RESEARCH ARTICLE

PROTEIN FOLDING

Accurate prediction of protein structures and interactions using a three-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[7], R. Dustin Schaeffer[6], Claudia Millán[8], Hahnbeom Park[1,2], Carson Adams[1,2], Caleb R. Glassman[9,10,11], 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[9,10,11,20], Nick V. Grishin[6,7,21], Paul D. Adams[12,22], Randy J. Read[8], David Baker[1,2,23]*

DeepMind presented notably accurate predictions at the recent 14th Critical Assessment of Structure Prediction (CASP14) conference. We explored network architectures that incorporate related ideas and obtained the best performance with a three-track network in which information at the one-dimensional (1D) sequence level, the 2D distance map level, and the 3D coordinate level is successively transformed and integrated. The three-track network produces structure predictions with accuracies approaching those of DeepMind in CASP14, enables the rapid solution of challenging x-ray crystallography and cryo–electron microscopy 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 that require modeling of individual subunits followed by docking. We make the method available to the scientific community to speed biological research.

The prediction of protein structure from amino acid sequence information alone has been a long-standing challenge. The biannual Critical Assessment of Structure Prediction (CASP) meetings have demonstrated that deep-learning methods such as AlphaFold (*1*, *2*) and trRosetta (*3*), which extract information from the large database of known protein structures in the Protein Data Bank (PDB), outperform more traditional approaches that explicitly model the folding process. The outstanding performance of DeepMind's AlphaFold2 in the recent 14th CASP (CASP14) meeting (https://predictioncenter.org/casp14/zscores_final.cgi) left the scientific community eager to learn details beyond the overall framework that was 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 (i) starting from multiple sequence alignments (MSAs) rather than from more-processed features such as inverse covariance matrices derived from MSAs, (ii) replacement of two-dimensional (2D) convolution with an attention mechanism that better represents interactions between residues distant along the sequence, (iii) 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, (iv) 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 (v) 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 that incorporate 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. By 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 because the three-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

[1]Department of Biochemistry, University of Washington, Seattle, WA 98195, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA 98195, USA. [3]Faculty of Arts and Sciences, Division of Science, Harvard University, Cambridge, MA 02138, 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]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA. [8]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK. [9]Program in Immunology, Stanford University School of Medicine, Stanford, CA 94305, USA. [10]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA 94305, USA. [11]Department of 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, BC, Canada. [19]Life Sciences Institute, Department of Biochemistry and Molecular Biology, The University of British Columbia, Vancouver, BC, Canada. [20]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305, USA. [21]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX, USA. [22]Department of Bioengineering, University of California, Berkeley, Berkeley, CA 94720, USA. [23]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA.
*Corresponding author. Email: dabaker@uw.edu

