloss=-0.07) as well as top L long-range contact accuracy (Δaccuracy=2%p).

For the row attention (attention over residue dimension), we tested two different attention methods: 1) un-tied attention and 2) softly tied attention inspired by MSA Transformer architecture (*40*). In MSA Transformer, the tied attention idea for residue-wise attention was first introduced because the homologous sequences in the MSA should have similar structures. Here, we modified this tied attention idea to reduce contributions from unaligned regions by introducing a learned position-wise weight factor (see Algorithm 1) to combine attention signals from sequences in MSA. We defined the soft-tied attention as Eq. (1), where N is the number of sequences in MSA, Q_n and K_n are the matrix of queries and keys for the n-th sequence of input, and W_n is the position-wise weight factor for the corresponding sequence.

$$attention = softmax(\sum_{n=1}^{N} W_n Q_n K_n^T) \quad \text{Eq. (1)}$$

In our experiments with small 2-track models, this soft-tied attention improves the top L long-range contact prediction performance by 2%p compared to the un-tied version (Table S1, architecture 6 vs 7). Interestingly, the soft-tied residue-wise attention maps showed correlations to the true contact map as shown in Fig. S12 (panel A and B). The final architecture used in RoseTTAFold is illustrated in Fig. S1A.

Algorithm 1. Position-wise weight factor calculation
Input: • Q: embedding of query sequence (batch,1,L,d_{msa}) • M: MSA embeddings (batch, N, L, d_{msa}) • H: the number of attention heads for subsequent tasks Get a query and key from given embeddings Query = Linear(d_{msa}, d_{msa})(Q) Key = Linear(d_{msa}, d_{msa})(M) Permute & reshape Query and Key to calculate cross-attention maps over sequence dimension Query = permute_and_reshape(Query) # (batch, L, H, 1, d_{msa}//H) Key = permute_and_reshape(Key) # (batch, L, H, N, d_{msa}//H) Calculate attention maps between Query and Key Attention = Query@Key.T # (batch, L, H, 1, N) Take softmax for the last dimension W = Softmax(Attention, dim=-1) Output: • W: positional weight for sequences

Update pair features with coevolution signal derived from MSA features

To extract residue pairwise interaction information from given MSA features, we adopted the outer product and aggregation idea from the CopulaNet method (*41*). The outer product can capture the correlation between two residues in each sequence. By aggregating the signals from all sequences in MSA, we can measure the strength of covariation. For example, in the simplest case with one-hot encoded embedding for sequences, we get a 21x21 substitution table for each pair of positions including gaps. When we take the average of the substitution tables from all sequences, the resulting 21x21 features will show different distributions depending on whether they interact with each other in 3D space or not. The broadly distributed 21x21 features indicate random uncorrelated mutations, and it means that those two residues are less likely to make contact in 3D space. On the other hand, if the aggregated features have sharp distributions (indicating correlated mutations), they will have a higher chance of interacting directly. In practice, the learned MSA embeddings through the network are used instead of one-hot encoding.

As outer products could require a huge memory ($O(d^2)$), the MSA embeddings are first projected down to the smaller hidden dimensions (32 features in this case) to reduce the memory requirements. After taking the outer product of embeddings derived from each sequence in MSA for any two residues, it calculates weighted averages of the outer products from all sequences with position-wise sequence weights. These aggregated coevolution features are then combined with 1D features (weighted average of MSA features) and residue-wise attention maps from the previous MSA update. They are projected down to match the hidden dimension for pair features.

To combine newly extracted pair features and previous pair features, we tested two different approaches: 1) Adding two pair features followed by feed-forward network and 2) concatenating two pair features followed by a single residual block of 2D convolutional network. As shown in Table S1 (architecture 5 vs 6), feature concatenation and 2D convolution clearly showed better performances, and we used this approach as outlined in Fig. S1B for our final model.

Refine pair features via row and column-wise self-attention

The updated pair features based on coevolution signals from MSAs are further refined by axial attention (*38*) as shown in Fig. S1C. Using axial attention instead of 2D convolution gave a clear improvement in inter-residue geometry predictions (Δloss=-0.35) with additional contact accuracy gain (Δaccuracy=2%p) even with single track architecture having sequential MSA and pair feature processing (Table S1, architecture 2 vs 3). This recapitulates one of DeepMind's observations that the attention mechanism is more suitable for protein structure prediction as it can directly learn the relationship between two residues distant in sequence.

In addition to the axial attention, we used Performer architecture (*39*) for the attention algorithm to further reduce memory usage so that the larger architecture could fit on the GPU for experiments as larger architecture showed better performance (Table S1, architecture 7 vs 8).

Update MSA features based on structure information encoded in pair features

The most distinctive feature of AlphaFold2 architecture is that MSA features are updated based on pairwise features. We experimented with two different ways to update MSAs based on given pair features: 1) taking cross-attention (or encoder-decoder attention) (*37*) between MSA and pair features by considering pair to MSA updates as a kind of encode-and-decode process and 2) applying attention maps derived from pair features to MSA features (named direct-attention here) so that MSA features can be updated by attending positions close in 3D space that encoded in pairwise features. As shown in Table S1 (architecture 4 vs 5), direct-attention showed clearly better performance (Δloss=-0.4, Δcontact accuracy=4%p). The attention maps derived from pairwise features showed a good agreement with the true contact map (Fig. S12, panel A and C). The final architecture based on direct-attention is outlined in Fig. S1D.

Initial 3D structure prediction

We employed Graph Transformer-based architecture (*42*) (shown in Fig. S1E) to generate initial backbone coordinates for the 3D track (structure track). The input is defined as a fully connected graph with nodes representing the residues in the protein. The node and edge embeddings are

learned from the averaged MSA features combined with a one-hot encoded query sequence and the pair features along with sequence separation, respectively. The backbone coordinates are estimated using a stack of four Graph Transformer layers followed by a simple linear transformation to predict Cartesian coordinates of N, C_α, C atoms for each residue node.

Structure updates through SE(3)-Transformer

SE(3)-Transformer (*6*) is employed to refine given 3D coordinates based on updated MSA and pair features in the 3-track model (Fig. S1F). The protein graph is defined with nodes representing C_α atoms, and each node is connected to the K-nearest neighbors. The positions of N and C atoms are encoded by including displacement vectors to the corresponding C_α atoms as the degree 1 node features (vector node features). The node embeddings derived from averaged MSA features and the one-hot encoded query sequence are used as degree 0 node features (scalar node features). Pair features corresponding to the graph edges are also included as input features for SE(3)-Transformer. SE(3)-Transformer predicts shifts of C_α atoms and new displacement vectors for N and C atoms to the updated C_α positions. It also gives degree 0 node features (called state features here) that are used to calculate attention maps for structure-based MSA updates described in the next section.

Update MSA features based on a 3D structure

Similar to the MSA updates based on pair features in the 2-track model, MSA features are updated based on attention maps derived from the current 3D structures. Four attention maps are calculated based on the state features, and they are masked based on the C_α distances with four different cutoffs (8, 12, 16, and 20 Å) so that it only attends to the neighbors in 3D space. The same attention maps are applied to all the sequences in the MSA. A pointwise feed-forward layer further processes the outputs from the masked multi-head attention. The entire process is outlined in Fig. S1G.

Definition of 2-track and 3-track feature processing blocks

We defined 2-track blocks with four arrows in Fig. 1A (orange box). It first updates MSA features through self-attention, extracts coevolution features from MSA, and combines them with the previous pair features. Pair features are further optimized by axial attention, and MSA features are updated based on the structural information encoded in the current pair features. For 3-track blocks (blue box in Fig. 1A), we found that the order of communication between tracks is important. We experimented with two different ways to communicate 1D, 2D, and 3D tracks: updating structures before and after synchronizing MSA and pair features as shown in Fig. S13. The 3D coordinate updates based on synchronized MSA and pair features showed clearly better performance (Table S1, architecture 10 vs 11).

Residue pairwise distance and orientation prediction

The inter-residue geometry representations (shown in Fig. S14) are predicted through a single residual block consisting of two 2D convolution layers with 3x3 filters followed by convolution

with 1x1 filters and softmax activation. Since maps for C_β-C_β distances and dihedral angles along pseudo C_β-C_β bonds are symmetric, we enforce symmetry in the network by using averages of transposed and untransposed feature maps as inputs for those predictions.

Additional structure module for iterative refinement through the network

Although structures are explicitly sampled in 3-track blocks, an additional structure module is introduced to build a model based on combined 1D features and 2D inter-residue geometry predictions for inference with multiple discontinuous crops. Initial coordinates for backbone N, C_α, C atoms are generated using simple graph-based architecture (see Initial 3D structure prediction section above) with node and edge features derived from averaged MSA features and 2D distance and orientation distributions. These coordinates are further refined with multiple SE(3)-Transformer layers (*6*) by taking the same node and edge features used to generate initial coordinates. At the end of SE(3)-Transformer layers, the residue-wise C_α-lDDT (*43*) is also estimated based on the degree 0 features from the final SE(3)-Transformer layer.

We didn't use any iteration during the training, and the parameters were optimized through a single pass of the network. However, we found that we could use this structure module as an iterative refinement tool by feeding the output coordinates of the final SE(3)-Transformer layer to the first SE(3) layer as inputs at inference time (Fig. S15). The predicted C_α-lDDT is used as a scoring function to decide when to stop the iteration and select the final model from all the sampled structures.

Comparison between 2-track end-to-end model and 3-track model

AlphaFold2 passed information from the 2-track trunk model into a 3D equivariant network operating on 3D coordinates directly. AlphaFold2 also employed end-to-end training, updating all model parameters by backpropagation from a loss function computed on 3D coordinates after many SE(3)-equivariant layers. As an experiment, we built a model with SE(3)-Transformer layers on top of the graph-based initial coordinate generation following the 2-track model. We found that adding SE(3)-Transformer layers improved the accuracy of structures generated by the simple graph-based network (Fig. S16), but this 2-track end-to-end model was not as good as the 3-track end-to-end model (Table S1, architecture 9 vs 12).

Training details

The extended trRosetta training set (containing 22,922 clusters with sequence identity cutoff 30%, 208,659 protein chains released in the PDB as of 02/17/2020) was used to train RoseTTAFold. We cycled through all sequence clusters every training epoch by picking a random protein chain from each cluster. For each selected protein chain, a subsampled MSA (having maximum $NxL=2^{14}$ tokens) and up to 10 randomly selected templates were used to augment training data. During training, protein chains over 260 residues in length were cropped to fit into GPU memory.

The loss function used to train the model consists of 1) distance and orientation prediction loss (cross entropy) with 0.5 Å and 10° bins, 2) coordinate and distance RMSD of predicted coordinates, and 3) mean squared error of predicted C_α-lDDT score. During training,

weights for coordinate and distance RMSD were ramped up from 0.05 to 0.2. For the other loss terms, weights are set to 1.0.

We train 130M parameters models having eight 2-track blocks and five 3-track blocks. Using eight 32GB V100 GPUs, it took about four weeks to train the model up to 200 epochs. The following hyper-parameters were used:

- MSA, pair, template embedding size: 384, 288, and 64, respectively
- The number of attention heads for self-attention on MSA, pair, and template: 12, 8, and 4
- The number of attention heads for MSA updates based on pair features: 4
- Size of node and edge features for initial coordinate generation: 64
- The number of attention heads for initial coordinate generation: 4
- Size of input node and edge features for SE(3)-Transformer: 32
- SE(3)-Transformer architecture: 2 layers with 16 channels, 4 attention heads, and up to representation degree 1 (l=0 and 1 features were used)
- The number of closest residues to define graph for SE(3)-Transformer: 128 for first two 3-track blocks, 64 for last three 3-track blocks
- Learning rate: 0.0005 with linear learning rate decay after 16000 warm up steps
- Effective batch size: 64 in total (8 V100 GPUs, single training example per GPU, 8 gradient accumulation steps)
- Weight decay: 0.0001

RoseTTAFold modeling pipeline

We built a fully automated modeling pipeline based on RoseTTAFold. It first iteratively searches homologous sequences against UniRef30 (*44*) and BFD (*45*) sequence databases using HHblits (*13*). The E-value cutoff for sequence search is gradually relaxed until the resulting MSA has at least 2000 sequences with 75% coverage or 5000 sequences with 50% coverage (both at 90% sequence identity cutoff). The generated MSA is used to perform template searches against the PDB100 database with HHsearch (*13*).

