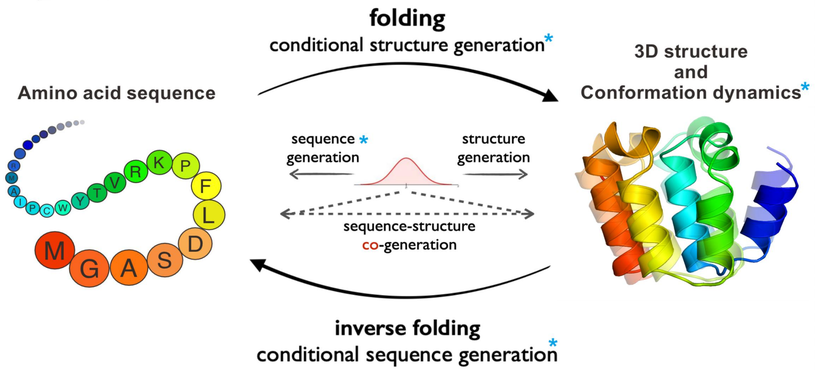
ProteinBench: A Holistic Evaluation of Protein Foundation Models

Fei Ye, Zaixiang Zheng*, Dongyu Xue*, Yuning Shen*, Lihao Wang*, Yiming Ma, Yan Wang, Xinyou Wang, Xiangxin Zhou and Quanquan Gu#*

** Equal Contribution # Corresponding Author*

Motivation



1. A taxonomic classification of tasks encompassing the main generative challenges in the protein domain.
2. A multi-metric evaluation approach assessing performance across four key dimensions: quality, novelty, diversity, and robustness.
3. In-depth analyses from various user objectives, providing a holistic view of model performance.
4. Leaderboard and code framework.

Conclusions

1. Valid evaluation of protein foundation models necessitates the **use of correct and comprehensive evaluation metrics**. Certain tasks may still lack sufficiently accurate evaluation methods.
2. No single model currently excels across all protein design objectives. The choice of model should be **carefully aligned with the intended applications**.
3. While generative models extended from classic folding models have shown ability to sample protein conformations, **challenges remain in both multiple-state prediction and distribution prediction.**

Limitations and Future work

1. The selection of foundation models may not be exhaustive. Future iterations should incorporate **additional foundation models** to provide a more comprehensive comparison.
2. Inconsistencies in training data across models currently hinder direct comparisons of different model architectures. Future work could address this by **standardizing datasets**, allowing for more accurate comparisons of architectural performance.
3. The benchmark could be expanded to include **a wider range of tasks**, further broadening its scope and utility.

Protein Design

1. Inverse Folding

Model	Fitting Evolution Distribution		De novo backbones based sequence design									
	CASP AAR ↑	CAMEO AAR ↑	length 100 scTM ↑	length 100 pLDDT ↑	length 200 scTM ↑	length 200 pLDDT ↑	length 300 scTM ↑	length 300 pLDDT ↑	length 400 scTM ↑	length 400 pLDDT ↑	length 500 scTM ↑	length 500 pLDDT ↑
ProteinMPNN	0.450	0.468	**0.962**	**94.14**	**0.945**	**89.34**	**0.962**	**90.28**	**0.875**	**83.76**	**0.568**	**67.09**
ESM-IF1	N/A	N/A	0.810	88.83	0.635	69.67	0.336	74.36	0.449	64.59	0.462	58.97
LM-Design	**0.516**	**0.570**	0.834	78.45	0.373	58.41	0.481	69.86	0.565	59.87	0.397	56.35
ESM3	N/A	N/A	0.942	86.60	0.486	60.69	0.632	70.78	0.564	62.63	0.452	59.37

