NEURAL INFORMATION
PROCESSING SYSTEMS

Protein Design with Dynamic Protein Vocabulary

Nuowei Liu*, Jiahao Kuang*, Yanting Liu

Tao Ji, Changzhi Sun, Man Lan, Yuanbin Wu

✉ nwliu@stu.ecnu.edu.cn

Correspondence to: Yuanbin Wu <ybwu@cs.ecnu.edu.cn>, Tao Ji <taoji@fudan.edu.cn>, Changzhi Sun and Man Lan

Background

Problem Design novel proteins that exhibit user-specified functions

$$p(P|t) = p((x_1, x_2, \cdots, x_k)|t, \forall i, x_i \in A)$$

Challenges

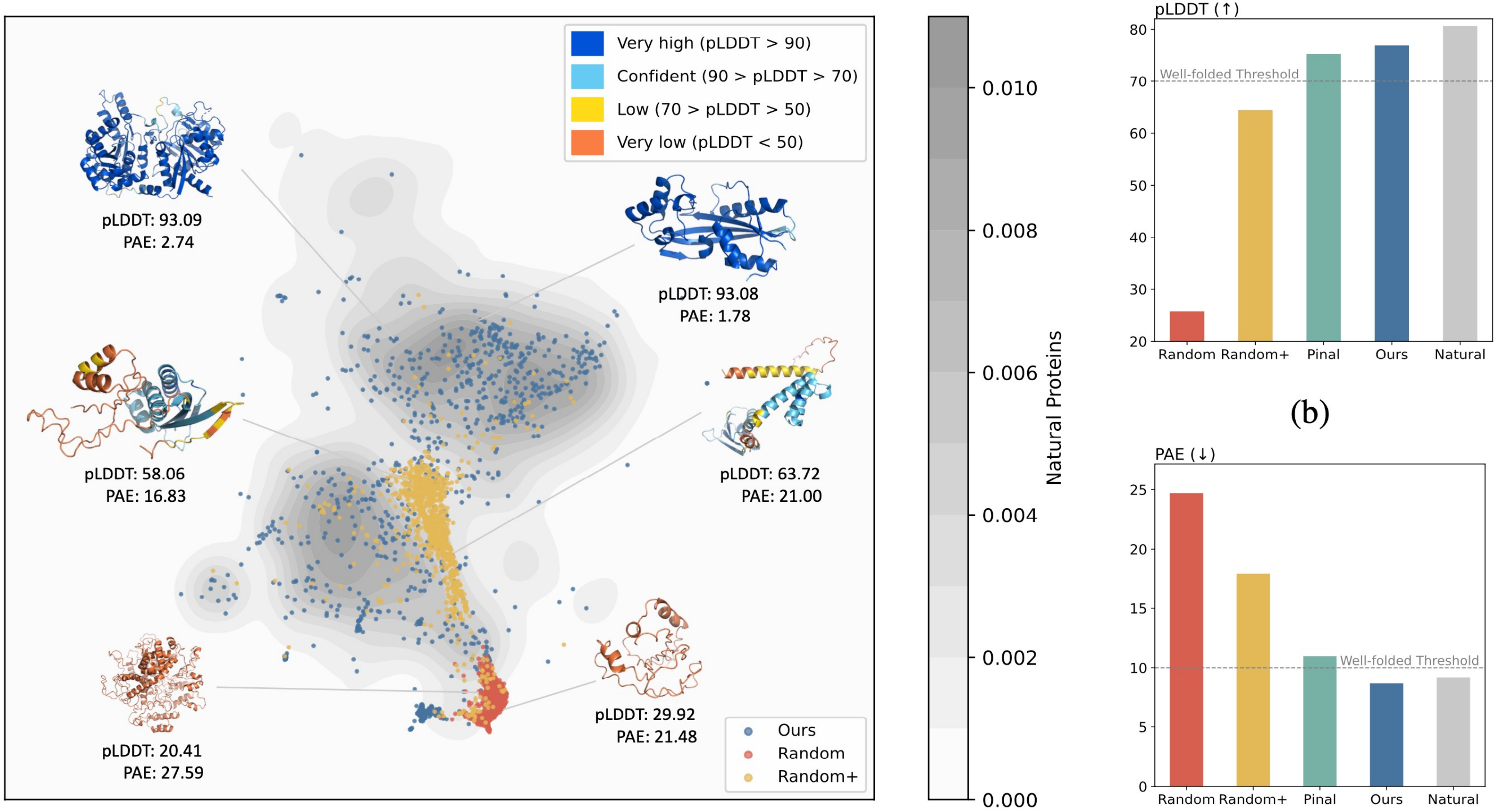
- Satisfying the requirements of **input textual descriptions**
- The designed proteins should be able to fold into **stable 3D structures**

