Boulder

Computational Biology

Maziar Raissi

Assistant Professor

Department of Applied Mathematics

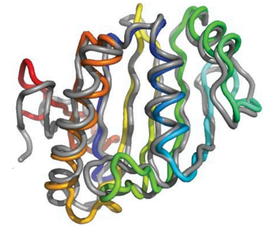
University of Colorado Boulder

maziar.raissi@colorado.edu

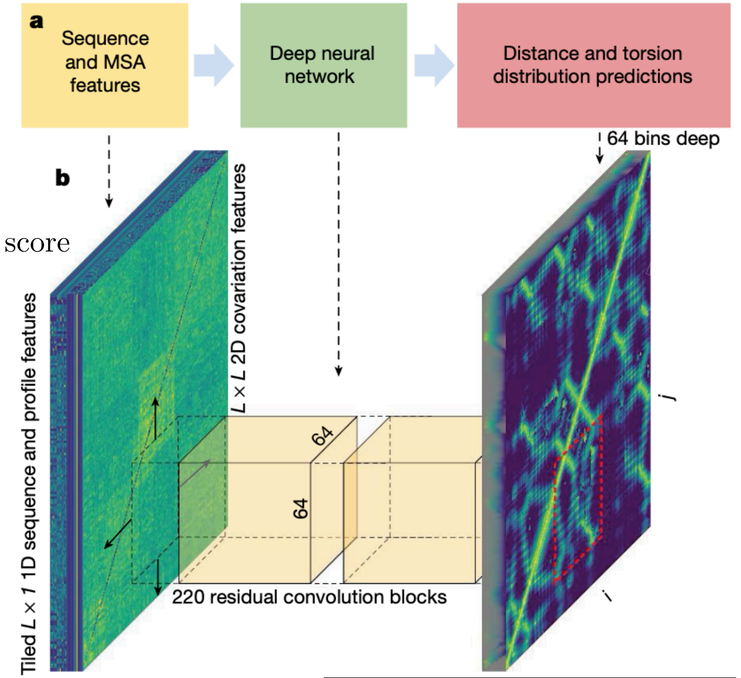
Improved Protein Structure Prediction using Potentials from Deep Learning

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

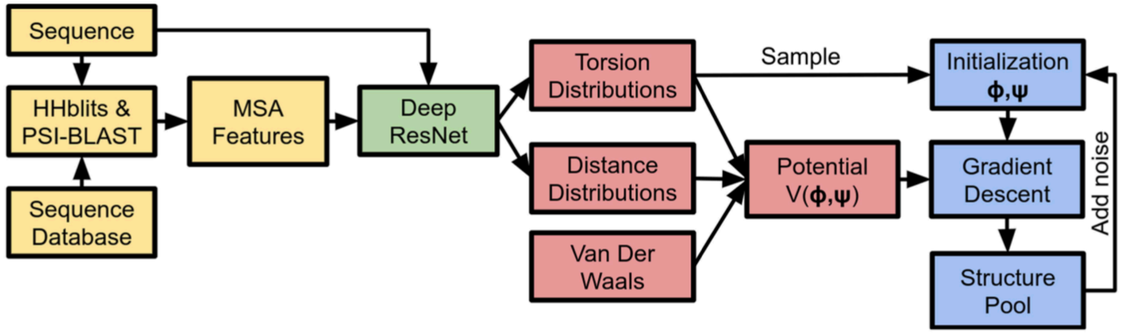


Template Modelling (TM) score



$x = G(\varphi, \psi) \rightarrow$ 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)

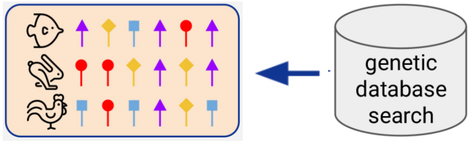
Senior, Andrew W., et al. "Improved protein structure prediction using potentials from deep learning." *Nature* 577.7792 (2020): 706-710.

Highly Accurate Protein Structure Prediction with AlphaFold

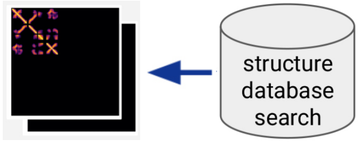
AlphaFold receives input features derived from the amino-acid sequence, MSA, and templates and outputs features including atom coordinates, the distogram, and per-residue confidence scores.

 → amino-acid sequence

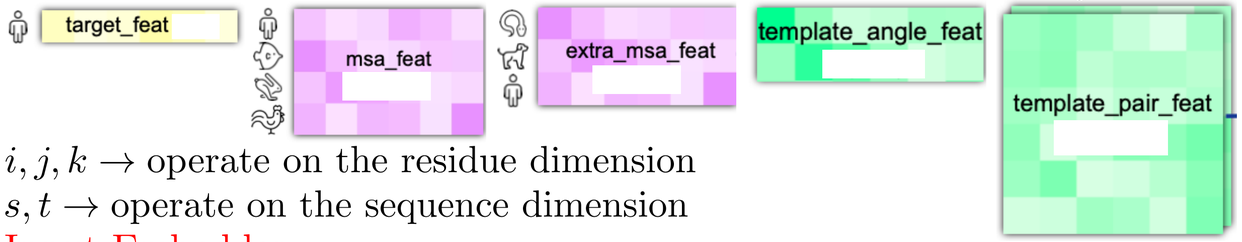
MSA (multiple sequence alignment): sequences of evolutionary related proteins



Templates: 3D atom coordinates of a small number of homologous structures



$\{\mathbf{f}_i^{\text{target\_feat}}\}, \{\mathbf{f}_{si}^{\text{msa\_feat}}\}, \{\mathbf{f}_{si}^{\text{extra\_msa\_feat}}\}, \{\mathbf{f}_{si}^{\text{template\_angle\_feat}}\}, \{\mathbf{f}_{sij}^{\text{template\_pair\_feat}}\}$



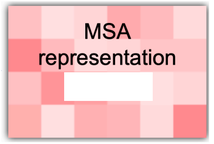
$i, j, k \rightarrow$ operate on the residue dimension

$s, t \rightarrow$ operate on the sequence dimension

Input Embedder

$\{\mathbf{f}_i^{\text{target\_feat}}\}, \{\mathbf{f}_{si}^{\text{msa\_feat}}\} \mapsto \{\mathbf{m}_{si}\}, \{\mathbf{z}_{ij}\}$

$\{\mathbf{m}_{si}\} \rightarrow$ MSA representation $\{\mathbf{z}_{ij}\} \rightarrow$ pair representation



Embed Templates

$\{\mathbf{f}_{si}^{\text{template\_angle\_feat}}\}, \{\mathbf{f}_{sij}^{\text{template\_pair\_feat}}\}, \{\mathbf{z}_{ij}\} \mapsto \{\mathbf{z}_{ij}\}$

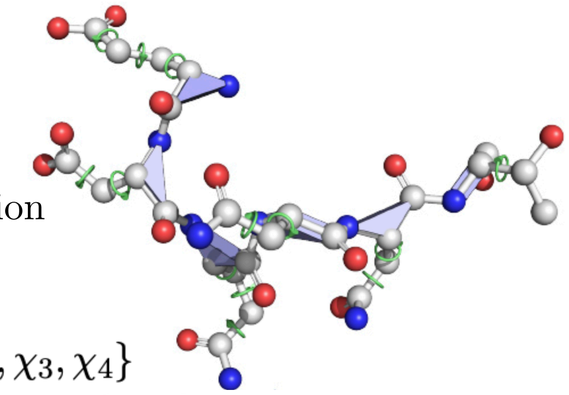
Embed extra MSA features

$\{\mathbf{f}_{si}^{\text{extra\_msa\_feat}}\}, \{\mathbf{z}_{ij}\} \mapsto \{\mathbf{z}_{ij}\}$

Evoformer

$\{\mathbf{m}_{si}\}, \{\mathbf{z}_{ij}\} \mapsto \{\mathbf{m}_{si}\}, \{\mathbf{z}_{ij}\}, \{\mathbf{s}_i\}$

$\{\mathbf{s}_i\} \rightarrow$ abstract single representation



Structure Module

$\{\mathbf{z}_{ij}\}, \{\mathbf{s}_i\} \mapsto \{T_i^f\}, \{\mathbf{x}_i^a\}$

$f \in \mathcal{S}_{\text{torsion names}} = \{\omega, \phi, \psi, \chi_1, \chi_2, \chi_3, \chi_4\}$

$a \in \mathcal{S}_{\text{atom names}} = \{\text{N}, \text{C}^\alpha, \text{C}, \text{O}, \text{C}^\beta, \text{C}^\gamma, \text{C}^{\gamma 1}, \text{C}^{\gamma 2}, \dots\}$

$T_i = (R_i, \mathbf{t}_i) \rightarrow$ backbone frames $\quad \mathbf{x}_i^a \rightarrow$ atom coordinates

$\vec{\mathbf{x}}_{\text{global}} = T_i \circ \vec{\mathbf{x}}_{\text{local}} = R_i \vec{\mathbf{x}}_{\text{local}} + \vec{\mathbf{t}}_i$

Frame Aligned Point Error (FAPE)

def computeFAPE($\{T_i\}, \{\vec{\mathbf{x}}_j\}, \{T_i^{\text{true}}\}, \{\vec{\mathbf{x}}_j^{\text{true}}\}, Z = 10\text{Å}, d_{\text{clamp}} = 10\text{Å}, \epsilon = 10^{-4}\text{Å}^2$) :

$T_i, T_i^{\text{true}} \in (\mathbb{R}^{3\times 3}, \mathbb{R}^3)$

$\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_j^{\text{true}} \in \mathbb{R}^3,$

$i \in \{1, \dots, N_{\text{frames}}\}, j \in \{1, \dots, N_{\text{atoms}}\}$

Chirality property!

1: $\vec{\mathbf{x}}_{ij} = T_i^{-1} \circ \vec{\mathbf{x}}_j$ $\quad \vec{\mathbf{x}}_{ij} \in \mathbb{R}^3$

2: $\vec{\mathbf{x}}_{ij}^{\text{true}} = T_i^{\text{true}\,-1} \circ \vec{\mathbf{x}}_j^{\text{true}}$ $\quad \vec{\mathbf{x}}_{ij}^{\text{true}} \in \mathbb{R}^3$

3: $d_{ij} = \sqrt{\|\vec{\mathbf{x}}_{ij} - \vec{\mathbf{x}}_{ij}^{\text{true}}\|^2 + \epsilon}$ $\quad d_{ij} \in \mathbb{R}$

4: $\mathcal{L}_{\text{FAPE}} = \frac{1}{Z} \text{mean}_{i,j}(\text{minimum}(d_{\text{clamp}}, d_{ij}))$

Self-distillation procedure, (BERT)-style objective, iterative refinement ("recycling"), predict model confidence, and structural violations penalty.

Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." *Nature* (2021): 1-11.

Boulder

Questions?