STELLA: LEVERAGING STRUCTURAL REPRESENTATIONS TO ENHANCE PROTEIN UNDERSTANDING WITH MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein biology centers on the intricate relationships among sequence, structure, and function (text), with structure understanding being a crucial aspect for uncovering protein biological functions. Traditional methods based on protein language models (pLMs) often focus on specific aspects of biological function prediction but do not account for the broader, dynamic context of protein research—an important component for addressing the complexity of protein biology. Modern large language models (LLMs) excel in human-machine interaction, language understanding and generation, at a human-like level. By bridging structural representations with the contextual knowledge encoded within LLMs, STELLA leverages the strengths of LLMs to enable versatile and accurate predictions in protein-related tasks. It showcases the transformative potential of multimodal LLMs as a novel paradigm besides pLMs in advancing protein biology research by achieving state-of-the-art performance in both functional description and enzyme-catalyzed reaction prediction tasks. This study not only establishes an innovative LLM-based paradigm to understand proteins, but also expands the boundaries of LLM capabilities in protein biology. To foster collaboration and inspire further innovation, the codes, datasets, and pre-trained models are made publicly available at the anonymous GitHub repository `https://anonymous.4open.science/r/STELLA-DF00`.

1 INTRODUCTION

Protein biology revolves around the interplay of three data modalities: sequence, structure, and function (text). The principle "sequence determines structure, and structure determines function" underscores the critical link between a protein's amino acid sequence, its tertiary structure, and its biological role, such as its main functions and enzyme-catalyzed reactions. Structural data offer significant insights into how a protein's three-dimensional conformation, including features such as active sites and binding pockets, enables and regulates its core biological functions. Accurate understanding of these biological functions plays a pivotal role in advancing disease research, drug discovery, metabolic pathway analysis, and the design of enzymes for medical and biotechnological applications.

Although extensive structural data have been accumulated through decades of protein science research, including experimentally determined structures in the RCSB Protein Data Bank (PDB) [1] (Berman et al., 2000) and computationally predicted structures in the AlphaFold Protein Structure Database (AFDB) [2] (Varadi et al., 2021) by AlphaFold 2 (AF2) (Jumper et al., 2021), further efforts are needed to leverage these resources for deeper understanding of protein biological functions. The PDB, as one of the most comprehensive repositories of experimentally determined protein structures, has long served as a cornerstone of structural biology and biology computational models, such as AlphaFold 3 (Abramson et al., 2024) and ESM3 (Hayes et al., 2024). Similarly, the AFDB has dramatically increased access to high-quality predicted protein structures. These vast structural datasets provide a valuable foundation for advancing protein science, offering new opportunities to deepen our understanding of proteins. However, fully realizing their potential requires bridging

[1] `https://www.rcsb.org/`
[2] `https://alphafold.ebi.ac.uk/`

the gap between structural data and the functional and biochemical insights essential for practical applications in both research and industry.

Understanding structures is crucial for uncovering biological functions, such as protein functional descriptions, elucidating enzyme-catalyzed reactions, and addressing fundamental biological questions, as emphasized in this study. Previous efforts in protein function analysis have included methods such as clustering methods based on protein structure similarity (Barrio-Hernandez et al., 2023; Huang et al., 2023) and text generation methods (Abdine et al., 2023). While these approaches have contributed valuable insights, they often fall short of fully capturing the intricate and multidimensional relationships between protein structure and function, limiting their ability in addressing the complexity of protein biology. Furthermore, these methods typically lack iterative feedback mechanisms from domain experts, which are essential for aligning results with their diverse research objectives. Predicting enzyme-catalyzed reactions is another complex task in protein science, attracting significant research attention (Derevyanko et al., 2018; Steinegger et al., 2019; Hermosilla et al., 2021; Zhang et al., 2022; Hermosilla and Ropinski, 2022; Fan et al., 2022). Although progress has been made, existing methods often approach enzyme prediction as a multi-label classification task, which still struggles with accurately predicting enzyme classes that have not been thoroughly explored in high-throughput proteomics studies. These limitations hinder the ability to fully understand the functions of such enzymes, underscoring the need for more advanced approaches to achieve reliable predictions and accelerate research in this area.

To address this challenge, innovative approaches that integrate structural data with cutting-edge computational tools are urgently needed. Recent advancements, including Prot2Text (Abdine et al., 2023), ProteinGPT (Xiao et al., 2024), and ProtChatGPT (Wang et al., 2024a), have explored the utilization of multimodal LLMs in protein biology. These models typically integrate protein sequence and structure data using a late fusion strategy, where each modality is encoded separately before being aligned or combined. However, late fusion approaches have certain limitations, such as the potential loss of cross-modal relationships and increased complexity of encoder modules. In contrast, the early fusion strategy—where different modalities are jointly represented and fused into a unified representation at encoding stage—has the potential to both preserve the intrinsic relationships between modalities and improve computational efficiency. Motivated by the aforementioned perspectives, this work investigates the advantages of early fusion for multimodal LLMs modeling in protein biology.

To leverage the potential of multimodal LLMs with an early fusion strategy in protein biology, this study introduces STELLA, a multimodal LLM designed to bridge protein language and natural language, enabling the learning of complex structure-function relationships from multimodal data. Unlike previous approaches that use late fusion strategies, STELLA utilizes ESM3 encoder (esm3_sm_open_v1) (Hayes et al., 2024), which inherently implements an early fusion mechanism, where protein sequence and structure are jointly represented in a unified encoding process. By leveraging these fused structural representations—integrating both sequence and structural information—STELLA enhances protein understanding through the power of LLMs, enabling it to interpret protein tertiary structures and predict functional descriptions and enzyme-catalyzed reactions from diverse and versatile user prompts. Apart from the advancement of protein language models (pLMs), STELLA highlights the transformative potential of multimodal LLMs in advancing protein biology research by achieving state-of-the-art performance in both tasks. In doing so, it offers a new paradigm for understanding proteins and extends the capabilities of general-purpose LLMs in the field of protein biology. The key contributions of this study include:

1. By inheriting the early fusion mechanism of ESM3, STELLA achieves state-of-the-art performance in protein functional description and enzyme-catalyzed reaction prediction tasks.

2. This study constructs a large-scale multimodal instruction tuning dataset, OPI-Struc, to support training of multimodal LLMs for protein-related tasks.

3. This study presents the methodology, architecture, and performance of STELLA, alongside the open source code, data, and pre-trained models to encourage collaboration and further advancements in the field.

We anticipate that this study will help drive the advancement of protein science and computational biology through LLM-based approaches, establishing a new paradigm beyond the pLM-based paradigms.

2 RELATED WORK

2.1 PROTEIN-TEXT MODELING

The long-term goal of protein representation learning is to extract biologically relevant information from diverse data modalities, including amino acid sequences and tertiary structures (i.e., protein language) as well as relevant texts in natural language that encapsulate protein related knowledge. Aligning the protein language and natural language has emerged as a crucial aspect of advancing protein representation learning, and attracted much attention in the research community. For instances, ProtST (Xu et al., 2023) utilizes contrastive learning to align amino acid sequences with biomedical texts, aiming to obtain biologically informative protein embeddings that can be applied to various downstream tasks. Besides protein representation learning, ProteinDT (Liu et al., 2023a) leverages textual data to enhance protein design in text-to-sequence generation tasks. Additionally, Prot2Text (Abdine et al., 2023) proposes a method of aligning protein structures and function description texts by using a fused multimodal encoder-decoder framework. In Prot2Text, the encoder is composed of a Relational Graph Convolutional Neural Network (RGCN) for encoding protein structures and a ESM2-35M (Lin et al., 2022) for encoding amino acid sequences and the decoder is a pretrained GPT-2 model to generate protein function annotations. Before the prevalence of LLMs, protein representation learning mainly focuses on single modality like amino acid sequences, or sequence-text alignment by contrastive learning. Hardly any research engages in how to effectively bridge biological language (e.g., protein tertiary structures) to the massive knowledge embedded in natural language that plays a pivotal role in both scientific communication and discovery. As we all know, the process of scientific discovery is a procedure propelled by communication among domain experts and iterative experimentation. Therefore, the excellent conversation and reasoning abilities of LLMs are highly expected to empower the process of scientific discovery.

