AlphaFold

Improved protein structure prediction using potentials from deep learning

Report

Vinayak Gupta 25th June 2021

Introduction

Protein folding is one of the holy grail problems in biology. Predicting protein structure from the amino acids chain was one of the biggest challenges and it was always thought that AI should be able to solve it.

They trained a neural network to predict the distance between the residues and the torsional angles from the MSA(multiple sequence alignment) features of the protein. Using the distance and the torsional angles, using just simple gradient descent they were able to achieve state-of-the-art performance.

AlphaFold

Using the sequence database from the Protein Data Bank, we create the MSA features for the corresponding amino acid sequence. Using the MSA features it would be easy to find the contact positions, because if there is a change at 2 places in the amino acid sequence, then we can guarantee that there will be a contact position between those 2 positions

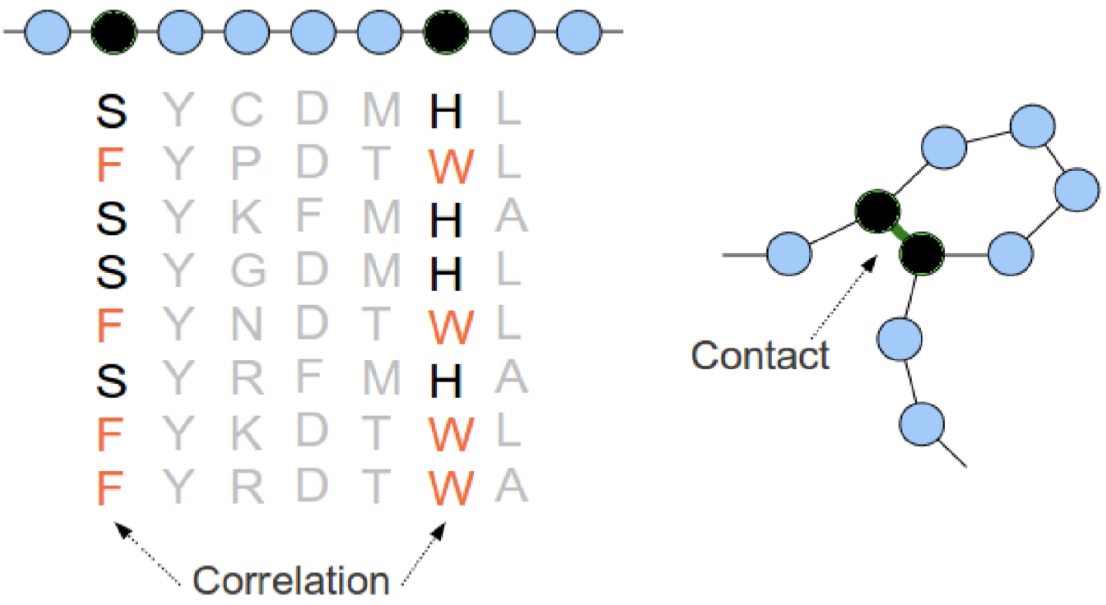
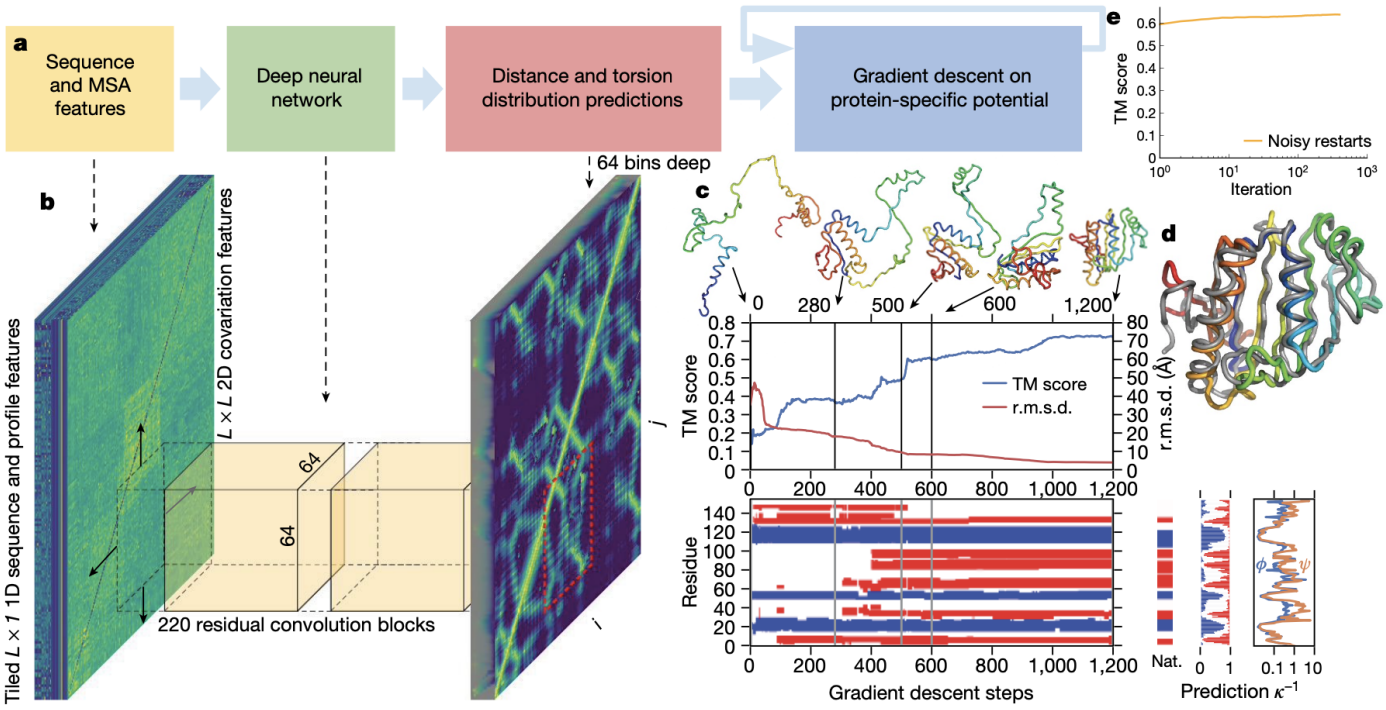


FIG. 1. (Color online) Left panel: small MSA with two positions of correlated amino-acid occupancy. Right panel: hypothetical corresponding spatial conformation, bringing the two correlated positions into direct contact.

Using the MSA features, we use deep neural networks to learn the distance matrix and the torsional angles(phi and psi). To avoid memory overload and overfitting we predict these features only using a 64x64 window like as shown in the figure below. It is also shown that contact predictions need only a small local window.

To find the distance matrix for all the LxL residue pairs, we combine these individual 64x64 maps. We use 220 residual blocks with dilated convolutions to predict the features. We use ELU non-linearity and stochastic gradient descent



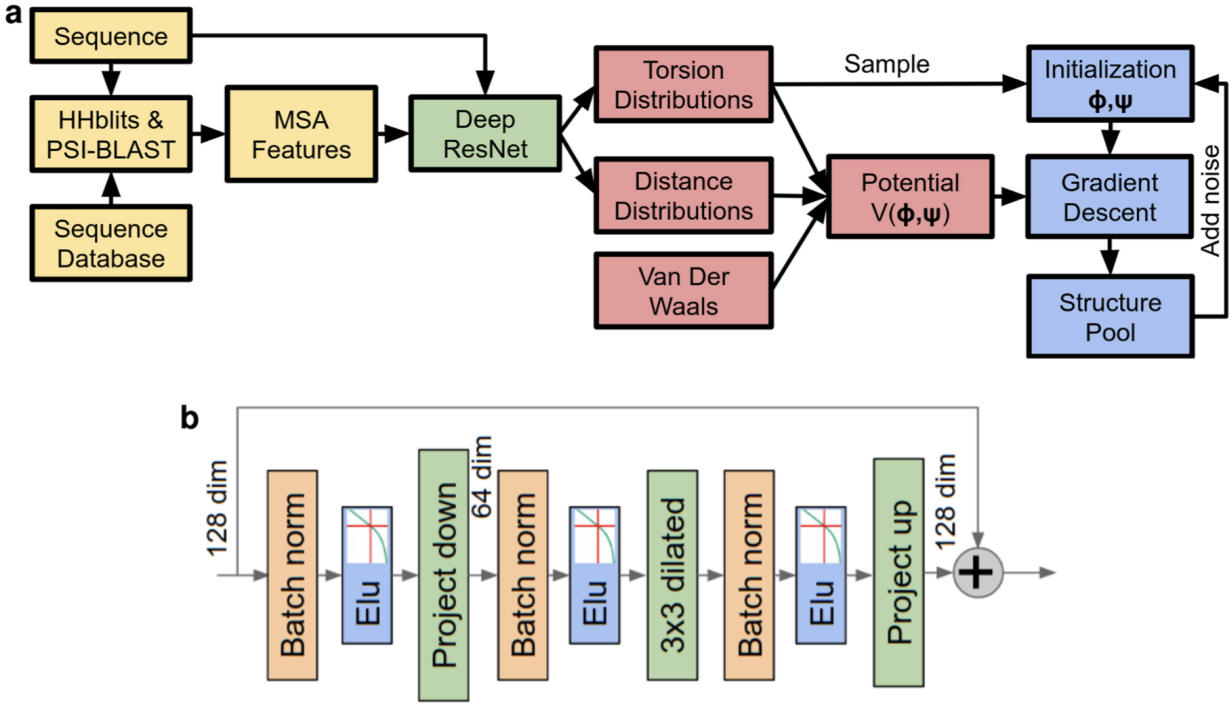
After predicting the distance matrix and the torsional distributions, now it's time to predict the 3d protein structure. This is done by minimizing the potential using gradient descent. They define 3 different types of potentials: Distance Potential, Torsional Potential and Vanderwaal Potential.

The distance potential is the negative log-likelihood of the distance summed over all the residue pairs. The torsional potential is also modelled as negative log-likelihood. Since we predict the marginal distribution, we fitted a unimodal von Mises distribution to the marginal predictions. Also taking into account the steric repulsions between the amino acids, we add a vanderwaal's term to take care of that

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

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

Now after defining the potentials, they initialise a 3d protein structure by sampling from the torsion distribution. They run gradient descent(L-BFGS) algorithm and they obtain the 3d protein structure



To generalise and to increase randomness in the dataset, we use different offsets of 64x64 map from the distogram which in turn creates new training data. Also after predicting the final 3d structure we add a noise to the coordinates and run gradient descent again. This creates randomness to the training set and hence reduces overfitting

Accuracy Methods

To score the final 3d model, they have used different metrics like TM score, GDT(Global Distance Test) and RMS Distance. These require geometric alignment. So they used a new metric called IDDT

To know more about AlphaFold, check out the paper by DeepMind: AlphaFold

To know more about the implementation of AlphaFold, check out their GitHub Repository Code