nature

https://doi.org/10.1038/s41586-021-03828-1

Accelerated Article Preview

Highly accurate protein structure prediction for the human proteome

Received: 11 May 2021

Accepted: 16 July 2021

Accelerated Article Preview Published online 22 July 2021

Cite this article as: Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* https://doi.org/10.1038/s41586-021-03828-1 (2021).

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J. Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A. A. Kohl, Anna Potapenko, Andrew J. Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W. Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper & Demis Hassabis

Article

Highly accurate protein structure prediction for the human proteome

https://doi.org/10.1038/s41586-021-03828-1

Received: 11 May 2021

Accepted: 16 July 2021

Published online: 22 July 2021

Kathryn Tunyasuvunakool[1]✉, Jonas Adler[1], Zachary Wu[1], Tim Green[1], Michal Zielinski[1], Augustin Žídek[1], Alex Bridgland[1], Andrew Cowie[1], Clemens Meyer[1], Agata Laydon[1], Sameer Velankar[2], Gerard J. Kleywegt[2], Alex Bateman[2], Richard Evans[1], Alexander Pritzel[1], Michael Figurnov[1], Olaf Ronneberger[1], Russ Bates[1], Simon A. A. Kohl[1], Anna Potapenko[1], Andrew J. Ballard[1], Bernardino Romera-Paredes[1], Stanislav Nikolov[1], Rishub Jain[1], Ellen Clancy[1], David Reiman[1], Stig Petersen[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Ewan Birney[2], Pushmeet Kohli[1], John Jumper[1,3]✉ & Demis Hassabis[1,3]✉

Protein structures can provide invaluable information, both for reasoning about biological processes and for enabling interventions such as structure-based drug development or targeted mutagenesis. After decades of effort, 17% of the total residues in human protein sequences are covered by an experimentally-determined structure[1]. Here we dramatically expand structural coverage by applying the state-of-the-art machine learning method, AlphaFold[2], at scale to almost the entire human proteome (98.5% of human proteins). The resulting dataset covers 58% of residues with a confident prediction, of which a subset (36% of all residues) have very high confidence. We introduce several metrics developed by building on the AlphaFold model, and use them to interpret the dataset, identifying strong multi-domain predictions as well as regions likely to be disordered. Finally, we provide some case studies illustrating how high-quality predictions may be used to generate biological hypotheses. Importantly, we are making our predictions freely available to the community via a public database (hosted by the European Bioinformatics Institute at https://alphafold.ebi.ac.uk/). We anticipate that routine large-scale and high-accuracy structure prediction will become an important tool, allowing new questions to be addressed from a structural perspective.

The monumental success of the human genome project revealed new worlds of protein coding genes, and many researchers set out to map these proteins to their structures[3,4]. Thanks to the efforts of individual labs and dedicated structural genomics initiatives, over 50,000 human protein structures have now been deposited, making *Homo sapiens* by far the best represented species in the Protein Data Bank (PDB)[5]. Despite this intensive study, only 35% of human proteins map to a PDB entry, and in many cases the structure covers just a fragment of the sequence[6]. Experimental structure determination requires overcoming many time-consuming hurdles: the protein must be produced in sufficient quantities, purified, appropriate sample preparation conditions chosen, and high-quality datasets collected. A target may prove intractable at any stage, and depending on the chosen method, properties like protein size, the presence of transmembrane regions, presence of disorder, or susceptibility to conformational change can be a hindrance[7,8]. As such, full structural coverage of the proteome remains an outstanding challenge.

Protein structure prediction contributes to closing this gap by providing actionable structural hypotheses quickly and at scale. Prior large-scale structure prediction work has addressed protein families[9–12], specific functional classes[13,14], domains identified within whole proteomes[15], and in some cases full chains or complexes[16,17]. In particular, projects like SWISS-MODEL Repository, Genome3D, and ModBase have made valuable contributions by providing access to large numbers of structures and encouraging their free use by the community[17–19]. Related protein bioinformatics fields have developed alongside structure prediction, including protein design[20,21], function annotation[22–24], disorder prediction[25], and domain identification and classification[26–28]. While some of our analyses are inspired by this prior work, here we focus mainly on structural investigations for which scale and accuracy are particularly enabling.

Structure prediction has seen substantial progress in recent years, as evidenced by the results of the biennial Critical Assessment of protein Structure Prediction (CASP)[29,30]. In particular, the latest version of AlphaFold was entered in CASP14 under the team name *"AlphaFold2"*. This system used a completely different model from our CASP13 entry[31], and demonstrated a considerable improvement over previous methods in terms of providing routinely high accuracy[29,30]. Backbone predictions with sub-Ångström Root Mean Square Distance (RMSD-Cα) are now common, and side chains are increasingly accurate[2]. Good results can often be achieved even for challenging proteins without a template structure in the PDB, or with relatively few related sequences

[1]DeepMind, London, UK. [2]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. [3]These authors contributed equally: John Jumper, Demis Hassabis.
✉e-mail: ktkool@deepmind.com; jumper@deepmind.com; dhcontact@deepmind.com

to build a Multiple Sequence Alignment (MSA)[2]. These improvements are significant because more accurate models permit a wider range of applications: not only homology search and putative function assignment, but also molecular replacement and druggable pocket detection, for instance[32–34]. In light of this we applied the current state-of-the-art method, AlphaFold, to the human proteome.

Model confidence and added coverage

We predicted structures for the UniProt human reference proteome (one representative sequence per gene), with an upper length limit of 2700 residues[6]. The final dataset covers 98.5% of human proteins with a full-chain prediction.

For the resulting predictions to be practically useful, they must come with a well-calibrated and sequence-resolved confidence measure. The latter point is particularly important when predicting full chains, as we expect to see high confidence on domains but low confidence on linkers and unstructured regions (see Extended Data Fig. 1 for examples). To this end, AlphaFold produces a per-residue confidence metric called predicted lDDT-Cα (pLDDT) on a scale from 0 to 100. pLDDT estimates how well the prediction would agree with an experimental structure based on the Local Distance Difference Test[35]. It has been shown to be well-calibrated (Fig. 1a; see also Extended Data Fig. 2, Extended Data Table 1) and full details on how pLDDT is produced are given in the Supplementary Information (SI) of the AlphaFold paper[2].

We consider a prediction highly accurate when, in addition to a good backbone prediction, the side chains are frequently correctly oriented. On this basis, pLDDT > 90 is taken as the high accuracy cutoff, above which AlphaFold Chi-1 rotamers are 80% correct on a recent PDB test set (Extended Data Fig. 3). A lower cutoff of pLDDT > 70 corresponds to a generally correct backbone prediction (Extended Data Table 2). AlphaFold's accuracy on an example protein within a number of pLDDT bands is illustrated in Fig. 1b.

On the human proteome, 35.7% of total residues fall within the highest accuracy band (corresponding to 38.6% of residues for which a prediction was produced; Fig. 1c). This is double the number of residues covered by an experimental structure. 58.0% of total residues are predicted confidently (pLDDT > 70), implying that we also add substantial coverage for sequences without a good template in PDB (sequence identity below 30%). At the per-protein level, 43.8% of proteins have a confident prediction on at least three quarters of their sequence, while 1,290 proteins contain a substantial region (over 200 residues) with pLDDT ≥ 70 and no good template.

The dataset adds high-quality structural models across a broad range of Gene Ontology (GO) terms[36,37], including pharmaceutically-relevant classes such as enzymes and membrane proteins[38] (Fig. 1d). Membrane proteins in particular are generally underrepresented in the PDB because they have historically been challenging experimental targets. This shows that AlphaFold is able to produce confident predictions even for protein classes that are not abundant wthin its training set.

We note that AlphaFold's accuracy was validated in CASP14[2], which focuses on challenging proteins dissimilar to the structures already in the PDB. By contrast, many human proteins have templates with high sequence identity. To evaluate AlphaFold's applicability in this regime, we predicted structures for one year of targets from the Continuous Automated Model Evaluation (CAMEO) benchmark[39,40], a structure-prediction assessment that measures a wider range of difficulties. We find that AlphaFold adds substantial accuracy over CAMEO's *BestSingleStructuralTemplate* baseline across a wide range of levels of template identity (Extended Data Fig. 4).

Full chain prediction

Many previous large-scale structure prediction efforts have focused on domains—regions of the sequence that fold independently[9–11,15]. Here we process full-length protein chains. There are several motivations. Restricting the prediction to pre-identified domains risks missing structured regions that have yet to be annotated. It also discards contextual information from the rest of the sequence, which might be useful where two or more domains interact substantially. Finally, the full-chain approach lets the model attempt an interdomain packing prediction.

Interdomain accuracy was assessed at CASP14, and AlphaFold outperformed other methods[41]. However, the assessment was based on a small target set. To further evaluate AlphaFold on long multi-domain proteins, we compiled a test set of recent PDB chains not in the model's training set. Only chains with > 800 resolved residues were included, and a template filter was applied (see Methods). Performance on this set was evaluated using the Template Modelling score (TM-score[42]), which should better reflect global as opposed to per-domain accuracy. The results were encouraging, with 70% of predictions having a TM-score > 0.7 (Fig. 2a).

The AlphaFold paper SI describes how a variety of useful predictors can be built on top of the main model. In particular, we can predict the residues likely to be experimentally resolved, and use them to produce a predicted TM-score (pTM), where each residue's contribution is weighted by the probability of it being resolved (Suppl. Methods 1). The motivation for the weighting is to downweight unstructured parts of the prediction, producing a metric that better reflects the model's confidence about the packing of the structured domains that are present. On the same recent PDB test set, pTM correlates well with the actual TM-score (Pearson's r = 0.84, Fig. 2b). Interestingly, while some outliers on this plot are genuine failure cases, others appear to be plausible alternate conformations (Fig. 2b; 6OFS_A[43]).

We computed pTM scores for the human proteome, in an effort to identify multi-domain predictions that might feature novel domain packings. The criteria applied were a pLDDT > 70 on at least 600 residues constituting over half the sequence, with no template hit covering over half the sequence. The distribution of pTM scores after applying the above filters is shown in Fig. 2c. Note that we would not expect uniformly high TM scores to be achievable on this set, since some proteins will contain domains mobile relative to each other, with no fixed packing. Of the set, 187 proteins have pTM > 0.8 and 343 have pTM > 0.7. While we expect AlphaFold's interdomain accuracy to be lower than its within-domain accuracy, this set should nonetheless be enriched for interesting multi-domain predictions, suggesting that the dataset provides on the order of hundreds of these. Four examples, the predictions with the highest number of confident residues subject to pTM > 0.8, are shown in Fig. 2d.

Highlighted predictions

We next discuss some case study predictions and the insights they may provide. All predictions presented are *de novo*, lacking any template with 25% sequence identity or more covering 20% of the sequence. Our discussion concerns biological hypotheses, which would ultimately need to be confirmed by experimental studies.

