AlphaFold: Improved protein structure prediction using potentials from deep learning

Andrew W. Senior[1*], Richard Evans[1*], John Jumper[1*], James Kirkpatrick[1*], Laurent Sifre[1*], Tim Green[1], Chongli Qin[1], Augustin Žídek[1], Alexander W. R. Nelson[1], Alex Bridgland[1], Hugo Penedones[1], Stig Petersen[1], Karen Simonyan[1], Steve Crossan[1], Pushmeet Kohli[1], David T. Jones[2,3], David Silver[1], Koray Kavukcuoglu[1], Demis Hassabis[1]

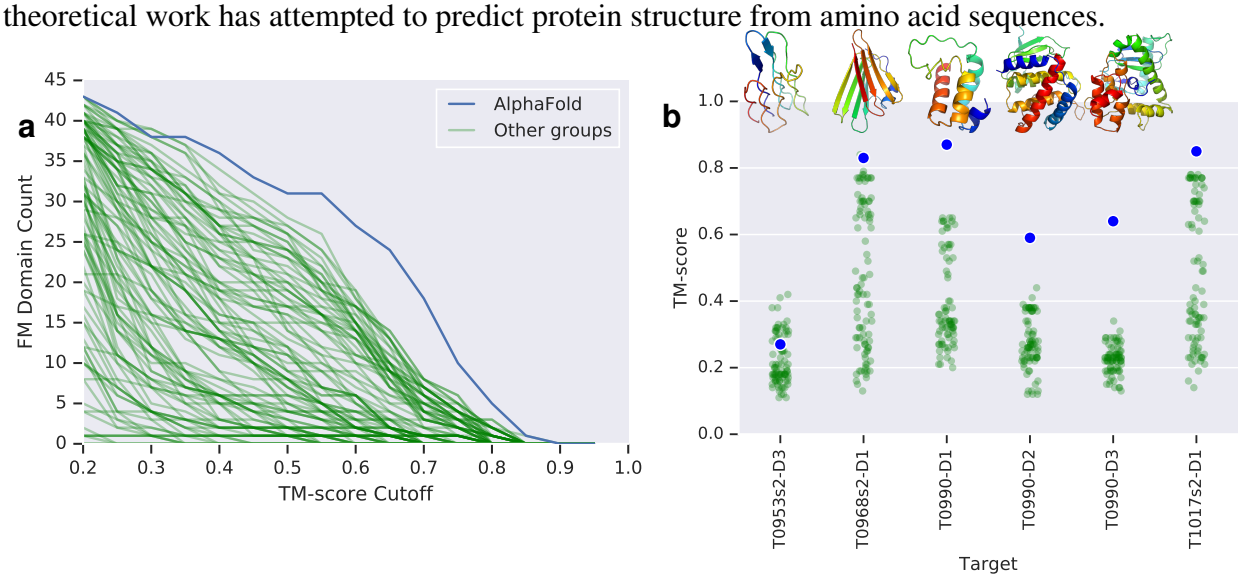
[1]*DeepMind, London, UK*
[2]*The Francis Crick Institute, London, UK*
[3]*University College London, London, UK*
[*]*These authors contributed equally to this work.*

Protein structure prediction aims to determine the three-dimensional shape of a protein from its amino acid sequence[1]. This problem is of fundamental importance to biology as the structure of a protein largely determines its function[2] but can be hard to determine experimentally. In recent years, considerable progress has been made by leveraging genetic information: analysing the co-variation of homologous sequences can allow one to infer which amino acid residues are in contact, which in turn can aid structure prediction[3]. In this work, we show that we can train a neural network to accurately predict the distances between pairs of residues in a protein which convey more about structure than contact predictions. With this information we construct a potential of mean force[4] that can accurately describe the shape of a protein. We find that the resulting potential can be optimised by a simple gradient descent algorithm, to realise structures without the need for complex sampling procedures. The resulting system, named AlphaFold, has been shown to achieve high accuracy, even for sequences with relatively few homologous sequences. In the most recent Critical Assessment of Protein Structure Prediction[5] (CASP13), a blind assessment of the state of the field of protein structure prediction, AlphaFold created high-accuracy structures (with TM-scores[†] of 0.7 or higher) for 24 out of 43 free modelling domains whereas the next best method, using sampling and contact information, achieved such accuracy for only 14 out of 43 domains. AlphaFold represents a significant advance in protein structure prediction. We expect the increased accuracy of structure predictions for proteins to enable insights in understanding the function and malfunction of these proteins, especially in cases where no homologous proteins have been experimentally determined[7].

Proteins are at the core of most biological processes. Since the function of a protein is dependent on its structure, understanding protein structure has been a grand challenge in biology for decades. While several experimental structure determination techniques have been developed

[†]Template Modelling score[6], between 0 and 1, measures the degree of match of the overall (backbone) shape of a proposed structure to a native structure.

and improved in accuracy, they remain difficult and time-consuming[2]. As a result, decades of theoretical work has attempted to predict protein structure from amino acid sequences.



c

Contact precisions		L long			L/2 long			L/5 long		
Set	N	AF	498	032	AF	498	032	AF	498	032
FM	31	**45.5**	42.9	39.8	**58.0**	55.1	51.7	**70.1**	67.3	61.6
FM/TBM	12	**59.1**	53.0	48.9	**74.2**	64.5	64.2	**85.3**	81.0	79.6
TBM	61	**68.3**	65.5	61.9	**82.4**	80.3	76.4	**90.6**	90.5	87.1

Fig. 1 | AlphaFold's performance in the CASP13 assessment. (a) Number of free modelling (FM + FM/TBM) domains predicted to a given TM-score threshold for AlphaFold and the other 97 groups. (b) For the six new folds identified by the CASP13 assessors, AlphaFold's TM-score compared with the other groups, with native structures. The structure of T1017s2-D1 is unavailable for publication. (c) Precisions for long-range contact prediction in CASP13 for the most probable L, $L/2$ or $L/5$ contacts, where L is the length of the domain. The distance distributions used by AlphaFold (AF) in CASP13, thresholded to contact predictions, are compared with submissions by the two best-ranked contact prediction methods in CASP13: 498 (RaptorX-Contact[8]) and 032 (TripletRes[9]), on "all groups" targets, excluding T0999.

CASP[5] is a biennial blind protein structure prediction assessment run by the structure prediction community to benchmark progress in accuracy. In 2018, AlphaFold joined 97 groups from around the world in entering CASP13. Each group submitted up to 5 structure predictions for each of 84 protein sequences whose experimentally-determined structures were sequestered. Assessors divided the proteins into 104 domains for scoring and classified each as being amenable to *template-based modelling* (TBM, where a protein with a similar sequence has a known structure, and that homologous structure is modified in accordance with the sequence differences) or requiring *free modelling* (FM, when no homologous structure is available), with an intermediate (FM/TBM) category. Figure 1a shows that AlphaFold stands out in performance above the other entrants, predicting more FM domains to high accuracy than any other system, particularly in the

0.6–0.7 TM-score range. The assessors ranked the 98 participating groups by the summed, capped z-scores of the structures, separated according to category. AlphaFold achieved a summed z-score of 52.8 in the FM category (best-of-5) vs 36.6 for the next closest group (322)[‡]. Combining FM and TBM/FM categories, AlphaFold scored 68.3 vs 48.2. AlphaFold is able to predict previously unknown folds to high accuracy as shown in Figure 1b. Despite using only free modelling techniques and not using templates, AlphaFold also scored well in the TBM category according to the assessors' formula 0-capped z-score, ranking fourth by the top-1 model or first by the best-of-5 models. Much of the accuracy of AlphaFold is due to the accuracy of the distance predictions, which is evident from the high precision of the contact predictions of Table 1c.

The most successful free modelling approaches so far[10–12] have relied on *fragment assembly* to determine the shape of the protein of interest. In these approaches a structure is created through a stochastic sampling process, such as simulated annealing[13], that minimises a statistical potential derived from summary statistics extracted from structures in the Protein Data Bank (PDB[14]). In fragment assembly, a structure hypothesis is repeatedly modified, typically by changing the shape of a short section, retaining changes which lower the potential, ultimately leading to low potential structures. Simulated annealing requires many thousands of such moves and must be repeated many times to have good coverage of low-potential structures.

In recent years, structure prediction accuracy has improved through the use of evolutionary covariation data[15] found in sets of related sequences. Sequences similar to the target sequence are found by searching large datasets of protein sequences derived from DNA sequencing and aligned to the target sequence to make a *multiple sequence alignment* (MSA). Correlated changes in two amino acid residue positions across the sequences of the MSA can be used to infer which residues might be in contact. Contacts are typically defined to occur when the β-carbon atoms of two residues are within 8 Ångström of one another. Several methods have been used to predict the probability that a pair of residues is in contact based on features computed from MSAs[16–19] including neural networks[20–23]. Contact predictions are incorporated in structure prediction by modifying the statistical potential to guide the folding process to structures that satisfy more of the predicted contacts[12,24]. Previous work[25,26] has made predictions of the distance between residues, particularly for distance geometry approaches[8,27–29]. Neural network distance predictions without covariation features were used to make the EPAD potential[26] which was used for ranking structure hypotheses and the QUARK pipeline[12] used a template-based distance profile restraint for template-based modelling.