2.2 LLMS FOR PROTEIN BIOLOGY

Recent studies have highlighted the potential of LLMs in advancing biomedical research, spanning molecules, proteins, and RNA. In the specific 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 3D protein 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 utilized to fine-tune LLMs for aligning protein sequences with natural language, enabling bidirectional tasks such as predicting functions from sequences and generating protein sequences from natural language prompts. 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). ProteinChat (Huo et al., 2024) represents a more recent multi-modal LLM designed to predict protein functions. It integrates a protein sequence encoder, xTrimoPGLM (Chen et al., 2024), with the Vicuna-13B model (Zheng et al., 2023) through a linear layer adapter. Trained on over 1.5 million protein-related (protein, prompt, answer) triplets from the Swiss-Prot dataset, ProteinChat covers a wide range of protein functions. By taking an amino acid sequence as input, it generates comprehensive narratives detailing the functional properties of the given protein.

3 A FIRST LOOK AT STELLA'S CAPABILITIES THROUGH CASE STUDIES

STELLA demonstrates outstanding performance in protein understanding by integrating structural representations into LLMs. As illustrated in Figure 1 (left), STELLA excels in following natural language instructions and providing responses that align with the research goals of human specialists. During the interaction, STELLA correctly identified the primary function of the newly reviewed protein G1TFE0 in the Swiss-Prot database, accurately recognizing it as a component of the large ribosomal subunit. As the dialogue progressed, STELLA elaborated on the core constituents of the ribonucleoprotein complex, highlighting its extensive domain knowledge. Additionally, STELLA showcased its reasoning capabilities by linking the loss of ribosomal function to cellular dysfunctions, demonstrating its ability to establish connections between complex biological processes. In the right

panel of Figure 1, STELLA accurately predicted the function of another newly characterized protein, A0A1D0BR98. Upon further inquiry from the user, it explained details about the mechanisms of the toxin and provided practical treatments. Both case studies illustrate STELLA's ability to predict protein functions from structural data and to deliver informative, contextually relevant responses tailored to diverse research objectives. Moreover, STELLA's robust reasoning abilities enable it to assist domain experts in uncovering connections between protein functions and disease mechanisms, further highlighting its potential to drive scientific discovery and innovation. More example demonstrations are shown in Figure 4 of Appendix A.1.

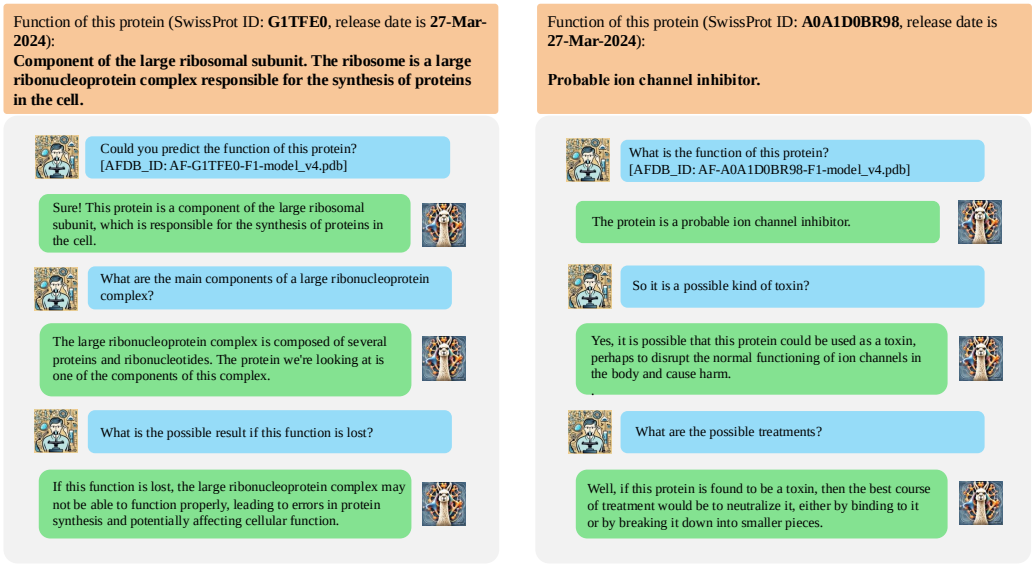


Figure 1: **STELLA's ability to engage in conversation (Model used: STELLA-ESM3-Llama-3.1-8B-Instruct).** The protein G1TFE0 and A0A1D0BR98 are from newly release **2024_02** of the Swiss-Prot database. **Left**: protein G1TFE0. **Right**: protein A0A1D0BR98. **Orange box**: ground truth of the function. **Blue Box**: inquiry from the user. **Green box**: output of the model. Images indicating the user and assistant were generated by AI tools.

4 METHODOLOGY

4.1 STELLA MODEL ARCHITECTURE

Overview. STELLA is a multimodal LLM for protein modeling, drawing inspiration from LLaVA (Liu et al., 2023b), a prominent multimodal architecture designed for vision-language tasks 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 paradigm employed by LLaVA and other multimodal LLMs such as Bunny (He et al., 2024), STELLA adopts a two-stage multimodal instruction tuning (MMIT) approach, 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 A.2, and hyperparameters in Table 6 (Appendix A.3).

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 (Hayes et al., 2024), 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 representation 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 acts as a bridge between the structural representations derived from the protein structure encoder and the natural language embeddings, such as function descriptions. In this implementation, a simple linear layer is employed as the adapter, which has proven effective, as demonstrated in previous works like LLaVA (Liu et al., 2023b).

LLM. The LLM integrated into 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., 2023c), 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 across sensitive applications.

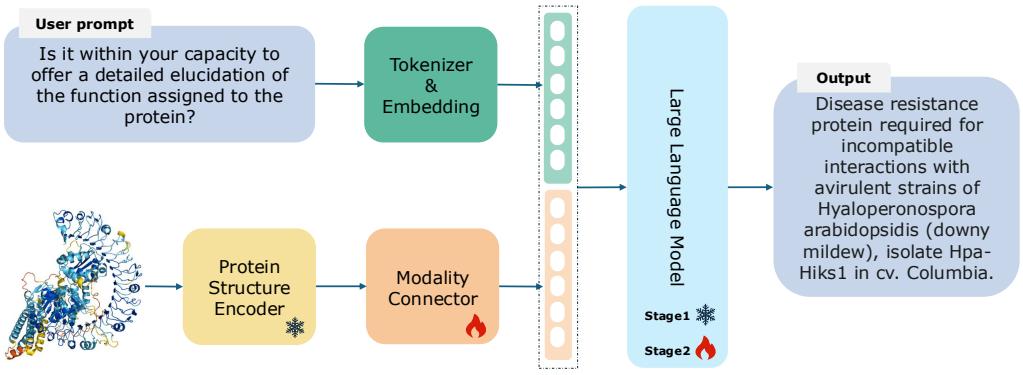


Figure 2: **The architecture of STELLA**. **Stage1 of MMIT:** to fine-tune the modality connector using the OPI-Struc dataset by freezing the protein structure encoder and LLM. **Stage2 of MMIT:** to continually fine-tune the modality connector and the LLM simultaneously with different learning rates, by freezing the protein structure encoder. **Flame**: model is trainable; **Snowflake**: model is frozen. Protein image credits: AFDB.

4.2 Task definition

Functional description prediction (FP). Through multimodal instruction tuning, STELLA effectively aligns protein structural representations with natural language, enabling the accurate prediction of protein functions from tertiary structures. By leveraging multimodal instruction data, STELLA can uncover novel functional associations, substantially reducing the labor-intensive process of manual annotation. This approach offers a powerful and flexible tool for protein functional description prediction. Furthermore, the integration of LLM-based multi-turn dialogues supports iterative interactions with researchers, facilitating continuous refinement of predictions. This adaptive learning process, driven by expert feedback, not only enhances the model's performance but also allows for tailored adjustments to meet specific research objectives.

