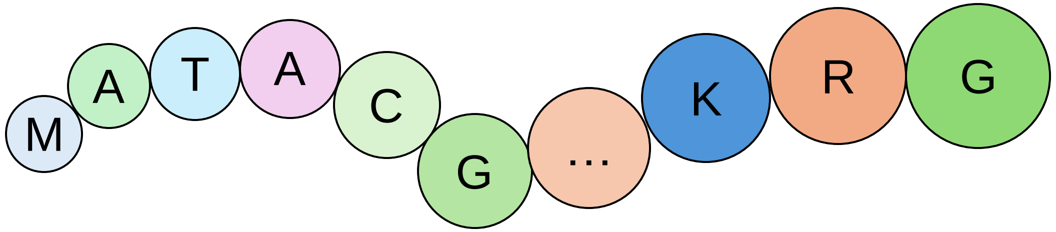
ProSST: Protein Language Modeling with Quantized Structure and Disentangled Attention

Mingchen Li[1,2,3], Yang Tan [1,2,3], Xinzhu Ma [2], Bozitao Zhong [1], Huiqun Yu [3], Ziyi Zhou [1], Wanli Ouyang [2]
Bingxin Zhou [1], Pan Tan [1,2,3], Liang Hong [1,2]

[1]Shanghai Jiao Tong University
[2] Shanghai Artificial Intelligence Laboratory
[3] East China University of Science and Technology

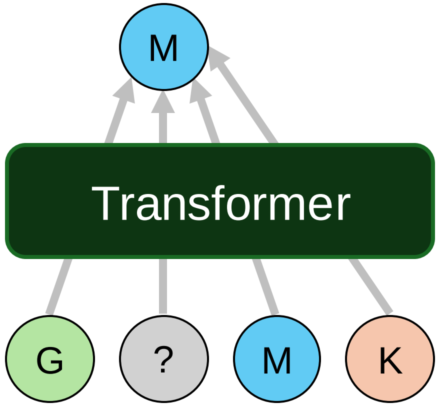
01 Introduction

- Proteins can be represented as sequences of tokens composed of 20 types of amino acids.

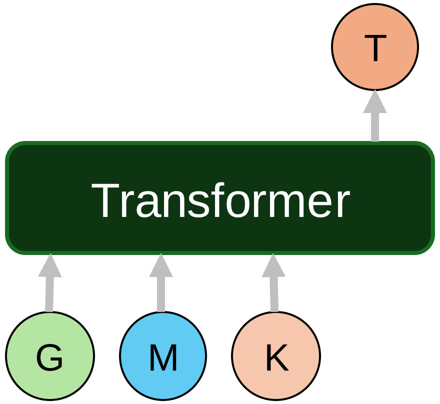


Protein Sequence (Amino acid string)

- Protein language models, pre-trained on databases with millions of protein sequences with BERT or GPT tasks, have become fundamental tools for protein function prediction.



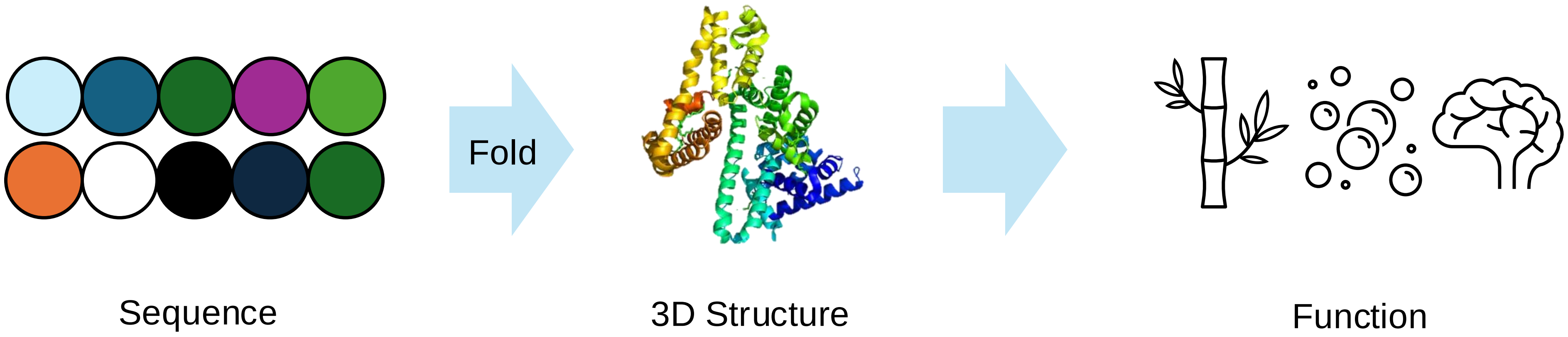
BERT-Style Pre-training (Masked token prediction)



GPT-Style (Next token prediction)

02 Introduction

- However, an essential property of proteins is that they form 3D structures, and this structure determines the protein's function.
- Only using amino acid token sequences may be insufficient.



