

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

Open access

 Check for updates

Kathryn Tunyasuvunakool¹✉, Jonas Adler¹, Zachary Wu¹, Tim Green¹, Michal Zielinski¹, Augustin Žídek¹, Alex Bridgland¹, Andrew Cowie¹, Clemens Meyer¹, Agata Laydon¹, Sameer Velankar², Gerard J. Kleywegt², Alex Bateman², Richard Evans¹, Alexander Pritzel¹, Michael Figurnov¹, Olaf Ronneberger¹, Russ Bates¹, Simon A. A. Kohl¹, Anna Potapenko¹, Andrew J. Ballard¹, Bernardino Romera-Paredes¹, Stanislav Nikolov¹, Rishabh Jain¹, Ellen Clancy¹, David Reiman¹, Stig Petersen¹, Andrew W. Senior¹, Koray Kavukcuoglu¹, Ewan Birney², Pushmeet Kohli¹, 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 structure¹. Here we markedly expand the structural coverage of the proteome by applying the state-of-the-art machine learning method, AlphaFold², at a scale that covers 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 that are likely to be disordered. Finally, we provide some case studies to illustrate how high-quality predictions could be used to generate biological hypotheses. We are making our predictions freely available to the community and anticipate that routine large-scale and high-accuracy structure prediction will become an important tool that will allow 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 laboratories and dedicated structural genomics initiatives, more than 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 only a fragment of the sequence⁶. Experimental structure determination requires overcoming many time-consuming hurdles: the protein must be produced in sufficient quantities and 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 such as 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. Previous large-scale structure prediction studies have addressed protein families^{9–12}, specific functional classes^{13,14}, domains identified within whole proteomes¹⁵ and, in some cases, full chains or complexes^{16,17}. In

particular, projects such as the 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 prediction²⁵, and domain identification and classification^{26–28}. Although some of our analyses are inspired by these previous studies, here we focus mainly on structural investigations for which scale and accuracy are particularly beneficial.

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 entry³¹, 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 deviation ($\text{C}\alpha$ r.m.s.d.) are now common, and side chains are increasingly accurate². Good results can often be achieved even for challenging proteins without a template structure in the PDB, or with relatively few related sequences to build a multiple sequence alignment (MSA)². These improvements are important, because more accurate models permit a wider range of

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

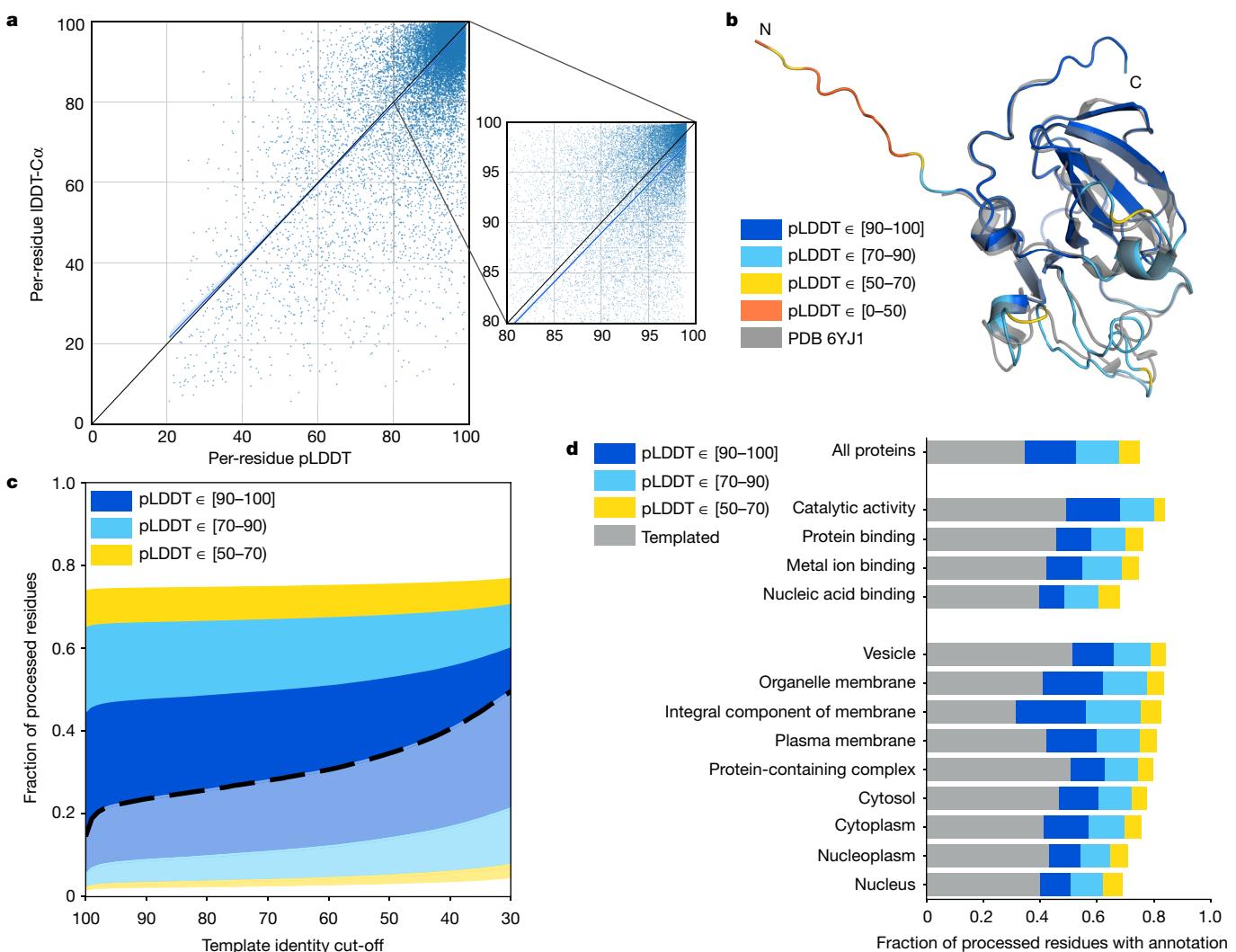


Fig. 1 | Model confidence and added coverage. **a**, Correlation between per-residue pLDDT and IDDT-C α . Data are based on a held-out set of recent PDB chains (Methods) filtered to those with a reported resolution of $<3\text{ \AA}$ ($n=10,215$ chains and 2,756,569 residues). The scatterplot shows a subsample (1% of residues), with the blue line showing a least-squares linear fit and the shaded region a 95% confidence interval estimated with 1,000 bootstrap samples. The black line shows $x=y$, for comparison. The smaller plot is a magnified region of the larger one. On the full dataset, the Pearson's $r=0.73$ and the least-squares linear fit is $y=(0.967 \pm 0.001) \times x + (1.9 \pm 0.1)$. **b**, AlphaFold prediction and experimental structure for a CASP14 target (PDB: 6YJ1)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 and a heavy dashed line separates these from residues without a template. **d**, Added residue-level coverage of the proteome for high-level GO terms, on top of residues covered by a template with sequence identity of more than 50%. Based on the same human proteome dataset as in **c** ($n=10,537,122$ residues).

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. All of our predictions can be accessed freely at <https://alphafold.ebi.ac.uk/>, hosted by the European Bioinformatics Institute.

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 2,700 residues⁶. 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 (Extended Data Fig. 1). To this end, AlphaFold produces a per-residue confidence metric called predicted local distance difference test (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 C α (IDDT-C α)³⁵. It has been shown to be well-calibrated (Fig. 1a, Extended Data Fig. 2 and Extended Data Table 1) and full details on how the pLDDT is produced are given in the supplementary information of the companion AlphaFold paper².