Glucose-6-phosphatase

G6Pase-α (UniProt P35575) is a membrane-bound enzyme that catalyzes the final step in glucose synthesis; it is therefore of critical importance to maintaining blood sugar levels. No experimental structure exists, but previous studies have attempted to characterize the transmembrane topology[44] and active site[45]. Our prediction has very high confidence (median pLDDT 95.5) and gives a nine-helix topology with the putative active site accessible via an entry tunnel roughly in line with the endoplasmic reticulum (ER) surface (Fig. 3a, Suppl. Video 1). Positively-charged residues in our prediction (median pLDDT 96.6) align closely with the previously identified active site homologue in a fungal vanadium chloroperoxidase (PDB 1IDQ; RMSD 0.56 Å, 49/51 aligned atoms)[46]. Since these enzymes have distinct functions, we

probed our prediction for clues about substrate specificity. In the G6Pase-α binding pocket face, opposite the residues shared with the chloroperoxidase, we predict a conserved glutamate (Glu110) that is also present in our G6Pase-β prediction (Glu105) but not the chloroperoxidase (Fig. 3a). The glutamate could stabilize the binding pocket in a closed conformation, forming salt bridges with positively charged residues there. It is also the most solvent-exposed residue of the putative active site, suggesting a possible gating function. To our knowledge this residue has not been discussed previously and illustrates the novel mechanistic hypotheses that can be obtained from high quality structure predictions.

Diacylglycerol O-acyltransferase 2

Triacylglycerol synthesis is responsible for storing excess metabolic energy as fat in adipose tissue. DGAT2 (UniProt Q96PD7) is one of two essential acyltransferases catalyzing the final acyl addition in this pathway, and inhibiting DGAT2 has been shown to improve liver function in mouse models of liver disease[47]. With our highly confident predicted structure (median pLDDT 95.9), we set out to identify the binding pocket for a known inhibitor, PF-06424439 (ref. [48]). We identified a pocket (median pLDDT 93.7) in which we were able to dock the inhibitor and observe specific interactions (Fig. 3b) that were not recapitulated in a negative example[49] (Suppl. Methods 2, Extended Data Fig. 5). DGAT2 has an evolutionarily-divergent but biochemically similar analog, diacylglycerol O-acyltransferase 1 (DGAT1)[50]. Within DGAT2's binding pocket, we identified residues (Glu243, His163; Fig. 3b) analogous to the proposed catalytic residues in DGAT1 (His415, Glu416)[51], although we note that the nearby Ser244 in DGAT2 may present an alternative mechanism via an acyl-enzyme intermediate. Previous experimental work with DGAT2 has shown that mutating His163 has a stronger deleterious effect than mutating a histidine two residues away[52]. Additionally, Glu243 and His163 are conserved across species[50], supporting this hypothesized catalytic geometry.

Wolframin

Wolframin (UniProt O76024) is a transmembrane protein localized to the ER. Mutations in the *WFS1* gene are associated with Wolfram syndrome 1, a neurodegenerative disease characterized by early-onset diabetes, gradual visual and hearing loss, and early death[53,54]. Given the lower confidence in our full prediction (median pLDDT 81.7, Fig. 3c), we proposed identifying regions unique to this structure. A recent evolutionary analysis suggested domains for wolframin, which our prediction largely supports[55]. An interesting distinction is the incorporation of a cysteine-rich domain (Fig. 3c, yellow) to the oligonucleotide binding (OB) fold (Fig. 3c, green and yellow) as the characteristic β1 strand[56]. The cysteine-rich region then forms an extended L12 loop with two predicted disulfide bridges, before looping back to the prototypical β2 strand. Comparing our prediction for this region (median pLDDT 86.0) against existing PDB chains using TM-align[42,57] identified 3F1Z[58] as the most similar known chain (TM-score 0.472; Fig. 3c, magenta). Despite being the most similar chain, 3F1Z lacks the cysteines present in wolframin, which could form disulfide cross-links in the ER[59]. As this region is hypothesized to recruit other proteins[55], these structural insights are likely important to understanding its partners.

Regions without a confident prediction

Since we are applying AlphaFold to the entire human proteome, we would expect a significant percentage of residues to be contained in regions that are always or sometimes disordered in solution. Disorder is common in the proteomes of eukaryotes[60,61], and one prior work[62] estimates the percentage of disordered residues in the human proteome at 37-50%. Thus disorder will play a large role when we consider a comprehensive set of predictions covering an entire proteome.

Furthermore, we observed a large difference in the pLDDT distribution between resolved and unresolved residues in PDB sequences (Fig. 4a). To investigate this connection, we evaluated pLDDT as a disorder predictor on the Critical Assessment of protein Intrinsic Disorder prediction (CAID) benchmark set[25]. The results showed pLDDT to be a competitive disorder predictor compared to the current state of the art (SPOT-Disorder2[63]), with an AUC of 0.897 (Fig. 4b). Moreover, the AlphaFold paper SI describes an "experimentally resolved head", which is specifically trained for the task of predicting whether a residue will be resolved in an experimental structure. The experimentally resolved head performed even better on the CAID benchmark, with an AUC of 0.921.

These disorder-prediction results suggest that a significant percentage of low-confidence residues may be explained by some form of disorder, but we caution that this could encompass both regions that are intrinsically disordered and regions that are structured only in complex. A potential example of the latter scenario drawn from recent PDB is shown in Fig. 4c; chain C interacts extensively with the rest of the complex, such that the interface region would be unlikely to adopt the same structure outside of this context. In a systematic analysis of recent PDB chains, we observed that AlphaFold has much lower accuracy for regions where the chain has a high percentage of heterotypic, cross-chain contacts (see Fig. 4d).

In summary, our current interpretation of regions where AlphaFold exhibits low pLDDT is that they have significant likelihood of being unstructured in isolation. In the current dataset, long regions with pLDDT < 50 adopt a readily identifiable ribbon-like appearance, and should not be interpreted as structures but rather as a prediction of disorder.

Discussion

In this work we generated comprehensive, state-of-the-art structure predictions for the human proteome. The resulting dataset makes a large contribution to structural coverage of the proteome; particularly for tasks where high accuracy is advantageous like molecular replacement or the characterisation of binding sites. We have also applied several metrics produced by building on the AlphaFold architecture - pLDDT, pTM, and the experimentally resolved head - demonstrating how they can be used to interpret our predictions.

While we have presented several case studies to illustrate the type of insights that may be gained from these data, we recognize that there is still much more to uncover. By making our predictions available to the community we hope to enable exploration of new directions in structural bioinformatics.

The parts of the human proteome still without a confident prediction represent directions for future research. Some proportion of these will be genuine failures, where a fixed structure exists but the current version of AlphaFold does not predict it. In many other cases, where the sequence is unstructured in isolation, the problem arguably falls outside the scope of single-chain structure prediction. It will be crucial to develop new methods that can address the biology of these regions, for example, by predicting the structure in complex or by predicting a distribution over possible states in the cellular milieu.

Finally, we note that the importance of the human proteome for health and medicine has led to it being intensively studied from a structural perspective. Other organisms are much less well represented in the PDB, including biologically significant, medically relevant, or economically important species. Structure prediction may have a more profound impact on the study of these organisms, for which fewer experimental structures are available. Looking beyond the proteome scale, the UniProt database contains hundreds of millions of proteins that have so far been addressed mainly by sequence-based methods, and for which the easy availability of structures might open up entirely new avenues of investigation. By providing scalable structure prediction with hitherto unprecedented accuracy, AlphaFold could enable

an exciting shift toward structural bioinformatics, further illuminating protein space.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03828-1.