In this work we present a new, deep-learning, approach to protein structure prediction, whose stages are illustrated in Figure 2a. We show that it is possible to construct a learned, protein-specific potential by training a neural network (Fig. 2b) to make accurate predictions about the structure of the protein given its sequence, and to predict the structure itself accurately by minimising the

[‡]Results from `http://predictioncenter.org/casp13/zscores_final.cgi?formula=assessors`

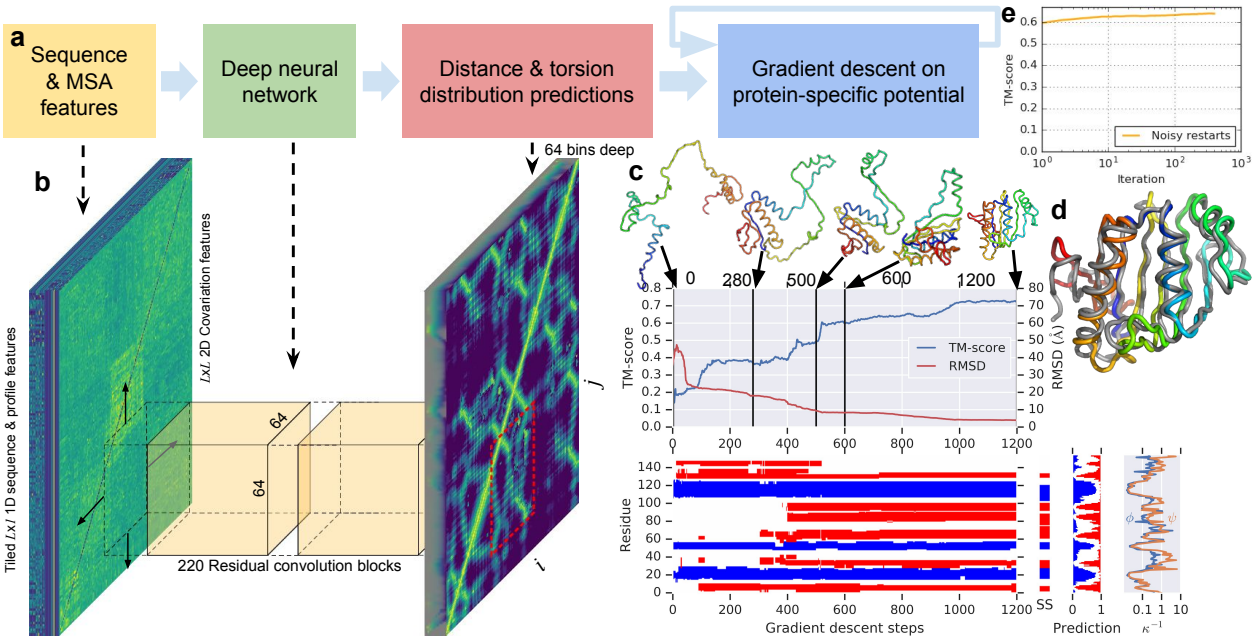


Fig. 2 | The folding process illustrated for CASP13 target T0986s2. (Length $L = 155$) (a) Steps of structure prediction. (b) The neural network predicts the entire $L \times L$ distogram based on MSA features, accumulating separate predictions for 64×64-residue regions. (c) One iteration of gradient descent (1 200 steps) is shown, with TM-score and RMSD plotted against step number with five snapshots of the structure. The secondary structure (from SST[30]) is also shown (helix in blue, strand in red) along with the the native secondary structure (SS), the network's secondary structure prediction probabilities and the uncertainty in torsion angle predictions (as κ^{-1} of the von Mises distributions fitted to the predictions for ϕ and ψ). While each step of gradient descent greedily lowers the potential, large global conformation changes are effected, resulting in a well-packed chain. (d) shows the final first submission overlaid on the native structure (in grey). (e) shows the average (across the test set, $n = 377$) TM-score of the lowest-potential structure against the number of repeats of gradient descent (log scale).

potential by gradient descent (Fig. 2c). The neural network predictions include backbone torsion angles and pairwise distances between residues. Distance predictions provide more specific information about the structure than contact predictions and provide a richer training signal for the neural network. Predicting distances, rather than contacts as in most prior work, models detailed interactions rather than simple binary decisions. By jointly predicting many distances, the network can propagate distance information respecting covariation, local structure and residue identities to nearby residues. The predicted probability distributions can be combined to form a simple, principled protein-specific potential. We show that with gradient descent, it is simple to find a set of torsion angles that minimise this protein-specific potential using only limited sampling. We also show that whole chains can be optimised together, avoiding the need for segmenting long proteins into hypothesised domains which are modelled independently.

The central component of AlphaFold is a convolutional neural network which is trained on PDB structures to predict the distances d_{ij} between the C_β atoms of pairs, ij, of a protein's residues. Based on a representation of the protein's amino acid sequence, $\mathcal{S}$, and features derived from the sequence's MSA, the network, similar in structure to those used for image recognition tasks[31], predicts a discrete probability distribution $P(d_{ij} \mid \mathcal{S}, \mathrm{MSA}(\mathcal{S}))$ for every ij pair in a 64×64 residue region, as shown in Fig. 2b. The full set of distance distribution predictions is constructed by averaging predictions for overlapping regions and is termed a *distogram* (from distance histogram). Figure 3 shows an example distogram prediction for one CASP protein, T0955. The modes of the distribution (Fig. 3c) can be seen to closely match the true distances (Fig. 3b). Example distributions for all distances to one residue (29) are shown in Fig. 3c. Further analysis of how the network predicts the distances is shown in Methods Figure 14.

In order to realise structures that conform to the distance predictions, we construct a smooth potential V_{distance} by fitting a spline to the negative log probabilities, and summing across all the residue pairs. We parameterise protein structures by the backbone torsion angles ($\boldsymbol{\phi}$, $\boldsymbol{\psi}$) of all residues and build a differentiable model of protein geometry $\mathbf{x} = G(\boldsymbol{\phi}, \boldsymbol{\psi})$ to compute the C_β coordinates, $\mathbf{x}$, and thus the inter-residue distances, $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$, for each structure, and express V_{distance} as a function of $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$. For a protein with L residues, this potential accumulates L^2 terms from marginal distribution predictions. To correct for the over-representation of the prior we subtract a *reference distribution*[32] from the distance potential in the log domain. The reference distribution models the distance distributions $P(d_{ij} \mid \mathrm{length})$ independent of the protein sequence and is computed by training a small version of the distance prediction neural network on the same structures, without sequence or MSA input features. A separate output head of the contact prediction network is trained to predict discrete probability distributions of backbone torsion angles $P(\phi_i, \psi_i \mid \mathcal{S}, \mathrm{MSA}(\mathcal{S}))$. After fitting a von Mises distribution, this is used to add a smooth torsion modelling term $V_{\mathrm{torsion}} = -\sum \log p_{\mathrm{vonMises}}(\phi_i, \psi_i \mid \mathcal{S}, \mathrm{MSA}(\mathcal{S}))$ to the potential. Finally, to prevent steric clashes, we add Rosetta's $V_{\texttt{score2\_smooth}}$[10] to the potential, as this incorporates a van der Waals term. We used multiplicative weights for each of the three terms in the potential, but no weighting noticeably outperformed equal weighting.

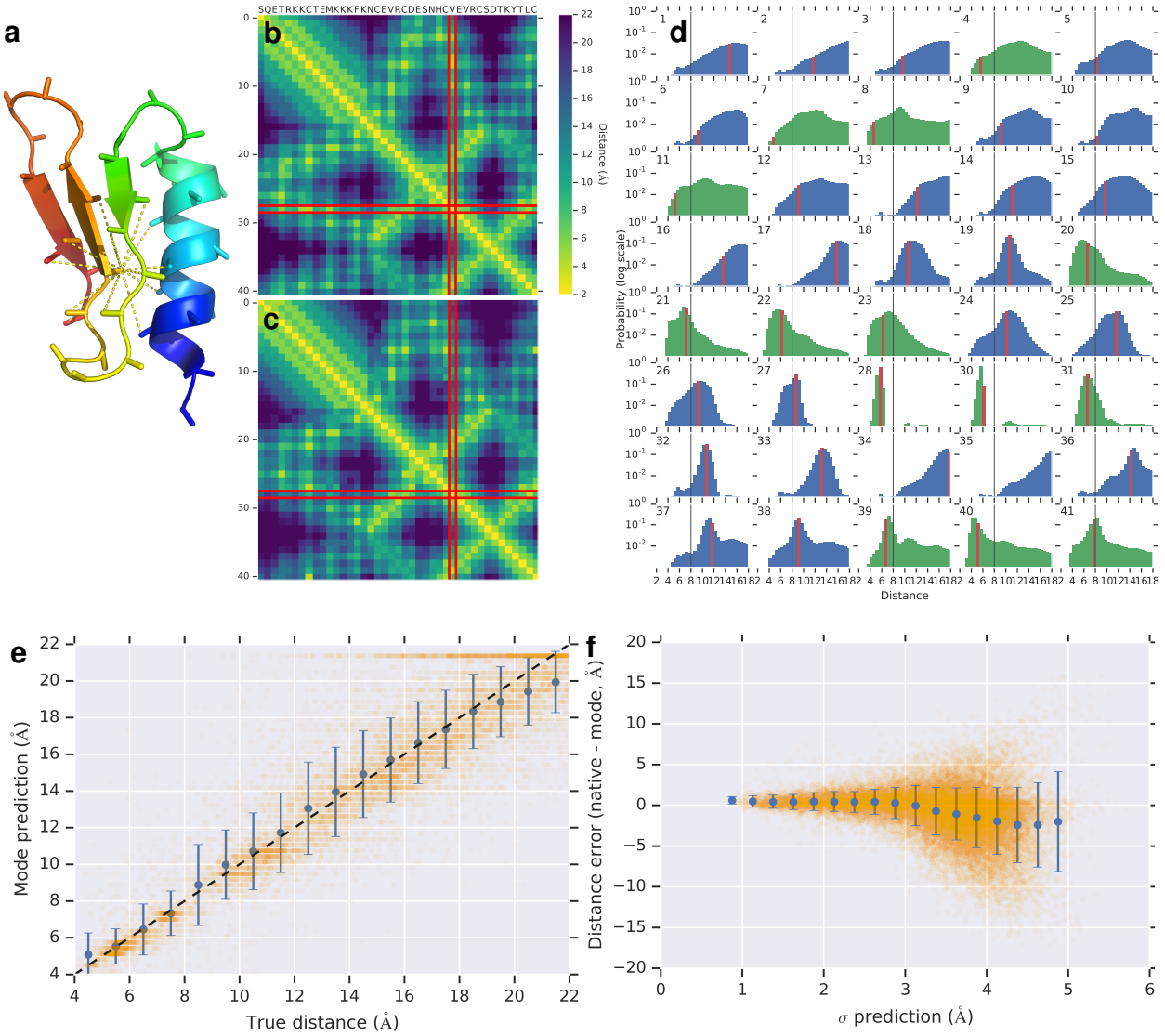


Fig. 3 | Predicted distance distributions compared with true distances. Above, for CASP target T0955 ($L = 41$): (a) Native structure showing distances under 8 Å from C_β of residue 29. (b) Native inter-residue distances and (c) the mode of the distance predictions, highlighting residue 29. (d) The predicted probability distributions for distances of residue 29 to all other residues. The bin corresponding to the native distance is highlighted in red, 8 Å drawn in black. True contacts' distributions are plotted in green, non-contacts in blue. Below, for CASP target T0990 ($L = 552$): (e) the mode of the predicted distance plotted against the true distance for all residue pairs with distances $\leq$22 Å, excluding distributions with standard deviation > 3.5 Å. The blue error bars show mean and standard deviation calculated for 1 Å bins. (f) The error of the mode distance prediction vs the standard deviation of the distance distributions, excluding pairs with native distances >22 Å). Mean and standard deviations are shown for 0.25 Å bins. The distogram is shown in Figure 7 in Methods.