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 cut-off, above which AlphaFold χ_1 rotamers are 80% correct for a recent PDB

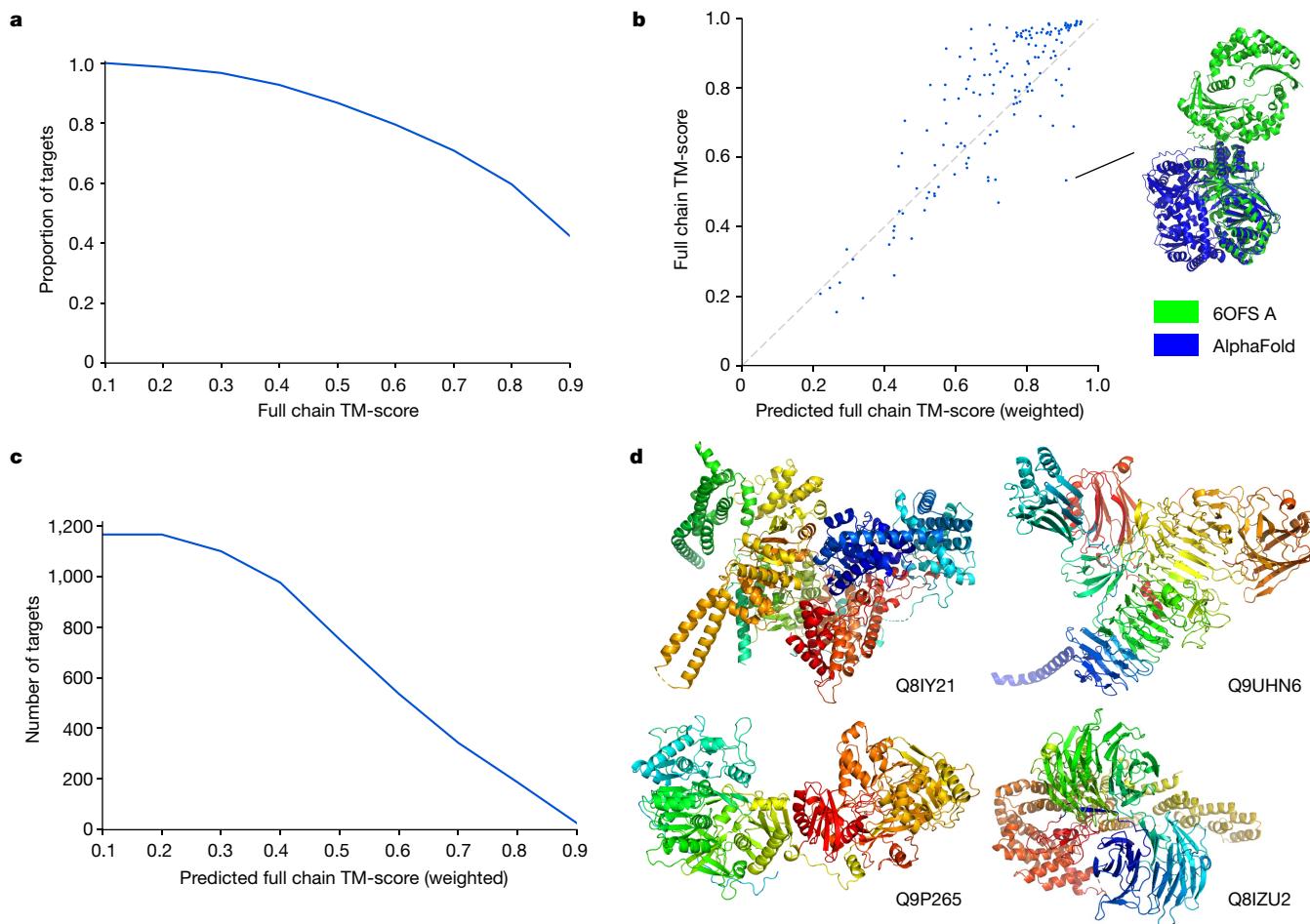


Fig. 2 | Full chain structure prediction. **a**, TM-score distribution for AlphaFold evaluated on a held-out set of template-filtered, long PDB chains ($n=151$ chains). Includes recent PDB proteins with more than 800 resolved residues and best 50% coverage template below 30% identity. **b**, Correlation 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 (PDB: 6OFS, chain A). **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). Human proteome predictions comprise more than 600 confident residues (more than 50% coverage) and no proteins with 50% coverage templates. **d**, Four of the top hits from the set shown in **c**, filtering by $pTM > 0.8$ and sorting by number of confident residues. Proteins are labelled by their UniProt accession. For clarity, regions with $pLDDT < 50$ are hidden, as are isolated smaller regions that were left after this cropping.

test dataset (Extended Data Fig. 3). A lower cut-off of $pLDDT > 70$ corresponds to a generally correct backbone prediction (Extended Data Table 2). The accuracy of AlphaFold within a number of $pLDDT$ bands is illustrated for an example protein in Fig. 1b.

Of 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. In total, 58.0% of residues were predicted confidently ($pLDDT > 70$), indicating that we also add substantial coverage for sequences without a good template in PDB (with a 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 (more than 200 residues) with $pLDDT \geq 70$ 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 proteins³⁸ (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 within its training set.

We note that the accuracy of AlphaFold was validated in CASP14², which focuses on challenging proteins that are dissimilar to structures already available in the PDB. By contrast, many human proteins have templates with high sequence identity. To evaluate the applicability of AlphaFold to this collection, we predicted structures for 1 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 the BestSingleStructuralTemplate baseline of CAMEO across a wide range of levels of template identity (Extended Data Fig. 4).

Prediction of full-length protein chains

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 for this. 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 in cases in which two or more domains interact substantially.