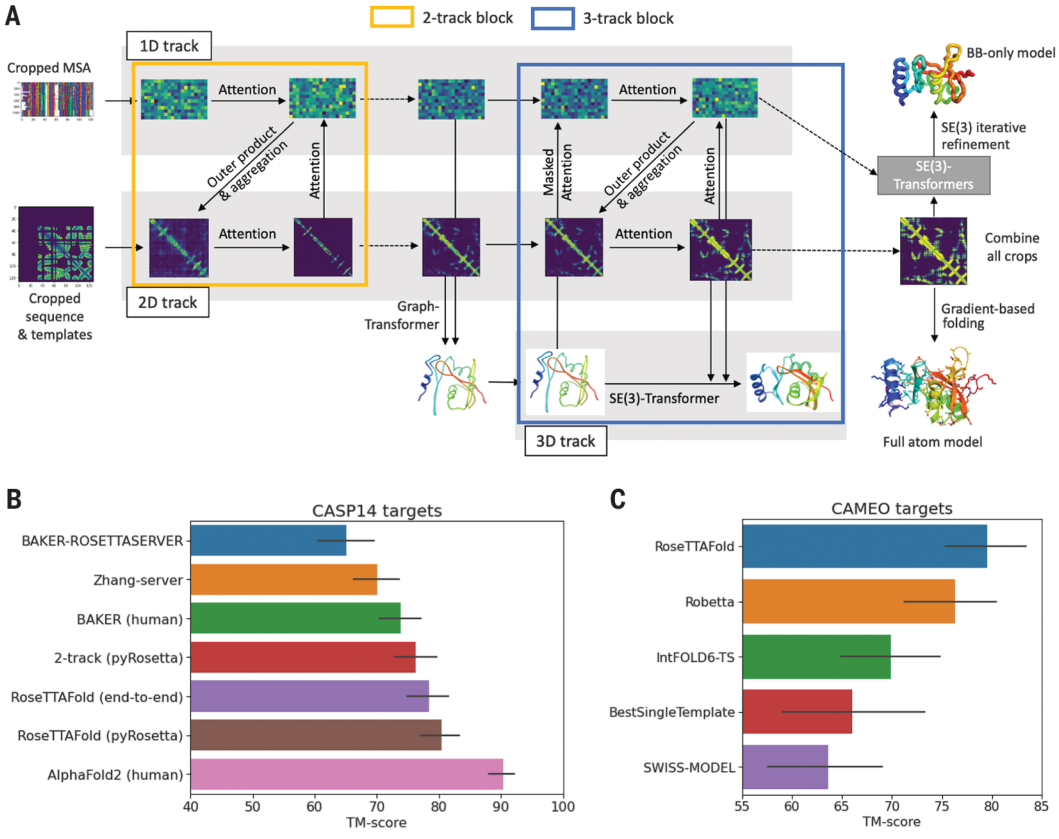

Fig. 1. Network architecture and performance. (**A**) RoseTTAFold architecture with 1D, 2D, and 3D attention tracks. Multiple connections between tracks allow the network to simultaneously learn relationships within and between sequences, distances, and coordinates (see methods and fig. S1 for details). (**B**) Average TM-score of prediction methods on the CASP14 targets. Zhang-server and BAKER-ROSETTASERVER were the top two server groups, whereas AlphaFold2 and BAKER were the top two human groups in CASP14; BAKER-ROSETTASERVER and BAKER predictions were based on trRosetta. Predictions with the two-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/). In (B) and (C), the error bars represent a 95% confidence interval.

features are fed into a final SE(3)-equivariant layer (*6*), and, after 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 advantages of requiring lower-memory graphics processing units (GPUs) at inference time [for proteins with more than 400 residues, 8 gigabytes (GB) rather than 24 GB] and of producing full side-chain models but requires central processing unit (CPU) time for the pyRosetta structure modeling step.

The three-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 two server groups (Zhang-server and BAKER-ROSETTASERVER), BAKER human group (ranked second among all groups), and our two-track attention models on CASP14 targets. As in the case of AlphaFold2, the correlation between MSA depth and model accuracy is lower for RoseTTAFold than for trRosetta and other methods tested at CASP14 (fig. S2). The performance of the three-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; after sequence and template search (~1.5 hours), the end-to-end version of RoseTTAFold requires ~10 min on an RTX2080 GPU to generate backbone coordinates for proteins with fewer than 400 residues, and the pyRosetta version requires 5 min 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 with the pyRosetta version (Fig. 1B; the latter incorporates side-chain information at the all-atom relaxation stage); because SE(3)-equivariant layers are used in the main body of the three-track model, the added gain from the final SE(3) layer is likely less than that 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 three-track models over the two-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 MSA, distance map, and 3D coordinate representations can more effectively extract sequence-structure relationships than reasoning over only MSA and distance map information. The relatively low computational cost makes it straightforward to incorporate the methods in a public server and predict structures for large sets of proteins, for example, all human G protein–coupled receptors (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 2 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 15 May 2021 on CAMEO; over the 69 medium and hard targets released during this time (15 May 2021 to 19 June 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. Because the network can take templates of known structures as input, 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 MSAs to sample a broader range of models. These approaches generated ensembles that contained 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 we 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 that are 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 MSA 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). As of now, RoseTTAFold only uses the top 1000 sequences because of memory limitations; with this addition, all available sequence information can be used (often more than 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 (cryo-EM) and to build models that provide 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 compared with 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 enabled solution of the structures by MR [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; 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, whereas 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 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 (root mean square deviation) between the predicted and the final refined structure is 3.0 Å for the core β 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, GPCRs 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