Since all the terms in the combined potential $V_{\text{total}}(\boldsymbol{\phi}, \boldsymbol{\psi})$ are differentiable functions of $(\boldsymbol{\phi}, \boldsymbol{\psi})$, it can be optimised with respect to these variables by gradient descent. Here we use L-BFGS[33]. Structures are initialised by sampling torsion values from $P(\phi_i, \psi_i \mid \mathcal{S}, \mathrm{MSA}(\mathcal{S}))$. Figure 2c illustrates a single gradient descent trajectory minimising the potential, showing how this greedy optimisation process leads to increasing accuracy and large-scale conformation changes. Secondary structure is partly set by the initialisation, since some areas of secondary structure are predicted accurately, leading to low-variance torsion angle distributions. Overall accuracy (TM-score) improves quickly and after a few hundred steps of gradient descent has converged to a local optimum.

We repeat the optimisation from sampled initialisations, leading to a pool of low potential structures from which further structure initialisations are sampled, with added backbone torsion noise ('noisy restarts'), leading to more structures to be added to the pool. After only a few hundred cycles the optimisation converges and the lowest potential structure is chosen as the best candidate structure. Figure 2e shows the progress in the accuracy of the best-scoring structures over multiple restarts of the gradient descent process, showing that after a few iterations the optimisation has converged. Noisy restarts enable slightly higher TM-score structures to be found than when continuing to sample from the predicted torsion distributions (average of 0.641 vs 0.636 on our test set).

A key component of AlphaFold's overall accuracy is that accurate distance predictions convey more information about structure than contact predictions. Figure 3e shows that the predictions of distance correlate well with the true distance. It can be seen from Figure 3f that the network is also modelling the uncertainty in its predictions. When the standard deviation of the predicted distribution is low, the predictions are more accurate. This is also evident in the predicted distributions of Figure 3d, where more confident predictions of the distance distribution (higher peak and lower standard deviation of the distribution) tend to be more accurate, with the true distance close to the peak. Broader, less-confidently-predicted distributions still assign probability to the correct value even when it is not close to the peak. The high accuracy of the distance predictions and consequently the contact predictions (Table 1c) comes from a combination of factors in the design of the neural network and its training, including predicting distances instead of contacts, data augmentation, feature representation, auxiliary losses, cropping and data curation. (See Methods section.)

Figure 4a shows that the distogram accuracy (measured by distogram lDDT_{12}, defined in Methods) correlates well with the TM-score of the final realised structures. Figure 4b shows the effect of changing the construction of the potential. Removing the distance potential entirely gives a TM-score of 0.266. Reducing the resolution of the distogram representation below 6 bins by averaging adjacent bins causes the TM-score to degrade. Removing the torsion potential, reference correction or $V_{\texttt{score2\_smooth}}$ degrade the accuracy only slightly. A final 'relaxation' (side-chain packing interleaved with gradient descent) with Rosetta[10], using a combination of the `Talaris2014` potential and a spline fit of our reference-corrected distance potential adds side-chain atom coor-

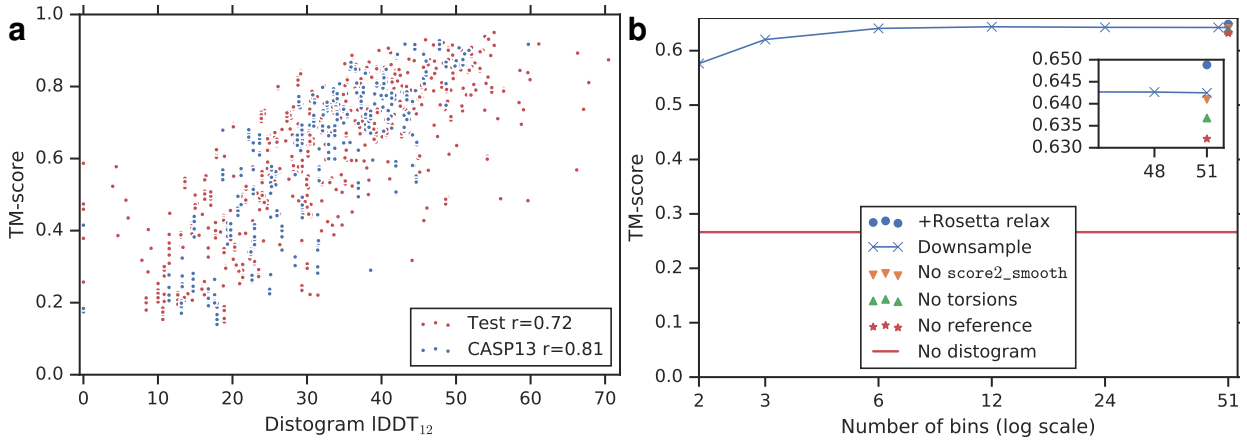


Fig. 4 | TM-scores vs the accuracy of the distogram, and the TM scores' dependency on different components of the potential. (a) TM-score vs distogram lDDT$_{12}$ with Pearson's correlation coefficients, for both CASP13 ($n = 108$) and test ($n = 377$) datasets. (b) Average TM-score over the test set ($n = 377$) vs number of histogram bins used when downsampling the distogram, compared with removing different components of the potential, or adding Rosetta relaxation.

dinates, and yields a small average improvement of 0.007 TM-score.

We have shown that a carefully designed deep-learning system can provide accurate predictions of inter-residue distances and be used to construct a protein-specific potential which represents protein structure. Furthermore we have shown that this potential can be simply optimised with gradient descent to achieve accurate structure predictions. While free modelling predictions only rarely approach the accuracy of experimental structures, the CASP13 assessment shows that the AlphaFold system achieves unprecedented free modelling accuracy and that this free modelling method can match the performance of template modelling approaches without using templates and is starting to reach the accuracy needed for biological understanding (see Methods). We hope that the methods we have described can be developed further and applied to benefit all areas of protein science with more accurate predictions for sequences of unknown structure.

References

1. Dill, K., Ozkan, S. B., Shell, M. & Weikl, T. The protein folding problem. *Annu. Rev. Biophys.* **37,** 289–316 (2008).
2. Dill, K. & MacCallum, J. The protein-folding problem, 50 years on. *Science* **338,** 1042–1046 (2012).
3. Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A. & Bonvin, A. M. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins* **86,** 51–66 (2018).

4. Kirkwood, J. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **3,** 300–313 (1935).

5. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins: Structure, Function, and Bioinformatics* **86,** 7–15 (2018).

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

7. Zhang, Y. Protein structure prediction: when is it useful? *Current opinion in structural biology* **19,** 145–155 (2009).

8. Xu, J. *Protein structure modeling by predicted distance instead of contacts* in *CASP13 Abstracts* Dec. 1, 2018 (2018), 146–7.

9. Zhang, C., Li, Y., Yu, D. & Zhang, Y. *Contact map prediction by deep residual fully convolutional neural network with only evolutionary coupling features derived from deep multiple sequence alignment* in *CASP13 Abstracts* Dec. 1, 2018 (2018), 181–2.

10. Das, R. & Baker, D. Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.* **77,** 363–382 (2008).

11. Jones, D. T. Predicting novel protein folds by using FRAGFOLD. *Proteins* **45,** 127–132 (2001).

12. Zhang, C., Mortuza, S., He, B., Wang, Y. & Zhang, Y. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins* **86,** 136–151 (2018).

13. Kirkpatrick, S., Gelatt, C. & Vecchi, M. Optimization by simulated annealing. *Science* **220,** 671–680 (1983).

14. Gilliland, G. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

15. Altschuh, D., Lesk, A., Bloomer, A. & Klug, A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* **193,** 693–707 (1987).

16. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife* **3,** e02030 (2014).

17. Seemayer, S., Gruber, M. & Söding, J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* **30,** 3128–3130 (2014).

18. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS* **108,** E1293–E1301. ISSN: 0027-8424 (2011).

19. Jones, D. T., Buchan, D. W., Cozzetto, D. & Pontil, M. PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28,** 184–190 (2011).

20. Skwark, M., Raimondi, D., Michel, M. & Elofsson, A. Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. *PLOS Computational Biology* **10,** 1–14 (2014).

21. Jones, D., Singh, T., Kosciolek, T. & Tetchner, S. MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **31,** 999–1006 (2015).

22. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput. Biol.* **13,** 999–1006 (2017).

23. Jones, D. T. & Kandathil, S. M. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics* **1,** 8 (2018).

24. Ovchinnikov, S. *et al.* Improved de novo structure prediction in CASP 11 by incorporating coevolution information into Rosetta. *Proteins* **84,** 67–75 (2016).

25. Aszódi, A. & Taylor, W. R. Estimating polypeptide α-carbon distances from multiple sequence alignments. *J. Math. Chem.* **17,** 167–184 (1995).

26. Zhao, F. & Xu, J. A position-specific distance-dependent statistical potential for protein structure and functional study. *Structure* **20,** 1118–1126 (2012).

27. Aszodi, A., Gradwell, M. & Taylor, W. Global fold determination from a small number of distance restraints. *J. Mol. Biol.* **251,** 308–326 (1995).

28. Kandathil, S., Greener, J. & Jones, D. *DMPfold: a new deep learning-based method for protein tertiary structure prediction and model refinement* in *CASP13 Abstracts* Dec. 1, 2018 (2018), 84–5.

29. Xu, J. Distance-based Protein Folding Powered by Deep Learning. *arXiv preprint arXiv:1811.03481* (2018).

30. Konagurthu, A. S., Lesk, A. M. & Allison, L. Minimum message length inference of secondary structure from protein coordinate data. *Bioinformatics* **28,** i97–i105 (2012).

31. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* **abs/1512.03385** (2015).

32. Simons, K., Kooperberg, C., Huang, E. & Baker, D. Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions. *J. Mol. Biol.* **268,** 209–225 (1997).

33. Liu, D. C. & Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **45,** 503–528 (1989).

1 Methods

Figure 5 shows the steps involved in MSA construction, feature extraction, distance prediction, potential construction and structure realisation.