Intuition Classical methods leverage natural protein structures

- Rational Design
- Directed Evolution

Whether well-folded novel proteins with user-specified functions can be directly assembled by utilizing fragments of natural proteins (e.g., motifs, functional sites, etc.) and their extensive functional annotations?

Background

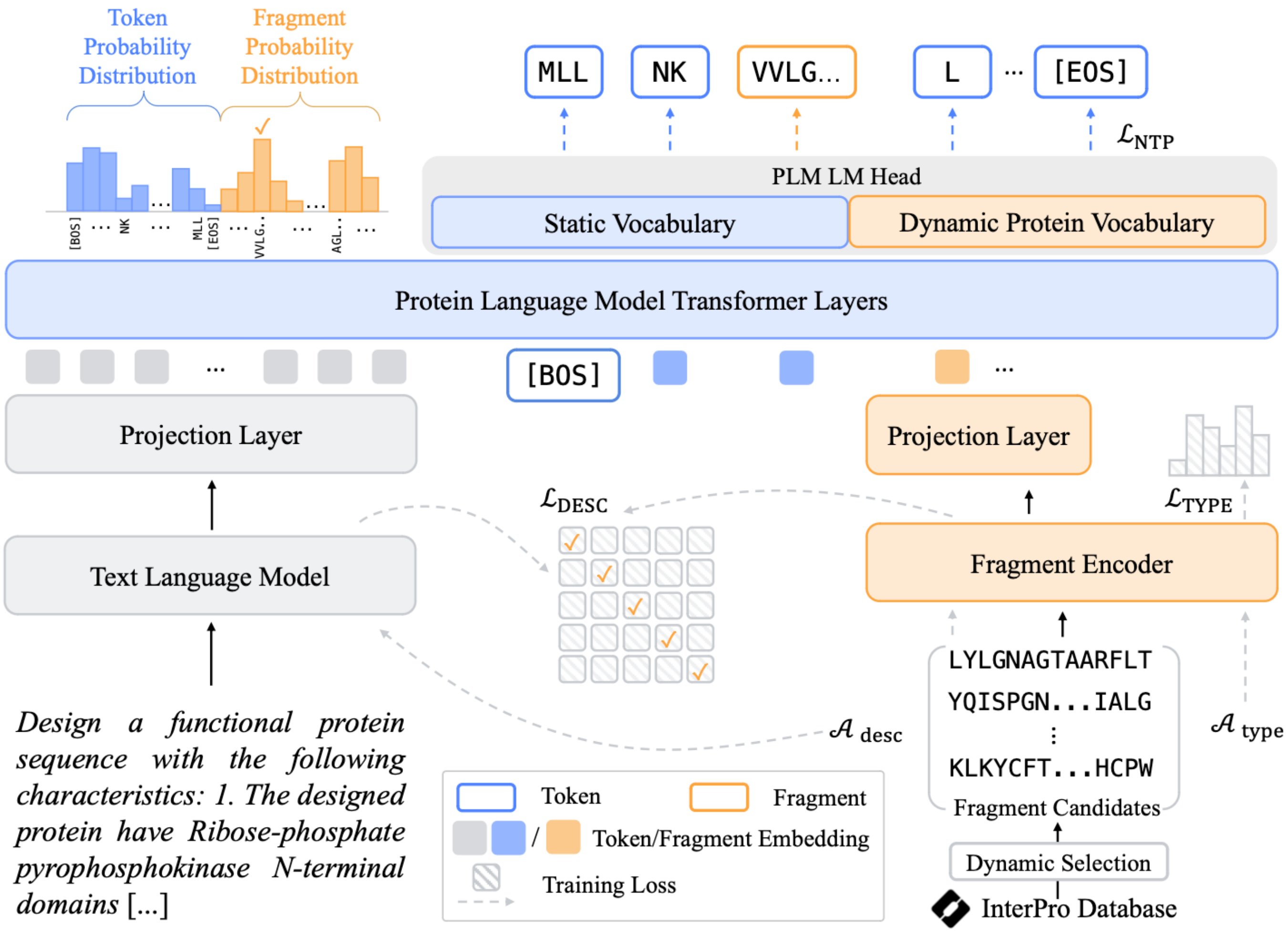


(a)

(b)

(c)