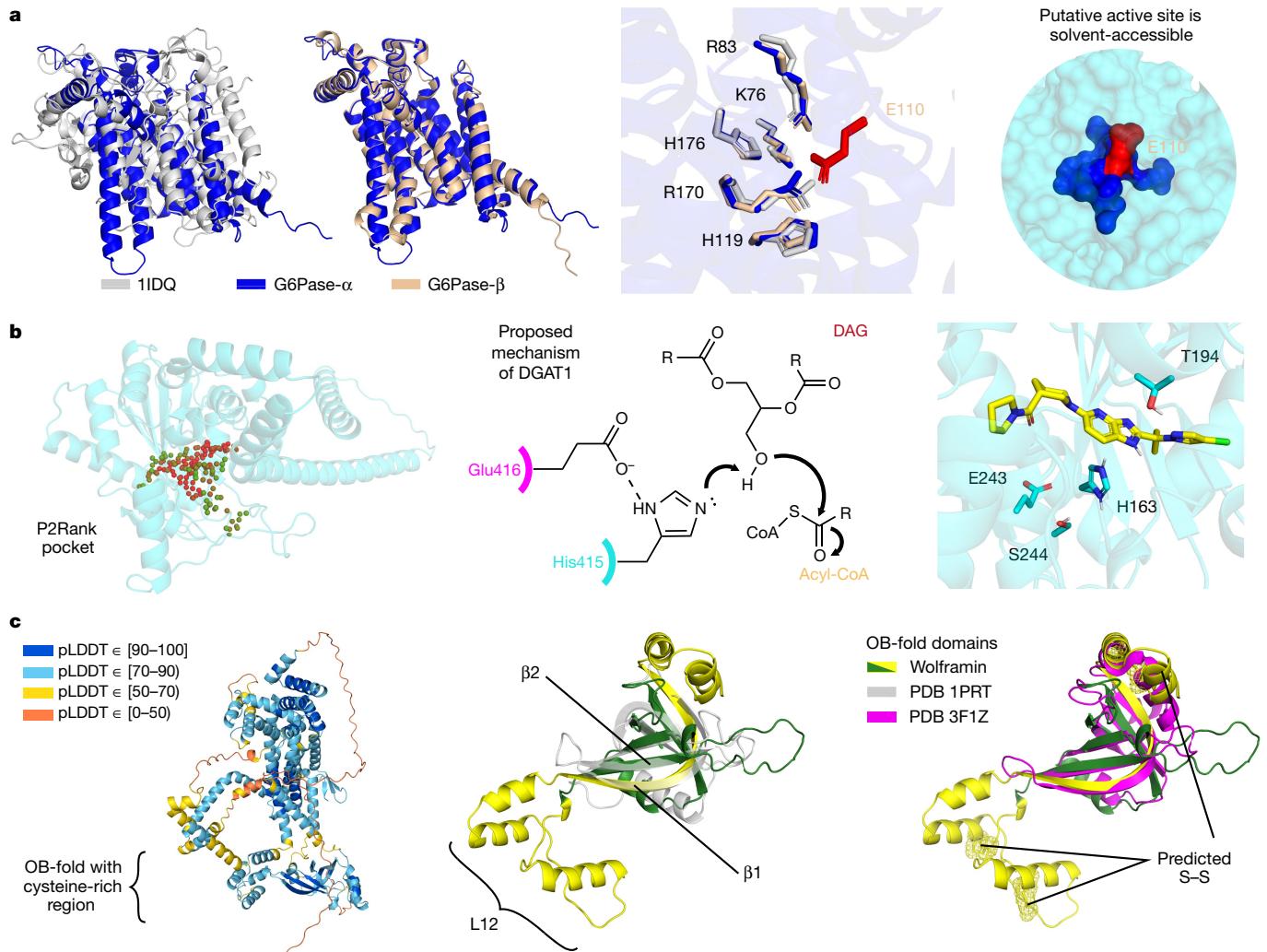


Fig. 3 | Highlighted structure predictions. **a**, Left, comparison of the active sites of two G6Pases (G6Pase- α and G6Pase- β) and a chloroperoxidase (PDB 1IDQ). The G6Pases are glucose-forming enzymes that contain a conserved, solvent-accessible glutamate (red; right) opposite the shared active-site residues (middle). **b**, Left, pocket prediction (P2Rank⁶⁵) identifies a putative binding pocket for DGAT2, which is involved in body-fat synthesis. Red and green spheres represent the ligandability scores by P2Rank of 1 and 0, respectively. Middle, a proposed mechanism for DGAT1⁵¹ activates the substrate with Glu416 and His415, which have analogous residues in the DGAT2 substrate. DAG: R-C(=O)O-CH2-CH2-O-C(=O)R. Acyl-CoA: R-C(=O)SCoA. Right: Structure of the DGAT2 active site showing residues T194, E243, H163, S244. **c**, Predicted structure of wolframin, mutations in which cause Wolfram syndrome. Although there are regions in wolframin with low pLDDT (left), we could identify an OB-fold region (green/yellow), with a comparable core to a prototypical OB-fold (grey; middle). However, the most similar PDB chain (magenta; right) lacks the conserved cysteine-rich region (yellow) of our prediction. This region forms the characteristic β 1 strand and an extended L12 loop, and is predicted to contain three disulfide bridges (yellow mesh).

pocket. The docked inhibitor is well placed for polar interactions with His163 and Thr194 (right). The chemical structure (middle) is adapted from ref.⁵¹. **c**, Predicted structure of wolframin, mutations in which cause Wolfram syndrome. Although there are regions in wolframin with low pLDDT (left), we could identify an OB-fold region (green/yellow), with a comparable core to a prototypical OB-fold (grey; middle). However, the most similar PDB chain (magenta; right) lacks the conserved cysteine-rich region (yellow) of our prediction. This region forms the characteristic β 1 strand and an extended L12 loop, and is predicted to contain three disulfide bridges (yellow mesh).

Finally, the full chain approach lets the model attempt an inter-domain packing prediction.

Inter-domain accuracy was assessed at CASP14, and AlphaFold outperformed other methods⁴¹. However, the assessment was based on a small target set. To further evaluate AlphaFold on long multi-domain proteins, we compiled a test dataset of recent PDB chains that were not in the training set of the model. Only chains with more than 800 resolved residues were included, and a template filter was applied (Methods). Performance on this set was evaluated using the template modelling score (TM-score⁴²), 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 supplementary information of the companion AlphaFold paper² describes how a variety of useful predictors can be built on top of the main model. In particular, we can predict the residues that are likely to be experimentally resolved, and use them to produce a predicted

TM-score (pTM), in which the contribution of each residue is weighted by the probability of it being resolved (Supplementary Methods 1). The motivation for the weighting is to downweight unstructured parts of the prediction, producing a metric that better reflects the confidence of the model about the packing of the structured domains that are present. On the same recent PDB test dataset, pTM correlates well with the actual TM-score (Pearson's $r = 0.84$) (Fig. 2b). Notably, although some outliers in this plot are genuine failure cases, others appear to be plausible alternate conformations (for example, 6OFS chain A⁴³ in Fig. 2b).

We computed pTM scores for the human proteome, in an effort to identify multi-domain predictions that could 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 more than half the sequence. The distribution of pTM scores after applying the above filters is shown in Fig. 2c. Note that we would not

