

STELLA: Towards Protein Function Prediction with Multimodal LLMs Integrating Sequence-Structure Representations

Anonymous ACL submission

Abstract

Protein biology focuses on the intricate relationships among sequences, structures, and functions. Deciphering protein functions is crucial for understanding biological processes, advancing drug discovery, and enabling synthetic biology applications. Since protein sequences determine tertiary structures, which in turn govern functions, integrating sequence and structure information is essential for accurate prediction of protein functions. Traditional protein language models (pLMs) have advanced protein-related tasks by learning representations from large-scale sequence and structure data. However, pLMs are limited in integrating broader contextual knowledge, particularly regarding functional modalities that are fundamental to protein biology. In contrast, large language models (LLMs) have exhibited outstanding performance in contextual understanding, reasoning, and generation across diverse domains. Leveraging these capabilities, STELLA is proposed as a multimodal LLM integrating protein sequence-structure representations with general knowledge to address protein function prediction. Through multimodal instruction tuning (MMIT) using the proposed OPI-Struc dataset, STELLA achieves state-of-the-art performance in two function-related tasks—functional description prediction (FP) and enzyme-catalyzed reaction prediction (EP). This study highlights the potential of multimodal LLMs as an alternative paradigm to pLMs to advance protein biology research. The project codes can be accessed via <https://anonymous.4open.science/r/STELLA-DF00>.

1 Introduction

Protein biology revolves around the interaction of three data modalities: sequences, structures, and functions (texts). The principle ‘sequences determine structures, and structures determine functions’ highlights the critical relationship among protein sequences, tertiary structures, and biological func-

tions. Specifically, a protein’s tertiary structure dictates how it interacts with other molecules, such as substrates or inhibitors, directly affecting its biological activity, including enzymatic catalysis and molecular recognition. Accurate understanding of these biological functions is significant for advancing research on disease pathology, where protein dysfunctions are often implicated. It also plays an important role in target identification for drug discovery, metabolic pathway analysis for engineered biosynthesis, and the design of custom enzymes for industrial and biotechnological applications.

Decades of research in protein science have led to the accumulation of extensive protein structures, including experimentally determined ones in the RCSB Protein Data Bank (PDB)¹ (Berman et al., 2000) and computationally predicted ones in the AlphaFold Protein Structure Database (AFDB)² (Varadi et al., 2021) generated by AlphaFold 2 (AF2) (Jumper et al., 2021). Although these datasets offer invaluable resources to the community, most protein functions are unknown. Therefore, more efforts are needed to leverage them to decipher protein functions, which includes understanding of molecular functions, biological processes, subcellular localization, enzymatic activities, etc. In this work, we focus on two representative tasks within the broader landscape of protein function prediction: functional description prediction (FP) and enzyme-catalyzed reaction prediction (EP), which together reflect both general and enzymatic aspects of protein functionality. FP aims to predict the comprehensive description of the biological roles proteins play, such as DNA repair mechanisms, signal transduction pathways, and viral replication processes. EP specifically focuses on identifying the catalytic functions of enzymes by predicting the types of reactions they catalyze.

¹<https://www.rcsb.org/>

²<https://alphafold.ebi.ac.uk/>

To address protein function prediction, recent pLMs have been proposed to learn joint sequence-structure representations from large-scale data (Su et al., 2023; Li et al., 2025). Although these methods have shown effectiveness in protein-related tasks, they might struggle to capture the representation of protein functions. To address the limitation, recent studies, including Prot2Text (Abdine et al., 2023), ProteinGPT (Xiao et al., 2024), and ProtChatGPT (Wang et al., 2024a), have explored to apply multimodal large language models (LLMs) to bridge protein sequence-structure representation and function representation. These models employ separate pre-trained encoders to represent protein sequences and structures and subsequently feed the joint representation into language models. Specifically, Prot2Text takes ESM-2 as the sequence encoder and a Relational Graph Convolutional Network (RGCN) for structural encoding. ProteinGPT adopts ESM-2 for sequence representation and takes a GVP-GNN (Jing et al., 2020) to encode structural information. ProtChatGPT employs ESM-1b (Rives et al., 2021b) for primary structure encoding, NetSurfP (Klausen et al., 2019) for secondary structure encoding, and ESM-IF1 (Hsu et al., 2022) for tertiary structure representation. Such approaches introduce architecture complexity, as they rely on the integration of multiple encoders to handle different protein modalities. This not only increases computational overhead, but also complicates model design and optimization. To remedy this, we investigate the feasibility of utilizing ESM3 (Hayes et al., 2024), a recent protein language model (pLM), as a unified protein encoder within the multimodal LLM framework. ESM3 represents sequences and structures in a unified embedding space, thereby offering a promising avenue to simplify model architecture while maintaining representation ability.

This study introduces STELLA, a multimodal LLM that bridges protein language and natural language to enable the learning of complex sequence-structure and function relationships from multimodal data. STELLA employs the open source esm3_sm_open_v1 (1.4B) protein encoder (Hayes et al., 2024) and empowers protein function prediction by leveraging ESM3’s unified encoding ability and the understanding ability of LLMs. This mechanism enables STELLA to interpret protein tertiary structures as well as predict functional descriptions and enzyme-catalyzed reactions based on user prompts. It offers a new paradigm for un-

derstanding proteins and extends the capabilities of LLMs in the field of protein biology. STELLA achieves state-of-the-art performance in both tasks, highlighting the transformative potential of multimodal LLMs to advance protein biology besides the development of pLMS. The primary contributions of this study are outlined below.

1. By leveraging ESM3, STELLA streamlines the model architecture and improves the applicability of multimodal LLMs to protein function prediction, achieving state-of-the-art performance in protein functional description prediction and enzyme-catalyzed reaction prediction tasks.

2. We curate a large-scale multimodal instruction tuning dataset, OPI-Struc, to facilitate the training and evaluation of multimodal LLMs in protein-related tasks. This dataset contains diverse protein structures, functional annotations, and enzyme-catalyzed reaction data.

3. This study introduces the methodology, dataset, model architecture, training, and evaluation of STELLA. It demonstrates the novelty and comprehensiveness in establishing a new paradigm in computational protein science (Fan et al., 2025) based on multimodal LLMs, moving beyond the traditional reliance on pLMs.

2 A first look at STELLA’s capabilities through case studies

STELLA demonstrates satisfactory performance in protein function prediction by integrating sequence-structure representations into LLMs. As illustrated in Figure 1 (left), STELLA excels in following natural language instructions and generating appropriate responses for users. In the example, STELLA correctly identifies the main function—a component of the large ribosomal subunit responsible for the synthesis of proteins in the cell—of a newly reviewed protein G1TFE0 in Swiss-Prot. Additionally, STELLA elaborates on the core constituents of the ribonucleoprotein complex, highlighting its advantage in grasping general knowledge. Furthermore, STELLA showcases its reasoning ability by linking loss of ribosomal function to cellular dysfunctions. In Figure 1 (right), STELLA accurately predicts the function of another newly characterized protein in Swiss-Prot, A0A1D0BR98. Upon further inquiry from the user, it explains the details of the toxin mechanisms and provides treatment suggestions. Both examples demonstrate STELLA’s ability in protein function prediction,

such as delivering informative, contextually relevant responses to diverse user prompts. Moreover, STELLA shows reasoning ability, which enables it to assist domain experts in uncovering connections between protein functions and disease mechanisms. These results highlight its potential to advance protein biology. More examples are shown in Figure 3 of Appendix A.

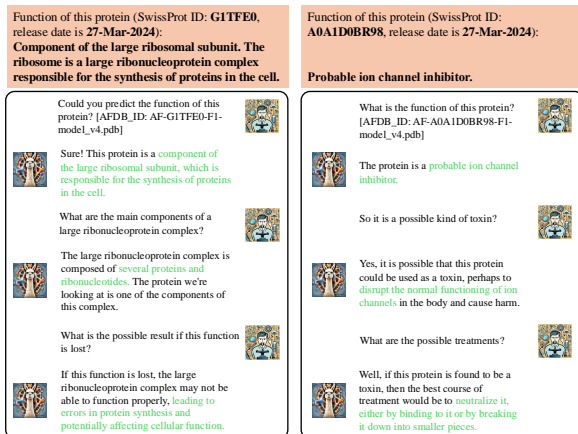


Figure 1: **Demo capability of STELLA. (STELLA-ESM3-Llama-3.1-8B-Instruct).** The examples involve two proteins—**G1TFE0** (left) and **A0A1D0BR98** (right)—sourced from the newly released Swiss-Prot 2024_02. The **orange box** indicates the ground-truth functional annotation. Text highlighted in **green** denotes critical and correct functional information generated by STELLA. User and assistant icons are AI-generated.

3 Related work

3.1 Protein-text modeling

Protein representation learning aims to extract biologically relevant information from diverse data modalities, including protein language (e.g., protein sequences and structures) and natural language (e.g., functional description). Aligning both modalities has emerged as an important aspect in protein biology. For instances, ProtST (Xu et al., 2023) employs contrastive learning to align protein sequence and text modalities, aiming to obtain biologically informative protein embeddings that are applicable to a variety of downstream protein-related tasks. Besides representation learning, ProteinDT (Liu et al., 2023c) leverages textual data to enhance protein design in text-to-sequence generation tasks. Prot2Text (Abdine et al., 2023) proposes a method to align protein structures and functional descriptions using a fused multimodal encoder-decoder framework. Before the prevalence of LLMs, protein representation learning mainly fo-

cuses on single modality like amino acid sequences, or sequence-text alignment by contrastive learning. Rarely research focuses on how to effectively bridge protein tertiary structures and general knowledge encoded within LLMs, which might play an important role in protein biology.

3.2 LLMs for protein biology

Recent studies have brought attention to the potential of LLMs in advancing biomedical research, spanning molecules, proteins, and RNA. In the domain of protein biology, several notable developments have emerged. ProTokens (Lin et al., 2023) employs discrete and compressed protein tokens that encode rich structural information for LLMs. These tokens are learned through an autoencoder framework, with both the input and output consisting of protein tertiary structures. InstructProtein (Wang et al., 2023) constructs instruction datasets derived from a knowledge graph to address the annotation imbalance present in previous protein-text datasets. This dataset is used to fine-tune LLMs for aligning protein sequences and natural language, enabling bidirectional tasks—predicting functions from sequences and generating sequences from natural language. BioMedGPT (Luo et al., 2023) employs a fully-connected layer to connect an amino acid sequence encoder, ESM-2-3B (Lin et al., 2022), and Llama2-Chat-7B (Touvron et al., 2023), which has been incrementally pretrained on biomedical literature from S2ORC (Lo et al., 2020). ProteinGPT (Xiao et al., 2024), ProtChatGPT (Wang et al., 2024a) and ProteinChat (Huo et al., 2024) represent recent multi-modal LLMs designed to predict protein functions or beyond. ProteinChat combines the protein sequence encoder xTrimoPGLM (Chen et al., 2024) and the Vicuna-13B model (Zheng et al., 2023) through a linear adapter and was trained with over 1.5 million protein data triplets (protein, prompt, answer) from Swiss-Prot. By taking a sequence as input, it predicts functional properties of the given protein.

4 Methodology

4.1 STELLA model architecture

Overview. STELLA is a multimodal LLM for protein function prediction, drawing inspiration from LLaVA (Liu et al., 2023a), a prominent vision-language model architecture that integrates vision encoders with LLMs. As illustrated in Figure 2,

STELLA is composed of three key components: a **protein structure encoder**, a **modality connector**, and a **LLM**. Similar to the typical two-stage training strategy employed by LLaVA and other multimodal LLMs such as Bunny (He et al., 2024), STELLA adopts two-stage multimodal instruction tuning (MMIT), which has proven effective in this study. What differs is that STELLA’s two stages of training utilize the same datasets, due to the extreme scarcity of protein instruction data. The prompt templates for training are provided in B, and hyperparameters in Table 7 (Appendix D).