Fold

Sequence

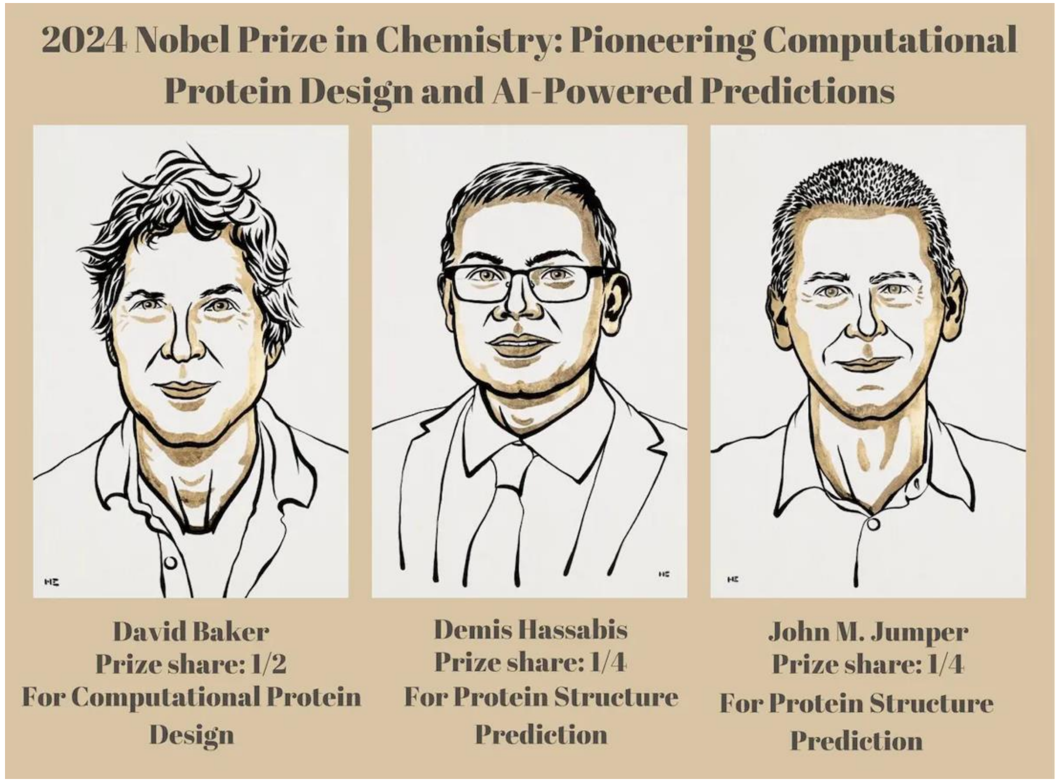
3D Structure

Function

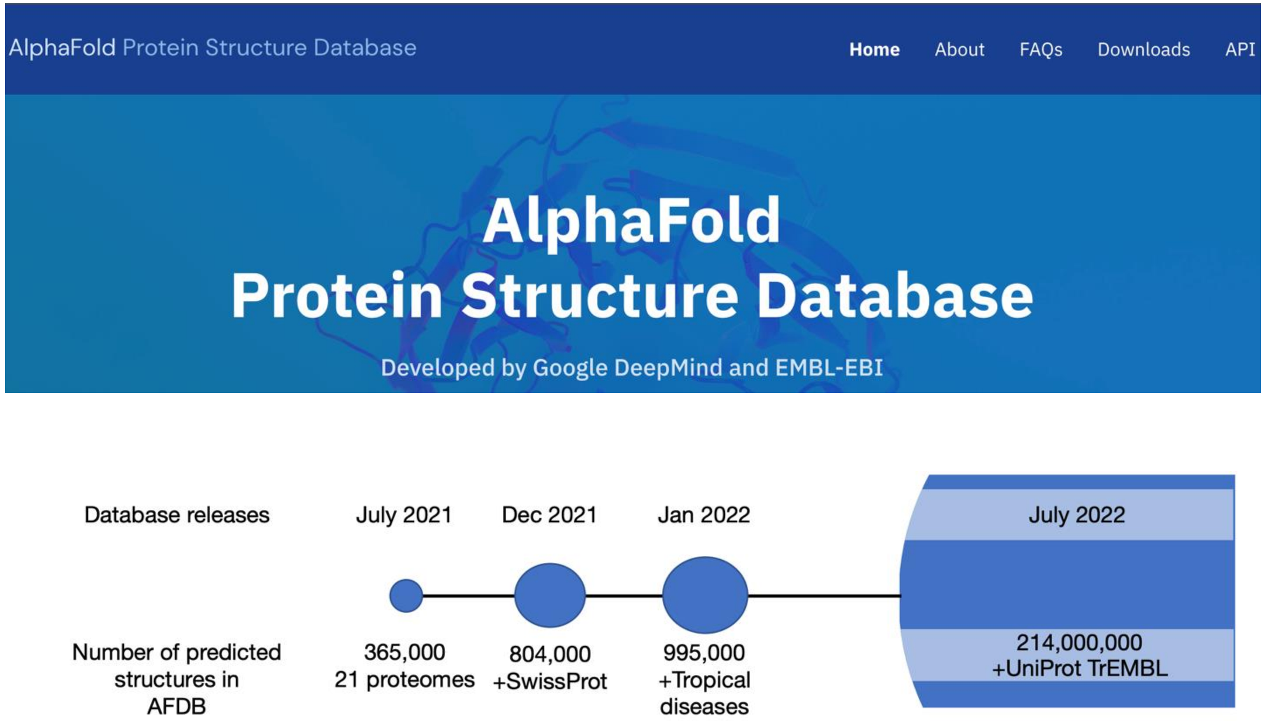
- Previous protein language models did not consider the 3D structure because structure data is hard to gather.

03 Introduction

- Luckily, AlphaFold 2 (which has won the 2024 Nobel Prize in Chemistry) can predict protein structures and has increased the protein structure database to millions, making it possible to develop structure-aware pre-traind protein language models.


2024 Nobel Prize in Chemistry: Pioneering Computational Protein Design and AI-Powered Predictions
David Baker
Prize share: 1/2
For Computational Protein Design
Demis Hassabis
Prize share: 1/4
For Protein Structure Prediction
John M. Jumper
Prize share: 1/4
For Protein Structure Prediction

2024 Nobel Prize in Chemistry



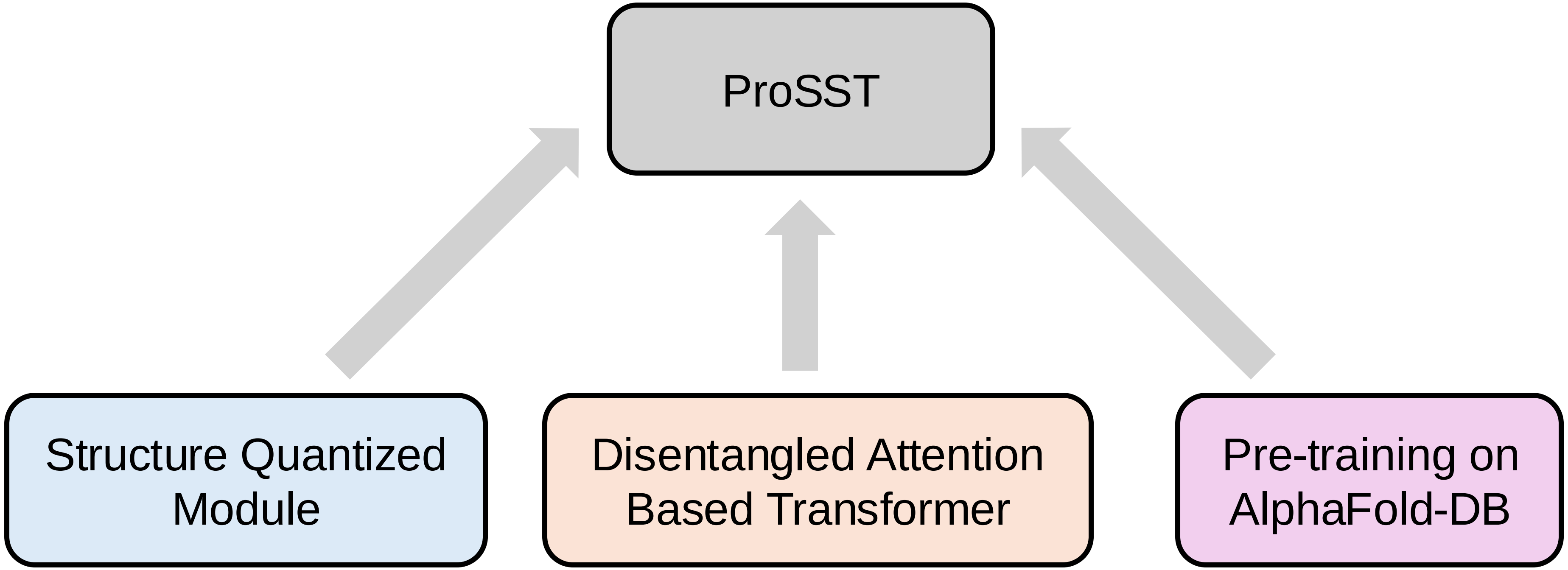
Barrio-Hernandez, et al. Nature. 2023.

