NLP for Science

Protein Design

Yi Hong Liu

2023/09/13

目录

- Background

- Method

- Conclusion

目 录

- Background

- Method

- Conclusion

Background

一、蛋白质与人类语言的共同之处

1.人类语言中，字母组成单词，单词连接构成有意义的句子，而在蛋白质中，氨基酸组合成二级结构元素或保守的蛋白质片段，片段可以组合成具有不同功能的蛋白质结构。

2. 所有语言都起源于50000-70000年前在中非使用的一种共同的祖先语言，在蛋白质中，所有生物的共同祖先是一种来自40亿年前的微生物，其中已经包含了通过进化发展起来的大多数现代蛋白质结构域。

3.在人类语言中，单词与相邻的单词有联系并相互作用，同时呈现长距离依存，在蛋白质结构中，序列中相距较远的氨基酸可以在3D结构中相互作用。

Background

二、蛋白质与人类语言的不同之处（需要攻克的难点）

1.首先，许多人类语言在书面文本中提供了清晰可辨的单词定义(中文是一个明显的例外)，但“单词边界”在蛋白质中不那么明显。
2.对蛋白质结构缺乏了解，就像我们目前对许多灭绝的语言缺乏了解一样。尽管我们有训练蛋白质结构的语料库，但正确解释生成的序列仍将是一个挑战，需要大量的实验测试来破译它们的功能。
3.蛋白质的进化也明显不同于语言的进化，它受到随机性和环境压力的影响，其语法不可避免地会包含许多不规则性。蛋白质的序列必须与折叠成三维结构相兼容，从而使蛋白质语言模型必须学习的模式产生偏差。

目录

- Background

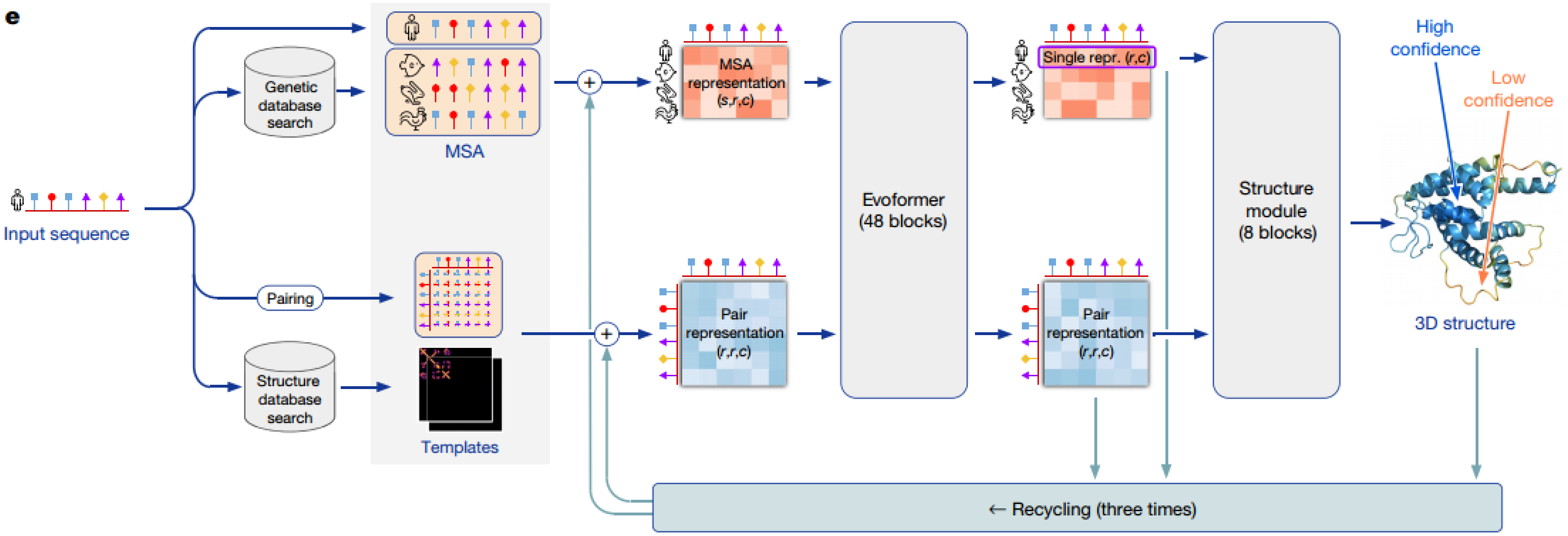
- Model

- Conclusion

一、AlphaFold2

AlphaFold2

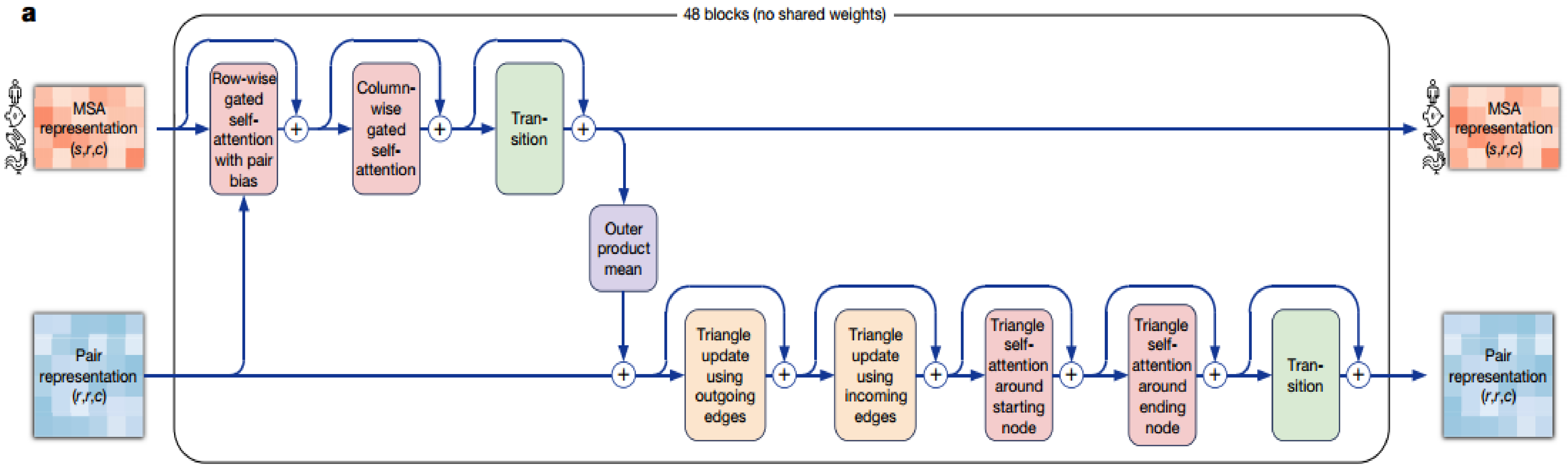
1.Model



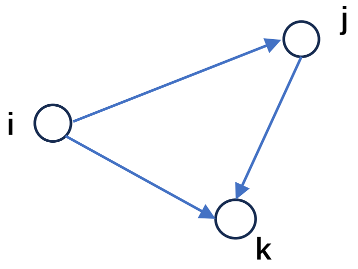
e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

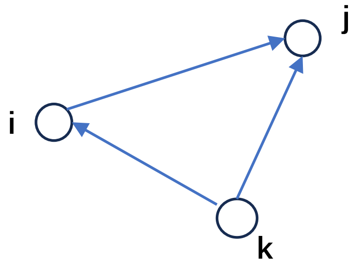
AlphaFold2

2.Encoder



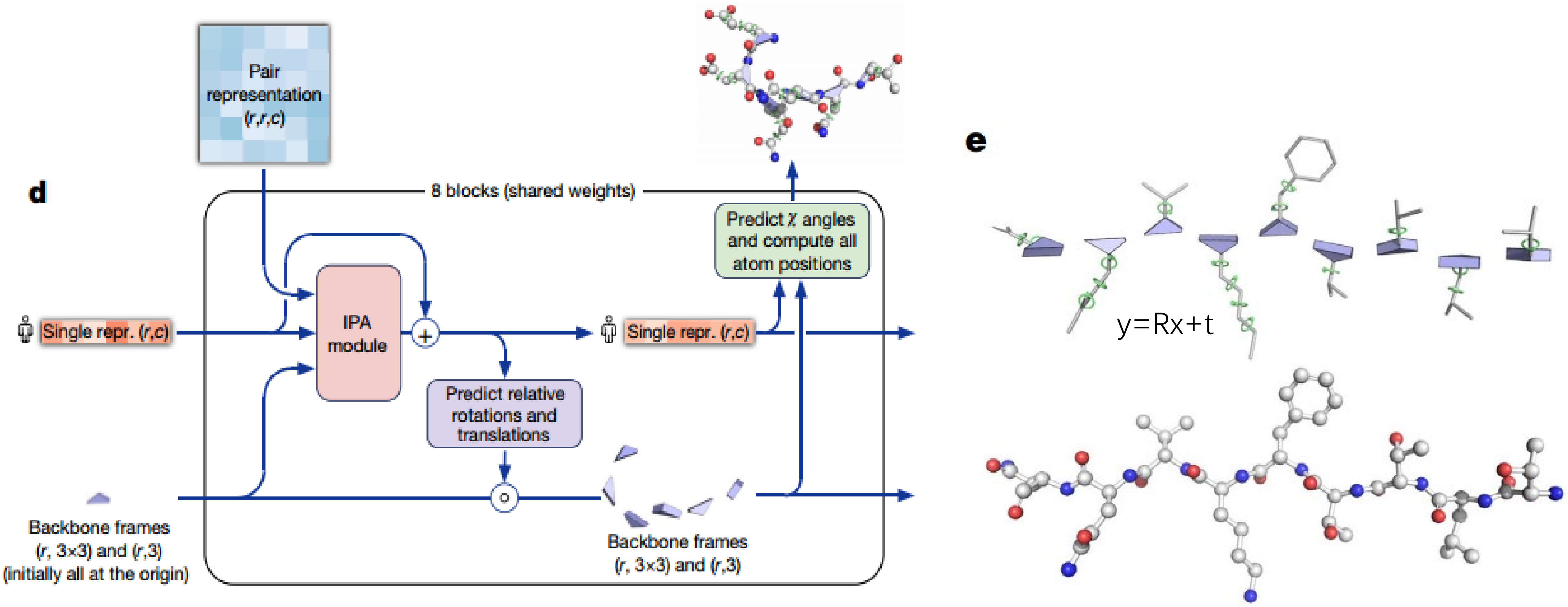
a
48 blocks (no shared weights)
MSA representation (s,r,c)
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Tran-sition
Outer product mean
Pair representation (r,r,c)
Triangle update using outgoing edges
Triangle update using incoming edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Tran-sition
MSA representation (s,r,c)
Pair representation (r,r,c)





