

2024-07-17

TODO

- ☒ ~~Analizar artículo citas TAPE y PEER CASP~~
- ☒ ~~Revisar TAPE PEER CASP~~
- ☒ ~~Meterse a fondo en las tareas establecidas~~
- ☐ Buscar resultados bajos:razones
- ☒ ~~Revisión de participación en CASP: todo bien o una única cosa bien~~
- ☐ Métodos relevantes (?)
- ☒ ~~Leer AF2 y AF3~~
- ☐ 5 Benchmarks Ahondar en trabajos

citas TAPE y PEER CASP

Unified Benchmarks and Evaluation Protocols

Initiatives like TAPE [145], ProteinGym [303], ProteinShake [315], PEER [316], ProteinInvBench [317], ProteinWorkshop [318], exemplify the critical role of comprehensive benchmarking in furthering innovation. Moreover, the Critical Assessment of Protein Structure Prediction (CASP) experiments act as a crucial platform for evaluating the most recent advancements in PSP [Survey 2024]

Revisar TAPE y PEER CASP

TAPE

Task Assessing Preprotein embeddings is a set of 5 biologically relevant Semisupervised learning Tasks spread across different domains of protein biology:

First attempt at systematically evaluating semisupervised learning on protein sequences.

Los modelos se preentrenaron utilizando un corpus no etiquetado de secuencias de proteínas del conjunto de datos Pfam.

Para cada tarea se utilizó una arquitectura (drawing from state-of-the-art where available) y el fine-tuning en todos los modelos fue el mismo.

1. Protein Structure Prediction
2. Evolutionary Understanding

3. Protein Engineering

Table 2: Results on downstream supervised tasks

Method		Structure		Evolutionary	Engineering	
		SS	Contact	Homology	Fluorescence	Stability
No Pretrain	Transformer	0.70	0.32	0.09	0.22	-0.06
	LSTM	0.71	0.19	0.12	0.21	0.28
	ResNet	0.70	0.20	0.10	-0.28	0.61
Pretrain	Transformer	0.73	0.36	0.21	0.68	0.73
	LSTM	0.75	0.39	0.26	0.67	0.69
	ResNet	0.75	0.29	0.17	0.21	0.73
Supervised [11] UniRep [12]	LSTM	0.73	0.40	0.17	0.33	0.64
	mLSTM	0.73	0.34	0.23	0.67	0.73
Baseline	One-hot	0.69	0.29	0.09	0.14	0.19
	Alignment	0.80	0.64	0.09	N/A	N/A

PEER

Peer benchmark includes 14 biologically relevant tasks (Now 17, <https://torchprotein.ai/benchmark>) that covers diverse aspects of protein understanding, including:

1. Protein Function Prediction
2. Protein Localization Prediction
3. Protein Structure Prediction
4. Protein-Protein interaction Prediction
5. Protein-Ligand Prediction

Para cada tarea se evaluó el performance de diferentes tipos de:

sequence-based approaches

sequence encoders: CNNs LSTMs Transformers pLLM

Se evaluaron diferentes approaches under the multitask learning setting

En cada benchmark task cada data set se diseñó adecuadamente para la tarea en cuestión

Also evaluate different approaches under multi-task learning setting

Task (Acronym)	Task Category	Data Source	#Protein	Seq. len.	#Train/Validation/Test	Metric
Function Prediction						
GB1 fitness prediction (GB1)	Protein-wise Reg.	FLIP [16]	8,733	378.6 _(0.9)	381/43/8,309	Spearman's ρ
AAV fitness prediction (AAV)	Protein-wise Reg.	FLIP [16]	82,583	1033.0 _(3.4)	28,626/3,181/50,776	Spearman's ρ
Thermostability prediction (Thermo)	Protein-wise Reg.	FLIP [16]	7,158	880.6 _(974.2)	5,149/643/1,366	Spearman's ρ
Fluorescence prediction (Flu)	Protein-wise Reg.	Sarkisyan's dataset [71]	54,025	343.3 _(1.3)	21,446/5,362/27,217	Spearman's ρ
Stability prediction (Sta)	Protein-wise Reg.	Rocklin's dataset [66]	68,934	66.6 _(5.2)	53,571/2,512/12,851	Spearman's ρ
β -lactamase activity prediction (β -lac)	Protein-wise Reg.	Envision [25]	5,198	396.1 _(0.7)	4,158/520/520	Spearman's ρ
Solubility prediction (Sol)	Protein-wise Cls.	DeepSol [39]	71,419	424.1 _(225.9)	62,478/6,942/1,999	Acc
Localization Prediction						
Subcellular localization prediction (Sub)	Protein-wise Cls.	DeepLoc [2]	13,961	665.3 _(395.3)	8,945/2,248/2,768	Acc
Binary localization prediction (Bin)	Protein-wise Cls.	DeepLoc [2]	8,634	636.5 _(396.5)	5,161/1,727/1,746	Acc
Structure Prediction						
Contact prediction (Cont)	Residue-pair Cls.	ProteinNet [3]	25,563	320.0 _(275.2)	25,299/224/40	L/5 precision
Fold classification (Fold)	Protein-wise Cls.	DeepSF [31]	13,766	235.4 _(155.1)	12,312/736/718	Acc
Secondary structure prediction (SSP)	Residue-wise Cls.	NetSurfP-2.0 [41]	11,361	360.5 _(229.3)	8,678/2,170/513	Acc
Protein-Protein Interaction Prediction						
Yeast PPI prediction (Yst)	Protein-pair Cls.	Guo's dataset [26]	1,707	726.3 _(432.0)	1,668/131/373	Acc
Human PPI prediction (Hum)	Protein-pair Cls.	Pan's dataset [59]	5,553	727.7 _(438.2)	6,844/277/227	Acc
PPI affinity prediction (Aff)	Protein-pair Reg.	SKEMPI [56]	627	304.9 _(193.8)	2,127/212/343	RMSE
Protein-Ligand Interaction Prediction						
Affinity prediction on PDBbind (PDB)	Protein-ligand Reg.	PDBbind [49]	10,607	414.9 _(234.3)	16,436/937/285	RMSE
Affinity prediction on BindingDB (BDB)	Protein-ligand Reg.	BindingDB [47]	1,006	799.8 _(417.0)	7,900/878/5,230	RMSE