AlphaFold Database

04

What is ProSST?

ProSST (**Pro**tein **S**equence-**S**tructure **T**ransformer) is a structure-aware protein language model with structure quantization and disentangled attention.



ProSST

Structure Quantized Module

Disentangled Attention Based Transformer

Pre-training on AlphaFold-DB

05

Protein Structure Quantization

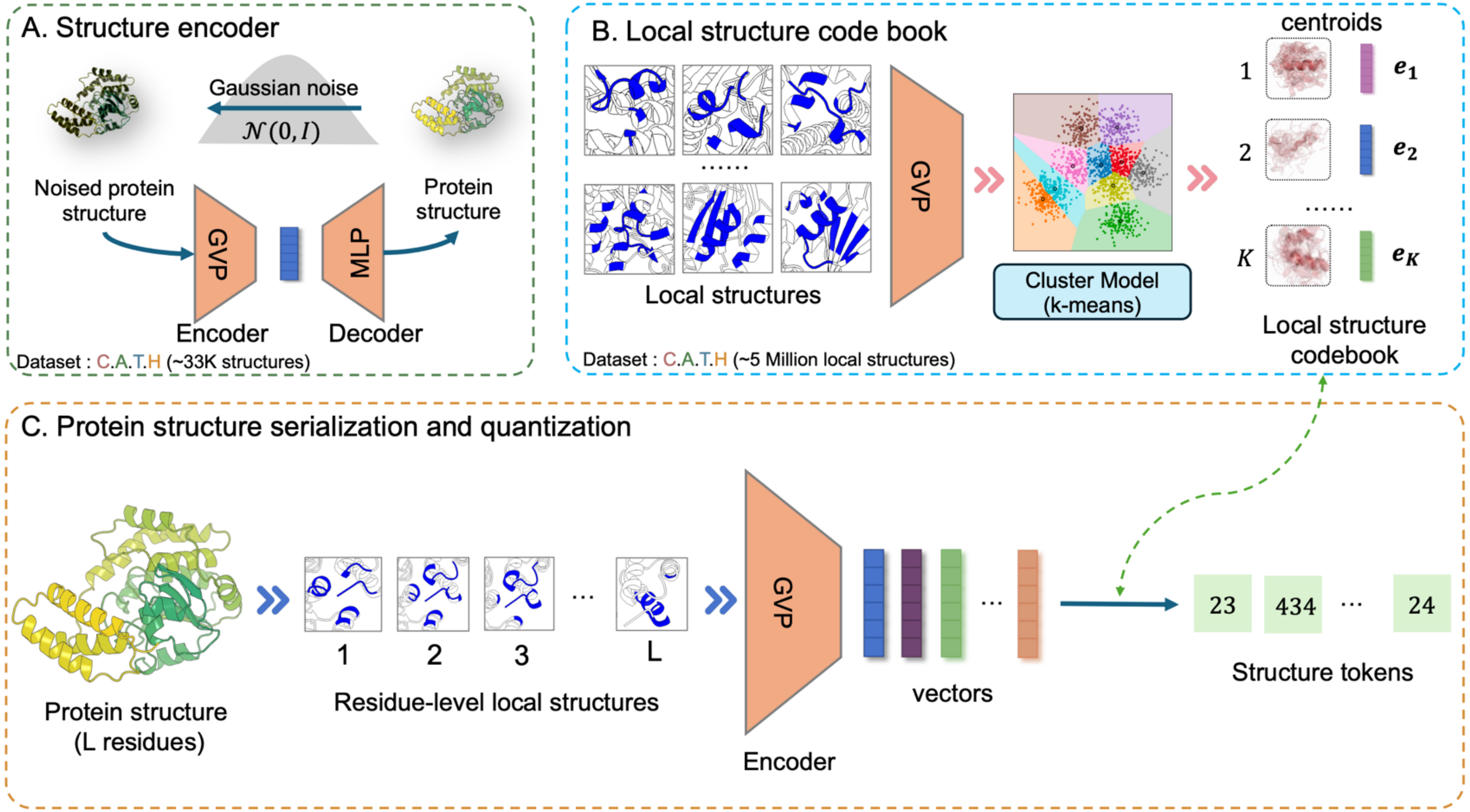
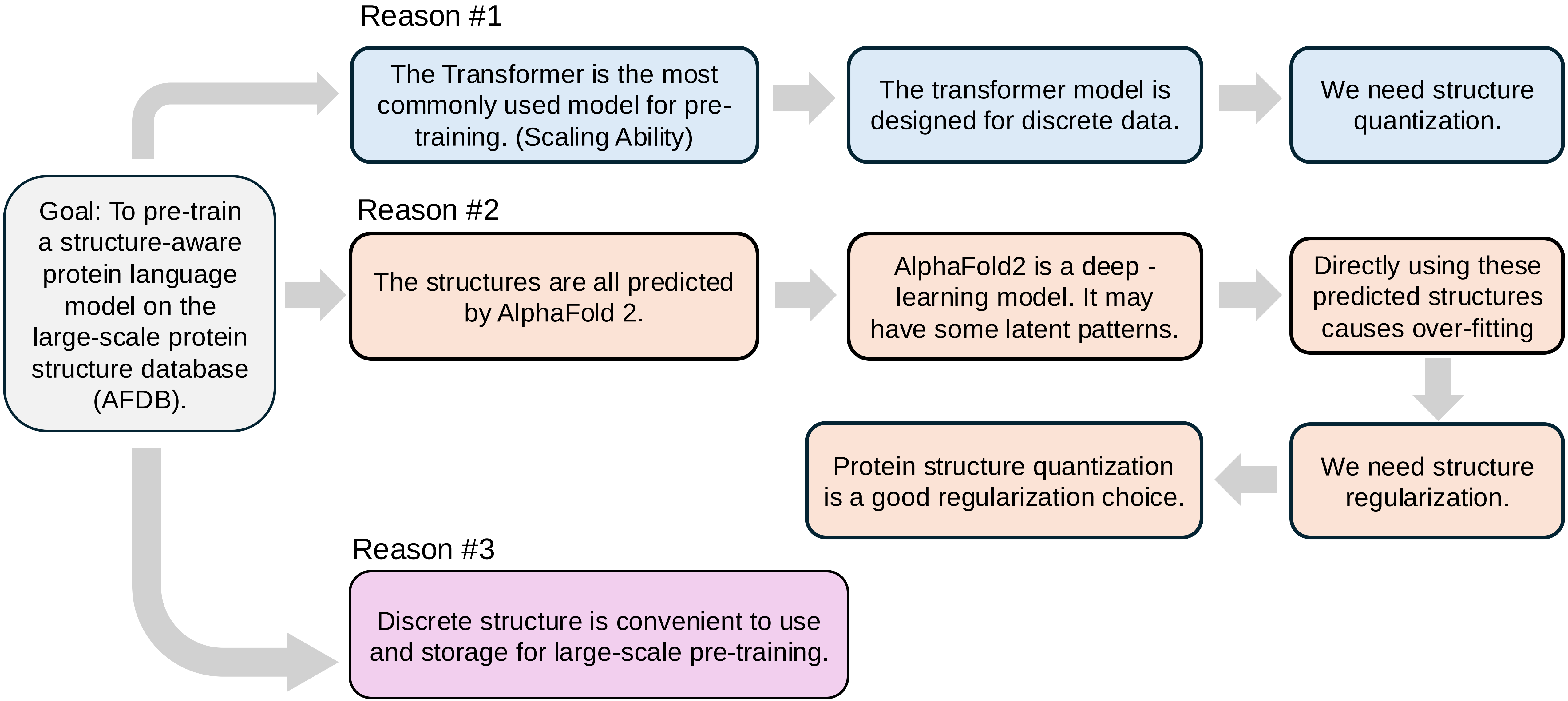


