BLOG POST
RESEARCH

15 JAN 2020

AlphaFold: Using AI for scientific discovery

Article

Improved protein structure prediction using potentials from deep learning

https://doi.org/10.1038/s41586-019-1923-7

Received: 2 April 2019

Accepted: 10 December 2019

Published online: 15 January 2020

Andrew W. Senior[1,4]*, Richard Evans[1,4], John Jumper[1,4], James Kirkpatrick[1,4], Laurent Sifre[1,4], 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]

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

RESEARCH ARTICLE | Open Access

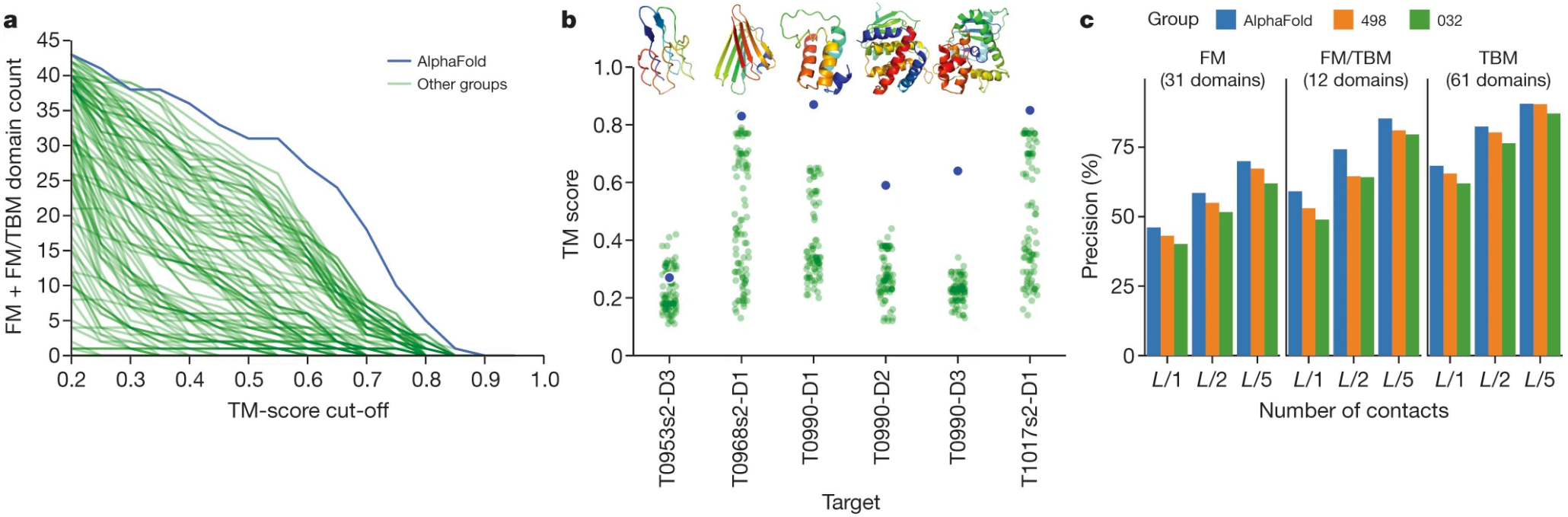
Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones ... See all authors

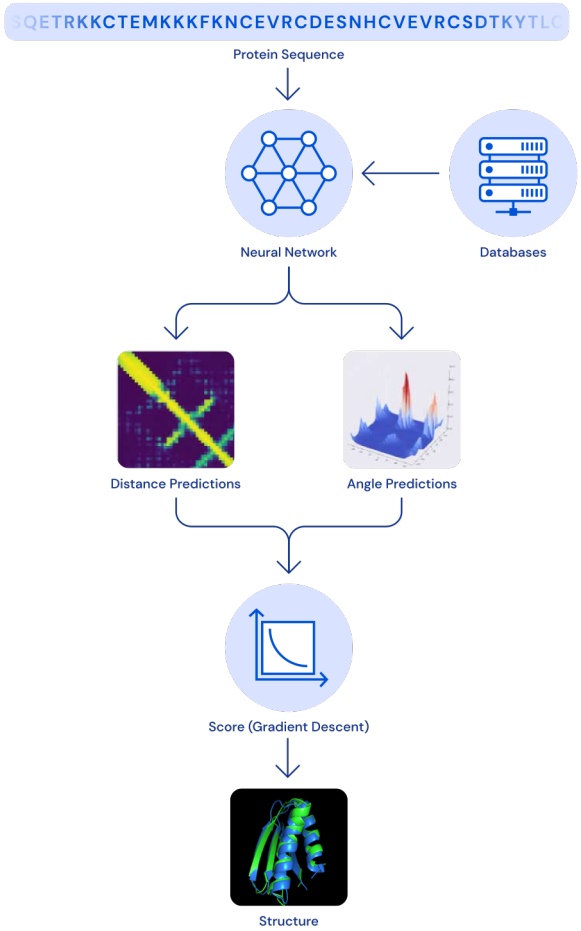
First published:10 October 2019 | https://doi.org/10.1002/prot.25834 | Citations: 9

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, and Laurent Sifre should be considered joint first authors

The performance of AlphaFold in the CASP13 assessment



A SCHEMATIC OF THE ARCHITECTURE OF THE ALPHAFOLD SYSTEM PREDICTING STRUCTURE FROM PROTEIN SEQUENCE

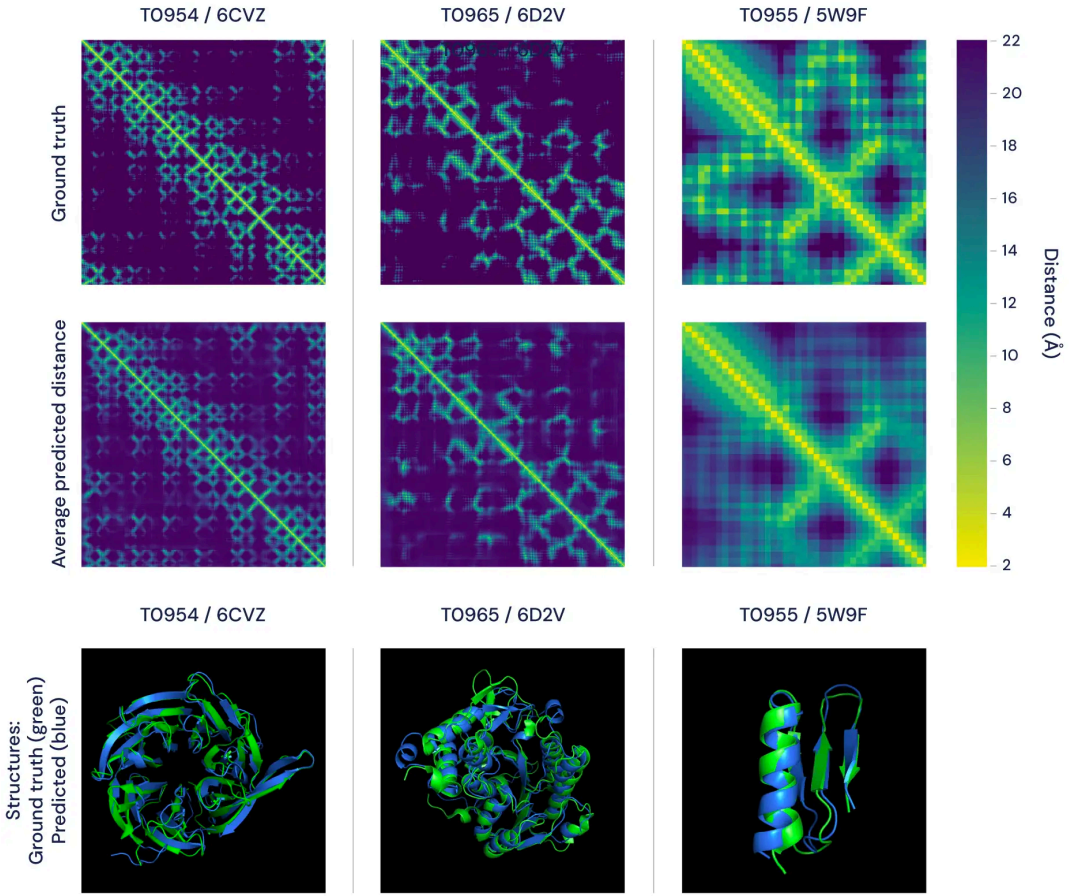


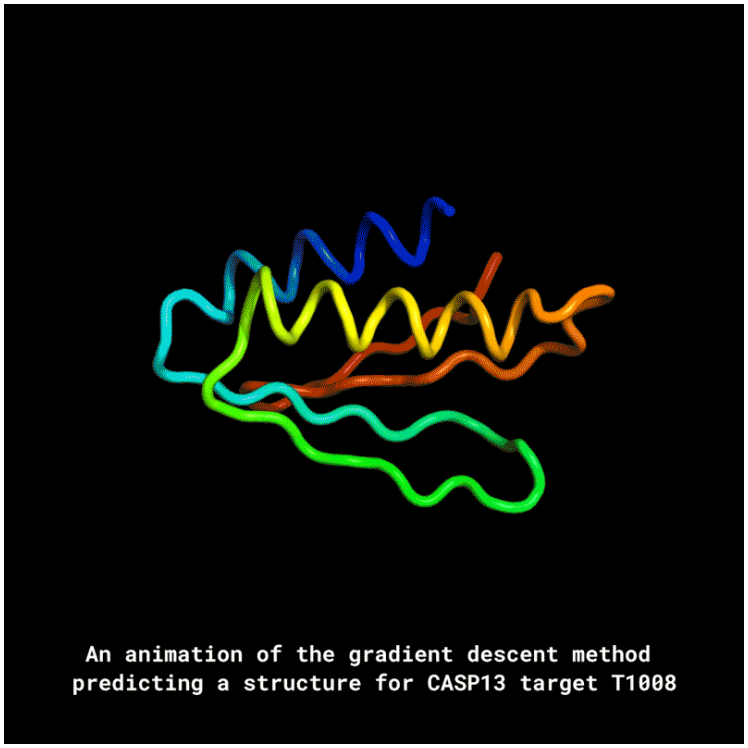
Using neural networks to predict physical properties

- Our team focused specifically on the problem of modelling target shapes from scratch, without using previously solved proteins as templates. We achieved a high degree of accuracy when predicting the physical properties of a protein structure, and then used two distinct methods to construct predictions of full protein structures.
- Both of these methods relied on deep neural networks that are trained to predict properties of the protein from its genetic sequence. The properties our networks predict are: (a) the distances between pairs of amino acids and (b) the angles between chemical bonds that connect those amino acids. The first development is an advance on commonly used techniques that estimate whether pairs of amino acids are near each other.
- We trained a neural network to predict a distribution of distances between every pair of residues in a protein. These probabilities were then combined into a score that estimates how accurate a proposed protein structure is. We also trained a separate neural network that uses all distances in aggregate to estimate how close the proposed structure is to the right answer.
- Using these scoring functions, we were able to search the protein landscape to find structures that matched our predictions. Our first method built on techniques commonly used in structural biology, and repeatedly replaced pieces of a protein structure with new protein fragments. We trained a generative neural network to invent new fragments, which were used to continually improve the score of the proposed protein structure.
- The second method optimised scores through gradient descent—a mathematical technique commonly used in machine learning for making small, incremental improvements—which resulted in highly accurate structures. This technique was applied to entire protein chains rather than to pieces that must be folded separately before being assembled into a larger structure, to simplify the prediction process.
- The code is available on Github for anyone interested in learning more, or replicating our protein folding results.