1. SWISS-MODEL *Homo sapiens* (Human). https://swissmodel.expasy.org/repository/species/9606 (2021).
2. Jumper, J. *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature* https://doi.org/10.1038/s41586-021-03819-2 (2021).
3. Lander, E. S. *et al*. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
4. Venter, J. C. *et al*. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
5. wwPDB Consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res*. **47**, D520–D528 (2018).
6. Bateman, A. *et al*. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res*. (2020).
7. Slabinski, L. *et al*. The challenge of protein structure determination—lessons from structural genomics. *Protein Sci*. **16**, 2472–2482 (2007).
8. Elmlund, D., Le, S. N. & Elmlund, H. High-resolution cryo-EM: the nuts and bolts. *Curr. Opin. Struct. Biol*. **46**, 1–6 (2017).
9. Yang, J. *et al*. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl Acad. Sci. USA* **117**, 1496–1503 (2020).
10. Greener, J. G., Kandathil, S. M. & Jones, D. T. Deep learning extends *de novo* protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat. Commun*. **10**, 1–13 (2019).
11. Michel, M., Menéndez Hurtado, D., Uziela, K. & Elofsson, A. Large-scale structure prediction by improved contact predictions and model quality assessment. *Bioinformatics* **33**, i23–i29 (2017).
12. Ovchinnikov, S. *et al*. Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife* **4**, e09248 (2015).
13. Zhang, J., Yang, J., Jang, R. & Zhang, Y. GPCR-I-TASSER: a hybrid approach to G protein-coupled receptor structure modeling and the application to the human genome. *Structure* **23**, 1538–1549 (2015).
14. Bender, B. J., Marlow, B. & Meiler, J. Improving homology modeling from low-sequence identity templates in Rosetta: A case study in GPCRs. *PLoS Comput. Biol*. **16**, e1007597 (2020).
15. Drew, K. *et al*. The Proteome Folding Project: proteome-scale prediction of structure and function. *Genome Res*. **21**, 1981–1994 (2011).
16. Xu, D. & Zhang, Y. *Ab initio* structure prediction for *Escherichia coli*: towards genome-wide protein structure modeling and fold assignment. *Sci. Rep*. **3**, 1–11 (2013).
17. Waterhouse, A. *et al*. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res*. **46**, W296–W303 (2018).
18. Sillitoe, I. *et al*. Genome3D: Integrating a collaborative data pipeline to expand the depth and breadth of consensus protein structure annotation. *Nucleic Acids Res*. **48**, D314–D319 (2020).
19. Pieper, U. *et al*. MODBASE: A database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res*. **42**, D336–D346 (2014).
20. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of *de novo* protein design. *Nature* **537**, 320–327 (2016).
21. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol*. **20**, 681–697 (2019).
22. Consortium, G. O. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res*. **47**, D330–D338 (2019).
23. Zhou, N. *et al*. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol*. **20**, 1–23 (2019).
24. Gligorijević, V. *et al*. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun*. **12**, 1–14 (2021).
25. Necci, M., Piovesan, D. & Tosatto, S. C. E. Critical assessment of protein intrinsic disorder prediction. *Nat. Methods* 1–10 (2021).
26. Sillitoe, I. *et al*. CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res*. **47**, D280–D284 (2019).
27. Andreeva, A., Kulesha, E., Gough, J. & Murzin, A. G. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res*. **48**, D376–D382 (2020).
28. Mistry, J. *et al*. Pfam: The protein families database in 2021. *Nucleic Acids Res*. **49**, D412–D419 (2021).
29. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Struct. Funct. Bioinf*. **87**, 1011–1020 (2019).
30. Lupas, A. N., Pereira, J. & Hartmann, M. D. High Accuracy Assessment in CASP14. https://predictioncenter.org/casp14/doc/presentations/2020_11_30_HighAccuracy_assessment_Lupas-Pereira-Hartmann.pdf (2020).
31. Senior, A. W. *et al*. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
32. Zhang, Y. Protein structure prediction: when is it useful? *Curr. Opin. Struct. Biol*. **19**, 145–155 (2009).
33. Flower, T. G. & Hurley, J. H. Crystallographic molecular replacement using an in silico-generated search model of SARS-CoV-2 ORF8. *Protein Sci*. **30**, 728–734 (2021).
34. Egbert, M. & Vajda, S. CASP14 Functional Assessment: Conservation of Binding Properties in Modeled Proteins. https://predictioncenter.org/casp14/doc/presentations/2020_12_03_Function_Assessment_VajdaLab_KozakovLab.pdf (2020).
35. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).
36. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat. Genet*. **25**, 25–29 (2000).
37. Carbon, S. *et al*. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res*. **49**, D325–D334 (2021).
38. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat. Rev. Drug Discov*. **1**, 727–730 (2002).
39. Haas, J. *et al*. Introducing 'best single template' models as reference baseline for the Continuous Automated Model *Evaluation (CAMEO). Proteins: Struct. Funct. Bioinf*. **87**, 1378–1387 (2019).
40. Haas, J. *et al*. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins: Struct. Funct. Bioinf*. **86**, 387–398 (2018).
41. Schaeffer, R. D., Kinch, L. & Grishin, N. CASP14: InterDomain Performance. https://predictioncenter.org/casp14/doc/presentations/2020_12_02_Interdomain_assessment1_Schaeffer.pdf (2020).
42. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Struct. Funct. Bioinf*. **57**, 702–710 (2004).
43. Grinter, R. *et al*. Protease-associated import systems are widespread in Gram-negative bacteria. *PLoS Genet*. **15**, e1008435 (2019).
44. Pan, C.-J., Lei, K.-J., Annabi, B., Hemrika, W. & Chou, J. Y. Transmembrane topology of glucose-6-phosphatase. *J. Biol. Chem*. **273**, 6144–6148 (1998).
45. Van Schaftingen, E. & Gerin, I. The glucose-6-phosphatase system. *Biochem. J*. **362**, 513–532 (2002).
46. Messerschmidt, A., Prade, L. & Wever, R. Implications for the catalytic mechanism of the vanadium-containing enzyme chloroperoxidase from the fungus *Curvularia inaequalis* by X-ray structures of the native and peroxide form. *Biol. Chem*. **378**, 309–315 (1997).
47. Amin, N. B. *et al*. Targeting diacylglycerol acyltransferase 2 for the treatment of nonalcoholic steatohepatitis. *Sci. Transl. Med*. **11**, (2019).
48. Futatsugi, K. *et al*. Discovery and optimization of imidazopyridine-based inhibitors of diacylglycerol acyltransferase 2 (DGAT2). *J. Med. Chem*. **58**, 7173–7185 (2015).
49. Birch, A. M. *et al*. Discovery of a potent, selective, and orally efficacious pyrimidinooxazinyl bicyclooctaneacetic acid diacylglycerol acyltransferase-1 inhibitor. *J. Med. Chem*. **52**, 1558–1568 (2009).
50. Cao, H. Structure-function analysis of diacylglycerol acyltransferase sequences from 70 organisms. *BMC Res. Notes* **4**, 1–24 (2011).
51. Wang, L. *et al*. Structure and mechanism of human diacylglycerol O-acyltransferase 1. *Nature* **581**, 329–332 (2020).
52. Stone, S. J., Levin, M. C. & Farese, R. V., Jr Membrane topology and identification of key functional amino acid residues of murine acyl-CoA: diacylglycerol acyltransferase-2. *J. Biol. Chem*. **281**, 40273–40282 (2006).
53. Rigoli, L., Lombardo, F. & Di Bella, C. Wolfram syndrome and *WFS1* gene. *Clin. Genet*. **79**, 103–117 (2011).
54. Urano, F. Wolfram syndrome: diagnosis, management, and treatment. *Curr. Diab. Rep*. **16**, 6 (2016).
55. Schäffer, D. E., Iyer, L. M., Burroughs, A. M. & Aravind, L. Functional innovation in the evolution of the calcium-dependent system of the eukaryotic endoplasmic reticulum. *Front. Genet*. **11**, 34 (2020).
56. Guardino, K. M., Sheftic, S. R., Slattery, R. E. & Alexandrescu, A. T. Relative stabilities of conserved and non-conserved structures in the OB-fold superfamily. *Int. J. Mol. Sci*. **10**, 2412–2430 (2009).
57. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*. **33**, 2302–2309 (2005).
58. Das, D. *et al*. The structure of KPN03535 (gi|152972051), a novel putative lipoprotein from *Klebsiella pneumoniae*, reveals an OB-fold. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun*. **66**, 1254–1260 (2010).
59. Fass, D. & Thorpe, C. Chemistry and enzymology of disulfide cross-linking in proteins. *Chem. Rev*. **118**, 1169–1198 (2018).
60. Basile, W., Salvatore, M., Bassot, C. & Elofsson, A. Why do eukaryotic proteins contain more intrinsically disordered regions? *PLoS Comput. Biol*. **15**, e1007186 (2019).
61. Bhowmick, A. *et al*. Finding our way in the dark proteome. *J. Am. Chem. Soc*. **138**, 9730–9742 (2016).
62. Oates, M. E. *et al*. D2P2: database of disordered protein predictions. *Nucleic Acids Res*. **41**, D508–D516 (2012).
63. Hanson, J., Paliwal, K. K., Litfin, T. & Zhou, Y. SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning. *Genomics Proteomics Bioinformatics* **17**, 645–656 (2019).
64. Dunne, M., Ernst, P., Sobieraj, A., Pluckthun, A. & Loessner, M. J. The M23 peptidase domain of the Staphylococcal phage 2638A endolysin. https://doi.org/10.2210/pdb6YJ1/pdb (2020).
65. Krivák, R. & Hoksza, D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J. Cheminform*. **10**, 1–12 (2018).
66. Li, Y.-C. *et al*. Structure and noncanonical Cdk8 activation mechanism within an Argonaute-containing Mediator kinase module. *Sci. Adv*. **7**, eabd4484 (2021).

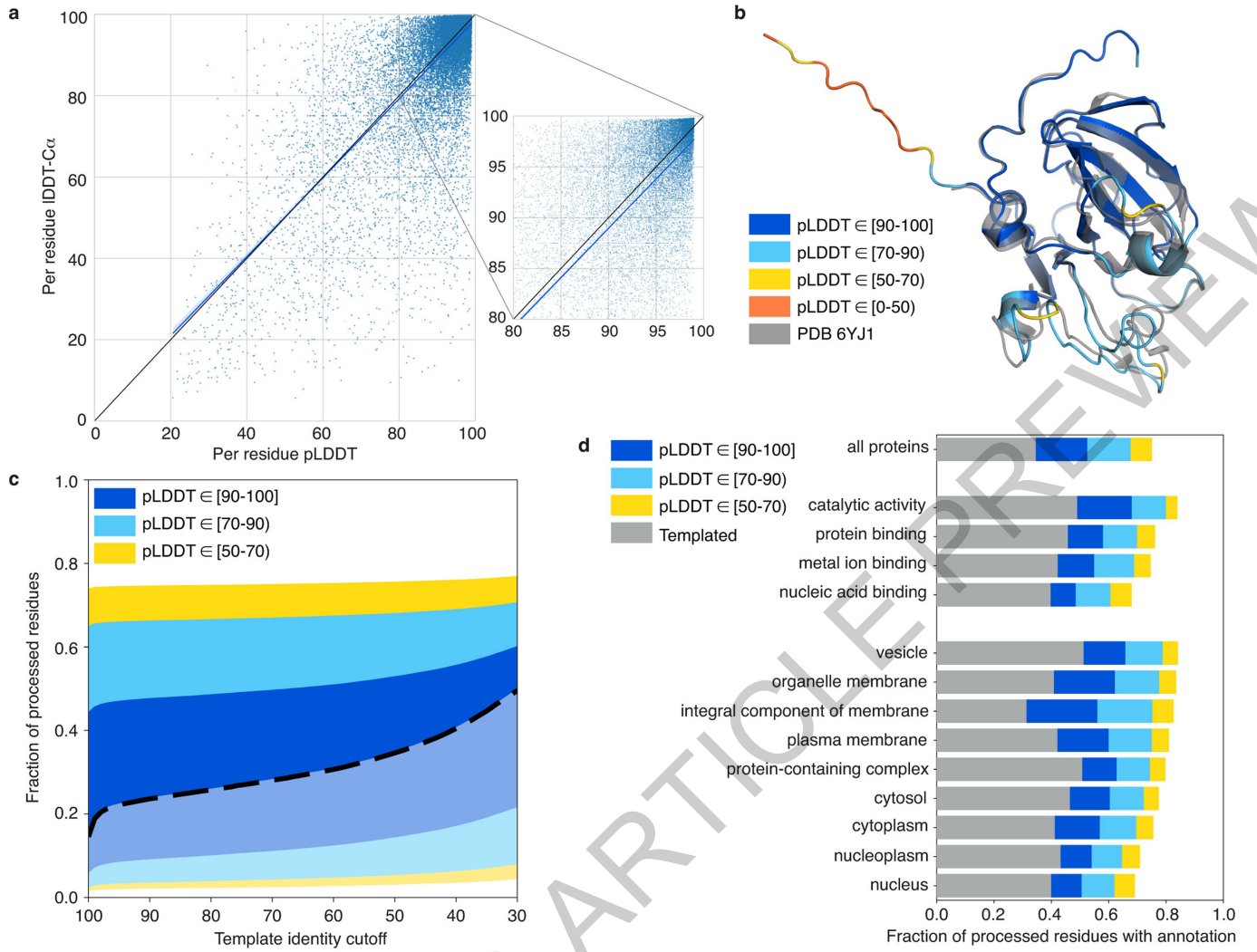


Fig. 1 | Model confidence and added coverage. a, Correlation between per-residue pLDDT and lDDT-Cɑ. Based on a held-out set of recent PDB chains (see Methods) filtered to those with a reported resolution < 3.5Å (N = 10,215 chains, 2,756,569 residues). Scatterplot shows a subsample (1% of residues), while correlation (Pearson's r = 0.73) is computed on the full dataset. Least-squares linear fit for the full dataset: $y = (0.967 +/- 0.001) * x + (1.9 +/- 0.1)$. Shaded region of the linear fit represents a 95% confidence interval estimated with 1,000 bootstrap samples. **b**, AlphaFold prediction and experimental structure for a CASP14 target[64]. The prediction is coloured by model confidence band, and the N terminus is an expression tag included in CASP but unresolved in the PDB structure. **c**, AlphaFold model confidence on all residues for which a prediction was produced (N = 10,537,122 residues). Residues covered by a template at the specified identity level are shown in a lighter colour. **d**, Added residue-level coverage of the proteome for high-level GO terms, on top of residues covered by a template with sequence identity > 50%. Based on the same dataset (N = 10,537,122 residues).

