ProtReason: A Reasoning-Based Framework for Interpretable Protein Function Prediction

by

Sartiz Alam Ayon
24341270
Alvi Sakib Orin
24141164
Arpon Biswas
24141233
Fahad Al Shahid
24141187
Shaikh Faiyaz Shahriyer
24141235

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
BRAC University
June 2025.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

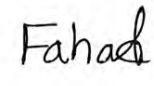
Student's Full Name & Signature:

Sartiz Alam Ayon
24341270

Arpon Biswas
24141233

Alvi Sakib Orin
24141164

Shaikh Faiyaz Shahriyer
24141235



Fahad Al Shahid
24141187

Approval

The thesis/project titled "ProtReason: Reasoning-Based Multimodal Framework for Interpretable Protein Function Prediction" submitted by

1. Sartiz Alam Ayon (24341270)
2. Alvi Sakib Orin (24141164)
3. Arpon Biswas (24141233)
4. Fahad Al Shahid (24141187)
5. Shaikh Faiyaz Shahriyer (24141235)

of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Spring'25.

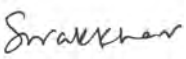
Examining Committee:

Supervisor:
(Member)

Dr. Farig Yousuf Sadeque

Associate Professor
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)



Dr. Swakkhar Shatabda

Professor
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Understanding how protein sequence determines function remains a central challenge in computational biology. While some protein language models have advanced function prediction but most of them produce outputs without any justification or explainability. Protein function can be justified by connecting biological evidence to functional conclusions. We present ProtReason: A reasoning-augmented framework that generates interpretable protein function predictions with structured reasoning traces. In this study, a curated dataset of 87K proteins is constructed which is enriched with protein domain motifs, localization predictions and structural features transformed into reasoning traces linked to functional labels. ProtReason employs a two-stage architecture that first aligns protein sequence embeddings with textual representations and then generates structured outputs including reasoning traces, functional descriptions, and confidence scores. Compared to a sequence-to-function baseline without reasoning, ProtReason achieves significantly improved BERT F1 scores, demonstrating the benefit of incorporating reasoning prior to function prediction. A systematic ablation study with 16 model variants shows the best design principles: a single unified reasoning path is better than a multi-step chain of reasoning and generating reasoning before function prediction yields superior performance. ProtReason performs competitively on standard benchmarks while providing biologically interpretable explanations with calibrated confidence estimates.

Keywords: Computational biology, biological evidence, Embeddings, protein language models, functional descriptions, Chain of thought reasoning.

Acknowledgement

We would like to thank the Almighty most sincerely whose unmeasurable grace, power, and wisdom helped us to go through all delegates of this research. Without His divine blessings, this would not have been able to be done.
Our supervisor, Dr. Farig Yousuf Sadeque, made us refine our work and improve it through his helpful recommendations. His continuous encouragement and constant motivation were a great help to us.
We would like to preciously thank our respected co-supervisor, Dr. Swakkhar Shatabda, who gave us exceptional guidance, excellent feedback and the constant support he had provided us during the course of this study. The experience and guidance that he provided were critical in defining the path and standard of our research.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The first chapter establishes the scientific context, motifs, and intentions of this work. We commence by pointing out recent developments in the representation learning of proteins, next state the practical and theoretical gaps that spur our study, and define the research problem and objectives, the proposed methodology, and the scope and expected challenges.

1.1 Background

Proteins are biomolecules that are used in nearly every biological process, from catalyzing reactions, drug interactions and transporting molecules to regulating signaling networks. Understanding and accurately annotating their functions is essential for drug discovery, enzyme engineering, improving our understanding of evolution and sustainable biotechnology. Classical computational approaches for function annotation primarily relied on sequence or structural homology, where function is inferred by sequence similarity. However, this method breaks down rapidly when sequence identity falls for divergent proteins (identical proteins can have different functions), leaving a vast portion of proteins functionally uncharacterized. So despite rapid growth in sequencing data, functional annotation lags far behind, leaving many proteins labeled as "hypothetical" or "uncharacterized". Recent advances in protein language models have reshaped the field by learning from massive unlabeled data. For example, S-PLM [1] employs multi-view contrastive learning between sequences and structures to capture biochemical signals beyond raw homology and secondary structure prediction. At the same time, large-scale models like Prot2Text[2] integrate protein sequences with natural-language function descriptions. Which aligned embeddings through contrastive learning and then fine-tuning a language model (LLaMA) to generate text-based annotations. This enables sequence embeddings to human-readable biochemical functions. Other approaches explicitly incorporate structure and language. ProtTeX [3] encodes proteins as unified discrete tokens spanning sequence, structure, and text, which can be processed as intermediate tokens for multimodal reasoning. ProteinGPT [4] demonstrates the potential of Q/A-style systems by curating over 132K proteins, fuse sequence and structure encoders with an LLM . Their Protein entries will be enriched with property tags and Q/A pairs, which will be used for instruction tuning to train the LLM to answer questions about protein properties.Even methods with different formulations contribute for instance, ProtDETR [5] frames protein function prediction as a residue