Resultados importantes: Los modelos preentrenados como ProtBert y ESM-1b tuvieron el mejor rendimiento en la mayoría de las tareas individuales, y Entrenar múltiples tareas conjuntamente mejoró el rendimiento de los modelos

CASP

The most well-known protein Benchmark. Focuses in structure modeling

Se centra en Protein Structure Prediction y nada más

Se realiza cada dos años, el más actual es CASP XV

Comenzó en 1994

Alpha Fold 2: supera a todos pero se combina con otros métodos para alcanzar buenos resultados. Usando los parámetros estándar del modelo, sólo produce los mejores resultados para 2/3 de los targets

El segundo mejor performance es rosettafold

En general, los modelos mostraron un rendimiento inferior en varios casos específicos, incluyendo **proteínas con baja homología, complejos de proteínas, proteínas con alta movilidad estructural y proteínas de membrana**. → Hay que mejorar los métodos de predicción para abordar la complejidad y variabilidad de las estructuras proteicas en diferentes contextos.

Resumen

Tasks

Aa Tarea	Task group	Definición	Paper	Métrica	State-of-art	Property
Secondary Structure Prediction	Structure Prediction	Predecir la estructura secundaria	PEER TAPE	Accuracy		sequence-to-sequence






Aa Tarea	Task group	Definición	Paper	Métrica	State-of-art	Property
		(hélice, hoja o otra) de cada aminoácido en una secuencia de proteína.				
Contact Prediction	Structure Prediction	Predecir si pares de aminoácidos en una secuencia de proteína están en contacto (a menos un δ de distancia)	PEER TAPE	L/5		binary clasification $(x_i, x_j) \rightarrow y_{\{i,j\}} \in \{0,1\}$
Remote Homology Detection	evolutionary understanding	Clasificar secuencias de proteínas en una de 1195 posibles estructuras de pliegues	TAPE	Accuracy		multi clasification $x \rightarrow y$
Fluorescence (Landscape). Prediction	Function Prediction engineering	(TAPE)Determina la intensidad de la log-fluorescence (PEER)Predice la intensidad de fluorescencia de mutantes de la proteína verde fluorescente	PEER TAPE	spearman		regression $x \rightarrow y$
Stability (Landscape). Prediction	Function Prediction engineering	(TAPE)cada proteína de entrada x se mapea a una etiqueta y que mide las circunstancias más extremas en las que la prot mantiene su estructura por encima de un umbral de concentración (PEER) Evalúa la estabilidad de las proteínas bajo condiciones naturales. y indica la medida	PEER TAPE	spearman		regression $x \rightarrow y$

Aa Tarea	Task group	Definición	Paper	Métrica	State-of-art	Property
		experimental de estabilidad				
<u>β-Lactamase Activity Prediction</u>	Function Prediction		PEER			
<u>Solubility Prediction</u>	Function Prediction		PEER			
<u>Subcellular Localization Prediction</u>	Localization Prediction		PEER			
<u>Binary Localization Prediction</u>	Localization Prediction		PEER			
<u>Fold Clasification</u>	Structure Prediction		PEER			
<u>Untitled</u>			PEER			
<u>Untitled</u>			PEER			
<u>Untitled</u>			PEER			
<u>Untitled</u>			PEER			
<u>Untitled</u>			PEER			
<u>Untitled</u>			PEER			
<u>Untitled</u>			PEER			

Métricas	Descripción	Paper
GDT_TS	evalúa la fracción de residuos en la estructura predicha que se encuentran dentro de ciertos umbrales de distancia (1, 2, 4 y 8 Å) de la estructura experimental.	CASP

Otros Benchmarks

Papers	Link	Fecha	descripcion	Paper ref
<u>PEER</u>				Surver24
<u>ProteinWorkshop</u>				Surver24

 Papers	 Link	 Fecha	 descripcion	 Paper ref
ProteinInvBench				Surver24
ProteinShake				Surver24
ProteinGym				Surver24
Proteinglue			uilds a benchmark containing 7 downstream tasks for evaluating self-supervised protein representation learning	PEER
ATOM3D			provides benchmark datasets for 3D structure based biomolecule understanding.	PEER
TDA			contains protein-related datasets and tasks for drug discovery.	PEER
functional properties (nature).			focuses on the evaluation of unsupervised protein representations and evaluates 23 typical methods	PEER
FLIP			proposes three protein landscape benchmarks for fitness prediction evaluation	PEER
TAPE			comprehensively compare different machine learning methods, is built on five tasks spread across different domains of protein biology and evaluate the performance of protein sequence encoders	PEER Surver24
CAFA			Is held for the evaluation of PFP	PEER
CASP			Focuses on PSP (golden standard)	PEER Surver24

Revisión de participación en casp: todo bien o una única cosa bien