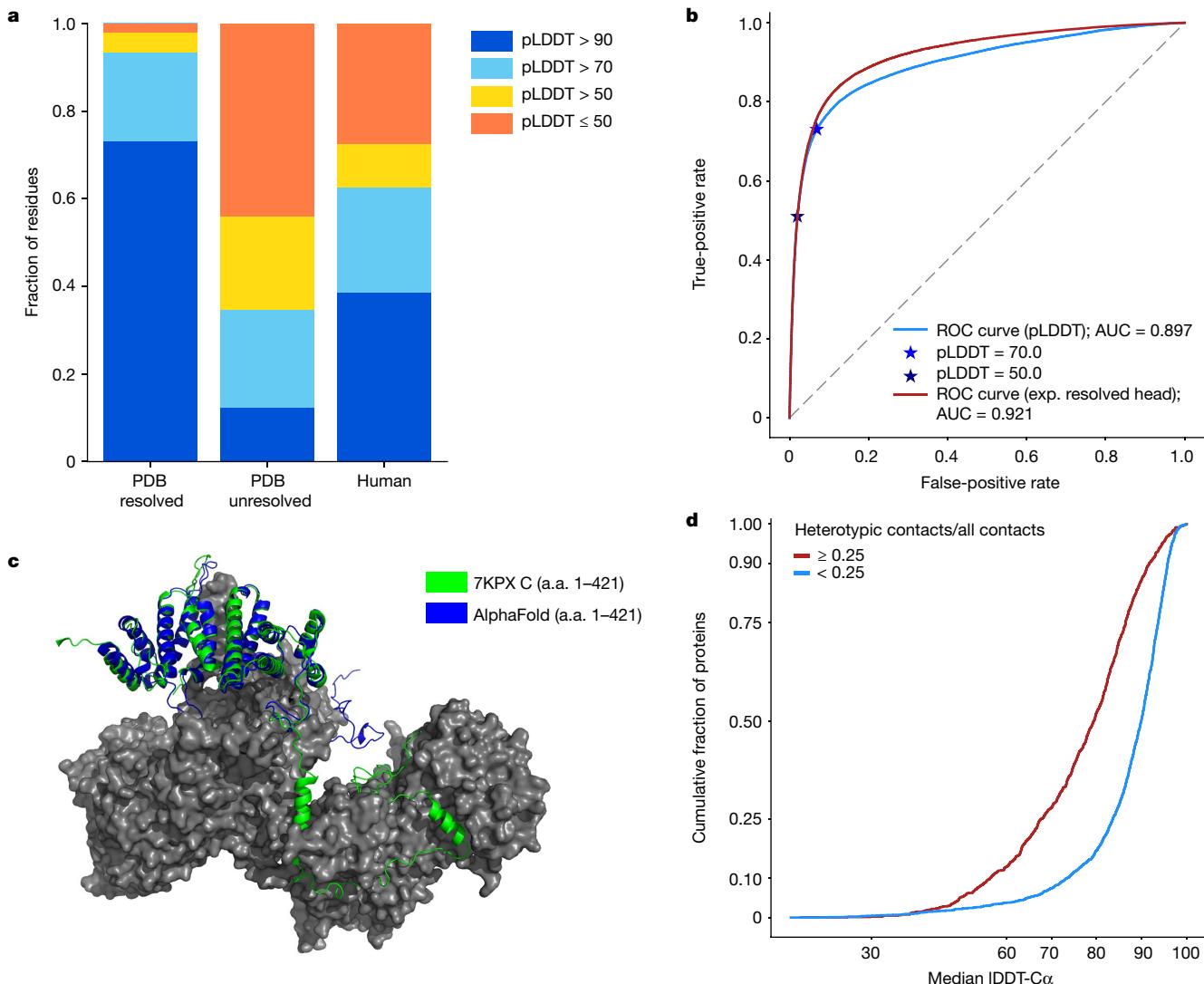


Fig. 4 | Low-confidence regions. **a**, pLDDT distribution of the resolved parts of PDB sequences ($n=3,440,359$ residues), the unresolved parts of PDB sequences ($n=589,079$ residues) and the human proteome ($n=10,537,122$ residues). **b**, Performance of pLDDT and the experimentally resolved head of AlphaFold as disorder predictors on the CAID Disprot-PDB benchmark dataset ($n=178,124$ residues). **c**, An example low-confidence prediction aligned to the corresponding PDB submission (7KPX chain C)⁶⁶. The globular domain is

well-predicted but the extended interface exhibits low pLDDT and is incorrect apart from some of the secondary structure. a.a., amino acid. **d**, A high ratio of heterotypic contacts is associated with a lower AlphaFold accuracy on the recent PDB dataset, restricted to proteins with fewer than 40% of residues with template identity above 30% ($n=3,007$ chains) (Methods). The ratio of heterotypic contacts is defined as: heterotypic/(intra-chain + homomeric + heterotypic).

expect uniformly high TM-scores to be achievable for this set, as some proteins will contain domains that are mobile relative to each other, with no fixed packing. Of the set, 187 proteins have $pTM > 0.8$ and 343 have $pTM > 0.7$. Although we expect the inter-domain accuracy of AlphaFold 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 that 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 catalyses the final step in glucose synthesis; it is therefore of critical importance to maintaining blood sugar levels. To our knowledge, no experimental structure exists, but previous studies have attempted to characterize the transmembrane topology⁴⁴ and active site⁴⁵. Our prediction has very high confidence (median pLDDT of 95.5) and gives a nine-helix topology with the putative active site accessible via an entry tunnel that is roughly in line with the surface of the endoplasmic reticulum (Fig. 3a and Supplementary Video 1). Positively charged residues in our prediction (median pLDDT of 96.6) align closely with the previously identified active site homologue in a fungal vanadium chloroperoxidase (PDB 1IDQ; r.m.s.d. of 0.56 Å; 49 out of 51 aligned atoms)⁴⁶. As these enzymes have distinct functions, we investigated 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 in 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 catalysing the final acyl addition in this pathway, and inhibiting DGAT2 has been shown to improve liver function in mouse models of liver disease⁴⁷. With our highly confident predicted structure (median pLDDT of 95.9), we set out to identify the binding pocket for a known inhibitor, PF-06424439 (ref. ⁴⁸). We identified a pocket (median pLDDT of 93.7) in which we were able to dock the inhibitor and observe specific interactions (Fig. 3b) that were not recapitulated in a negative example⁴⁹ (Extended Data Fig. 5 and Supplementary Methods 2). DGAT2 has an evolutionarily divergent but biochemically similar analogue, diacylglycerol *O*-acyltransferase 1 (DGAT1)⁵⁰. Within the binding pocket of DGAT2, we identified residues (Glu243 and His163) (Fig. 3b) that are analogous to the proposed catalytic residues in DGAT1 (His415 and Glu416)⁵¹, although we note that the nearby Ser244 in DGAT2 may present an alternative mechanism through an acyl-enzyme intermediate. Previous experimental research with DGAT2 has shown that mutating His163 has a stronger deleterious effect than mutating a histidine that is two residues away⁵². Additionally, Glu243 and His163 are conserved across species⁵⁰, supporting this hypothesized catalytic geometry.

Wolframin

Wolframin (UniProt O76024) is a transmembrane protein localized to the endoplasmic reticulum. 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 of 81.7) (Fig. 3c), we proposed identifying regions that are unique to this structure. A recent evolutionary analysis suggested domains for wolframin, which our prediction largely supports⁵⁵. 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 β 1 strand⁵⁶. The cysteine-rich region then forms an extended L12 loop with two predicted disulfide bridges, before looping back to the prototypical β 2 strand. Comparing our prediction for this region (median pLDDT of 86.0) to existing PDB chains using TM-align^{42,57} identified 3F1Z⁵⁸ as the most similar known chain (TM-score of 0.472) (Fig. 3c, magenta). Despite being the most similar chain, 3F1Z lacks the cysteines that are present in wolframin, which could form disulfide cross-links in the endoplasmic reticulum⁵⁹. As this region is hypothesized to recruit other proteins⁵⁵, these structural insights are probably important to understanding its partners.

Regions without a confident prediction

As we are applying AlphaFold to the entire human proteome, we would expect a considerable 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 previous study⁶² estimated that the percentage of disordered residues in the human proteome is between 37% and 50%. Thus disorder will have a large role when we consider a comprehensive set of predictions that covers 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 dataset²⁵. The results showed pLDDT to be a competitive disorder predictor compared with the current state of the art (SPOT-Disorder2⁶³), with an area under the curve (AUC) of 0.897 (Fig. 4b). Moreover, the supplementary information of the companion AlphaFold paper² 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 considerable 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 a recent PDB structure 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 in which the chain has a high percentage of heterotypic, cross-chain contacts (Fig. 4d).

In summary, our current interpretation of regions in which AlphaFold exhibits low pLDDT is that they have high 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 study, we generated comprehensive, state-of-the-art structure predictions for the human proteome. The resulting dataset makes a large contribution to the structural coverage of the proteome; particularly for tasks in which high accuracy is advantageous, such as molecular replacement or the characterization of binding sites. We also applied several metrics produced by building on the AlphaFold architecture—pLDDT, pTM and the experimentally resolved head—to demonstrate how they can be used to interpret our predictions.

Although we present 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 via <https://alphafold.ebi.ac.uk/>, we hope to enable exploration of new directions in structural bioinformatics.