detection problem, using learnable queries to extract local features for each function label, which boosts accuracy and interpretability Despite this progress most methods focus on either classification (e.g., EC numbers), embedding alignment or generative descriptions. Few tackle reasoning oriented tasks that explicitly require connecting structural motifs to biochemical roles through natural language. Inspired by reasoning-based frameworks in other domains, such as ProLLM [6] (which uses chain-of-thought prompts for protein interactions) and CellReasoner [7] (which applies reasoning-based Q/A to cellular processes), We propose a reasoning based dataset for protein function prediction by aligning sequence and textual knowledge within a framework. The goal is to push beyond prediction into interpretable, more generalizable function representations which biologists can verify.

1.2 Rational of the Study or Motivation

The motivation for this study arises from the gap between interpretable prediction with reasoning in protein function annotation. A Reasoning dataset will train models to integrate protein sequence and textual knowledge in a biological backed reasoning which encourages interpretable reasoning about function. Existing models hint at this potential like Prot2Text [2] generates detailed function descriptions from sequence embeddings and ProteinGPT [4] demonstrates that a language model can answer protein queries when trained on Q/A pairs. However, neither was explicitly designed a system with a reasoning chain. For instance, a model may correctly predict "oxidoreductase activity" from sequence, but it cannot easily explain which localization, motifs, or structural features support this conclusion. Our motive is to create interpretable reasoning steps by connecting structural evidence and biochemical descriptions to annotated the function of protein. This approach encourages models to not only provide a label but also justify their predictions in natural language with a cofidence score. This study is also motivated by evidence from related fields like Bioreason [**?**] shows that a reasoning datasets significantly improve performance in biomedical LLMs. ProLLM [6] shows that translating biological pathways into chain-of-thought prompts improves prediction of protein function. ProteinReasoner [8] and ProteinVec+Conformal [9] highlight the importance of interpretability and calibration in protein-related predictions. Building on these ideas, our dataset will be tailored with reasoning teaching models to justify predictions with natural-language explanations to improve protein function prediction by interpreting the prediction with biological knowledge. By leveraging complementary modalities and explicit reasoning, we will build a dataset which will be able to interpret function and also improve performance significantly.

1.3 Problem Statement

Current protein function annotation tools process sequence, structure, and textual information independently, lacking a unified reasoning framework. Approaches such as Prot2Text [2] generate textual descriptions aligned with sequence embeddings but do not enforce explicit logical reasoning chains. ProteinGPT [4], though question–answer oriented, emphasizes answer generation without explicit supervision over reasoning processes. Consequently:

1. Models cannot be directly evaluated on reasoning quality.
2. Predictions lack traceable links to underlying molecular evidence.
3. There is no systematic understanding of whether reasoning improves prediction performance.

Without the reasoning elevated training, it is not possible to test or compare the role of explicit reasoning traces on predicting protein functionality. This research closes that gap by:
a) Creatng a database with reasoning that helps to connect the functional annotation with molecular evidence.
b) Creating a base model (baseline) that have no reasoning traces present.
c) Lastly, by showing that our reasoning traces show superior output than the baseline mode.

1.4 Objective

Our main objective in this research is to build a protein function prediction framework that has reasoning traces and it adds protein sequence information with biological knowledge (textual) to show that reasoning helps to improve prediction performance.

Specific Objectives

1. We designed a 4 component reasoning structure which captures many aspects of protein function reasoning:
    - **Functional Reasoning - R1:** Show us the information about domains, motifs and sequence features on the predicted molecular function.
    - **Localization Reasoning - R2:** It shows how the subcellular location signals and predictions targeting combine with biological role.
    - **Structural Reasoning - R3:** Collects information about membrane topology and structural features wth functional annotations.
    - **Integrated Reasoning - R4:** Provides a confidence score combining evidence from multiple sources and turning it into a mechanistic hypothesis.
2. We created a protein dataset that included high quality reasoning traces. It linked molecular evidence to functional annotations. It further enabled training (supervised) and the evaluation of reasoning enabled model.
3. We used a standard dataset split mechanism using MMseqs2 clustering. We set it at $\leq 45\%$ sequence identity in order to remove redundancy and at the same time, preserve representative diversity.
4. We developed a reasoning architecture (finetuned) which combines protein sequence embeddings with biological representations (textual) and gives structured reasoning outputs.
5. We established a baseline model with no reasoning supervision. It helped us to demonstrate that our reasoning variants presented improved outputs.