Figure 1: **The pipeline of structure quantization.** (A) Training of the structure encoder. (B) Local structure clustering and labeling. (C) Converting a protein structure to structure token sequence.

ProSST vs Foldseek (Kempen et al. 2024.)

	Foldseek	ProSST
Structure vocab size	20	2048
Local structure	3 residues	Up to 40 residues
Network	MLP	GVP-GNN
Training	VQ-VAE	DAE + k-means

06 Why We Do Structure Quantization?



Goal: To pre-train a structure-aware protein language model on the large-scale protein structure database (AFDB).

Reason #1

The Transformer is the most commonly used model for pre-training. (Scaling Ability)

The transformer model is designed for discrete data.

We need structure quantization.

Reason #2

The structures are all predicted by AlphaFold 2.

AlphaFold2 is a deep - learning model. It may have some latent patterns.

Directly using these predicted structures causes over-fitting

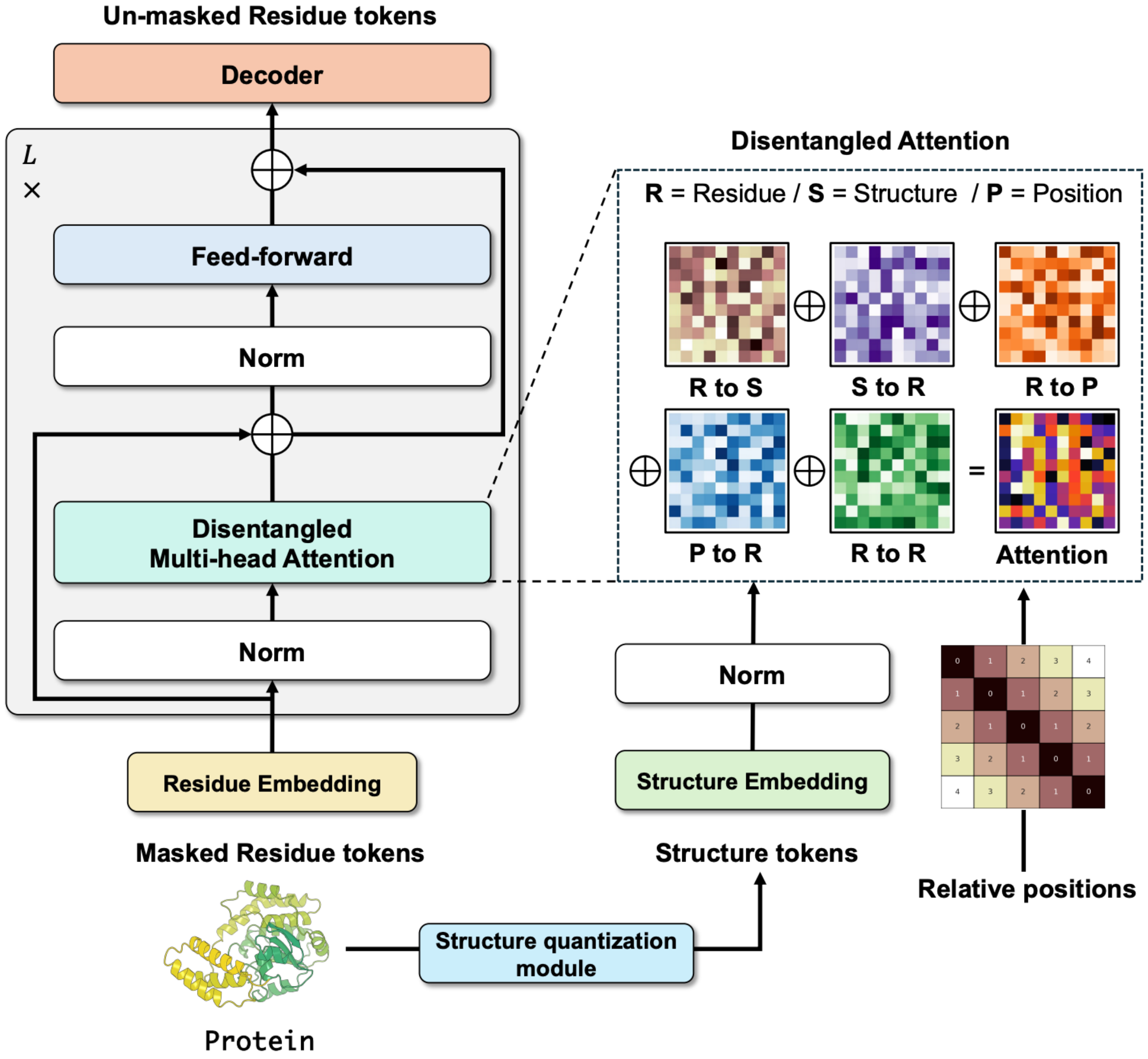
We need structure regularization.