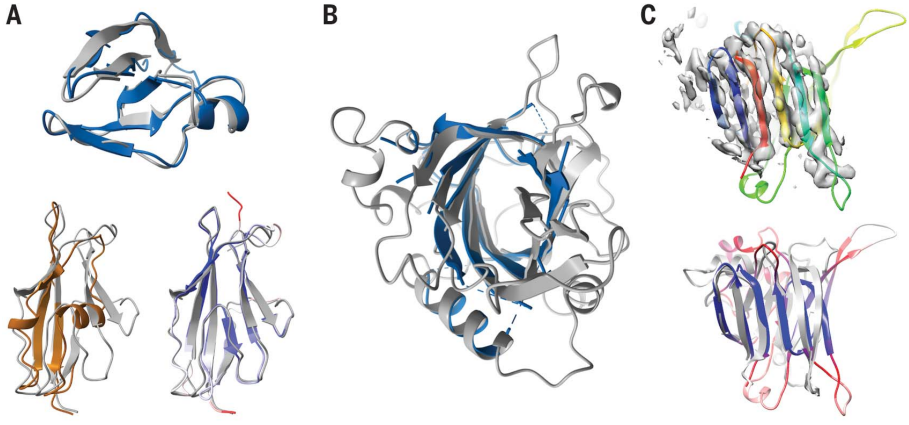

Fig. 2. Enabling experimental structure determination with RoseTTAFold. (**A** and **B**) Successful molecular replacement with RoseTTAFold models. SLP is shown in (A). The C-terminal domain is shown at the top, with a comparison of final refined structure (gray) to RoseTTAFold model (blue); there are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

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 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. More than one-third of these models have a predicted local distance difference test (lDDT) >0.8, which corresponded to an average C_α-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 endoplasmic reticulum (*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 may function as an enzyme that hydrolyzes 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 [Arg^{26}→Lys (R26K), Arg^{32}→Gln (R32Q), and Leu^{50}→Pro (L50P), near the active site] or produce steric clashes [Gly^{154}→Arg (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 (a disintegrin and metalloproteinase with thrombospondin motifs) families of metalloproteases are encoded by more than 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 the fold of the prodomain, which has no homologs of known structure, is not. The RoseTTAFold-predicted structure of the ADAM33 prodomain has a lipocalin-like β-barrel fold (Fig. 3C) that belongs to an extended superfamily that includes metalloprotease inhibitors (*25*). There is a cysteine in an extension following the predicted prodomain barrel; taken together, these data are consistent with experimental data that suggest 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).

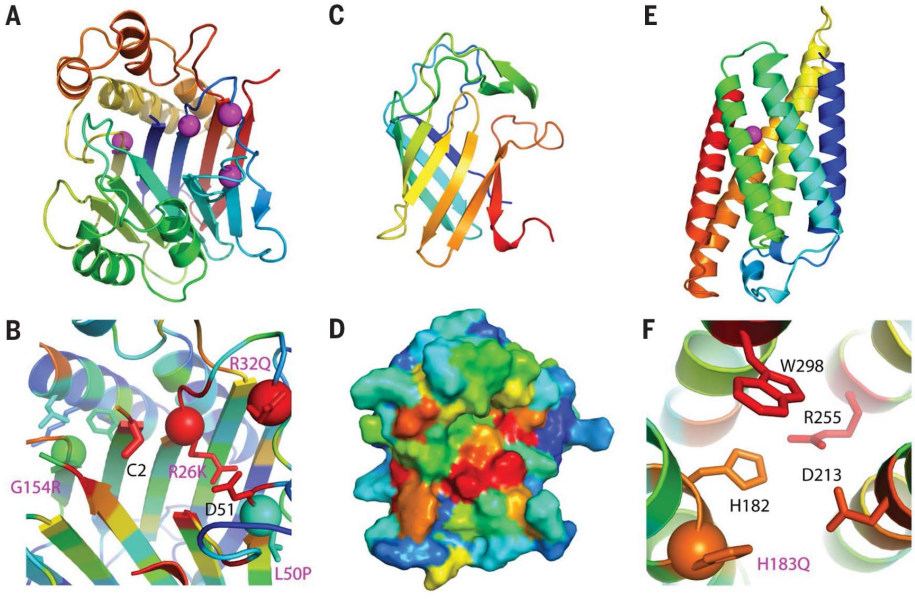

Fig. 3. RoseTTAFold models provide insights into function. (**A**) TANGO2 model, colored in a rainbow from the N terminus (blue) to the C terminus (red), adopts an Ntn hydrolase fold. Pathogenic mutation sites are represented by magenta spheres. (**B**) Predicted TANGO2 active site colored by ortholog conservation from variable (blue) to conserved (red), with conserved residues shown in stick format and labeled. Pathogenic mutations (spheres with wild-type side chains in sticks) are labeled in magenta; select neighboring residues are depicted in sticks. (**C**) ADAM33 prodomain adopts a lipocalin-like barrel (blue, N terminus; red, C terminus). (**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 colored from N terminus (blue) to C terminus (red), with a pathogenic mutation in TMH2 near a central cavity represented by a magenta sphere. (**F**) Zoom-in of the CERS1 active site, with residues colored by ortholog conservation from variable (blue) to conserved (red). Residues that contribute to catalysis (His^{182} and Asp^{213}) or are conserved (Trp^{298} and Asp^{213}) line the cavity. The conserved pathogenic mutation H183Q is adjacent to the active site. D, Asp; H, His; Q, Gln; R, Arg; W, Trp.