6. We used standard text generation metrics like ROGUE, BLEU, BERTScore to evaluate our predictive accuracy and reasoning standard. We also used calibration analysis to check functional reliability and reasoning exactness.

7. Lastly, we showed and analyzed our explicit reasoning contribution using a structured ablation study which examined reasoning structure, depth and output ordering.

1.5 Methodology Overview

The study constructs a logic-based multimodal protein predictor of functions to overcome the shortcomings of known predictors which utilize sequence homology or a classification nomenclature. The strategy combines the use of biotools data and reasoning to enhance accuracy and probability.

The data Prot2Text [2] is used as the startup point of the work, and it consists of 250K protein sequences enriched by Pfam, InterPro, ProSite and DeepLoc etc. annotations. MMseqs2 can cluster the dataset to bring it to 87K entries and ensure biological diversity. Each protein entry is added with an evidence pack which consists of sequence, domain and structural information. Then it is passed down to GPT-OSS 20B to create reasoning traces (R1 to R4) in order to obtain functional predictions.

Training our our model was completed in 2 phases. The initial step involves the MSSP-128 model, comprising of LLaMA-16 embeddings, which are done through weighted pooling strategy. In the second step our model is optimized using LLaMA 3.1 (8B) with help of soft prompting and reasoning based decoding in order to fine tune predictions. Finally, unfunctional precision and integrative chain of thought were all tested in the system using benchmark and measures of causal-consistency, which demonstrate predictable ways of understanding the functions of proteins.

1.6 Scopes and Challenges

Scope

The thesis aims to build a framework based on reasoning to predict the functions of the proteins, using sequence data along with other contextual data given by bioinformatics tools. Contrary to the old-fashioned approaches based on sequence homology or classification labels, this model combines the reasoning traces to offer understandable predictions alongside the demonstrated justification.

The study explores:

1. The methodology behind constructing a reasoning dataset,

2. Reasoning architecture implementation to achieve structured outputs,

3. Multimodal architecture design for structured output generation,

4. Systematic ablation to identify optimal reasoning configurations, and
5. Empirical comparison between baseline and reasoning-augmented models.

Challenges

- Functional descriptions in UniProt are often inconsistent or paraphrased, requiring robust handling of synonyms and domain-specific terminology in reasoning chain construction.
- High-quality reasoning traces require well-evidenced data; proteins with limited annotations pose challenges for grounded reasoning generation.
- Mapping structural features to linguistic tokens requires careful alignment between protein embeddings and language model representations.
- Beyond accuracy, reasoning requires faithfulness to underlying biological evidence, proper calibration of uncertainty, and interpretability.
- Training encoders alongside large language models is resource-intensive; lightweight strategies such as LoRA, quantization, and frozen encoders are essential for efficiency.
- Establishing that reasoning improves performance requires careful experimental design with matched baselines and controlled ablation.

Chapter 2

Literature Review

Chapter 2 looks at almost 20 studies that combine protein shapes, sequences,structures, molecules, and language to better predict protein functions. It explains how mixing these different types of information helps. The research shows that when protein sequences and their descriptions match well, predictions get better.

2.1 Preliminaries

Traditional methods often treat protein sequences and structures separately, which limits their ability to predict protein functions, particularly they have no ability to do function reasoning. This work, suggests a reasoning-based approach that leverages protein sequences coupled with the text description of biochemical processes, based on reasoning traces produced by different bioinformatics tools. The framework combines sequence data with functional annotations in a common space, which is oriented towards reasoning chains based on biotools. It is a more holistic method to understand protein functionality with more accuracy and can be predictive as well as interpretable.

Key Concepts:

1. **Multimodal Feature Extraction:** The sequence of amino acids of each protein entry in UniProt is encoded with pretrained encoder like ESM-2 (3B). The Llama-3.1 (8B) is used to process complementary natural language data such as function-related queries, reasoning, as well as responses. This two-fold representation provides biochemical properties and semantic background on the reasoning of protein functions.