2. Structure Design

Model	length 300					length 500				
	Quality		Novelty	Diversity		Quality		Novelty	Diversity	
	scTM ↑	scRMSD ↓	Max TM ↓	pairwise TM ↓	Max Clust. ↑	scTM ↑	scRMSD ↓	Max TM ↓	pairwise TM ↓	Max Clust. ↑
Native PDBs	0.97	0.82	N/A	0.28	0.77	0.97	1.07	N/A	0.29	0.80
RFdiffusion	**0.96**	**1.03**	0.64	**0.36**	0.65	**0.79**	**5.60**	0.62	0.33	0.89
FrameFlow	0.92	1.95	0.65	0.43	0.88	0.61	7.92	0.61	0.40	0.92
Chroma	0.87	2.47	0.66	**0.36**	0.67	0.72	6.71	0.60	**0.29**	**0.99**
FrameDiff(latest)	0.87	2.73	0.69	0.48	0.21	0.63	9.52	**0.58**	0.40	0.52
FoldFlow1(sfm)	0.45	9.04	0.54	0.39	1.00	0.37	13.04	0.53	0.37	1.00
FoldFlow1(base)	0.43	9.56	0.54	0.39	0.98	0.35	13.20	0.52	0.39	1.00
FoldFlow1(ot)	0.54	8.21	**0.58**	0.41	**0.94**	0.37	12.48	0.51	0.35	1.00
Genie	0.27	20.37	0.30	0.23	1.00	0.25	26.08	0.22	0.23	1.00

3: Sequence Design

Model	length 300					length 500				
	Quality		Diversity		Novelty	Quality		Diversity		Novelty
	ppl ↓	pLDDT ↑	pairwise TM ↓	Max Clust. ↑	Max TM ↓	ppl ↓	pLDDT ↑	pairwise TM ↓	Max Clust. ↑	Max TM ↓
Native Seqs		61.49	0.51	0.85	N/A		62.95	0.51	0.78	N/A
Progen 2 (700M)	6.25	65.69	0.42	0.93	**0.66**	4.27	61.45	0.32	0.95	0.68
EvoDiff	17.13	45.14	**0.31**	**1.00**	0.68	16.51	43.14	**0.31**	**1.00**	0.69
DPLM (650M)	**3.47**	**93.07**	0.57	0.63	0.91	**3.33**	**87.73**	0.43	0.85	0.85
ESM3 (1.4B)	14.59	48.08	0.32	**1.00**	0.75	11.10	52.17	0.30	**1.00**	**0.54**

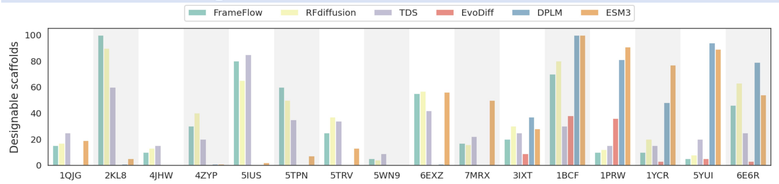
4: Sequence and structure Co-Design

Model	Length 300				Length 500			
	Quality		Diversity	Novelty	Quality		Diversity	Novelty
	scTM ↑	scRMSD ↓	Max Clust. ↑	Max TM ↓	scTM ↑	scRMSD ↓	Max Clust. ↑	Max TM ↓
Native PDBs	0.92	3.94	0.75	NaN	0.9	9.64	0.8	NaN
ProteinGenerator	0.81	9.26	0.22	0.71	0.41	33.91	0.18	0.73
ProtPardelle*	0.69	14.91	0.04	0.72	0.4	41.23	0.6	0.69
Multiflow	0.96	2.14	0.58	0.71	0.83	8.48	0.67	0.68
ESM3 (1.4B)*	0.59	25.5	0.52	0.73	0.54	33.7	0.37	0.77

5: Antibody Design

Method	Accuracy		Functionality		Specificity	
	AAR ↑	RMSD ↓	TM-score ↑	Binding Energy ↓	SeqSim-outer ↓	SeqSim-inner ↑	PHR ↓
RAbD (natural)	100.00%	0.00	1.00	-15.33	0.26	N/A	45.78%
HERN	33.17%	9.86	0.16	1242.77	0.41	N/A	39.83%
MEAN	33.47%	**1.82**	0.25	263.90	0.65	N/A	40.74%
dyMEAN	**40.95%**	2.36	0.36	889.28	0.58	N/A	42.04%
*dyMEAN-FixFR	40.05%	2.37	0.35	612.75	0.60	**0.96**	43.75%
*DiffAb	35.04%	2.53	**0.37**	489.42	**0.37**	0.45	40.68%
*AbDPO	31.29%	2.79	0.35	**116.06**	0.38	0.60	69.69%
*AbDPO++	36.25%	2.48	0.35	223.73	0.39	0.54	44.51%