Method

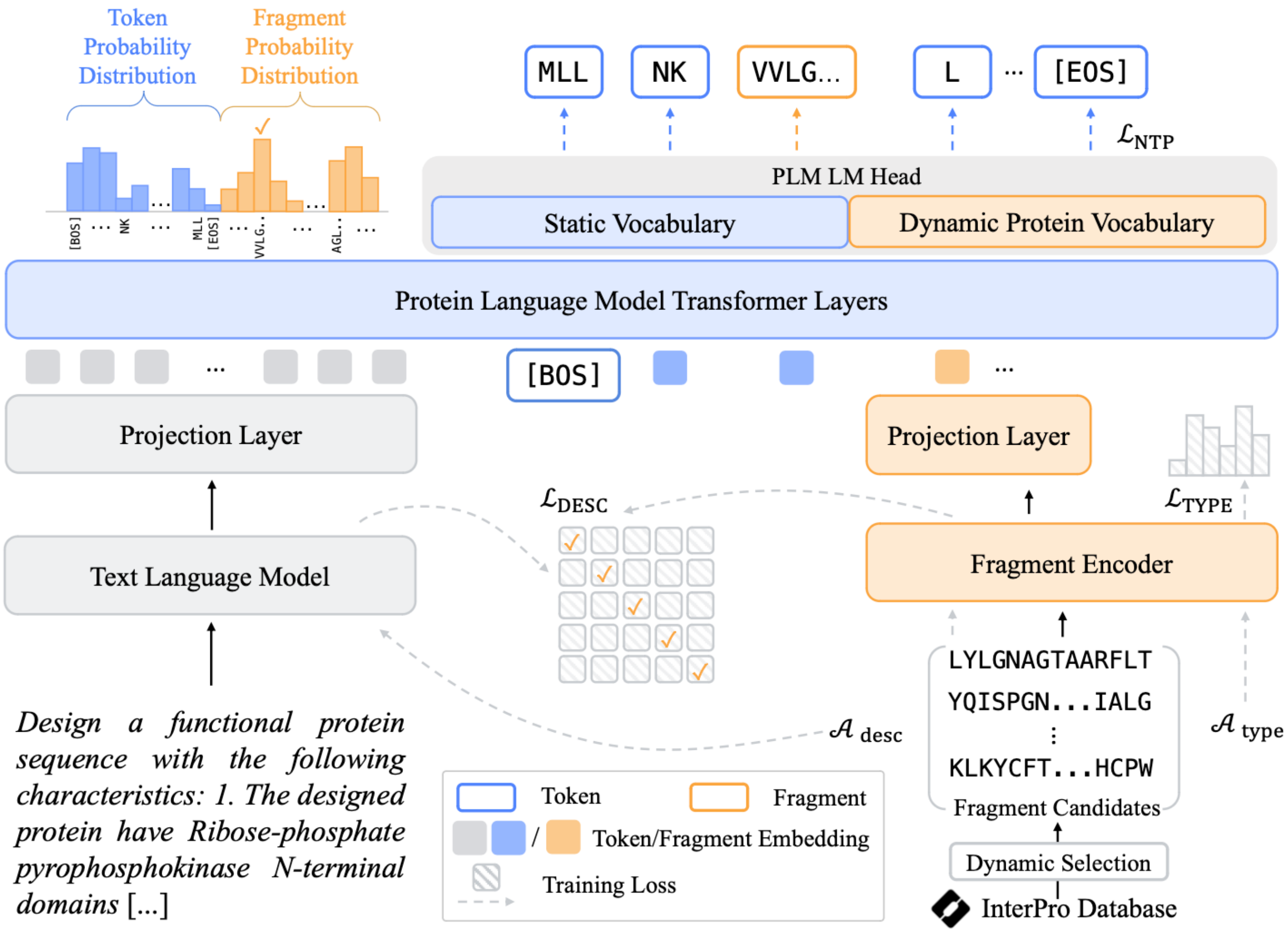

Token Probability Distribution
Fragment Probability Distribution
[BOS]
NK
MLL
[EOS]
VVLG...
AGL...
MLL
NK
VVLG...
L
[EOS]
$\mathcal{L}_{\text{NTP}}$
PLM LM Head
Static Vocabulary
Dynamic Protein Vocabulary
Protein Language Model Transformer Layers
[BOS]
Projection Layer
Projection Layer
Text Language Model
$\mathcal{L}_{\text{DESC}}$
$\mathcal{L}_{\text{TYPE}}$
Fragment Encoder
LYLGNAGTAARFLT
YQISPGN...IALG
KLKYCFT...HCPW
Fragment Candidates
Dynamic Selection
InterPro Database
$\mathcal{A}$ desc
$\mathcal{A}$ type
Design a functional protein sequence with the following characteristics: 1. The designed protein have Ribose-phosphate pyrophosphokinase N-terminal domains [...]
Token
Fragment
Token/Fragment Embedding
Training Loss

Model Architecture

- Text Language Model
- Protein Language Model
- Fragment Encoder

ProDVa (**Pro**tein Design with **D**ynamic Protein **Va**bulary)

Method



Design a functional protein sequence with the following characteristics: 1. The designed protein have Ribose-phosphate pyrophosphokinase N-terminal domains [...]

ProDVa (**Pro**tein Design with **D**ynamic Protein **Va**bulary)