Transmembrane spanning ceramide synthase (CERS1) is a key enzyme in sphingolipid metabolism that uses acyl–coenzyme A (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 ceramide synthase 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 (His^{182} and Asp^{213}) (*30*) or conserved (Trp^{298}), as well as a pathogenic mutation [His^{183}→Gln (H183Q)] found in progressive myoclonus epilepsy and dementia that decreases ceramide levels (*31*). This active-site composition (His^{182}, Asp^{213}, and potentially a neighboring Ser^{212}) 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 three-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 with 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 being 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

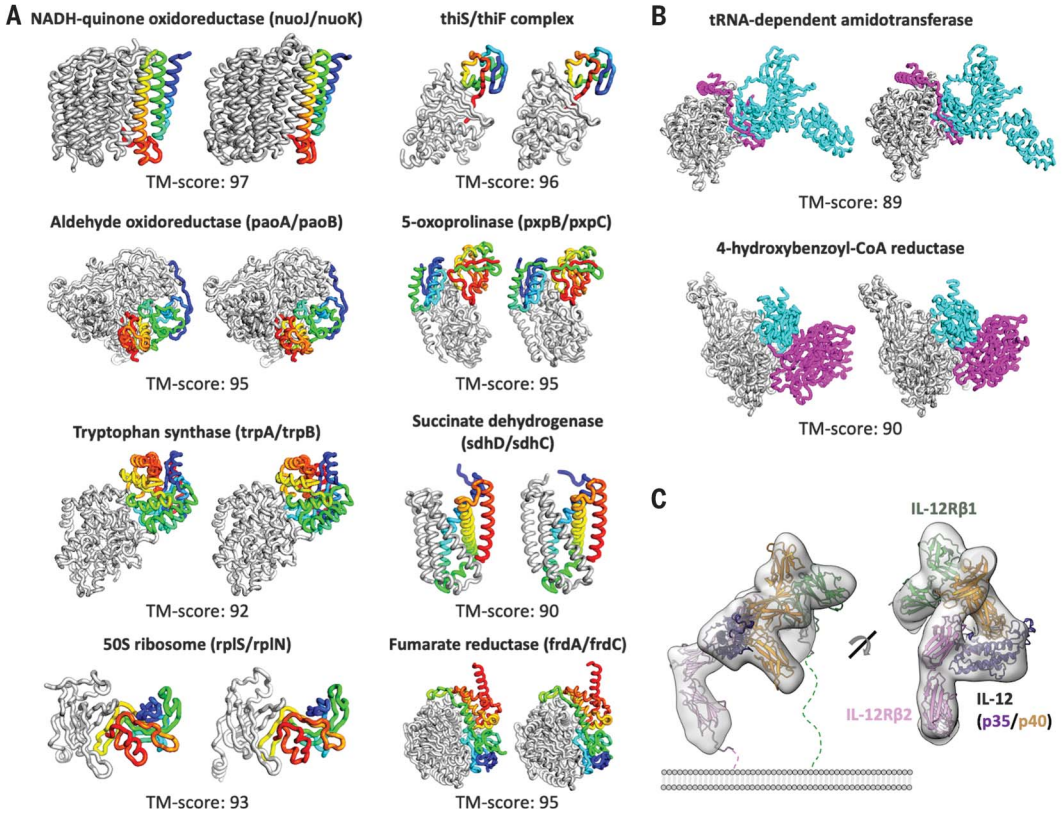

Fig. 4. Complex structure prediction using RoseTTAFold. (**A** and **B**) Prediction of structures of *Escherichia coli* protein complexes from sequence information. Experimentally determined structures are on the left, and RoseTTAFold models are on the right; the TM-scores below indicate the extent of structural similarity. Two chain complexes are shown in (A). The first subunit is colored in gray, and the second subunit is colored in a rainbow from blue (N terminus) to red (C terminus). Three chain complexes are shown in (B). 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).