2. **Integrated Architectures:** The framework combines sequence embeddings from protein encoders with textual embeddings via a learned projector module (MSPP-128). This projector maps the variable-length representations of proteins into fixed-length prefix tokens with language model input. The system is able to infer protein functions with explicit justification even in proteins with minimal homology to training proteins by allowing the correspondence of low-level molecular patterns sensed by protein encoders to high-level linguistic descriptions generated by language models.

3. **Contrastive Learning and Reasoning-Based Supervision:** During the

alignment stage, the protein embeddings are drawn towards their respective text embeddings and push away the non-matching pairs using contrastive objectives (InfoNCE loss). This results in many proteins of similar functions coming together in a common semantic space. During the generation stage, the language model learns to produce R1–R4 stepwise reasoning traces that explain functional outcomes by connecting specific evidence items such as- domains, localization signals, structural features to functional conclusions. This aspect allows the model to make evidence-based predictions that are interpretable instead of black-box predictions.

4. **Transfer Learning & Coputational Efficiency:** Leveraging pretrained encoders (ESM-2 3B) and language models (LLaMA-3.1 8B) provides strong biological and linguistic priors that would be prohibitively expensive to learn from scratch. Efficient fine-tuning via QLoRA (Quantized Low-Rank Adaptation) enables training on modest hardware (dual T4 GPUs) while maintaining prediction quality. The protein encoder remains frozen during training, reducing memory requirements and preventing catastrophic forgetting of learned biological representations.

2.2 Review of Existing Research

This section surveys state-of-the-art approaches across three categories: structure-centric graph neural network frameworks, sequence-based contrastive and retrieval methods, and multimodal generative models. However, there are difficulties left: models continue to rely heavily on experimental structures which are limited, over-fit on the new folds of proteins popular in research, are not consistently appropriately calibrated to predict confidence in their accuracy

However, a critical gap emerges from this survey: **no existing work systematically compares reasoning-augmented generation against matched non-reasoning baselines to quantify the specific contribution of explicit reasoning to prediction quality.** While several papers generate natural language outputs and some incorporate reasoning-like components, none establish controlled experiments isolating reasoning's effect. This gap is particularly notable given that reasoning-enhanced approaches in related biological domains (genomics, cell biology) have demonstrated substantial performance improvements.

2.2.1. Structure-centric GNN Frameworks

Graph neural networks that operate directly on three-dimensional atomic graphs to capture local catalytic geometry and overcome fold bias in enzyme classification. Relevant representative papers are Structure-Aligned pLM [1]

Chen et al. (2025) [1] introduced a structure-aligned Protein Language Model (pLMs) that incorporates structural awareness into sequence-based models using a dual-task learning architecture. This framework integrates protein language models (pLMs) with Graph Neural Network(GNN) embeddings derived from pre-trained models like GearNet, so that the model can learn inter-protein spatial relationships with the discussion of latent-level contrastive learning. It parallely extracts intra-protein spatial information by predicting structural tokens and encoding local residue geom-

etry. A central contributive upgrade is a residue loss selection routine which fortifies training, focusing on residues of privileged PDB areas, eliminating noisy as well as uninformative instances relying on surplus loss compared to a quotient model. Applying to ESM2 and AMPLIFY, the derived structure-corrected variants (SaESM2 and SaAMPLIFY) show significant gains in structure based prediction tasks such as contact prediction (P@L/5 = 61.02), fold assignment, secondary structure assignment and mutation fitness prediction (Sp = 0.951). The model also attains a EC prediction Fmax of 0.861 that is far above baseline sequence-only and structure-aware models such as the ESM2-s and ISM. These findings justify the usefulness of such structural enhancement in terms of both inter- and intra-chain architectural elements to achieve improved performance in predicting downstream functional properties, providing an attractive basis in pLM architectures with a more charismatic representation of the structural context.