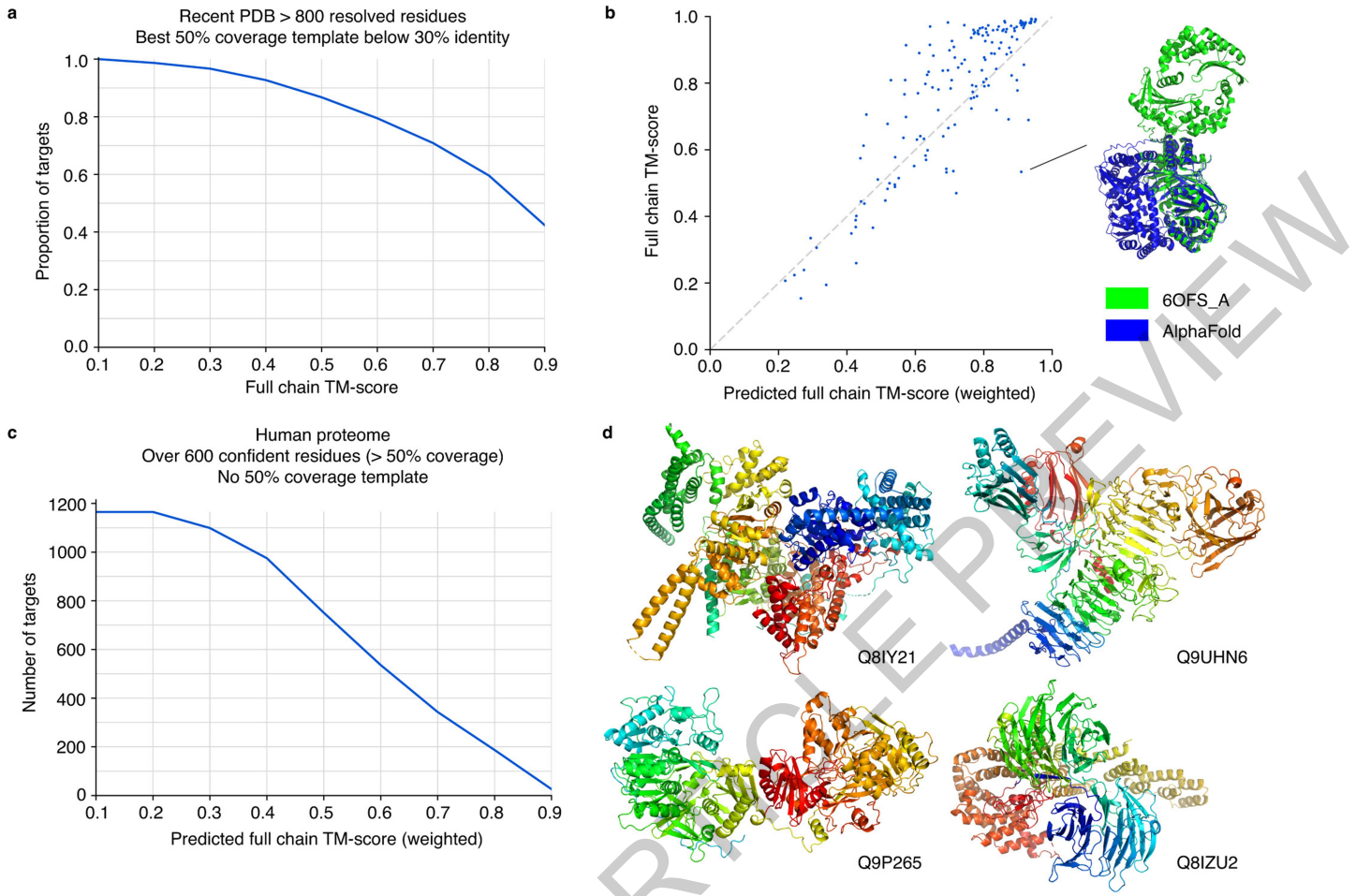


Fig. 2 | Full chain structure prediction. a, TM-score distribution for AlphaFold evaluated on a held-out set of template-filtered, long PDB chains (N = 151 chains). **b**, Correlation between full chain TM-score and pTM on the same set (N = 151 chains), Pearson's r = 0.84. The ground truth and predicted structure are shown for the most over-optimistic outlier. **c**, pTM distribution on a subset of the human proteome that we expect to be enriched for structurally-novel multidomain proteins (N = 1,165 chains). **d**, Four of the top hits from the set in panel **c**, filtering by pTM > 0.8 and sorting by number of confident residues. For clarity, regions with pLDDT < 50 are hidden, as are isolated smaller regions left after this cropping.

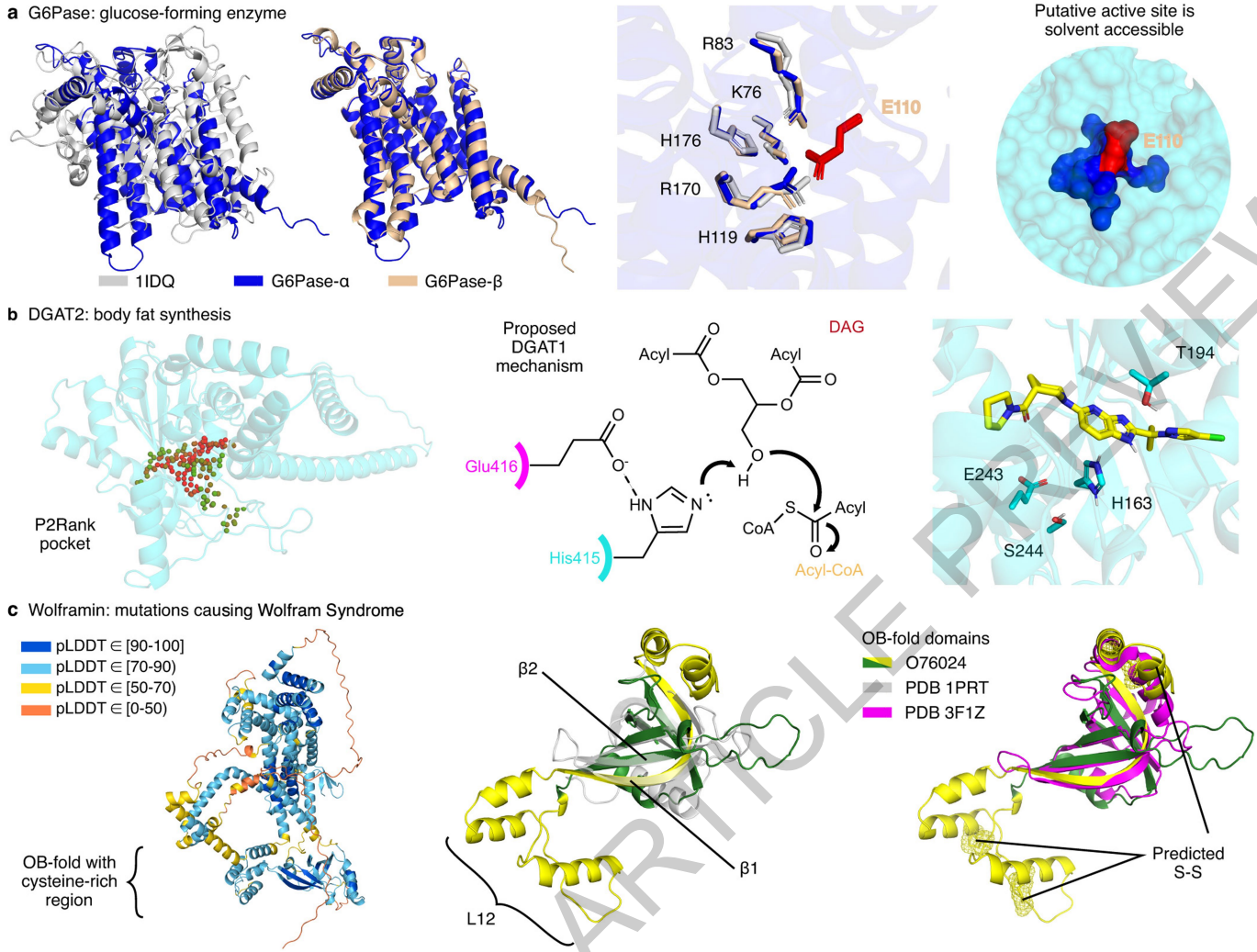


Fig. 3 | Highlighted structure predictions. a, Comparison of the active sites of two G6Pases and a chloroperoxidase. The G6Pases contain a conserved, solvent-accessible glutamate (red) opposite the shared active site residues. **b**, Pocket prediction (P2Rank[65]) identifies a putative binding pocket for DGAT2, where red/green spheres represent P2Rank's ligandability scores of 1 and 0, respectively. A proposed mechanism for DGAT1[51] activates the substrate with Glu416 and His415, which have analogous residues in the DGAT2 pocket. The docked inhibitor is well placed for polar interactions with His163 and Thr194. Middle figure reprinted by permission from Ming Zhou: Springer Nature.[51] **c**, While there are regions in wolframin with low pLDDT (leftmost panel), we could identify an OB-fold region (green/yellow), with a comparable core to a prototypical OB-fold (grey). However, the most similar PDB chain (magenta) lacks the conserved cysteine-rich region (yellow) of our prediction. This region forms the characteristic β1 strand and an extended L12 loop, and is predicted to contain three disulfide bridges (yellow mesh).