Protein structure quantization is a good regularization choice.

Reason #3

Discrete structure is convenient to use and storage for large-scale pre-training.

07 Disentangled Attention-based Transformer



Un-masked Residue tokens

Decoder

L ×

Feed-forward

Norm

Disentangled Multi-head Attention

Norm

Residue Embedding

Masked Residue tokens

Protein

Structure quantization module

Structure tokens

Structure Embedding

Norm

Disentangled Attention

R = Residue / S = Structure / P = Position

R to S ⊕ S to R ⊕ R to P ⊕ P to R ⊕ R to R = Attention

0	1	2	3	4
1	0	1	2	3
2	1	0	1	2
3	2	1	0	1
4	3	2	1	0

Relative positions

$$\hat{A}_{i,j} = \underbrace{Q_i^r K_j^{r\top}}_{\text{(a) residue-to-residue}} + \underbrace{Q_i^r K_j^{s\top}}_{\text{(b) residue-to-structure}} + \underbrace{Q_i^r {K_{\delta(i,j)}^{p}}^{\top}}_{\text{(c) residue-to-position}} + \underbrace{K_j^r Q_i^{s\top}}_{\text{(d) structure-to-residue}} + \underbrace{K_j^r {Q_{\delta(j,i)}^{p}}^{\top}}_{\text{(e) position-to-residue}}$$

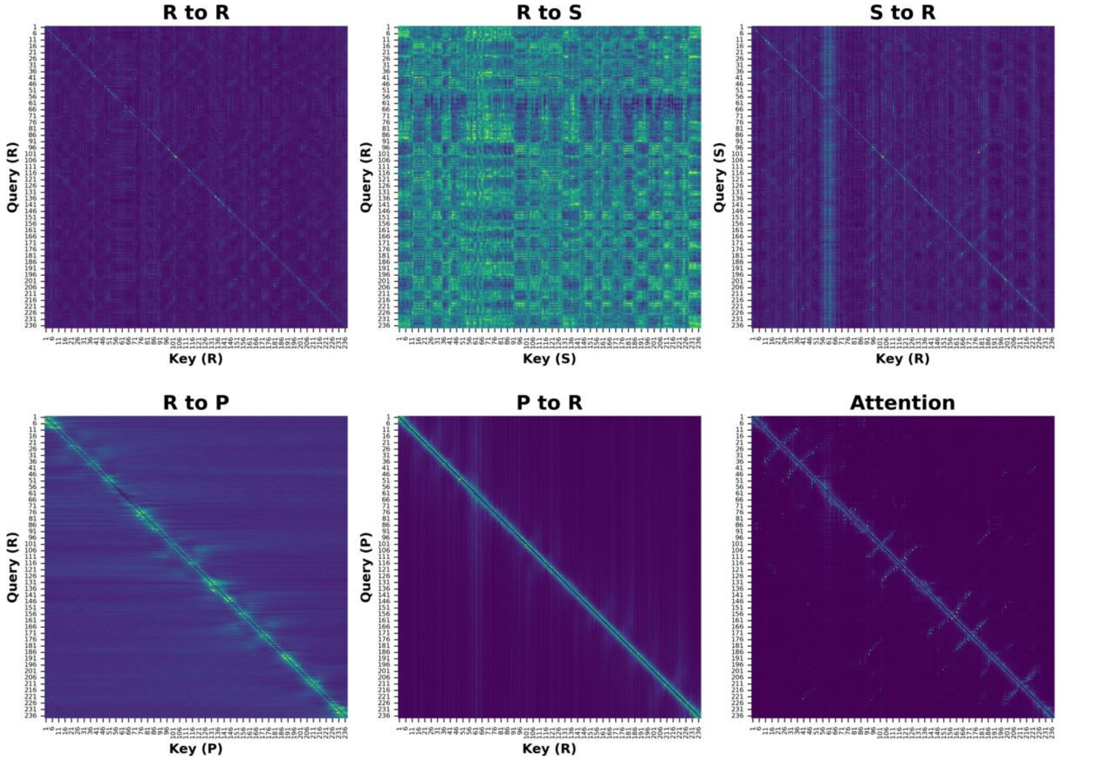
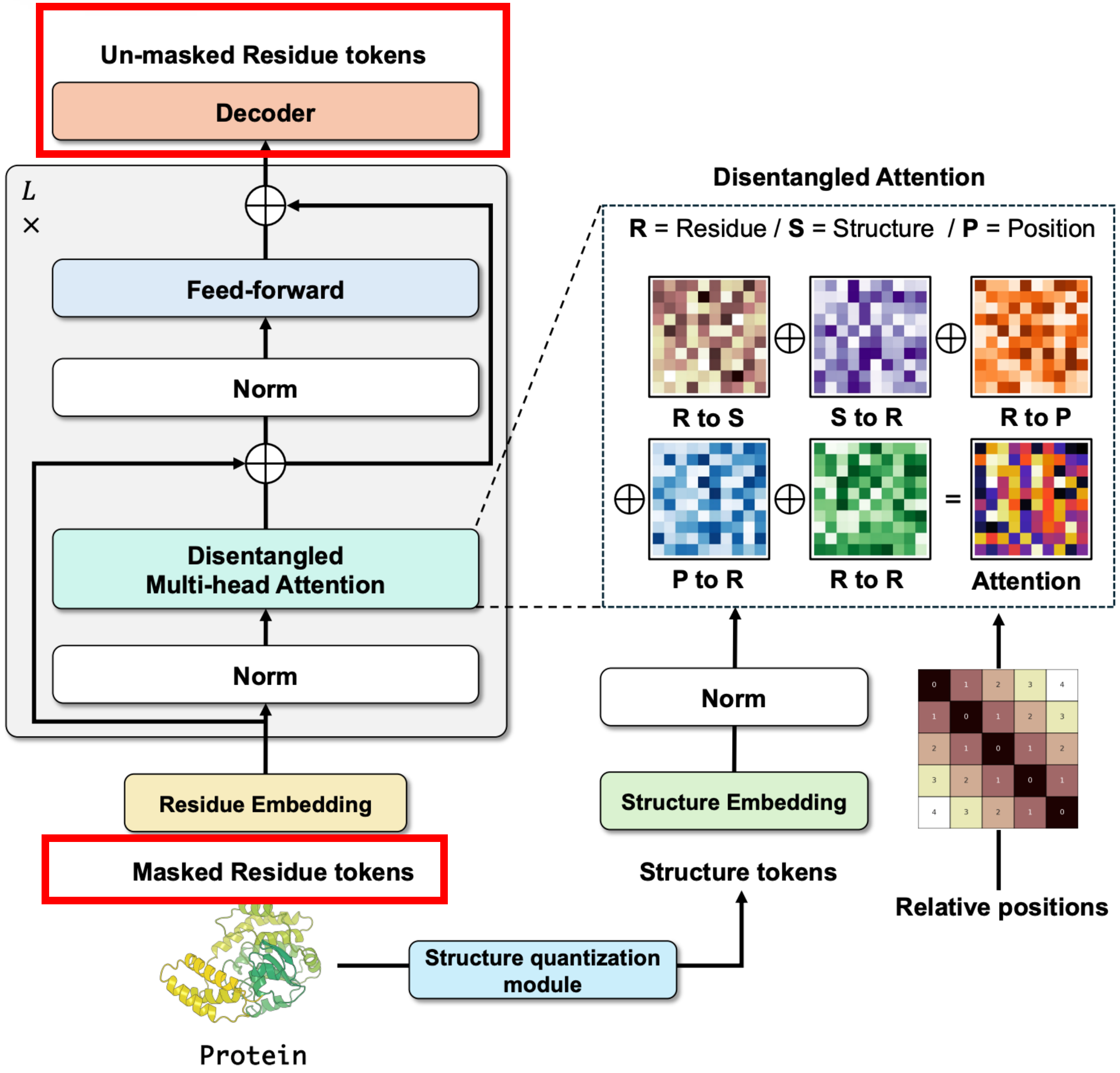
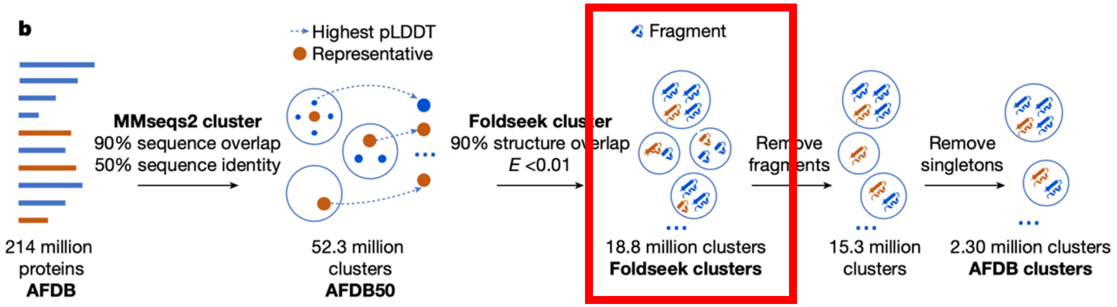