AlphaFold2

3.Decoder


d
Pair representation (r,r,c)
8 blocks (shared weights)
Single repr. (r,c)
IPA module
+
Predict relative rotations and translations
○
Single repr. (r,c)
Predict χ angles and compute all atom positions
Backbone frames (r, 3×3) and (r,3) (initially all at the origin)
Backbone frames (r, 3×3) and (r,3)
e
y=Rx+t

AlphaFold2

4.Train

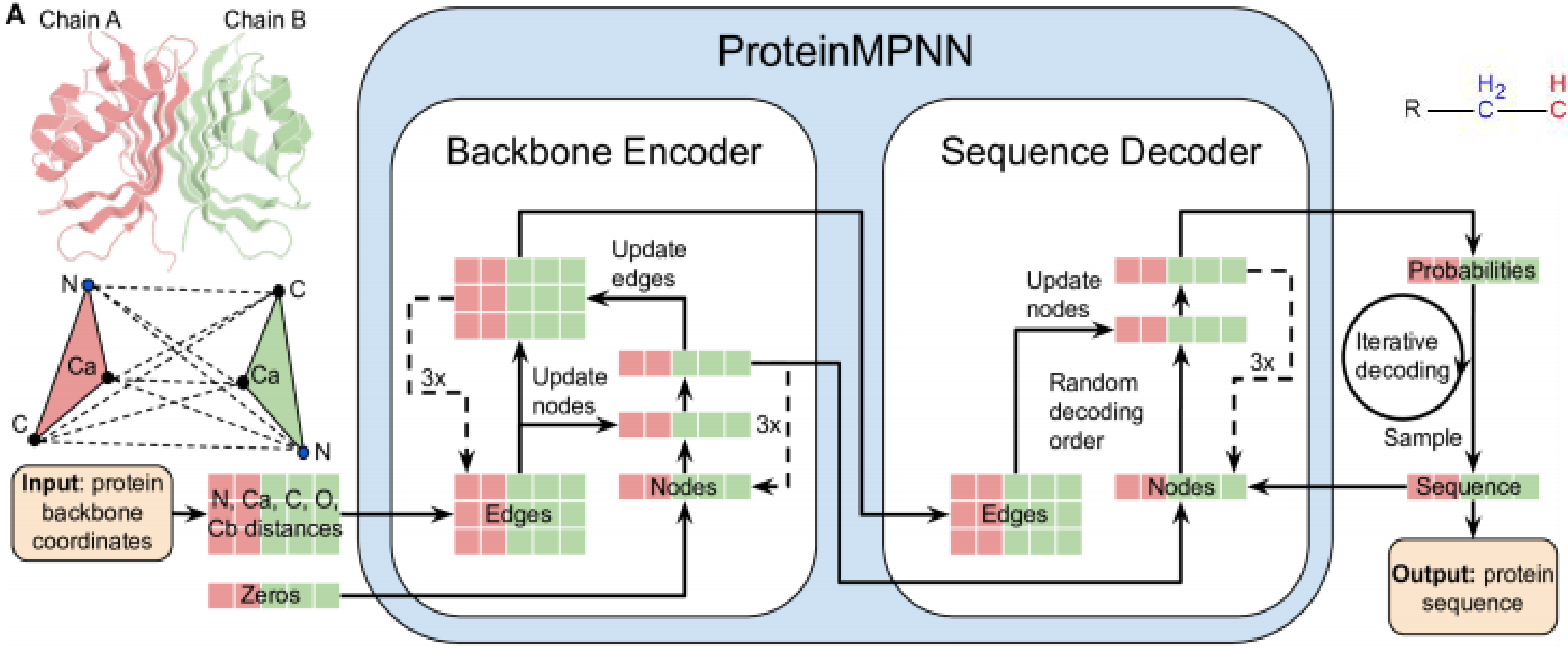
The AlphaFold architecture is able to train to high accuracy using only supervised learning on PDB data, but we are able to enhance accuracy using an approach similar to noisy student self-distillation35. In this procedure, we use a trained network to predict the structure of around 350,000 diverse sequences from Uniclust3036 and make a new dataset of predicted structures filtered to a high-confidence subset. We then train the same architecture again from scratch using a mixture of PDB data and this new dataset of predicted structures as the training data, in which the various training data augmentations such as cropping and MSA subsampling make it challenging for the network to recapitulate the previously predicted structures. This self-distillation procedure makes effective use of the unlabelled sequence data and considerably improves the accuracy of the resulting network.Additionally, we randomly mask out or mutate individual residues within the MSA and have a Bidirectional Encoder Representations from Transformers (BERT)-style37 objective to predict the masked elements of the MSA sequences. This objective encourages the network to learn to interpret phylogenetic and covariation relationships without hardcoding a particular correlation statistic into the features. The BERT objective is trained jointly with the normal PDB structure loss on the same training examples and is not pre-trained, in contrast to recent independent work

二、ProtMPNN

ProtMPNN

1.Model

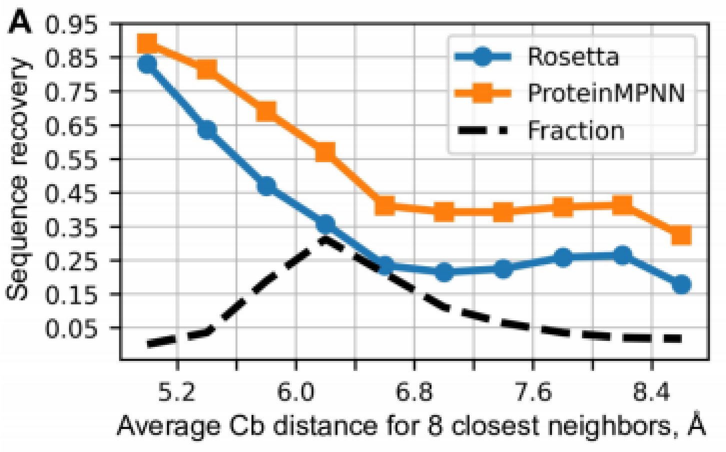
R
H2
C
H2
C
CHO



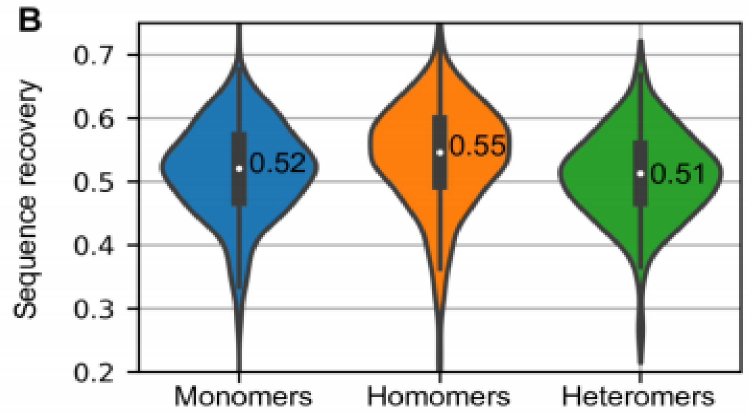
Dataset:CATH

ProtMPNN

2.Experiment

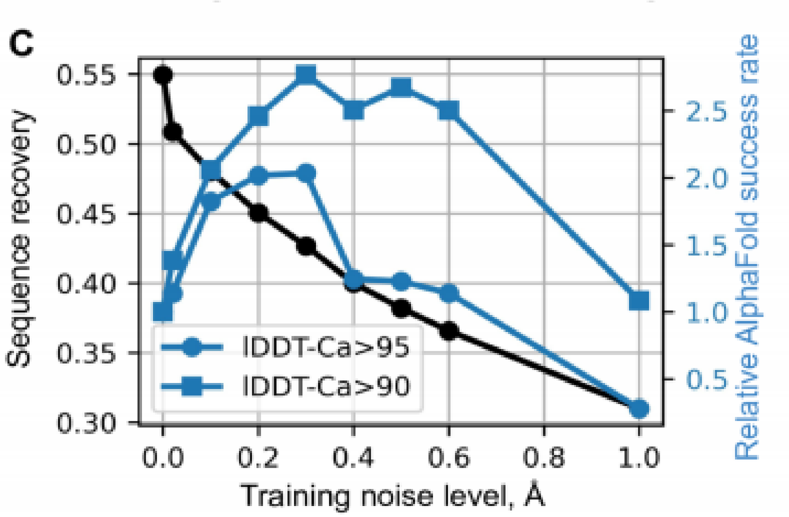


Average sequence recovery for ProteinMPNN was 52.4%, compared to 32.9% for Rosetta.