Enzyme-catalyzed reaction prediction (EP). Predicting enzyme-catalyzed reactions aim at forecasting the biochemical outcomes facilitated by enzymes. Enzymes, as protein-based biological catalysts, are essential for accelerating chemical reactions by lowering activation energy barriers. Accurate prediction of enzyme-catalyzed reactions holds substantial value across various domains, including drug discovery, metabolic engineering, and synthetic biology. In this study, enzyme-catalyzed reactions were mapped to their corresponding enzyme names, which serve as proxies for the reactions in which the associated proteins are involved. This approach allows for more seamless integration with LLMs, ensuring the EP task effectively captures the biological functions of enzymes in a way that aligns with the capabilities of LLMs.

4.3 OPI-STRUC DATASET

Overview. The **Open Protein Instructions for Structures (OPI-Struc)** dataset was specifically curated to support multimodal instruction tuning (MMIT) in this study, by integrating both protein structural and textual modalities. Corresponding to the **FP** and **EP** task, OPI-Struc is organized into two main categories: **Function** and **Enzyme** (see Appendix A.7, example ④). The **Function** dataset is further divided into two subcategories: **Func**$_{ft}$ (see Appendix A.7, example ①) and **Func**$_{mc}$ (see Appendix A.7, example 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 function annotations generated through conversations using Llama-2-13B-Chat, forming the **Func**$_{ft\_train\_aug}$ dataset (see Appendix A.7, example ②). The splitted training and testing sets and corresponding statistics are presented in Table 1.

Table 1: **Statistics of OPI-Struc**. The FP task is composed of two subtasks: FP$_{ft}$ and FP$_{mc}$. In the FP$_{ft}$ task, besides the hold-out testing set Func$_{ft_test}$, a newer release of Swiss-Prot v2024_01 (v2401) was utilized to construct Func$_{ft_test_v2401}$ that aims to assess STELLA's performance on unseen data. In the FP$_{mc}$ task, we designed two versions of testing sets: Func$_{mc_test_1x}$ (options w/o permutation) and Func$_{mc_test_4x}$ (options w/ permutation). See Appendix A.7 for data examples ①, ②, ③ and ④.

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

Data explanation. Each sample of the OPI-Struc dataset consists of a protein tertiary structure (sourced from either AFDB or PDB), task-specific natural language instructions formatted as conversations, and corresponding labels. In the **Function** dataset, protein structures are derived from AFDB, while the labels (i.e., protein function descriptions) are from the release 2022_04 [3] of Swiss-Prot [4]. In addition, when curating **Func**$_{mc\_train}$, the four answer options (A, B, C, D) were randomly permuted within the training set to introduce variability and mitigate answer bias. For the testing set **Func**$_{mc\_test}$, two versions were generated: one without permuted answer options ($1x$) and another with permutation ($4x$), ensuring a more robust evaluation by accounting for both consistent and variable answer configurations. The **Enzyme** dataset was obtained from the SIFTS database (Dana et al., 2018), and the original labels, defined by Enzyme Commission (EC) numbers, were mapped to enzyme names using the BRENDA Enzyme Database [5] (e.g., *1.1.1.10* $\rightarrow$ *L-xylulose reductase*). To ensure consistency and accuracy, the OPI-Struc dataset underwent a rigorous preprocessing pipeline following established data cleaning protocols. In addition, detailed analysis of various dataset characteristics were conducted to highlight its comprehensiveness and potential implications for model performance. For instance, the distribution of protein sequence lengths, which correlates with the complexity of protein structures, was examined (see Figure 5, Appendix A.4). These variations underscore the dataset's coverage of a wide range of structural complexities, which is crucial for training models that can generalize effectively across both simple and complex protein structures. Furthermore, the label distribution was analysed, including the length distribution of function descriptions and the frequency of enzyme names, as shown in Figure 6 (Appendix A.4). These insights emphasize the importance of ensuring model robustness across diverse structural and functional complexities to achieve reliable and consistent performance during evaluation.

Instruction preparation. The raw data were transformed into an instruction-based format to support learning tasks by providing diverse and structured task instructions. To achieve variation in instruction phrasing, ChatGPT (GPT-3.5) was employed via a web interface to generate rephrased instructions.

[3] https://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/

For instance, 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 Appendix A.5 for a detailed list). Each generated instruction was carefully reviewed for accuracy and relevance, ensuring that only high-quality variations were included in the final **Function** dataset. During the augmentation process for the Function-aug$_{train_FTQA}$ dataset, the Llama-2-13B-Chat model (Touvron et al., 2023) was utilized to generate dialogic interactions based on protein function descriptions sourced from Swiss-Prot. The prompt used for this augmentation was: *"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':]."* By integrating diverse instructions, this approach facilitated a more dynamic and engaging bridge between protein structural and textual modalities, thereby enriching the OPI-Struc dataset and improving its adaptability and effectiveness for addressing a wide range of research objectives.

Data split. (1) The **Function** dataset was divided according to the data split method used in (Abdine et al., 2023), maintaining less than 40% sequence similarity between the protein sequences in the training and testing sets to ensure a rigorous evaluation. (2) The **Enzyme** dataset was partitioned following the same split method as in (Hermosilla et al., 2021).

5 Evaluation of STELLA model

This study is critical for advancing our understanding of how multimodal LLMs can effectively leverage protein structural representations to address protein-related tasks and extend beyond these applications. By systematically evaluating the STELLA model across the **FP** and **EP** tasks, we seek to elucidate both the strengths and limitations of structural representations in the context of building multimodal LLMs for protein modeling. For these tasks, we 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**$_{eval}$. The hyperparameters for evaluation are presented in Appendix A.3, while the user prompts for evaluation are listed in Table 7 (Appendix A.6).

Experimental results demonstrate that STELLA is a robust and highly adaptable multimodal LLM. By integrating protein structural representations and LLMs, STELLA exhibits enhanced flexibility and scalability across diverse protein-related tasks, consistently delivering accurate and contextually appropriate outputs. In addition to these strengths, STELLA achieves competitive performance in function and enzyme prediction tasks, rivalling existing specialized models. These results underscore STELLA's potential as a powerful tool for advancing protein science, offering new possibilities for the broader field of computational biology.

5.1 Evaluation metrics

Multiple typical metrics for natural language processing (NLP) tasks, including BLEU, BERT-score, and ROUGE, were employed for comprehensive evaluation in the **FP** task. However, given the specialization and complexity of biological function descriptions, the quality of LLM responses cannot be fully captured by solely NLP metrics. Recognizing the limitations of such conventional NLP metrics in protein-related tasks, we intentionally designed the **multiple-choice QA (MCQA) subtask, FP**$_{mc}$, which adopted Accuracy as metrics, to objectively assess STELLA's performance. BLEU, typically applied in machine translation, is used to assess the similarity between two sequences. Particularly, BLEU-4, which measures the overlap of 4-grams between the generated and reference text, was adopted in this study. BERT-score evaluates the token-level similarity between a generated sentence and a reference sentence. ROUGE, a set of metrics traditionally used for automatic text summarization and machine translation, compares generated text to reference texts to calculate the degree of overlap. It includes ROUGE-1, ROUGE-2, and ROUGE-L, which are based on different n-gram methods. ROUGE-L, which focuses on the longest common subsequence, is particularly effective in evaluating summarization and translation quality by considering overall sentence structure. In addition, the **EP** task adopted Accuracy as metrics.

5.2 EVALUATION RESULTS

5.2.1 RESULTS OF FUNCTION DESCRIPTION PREDICTION

In order to assess STELLA's capability to predict protein functional descriptions based on tertiary structures, we designed the $\mathbf{FP}_{ft\_eval}$, using the hold-out testing set $\mathbf{Func}_{ft\_test}$, which was also utilized for evaluation in Prot2Tex (Abdine et al., 2023). As shown in Table 2, STELLA demonstrated state-of-the-art (SOTA) overall performance, surpassing Prot2Text_{BASE} and Prot2Text_{LARGE} (Abdine et al., 2023) in the $\mathbf{FP}_{ft\_eval}$.