With MSA and top 10 templates as input, the RoseTTAFold network predicts inter-residue geometries (probability distributions of 6D coordinates described in Fig. S14) for many 300×300 discontinuous crops (150 residues per each segment) and combined them by taking weighted averages based on predicted C_α-lDDT values. We used two different strategies to generate final structure model with this combined 6D coordinate distribution: 1) gradient-based folding using pyRosetta (*5*) script and 2) a structure module based on SE(3)-Transformer architecture described above (see *Additional structure module for iterative refinement through the network* section). The first method doesn't require a large memory GPU as it predicts 300×300 sizes of 6D coordinates only and gives a full-atom model at the end, but it requires more CPU cores and time to run multiple trajectories (15 in total) of gradient-based folding from scratch. The second method can model backbone coordinates much faster than gradient-based folding (with a similar accuracy level), but it requires a large memory GPU (e.g. TITAN RTX) for proteins having more than 400 residues.

For the pyRosetta-based modeling protocol, the five models out of 15 sampled structures are selected based on predicted lDDT of DeepAccNet (*12*) after clustering. The C_α RMS error is estimated by converting predicted non-local C_α-lDDT (only considering residue pairs having sequence separation > 12) using Eq. (2). This pyRosetta-based protocol is implemented in the Robetta server.

$$C_\alpha\ RMS\ error = 1.5e^{4\times(0.7-lDDT)} \quad \text{Eq. (2)}$$

Both pyRosetta and end-to-end versions are available at https://github.com/RosettaCommons/RoseTTAFold. The following tutorial shows how to install and run the RoseTTAFold method.

Tutorial. How to install and use the RoseTTAFold method to predict protein structures

Installation

1. Clone the package
   ```
   git clone https://github.com/RosettaCommons/RoseTTAFold
   cd RoseTTAFold
   ```
2. Create conda environments using `RoseTTAFold-linux.yml` file and `folding-linux.yml` file. The latter is required to run the pyRosetta version only (run_pyrosetta_ver.sh).
   ```
   conda env create -f RoseTTAFold-linux.yml
   conda env create -f folding-linux.yml
   ```
3. Download network weights (under Rosetta-DL Software license -- please see below) While the code is licensed under the MIT License, the trained weights and data for RoseTTAFold are made available for non-commercial use only under the terms of the Rosetta-DL Software license. You can find details at https://files.ipd.uw.edu/pub/RoseTTAFold/Rosetta-DL_LICENSE.txt
   ```
   wget https://files.ipd.uw.edu/pub/RoseTTAFold/weights.tar.gz
   tar xfz weights.tar.gz
   ```
4. Download and install third-party software if you want to run the entire modeling script (run_pyrosetta_ver.sh)
   ```
   ./install_dependencies.sh
   ```
5. Download sequence and structure databases (UniRef30, BFD, and pdb100)
   ```
   # uniref30 [46G]
   wget
http://wwwuser.gwdg.de/~compbiol/uniclust/2020_06/UniRef30_2020_06_hhsuite.tar.gz
   mkdir -p UniRef30_2020_06
   tar xfz UniRef30_2020_06_hhsuite.tar.gz -C ./UniRef30_2020_06

   # BFD [272G]
   wget
https://bfd.mmseqs.com/bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_
   ```

```
opt.tar.gz
      mkdir -p bfd
      tar xfz
bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt.tar.gz -C ./bfd

      # structure templates [10G]
      wget https://files.ipd.uw.edu/pub/RoseTTAFold/pdb100_2021Mar03.tar.gz
      tar xfz pdb100_2021Mar03.tar.gz
```

6. Obtain a PyRosetta licence and install the package in the newly created `folding` conda environment (only for pyRosetta version).

Usage

```
cd example
../run_pyrosetta_ver.sh input.fa .  # running pyrosetta version
../run_e2e_ver.sh input.fa .        # running end-to-end version
```

Expected outputs

For the pyRosetta version, users will get five final models having estimated CA rms error at the B-factor column (model/model_[1-5].crderr.pdb). For the end-to-end version, there will be a single PDB output with estimated residue-wise CA-lDDT at the B-factor column (t000_.e2e.pdb).

Molecular replacement calculations

Structure of glycine N-acyltransferase

The structure of glycine N-acyltransferase (GLYAT) from *Bos taurus* had evaded numerous attempts at solution, despite the availability of excellent data from three crystal forms. Structures of homologues were found using HHpred (*46*), which revealed that the only known structures were from distant relatives, almost all with low coverage of the target. Only 3 homologues (including the top hit) had greater than 60% coverage; these were only 12% identical in sequence. The top 5 hits were prepared for molecular replacement trials by pruning non-conserved side chains and loops using phenix.sculptor (*47*). In addition, an ensemble model was prepared by superimposing the individual homologues in phenix.ensembler (*48*) and trimming parts of the ensemble that are poorly conserved to leave a small conserved core. Molecular replacement trials with Phaser (*49*), MoRDa (*50*) and I-TASSER-MR (*51*) on all three crystal forms, using individual models, ensemble models and domain models, failed to yield any convincing results. Models made with trRosetta (*3*) also failed in MR calculations with Phaser.

In contrast, molecular replacement was straightforward for all three crystal forms when using the RoseTTAFold models, whether as individual models or trimmed ensembles. An estimate of the effective RMS error is required to calibrate the likelihood target, and a value of 1.2 Å was used for these models.

A post mortem analysis was carried out to verify that model quality was the limiting factor for molecular replacement with models derived from the PDB. This analysis concentrated on a tetragonal crystal form, which diffracts to 1.5 Å resolution and has a single copy in the asymmetric unit. The other two crystal forms each have two copies of the protein in the asymmetric unit.

In the likelihood-based molecular replacement algorithm implemented in Phaser, the log-likelihood-gain (LLG) score is an excellent predictor of success. If LLG scores of 60 or more are achieved in placing a single copy, the solution is almost always correct (*52*). In contrast, scores below 30 are more likely to correspond to random incorrect placements. By correctly positioning a molecular replacement model and carrying out a rigid-body refinement in Phaser, we can evaluate the score that could have been achieved in the search. This calculation shows that none of the available models came close to providing sufficient signal to solve the structure, giving LLG scores of only 7 to 11 when correctly placed. The best model (with a score of 11) was the top hit in HHpred, PDB entry 1sqh. A full molecular replacement search with this model yielded a top LLG score of 22 for incorrect placement. The high quality of the RoseTTAFold model, especially compared to the model derived from 1sqh, can be seen in Fig. S5A. For this figure, the experimental structure is illustrated using chain A from the current model of the hexagonal crystal form, in which the poorly ordered loop is most clearly defined. Table S2 summarizes the refinement statistics for this structure, as well as other crystal structures discussed below.

Value added by coordinate error estimates for GLYAT structure determination

LLG scores obtained with the RoseTTAFold models were compared, either ignoring the estimates provided for the RMS error of each amino acid or using it to weight each atom's contribution by providing a B-factor equal to $(8\pi^2/3)RMS^2$ (*53*). Before applying the B-factor weighting, the LLG scores ranged from 88 to 148 for the 5 alternative models. After applying the weighting, the LLG scores ranged between 117 and 188. Similarly, the LLG score for the trimmed ensemble model increased from 191 to 244. In a more marginal case, such weighting could well be pivotal to success. Fig. S5A illustrates the correlation between predicted and actual errors in the RoseTTAFold model, especially in the poorly ordered loop which has the highest predicted errors.

Structure of a bacterial oxidoreductase

The structure of an oxidoreductase from a bacterial source wasn't solved by molecular replacement using related structures available from the PDB, identified using HHpred (*46*). These efforts were likely unsuccessful because available structures had low sequence identity and only moderate sequence coverage - the best structures had an identity of ~33% for the first 40% of the sequence, or ~25% identity for the first 60% of the sequence. In addition, the 2 crystal forms were expected to have 6 or 12 molecules in the crystallographic asymmetric unit based on the most probable solvent content. The top 5 HHpred structures were prepared for molecular replacement trials by pruning non-conserved side chains and loops using phenix.sculptor (*47*). In addition, an ensemble model was prepared by superimposing the individual homologues in phenix.ensembler (*48*) and trimming parts of the ensemble that are poorly conserved to leave a small conserved core. Molecular replacement trials with Phaser (*49*) did not produce correct solutions as judged by significant overlaps between placed molecules, and a modest TFZ score of 7.4 in the lower probability P2 space group.

The top 5 RoseTTAFold models were superimposed using phenix.ensembler and parts of the ensemble that are poorly conserved were automatically trimmed. Atomic B-factors were calculated from the estimated RMS error as described above. Molecular replacement trials with Phaser produced a solution in the more likely $P2_1$ space group, albeit with a modest TFZ score of 6.9. Manual inspection of the solution revealed that 4 of the molecules formed 2 dimers, which were expected on the basis of the closest homologue structures and biophysical data. One dimer was extracted from the model and used in a new MR trial, which produced a very clear solution with a TFZ of 17.2. Comparison of the 2 molecular placement trials showed that the initial search had placed 5 molecules correctly but the 6th incorrectly. The successful dimer-based solution was used as the starting point for phase improvement using statistical density modification methods (*54*) in Phenix (*55*). The resulting map showed unambiguous density for the protein including many regions where the search model was locally different from the true structure. The structure could be completed by the application of automated model building methods in phenix.phase_and_build and phenix.autosol (*56*), followed by manual model rebuilding in coot (*57*) in combination with refinement in phenix.refine (*58*) and validation with MolProbity (*59*).

Structure of bacterial surface layer protein (SLP)

Excellent diffraction data were available for SLP, but a search for homologues in the PDB using HHpred (*46*) yielded only one hit at a low significance level (E-value of 6.1, sequence identity of 19%) covering only 38% of the protein sequence. Considering that the crystal contains 4 copies of SLP in the asymmetric unit, it was not surprising that molecular replacement attempts failed before the RoseTTAFold models were available.

Initial attempts to solve the structure using an ensemble made from models of the entire protein were partially successful but failed because of crystal packing clashes. However, when the models were divided into two domains, searches with four copies of an ensemble model for the N-terminal domain gave a clear solution with good signal. This turned out to be sufficient to complete the structure if weak phase information from a mercury derivative was added by MR-SAD (*60*). Alternatively, the structure could be solved purely by molecular replacement, by adding four copies of an ensemble model for the C-terminal domain, in which B-factors were computed from the estimated RMS errors and residues with a predicted error greater than 1.3 Å were removed. Automated building procedures were sufficient to complete the structure from this point. As a control, further molecular replacement calculations were carried out using models obtained with trRosetta (*3*), IntFOLD6 (*8*), RaptorX (*61*), I-TASSER (*62*) and QUARK (*63*), but none of these succeeded.

Structure of secreted fungal protein Lrbp

Diffraction data were available to 1.53 Å resolution, but no significant hits were found in a search for homologues in the PDB using HHpred (*46*) as the top hit had an E-value of 110. Attempts over the course of 4 years to solve this structure, using a variety of predicted models and small fragments of regular secondary structure had failed.

The initial MR searches using RoseTTAFold models prepared with the default protocol also failed. However, the diversity of the models was increased by varying the selection criteria for the MSA, and the estimated RMS errors were used to delete residues with errors estimated to

be greater than 1.3 Å. To generate more diverse models, we collected 8 different MSAs with E-value cutoff of 1e-40, 1e-30, 1e-20, and 1e-10 and sequence coverage cutoff of 50% and 75%. With this strategy, clear solutions for the two copies in the asymmetric unit emerged, leading to a high quality model. As seen in Fig. S5C, the error estimates give a reliable indication of where confidence should be placed in the model.

Modeling of GPCR structures

GPCR modeling benchmark set construction and evaluation

A benchmark set of 27 GPCR sequences with experimentally determined structures that were not included in the RoseTTAFold training set was constructed. X-ray and cryo-EM structures determined with resolution higher than 4 Å were excluded. Annotations in the GPCRdb (*14*) were used to classify GPCR sequences, structures, active states, and the transmembrane region residues for analyses. All predicted models were evaluated for the transmembrane regions only. The reference experimental structures were also truncated to the corresponding transmembrane regions, and the TM-score software (*33*) was used to calculate Cα-RMSD of the models. To check if templates with similar sequences were available, the sequence identities between the target transmembrane region sequence and the aligned sequences were re-calculated. From the HHblits template search, results with e-value less than 1e-10 were considered, if they were found. The highest sequence identity among the alignments that have transmembrane region coverage higher than 80% was used for analysis. The estimated model accuracy (DAN-lDDT) was predicted by applying the DeepAccNet (*12*) on each truncated model.

Modeling active and inactive states of GPCRs

For each target sequence, active and inactive state GPCR template sets were separately provided to two parallel predictions, each generating the corresponding state models. When a template structure in a certain state was not available, models were not predicted for that state. For the benchmark test, templates with sequence identities higher than 70% from HHsearch (*13*) results were excluded to construct the test more fairly.