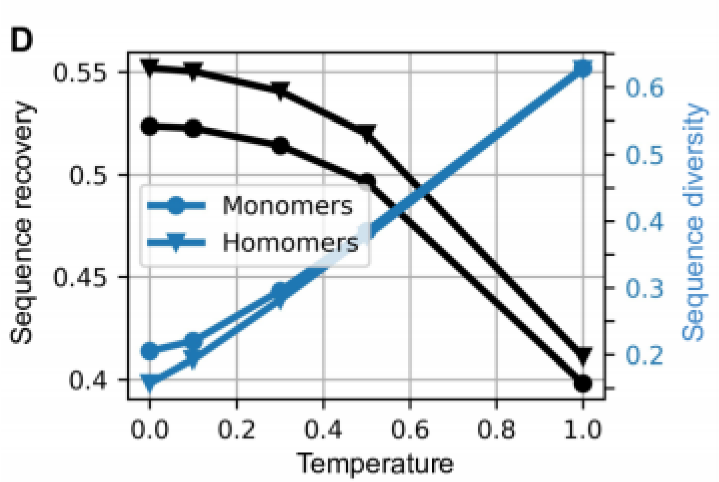


ProteinMPNN has similarly high sequence recovery for monomers(单体), homo-oligomers(均聚物), and hetero-oligomers(异聚物)

ProtMPNN

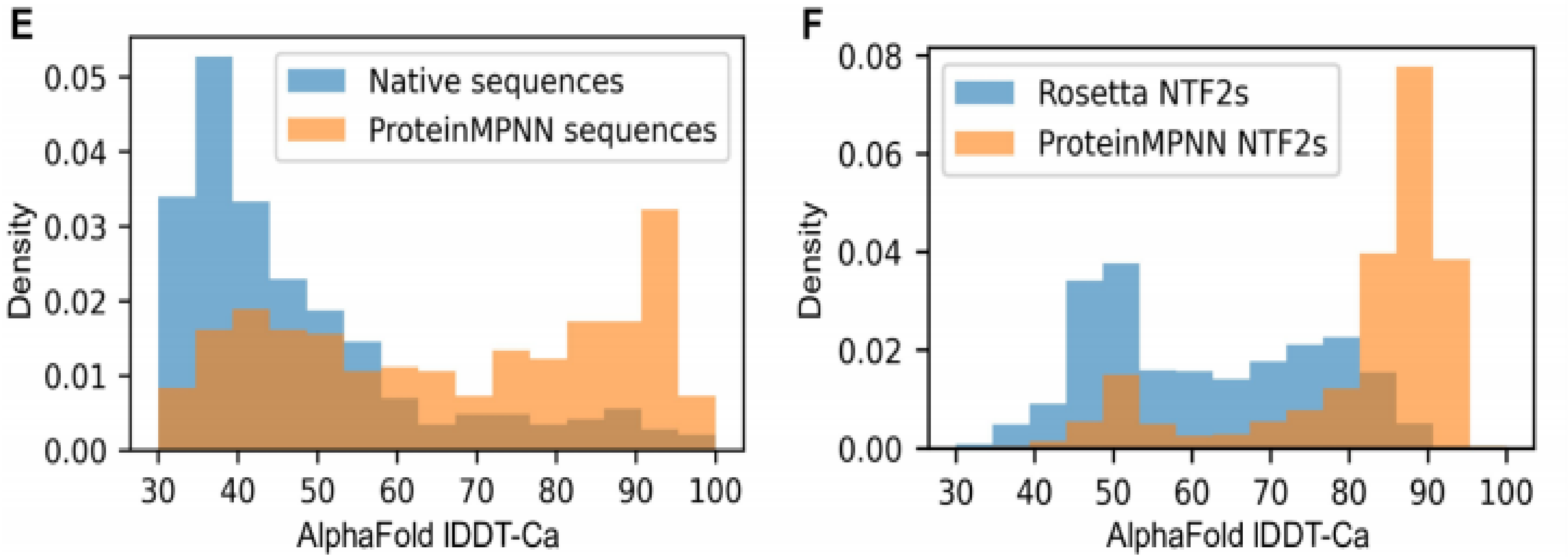


Sequence recovery (black) and relative AlphaFold success rates (blue) as a function of training noise level.



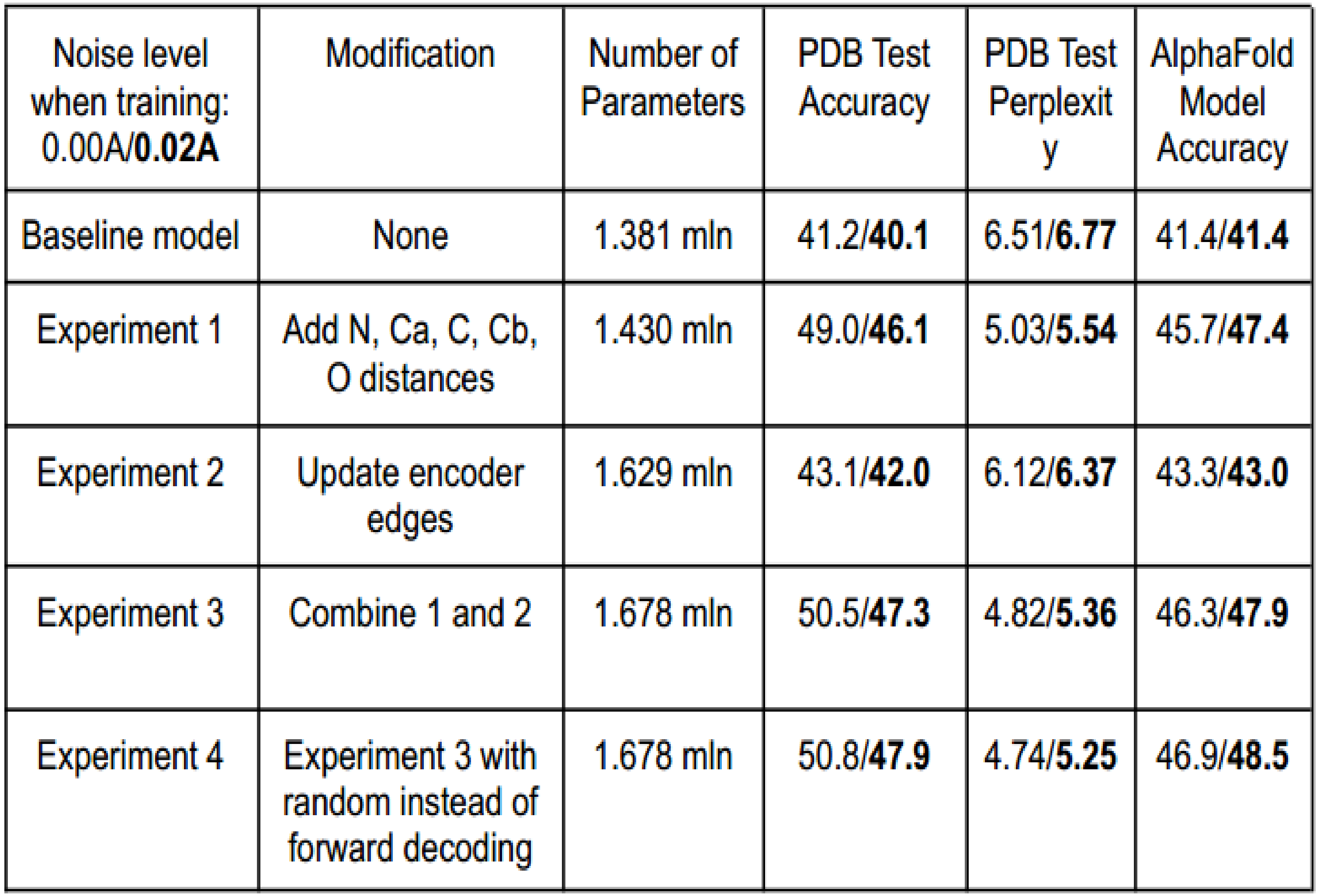
Sequence recovery and diversity as a function of sampling temperature.

ProtMPNN



ProteinMPNN redesign of previous Rosetta designed NTF2 fold proteins (3,000 backbones in total) results in considerably improved AlphaFold single sequence prediction accuracy.

ProteMPNN



Noise level when training: 0.00A/**0.02A**	Modification	Number of Parameters	PDB Test Accuracy	PDB Test Perplexity	AlphaFold Model Accuracy
Baseline model	None	1.381 mln	41.2/**40.1**	6.51/**6.77**	41.4/**41.4**
Experiment 1	Add N, Ca, C, Cb, O distances	1.430 mln	49.0/**46.1**	5.03/**5.54**	45.7/**47.4**
Experiment 2	Update encoder edges	1.629 mln	43.1/**42.0**	6.12/**6.37**	43.3/**43.0**
Experiment 3	Combine 1 and 2	1.678 mln	50.5/**47.3**	4.82/**5.36**	46.3/**47.9**
Experiment 4	Experiment 3 with random instead of forward decoding	1.678 mln	50.8/**47.9**	4.74/**5.25**	46.9/**48.5**