6: Motif Scaffolding



Conformation Prediction

1. Single State Prediction

Method	Accuracy				Quality	
	TM-score ↑	RMSD ↓	GDT-TS ↑	lDDT ↑	CA clash (%) ↓	PepBond break (%) ↓
AlphaFold2	**0.871**	**3.21**	**0.860**	**0.904**	**0.3**	4.8
OpenFold	0.870	**3.21**	0.856	0.899	0.4	**2.0**
RoseTTAFold2	0.859	3.52	0.845	0.892	**0.3**	5.5
ESMFold	0.847	3.98	0.826	0.870	**0.3**	4.7
EigenFold	0.743	7.65	0.703	0.737	8.0	N/A

2. Multi-state Prediction

Method	RMSDens ↓			RMSD Cluster 3 ↓			Diversity	Quality	
	N=10	N=100	N=1000	N=10	N=100	N=1000	Pairwise RMSD	CA clash (%) ↓	PepBond break (%) ↓
EigenFold	**1.56**	1.50	1.46	2.54	2.48	2.46	0.85	1.4	N/A
MSA-depth32	1.66	1.54	1.41	**2.43**	**2.19**	**1.85**	2.14	0.6	10.6
Str2Str-ODE (Tmax=0.15)	2.40	2.20	2.09	3.00	2.73	2.58	1.86	**0.0**	13.9
ESMFlow-MD	1.68	1.47	1.39	2.44	2.27	2.18	1.17	**0.0**	14.3
ConfDiff-ESM-Force	1.58	**1.43**	**1.36**	2.44	2.35	2.24	1.76	0.1	**8.9**

Method	Accuracy			Diversity	Quality	
	apo-TM ↑	*holo*-TM ↑	TMens ↑	Pairwise TM	CA clash (%) ↓	PepBond break (%) ↓
***apo* model**	1.000	0.790	0.895	N/A	N/A	N/A
EigenFold	0.831	0.864	0.847	0.907	3.6	N/A
MSA-depth256	0.845	0.889	0.867	0.978	**0.2**	**4.6**
Str2Str-ODE (Tmax=0.3)	0.766	0.781	0.774	0.872	**0.2**	14.7
AlphaFlow-PDB	**0.855**	**0.891**	**0.873**	0.924	0.3	6.6
ConfDiff-Open-PDB	0.847	0.886	0.867	0.909	0.5	5.5

3. Distribution Prediction

Method	Diversity	Flexibility: *Pearson* r on				Distributional accuracy			
	Pairwise RMSD	*RMSF	Pairwise RMSD ↑	*Global RMSF ↑	*Per target RMSF ↑	*RMWD ↓	MD PCA W2 ↓	Joint PCA W2 ↓	PC sim > 0.5 % ↑
MD iid	2.76	1.63	0.96	0.97	0.99	0.67	0.73	0.71	93.9
MD 2.5ns	1.54	0.98	0.89	0.85	0.85	2.22	1.55	1.89	36.6
EigenFold	5.96	N/A	-0.03	N/A	N/A	N/A	2.31	7.96	12.2
MSA-depth256	0.83	0.53	0.25	0.34	0.59	3.60	1.79	2.91	29.3
Str2Str-ODE (Tmax=0.1)	1.66	N/A	0.13	N/A	N/A	N/A	2.14	4.39	6.1
AlphaFlow-MD	2.87	1.63	0.53	0.66	**0.85**	**2.64**	1.55	2.29	**39.0**
ConfDiff-Open-MD	3.43	2.21	**0.59**	**0.67**	**0.85**	2.75	**1.41**	**2.27**	35.4

Method	Ensemble observables				Quality	
	Weak contacts J ↑	Transient contacts J ↑	*Exposed residue J ↑	*Exposed MI matrix ρ ↑	CA clash % ↓	*PepBond break % ↓
MD iid	0.90	0.80	0.93	0.56	0.0	3.4
MD 2.5ns	0.62	0.45	0.64	0.25	0.0	3.4
EigenFold	0.36	0.19	N/A	N/A	5.6	N/A
MSA-depth256	0.30	0.29	0.36	0.06	**0.0**	**5.5**
Str2Str-ODE (Tmax=0.1)	0.42	0.18	N/A	N/A	**0.0**	12.1
AlphaFlow-MD	0.62	**0.41**	**0.69**	**0.35**	**0.0**	22.2
ConfDiff-Open-MD	**0.63**	0.39	0.65	0.33	0.5	6.5