GPCR benchmark test performance

Models with highest estimated accuracy values (DAN-lDDT) were selected for each active and inactive state. RoseTTAFold could predict highly accurate models of both active and inactive states. Examples of good predictions are shown in Fig. S7 panel A and B.

Template-based models of the benchmark set targets were collected from available GPCR model databases. Active state models were brought from GPCRdb (*14*) and inactive state models were downloaded from the Meiler group modeling database (*15*). Targets that could have been modeled easily using any template with sequence identity > 70% in the same state were excluded for analysis. The accuracies of the RoseTTAFold model and corresponding homology model are compared in Fig. S7C. For most of the targets, RoseTTAFold could predict higher TM-score structures.

The best template sequence identity values for each GPCR sequence are reported with estimated model accuracy (DAN-lDDT) and actual accuracy in Fig. S7D. When multiple

reference experimental structures existed for the corresponding state, the best Cα-RMSD was reported with color representing model accuracy. RoseTTAFold prediction results on the GPCR benchmark set didn't have a high correlation with the best template sequence identity. This again corroborates that the deep-learned network of RoseTTAFold can predict models with accuracies beyond that which can be achieved only with homology information. However, generating highly accurate active state models (Cα-RMSD < 1.5 Å) was more feasible when templates with higher sequence identities were available.

The DAN-lDDT of 0.80 can roughly be used as a threshold to discriminate between accurate (Cα-RMSD < 1.5 Å or TM-score > 0.9, Fig. S7D) and inaccurate models. Using this guideline to estimate model accuracy could be better applied to inactive state models (Fig. S7D). The active state models turned out to have lower DAN-lDDT than their actual accuracy. The DeepAccNet was trained on monomeric structures only, and the receptor chain in an active state, which would require other chains such as G-proteins as interacting partners, could have been underestimated.

GPCR models of unknown states

In the GPCR benchmark set we constructed, 25 targets (as of May 14th, 2021) didn't have known structures of one state, either inactive or active. We predicted models of the unknown state for each target, and models with DAN-lDDT higher than 0.75 were achieved for all targets. These models are provided in http://files.ipd.uw.edu/pub/RoseTTAFold/GPCR_benchmark_one_state_unknown_models.tar.gz

Human GPCR model generation

We collected a set of 298 human GPCR sequences without known experimental structures as of May 14th, 2021. Models both in active and inactive states were predicted by applying RoseTTAFold. The best template sequence identity and the estimated accuracy (DAN-lDDT) of the models are reported in Fig. S7D. All models with DAN-lDDT values higher than 0.75 are provided in http://files.ipd.uw.edu/pub/RoseTTAFold/all_human_GPCR_unknown_models.tar.gz. The DAN-lDDT metric can be used to estimate the reliability of each model, and the relative per-residue quality estimation information can be found in the B-factor column.

Modeling of structurally uncharacterized domains from human proteins

We selected human proteins of biomedical importance based on the number (>50) of literature that are linked to them in Uniprot (*64*) and whether mutations in them are known to cause human diseases according to the DBSAV database (*65*). 7,639 human proteins were selected and domains were predicted using the HMMER (*66*) search against the Pfam database (*67*). A total of 18,233 domains were detected (e-value < 1e-5) in these proteins. The majority of these domains can be modeled confidently by homologous structure in PDB (*68*), and out of the structurally uncharacterized domains, over half of them include a considerably large (> 25%) fraction of residues that are predicted to be disordered (*69*). Excluding domains that are disordered or can be modeled by homology, we removed redundancy, i.e. domains that were mapped to the same Pfam, in the remaining 2,083 domains, resulting in 693 targets to model

with our method. We obtained high-quality (estimated lDDT with DeepAccNet (*12*) > 0.8) models for 245 targets (provided in http://files.ipd.uw.edu/pub/RoseTTAFold/human_prot.tar.gz). Only 28 out of 693 targets have predicted lDDT lower than 0.5, and half of them are turned out to be potential disordered proteins (predicted by SPOT-Disorder2 (*70*)). For the rest of the targets (420 targets having predicted lDDT between 0.5 and 0.8), it failed to predict high-accuracy structures due to the several factors, including 1) local inaccuracies come from the local regions that might be stabilized by interactions to its binding partner (other proteins or nucleic acids), 2) having disordered local regions, or 3) limitations of the method itself. The 245 high-quality models were manually inspected to reveal biological insights with the help of literature, sequence conservation, and remote homology that can be detected by searching structurally similar proteins.

For three RoseTTAFold structure models that provided insight into their biological function, their sequences (Q6ICL3:1-259 for TANGO2, P27544:98-304 for CERS1, and Q9BZ11:39-167 for ADAM33 prodomain) were checked against the SWISS-MODEL repository (*68*) for homology models. Their sequences were also submitted to the HHpred server (*71*) for search against the PDB database (PDB_mmCIF70_17_May) and the ECOD (*72*) domain database (ECOD_F70_20200717) using default parameters. For the CERS1 example, where no confident hits were identified, a second MSA generation method using PSI-BLAST against nr70 was used to identify possible template homologs. HHpred results are summarized in Table S3, omitting hits below rank 5. To identify related folds for the examples, RoseTTAFold models were used as queries to search the ECOD database with RUPEE (default settings) (*73*). Potential functional sites for the models were mapped with AL2CO (*74*) using conservations from multiple sequence alignment (MAFFT, default settings (*75*)) of orthologs collected from the OMA database (*76*).

The SWISS-MODEL repository could only generate low-quality models for the TANGO2 sequence. However, HHpred generated alignments for several Ntn templates with high confidence (Table S3). We chose the top two templates (3gvz and 2x1d) to generate homology models using the SWISS-MODEL workspace alignment mode (*68*). Each of the homology models was of poor quality based on QMEAN scores (*77*) (-6.12 and -6.11, respectively). These homology models were compared to the RoseTTAFold structure using pairwise DaliLite (*78*) superpositions (DaliZ 19.1 and 17.9, respectively). Compared to the RoseTTAFold structure (Fig. S9A), each of the homology models displays shifts in alignment and relatively poorly structured loops. Some of the conserved residues that form the RoseTTAFold active site (Fig. S9A, colored red) shifts further away from the active site in each of the homology models: R86, G87, and K166 in the 3gvz model (Fig. S9B) and G49, G51, and K166 in the 2x1d model (Fig. S9C).

Template search for the ADAM33 prodomain confidently identified an incorrect template (4on1_B) corresponding to a fragilysin-3 prodomain fold. While each of the structures possesses a similar four-stranded beta-meander, the alpha + beta C-terminus of the fragilysin prodomain extends the beta-meander into a longer sheet (Fig. S17A). Alternately, the N-terminus of ADAM33 continues the beta-meander to form a beta-barrel fold similar to that of lipocalin (Fig. S17B). The HHpred alignment for the fragilysin template incorrectly extended the metalloprotease domain present in both ADAM33 and fragilysin into the prodomain (aligned portions of the prodomains in a rainbow, Fig. S17). HHpred search with the ADAM33

prodomain sequence of templates from the ECOD domain database, which separates the prodomain from the metalloprotease, avoids this multi-domain problem.

Hetero-complex structure prediction using RoseTTAFold

Despite RoseTTAFold being trained on single protein chains, we deployed its ability to make inferences on discontinuous sequence segments to the hetero-oligomer complexes. The only modification we introduced for hetero-complex structure prediction was a change in the positional encoding. We added 200 to the residue numbers of the following subunits to let the network know that it has chain breaks between each subunit.

As a benchmark, we predicted the hetero-oligomer structures of *E. coli.* proteins from the PDB benchmark set (*32*). Among 868 pairs in the PDB benchmark set, we selected 68 interaction pairs having known complex structures of identical or close homologous proteins (sequence identity > 90%) in the PDB and having interface area (calculated by naccess (*79*)) larger than 1,500 $Å^2$. The list of 68 interaction pairs and the accuracy of predicted complex models are provided in Table S4. The complex model accuracy is evaluated based on the Interface Contact Similarity (ICS) score (*80*) and complex TM-score (*33*). To see whether RoseTTAFold can predict higher-order oligomer structures, we also tried to predict hetero-trimer complex structures of bacterial proteins shown in Fig. 4B. For both cases (dimer and trimer prediction), the prediction was made based on a paired alignment of the target complex without any template information.

We generated paired alignment for human IL-12R/IL-12 complex structure prediction by simply pairing the sequences with the same taxonomy ID. Based on the paired sequence alignments and the template structure (IL-23R/IL-23 complex structure; PDB 6wdq), the backbone coordinates were predicted using RoseTTAFold. The full-atom structures were generated by FastRelax (*81*) with restraints derived from predicted distances and orientations.

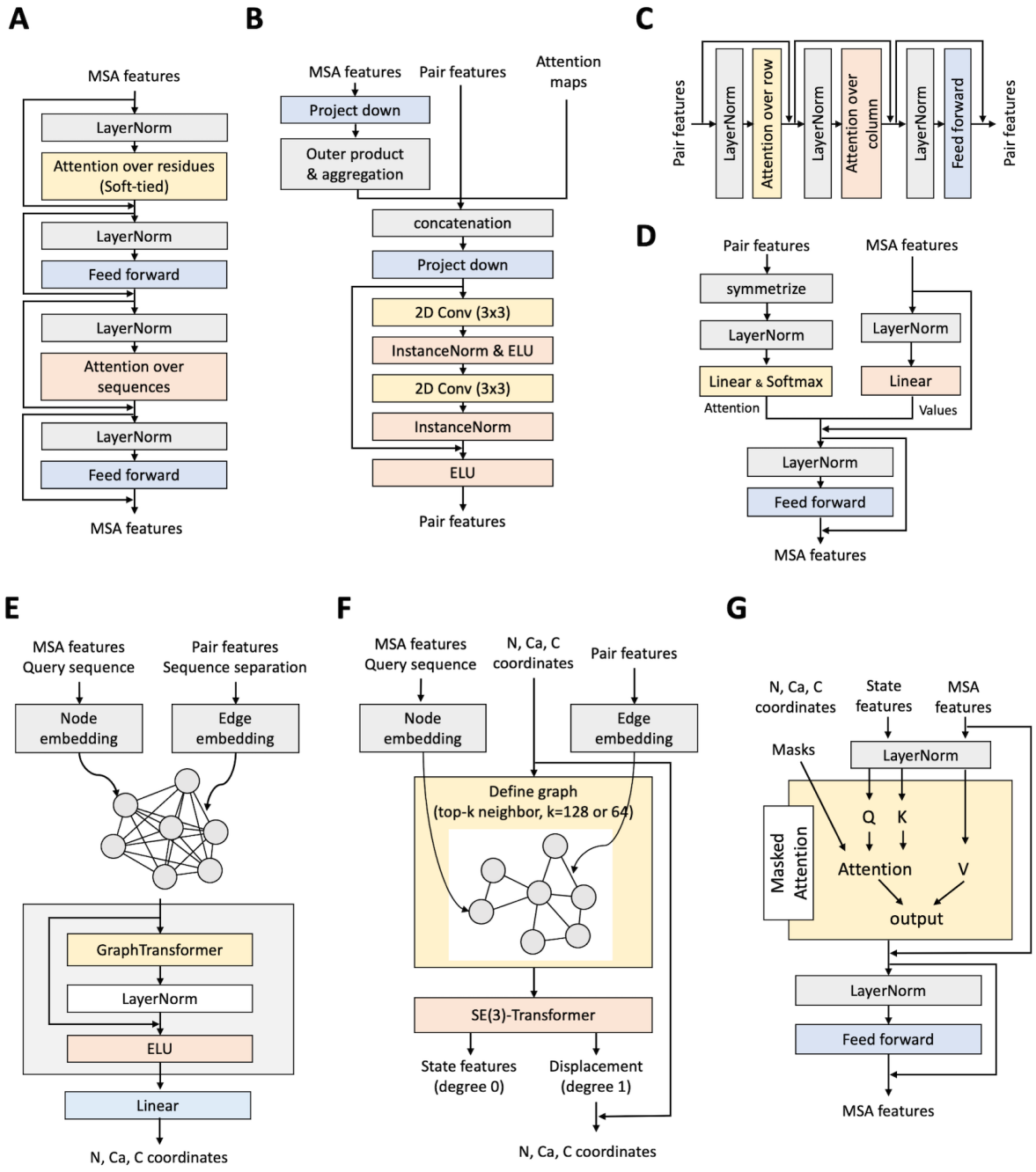

Fig. S1. Detailed architecture of each component of RoseTTAFold. (A) MSA updates via self-attention on MSA features. The attention maps over residues are softly tied. **(B)** Pair feature updates based on co-evolution signals derived from MSA features by taking outer-products and weighted averages. **(C)** Pair feature refinement through axial attention. **(D)** MSA feature updates based on attention maps derived from given pair features. **(E)** Initial N, C_α, C coordinate generation using Graph Transformer architecture. **(F)** 3D coordinate refinements with SE(3)-Transformer. **(G)** MSA feature updates based on given 3D structures using masked attention maps.