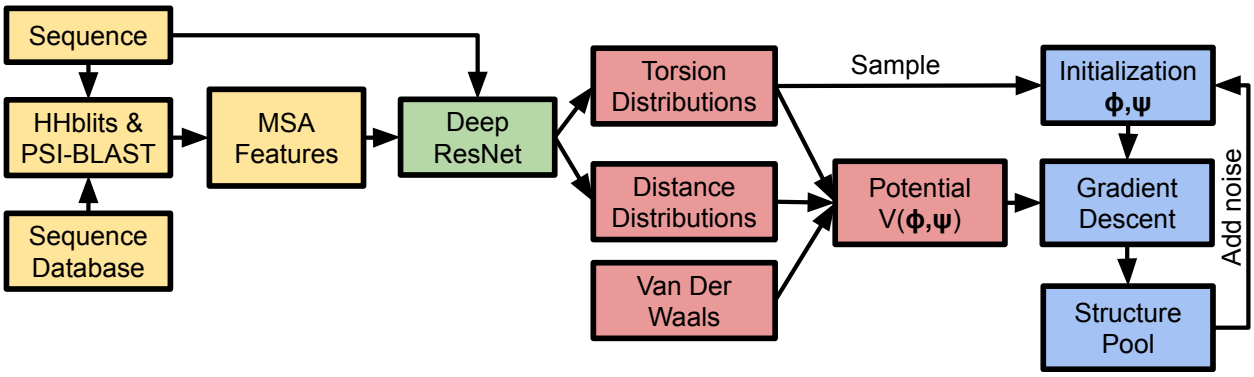


Fig. 5 | A schematic of the folding system. Feature extraction stages are shown in yellow, structure-prediction neural network in green, potential construction in red and structure realisation in blue.

Data Our models are trained on structures extracted from the Protein Data Bank[1]. We extract non-redundant domains by utilising the CATH[2] 35% sequence similarity cluster representatives. This gives 31 247 domains, which are split into train, and test sets (29 427 and 1 820 proteins respectively) keeping all domains from the same homologous superfamily (H-level in the CATH classification) in the same partition. The CATH superfamilies of FM domains from CASP11 and CASP12 were also excluded from the training set. From the test set, we take a single domain per homologous superfamily to create the 377 domain subset used for the results presented here. We note that accuracies for this set are higher than for the CASP13 test domains.

CASP13 submission results are drawn from the CASP13 results pages with additional results shown for the CASP13 dataset for "all groups" chains, scored on CASP13 PDB files, by CASP domain definitions. Contact prediction accuracies are recomputed from the group 032 and 498 submissions (as RR files), compared with the distogram predictions used by AlphaFold for CASP13 submissions. Contact prediction probabilities are obtained from the distograms by summing the probability mass in each distribution below 8 Å.

For each training sequence, we search for and align similar protein sequences in the Uniclust30[3] dataset with HHblits[4] and use the returned MSA to generate *profile* features with the position-specific substitution probabilities for each residue as well as covariation features — the parameters of a regularised pseudolikelihood-trained Potts model similar to CCMPred[5]. CCMPred uses the Frobenius norm of the parameters, but we feed both this norm (1 feature) as well as the raw parameters (484 features) into the network for each residue pair ij. In addition we provide the network with features explicitly representing gaps and deletions in the MSA. To make the network better able to make predictions for shallow MSAs, and as a form of data augmentation, we take a sample of half the sequences from the the HHblits MSA before computing the MSA-based features. Our training set contains 10 such samples for each domain. We extract additional profile features using PSI-BLAST[6].

The distance prediction neural network was trained with the following input features (with number of features).

- Number of HHblits alignments (1D scalar)
- Sequence-length features: 1-hot amino acid type (21D), Profiles: PSI-BLAST (21D), HHblits profile (22D), non-gapped profile (21D), HHblits bias, HMM profile (30D) Potts model bias (22D); Deletion probability (1D); residue index (integer index of residue number, consecutive except for multi-segment domains, encoded as 5 least-significant bits and a scalar);
- Sequence-length-squared features: Potts model parameters (484D, fitted with 500 iterations of gradient descent using Nesterov momentum 0.99, without sequence reweighting); Frobenius norm (1D); Gap matrix (1D)

Distogram prediction The inter-residue distances are predicted by a deep neural network. The architecture is a deep two-dimensional dilated convolutional residual network. Xu *et al.*[7] used a two-dimensional residual network preceded by one-dimensional embedding layers for contact prediction. Our network is two-dimensional throughout and uses 220 residual blocks[8] with dilated convolutions[9]. Each residual block, illustrated in Figure 6 consists of a sequence of neural network layers[10], interleaving three batchnorm layers; two 1×1 projection layers; a 3×3 dilated convolution layer and ELU[11] nonlinearities. Successive layers cycle through dilations of $1, 2, 4, 8$ pixels to allow propagation of information quickly across the cropped region. At the final layer, a position-specific bias was used, so the biases were indexed by residue-offset (capped at 32) and bin number.

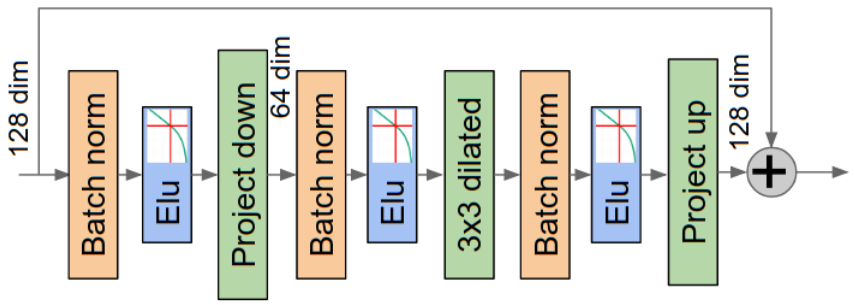


Fig. 6 | The layers used in one block of the deep residual convolutional network. The dilated convolution is applied on reduced-dimensional features. The output of the block is added to the representation from the previous layer. The residual network's bypass connections allow gradients to pass back through the network undiminished, permitting the training of very deep networks.

The network is trained with stochastic gradient descent using a cross-entropy loss. The target is a quantisation of the distance between the residues' C_β atoms (C_α for glycine). We quantise the range 2–22 Å into 64 equal bins. The input to the network consists of a two-dimensional array of features where each i, j feature is the concatenation of the 1-dimensional features for both i and j

as well as the two-dimensional features for i, j.

Individual training runs were cross-validated with early stopping using 27 CASP11 FM domains as a validation set. Models were selected by cross-validation on 27 CASP12 FM domains.

Neural network hyperparameters

- 7×4 Blocks with 256 channels, cycling through dilations 1, 2, 4, 8
- 48×4 Blocks with 128 channels, cycling through dilations 1, 2, 4, 8
- Optimisation: Synchronized stochastic gradient descent
- Batch size: batch of 4 crops on each of 8 GPU workers
- 0.85 Dropout keep probability
- Nonlinearity: ELU
- Learning rate 0.06
- Auxilliary loss weights: Secondary structure: 0.005; Accessible surface area: 0.001. These auxilliary losses were cut by a factor 10 after 100 000 steps.
- Learning rate decayed by 50% at 150 000, 200 000, 250 000 and 350 000 steps.
- Training time: about 5 days for 600 000 steps

To constrain memory usage and avoid overfitting, the network is always trained on 64×64 regions of the distance matrix, that is the pairwise distances between 64 consecutive residues and another group of 64 consecutive residues. For each training domain, the entire distance matrix is split into non-overlapping 64×64 crops. By training off-diagonal crops, the interaction between residues further apart than 64 residues can be modelled. Each crop consists of the distance matrix which represents the juxtaposition of two 64-residue fragments. Jones and Kandathil[12] have shown that contact prediction needs only a limited context window. We note that the distance predictions close to the diagonal $i = j$, encode predictions of the local structure of the protein, and for any cropped region the distances are governed by the local structure of the two fragments represented by the i and j ranges of the crop. Augmenting the inputs with the on-diagonal 2D input features that correspond to both the i and j ranges provides additional information to predict the structure of each fragment and thus distances between them. It can be seen that if the fragment structures can be well predicted (for instance if they are confidently predicted as helices or sheets) then prediction of a single contact between the fragments will strongly constrain the distances between all other pairs

Randomising the offset of the crops each time a domain is used in training leads to a form of data augmentation where a single protein can generate many thousands of different training examples. This is further enhanced by adding noise to the atom coordinates, proportional to the ground truth resolution leading to variation in the target distances. Data augmentation (MSA subsampling and coordinate noise), together with dropout, prevents the network from overfitting to the training data.

To predict the distance distribution for all $L \times L$ residue pairs, many 64×64 crops are combined. To avoid edge effects, several such tilings are produced with different offsets and averaged together, with a heavier weighting for the predictions near the centre of the crop. To improve accuracy further, predictions from an ensemble of four separate models, trained independently with slightly different hyperparameters, are averaged together. Figure 7 shows an example of true distances (a) and the mode of the distogram prediction (b) for a three-domain CASP13 target.

