Improved protein structure prediction using potentials from deep learning

Mingyu Derek Ma

CM229 22S Paper Presentation

April 25, 2022

The Task: Protein Folding

High-level Idea of AlphaFold

Methods

Evaluation

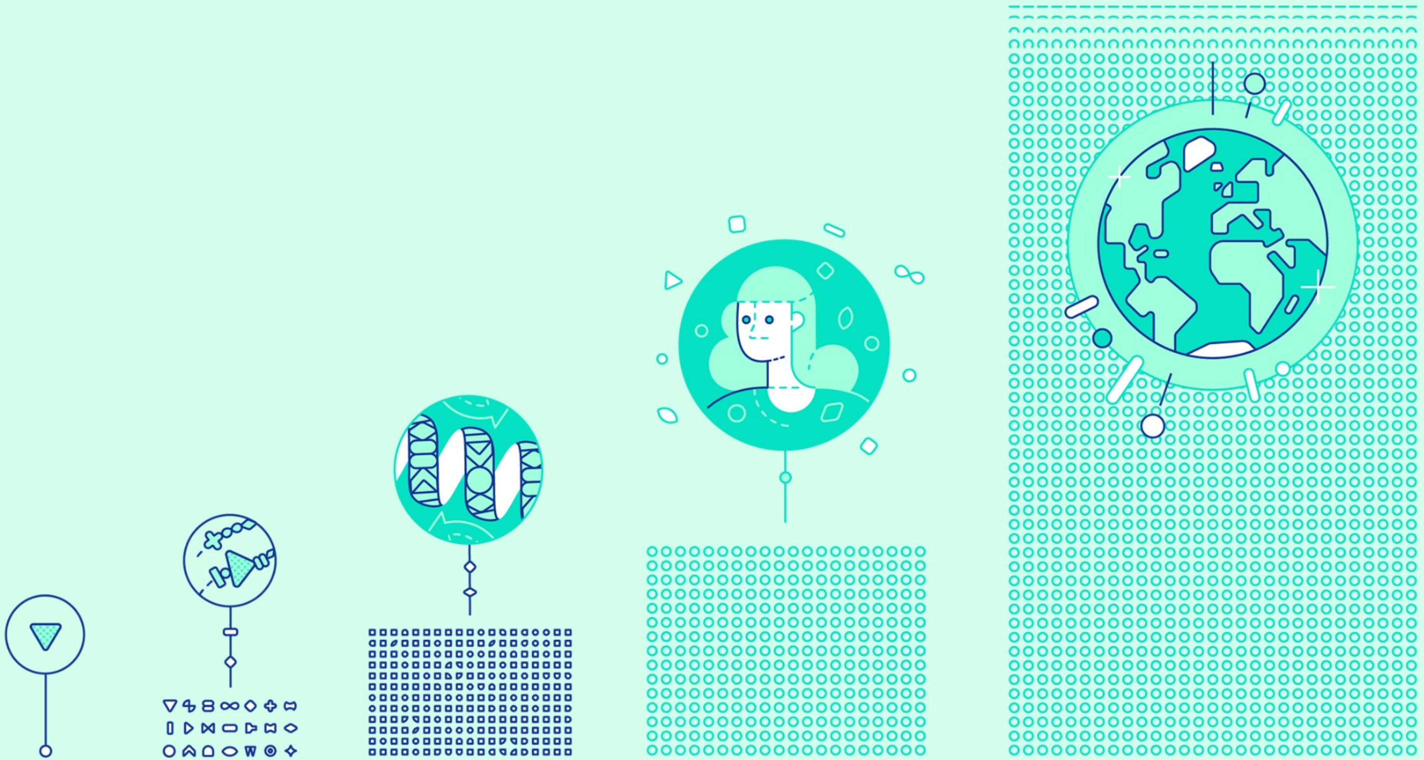
Limitation

Conclusion

The Task: Protein Folding

Protein

- Proteins are the building blocks of life, they are large, complex molecules essential to nearly every function that our body performs
- There are around 100 million known distinct proteins, each one has a unique 3D shape that determines how it works and what it does
- Figuring out the exact structure of a protein: expensive, time-consuming



1
AMINO ACID

20
AMINO ACIDS
IN A STRING

100's
AMINO ACIDS
IN A PROTEIN

20,000
PROTEINS IN
IN HUMAN BODY

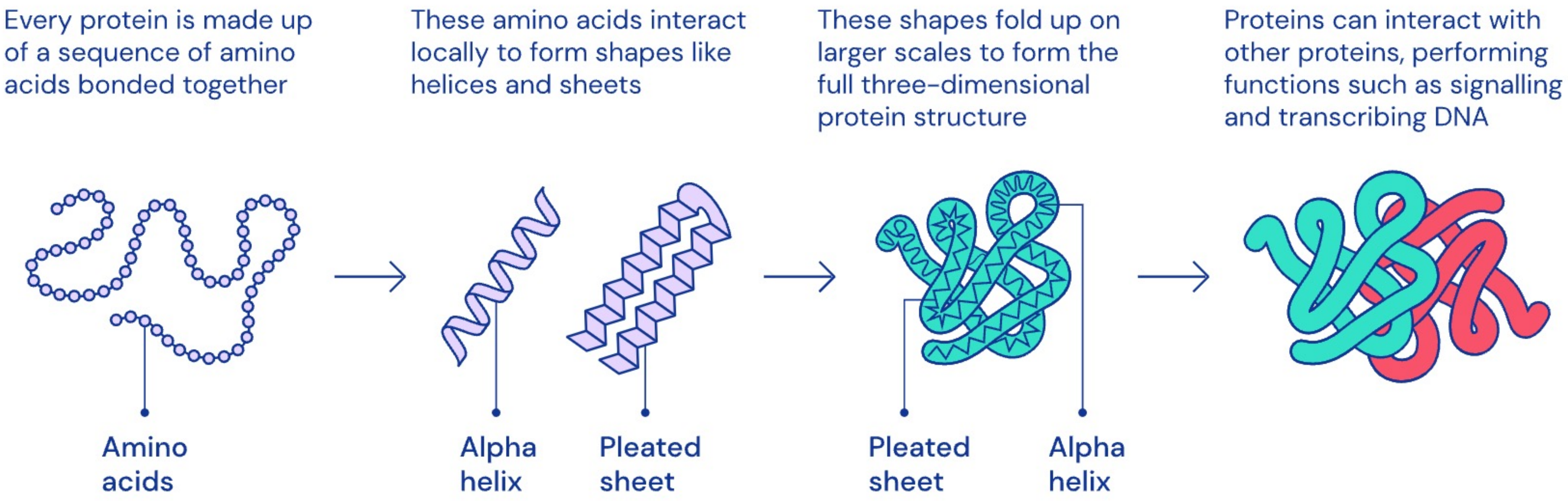
100,000,000
KNOWN PROTEINS
FOUND ON EARTH

[AlphaFold (deepmind.com)]()

Task: protein folding

- What any given protein can do depends on its unique 3D structure
- Recipes for proteins (genes) are encoded in DNA
- Proteins are comprised of chains of amino acids
- Given those sequence, we want to know how **chains of amino acids** fold into the intricate **3D structure** -> protein folding problem
- Why protein folds?
  - Attraction and repulsion between the 20 different types of amino acids cause the string to fold in a feat of “spontaneous origami”
  - Form the intricate curls, loops and pleats of a protein’s 3D structure

Task: protein folding



Every protein is made up of a sequence of amino acids bonded together

These amino acids interact locally to form shapes like helices and sheets

These shapes fold up on larger scales to form the full three-dimensional protein structure

Proteins can interact with other proteins, performing functions such as signalling and transcribing DNA

AlphaFold: Using AI for scientific discovery (deepmind.com)

Experimental techniques

- Experimental techniques
  - Cryo-electron microscopy
  - Nuclear magnetic r esonance
  - X-ray crystallography
- Disadvantages
  - Each method depends on a lot of trial and error
  - Take years of work
  - Take millions of dollars

AI needs data to train... Is data ready?

- Huge data available thanks to experimental structure techniques
  - 150,000 Protein Data Banks entries
  - Highly redundant, compare with scale of many other problems
    - 10s of millions of utterances for speech
    - 15 million labelled images in ImageNet for computer vision

Testbed: CASP13

- 13th Critical Assessment of protein Structure Prediction
- Biennial Critical Assessment of Protein Structure Prediction
  - **Blind** structure prediction of 82 newly-solved structured
  - For each chain, 3 weeks to return up to 5 structure predictions
  - 90+ groups from labs around the world
- Post-hoc scoring relative to ground-truth
  - Chains are partitioned into domains
  - Domains
    - Free-Modelling (FM): no homologous structure is available
    - Template-Based Modelling (TBM): a solved protein can be found that has a similar sequence and used to infer the shape
    - FM/TBM: intermediate category
  - Metrics are chosen post-hoc based on backbone alignment metric GDT_TS
- AlphaFold do exclusively free-modelling

Home – Prediction Center

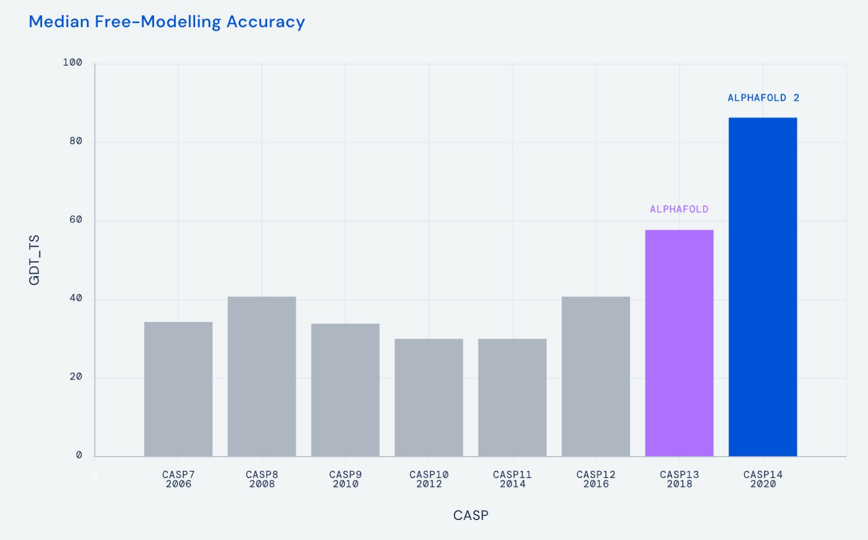
Existing FM approaches

- Relied on fragment assembly
    - A structure hypothesis is repeatedly modified, typically by changing the shape of a short section while retaining changes that lower the potential
    - This requires many thousands of such moves and must be repeated many times to have good coverage of low-potential structures
- Existing works predict contact probability between residues
    - Distance between two residue are within a certain threshold
- There is neural network approach to predict distance between residues without covariation features

High-level Idea of AlphaFold

Timeline of AlphaFold

- 2016: AlphaGo for Go
- 2018: AlphaFold first public test
  - Benchmarked in the 13th Critical Assessment of Protein Structure Prediction (CASP13), ranked first
- 2020: AlphaFold 2
  - Huge margin and win the CASP14
  - Three times more accurate than the next best system and comparable to experimental methods

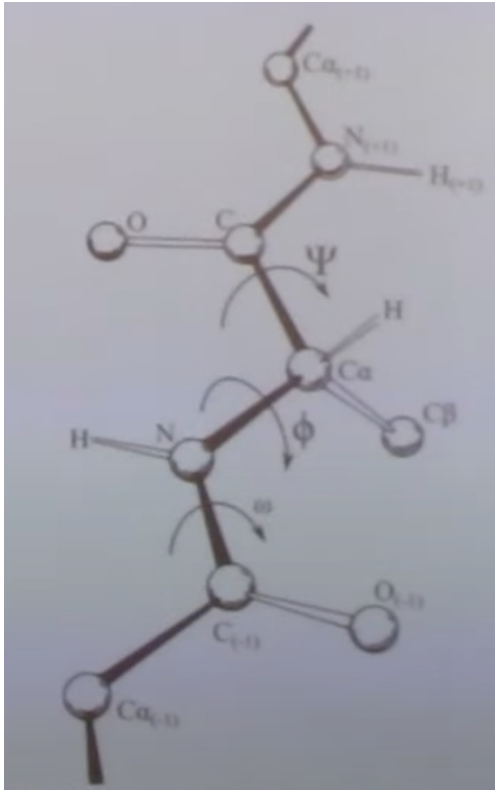


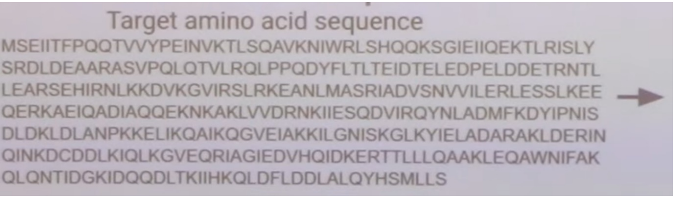
Goal of AlphaFold

- Modeling protein shapes from scratch
- Without using previously solved proteins as templates
- Input data is the genetic sequence of the protein

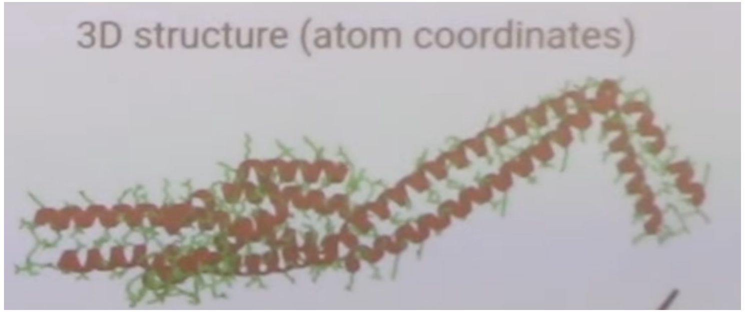
Representing “shape”

- Amino acid residues connected in a chain with a repeating -N-C-C- backbone
    - Side chains connected to the C-alpha determine structure
    - Goal is to predict the coordinates of every atom, particularly the backbone
- Torsion angles (Φ, Ψ) for each residue are a complete parameterization of backbone geometry
    - L-length sequence -> 2L parameters
- Another function to map torsion angles to atom coordinates




Target amino acid sequence

MSEIITFPQQTVVYPEINVKTLSQAVKNIWRLSHQQKSGIEIIQEKTLRISLY
SRDLDEAARASVPQLQTVLRQLPPQDYFLTLTEIDTELEDPELDDETRNTL
LEARSEHIRNLKKDVKGVIRSLRKEANLMASRIADVSNVVILERLESSLKEE
QERKAEIQADIAQQEKNKAKLVVDRNKIIESQDVIRQYNLADMFKDYIPNIS
DLDKLDLANPKKELIKQAIKQGVEIAKKILGNISKGLKYIELADARAKLDERIN
QINKDCDDLKIQLKGVEQRIAGIEDVHQIDKERTTLLLQAAKLEQAWNIFAK
QLQNTIDGKIDQQDLTKIIHKQLDFLDDLALQYHSMLLS


3D structure (atom coordinates)

AlphaFold: improved protein structure prediction using potentials from deep learning - YouTube

Predict distances and torsion with neural networks

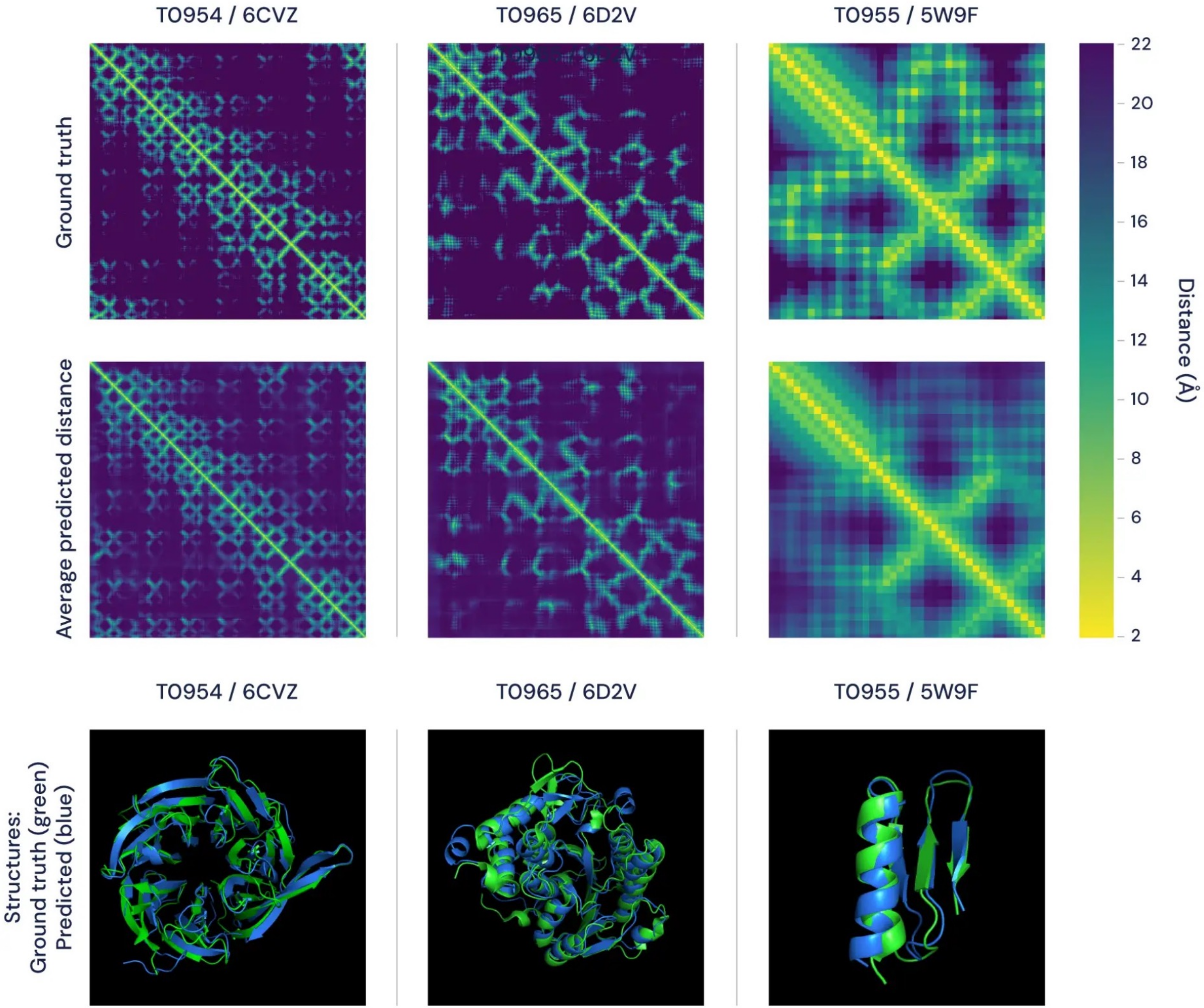
- Predictions to make
    - The **distances** between pairs of amino acids
    - The angles between chemical bonds that connect those amino acids -> **torsion angles**
- Train a neural network to predict a distribution of distances between **every pair** of residues in a protein
- Pair-wise probabilities were then combined into a score that estimates how accurate a proposed protein structure is
- The system primarily rely on distance prediction, get little gain from torsion angle prediction

Predict distances and torsion with neural networks

- Previous works predict contact, which only contain very small distance scenarios
    - Contact prediction -> distance prediction
    - 2 bins to many more bins
- Benefits of distance prediction
    - Much richer and specific training signal
    - Fine-grained detailed signal
    - Network can propagate distance information that respects covariation, local structure and residue identities of nearby residues

Distances vs 3D structures

- Three protein examples
- Brighter = closer
- Bright pixel far away from diagonal = residues which are distance along the sequence but are close in 3D structure



What we can read from the distogram

1: a helix self-contact

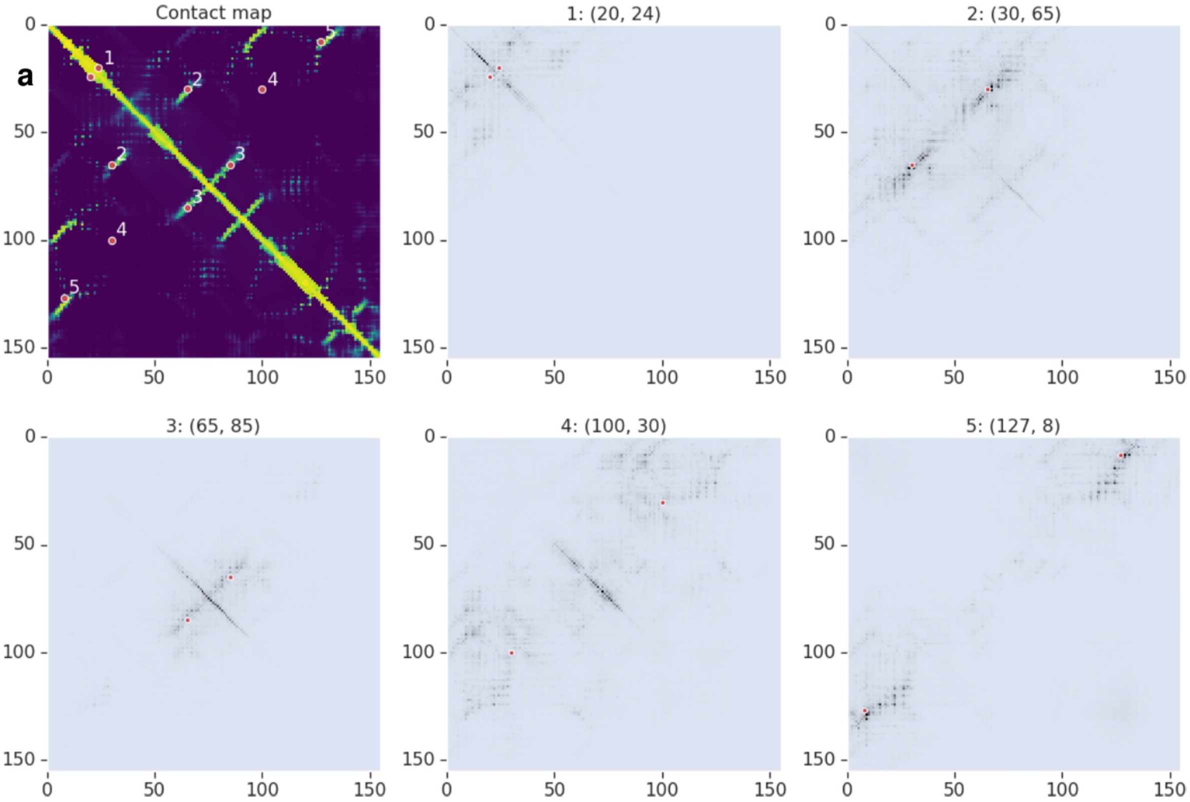
2: a long-range strand-strand contact

3: a medium-range strand-strand contact

4: a non-contact

5: a very long range strand-strand contact

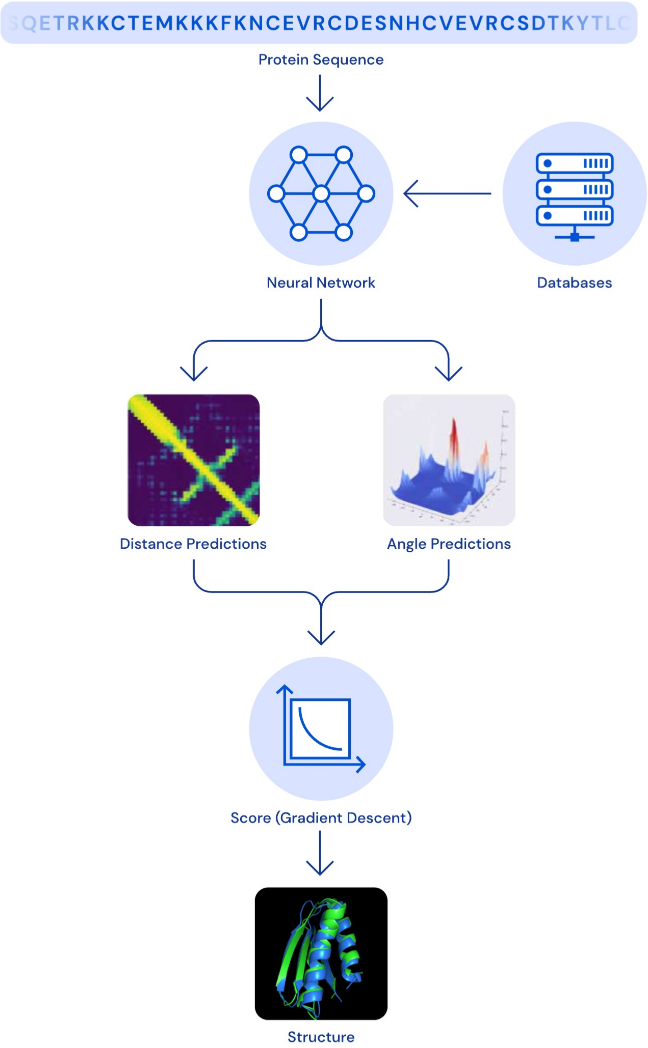
Darker colors indicate a higher attribution weight



Extended Data Fig. 9

Methods

Components



SQETRKKCTEMKKKFKNCEVRCDESNHCVEVRCSDTKYTLC

Protein Sequence

Neural Network

Databases

Distance Predictions

Angle Predictions

Score (Gradient Descent)

Structure

Training data

- Structures: Protein Data Bank (PDB)
- Sequences: Uniclust30
- Training data includes 29,400 data points

MSA features

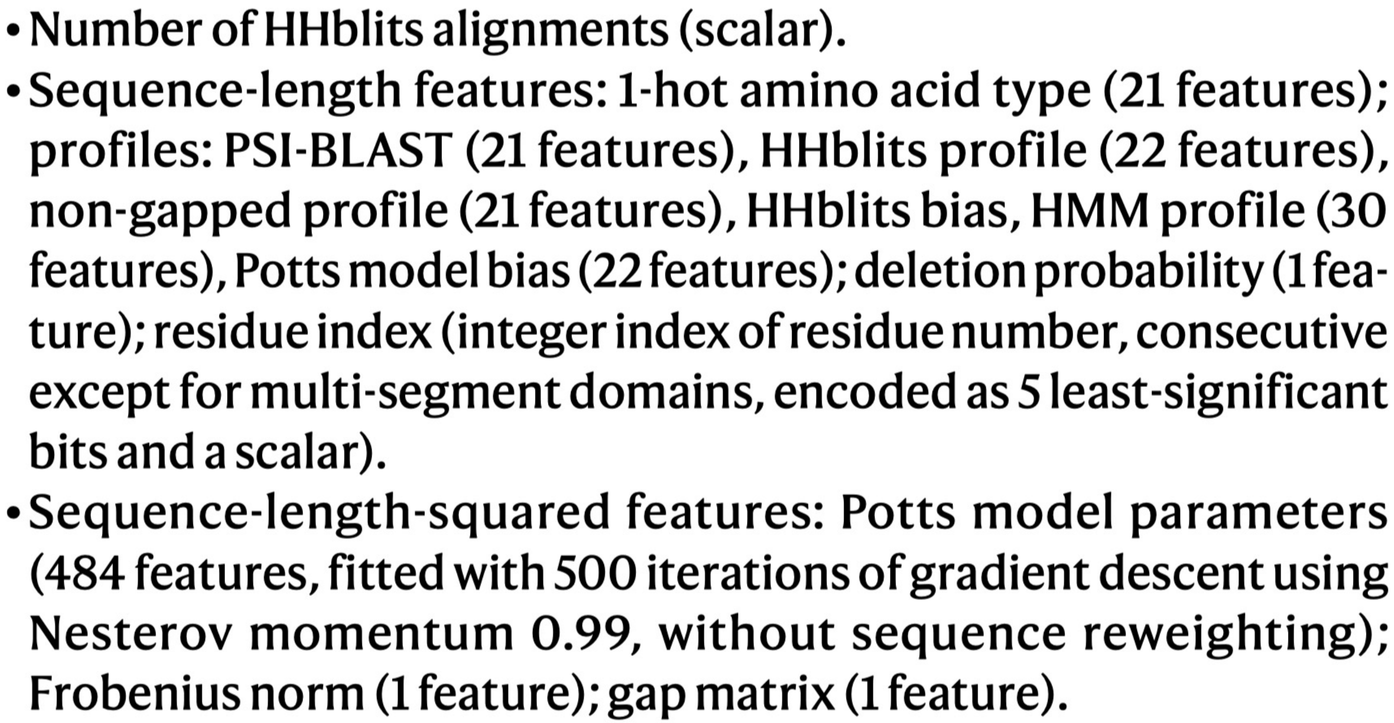
Intuition

- Similar sequence tend to lead to similar 3D structure
- Use coevolutionary features

MSA Features

- Multiple Sequence Alignment (MSA): sequences that are similar to the target sequence
  - Use HHBlits and PSIBLAST profiles to find similar sequences
- Extract features (already used in previous works)
  - 2D features from Potts model fit in TensorFlow
    - Frobenius norm (L x L x 1): covariation between pairs of amino acids
    - Raw parameters (L x L x 22 x 22)

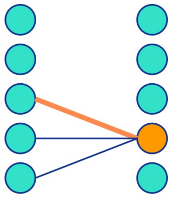
1D input feature for each residue



- 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).

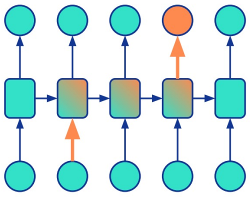
Deep neural network

- Capable of modelling complex data
  - Long range, subtle patterns, with redundancy, needing generalization
  - Structure of the network gives **inductive bias** to certain kinds of modelling
- Inductive bias examples



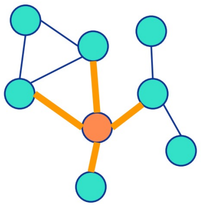
Convolutional Networks (e.g. computer vision)

- data in regular grid
- information flow to local neighbours



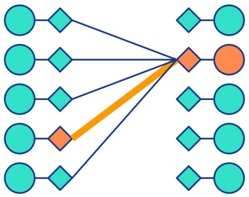
Recurrent Networks (e.g. language)

- data in ordered sequence
- information flow sequentially



Graph Networks (e.g. recommender systems or molecules)

- data in fixed graph structure
- information flow along fixed edges



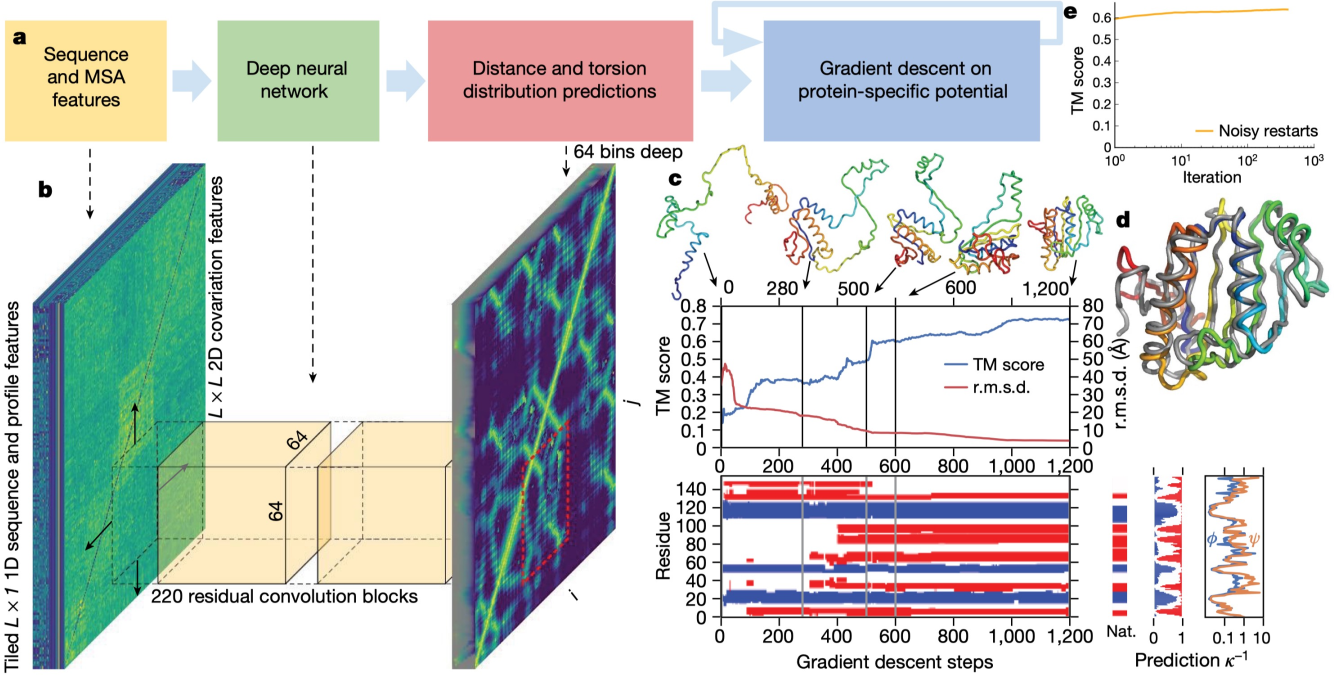
Attention Module (e.g. language)

- data in unordered set
- information flow dynamically controlled by the network (via keys and queries)

2020_12_01_TS_predictor_AlphaFold2.pdf (predictioncenter.org)

Deep distance distribution network

- Central component: convolutional neural network
- Target: Predict distance distributions between pairs of residues of a protein

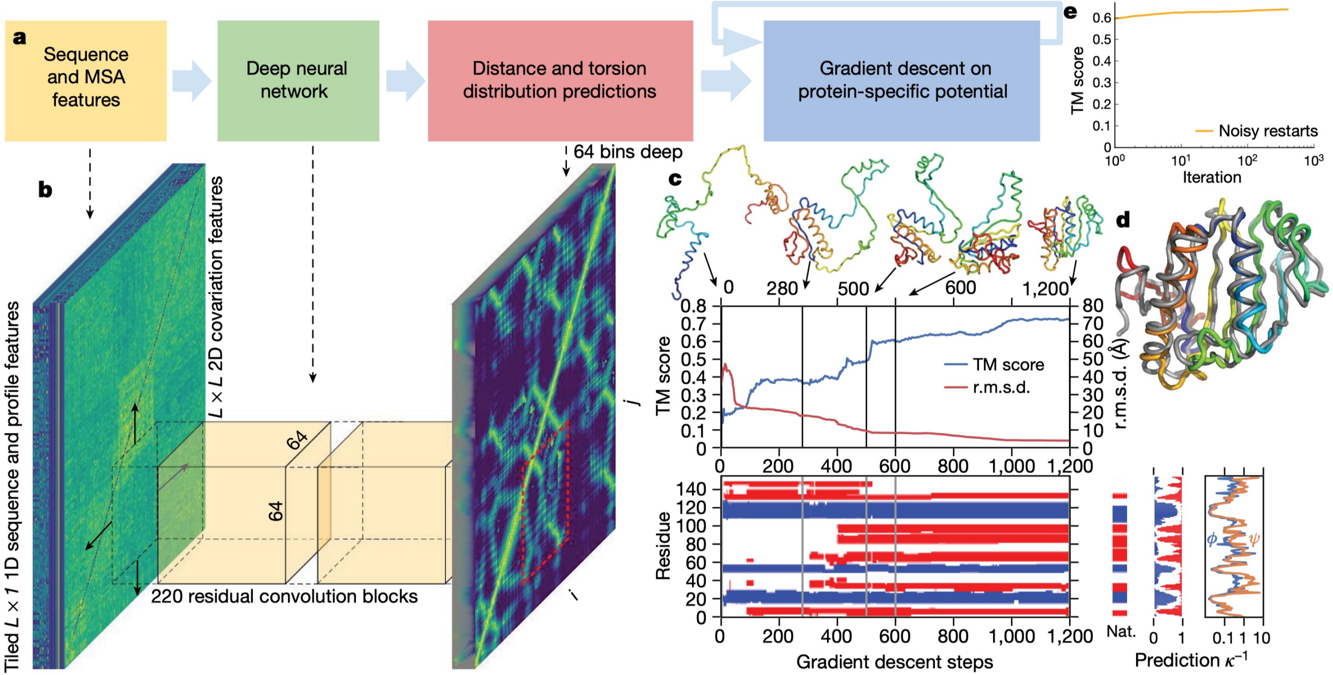

a
Sequence and MSA features
Deep neural network
Distance and torsion distribution predictions
Gradient descent on protein-specific potential
b
Tiled $L \times 1$ 1D sequence and profile features
$L \times L$ 2D covariation features
64
64
220 residual convolution blocks
64 bins deep
j
i
c
0
280
500
600
1,200
TM score
r.m.s.d. (Å)
TM score
r.m.s.d.
Residue
Gradient descent steps
e
TM score
Noisy restarts
Iteration
d
Nat.
Prediction
κ^{-1}
ϕ
ψ

Deep neural network

- Input: two-dimensional array of features, concatenating:
  - The one-dimensional feature for i
  - The one-dimensional feature for j
  - Two-dimensional feature for i, j
- Loss: cross entropy between predicted and ground-truth distance
- Output: softmax probability distribution for each i, j pair
  - Produce a distance histograms -> “distograms”
- Optimization: stochastic gradient descent

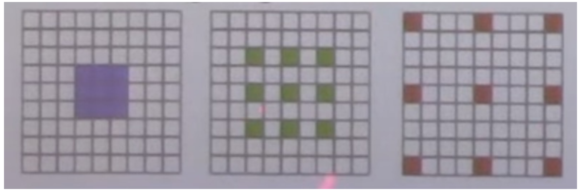
Deep neural network

- Takes in any 64 x 64 region of the entire distance matrix
- Produce 64 bin distance histogram

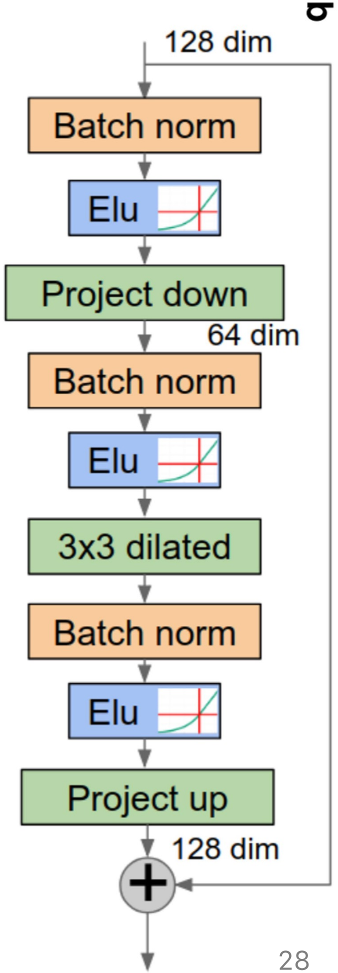


Deep dilated convolutional residual network

- 220 residual blocks: repeat 220 times
- Each residual block consists of a sequence of neural network layers that interleave
  - Three batch norm layers
  - Two 1x1 projection layers
  - A 3x3 dilated convolution layer
  - Exponential linear unit (ELU) nonlinearities
- Dilated convolution
  - 3x3
  - At each stage only look at 9 pixels
- 21 million parameters



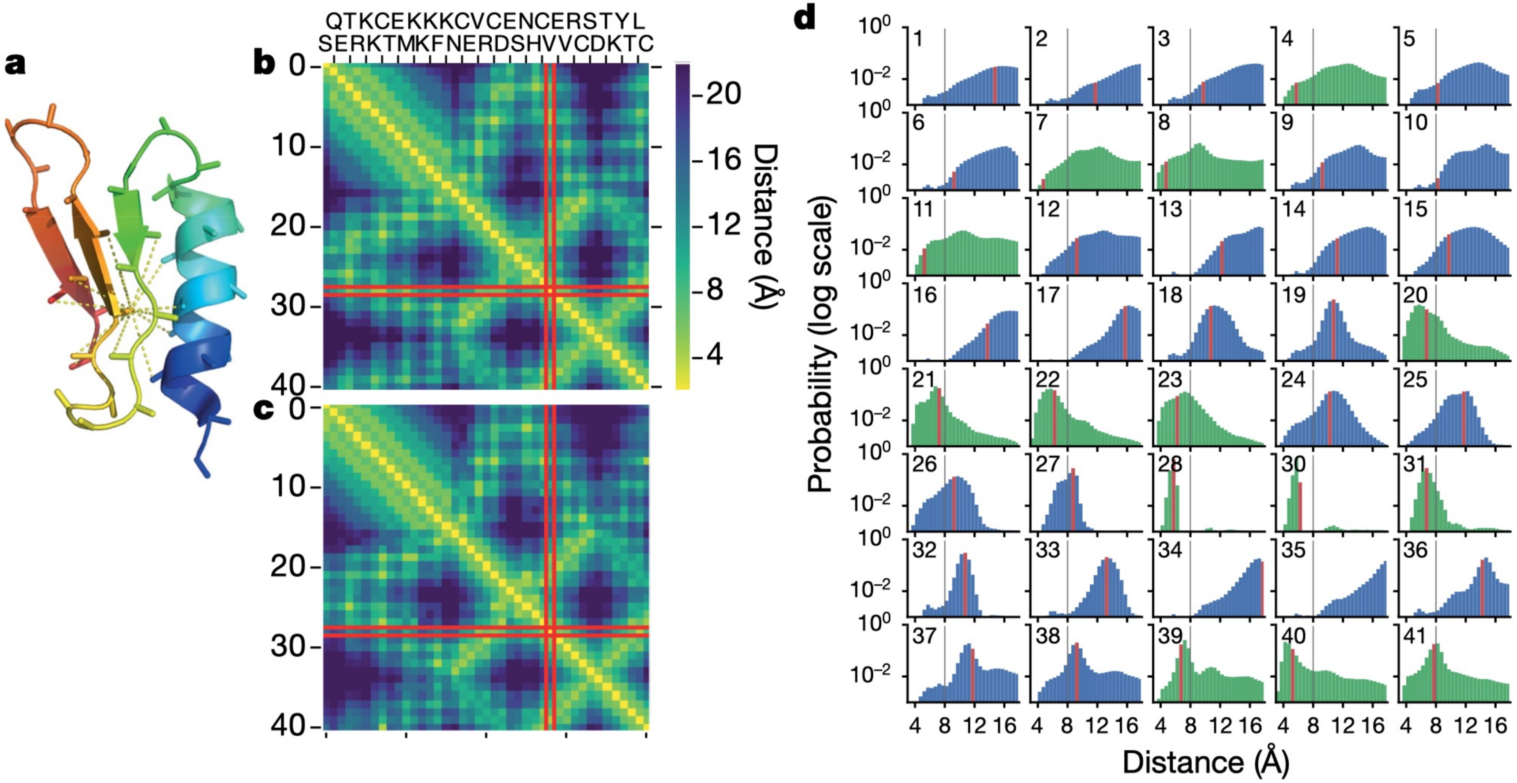
Architecture of a single residual block



Cropping and tiling

- Always training and predicting on a pair of 64 consecutive residues
- Use 64 x 64 crops from the protein’s distance map
  - Consistent size
- Benefit
  - Efficient to train, especially distributed training
  - The model will not have inconsistency between long and short protein prediction
  - Each protein now gives rise to thousands of training examples -> helps avoid overfitting by data augmentation
- At test time
  - Average of all different versions of tiling

Distance and torsion prediction result



CASP protein T0955

b: ground-truth

c: prediction

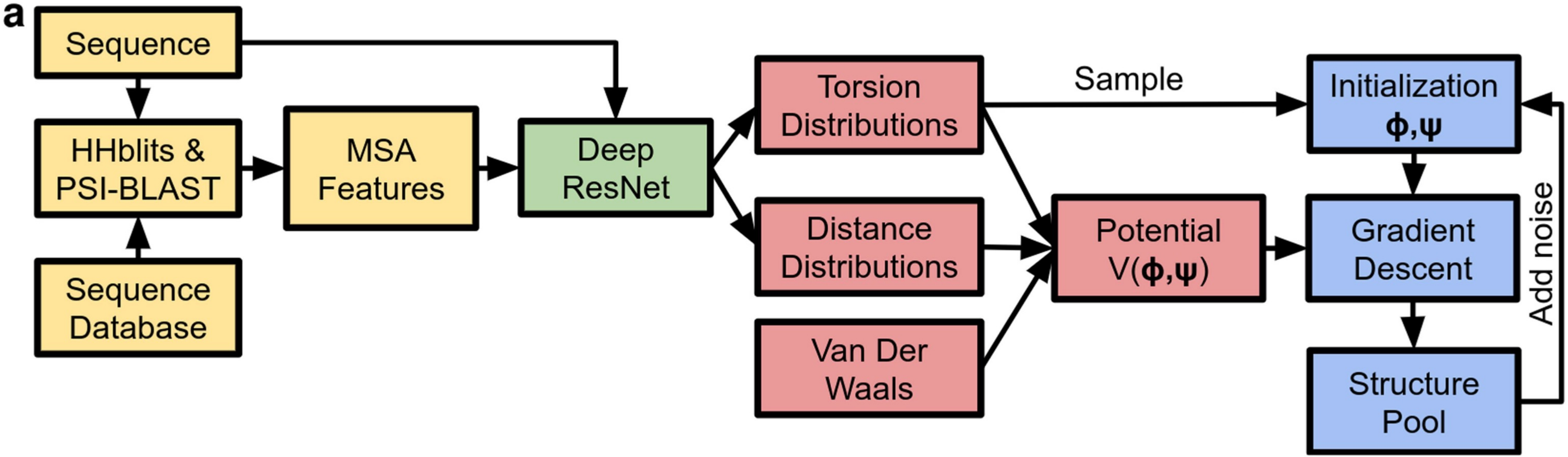
d: predicted distance to residue 29

Red: ground-truth value

Green: consider as a contact

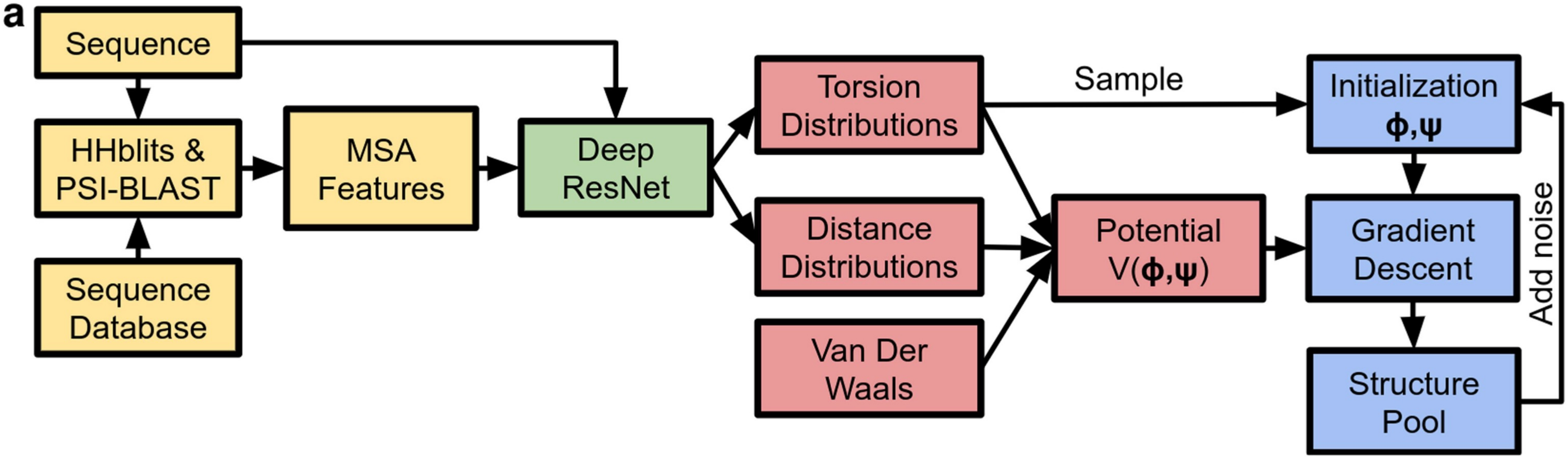
Two distributions to potential

- Reference potential: distance distribution given length, independent of sequence
- Distance potential: negative log likelihood of the distances, summed over all pairs
- Torsion potential: negative log likelihood of the torsion predicted
- Add Van Der Walls term to prevent steric clashes



Structure realization by gradient descent

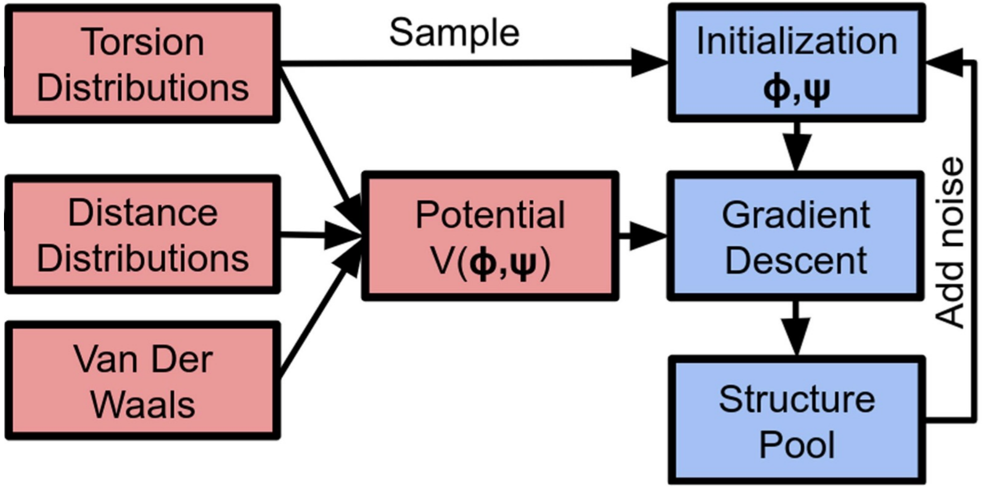
- Function between torsion angles to atom coordinates
- Target: minimize potential (the sum of the distance, torsion and score2_smooth)



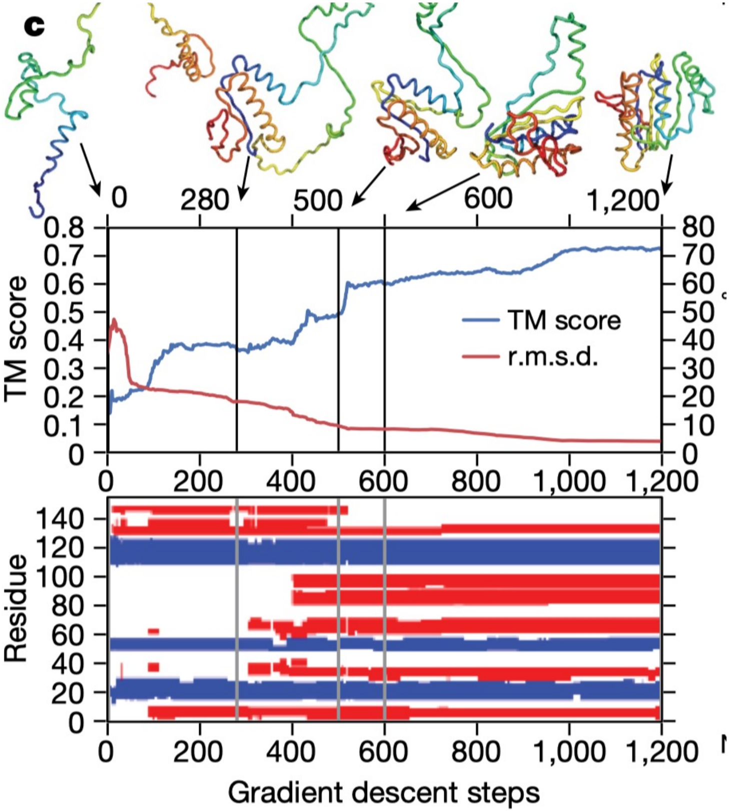
Structure realization by gradient descent

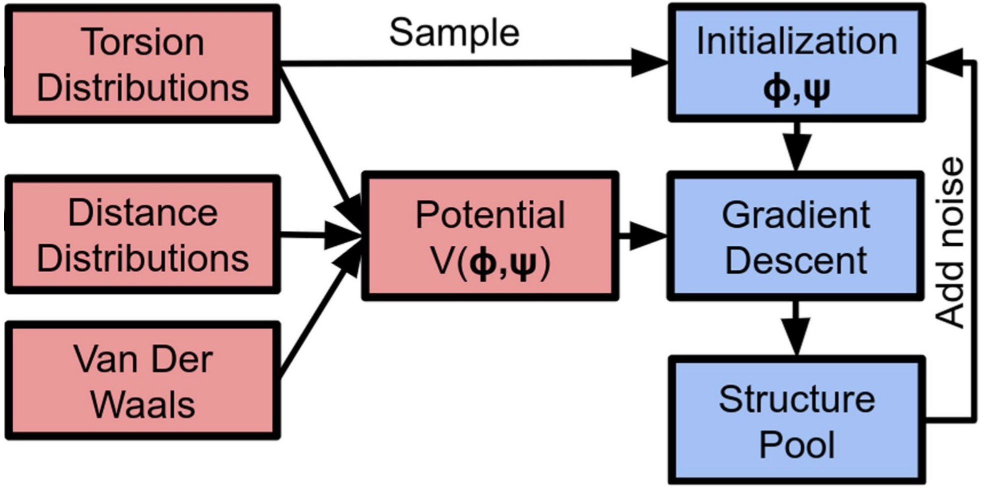
Training process

- Repeated the optimization from sampled initialization
- Produce low-potential structures
- Then sample from the low-potential structure pools as new set of initialization to optimize
- After a few hundred cycles, the optimization converges and the lowest potential structure is chosen as the best candidate structure to output



Structure realization by gradient descent





Evaluation

How to evaluate the prediction?

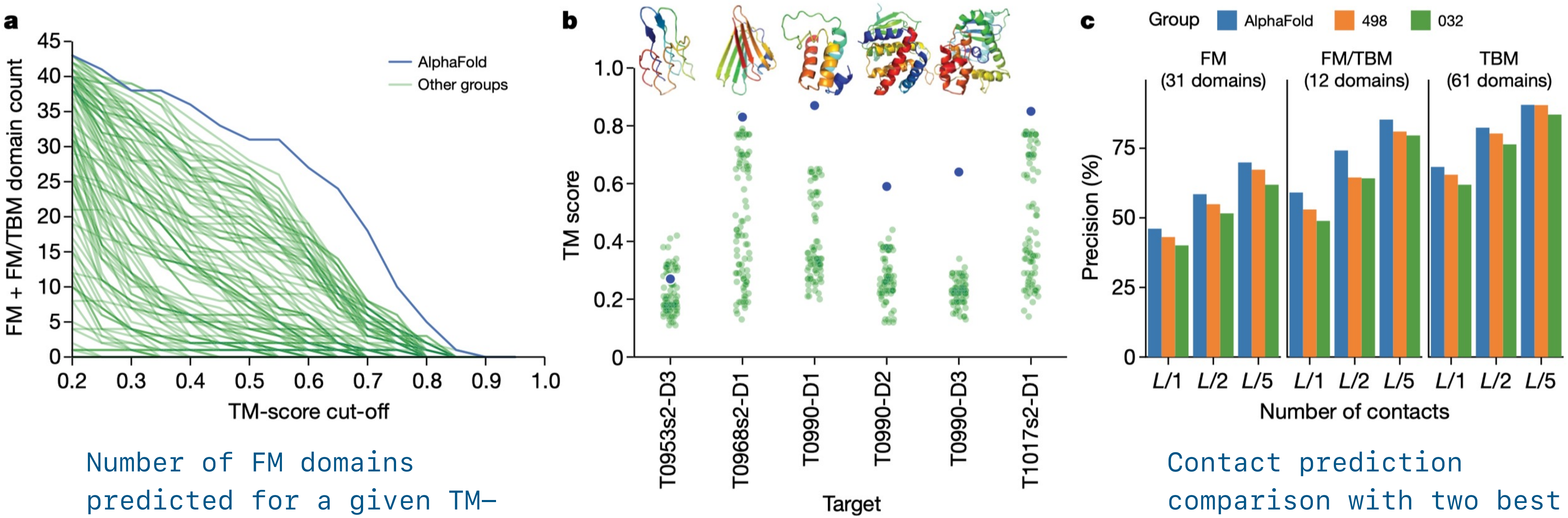
- Assessors divided the proteins into 104 domains for scoring and classified each as:
  - Being amenable to template based modeling (TBM)
    - Protein with a similar sequence has a known structure, and that homologous structure is modified in accordance with the sequence differences
  - Requiring free modeling (FM)
    - No homologous structure is available
  - FM/TBM: intermediate category
- For each domain, use accuracy, precision

How to evaluate the prediction?

- Compare the final structure to the experimentally determined structures
  - TM score
  - GDT_TS (Global distance test, total score)
- Alternative accuracy without requiring geometry alignment
  - IDDT: percentage of native pairwise distances
  - Distogram IDDT (DLDDT)

AlphaFold in the CASP13 assessment

- AlphaFold predicts more FM domains with high accuracy than any other system



Number of FM domains predicted for a given TM-score threshold

Contact prediction comparison with two best ranked methods

Fig. 1

Ablation study and the effect of number of bins

- Number of bins does not need to very large to have good performance
- Change construction of the potential
  - Distance prediction is the primary contribution for the potential
  - Removing torsion potential, reference correction or score2_smooth degrades the accuracy only slightly
- Adding Rosetta relax (side chain packing) is slightly helpful

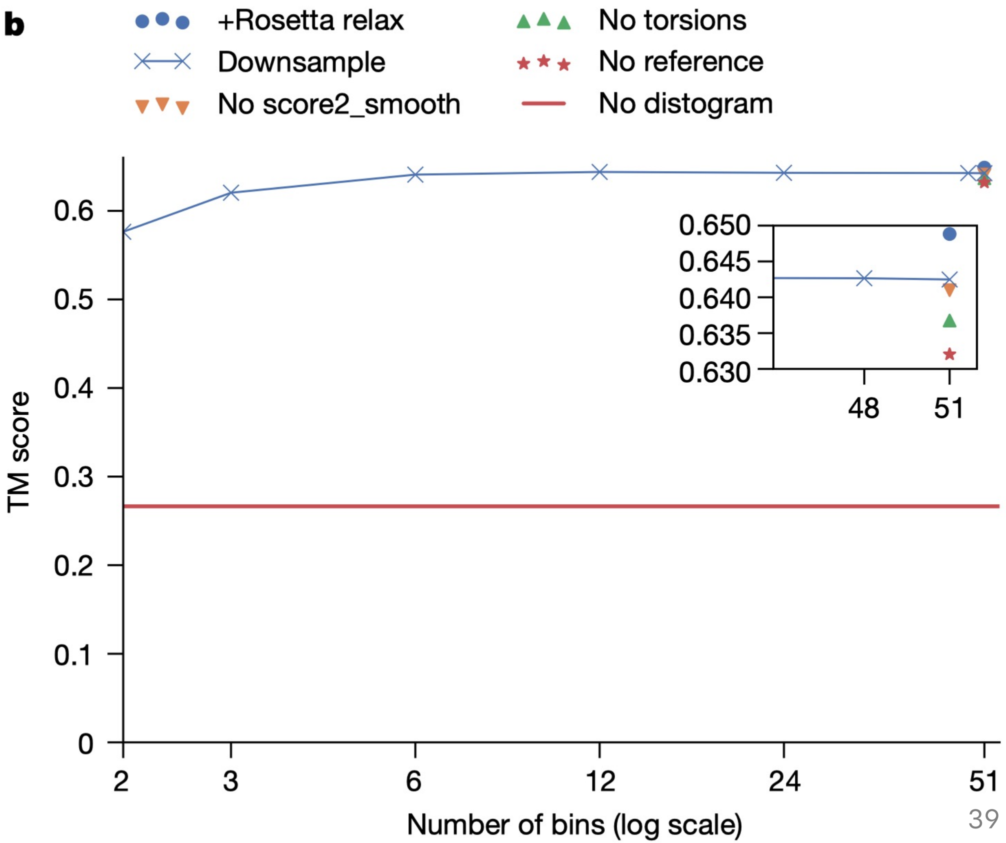
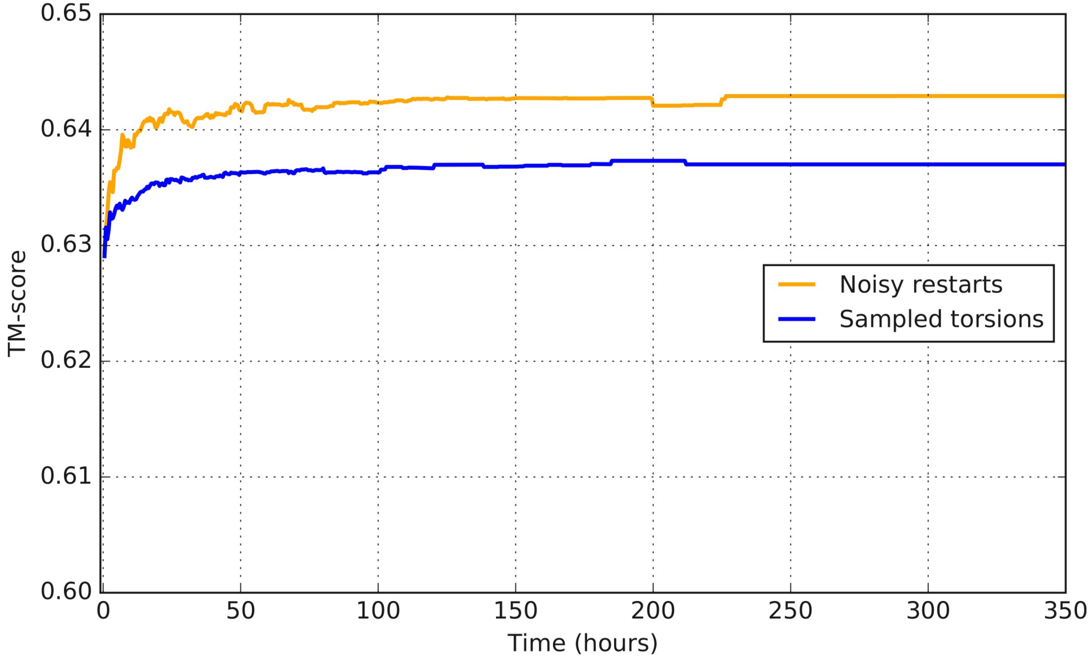


Fig 4b

Computation time for structure realization

- Around 100 node hours are enough
- Noisy restarts is helpful



Extended Data Fig. 4

Limitation

Limitation

- Experiments show the method misses some templates with huge margin
- Not directly predict side chains, the side chain prediction is replied on external tools
- Not use existing templates, and solved protein structure
- Interpretability and robustness of the model
- Heavily rely on MSA features and similar sequence

Conclusion

Conclusion

- AlphaFold represents a considerable advance in protein-structure prediction
- Train a neural network to make accurate predictions of the distances between pairs of residues, which convey more information about the structure than contact predictions
- Construct a potential of mean force that can accurately describe the shape of a protein
- Resulting potential can be optimized by a simple gradient descent algorithm
- The resulting system achieves high accuracy, even for sequences with fewer homologous sequences

Thanks!