Figure A6: Different types of attentions on Green Fluorescent Protein (GFP). These attentions are the average of each head in the final layer of the Transformer.

08 Pre-training ProSST on AFDB



Pre-training Data (18 Million Structures)



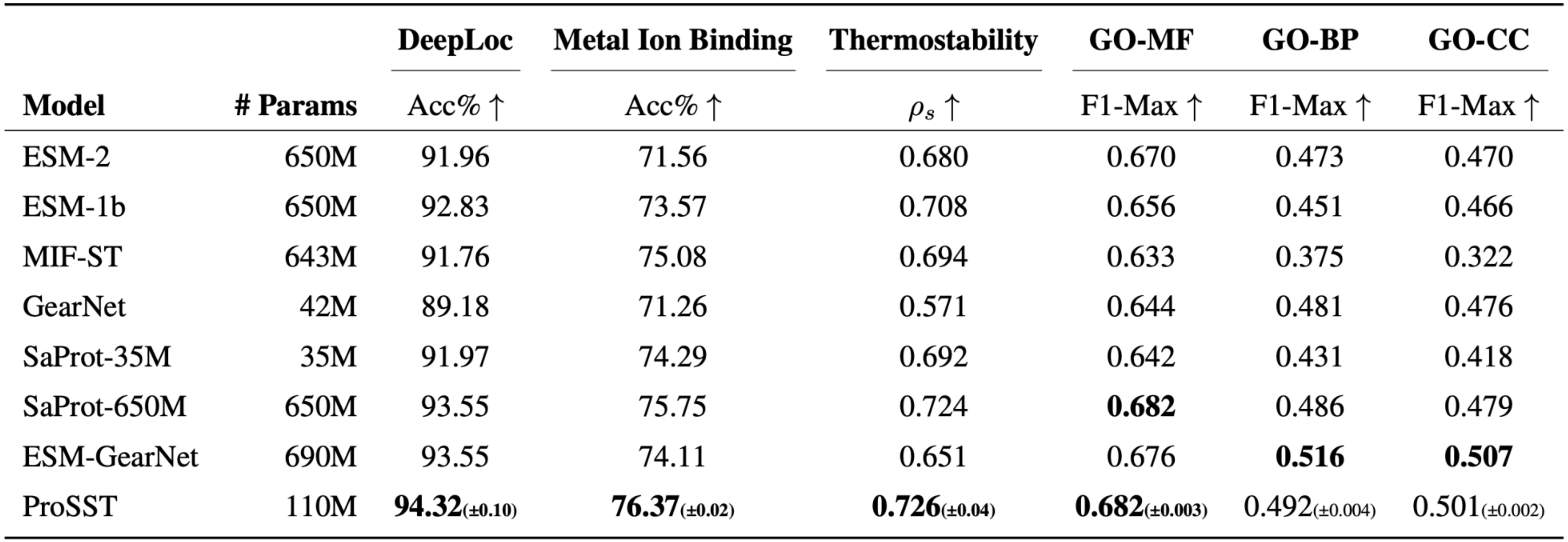
Barrio-Hernandez, et al. Nature. 2023.