The parts of the human proteome that are still without a confident prediction represent directions for future research. Some proportion of these will be genuine failures, in which a fixed structure exists but the current version of AlphaFold does not predict it. In many other cases, in which 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 important, medically relevant or economically important species. Structure prediction may have a more profound effect 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 could open up entirely new avenues of investigation. By providing scalable structure prediction with very high accuracy, AlphaFold could enable an exciting shift towards 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* **391**, 1038–1042 (2021).
3. International Human Genome Sequencing Consortium. 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. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
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. Greer, 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**, 3977 (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**, 1895 (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. The Gene Ontology Consortium. 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**, 244 (2019).
24. Gligorjević, V. et al. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**, 3168 (2021).
25. Necci, M., Piovesan, D., CAID Predictors, DisProt Curators & Tosatto, S. C. E. Critical assessment of protein intrinsic disorder prediction. *Nat. Methods* **18**, 472–481 (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* **87**, 1011–1020 (2019).
30. Pereira, J. et al. High-accuracy protein structure prediction in CASP14. *Proteins* <https://doi.org/10.1002/prot.26171> (2021).
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. et al. Functional assessment. https://predictioncenter.org/casp14/doc/presentations/2020_12_03_Function_Assessment_VajdaLab_KozakovLab.pdf (2020).
35. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).
36. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
37. The Gene Ontology Consortium. 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* **87**, 1378–1387 (2019).
40. Haas, J. et al. Continuous Automated Model Evaluation (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins* **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* **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**, eaav9701 (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 pyrimidinoxazinyl 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**, 249 (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 KPN0353 (gi|52972051), a novel putative lipoprotein from *Klebsiella pneumoniae*, reveals an OB-fold. *Acta Crystallogr. F* **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 (2013).
63. Hanson, J., Palitwa, 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**, 39 (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).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Methods

Structure prediction (human proteome)

Sequences for the human reference proteome were obtained from UniProt release 2021_02⁶. Structure prediction was attempted for all sequences with 16–2,700 amino acids; sequences with residue codes B, J, O, U, Z or X were excluded. The length ceiling of 2,700 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 paper², 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 file containing the first non-consensus sequence from each BFD a3m alignment. 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 with more than 2,700 residues account for the majority of exclusions. This amounts to 10,537,122 residues (92.5% of residues).

Structure prediction (recent PDB dataset)

For structure predictions of recent PDB sequences, we used a copy of the PDB downloaded on 15 February 2021. Structures were filtered to those with a release date after 30 April 2018 (the date limit for inclusion in the training set). Chains were then further filtered to remove sequences that consisted 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 fewer than 16 resolved residues, with unknown residues and structures solved by NMR methods 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 April 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, producing a dataset of 12,494 chains. This is referred to as the recent PDB dataset.

Computational resources

Inference was run on V100 graphics processing units (GPUs), with each sequence inferred 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 1,401–2,000 residues were processed by workers with two GPUs, and those of length 2,001–2,700 residues by workers with four GPUs (further details of unified memory on longer proteins are provided in the companion paper²; 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; around 190 GPU days would be sufficient to inference each protein once. Long sequences had a disproportionate effect owing to the multi-GPU workers described above. Approximately 250 GPU days would have been sufficient to produce five models for all proteins shorter than 1,400 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 (for example, 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 Figure 7 shows 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 on 15 February 2021) using default flags⁶⁷. The prior template coverage at residue *i* is the maximum percentage sequence identity of all hits covering residue *i*, regardless of whether the hit residue is experimentally resolved. For the recent PDB analysis, only template hits corresponding to a structure released before 30 April 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 more than 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 change log <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 more than 3,000 annotations, and from those selecting only moderately specific terms (a term cannot have a child with more than 3,000 annotations). The remaining terms in the ‘molecular function’ and ‘cellular component’ ontologies are shown in Fig. 1d.

Structure analysis

Structure images were created in PyMOL⁶⁹, and PyMOL align was used to compute r.m.s.d.s (outlier rejection is described in the text where applicable).

For docking against DGAT2, P2Rank⁶⁵ was used to identify ligand-binding pockets in the AlphaFold structure. AutoDockTools⁷⁰ 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 Vina⁷¹ was run with an exhaustiveness parameter of 32, a seed of 0 and a docking search space of $25 \times 25 \times 25 \text{ \AA}^3$ centred at the point identified by P2Rank.

For identifying the most similar structure to wolframin, TM-align⁴² was used to compare against all PDB chains (downloaded 15 February 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 supplementary information section 1.9.7 of the companion AlphaFold paper² and the implementation of the experimentally resolved head is described in supplementary information section 1.9.10 of the companion AlphaFold paper². The weighted version of pTM is described in Supplementary Methods 1.

Analysis of low-confidence regions

For evaluation on CAID, the target sequences and ground-truth labels for the Disprot-PDB dataset were downloaded from <https://idpcentral>.

Article

org/. Structure prediction was performed as described above for the recent PDB dataset, with a template cut-off of 30 April 2018. To enable complete coverage, two sequences containing non-standard residues (X, U) had these remapped to G (glycine). Sequences longer than 2,000 residues were split into two segments: 1–2,000 and 2,000–end, and the pLDDT and experimentally resolved head arrays were concatenated for evaluation. The two evaluated disorder predictors were taken to be $1 - 0.01 \times \text{pLDDT}$ and $1 - \text{predicted resolvability for C}\alpha\text{ atoms}$.

To obtain the ratio of heterotypic contacts to all contacts (Fig. 4d), two residues are considered in contact if their C β atoms (or C α for glycine) are within 8 Å 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.

Comparison with BestSingleStructuralTemplate

CAMEO data for the period 21 March 2020 to 13 March 2021 were 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 maximum template date of 1 March 2020. Predictions were scored against the CAMEO ground truth using IDDT-C α . For BestSingleStructuralTemplate, IDDT-C α 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 the PDB downloaded on 15 February 2021 with a template search performed using Hmmsearch. Templates were filtered to those with at least 70% coverage of the sequence and a release date before 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 Figure 4 shows the distribution of IDDT-C α for each model within each bin as a box plot (horizontal line at the median, box spanning from the lower to the upper quartile, whiskers extending to the minimum and maximum). In total 428 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

Structure predictions by AlphaFold 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 v.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 v.2020_03 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2020_03/uniref/), MGnify clusters v.2018_12 (https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2018_12/) and a reduced version of BFD (produced as outlined in the Methods using the 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 on 15 February 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 were drawn from the same copy of the PDB; we show structures with accessions 6YJ1⁶⁴, 6OFS⁴³, IIDQ⁴⁶, 1PRT⁷², 3F1Z⁵⁸, 7KPx⁶⁶ and 6VPO⁵¹. The template search used PDB70, downloaded on 10 February 2021 (http://wwwuser.gwdg.de/~combiol/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 on 17 March 2021 at https://www.cameo3d.org/static/downloads/modeling/1-year/raw_targets-1-year.public.tar.gz. A copy of the current Gene Ontology database was downloaded on 29 April 2021 from <http://current.geneontology.org/ontology/go.obo>. Source data are provided with this paper.

Code availability

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 v.1 (<https://github.com/tensorflow/tensorflow>), Sonnet v.1 (<https://github.com/deepmind/sonnet>), JAX v.0.1.69 (<https://github.com/google/jax/>) and Haiku v.0.0.4 (<https://github.com/deepmind/dm-haiku>).

For MSA search on UniRef90, MGnify clusters and the reduced BFD, we used jackhmmer and for the template search on the PDB SEQRES we used hmmsearch, both from HMMER v.3.3 (<http://eddylab.org/software/hmmr/>). For the template search against PDB70, we used HHsearch from HH-suite v.3.0-beta.3 14/07/2017 (<https://github.com/soedinglab/hh-suite>). For constrained relaxation of structures, we used OpenMM v.7.3.1 (<https://github.com/openmm/openmm>) with the Amber99sb force field.