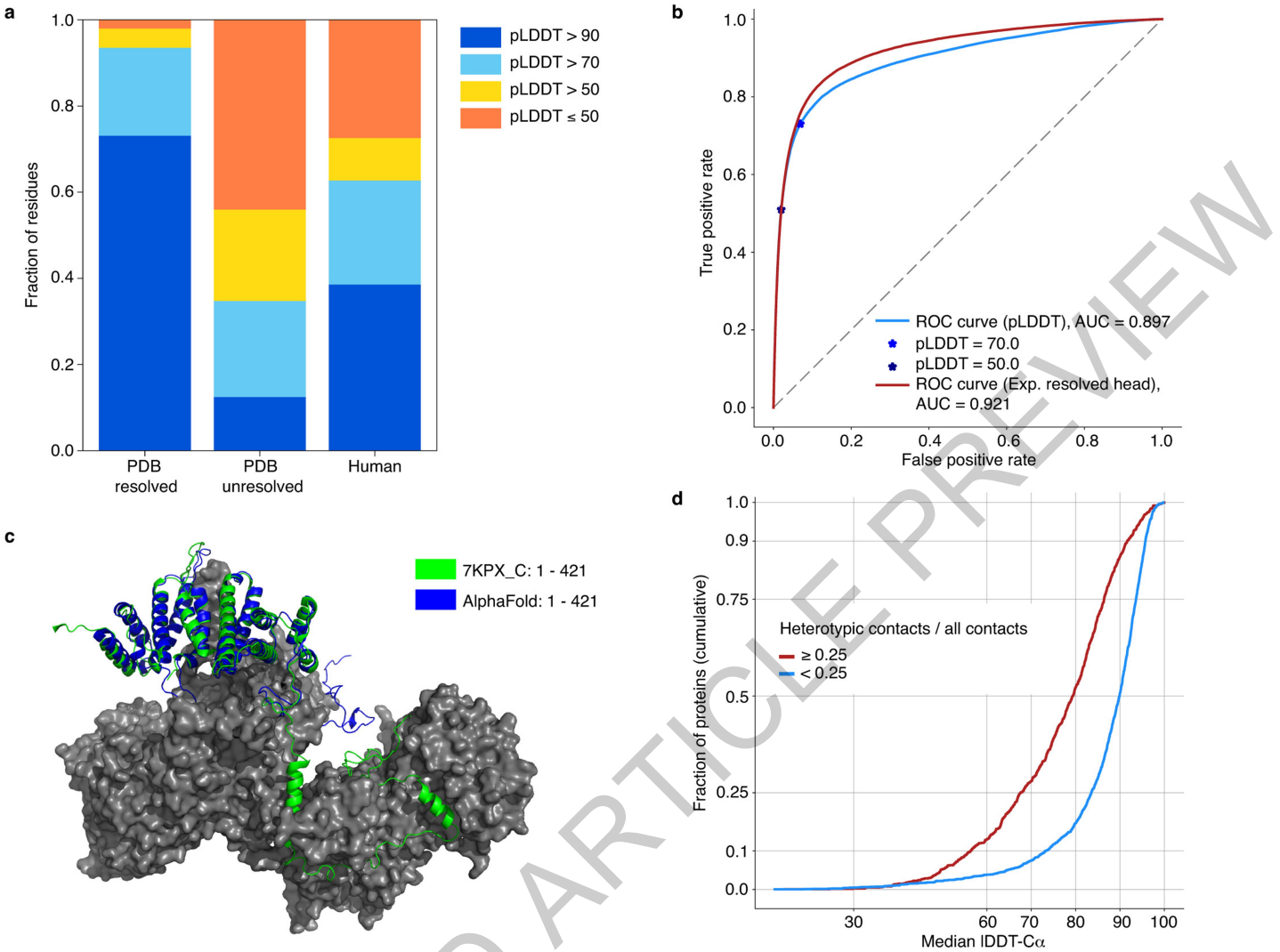


Fig. 4 | Low confidence regions. a, pLDDT distribution on the resolved parts of PDB sequences (N = 3,440,359 residues), the unresolved parts of PDB sequences (N = 589,079 residues), and the human proteome (N = 10,537,122 residues). **b**, Performance of pLDDT and AlphaFold's experimentally-resolved head as disorder predictors on the CAID Disprot-PDB benchmark set (N = 178,124 residues). **c**, An example low-confidence prediction from recent PDB[66]. The globular domain is well-predicted but the extended interface exhibits low pLDDT and is incorrect apart from some secondary structure. **d**, A high ratio of heterotypic contacts is associated with lower AlphaFold accuracy on the recent PDB set, restricted to proteins with < 40% of residues with template identity above 30 (N = 3,007 chains, see Methods). The ratio of heterotypic contacts is defined as: heterotypic / (intra-chain + homomeric + heterotypic).

Methods

Structure prediction (human proteome)

Sequences for the human reference proteome were obtained from UniProt release 2021_02[6]. Structure prediction was attempted for all sequences with 16–2700 amino acids, and sequences with residue codes B, J, O, U, Z or X are also excluded. The length ceiling of 2700 residues does not represent an absolute limit for the method, but was chosen to keep run times manageable. The structure prediction process was largely as described in the AlphaFold paper[2], consisting of five steps: MSA construction, template search, inference with five models, model-ranking based on mean pLDDT, and constrained relaxation of the predicted structures. The following differences were introduced for the proteome-scale pipeline. First, the search against the metagenomics database Big Fantastic Database (BFD) was replaced with a search against "Reduced BFD" using Jackhmmer from HMMER3[67,68]. Reduced BFD consists of a multiline FASTA containing the first non-consensus sequence from each BFD a3m. Second, the amount of ensembling was reduced by a factor of eight. At least four relaxed full-chain models were successfully produced for 20,296 sequences out of 20,614 FASTA entries, covering 98.5% of proteins. Sequences > 2700 residues account for the majority of exclusions. This amounts to 10,537,122 residues (92.5% of residues).

Structure prediction (recent PDB set)

For structure prediction on recent PDB sequences, we used a copy of the PDB downloaded 15/02/2021. Structures were filtered to those with a release date after 30/04/2018 (the date limit for inclusion in AlphaFold's training set). Chains were then further filtered to remove sequences consisting of a single amino acid, sequences with an ambiguous chemical component at any residue position, and sequences without a PDB 40% sequence clustering. Exact duplicates were removed by choosing the chain with the most resolved Cα atoms as the representative sequence. Then, structures with < 16 resolved residues, with unknown residues, and structures solved by nuclear magnetic resonance methods (NMR) were filtered out. Structure prediction then followed the same procedure as for the human proteome with the same length and residue limits, except that templates with a release date after 30/04/2018 were disallowed. Finally the dataset was redundancy reduced, by taking the chain with the best non-zero resolution from each cluster in the PDB 40% sequence clustering giving a dataset of 12,494 chains. This is referred to as the recent PDB set.

Compute resources

Inference was run on V100 graphics processing units (GPUs), with each sequence inferenced five times to produce five inputs to model selection. To prevent out of memory errors, long sequences were assigned to multi-GPU workers. Specifically, sequences of length 1401-2000 residues were processed by workers with two GPUs, and those of length 2001-2700 residues by workers with four GPUs (see companion paper for details of unified memory on longer proteins; it is possible higher memory workers could be used without additional GPUs).

The total resources used for inference were logged and amounted to 930 GPU days. This accounts for generating five models per protein; ~190 GPU days would be sufficient to inference each protein once. Long sequences had a disproportionate impact due to the multi-GPU workers described above. Approximately 250 GPU days would have been sufficient to produce five models for all proteins shorter than 1400 residues. For reference, Extended Data Fig. 6 shows the relationship between sequence length and inference time.

All other stages of the pipeline (MSA search, template search and constrained relaxation) ran on the central processing unit (CPU) and used standard tools. Our human proteome run made use of some cached intermediates (*e.g.* stored MSA search results). However, we estimate the total cost of running these stages from scratch at 510 core days. This estimate is based on taking a sample of 240 human proteins stratified by length, timing each stage when run with empty caches, fitting a quadratic relationship between sequence length and run time, then applying that relationship to the sequences in the human proteome. Extended Data Fig. 7 presents the data used to make this estimate.

Template coverage

Except where otherwise noted, template coverage was estimated on a per-residue basis as follows. Hmmsearch was run against a copy of the PDB SEQRES (downloaded 15/02/2021) using default flags[67]. The prior template coverage at residue *i* is the max % sequence identity of all hits covering residue *i*, regardless of whether the hit residue is experimentally resolved. For recent PDB analysis, only template hits corresponding to a structure released before 30/04/2018 were accepted.

In the section on full-chain prediction, template filtering is based on the highest sequence identity of any *single* Hmmsearch hit with > 50% coverage. This is because high coverage templates are particularly relevant when considering whether a predicted domain packing is novel.

GO term breakdown

GO annotations were taken from the XML metadata for the UniProt human reference proteome and were matched to the Gene Ontology in obo format[36,37]. One erroneous *is_a* relationship was manually removed (GO:0071702 *is_a* GO:0006820, see changelog https://www.ebi.ac.uk/QuickGO/term/GO:0071702). The ontology file was used to propagate the GO annotations using *is_a* and *part_of* relations to assign parent-child relationships, and accounting for alternative IDs.

GO terms were then filtered to a manageable number for display, first by filtering for terms with > 3000 annotations, and from those selecting only moderately specific terms (a term cannot have a child with > 3000 annotations). The remaining terms in the Molecular Function and Cellular Component ontologies are displayed in Fig. 1d.

Structure analysis

Structure images were created in PyMOL[69], and PyMOL align was used to compute RMSDs (outlier rejection is described in the text where applicable).

For docking against DGAT2, P2Rank[65] was used to identify ligand-binding pockets in the AlphaFold structure. AutoDockTools[70] was used to convert the AlphaFold prediction to PDBQT format. For the ligands, DGAT2-specific inhibitor (CAS Number: 1469284-79-4) and DGAT1-specific inhibitor (CAS Number: 942999-61-3) were also prepared in PDBQT format using AutoDockTools. AutoDock Vina[71] was run with an exhaustiveness parameter of 32, a seed of 0, and a docking search space of 25 × 25 × 25 Å^3 centred at the point identified by P2Rank.

For identifying the most similar structure to wolframin, TM-align[42] was used to compare against all PDB chains (downloaded 15/02/2021) with our prediction as the reference. This returned 3F1Z with a TM-score of 0.472.

Additional metrics

The implementation of pTM is described in the AlphaFold paper SI section 1.9.7, and the implementation of the experimentally resolved head is described in section 1.9.10. The weighted version of pTM is described in Suppl. Methods 1.

Analysis of low-confidence regions

For evaluation on CAID, the target sequences and ground truth labels for the Disprot-PDB set were downloaded from idpcentral.org. Structure prediction was performed as described above for recent PDB, with a template cutoff of 30/04/2018. To enable complete coverage, two sequences containing non-standard residues (X, U) had these remapped to G, glycine. Sequences longer than 2000 residues were split into two segments: 1-2000 and 2000-end, and the

pLDDT and experimentally-resolved head arrays were concatenated for evaluation. The two evaluated disorder predictors were taken to be 1 - 0.01 * pLDDT and 1 - predicted resolvability for Cα atoms, respectively.

To obtain the ratio of heterotypic contacts to all contacts (Fig. 4d), two residues are considered in contact if their Cβ atoms (or Cα for glycine) are within 8 Å[29] and if they are separated in primary sequence by at least three other residues (to exclude contacts within an α-helix). Heteromers are identified as protein entities with a different *entity_id* in the structure mmCIF file.