Pre-training Objective:

$$\mathcal{L}_{MLM} = E_{x\sim X} E_{M} \sum_{i\in M} -\log p(x_i | x_{/M}, s)$$

Masked language modeling on the residue tokens.

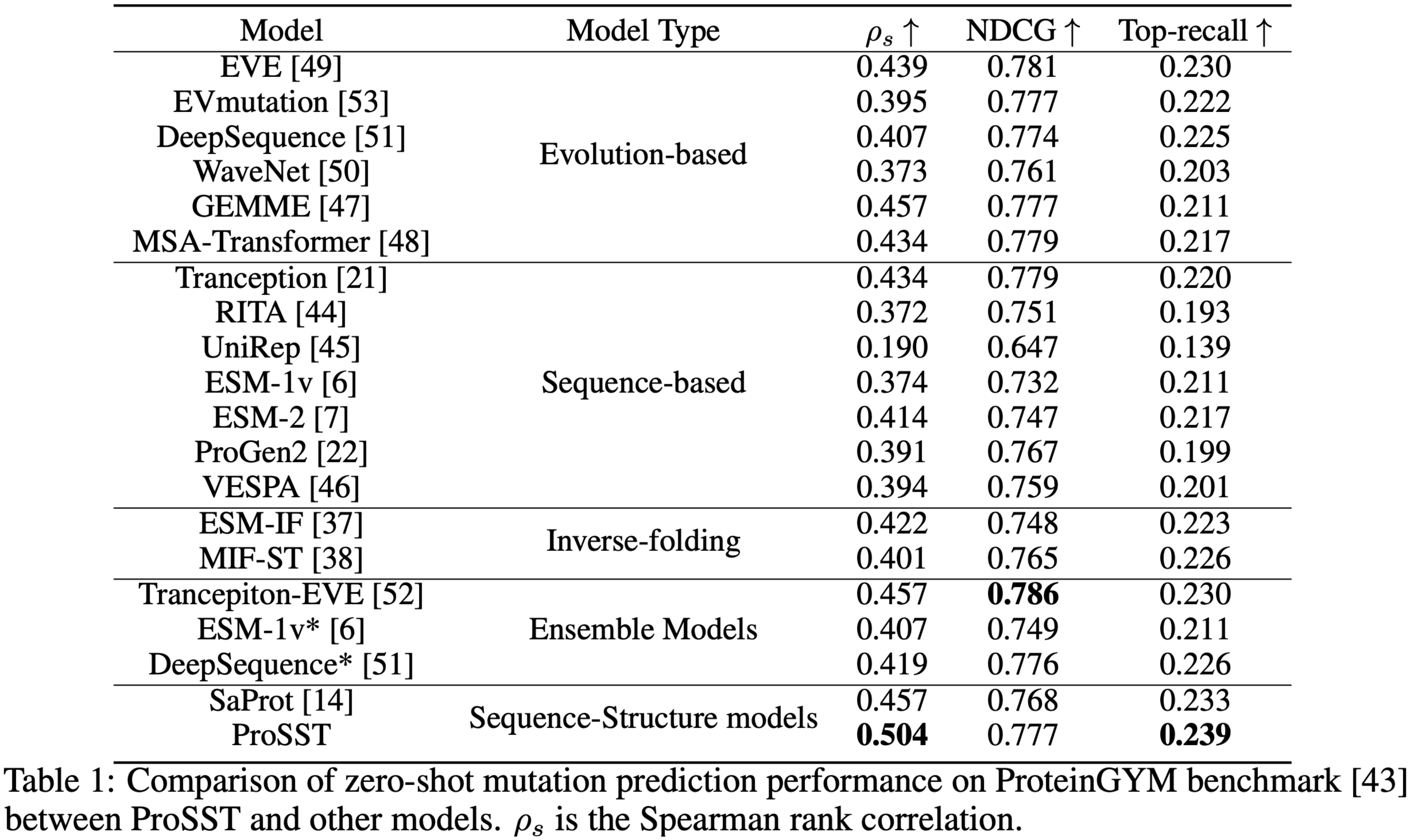
Results (Transfer Learning)



		DeepLoc	Metal Ion Binding	Thermostability	GO-MF	GO-BP	GO-CC
Model	**# Params**	Acc% ↑	Acc% ↑	ρ_s ↑	F1-Max ↑	F1-Max ↑	F1-Max ↑
ESM-2	650M	91.96	71.56	0.680	0.670	0.473	0.470
ESM-1b	650M	92.83	73.57	0.708	0.656	0.451	0.466
MIF-ST	643M	91.76	75.08	0.694	0.633	0.375	0.322
GearNet	42M	89.18	71.26	0.571	0.644	0.481	0.476
SaProt-35M	35M	91.97	74.29	0.692	0.642	0.431	0.418
SaProt-650M	650M	93.55	75.75	0.724	**0.682**	0.486	0.479
ESM-GearNet	690M	93.55	74.11	0.651	0.676	**0.516**	**0.507**
ProSST	110M	**94.32**(**±0.10**)	**76.37**(**±0.02**)	**0.726**(**±0.04**)	**0.682**(**±0.003**)	0.492(±0.004)	0.501(±0.002)

Table 2: Comparison of supervised fine-tuning on downstream tasks. ρ_s denotes the Spearman correlation coefficient.