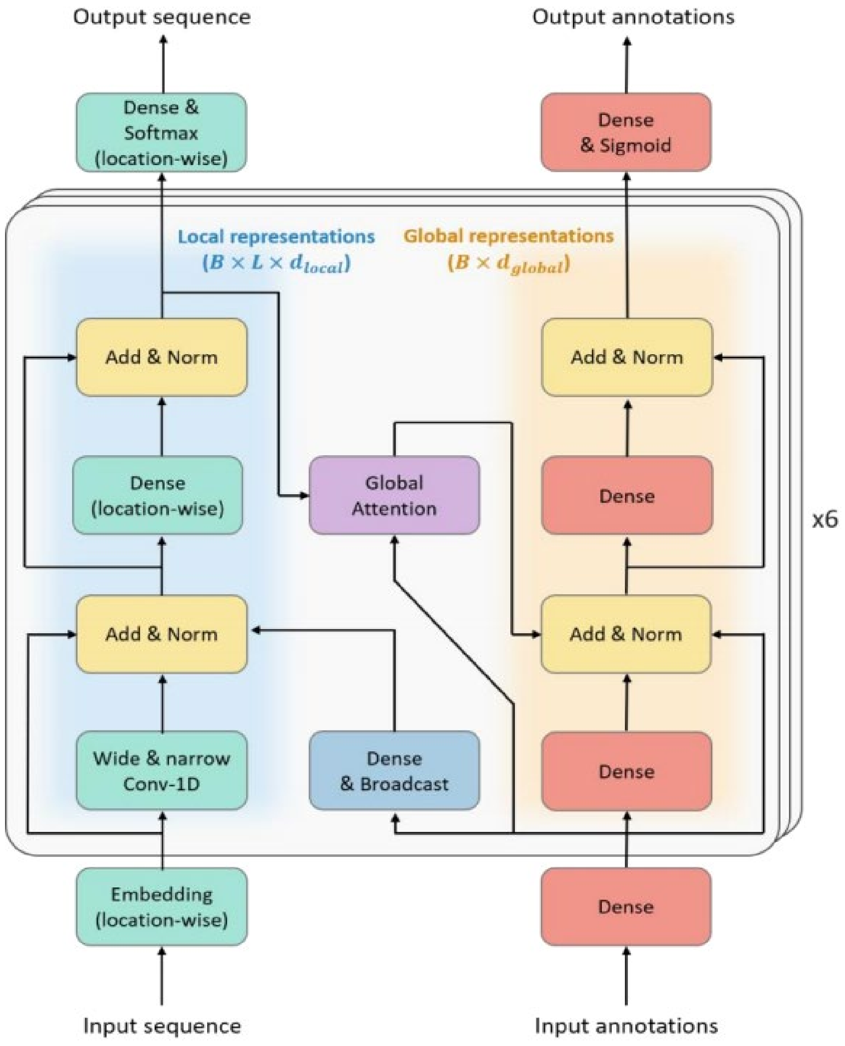
Test accuracy (percentage of correct amino amino acids recovered) and test perplexity (exponentiated categorical cross entropy loss per residue) are reported for models trained on the native backbone coordinates (left, normal font) and models trained with Gaussian noise (std=0.02Å) added to the backbone coordinates (right, bold font); all test evaluations are with no added noise. The final column shows sequence recovery on 5,000 AlphaFold protein backbone models with average pLDDT > 80.0 randomly chosen from UniRef50 sequences.

三、ProtBert

Model

三、ProteinBert

1.Model



ProteinBERT’s architecture is inspired by BERT. Unlike standard Transformers, ProteinBERT supports both local (sequential) and global data. The model consists of 6 transformer blocks manipulating local (left side) and global (right side) representations. Each such block manipulates these representations by fully-connected and convolutional layers (in the case of local representations), with skip connections and normalization layers between them. The local representations affect the global representations through a global attention layer, and the global representations affect the local representations through a broadcast fully-connected layer.

Dataset:UniProtKB/UniRef90

ProteinBert

2.Loss Function

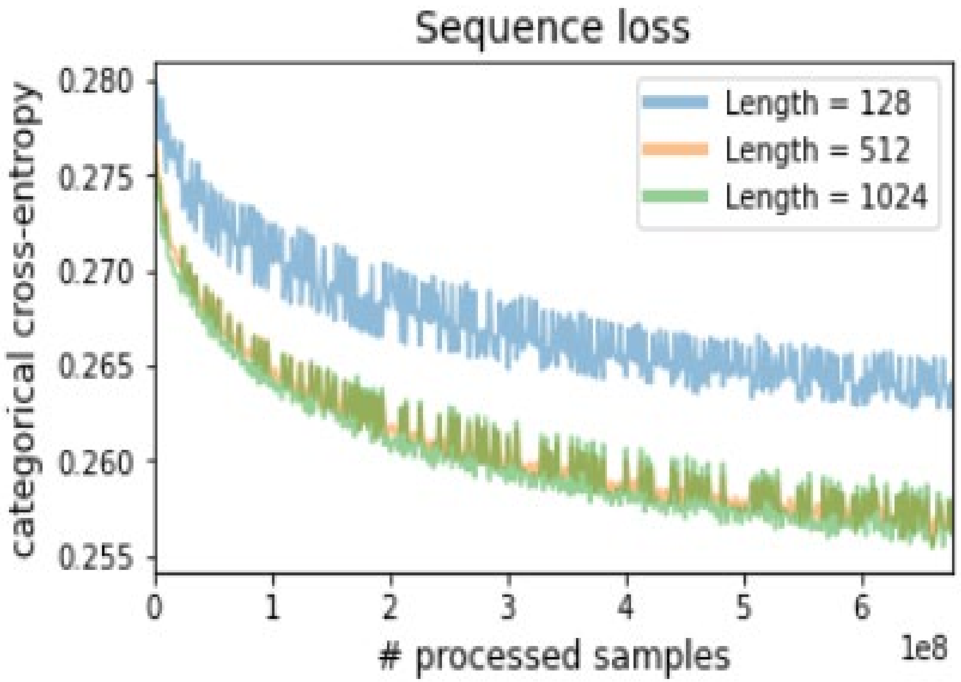
We considered only the 8,943 most frequent GO annotations that occurred at least 100 times in UniRef90

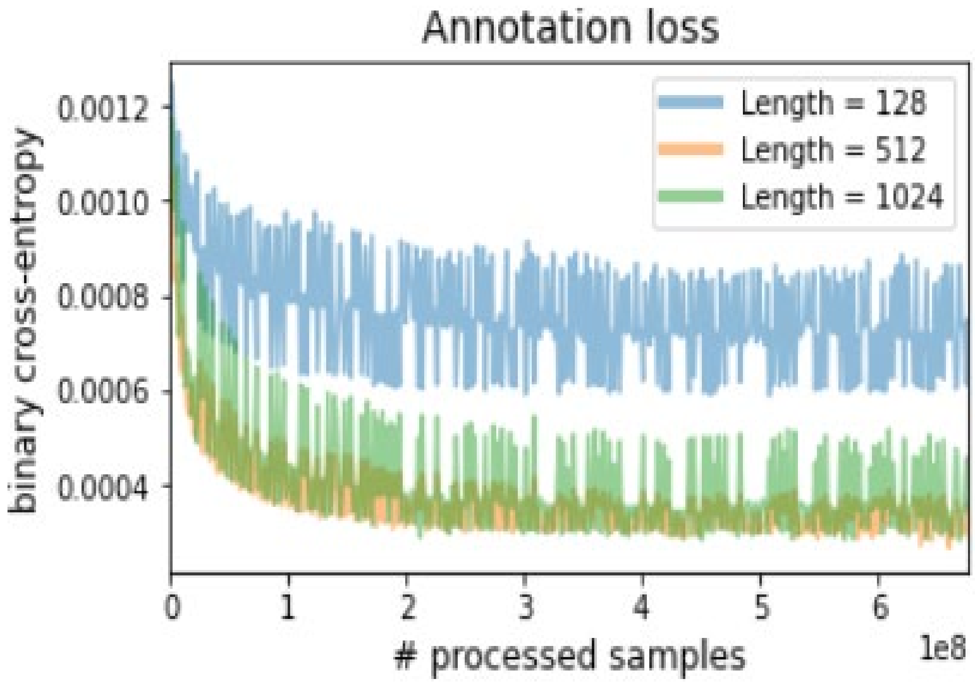
We used 26 unique tokens representing the 20 standard amino acids, selenocysteine (U), an undefined amino-acid (X), another amino acid (OTHER), and 3 additional tokens (START, END and PAD).

$$\mathcal{L} = -\sum_{i=1}^{l} log\left(\hat{S}_{i,S_i}\right) - \sum_{j=1}^{8943}\left(A_j \cdot log\left(\hat{A}_j\right) + \left(1 - A_j\right) \cdot log\left(1 - \hat{A}_j\right)\right)$$

where l is the sequence length, $S_i \in \{1, \cdots, 26\}$ is the sequence's true token at position i, $\hat{S}_{i,k} \in [0,1]$ is the predicted probability that position i has the token k (for any $k \in \{1, \cdots, 26\}$), $A_j \in \{0,1\}$ is the true indicator for annotation j (for any $j \in \{1, \cdots 8943\}$), and $\hat{A}_j \in [0,1]$ is the predicted probability for the protein to have annotation

ProteinBert

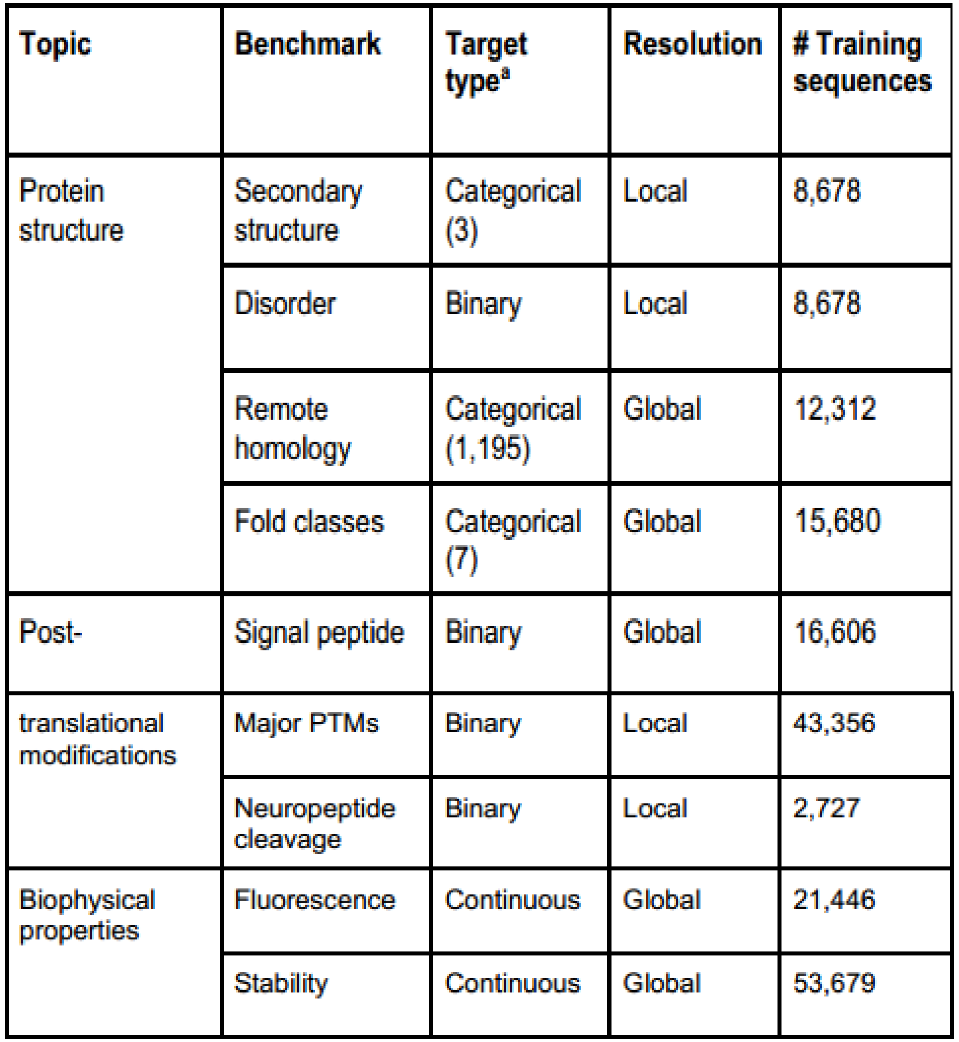




ProteinBERT was pretrained on ~106M UniRef90 records for ~6.4 epochs. We see that the language modeling loss continues to improve on the training set (i.e does not saturate), even after multiple epochs