Comparison with Foldseek. We adopt Foldseek as baseline comparison, indulging two steps: structure retrieval using Foldseek (Van Kempen et al., 2024) and function mapping from Swiss-Prot. In the first step, for the 4,203 structures in our testing set, we used the Foldseek easy-search [6] command with default parameters to search for similar protein structures within the training set for each test protein. For the e-value parameter, only matches with an e-value below 0.001 are considered and returned. In the second step, the corresponding functional description is determined based on the top-1 retrieved protein from the Swiss-Prot database. The median e-value of the top-1 retrieved proteins is 2.723e-20, indicating a high confidence in the retrieval results by Foldseek.

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 ↑	BERT Score ↑	ROUGE Score ↑		
				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

Furthermore, it is noteworthy that $\mathbf{FP}_{ft\_eval}$ may be impacted by linguistic variability, where model-generated responses with correct meanings differ in expression from the reference. Therefore, we designed $\mathbf{FP}_{mc\_eval\_1x}$ and $\mathbf{FP}_{mc\_eval\_4x}$ to eliminate ambiguity by providing predefined answer choices, which enables more objective and standardized evaluation. This method requires STELLA to not only identify the correct answer but also engage in reasoning and option filtering based on contextual knowledge, thus providing a more comprehensive assessment of its reasoning capabilities. This provides a more robust evaluation for STELLA. Our experiments demonstrated that STELLA exhibits strong reasoning capabilities by achieving accuracies at $\mathbf{Acc@FP}_{mc\_eval\_1x}$ **= 0.8056** and $\mathbf{Acc@FP}_{mc\_eval\_4x}$ **= 0.7618**. Notably, without integrating with LLMs, baseline models like vanilla ESM3 and Prot2Text are unable to produce outputs in a MCQA format.

5.2.2 RESULTS OF ENZYME NAME PREDICTION

$\mathbf{EP}_{eval}$ aims to assess STELLA's ability in enzyme name prediction. On top of the original $\mathbf{Enzyme}_{train}$ set, we excluded 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 **Accuracy = 0.8806** to **Accuracy = 0.8885**, 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 SOTA **Accuracy = 0.8850** in CDConv (Fan et al., 2022).

5.3 ABLATION STUDY

5.3.1 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 3. The visualization reveals that for the five most frequently occurring functions in the testing set, proteins with the same

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

Table 3: **Evaluation results of the EP task**. Accuracy is a metric that means the predict answer totally matches the target. **Single:** Enzyme$_{train}$ dataset, **mix3:** Func$_{ft_train}$ + Func$_{mc_train}$ + Enzyme$_{train}$. **Bold** and underline indicate the best and the runner-up performance.

Model	Training manner	Acc@EP ↑
DeepFRI (Gligorijević et al., 2021)	w/ pretrain	0.6330
UniRep (Alley et al., 2019)	w/o pretrain	0.7290
3DCNN (Derevyanko et al., 2018)	w/o pretrain	0.7880
HH-suite3 (Steinegger et al., 2019)	w/o pretrain	0.8260
ESM-1b (Rives et al., 2021)	w/ pretrain	0.8310
GearNet-Edge-IEConv (Zhang et al., 2022)	w/o pretrain	0.8530
IEConv (Hermosilla et al., 2021)	w/o pretrain	0.8720
GearNet-Multiview-Contrast (Zhang et al., 2022)	w/ pretrain	0.8750
New IEConv (Hermosilla and Ropinski, 2022)	w/ pretrain	0.8810
CDConv (Fan et al., 2022)	w/o pretrain	0.8850
STELLA-ESM3-Llama-3.1-8B-Instruct(single,two-stage,e3+e3)	MMIT	0.8806
STELLA-ESM3-Llama-3.1-8B-Instruct(single,two-stage,e3+e6)	MMIT	**0.8885**

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 Table 8 (Appendix A.8). Furthermore, several leading LLMs, outlined in Table 9 (Appendix A.9), were integrated into the STELLA framework, enabling an analysis of their impact on STELLA's performance. The ablation results in Table 4 indicate that the combination of the ESM3 encoder with the Llama-3.1 model yielded the best performance in protein function prediction tasks. Moreover, the results underscore the strong overall performance of Llama models across various encoders, reaffirming the effectiveness of combining protein structural information with LLM-based reasoning capabilities.

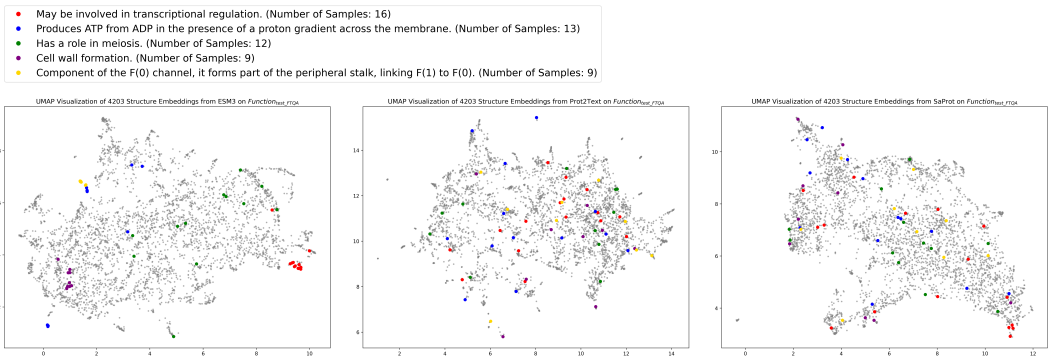


Figure 3: **UMAP visualization of 4,203 protein structure embeddings in the testing set Func**$_{ft\_test}$ **generated by ESM3, Prot2Text, and SaProt.** Each plot illustrates the clustering of protein structures based on their embeddings, revealing the representational differences among the three encoders. The highlighted proteins belong to specific functions as detailed in the legend. ESM3 demonstrates the strongest representative ability.

5.3.2 Ablation of training data mix and training epochs

An ablation study was conducted to evaluate model performance across varying training data mixes and training epochs. The results, presented in Table 5, indicate that increasing training epochs consistently enhances performance across all data mix configurations. Notably, the model trained exclusively on the **Func**$_{ft\_train}$ dataset achieved the highest evaluation scores when trained for three epochs (e3+e3), suggesting that a longer training duration significantly improves its capability to generate accurate and contextually relevant responses. Incorporating the **Func**$_{mc\_train}$ dataset endowed STELLA with multi-choice Q&A capabilities, while causing only a slight decline in its predictive performance on **FP**$_{ft\_eval}$, as both datasets belong to the same overarching task domain. However, the inclusion of the **Enzyme**$_{train}$ dataset in the mix3 configuration led to superior enzyme prediction performance but caused a noticeable decline in function prediction capability, highlighting the challenges inherent in designing high-quality multitask datasets. Furthermore, during the mix3 training, all metrics demonstrated consistent improvement with extended training, progressing from

Table 4: **Ablation of protein encoders and LLMs in the FP$_{ft_eval}$.** Training recipes: single Func$_{ft_train}$ dataset, epochs of two stages (e3+e3). **Bold** and underline indicate the best and the runner-up performance.

Evaluation	Model	BLEU-4 ↑	BERT Score ↑	ROUGE Score ↑		
				ROUGE-1	ROUGE-2	ROUGE-L
FP$_{ft_eval}$	STELLA-ESM3-Llama-3.1-8B-Instruct	**0.4024**	0.8496	0.5218	**0.4487**	**0.5041**
	STELLA-ESM3-Llama-3-8B-Instruct	0.4020	0.8503	0.5138	0.4478	0.5001
	STELLA-ESM3-Phi-3-mini-128k-instruct	0.3807	0.8435	0.4991	0.4273	0.4839
	STELLA-Prot2Text-Llama-3.1-8B-Instruct	0.4009	0.8497	**0.5284**	0.4454	0.5031
	STELLA-Prot2Text-Llama-3-8B-Instruct	0.3892	0.8456	0.5177	0.4329	0.4915
	STELLA-Prot2Text-Phi-3-mini-128k-instruct	0.3771	0.8426	0.5058	0.4210	0.4799
	STELLA-Prot2Text-Mistral-7B-Instruct-v0.2	0.3889	0.8525	0.5224	0.4359	0.4949
	STELLA-Prot2Text-BioMedGPT-LM-7B	0.3999	0.8488	0.5282	0.4447	0.5020
	STELLA-Prot2Text-BioMistral-7B-DARE	0.3870	**0.8533**	0.5241	0.4357	0.4980
	STELLA-SaProt-Llama-3-8B-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
FP$_{ft_eval_v2401}$	STELLA-ESM3-Llama-3.1-8B-Instruct	0.0489	0.7565	0.2210	**0.1085**	0.1867
	STELLA-Prot2Text-Llama-3.1-8B-Instruct	0.0425	0.7555	0.2454	0.1020	0.1919
	STELLA-Prot2Text-Llama-3-8B-Instruct	**0.0510**	0.7605	0.2486	0.1062	0.1918
	STELLA-Prot2Text-Mistral-7B-Instruct-v0.2	0.0440	**0.7685**	**0.2529**	0.1046	**0.1975**

(e3+e1) to (e3+e3), as illustrated in Figure 7 (AppendixA.10). This trend underscores the positive effect of prolonged training on model performance and emphasizes the significance of meticulous dataset selection and appropriate training duration to optimize predictive performance. Additionally, the **FP**$_{ft\_eval\_v2401}$ was designed to assess STELLA's generalization capability on newly released proteins, using the testing set **Func**$_{ft\_test\_v2401}$.

Table 5: **Ablation of training data mix and training epochs across FP$_{ft_eval}$, FP$_{mc_eval_1x}$, FP$_{mc_eval_4x}$ and EP$_{eval}$ for STELLA-ESM3-Llama-3.1-8B-Instruct. single:** Func$_{ft_train}$, **mix2:** Func$_{ft_train}$ + Func$_{mc_train}$, **mix3:** Func$_{ft_train}$ + Func$_{mc_train}$ + Enzyme$_{train}$. The 2nd column indicates the training epochs of two stages. **Bold** indicates the best performance in each configuration.

Data mix	Training epochs	BLEU-4 ↑	BERT Score ↑	ROUGE Score ↑			Acc@FP$_{mc_eval}$ ↑		Acc@EP$_{eval}$ ↑
				ROUGE-1	ROUGE-2	ROUGE-L	1x	4x	
single	(e3+e1)	0.2653	0.8065	0.3938	0.3097	0.3770	-	-	-
	(e3+e2)	0.3574	0.8363	0.4790	0.4028	0.4617	-	-	-
	(e3+e3)	**0.4024**	**0.8496**	**0.5218**	**0.4487**	**0.5041**	-	-	-
mix2	(e3+e1)	0.2397	0.8003	0.3624	0.2861	0.3505	0.6936	0.5893	-
	(e3+e2)	0.3411	0.8330	0.4554	0.3878	0.4428	0.7940	0.7428	-
	(e3+e3)	0.4020	0.8491	0.5119	0.4465	0.4980	**0.8056**	**0.7618**	-
mix3	(e3+e1)	0.1092	0.7665	0.1749	0.1352	0.1747	0.7345	0.6460	0.7972
	(e3+e2)	0.1948	0.7898	0.2754	0.2254	0.2687	0.7904	0.7307	0.8666
	(e3+e3)	0.2394	0.8025	0.3233	0.2720	0.3151	0.7956	0.7402	**0.8809**

6 CONCLUSION AND FUTURE WORK

This study presented STELLA, a novel multimodal LLM in an early fusion strategy inherited from ESM3, which integrates protein structural representations with the contextual knowledge embedded within LLMs. Through multimodal instruction tuning using the OPI-Struc dataset, STELLA achieves SOTA performance in two critical tasks: protein functional description prediction and enzyme-catalyzed reaction prediction. Beyond excelling in protein understanding, STELLA establishes an innovative LLM-based paradigm for protein-related research, complementing traditional pLM-based approaches. This work highlights the transformative potential of STELLA as a powerful computational tool for advancing protein biology. Looking ahead, future efforts should focus on encompassing more downstream tasks by expanding the OPI-Struc dataset, as well as exploring advanced techniques such as retrieval-augmented generation (RAG) and agent-based systems to further enhance STELLA's capabilities. These advancements will unlock its full potential as a pioneering tool in computational biology, solidifying its role in driving the next generation of innovations in protein science and beyond.

REFERENCES

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235. URL https://doi.org/10.1093/nar/28.1.235.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1061.

John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David A. Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

Tomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, pages 2024–07, 2024.

Inigo Barrio-Hernandez, Jingi Yeo, Jürgen Jänes, Milot Mirdita, Cameron LM Gilchrist, Tanita Wein, Mihaly Varadi, Sameer Velankar, Pedro Beltrao, and Martin Steinegger. Clustering predicted structures at the scale of the known protein universe. *Nature*, 622(7983):637–645, 2023. doi: 10.1038/s41586-023-06510-w.

Jiaying Huang, Qiupeng Lin, Hongyuan Fei, Zixin He, Hu Xu, Yunjia Li, Kunli Qu, Peng Han, Qiang Gao, Boshu Li, et al. Discovery of deaminase functions by structure-based protein clustering. *Cell*, 186(15):3182–3195, 2023. doi: 10.1016/j.cell.2023.05.041.

Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. Prot2text: Multimodal protein's function generation with gnns and transformers. *arXiv preprint arXiv:2307.14367*, 2023.

Georgy Derevyanko, Sergei Grudinin, Yoshua Bengio, and Guillaume Lamoureux. Deep convolutional networks for quality assessment of protein folds. *Bioinformatics*, 34(23):4046–4053, 2018.

Martin Steinegger, Markus Meier, Milot Mirdita, Harald Vöhringer, Stephan J Haunsberger, and Johannes Söding. Hh-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics*, 20:1–15, 2019.

Pedro Hermosilla, Marco Schäfer, Matej Lang, Gloria Fackelmann, Pere-Pau Vázquez, Barbora Kozlikova, Michael Krone, Tobias Ritschel, and Timo Ropinski. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. In *International Conference on Learning Representations*, 2021.

Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*, 2022.

Pedro Hermosilla and Timo Ropinski. Contrastive representation learning for 3d protein structures. *arXiv preprint arXiv:2205.15675*, 2022.

Hehe Fan, Zhangyang Wang, Yi Yang, and Mohan Kankanhalli. Continuous-discrete convolution for geometry-sequence modeling in proteins. In *The Eleventh International Conference on Learning Representations*, 2022.

Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. ProteinGPT: Multimodal llm for protein property prediction and structure understanding. *arXiv preprint arXiv:2408.11363*, 2024.

Chao Wang, Hehe Fan, Ruijie Quan, and Yi Yang. ProtChatGPT: Towards understanding proteins with large language models. *arXiv preprint arXiv:2402.09649*, 2024a.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. ProtST: Multi-modality learning of protein sequences and biomedical texts. *arXiv preprint arXiv:2301.12040*, 2023.

Shengchao Liu, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Anthony Gitter, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. A text-guided protein design framework. *arXiv preprint arXiv:2302.04611*, 2023a.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

Xiaohan Lin, Zhenyu Chen, Yanheng Li, Xingyu Lu, Chuanliu Fan, Ziqiang Cao, Shihao Feng, Yi Qin Gao, and Jun Zhang. Protokens: A machine-learned language for compact and informative encoding of protein 3d structures. *bioRxiv*, pages 2023–11, 2023.

Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin, Xiang Zhuang, Xiaotong Li, and Huajun Chen. InstructProtein: Aligning human and protein language via knowledge instruction. *arXiv preprint arXiv:2310.03269*, 2023.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. doi: 10.48550/arXiv.2307.09288.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://www.aclweb.org/anthology/2020.acl-main.447.

Mingjia Huo, Han Guo, Xingyi Cheng, Digvijay Singh, Hamidreza Rahmani, Shen Li, Philipp Gerlof, Trey Ideker, Danielle A. Grotjahn, Elizabeth Villa, Le Song, and Pengtao Xie. Multi-modal large language model enables protein function prediction. *bioRxiv*, 2024. doi: 10.1101/2024.08.19.608729. URL https://www.biorxiv.org/content/early/2024/08/20/2024.08.19.608729.

Bo Chen, Xingyi Cheng, Pan Li, Yangli-ao Geng, Jing Gong, Shen Li, Zhilei Bei, Xu Tan, Boyan Wang, Xin Zeng, et al. xtrimopglm: unified 100b-scale pre-trained transformer for deciphering the language of protein. *arXiv preprint arXiv:2401.06199*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b. doi: 10.48550/arXiv.2304.08485.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective, 2024. URL https://arxiv.org/abs/2402.11530.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021a. URL https://arxiv.org/abs/2009.03300.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024b. URL https://arxiv.org/abs/2406.01574.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

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. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021b. URL https://arxiv.org/abs/2103.03874.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Mark Chen and Jerry Tworek et al. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation, 2023c. URL https://arxiv.org/abs/2305.01210.

Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Berkeley function calling leaderboard. 2024.

Venkat Krishna Srinivasan, Zhen Dong, Banghua Zhu, Brian Yu, Hanzi Mao, Damon Mosk-Aoyama, Kurt Keutzer, Jiantao Jiao, and Jian Zhang. Nexusraven: a commercially-permissive language model for function calling. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. URL https://openreview.net/forum?id=Md6RUrGz67.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞bench: Extending long context evaluation beyond 100k tokens, 2024. URL https://arxiv.org/abs/2402.13718.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners, 2022. URL https://arxiv.org/abs/2210.03057.

Jose M Dana, Aleksandras Gutmanas, Nidhi Tyagi, Guoying Qi, Claire O'Donovan, Maria Martin, and Sameer Velankar. 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, 11 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1114. URL https://doi.org/10.1093/nar/gky1114.

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12 (1):3168, 2021.

Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein language modeling with structure-aware vocabulary. *bioRxiv 2023.10.01.560349*, 2023. doi: 10.1101/2023.10.01.560349.

AI@Meta. Llama 3 model card, 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Marah Abdin, Jyoti Aneja, and et al Hany Awadalla. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch, 2024.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains, 2024.

A Appendix

A.1 Example demonstration of STELLA's capabilities through case studies

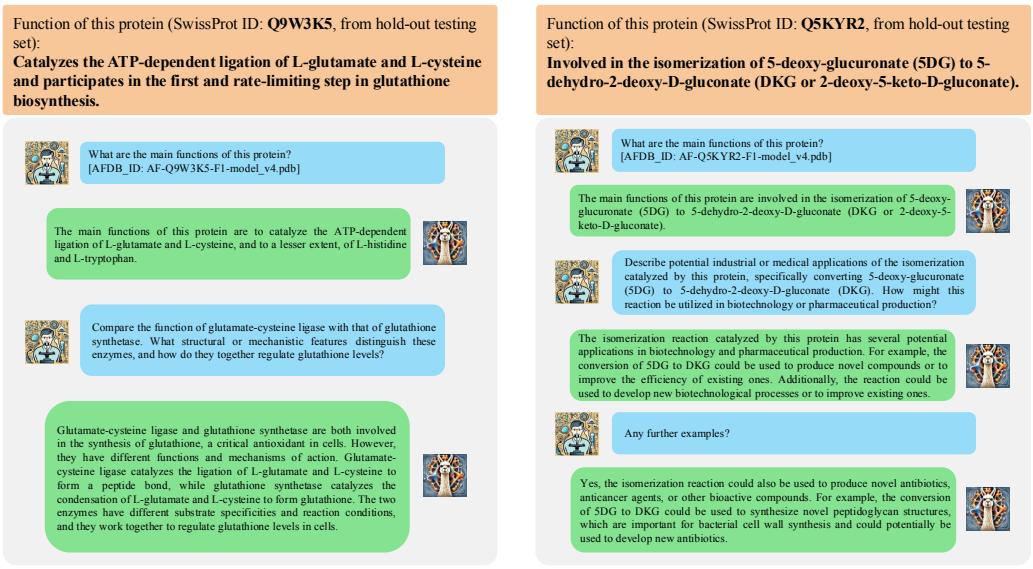


Figure 4: **STELLA's ability to engage in conversation (Model used: STELLA-ESM3-Llama-3.1-8B-Instruct).** The protein Q9W3K5 and Q5KYR2 are from the hold-out testing set of OPI-Struc. **Left**: protein Q9W3K5. **Right**: protein Q5KYR2. **Orange box**: ground truth of the function. **Blue Box**: inquiry from the user. **Green box**: output of the model. Images indicating the user and assistant were generated by AI tools.

A.2 Prompt template for training

The prompt template of STELLA-Prot2Text-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|>

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>

A.3 Hyperparameters for training and evaluation

Stage1 aims to align a protein structure embedding space and a plain-text embedding space. In this stage, the modality connector trainable, while both the protein structure encoder and the LLM are frozen. Stage2 is dedicated to teach 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 trainable with different learning rates, while the protein structure encoder is still frozen.

Both stages use the same training datasets. The prompts templates for training follow the examples shown in Appendix A.2.

Hyperparameters in PT stage and IT stage are summarized in Table 6. It is noteworthy that we adopt different learning rates for each different components of STELLA to finely control the training process. Especially, in the IT stage, we set the learning rate of the modality connector larger than LLM backbone, to improve LLMs' training convergence.

Table 6: **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	(β_1,β_2)=(0.9, 0.999), eps=1e-8	(β_1,β_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

A.4 Analysis of data label distribution of the OPI-Struc dataset

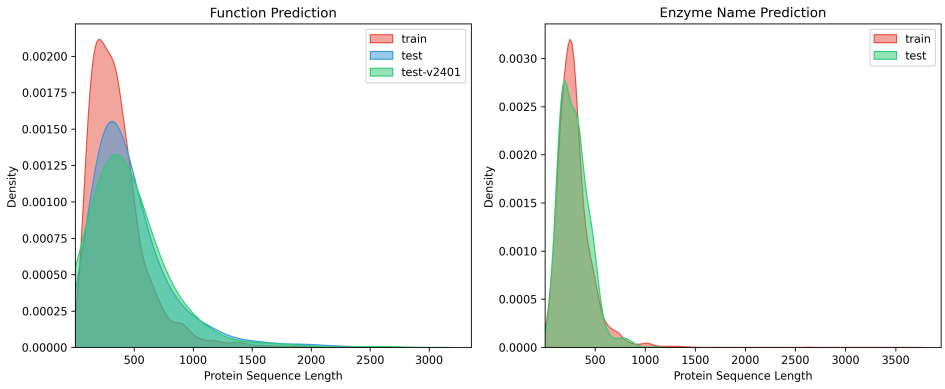
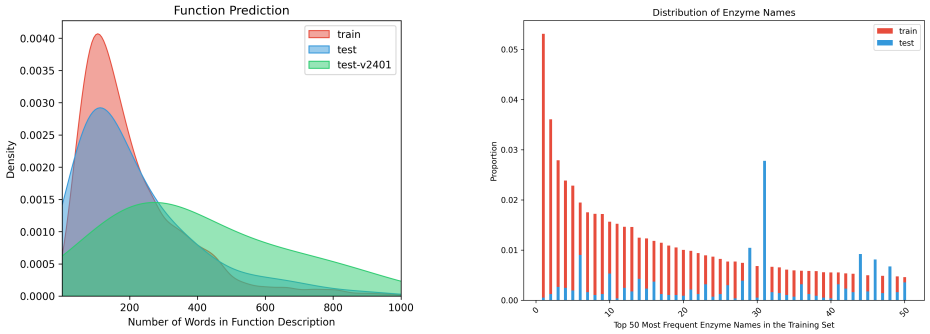


Figure 5: **Distribution of protein sequence lengths across the FP (left) and EP (right) tasks for training and testing sets.** The variation in sequence length distribution between the training and testing sets ensures model robustness across proteins with diverse structural complexities.



(a) (b)

Figure 6: **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.

A.5 Expanded instructions by ChatGPT (GPT-3.5)

Expanded instructions by ChatGPT (GPT-3.5)

- May I request an elaborate 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?
- Can you supply a detailed breakdown of the function ascribed to the protein?
- May I request a comprehensive depiction of the function pertaining to the protein?
- May I request a comprehensive account outlining the function of the protein?
- Is it possible for you to furnish a comprehensive breakdown of the function associated with the protein?
- May I request a comprehensive breakdown outlining the function linked to the protein?'
- Could you share a detailed elucidation of the function assigned to the protein?'
- Would you mind giving me a detailed breakdown of the function associated with the protein?
- Is it within your capacity to provide a comprehensive overview of the function linked to the protein?
- Could you supply an extensive description of the function ascribed to the protein?
- Can you furnish a comprehensive elucidation of the function ascribed to the protein?
- Is it feasible for you to offer a comprehensive analysis regarding the function of the protein?
- Would it be possible for you to offer a thorough breakdown of the function ascribed 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?
- May I inquire about a comprehensive explanation encompassing the function of the protein?
- Can you furnish a comprehensive description of the function ascribed to the protein?
- Would you mind providing a comprehensive overview of the function attributed to the protein?
- Could you share an elaborate overview of the function linked to the protein?
- Could you share a comprehensive overview of the function encompassing the protein?

- Could you offer a comprehensive elucidation of the function assigned to the protein?
- May I request a comprehensive breakdown outlining the function associated with the protein?
- Would you mind giving me a comprehensive analysis of the function attributed to the protein?
- Is it within your capacity to offer a detailed elucidation of the function assigned to the protein?
- Can you supply a comprehensive explanation of the function related to the protein?
- Can you give me a comprehensive explanation of the function ascribed to the protein?
- Is it possible for you to provide a detailed description of the function ascribed to the protein?
- Could you share a comprehensive description of the function encompassing the protein?
- Would you mind providing a thorough explanation of the function related to the protein?
- Can you offer a comprehensive analysis of the function attributed to the protein?
- Can you supply a comprehensive depiction of the function related to the protein?
- May I request a detailed overview of the function associated with the protein?
- May I request 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?
- Is it feasible for you to offer a comprehensive explanation regarding the function of the protein?
- Is it within your capacity to provide a comprehensive explanation of the function related to the protein?
- Would it be possible for you to provide a comprehensive analysis of the function attributed to the protein?
- May I inquire about a thorough account of the function related to the protein?
- May I request a comprehensive 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?
- Could you provide a detailed elucidation of the function encompassing the protein?
- Would it be possible for you to offer a comprehensive depiction encompassing the function of the protein?
- Is it feasible for you to offer a comprehensive account of the function ascribed to the protein?
- Is it within your capacity to provide a comprehensive breakdown of the function linked to the protein?
- Could you share a comprehensive breakdown of the function linked to the protein?
- May I inquire about a comprehensive depiction of the function encompassing the protein?
- Is it within your capacity to provide a comprehensive overview of the function assigned to the protein?
- May I inquire about a comprehensive account of the function associated with the protein?
- Could you provide a detailed account of the function assigned to the protein?
- Could you furnish a detailed depiction of the function encompassing the protein?
- Can you provide a detailed description of the function ascribed to the protein?
- May I inquire about a comprehensive explanation outlining the function of the protein?
- May I request a comprehensive overview of the function ascribed to the protein?
- Could you provide a detailed elucidation outlining the function associated with the protein?

- Can you provide a comprehensive elucidation of the function assigned to the protein?
- Would it be possible for you to offer a comprehensive explanation of the function associated with the protein?
- Would you mind giving me a comprehensive account of the function attributed to the protein?
- May I inquire about a comprehensive breakdown of the function assigned 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?
- Is it possible for you to furnish a comprehensive depiction of the function encompassing the protein?
- Can you supply a comprehensive breakdown of the function associated with the protein?
- Can you furnish a detailed overview of the function linked to the protein?
- May I inquire about a thorough explanation of the function related to the protein?
- Could you share a detailed analysis of the function attributed to the protein?
- Would you be able to furnish a detailed explanation of the function encompassing the protein?
- Is it feasible for you to provide an elaborate account of the function attributed to the protein?
- May I inquire about a comprehensive analysis of the function assigned to the protein?
- Would you be able to provide a detailed elucidation of the function assigned to the protein?
- May I request a detailed breakdown of the function associated with the protein?
- Would it be possible for you to offer a comprehensive depiction of the function ascribed to the protein?
- May I inquire about a detailed account of the function assigned to the protein?
- Could you provide an in-depth explanation of the function associated with the protein?
- May I inquire about a detailed description of the function ascribed to the protein?
- Would you be able to provide a comprehensive account of the function pertaining to the protein?
- Can you furnish a comprehensive description outlining the function associated with the protein?
- Can you supply a comprehensive analysis of the function linked to the protein?
- Would it be possible for you to offer a comprehensive analysis of the function related to the protein?
- Could you offer a comprehensive breakdown of the function associated with the protein?
- Could you supply a thorough explanation of the function related to the protein?
- Is it feasible for you to supply a thorough explanation of the function related to the protein?
- Would it be possible for you to offer an in-depth description of the function of the protein?
- Is it within your capacity to provide a comprehensive depiction of the function related to the protein?
- Could you provide a detailed description outlining the function of the protein?
- Can you share a comprehensive account of the function pertaining to the protein?
- Would it be possible for you to provide an extensive description of the function ascribed to the protein?
- Could you share a comprehensive depiction of the function pertaining to the protein?
- Could you provide a detailed analysis of the function ascribed to the protein?
- Is it within your capacity to provide a comprehensive elucidation of the function associated with the protein?

- Would you mind giving me a comprehensive depiction of the function pertaining to the protein?
- Could you share a comprehensive overview of the function ascribed to the protein?
- Is it within your capability to offer a detailed account of the function pertaining to the protein?
- Can you supply a comprehensive account of the function linked to the protein?
- Could you share a comprehensive breakdown of the function ascribed to the protein?
- Would it be possible for you to offer a comprehensive account linked to the function of the protein?
- Can you supply a comprehensive explanation of the function assigned to the protein?
- Is it possible for you to provide a comprehensive analysis of the function attributed to the protein?
- Is it feasible for you to offer a comprehensive description of the function attributed to the protein?

A.6 PROMPT TEMPLATE FOR EVALUATION

Table 7 presents the user prompts used in the evaluation of three tasks. Notably, we designed the prompt to ensure that the model outputs only one of the four options (A, B, C, or D) in the FP_{MCQA} task, facilitating assessment.

Table 7: **User prompts for evaluation**.

Task	Testing set	Answer formatting prompts
FP_{ft}	Func_{ft_test} $\text{Func}_{ft\_test\_v2401}$	What are the main functions of this protein?
FP_{mc}	$\text{Func}_{mc\_text\_1x}$ $\text{Func}_{mc\_text\_4x}$	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?

A.7 EXAMPLES OF THE OPI-STRUC DATA

① An example of Func_{ft_train} data


```
[
  {
    "swissprot_id": "Q0BWM9",
    "sequence": "
        MFNKQSVSLEWAGRTLTIETGQVARQADGAVMVQYGDTIVLATAVFAKEAKPGQDFFPLTV
       NYQEKYFASGRIPGGFFKREGRPTEKETLTSRLIDRPIRPLFVDGFKHEVQVVVTTLSYDL
       ENDADIIALVGASAALVLSGAPFMGPIGAARVGYKDGQYIINPTIAELEESELDLVVAGTT
       DAVMMVESQAAELSEDVMLGAVVAGHDAMQPVIDAIIALAEKAAKEPFAYEPPDHSAALKS
       VIDTVGADLSKAYKITAKGERYAAIGAAKDKAKAALLGTDEAPGVMTPEVFKTVFKEAEAS
       VVRGDILKTGQRIDGRKLDQIRPIVAEAGFLPRTHGSSLFTRGETQAICVATLGTSDDEQY
       IDGLDGTKKEKFMLHYNFPPYSVGETGRMGGAGRREIGHGKLAWRALKAVLPKHEDFPYTI
       RMVSEITESNGSSSMATVCGCSLAMMDAGVPLTRPVSGIAMGLILEGSEFAVLSDILGDED
       HLGDMDFKVAGTENGVTSLQMDIKVAGITKDIMGKALEQAKGGRMHILGEMGKALTASRGQ
       LSENAPQMEIIKVPTDKIRDVIGSGGKVIRGIVDETGAKVNIDDDGTVQISAMDRKSIDAA
       IKMIKGITAEAEVGEIYEGKVVSMKDFGIFVNFFGPKDGLVHVSQMANKRIGHPKEMVKEG
       DKVWVKLMGFDERGKVRLSMKVVDQETGKELAEEAGDDASED",
    "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"
  },
]
```


② An example of Func$_{ft_train_aug}$ dataset


```
[
  {
    "swissprot_id": "Q1WT60",
    "sequence": "
        MENKKLHIALLFGGNSSEHDVSKRSAHNIYDALDKDKYDVSVFMFTKKGFLLGNKDSMRIF
      DGENEDDVVTEVIKDVDFSNPLANIQNLAEVKDVDVFYPVIHGNMGEDGTVQGLFRLLNKPW
      IGSGVASSGVSFDKDLTKKLLTLNGIRNTKYVLVTPENKADYPYAKVAEELGETLFVKPARQ
      GSSVGIHKVRNEEEYNAALEDGFKYDYKILVEEAIKNPREVECSVLGNRDIKASKLGAIRIP
      ESDDFYDYNNKFVDASGVVFEMPIKLPEKLTKEIQQMSLDAFRALDNRGLARMDFLVDKNDV
      PYFGEVNTLPGFTNISLYPQLWEVSGISYSELIDQLIQLAIDEFNDNAKIHYDFTKLGTEKV
      GKKIIGE",
    "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"
  },
]
```


③ An example of Func$_{mc_train}$ data


```
[
  {
   "swissprot_id": "P62877",
   "sequence": "
       MAAAMDVDTPSGTNSGAGKKRFEVKKWNAVALWAWDIVVDNCAICRNHIMDLCIECQANQ
   ASATSEECTVAWGVCNHAFHFHCISRWLKTRQVCPLDNREWEFQKYGH",
   "conversations": [
    {
      "from": "human",
      "value": "<structure>\nWhat are the main functions of this
          protein?\nA. 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.\nB. Part of the
          MIS12 complex which is required for normal chromosome
          alignment and segregation and kinetochore formation during
           mitosis.\nC. 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.\nD. Pair-
          rule protein expressed in a segmentally repeating pattern
          to define the polarity of embryonic segments. Capable of
```

```
          sequence-specific DNA-binding.\nAnswer with the option's
          letter from the given choices directly."
      },
      {
        "from": "gpt",
        "value": "A"
      }
    ],
    "AFDB_id": "AF-P62877-F1-model_v4"
  },
]
```


④ An example of Enzyme$_{train}$ data


```
[
  {
    "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"
      }
    ]
  },
]
```


A.8 Comparison of protein structure encoders

Difference among three representative protein structure encoders employed in this study, ESM3, Prot2Text and SaProt, are presented in Table 8.

A.9 Different composition of protein structure encoders and LLMs

According to the architecture of STELLA, it is flexible and customizable to integrate various protein encoders and LLMs to form STELLA variants. In order to delve into the effectiveness of different composition of protein encoders and LLMs, we elaborately choose different protein encoders and foundation LLMs, as shown in Table 9.

A.10 Ablation of training epochs for mix3 training

Each graph in Figure 7 shows how the scores for BLEU-4, BERT Score, ROUGE Scores, and Accuracy change over the training periods labeled as (e3+e1), (e3+e2), and (e3+e3). All the metrics improve as training epochs increase, suggesting better performance with more training.

Table 8: **Comparison of three representative protein structure encoders.**

Protein encoder	Modality	Modality fusion methods
ESM3	Sequence, Structure, Function	ESM3 is a multimodal model pretrained on massive sequence, structure and function tokens via masked language modeling (MLM). It encodes these modalities as discrete token tracks, which are fused into a unified representation space using several transformer blocks, with geometric attention in the first block to incorporate atomic information.
Prot2Text	Sequence, Structure, Function	Prot2Text is a multimodal model incorporating a Relational Graph Convolution Network (RGCN), ESM-2 and GPT-2 to generate protein function annotation. It is designed to integrate information from two sources: the output of the RGCN and the protein sequence data processed by ESM-2. The RGCN receives all-atom protein structures as its input, providing detailed structural information. Subsequently, the Prot2Text encoder aligns this integrated data with functional annotation through a generative alignment approach using a text decoder. Prot2Text serve as a method for protein structure-text feature alignment.
SaProt	Sequence, Structure	SaProt is a large-scale pre-trained model using about 40 million protein sequences and structures with structure-aware vocabulary which integrates residue tokens with structure tokens simultaneously. It adopts an ESM-based architecture that takes inputs as structure-aware protein sequences, which combine the protein sequence residue tokens and discrete structural tokens encoded using foldseek. This encoder is not aligned with functional annotation text.

Table 9: **Specifications of STELLA composition of various protein structure encoders and foundation LLMs.**

Protein encoder	Foundation LLM	Note	Composed 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/openftlist/

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

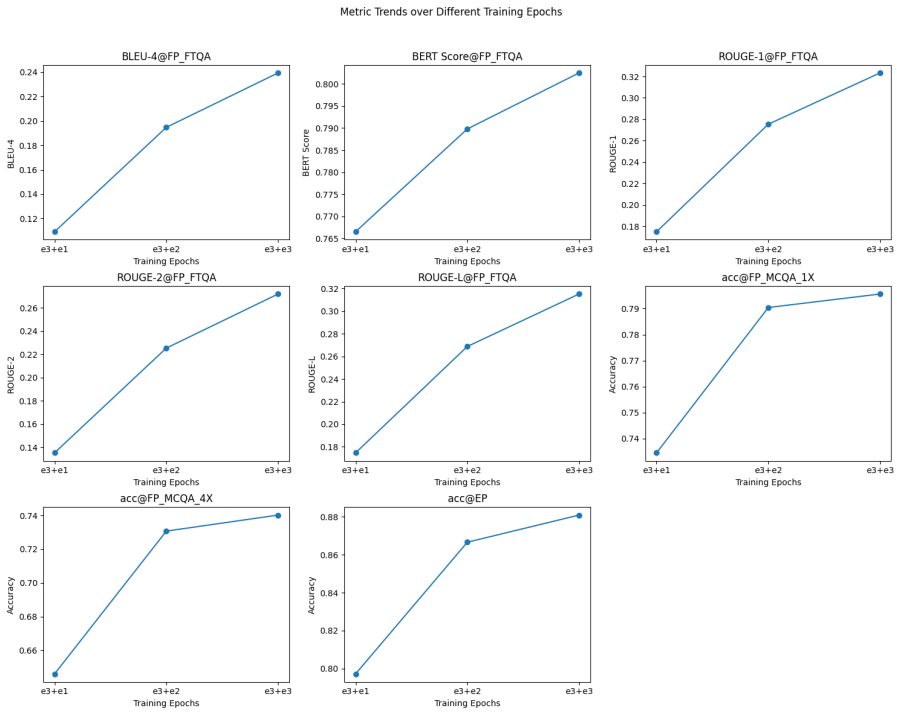


Figure 7: **The trend lines for the various metrics across different training epochs.**