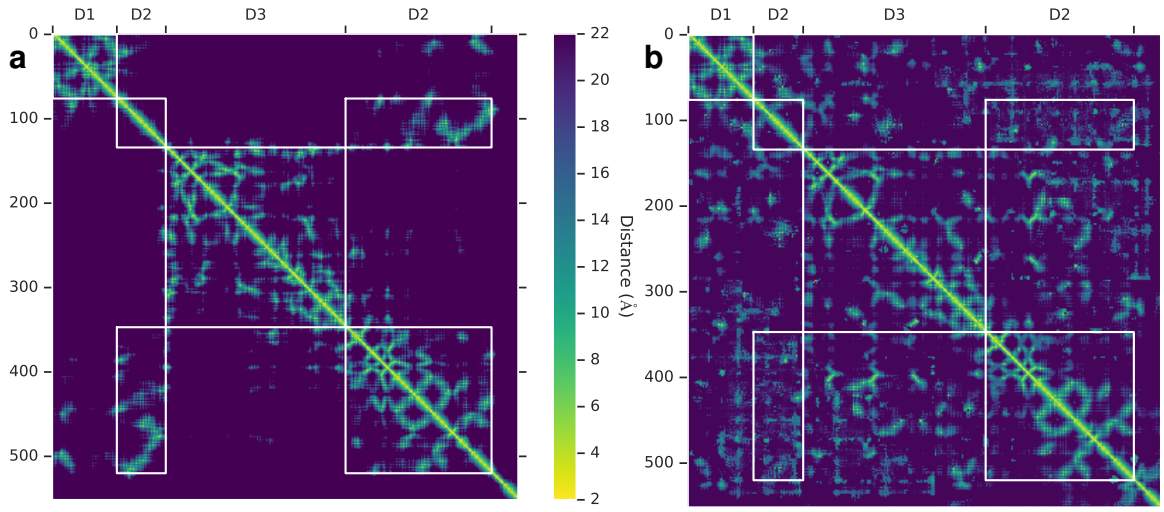
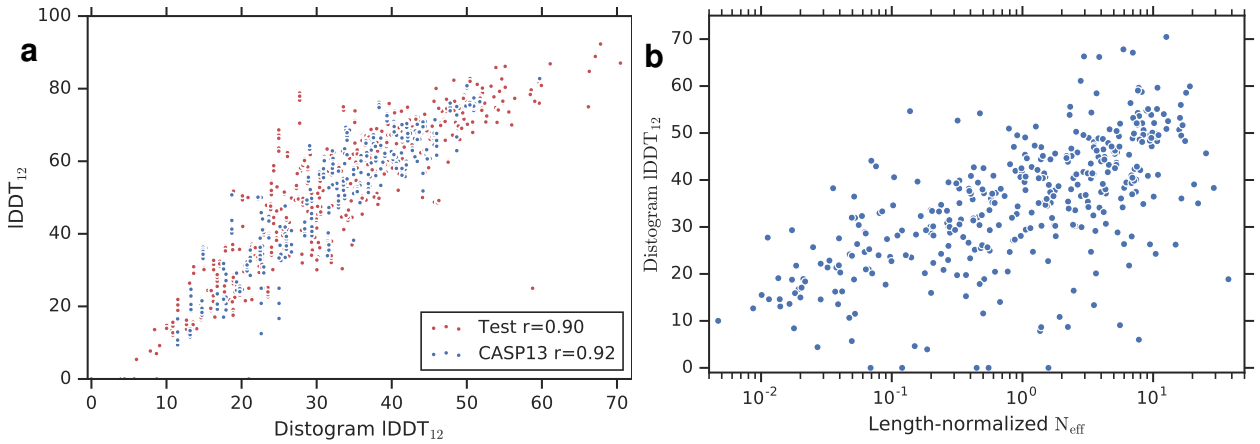


Fig. 7 | True distances (a) and modes of the predicted distogram (b) for CASP13 target T0990. CASP divides this chain into 3 domains as shown (D3 is inserted in D2) for which there are 39, 36 and 42 HHblits alignments respectively (from the CASP website).

Since the network has a rich representation capable of incorporating both profile and covariation features of the MSA, we argue that the network can be used to predict secondary structure directly. By mean- and max-pooling the 2D activations of the penultimate layer of the network separately in both i and j, we add an additional 1-dimensional output head to the network which predicts 8-class secondary structure labels as computed by DSSP[13] for each residue in j and i. The resulting Q3 (distinguishing the three helix / sheet / coil classes) predictions' accuracy is 84% which is comparable to the state-of-the-art[14]. The relative accessible surface area (ASA) of each residue can also be predicted.

The 1-dimensional pooled activations are also used to predict the marginal Ramachandran distributions: $P(\phi_i, \psi_i \mid \mathcal{S}, \mathrm{MSA}(\mathcal{S}))$, independently for each residue, as a discrete probability distribution quantised to $10°$ (1296 bins). In practice during CASP13 we used distograms from a network that was trained to predict distograms, secondary structure and ASA with torsions from a second, similar network trained to predict distograms, secondary structure, ASA and torsions, since the former had been more thoroughly validated.

Figure 8b shows that an important factor in the accuracy of the distograms (as has previously been found with contact prediction systems) is $\mathrm{N_{eff}}$, the effective number of sequences in the



c

Potential	Bins	TM-score	GDT_TS	lDDT	RMSD (Å)
Full + relax	51/64	0.649	65.8	54.2	5.94
Full	51/64	0.642	65.0	53.9	5.91
W/o reference	51/64	0.632	64.3	50.0	6.64
W/o `score2_smooth`	51/64	0.641	64.8	53.7	5.93
W/o torsions	51/64	0.637	64.3	53.6	6.04
W/o distogram	51/64	0.266	29.1	19.1	14.88
Full	48/64	0.643	65.0	54.1	5.90
Full	24/32	0.643	65.0	53.8	5.89
Full	12/16	0.644	65.1	53.9	5.85
Full	6/8	0.641	64.6	53.7	5.94
Full	3/4	0.620	62.4	52.8	6.22
Full	2/3	0.576	58.2	49.3	8.38

Fig. 8 | Analysis of structure accuracies. (a) lDDT vs distogram lDDT_{12} (Defined below under 'Accuracy'). The distogram accuracy predicts the realised structure's lDDT (as well as TM-score as shown in Fig. 4a) well for both CASP13 ($n = 108$) and test ($n = 377$) datasets. Shown with Pearson's correlation coefficients. (b) DLDDT_{12} against effective number of sequences in the MSA (N_{eff}) normalised by sequence length ($n = 377$). The number of effective sequences correlates with this measure of distogram accuracy ($r = 0.634$). (c) Structure accuracy measures, computed on the test set, for gradient descent optimisation of different forms of the potential. Above: Removing terms in the potential, also showing the effect of following optimisation with Rosetta relax. Bins shows the number of bins fitted by the spline before extrapolation and the number in the full distribution. In CASP13 splines were fitted to the first 51 of 64 bins. Below, reducing the resolution of the distogram distributions. The original 64-bin distogram predictions are repeatedly downsampled by a factor 2 by summing adjacent bins, in each case with constant extrapolation beyond 18 Å (the last $\frac{1}{4}$ of the bins). The final row's two-level potential, designed to compare to contact predictions, is constructed by summing the probability mass below 8 Å and between 8–14 Å, with constant extrapolation beyond 14 Å. The TM-scores in this table are plotted in Figure 4 (b) Accuracy measures, computed on the test set ($n = 377$), for gradient descent optimisation with differently constructed potentials. The TM-scores in this table are plotted in Figure 4b.

MSA[15]. This is the number of sequences found in the MSA, discounting redundancy at the 62% sequence identity level, which we then divide by the number of residues in the target, and is an indication of the amount of covariation information in the MSA.

Distance potential The distogram probabilities are estimated for discrete distance bins, so to construct a differentiable potential the distribution is interpolated with a cubic spline. Because the final bin accumulates probability mass from all distances beyond 22 Å, and since greater distances are harder to predict accurately, the potential is only fit up to 18 Å (determined by cross-validation), with a constant extrapolation thereafter.

To predict a reference distribution, a similar model is trained on the same dataset. The reference distribution is not conditioned on the sequence, but to account for the atoms between which we are predicting distances, we do provide a feature $\delta_{\alpha\beta}$ to indicate if the residue is glycine (C_α atom) or not (C_β) and the overall length of the protein.

A distance potential is created from the negative log likelihood of the distances, summed over all pairs of residues i, j.

$$V_{\text{distance}}(\mathbf{x}) \;=\; -\sum_{i,j,\ i\neq j} \log P(d_{ij} \mid \mathcal{S}, \text{MSA}(\mathcal{S})) \tag{1}$$

With a reference state this becomes the log likelihood ratio of the distances under the full conditional model and under the background model:

$$V_{\text{distance}}(\mathbf{x}) \;=\; -\sum_{i,j,\ i\neq j} \log P(d_{ij} \mid \mathcal{S}, \text{MSA}(\mathcal{S})) - \log P(d_{ij} \mid \text{length}, \delta_{\alpha\beta}) \tag{2}$$

Torsions are modelled as a negative log likelihood under the predicted torsion distributions. Since we have marginal distribution predictions, each of which can be multimodal, it can be difficult to jointly optimise the torsions. To unify all the probability mass, at the cost of modelling fidelity of multimodal distributions, we fit a unimodal von Mises distribution to the marginal predictions. The potential is summed over all residues i.

$$V_{\text{torsion}}(\boldsymbol{\phi}, \boldsymbol{\psi}) \;=\; -\sum_{i} \log p_{\text{vonMises}}(\phi_i, \psi_i \mid \mathcal{S}, \text{MSA}(\mathcal{S})) \tag{3}$$

Finally, to prevent steric clashes, a van der Waals term is introduced through the use of Rosetta's $V_{\texttt{score2\_smooth}}$.

Structure realisation by gradient descent To realise structures which minimise the constructed potential, we create a differentiable model of ideal protein backbone geometry, giving backbone

atom coordinates as a function of the torsion angles $(\boldsymbol{\phi}, \boldsymbol{\psi})$: $\mathbf{x} = G(\boldsymbol{\phi}, \boldsymbol{\psi})$. The complete potential to be minimised is then*:

$$V_{\text{total}}(\boldsymbol{\phi}, \boldsymbol{\psi}) \quad = \quad V_{\text{distance}}(G(\boldsymbol{\phi}, \boldsymbol{\psi})) + V_{\text{torsion}}(\boldsymbol{\phi}, \boldsymbol{\psi}) + V_{\texttt{score2\_smooth}}(G(\boldsymbol{\phi}, \boldsymbol{\psi})). \tag{4}$$

Since every term in V_{total} is differentiable with respect to the torsion angles, given an initial set of torsions $\boldsymbol{\phi}, \boldsymbol{\psi}$ which can be sampled from the predicted torsion marginals, we can minimise V_{total} using a gradient descent algorithm, such as L-BFGS[16]. The optimised structure is dependent on the initial conditions, so we repeat the optimisation multiple times with different initialisations. A pool of the 20 lowest-potential structures is maintained and once full, we initialise 90% of trajectories from those with 30° noise added to the backbone torsions (the remaining 10% being sampled from the predicted torsion distributions). In CASP13 we made 5000 optimisation runs for each chain. Figure 2 shows the change in TM-score against the number of restarts. Since longer chains take longer to optimise, this work load was balanced across $(50+L)/2$ parallel workers. Figure 9 shows

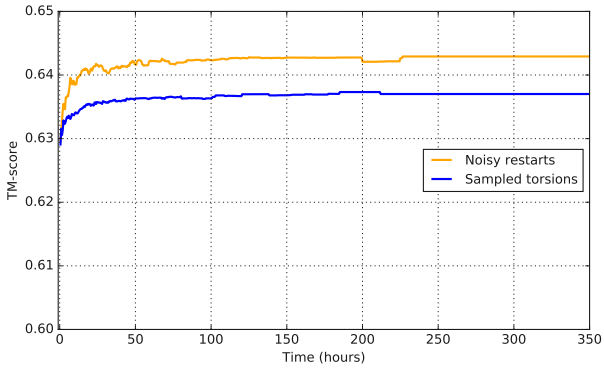


Fig. 9 | TM-score vs per-target computation time computed as an average over the test set ($n = 377$). Full optimisation with noisy restarts (orange) is compared with initialisation from sampled torsions (blue). Computation is measured as the product of the number of (CPU-based) machines and time elapsed and can be largely parallelised. Longer targets take longer to optimise.

that this is achieved with a moderate computation budget, which can be parallelised over multiple machines.