During pretraining, we periodically changed the sequence length used to encode the input and output protein sequences (128, 512 or 1024 tokens). We observe somewhat lower performance for the 128-token encoding, but similar for 512 and 1024

3.Experiment



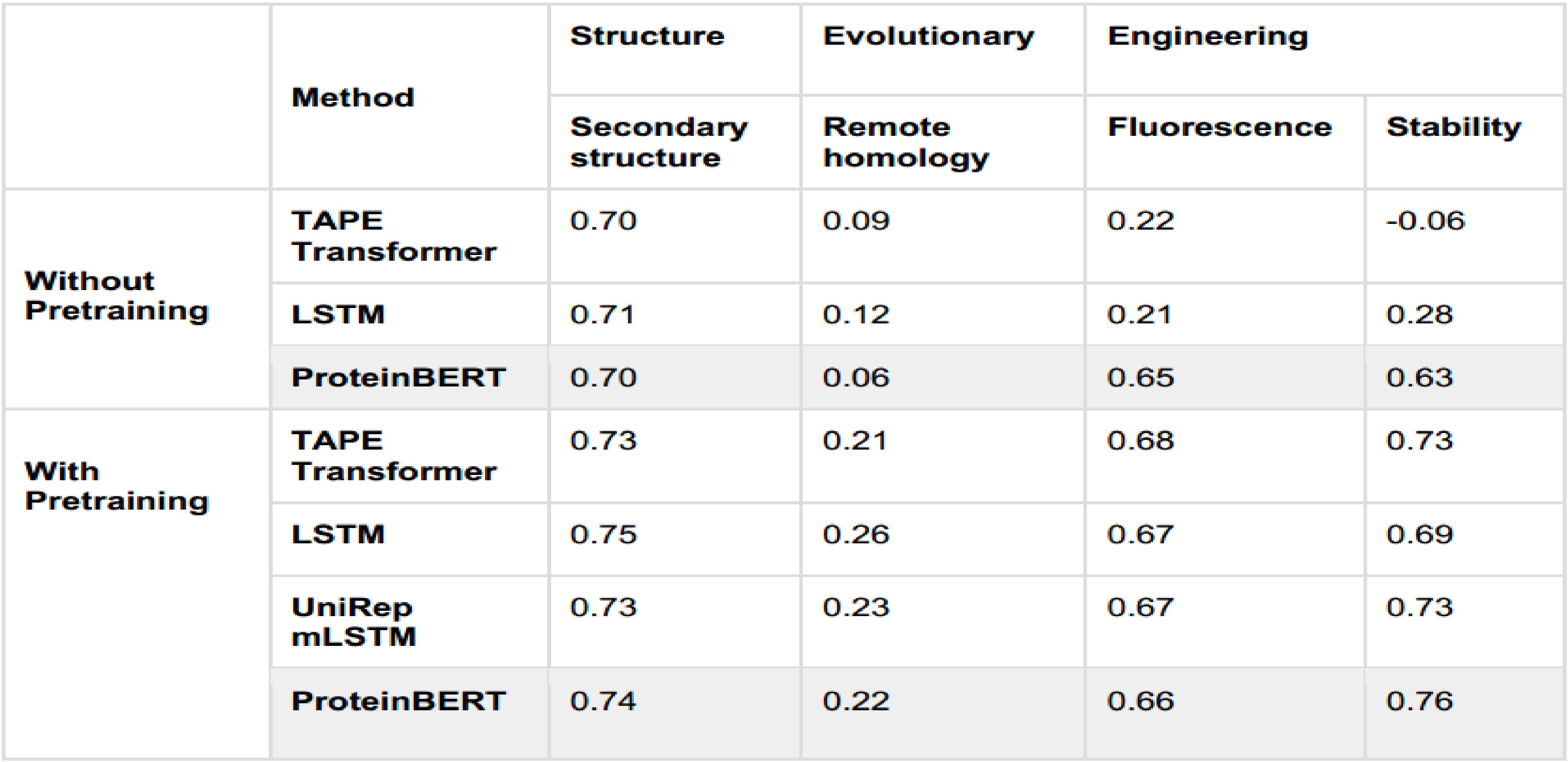
Topic	Benchmark	Target type[a]	Resolution	# Training sequences
Protein structure	Secondary structure	Categorical (3)	Local	8,678
	Disorder	Binary	Local	8,678
	Remote homology	Categorical (1,195)	Global	12,312
	Fold classes	Categorical (7)	Global	15,680
Post-	Signal peptide	Binary	Global	16,606
translational modifications	Major PTMs	Binary	Local	43,356
	Neuropeptide cleavage	Binary	Local	2,727
Biophysical properties	Fluorescence	Continuous	Global	21,446
	Stability	Continuous	Global	53,679

To evaluate ProteinBERT, we tested it on nine benchmarks concerning all major facets of protein research, covering protein structure, post-translational modifications and biophysical properties.

ProteinBert

4.Result

Table 2: TAPE benchmark results

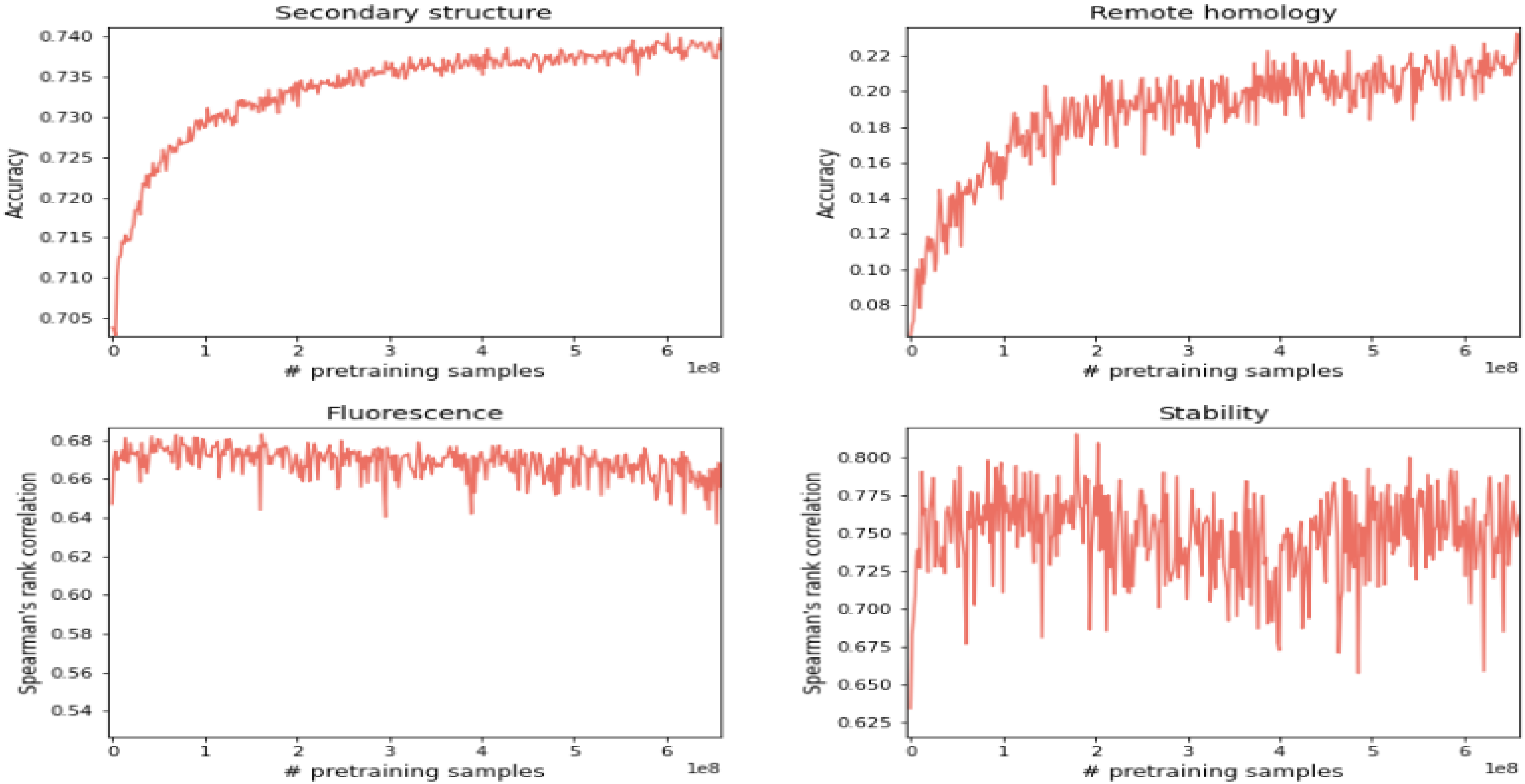


	Method	Structure	Evolutionary	Engineering	
		Secondary structure	Remote homology	Fluorescence	Stability
Without Pretraining	TAPE Transformer	0.70	0.09	0.22	-0.06
	LSTM	0.71	0.12	0.21	0.28
	ProteinBERT	0.70	0.06	0.65	0.63
With Pretraining	TAPE Transformer	0.73	0.21	0.68	0.73
	LSTM	0.75	0.26	0.67	0.69
	UniRep mLSTM	0.73	0.23	0.67	0.73
	ProteinBERT	0.74	0.22	0.66	0.76

Notably, the compared deep learning models have ~38M parameters, in contrast to ~16M parameters in ProteinBERT. We evaluated ProteinBERT with and without pretraining, observing that pretraining has a major, positive effect on performance in all tasks. Across these benchmarks, ProteinBERT shows performance comparable, or that exceeds similar, larger models, such as the Transformer used in TAPE.