BestSingleStructuralTemplate comparison

CAMEO data for the period 21/03/2020 to 13/03/2021 was downloaded from the CAMEO website. AlphaFold predictions were produced for all sequences in the target.fasta files, using the same procedure detailed above but with a max template date of 01/03/2020. Predictions were scored against the CAMEO ground truth using lDDT-Cα. For *BestSingleStructuralTemplate* lDDT-Cα scores were taken from the CAMEO JavaScript Object Notation (JSON) files provided. Structures solved by solution NMR and solid-state NMR were filtered out at the analysis stage. To determine the template identity, templates were drawn from a copy of PDB downloaded on 15/02/2021 with template search performed using Hmmsearch. Templates were filtered to those with at least 70% coverage of the sequence and a release date prior to the query. The template with the highest e-value after filtering was used to compute the template identity. Targets were binned according to template identity, with width 10 bins ranging from 30 to 90. Extended Data Fig. 4 shows the distribution of lDDT-Cα for each model within each bin as a boxplot (horizontal line at the median, box spanning from the lower to the upper quartile, whiskers extending to the minimum and maximum. Four hundred and twenty-eight targets were included in the analysis.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

AlphaFold structure predictions for the human proteome are available under a CC-BY-4.0 license at https://alphafold.ebi.ac.uk/.

All input data are freely available from public sources. The human reference proteome together with its xml annotations was obtained from UniProt 2021_02 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2021_02/knowledgebase/). At prediction time, MSA search was performed against UniRef90 2020_03 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2020_03/uniref/), MGnify clusters 2018_12 (https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2018_12/), and a reduced version of BFD (produced from as outlined in the Methods from BFD https://bfd.mmseqs.com/). Template structures, the SEQRES fasta file, and the 40% sequence clustering were taken from a copy of the PDB downloaded 15/2/2021 (https://www.wwpdb.org/ftp/pdb-ftp-sites; see also https://ftp.wwpdb.org/pub/pdb/derived_data/ and https://cdn.rcsb.org/resources/sequence/clusters/bc-40.out for sequence data). Experimental structures are drawn from the same copy of the PDB; we show structures with accessions 6YJ1[64], 6OFS[43,46], 1IDQ[46], 1PRT[72], 3F1Z[58], 7KPX[66] and 6VP0[51]. Template search used PDB70, downloaded 10/02/2021 (http://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/). The CAID dataset was downloaded from (https://idpcentral.org/caid/data/1/reference/disprot-disorder-pdb-atleast.txt). CAMEO data was accessed 17/03/2021 from (https://www.cameo3d.org/static/downloads/modeling/1-year/raw_targets-1-year.public.tar.gz). A copy of the current Gene Ontology was downloaded 29/04/2021 from (http://current.geneontology.org/ontology/go.obo).

Code availability

Source code for the AlphaFold model, trained weights, and an inference script is available under an open-source license at https://github.com/deepmind/alphafold. Neural networks were developed with TensorFlow v1 (https://github.com/tensorflow/tensorflow), Sonnet v1 (https://github.com/deepmind/sonnet), JAX v0.1.69 (https://github.com/google/jax/), and Haiku v0.0.4 (https://github.com/deepmind/dm-haiku).

For MSA search on UniRef90, MGnify clusters, and reduced BFD we used jackhmmer and for template search on the PDB SEQRES we used hmmsearch, both from HMMER v3.3 (http://eddylab.org/software/hmmer/). For template search against PDB70, we used HHsearch from HH-suite v3.0-beta.3 14/07/2017 (https://github.com/soedinglab/hh-suite). For constrained relaxation of structures, we used OpenMM v7.3.1 (https://github.com/openmm/openmm) with the Amber99sb force field.

Docking analysis on DGAT used P2Rank v2.1 (https://github.com/rdk/p2rank), MGLTools v1.5.6 (https://ccsb.scripps.edu/mgltools/) and AutoDockVina v1.1.2 (http://vina.scripps.edu/download/) on a workstation running Debian GNU/Linux rodete 5.10.40-1rodete1-amd64 x86_64.

Data analysis used Python v3.6 (https://www.python.org/), NumPy v1.16.4 (https://github.com/numpy/numpy), SciPy v1.2.1 (https://www.scipy.org/), seaborn v0.11.1 (https://github.com/mwaskom/seaborn), scikit-learn v0.24.0 (https://github.com/scikit-learn/), Matplotlib v3.3.4 (https://github.com/matplotlib/matplotlib), pandas v1.1.5 (https://github.com/pandas-dev/pandas), and Colab (https://research.google.com/colaboratory). TM-align v20190822 (https://zhanglab.dcmb.med.umich.edu/TM-align) was used for computing TM-scores. Structure analysis used Pymol v2.3.0 (https://github.com/schrodinger/pymol-open-source).

67. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Informatics 2009: Genome Informatics Series* **23**, 205–211 (World Scientific, 2009).
68. Steinegger, M., Mirdita, M. & Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* **16**, 603–606 (2019).
69. Schrödinger. *The PyMOL Molecular Graphics System, Version* **1**.8. (2015).
70. Morris, G. M. *et al*. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem*. **30**, 2785–2791 (2009).
71. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem*. **31**, 455–461 (2010).
72. Stein, P. E. *et al*. The crystal structure of pertussis toxin. *Structure* **2**, 45–57 (1994).
73. Necci, M., Piovesan, D., Clementel, D., Dosztányi, Z. & Tosatto, S. C. E. MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavors in proteins. *Bioinformatics* (2020).
74. Dyson, H. J. Roles of intrinsic disorder in protein-nucleic acid interactions. *Mol. Biosyst*. **8**, 97–104 (2012).
75. Dunbrack, R. L., Jr & Karplus, M. Backbone-dependent rotamer library for proteins application to side-chain prediction. *J. Mol. Biol*. **230**, 543–574 (1993).
76. Fischer, A., Smiesko, M., Sellner, M. & Lill, M. A. Decision making in structure-based drug discovery: visual inspection of docking results. *J. Med. Chem*. **64**, 2489-2500 (2021).

Acknowledgements We would like to thank Antonia Paterson, Caroline Low, Craig Donner, David Evans, Fan Yang, Jeff Stanway, Joe Stanton, Louise Deason, Natasha Latysheva, Nicole Hobbs, Raia Hadsell, Richard Green, Sasha Brown, Vijay Bolina, Žiga Avsec, and the Research Platform Team for their contributions; Rupert Kemp for help in managing the project, and our colleagues at DeepMind, Google and Alphabet for their encouragement and support. We thank Dr Emile Van Schaftingen (UC Louvain), Dr Ming Zhou (Baylor College of Medicine), and Dr Fumihiko Urano (Washington University School of Medicine) for reading and commenting on our discussion of glucose-6-phosphatase, diacylglycerol O-acyltransferase 2, and wolframin, respectively. We also thank the team at EMBL-EBI for their work on making AlphaFold structure predictions available, in particular David Yuan, Mandar Deshpande, Mihaly Varadi, Sreenath Sasidharan Nair, Stephen Anyango, and Edith Heard.

Author contributions K.T., J.J. and D.H. led the research. D.H., K.K., P.K., C.M. and E.C. managed the research. T.G. developed the proteome-scale inference system. K.T., J.A., Z.W., M.Z., R.E., M.F., Al.Br. and A.C. generated and analysed the structure predictions. J.J., M.F., S.K. and O.R. developed the metrics used to interpret predictions. A.Z., S.P., T.G., A.C. and K.T. developed the data processing pipelines to produce the AlphaFold protein structure database. S.V., A.L., Al.Ba., G.K., D.H. and E.B. managed the work to make AlphaFold predictions available via EMBL-EBI-hosted resources. S.V., G.K. and Al.Ba. provided scientific advice on how predictions

should be displayed. J.J., R.E., Al.P., M.F., O.R., R.B., An.P., S.K., B.R.-P., J.A., A.S., T.G., A.Z., K.T., Al.Br., An.B., A.C., S.N., R.J., D.R. and M.Z. developed the network and associated infrastructure used in inferencing the proteome. K.T., J.A., Z.W., J.J., M.F., M.Z., C.M. and D.H. wrote the paper.

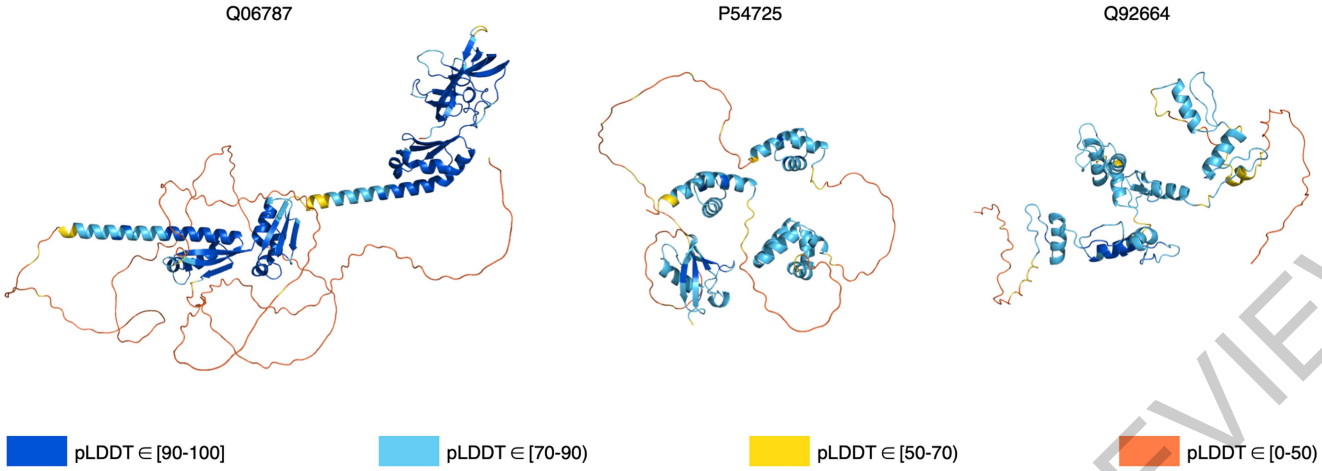
Competing interests J.J., R.E., Al.P., T.G., M.F., O.R., R.B., A.B., S.K., D.R. and A.S. have filed non-provisional patent applications 16/701,070, PCT/EP2020/084238, and provisional patent applications 63/107,362, 63/118,917, 63/118,918, 63/118,921 and 63/118,919, each in the name of DeepMind Technologies Limited, each pending, relating to machine learning for predicting protein structures. E.B. is a paid consultant to Oxford Nanopore and Dovetail Inc, genomics companies. The remaining authors declare no competing interests.