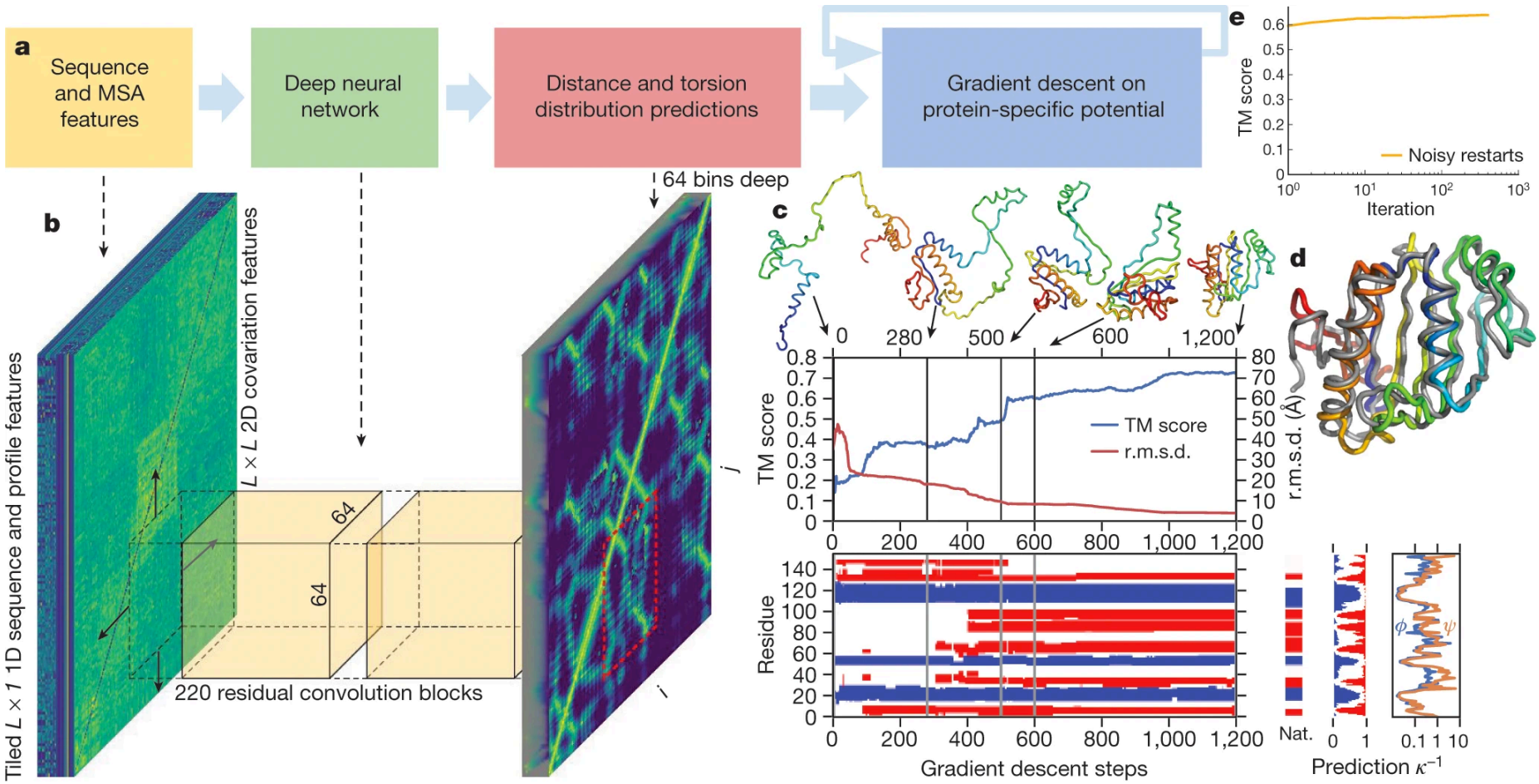
THE TOP FIGURE FEATURES THE DISTANCE MATRICES FOR THREE PROTEINS. THE BRIGHTNESS OF EACH PIXEL REPRESENTS THE DISTANCE BETWEEN THE AMINO ACIDS IN THE SEQUENCE COMPRISING THE PROTEIN–THE BRIGHTER THE PIXEL, THE CLOSER THE PAIR.
SHOWN IN THE TOP ROW ARE THE REAL, EXPERIMENTALLY DETERMINED DISTANCES AND, IN THE BOTTOM ROW, THE AVERAGE OF ALPHAFOLD'S PREDICTED DISTANCE DISTRIBUTIONS. IMPORTANTLY, THESE MATCH WELL ON BOTH GLOBAL AND LOCAL SCALES.

THE BOTTOM PANELS REPRESENT THE SAME COMPARISON USING 3D MODELS, FEATURING ALPHAFOLD'S PREDICTIONS (BLUE) VERSUS GROUND-TRUTH DATA (GREEN) FOR THE SAME THREE PROTEINS.



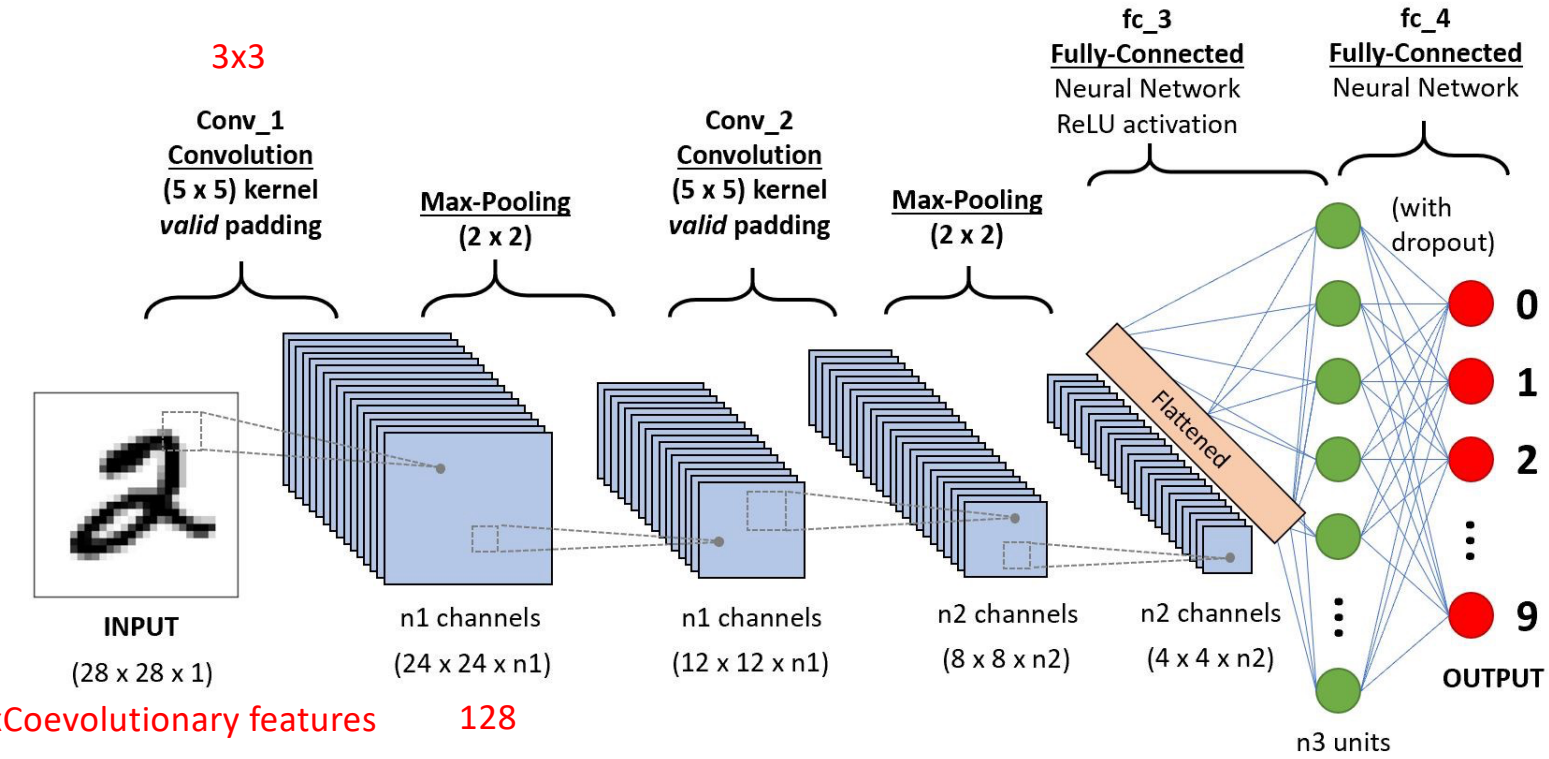

An animation of the gradient descent method
predicting a structure for CASP13 target T1008

The folding process illustrated for CASP13 target T0986s2.