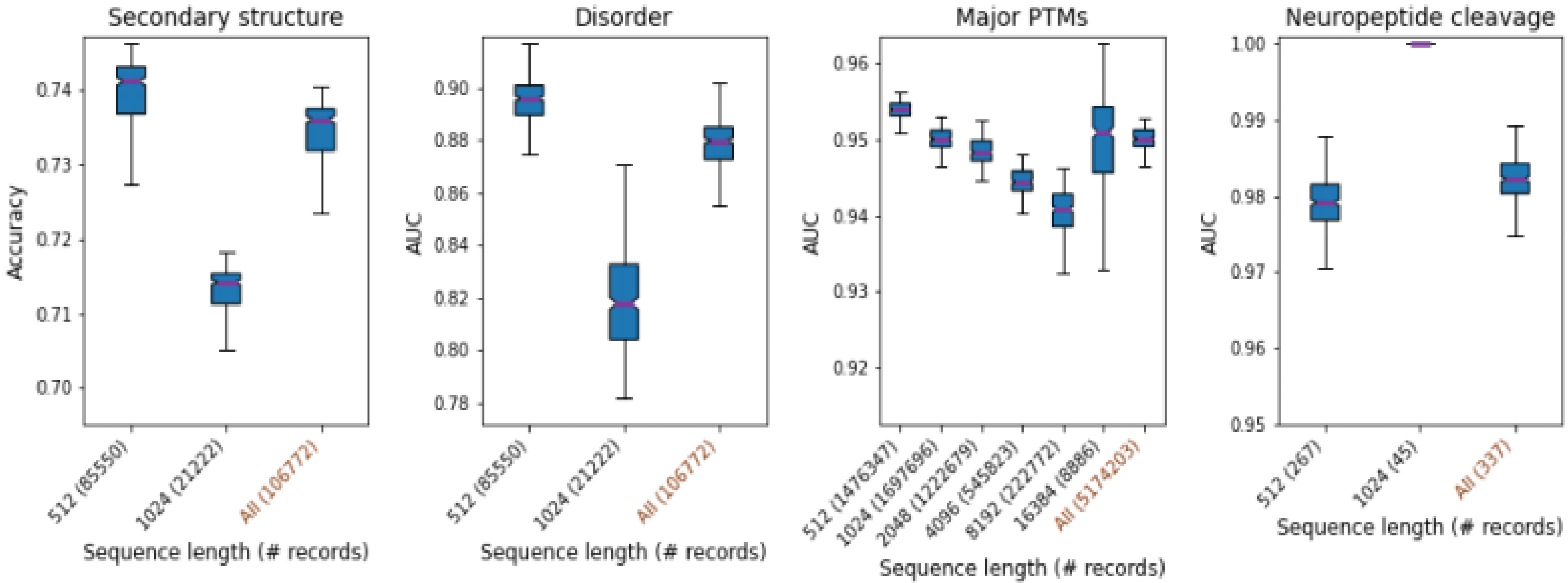
ProteinBert

4.Result



ProteinBert

4.Result



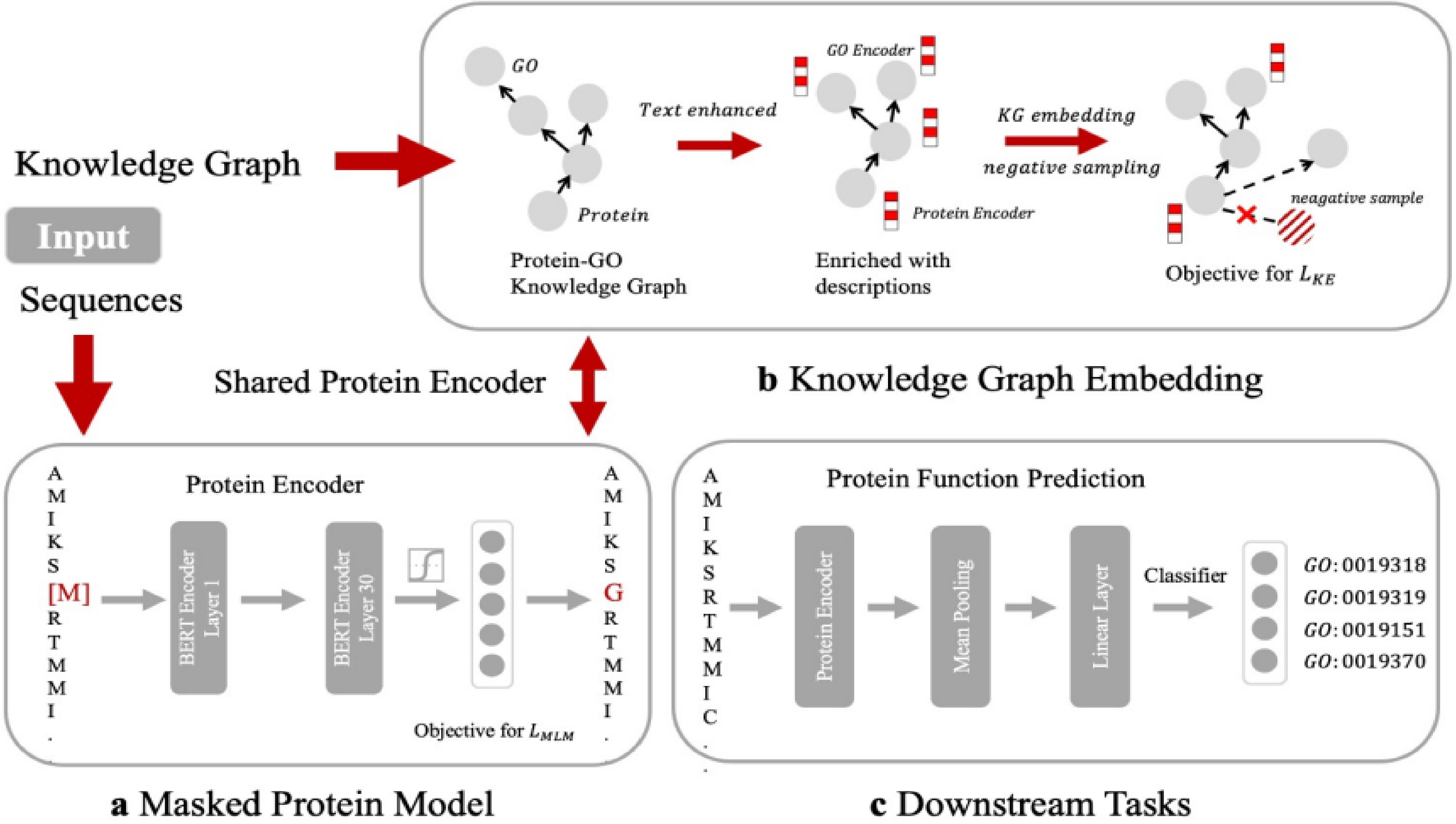
We observe that in most cases ProteinBERT performs slightly worse for longer sequences, but only modestly, showing that it indeed generalizes across a very wide range of protein lengths.

四、Ontoprotein

Model

四、OntoProtein

1.Model



Sequences dataset: UniRef100

KG dataset: ProteinKG25

2.Loss Function

We define a triplet as (h; r; t), where h and t are head and tail entities, r is the relation whose type usually is pre- defined in the schema

$$\ell_{KE} = -\log \sigma(\gamma - d(h,t)) - \sum_{i=1}^{n} \frac{1}{n} \log \sigma(d(h_i', t_i') - \gamma)$$

$$d_r(h,t) = \|h + r - t\|$$

$$\ell = \alpha \ell_{KE} + \ell_{MLM}$$

3.Experiment

1. Pre-training Dataset： To incorporate Gene Ontology knowledge into language models, we build a new pre-training dataset called ProteinKG256, which is a large-scale KG dataset with aligned descriptions and protein sequences respectively to GO terms7 and proteins entities.