Protein structure encoder. The protein structure encoder is responsible for translating protein tertiary structures into high-dimensional structural representations. In this study, we utilize ESM3, a leading model pretrained on multiple modalities, including sequence, structure, and function tokens. ESM3 encodes these distinct modalities as discrete token tracks and integrates them into a unified embedding space through transformer blocks. Notably, the model incorporates geometric attention in its initial transformer block, effectively capturing atomic-level details of proteins.

Modality connector. The modality connector bridges structural representations derived from the protein structure encoder and protein functional descriptions encoded into text embeddings. In this implementation, a simple linear layer is employed as the adapter, which has also proven effective in previous works like LLaVA (Liu et al., 2023a).

LLM. The LLM integrated in STELLA is Llama-3.1-8B-Instruct (Dubey et al., 2024), a highly capable model that excels across multiple benchmarks, including general knowledge (Hendrycks et al., 2021a; Wang et al., 2024b; Zhou et al., 2023), mathematics (Cobbe et al., 2021; Hendrycks et al., 2021b; Rein et al., 2023; Clark et al., 2018), code generation (Chen and et al., 2021; Liu et al., 2023b), tool-use (Yan et al., 2024; Srinivasan et al., 2023), long context tasks (Zhang et al., 2024) and multilingual ability (Shi et al., 2022). Additionally, the model exhibits strong safety features, supported by Llama Guard 3, ensuring reliable performance in sensitive tasks.

4.2 Task definition

Functional description prediction (FP). STELLA effectively aligns protein language and natural language representations via MMIT, enabling accurate prediction of protein function based on tertiary structures. The integration of LLM supports

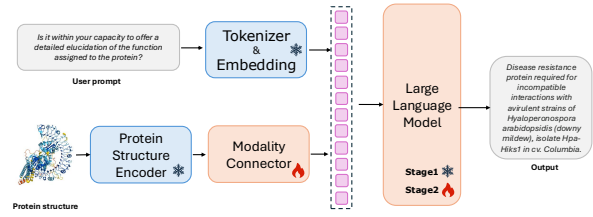


Figure 2: **Overview of STELLA architecture.** Stage1 of MMIT fine-tunes the modality connector on the OPI-Struc dataset while keeping the protein encoder and LLM frozen. Stage2 continues training by jointly fine-tuning the modality connector and LLM with different learning rates, while the protein encoder remains frozen. Flame indicates trainable components; Snowflake indicates frozen components. Protein credits: AFDB.

diagonal-based interactions between STELLA and users. This approach offers a flexible way to predict protein functions.

Enzyme-catalyzed reaction prediction (EP).

In this study, enzyme-catalyzed reactions are mapped to their corresponding enzyme names, which serve as proxies for reactions in which the associated proteins are involved. This approach allows for more seamless integration with LLMs, ensuring that enzyme functions are captured in a way that aligns with the abilities of STELLA.

4.3 OPI-Struc dataset

Data overview. The **Open Protein Instructions for Structures (OPI-Struc)** dataset was curated to support multimodal instruction tuning (MMIT), by integrating protein structural and textual modalities. Corresponding to the **FP** and **EP** task, OPI-Struc was organized into two main categories: **Function** and **Enzyme** (see Appendix L, example L.4). The **Function** dataset was further divided into two sub-categories: **Func_{ft}** (see Appendix L, example L.1) and **Func_{mc}** (see Appendix L, example L.3) based on label formats: free-text question-answer (ft) and multiple-choice question-answer (mc), respectively. Additionally, to reflect the iterative nature of scientific discovery, 20% (49,663 samples) of the **Func_{ft_train}** dataset were randomly selected to be augmented with enriched questions and responses generated by Llama-2-13B-Chat, forming the **Func_{ft_train_aug}** dataset (see Appendix L, example L.2). The data augmentation methods are detailed in Appendix J. The training and testing sets and corresponding statistics are presented in Table 1. To clarify, while ESM3’s training data includes protein-related functional keywords, this

work utilizes free-text functional descriptions of proteins. This distinction ensures that our testing data are not explicitly included in ESM3’s pre-training data, namely data leakage.

Table 1: **Statistics of OPI-Struc.** The FP task includes FP_{ft} and FP_{mc} subtasks. FP_{ft} uses both a hold-out test set and a newer Swiss-Prot release (v2401) for evaluation. FP_{mc} includes two test variants (w/ and w/o permutation). See Appendix L for examples L.1–L.4.

Task	FP_{ft}
Training set/size	Func _{ft_train} (+aug)/248,315 (+49,663)
Testing set/size	Func _{ft_test} /4,203, Func _{ft_test_v2401} /270
Metrics	BLEU-4, BERT score, ROUGE
Protein source	AFDB
Task	FP_{mc}
Training set/size	Func _{mc_train} /24,000
Testing set/size	Func _{mc_test_1x} /4,203, Func _{mc_test_4x} /16,812
Metrics	Accuracy
Protein source	AFDB
Task	EP
Training set/size	Enzyme _{train} /29,205
Testing set/size	Enzyme _{test} /5,651
Metrics	Accuracy
Protein source	PDB

Data explanation. Each sample of the OPI-Struc dataset comprises a protein tertiary structure-sourced from either AFDB or PDB-paired with task-specific natural language instructions formatted as conversations, along with corresponding response labels. For the **Function** dataset, protein structures are obtained from AFDB, while protein functional descriptions are derived from the release 2022_04³ of Swiss-Prot⁴. To mitigate answer bias during model training, the multiple-choice training set (**Func**_{mc_train}) is constructed by randomly permuting the four answer options (A, B, C, D) for each sample. For evaluation, we provide two variants of the testing set **Func**_{mc_test}—one with a fixed answer order (1x) and one with randomized answer permutations (4x), ensuring a more robust assessment of model generalization across consistent and variable answer configurations. The **Enzyme** dataset is sourced from the SIFTS database (Dana et al., 2018), with original labels annotated using Enzyme Commission (EC) numbers. These EC numbers are subsequently mapped to enzyme names via the BRENDA En-

zyme Database⁵ (e.g., *1.1.1.10* → *L-xylulose reductase*). To ensure data quality and consistency, the entire OPI-Struc dataset underwent a rigorous preprocessing pipeline, adhering to established data cleaning methods. All metadata, including PubMed IDs, ECO IDs, and auxiliary annotations, were systematically removed to enhance dataset purity and reliability. This preprocessing step is essential for minimizing noise and improving the robustness of downstream modeling. We further conducted detailed statistical analysis to characterize the dataset and highlight its comprehensiveness. For instance, we examined the distribution of protein sequence lengths—a proxy for structural complexity—as shown in Figure 4 in Appendix E. The broad variation in sequence lengths reflects the dataset’s extensive coverage across proteins of varying structural complexities. In addition, we analyzed label characteristics such as the distribution of functional description lengths and the frequency of enzyme names (see Figure 5 in Appendix E). These analysis underscores that it is important to design models that are robust across diverse structural and functional complexities to ensure consistent and reliable performance during evaluation.

Data split. To ensure consistency, (1) the **Function** dataset was split following the same strategy adopted by (Abdine et al., 2023), with a maximum of 40% sequence similarity allowed between the training and testing sets to ensure a more rigorous evaluation; (2) the **Enzyme** dataset was split according to the method described in (Hermosilla et al., 2021).

Diversification of instructions The raw data were transformed into an instruction-based format to support instruction tuning. To achieve variation in instruction phrasing, we used web-based ChatGPT (GPT-3.5) to generate rephrased instructions. For example, using the query: “*Could you provide 100 alternative ways to rephrase the prompt ‘Please describe the function of the protein’?*”, approximately 100 distinct variations of task instructions were produced (see the instruction list K.1, K.2 and K.3 in Appendix K). Each generated instruction was manually reviewed for accuracy and relevance, ensuring that only high-quality variations were included in the final **Function** dataset. We took a similar way to diversify its instructions (see the instruction list K.4 and K.5 in Appendix K) for the **Enzyme** dataset.

³https://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2022_04/knowledgebase/UniProtKB_SwissProt-relstat.html

⁴<https://www.uniprot.org/uniprotkb?query=reviewed:true>

⁵<https://www.brenda-enzymes.org/>

5 Evaluation of STELLA

This study sheds light on applying multimodal LLMs to address protein function prediction tasks. By systematically evaluating STELLA’s performance in the **FP** and **EP** tasks, we seek to elucidate both the strengths and limitations of this multimodal LLM in protein biology. We have designed **five distinct evaluations** based on the corresponding testing sets detailed in Table 1, including **FP_{ft_eval}**, **FP_{ft_eval_v2401}**, **FP_{mc_eval_1x}**, **FP_{mc_eval_4x}**, **EP_{eval}**. Hyperparameters for evaluation are presented in Appendix D, and user prompts for evaluation are listed in Table 6 (Appendix C).

5.1 Evaluation metrics

Multiple typical metrics for natural language processing (NLP) tasks, including BLEU, BERT score, and ROUGE, are employed for comprehensive evaluation of the **FP** task. BLEU is typically used to assess the similarity between two sequences in machine translation. In particular, this study employs BLEU-4, a metric that evaluates the degree of 4-gram overlap between the generated and reference texts. BERT score evaluates the token-level similarity between a generated sentence and a reference sentence. ROUGE is a suite of evaluation metrics traditionally used to compare generated text with reference texts to evaluate their overlap. It includes ROUGE-1, ROUGE-2, and ROUGE-L, each based on different n-gram strategies. Among these, ROUGE-L is particularly effective for assessing summarization and translation quality, as it captures overall sentence structure by emphasizing the longest common subsequence of the original and generated texts. Currently, there are no universally established evaluation metrics specifically tailored for biological text generation. Therefore, BLEU, BERT score, and ROUGE have also been widely adopted in prior works to assess text generation tasks in biological contexts. However, we recognize the limitations of these NLP metrics in aligning with biological accuracy. Therefore, besides the NLP metric and as part of our evaluation, we design a **multiple-choice QA (MCQA) subtask**, **FP_{mc}**, which adopts Accuracy as metrics, to objectively assess STELLA’s performance. In addition, the **EP** task takes Accuracy as metrics.

5.2 Results of Functional Description Prediction

The **FP_{ft_eval}** is designed using the hold-out testing set **Func_{ft_test}**, which is also utilized for eval-

uation in Prot2Text (Abdine et al., 2023). It aims to assess STELLA’s capability to predict protein functional descriptions. STELLA demonstrates state-of-the-art overall performance, surpassing Prot2Text_{BASE} and Prot2Text_{LARGE} in the **FP_{ft_eval}**, as shown in Table 2.

Comparison between STELLA and Foldseek.

We took Foldseek as a baseline, which includes two steps: structure retrieval using Foldseek (Van Kempen et al., 2024) and function mapping from Swiss-Prot. In the first step, we used the Foldseek easy-search⁶ tool with default parameters to search similar protein structures within the training set for each of the 4,203 testing samples. Only the matched results with an e-value < 0.001 were considered and returned. In the second step, the corresponding functional description was determined based on the top-1 retrieved protein from Swiss-Prot. The median e-value of the top-1 retrieved proteins is 2.723e-20, indicating a high confidence in the retrieval results by Foldseek. As shown in Table 2, STELLA achieves a 14.6% improvement over Foldseek in terms of ROUGE-L, demonstrating STELLA’s advantage compared to retrieval-based methods.

