SCIENCE CHINA
Information Sciences

• Supplementary File •

Large Language Model Transforms Biological Research: From Architecture to Utilization

Tao WANG[1,2*] & Zeyu LUO[3]

[1]*School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China*
[2]*Key Laboratory of Big Data Storage and Management, Ministry of Industry and Information Technology, Northwestern Polytechnical University, Xi'an 710072, China*
[3]*College of Computer and Control Engineering, Northeast Forestry University, Harbin 150040, China*

Appendix A Supplementary Table

Table A1 The summary of represented LLMs in biological field.

Model name	Parameters	Base model	Fitting data	Prediction Capability	Science field
DNABERT [1]	110M	BERT	DNA sequence	Promoters, Splice sites, Transcription factor binding sites, etc.	
DNABERT2 [2]	117M	BERT	DNA sequence	Promoters, Splice sites, Transcription factor binding sites, etc.	
Enformer [3]	240M	BERT	DNA sequence	Gene expression	Genomic and Transcriptomic
miProBERT [4]&	110M	DNABERT	DNA sequence (Promoter area)	TATA-Promoter identification	
GROVER [5]	–	BERT	DNA sequence	Promoter identification and scanning, CTCF motif binding.	
BERT-Promoter [6]&	110M	BERT	DNA sequence	Promoter strength	
BERT6mA [7]	Millions	BERT	DNA sequence	6mA methylation sites prediction	
iDNA-ABF [8]&	110M	DNABERT	DNA sequence	Multiple type methylation sites prediction	
iDNA-ABT [9]	–	BERT	DNA sequence	Multiple type methylation sites prediction	
MuLan-Methyl [10]	110M, 66M, 12M, Millions, Millions	BERT, DistilBERT [11], ALBERT [12], XLNet [13], ELECTRA [14]	DNA sequence	Multiple type methylation sites prediction.	

Continued on next page

* Corresponding author (email: twang@nwpu.edu.cn)

Table A1 – continued from previous page

Model name	Parameters	Base model	Fitting data	Prediction Capability	Science field
EPiGePT [15]	–	Transformer	DNA sequence	General epigenetic tasks fulfilling	
DNAGPT [16]	100M	GPT	DNA sequence	Genomic signals and non-coding regions identification, mRNA expression, etc.	
PromoGen [17]	Millions	GPT	DNA sequence (Promoter area)	Promoter sequence design	
ENBED [18]	580M-1.2B	Transformer	DNA sequence	Function prediction	
UNI-RNA [19]	25-400M	BERT	RNA sequence	RNA secondary and tertiary structure, RNA modification site, etc.	
RNA-FM [20]	Millions	BERT	RNA sequence	Evolutionary trend analysis of lncRNA, RNA structure, Function prediction, etc.	
RNA-MSM [21]	Millions	MSA-Transformer	RNA sequence	RNA structure, RNA solvent accessibility prediction.	
ERNIE-RNA [22]	86M	BERT	RNA sequence	RNA structure, RNA-protein binding, modification.	
scBERT [23]	Millions	BERT[e]	Single-cell RNA-seq data	Cell type annotation, cell trajectory, marker gene, etc.	
scGPT [24]	–	GPT[e]	Single-cell RNA-seq data, scATAC-seq data	Cell cluster, Batch correction, Cell type annotation,	
etc.					
scFoundation [25][&]	100M	xTrimoGene [26]	Single-cell RNA-seq data	Gene expression enhancement, Medical-drug Response Prediction, etc.	Proteomic
ESM-1b [27]	650M	BERT	Protein sequence	Secondary struct, Localization, Contact, etc.	
ESM2 [28]	8M-15B	RoBEATa	Protein sequence	Protein structure Localization, PPI, etc.	