Accuracy We compare the final structures to the experimentally determined structures to measure their accuracy using metrics such as TM-score, GDT_TS[17] and RMSD. All of these accuracy measures require geometric alignment between the candidate structure and the experimental structure. An alternative accuracy measure which requires no alignment is the Local Distance Difference Test (lDDT[18]) which measures the percentage of native pairwise distances D_{ij} under 15 Å, with sequence offsets $\geq r$ residues, that are realised in a candidate structure (as d_{ij}) within a tolerance

*While there is no guarantee that these potentials have equivalent scale, scaling parameters on the terms were introduced and chosen by cross-validation on CASP12 FM domains. In practice equal weighting for all terms was found to lead to the best results.

of the true value, averaging across tolerances of 0.5, 1, 2 and 4 Å (without stereochemical checks).

$$lDDT_r = \frac{100}{4L} \sum_{t \in \{0.5,1,2,4\}} \sum_{i=1}^{L} \frac{\sum_{j,|i-j|\geq r, D_{ij}<15} \mathbb{1}(|D_{ij} - d_{ij}| < t)}{\sum_{j,|i-j|\geq r, D_{ij}<15} 1}. \quad (5)$$

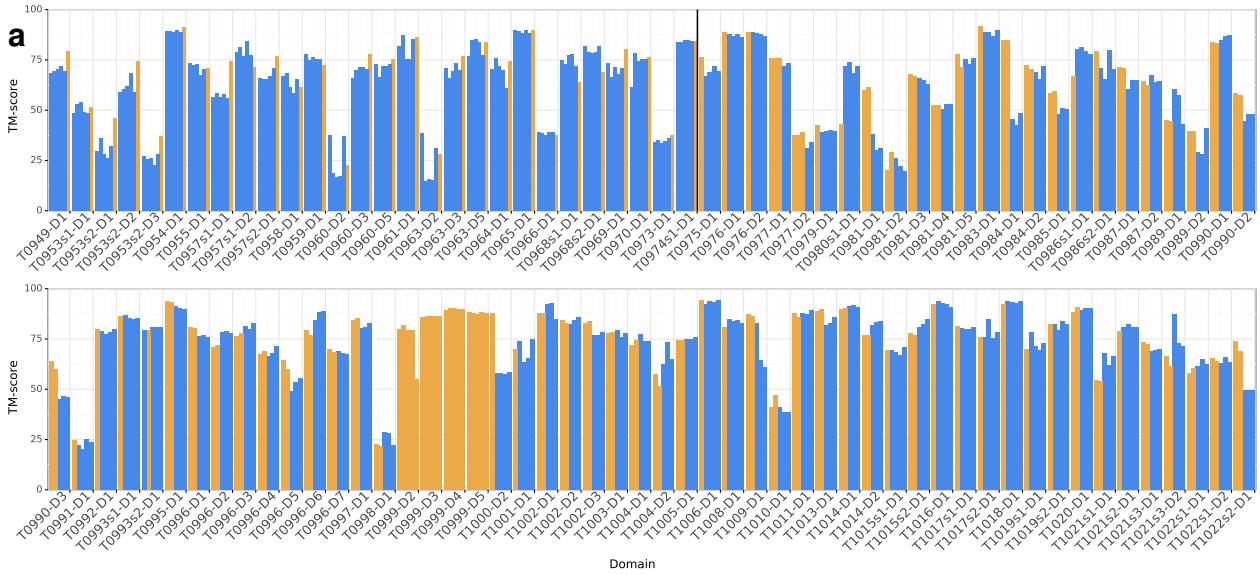
Since the distogram predicts pairwise distances, we can introduce *distogram lDDT* (DLDDT), a measure like lDDT computed directly from the distograms' probabilities.

$$DLDDT_r = \frac{100}{4L} \sum_{t \in \{0.5,1,2,4\}} \sum_{i=1}^{L} \frac{\sum_{j,|i-j|\geq r, D_{ij}<15} P(|D_{ij} - d_{ij}| < t \mid \mathcal{S}, \mathrm{MSA}(\mathcal{S}))}{\sum_{j,|i-j|\geq r, D_{ij}<15} 1} \quad (6)$$

Since distances between residues nearby in the sequence are often short, easier to predict and are not critical in determining the overall fold topology, we set $r = 12$, considering only those distances for residues with a sequence separation ≥ 12. Since we predict C_β distances, for this work we compute both lDDT and DLDDT using the C_β distances.

Full chains without domain segmentation Parameterising proteins of length L by two torsion angles per residue, the dimension of space of structures grows as $2L$, so searching for structures of large proteins becomes much harder. Traditionally this problem is addressed by splitting longer protein chains into pieces, termed domains, which fold independently. However, the problem of domain segmentation from the sequence alone is itself difficult and error-prone. For this work, we avoided domain segmentation and folded entire chains. Typically multiple sequence alignments are based upon a given domain segmentation, but we used a *sliding windows* approach, computing a full-chain multiple sequence alignment to predict a baseline full-sequence distogram. We then compute MSAs for subsequences of the chain, trying windows of size 64, 128, 256 with offsets at multiples of 64. Each of these MSAs gives rise to an individual distogram corresponding to an on-diagonal square of the full-chain distogram. We average all these distograms together, weighted by the number of sequences in the MSA to produce an average full-chain distogram which is more accurate in regions where many alignments can be found.

CASP13 results In CASP13 the 5 AlphaFold submissions were from 3 different systems, all using potentials based on the neural network distance predictions. Before T0975, two systems based on simulated annealing and fragment assembly (and using 40 bin distance distributions) were used. From T0975 on, newly trained 64-bin distogram predictions were used and structures were generated by the gradient descent system described here (3 independent runs) as well as one of the fragment assembly systems (5 independent runs). 5 submissions were chosen from these 8 structures (the lowest potential structure generated by each independent run) with the first submission ('top-1') being the lowest-potential structure generated by gradient descent. The remaining four submissions were the four best other structures, with the fifth being a gradient descent structure if none had been chosen for position 2, 3 or 4. All submissions for T0999 were generated by gradient descent. Figure 10a shows the methods used for each submission, comparing with 'back-fill' structures generated by a single run of gradient descent for targets before T0975. Table 10b shows that



b

Method	FM	TBM/FM	TBM	All
Top-1	58.0	68.1	76.2	69.9
Best-of-5	62.6	73.6	78.6	73.2
1× gradient descent	58.4	71.6	76.3	70.4
1× fragment assembly	54.3	69.9	74.5	68.0

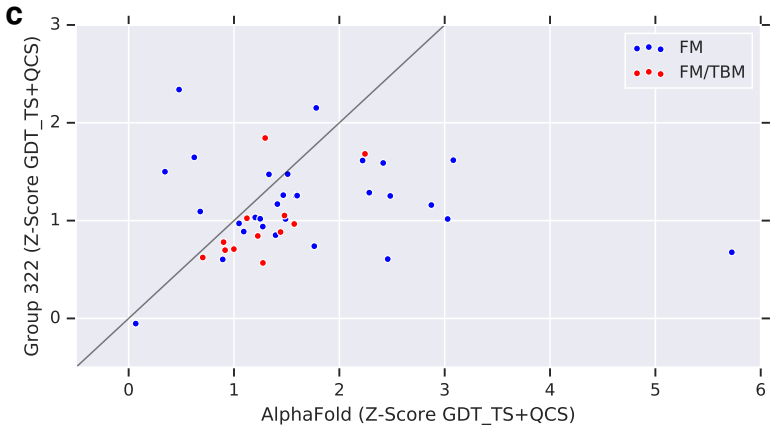


Fig. 10 | AlphaFold CASP13 results. (a) The TM-score for each of the 5 AlphaFold CASP13 submissions are shown. Simulated annealing with fragment assembly entries are shown in blue. Gradient-descent entries are shown in yellow. Gradient descent was only deployed for targets T0975 and later, so to the left of the black line we also show the results for a single, 'back-fill', run of gradient descent for each earlier target using the deployed system. T0999 (1589 residues) was manually segmented based on HHpred[19] homology matching. (b) Average TM-scores of the AlphaFold CASP13 submissions ($n = 104$ domains), comparing the first model submitted, the best-of-5 model (submission with highest GDT), a single run of full-chain gradient descent (a CASP13 run for T0975 and later, back-fill for earlier targets) and a single CASP13 run of fragment assembly with domain segmentation (using a gradient descent submission for T0999). (c) Assessors' formula standardised (z) scores of GDT_TS + QCS[20], best-of-5 for CASP FM ($n = 31$) and FM/TBM ($n = 12$) domains comparing AlphaFold with the closest competitor (group 322), coloured by domain category.

the gradient descent method deployed later in CASP performed better than the fragment assembly method, in each category. Figure 10c compares the accuracy of the AlphaFold submissions for FM and FM/TBM domains with the next best group 322. For the CASP13 assessment full chains were relaxed with Rosetta relax with a potential of $V_{\texttt{Talaris2014}} + 0.2V_{\text{distance}}$ (weighting determined by cross-validation) and submissions from all the systems were ranked based on this potential.

Biological relevance There is a wide range of uses of predicted structures, all with different accuracy requirements, from generally understanding the fold shape to understanding detailed side-chain configurations in binding regions. Contact predictions alone can guide biological insight[21], for instance targeting mutations to destabilise the protein. The accuracy of the contact predictions shown in Table 1c indicates that the AlphaFold contact predictions exceed the state of the art. Here we present further results which indicate that AlphaFold's accuracy improvements lead to more accurate interpretation of function; better interface prediction for protein-protein interaction; better binding pocket prediction and improved molecular replacement in crystallography.