Evaluation with incomplete protein structures. Incomplete protein data pose challenges in protein biology. To address this, we conducted an additional experiment to evaluate STELLA’s ability to handle incomplete protein structures. Specifically, for the testing data, we cut away the terminal 10% of the protein structures to simulate incomplete structural information and assess the model’s performance under these conditions. STELLA’s performance see a slight decrease from 0.5257 to 0.4805 on ROUGE-L. Considering that the training procedure was conducted using complete protein structures, this slight decrease due to inconsistency still demonstrates the robustness of STELLA, indicating its applicability in more challenging tasks.

Additional evaluation in the MCQA subtask.

We recognize that **FP_{ft_eval}** may be impacted by linguistic variability, where generated responses with correct meanings differ in expression from the reference. Therefore, we designed **FP_{mc_eval_1x}** and **FP_{mc_eval_4x}** to eliminate ambiguity by providing predefined answer choices, which allow for a more objective and standardized evaluation. This method requires STELLA to not only identify

⁶<https://github.com/steineggerlab/foldseek?tab=readme-ov-file#search>

Table 2: **Evaluation results of the FP task, comparing with existing work.** Training recipes for STELLA-ESM3-Llama-3.1-8B-Instruct: Func_{ft_train} dataset. **Bold** and underline indicate the best and the runner-up performance.

Evaluation	Model/Method	BLEU-4 \uparrow	BERT Score \uparrow	ROUGE Score \uparrow		
				ROUGE-1	ROUGE-2	ROUGE-L
FP_{ft_eval}	Prot2Text _{BASE}	0.3511	0.8430	0.5059	0.4271	0.4849
	Prot2Text _{LARGE}	0.3629	<u>0.8520</u>	<u>0.5368</u>	<u>0.4560</u>	<u>0.5140</u>
	STELLA-ESM3-Llama-3.1-8B-Instruct (e3+e3)	<u>0.4024</u>	0.8496	0.5218	0.4487	0.5041
	STELLA-ESM3-Llama-3.1-8B-Instruct (e3+e6)	0.4300	0.8564	0.5423	0.4747	0.5257
	Foldseek	0.3627	0.8358	0.4799	0.4027	0.4586
	Incomplete protein	0.3793	0.8421	0.4991	0.4241	0.4805

the correct answer but also engage in reasoning and option filtering based on contextual knowledge, thus providing a more comprehensive and robust assessment of its capabilities. STELLA achieves accuracies at $\text{Acc@FP}_{mc_eval_1x} = 80.56$ and $\text{Acc@FP}_{mc_eval_4x} = 76.18$, exhibiting advantageous capabilities in terms of instruction following and protein understanding. Notably, without integrating LLMs, baseline models like vanilla ESM3 and Prot2Text cannot response in a MCQA format.

5.3 Results of Enzyme Name Prediction

EP_{eval} aims to assess STELLA’s ability in enzyme name prediction. On top of the original Enzyme_{train} set, we exclude 10 samples due to their associated PDB files lacking certain atom coordinates necessary for embedding extraction with the protein structure encoder in STELLA. As shown in Table 3, we witnessed the performance from $\text{Accuracy} = 88.06$ to $\text{Accuracy} = 88.85$, by increasing the training epoch from 3 to 6 in stage-2 training. At last, STELLA achieved a state-of-the-art result in the EP task, surpassing previous best $\text{Accuracy} = 88.50$ in CDConv (Fan et al., 2022) and Sable (Li et al., 2025).

5.4 Ablation study

5.4.1 Ablation of protein encoders and LLMs

To further investigate the representative ability of different protein encoders, we visualized 4,203 protein structure embeddings from the testing set, Func_{ft_test} , generated by ESM3, Prot2Text (Abdine et al., 2023), and SaProt (Su et al., 2023), using UMAP, as illustrated in Figure 6. The figures reveal that for the five most frequently occurring functions in the testing set, proteins with the same function tend to form more compact clusters in the ESM3 representation space compared to the other two encoders. A detailed description of the three encoders is provided in Appendix H. Furthermore, several leading LLMs, outlined in Table 8

(Appendix F), were experimented via the STELLA architecture, enabling an analysis of their impact on STELLA’s performance. The ablation results in Table 4 indicate that the integration of ESM3 and Llama-3.1 yielded the best performance in the FP and EP tasks. Regarding the performance in the FP task, we observe a drop in $\text{FP}_{ft_eval_v2401}$ on the newly Swiss-Prot release, where new characterized proteins may introduce novel structural or functional patterns that are underrepresented in earlier training data, reflecting distributional shifts. This phenomenon is a recognized challenge in representation learning, as biological databases continually evolve. We acknowledge the importance of improving generalization to newly characterized proteins. In the future, we plan to explore efficient continual learning strategies and leverage external functional annotations to enhance STELLA’s adaptability.

5.4.2 Ablation of training strategies

Training STELLA centers around two objectives: cross-modality feature alignment and instruction tuning. A two-stage training strategy decouples these objectives, avoiding potential conflicts during optimization. Stage-1 focuses on aligning protein embeddings with textual embeddings by training a modality connector, enabling the LLM to interpret protein features and mitigating the interference caused by the distribution gap between modalities. Stage-2 emphasizes instruction tuning and task-specific modeling, further enhancing the LLM’s generation capabilities and generalization performance. Moreover, two-stage training might help prevent catastrophic forgetting—without the first-stage proper modality alignment, the model may overfit to textual input and neglect protein features during instruction tuning. Finally, the two-stage training strategy allows for higher flexibility in terms of response-format prompts (e.g., MCQA templates) in the second stage, which facilitates more controlled and task-aligned output genera-

Table 3: **Evaluation results of the EP task.** Acc@EP: predictions that exactly match the ground truth. **Single:** Enzyme_{train} dataset. **Bold** and underline indicate the best and the runner-up performance.

Model	Training manner	Acc@EP \uparrow
UniRep (Alley et al., 2019)	w/o pretrain	72.90
3DCNN (Derevyanko et al., 2018)	w/o pretrain	78.80
TAPE-LSTM (Rao et al., 2019)	w/o pretrain	79.90
HH-suite3 (Steinegger et al., 2019)	w/o pretrain	82.60
GearNet-Edge-IEConv (Zhang et al., 2022)	w/o pretrain	85.30
IEConv (Hermosilla et al., 2021)	w/o pretrain	87.20
New IEConv (Hermosilla and Ropinski)	w/o pretrain	87.20
CDCConv (Fan et al., 2022)	w/o pretrain	<u>88.50</u>
DeepFRI (Gligorijević et al., 2021)	w/ pretrain	63.30
ProtBERT-BFD (Elnaggar et al., 2022)	w/ pretrain	72.20
ESM-1b (Rives et al., 2021a)	w/ pretrain	83.10
GearNet-Multiview-Contrast (Zhang et al., 2022)	w/ pretrain	87.50
New IEConv (Hermosilla and Ropinski, 2022)	w/ pretrain	88.10
Sable (Li et al., 2025)	w/ pretrain	<u>88.50</u>
STELLA-ESM3-Llama-3.1-8B-Instruct(single,two-stage,e3+e3)	MMIT	88.06
STELLA-ESM3-Llama-3.1-8B-Instruct(single,two-stage,e3+e6)	MMIT	88.85

Table 4: **Ablation of protein encoders and LLMs.** Training recipes: using single Func_{ft_train} dataset, training epochs of two stages are both e3. **Bold** and underline indicate the best and runner-up result.

Model	BLEU-4 \uparrow	BERT Score \uparrow	ROUGE Score \uparrow		
			ROUGE-1	ROUGE-2	ROUGE-L
Evaluation - FP_{ft_eval}					
STELLA-ESM3-Llama-3.1-8B-Instruct	0.4024	0.8496	0.5218	0.4487	0.5041
STELLA-ESM3-Llama-3.1-8B-Instruct	<u>0.4020</u>	0.8503	0.5138	<u>0.4478</u>	0.5001
STELLA-ESM3-Phi-3-mini-128k-instruct	0.3807	0.8435	0.4991	0.4273	0.4839
STELLA-ProtText-Llama-3.1-8B-Instruct	0.4009	0.8497	0.5284	0.4454	<u>0.5031</u>
STELLA-ProtText-Llama-3.1-8B-Instruct	0.3892	0.8456	0.5177	0.4329	0.4915
STELLA-ProtText-Phi-3-mini-128k-instruct	0.3771	0.8426	0.5058	0.4210	0.4799
STELLA-ProtText-Mistral-7B-Instruct-v0.2	0.3889	<u>0.8525</u>	0.5224	0.4359	0.4949
STELLA-ProtText-BioMedGPT-LM-7B	0.3999	0.8488	<u>0.5282</u>	0.4447	0.5020
STELLA-ProtText-BioMistral-7B-DARE	0.3870	0.8533	0.5241	0.4357	0.4980
STELLA-SaProt-Llama-3.1-8B-Instruct	0.3588	0.8276	0.4685	0.3965	0.4523
STELLA-SaProt-Mistral-7B-Instruct-v0.2	0.3514	0.8251	0.4607	0.3894	0.4455
Evaluation - FP_{ft_eval}, v2401					
STELLA-ESM3-Llama-3.1-8B-Instruct	<u>0.0489</u>	0.7565	0.2210	0.1085	0.1867
STELLA-ProtText-Llama-3.1-8B-Instruct	0.0425	0.7555	0.2454	0.1020	<u>0.1919</u>
STELLA-ProtText-Llama-3.1-8B-Instruct	0.0510	<u>0.7605</u>	<u>0.2486</u>	<u>0.1062</u>	0.1918
STELLA-ProtText-Mistral-7B-Instruct-v0.2	0.0440	0.7685	0.2529	0.1046	0.1975

Table 5: **Ablation of training strategies for STELLA-ESM3-Llama-3.1-8B-Instruct in FP_{ft_eval}.** Training dataset: Func_{ft_train}. **Bold** indicates the best result.

Training strategy	Stage1 epoch	Stage2 epoch	BLEU-4 \uparrow	BERT Score \uparrow	ROUGE Score \uparrow		
					ROUGE-1	ROUGE-2	ROUGE-L
one-stage	-	e1	0.2233	0.7885	0.3530	0.2631	0.3350
one-stage	-	e2	0.3099	0.8199	0.4346	0.3522	0.4160
one-stage	-	e3	0.3642	0.8363	0.4840	0.4073	0.4660
two-stage	e3	e1	0.2653	0.8065	0.3938	0.3097	0.3770
two-stage	e3	e2	0.3574	0.8363	0.4790	0.4028	0.4617
two-stage	e3	e3	0.4024	0.8496	0.5218	0.4487	0.5041

tion. Empirically, our experimental results validate the effectiveness of the two-stage strategy. As shown in Table 5, two-stage training consistently outperforms their one-stage counterparts across all evaluation metrics in the FP_{ft_eval}. For the two-stage training, although both stages utilize the same data, the differences in initialization and learning

rates between the two stages contribute to improved model convergence and overall performance.

6 Conclusion and future work

This work proposes STELLA, a multimodal LLM leveraging the integrated encoding ability of ESM3, which bridges protein sequence-structure representations with general knowledge within LLMs. Through multimodal instruction tuning using the proposed OPI-Struc dataset, STELLA achieves state-of-the-art performance in two tasks—protein functional description prediction and enzyme-catalyzed reaction prediction. Beyond excelling in both tasks, STELLA establishes an innovative paradigm based on multimodal LLMs for protein-related tasks, complementing traditional pLM-based approaches. This work demonstrates the transformative potential of STELLA as a novel computational method to assist protein biology research. Looking further, future efforts should focus on designing more effective specialized tokenizers for sequences and structures, encompassing more downstream tasks tailored to multimodal LLMs, and exploring retrieval-augmented generation or agent-based techniques to enhance model capabilities. These advances will unlock the great potential of multimodal LLMs and solidify their role in driving the next generation innovation of computational approaches in protein biology and beyond.

Limitations

While STELLA demonstrates significant advances in protein-related tasks as a novel computational approach, it still faces limitations in generalizing to more diverse tasks. Future improvements could be achieved through the development of specialized tokenizers and broader integration of retrieval-augmented generation and agent-based techniques to enhance the model’s versatility. These advancements will further solidify the role of multimodal LLMs as promising computational models in the field of protein biology.

Ethics Statement

To the best of our knowledge, the proposed OPI-Struc dataset is constructed from public databases and does not involve any sensitive information.

References

- Marah Abdin, Jyoti Aneja, and et al Hany Awadalla. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *Preprint*, arXiv:2404.14219.
- Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. 2023. [Prot2Text: Multimodal protein’s function generation with gnns and transformers](#). *arXiv preprint arXiv:2307.14367*.
- AI@Meta. 2024. [Llama 3 model card](#).
- Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. 2019. [Unified rational protein engineering with sequence-based deep representation learning](#). *Nature methods*, 16(12):1315–1322.
- Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. [The Protein Data Bank](#). *Nucleic Acids Research*, 28(1):235–242.
- Bo Chen, Xingyi Cheng, Pan Li, Yangli-ao Geng, Jing Gong, Shen Li, Zhilei Bei, Xu Tan, Boyan Wang, Xin Zeng, and 1 others. 2024. [xtrimopglm: unified 100b-scale pre-trained transformer for deciphering the language of protein](#). *arXiv preprint arXiv:2401.06199*.
- Mark Chen and Jerry Tworek et al. 2021. [Evaluating large language models trained on code](#). *Preprint*, arXiv:2107.03374.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#). *Preprint*, arXiv:1803.05457.

- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Jose M Dana, Aleksandras Gutmanas, Nidhi Tyagi, Guoying Qi, Claire O’Donovan, Maria Martin, and Sameer Velankar. 2018. [SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins](#). *Nucleic Acids Research*, 47(D1):D482–D489.
- Georgy Derevyanko, Sergei Grudinin, Yoshua Bengio, and Guillaume Lamoureaux. 2018. [Deep convolutional networks for quality assessment of protein folds](#). *Bioinformatics*, 34(23):4046–4053.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Ahmed Elnaggar, Michael Heinzinger, Christian Dal-lago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. 2022. [ProtTrans: Toward understanding the language of life through self-supervised learning](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127.
- Hehe Fan, Zhangyang Wang, Yi Yang, and Mohan Kankanhalli. 2022. [Continuous-discrete convolution for geometry-sequence modeling in proteins](#). In *The Eleventh International Conference on Learning Representations*.
- Wenqi Fan, Yi Zhou, Shijie Wang, Yuyao Yan, Hui Liu, Qian Zhao, Le Song, and Qing Li. 2025. [Computational protein science in the era of large language models \(llms\)](#). *arXiv preprint arXiv:2501.10282*.
- Vladimir Gligoričević, P Douglas Renfrew, Tomasz Kosciółek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, and 1 others. 2021. [Structure-based protein function prediction using graph convolutional networks](#). *Nature communications*, 12(1):3168.
- Tomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, and 1 others. 2024. [Simulating 500 million years of evolution with a language model](#). *bioRxiv*, pages 2024–07.
- Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yuezhe Wang, Tiejun Huang, and Bo Zhao. 2024. [Efficient multimodal learning from data-centric perspective](#). *Preprint*, arXiv:2402.11530.

744	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou,	Morten Nielsen, Bent Petersen, and 1 others. 2019.	800
745	Mantas Mazeika, Dawn Song, and Jacob Steinhardt.	NetSurfP-2.0: Improved prediction of protein struc-	801
746	2021a. Measuring massive multitask language under-	tural features by integrated deep learning . <i>Proteins:</i>	802
747	standing . <i>Preprint</i> , arXiv:2009.03300.	<i>Structure, Function, and Bioinformatics</i> , 87(6):520–	803
		527.	804
748	Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul	Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-	805
749	Arora, Steven Basart, Eric Tang, Dawn Song, and	Antoine Gourraud, Mickael Rouvier, and Richard	806
750	Jacob Steinhardt. 2021b. Measuring mathematical	Dufour. 2024. Biomistral: A collection of open-	807
751	problem solving with the math dataset . <i>Preprint</i> ,	source pretrained large language models for medical	808
752	arXiv:2103.03874.	domains . <i>Preprint</i> , arXiv:2402.10373.	809
753	Pedro Hermosilla and Timo Ropinski. Contrastive rep-	Jiashan Li, Xi Chen, He Huang, Mingliang Zeng,	810
754	resentation learning for 3D protein structures .	Jingcheng Yu, Xinqi Gong, and Qiwei Ye. 2025.	811
755	Pedro Hermosilla and Timo Ropinski. 2022. Con-	Sable: bridging the gap in protein structure un-	812
756	trastive representation learning for 3D protein struc-	derstanding with an empowering and versatile pre-	813
757	tures . <i>arXiv preprint arXiv:2205.15675</i> .	training paradigm . <i>Briefings in Bioinformatics</i> ,	814
758	Pedro Hermosilla, Marco Schäfer, Matej Lang, Gloria	26(2):bbaf120.	815
759	Fackelmann, Pere-Pau Vázquez, Barbora Kozlikova,	Xiaohan Lin, Zhenyu Chen, Yanheng Li, Xingyu Lu,	816
760	Michael Krone, Tobias Ritschel, and Timo Ropinski.	Chuanliu Fan, Ziqiang Cao, Shihao Feng, Yi Qin Gao,	817
761	2021. Intrinsic-extrinsic convolution and pooling for	and Jun Zhang. 2023. ProTokens: A machine-learned	818
762	learning on 3D protein structures . In <i>International</i>	language for compact and informative encoding of	819
763	<i>Conference on Learning Representations</i> .	protein 3D structures . <i>bioRxiv</i> , pages 2023–11.	820
764	Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin,	Zeming Lin, Halil Akin, Roshan Rao, Brian Hie,	821
765	Brian Hie, Tom Sercu, Adam Lerer, and Alexander	Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa,	822
766	Rives. 2022. Learning inverse folding from millions	Maryam Fazel-Zarandi, Tom Sercu, Sal Candido,	823
767	of predicted structures . In <i>International conference</i>	and 1 others. 2022. Language models of protein	824
768	on machine learning , pages 8946–8970. PMLR.	sequences at the scale of evolution enable accurate	825
769	Mingjia Huo, Han Guo, Xingyi Cheng, Digvijay	structure prediction . <i>BioRxiv</i> , 2022:500902.	826
770	Singh, Hamidreza Rahmani, Shen Li, Philipp Gerlof,	Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae	827
771	Trey Ideker, Danielle A. Grotjahn, Elizabeth Villa,	Lee. 2023a. Visual instruction tuning . <i>arXiv preprint</i>	828
772	Le Song, and Pengtao Xie. 2024. Multi-modal large	<i>arXiv:2304.08485</i> .	829
773	language model enables protein function prediction .	Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and	830
774	<i>bioRxiv</i> .	Lingming Zhang. 2023b. Is your code generated	831
775	Albert Q. Jiang, Alexandre Sablayrolles, Arthur Men-	by chatgpt really correct? rigorous evaluation of	832
776	sch, Chris Bamford, Devendra Singh Chaplot, Diego	large language models for code generation . <i>Preprint</i> ,	833
777	de las Casas, Florian Bressand, Gianna Lengyel, Guil-	arXiv:2305.01210.	834
778	laume Lample, Lucile Saulnier, L��lio Renard Lavaud,	Shengchao Liu, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili	835
779	Marie-Anne Lachaux, Pierre Stock, Teven Le Scao,	Nie, Anthony Gitter, Chaowei Xiao, Jian Tang,	836
780	Thibaut Lavril, Thomas Wang, Timoth��e Lacroix,	Hongyu Guo, and Anima Anandkumar. 2023c. A	837
781	and William El Sayed. 2023. Mistral 7b . <i>Preprint</i> ,	text-guided protein design framework . <i>arXiv preprint</i>	838
782	arXiv:2310.06825.	<i>arXiv:2302.04611</i> .	839
783	Bowen Jing, Stephan Eismann, Patricia Suriana,	Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kin-	840
784	Raphael JL Townshend, and Ron Dror. 2020. Learn-	ney, and Daniel Weld. 2020. S2ORC: The semantic	841
785	ing from protein structure with geometric vector per-	scholar open research corpus . In <i>Proceedings of the</i>	842
786	ceptrons . <i>arXiv preprint arXiv:2009.01411</i> .	<i>58th Annual Meeting of the Association for Compu-</i>	843
787	John M. Jumper, Richard Evans, Alexander Pritzel, Tim	<i>tational Linguistics</i> , pages 4969–4983, Online. Asso-	844
788	Green, Michael Figurnov, Olaf Ronneberger, Kathryn	ciation for Computational Linguistics.	845
789	Tunyasuvunakool, Russ Bates, Augustin Z��dek, Anna	Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang,	846
790	Potapenko, Alex Bridgland, Clemens Meyer, Simon	Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023.	847
791	A A Kohl, Andy Ballard, Andrew Cowie, Bernardino	Biomedgpt: Open multimodal generative pre-	848
792	Romera-Paredes, Stanislav Nikolov, Rishub Jain,	trained transformer for biomedicine . <i>Preprint</i> ,	849
793	Jonas Adler, and 15 others. 2021. Highly accurate	arXiv:2308.09442.	850
794	protein structure prediction with alphafold . <i>Nature</i> ,	Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan	851
795	596:583–589.	Duan, Peter Chen, John Canny, Pieter Abbeel, and	852
796	Michael Schantz Klausen, Martin Closter Jes-	Yun Song. 2019. Evaluating protein transfer learn-	853
797	persen, Henrik Nielsen, Kamilla Kjaergaard Jensen,	ing with TAPE . In <i>Advances in Neural Information</i>	854
798	Vanessa Isabell Jurtz, Casper Kaae Soenderby,	<i>Processing Systems</i> , volume 32. Curran Associates,	855
799	Morten Otto Alexander Sommer, Ole Winther,	Inc.	856

857	David Rein, Betty Li Hou, Asa Cooper Stickland,	2021. AlphaFold Protein Structure Database: mas-	913
858	Jackson Petty, Richard Yuanzhe Pang, Julien Di-	sively expanding the structural coverage of protein-	914
859	rani, Julian Michael, and Samuel R. Bowman. 2023.	sequence space with high-accuracy models. <i>Nucleic</i>	915
860	Gpqa: A graduate-level google-proof q&a bench-	<i>Acids Research</i> , 50(D1):D439–D444.	916
861	mark. <i>Preprint</i> , arXiv:2311.12022.		
862	Alexander Rives, Joshua Meier, Tom Sercu, Siddharth	Chao Wang, Hehe Fan, Ruijie Quan, and Yi Yang.	917
863	Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle	2024a. ProtChatGPT: Towards understanding pro-	918
864	Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fer-	teins with large language models. <i>arXiv preprint</i>	919
865	gus. 2021a. Biological structure and function emerge	<i>arXiv:2402.09649.</i>	920
866	from scaling unsupervised learning to 250 million		
867	protein sequences. <i>Proceedings of the National</i>	Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni,	921
868	<i>Academy of Sciences</i> , 118(15):e2016239118.	Abhranil Chandra, Shiguang Guo, Weiming Ren,	922
869	Alexander Rives, Joshua Meier, Tom Sercu, Siddharth	Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max	923
870	Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott,	Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue,	924
871	C Lawrence Zitnick, Jerry Ma, and 1 others. 2021b.	and Wenhui Chen. 2024b. Mmlu-pro: A more robust	925
872	Biological structure and function emerge from scal-	and challenging multi-task language understanding	926
873	ing unsupervised learning to 250 million protein se-	benchmark. <i>Preprint</i> , arXiv:2406.01574.	927
874	quences. <i>Proceedings of the National Academy of</i>		
875	<i>Sciences</i> , 118(15):e2016239118.	Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin,	928
876	Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang,	Xiang Zhuang, Xiaotong Li, and Huajun Chen. 2023.	929
877	Suraj Srivats, Soroush Vosoughi, Hyung Won Chung,	InstructProtein: Aligning human and protein lan-	930
878	Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan	guage via knowledge instruction. <i>arXiv preprint</i>	931
879	Das, and Jason Wei. 2022. Language models are	<i>arXiv:2310.03269.</i>	932
880	multilingual chain-of-thought reasoners. <i>Preprint</i> ,		
881	arXiv:2210.03057.	Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and	933
882	Venkat Krishna Srinivasan, Zhen Dong, Banghua Zhu,	Wei Wang. 2024. ProteinGPT: Multimodal LLM for	934
883	Brian Yu, Hanzi Mao, Damon Mosk-Aoyama, Kurt	protein property prediction and structure understand-	935
884	Keutzer, Jiantao Jiao, and Jian Zhang. 2023. Nexus-	ing. <i>arXiv preprint arXiv:2408.11363.</i>	936
885	raven: a commercially-permissive language model		
886	for function calling. In <i>NeurIPS 2023 Workshop on</i>	Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian	937
887	<i>Instruction Tuning and Instruction Following.</i>	Tang. 2023. ProtST: Multi-modality learning of pro-	938
888	Martin Steinegger, Markus Meier, Milot Mirdita, Har-	tein sequences and biomedical texts. <i>arXiv preprint</i>	939
889	ald Vöhringer, Stephan J Haunsberger, and Johannes	<i>arXiv:2301.12040.</i>	940
890	Söding. 2019. HH-suite3 for fast remote homology	Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun	941
891	detection and deep protein annotation. <i>BMC bioin-</i>	Zhang, Shishir G. Patil, Ion Stoica, and Joseph E.	942
892	<i>formatics</i> , 20:1–15.	Gonzalez. 2024. Berkeley function calling leader-	943
893	Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan,	board.	944
894	Xibin Zhou, and Fajie Yuan. 2023. SaProt: Protein	Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin	945
895	language modeling with structure-aware vocabulary.	Li. 2024. Language models are super mario: Absorb-	946
896	<i>bioRxiv 2023.10.01.560349.</i>	ing abilities from homologous models as a free lunch.	947
897	Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-	<i>Preprint</i> , arXiv:2311.03099.	948
898	bert, Amjad Almahairi, and 1 others. 2023. Llama 2:	Xinrong Zhang, Yingfa Chen, Shengding Hu, Zi-	949
899	Open foundation and fine-tuned chat models. <i>arXiv</i>	hang Xu, Junhao Chen, Moo Khai Hao, Xu Han,	950
900	<i>preprint arXiv:2307.09288.</i>	Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and	951
901	Michel Van Kempen, Stephanie S Kim, Char-	Maosong Sun. 2024. ∞bench: Extending long	952
902	lotte Tumescheit, Milot Mirdita, Jeongjae Lee,	context evaluation beyond 100k tokens. <i>Preprint</i> ,	953
903	Cameron LM Gilchrist, Johannes Söding, and Martin	arXiv:2402.13718.	954
904	Steinegger. 2024. Fast and accurate protein struc-	Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chen-	955
905	ture search with foldseek. <i>Nature biotechnology</i> ,	thamarakshan, Aurelie Lozano, Payel Das, and	956
906	42(2):243–246.	Jian Tang. 2022. Protein representation learning	957
907	Mihaly Varadi, Stephen Anyango, Mandar Deshpande,	by geometric structure pretraining. <i>arXiv preprint</i>	958
908	Sreenath Nair, Cindy Natassia, Galabina Yordanova,	<i>arXiv:2203.06125.</i>	959
909	David Yuan, Oana Stroe, Gemma Wood, Agata Lay-	Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan	960
910	don, Augustin Židek, Tim Green, Kathryn Tunyasu-	Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,	961
911	vunakool, Stig Petersen, John Jumper, Ellen Clancy,	Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang,	962
912	Richard Green, Ankur Vora, Mira Lutfi, and 8 others.	Joseph E. Gonzalez, and Ion Stoica. 2023. Judg-	963
		ing llm-as-a-judge with mt-bench and chatbot arena.	964
		<i>Preprint</i> , arXiv:2306.05685.	965
		Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha	966
		Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and	967
		Le Hou. 2023. Instruction-following evaluation for	968
		large language models. <i>Preprint</i> , arXiv:2311.07911.	969

A Example demonstration of STELLA’s capabilities through case studies

Figure 3 shows two examples of STELLA-ESM3-Llama-3.1-8B-Instruct to uncover protein functions and related properties.

B Prompt template for training

Box B.1: The prompt template of STELLA-ESM3-Llama-3.1-8B-Instruct

```
<|begin\_of\_text|><|start\_header\_id|>user
<|end\_header\_id|>

<structure>
May I request a comprehensive breakdown
outlining the function linked to the
protein?

<|eot\_id|><|start\_header\_id|>assistant<|
end\_header\_id|>

Involved in the gluconeogenesis. Catalyzes
stereospecifically the conversion of
dihydroxyacetone phosphate (DHAP) to D-
glyceraldehyde-3-phosphate (G3P). <|eot\_
id|><|end\_of\_text|>
```

Box B.2: The prompt template of STELLA-Prot2Text-Mistral-7B-Instruct-v0.2

```
<s>[INST] <structure>
May I request a comprehensive breakdown
outlining the function linked to the
protein? [/INST]Involved in the
gluconeogenesis. Catalyzes
stereospecifically the conversion of
dihydroxyacetone phosphate (DHAP) to D-
glyceraldehyde-3-phosphate (G3P)</s>
```

C Prompt template for evaluation

Table 6 presents the user prompts used in the evaluation process. Notably, we design the prompt to ensure that the model outputs only one of the four options (A, B, C, or D) in the FP_{mcqa} task.

D Hyperparameters for training and evaluation

Stage1 aims to align a protein structure embedding space and a text embedding space. In this stage, the modality connector is trainable, while both the protein structure encoder and the LLM are frozen. Stage2 is dedicated to enabling STELLA to follow complicated natural language instructions and generate response dedicated to protein tasks. In this stage, both the modality connector and the

LLM are trained with different learning rates, while the protein structure encoder is still frozen. Both stages use the same training datasets. The training prompt templates follow the examples shown in Appendix B.

The hyperparameters in two stages are summarized in Table 7. It should be noted that we adopt different learning rates for each different components of STELLA to finely control the training process. Especially, in stage2, we set the learning rate of the modality connector larger than LLM backbone, to improve LLMs’ training convergence.

E Analysis of data label distribution of the OPI-Struc dataset

Figures 4 illustrates the distribution of protein sequence lengths across the FP (left) and EP (right) tasks for training and testing sets. Figure 5 shows (a) the length distribution of functional descriptions in the Function dataset and (b) the frequency of enzyme names in the Enzyme dataset.

F Different composition of protein structure encoders and LLMs

The architecture of STELLA is flexible and customizable to integrate various protein encoders and LLMs to form variants. We elaborately choose different protein encoders and LLMs to investigate the effectiveness of different composition of these components, as shown in Table 8.

G Embedding visualization for different protein encoders

Figure 6 illustrates the UMAP visualization of protein structure embeddings, generated by ESM3, Prot2Text, and SaProt, of the 4,203 testing samples in $Func_{ft_test}$.

H Comparison of protein structure encoders

STELLA employs three different encoders ESM3 (Hayes et al., 2024), Prot2Text (Abdine et al., 2023), and SaProt (Su et al., 2023) for ablation experiments. ESM3 and Prot2Text model the interplay of sequence, structure, and function, while SaProt only models the sequence and structure modalities. This setup allows us to assess the impact of each encoder on the overall performance, providing insights into the

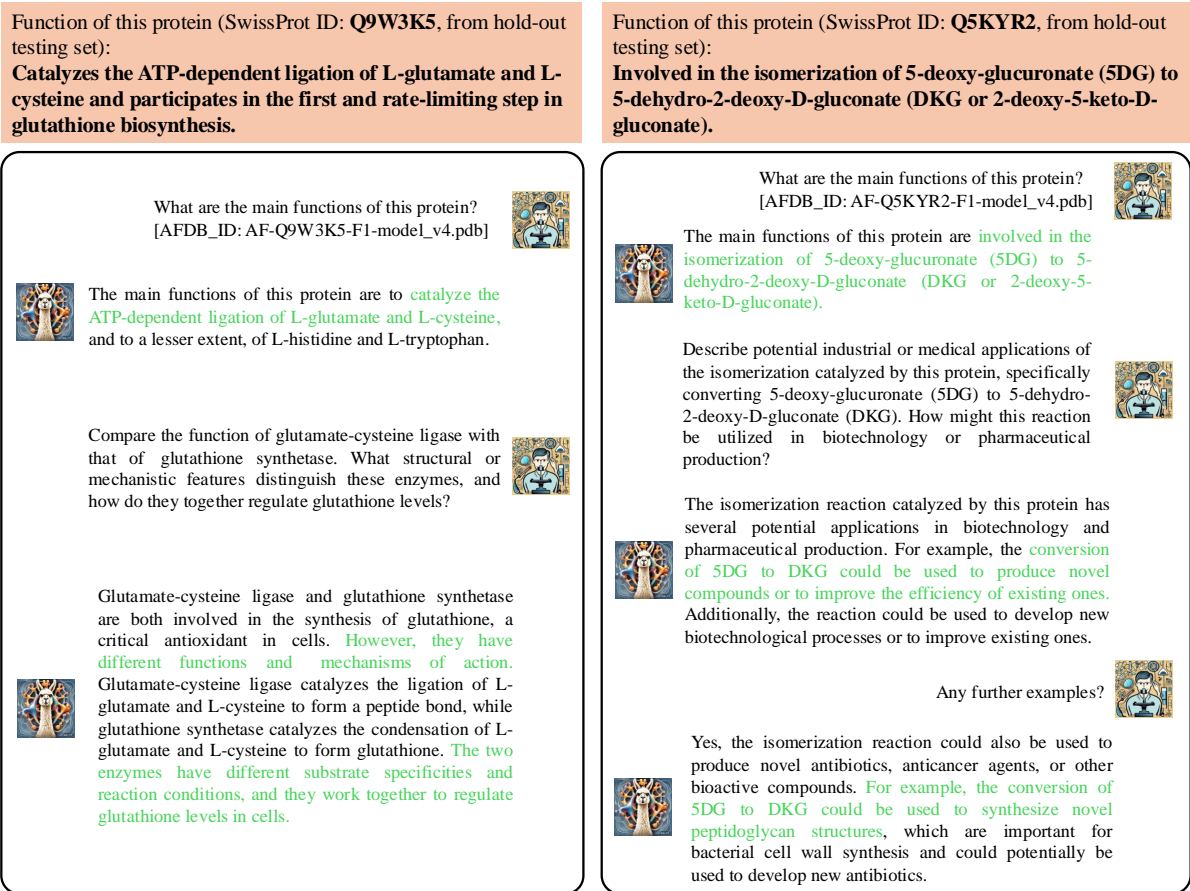


Figure 3: **STELLA’s ability to engage in conversation (Model used: STELLA-ESM3-Llama-3.1-8B-Instruct).** The protein Q9W3K5 and Q5KYR2 are from the hold-out testing set of OPI-Struc. **Left:** protein Q9W3K5. **Right:** protein Q5KYR2. **Orange box:** ground truth of the function. The content highlighted in green font represents the correct and critical output of STELLA. User and assistant icons are AI-generated.

Table 6: **User prompts for evaluation.**

Task	Testing set	Response formatting prompts
FP _{ft}	Func _{ft_test} Func _{ft_test_v2401}	What are the main functions of this protein?
FP _{mc}	Func _{mc_text_1x} Func _{mc_text_4x}	Answer with the option’s letter from the given choices directly. Please respond to the question with an answer choice, which is either A, B, C or D.
EP	Enzyme _{test}	What is the enzyme name linked to this protein?

contributions of different components to the model’s effectiveness.

ESM3 is a large multimodal model pretrained on massive sequence, structure, and function tokens using masked language modeling (MLM). It encodes these modalities as discrete token tracks, which are fused into a unified representation space via several transformer blocks, with geometric attention in the first block to incorporate atomic in-

formation.

Prot2Text is a multimodal model that integrates a Relational Graph Convolution Network (RGCN), ESM-2, and GPT-2 to generate protein function annotation. It combines two sources of information: the output of the RGCN, which processes all-atom protein structures to provide detailed structural representations, and protein sequences processed by ESM-2. The Prot2Text encoder aligns these inte-

Table 7: **Hyperparameters for stage1 training, stage2 training and testing.** FFT: Full Fine-tuning; LoRA: LoRA Tuning

Config	Stage1	Stage2	Testing
DeepSpeed ZeRO Stage	2	3	NA
optimizer	AdamW	AdamW	NA
optimizer hyperparameters	$(\beta_1, \beta_2)=(0.9, 0.999)$, eps=1e-8	$(\beta_1, \beta_2)=(0.9, 0.999)$, eps=1e-8	NA
per_device_train_batch_size	2	1(FFT)/2(LoRA)	NA
gradient_accumulation_steps	4	2(FFT)/4(LoRA)	NA
gradient_checkpointing	True	True	NA
learning rate (lr)	2e-5 (Connector)	2e-4 (Connector), 2e-5 (LLM)	NA
weight decay	0.0	0.0	NA
warmup steps	48	-	NA
warmup ratio	-	0.03	NA
lr scheduler type	cosine	cosine	NA
training epochs	3	3	NA
GPU	4*A100	8*A100(FFT)/4*A100(LoRA)	1*A100
temperature	NA	NA	0.2
top_k	NA	NA	50
top_p	NA	NA	0.75
num_beams	NA	NA	1
max_new_tokens	NA	NA	1000
use_cache	NA	NA	True
do_sample	NA	NA	True

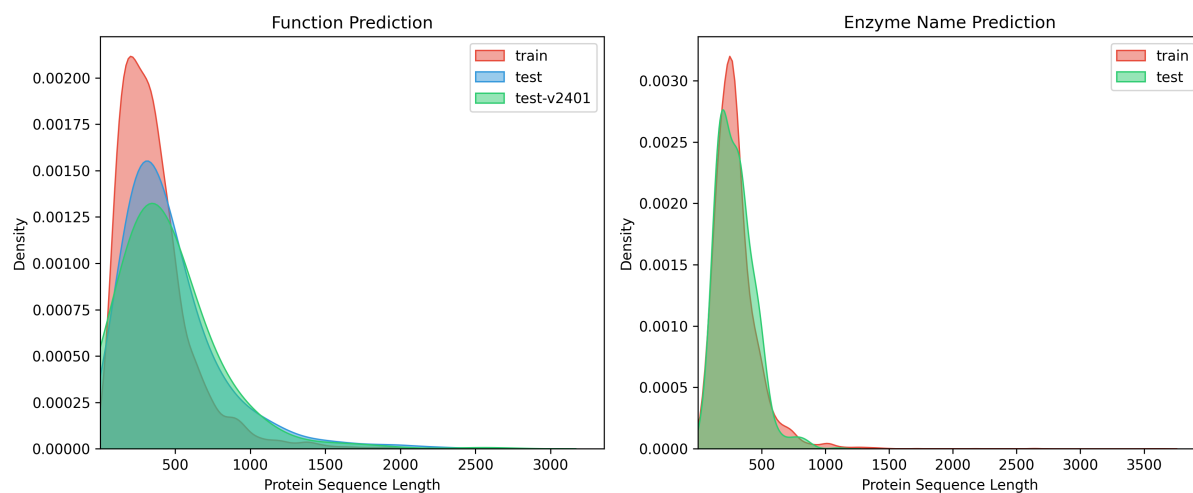


Figure 4: **Distribution of protein sequence lengths across the FP (left) and EP (right) tasks for training and testing sets.** The variation in sequence length distribution between the training and testing sets ensures model robustness across proteins with diverse structural complexities.

grated data with functional annotation through a generative alignment approach using a text decoder.

SaProt is a large-scale pre-trained model utilizing approximately 40 million protein sequences and structures, with a structure-aware vocabulary that integrates residue tokens and structural tokens simultaneously. It adopts an ESM-based architecture that takes structure-aware protein sequences as input, which combine protein sequence residue

tokens and discrete structural tokens encoded using foldseek (Van Kempen et al., 2024). However, this encoder is not aligned with functional annotation text.

Table 8: Specifications of STELLA composition of various protein structure encoders and foundation LLMs.

Protein encoder	LLM	Note of LLM	STELLA variant
ESM3 (Hayes et al., 2024)	Llama-3.1-8B-Instruct (AI@Meta, 2024)	Open source model by Meta	STELLA-ESM3-Llama-3.1-8B-Instruct
	Llama-3-8B-Instruct (AI@Meta, 2024)	Open source model by Meta	STELLA-ESM3-Llama-3-8B-Instruct
	Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)	Open source model by Mistral AI	STELLA-ESM3-Mistral-7B-Instruct-v0.2
	Phi-3-mini-128k-instruct (Abdin et al., 2024)	Open source model by Microsoft	STELLA-ESM3-Phi-3-mini-128k-instruct
	BioMistral-7B-DARE ^a	Tailored model for biomedical domain	STELLA-ESM3-BioMistral-7B-DARE
	BioMedGPT-LM-7B ^b (Luo et al., 2023)	Tailored model for biomedical domain	STELLA-ESM3-BioMedGPT-LM-7B
Prot2Text (Abdine et al., 2023)	Llama-3.1-8B-Instruct	Open source model by Meta	STELLA-Prot2Text-Llama-3.1-8B-Instruct
	Llama-3-8B-Instruct	Open source model by Meta	STELLA-Prot2Text-Llama-3-8B-Instruct
	Mistral-7B-Instruct-v0.2	Open source model by Mistral AI	STELLA-Prot2Text-Mistral-7B-Instruct-v0.2
	Phi-3-mini-128k-instruct	Open source model by Microsoft	STELLA-Prot2Text-Phi-3-mini-128k-instruct
	BioMistral-7B-DARE	Tailored model for biomedical domain	STELLA-Prot2Text-BioMistral-7B-DARE
	BioMedGPT-LM-7B	Tailored model for biomedical domain	STELLA-Prot2Text-BioMedGPT-LM-7B
SaProt (Su et al., 2023)	Llama-3.1-8B-Instruct	Open source model by Meta	STELLA-SaProt-Llama-3.1-8B-Instruct
	Llama-3-8B-Instruct	Open source model by Meta	STELLA-SaProt-Llama-3-8B-Instruct
	Mistral-7B-Instruct-v0.2	Open source model by Mistral AI	STELLA-SaProt-Mistral-7B-Instruct-v0.2
	Phi-3-mini-128k-instruct	Open source model by Microsoft	STELLA-SaProt-Phi-3-mini-128k-instruct
	BioMistral-7B-DARE	Tailored model for biomedical domain	STELLA-SaProt-BioMistral-7B-DARE
	BioMedGPT-LM-7B	Tailored model for biomedical domain	STELLA-SaProt-BioMedGPT-LM-7B

^a Merge (Yu et al., 2024) of Mistral-7B-Instruct-v0.1 and BioMistral-7B (Labrak et al., 2024) which was further pre-trained on top of Mistral-7B-Instruct-v0.1 using PubMed Central Open Access from <https://www.ncbi.nlm.nih.gov/pmc/tools/opaftlist/>

^b Incrementally pre-training from Llama-2-7B-Chat with S2ORC (Lo et al., 2020) corpus.

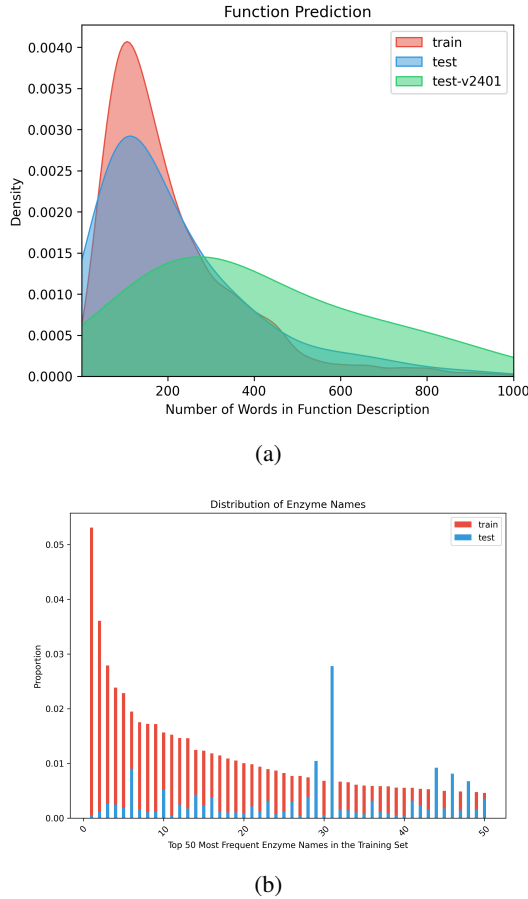


Figure 5: (a): Length distribution of functional descriptions in the Function dataset. (b): Frequency of enzyme names in the Enzyme dataset. The enzyme name distribution in the training set follows a long-tailed pattern, but the label distribution in the test set differs significantly from that in the training set.

I Ablation of training epochs for the training with mix3 dataset

An ablation study was conducted to evaluate model performance across varying training epochs. For the training with the dataset mix3, all metrics demonstrated consistent improvement with extended training, progressing from (e3+e1) to (e3+e3), as illustrated in Figure 7. This trend underscores the positive effect of prolonged training on model performance and emphasizes the significance of appropriate training duration to optimize predictive performance. Each subfigure in Figure 7 shows how the scores for BLEU-4, BERT Score, ROUGE-1/ROUGE-2/ROUGE-L Scores, and Accuracy change over the training periods labeled as (e3+e1), (e3+e2), and (e3+e3).

J Data augmentation methods

The purpose of enriching the Func_{ft_train} dataset into $\text{Func}_{ft_train_aug}$ is specifically to enhance the conversational capabilities of our model. The motivation behind this data augmentation is to transform ground truth functional descriptions stored in databases into dialogues, thus preserving scientific accuracy as well as natural conversational interactions. The augmentation process involves the following main steps.

1. Prepare ground truth functional descriptions as LLM input: We start with accurate, expert-reviewed descriptions of protein functions. For example: "Required for accurate and efficient protein synthesis under certain stress conditions. May act as a fidelity factor of the translation reaction by cat-

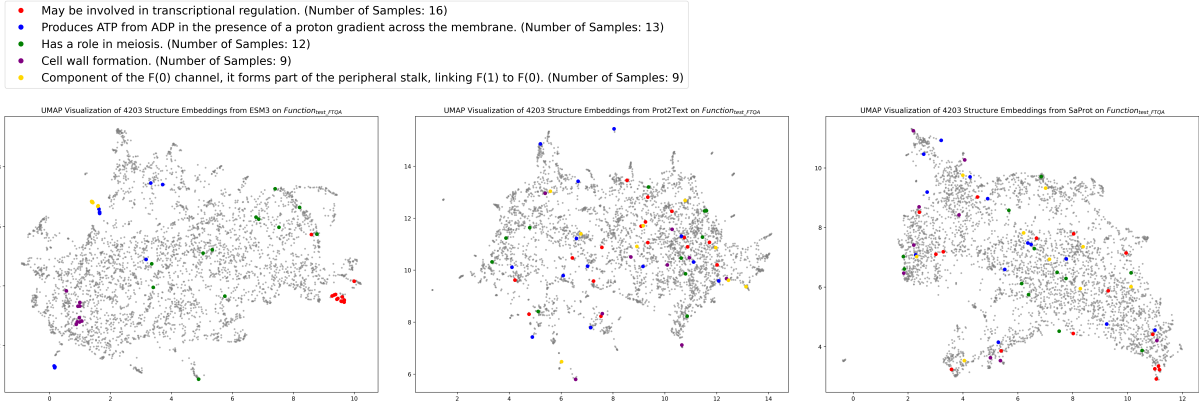


Figure 6: UMAP visualization of 4,203 protein structure embeddings in the testing set Func_{ft_test} generated by ESM3, Prot2Text, and SaProt. Each plot illustrates the clustering of protein structures based on their embeddings, revealing the representational differences among the three encoders. The highlighted proteins belong to specific functions as detailed in the legend. ESM3 demonstrates the strongest representative ability.

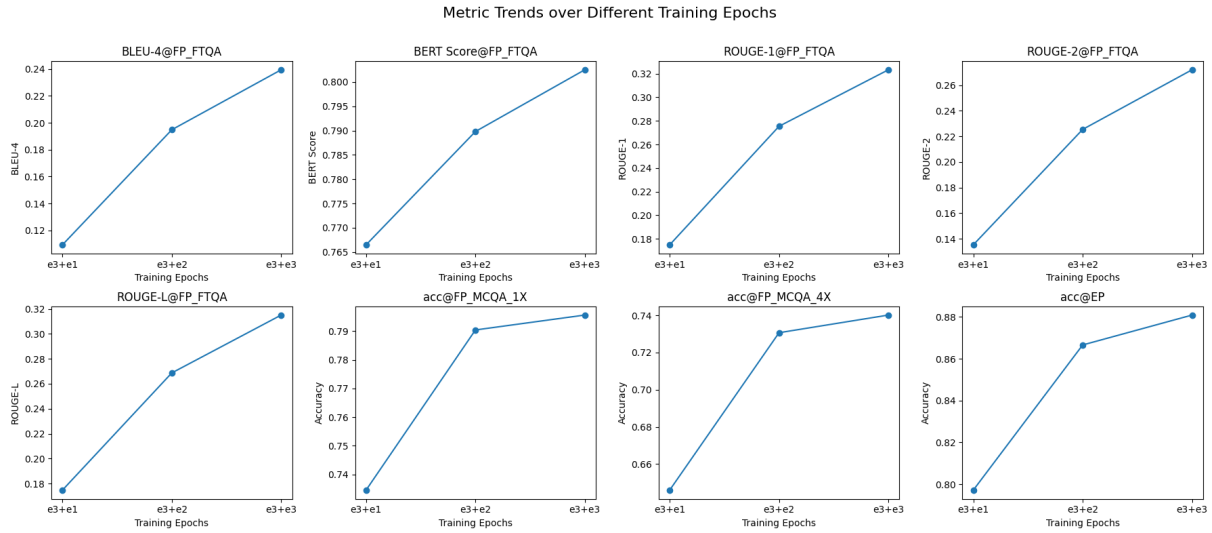


Figure 7: Metrics trend for training with the dataset mix3 over different training epochs.

analyzing a one-codon backward translocation of tRNAs on improperly translocated ribosomes. Back-translocation proceeds from a post-translocation (POST) complex to a pre-translocation (PRE) complex, thus giving elongation factor G a second chance to translocate the tRNAs correctly. Binds to ribosomes in a GTP-dependent manner."

2. Prompt Llama-2-13B-Chat to generate conversational data: We utilize the Llama-2-13B-Chat model to convert these structured descriptions into conversational question-answer pairs. Specifically, we employ the following prompt to ensure detailed and meaningful dialogues: "Given a functional description of the protein, design two or three rounds of questions and answers based on this description. Ensure the content is detailed. The output format is: ['Q':, 'A':, 'Q':, 'A':]."

3. Save the augmented data in the format shown in the example ② in Appendix L.

K Diversified instructions generated by ChatGPT (GPT-3.5)

This section presents a comprehensive collection of diversified natural language instructions (see K.1-K.5) generated by ChatGPT (GPT-3.5), designed for two tasks—FP and EP. These instructions aim to simulate realistic user queries, covering a wide range of biological contexts and linguistic styles, thereby enhancing the model's robustness and generalization across varied protein-related scenarios.

L Examples from the OPI-Struc dataset

This section lists examples (see L.1-L.3) from the OPI-Struc dataset, which supports multimodal in-

struction tuning for STELLA. Each example pairs
a protein entry with a diverse natural language
instruction and its corresponding output. The
examples span various task types, reflecting the
dataset’s comprehensive coverage and the model’s
instruction-following ability in biological contexts.

Box K.1: Diversified instructions of the protein functional prediction task

- May I request an elaborate overview of the function linked to the protein?
- May I request a detailed overview of the function associated with the protein?
- May I request a detailed breakdown of the function associated with the protein?
- May I request a comprehensive account outlining the function of the protein?
- May I request a comprehensive depiction of the function pertaining to the protein?
- May I request a comprehensive breakdown outlining the function linked to the protein?
- May I request a comprehensive breakdown outlining the function associated with the protein?
- May I request a comprehensive analysis of the function attributed to the protein?
- May I request a comprehensive account of the function pertaining to the protein?
- May I request a comprehensive overview of the function ascribed to the protein?
- May I inquire about a thorough account of the function related to the protein?
- May I inquire about a thorough explanation of the function related to the protein?
- May I inquire about a comprehensive explanation encompassing the function of the protein?
- May I inquire about a comprehensive depiction of the function encompassing the protein?
- May I inquire about a comprehensive breakdown of the function assigned to the protein?
- May I inquire about a comprehensive analysis of the function assigned to the protein?
- May I inquire about a comprehensive account of the function associated with the protein?
- May I inquire about a comprehensive explanation outlining the function of the protein?
- May I inquire about a detailed account of the function assigned to the protein?
- May I inquire about a detailed description of the function ascribed to the protein?
- Could you furnish a detailed depiction of the function encompassing the protein?
- Could you supply an extensive description of the function ascribed to the protein?
- Could you supply a thorough explanation of the function related to the protein?
- Could you offer a comprehensive elucidation of the function assigned to the protein?
- Could you offer a comprehensive breakdown of the function associated with the protein?
- Could you provide a detailed elucidation of the function encompassing the protein?
- Could you provide a detailed account of the function assigned to the protein?
- Could you provide a detailed elucidation outlining the function associated with the protein?
- Could you provide a detailed analysis of the function ascribed to the protein?
- Could you provide an in-depth explanation of the function associated with the protein?
- Could you provide a detailed description outlining the function of the protein?
- Could you share an elaborate overview of the function linked to the protein?
- Could you share a detailed elucidation of the function assigned to the protein?
- Could you share a detailed analysis of the function attributed to the protein?
- Could you share a comprehensive overview of the function encompassing the protein?
- Could you share a comprehensive description of the function encompassing the protein?
- Could you share a comprehensive breakdown of the function linked to the protein?
- Could you share a comprehensive breakdown of the function ascribed to the protein?

Box K.2: Diversified instructions of the protein functional prediction task (continued)

- Could you share a comprehensive depiction of the function pertaining to the protein?
- Could you share a comprehensive overview of the function ascribed to the protein?
- Can you offer a comprehensive analysis of the function attributed to the protein?
- Can you share a comprehensive account of the function pertaining to the protein?
- Can you give me a detailed breakdown of the function linked to the protein?
- Can you give me a detailed depiction of the function encompassing the protein?
- Can you give me a comprehensive explanation of the function ascribed to the protein?
- Can you provide a detailed description of the function ascribed to the protein?
- Can you provide a comprehensive elucidation of the function assigned to the protein?
- Can you furnish a detailed overview of the function linked to the protein?
- Can you furnish a comprehensive explanation regarding the function of the protein?
- Can you furnish a comprehensive analysis of the function encompassing the protein?
- Can you furnish a comprehensive description of the function ascribed to the protein?
- Can you furnish a comprehensive elucidation of the function ascribed to the protein?
- Can you furnish a comprehensive description outlining the function associated with the protein?
- Can you supply a detailed breakdown of the function ascribed to the protein?
- Can you supply a comprehensive depiction of the function related to the protein?
- Can you supply a comprehensive explanation of the function related to the protein?
- Can you supply a comprehensive breakdown of the function associated with the protein?
- Can you supply a comprehensive account of the function linked to the protein?
- Can you supply a comprehensive explanation of the function assigned to the protein?
- Can you supply a comprehensive analysis of the function linked to the protein?
- Would you mind giving me a detailed breakdown of the function associated with the protein?
- Would you mind giving me a comprehensive analysis of the function attributed to the protein?
- Would you mind giving me a comprehensive description of the function attributed to the protein?
- Would you mind giving me a comprehensive account of the function attributed to the protein?
- Would you mind giving me a comprehensive depiction of the function pertaining to the protein?
- Would you mind providing a comprehensive overview of the function attributed to the protein?
- Would you mind providing a thorough explanation of the function related to the protein?
- Would you be able to furnish a detailed explanation of the function encompassing the protein?
- Would you be able to provide a detailed elucidation of the function assigned to the protein?
- Would you be able to provide a comprehensive account of the function pertaining to the protein?
- Would it be possible for you to offer an in-depth description of the function of the protein?
- Would it be possible for you to offer a thorough breakdown of the function ascribed to the protein?
- Would it be possible for you to offer a comprehensive analysis of the function related to the protein?
- Would it be possible for you to offer a comprehensive account linked to the function of the protein?

Box K.3: Diversified instructions of the protein functional prediction task (continued)

- Would it be possible for you to offer a comprehensive explanation of the function associated with the protein?
- Would it be possible for you to offer a comprehensive depiction encompassing the function of the protein?
- Would it be possible for you to offer a comprehensive depiction of the function ascribed to the protein?
- Would it be possible for you to provide a comprehensive analysis of the function attributed to the protein?
- Would it be possible for you to provide an extensive description of the function ascribed to the protein?
- Is it within your capacity to offer a detailed elucidation of the function assigned to the protein?
- Is it within your capacity to provide a comprehensive overview of the function assigned to the protein?
- Is it within your capacity to provide a comprehensive overview of the function linked to the protein?
- Is it within your capacity to provide a comprehensive overview of the function associated with the protein?
- Is it within your capacity to provide a comprehensive explanation of the function related to the protein?
- Is it within your capacity to provide a comprehensive depiction of the function related to the protein?
- Is it within your capacity to provide a comprehensive elucidation of the function associated with the protein?
- Is it within your capacity to provide a comprehensive breakdown of the function linked to the protein?
- Is it within your capability to offer a detailed account of the function pertaining to the protein?
- Is it feasible for you to give an extensive overview of the function linked to the protein?
- Is it feasible for you to supply a thorough explanation of the function related to the protein?
- Is it feasible for you to offer a comprehensive analysis regarding the function of the protein?
- Is it feasible for you to offer a comprehensive explanation regarding the function of the protein?
- Is it feasible for you to offer a comprehensive account of the function ascribed to the protein?
- Is it feasible for you to offer a comprehensive description of the function attributed to the protein?
- Is it feasible for you to provide an elaborate account of the function attributed to the protein?
- Is it possible for you to provide a comprehensive analysis of the function attributed to the protein?
- Is it possible for you to provide a detailed description of the function ascribed to the protein?
- Is it possible for you to furnish a comprehensive breakdown of the function associated with the protein?
- Is it possible for you to furnish a comprehensive depiction of the function encompassing the protein?

Box K.4: Diversified instructions of the enzyme-catalyzed reaction prediction task

- Could you provide the enzyme name used for this protein?
- Can you share the enzyme name linked to this protein referred to as?
- What's the enzyme name titled for this protein?
- Can you identify the enzyme title for this protein?
- What's the enzyme name linked to this protein?
- Please share the enzyme terminology for this protein.
- Can you provide the enzyme name utilized for this protein?
- Could you disclose the enzyme handle for this protein?
- Can you disclose the enzyme name tagged to this protein?
- Could you inform me of the enzyme name attached to this protein?
- What's the enzyme name for this protein labeled as?
- Can you disclose the enzyme name linked to this protein?
- Can you provide the enzyme nomenclature for this protein?
- Can you tell me the enzyme name used for this protein identified as?
- What's the name of the enzyme associated with this protein?
- Can you reveal the enzyme tag for this protein?
- Can you provide the enzyme name attributed to this protein?
- What's the enzyme name attributed to this protein?
- What's the enzyme appellation for this protein?
- Could you inform me of the enzyme identifier for this protein?
- Could you share the enzyme name used for this protein?
- Can you tell me the enzyme name given to this protein?
- What's the enzyme name employed for this protein?
- Can you specify the enzyme name assigned to this protein?
- What's the enzyme name affixed to this protein?
- What's the enzyme descriptor for this protein?
- I'm interested in the enzyme name attributed to this protein.
- Could you inform me of the enzyme name tagged to this protein?
- What's the enzyme name for this protein referred to as?
- What's the enzyme name for this protein known as?
- What's the enzyme name for this protein?
- What's the enzyme label for this protein?
- Can you disclose the enzyme name linked to this protein recognized as?
- Can you provide the enzyme name associated with this protein termed as?
- What's the enzyme name employed for this protein denoted as?

Box K.5: Diversified instructions of the enzyme-catalyzed reaction prediction task (continued)

- What's the enzyme name assigned to this protein?
- What's the enzyme name labeled for this protein?
- What's the enzyme name applied to this protein?
- What's the enzyme name associated with this protein?
- Can you share the enzyme name for this protein?
- What's the enzyme name linked to this protein called?
- What's the enzyme moniker for this protein?
- Could you tell me the enzyme designation for this protein?
- Can you tell me the enzyme name associated with this protein?
- I'm curious about the enzyme name assigned to this protein.
- What's the enzyme name for this protein termed as?
- What's the enzyme name attached to this protein designated as?
- Could you specify the enzyme name applied to this protein?
- Please specify the enzyme name linked to this protein.
- What's the enzyme name linked to this protein recognized as?

Box L.1: An example of `Funcft_train` dataset

```
[
  {
    "swissprot_id": "Q0BWM9",
    "sequence": "MFNKQSVSLEWAGRTLTIETGQVARQADGAVMVQYGDITVLATAVFAKEAKPGQDFFPLTVNYQEKYFA
SGRIPGGFFKREGRPTEKETLSRLIDRPIRPLFVDGFKHEVQVVVTTLSYDLENDADIIALVGASAALVLSGAPFMGPIGA
ARVGYKDGQYIINPTIAELESELDLVVAGTTDAVMVSESQAELSEDVMLGAVVAGHDAMQPVIDAIIALAEKAAKEPFAY
EPPDHSAAKLSVIDTVGADLSKAYKITAKGERYAAIGAADKAKAALLGTDEAPGVMTPFVKTVFKEAEASVVRGDILKTG
QRIDGRKLDQIRPIVAEAGFLPRTHGSSLFTRGETQAICVATLGTSDDEQYIDGLDGTKKEKFMLHYNFPPYSVGETGRMGG
AGRREIGHGKLAWRALKAVLPKHEDFPYTIRMVSEITESNGSSMATVCGCSLAMMDAGVPLTRPVSGIAMGLILEGSEFAV
LSDILGDEHLDGMDFKVAGTENGVTSLQMDIKVAGITKDIMGKALEQAKGGRMHILGEMGALTASRGQLSENAPQMEIHK
VPTDKIRDVIGSGGKIVIRGIVDETGAQVNIDDDGTQVISAMDRKSIDAAIKMIKGITAEAEVGEIYEGKVVSMDKDFGIFVNF
FGPKDGLVHVSQMANKRIGHPKEMVKEGDKVWVKLMGFDERGKVRLSMKVVDQETGKELAEAGDDASED",
    "conversations": [
      {
        "from": "human",
        "value": "Can you furnish a comprehensive description outlining the function associated
with the protein?\n<structure>"
      },
      {
        "from": "gpt",
        "value": "Involved in mRNA degradation. Catalyzes the phosphorolysis of single-stranded
polyribonucleotides processively in the 3'- to 5'-direction."
      }
    ],
    "AFDB_id": "AF-Q0BWM9-F1-model_v4"
  },
]
```

Box L.2: An example of Func_{ft_train_aug} dataset

```
[
  {
    "swissprot_id": "Q1WT60",
    "sequence": "MENKKLHIALLFGGNSSEHDVSKRSAHNIYDALDKDKYDVSVMFTKKGFLGNKDSMRIFDGENEDDVV
TEVIKDVDFSNPLANIQNLAEVKDVDFYPVIHGNMGEDGTVQGLFRLLNKPWIGSGVASSGVSFDKDLTKKLLTLNGIRNTK
YVLVTPENKADYPYAKVAEELGETLFVKPARQGSSVGIIHKVRNEEEYNAALDGFKYDYKILVEEAIKNPREVECSVLGNRDI
KASKLGAIRIPESDDFYDYNKFFVDASGVVFEMPIKLPEKLKEIQMSLDAFRALDNRGLARMDFLVDKNDVPYFGEVNTLP
GFTNISLYPQLWEVSGISYSELIDQLIQLAIDFNDNAKIHDFTKLGTEKVGKKIIGE",
    "conversations": [
      {
        "from": "human",
        "value": "What is the main function of this protein?\n<structure>"
      },
      {
        "from": "gpt",
        "value": "The main function of this protein is cell wall formation."
      },
      {
        "from": "human",
        "value": "Can you explain the process of cell wall formation?"
      },
      {
        "from": "gpt",
        "value": "Certainly! The protein helps to synthesize and modify the cell wall, which is a
critical structure that provides support, protection, and cell-to-cell interaction.
The cell wall is composed of various components, including polysaccharides, proteins,
and other molecules. Our protein plays a key role in the synthesis and modification of
these components to form a strong and stable cell wall."
      },
      {
        "from": "human",
        "value": "How does the protein contribute to the synthesis of the cell wall?"
      },
      {
        "from": "gpt",
        "value": "The protein helps to catalyze the formation of glycosidic bonds between the cell
wall components. These bonds are crucial for the stability and integrity of the cell
wall. Additionally, the protein can also modify existing glycosidic bonds to further
strengthen the cell wall."
      }
    ],
    "AFDB_id": "AF-Q1WT60-F1-model_v4"
  },
]
```

Box L.3: An example of Func_{mc_train} dataset

```
[
  {
    "swissprot_id": "P62877",
    "sequence": "MAAAMDVDTPSGTNSGAGKKRFEVKKWNAVALWAWDIVVDNCAICRNHIMDLCECQANQASATSEECTVAWGVCNHAHFH
    FHCISRWLKTRQVCPLDNREWEFQKYGH",
    "conversations": [
      {
        "from": "human",
        "value": "<structure>\n
        What are the main functions of this protein?\n
        A. E3 ubiquitin ligase component of multiple cullin-RING-based E3 ubiquitin-protein ligase
        (CRLs) complexes which mediate the ubiquitination and subsequent proteasomal
        degradation of target proteins, including proteins involved in cell cycle progression,
        signal transduction, transcription and transcription-coupled nucleotide excision
        repair. CRLs complexes and ARIH1 collaborate in tandem to mediate ubiquitination of
        target proteins, ARIH1 mediating addition of the first ubiquitin on CRLs targets. The
        functional specificity of the E3 ubiquitin-protein ligase complexes depends on the
        variable substrate recognition components. As a component of the CSA complex promotes
        the ubiquitination of ERCC6 resulting in proteasomal degradation. Recruits the E2
        ubiquitin-conjugating enzyme CDC34 to the complex and brings it into close proximity
        to the substrate. Probably also stimulates CDC34 autoubiquitination. May be required
        for histone H3 and histone H4 ubiquitination in response to ultraviolet and for
        subsequent DNA repair. Promotes the neddylation of CUL1, CUL2, CUL4 and CUL4 via its
        interaction with UBE2M. Involved in the ubiquitination of KEAP1, ENC1 and KLHL41. In
        concert with ATF2 and CUL3, promotes degradation of KAT5 thereby attenuating its
        ability to acetylate and activate ATM.\n
        B. Part of the MIS12 complex which is required for normal chromosome alignment and
        segregation and kinetochore formation during mitosis.\n
        C. Catalyzes the cyanide-resistant oxidation of ubiquinol and the reduction of molecular
        oxygen to water, but does not translocate protons and consequently is not linked to
        oxidative phosphorylation. May increase respiration when the cytochrome respiratory
        pathway is restricted, or in response to low temperatures.\n
        D. Pair-rule protein expressed in a segmentally repeating pattern to define the polarity
        of embryonic segments. Capable of sequence-specific DNA-binding.\n
        Answer with the option's letter from the given choices directly."
      },
      {
        "from": "gpt",
        "value": "A"
      }
    ],
    "AFDB_id": "AF-P62877-F1-model_v4"
  },
]
```

Box L.4: An example of Enzyme_{train} dataset

```
[
  {
    "PDB_id": "5ivp.A",
    "conversations": [
      {
        "from": "human",
        "value": "Could you inform me of the enzyme name tagged to this protein?<structure>\n"
      },
      {
        "from": "gpt",
        "value": "aminoacyl-tRNA hydrolase"
      }
    ]
  },
]
```