Training Objectives

- Learning Next Token/Fragment Prediction
- Learning Functional Annotations

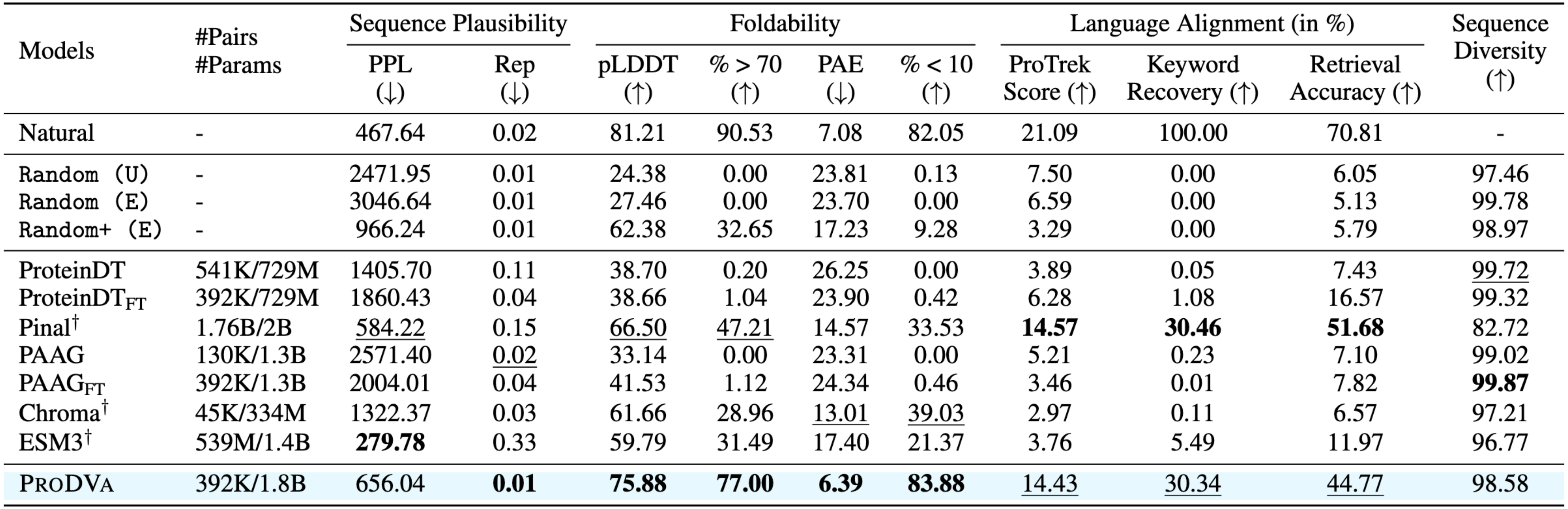
$$\mathcal{L} = \mathcal{L}_{\text{NTP}} + \alpha\mathcal{L}_{\text{TYPE}} + \beta\mathcal{L}_{\text{DESC}}$$

Inference

- Retrieving the top K most relevant descriptions
- Constructing the fragment candidates

Experiments

Designing Proteins from Function Keywords



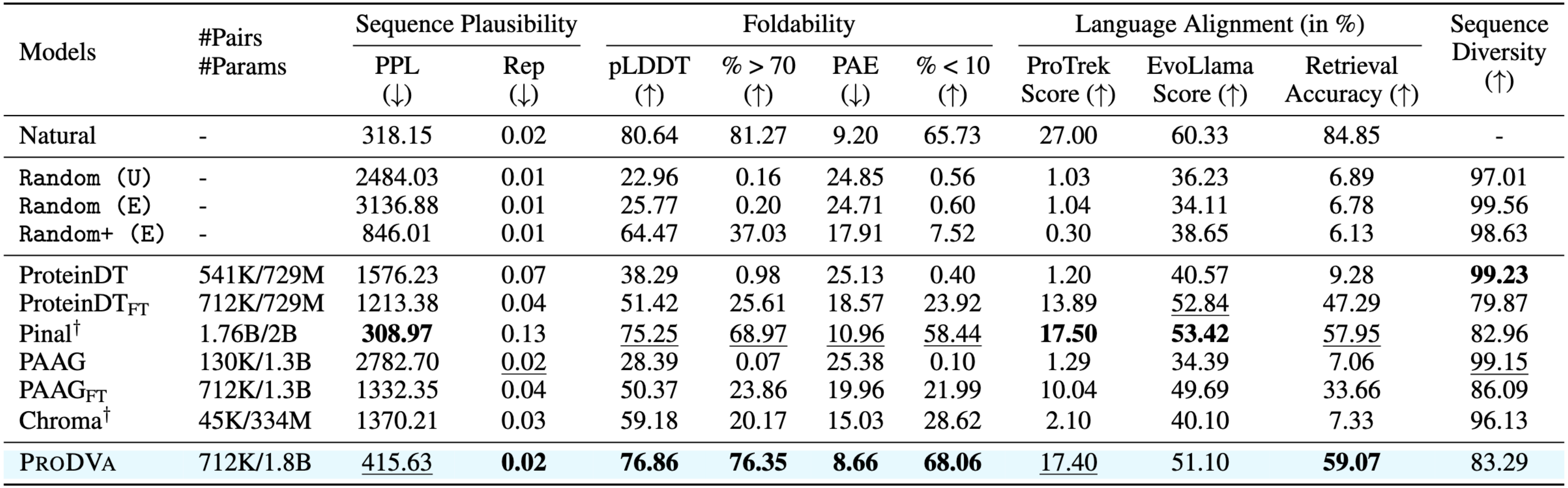
Models	#Pairs #Params	Sequence Plausibility		Foldability				Language Alignment (in %)			Sequence Diversity (↑)
		PPL (↓)	Rep (↓)	pLDDT (↑)	% > 70 (↑)	PAE (↓)	% < 10 (↑)	ProTrek Score (↑)	Keyword Recovery (↑)	Retrieval Accuracy (↑)	
Natural	-	467.64	0.02	81.21	90.53	7.08	82.05	21.09	100.00	70.81	-
`Random (U)`	-	2471.95	0.01	24.38	0.00	23.81	0.13	7.50	0.00	6.05	97.46
`Random (E)`	-	3046.64	0.01	27.46	0.00	23.70	0.00	6.59	0.00	5.13	99.78
`Random+ (E)`	-	966.24	0.01	62.38	32.65	17.23	9.28	3.29	0.00	5.79	98.97
ProteinDT	541K/729M	1405.70	0.11	38.70	0.20	26.25	0.00	3.89	0.05	7.43	99.72
ProteinDT$_{FT}$	392K/729M	1860.43	0.04	38.66	1.04	23.90	0.42	6.28	1.08	16.57	99.32
Pinal†	1.76B/2B	584.22	0.15	66.50	47.21	14.57	33.53	**14.57**	**30.46**	**51.68**	82.72
PAAG	130K/1.3B	2571.40	0.02	33.14	0.00	23.31	0.00	5.21	0.23	7.10	99.02
PAAG$_{FT}$	392K/1.3B	2004.01	0.04	41.53	1.12	24.34	0.46	3.46	0.01	7.82	**99.87**
Chroma†	45K/334M	1322.37	0.03	61.66	28.96	13.01	39.03	2.97	0.11	6.57	97.21
ESM3†	539M/1.4B	**279.78**	0.33	59.79	31.49	17.40	21.37	3.76	5.49	11.97	96.77
ProDVa	392K/1.8B	656.04	**0.01**	**75.88**	**77.00**	**6.39**	**83.88**	14.43	30.34	44.77	98.58

Key Findings

- Under the same training data setting, ProDVa consistently surpasses both ProteinDT and PAAG
- ProDVa remains within a **reasonable PPL range** and demonstrates the capability to **design well-folded proteins**
- ProDVa uses only **0.02%** of the text-protein pairs used to train Pinal, yet achieves competitive performance

Experiments

Designing Proteins from Textual Descriptions



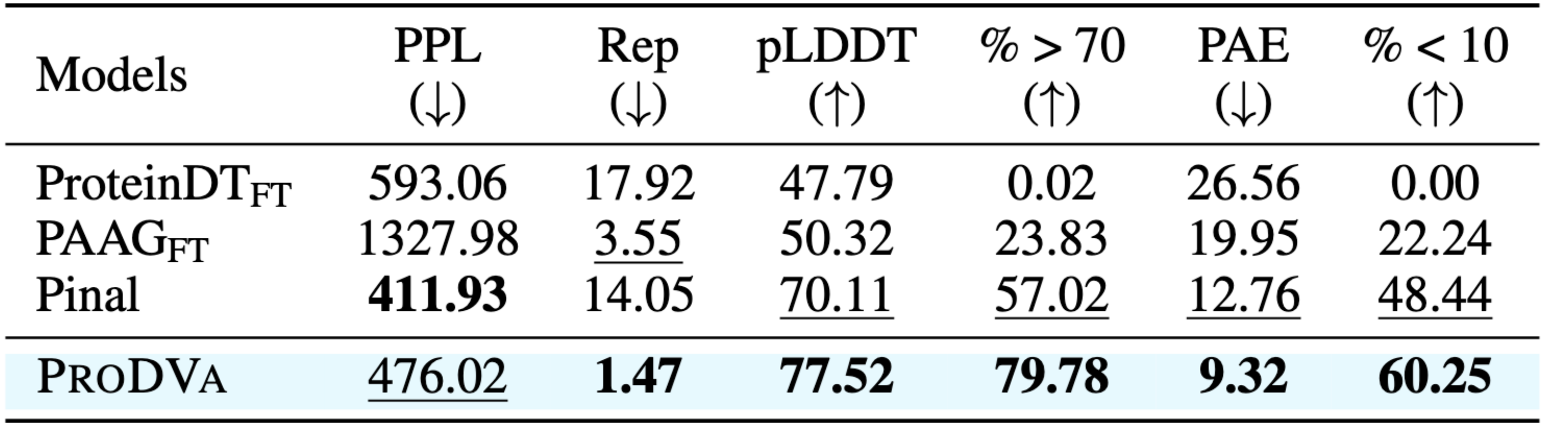
Models	#Pairs #Params	Sequence Plausibility		Foldability				Language Alignment (in %)			Sequence Diversity (↑)
		PPL (↓)	Rep (↓)	pLDDT (↑)	% > 70 (↑)	PAE (↓)	% < 10 (↑)	ProTrek Score (↑)	EvoLlama Score (↑)	Retrieval Accuracy (↑)	
Natural	-	318.15	0.02	80.64	81.27	9.20	65.73	27.00	60.33	84.85	-
`Random (U)`	-	2484.03	0.01	22.96	0.16	24.85	0.56	1.03	36.23	6.89	97.01
`Random (E)`	-	3136.88	0.01	25.77	0.20	24.71	0.60	1.04	34.11	6.78	99.56
`Random+ (E)`	-	846.01	0.01	64.47	37.03	17.91	7.52	0.30	38.65	6.13	98.63
ProteinDT	541K/729M	1576.23	0.07	38.29	0.98	25.13	0.40	1.20	40.57	9.28	**99.23**
ProteinDT$_{\text{FT}}$	712K/729M	1213.38	0.04	51.42	25.61	18.57	23.92	13.89	52.84	47.29	79.87
Pinal†	1.76B/2B	**308.97**	0.13	75.25	68.97	10.96	58.44	**17.50**	**53.42**	57.95	82.96
PAAG	130K/1.3B	2782.70	0.02	28.39	0.07	25.38	0.10	1.29	34.39	7.06	99.15
PAAG$_{\text{FT}}$	712K/1.3B	1332.35	0.04	50.37	23.86	19.96	21.99	10.04	49.69	33.66	86.09
Chroma†	45K/334M	1370.21	0.03	59.18	20.17	15.03	28.62	2.10	40.10	7.33	96.13
PRODVA	712K/1.8B	415.63	**0.02**	**76.86**	**76.35**	**8.66**	**68.06**	17.40	51.10	**59.07**	83.29

Key Findings

- Most baselines struggle to design proteins that are **both well-folded and well-aligned**
- Incorporating additional data may potentially improve performance, particularly in terms of language alignment
- ProDVa demonstrates **competitive sequence diversity** compared to other baselines

Experiments

Unconditional Protein Generation



Models	PPL (↓)	Rep (↓)	pLDDT (↑)	% > 70 (↑)	PAE (↓)	% < 10 (↑)
$\text{ProteinDT}_{\text{FT}}$	593.06	17.92	47.79	0.02	26.56	0.00
PAAG_{FT}	1327.98	3.55	50.32	23.83	19.95	22.24
Pinal	**411.93**	14.05	70.11	57.02	12.76	48.44
PRODVA	476.02	**1.47**	**77.52**	**79.78**	**9.32**	**60.25**

- Fixing the input instruction to `Design a novel protein sequence`
- Replacing the retrieval method with the random selection of fragments

Key Findings

- ProDVa **outperforms all baseline models** on the unconditional protein generation task
- Compared to other fine-tuned models, ProDVa achieves **substantially superior performance**

Thank You!

GitHub Repo

🤗 HuggingFace