A Fine-tuning Dataset and Benchmark for Large Language Models for Protein Understanding

Yiqing Shen[1,2,†], Zan Chen[1,†], Michail Mamalakis[3,†], Luhan He[1,†,§], Haiyang Xia[1,4,†], Tianbin Li[5], Yanzhou Su[5], Junjun He[5], Yu Guang Wang[1,5,6,7,*]

[1]*Toursun Synbio*, Shanghai, China

[2]*Department of Computer Science*, *Johns Hopkins University*, Baltimore, USA

[3]*Department of Computer Science and Technology*, *University of Cambridge*, Cambridge, UK

[4]*Shanghai Institute for Biomedical and Pharmaceutical Technologies*, Shanghai, China

[5]*Shanghai AI Laboratory*, Shanghai, China

[6]*Institute of Natural Sciences*, *Shanghai Jiao Tong University*, Shanghai, China

[7]*UNSW Sydney*, Sydney, Australia

[†]Equal contribution. [§]Work done during the internship at Toursun Synbio. [*]Corresponding author.

yshen92@jhu.edu yuguang.wang@sjtu.edu.cn

***Abstract*—The high similarities between protein sequences and natural language, particularly in their sequential data structures, have driven parallel advancements in deep learning models for both domains. In natural language processing (NLP), large language models (LLMs) have achieved remarkable success in tasks such as text generation, translation, and conversational agents, owing to their extensive training on diverse datasets that enable them to capture complex language patterns and generate human-like text. Inspired by these advancements, researchers have attempted to adapt LLMs for protein understanding by integrating a protein sequence encoder with a pre-trained LLM, following designs like LLaVa. However, this adaptation raises a fundamental question: *"Can LLMs, originally designed for NLP, effectively comprehend protein sequences as a form of language?"* Current datasets fall short in addressing this question due to the lack of a direct correlation between protein sequences and corresponding text descriptions, limiting the ability to train and evaluate LLMs for protein understanding effectively. To bridge this gap, we introduce `ProteinLMDataset`, a dataset specifically designed for further self-supervised pretraining and supervised fine-tuning (SFT) of LLMs to enhance their capability for protein sequence comprehension. Specifically, `ProteinLMDataset` includes 17.46 billion tokens for pretraining and 893K instructions for SFT. Additionally, we present `ProteinLMBench`, the first benchmark dataset consisting of 944 manually verified multiple-choice questions for assessing the protein understanding capabilities of LLMs. `ProteinLMBench` incorporates protein-related details and sequences in multiple languages, establishing a new standard for evaluating LLMs' abilities in protein comprehension. The large language model InternLM2-7B, pretrained and fine-tuned on the `ProteinLMDataset`, outperforms GPT-4 on `ProteinLMBench`, achieving the highest accuracy score. The dataset and the benchmark are available at https://huggingface.co/datasets/tsynbio/ProteinLMDataset/ and https://huggingface.co/datasets/tsynbio/ProteinLMBench. The code is available at https://github.com/tsynbio/ProteinLMDataset/.**

***Index Terms*—Deep Learning, Dataset, Benchmark, Large Language Model, Protein Engineering.**

I. Introduction

Protein science offers insights into biological processes at the molecular level, driving progress in medicine, biotechnology, and our understanding of life. Given the similar sequential data structures of protein sequences and natural language, there have been parallel advancements in deep learning models, such as protein language models [1] and large language models (LLMs) [2]. LLMs have already demonstrated their strong capabilities in text understanding across various tasks, prompting researchers to explore their potential in protein understanding from the perspective of multi-modal language models (MMLMs) by integrating protein sequences or structures with textual content using separate encoders. Specifically, they encode each modality (protein sequence or protein structure) independently before combining them with a frozen LLM [3], [4]. However, this approach poses challenges in fully leveraging the intricate connections between protein sequences and textual information. As far as we can ascertain, there is no comprehensive dataset that seamlessly integrates protein sequences and their corresponding textual descriptions, enabling effective training of language models to comprehend protein information. This gap limits the ability to train LLMs to fully utilize the intricate connections between protein sequences and textual information.

