STELLA: A Multimodal LLM for Protein Functional Annotation via Unified Sequence-Structure Encoding

Anonymous ACL submission

Abstract

Understanding the intricate interplay among sequence, structure, and function remains a fundamental challenge in proteomics. The sequence-structure-function paradigm posits that biological roles are governed by the tertiary geometric conformations encoded within primary sequences; consequently, integrating these multi-modal descriptors is imperative for accurate functional annotation. While protein language models (pLMs) have achieved significant progress via representation learning on massive sequence data, they often lack the capacity to incorporate high-resolution structural information and the rich textual context that characterizes protein roles. In this work, we present STELLA, a multimodal LLM that synergistically aligns bimodal (sequence-structure) representations with the textual modality to advance protein functional annotation. By leveraging ESM3 for unified bimodal encoding and Llama-3.1-8B-Instruct for natural language modeling, STELLA achieves state-of-the-art performance in two critical tasks: Functional Description Prediction and Enzyme-catalyzed Reaction Prediction. This study demonstrates that multimodal LLMs represent a paradigm shift beyond pure pLMs, offering a new frontier for protein biology and biomedical discovery. The codes can be accessed via https://anonymous.4open.science/r/STELLA-DF00.

1 Introduction

Protein biology centers on the intricate interplay among three fundamental modalities: sequence, structure, and function. The central tenet of structural biology—that sequences dictate structures, which in turn govern functions—underscores the deterministic relationship between a protein's primary sequence and its biological function. Specifically, the tertiary topology of a protein defines its interaction landscape with ligands, substrates, or inhibitors, thereby mediating essential activities such as enzymatic catalysis and molecular recognition. Deciphering these functional mechanisms is paramount for elucidating disease pathology—where protein dysfunctions are frequently the primary drivers—and for accelerating drug discovery, metabolic engineering, and the design of novel biocatalysts for industrial biotechnology.

Recent decades have witnessed an explosion of structural data, characterized by the growth of experimentally solved structures in the RCSB Protein Data Bank (PDB)[1] (Berman et al., 2000) and the vast repository of high-confidence predictions in the AlphaFold Database (AFDB)[2] (Varadi et al., 2021). While advancements in structural proteomics (e.g., AlphaFold 2 (Jumper et al., 2021), AlphaFold3 (Abramson et al., 2024), Chai-1 (Chai-Discovery-Team et al., 2024), Boltz-1 (Wohlwend et al., 2025) and OpenComplex2 (OC-Team and Ye, 2025) have reached unprecedented levels of accuracy, our understanding of protein functions has not kept pace with this structural revolution. A profound functional annotation gap remains, as the biological roles of the majority of sequenced and folded proteins remain elusive. Consequently, despite the maturation of structure prediction, the challenge has shifted toward harnessing these structural insights to facilitate functional elucidation—a task that requires moving beyond static geometry to understand complex biological processes, subcellular dynamics, and context-dependent activities.

To bridge this gap, protein language models (pLMs) have been developed to learn joint sequence-structure representations (Su et al., 2023; Li et al., 2025). However, while these models excel at capturing biophysical attributes, they often struggle to integrate the textual modality—the descriptive knowledge essential for defining biological function. Emerging multimodal large language models (MLLMs), such as Prot2Text (Abdine et al.,

[1] https://www.rcsb.org/

[2] https://alphafold.ebi.ac.uk/

2023), ProteinGPT (Xiao et al., 2024), ProtChatGPT (Wang et al., 2024), ProteinChat (Huo et al., 2024), have attempted to integrate protein data with natural language. Nevertheless, these frameworks typically rely on disparate pre-trained encoders for sequence and structure, necessitating complex fusion layers that increase computational overhead and complicate gradient-based optimization. This architectural fragmentation motivates our investigation into ESM3 (Hayes et al., 2024), a frontier pLM, as a unified protein encoder. By embedding sequence and structure into a cohesive latent space, ESM3 offers a streamlined yet potent foundation for multimodal integration within an LLM.

In this work, we present STELLA, a multimodal LLM designed to synergize sequence-structure with natural language (function). STELLA integrates the esm3_sm_open_v1 (1.4B) encoder with the Llama-3.1-8B-Instruct model (Dubey et al., 2024), establishing a new paradigm that leverages the unified encoding capacity of pLMs and the superior generation power of generative LLMs. We demonstrate STELLA's efficacy on two critical tasks: **Functional Description Prediction (FP)** and **Enzyme-catalyzed Reaction Prediction (EP)**, representing the **global and biochemical dimensions of protein functionality**, respectively. While FP requires the generation of comprehensive narratives describing biological roles (e.g., signal transduction or DNA repair), EP demands the precise identification of catalytic specificity. STELLA achieves state-of-the-art performance across both tasks, underscoring the potential of multimodal LLMs to serve as a transformative tool for protein biology and biomedical discovery, transcending the limitations of traditional pLMs. Contributions include:

1. We develop STELLA, a multimodal LLM that simplifies protein-to-text integration via a unified bimodal encoder, setting a new state-of-the-art benchmark for FP and EP.

2. We release OPI-Struc, a large-scale multimodal instruction-tuning dataset that bridges the gap between protein structures and natural language annotations, providing a foundational resource for the protein-LLM community.

3. We establish an innovative paradigm in computational protein science, demonstrating that multimodal LLMs can synergize with pLM-based representations to achieve high-fidelity, context-aware functional characterization.

2 Related Work

2.1 Protein Representation Learning

Protein representation learning seeks to develop models capable of extracting biologically meaningful features from diverse modalities, including sequences, structures, and functions. Early foundations were laid by unimodal sequence models such as ProtBERT (Elnaggar et al., 2022), ESM-2 (Lin et al., 2023), and ProtGPT2 (Ferruz et al., 2022), which capture the "grammar" of amino acids through large-scale pre-training. Expanding beyond sequences, research has branched into two primary trajectories. One trajectory focuses on cross-modal alignment between sequences and natural language; for instance, ProtST (Xu et al., 2023) employs contrastive learning to align protein representations with textual descriptors, while ProteinDT (Liu et al., 2023b) leverages text-conditioned diffusion for protein design. Another one emphasizes structural integration, exemplified by SaProt (Su et al., 2023), which pioneered a structure-aware vocabulary by encoding residues and Foldseek-based tokens into a unified ESM architecture. Despite these multi-modal efforts, these models are typically optimized for latent space alignment or specific design tasks, often lacking the generative reasoning capacity required for nuanced functional characterization.

2.2 LLMs for Protein Biology

The emergence of Large Language Models (LLMs) has provided a powerful framework for integrating protein-specific data with the extensive "world knowledge" embedded in natural language models. Recent studies have signaled the transformative potential of LLMs in proteomics. Prot2Text first proposed an encoder-decoder architecture to align protein structures with functional narratives by combining ESM-2 and GPT-2. To facilitate biological reasoning, BioMedGPT (Luo et al., 2023) and InstructProtein (Wang et al., 2023) connected sequence encoders with Llama-based models for protein-text generation and question-answering. To incorporate structural insights, ProteinGPT (Xiao et al., 2024) and ProtChatGPT (Wang et al., 2024) utilized specialized encoders such as GVP-GNN and ESM-IF1 to capture tertiary topologies. More recently, ProteinChat (Huo et al., 2024) integrated the xTrimoPGLM (Chen et al., 2024) encoder with Vicuna-13B (Zheng et al., 2023) to enable dialogue-based functional prediction, though it re-

lies solely on sequence inputs. Prot2Chat (Wang et al., 2025b) further extended this by incorporating both sequence and structure through LoRA-based fine-tuning. A common characteristic of these contemporary approaches is the reliance on disparate pre-trained encoders for different modalities, which often introduces architectural redundancy and complicates cross-modal optimization. This sets the stage for our investigation into a more streamlined, unified encoding paradigm.