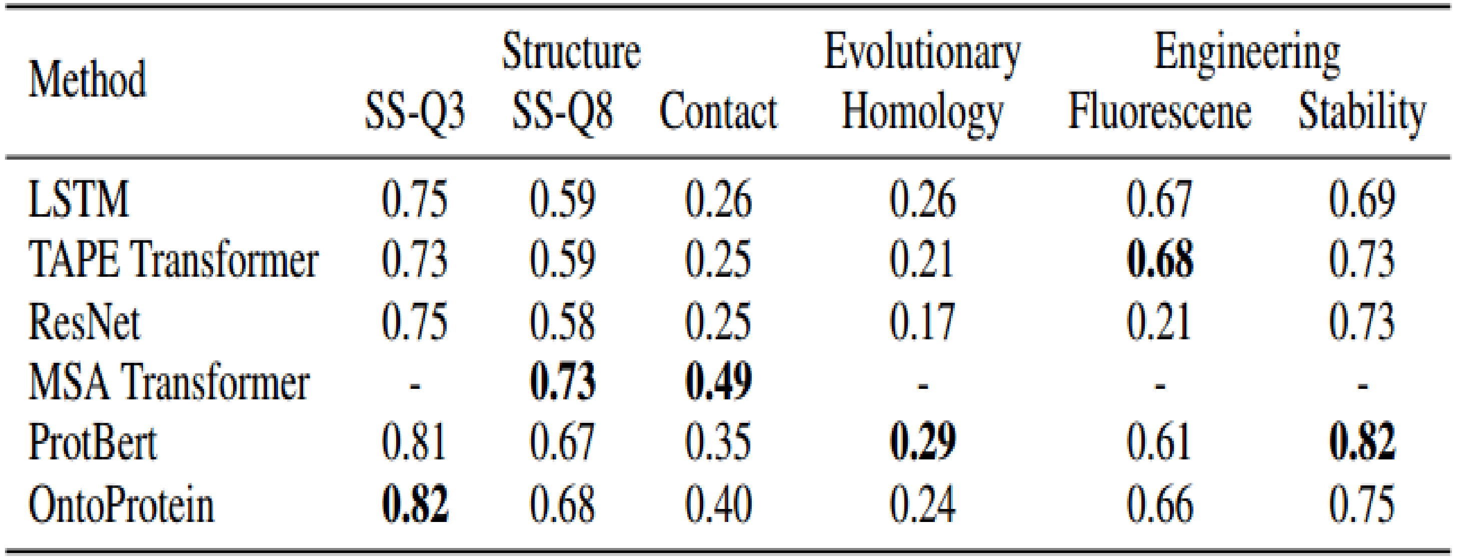
2. Downstream Task Dataset： We use TAPE as the benchmark to evaluate protein representation learning. There are three types of tasks in TAPE, including **structure, evolutionary, and engineering for proteins**. Following Rao et al. we select 6 representative datasets including secondary structure (SS), contact prediction to evaluate OntoProtein.

Protein-protein interactions (PPI) are physical contacts of high specificity established between two or more protein molecules; we regard PPI as a sequence classification task and use three datasets with different sizes for evaluation. STRING is built by Lv et al， which contains 15,335 proteins and 593,397 PPIs. We also use SHS27k and SHS148k, which are generated by Chen et al.

Protein function prediction aims to assign biological or biochemical roles to proteins, and we also regard this task as a sequence classification task. We build a new evaluation dataset based on our ProteinKG25 following the standard CAFA protocol.

OntoProtein

4.Result



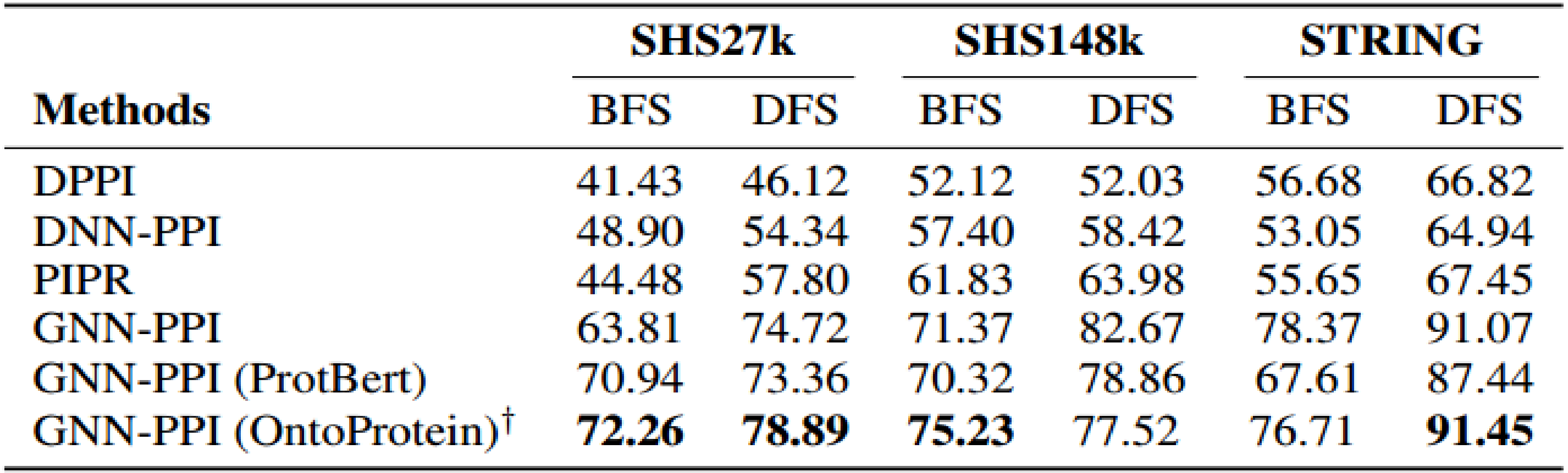
Method	Structure SS-Q3	 SS-Q8	 Contact	Evolutionary Homology	Engineering Fluorescene	 Stability
LSTM	0.75	0.59	0.26	0.26	0.67	0.69
TAPE Transformer	0.73	0.59	0.25	0.21	**0.68**	0.73
ResNet	0.75	0.58	0.25	0.17	0.21	0.73
MSA Transformer	-	**0.73**	**0.49**	-	-	-
ProtBert	0.81	0.67	0.35	**0.29**	0.61	**0.82**
OntoProtein	**0.82**	0.68	0.40	0.24	0.66	0.75

Table 1: Results on TAPE Benchmark. SS is a secondary structure task that evaluates in CB513. In contact prediction, we test medium- and long-range using P@L/2 metrics. In protein engineering tasks, we test fluorescence and stability prediction using spearman's ρ metric.

We detail the experimental result on TAPE in Table 1. Concretely, we notice that OntoProtein yields better performance in all token level tests. For the second structure and contact prediction, OntoProtein outperforms TAPE Transformer and ProtBert, showing that it can benefit from those informative biology knowledge graphs in pre-training. Moreover, OntoProtein can achieve comparable performance with MSA transformer. However, with external gene ontology knowledge injection, OntoProtein can obtain promising performance. In sequence level tasks, OntoProtein can achieve better performance than ProtBert in fluorescence prediction. However, we observe that OntoProtein does not perform well in protein engineering, homology, and stability prediction, which are all regression tasks. We think this is due to the lack of sequence-level objectives in our pre-training object.

OntoProtein

4.Result

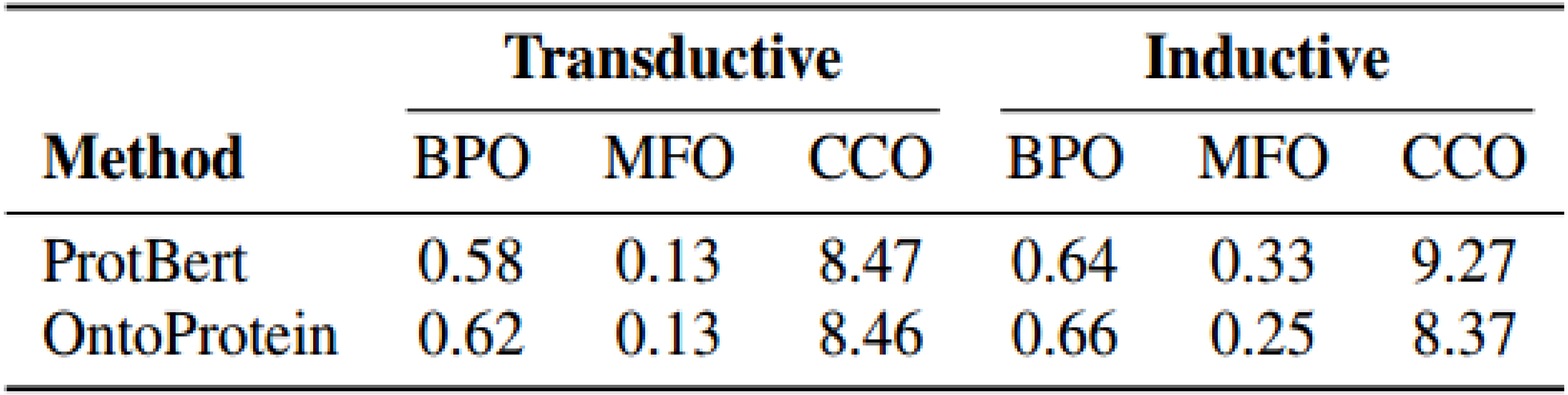


Methods	SHS27k		SHS148k		STRING	
	BFS	DFS	BFS	DFS	BFS	DFS
DPPI	41.43	46.12	52.12	52.03	56.68	66.82
DNN-PPI	48.90	54.34	57.40	58.42	53.05	64.94
PIPR	44.48	57.80	61.83	63.98	55.65	67.45
GNN-PPI	63.81	74.72	71.37	82.67	78.37	91.07
GNN-PPI (ProtBert)	70.94	73.36	70.32	78.86	67.61	87.44
GNN-PPI (OntoProtein)†	**72.26**	**78.89**	**75.23**	77.52	76.71	**91.45**

Table 2: Protein-Protein Interaction Prediction Results. Breath-First Search (BFS) and Depth-First Search (DFS) are strategies that split the training and testing PPI datasets.

From Table 2, we observe that the performance of OntoProtein is better than PIPR, which demonstrates that external structure knowledge can be beneficial for protein-protein interaction prediction. We also notice th at our method can achieve promising improvement in smaller dataset SHS2K, even outperforming GNN-PPI and GNN-PPI (ProtBert). With a larger size of datasets, OntoProtein can still obtain comparable performance to GNN-PPI and GNN-PPI (ProtBert).