Docking analysis on DGAT used P2Rank v.2.1 (<https://github.com/rdk/p2rank>), MGLTools v.1.5.6 (<https://ccsb.scripps.edu/mgltools/>) and AutoDockVina v.1.1.2 (<http://vina.scripps.edu/download/>) on a workstation running Debian GNU/Linux rodate 5.10.40-1rodate1-amd64 x86_64.

Data analysis used Python v.3.6 (<https://www.python.org/>), NumPy v.1.16.4 (<https://github.com/numpy/numpy>), SciPy v.1.2.1 (<https://www.scipy.org/>), seaborn v.0.11.1 (<https://github.com/mwaskom/seaborn>), scikit-learn v.0.24.0 (<https://github.com/scikit-learn/>), Matplotlib v.3.3.4 (<https://github.com/matplotlib/matplotlib>), pandas v.1.1.5 (<https://github.com/pandas-dev/pandas>) and Colab (<https://research.google.com/colaboratory>). TM-align v.20190822 (<https://zhanglab.dcmbl.med.umich.edu/TM-align>) was used for computing TM-scores. Structure analysis used PyMol v.2.3.0 (<https://github.com/schrodinger/pymol-open-source>).

67. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).
68. Steinegger, M., Mirdita, M. & Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manifold. *Nat. Methods* **16**, 603–606 (2019).
69. Schrödinger. The PyMOL Molecular Graphics System v1.8 (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 flavours in proteins. *Bioinformatics* **36**, 5533–5534 (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).

Acknowledgements We thank A. Paterson, C. Low, C. Donner, D. Evans, F. Yang, J. Stanway, J. Stanton, L. Deason, N. Latysheva, N. Hobbs, R. Hadsell, R. Green, S. Brown, V. Bolina, Ž. Avsec and the Research Platform Team for their contributions; R. Kemp for help in managing the project and our colleagues at DeepMind, Google and Alphabet for their encouragement and support; E. van Schaftingen, M. Zhou and F. Urano for reading and commenting on our discussion of glucose-6-phosphatase, diacylglycerol O-acyltransferase 2 and wolframin, respectively; the team at EMBL-EBI for their work on making AlphaFold structure predictions available, in particular M. Varadi, M. Deshpande, S. Sasidharan Nair, S. Anyango, G. Yordanova, C. Natassia, D. Yuan and E. 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., A. Bridgland and A.C. generated and analysed the structure predictions. J.J., M.F., S.A.A.K.

and O.R. developed the metrics used to interpret predictions. A.Ž., 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., A. Bateman, G.J.K., D.H. and E.B. managed the work to make AlphaFold predictions available via EMBL-EBI-hosted resources. S.V., G.J.K. and A. Bateman provided scientific advice on how predictions should be displayed. J.J., R.E., A. Pritzel, M.F., O.R., R.B., A. Potapenko, S.A.A.K., B.R.-P., J.A., A.W.S., T.G., A.Ž., K.T., A. Bridgland, A.J.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., A. Pritzel, T.G., M.F., O.R., R.B., A. Bridgland, S.A.A.K., D.R. and A.W.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, which are genomics companies. The other authors declare no competing interests.

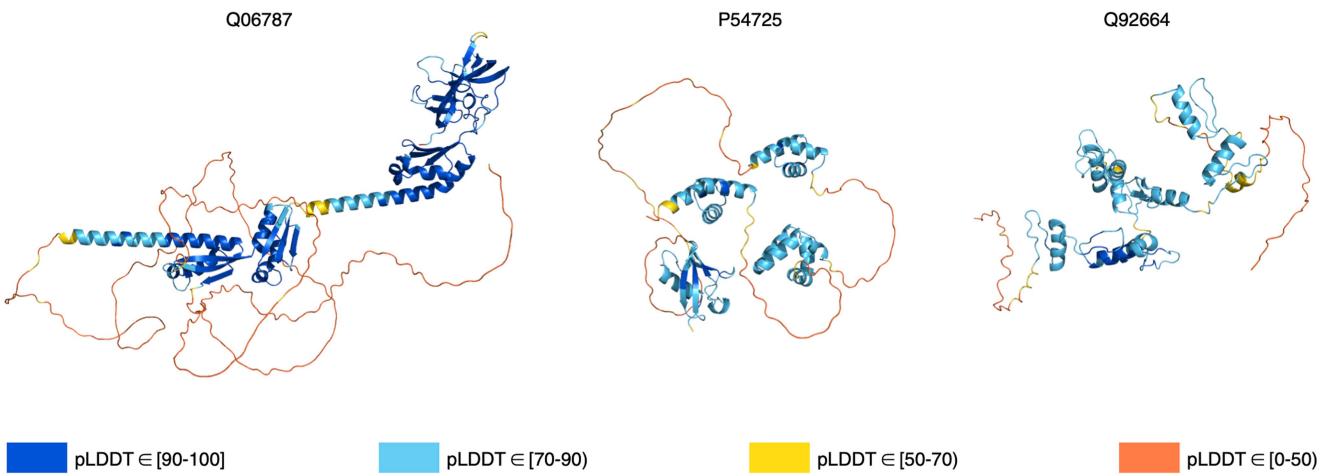
Additional information

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.

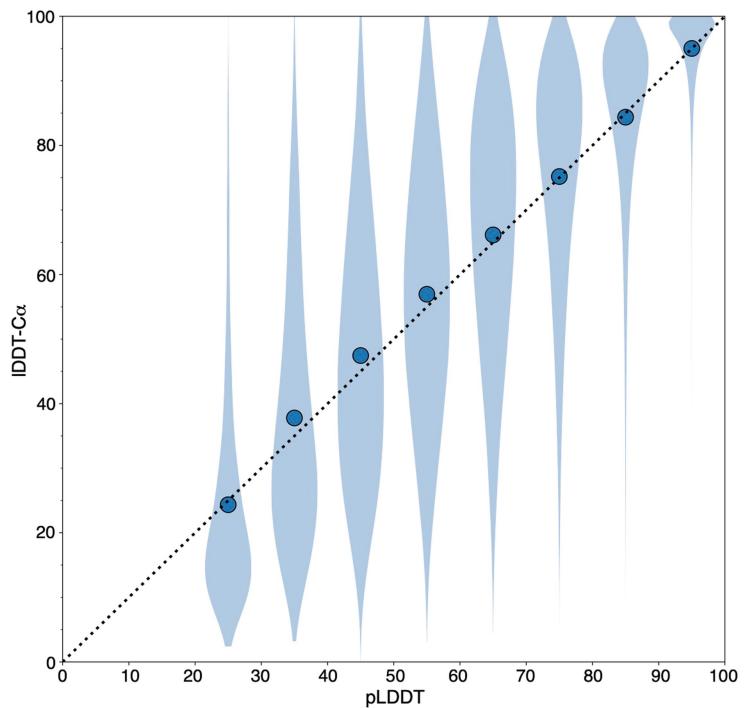
Peer review information *Nature* thanks Mohammed AlQuraishi, Yang Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



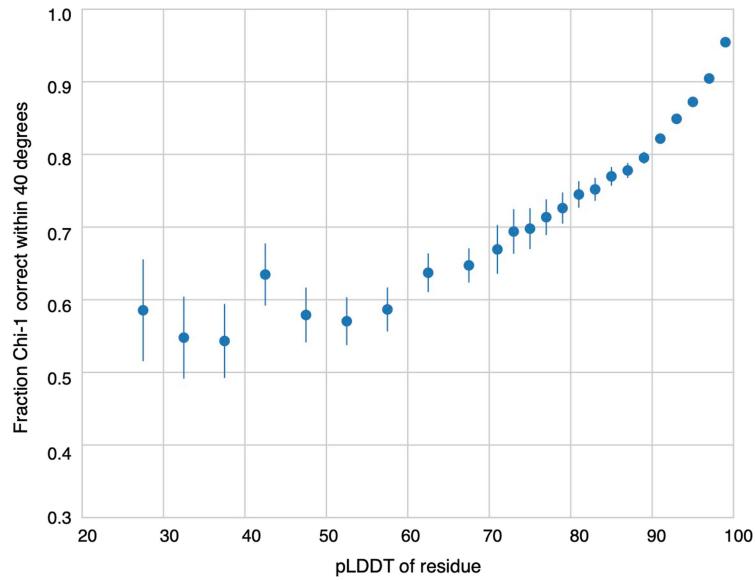
Extended Data Fig. 1 | Example full chain outputs containing both high- and low-confidence regions. Q06787 (synaptic functional regulator FMRI) and P54725 (UV excision repair protein RAD23 homologue A) are predicted to be

disordered outside the experimentally determined regions by MobiDB⁷³. Q92664 (transcription factor IIIA) has been described as ‘beads on a string’, consisting of zinc-finger domains joined by flexible linkers⁷⁴.