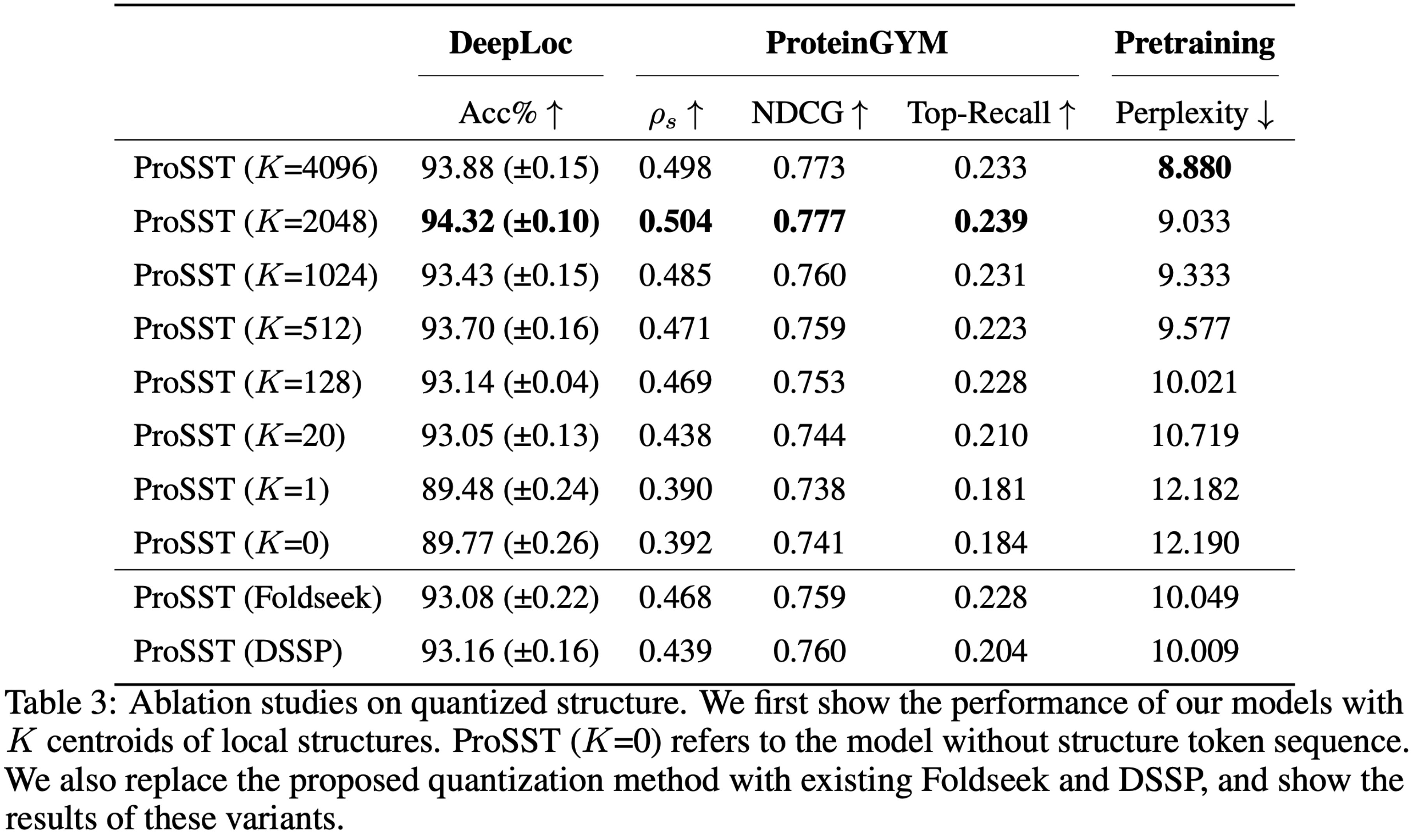
Results (Zero-shot mutant effect prediction)



Model	Model Type	$\rho_s \uparrow$	NDCG ↑	Top-recall ↑
EVE [49]	Evolution-based	0.439	0.781	0.230
EVmutation [53]		0.395	0.777	0.222
DeepSequence [51]		0.407	0.774	0.225
WaveNet [50]		0.373	0.761	0.203
GEMME [47]		0.457	0.777	0.211
MSA-Transformer [48]		0.434	0.779	0.217
Tranception [21]	Sequence-based	0.434	0.779	0.220
RITA [44]		0.372	0.751	0.193
UniRep [45]		0.190	0.647	0.139
ESM-1v [6]		0.374	0.732	0.211
ESM-2 [7]		0.414	0.747	0.217
ProGen2 [22]		0.391	0.767	0.199
VESPA [46]		0.394	0.759	0.201
ESM-IF [37]	Inverse-folding	0.422	0.748	0.223
MIF-ST [38]		0.401	0.765	0.226
Trancepiton-EVE [52]	Ensemble Models	0.457	**0.786**	0.230
ESM-1v* [6]		0.407	0.749	0.211
DeepSequence* [51]		0.419	0.776	0.226
SaProt [14]	Sequence-Structure models	0.457	0.768	0.233
ProSST		**0.504**	0.777	**0.239**

Table 1: Comparison of zero-shot mutation prediction performance on ProteinGYM benchmark [43] between ProSST and other models. ρ_s is the Spearman rank correlation.

Ablation Results (Quantized structure)



	DeepLoc	ProteinGYM			Pretraining
	Acc% ↑	ρ_s ↑	NDCG ↑	Top-Recall ↑	Perplexity ↓
ProSST (K=4096)	93.88 (±0.15)	0.498	0.773	0.233	**8.880**
ProSST (K=2048)	**94.32 (±0.10)**	**0.504**	**0.777**	**0.239**	9.033
ProSST (K=1024)	93.43 (±0.15)	0.485	0.760	0.231	9.333
ProSST (K=512)	93.70 (±0.16)	0.471	0.759	0.223	9.577
ProSST (K=128)	93.14 (±0.04)	0.469	0.753	0.228	10.021
ProSST (K=20)	93.05 (±0.13)	0.438	0.744	0.210	10.719
ProSST (K=1)	89.48 (±0.24)	0.390	0.738	0.181	12.182
ProSST (K=0)	89.77 (±0.26)	0.392	0.741	0.184	12.190
ProSST (Foldseek)	93.08 (±0.22)	0.468	0.759	0.228	10.049
ProSST (DSSP)	93.16 (±0.16)	0.439	0.760	0.204	10.009

Table 3: Ablation studies on quantized structure. We first show the performance of our models with K centroids of local structures. ProSST (K=0) refers to the model without structure token sequence. We also replace the proposed quantization method with existing Foldseek and DSSP, and show the results of these variants.

Conclusion & Future work

- We propose a protein structure quantization module, which can convert a protein structure into a sequence of discrete tokens

- We propose a disentangled attention Transformer to learn the relationship between protein structure and sequence.

- We pre-train our model on 18 millions of protein structures and it has achieved good performance in multiple tasks.

Future work

- ➢ Develop larger model with larger database.

- ➢ Study the structure search ability of our quantization module.