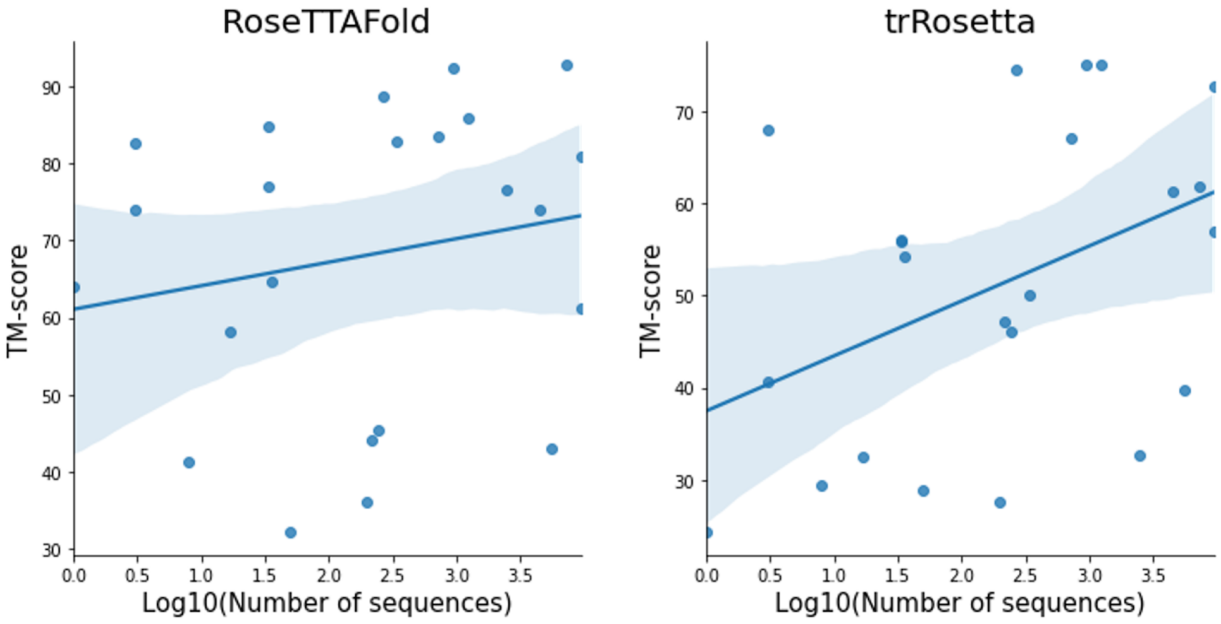

Fig. S2. A correlation between the number of sequences in multiple sequence alignments (MSA) and model accuracy. RoseTTAFold shows a weaker correlation compared to trRosetta.

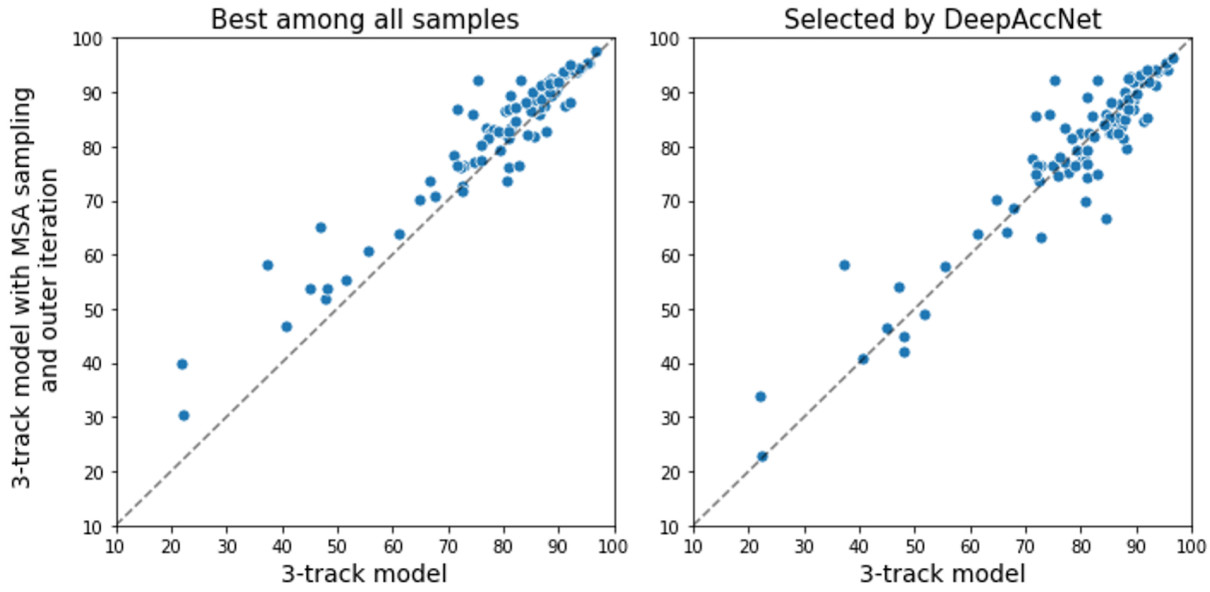

Fig. S3. Model accuracy changes upon an intensive use of the network for inference. By sampling MSAs randomly and providing predicted structures as templates (y-axis), the 3-track end-to-end model was able to sample much better model structures than the single-pass (x-axis) as shown in the left panel. DeepAccNet was able to select improved structures for some cases (right), but there is still room for improvement in model accuracy estimation. The model accuracy is measured by TM-score.

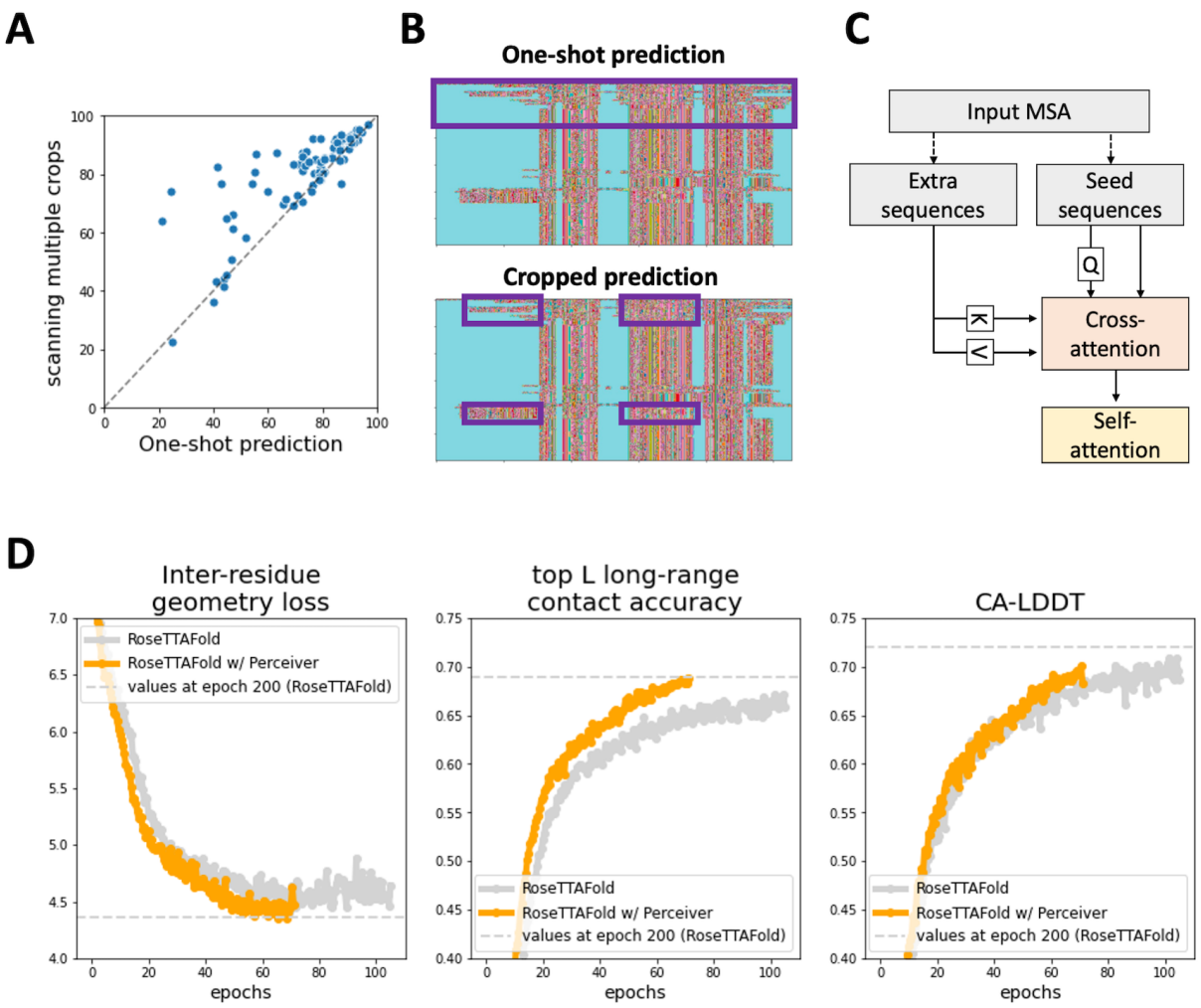

Fig. S4. Experiments on network architecture modification with Perceiver for efficient MSA encoder. (A) Model accuracy comparison (in terms of TM-score) between predicting the entire structures (one-shot prediction) and combining predictions from multiple discontinuous crops. Scanning multiple crops generated more accurate predictions. **(B)** Differences in the subset of sequences used for one-shot and cropped prediction. Due to the memory limitation, only up to 1,000 sequences were used during the prediction. For the one-shot prediction, the top 1,000 sequences were selected, while 1,000 sequences having sequence coverage over 50% were selected for the cropped prediction. **(C)** A new MSA update process based on Perceiver architecture. It keeps accessing the extra sequences having richer information at every iteration and extracting meaningful information through cross-attention. **(D)** Training curve of RoseTTAFold model and the model with Perceiver architecture. The inter-residue geometry loss, top L long-range contact accuracy, and CA-LDDT for validation set are shown. The horizontal dashed line colored in gray showed the value of each metric at the last epoch (epoch 200).