To address these limitations, we introduce a comprehensive large-scale protein sequence and text (seq-text) hybrid dataset named `ProteinLMDataset`. This dataset is designed to train LLMs to learn the correspondence between textual descriptions and protein sequences, enabling them to effectively understand and leverage the intricate connections between these two forms of data. Furthermore, we propose a novel benchmark `ProteinLMBench` comprising meticulously curated multiple-choice questions to rigorously evaluate the LLMs' proficiency to comprehend protein sequences. To the best of our knowledge, this is the first dataset of its kind, combining the largest dataset for both self-supervised learning (SSL) and supervised fine-tuning (SFT) specifically in the protein science domain, along with the first manually annotated multiple-choice questions benchmark for evaluating protein understanding capability with LLMs.

The `ProteinLMDataset` comprises 17.46 billion tokens

for self-supervised learning and 893K instructions for supervised fine-tuning. `ProteinLMBench` contains 944 manually annotated multiple-choice questions for evaluation. The self-supervised dataset is structured into three segments: 0.69% Chinese-English text pairs in protein science, 41.51% protein sequence-English text pairs, and 57.80% protein-related English text. The supervised fine-tuning dataset includes seven tasks: a novel enzyme chain of thought (ECoT), protein functionality, induction, disease involvement, post-translational modifications, sub-unit structure, and tissue specificity. The ECoT approach aims to innovatively enable LLMs to think step-by-step and generate well-founded protein knowledge, thereby enhancing the reliability and accuracy of outputs in protein science and engineering [5]. Unlike the plain chain of thought, ECoT incorporates prior knowledge of protein science by guiding LLMs to infer protein functions based on the reactions they are involved in. This approach enhances logical reasoning capabilities in the protein domain. `ProteinLMDataset` aims to bridge existing dataset gaps and unlock the potential of LLMs in protein data analysis by seamlessly integrating protein sequences with their relevant textual information.

The major contributions of this work are three-fold. Firstly, we present the first large-scale protein-text dataset designed to enable LLMs to comprehend protein sequences without an extra encoder. Secondly, we introduce the Enzyme Chain of Thought (ECoT), a novel mechanism specifically for protein understanding, allowing LLMs to generate reliable and accurate protein knowledge step-by-step. Thirdly, we introduce `ProteinLMBench`, the first comprehensive, manually annotated benchmark for thoroughly evaluating LLMs' comprehension and representation of protein sequences.

II. Dataset and Benchmark Composition

A. ProteinLMDataset

The `ProteinLMDataset` includes both self-supervised and supervised fine-tuning components. The self-supervised part is strategically categorized into three primary segments, each contributing distinct dimensions to our objectives of empowering the LLMs to understand protein sequence. The first segment focuses on comparing biological science in Chinese and English, encompassing 7,000 entries and yielding over 120 million tokens for training and tuning purposes. The second segment contains protein sequence and English text pairs, derived from a weakly linked PMC full-text dataset. Associations are established using the Structure Integration with Function, Taxonomy, and Sequence (SIFTS) database (https://www.ebi.ac.uk/pdbe/docs/sifts/), which maps PDB to UniProt IDs and subsequently to protein names, guiding our search for relevant literature within the PMC database. The full-text component comprises 320,000 entries, providing an extensive corpus exceeding 2.9 billion tokens. Moreover, the PubMed abstract contributes over 30 million entries and 7.0 billion tokens for comprehensive training and tuning. For protein sequence and text pairs, we leverage a highly relevant PMC full-text dataset totaling 18,000 entries and surpassing 195 million tokens for nuanced training and tuning. The integration of the UniProtKB Swiss-Prot dataset further enriches this category with an additional 349 million tokens. The final segment involves extracting insights from a PubMed abstract dataset, boasting more than 3 million entries, and generating over 7.0 billion tokens for robust training and tuning. This strategic categorization ensures a diverse and versatile dataset, addressing the limitations of existing benchmarks and facilitating advancements in cross-lingual language models and protein sequence understanding.