4.Result

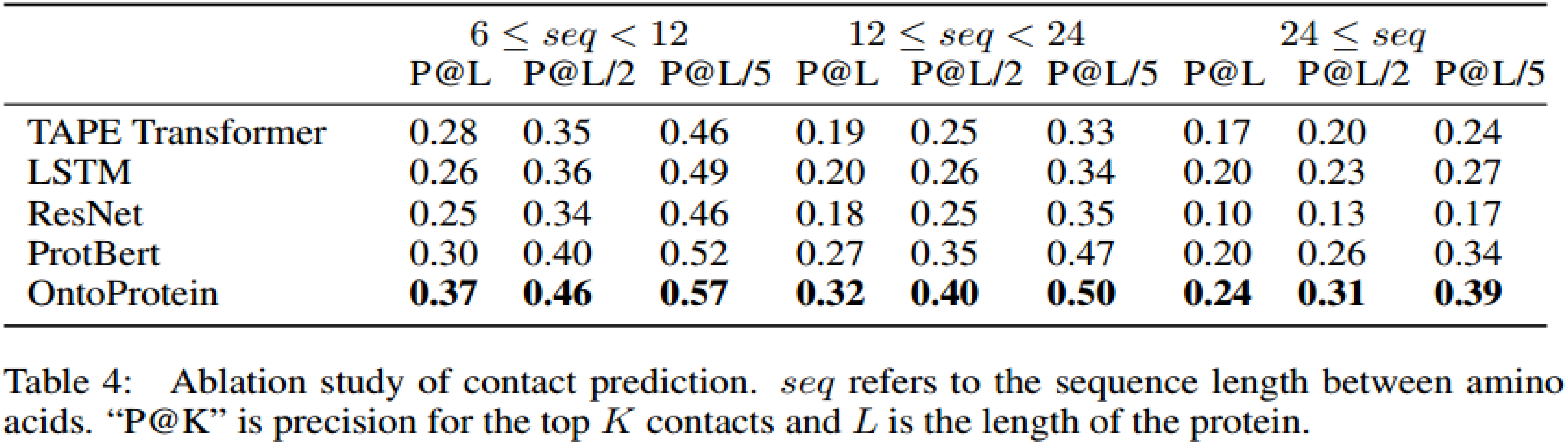


	Transductive			Inductive		
Method	BPO	MFO	CCO	BPO	MFO	CCO
ProtBert	0.58	0.13	8.47	0.64	0.33	9.27
OntoProtein	0.62	0.13	8.46	0.66	0.25	8.37

We split the test sets into three subsets (BPO, MFO, and CCO) and evaluate the performance of models separately. We notice that our OntoProtein can yield a 4% improvement with transductive setting and 2% advancement with inductive setting in BPO, further demonstrating the effectiveness of our proposed approach. We also observe that OntoProtein obtain comparable performance in other subsets. Note that there exists a severe long-tail issue in the dataset, and knowledge injecting may affect the representation learning for the head but weaken the tail representation.

MFO：分子功能, CCO： 细胞成分，BPO： 生物过程.

5.Analysis



	$6 \leq seq < 12$			$12 \leq seq < 24$			$24 \leq seq$		
	P@L	P@L/2	P@L/5	P@L	P@L/2	P@L/5	P@L	P@L/2	P@L/5
TAPE Transformer	0.28	0.35	0.46	0.19	0.25	0.33	0.17	0.20	0.24
LSTM	0.26	0.36	0.49	0.20	0.26	0.34	0.20	0.23	0.27
ResNet	0.25	0.34	0.46	0.18	0.25	0.35	0.10	0.13	0.17
ProtBert	0.30	0.40	0.52	0.27	0.35	0.47	0.20	0.26	0.34
OntoProtein	**0.37**	**0.46**	**0.57**	**0.32**	**0.40**	**0.50**	**0.24**	**0.31**	**0.39**

Table 4: Ablation study of contact prediction. *seq* refers to the sequence length between amino acids. “P@K” is precision for the top K contacts and L is the length of the protein.

Table 4 illustrates a detailed experimental analysis on the contact prediction. To further analyze the model’s performance, we conduct experiments to probe the performance of different sequences. Specifically, protein sequence lengths from short-range (6 ⩽ seq < 12) to long-range (24 ⩽ seq) are tested with three metrics (P@L, P@L/2, P@L/5). We choose several basic algorithms such as LSTM and TAPE transformer as baselines. For fairness, ProtBert is also leveraged for comparison. It can be seen that the performance of OntoProtein exceeds all other methods in all test settings, which is reasonable because the knowledge injected from Gene Ontology is beneficial.

6.Conclusion

In this paper, we take the first step to integrating external factual knowledge from gene ontology into protein language models. We present protein pretraining with gene ontology embedding (OntoProtein), which is the first general framework to integrate external knowledge graphs into protein pre-training. Experimental results on widespread protein tasks demonstrate that efficient knowledge injection helps understand and uncover the grammar of life. Besides, OntoProtein is compatible with the model parameters of lots of pre-trained protein language models, which means that users can directly adopt the available pre-trained parameters on OntoProtein without modifying the architecture. These positive results point to future work in **(1) improving OntoProtein by injecting more informative knowledge with gene ontology selection; (2) extending this approach to sequence generating tasks for protein design.**

五、ProtGPT

Model

五、ProtGPT

Introduction: ProtGPT2, an autoregressive Transformer model with 738 million parameters capable of generating de novo protein sequences in a high-throughput fashion.

Dataset:UniRef50 (UR50) (version 2021_04)

we trained a Transformer to produce a model that generates protein sequences. Language models are statistical models that assign probabilities to words and sentences. We are interested in a model that assigns high probability to sentences (W) that are semantically and syntactically correct or fit and functional, in the case of proteins. Because we are interested in a generative language model, we trained the model using an autoregressive strategy. In autoregressive models, the probability of a particular token or word (w_i) in a sequence depends solely on its context, namely the previous tokens in the sequence. The total probability of a sentence (W) is the combination of the individual probabilities for each word (w_i):

$$p(W) = \prod_{i}^{n} p(w_i | w_{<i})$$

We trained the Transformer by minimizing the negative loglikelihood over the entire dataset. More intuitively, the model must learn the relationships between a word wi —or amino acid—and all the previous ones in the sequence, and must do so for each sequence k in dataset (D):

$$\mathscr{L}_{\mathrm{CLM}} = -\sum_{k=1}^{D} log\, p_\theta\left(w_i^k | w_{<i}^k\right)$$

ProtGPT

2.Effect

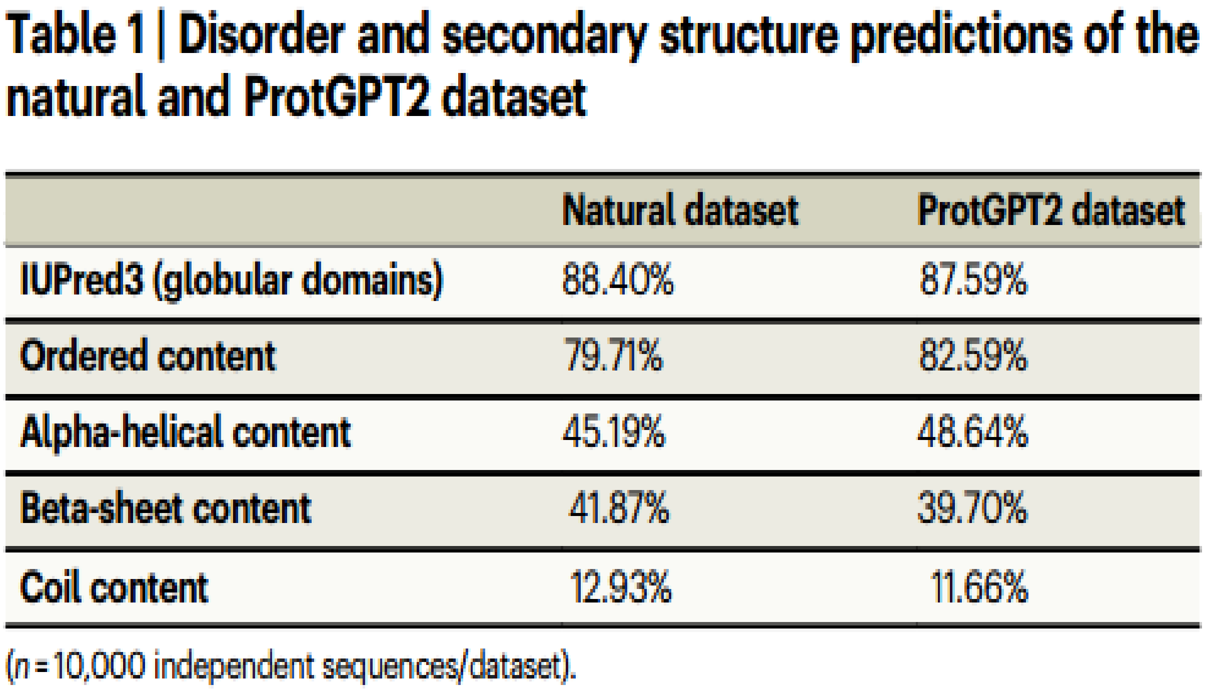


Table 1 | Disorder and secondary structure predictions of the natural and ProtGPT2 dataset

	Natural dataset	ProtGPT2 dataset
IUPred3 (globular domains)	88.40%	87.59%
Ordered content	79.71%	82.59%
Alpha-helical content	45.19%	48.64%
Beta-sheet content	41.87%	39.70%
Coil content	12.93%	11.66%

(*n* = 10,000 independent sequences/dataset).

Interestingly, our analysis shows a similar number of globular domains among the ProtGPT2-generated sequences (87.59%) and natural sequences (88.40%). Several methods have been reported that detect short intrinsically disorder regions36. Since our goal is to provide highlevel comparisons of globularity and prevalent disorder across datasets, we further performed an analysis of the protein sequences at the amino acid level using IUPred3. Remarkably, our results show a similar distribution of ordered/disordered regions for the two datasets, with 79.71 and 82.59% of ordered amino acids in the ProtGPT2 and natural datasets, respectively (Table 1). These results indicate that ProtGPT2 generates sequences that resemble globular domains whose secondary structure contents are comparable to those found in the natural space.

目录

- Background
- Model
- Conclusion

Conclusion

Through this survey, we can find that NLP technology for the design and monitoring of protein sequences is becoming more and more mature (equivalent to completing the development process of NLP for more than ten years in these three or four years). From the ProtGPT2 model just now, we can see that the large model is also about to be used to solve the task of protein sequence design. With the release of GPT4 this year, we can clearly perceive that the sequence design of large models for proteins will become a hot spot in the past two years. How to use NLP technology to design proteins and predict the structure of proteins is the direction of my efforts in the next period.

Thank you for listening