2.2.2 Sequence-Based Contrastive and Retrieval Methods

Protein language models whose sequence embeddings are contrastively aligned with structural information, yielding "structure-aware" representations even when 3-D data are absent at inference. Relevant representative papers are S-PLM [1], CLEAN-Contact [10], ProtDETR [5], Protein-Vec+Conformal [9]. Yang et al. (2024) [10] introduce CLEAN-Contact, a multimodal contrastive learning framework that performs enzyme function prediction by combining both sequence information with inferred structure information. It uses the powerful transformer-based protein language model (ESM-2) and a convolutional neural network (ResNet-50), These modality-specific representations are orientated in a common latent space using contrastive loss which is optimised using triplet sampling. Two test datasets New-392 and Price-149 are used to evaluate CLEAN-Contact at different levels of the EC hierarchy using standard metrics such as Precision, Recall, F1-score, and AUROC. The model outperforms its predecessor CLEAN and other models with precision and F1-score improvements of up to 16.2% and 12.3%, respectively. It is also great at generalizing, especially on the rare classes of ECs and proteins whose average sequence identity to the training set is less than 50%. One of those recent advancements is its adaptive EC-level prediction process in which its confidence scoring is based on Gaussian Mixture Model (GMM) which makes the predictions of hierarchical higher EC levels to be more reliable. The capability of CLEAN-Contact to detect structural noise and retrieve literal functional annotation in low homology context in the fields of enzyme discovery, microbial pathway design as well as large-scale metabolic modelling. The authors suggest future studies by integrating complete 3D geometries and hierarchical contrastive loss functions to further increment accuracy and performance. This article also emphasizes how multimodal contrastive alignment is important in solving the ambiguity of the functional predictions of sequences.

Yang et al. (2025) [5] proposed ProtDETR, a new Transformer-based model that redefines the problem of enzyme function prediction as a detection task. It is applicable by detecting particular residue units which are causes of specific Enzyme Commission (EC) numbers. The model suggests a mini list of learnable queries referring to functions that are mingled with residue level features that are obtained out of ESM-1b through cross-attention. ProtDETR accomplishes the state-of-the-art

upon benchmark data, such as New-392 and Price-149, improving F1 by as much as 25% compared to the previous approaches, like CLEAN. In monofunctional settings, it also has better performance on EnzBert and ECPred on all the levels of EC hierarchy on the ECPred40 data. Notably, ProtDETR has good interpretability, indicated by attention maps that point out residue fragments that are involved in each of the predicted functions, providing a view into the catalytic sensibility of monofunctional and a multifunctional enzyme. The model can generalize functions when low or little similarity is found, and it can also keep reliable recall even of infrequent enzyme functions. Due to adaptive, residue-based functionality detection capability, ProtDETR represents a viable modality in explainable functional annotation in metagenomics, protein design and evolutionary studies.

Boger et al. (2025) [9] suggested a statistically justified framework, and the result of the new adapted approach to protein normal mining was the use of conformal predictions that deals with a lack of confidence in a traditional homology search strategy. Their work involved calibration of protein search models such as Protein-Vec where the similarity scores were transformed into probability sets through isotonic regression and/or Venn predictors-Abers, and also contains the hierarchical loss functions in line with annotation ontologies (EC, SCOPe, Pfam) to aid user-defined control risk (e.g., FDR subsets, FNR). This model has been tested on the data spaces such as Pfam-annotated UniPort, minimal JCVI Syn3.0 genome, and enzyme function benchmark (New-392, Price-149) and the structural alignment judging/prefiltering is supported by AlphaFold and SCOPe entries. Conformal methods restored the previous unannotated Syn3.0 genes that were exactly matched with the functional counterparts, and it bested methods such as CLEAN because it had 57.65% F1 and 81.50% AUC during enzyme annotation. Pre-filtering of Protein-Vec also slimmed the AlphaFold database by 31.5 percent with 82.8 percent of the high-confidence DALI hits preserved. This gave revolutionary calibration and interpretability and the ability to produce empty result sets when the confidence is insufficient. It is more than CLEAN and structural alignment tools. Such technique maximizes the functional annotation accuracy in metagenomics, large scale structural analyses, as well as in enzyme mining. The author demonstrated the prospect of how future plans lay ahead as they mentioned a proposal to expand the framework with conformal quantile regression and improve the worth of the sequence perplexity of the protein language models to render annotation under distributional shifts further.

2.2.3 Multimodal Generative & Text-Aligned Models

Large language-model platforms that fuse protein encoders with generative or instruction following LLMs to support free-text annotation, Q&A or sequence editing. Here the representative papers are Prot2Text [2], STELLA [11], ProteinGPT [4], ProtET [12], InstructProtein [13], ProtTeX [3], ProteinReasoner [8], ProLLM [6], CellReasoner [7], SCI-Reason [14],BioMol-MQA [15], Thought Graph [16]