Additional information

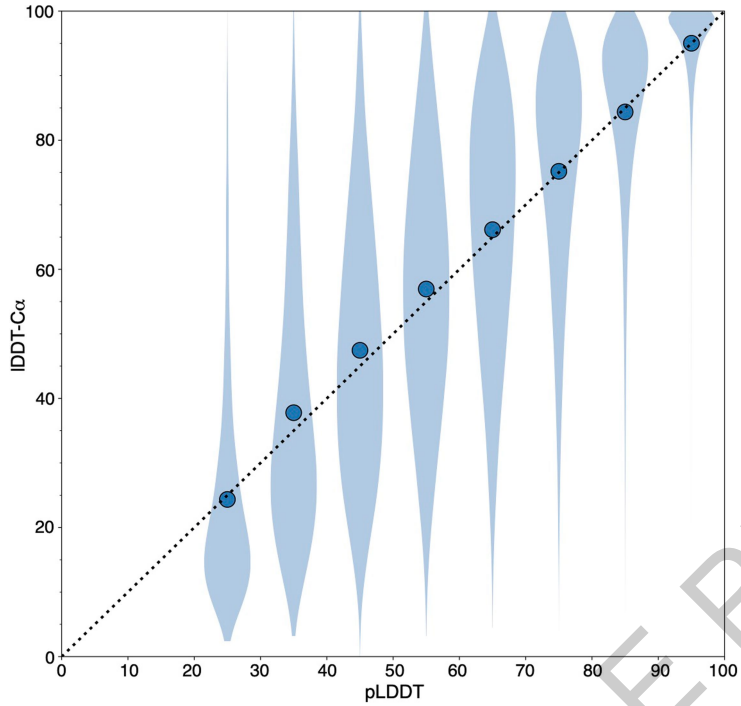
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03828-1.

Correspondence and requests for materials should be addressed to K.T., J.J. or D.H.

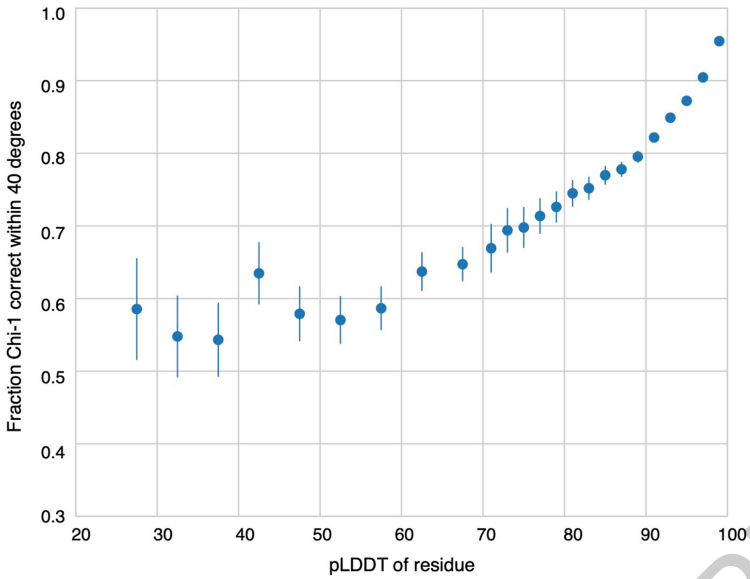
Peer review information *Nature* thanks Mohammed AlQuraishi, Yang Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



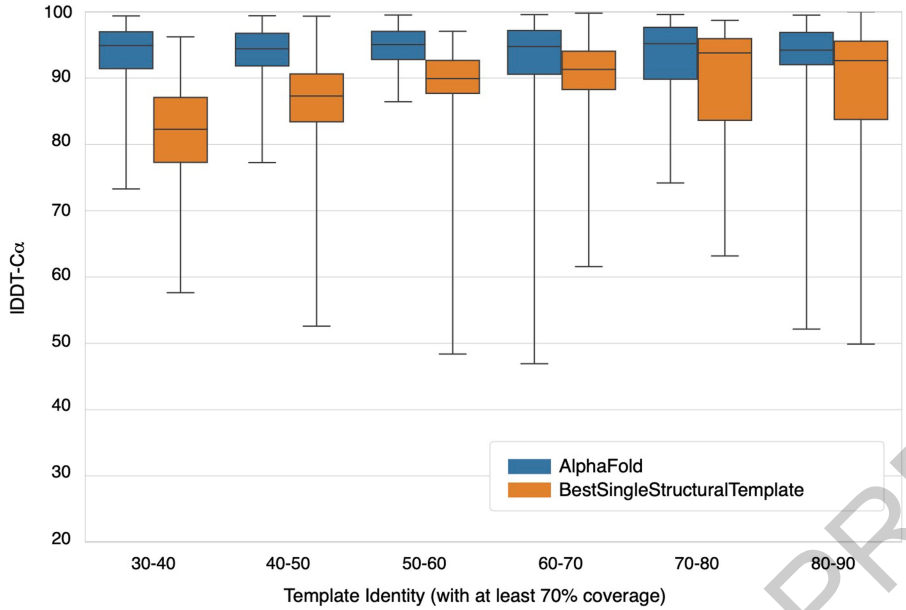
Extended Data Fig. 1 | Example full-chain outputs containing both high- and low-confidence regions. Q06787 (synaptic functional regulator FMR1) and P54725 (UV excision repair protein RAD23 homolog A) are predicted to be disordered outside the experimentally-determined regions by MobiDB[73]. Q92664 (transcription factor IIIA) has been described as "beads on a string", consisting of zinc finger domains joined by flexible linkers[74].



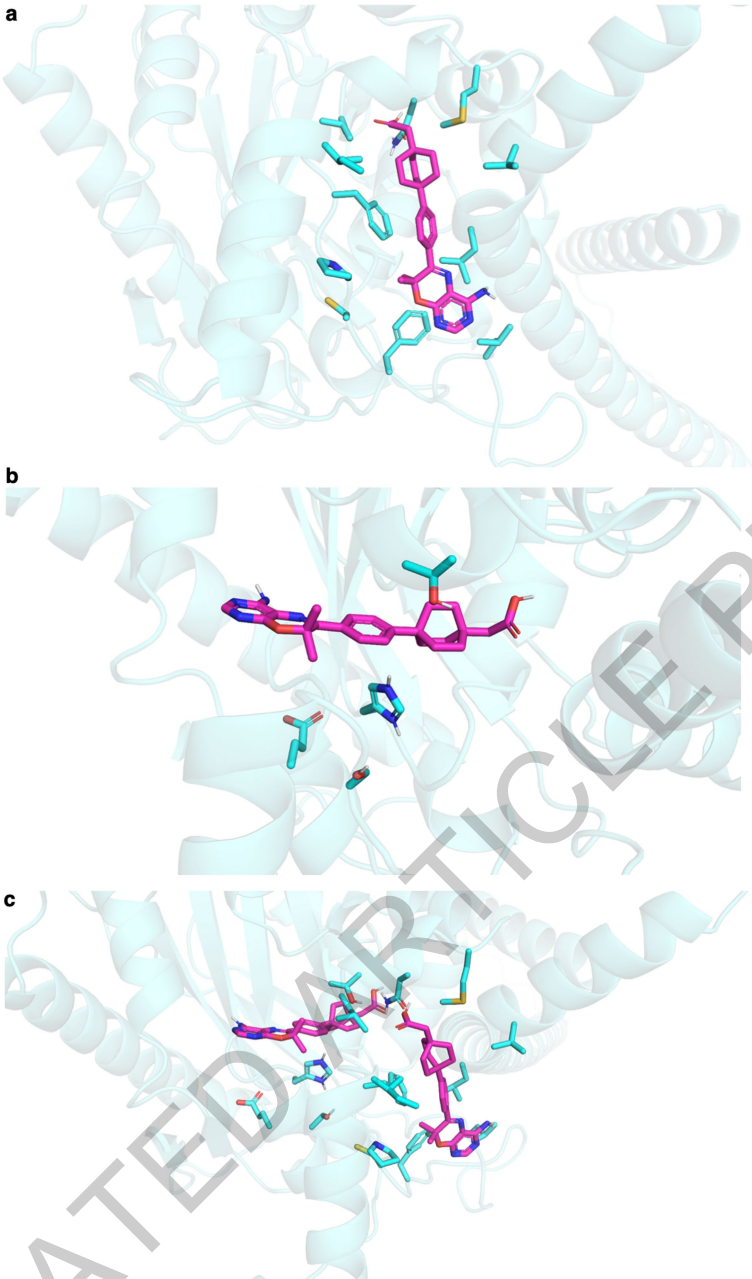
Extended Data Fig. 2 | Distribution of per-residue lDDT-Cα within eight pLDDT bins. This represents an alternative visualisation to Fig. 1a that does not sample the data. It uses the recent PDB set (see Methods) restricted to structures with a reported resolution < 3.5 Å (N = 2,756,569 residues). Residues were assigned to bins of width 10 based on their pLDDT (min 20, max 100). Markers show the mean lDDT-Cα within each bin, while the lDDT-Cα distribution is visualized as a Matplotlib violin plot (kernel density estimate bandwidth 0.2). Smallest sample size for the corresponding violin is 5,655 residues for the left-most bin.



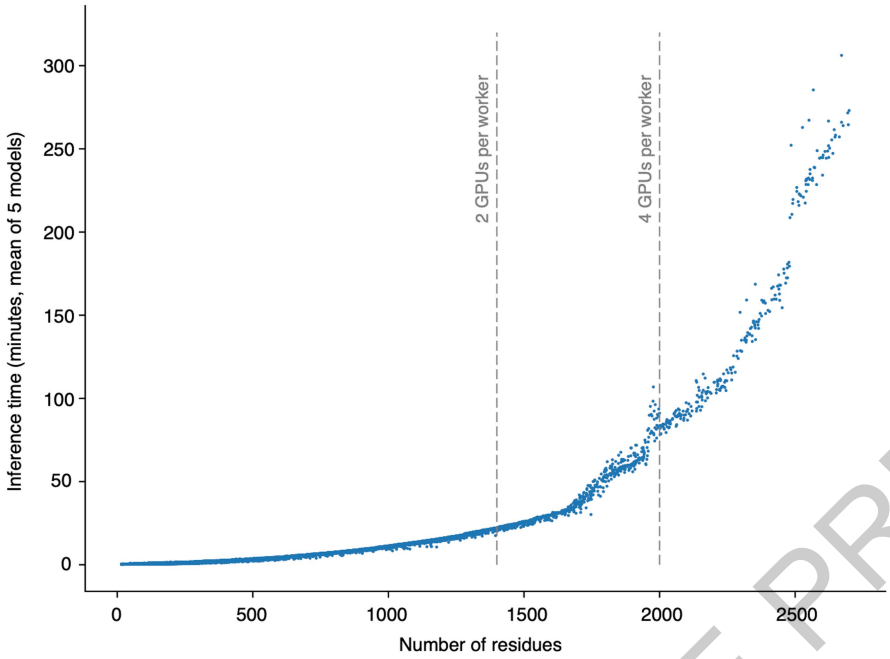
Extended Data Fig. 3 | Relationship between pLDDT and side-chain Chi-1 correctness. Evaluated on the recent PDB set (see Methods) restricted to structures with a reported resolution < 2.5 Å (N = 5,983 chains) and residues with a B-factor < 30 Å^2 (N = 609,623 residues). Residues are binned by pLDDT, with bin width 5 between 20 and 70 pLDDT and width 2 above 70 pLDDT. A Chi-1 angle is considered correct if it is within 40° of its value in the PDB structure[75,76]. Markers show the proportion of correct Chi-1 angles within each bin, while error bars indicate the 95% confidence interval (Student's t, two-sided). Smallest sample size for the error bars is 193 residues for the left-most bin.