3 Methodology of STELLA

3.1 Model Architecture

Overview. The architectural design of STELLA is inspired by the LLaVA framework(Liu et al., 2023a), adapting the established vision-language paradigm to the domain of protein biology. As illustrated in Figure 1, STELLA comprises three core modules: a **protein structure encoder**, a **modality connector**, and a **LLM**. Following the prevailing training methodology in MLLMs (He et al., 2024), STELLA undergoes a two-stage Multimodal Instruction Tuning (MMIT) process. Notably, diverging from the standard LLaVA approach—which typically uses distinct datasets for alignment and tuning—STELLA utilizes the same curated dataset for both stages. This strategy is necessitated by the acute scarcity of high-quality protein-text instruction pairs, ensuring maximal data efficiency for both modality alignment and functional reasoning. Detailed prompt templates and hyperparameter configurations are documented in Appendix A and Table 6 of Appendix E, respectively.

Protein structure encoder. The protein structure encoder is tasked with projecting tertiary topologies into high-dimensional latent representations. In this study, we employ ESM3, a frontier multimodal protein language model pretrained on a vast repertoire of sequence, structure, and functional descriptors. ESM3 conceptualizes these disparate modalities as discrete token tracks, which are integrated into a unified embedding space via a transformer-based architecture. A pivotal feature of ESM3 is the incorporation of geometric attention within its initial transformer blocks, which provides a strong inductive bias for capturing fine-grained atomic-level spatial details—an essential requirement for accurate functional inference.

Modality connector. The modality connector serves as the neural bridge that aligns the protein-centric representations with the textual latent space of the LLM. In our implementation, we utilize a single linear projection layer as the adapter. Despite its structural simplicity, this linear bottleneck has demonstrated remarkable efficacy in established multimodal frameworks (Liu et al., 2023a; He et al., 2024), offering a computationally efficient yet robust mechanism for cross-modal feature mapping without compromising representational fidelity.

LLM. For the generative and reasoning backbone of STELLA, we employ Llama-3.1-8B-Instruct which represents the state-of-the-art in open-source LLMs, exhibiting superior performance across a diverse spectrum of benchmarks, including general knowledge (Hendrycks et al., 2021), mathematical reasoning (Cobbe et al., 2021), code generation (Chen and et al., 2021), tool-use (Yan et al., 2024; Srinivasan et al., 2023), long context tasks (Zhang et al., 2024) and multilingual ability (Shi et al., 2022). Its sophisticated instruction-following capability ensures that the model can interpret nuanced biological prompts, facilitating high-fidelity functional annotation. Furthermore, the integration of Llama Guard 3 ensures the model maintains rigorous safety and reliability standards during text generation.

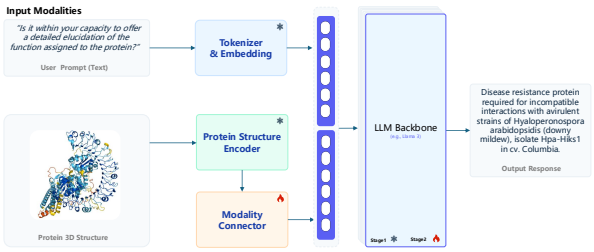


Figure 1: **The STELLA architecture and MMIT workflow.** (a) **Stage 1:** Fine-tuning the modality connector on OPI-Struc (protein encoder and LLM frozen). (b) **Stage 2:** Joint fine-tuning of the connector and LLM with distinct learning rates, while maintaining a frozen protein encoder. Trainable (🔥) and frozen (❄) modules are indicated. Protein structure credits: AFDB.

3.2 Multimodal Instruction Dataset

Data overview. To facilitate MMIT, we curated the Open Protein Instructions for Structures (OPI-Struc) dataset, which synergistically integrates protein structural modalities with descriptive natural language. It differs with prior sequence-only (Mol-Instructions (Fang et al.)) or property-prediction (PEER (Xu et al., 2022)) datasets. Aligned with the FP and EP tasks, OPI-Struc is stratified into two primary domains: **`Function`** and **`Enzyme`** (representative examples are provided in Appendix L).

Within the Function domain, samples are further categorized based on their linguistic format: Func_ft, comprising free-text question-answer pairs for generative evaluation, and Func_mc, employing a multiple-choice framework for discriminative assessment. To emulate the iterative and conversational nature of scientific inquiry, we performed synthetic data augmentation on a 20% subset (49,663 samples) of the Func_ft_train corpus. Utilizing Llama-2-13B-Chat, we generated enriched inquiry-response pairs to form the Func_ft_train_aug dataset, thereby enhancing the model's linguistic diversity and reasoning depth (see Appendix J for methodological details). Comprehensive dataset statistics and split are detailed in Figure 2.

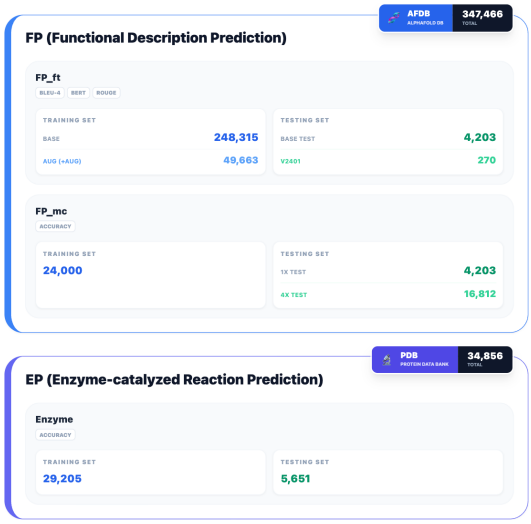


Figure 2: **OPI-Struc dataset statistics.** The FP task includes FP_ft (evaluated on 2022_04 hold-out and 2024_01 OOD sets) and FP_mc (with/without option permutation). Example instructions are provided in Appendix Box L.4–L.3.

Data split. To ensure a fair comparison with existing benchmarks, we implemented established splitting protocols for both domains. (1) For the Function dataset, we adopted the partitioning strategy established by Prot2Text. Specifically, a stringent 40% sequence identity threshold was enforced between the training and testing sets to mitigate potential data leakage and ensure rigorous evaluation of STELLA's generalization capabilities. (2) For the Enzyme dataset, we followed the methodology described in IEConv (Hermosilla et al., 2021), maintaining consistency with prior studies.

Data explanation. Each sample in the OPI-Struc dataset integrates a protein tertiary topology—sourced from the AFDB or PDB—with task-specific conversational instructions and their corresponding ground-truth responses. For the Function domain, structural data are paired with functional description derived from the UniProtKB/Swiss-Prot release 2022_04. To rigorously assess STELLA's generalization to newly characterized proteins, we utilized the 2024_01 release as a temporal hold-out set for zero-shot evaluation. To mitigate option bias during multimodal training, the multiple-choice training set (Func_mc_train) was constructed by randomly permuting the four response options (A, B, C, D) for each inquiry. For evaluation, we provide two testing variants: a fixed-order version (1x) and a permuted-order version (4x). The latter employs four randomized permutations per sample to ensure a rigorous assessment of STELLA to generalize across diverse response configurations. The Enzyme dataset is curated from the SIFTS database (Dana et al., 2018), where original Enzyme Commission (EC) numbers are mapped to enzyme names via the BRENDA Enzyme Database(e.g., *1.1.1.10* → *L-xylulose reductase*).

Data Integrity and Contamination Mitigation. We emphasize a critical distinction between the pre-training objective of ESM3 and our evaluation framework. While ESM3's training involved coarse-grained protein-related keywords, OPI-Struc leverages nuanced, free-text functional narratives. This transition from keyword-level association to high-dimensional natural language descriptions ensures that the OPI-Struc test suites represent distinct evaluative scenarios. Consequently, this design effectively mitigates the risk of data contamination, ensuring that the benchmarks are not explicitly contained within ESM3's pre-existing structural or functional priors.

Preprocessing and Statistical Profiling. To ensure dataset purity and reliability, OPI-Struc underwent a stringent preprocessing pipeline adhering to established data-cleaning protocols. All auxiliary metadata—including PubMed IDs, Evidence Code Ontology (ECO) IDs, and non-functional annotations—were systematically removed. We performed extensive statistical analyses to characterize the dataset's comprehensiveness. Protein sequence lengths were utilized as a proxy for structural complexity, with their distribution (illustrated in Figure 6, Appendix C) demonstrating broad coverage across diverse folding scales. Furthermore, we analyzed the density of functional description lengths and the frequency of enzymatic labels (Figure 7,

Appendix C). These multifaceted distributions underscore the necessity for models that remain robust across varying scales of structural and functional complexity, ensuring consistent performance across the broad landscape of protein biology.

Instruction Synthesis and Diversification. To facilitate robust multimodal instruction tuning, the raw data were transformed into a conversational format. We sought to enhance the lexical and structural diversity of the task prompts by leveraging ChatGPT to synthesize semantically equivalent variations of the core instructions. For instance, the baseline prompt—"Please describe the function of the protein."—was expanded into approximately 100 distinct linguistic variations. A representative list of these expanded prompts is provided in Box K.1 (Appendix K). To ensure scientific accuracy and relevance, each generated variation underwent rigorous manual curation before being integrated into the final `Function` dataset. A parallel diversification process was applied to the `Enzyme` dataset (see Box K.2, Appendix K). The complete diversified instructions are documented in our open-source repository.

3.3 Evaluation Tasks

Functional description prediction (`FP`). This task evaluates STELLA's capacity to decode protein tertiary topologies into detailed functional narratives. By synergistically aligning protein structural descriptors with the natural language space through MMIT, STELLA facilitates the precise generation of comprehensive biological descriptions based on 3D folding patterns. The integration of a generative LLM backbone enables dialogue-driven interactions, providing an intuitive and versatile interface for context-aware protein function elucidation.

Enzyme-catalyzed reaction prediction (`EP`). In this task, complex enzyme-catalyzed reactions are formulated as the prediction of their canonical enzyme names, which serve as semantic proxies for specific biochemical transformations mediated by the enzyme. This mapping leverages LLM's extensive knowledge of biochemical nomenclature and reaction mechanisms, ensuring that enzymatic functions are captured in a format that aligns with STELLA's generative inference strengths. This approach bridges structured biochemical data with natural language, facilitating high-fidelity inference of catalytic roles.

3.4 Evaluation Metrics

To comprehensively evaluate STELLA's performance on the `FP` task, we employ a suite of metrics from natural language processing (NLP) to quantify the linguistic and semantic fidelity of the generated functional descriptions. Specifically, we utilize BLEU-4 (Papineni et al., 2002) to assess n-gram lexical overlap between the generated and reference sequences, and ROUGE scores (1/2/L) (Lin, 2004) to evaluate the recall and structural preservation of biological narratives. Among these, ROUGE-L is particularly informative for functional characterization as it identifies the longest common subsequence, capturing the overarching structure of biological descriptions. To supplement these lexical-based metrics, BERTScore (Zhang et al., 2019) is employed to measure token-level semantic similarity using contextual embeddings, providing a more robust assessment of functional equivalence beyond surface-level word matching.

Despite their prevalence, we recognize that standard metrics may not fully reflect the biological precision required for functional annotation. Given the current absence of universally established metrics tailored specifically for biological text generation, these metrics remain the most rigorous available proxies and have been widely adopted in prior literature. To address their inherent limitations, we further introduce MCQA for the `FP` task. This subtask utilizes Accuracy as an objective metric to evaluate STELLA's ability to discern correct functional roles from decoys. Similarly, for the `EP` task, which involves the precise identification of catalytic roles, Accuracy is employed as the primary metric to ensure the reliability of the predicted enzyme-catalyzed reactions.

4 Performance Evaluation of STELLA

STELLA is benchmarked against five diverse scenarios—`FP_ft_eval`, `FP_ft_eval_v2401`, `FP_mc_eval_1x`, `FP_mc_eval_4x`, and `EP_eval`—to systematically evaluate its functional prediction capabilities. The characteristics of these test sets are summarized in Figure 2. Comprehensive documentation of the evaluation prompts and hyperparameters can be found in Appendices B and E.

4.1 Evaluation of FP Performance

Hold-out Benchmark. To establish a baseline of STELLA's performance, we first evaluated the model on the independent hold-out test set,

Func_ft_test, following the experimental protocol established by Prot2Text. As summarized in Figure 3 (see Table 8, Appendix M for detailed results), STELLA (e3+e6) achieves state-of-the-art results, consistently outperforming prior baselines.

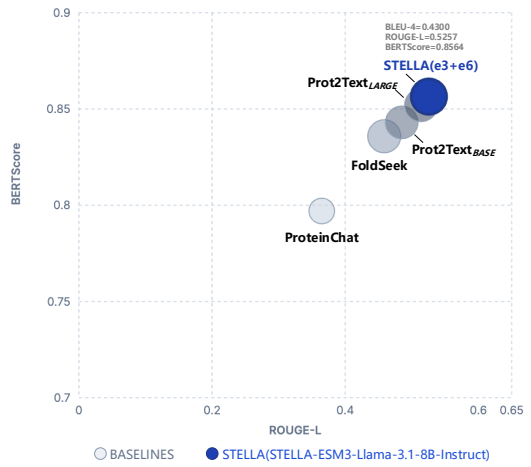


Figure 3: **Hold-out evaluation of FP performance across STELLA, baselines and state-of-the-art methods on the FP_ft_eval benchmark.** STELLA is fine-tuned on the Func_ft_train dataset using a two-stage strategy (3+6 epochs).

Zero-shot Temporal Generalization. We further assessed STELLA's robustness against distributional shifts using FP_ft_test_v2401, a temporal hold-out set derived from a newer Swiss-Prot release unseen during training. As shown in Figure 4, we observed a performance decrement of zero-shot performance relative to the hold-out evaluation (see Table 9, Appendix N for more detailed results). This is likely attributable to newly characterized proteins possessing novel structural or functional motifs that remain underrepresented in the training corpus—a common challenge in protein representation learning due to the continuous expansion of biological knowledge. To address this Out-of-Distribution (OOD) gap, future iterations will explore RAG strategies and the integration of external functional metadata.

Robustness to Structural Degradation. Inherent noise and incomplete coordinates are prevalent in experimental protein structures. To evaluate STELLA's resilience, we simulated structural perturbations by truncating the terminal 10% of residues for each protein in the test set. As detailed in Table 1, STELLA demonstrates remarkable stability under these degraded conditions. Notably, for the e3+e6 configuration, the ROUGE-L score decreased by only 4.1% (from 0.5257 to 0.4915),

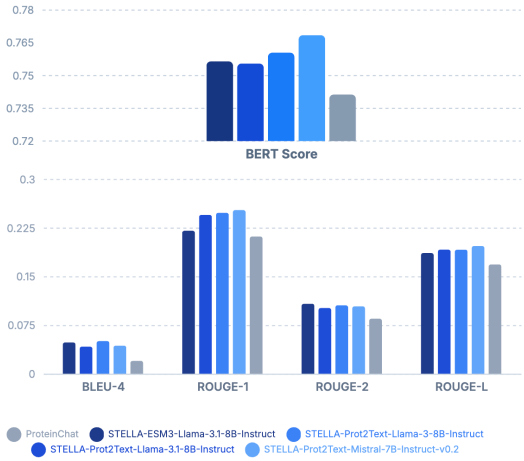


Figure 4: **Zero-shot temporal evaluation of STELLA on the FP_ft_eval_v2401 benchmark.** STELLA models are fine-tuned on the Func_ft_train dataset using a two-stage strategy (3+3 epochs).

while Prot2Text$_{LARGE}$ experienced a more substantial decline of 13.7%. Given that all models were trained on complete structures, STELLA's superior retention of functional information highlights its representational robustness and applicability to partially resolved structural data.

Table 1: **Performance evaluation on structural degradation.** ROUGE-L are reported for each model on complete and incomplete protein. STELLA (STELLA-ESM3-Llama-3.1-8B-Instruct) is fine-tuned on the Func_ft_train dataset using a two-stage strategy.

Model	Complete	Incomplete	Perf. Drop
Prot2Text$_{LARGE}$	0.5140	0.4438	13.7%
STELLA (e3+e3)	0.5041	0.4805	4.7%
STELLA (e3+e6)	0.5257	0.4915	4.1%

Comparison with Alignment-based Retrieval. We compared STELLA against a representative retrieval-based baseline utilizing Foldseek (Van Kempen et al., 2024), which includes two steps: structure retrieval using Foldseek and function mapping from Swiss-Prot. This pipeline involves (1) structural retrieval via Foldseek (easy-search mode, $e\text{-}value < 0.001$) within the training set, followed by (2) functional mapping from the top-1 retrieved Swiss-Prot hit. The median $e\text{-}value$ for the top hits was 2.723×10^{20}, reflecting high retrieval confidence. Nevertheless, as shown in Figure 3, STELLA outperforms the Foldseek-based method (by 14.6%) in ROUGE-L. This underscores STELLA's ability to generalize beyond simple homology-based lookups by learning the underlying sequence-structure-function mapping.

Objective Assessment via MCQA. To mitigate the impact of linguistic variability—where seman-

tically correct responses may diverge from reference texts in phrasing—we introduced the MCQA subtask. This format eliminates generative ambiguity and provides a standardized metric for discriminative reasoning. STELLA achieved accuracies of **80.56** and **76.18** on `FP_mc_eval_1x` and `FP_mc_eval_4x`, respectively. The results demonstrate STELLA's robust instruction-following capability and its capacity to perform precise functional reasoning under constrained choice sets.

Generalization Analysis Across Structural Homology Tiers. To assess structural generalization beyond sequence similarity, we quantified the representational density of structural motifs via Foldseek-based clustering. As detailed in Table 5, Appendix D, STELLA maintains robust performance even for proteins with minimal structural homology in the training set.

4.2 Evaluation of EP Performance

The `EP_eval` suite was employed to determine STELLA's precision in predicting enzyme-catalyzed reactions. During preprocessing, 10 samples were excluded due to missing atomic coordinates, ensuring the integrity of the structural embeddings. As illustrated in Table 2, extending the Stage 2 training duration from 3 to 6 epochs yielded a performance gain, increasing accuracy from **88.06** to **88.85**. Ultimately, STELLA establishes a new state-of-the-art for the `EP` task, surpassing leading models such as CDConv (Fan et al., 2022) and Sable (Li et al., 2025) (previous best: 88.50). This performance highlights STELLA's exceptional capability in bridging high-resolution structural features with precise biochemical nomenclature.

Table 2: **Evaluation of EP performance.** Accuracy measures exact matches with the ground truth. STELLA (STELLA-ESM3-Llama-3.1-8B-Instruct) is fine-tuned on the `Enzyme_train` dataset using a two-stage strategy. **Bold**: best; underline: runner-up.

Model	Accuracy ↑
w/o pretrain	
UniRep (Alley et al., 2019)	72.90
3DCNN (Derevyanko et al., 2018)	78.80
TAPE-LSTM (Rao et al., 2019)	79.90
HH-suite3 (Steinegger et al., 2019)	82.60
GearNet-Edge-IEConv (Zhang et al., 2022)	85.30
IEConv (Hermosilla et al., 2021)	87.20
New IEConv (Hermosilla and Ropinski, 2022a)	87.20
CDConv (Fan et al., 2022)	<u>88.50</u>
w/ pretrain	
DeepFRI (Gligorijević et al., 2021)	63.30
ProtBERT-BFD (Elnaggar et al., 2022)	72.20
ESM-1b (Rives et al., 2021)	83.10
GearNet-Multiview-Contrast (Zhang et al., 2022)	87.50
New IEConv (Hermosilla and Ropinski, 2022b)	88.10
Sable (Li et al., 2025)	<u>88.50</u>
MMIT (Two-stage training on `Enzyme_train` *dataset)*	
STELLA (e3+e3)	88.06
STELLA (e3+e6)	**88.85**

5 Comparative Analysis

5.1 Encoder Efficacy and LLM Synergy

We characterized the discriminative power of protein encoders by visualizing the embeddings of the `Func_ft_test` set via UMAP. Compared to Prot2Text and SaProt, ESM3 produces more discrete and compact functional clusters (Figure 8), reflecting a more biologically informative latent space. A comparative study of various LLMs within the STELLA framework reveals that the integration of Llama-3.1-8B-Instruct yields superior performance in `FP` and `EP` tasks (Table 3).

Additionally, our analysis of reasoning-centric LLMs and standard models indicates that while reasoning capabilities slightly bolster OOD robustness, the overall performance in these challenging regimes remains a frontier (Table 4). These findings are consistent with recent studies on the stability and generalization of reasoning models in specialized domains (Yao et al., 2025; Wang et al., 2025a; Huang et al., 2025), highlighting the ongoing challenge of achieving robust zero-shot reasoning in protein biology.

5.2 Stage-wise Training Strategy

The training procedure for STELLA is partitioned into two distinct phases: cross-modal representation alignment and multimodal instruction tuning. This decoupled strategy is designed to mitigate optimization conflicts and ensure a stabilized transition from protein-centric descriptors to complex natural language reasoning.

Stage 1 focuses on projecting bimodal protein embeddings into the LLM's latent space via a modality connector. By aligning protein features with textual semantics at this stage, we bridge the inherent cross-modal disparity, allowing the LLM to interpret biological features as coherent linguistic tokens. **Stage 2** emphasizes instruction tuning and task-specific refinement, enhancing the model's generative fidelity and zero-shot generalization. This progressive paradigm is instrumental in preventing catastrophic forgetting and representation collapse; specifically, it ensures that the model does not disproportionately overfit to textual priors at the expense of intrinsic protein features. Furthermore, Stage 2 accommodates diverse

Table 3: **Comparison of protein encoder efficacy and LLM synergy on the `FP_ft_eval` benchmark.** All models are fine-tuned on `Func_ft_train` using a two-stage strategy (3+3 epochs). **Bold**: best; underline: runner-up.

Model	BLEU-4 ↑	BERTScore ↑	ROUGE Score ↑		
			ROUGE-1	ROUGE-2	ROUGE-L
ESM3 encoder					
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
Prot2Text encoder					
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
SaProt encoder					
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

Table 4: **Comparison of Reasoning-centric vs. Standard LLMs.** All models utilize the STELLA-ESM3 architecture and are fine-tuned on the `Func_ft_train` dataset using a two-stage strategy (3+6 epochs). **Bold**: best; underline: runner-up. DS: DeepSeek.

LLM	BLEU-4 ↑	BERTScore ↑	ROUGE-L ↑
Evaluation on `FP_ft_eval` *(Hold-out)*			
Llama-3.1-8B-Instruct	**0.4300**	0.8564	**0.5257**
DS-R1-Distill-Qwen-1.5B	0.3869	0.8422	0.4853
DS-R1-Distill-Qwen-14B	0.4268	0.8549	0.5215
DS-R1-Distill-Llama-8B	0.4249	**0.8569**	0.5229
Evaluation on `FP_ft_eval_v2401` *(OOD)*			
Llama-3.1-8B-Instruct	0.0489	0.7565	0.1867
DS-R1-Distill-Qwen-1.5B	0.0468	0.7566	0.1774
DS-R1-Distill-Qwen-14B	0.0538	0.7588	0.1845
DS-R1-Distill-Llama-8B	**0.0542**	**0.7593**	**0.1879**

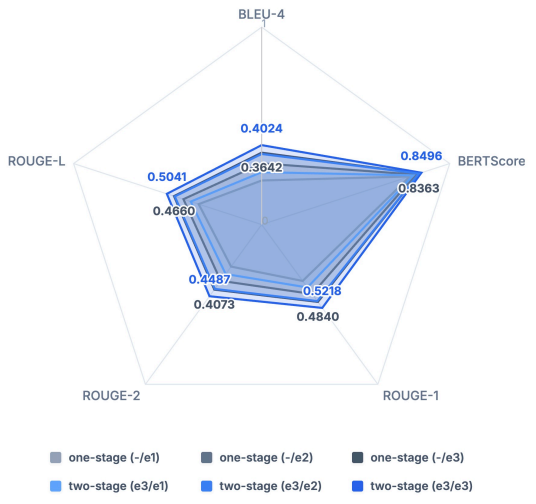


Figure 5: **Comparative analysis of stage-wise training strategy for STELLA-ESM3-Llama-3.1-8B-Instruct on `FP_ft_eval`.** The two-stage strategy consistently outperforms the single-stage training approach.

response schemata, such as MCQA templates, to facilitate more controlled and task-aligned outputs. While both stages utilize the same dataset, they employ differential learning rates and parameter freezing strategies to facilitate optimal convergence. As evidenced in Figure 5, this two-stage paradigm consistently yields superior performance compared to a single-stage joint-training approach across all evaluated metrics for the FP task (see Table 10, Appendix O for more detailed results).

6 Conclusion and Future Work

This study introduces STELLA, a multimodal LLM tailored for protein function elucidation, supported by the curated OPI-Struc dataset—a high-fidelity resource specifically designed for multimodal instruction tuning. By synergistically aligning bimodal sequence-structure representations with the extensive world knowledge embedded in LLMs, STELLA not only streamlines the multimodal training pipeline but also defines a new state-of-the-art benchmark in two critical domains: functional description and enzyme-catalyzed reaction prediction. Beyond its empirical performance, STELLA represents a significant paradigm shift in computational protein science (Fan et al., 2025), effectively bridging the chasm between generative AI and life sciences. Our findings demonstrate that the integration of structural topologies and natural language reasoning can transcend the limitations of traditional protein language models. Looking ahead, we envision that the development of domain-specific biomolecular tokenizers and the deployment of autonomous agentic AI will unlock unprecedented analytical capabilities in biomedical discovery.

Acknowledgment

1. Prior ARR submission: https://openreview.net/forum?id=Dknvz0M89b.
2. Some of the figures in this manuscript were refined using AI assistants to ensure optimal presentation.

Limitations

Despite its advancements, STELLA is not without limitations. Currently, the model's performance is bounded by the granularity of the structural **tokenization** and the inherent complexity of aligning high-dimensional geometric features with discrete linguistic tokens. While effective, the current framework primarily relies on general-purpose LLM reasoning, which may occasionally lack the hyper-specialized biochemical intuition required for *de novo* functional discovery. Furthermore, the current iteration does not yet incorporate **external knowledge integration**, which could provide real-time access to the rapidly expanding body of biological literature and structural databases. Future research will focus on developing high-resolution structural adapters, exploring RAG-enhanced workflows to mitigate hallucinations, and investigating multi-agent architectures to handle complex, multi-step biological reasoning tasks. In addtion, well-established **benchmark** suite remains limited. The aforementioned refinements will further solidify multimodal LLMs and agentic AI as indispensable tools for accelerating innovation within the vast and intricate landscape of life sciences.

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*.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, and 1 others. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500.

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.

Chai-Discovery-Team, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhonikov, and Kevin Wu. 2024. Chai-1: Decoding the molecular interactions of life. *BioRxiv*, pages 2024–10.

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.

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 Lamoureux. 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 Dallago, 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*.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-Instructions: A large-scale biomolecular instruction dataset for large language models. In *The Twelfth International Conference on Learning Representations*.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. 2022. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13.

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, 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, Yueze Wang, Tiejun Huang, and Bo Zhao. 2024. Efficient multimodal learning from data-centric perspective. *Preprint*, arXiv:2402.11530.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Pedro Hermosilla and Timo Ropinski. 2022a. Contrastive representation learning for 3D protein structures. *Preprint*, arXiv:2205.15675.

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

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

Shulin Huang, Linyi Yang, Yan Song, Shuang Chen, Leyang Cui, Ziyu Wan, Qingcheng Zeng, Ying Wen, Kun Shao, Weinan Zhang, and 1 others. 2025. ThinkBench: Dynamic out-of-distribution evaluation for robust llm reasoning. *arXiv preprint arXiv:2502.16268*.

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. 2024. Multi-modal large language model enables protein function prediction. *bioRxiv*.

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. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

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, and 15 others. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Jiashan Li, Xi Chen, He Huang, Mingliang Zeng, Jingcheng Yu, Xinqi Gong, and Qiwei Ye. 2025. Sable: bridging the gap in protein structure understanding with an empowering and versatile pretraining paradigm. *Briefings in Bioinformatics*, 26(2):bbaf120.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, and 1 others. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

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

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

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. BioMedGPT: Open multimodal generative pre-trained transformer for biomedicine. *Preprint*, arXiv:2308.09442.

OC-Team and Qiwei Ye. 2025. Towards unraveling biomolecular conformational landscapes with a generative foundation model. *bioRxiv*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. 2019. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

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. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

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

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

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

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

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, and 8 others. 2021. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444.

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

Ru Wang, Wei Huang, Selena Song, Haoyu Zhang, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. 2025a. Beyond in-distribution success: Scaling curves of cot granularity for language model generalization. *arXiv preprint arXiv:2502.18273*.

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

Zhicong Wang, Zicheng Ma, Ziqiang Cao, Changlong Zhou, Jun Zhang, and Yi Qin Gao. 2025b. Prot2Chat: protein large language model with early fusion of text, sequence, and structure. *Bioinformatics*, 41.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, and 1 others. 2025. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, pages 2024–11.

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

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

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. 2022. PEER: a comprehensive and multi-task benchmark for protein sequence understanding. *Advances in Neural Information Processing Systems*, 35:35156–35173.

Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. Berkeley function calling leaderboard. `https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html`.

Xinhao Yao, Ruifeng Ren, Yun Liao, and Yong Liu. 2025. Unveiling the mechanisms of explicit CoT training: How cot enhances reasoning generalization. *arXiv preprint arXiv:2502.04667*.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. *Preprint*, arXiv:2311.03099.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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. 2024. ∞bench: Extending long context evaluation beyond 100k tokens. *Preprint*, arXiv:2402.13718.

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

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. 2023. Judging LLM-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Appendix

A Prompt Templates for Training

Box A.1: The prompt template for 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 A.2: The prompt template for 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>
```

B Prompts for Evaluation

We design the following evaluation prompts to constrain the model output in specific tasks.

Box B.1: Evaluation prompt for FP_ft

```
<user>
What are the main functions of this protein?
```

Box B.2: Evaluation prompt for FP_mc

```
<user>
Please answer the question directly with the
    corresponding letter (A, B, C, or D) from
     the provided options.
```

Box B.3: Evaluation prompt for EP

```
<user>
What is the enzyme name linked to this
    protein?
```

C Analysis of Data Label Distribution of OPI-Struc

Figure 7 shows (a) the length distribution of functional descriptions in the Function dataset and (b) the frequency of enzyme names in the Enzyme dataset. Figures 6 illustrates the distribution of protein sequence lengths across the FP (left) and EP (right) tasks for training and testing sets.

D Performance Sensitivity to Structural Representational Density

To delineate the relationship between structural familiarity and predictive accuracy, we performed a clustering-based sensitivity analysis. We first partitioned the training set into structural clusters using Foldseek based on global fold similarity. For each protein in the test suite (Func_ft_test, N=4,203), we identified its corresponding structural cluster within the training corpus. The **representational density**—defined as the number of structurally homologous training samples within the matched cluster—serves as a proxy for structural novelty; a smaller cluster size indicates a more novel topology relative to the training distribution.

As summarized in Table 5, we observed a positive correlation between representational density and predictive performance. Specifically, the mean ROUGE-L score increases from 0.4323 for near-unique structures (cluster size ≤ 1) to 0.6691 for well-represented protein families (≥ 20 samples). Notably, STELLA retains substantial generative

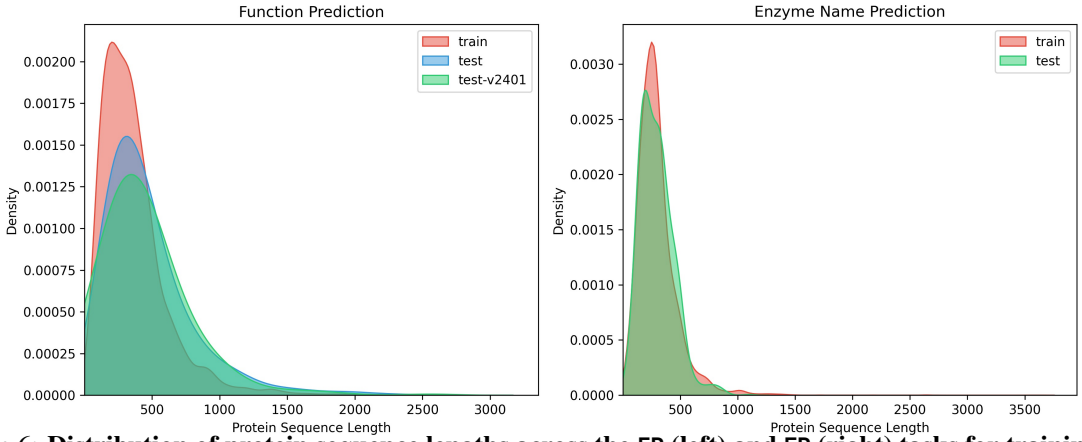


Figure 6: **Distribution of protein sequence lengths across the FP (left) and EP (right) tasks for training and testing sets.** The variation in sequence length distribution between the training and testing sets ensures model robustness across proteins with diverse structural complexities.

capacity even in the most challenging "structural dark matter" regime (0-1 similar training samples), underscoring its ability to generalize beyond simple structural memorization to intrinsic sequence-structure-function mappings.

This performance gradient suggests that while STELLA benefits from structural motifs encountered during training, its multimodal alignment enables the inference of biological roles for proteins with unseen global folds.

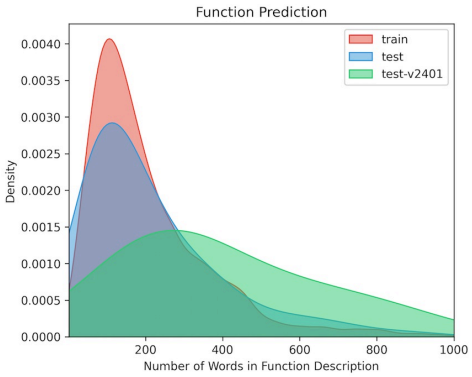
Table 5: Performance stratification by structural representational density. Test samples are grouped based on the number of structural homologs identified in the training set via Foldseek.

Representational Density (Train Samples per Cluster)	Testing samples	Mean ROUGE-L
$(0, 1]$	1202	0.4323
$[2, 5]$	1205	0.4918
$[6, 19]$	921	0.5558
$[20, 1134]$	875	0.6691

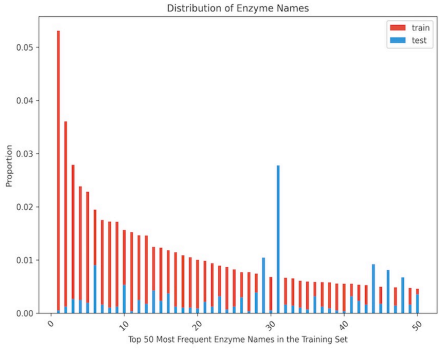
E Hyperparameters for Training and Evaluation

The stage1 training aims to align the embedding space of protein structures and texts. 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 A.

The hyperparameters in two stages are summarized in Table 6. 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.



(a) Length distribution of functional descriptions in the `Function` dataset.



(b) Frequency of enzyme names in the `Enzyme` dataset. The training set follows a long-tailed pattern, but the test set distribution differs significantly.

Figure 7: Analysis of label distributions across datasets.

Table 6: **Hyperparameters for stage1 training, stage2 training and testing.** FFT: Full Fine-tuning.

Config	Stage1	Stage2	Testing
DeepSpeed ZeRO Stage	2	3	N/A
optimizer	AdamW	AdamW	N/A
optimizer hyperparameters	(β_1,β_2)=(0.9, 0.999), eps=1e-8	(β_1,β_2)=(0.9, 0.999), eps=1e-8	N/A
per_device_train_batch_size	2	1(FFT)/2(LoRA)	N/A
gradient_accumulation_steps	4	2(FFT)/4(LoRA)	N/A
gradient_checkpointing	True	True	N/A
learning rate (lr)	2e-5 (Connector)	2e-4 (Connector), 2e-5 (LLM)	N/A
weight decay	0.0	0.0	N/A
warmup steps	48	-	N/A
warmup ratio	-	0.03	N/A
lr scheduler type	cosine	cosine	N/A
training epochs	3	3	N/A
GPU	4*A100	8*A100(FFT)/4*A100(LoRA)	1*A100
temperature	N/A	N/A	0.2
top_k	N/A	N/A	50
top_p	N/A	N/A	0.75
num_beams	N/A	N/A	1
max_new_tokens	N/A	N/A	1000
use_cache	N/A	N/A	True
do_sample	N/A	N/A	True

F Comparison of Protein Encoders

In terms of STELLA's architecture, we employ three protein encoders–ESM3 (Hayes et al., 2024), Prot2Text (Abdine et al., 2023), and SaProt (Su et al., 2023)–for comparative analysis. 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 investigate the impact of different encoders on the STELLA's overall performance, providing insights into the contributions of different components to the its capability.

ESM3 is a large multimodal model pretrained on massive sequence, structure, and function tokens using masked language modeling (MLM) strategy. 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 information.

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

SaProt is a large-scale pre-trained pLM utilizing around 40 million protein sequences and structures, with a structure-aware vocabulary integrating 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 via FoldSeek (Van Kempen et al., 2024). However, this encoder is not aligned with functional annotation text.

Figure 8 illustrates the UMAP visualization of protein embeddings generated by ESM3, Prot2Text, and SaProt for the 4,203 testing samples in `Func_ft_test`, from which it can be observed that ESM3 provides more distinct and informative protein feature representations.

G Different Composition of Protein 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 7.

H Ablation of Training Epochs with Hybrid Datasets

An ablation study was conducted to evaluate model performance across varying training epochs. For the training with the hybrid three training datasets, i.e., `Func_ft_train`, `Func_mc_train` and `Enzyme_train`, all metrics demonstrated consistent improvement with extended training, progressing from (e3+e1) to (e3+e3), as illustrated in

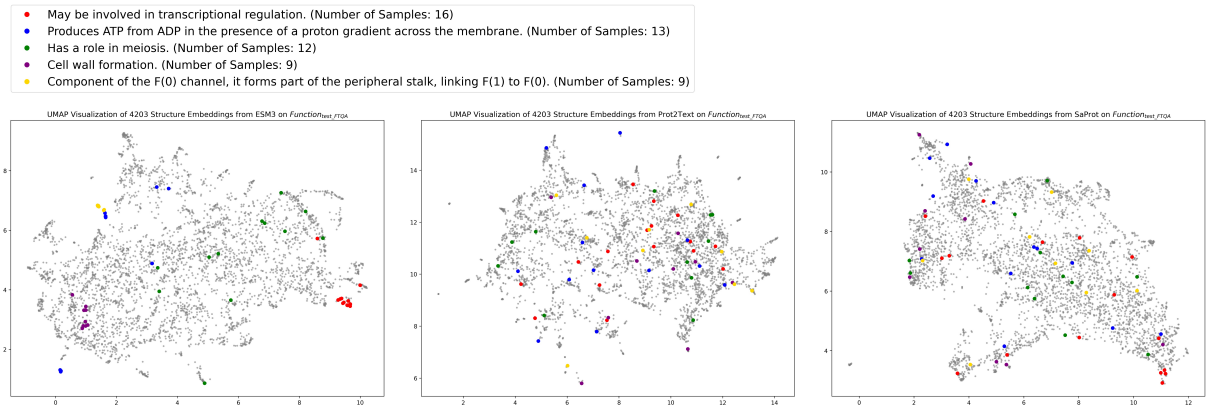


Figure 8: **UMAP visualization of 4,203 protein structure embeddings in the testing set `Func_ft_test` generated by ESM3, Prot2Text, and SaProt.** Each plot illustrates the clustering of protein structures based on their embeddings, revealing the representational differences among the three encoders. The highlighted proteins belong to specific functions as detailed in the legend. ESM3 demonstrates the strongest representative ability.

Table 7: **Specifications of STELLA composition of various protein structure encoders and foundation LLMs.**

Protein encoder	LLM	Note for LLM	STELLA variant
ESM3	Llama-3.1-8B-Instruct (AI@Meta, 2024)	Open source by Meta	STELLA-ESM3-Llama-3.1-8B-Instruct
	Llama-3-8B-Instruct (AI@Meta, 2024)	Open source by Meta	STELLA-ESM3-Llama-3-8B-Instruct
	Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)	Open source by Mistral AI	STELLA-ESM3-Mistral-7B-Instruct-v0.2
	Phi-3-mini-128k-instruct (Abdin et al., 2024)	Open source by Microsoft	STELLA-ESM3-Phi-3-mini-128k-instruct
	BioMistral-7B-DARE [a]	Tailored for biomedical domain	STELLA-ESM3-BioMistral-7B-DARE
	BioMedGPT-LM-7B [b] (Luo et al., 2023)	Tailored for biomedical domain	STELLA-ESM3-BioMedGPT-LM-7B
Prot2Text	Llama-3.1-8B-Instruct	Open source by Meta	STELLA-Prot2Text-Llama-3.1-8B-Instruct
	Llama-3-8B-Instruct	Open source by Meta	STELLA-Prot2Text-Llama-3-8B-Instruct
	Mistral-7B-Instruct-v0.2	Open source by Mistral AI	STELLA-Prot2Text-Mistral-7B-Instruct-v0.2
	Phi-3-mini-128k-instruct	Open source by Microsoft	STELLA-Prot2Text-Phi-3-mini-128k-instruct
	BioMistral-7B-DARE	Tailored for biomedical domain	STELLA-Prot2Text-BioMistral-7B-DARE
	BioMedGPT-LM-7B	Tailored for biomedical domain	STELLA-Prot2Text-BioMedGPT-LM-7B
SaProt	Llama-3.1-8B-Instruct	Open source by Meta	STELLA-SaProt-Llama-3.1-8B-Instruct
	Llama-3-8B-Instruct	Open source by Meta	STELLA-SaProt-Llama-3-8B-Instruct
	Mistral-7B-Instruct-v0.2	Open source by Mistral AI	STELLA-SaProt-Mistral-7B-Instruct-v0.2
	Phi-3-mini-128k-instruct	Open source by Microsoft	STELLA-SaProt-Phi-3-mini-128k-instruct
	BioMistral-7B-DARE	Tailored for biomedical domain	STELLA-SaProt-BioMistral-7B-DARE
	BioMedGPT-LM-7B	Tailored 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 9. 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 9 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).

I STELLA in Action: Case Studies of `FP` Task

STELLA demonstrates feasibility in protein function prediction by integrating sequence-structure represent-ations into LLMs. As illustrated in Figure 10 (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 10 (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 treat-

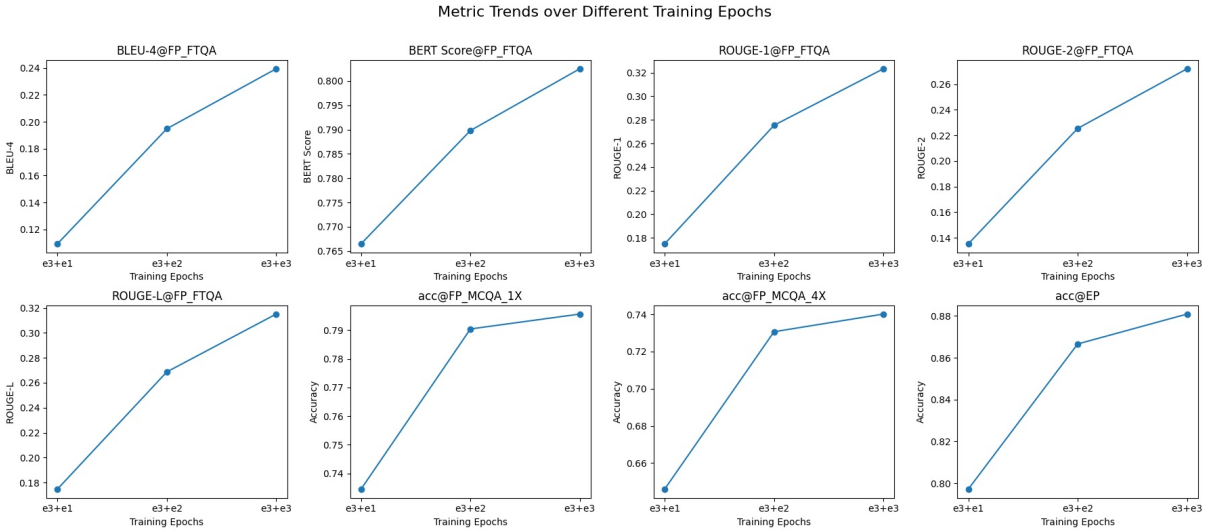


Figure 9: Metrics trend for training with the dataset mix3 over different training epochs.

ment 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.

In addition, Figure 11 presents two representative case studies demonstrating the capabilities of STELLA-ESM3-Llama-3.1-8B-Instruct in exploring and predicting protein functions as well as other biologically relevant properties. In these examples, STELLA behaves in a step-by-step conversational manner to respond to user prompts, highlighting its ability to reason over protein information to provide accurate and biologically meaningful explanations.

J Data Augmentation for `Function` Dataset

The purpose of enriching the `Func_ft_train` dataset into `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 catalyzing 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."[3]

2. Prompt Llama-2-13B-Chat, which maintains computational efficiency while ensuring diversity, 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 augmentated data in the format as shown in the example Box L.1 in Appexdix L.

K Diversified Instructions for Training Data Construction

This section presents some exampels of the diversified natural language instructions (see Box K.1-K.2) generated by ChatGPT, designed for two tasks–

[3] https://www.uniprot.org/uniprotkb/O67618/entry

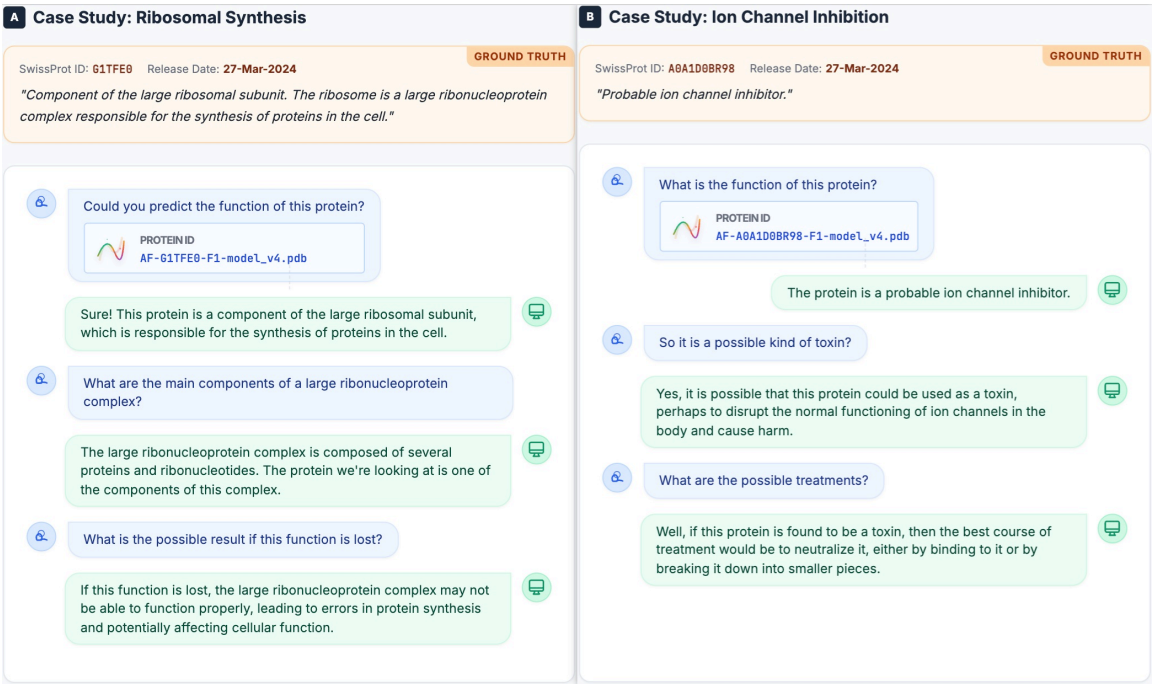


Figure 10: **Functional annotation examples generated by STELLA-ESM3-Llama- 3.1-8B-Instruct.** The proteins G1TFE0 (left) and A0A1D0BR98 (right) are sourced from Swiss-Prot 2024_02. The orange boxes indicate ground-truth annotations, while **green text** highlights correct and essential functional insights predicted by STELLA.

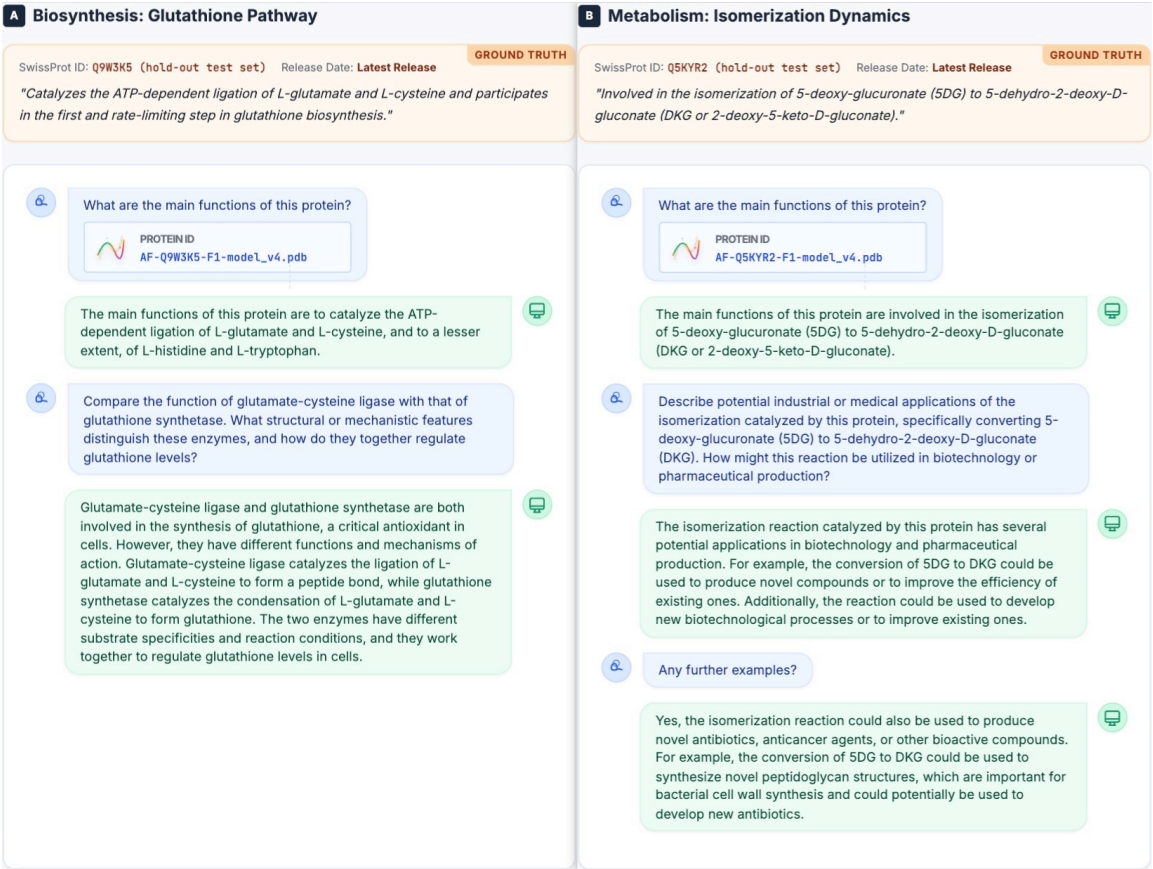


Figure 11: **Functional annotation examples generated by STELLA-ESM3-Llama- 3.1-8B-Instruct.** The proteins Q9W3K5 (left) and Q5KYR2 (right) are sourced from the hold-out test set. The orange boxes indicate ground-truth annotations, while **green text** highlights correct and essential functional insights predicted by STELLA.

FP and EP.

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?
- ...

Box K.2: 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?
- ...

L Data Examples of OPI-Struc

This section lists examples (see Box L.4-L.3) from the OPI-Struc dataset, which supports multimodal instruction 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 L.1: An example of `Func_ft_train_aug` dataset


```
[
  {
    "swissprot_id": "Q1WT60",
    "sequence":
    ↪  "MENKKLHIALLFGGN...YDFTKLGTEKVGKKIIGE",
    "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.2: 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"
      }
    ]
  },
]
```

Box L.3: An example of `Func_mc_train` dataset

```
[
  {
    "swissprot_id": "P62877",
    "sequence":
    ↪  "MAAAMDVDTPSGTNS...RQVCPLDNREWEFQKYGH",
    "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 the `Func_ft_train` dataset

```
{
  "swissprot_id": "Q0BWM9",
  "sequence":
  ↪  "MFNKQSVSLEWAGRT...DQETGKELAEEAGDDASED",
  "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"
}
```

M Hold-out Evaluation Results on the `FP_ft_eval` Benchmark

Table 8 presents a comparative analysis of STELLA against baselines and state-of-the-art methods using the `FP_ft_test` hold-out set. The results underscore STELLA's capability in generating high-fidelity functional descriptions.

Table 8: **Hold-out evaluation of FP performance. Bold**: best; underline: runner-up.

Model/Method	BLEU-4 ↑	BERTScore ↑	ROUGE Score ↑		
			ROUGE-1	ROUGE-2	ROUGE-L
Foldseek	0.3627	0.8358	0.4799	0.4027	0.4586
Prot2Text$_{BASE}$	0.3511	0.8430	0.5059	0.4271	0.4849
Prot2Text$_{LARGE}$	0.3629	0.8520	0.5368	0.4560	0.5140
ProteinChat	0.1918	0.7970	0.3957	0.2799	0.3648
STELLA (e3+e6)	**0.4300**	**0.8564**	**0.5423**	**0.4747**	**0.5257**

N Zero-shot Temporal Evaluation on the `FP_ft_eval_v2401` Benchmark

Table 9 illustrates the zero-shot Out-of-Distribution (OOD) generalization of STELLA on the `FP_ft_eval_v2401` benchmark. This temporal test set assesses the model's generalization on proteins characterized after the training data cutoff.

Table 9: **Zero-shot temporal OOD evaluation. Bold**: best; underline: runner-up.

Model	BLEU-4 ↑	BERTScore ↑	ROUGE Score ↑		
			ROUGE-1	ROUGE-2	ROUGE-L
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**
ProteinChat	0.0205	0.7413	0.2121	0.0855	0.1691

O Comparative Analysis of the Stage-wise Training Strategy

Table 10 compares the efficacy of the decoupled training strategy. By comparing single-stage and two-stage training for STELLA-ESM3-Llama-3.1-8B-Instruct, we highlight the benefit of two-stage optimization in multimodal instruction tuning.

Table 10: **Ablation of the stage-wise training strategy on the FP_ft_eval benchmark.** S1: Stage 1, S2: Stage 2. **Bold**: best.

Strategy	S1 Epoch	S2 Epoch	BLEU-4 ↑	BERTScore ↑	ROUGE Score ↑		
					ROUGE-1	ROUGE-2	ROUGE-L
Single-stage	-	e1	0.2233	0.7885	0.3530	0.2631	0.3350
Single-stage	-	e2	0.3099	0.8199	0.4346	0.3522	0.4160
Single-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**