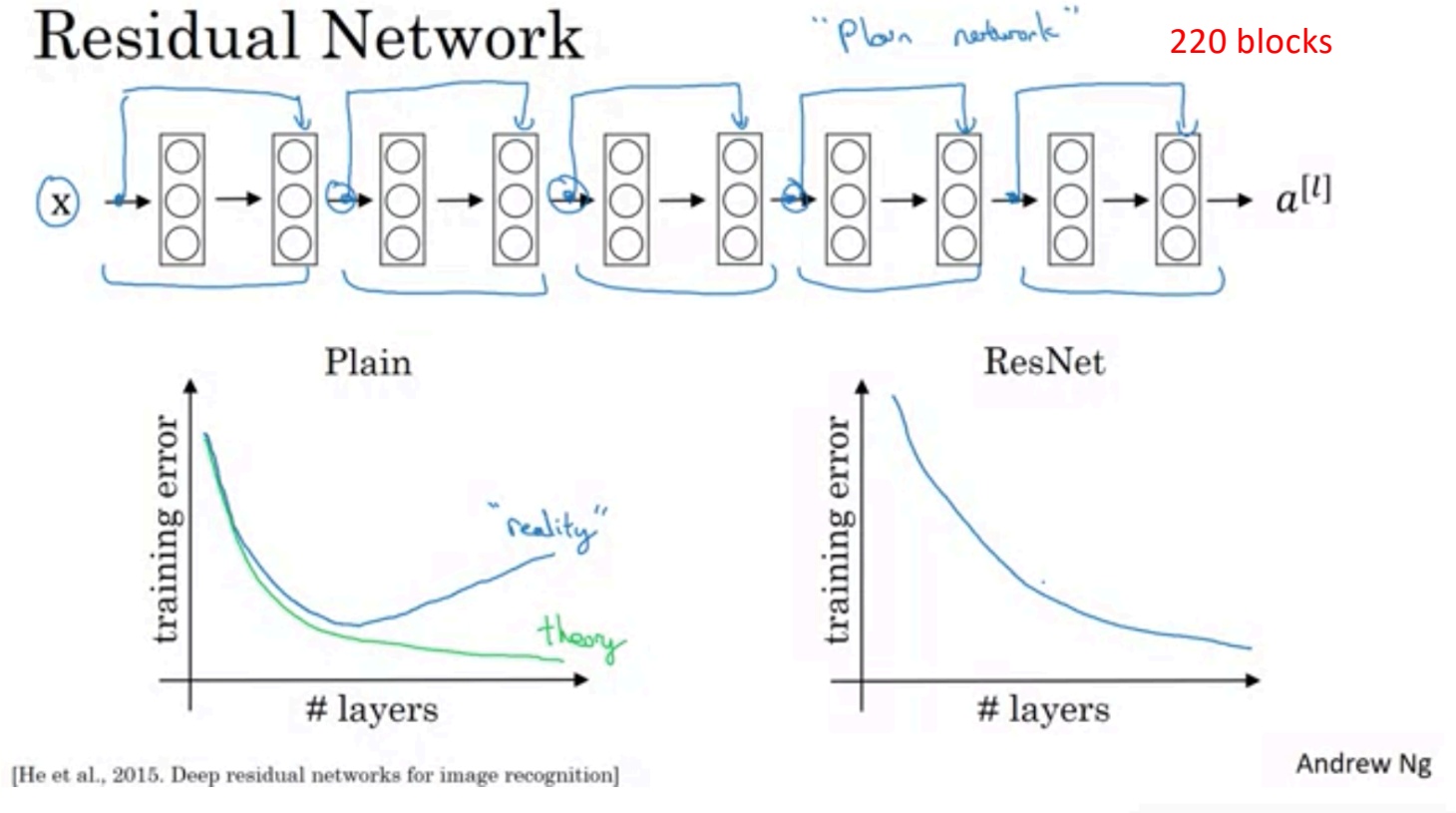


Convolutional Neural Networks (CNN)

In deep learning, a **convolutional neural network** (**CNN**, or **ConvNet**) is a class of deep neural networks, most commonly applied to analyzing images.

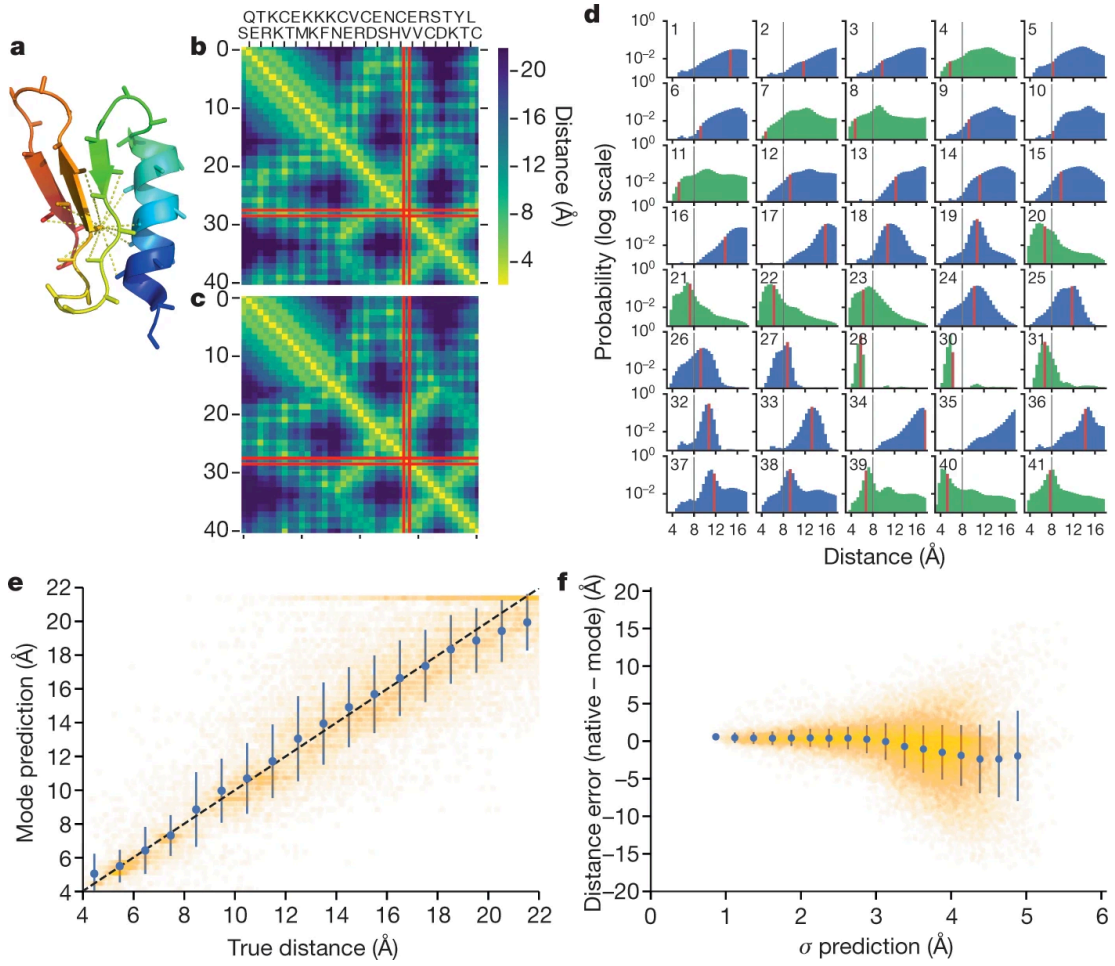


These channels function just like the RGB channels, but these channels are an abstract version of color, with each channel representing some aspect of information about the image.



Residual Network (ResNet) is a Convolutional Neural Network (CNN) architecture which can support hundreds or more convolutional layers. ResNet can add many layers with strong performance, while previous architectures had a drop off in the effectiveness with each additional layer.

Predicted distance distributions compared with true distances.

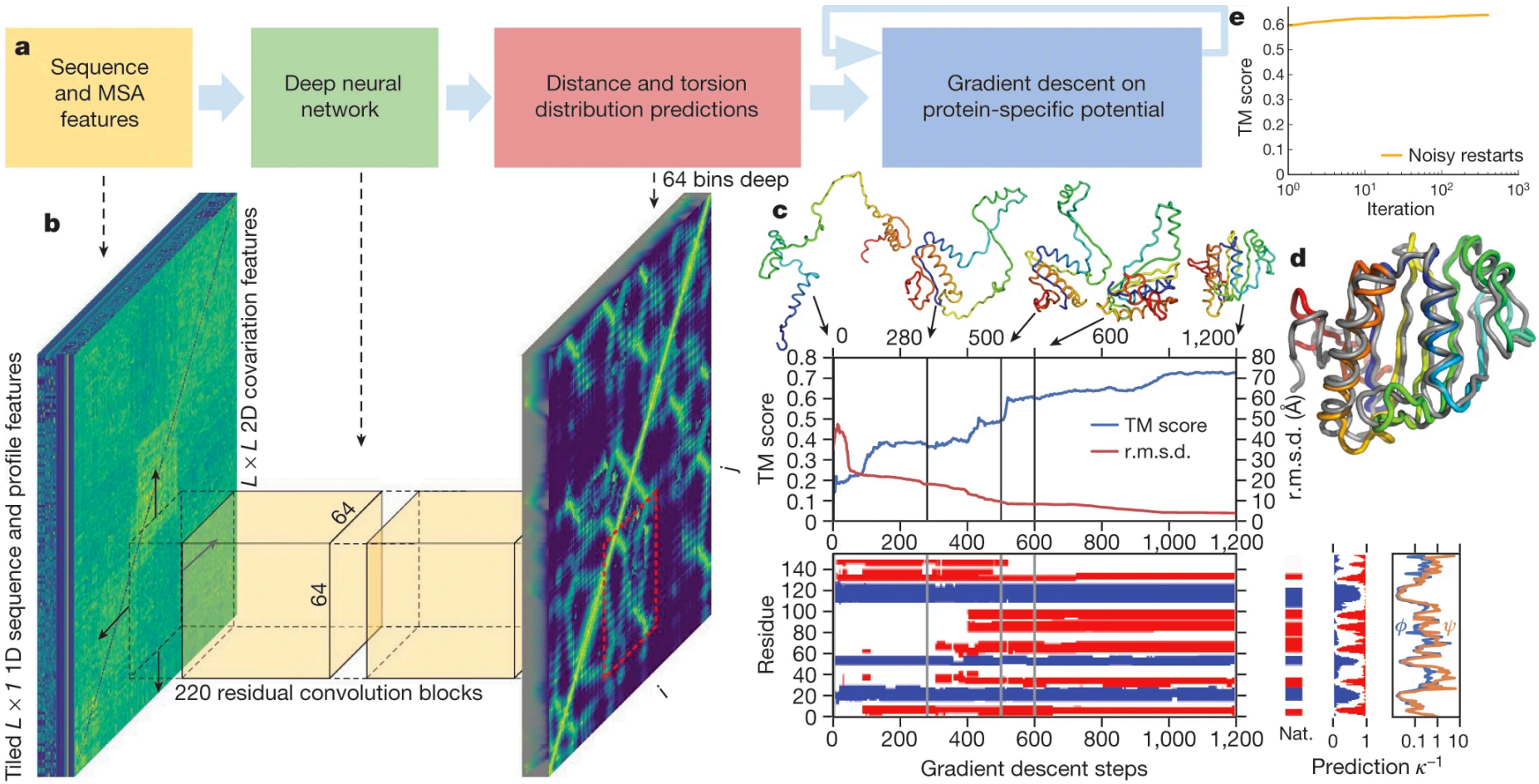


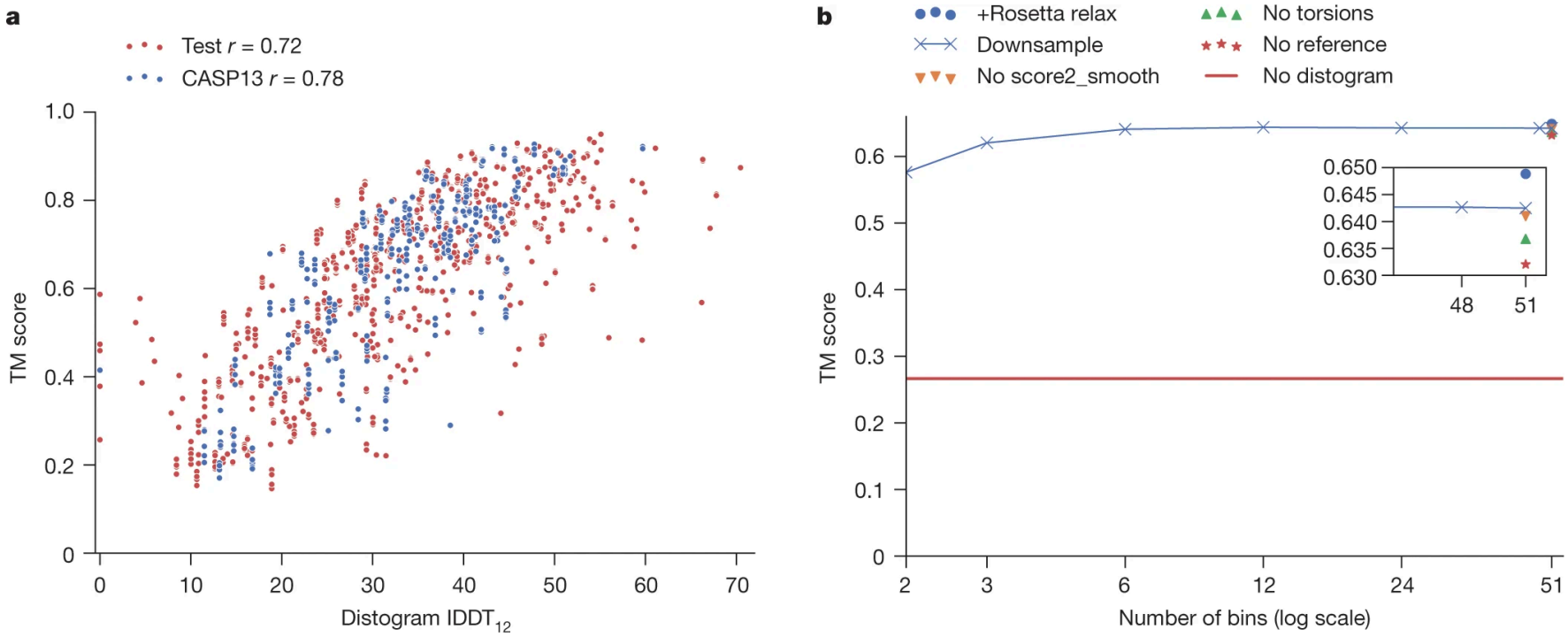
The predicted probability distributions for distances of residue 29 to all other residues. The bin corresponding to the native distance is highlighted in red, 8 Å is drawn in black. The distributions of the true contacts are plotted in green, non-contacts in blue.

We found that the predictions of the distance correlate well with the true distance between residues (Fig. 3e). Furthermore, the network also models the uncertainty in its predictions (Fig. 3f).

more confident predictions of the distance distribution (higher peak and lower s.d. of the distribution) tend to be more accurate, with the true distance close to the peak.

The folding process illustrated for CASP13 target T0986s2.




a
Test r = 0.72
CASP13 r = 0.78
TM score
Distogram IDDT12
b
+Rosetta relax
Downsample
No score2_smooth
No torsions
No reference
No distogram
TM score
Number of bins (log scale)

Data

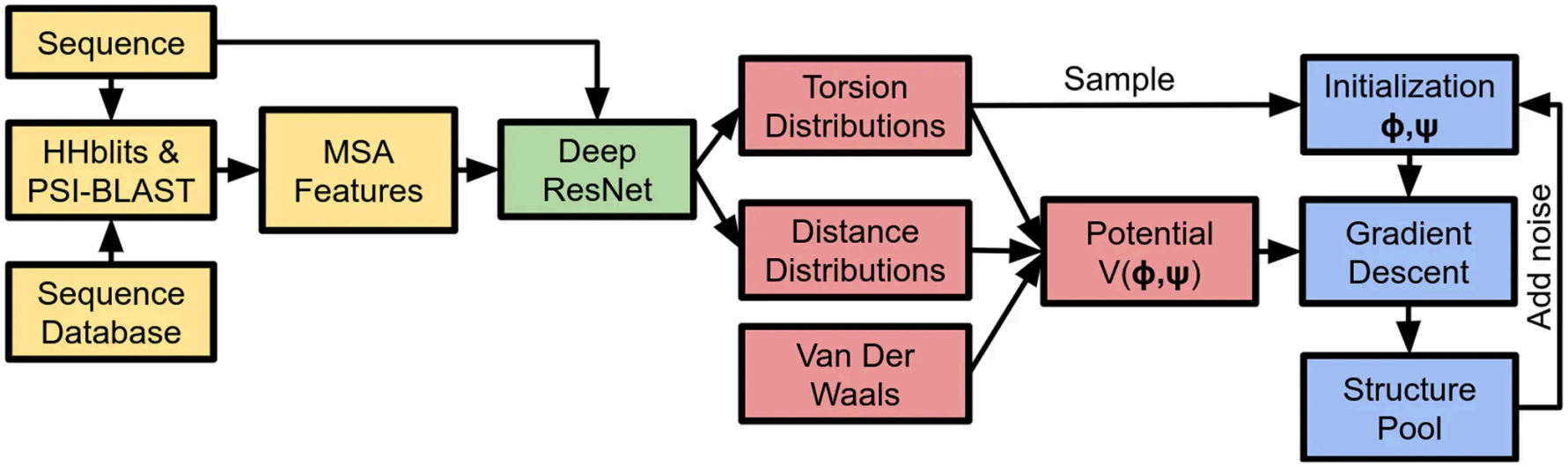
- Our models are trained on structures extracted from the PDB[13]. We extract non-redundant domains by utilizing the CATH[34] 35% sequence similarity cluster representatives. This generated 31,247 domains, which were split into train and test sets (29,427 and 1,820 proteins, respectively), keeping all 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 took—at random—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.

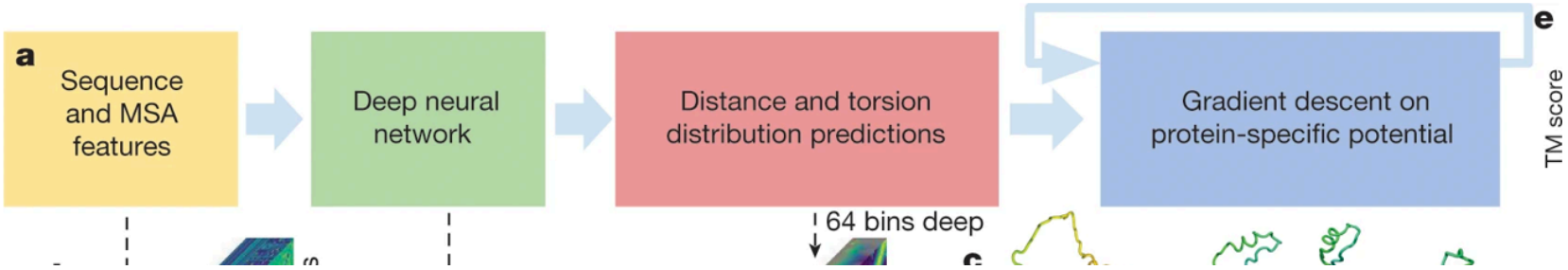
- For each training sequence, we searched for and aligned to the training sequence similar protein sequences in the Uniclust30[35] dataset with HHblits[36] and used the returned MSA to generate profile features with the position-specific substitution probabilities for each residue as well as covariation features—the parameters of a regularized pseudolikelihood-trained Potts model similar to CCMpred[16]. CCMPred uses the Frobenius norm of the parameters, but we feed both this norm (1 feature) and the raw parameters (484 features) into the network for each residue pair *ij*. In addition, we provide the network with features that explicitly represent 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[37].

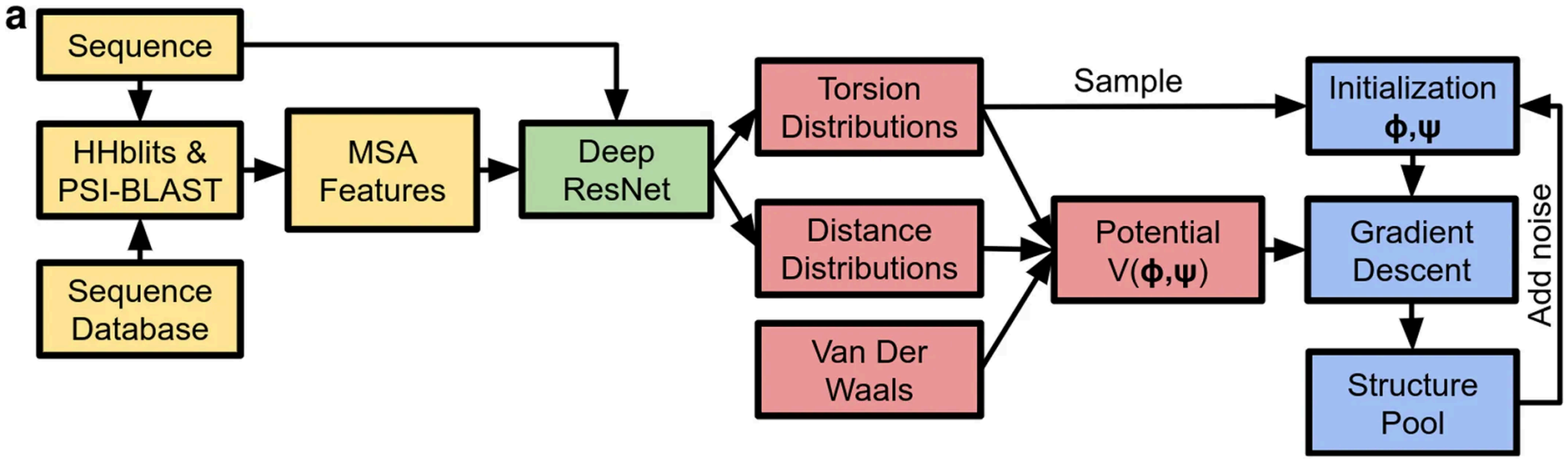
The distance prediction neural network was trained with the following input features (with the number of features indicated in brackets).

- Number of HHblits alignments (scalar).
- Sequence-length features: 1-hot amino acid type (21 features); profiles: PSI-BLAST (21 features), HHblits profile (22 features), non-gapped profile (21 features), HHblits bias, HMM profile (30 features), Potts model bias (22 features); deletion probability (1 feature); 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 (484 features, fitted with 500 iterations of gradient descent using Nesterov momentum 0.99, without sequence reweighting); Frobenius norm (1 feature); gap matrix (1 feature).

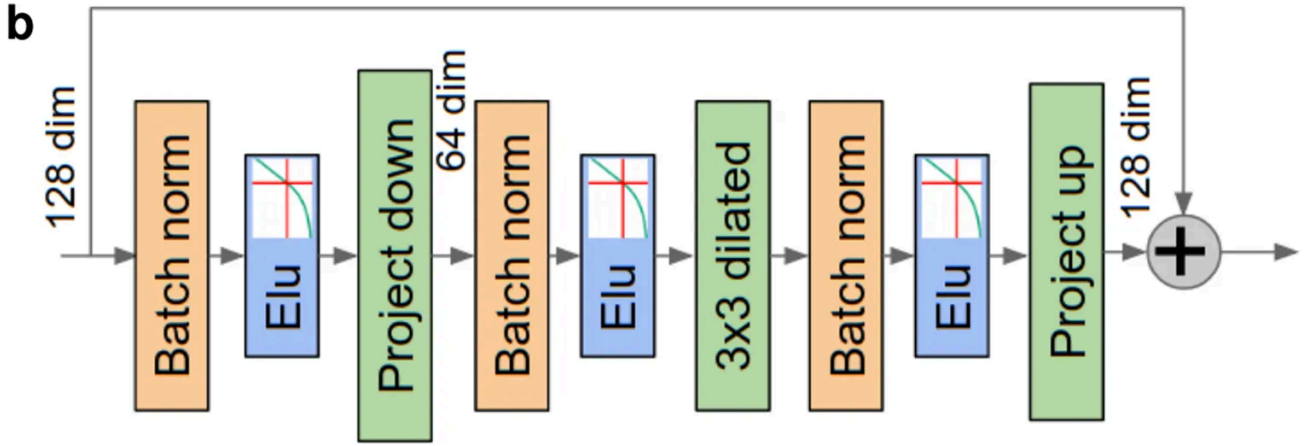
Feature extraction stages (constructing the MSA using sequence database search and computing MSA-based features) are shown in yellow; the structure-prediction neural network in green; potential construction in red; and structure realization in blue







one block of the deep residual convolutional network

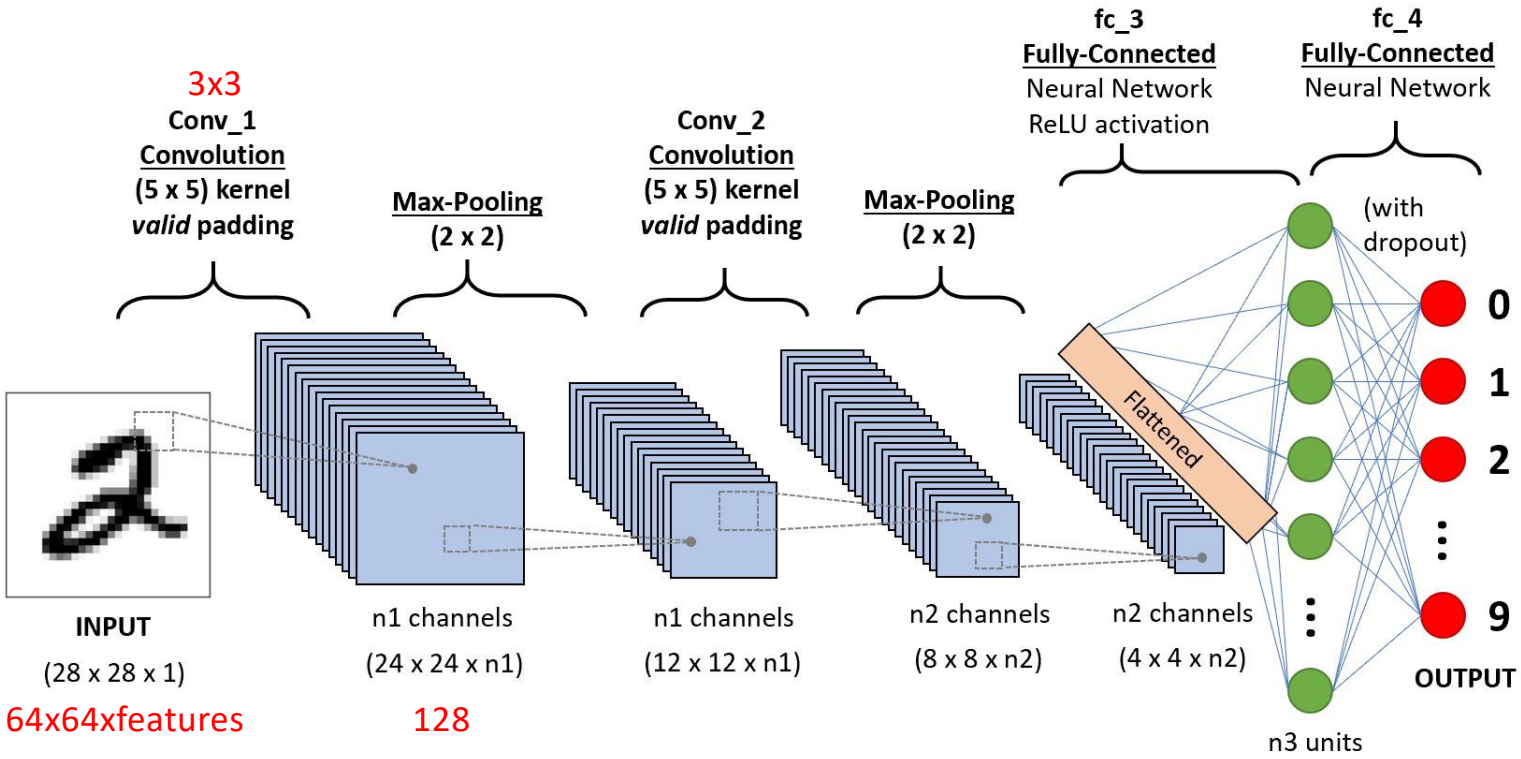


Structure realization by gradient descent

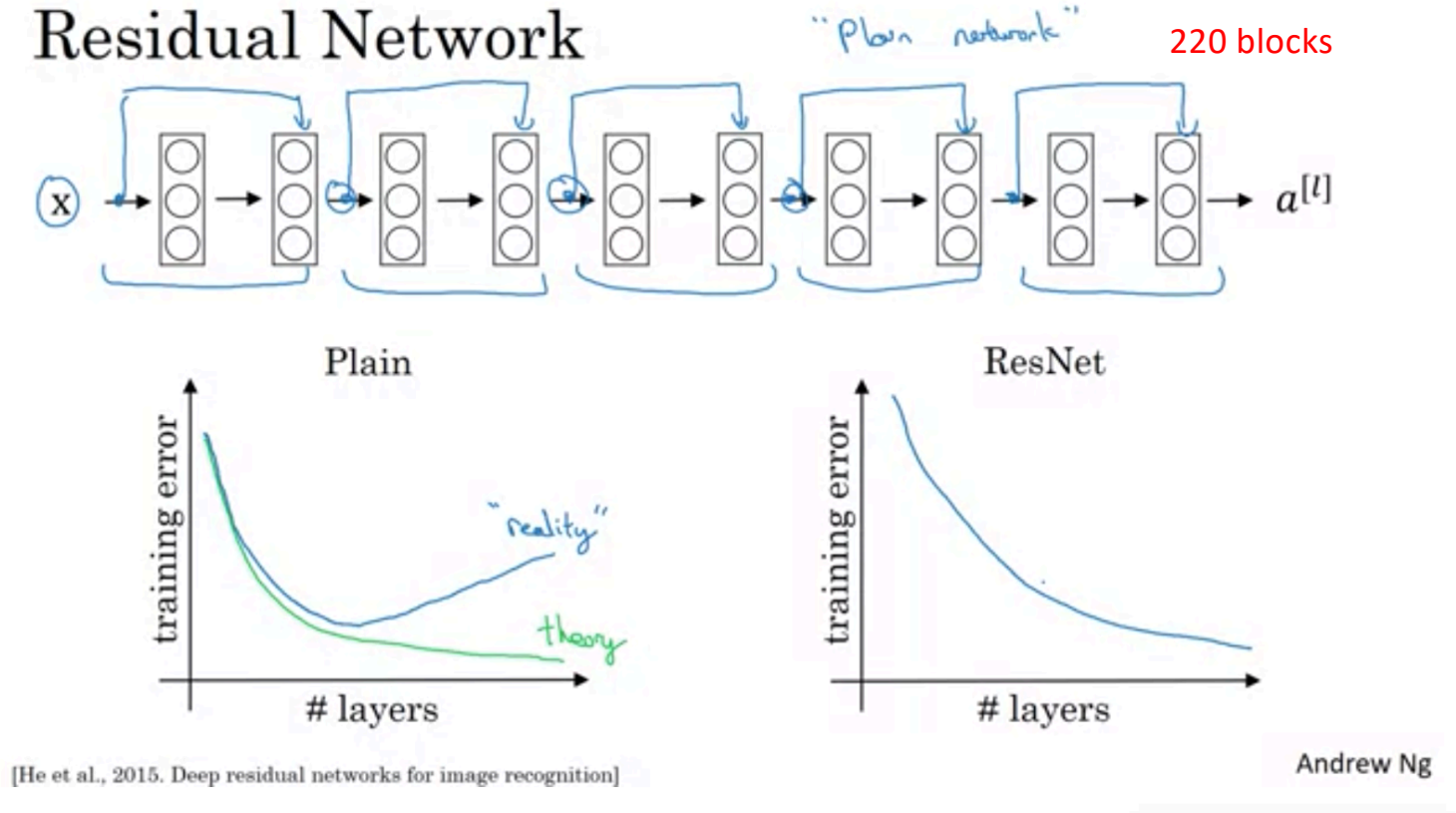
- To realize structures that minimize the constructed potential, we created a differentiable model of ideal protein backbone geometry, giving backbone atom coordinates as a function of the torsion angles (φ, ψ): $\mathbf{x} = G(\varphi, \psi)$. The complete potential to be minimized is then the sum of the distance, torsion and score2_smooth (Supplementary equation (4)). Although 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.
- As every term in V_{total} is differentiable with respect to the torsion angles, given an initial set of torsions φ, ψ, which can be sampled from the predicted torsion marginals, we can minimize V_{total} using a gradient descent algorithm, such as L-BFGS[31]. The optimized structure is dependent on the initial conditions, so we repeat the optimization multiple times with different initializations. A pool of the 20 lowest-potential structures is maintained and once full, we initialize 90% of trajectories from those with 30° noise added to the backbone torsions (the remaining 10% still being sampled from the predicted torsion distributions). In CASP13, we obtained 5,000 optimization runs for each chain. Figure 2c shows the change in TM score against the number of restarts per protein. As longer chains take longer to optimize, this work load was balanced across (50 + L)/2 parallel workers. Extended Data Figure 4 shows similar curves against computation time, always comparing sampling starting torsions from the predicted marginal distributions with restarting from the pool of previous structures.

Convolutional Neural Networks (CNN)

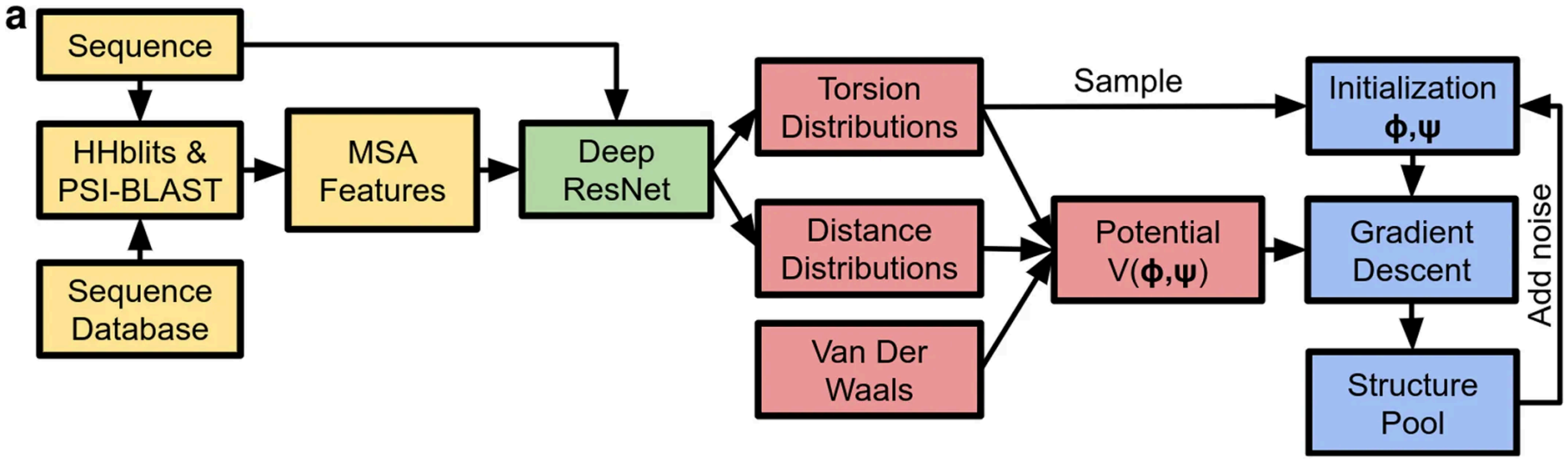
In deep learning, a **convolutional neural network** (**CNN**, or **ConvNet**) is a class of deep neural networks, most commonly applied to analyzing images.



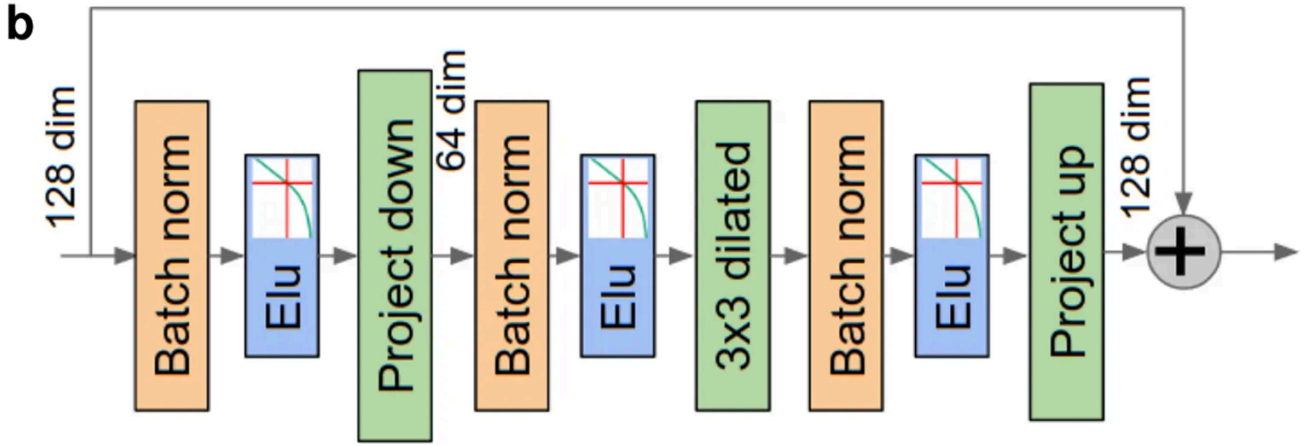
These channels function just like the RGB channels, but these channels are an abstract version of color, with each channel representing some aspect of information about the image.



Residual Network (ResNet) is a Convolutional Neural Network (CNN) architecture which can support hundreds or more convolutional layers. ResNet can add many layers with strong performance, while previous architectures had a drop off in the effectiveness with each additional layer.



one block of the deep residual convolutional network



RESEARCH ARTICLE | Open Access

Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones ... See all authors

First published:10 October 2019 | https://doi.org/10.1002/prot.25834 | Citations: 9

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, and Laurent Sifre should be considered joint first authors

The A7D system, called AlphaFold, used three deep-learning-based methods for free modeling (FM) protein structure prediction, without using any template-based modeling (TBM). These methods were based around combinations of three neural networks.

3 NNs

1. To predict the distances between pairs of residues within a protein (nature paper)
2. To directly estimate the accuracy of a candidate structure (termed the GDT-net)
3. To directly generate protein structures

The A7D submissions were generated by three methods which combined these algorithms:

A. Memory-augmented simulated annealing with neural fragment generation with GDT-net potential.
B. Memory-augmented simulated annealing with neural fragment generation with distance potential.
C. Repeated gradient descent of distance potential.

The main conclusions of this work are that the three systems performed similarly, with the GDT-net (A) and gradient descent (C) methods giving small improvements over B.

3 NNs

1. To predict the distances between pairs of residues within a protein(nature paper)

$$V_{\text{distance}}(\mathbf{x}) = \sum_{i,j,i \neq j} -\log P\left(d_{ij} \mid S, \text{MSA}(S)\right) - \log P\left(d_{ij} \mid \text{length}\right)$$

A distance potential is created from the negative log likelihood of the distances, summed over all pairs of residues *i*,*j*, and under a background model predicting the distance distributions $P(d_{ij} \mid \text{length})$ independent of sequence.

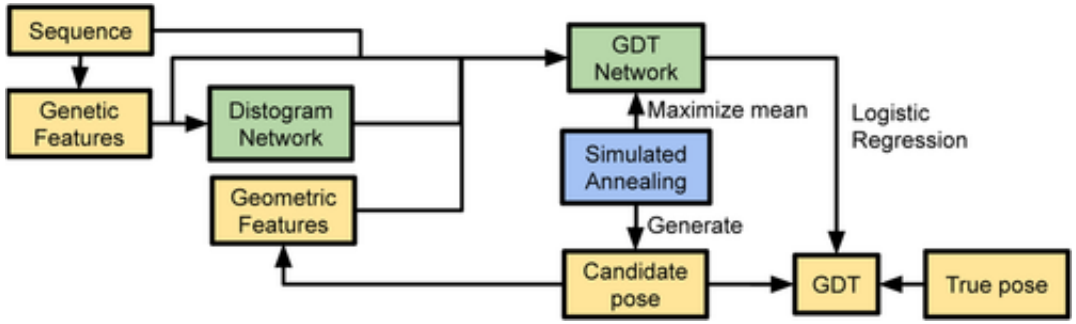
This distance-based potential is used for our fragment assembly system and our gradient descent system. We substitute it with a learned potential in the GDT-net model

2. To directly estimate the accuracy of a candidate structure (termed the GDT-net)

3. To directly generate protein structures

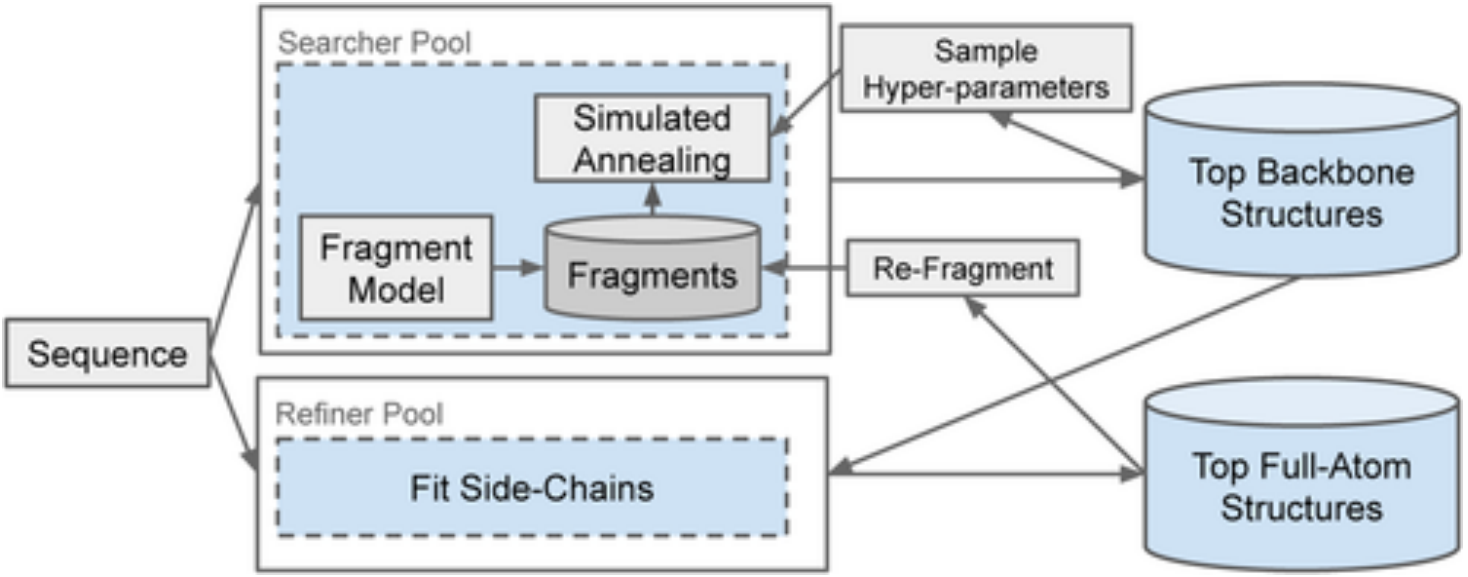
2. Estimate the accuracy of a candidate structure (termed the GDT-net)

- Predict GDT_TS (reference: AlphaGo reinforcement learning)



Mastering the game of Go without Human Knowledge

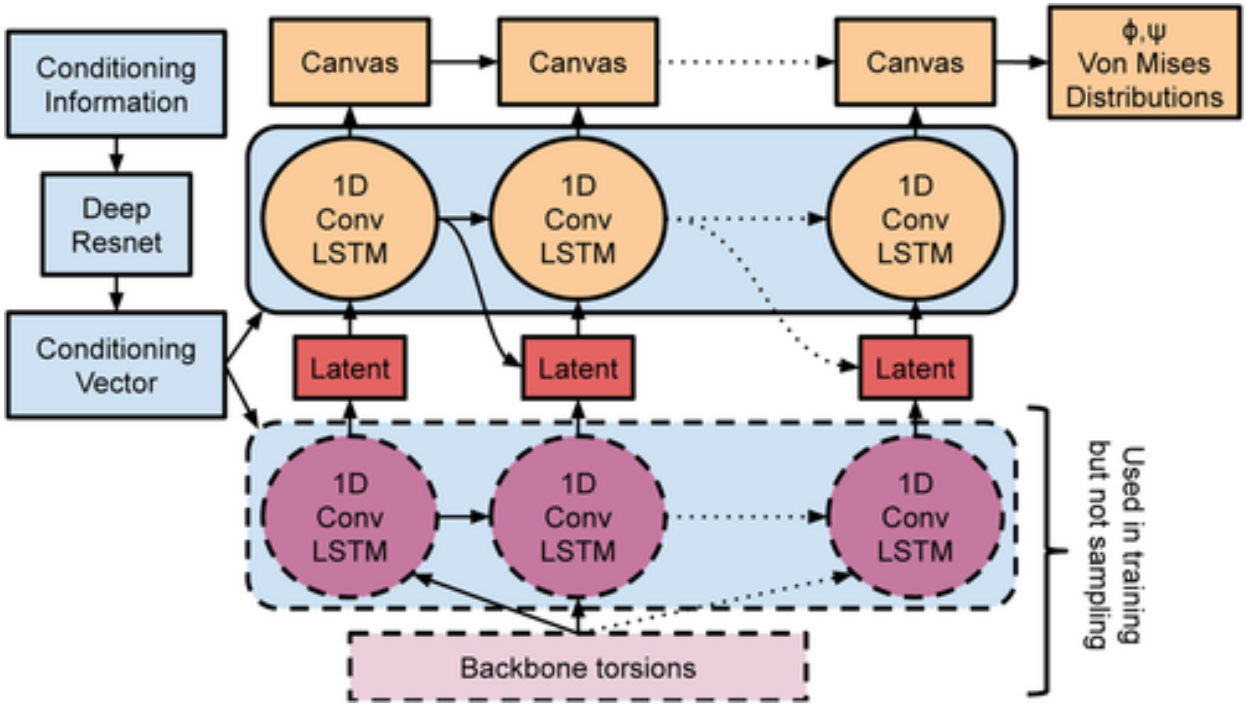
An overview of the simulated annealing framework. A pool of workers runs simulated annealing to optimize the backbone structure. Another pool refines these structures to add side-chain atoms. Fragments from these full-atom structures are reused in simulated annealing in a continuous fashion



3. Directly generate protein structures

a convolutional, autoregressive latent model

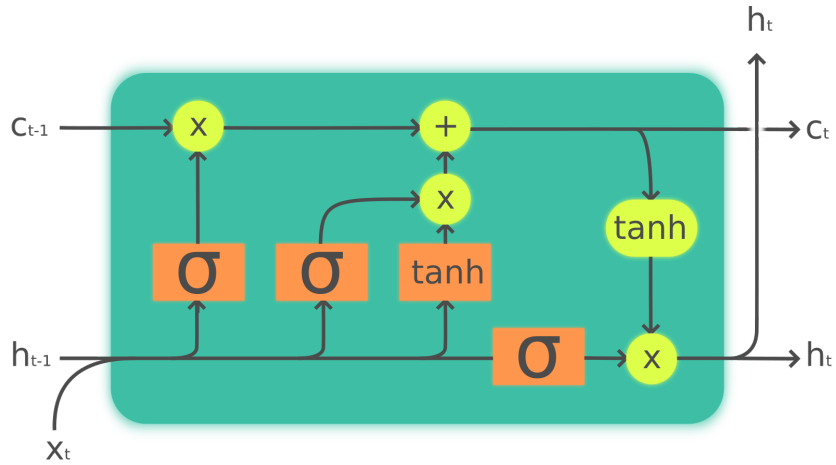
a convolutional, autoregressive latent model



A schematic of the fragment network 3. The blue parts of the network describe the conditioning network, the purple parts are the encoding network used to approximate the posterior, and the orange parts are the generative decoder

LSTM: long short-term memory (LSTM) – recurrent neural network (RNN)

LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series.



Legend:

Layer

Pointwize op

Copy

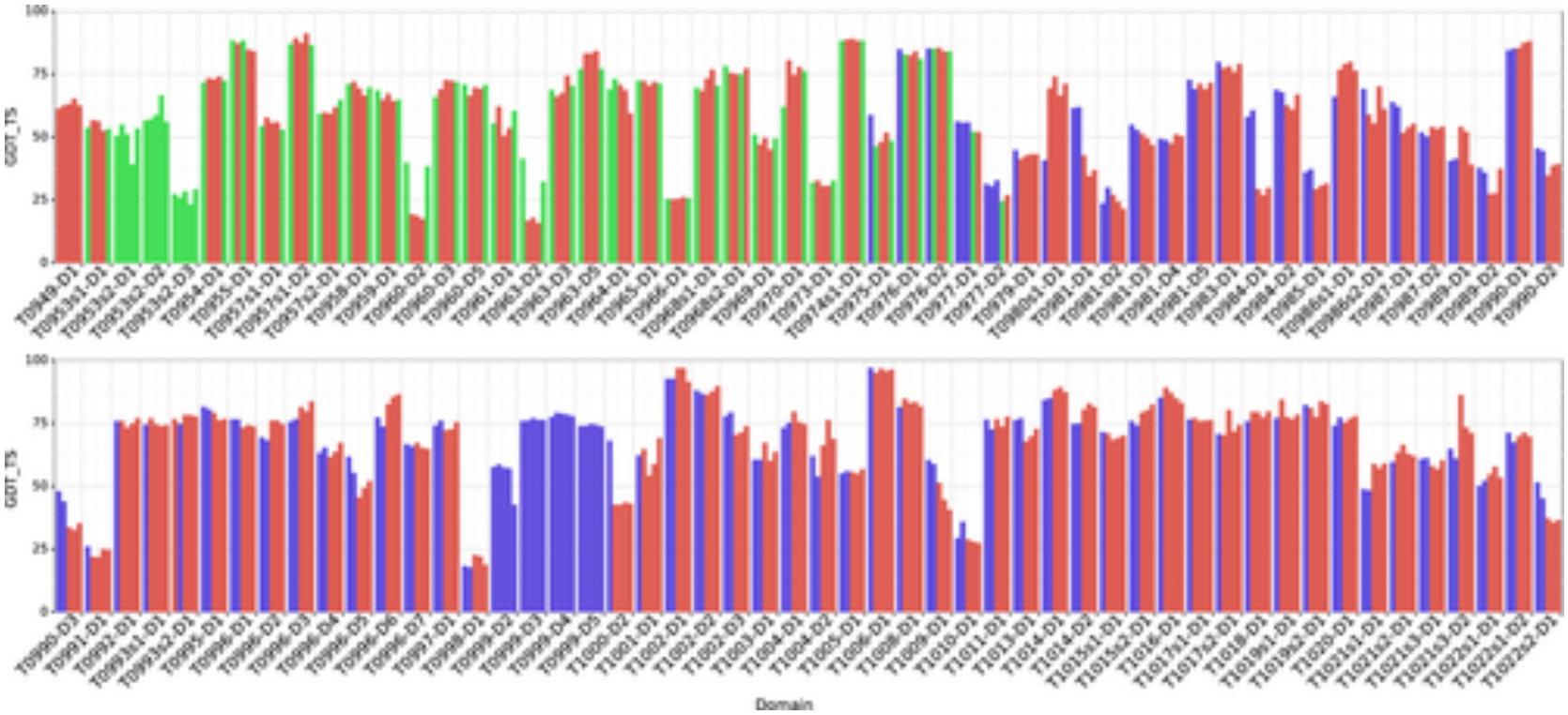
The Long Short-Term Memory (LSTM) cell can process data sequentially and keep its hidden state through time.

Repeated gradient descent

- Nature paper

Domain segmentation

- Simulated annealing is computationally expensive to run on long protein chains, particularly when using the GDT-net scoring.
- For this reason, we used a simple domain segmentation approach to partition a chain into pieces which are modeled independently in parallel.
- Our approach is based on assuming that domains will have many interresidue contacts whereas there will be fewer contacts between domains.
- We consider all possible partitions of a chain into two or three segments and score each segmentation, similar to the “Domain Guess by Size” method



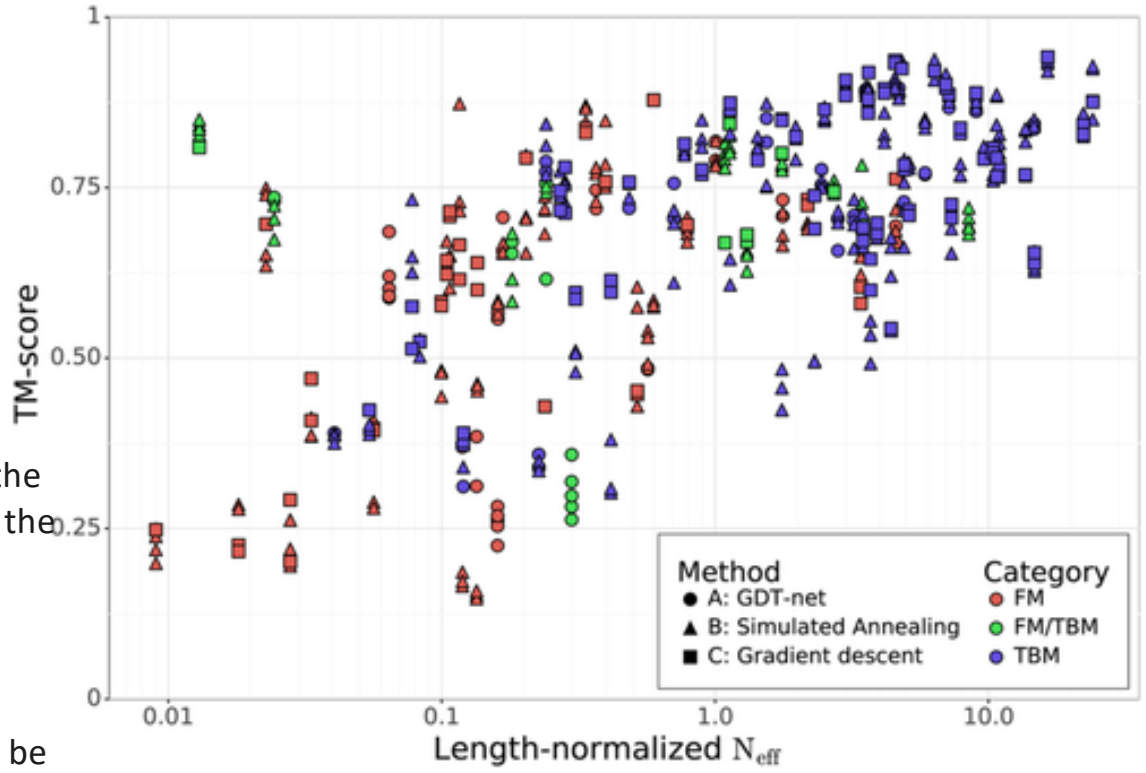
A7D CASP13 submission accuracies by domain. The GDT_TS for each of the five A7D CASP13 submissions are shown. Submissions are colored by method with fragment assembly submissions (B) colored red, GDT-net submissions (A) colored green, and gradient descent submissions (C) colored blue. T0999 (1589 residues) was manually segmented based on HHpred[28] homology matching

Decoy selection

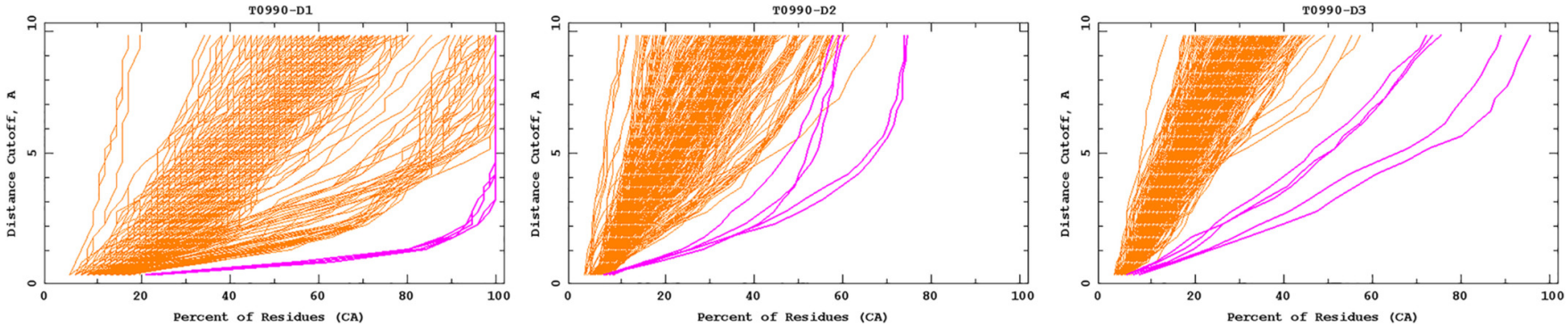
- For all but five of the targets in CASP13, we used exactly two of the three folding systems.
- Before target T0975, the two systems based on simulated annealing and fragment assembly (and using 40-bin distance distributions) were used (five independent runs with the distance potential, three with the GDT-net).
- From T0975 on, a newly trained 64-bin distance prediction network was used and structures were generated by the repeated gradient descent system (three independent runs) as well as the distance-potential fragment assembly system (five independent runs), while the GDT-net model was retired.
- Five submissions were chosen from the eight structures (the lowest potential structure generated by each independent run) with the first submission (“top-1”) being the lowest-potential structure generated by GDT-net (pre-T0975) or gradient descent (thereafter).
- The remaining four submissions were the four best other structures, with the fifth being a gradient descent structure/GDT-net if none had been chosen for position 2, 3, or 4.
- All submissions for T0999 were generated by gradient descent.

Table 1. A7D CASP13 accuracies by method. Average GDT_TS scores of the A7D CASP13 submissions broken down by method. Since the methods used changed after T0975, we show the means for these two sets separately. Domains in which only one method was used have been excluded to make the numbers comparable

	Mean GDT_TS for targets	
Method	**Before T0975**	**T0975 onwards**
Fragment assembly with GDT-net	63.8	N/A
Fragment assembly with distance potential	62.4	63.4
Gradient descent on distance potential	N/A	64.4



the system produces more accurate structures when the multiple sequence alignments are deeper, because of the distance predictor's dependence on coevolutionary information. Since the system does not search for templates, the performance on TBM targets is often worse than that for FM targets with similar N_{eff}. Performance on TBM targets with few alignments can be much worse than for systems which explicitly use templates (eg, T0973-D1 which was over 40 GDT_TS worse than the best submission). Interestingly, the low-alignment designed protein T0955-D1 was solved to high accuracy (GDT_TS 88.4) despite having no alignments, presumably because of its short length and because the design process ensured it had highly typical structure.



The A7D system was able to generate good structures for several hard targets, for instance, the three-domain protein T0990, shown in Figure 7. In this case, domain D3 is inserted into domain D2, so our domain segmentation algorithm, which only considers single-segment domains, was unable to generate a correct segmentation. It can be seen that the gradient descent method which does not use domain segmentation produced better results than the fragment assembly method. For T0980 s1-D1, on the other hand, fragment assembly produced better models than the repeated gradient descent, which failed to correctly assemble the beta sheet.

Discussion

- We have shown in this blind evaluation that deep-learning-based methods have excellent performance across a range of targets, including novel folds.
- All approaches rely heavily on a deep distance prediction neural network which uses coevolution information as inputs. We found that all three approaches performed similarly, but having the diversity of the different methods generating submissions for each target was useful.
- Despite the differences in the structure assembly methods, we did not find significant differences in accuracy arising from native contact order or other structural features. Our approaches tried to avoid heuristics and hand-crafted assumptions on the structure of proteins but for the fragment assembly approach we relied on a heuristic method to segment domains as described in Equation (3). In contrast, many fragment assembly approaches rely on secondary structure to limit the types of fragments available in certain regions and to modify the folding potential. The distance prediction network can express ambiguity by predicting distance distributions, which can represent secondary structure in short-range distances, in a way which is harder to do with three-way classification of secondary structure. The gradient descent algorithm has even fewer hyperparameters and assumptions than this and performs better.
- We note that our method performed well in the TBM + TBM/FM category, despite none of the methods explicitly using template information. This is because proteins that have a template in the PDB also tend to have rich coevolutionary information and can thus be well modeled by the distance prediction potential.
- The main weakness of our approach is that it still relies heavily on coevolution. When few alignments are available, distance predictions tend to be uninformative and poor structures are generated (Figure 6). Since there is no explicit template lookup, performance on TBM targets with few homologous sequences was much worse than template-based methods.
- Also, we did not attempt to propagate the uncertainty about distance into an uncertainty in residue positions. The B-factors submitted were incorrect leading to suboptimal performance when A7D models were used for molecular replacement.