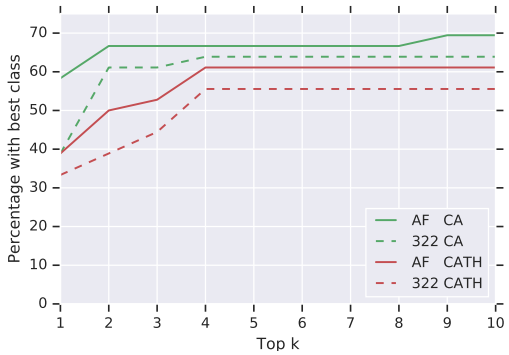


Fig. 11 | Correct fold identification by structural search in CATH. For each of the FM or TBM/FM domains, the top-1 submission and ground-truth are compared to all 30 744 CATH S40 non-redundant domains with TM-align[22]. For the 36 domains where there is a good ground-truth match (score > 0.5), we show the percentage of decoys where a domain with the same CATH code (in red, CA in green. CAT results are close to CATH results) as the top ground-truth match is in the at-most top-k matches with score > 0.5. Curves are shown for AlphaFold and the next-best group (322). AlphaFold predictions determine the matching fold more accurately. Determination of the matching CATH domain can give insight into the function of a new protein.

Often protein function can be inferred by finding homologous proteins of known function. Figure 11 shows that AlphaFold's FM predictions give greater accuracy in structure-based search for homologous domains in the CATH database.

Protein-protein interaction is an important domain for understanding protein function that has hitherto largely been limited to template-based models because of the need for high accuracy predictions, though there has been moderate success[23] in docking with predicted structures up to 6 Å RMSD. Figure 12 shows that AlphaFold's predictions improve accuracy in the interface

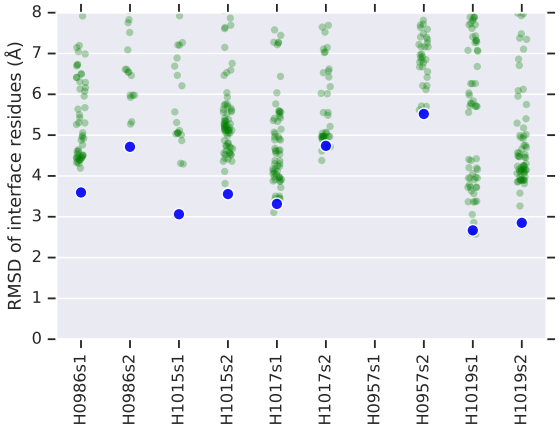


Fig. 12 | Accuracy of predictions for interfaces. For the five all-groups heterodimer CASP13 targets the full-atom RMSDs of the interface residues (residues with a ground truth inter-chain heavy atom distance < 10 Å) are computed for all groups' chain submissions, relative to the target complex. Results > 8 Å are not shown. AlphaFold achieves consistently high accuracy interface regions and for 4 out of 5 targets predicts both chains' interfaces below < 5 Å.

regions of chains in hetero-dimer structures and are likely better candidates for docking, though docking did not form part of the AlphaFold system and all submissions were for isolated chains rather than complexes.

Further evidence of AlphaFold reaching accuracy sufficient for biological relevance is shown in Figure 13. The images show the pocket in T1011 indicating that the accuracy gain in AlphaFold's structure prediction can leads to more accurate prediction of pocket geometry and thus the binding of ligands.

So far only template-based predictions have been able to deliver the most accurate predictions. While AlphaFold is able to match template-based modelling without using templates, and in some cases outperform other methods (e.g. T0981-D5, 72.8 GDT_TS, and T0957s1-D2, 88.0 GDT_TS, two TBM-hard domains where AlphaFold's top-1 model is 12 GDT_TS better than any other top-1 submission) accuracy for FM targets still lags that for TBM targets and can still not be relied upon for detailed understanding of hard structures. In an analysis of the performance of CASP13 TBM predictions for Molecular Replacement, Read et al.[25] reported that the AlphaFold predictions (raw coordinates, without B-factors) led to a marginally greater log-likelihood gain (LLG) than those of any other group, indicating that these improved structures can assist in phasing for X-ray crystallography.

Interpretation of distogram neural network We have shown that the deep distance prediction neural network achieves high accuracy, but we would like to understand how the network arrives at its distance predictions, and in particular to understand how the inputs to the model affect the final

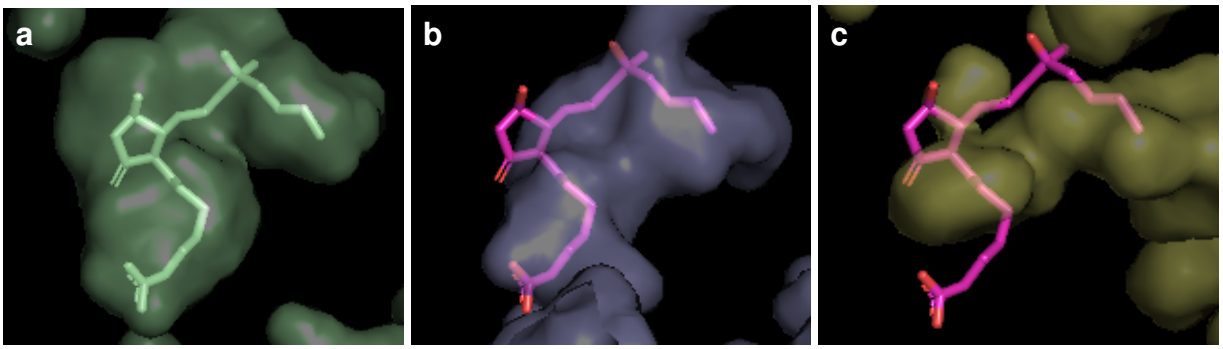


Fig. 13 | Ligand pocket visualizations for T1011 PDB 6M9T: EP3 receptor bound to misoprostol-FA[24] (a) the native structure showing the ligand in a pocket. (b) AlphaFold's submission 5 (78.0 GDT_TS) made without knowledge of the ligand shows a pocket more similar to the true pocket than that of (c) the best other submission (322 model 3, 68.7 GDT_TS). Both submissions are aligned to the native using the same subset of residues from the helices close to the ligand pocket and visualized with the interior pocket together with the native ligand position.

prediction. This might lead to understanding of the folding mechanisms or suggest improvements to the model. However, deep neural networks are complex non-linear functions of their inputs, and so this attribution problem is difficult, under-specified and an on-going topic of research. Even so, there are a number of methods for such analysis: here we apply Integrated Gradients [26] to our trained distogram network to indicate the location of input features which affect the network's predictions of a particular distance.

Given the expected value of the distance between any two residues I and J, $d^{I,J}(x)$, we can consider its derivatives with respect to the input features $x_{i,j,c}$, where i and j are residue indices and c is the feature channel index. The attribution function, as calculated using Integrated Gradients, of the expected distance between residues I and J with respect to the input features is then defined as

$$S^{I,J}_{i,j,c} = (x_{i,j,c} - x'_c) \int_{\alpha=0}^{1} d\alpha \frac{\partial d^{I,J}(\alpha x + (1-\alpha)x')}{\partial x_{i,j,c}}, \tag{7}$$

$$\text{s.t.} \sum_{i,j,c} S^{I,J}_{i,j,c} = d^{I,J}(x) - d^{I,J}(x'), \tag{8}$$

where x' is a reference set of features; in this case we average the input features spatially:

$$x'_c = \frac{1}{N^2} \sum_{i=0,j=0}^{N,N} x_{i,j,c}. \tag{9}$$

The derivatives of d can be calculated using backpropagation on the trained distogram network, and the integral over α is approximated as a numerical summation.

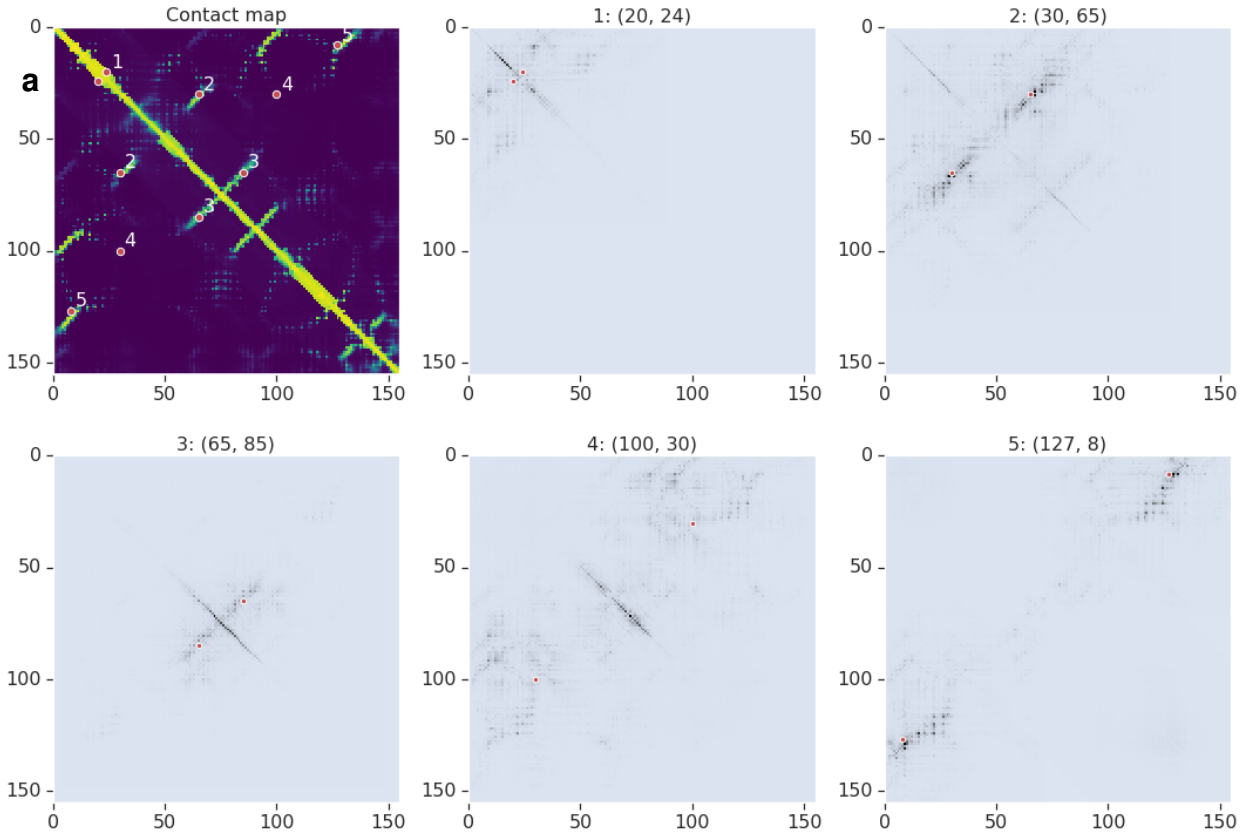


Fig. 14 | Attribution map of distogram network The contact probability map of T0986s2, and the summed absolute value of the Integrated Gradient, $\sum_c |S^{I,J}_{i,j,c}|$, of the input 2D features with respect to the expected distance between five different pairs of residues (I, J): (1) a helix self-contact, (2) a long-range stand-strand contact, (3) a medium-range strand-strand contact, (4) a non-contact and (5) a very long-range strand-strand contact. Each pair is shown as two red dots on the diagrams. Darker means higher attribution weight.