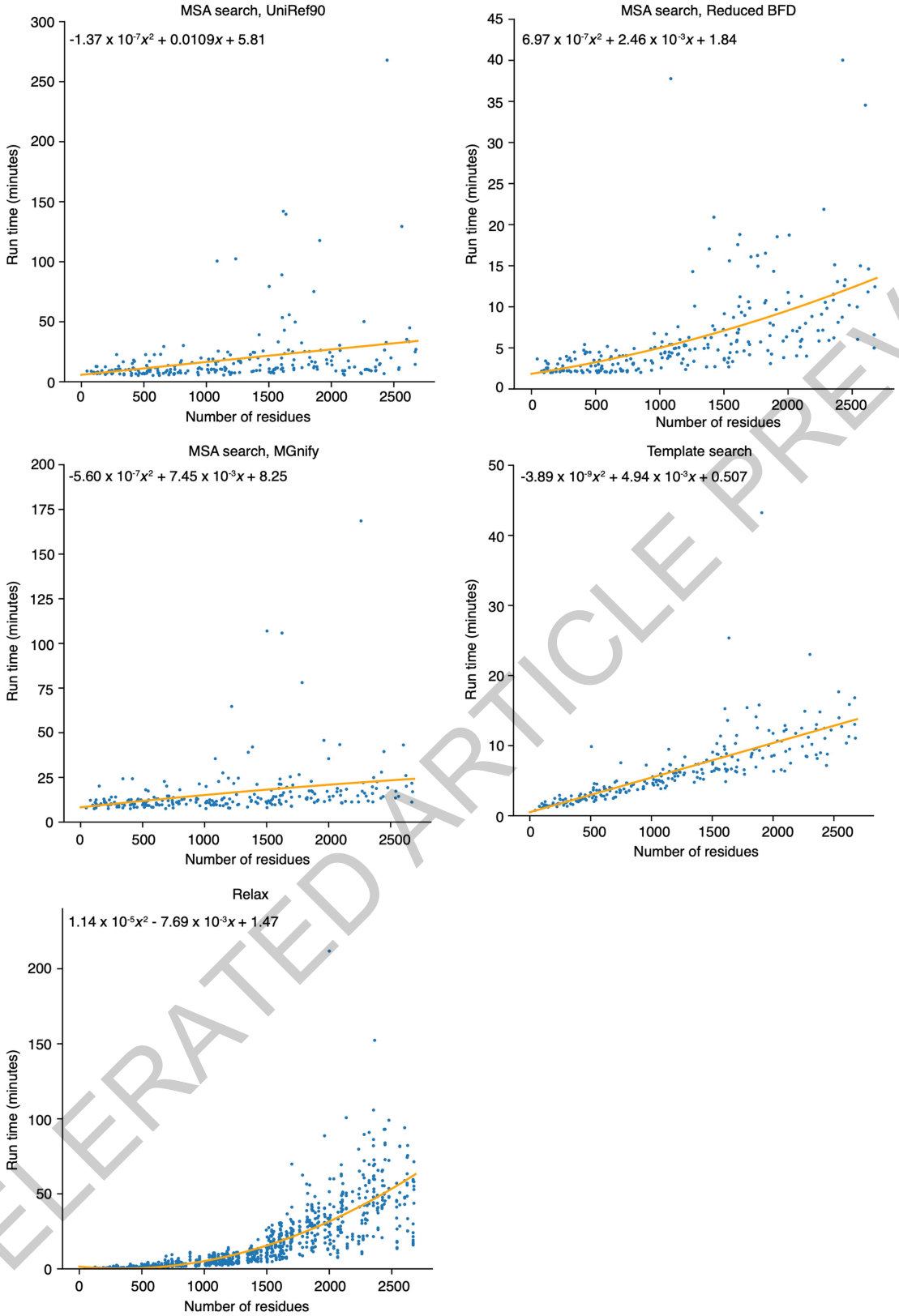
Extended Data Fig. 4 | AlphaFold performance at a range of template sequence identities. lDDT-Cɑ for AlphaFold and *BestSingleStructuralTemplate* on one year of CAMEO targets[39]. Targets are binned according to the sequence identity of the best template covering at least 70% of the target, and a boxplot is shown for each bin. A horizontal line indicates the median, boxes span from the lower to the upper quartile, and the whiskers extend from the minimum to the maximum. Four hundred twenty-eight targets are included in total (see Source Data file provided); the smallest number of targets in any bin is 18.



Extended Data Fig. 5 | Docking poses for a DGAT1-specific inhibitor in DGAT2. a, Top binding pose from Autodock Vina for a DGAT1-specific inhibitor in DGAT2, which does not match the predicted binding pocket for a DGAT2-specific inhibitor. **b**, Next best binding pose, which matches the binding pocket for the DGAT2-specific inhibitor, but does not contain components that satisfy polar side chains His163 and Thr194. **c**, Relative positions of both binding poses.



Extended Data Fig. 6 | Relationship between sequence length and inference time. Based on logs from our human proteome set. All processed proteins are shown (N = 20,296). Each point indicates the mean inference time for the protein over the models produced. Vertical lines show the length cutoffs above which sequences were processed by multi-GPU workers.



Extended Data Fig. 7 | Relationship between sequence length and run time for non-inference stages of the pipeline. Based on 240 human protein sequences, chosen by stratified sampling from the length buckets: [16, 500), [500, 1000), [1000, 1500), [1500, 2000), [2000, 2500), [2500, 2700]. The relax plot shows five times more points, since five relaxed models are generated per protein. Coefficients for the quadratic lines of best fit were computed with numpy polyfit.

Extended Data Table 1 | lDDT-Cα distribution in various pLDDT bins

	lDDT-Cα					
	Mean	Median	Q1	Q3	IQR	< lower bin edge - 5
pLDDT in (50-70]	62.5	63.8	48.6	77.8	29.2	20%
pLDDT in (70-90]	82.3	86.5	76.1	92.9	16.7	12%
pLDDT in (90-100]	95.0	97.4	93.8	99.3	5.6	7%

Based on per-residue lDDT-Cα and per-residue pLDDT on resolved regions. This table uses the recent PDB set (see Methods) restricted to structures with a reported resolution < 3.5 Å. The total number of chains included is 10,215.

Extended Data Table 2 | Relationship between pLDDT and TM-score

	TM-score					
	Mean	Median	Q1	Q3	IQR	≥ 0.5
pLDDT in (50-70]	0.44	0.43	0.29	0.58	0.29	37%
pLDDT in (70-90]	0.75	0.83	0.63	0.92	0.29	86%
pLDDT in (90-100]	0.93	0.97	0.92	0.98	0.06	99%

Binning is based on the mean pLDDT over each chain, weighted by the output of the experimentally resolved head. This table uses the recent PDB set (see Methods) restricted to structures with a reported resolution < 3.5 Å. The total number of chains included is 10,215.

nature research

Corresponding author(s): Kathryn Tunyasuvunakool, John Jumper, Demis Hassabis

Last updated by author(s): Jul 13, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Source code for the AlphaFold model, trained weights, and an inference script are available under an open-source license at https://github.com/deepmind/alphafold. Neural networks were developed with TensorFlow v1 (https://github.com/tensorflow/tensorflow), Sonnet v1 (https://github.com/deepmind/sonnet), JaX v0.1.69 (https://github.com/google/jax/), and Haiku v0.0.4 (https://github.com/deepmind/dm-haiku). For MSA search on UniRef90, MGnify clusters, and reduced BFD we used jackhmmer and for template search on the PDB seqres we used hmmsearch, both from HMMER v3.3 (http://eddylab.org/software/hmmer/). For template search against PDB70, we used HHsearch from HH-suite v3.0-beta.3 14/07/2017 (https://github.com/soedinglab/hh-suite). For constrained relaxation of structures, we used OpenMM v7.3.1 (https://github.com/openmm/openmm) with the Amber99sb force field. Docking analysis on DGAT used P2Rank v2.1 (https://github.com/rdk/p2rank), MGLTools v1.5.6 (https://ccsb.scripps.edu/mgltools/) and AutoDockVina v1.1.2 (http://vina.scripps.edu/download/) on a workstation running Debian GNU/Linux rodete 5.10.40-1rodete1-amd64 x86_64.
Data analysis	Data analysis used Python v3.6 (https://www.python.org/), NumPy v1.16.4 (https://github.com/numpy/numpy), SciPy v1.2.1 (https://www.scipy.org/), seaborn v0.11.1 (https://github.com/mwaskom/seaborn), scikit-learn v0.24.0 (https://github.com/scikit-learn/), Matplotlib v3.3.4 (https://github.com/matplotlib/matplotlib), pandas v1.1.5 (https://github.com/pandas-dev/pandas), and Colab (https://research.google.com/colaboratory). TM-align v20190822 (https://zhanglab.dcmb.med.umich.edu/TM-align) was used for computing TM-scores. Structure analysis used Pymol v2.3.0 (https://github.com/schrodinger/pymol-open-source).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

AlphaFold structure predictions for the human proteome are available under a CC-BY-4.0 license at https://alphafold.ebi.ac.uk/.

All input data are freely available from public sources. The human reference proteome together with its xml annotations was obtained from UniProt 2021_02 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2021_02/knowledgebase/).
At prediction time, MSA search was performed against UniRef90 2020_03 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2020_03/uniref/), MGnify clusters 2018_12 (https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2018_12/), and a reduced version of BFD (produced from as outlined in the Methods from BFD https://bfd.mmseqs.com/). Template structures, the SEQRES fasta file, and the 40% sequence clustering were taken from a copy of the PDB downloaded 15/2/2021 (https://www.wwpdb.org/ftp/pdb-ftp-sites; see also https://ftp.wwpdb.org/pub/pdb/derived_data/ and https://cdn.rcsb.org/resources/sequence/clusters/bc-40.out for sequence data). Experimental structures are drawn from the same copy of the PDB; we show structures with accessions 6YJ1, 6OFS, 1IDQ, 1PRT, 3F1Z, 7KPX and 6VP0. Template search used PDB70, downloaded 10/02/2021 (http://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/). The CAID dataset was downloaded from (https://idpcentral.org/caid/data/1/reference/disprot-disorder-pdb-atleast.txt). CAMEO data was accessed 17/03/2021 from (https://www.cameo3d.org/static/downloads/modeling/1-year/raw_targets-1-year.public.tar.gz). A copy of the current Gene Ontology was downloaded 29/04/2021 from (http://current.geneontology.org/ontology/go.obo).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study does not concern a sample; the main results concern the largest subset of the human reference proteome that it was easily practical to inference.
Data exclusions	Proteins were excluded from the analysis for length reasons (less than 16 or greater than 2700 residues), or because they contained a nonstandard single letter amino acid code in their fasta sequence.
Replication	Not applicable, no experimental work is described in this study. The results are the output of a computational method which will be made available.
Randomization	Not applicable, we are not drawing a comparison between two groups
Blinding	Not applicable, we are not drawing a comparison between two groups

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging