Fast uncovering of protein sequence diversity from structure

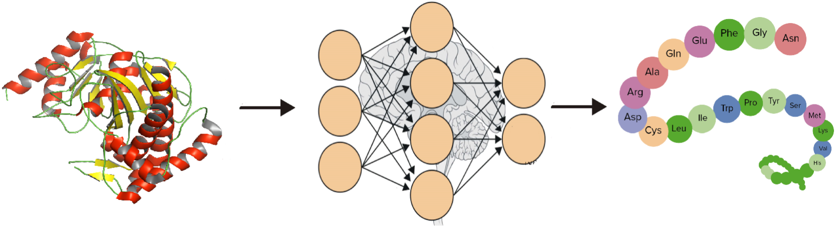
Luca Alessandro Silva, Barthelemy Meynard-Piganeau
Carlo Lucibello, Christoph Feinauer

Bocconi University, Genbio.AI

April 1, 2025

Protein inverse folding

Protein inverse folding: given a 3D structure X, find σ folding into X



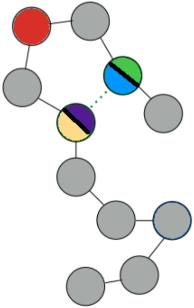
- Given the task's complexity, recently deep-learning methods emerged.
- Models such as ESM-IF1, Protein-MPNN though only map a $X \rightarrow \sigma_X$.

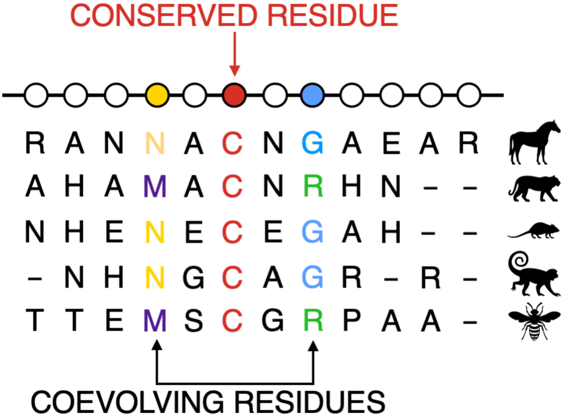
Sequence-structure data imbalance

- Determined sequences ~ 200 millions, determined structures $\sim 200k \implies 0.1\%$ of sequences have a determined structure.
- Mapping $X \to \sigma_X$ focuses on a very small part of sequences.
- Many proteins come from a common ancestor $\implies$ their sequences have some very conserved regions.
- We can cluster such sequences $\implies$ protein families.
- Such sequences evolve constrained by they function and hence structure.
- Crucial *many-to-one* nature of inverse folding

Protein Families

Function/structure conservation constraint evolution of **homologues** ⟹ statistical patterns in MSA.





CONSERVED RESIDUE

R A N N A C N G A E A R

A H A M A C N R H N – –

N H E N E C E G A H – –

– N H N G C A G R – R –

T T E M S C G R P A A –

COEVOLVING RESIDUES

Pairwise models

- MSA has long-range correlations $\Longrightarrow$ need a **global model**

- Probability of a sequence σ_i in a MSA is

$$p(\sigma_i|J,h) = \frac{1}{Z(J,h)} \exp\Big\{ -\Big[\sum_{i<j} J_{ij}(\sigma_i,\sigma_j) + \sum_i h_i(\sigma_i) \Big] \Big\}$$

- *Couplings* $J \in \mathbb{R}^{L\times q\times L\times q}$. $J_{i,j}(a,b)$ describe propensity of amino acids a, b to co-appear at position i, j.
- *Fields* $h \in \mathbb{R}^{L\times q}$. $h_i(a)$ describe the marginal propensity of an amino acid a to appear at position a.
- Potts models can replicate **first and second order** patterns of MSA

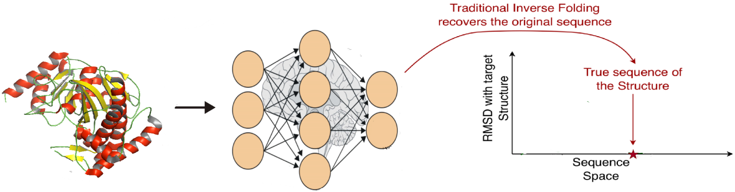
InvMSAFold architecture

- We propose a novel architecture, **InvMSAFold**, which forces the model to learn this variability

- **InvMSAFold** outputs the parameters of a light-weight low-rank approximation of the couplings J

  $$J_{i,j}[a,b] = \frac{1}{\sqrt{K}} V_i[a]^\top v_j[b],$$

  trained on the *augmented* dataset (X, M_X).



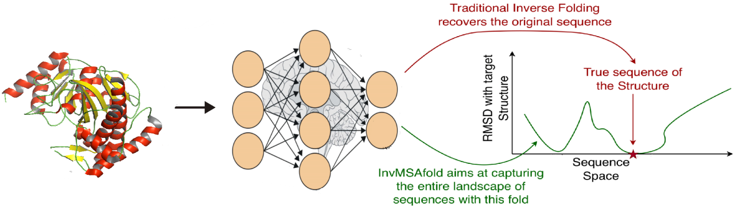
Traditional Inverse Folding recovers the original sequence
True sequence of the Structure
RMSD with target Structure
Sequence Space

InvMSAFold architecture

- We propose a novel architecture, **InvMSAFold**, which forces the model to learn this variability
- **InvMSAFold** outputs the parameters of a light-weight low-rank approximation of the couplings J

$$J_{i,j}[a,b] = \frac{1}{\sqrt{K}} V_i[a]^\top v_j[b],$$

trained on the *augmented* dataset (X, M_X).



Traditional Inverse Folding recovers the original sequence

True sequence of the Structure

InvMSAfold aims at capturing the entire landscape of sequences with this fold

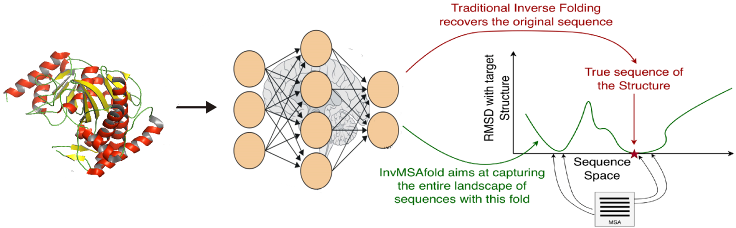
InvMSAFold architecture

- We propose a novel architecture, **InvMSAFold**, which forces the model to learn this variability

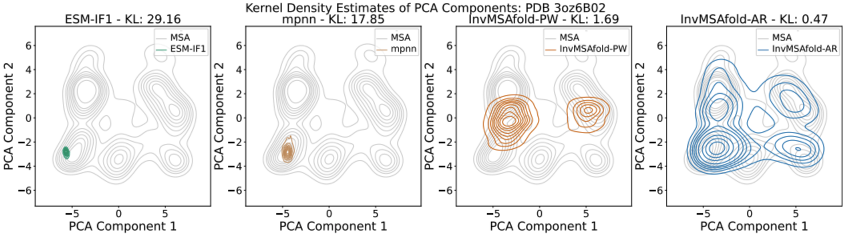
- **InvMSAFold** outputs the parameters of a light-weight low-rank approximation of the couplings J

$$J_{i,j}[a,b] = \frac{1}{\sqrt{K}} V_i[a]^\top v_j[b],$$

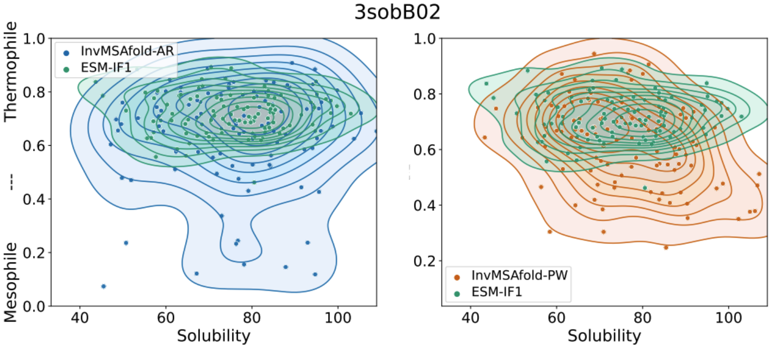
trained on the *augmented* dataset (X, M_X).


Traditional Inverse Folding recovers the original sequence
True sequence of the Structure
RMSD with target Structure
Sequence Space
InvMSAfold aims at capturing the entire landscape of sequences with this fold
MSA

PCA plots


Kernel Density Estimates of PCA Components: PDB 3oz6B02
ESM-IF1 - KL: 29.16
mpnn - KL: 17.85
InvMSAfold-PW - KL: 1.69
InvMSAfold-AR - KL: 0.47
MSA
ESM-IF1
mpnn
InvMSAfold-PW
InvMSAfold-AR
PCA Component 1
PCA Component 2
−6
−4
−2
0
2
4
6
−5
0
5

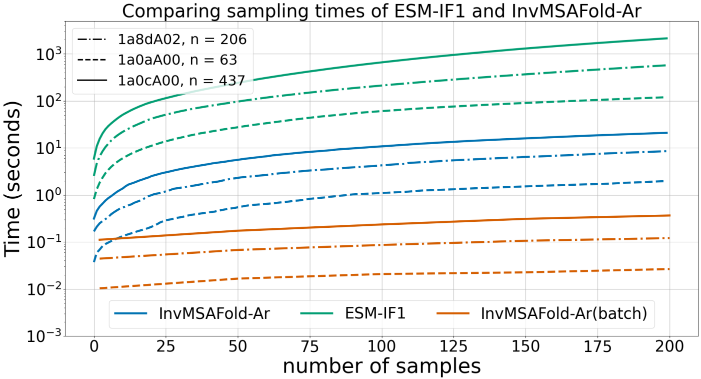
Protein property sampling



Sampling speed at generation

ESM-IF1, Protein-MPNN can be very expensive at inference

- Need $\Theta(L)$ passes through the transformer
- Inv-MSAFold-AR needs just **one**!



Thank you for you attention!

- Paper available at:
  `https://openreview.net/forum?id=1iuaxjssVp`
- Hope to see you all at the poster presentation:
  Fri 25 Apr 7 p.m. PDT 9:30 p.m. PDT!