(complex models are generated from sequence information in ~30 min on a 24-GB TITAN RTX GPU), this approach implements "flexible backbone" docking almost by construction because the structures of the chains are predicted in the context of each other. We tested the end-to-end three-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 [template modeling score (TM-score) (*33*) >0.8]. Information on residue-residue coevolution between the paired sequences likely contributes to the accuracy of the rigid-body placement because 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 interleukin-12 receptor–interleukin-12 (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 affecting IL-23 signaling. The RoseTTAFold model fits the experimental cryo-EM density well and identified a shared interaction between Tyr^{189} in IL-12p35 and Gly^{115} in IL-12Rβ2 analogous to the packing between Trp^{156} in IL-23p19 with Gly^{116} in IL-23R. In addition, the model suggests a role for the IL-12Rβ2 N-terminal peptide (residues 24 to 31) in IL-12 binding (IL-12Rβ2 Asp^{26} may interact with nearby Lys^{190} and Lys^{194} in IL-12p35), which may provide an avenue to specifically target the IL-12Rβ2–IL-12 interaction.

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 that incorporate 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 *et al.*, *Nature* **577**, 706–710 (2020).
2. J. Jumper *et al.*, in *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction: CASP14 Abstract Book* (Protein Structure Prediction Center, 2020), pp. 22–24.
3. J. Yang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 1496–1503 (2020).
4. I. Anishchenko, T. M. Chidyausiku, S. Ovchinnikov, S. J. Pellock, D. Baker, bioRxiv 2020.07.22.211482 [Preprint] (2020); https://doi.org/10.1101/2020.07.22.211482.
5. S. Chaudhury, S. Lyskov, J. J. Gray, *Bioinformatics* **26**, 689–691 (2010).
6. F. B. Fuchs, D. E. Worrall, V. Fischer, M. Welling, arXiv:2006.10503 [cs.LG] (2020).
7. J. Haas *et al.*, *Proteins* **86**, 387–398 (2018).
8. L. J. McGuffin *et al.*, *Nucleic Acids Res.* **47**, W408–W413 (2019).
9. J. Haas *et al.*, *Proteins* **87**, 1378–1387 (2019).
10. A. Waterhouse *et al.*, *Nucleic Acids Res.* **46**, W296–W303 (2018).
11. A. Jaegle *et al.*, arXiv:2103.03206 [cs.CV] (2021).
12. N. Hiranuma *et al.*, *Nat. Commun.* **12**, 1340 (2021).
13. M. Steinegger *et al.*, *BMC Bioinformatics* **20**, 473 (2019).
14. A. J. Kooistra *et al.*, *Nucleic Acids Res.* **49**, D335–D343 (2021).
15. B. J. Bender, B. Marlow, J. Meiler, *PLOS Comput. Biol.* **16**, e1007597 (2020).
16. L. S. Kremer *et al.*, *Am. J. Hum. Genet.* **98**, 358–362 (2016).
17. C. Rabouille, V. Kondylis, *Genome Biol.* **7**, 213 (2006).
18. M. P. Milev *et al.*, *J. Inherit. Metab. Dis.* **44**, 426–437 (2021).
19. S. R. Lalani *et al.*, *Am. J. Hum. Genet.* **98**, 347–357 (2016).
20. T. G. Wolfsberg, P. Primakoff, D. G. Myles, J. M. White, *J. Cell Biol.* **131**, 275–278 (1995).
21. T. Klein, R. Bischoff, *J. Proteome Res.* **10**, 17–33 (2011).

22. S. Zhong, R. A. Khalil, *Biochem. Pharmacol.* **164**, 188–204 (2019).
23. P. Orth *et al*., *J. Mol. Biol.* **335**, 129–137 (2004).
24. S. Takeda, T. Igarashi, H. Mori, S. Araki, *EMBO J.* **25**, 2388–2396 (2006).
25. D. R. Flower, A. C. North, C. E. Sansom, *Biochim. Biophys. Acta* **1482**, 9–24 (2000).
26. H. E. Van Wart, H. Birkedal-Hansen, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 5578–5582 (1990).
27. M. Levy, A. H. Futerman, *IUBMB Life* **62**, 347–356 (2010).
28. J. L. Kim, B. Mestre, S.-H. Shin, A. H. Futerman, *Cell. Signal.* **82**, 109958 (2021).
29. E. Winter, C. P. Ponting, *Trends Biochem. Sci.* **27**, 381–383 (2002).
30. S. Spassieva *et al*., *J. Biol. Chem.* **281**, 33931–33938 (2006).
31. N. Vanni *et al*., *Ann. Neurol.* **76**, 206–212 (2014).
32. Q. Cong, I. Anishchenko, S. Ovchinnikov, D. Baker, *Science* **365**, 185–189 (2019).
33. Y. Zhang, J. Skolnick, *Proteins* **57**, 702–710 (2004).
34. C. R. Glassman *et al*., *Cell* **184**, 983–999.e24 (2021).
35. E. F. Pettersen *et al*., *Protein Sci.* **30**, 70–82 (2021).
36. M. Baek *et al*., RoseTTAFold: The first release of RoseTTAFold. Zenodo (2021); https://zenodo.org/record/5068265.

ACKNOWLEDGMENTS

We thank E. Horvitz, N. Hiranuma, D. Juergens, S. Mansoor, and D. Tischer for helpful discussions; D. E. Kim for web-server construction; and L. Goldschmidt for computing resource management. T.P.-K. thanks B. Nidetzky and M. Monschein from Graz University of Technology for providing protein samples for crystallization. D.J.O. acknowledges assistance with data collection from scientists of Diamond Light Source beamline I04 under proposal mx20303. T.S., C.B., and T.P.-K. acknowledge the ESRF (ID30-3, Grenoble, France) and DESY (P11, PETRAIII, Hamburg, Germany) for provision of synchrotron-radiation facilities and support during data collection. P.D.A., J.H.P., A.D., and A.V.R. acknowledge support from the Joint BioEnergy Institute, which is supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research under contract no. DE-AC02-05CH11231 between LBNL and the US Department of Energy. **Funding:** This work was supported by Microsoft (M.B., D.B., and generous gifts of Azure compute time and expertise); Open Philanthropy (D.B. and G.R.L.); E. and W. Schmidt by recommendation of the Schmidt Futures program (F.D. and H.P.); The Washington Research Foundation (M.B., G.R.L., and J.W.); the National Science Foundation Cyberinfrastructure for Biological Research, award no. DBI 1937533 (I.A.); Wellcome Trust grant number 209407/Z/17/Z (R.J.R.); the National Institutes of Health, grant numbers P01GM063210 (P.D.A. and R.J.R.), DP5OD026389 (S.O.), R01-AI51321 (K.C.G.), and GM127390 (N.V.G.); the Mathers Foundation (K.C.G.); the Canadian Institute of Health Research (CIHR) Project Grant, grant numbers 168998 (J.E.B.) and 168907 (C.K.Y.); the Welch Foundation I-1505 (N.V.G.); the Global Challenges Research Fund (GCRF) through Science & Technology Facilities Council (STFC), grant number ST/R002754/1: Synchrotron Techniques for African Research and Technology (START) (D.J.O., A.A.v.D., and A.C.E.); and the Austrian Science Fund (FWF), projects P29432 and DOC50 (doc.fund Molecular Metabolism) (T.S., C.B., and T.P.-K.). **Author contributions:** M.B., F.D., and D.B. designed the research; M.B., F.D., I.A., J.D., S.O., and J.W. developed the deep-learning network; G.R.L. and H.P. analyzed GPCR modeling results; Q.C., L.N.K., R.D.S., and N.V.G. analyzed modeling results for proteins related to the human diseases; C.R.G. and K.C.G. analyzed modeling results for the IL-12R–IL-12 complex; P.D.A., R.J.R., C.A., F.D., and C.M. worked on structure determination; A.A.v.D., A.C.E., D.J.O., T.S., C.B., T.P.-K., M.K.R., U.D., C.K.Y., J.E.B., A.D., J.H.P., and A.V.R. provided experimental data; M.B., F.D., G.R.L., Q.C., L.N.K., H.P., C.R.G., P.D.A., R.J.R., and D.B. wrote the manuscript; and all authors discussed the results and commented on the manuscript. **Competing interests:** The 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. Coordinates for the full PI3K complex structure determined by cryo-EM are available at the PDB with accession code PDB: 7MEZ. Model structures used for molecular replacement are available at http://files.ipd.uw.edu/pub/RoseTTAFold/MR_models.tar.gz. The refined 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*). This research was funded in whole or in part by Wellcome Trust, grant #209407/Z/17/Z, a cOAlition S organization. The author will make the Author Accepted Manuscript (AAM) version available under a CC BY public copyright license.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6557/871/suppl/DC1
Materials and Methods
Figs. S1 to S17
Tables S1 to S4
References (*37*–*82*)
MDAR Reproducibility Checklist

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

7 June 2021; accepted 7 July 2021
Published online 15 July 2021
10.1126/science.abj8754