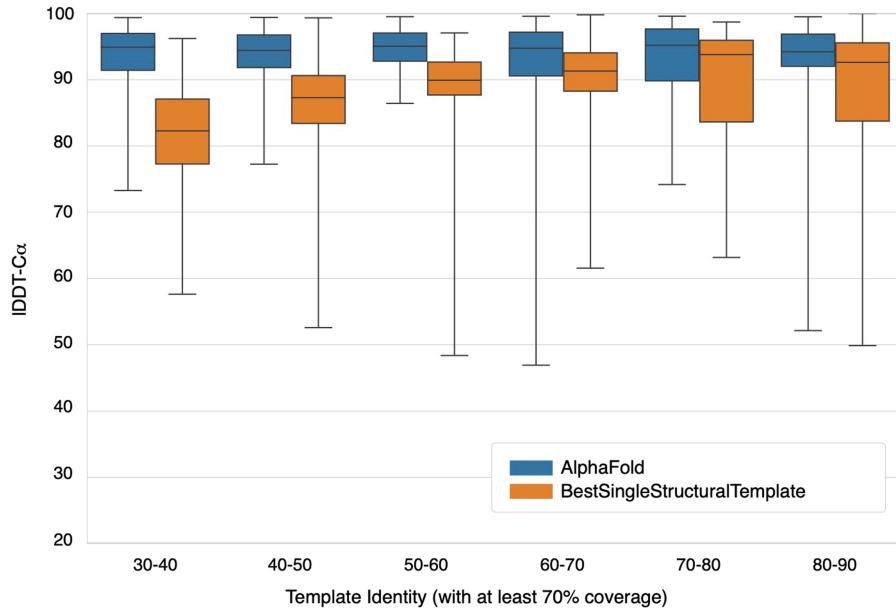
Extended Data Fig. 2 | Distribution of per-residue IDDT-C α within eight pLDDT bins. This represents an alternative visualization to Fig. 1a that does not sample the data. It uses the recent PDB dataset (Methods), which is restricted to structures with a reported resolution of <3.5 Å ($n=2,756,569$ residues). Residues were assigned to bins of width 10 based on their pLDDT

(minimum, 20; maximum, 100). Markers show the mean IDDT-C α within each bin, while the IDDT-C α distribution is visualized as a Matplotlib violin plot (kernel density estimate bandwidth, 0.2). The smallest sample size for the corresponding violin is 5,655 residues for the left-most bin.



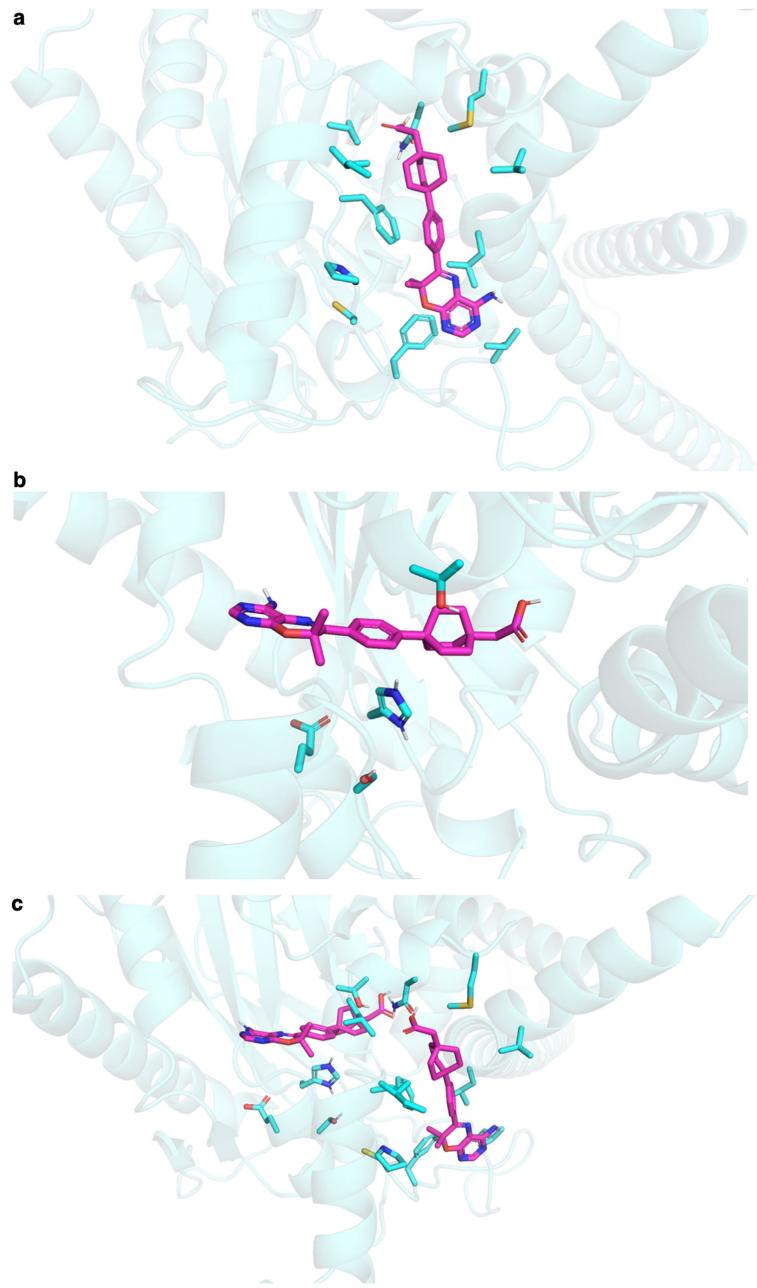
Extended Data Fig. 3 | Relationship between pLDDT and side-chain χ_1 correctness. Evaluated on the recent PDB dataset (Methods), which is restricted to structures with a reported resolution of $<2.5\text{ \AA}$ ($n=5,983$ chains) and residues with a B -factor of $<30\text{ \AA}^2$ ($n=609,623$ residues). Residues are binned by pLDDT, with bin width 5 between 20 and 70 pLDDT and bin width 2

above 70 pLDDT. A χ_1 angle is considered correct if it is within 40° of its value in the PDB structure⁷⁵. Markers show the proportion of correct χ_1 angles within each bin; error bars indicate the 95% confidence interval (two-sided Student's t -test). The smallest sample size for the error bars is 193 residues for the left-most bin.