Continued on next page

Table A1 – continued from previous page

Model name	Parameters	Base model	Fitting data	Prediction Capability	Science field
ESMFold [28]&	15B	ESM2	Protein sequence	Protein structure	
ESM3 [29]	1.4B-98B	Bidirectional Transformer[a]	Protein sequence, Structure, and Function tokens.	Protein structure, Protein design, etc.	
ESM-1v [30]&	650m-3.25B	ESM1-b	Protein sequence	Mutation effect predication	
ESM-IF1 [31,32]&	Millions	ESM family	Protein structure	Protein sequence design and scoring	
EMNGly [33]&	650M	ESM-1b	Protein sequence, Structure embedding.	Glycosylation sites	
ProteinBERT [34]	16M	BERT	Protein sequence, Go annotation.	Protein secondary structure, remote homology prediction, etc.	
antiBERTy [35]	26M	BERT	Protein sequence (antibody only)	Trajectory, Antibody binding site.	
ProtBERT [36]	420M	BERT	Protein sequence	Secondary structure, Molecular fingerprint, etc.	
DistilProtBert [37]&	230M	protBERT	Protein sequence	Secondary structure, Real protein identification.	
MFE [38]&	–	protBERT	Protein sequence, Molecular structure.	Protein-ligand binding affinity.	
TAPE-Transformer [39]	38M	Transformer	Protein sequence	Structure, Contact prediction, Remote homology, etc.	
ProGen [40]	1.2B	GPT	Bidirectional-Protein sequence, Property token.	Protein structure, Protein function, etc.	
ProGen2 [41]	151M-6.4B	GPT	Protein sequence, Protein structure.	Protein design, Protein function, etc.	
PoET [42]	201M	GPT	Protein sequence concatenated by family.	Variant function, Mutation scanning, Protein design.	
ProtT5 [36]	3B	T5	Protein sequence	Localization, PPI, Secondary structure, etc.	

Continued on next page

Table A1 – continued from previous page

Model name	Parameters	Base model	Fitting data	Prediction Capability	Science field
ProtST [43]&	–	ProtBert, ESM-1b, ESM-2, PubMedBERT	Protein sequences, Biomedical annotation text.	Protein property and functional protein retravel.	
ProstT5 [44]&	3B	T5, ProtT5	Protein sequence, Structure token.	3D structure, binding sites, interaction, etc.	
SaProt [45]&	650M	BERT	Protein sequence, Structure token.	Mutation effect, PPI, Ion Binding, etc.	
Deeploc2.0 [46]&	650M,3B	ESM2, ProtT5	Protein sequence	Localization	
ProtLoc-Mex [47]&	650M	ESM2	Protein sequence	Localization, Feature extraction tool	
AFP-MFL [48]&	3B	ProtT5	Protein sequence	Antifungal peptides identification	
DeepProSite [49]&	3B	ESMFold, ProtT5	Protein sequence	Protein binding site	
EasIFA [50]&	650M	ESM2, SaProt	Protein sequence, SMILES sequence [51], Structure graph	Enzyme site annotation	
EvoDiff [52]	–	Diffusion model[b]	Protein sequence, MSA input token.	Protein design, Functional scaffold design.	
AlphaFold2 [53]	Millions	Evoformer[c]	Protein sequence, structure data[d]	Protein structure	
AlphaFold3 [54]	–	Evoformer-Pairformer-Diffusion	Sequence (include DNA/RNA, Protein, Atom, etc.), Protein and molecular structure data.	Protein-Protein complexes structure, Protein-molecular complexes structure.	
CATHe [55]&	420M,3B	ProtBERT, ProtT5	Protein sequence	Remote homology searching	
PLMSearch [56]&	650M	ESM-1b	Protein sequence	Remote homology searching	

a: Structure tokenizer adopts VQ-VAE [57] as the backbone frames, and only applied for generating structure token. Encoder composed by stacking transformer block, Geometric attention are introduced in some transformer block. Decoder are adopted for protein structure design and generation.
b: one type of generation model differs from Transformer, generating data through denoising
c: Evoformer is similar to Transformer, based on attention block with adding special block, for instance MSA (Multiple Sequence Alignment) [58] and protein pair block. Typically, AlphaFold family does not belong to LLMs, but famous in the domain of biological science.

d: Protein sequence as input, then conduct searching in Genetic database for MSA representation. Meanwhile, combining sequence and information from structure database for pair representation.
e: Embedding layer apply gene expression bin token to replace positional token.

&: Represent domain Fine-Tune model based on pretrained LLMs.

References

1 Ji Y, Zhou Z, Liu H, et al. DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics, 2021, 37: 2112-2120
2 Zhou Z, Ji Y, Li W, et al. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. arXiv: 2306.15006, 2023
3 Avsec Ž, Agarwal V, Visentin D, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods, 2021, 18: 1196-1203
4 Wang X, Gao X, Wang G, et al. miProBERT: Identification of microRNA promoters based on the pre-trained model BERT. Brief Bioinform, 2023, 24: bbad123
5 Sanabria M, Hirsch J, Joubert P M, et al. DNA language model GROVER learns sequence context in the human genome. Nat Mach Intell, 2024, 6: 1-13
6 Le N Q K, Ho Q T, Nguyen V N, et al. BERT-Promoter: An improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection. Comput Biol Chem, 2022, 99: 107732
7 Tsukiyama S, Hasan M M, Deng H W, et al. BERT6mA: Prediction of DNA N6-methyladenine site using deep learning-based approaches. Brief Bioinform, 2022, 23: bbac053
8 Jin J, Yu Y, Wang R, et al. iDNA-ABF: Multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. Genome Biol, 2022, 23: 219
9 Yu Y, He W, Jin J, et al. iDNA-ABT: Advanced deep learning model for detecting DNA methylation with adaptive features and transductive information maximization. Bioinformatics, 2021, 37: 4603-4610
10 Zeng W, Gautam A, Huson D H. MuLan-Methyl: Multiple transformer-based language models for accurate DNA methylation prediction. Gigascience, 2023, 12: giad054
11 Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv: 1910.01108, 2019
12 Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations. arXiv: 1909.11942, 2019
13 Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding. Adv Neural Inf Process Syst, 2019, 32: 5754-5764
14 Clark K, Luong M T, Le Q V, et al. Electra: Pre-training text encoders as discriminators rather than generators. arXiv: 2003.10555, 2020
15 Gao Z, Liu Q, Zeng W, et al. EpiGePT: A pretrained transformer model for epigenomics. bioRxiv: 2023.05.01.538975, 2023
16 Zhang D, Zhang W, He B, et al. DNAGPT: A generalized pretrained tool for multiple DNA sequence analysis tasks. bioRxiv: 2023.07.11.548628, 2023
17 Xia Y, Du X, Liu B, et al. Species-specific design of artificial promoters by transfer-learning based generative deep-learning model. Nucleic Acids Res, 2024, 52: 6145-6157
18 Malusare A, Kothandaraman H, Tamboli D, et al. Understanding the natural language of DNA using encoder-decoder foundation models with byte-level precision. arXiv: 2306.14999, 2023
19 Wang X, Gu R, Chen Z, et al. UNI-RNA: Universal pre-trained models revolutionize RNA research. bioRxiv: 2023.07.11.548588, 2023
20 Chen J, Hu Z, Sun S, et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. arXiv: 2204.00300, 2022
21 Zhang Y, Lang M, Jiang J, et al. Multiple sequence alignment-based RNA language model and its application to structural inference. Nucleic Acids Res, 2024, 52: e3
22 Yin W, Zhang Z, He L, et al. ERNIE-RNA: An RNA language model with structure-enhanced representations. bioRxiv: 2024.03.17.585376, 2024
23 Yang F, Wang W, Wang F, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nat Mach Intell, 2022, 4: 852-866
24 Cui H, Wang C, Maan H, et al. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods, 2024, 21: 1-12
25 Hao M, Gong J, Zeng X, et al. Large-scale foundation model on single-cell transcriptomics. Nat Methods, 2024, 21: 1-10
26 Gong J, Hao M, Cheng X, et al. xTrimoGene: An efficient and scalable representation learner for single-cell RNA-seq data. Adv Neural Inf Process Syst, 2024, 36: 12345-12360
27 Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA, 2021, 118: e2016239118
28 Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 2023, 379: 1123-1130
29 Hayes T, Rao R, Akin H, et al. Simulating 500 million years of evolution with a language model. bioRxiv: 2024.07.01.600583, 2024
30 Meier J, Rao R, Verkuil R, et al. Language models enable zero-shot prediction of the effects of mutations on protein function. Adv Neural Inf Process Syst, 2021, 34: 29287-29303
31 Hsu C, Verkuil R, Liu J, et al. Learning inverse folding from millions of predicted structures. In: Proc 39th Int Conf Mach Learn. Baltimore: PMLR, 2022. 8946-8970
32 Shanker V R, Bruun T U, Hie B L, et al. Unsupervised evolution of protein and antibody complexes with a structure-informed language model. Science, 2024, 385: 46-53
33 Hou X, Wang Y, Bu D, et al. EMNGly: Predicting N-linked glycosylation sites using the language models for feature extraction. Bioinformatics, 2023, 39: btad650
34 Brandes N, Ofer D, Peleg Y, et al. ProteinBERT: A universal deep-learning model of protein sequence and function. Bioinformatics, 2022, 38: 2102-2110
35 Ruffolo J A, Gray J J, Sulam J. Deciphering antibody affinity maturation with language models and weakly supervised learning. arXiv: 2112.07782, 2021
36 Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: Toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 7112-7127
37 Geffen Y, Ofran Y, Unger R. DistilProtBert: A distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts. Bioinformatics, 2022, 38: ii95-ii98
38 Xu S, Shen L, Zhang M, et al. Surface-based multimodal protein-ligand binding affinity prediction. Bioinformatics, 2024, 40: btae123
39 Rao R, Bhattacharya N, Thomas N, et al. Evaluating protein transfer learning with TAPE. Adv Neural Inf Process Syst,

2019, 32: 9689-9701
40 Madani A, McCann B, Naik N, et al. Progen: Language modeling for protein generation. arXiv: 2004.03497, 2020
41 Nijkamp E, Ruffolo J A, Weinstein E N, et al. Progen2: Exploring the boundaries of protein language models. Cell Syst, 2023, 14: 968-978
42 Truong T F, Bepler T. PoET: A generative model of protein families as sequences-of-sequences. arXiv: 2306.06156, 2023
43 Xu M, Yuan X, Miret S, et al. Protst: Multi-modality learning of protein sequences and biomedical texts. In: Proc 40th Int Conf Mach Learn. Honolulu: PMLR, 2023. 38749-38767
44 Heinzinger M, Weissenow K, Sanchez J G, et al. Bilingual language model for protein sequence and structure. bioRxiv: 2023.07.23.550085, 2023
45 Su J, Han C, Zhou Y, et al. SaProt: Protein language modeling with structure-aware vocabulary. bioRxiv: 2023.10.01.560349, 2023
46 Thumuluri V, Almagro Armenteros J J, Johansen A R, et al. DeepLoc 2.0: Multi-label subcellular localization prediction using protein language models. Nucleic Acids Res, 2022, 50: W228-W234
47 Luo Z, Wang R, Sun Y, et al. Interpretable feature extraction and dimensionality reduction in ESM2 for protein localization prediction. Brief Bioinform, 2024, 25: bbad534
48 Fang Y, Xu F, Wei L, et al. AFP-MFL: Accurate identification of antifungal peptides using multi-view feature learning. Brief Bioinform, 2023, 24: bbac606
49 Fang Y, Jiang Y, Wei L, et al. DeepProSite: Structure-aware protein binding site prediction using ESMFold and pretrained language model. Bioinformatics, 2023, 39: btad718
50 Wang X, Yin X, Jiang D, et al. Multi-modal deep learning enables efficient and accurate annotation of enzymatic active sites. Nat Commun, 2024, 15: 7348
51 GDR H B, Sharon N, Australia E. Nomenclature and symbolism for amino acids and peptides. Pure Appl Chem, 1984, 56: 595-624
52 Alamdari S, Thakkar N, van den Berg R, et al. Protein generation with evolutionary diffusion: Sequence is all you need. bioRxiv: 2023.09.11.556673, 2023
53 Tunyasuvunakool K, Adler J, Wu Z, et al. Highly accurate protein structure prediction for the human proteome. Nature, 2021, 596: 590-596
54 Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024, 630: 493-500
55 Nallapareddy V, Bordin N, Sillitoe I, et al. CATHe: Detection of remote homologues for CATH superfamilies using embeddings from protein language models. Bioinformatics, 2023, 39: btad029
56 Liu W, Wang Z, You R, et al. PLMSearch: Protein language model powers accurate and fast sequence search for remote homology. Nat Commun, 2024, 15: 2775
57 Razavi A, Van den Oord A, Vinyals O. Generating diverse high-fidelity images with vq-vae-2. Adv Neural Inf Process Syst, 2019, 32: 14837-14847
58 Mirabello C, Wallner B. rawMSA: End-to-end deep learning using raw multiple sequence alignments. PLoS One, 2019, 14: e0220182