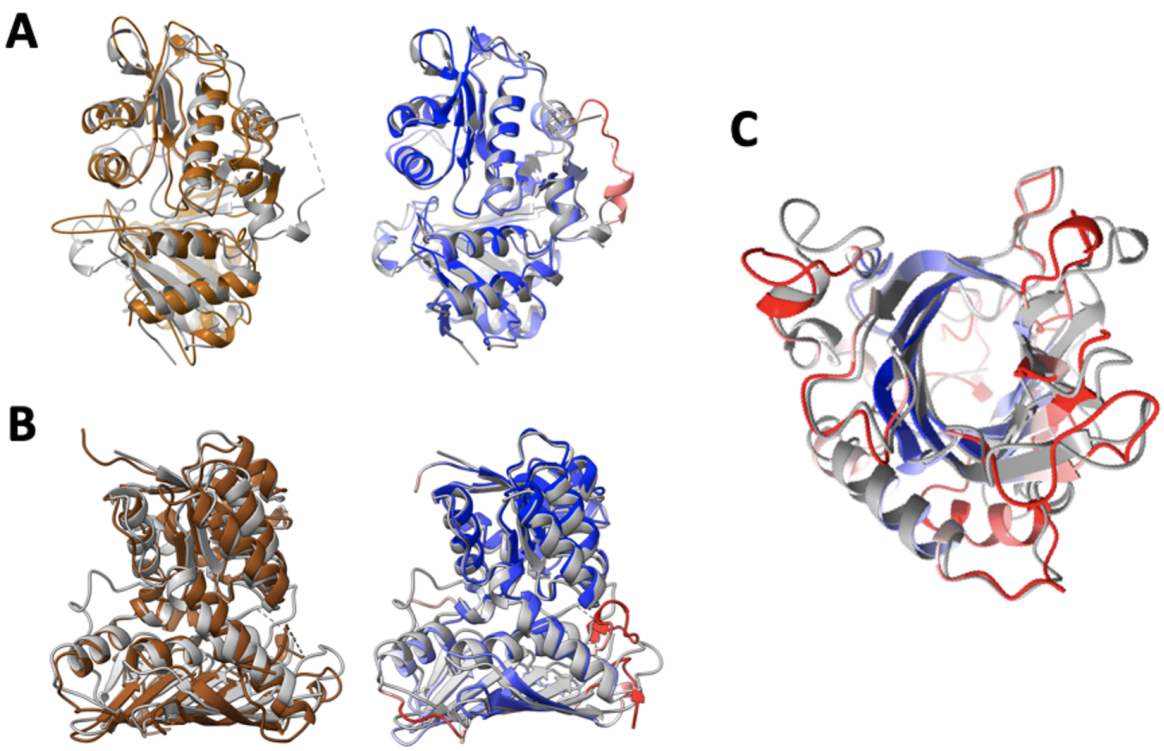

Fig. S5. Enabling experimental structure determination with RoseTTAFold for proteins having distant homologs. (A) The final structure of the hexagonal crystal form of GLYAT is shown in gray, with a dashed line representing the disordered loop. The best single template (left) of the known structure (PDB entry 1sqh) is shown in brown. The RoseTTAFold model (right) is colored based on estimated RMS error, ranging from blue for the minimum of 0.56 Å to red for 1.5 Å and higher. In these superpositions, 217 C_α-atoms of 1sqh match the experimental structure with a C_α-RMSD of 1.84 Å, whereas 283 of the RoseTTAFold model match with a C_α-RMSD of 1.27 Å. **(B)** Structure determination of an oxidoreductase. The final structure of the oxidoreductase is shown in gray, and the best template (PDB entry 4mkz) is shown in brown. Dashed lines indicate unmodelled residues. The RoseTTAFold model is colored based on estimated RMS error, ranging from blue for the minimum of 0.6 Å to red for 3.5 Å and higher. In these superpositions, 203 C_α atoms of 4mkz match the experimental structure with a C_α-RMSD of 1.8 Å, whereas 272 of the RoseTTAFold model match with a C_α-RMSD of 1.34 Å **(C)** The full RoseTTAFold model for Lrbp. The model structure is colored based on estimated RMS error, ranging from red for the minimum of 0.84 Å to red for 1.8 Å and higher. The refined structure is shown in gray.

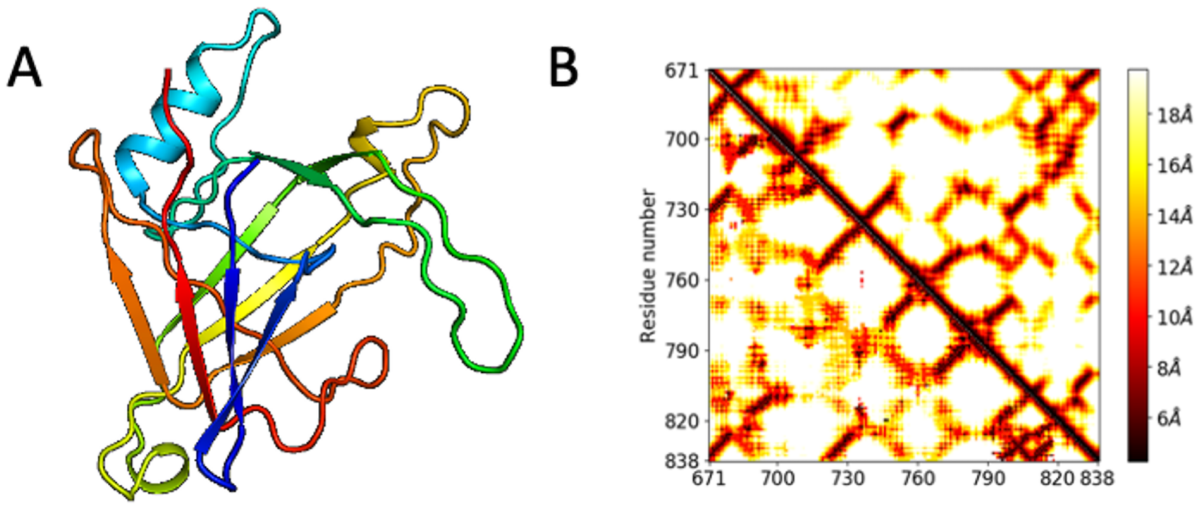

Fig. S6. **A trRosetta model for p101 GBD case. (A)** trRosetta predictions led to irregular all-beta topologies that were physically unrealistic and poorly matched to the resulting density. Six-dimensional density map searching did not yield a preferred placement. **(B)** The trRosetta contacts are ambiguous, particularly at longer sequence separations resulting in a totally different fold. The predicted contacts are shown on the lower left triangle, and the experimentally determined contacts are on the upper right triangle.

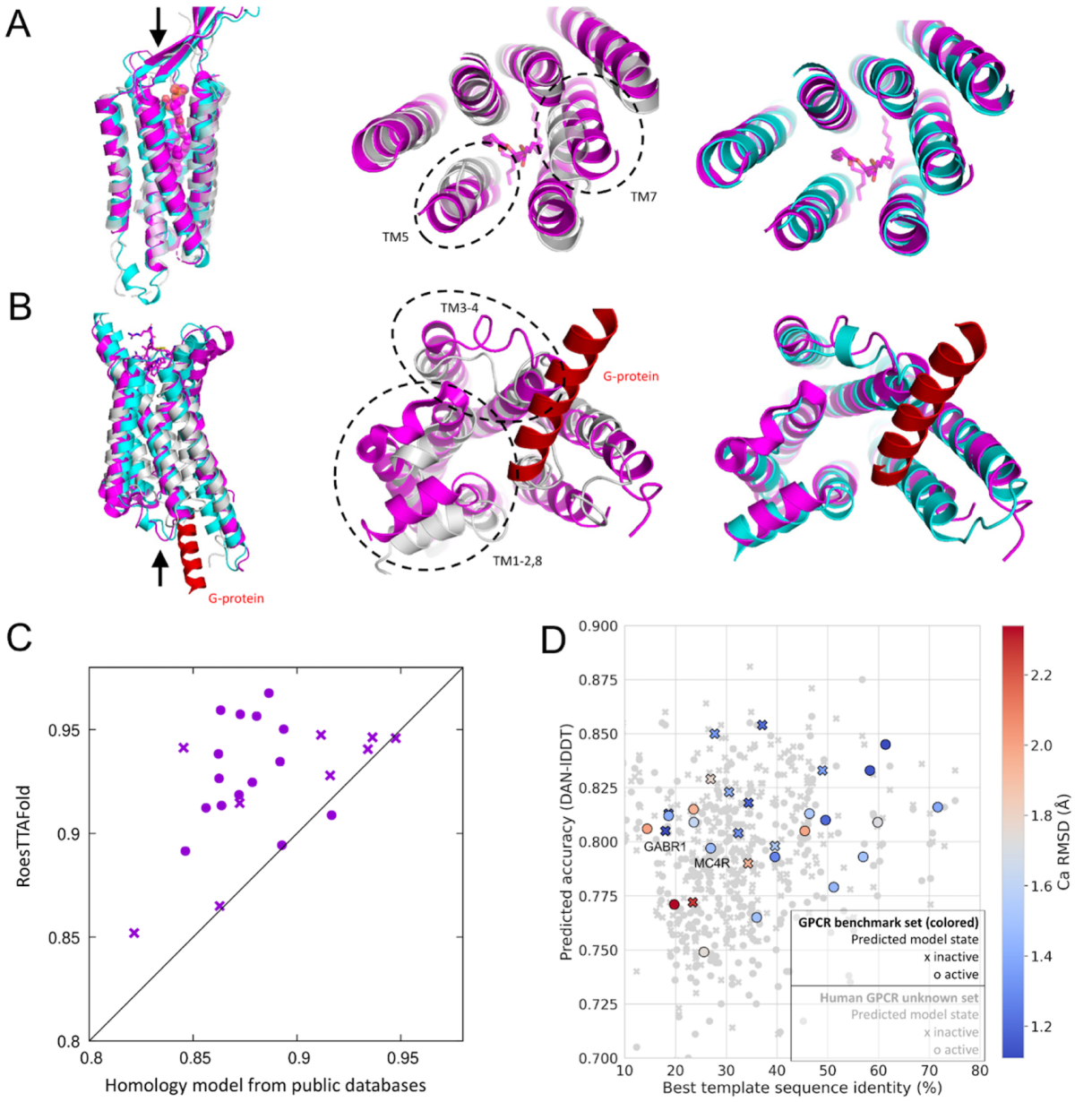

Fig. S7. GPCR modeling. (A, B) Models built for GPCRs not in the training set are compared to crystal structures. **(A)** The best DAN-lDDT (0.81) inactive state model of GABR1_HUMAN (cyan, Uniprot ID Q9UBS5) compared to the native (PDB 6w2y chain B, magenta) and the closest homolog of known structure (PDB 4or2 chain A, gray, seqID 18%). Transmembrane region C_α-RMSD was 1.14 Å. Middle and right panels focused on extracellular regions (top view). **(B)** The best DAN-lDDT (0.80) active state model of MC4R_HUMAN (cyan, Uniprot ID P32245) compared to the native (PDB 7aue chain R, magenta, G-protein helix in red) and the closest homolog of known structure (gray, PDB 3kj6 chain A, seqID 27%). Transmembrane region C_α-RMSD was 1.49 Å. Middle and right panels focused on intracellular regions (bottom view). **(C)** Accuracies (in TM-score) of RoseTTAFold models versus template-based models from public databases (*14*, *15*). Only transmembrane regions were considered. **(D)** For each

active (o) and inactive (x) state prediction, the best template sequence identity and predicted model accuracy (DAN-lDDT) are reported. The color gradient represents actual model accuracy in C_α-RMSD for the subset of proteins of known structure, ranging from 1.2 Å (accurate, blue) to 2.2 Å (inaccurate, red). The human GPCR set with unknown structures is shown in light gray. Data with DAN-lDDT between 0.7 and 0.9 are only shown.

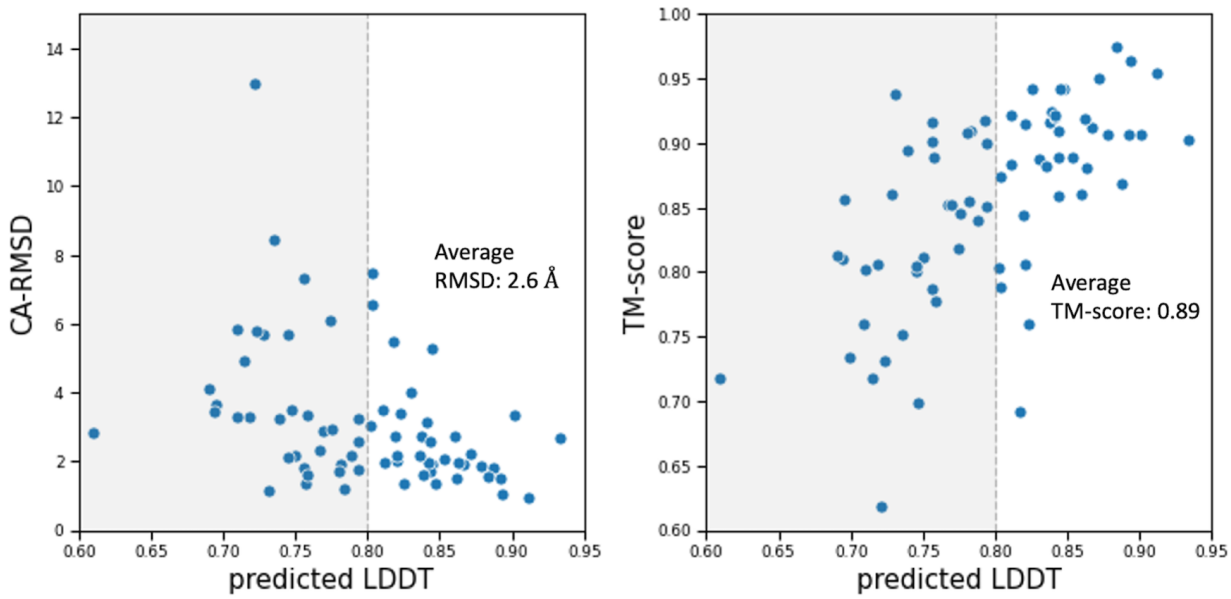

Fig. S8. A correlation between predicted lDDT by DeepAccNet and actual C_α-RMSD for CASP14 targets. The predicted lDDT of 0.80 can roughly be used as a threshold to discriminate between accurate (average C_α-RMSD of 2.6 Å and TM-score of 0.89) and inaccurate models.

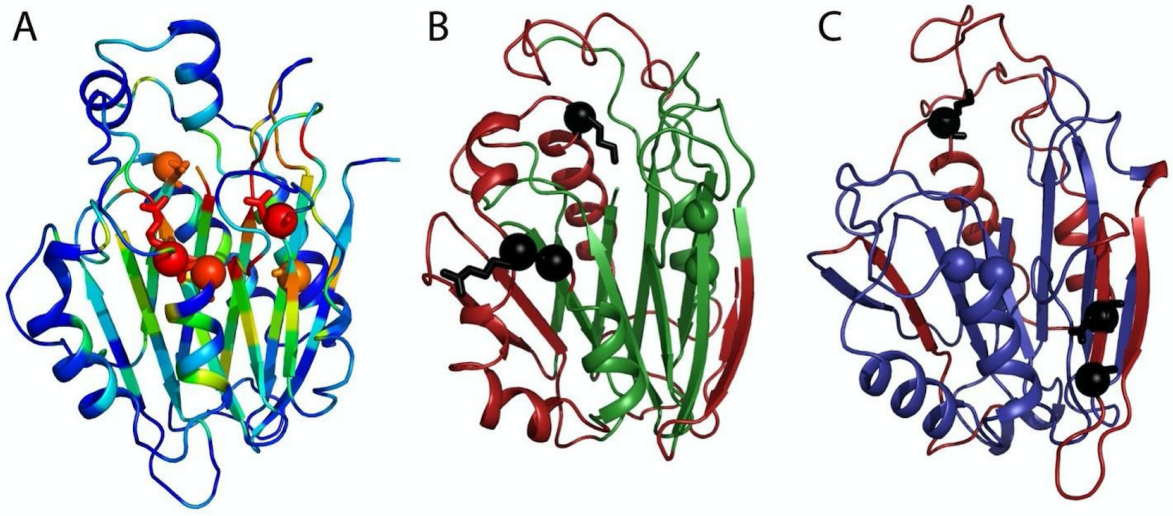

Fig. S9. RoseTTAFold structure for TANGO2 improves homology models. (A) TANGO2 RoseTTAFold structure is colored by ortholog conservation in the rainbow from variable (blue) to conserved (red). Shifted active site residues in either of the homology models are shown in stick with the C_α in the sphere. **(B)** The homology model based on the top HHpred hit to 3gvz template is colored green (aligned with the RoseTTAFold structure) or red (shifted alignment). Three conserved residues (black sphere and stick) shift away from the active site. **(C)** The homology model based on the next best HHpred hit to 2x1d template is colored blue (aligned with the RoseTTAFold structure) or red (shifted alignment). Three conserved residues (black sphere and stick) shift away from the active site.

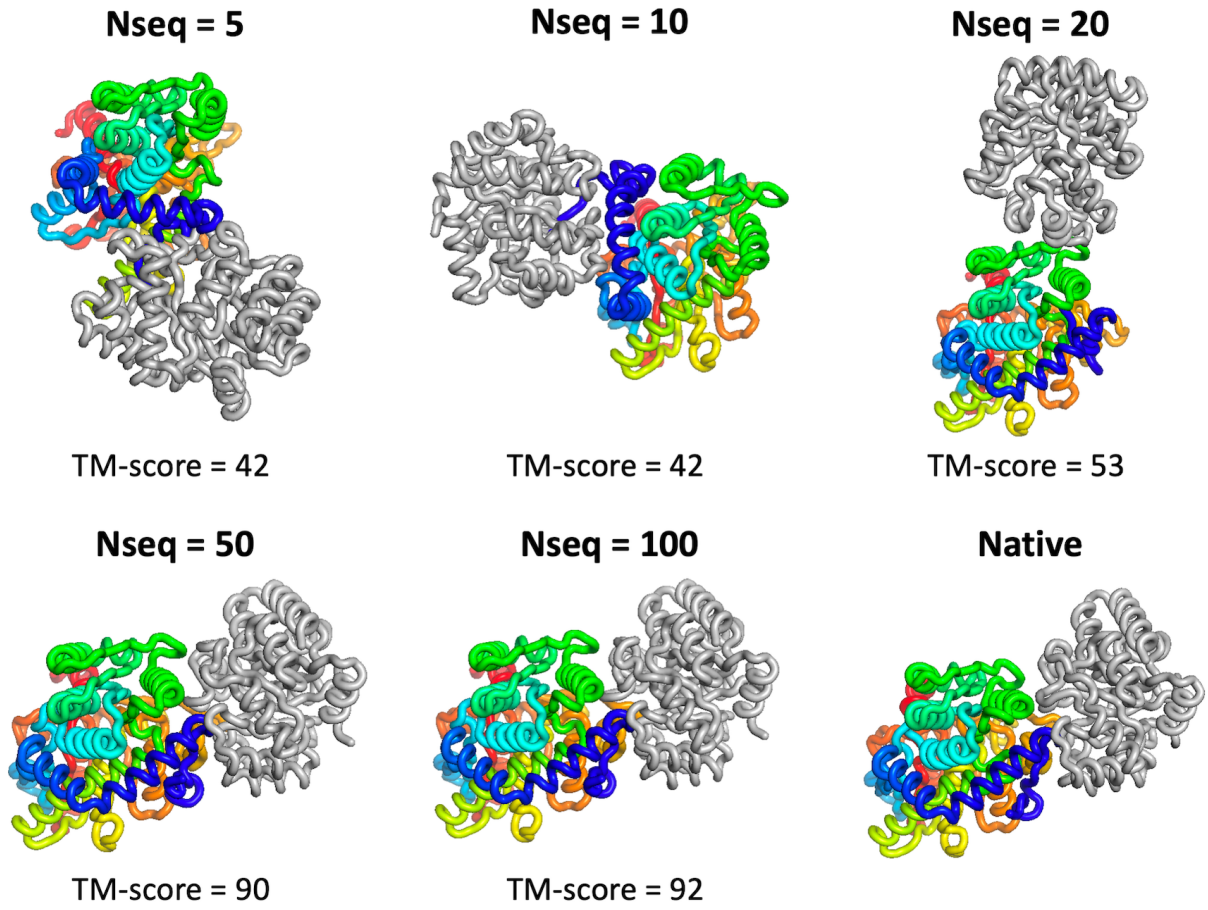

Fig. S10. **Complex model accuracy depends on the number of sequences in paired MSA**. Predicted complex structures of tryptophan synthase (trpA/trpB) with MSAs with various depths are shown. The number of sequences in the MSA is written on the top, and the complex TM-score of the predicted model is written on the bottom.

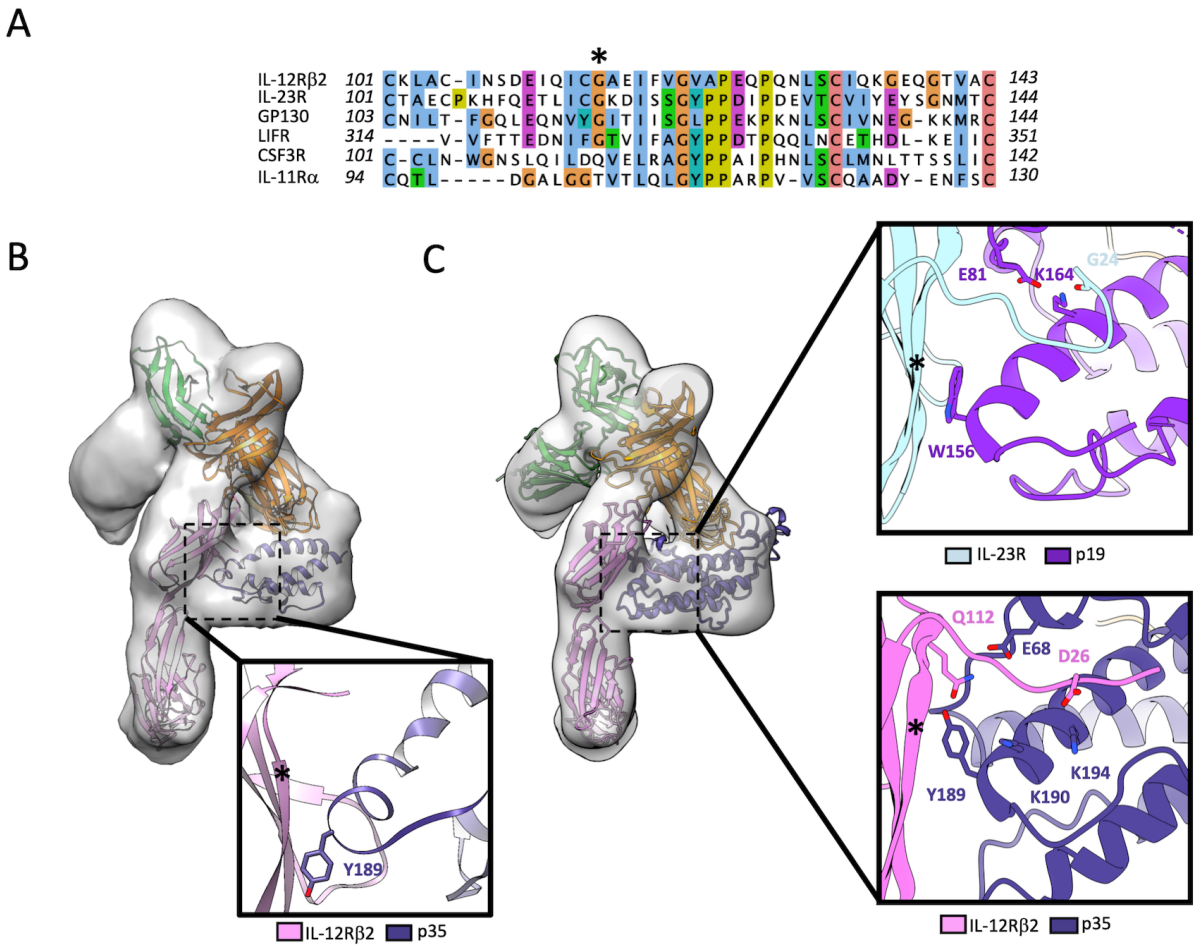

Fig. S11. Analysis of the complete IL-12 receptor complex model. (A) Multiple sequence alignment of gp130 family cytokines highlighting the conserved glycine residue in IL-12Rβ2 (G115), IL-23R (G116), GP130 (G117), and LIFR (G324). Residues were colored using ClustalX (*82*). **(B)** SWISS-MODEL based on the same template (PDB: 6wdq) failed to generate an accurate model. Inset shows the predicted interface between IL-12Rβ2 and IL-12p35. SWISS-MODEL failed to recapitulate the well-conserved interaction between G115 in IL-12Rβ2 and Y189 in IL-12p35. **(C)** Experimental cryo-EM density of the quaternary IL-12R/IL-12 complex (EMD-21645) fits with the RoseTTAFold model. Inset shows a comparison of the interaction between IL-23R and p19 (top, PDB: 6wdq) and IL-12Rβ2 and p35 (bottom, computational model). Star represents the position of glycine residue (G115 in IL-12Rβ2, G116 in IL-23R).

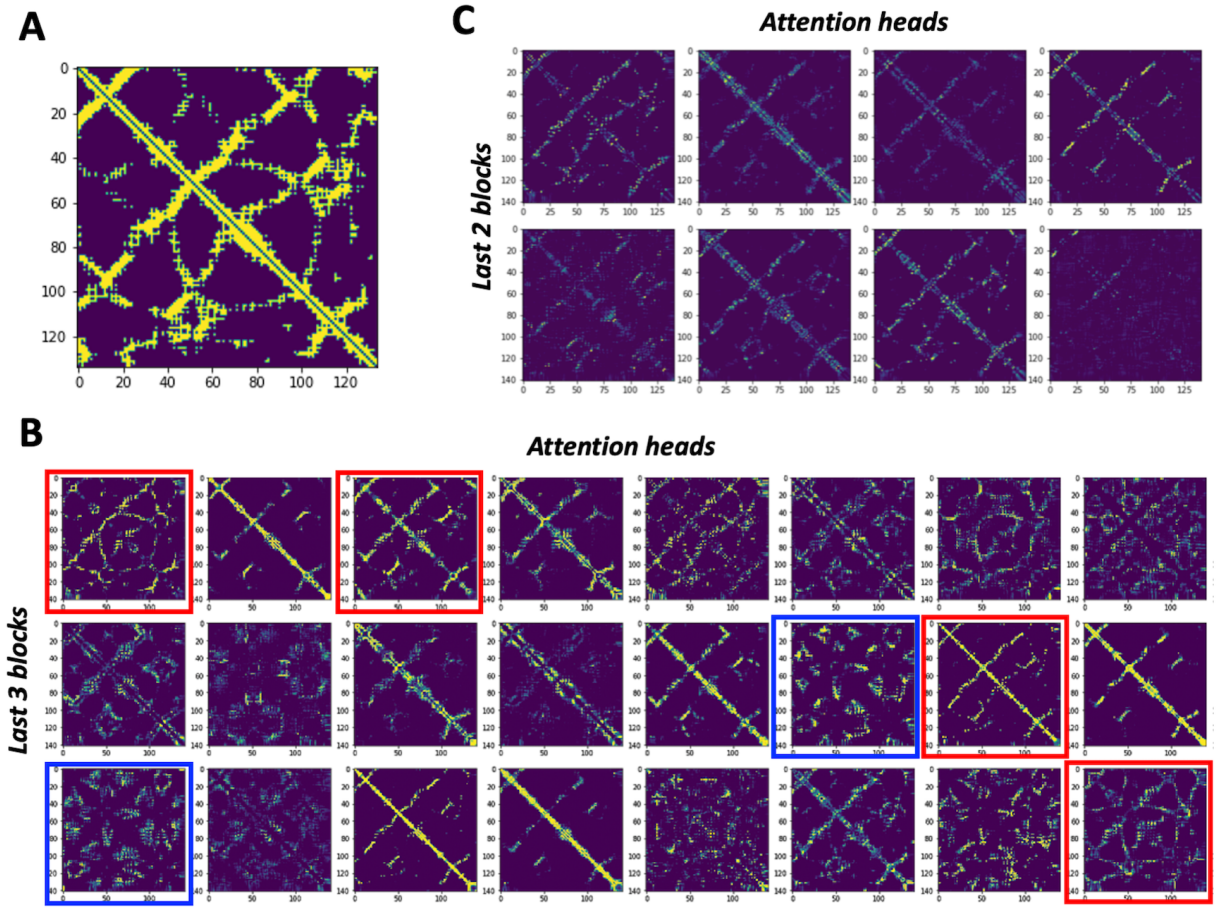

Fig. S12. Examples of attention maps used to update MSA. (A) True contact map of CASP14 target T1049. **(B)** Attention maps from self-attention on MSA features for the last three blocks of the 2-track model (76M parameter model). Some of the attention heads (red boxes) resemble a true contact map. Some cases (blue boxes) only attend to the positions not making the direct contacts. **(C)** Attention maps derived from pair features used to update MSA features. It also shows a similar pattern to the true contact map. The attention maps shown in this figure are symmetrized for clear visualization.

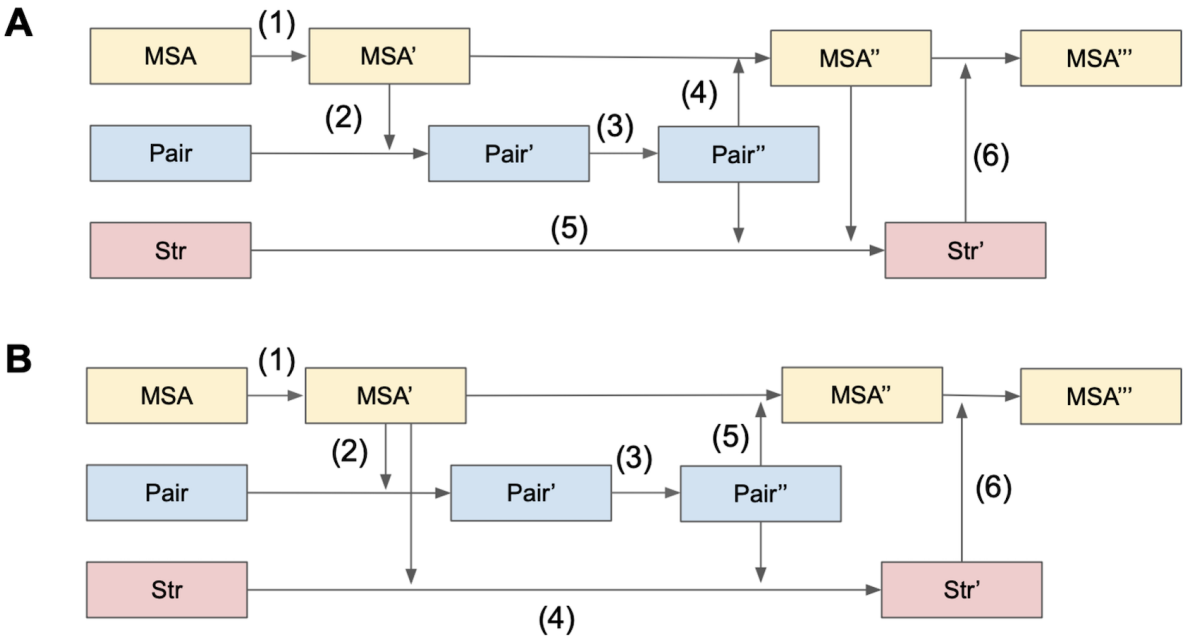

Fig. S13. Two different 3-track block definitions. (A) MSA and pair features are synchronized before structure updates. **(B)** The structure is updated based on unsynchronized MSA and pair features. The numbers in parentheses indicate the order of calculation.

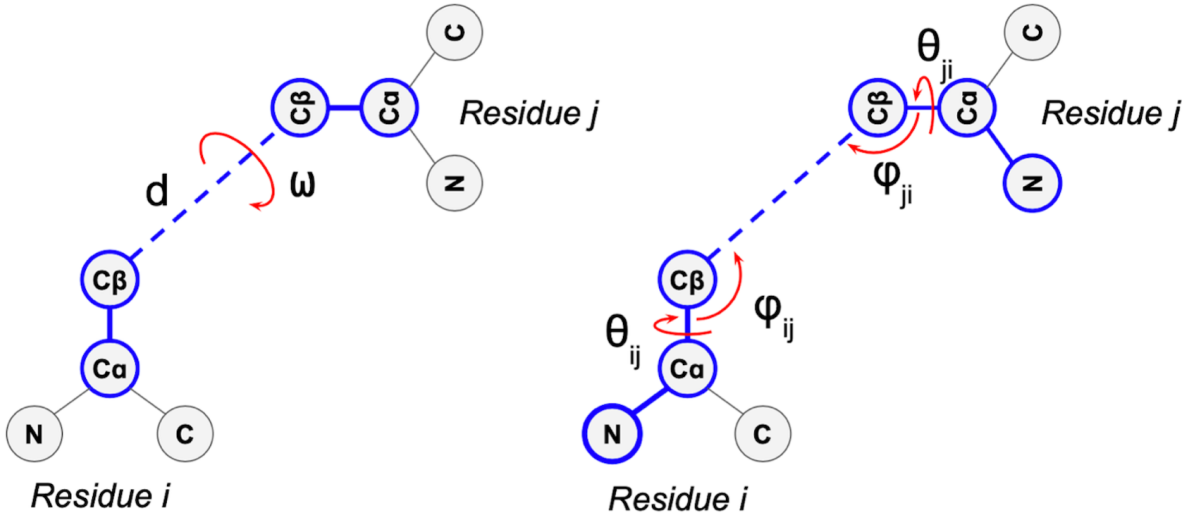

Fig. S14. 6D representation of rigid body transforms between two residues. It includes distance (d) between Cβ atoms, dihedral angle (w) along the virtual bond connecting two Cβ atoms, and two dihedral angles (θ_{ij}, θ_{ji}) and two pseudo-bond angles (φ_{ij}, φ_{ji}) specifying the direction of the Cβ atom of a residue in a reference frame centered on the other residue.

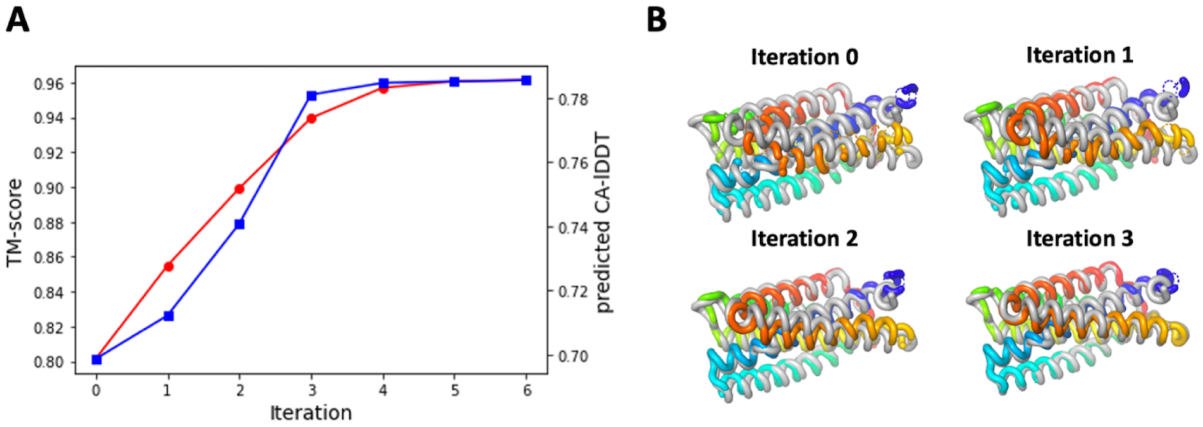

Fig. S15. An example (T1024-D1 from CASP14 targets) of Iterative refinement using SE(3)-Transformers. (A) Model accuracy (TM-score) is improved with iterative refinement. Predicted C_α-lDDT from the network shows a good correlation to the actual model accuracy. **(B)** The model structure at each iteration is shown. The RoseTTAFold models are colored in a rainbow (blue; N-terminal, red; C-terminal), and the native structures are colored in gray.

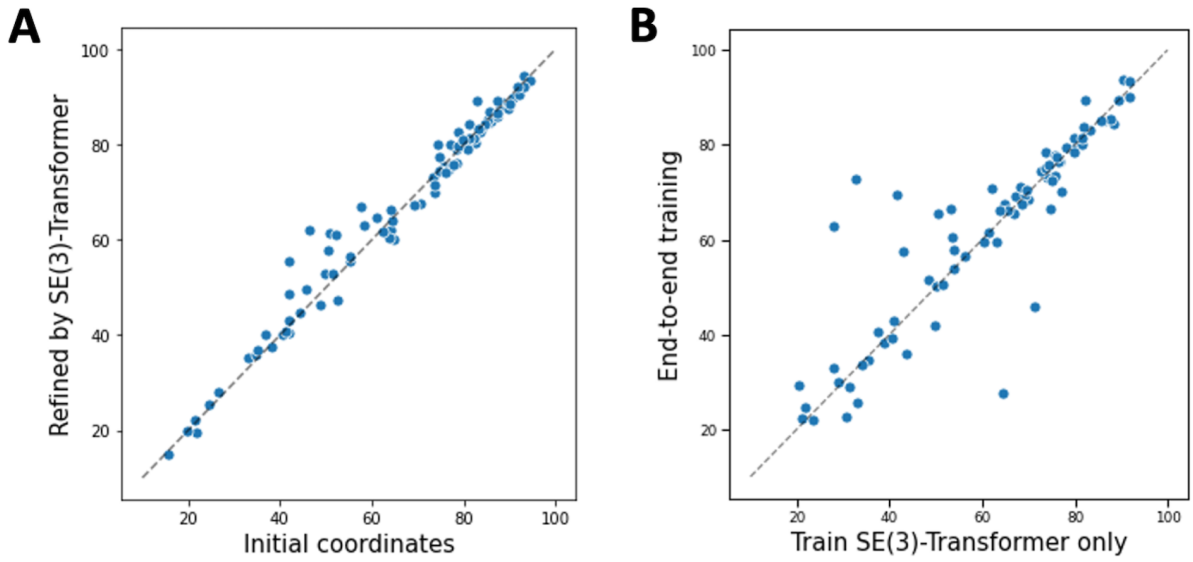

Fig. S16. Experiments with SE(3)-Transformer layers on top of the two-track model. (A) Model accuracy comparison between initial coordinates generated by the simple graph-based network and the refined models through SE(3)-Transformer. **(B)** Model accuracy comparison between networks trained in two different ways: SE(3)-Transformer trained separately with the frozen 2-track model (x-axis) and structure module having the same architecture trained together with 2-track model part (y-axis). Model accuracy is measured by TM-score.