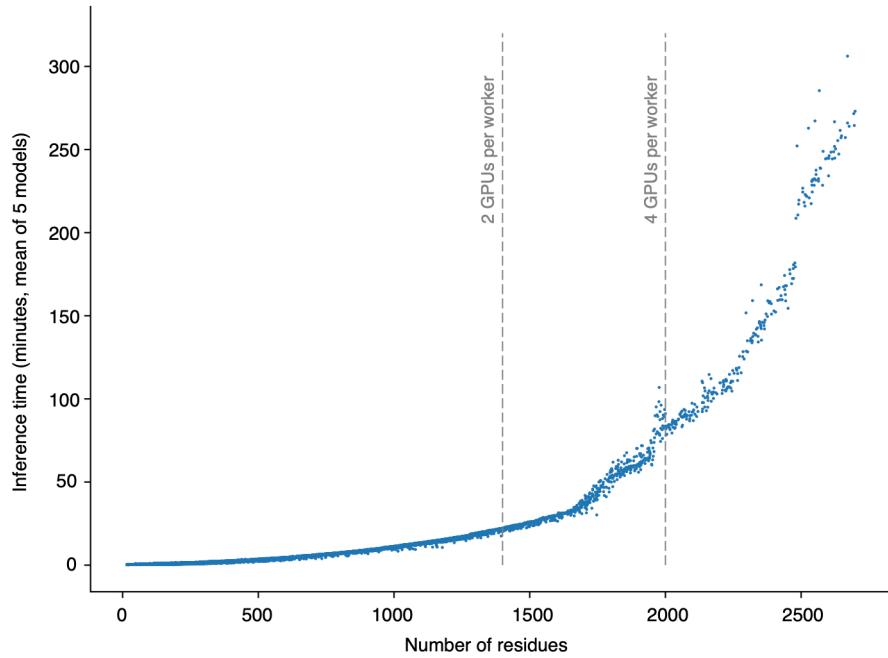
Extended Data Fig. 4 | AlphaFold performance at a range of template sequence identities. IDDT-C α for AlphaFold and BestSingleStructuralTemplate on 1 year of CAMEO targets³⁹. Targets are binned according to the sequence identity of the best template covering at least 70% of the target, and a box plot

is shown for each bin. The horizontal line indicates the median, boxes range from the lower to the upper quartile, and the whiskers extend from the minimum to the maximum. In total, 428 targets are included (see Source Data); the smallest number of targets in any bin is 18.



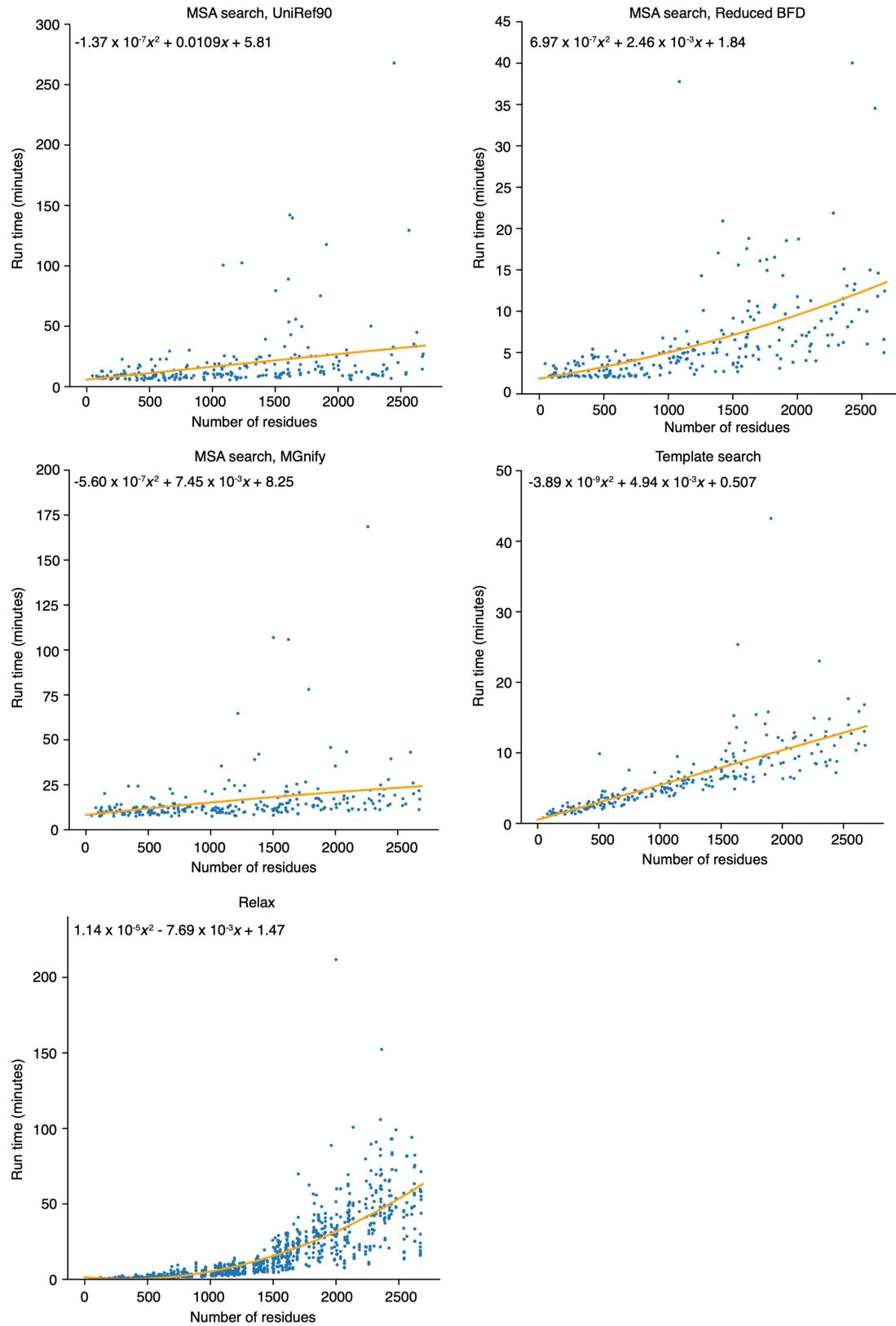
Extended Data Fig. 5 | Docking poses for a DGAT1-specific inhibitor in DGAT2. **a**, Top binding pose from Autodock Vina for a DGAT1-specific inhibitor in DGAT2, which does not match the predicted binding pocket for a DGAT2-specific inhibitor. **b**, Next best binding pose, which matches the

binding pocket for the DGAT2-specific inhibitor, but does not contain components that satisfy the polar side chains His163 and Thr194. **c**, Relative positions of both binding poses.



Extended Data Fig. 6 | Relationship between sequence length and inference time. On the basis of logs from our human proteome set. All of the 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 cut-offs above which sequences were processed by multi-GPU workers.



Extended Data Fig. 7 | Relationship between sequence length and run time for the non-inference stages of the pipeline. On the basis of 240 human protein sequences, chosen by stratified sampling from the length buckets: [16, 500), [500, 1,000), [1,000, 1,500), [1,500, 2,000), [2,000, 2,500] and [2,500, 2,700]. 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 | IDDT-C α distribution in various pLDDT bins

	IDDT-C α					
	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%

Data are based on the per-residue IDDT-C α and per-residue pLDDT of resolved regions. This table uses the recent PDB dataset (Methods), which is restricted to structures with a reported resolution of <3.5 Å. The total number of chains included is 10,215.

Article

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 dataset (Methods), which restricted to structures with a reported resolution of <3.5 Å. The total number of chains included is 10,215.

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/hmmmer/>). 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 6VPO. Template search used PDB70, downloaded 10/02/2021 (http://wwwuser.gwdg.de/~combiol/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
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging