S-PLM: Structure-aware Protein Language Model via Contrastive Learning between Sequence and Structure

Qing Shao

qshao@uky.edu

Department of Chemical and Materials Engineering

University of Kentucky

https://www.biorxiv.org/content/10.1101/2023.08.06.552203v1.full

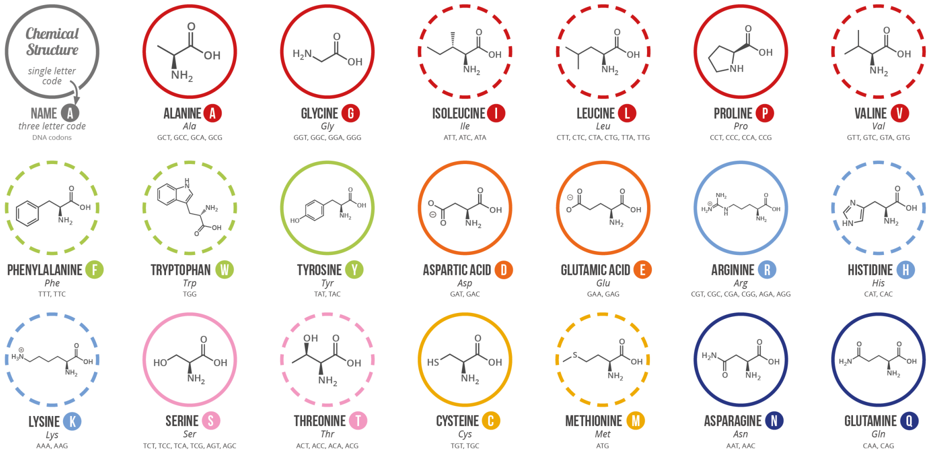
Acknowledgment

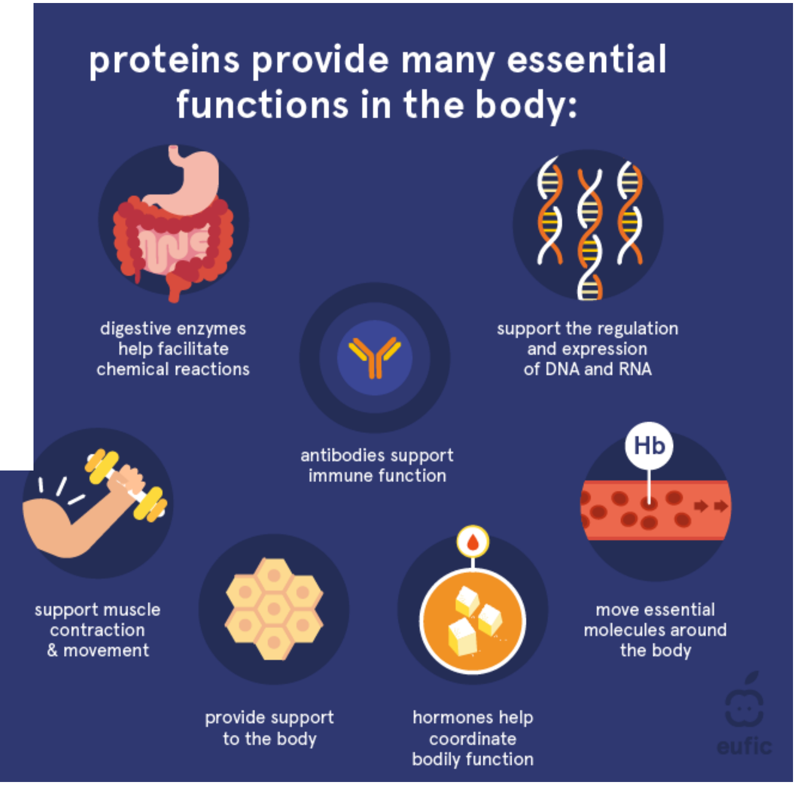
- Dr. Duolin Wang (U. Missouri)
- Dr. Dong Xu (U. Missouri)
- Dr. Jin Chen (Uky, currently U. Alabama)
- Usman Abbas (UKy)

- Funding Support:
- University of Kentucky Start-up Funds
- AI for Medicine Alliance Pilot

Why need protein language models?

- Proteins are biomolecules composed of twenty natural amino acids
- Proteins play a central role in human health




proteins provide many essential functions in the body:
digestive enzymes help facilitate chemical reactions
support the regulation and expression of DNA and RNA
antibodies support immune function
Hb
support muscle contraction & movement
move essential molecules around the body
provide support to the body
hormones help coordinate bodily function
eufic

Chemical structure of 20 amino acids

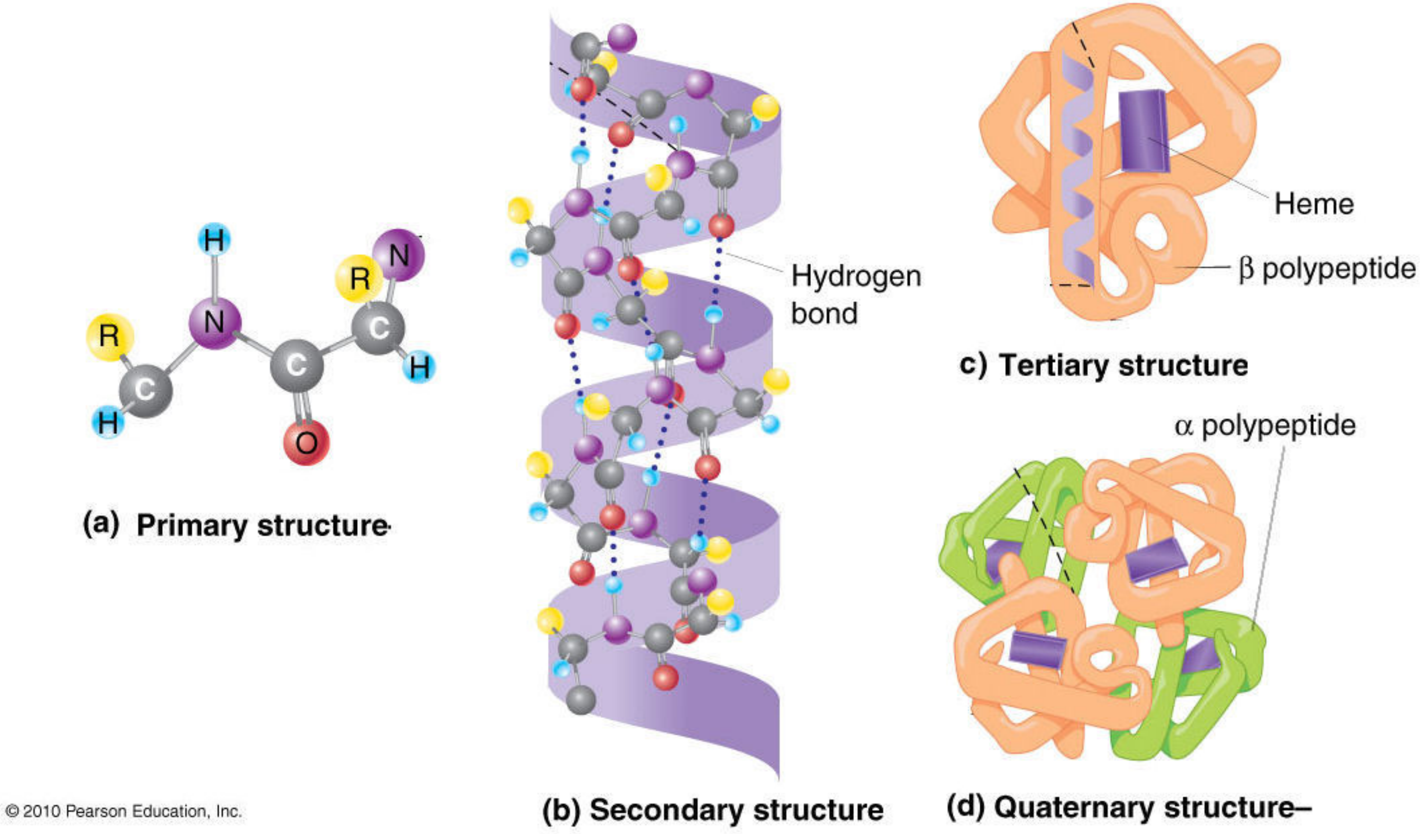
https://www.compoundchem.com/2014/09/16/aminoacids/
https://www.eufic.org/en/whats-in-food/article/what-are-proteins-and-what-is-their-function-in-the-body

Why need protein language models?

- The protein universe is huge.
- (One cell may have ~10 K different types of proteins)
- We need tools to provide reliable information about proteins in a quick way

Why need protein language models?

- Proteins structures determine their properties

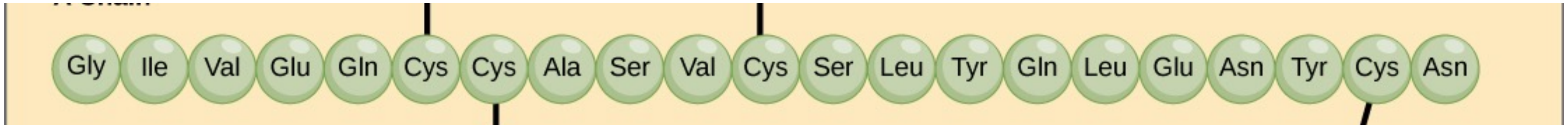


https://www.mun.ca/biology/scarr/iGen3_06-04.html

Why need protein language models?

- It is quite natural to find similarities between the protein primary structure and a sentence

Protein sequence



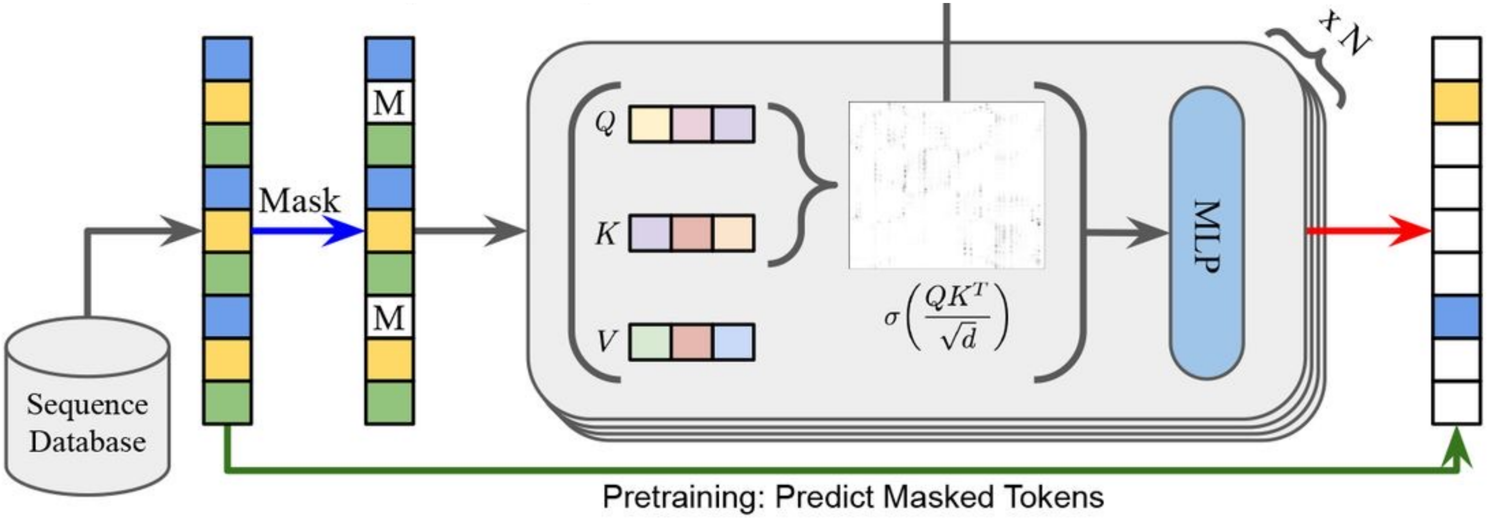
A sentence

I will present at the CCS 2023 summit.

Thus, protein language models can be the tool we are looking for.

https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure

A typical protein language model


Sequence Database
Mask
M
M
Q
K
V
$\sigma\left(\frac{QK^T}{\sqrt{d}}\right)$
MLP
x N
Pretraining: Predict Masked Tokens

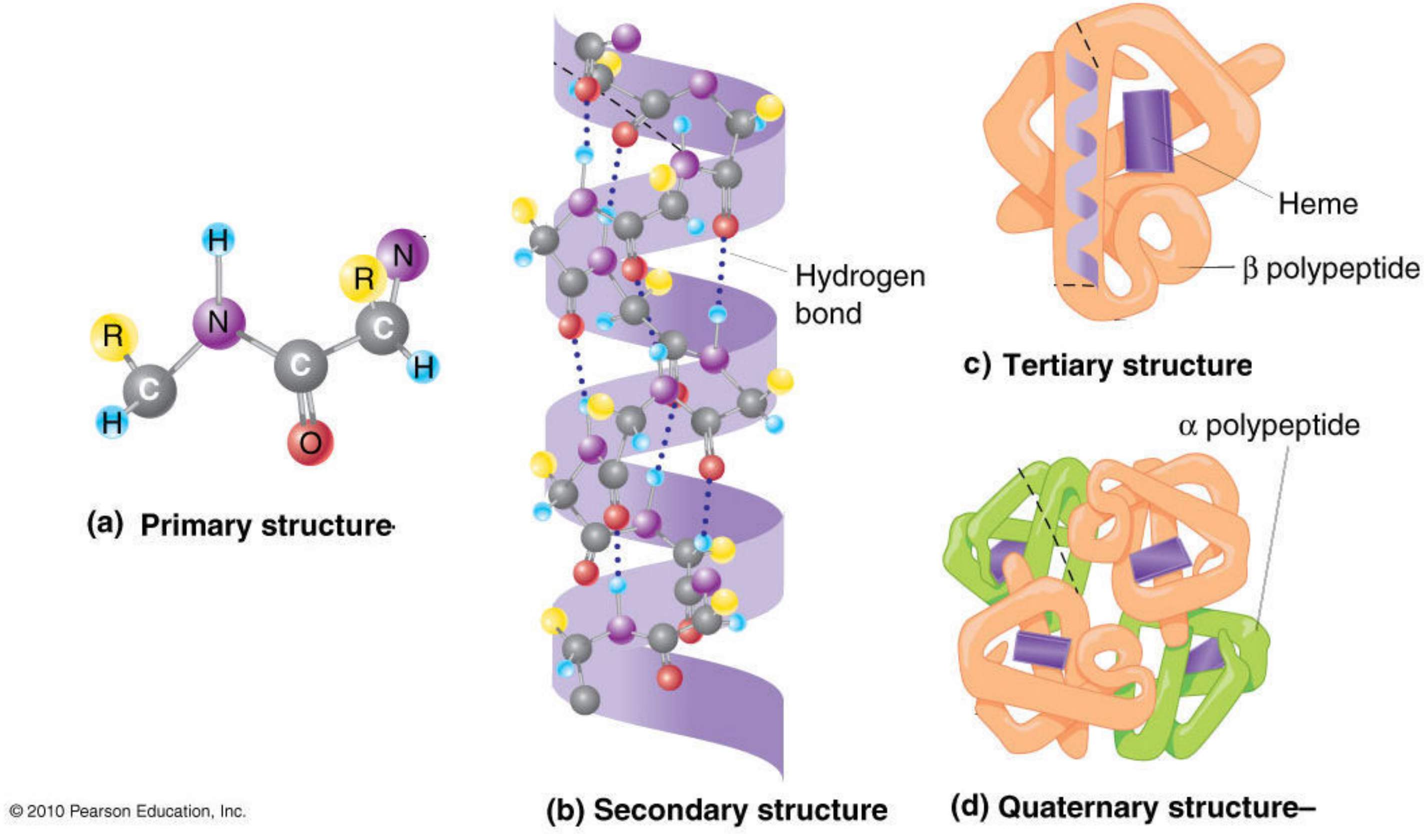
The development of a protein language model is similar to the development of a language model.
One of the most valuable products is the embedding of the language model.
We could use the embedding to develop downstream tasks for protein design and property prediction

What is embedded in the embedding is the key

https://www.biorxiv.org/content/10.1101/2020.12.15.422761v1.full

The current protein language models learn the primary structure

- Proteins structures determine their properties



(a) Primary structure

(b) Secondary structure

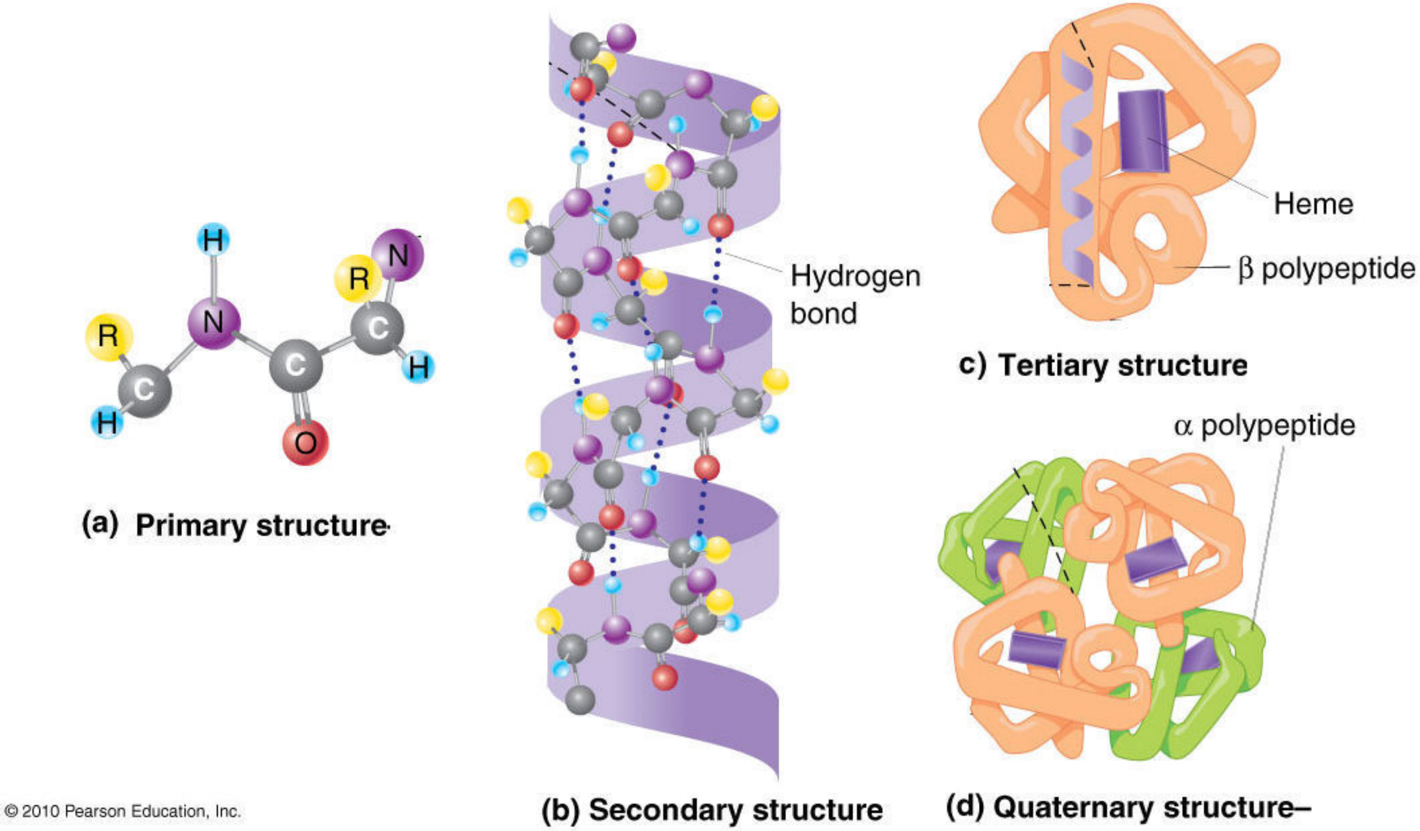
c) Tertiary structure

(d) Quaternary structure

© 2010 Pearson Education, Inc.

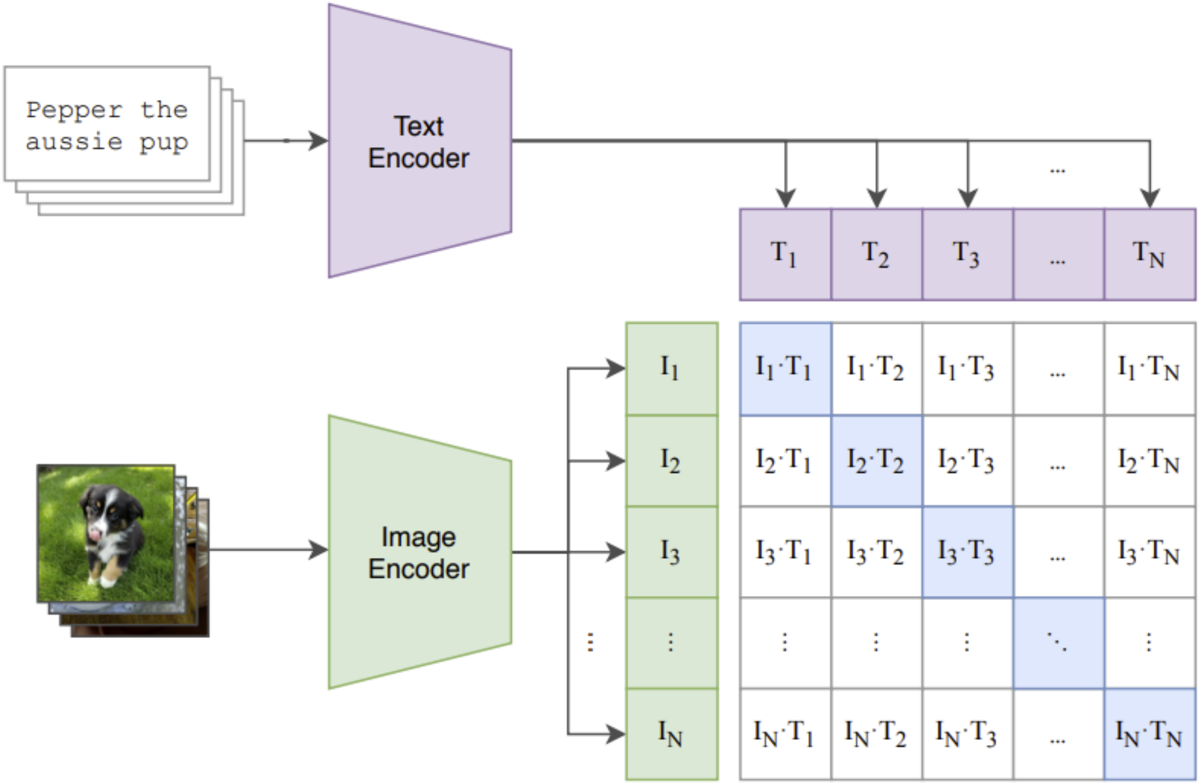
https://www.mun.ca/biology/scarr/iGen3_06-04.html

We must let the protein language model learn the other structures



https://www.mun.ca/biology/scarr/iGen3_06-04.html

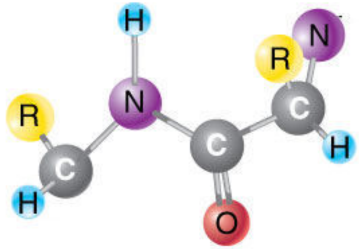
Multiview contrastive learning enables the fusion of information from different sources

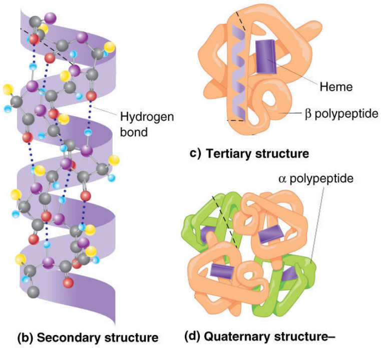


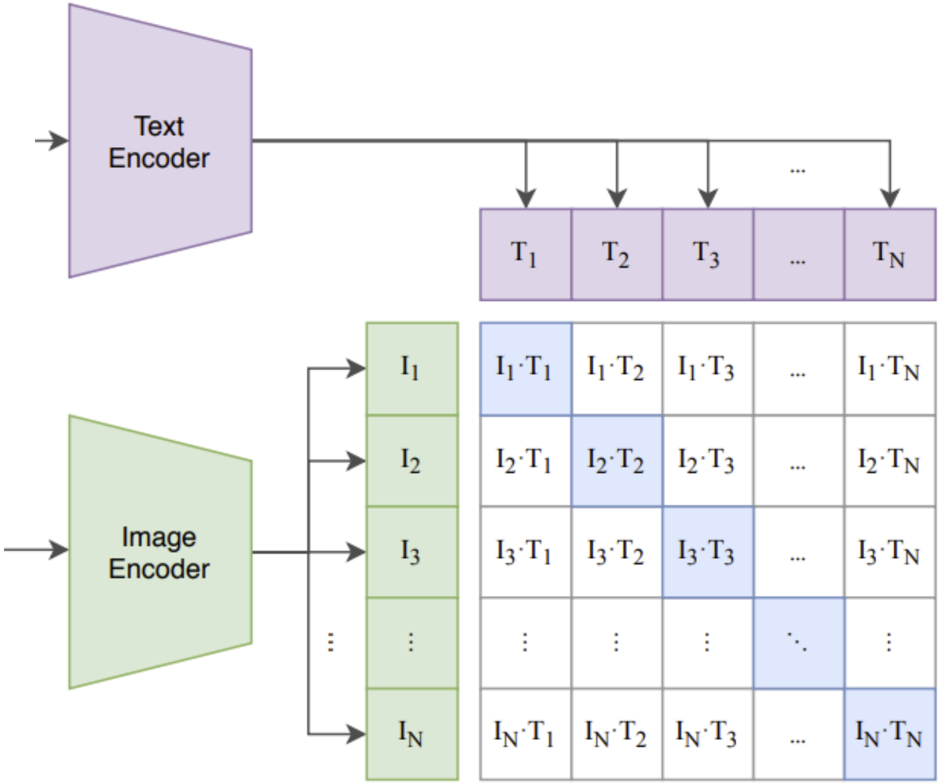
	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Proceedings of the 38th International Conference on Machine Learning, PMLR 139:8748-8763, 2021.

Multiview contrastive learning enables the fusion of information from different sources



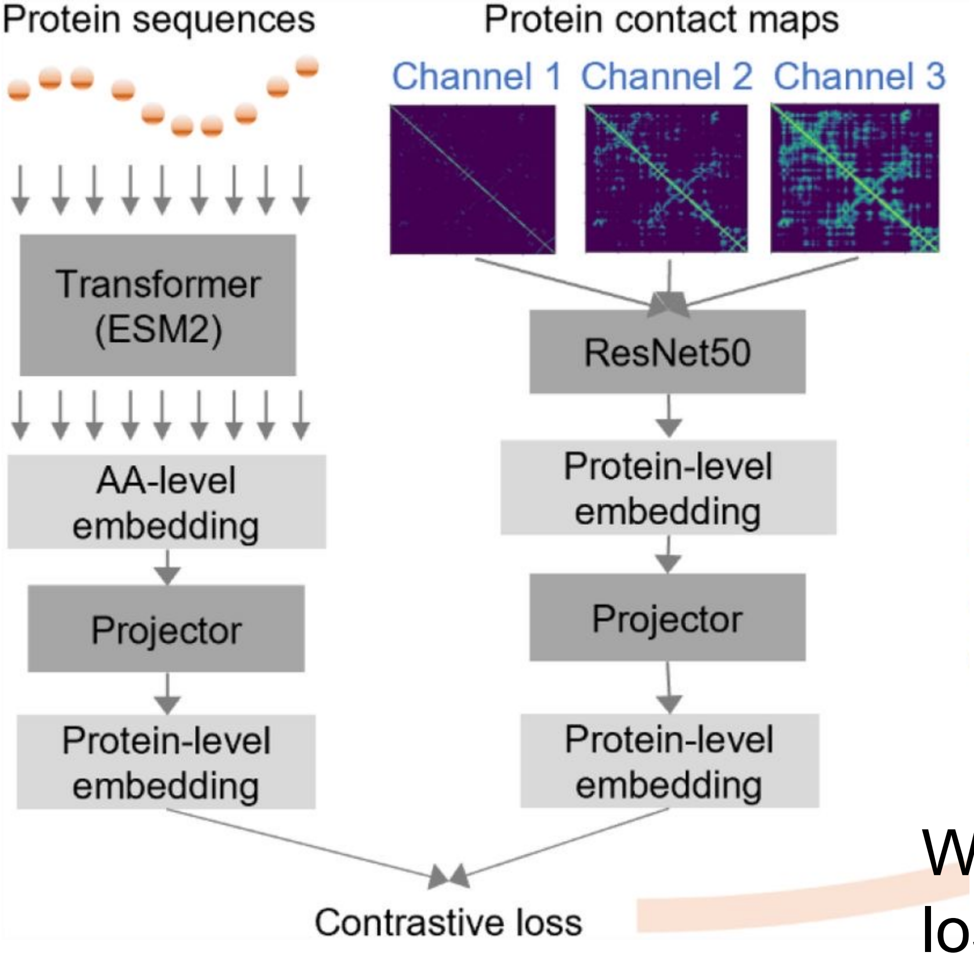




	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Proceedings of the 38th International Conference on Machine Learning, PMLR 139:8748-8763, 2021.

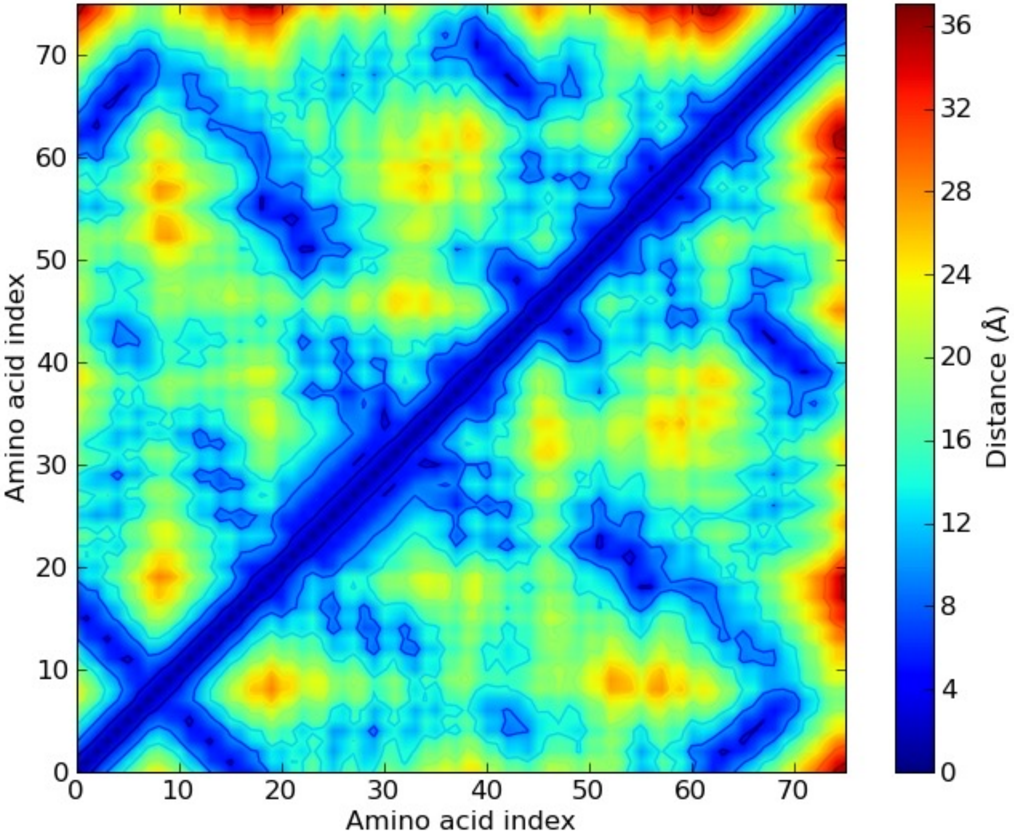
The architecture of our model



We adapt simCLR loss function

https://arxiv.org/abs/2002.05709

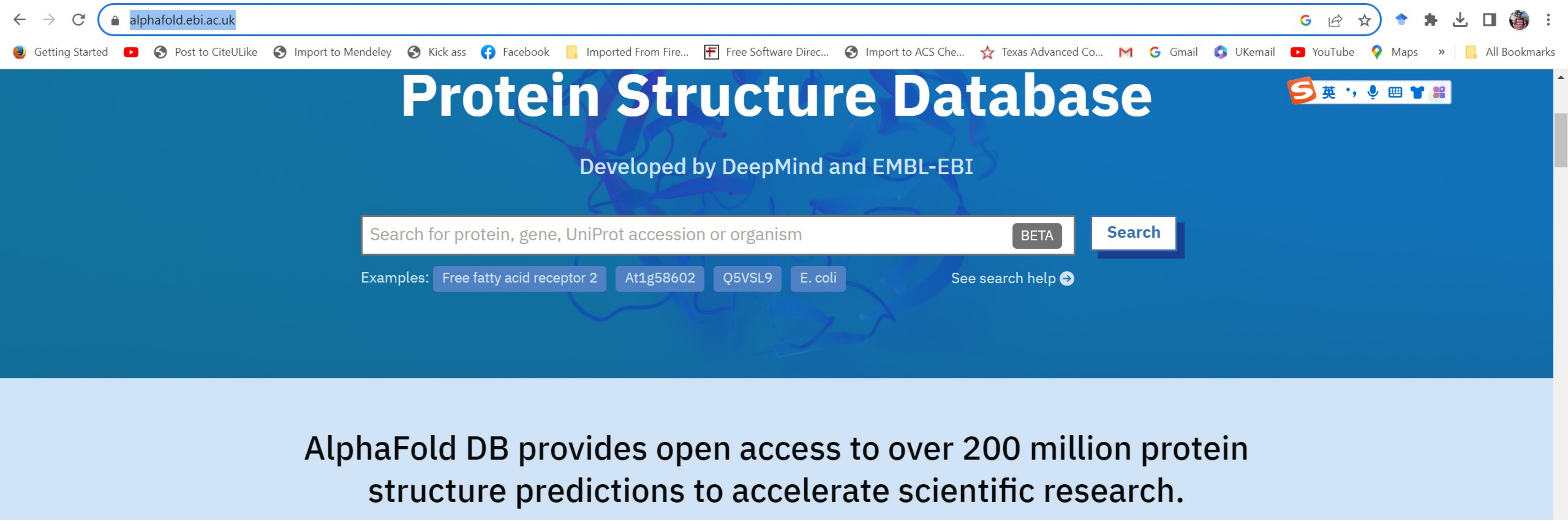
What is contact map?



Contact map is a 2D matrix.
The element (i,j) is the distance between amino acid i and j.
The distance is determined based on the Cα atom

Where to get the protein structure?

- We cannot do our work without the great AlphaFold2



Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism
BETA
Search
Examples: Free fatty acid receptor 2
At1g58602
Q5VSL9
E. coli
See search help

AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research.

https://alphafold.ebi.ac.uk/

Our data

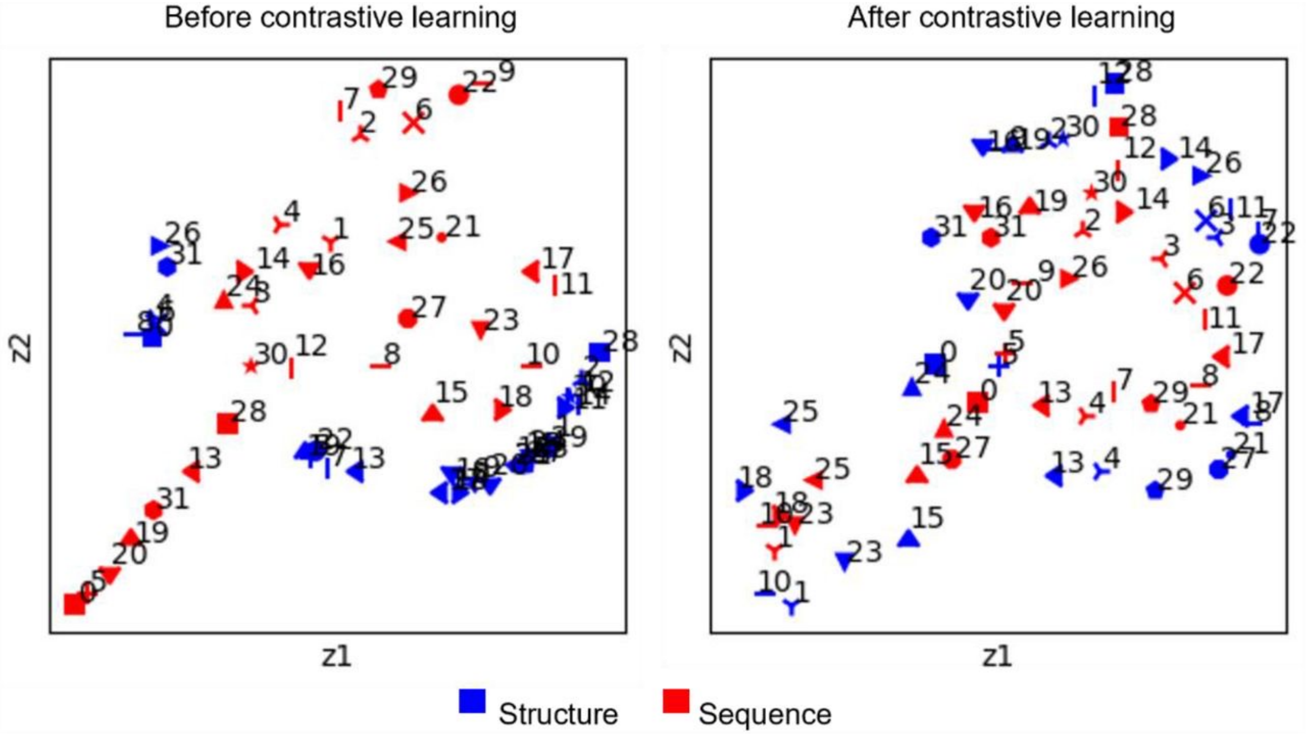
- Swiss-Prot Database (~ 580 K protein sequences)
- Gain the 3D structure from AlphaFold2 predictions
- Gain the contact map using self-developed Python code.

https://alphafold.ebi.ac.uk/

Some other useful detail

- 540 K proteins for training
- 40 K proteins for validation
- Play with hyperparameters
- Trained on our LCC V100
- Code developed based Pytorch

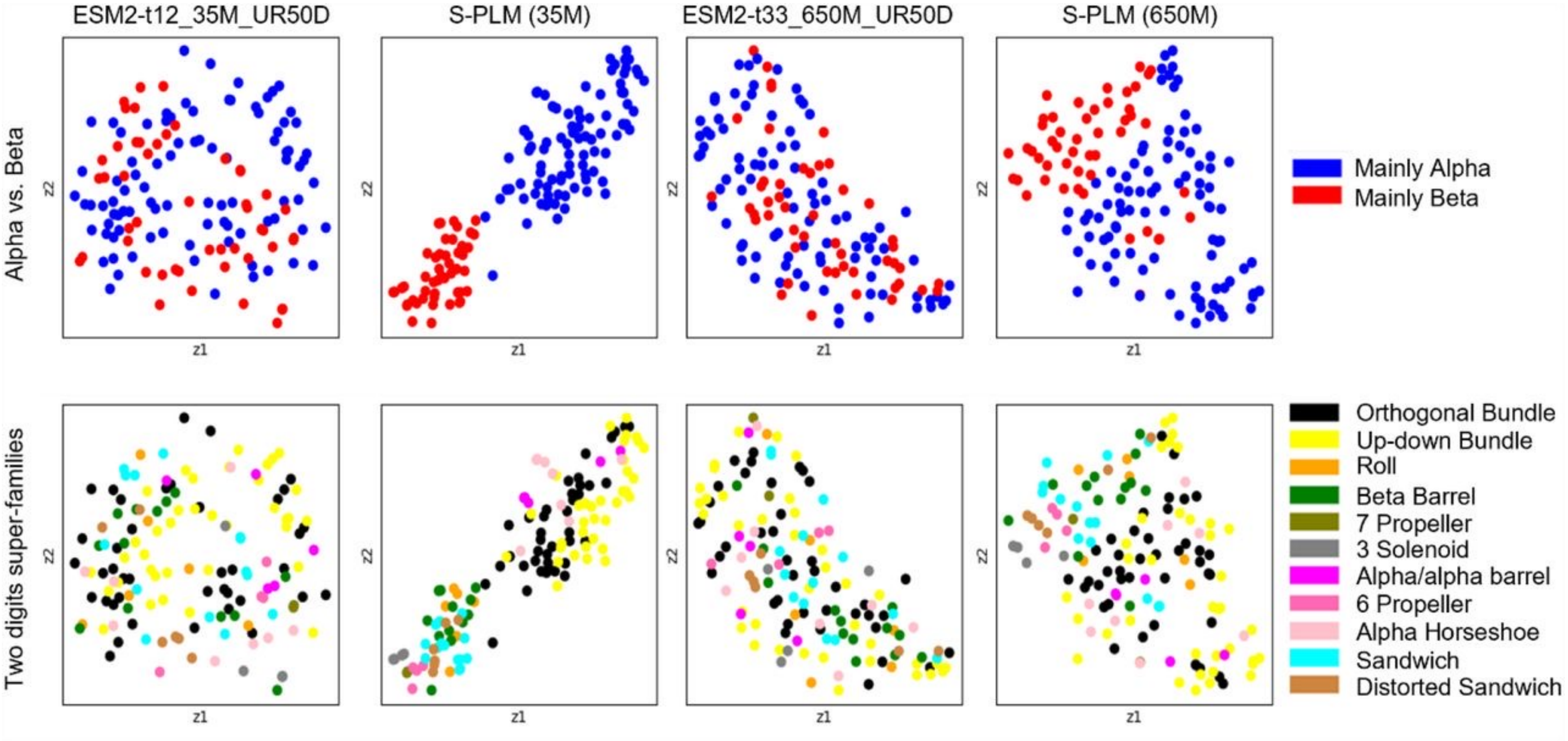
Some preliminary results



Contrastive learning enhances the alignment between sequence and structure embedding.
This enhancement implies that the embedding knows structure better

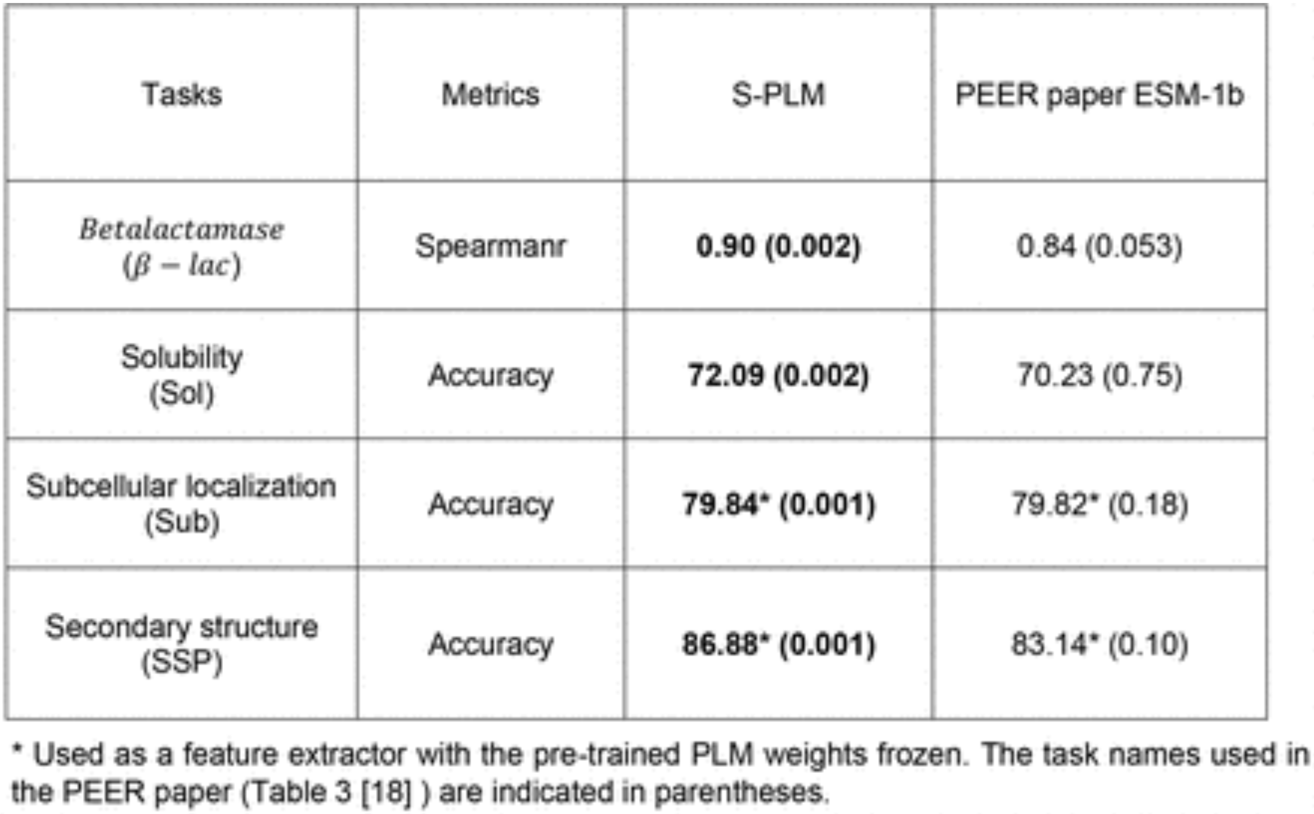
https://www.biorxiv.org/content/10.1101/2023.08.06.552203v1.full

CATH superfamily task



Our sequence embedding performs better on structure task than the original ESM2 model

Some preliminary results



Tasks	Metrics	S-PLM	PEER paper ESM-1b
$Betalactamase$ $(\beta - lac)$	Spearmanr	**0.90 (0.002)**	0.84 (0.053)
Solubility (Sol)	Accuracy	**72.09 (0.002)**	70.23 (0.75)
Subcellular localization (Sub)	Accuracy	**79.84* (0.001)**	79.82* (0.18)
Secondary structure (SSP)	Accuracy	**86.88* (0.001)**	83.14* (0.10)

* Used as a feature extractor with the pre-trained PLM weights frozen. The task names used in the PEER paper (Table 3 [18]) are indicated in parentheses.

https://arxiv.org/abs/2206.02096

Summary

- We have successfully implemented structure information into sequence embedding
- The developed structure-aware protein language models perform better in some downstream tasks