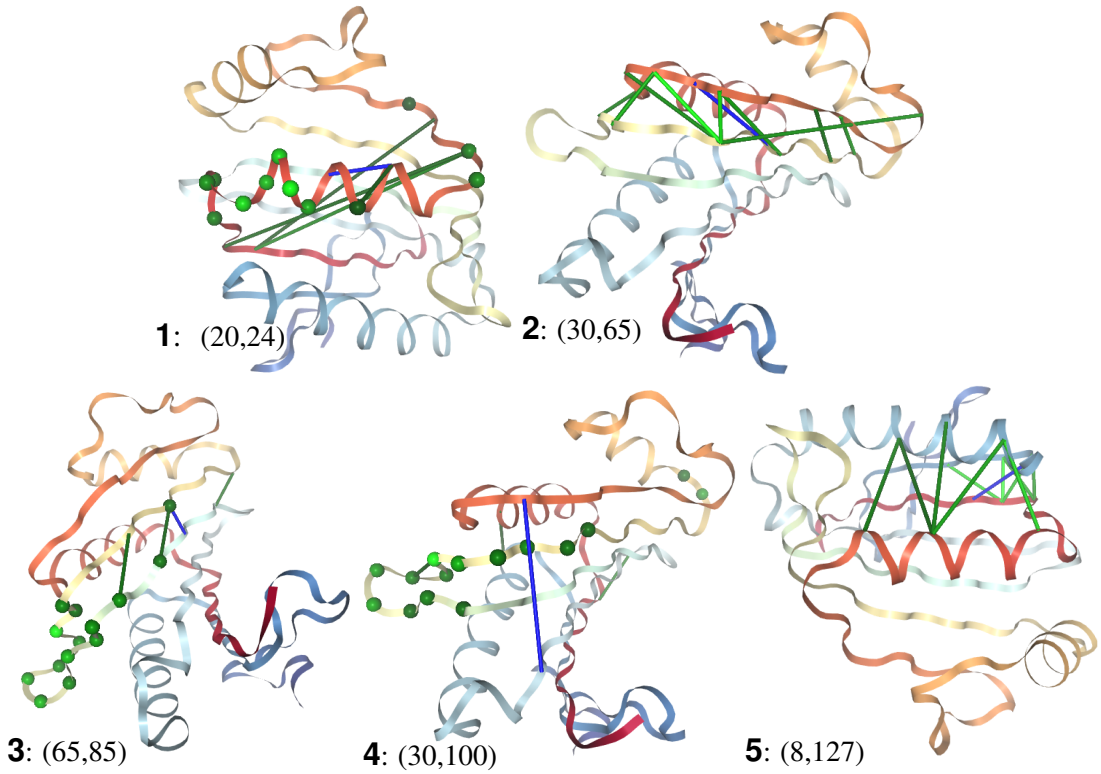


Fig. 15 | Attribution shown on predicted structure. For T0986s2 (TM-score 0.8), the top 10 input pairs, including self-pairs, with highest attribution weight for each of the five output pairs shown in Figure 14, are shown as lines (or spheres, for self-pairs) coloured by sensitivity, lighter green is more sensitive, and the output pair shown as a blue line.

In Figure 14, plots of summed absolute Integrated Gradient, $\sum_c |S^{I,J}_{i,j,c}|$, are shown for selected I, J output pairs in T0986s2 and in Figure 15, the top-10 highest attribution input pairs for each output pair are shown on top of AlphaFold's top-1 predicted structure. The attribution maps are sparse and highly structured, closely reflecting the predicted geometry of the protein. For the four in-contact pairs presented (1, 2, 3, 5), all the highest attribution pairs are pairs within or between the secondary structure that one or both the output pair are members of. In (1) the helix residues are important as well as connections between the strands which follow either end of the helix, which might indicate strain on the helix. In (2) all the most important residue pairs connect the same two strands, whereas in (3) a mix of inter-strand pairs and strand residues are most salient. In (5) the most important pairs involve the packing of nearby secondary structure elements to the strand and helix. For the non-contacting pair (4), the most important input pairs are the residues that are geometrically between I and J in the predicted protein structure. Furthermore, most of the high attribution input pairs are themselves in contact.

Since the network is tasked with predicting the spatial geometry, with no structure available at the input, these patterns of interaction indicate that the network is using intermediate predictions to discover important interactions and channelling information from related residues to refine the final prediction.

References

1. Gilliland, G. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
2. Dawt1son, N. *et al.* CATH: An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* (2017).
3. Mirdita, M. *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45,** D170–D176 (2016).
4. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9,** 173 (2012).
5. Seemayer, S., Gruber, M. & Söding, J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* **30,** 3128–3130 (2014).
6. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
7. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput. Biol.* **13,** 999–1006 (2017).
8. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* **abs/1512.03385** (2015).
9. Yu, F. & Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* (2015).

10. Oord, A. v. d. *et al.* Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

11. Clevert, D.-A., Unterthiner, T. & Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *arXiv preprint arXiv:1511.07289* (2015).

12. Jones, D. T. & Kandathil, S. M. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics* **1,** 8 (2018).

13. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637 (1983).

14. Yang, Y. *et al.* Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Briefings Bioinf.* **19,** 482–494. ISSN: 1467-5463 (2016).

15. Jones, D., Singh, T., Kosciolek, T. & Tetchner, S. MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **31,** 999–1006 (2015).

16. Liu, D. C. & Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **45,** 503–528 (1989).

17. Zemla, A., Venclovas, Č., Moult, J. & Fidelis, K. Processing and analysis of CASP3 protein structure predictions. *Proteins* **37,** 22–29 (1999).

18. 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).

19. Söding, J., Biegert, A. & Lupas, A. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research* **33,** W244–W248 (2005).

20. Cong, Q. *et al.* An automatic method for CASP9 free modeling structure prediction assessment. *Bioinformatics* **27,** 3371–3378 (2011).

21. Kayikci, M. *et al.* Protein contacts atlas: visualization and analysis of non-covalent contacts in biomolecules. *Nat. Struct. Mol. Biol.* **25,** 185–194 (2018).

22. Zhang, Y. & Skolnick, J. TM-align: A protein structure alignment algorithm based on TM-score. *Nucleic Acids Research* **33,** 2302–2309 (2005).

23. Tovchigrechko, A., Wells, C. & Vakser, I. Docking of protein models. *Protein Sci.* **11,** 1888–1896 (2002).

24. Audet, M. *et al.* Crystal structure of misoprostol bound to the labor inducer prostaglandin E 2 receptor. *Nature chemical biology* **15,** 11 (2019).

25. Read, R. Molecular replacement in CASP13. *In preparation* (2019).

26. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. *CoRR* **abs/1703.01365** (2017).

2 Acknowledgements

We thank Clemens Meyer for assistance in preparing the paper; Ben Coppin, Oriol Vinyals, Marek Barwinski, Ruoxi Sun, Carl Elkin, Peter Dolan, Matthew Lai and Yujia Li for their contributions and support; Olaf Ronneberger for reviewing the paper; the rest of the DeepMind team for their support; the CASP13 organisers and the experimentalists whose structures enabled the assessment.

3 Author contributions

R.E., J.J., J.K., L.S., A.S., C.Q., T.G., A.Z., A.B., H.P. and K.S. designed and built the AlphaFold system with advice from D.S., K.K. and D.H..
D.J. provided advice and guidance on protein structure prediction methodology.
S.P. contributed to software engineering.
S.C., A.N., K.K. and D.H. managed the project.
J.K., A.S., T.G., A.Z., A.B., R.E., P.K. and J.J. analysed the CASP results for the paper.
A.S., J.K. wrote the paper with contributions from J.J, R.E., L.S., T.G., A.Z., D.J., P.K., K.K. and D.H.
A.S. led the team.

4 Competing financial interests

A.S., J.K., L.S., R.E., H.P., C.Q., K.S. and J.J. have filed provisional patent application 62/734,757. A.S. and J.J. have filed provisional patent application 62/734,773. A.S., J.K., T.G., J.J. L.S., R.E. and C.Q. have filed provisional patent application 62/770,490. J.J., A.S., R.E., A.B., T.G. and A.Z. have filed provisional patent application 62/774,071. The remaining authors declare no competing financial interests.

5 Materials and correspondence

Correspondence and materials requests should be sent to Andrew Senior: `andrewsenior@google.com`

6 Reporting summary

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

7 Code availability

Upon publication we will make available source code for the distogram and torsion predictions, together with neural network weights and input data for the CASP13 targets.

We make use of several open-source libraries to conduct our experiments, particularly HHblits[1], PSI-BLAST[2] and the machine learning framework TensorFlow[§] along with the TensorFlow library Sonnet[¶] which provides implementations of individual model components[3]. We also used Rosetta[4] under license.

8 Data availability

The following public datasets were used in this work:

- PDB 2018-03-15
- CATH 2018-03-16
- Uniclust30 2017-10
- PSI-BLAST `nr` dataset (as of 2017-12-15)

We will make available our train/test split (CATH domain codes).

9 Extended data

The following tools and data set versions were used for the CASP system and for subsequent experiments.

- PDB 2018-03-15
- CATH 2018-03-16
- HHblits based on version 3.0-beta.3 (3 iterations, E-value 1e-3)
- HHpred web server
- Uniclust30 2017-10
- PSI-BLAST version 2.6.0 `nr` dataset (as of 2017-12-15) (3 iterations, E-value 1e-3)
- SST web server (March 2019)
- BioPython v1.65
- Rosetta v3.5
- PyMol for structure visualisation.
- TM-align 20160521

[§] `https://github.com/tensorflow/tensorflow`
[¶] `https://github.com/deepmind/sonnet`