Boulder

Computational Biology

Maziar Raissi

Assistant Professor

Department of Applied Mathematics

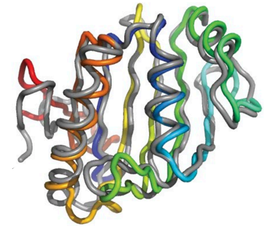
University of Colorado Boulder

maziar.raissi@colorado.edu

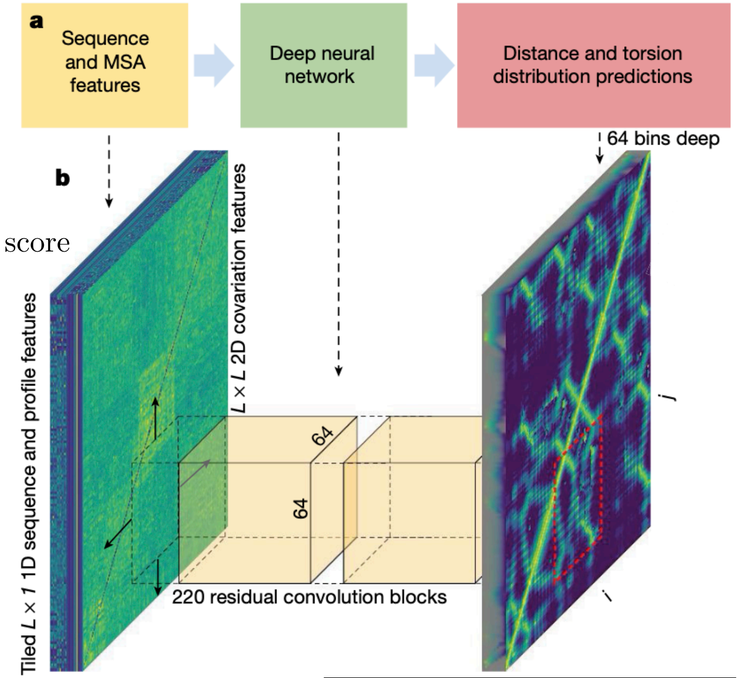
Improved Protein Structure Prediction using Potentials from Deep Learning

Protein folding problem: determine the three-dimensional shape of a protein from its amino acid sequence (21 amino acids)
SQETRKKCTEMKKKFKNCEVRCDESNHCVEVRCSDTKYTLC

Protein Data Bank (PDB)
https://www.rcsb.org/3d-view/5W9F
Critical Assessment of Protein Structure Prediction (CASP13)
$S = (s_1, \ldots, s_L) \rightarrow$ amino acid sequence of a protein
$s_i \rightarrow i$-th residue
$\text{MSA}(S) \rightarrow$ multiple sequence alignment features (HHblits & PSI-BLAST)
The input to the network consists of a two-dimensional array of features in which each i, j feature is the concatenation of the one-dimensional features for both i and j as well as the two-dimensional features for i, j.
$P(\varphi_i, \psi_i | S, \text{MSA}(S)) \rightarrow$ discrete probability distributions of backbone torsion angles
(Neural Network; backbone torsion angles: parametrize protein structure)
$x_i \rightarrow$ coordinates for residue i

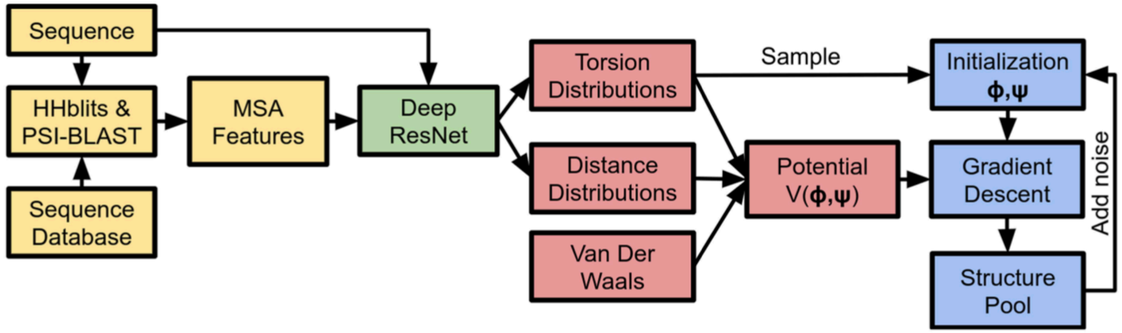


Template Modelling (TM) score



$x = G(\varphi, \psi) \rightarrow$ build a differentiable model G (Neural Network) of protein geometry
$d_{ij} = \|x_i - x_j\| \rightarrow$ inter-residue distances
$P(d_{ij} | S, \text{MSA}(S)) \rightarrow$ discrete probability distribution for every ij pair
(Neural Network)

To constrain memory usage and avoid overfitting, the network was always trained and tested on 64 × 64 regions of the distance matrix.



Potentials

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

$$V_{\text{distance}}(\mathbf{x}) = -\sum_{i,j,\ i \neq j} \log P(d_{ij} \mid \mathcal{S}, \text{MSA}(\mathcal{S})) - \log P(d_{ij} \mid \text{length}, \delta_{\alpha\beta})$$
(reference; glycine (C_α atom) or not (C_β))

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

$$V_{\text{total}}(\boldsymbol{\phi}, \boldsymbol{\psi}) = V_{\text{distance}}(G(\boldsymbol{\phi}, \boldsymbol{\psi})) + V_{\text{torsion}}(\boldsymbol{\phi}, \boldsymbol{\psi}) + V_{\texttt{score2\_smooth}}(G(\boldsymbol{\phi}, \boldsymbol{\psi}))$$
(a van der Waals term)

Senior, Andrew W., et al. "Improved protein structure prediction using potentials from deep learning." *Nature* 577.7792 (2020): 706-710.

Boulder

Questions?