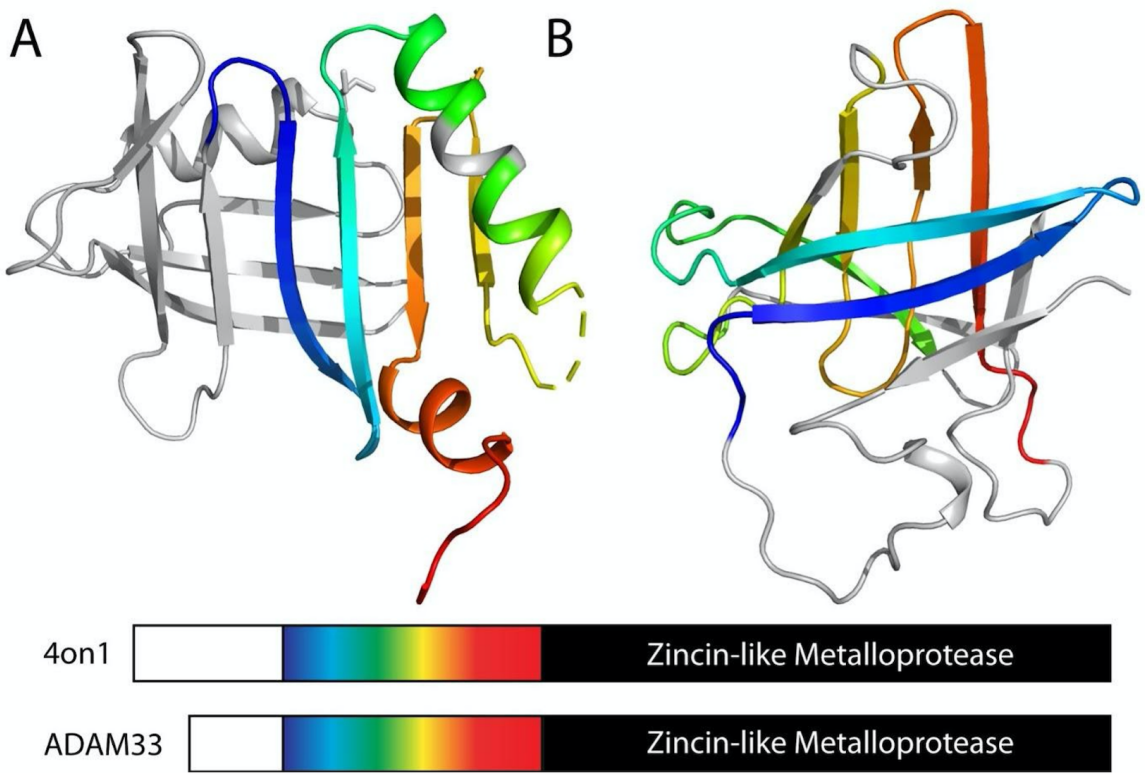

Fig. S17. RoseTTAFold structure avoids multi-domain problems. (A) The prodomain from HHpred template 4on1 is in ribbon and adopts a fragilysin-like α+β fold with a central 4-stranded beta-meander. The domain architecture below highlights a C-terminal metalloprotease that is also in ADAM33. The HHpred template alignment incorrectly extends into the prodomain (aligned sequence in a rainbow). **(B)** ADAM33 RoseTTAFold structure (oriented by its corresponding central beta-meander) adopts a lipocalin-like beta-barrel. The aligned beta-meander sequence (in a rainbow) is unrelated to the alpha + beta sequence from the template.

Table S1. Performance of different model architectures in terms of inter-residue geometry prediction loss (cross entropy), top L long-range contact accuracy and C_α-lDDT.

Architecture	Inter-residue geometry loss	Top L long-range contact accuracy	C_α-lDDT
Single Track (Sequential processing of MSA and pair feature)			
Architecture 1) Hand-crafted features + 2D convolution	5.56	54%	-
Architecture 2) MSA encoder + 2D convolution	5.49	56%	-
Architecture 3) MSA encoder + Axial attention	5.14	58%	-
2-track (Parallel track for MSA and pair features)			
Architecture 4) Untied + addition + cross	5.54	54%	-
Architecture 5) Untied + addition + direct	5.18	58%	-
Architecture 6) Untied + concat + direct	5.01	60%	-
Architecture 7) Soft-tied + concat + direct	4.84	62%	-
Architecture 8) architecture 7 + scale-up	4.50	67%	-
Architecture 9) architecture 8 + SE(3) structure module	4.54	67%	0.70
3-track (Parallel track for MSA, pair, and 3D coordinates)			
Architecture 10) Structure update w/ unsynchronized MSA and pair features (**Fig. S13B)**	4.63	64%	0.68
Architecture 11) Structure update w/ synchronized MSA and pair features (**Fig. S13A)**	4.36	69%	0.72
Architecture 12) architecture 11 + SE(3) structure module	4.39	69%	0.77

Table S2. Current refinement statistics for crystal structures

Crystal	GLYAT	Oxidoreductase	SLP	Lrbp
Space group	$P6_5$	$P2_1$	$P2_12_12_1$	$P2_1$
Cell dimensions				
a, b, c (Å)	97.18, 97.18, 144.63	79.15, 157.86, 95.01	63.16, 98.87, 155.12	50.10, 81.37, 78.47
α, β, γ (˚)	90, 90, 120	90, 114.45, 90	90, 90, 90	90, 107.57, 90
Resolution (Å)	1.65	2.34	2.18	1.53
No. non-H atoms	5002	14568	5463	4621
No. reflections	83145	87002	49958	89331
R_{work}, R_{free}	0.174, 0.200	0.283, 0.322	0.216, 0.250	0.248, 0.280

Table S3. HHpred Results Summary for TANGO2, CERS1, and ADAM33

Example	Sequence	MSA	Hit	Prob	Cols	Query	Temp.	Coverage
TANGO2	Q6ICL3:1-259	Uniref30	PDB: 3GVZ_A	98.9	219	1-252	25-256	0.846
TANGO2	Q6ICL3:1-259	Uniref30	ECOD: e3gvzA1	98.5	217	2-253	1-232	0.838
TANGO2	Q6ICL3:1-259	Uniref30	PDB: 2X1D_D	98.2	210	1-254	102-330	0.811
TANGO2	Q6ICL3:1-259	Uniref30	PDB: 3HBC_A	97.8	217	1-255	3-274	0.838
TANGO2	Q6ICL3:1-259	Uniref30	ECOD: e3hbcA1	97.7	212	1-247	3-268	0.819
CERS1	P27544:98-304	Uniref30	ECOD: e3nqwB1	8.5	53	64-116	12-65	0.256
CERS1	P27544:98-304	PDB70	PDB: 6TY2_A	17.7	22	103-124	27-48	0.106
ADAM33	Q9BZ11:39-167	Uniref30	PDB: 4ON1_B	96.3	90	24-117	89-184	0.698

Table S4. Performance of RoseTTAFold on 68 interacting pairs in the PDB benchmark set. The complex model accuracy is evaluated based on the Interface Contact Similarity (ICS) score and complex TM-score.

UniProtID	PDB ID	ICS	Complex TM-score	UniProtID	PDB ID	ICS	Complex TM-score
P77499_P77689	2zu0_C,2zu0_A	0.93	0.52	P0A9P4_P0AA25	1f6m_F,1f6m_H	0.36	0.69
P77165_P77489	5g5g_A,5g5g_C	0.91	0.95	P07014_P69054	2acz_B,2acz_C	0.35	0.75
P76077_P76079	3pw8_D,3pw8_B	0.90	0.96	P0A772_P37146	3n3b_D,3n3b_C	0.34	0.79
P0AAV4_P75745	5dud_B,5dud_A	0.89	0.95	P0A6E6_P0ABA6	3oaa_X,3oaa_W	0.33	0.81
P0AFE4_P0AFF0	3rko_J,3rko_I	0.89	0.97	P02358_P0A7T7	4v6l_J,4v6l_V	0.32	0.64
P00363_P0AC47	5vpn_E,5vpn_F	0.88	0.95	P11349_P11350	3ir7_B,3ir7_C	0.32	0.53
P77165_P77324	5g5g_A,5g5g_B	0.88	0.93	Q46898_Q46899	5cd4_H,5cd4_G	0.30	0.69
P0A8Q0_P0A8Q3	6awf_C,6awf_D	0.85	0.86	P0A7R9_P68679	4v6l_O,4v6l_Y	0.28	0.62
P0A7K6_P0ADY3	6c4i_Q,6c4i_L	0.85	0.93	P0A7M9_P62399	4v6l_AB,4v6l_FA	0.22	0.68
P0AC44_P69054	2acz_D,2acz_C	0.84	0.90	P0A9Q5_P0ABD5	2f9y_B,2f9y_A	0.19	0.52
P0A877_P0A879	4hn4_A,4hn4_B	0.84	0.92	P0A7V3_P0AG59	3ja1_S,3ja1_N	0.09	0.61
P0A6X7_P0A6Y1	5wfe_K,5wfe_L	0.83	0.85	P0AEJ6_P19636	3ao0_C,3ao0_D	0.05	0.65
O32583_P30138	1zud_D,1zud_C	0.83	0.96	P05719_P08957	2y7h_A,2y7h_C	0.00	0.44
P07014_P0AC41	2wu5_J,2wu5_I	0.82	0.94	P02916_P0AEX9	4ki0_D,4ki0_C	0.00	0.50
P28630_P28631	1xxi_F,1xxi_J	0.80	0.64	Q46897_Q46899	5cd4_A,5cd4_B	0.00	0.52
P30750_P31547	3tuz_D,3tuz_B	0.79	0.87	P68183_P68187	3puy_C,3puy_D	0.00	0.53
P69346_P69348	2a6q_C,2a6q_F	0.79	0.82	P45956_Q46896	5dqz_H,5dqz_D	0.00	0.70
P76014_P76015	3pnl_B,3pnl_A	0.77	0.94	P0C077_P0C079	4fxe_E,4fxe_B	0.00	0.50
P0A836_P0AGE9	1scu_D,1scu_C	0.77	0.91	P0AFE8_P0AFF0	3rko_H,3rko_I	0.00	0.51
P06609_P06611	4dbl_G,4dbl_I	0.76	0.78	P0AEX9_P68183	3puy_A,3puy_C	0.00	0.54
P30748_P30749	1fma_A,1fma_B	0.76	0.91	P0AA25_P17854	2o8v_B,2o8v_A	0.00	0.62
Q47149_Q47150	4q2u_J,4q2u_I	0.75	0.68	P0A988_P69931	5x06_B,5x06_F	0.00	0.60
P02916_P68183	3puy_B,3puy_C	0.72	0.71	P0A988_P28630	1jqj_A,1jqj_C	0.00	0.52
P76458_P76459	5dbn_G,5dbn_H	0.72	0.94	P0A7L0_P0A9W3	3j5s_E,3j5s_D	0.00	0.47
P0A7L3_P0AG48	6enu_LA,6enu_MA	0.71	0.87	P0A7I0_P0ACC1	2b3t_B,2b3t_A	0.00	0.52

P0AFE0_P0AFE4	3rko_L,3rko_J	0.69	0.85	P0A6P1_P0CE48	4pc2_D,4pc2_C	0.00	0.39
P0AFE0_P0AFF0	3rko_L,3rko_I	0.65	0.90	P0A6P1_P0CE47	3avy_A,3avy_A	0.00	0.39
P0AAJ3_P0AEK7	1kqg_B,1kqg_C	0.63	0.77	P0A6N4_P0A8N7	3a5z_D,3a5z_C	0.00	0.64
P08839_P0AA04	2xdf_B,2xdf_D	0.60	0.55	P06609_P37028	4fi3_B,4fi3_E	0.00	0.54
P0ADC1_P31554	4q35_B,4q35_A	0.48	0.83	P03007_P0A988	5m1s_D,5m1s_B	0.00	0.59
P0AF32_P11349	1q16_C,1q16_B	0.40	0.46	P02916_P68187	4jbw_A,4jbw_D	0.00	0.42
P0AG99_P0AGA2	3j45_C,3j45_A	0.37	0.78	P02413_7P0A7Q1	6c4i_M,6c4i_FA	0.00	0.34
P0AE70_P0AE72	1ub4_A,1ub4_C	0.37	0.64	P00634_P0AG86	5jtl_E,5jtl_D	0.00	0.24
P0AFK0_P0AGG8	5nj5_B,5nj5_A	0.36	0.63	P00363_P64559	6b58_C,6b58_D	0.00	0.79