# 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]

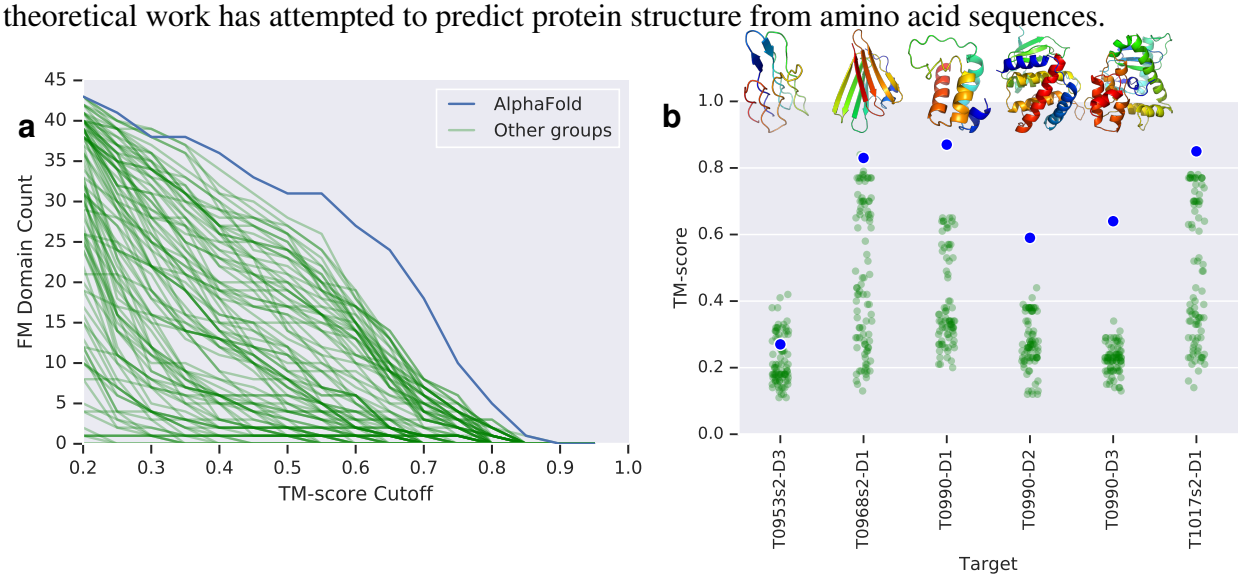
[1]*DeepMind, London, UK*
[2]*The Francis Crick Institute, London, UK*
[3]*University College London, London, UK*
[*]*These authors contributed equally to this work.*

**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 $\beta$-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`

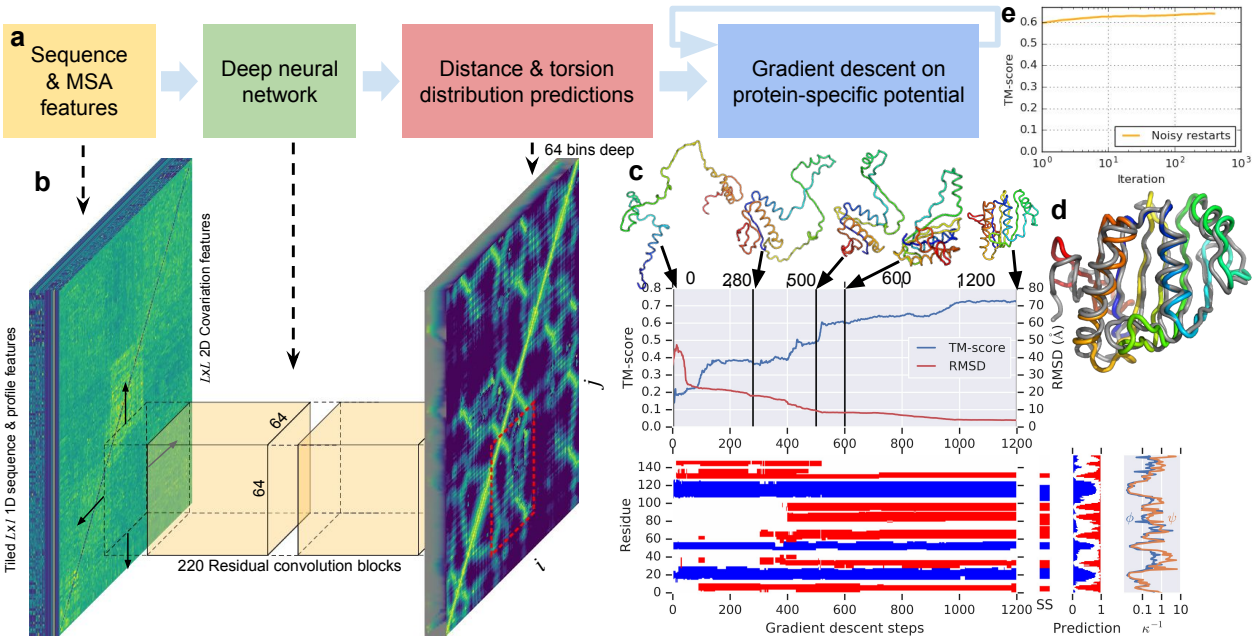



**Fig. 2 | The folding process illustrated for CASP13 target T0986s2.** (Length $L = 155$) (a) Steps of structure prediction. (b) The neural network predicts the entire $L \times L$ distogram based on MSA features, accumulating separate predictions for $64 \times 64$-residue regions. (c) One iteration 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 $\kappa^{-1}$ of the von Mises distributions fitted to the predictions for $\phi$ and $\psi$). 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 $C_\beta$ 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 $64 \times 64$ 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 $V_{\mathrm{distance}}$ 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 $C_\beta$ coordinates, $\mathbf{x}$, and thus the inter-residue distances, $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$, for each structure, and express $V_{\mathrm{distance}}$ 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.