The supervised fine-tuning component in `ProteinLMDataset` is a collection of 893,000 instructions, spanning seven different segments: enzyme Chain of Thought (10.8k), protein functionality (465k), induction (25.4k), disease involvement (5.58k), post-translational modifications (45.8k), sub-unit structure (291k), and tissue specificity (50.3k). It includes analytical instructions, input tokens, and ground truth prediction outputs for fine-tuning and performance score determination. The primary source of data collection is UniProtKB.

B. ProteinLMBench

Our proposed benchmark, `ProteinLMBench`, features an evaluation dataset consisting of 944 six-choice questions, each accompanied by an explanation of the correct answer. These questions cover a range of topics including protein-based property prediction, protein descriptions, and protein sequence understanding. This benchmark is designed to assess the ability of large language models to interpret and analyze protein sequences in conjunction with their associated textual descriptions.

III. Collection Methodology

A. The self-supervised dataset collection

a) Biology Chinese and English text pair collection.: We extracted Chinese-English text pairs from articles published in the *Chinese Journal of Biotechnology* and the *Chinese Journal of Cell Biology*. This extraction process ensures a diverse and authentic representation of scientific language. The selected time interval, 2013-2024, enhances the dataset's relevance by capturing contemporary scientific discourse in both Chinese and English languages.

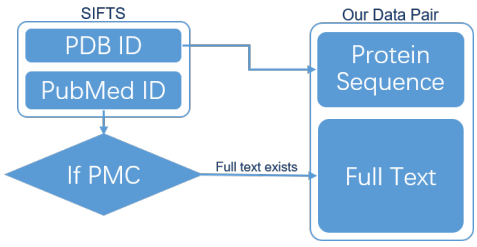

Fig. 1. Process for pairing proteins with the corresponding PMC articles' text

b) Protein sequence and English text pair.: We retrieved Protein Data Bank (PDB) IDs of proteins and their corresponding PubMed IDs from the SIFTS database (https://www.ebi.ac.uk/pdbe/docs/sifts/). Using a weakly linked PMC full-text dataset, we established associations through the *protein_name_to_pmcid.json* index file derived from *pdb_chain_uniprot.csv*. This mapping, connecting PDB to UniProt IDs and subsequently to protein names, guided our literature search within the PMC database. In this manner, we can associate a specific protein sequence with the complete texts of the most relevant articles in PMC. We examined the PubMed articles for PubMed Central (PMC) IDs; when available, we accessed the full text. This enabled us to acquire protein sequences using PDB IDs and retrieved detailed research articles on the proteins, creating pairs to describe the protein sequences in human language (Fig. 1).

Protein sequence collection. We employed BERN2 [6] which is a Named Entity Recognition (NER) model to extract NERs from the titles and abstracts of all PubMed articles, identifying nouns potentially referring to proteins or genes (since it is challenging to distinguish whether a single word refers to a protein or gene without context) and recorded their confidence scores for further analysis (Fig. 2). We then linked all extracted nouns to specific entities in the NCBI Gene database [7] through Named Entity Normalization (NEN) and obtained their NCBI Gene IDs. For these Gene IDs, we could map them to specific protein sequences via the NCBI's RefSeq database. Afterward, we extracted all article titles and the results from the NER & NEN models with a confidence score above a certain threshold (here, 98%), inputting them into a large language model to determine whether these nouns are referred to as proteins or genes in the given context. We assumed that protein names appearing in titles are likely the focus of the research and used the abstracts to help describe these proteins. Finally, we located the first occurrence of each noun in the abstracts that matched the protein names identified in the titles with the same NCBI Gene ID and inserted the previously obtained protein sequences after these nouns. This way, we created a dataset of textual descriptions wrapping protein sequences.

B. The supervised fine-tuning dataset collection

The primary source for the supervised fine-tuning dataset is UniProtKB. Through a comprehensive analysis of the locally downloaded raw data, we identified a substantial corpus of curated and verified protein information. This encompasses a diverse range of attributes, including protein functionality, induction, disease associations, post-translational modifications, sub-unit structures, and tissue specificity. For each protein, we crafted a dedicated template. The protein sequence was embedded within the question template, while its corresponding properties were incorporated into the answer template. This approach facilitated the creation of effective question-answer pairs to understand protein and its properties. In total, we generated approximately 883k such data entries.

To create the Chain of Thought (CoT) data, we utilized the IUBMB Enzyme Nomenclature as the raw source. We extracted the name, nickname, reaction formula, and comments (typically describing the enzyme's function) for each enzyme entry. The response generation process followed a systematic thinking process: first, the model outputted the reactions the enzyme would catalyze, and then provided the enzyme's comments, forming a chain of thought. During data creation, we generated multiple suitable templates and inserted the obtained information, resulting in an effective CoT supervised fine-tuning dataset comprising 10,800 entries.

C. The benchmark multi-choice questionnaire collection

The evaluation questionnaire, `ProteinLMBench`, comprises six choices for each question, accompanied by an explanation of the correct answer. The questions span various domains, including protein-based property prediction, protein descriptions, and protein sequence understanding. The questionnaire consists of a synthetic dataset generated by Retrieval-Augmented Generation (RAG; [8]) and GPT-4 [2]. We implemented a two-round verification scheme involving RAG to generate the initial questions, choices, and answers, followed by GPT-4 to validate the answers. Questions with inconsistencies between the two rounds were discarded, ensuring the consistency, robustness, and reproducibility of the questionnaire. Despite starting with 1,000 questions, the verification process yielded a final set of 944 questions.

D. Text input filtering and tokenization procedure

To preprocess the text inputs from three collections, we employed a filter cleaner to eliminate citations, references, tables, figures, commas, and any symbols unrelated to the task of interest. This filter was systematically applied to both titles and abstracts within the collected manuscripts. The resulting filtered outputs were stored in *".json"* format. Subsequently, this *json* files serve as input for extracting tokens of various sizes associated with the collection and the task of interest, such as protein sequences, pairs of English and Chinese text, protein descriptions, and design elements. The implementation source code for the filters is available on our GitHub repository at the following link: https://github.com/tsynbio/Protein_LM/blob/main/src/PMC_data_collector.py. The tokenization is performed using the open-source coding of the *Interlm* tokenizer based on byte-level byte-pair-encoding (https://huggingface.co/internlm/internlm-7b/blob/main/tokenization_internlm.py), which processes the filtered text inputs.

IV. Quality of Dataset and Benchmark

a) Diversity and statistics of ProteinLMDataset.: The diversity and multi-source curation of the dataset are presented in Tab. I, highlighting the document collection based on the token length. The three segments of the proposed dataset includes manuscripts, abstracts, protein sequences, and taxonomy, divided into five collections: 'Biology Chinese/English text pair', 'PMC full-text manuscripts', 'PubMed Abstracts',

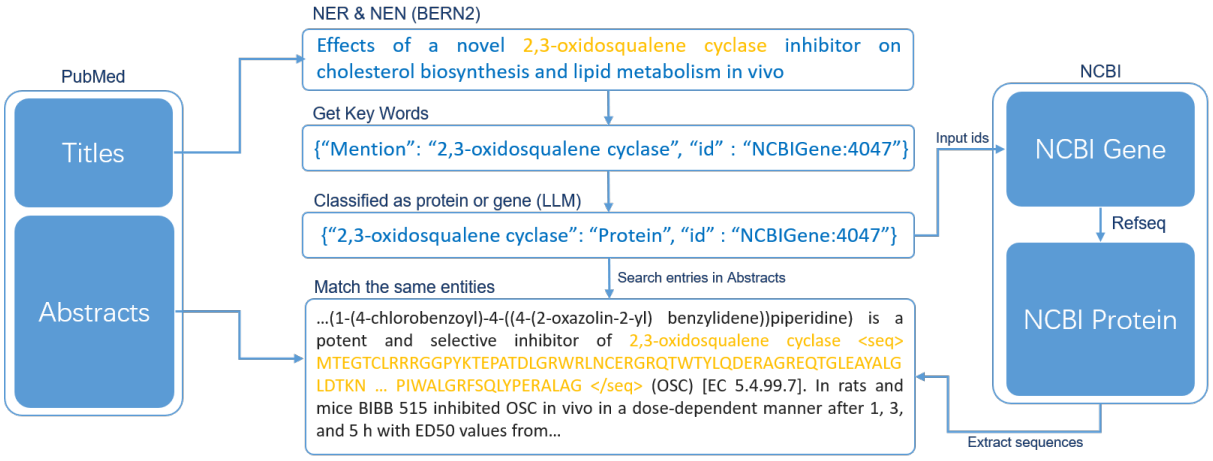

Fig. 2. Process for Inserting Protein Sequences into Text

'UniProtKB Swiss-Prot', and 'Protein Sequence/PubMed Abstracts' and a variability ranges from minimum lengths of 49 to 549 characters to maximum lengths of 39,239 to 8,390,807 characters. Tab. I highlights the extracted tokens from the documentation characters after the filtering and noise reduction process (citations, references, tables, etc. unrelated to the task of interest). The produced tokens exhibit variability, ranging from minimum lengths of 21 to 101 characters to maximum lengths of 1,017 to 2,351,194 characters. These variations across different tokens, texts and combinations validate the unbiased and comprehensive content of our dataset. Fig. 3 illustrates the statistics regarding the different sources and ratios within our three segment datasets. Our dataset predominantly represents protein sequences, with the primary source being PubMed abstracts (6.9B), followed by UniProtKB Swiss-Prot collection (349M). Regarding protein design and information, the primary resources are derived from PubMed abstracts of scientific manuscripts (7.0B), followed by weakly linked PMC full manuscripts and highly relevant PMC full texts (2.9B; see Fig. 3a.). Additionally, the Chinese-English collection comprises a substantial of 200M tokens. To delve deeper, the UniProtKB Swiss-Prot collection consists of 62% protein sequences and 38% text relevant to protein design and scientific biology documentation tokens, contrasting with the Protein Seq/PubMed Abstract collection, which includes 16% text documentation and 84% protein sequences tokens (see Fig. 3b).

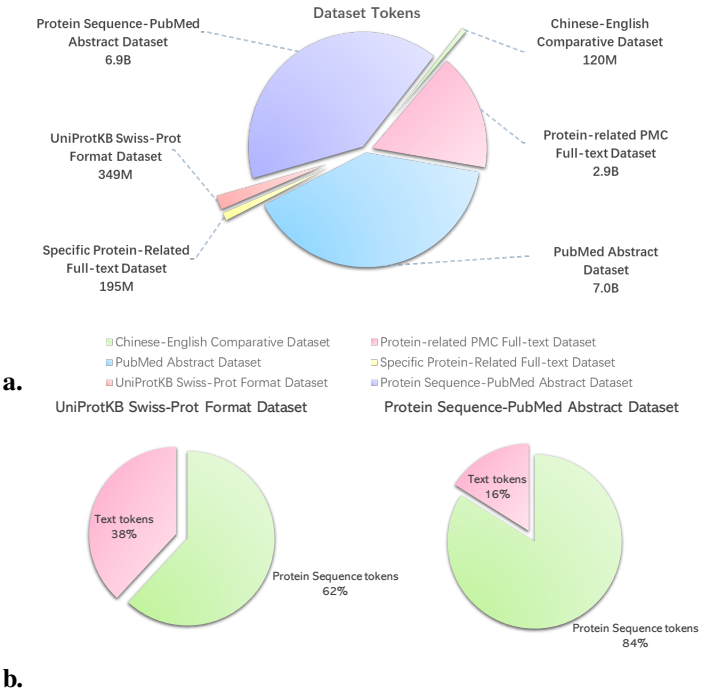

Fig. 3. Statistics of the sources and ratios of the three segments of `ProteinLMDataset`: Chinese-English pairs, protein sequences, and protein sequence-text pairs.

Tab. II presents the diversity of multi-design topics and curation of the fine-tuning dataset, highlighting the different tasks and analyzing the statistics based on token length. The proposed benchmark comprises seven segments: 'Enzyme CoT', 'UniProt Function', 'UniProt Induction', 'UniProt Involvement in disease', 'UniProt Post-translational modification', 'UniProt Sub-unit structure', and 'UniProt Tissue specificity'. These segments exhibit a wide range of token lengths, varying from a minimum of 65 to 88 characters to a maximum of 2,310 to 70,500 tokens. Tab. II highlights the extracted token counts for each segment. After filtering and noise reduction, the instructions and output of the benchmark are provided on the website https://huggingface.co/datasets/tsynbio/ProteinLMBench. The variations across different token lengths, fine-tuning tasks, and combinations validate the unbiased and comprehensive nature of our benchmark.

b) Diversity and statistics of the ProteinLMBench.: The evaluation dataset consists of 944 questions with a median string length of 98-124 tokens, with a minimum of 46 and a maximum of 305 tokens. There are six answer choices, with an average string length of 77-114 tokens, with a minimum of 3 and a maximum of 368 tokens. The answers are well-

TABLE I

THE OVERALL STATISTICS OF EXTRACTED TOKENS FROM THE PROPOSED SELF-SUPERVISED DATASET. THE NUMBERS ARE REPRESENTATION OF THE LENGTH (LEN.) TOKEN PER DOCUMENT. THE DATASET CLASSIFY IN THREE SEGMENTS-BIOLOGY CHINESE/ENGLISH (CHI/ENG) TEXT PAIR, PROTEIN SEQUENCE AND ENGLISH (SEQ/ENG) TEXT PAIR AND PROTEIN SEQUENCE (SEQ)-. THE STATISTICS ARE AGGREGATED OVER THE ENTIRE COHORT.

Tokens statistic of the self-supervised dataset				
Segment	Collection	Min Len.	Average Len.	Max Len.
Biology Chi/Eng text pair	Chinese/English text	54/48	245/267	1017/1165
Protein Seq/Eng text pair	PMC Full text	101	9954	2351194
	PubMed Abstract	21	282	13983
	UniProtKB Swiss-Prot	69	591	35751
Protein Seq	Protein Seq	65	1650	70500

TABLE II

THE OVERALL STATISTICS OF EXTRACTED TOKENS FROM THE PROPOSED FINE-TUNING DATASET REPRESENT THE TOKEN LENGTH PER DOCUMENT. THE DATASET IS CATEGORIZED INTO SEVEN SEGMENTS: ENZYME CHAIN OF THOUGHT (ECOT), PROTEIN FUNCTIONALITY, INDUCTION OF PROTEIN EXPRESSION, DISEASE INVOLVEMENT, POST-TRANSLATIONAL MODIFICATIONS, SUB-UNIT STRUCTURE, AND TISSUE SPECIFICITY. THESE STATISTICS ARE AGGREGATED ACROSS THE ENTIRE COHORT.

Tokens statistic of the fine-tuning dataset				
Segment	Scope	Min Len.	Average Len.	Max Len.
Enzyme CoT	molecule's expression	84	800	15300
UniProt Function	protein functionality	74	3610	70500
UniProt Induction	protein induction	82	1650	31300
UniProt Involvement in disease	protein diseases	88	3495	68800
Uniprot Post-translational modification	protein translation	88	361	6420
UniProt Sub-unit structure	protein structure	74	3560	70500
UniProt Tissue specificity	tissue gene expression	65	145	2310

balanced, with the correct answer being option 1 in 16.3% of the cases, option 2 in 17.9%, option 3 in 19.9%, option 4 in 16.1%, option 5 in 15.6%, and option 6 in 14.0% of the cases.

c) Safety during collection.: In ensuring the safety and reliability of our proposed dataset, we have implemented rigorous measures and levels of filtering to address potential issues related to data integrity and quality. Given the inherent imperfections in structural and textual data, we acknowledge the importance of safeguarding against corrupted or erroneous information. Our dataset curation process includes thorough validation checks and quality control procedures to identify and rectify any anomalies that may compromise the integrity of the data. We are committed to transparency in our approach, encouraging users to report any instances of suspected data corruption or safety concerns through our designated channels. This proactive engagement with the research community is integral to maintaining the credibility of the dataset and fostering a collaborative environment for continual improvement. Additionally, our strategic categorization approach facilitates the identification and isolation of potential outliers, contributing to the overall safety and reliability of the dataset by providing users with a clear understanding of the distinct dimensions and sources within each segment. The ethical statement of this work is presented in Appendix Section **??**.

d) Technical limitations.: Even when systematically collecting protein sequence data through scientific filtering and established resources, inherent technical limitations persist. Achieving perfection in structural data is rare, as experimental uncertainties stem from the limited resolution of techniques like X-ray crystallography or electron cryo-microscopy.

V. EXPERIMENT

a) Experiment setup.: We conducted three different experiments: no training, supervised fine-tuning training (SFT), and self-supervised learning combined with supervised fine-tuning training (SSL-SFT) using `ProteinLMDataset`. For the no training experiment, we employed various large language models, including Falcon-7b [9], Qwen1.5-7B [10], Moonshot [11], Mistral-7B-Instruct-v0.2 [12], Baichuan2-7B-Chat [13], Llama-2-7B-Chat-hf [14], InternLM-Chat-20B, InternLM2-Chat-7B, and variations [15], ChatGLM3-6B [16], Yi-6B-Chat [17], GPT3.5-turbo, and GPT4.0-turbo [2]. For the SFT and SSL-SFT experiments, we utilized the InternLM2-7B model, naming them InternLM2-Protein-7B (w/o SSL) and InternLM2-Protein-7B, respectively. All the experiments were evaluated for accuracy using our proposed `ProteinLMBench`. For detailed training procedures and LLM architectures, please refer to Appendix C.

b) Experiment results.: The results of all experiments are presented in Fig. 4. In the no-training experiment (depicted by green box plots), GPT-4.0-turbo achieved the highest accuracy with a correct rate of 57.94%, followed by InternLM2-20B with a score of 57.52%. Falcon-7B exhibited the poorest performance, achieving a correct rate of 19.17%. In the supervised fine-tuning experiment (illustrated by pink box plots), InternLM2-7B delivered accuracy scores of 58.26% (Fig. 4; InternLM2-Protein-7B (w/o SSL)). This is an improvement from the no training experiment, where its accuracy

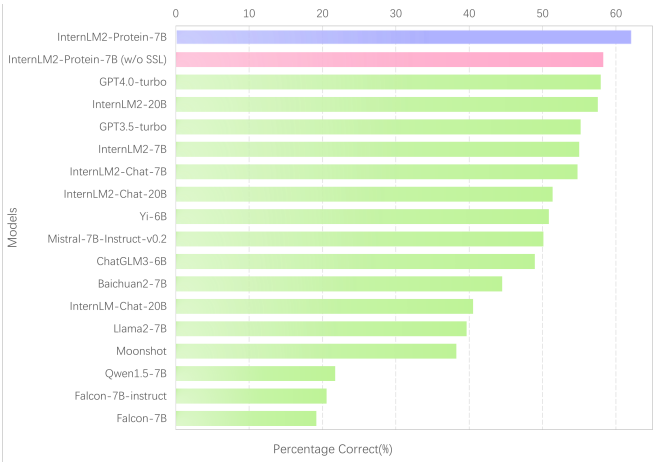

Fig. 4. Model performence of LLMs on `ProteinLMBench`. The purple box plot represents the model performance combined training on `ProteinLMDataset` SSL and SFT datasets. The pink box plot represents the model performance obtained by only conducting SFT without SSL training. The green box plots represent the performance of other comparative models.

was 54.98%. Notably, when we initially trained InternLM2-7B model on the self-supervised dataset and subsequently fine-tuned it (Fig. 4; InternLM2-Protein-7B), the accuracy increased significantly to 62.18%. This underscores the importance of the proposed `ProteinLMDataset` for more efficient and accurate protein understanding.

VI. Conclusion

Our proposed dataset-benchmark tandem offers a comprehensive platform for training, fine-tuning, and evaluating language models on both Chinese-English text pairs and protein sequences, addressing limitations in existing datasets. The strategic classification approach ensures diversity and adaptability, tackling constraints within protein science and language model benchmarking. By curating a diverse collection from various sources, we reinforce the impartiality of our dataset. Incorporating both Chinese and English in detailing protein characteristics introduces a novel cross-lingual perspective, mitigating biases and enhancing multilingual understanding. Moreover, our proposed fine-tuning dataset enhances language models' capabilities in understanding, designing, and analyzing protein sequences. Experimentation within our `ProteinLMBench` benchmark validates the efficacy of our dataset. The InternLM2-7B language model achieved an accuracy of less than 55.00% without further training on the `ProteinLMDataset`. When fine-tuned, its accuracy increased to 58.26%. Notably, when InternLM2-7B was trained on both the self-supervised and fine-tuning datasets of `ProteinLMDataset`, it exhibited an exceptional improvement in accuracy, reaching 62.18%. Our dataset and benchmark contribute to advancing artificial intelligence for biological sciences. It enables groundbreaking research and applications with the potential to transform diverse fields.

References

[1] A. Rives *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *PNAS*, 2019. [Online]. Available: https://www.biorxiv.org/content/10.1101/622803v4

[2] OpenAI *et al.*, "GPT-4 Technical Report," *arxiv preprint arXiv:2303.08774*, 2024.

[3] M. Xu, X. Yuan, S. Miret, and J. Tang, "ProtST: Multi-modality learning of protein sequences and biomedical texts," in *ICML*, vol. 202, 2023, pp. 38 749–38 767. [Online]. Available: https://proceedings.mlr.press/v202/xu23t.html

[4] H.-Y. Zhou, Y. Fu, Z. Zhang, B. Cheng, and Y. Yu, "Protein representation learning via knowledge enhanced primary structure reasoning," in *ICLR*, 2023. [Online]. Available: https://openreview.net/forum?id=VbCMhg7MRmj

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, brian ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," in *NeurIPS*, 2022. [Online]. Available: https://openreview.net/forum?id=_VjQlMeSB_J

[6] M. Sung, M. Jeong, Y. Choi, D. Kim, J. Lee, and J. Kang, "BERN2: an advanced neural biomedical named entity recognition and normalization tool," *Bioinformatics*, vol. 38, no. 20, p. 4837–4839, Sep. 2022. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btac598

[7] E. W. Sayers *et al.*, "Database resources of the national center for biotechnology information," *Nucleic Acids Research*, vol. 50, no. D1, pp. D20–D26, 12 2021. [Online]. Available: https://doi.org/10.1093/nar/gkab1112

[8] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2024.

[9] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, Étienne Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, and G. Penedo, "The Falcon Series of Open Language Models," *arXiv preprint arXiv:2311.16867*, 2023.

[10] J. Bai *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[11] D. J. Zhang, D. Li, H. Le, M. Z. Shou, C. Xiong, and D. Sahoo, "Moonshot: Towards controllable video generation and editing with multimodal conditions," *arXiv preprint arXiv:2401.01827*, 2024.

[12] A. Q. Jiang *et al.*, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023.

[13] Baichuan, "Baichuan 2: Open large-scale language models," *arXiv preprint arXiv:2309.10305*, 2023. [Online]. Available: https://arxiv.org/abs/2309.10305

[14] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv preprint arXiv:2307.09288*, 2023.

[15] Z. Cai *et al.*, "Internlm2 technical report," *arXiv preprint arXiv:2403.17297*, 2024.

[16] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "GLM: General Language Model Pretraining with Autoregressive Blank Infilling," in *ACL (Volume 1: Long Papers)*, 2022, pp. 320–335.

[17] 01.AI, :, A. Young, B. Chen, C. Li, C. Huang, G. Zhang, G. Zhang, H. Li, J. Zhu, J. Chen, J. Chang, K. Yu, P. Liu, Q. Liu, S. Yue, S. Yang, S. Yang, T. Yu, W. Xie, W. Huang, X. Hu, X. Ren, X. Niu, P. Nie, Y. Xu, Y. Liu, Y. Wang, Y. Cai, Z. Gu, Z. Liu, and Z. Dai, "Yi: Open Foundation Models by 01.AI," *arXiv preprint arXiv:2403.04652*, 2024.