Xiao et al. (2025) [11] present STELLA is a multimodal large language model (LLM) designed for protein function prediction using both protein sequence and tertiary structural information. To overcome the limitations of retrieval-based or sequence-only models, STELLA includes ESM3 a single encoder that can process 3D

structures and protein sequences. Also, Llama-3.1-8B-Instruct serves as the logical backbone for producing context aware functional outputs. The OPI-Struc dataset contains high-quality protein structures as well as descriptive functional annotations used for training using multimodal instruction tuning. STELLA is evaluated using metrics like BLEU-4, ROUGE and classification accuracy on two primary tasks such as functional property description and enzyme-catalyzed reaction prediction. The model consistently outperformed traditional protein language models and structure retrieval tools Foldseek in structurally unclear contexts. STELLA uses information without complete or noisy structural data, but bounds the prediction with high accuracy. Real-life applications require robustness which are usually structurally ambiguous. The dynamic user interface of this model permits natural language queries, and therefore, the generation of conceptually and biologically relevant results that add an important level of utility to its application base, as has been witnessed in the study of proteins through annotation in the field of research and design of enzymes. Compared to unimodal or retrieval based systems STELLA obtains a significant improvement in capturing complex functional patterns based on geometry and sequence. The authors also propose further improvements through continuous learning. The incorporation of additional biological priors, and the extension to broader protein interaction contexts by establishing STELLA as a powerful and generalisable framework for protein biology.

Xiao et al. (2024) [4] presents ProteinGPT a multimodal large language model (LLM) designed for protein level instruction understanding and conceptual understanding. The model tries to bridge the divide between the protein sequence structure data and the natural language queries by aligning multimodal protein embedding representations with the ones compatible with LLM. ProteinGPT generates sequence and structure embeddings using frozen ESM2 and ESM-IF1 encoders. Then projected into a unified space through soft prompt modality fusion a discreet method that does not require training the core LLM. The model is trained in two stages of protein representations aligned with textual descriptions, then the model is fine-tuned on instruction following tasks using QA pairs which are generated by GPT-4o. The authors produced ProteinQA a large dataset with around 3.7 million QA pairs across 132,000 proteins. ROUGE, BERTScore, GPTScore, and a GPT-4 accuracy test are used as evaluation benchmarks. ProteinGPT has a conceptual F1-score of 0.829 and more than 80% accuracy in fact. IT is outperforming general-purpose LLMs such as GPT-4 and GPT-3.5 on protein-specific tasks. ProteinGPT's architecture allows for context-specific, biologically accurate answers to complex molecular queries by making it a valuable tool in protein engineering, drug discovery and molecular biology research. The model's modular design allows easy extension to more complex biomolecular systems without retraining the LLM backbone. Future plans involve expansion to also cover protein complexes, multi-chains systems and non-protein ligands. IT will involve structural biological reasoning to establish increased robustness and insights in scientific applications.

Yin et al. (2024) [12] presented ProtET, a novel protein editing system that incorporates multi-modal contrastive learning which was inspired by CLIP. The major innovation of ProtET is that it is an effective combination of sequence data and textual instructions to be read by humans in order to strengthen protein engineering using ele-

ments of ProtET. This model uses a hierarchical two-stage learning pipeline. Initially, it pretrains models to align protein sequences with textual functional descriptions using two specialized large language models (LLMs): ESM-2 for encoding protein sequences and PubMedBERT for biological textual annotations. In the following step, the model adds to a Feature-wise Linear Modulation(FiLM) directed generative decoder and uses such aligned representations and editing directions to generate optimized protein sequences. Rigorous comparisons were made against the common protein editing strategies prevalent including Single-Mutant, AFP-DE, and EvoPlay and were specifically aimed towards enzyme catalytic activity, protein stability and antibody-specific binding activity. It was found that ProtET outperformed current methods dramatically by nearly 17% in terms of protein stability benchmarks, which means that it has a real applicability to biomedical applications in the real world including production of vaccines and gene therapy. ProtET enabled controllability and interpretability of protein editing processes by the factions of interactive, text-based instructions. The authors stated that the generalizability and predictive power of this methodology should be improved further by using bigger multi-modal data sets and more biological background in order to make it even more reliable and the authors also see a potential future in further research work in the area of protein modification.

The current issue with the application of Large Language Models (LLMs) to proteomics is that the semantic gap between the language of protein sequences and the human language of description is very large. This disconnection has constrained the usefulness of the LLMs where they cannot translate fluently between these two different languages. Wang et al. (2023) [13] solve this issue with InstructProtein a new bidirectional LLM. It is able to infer the function of a protein based on its sequence, and also to produce a sequence based on a natural language query. The most important innovation of the model is the training process. It is then supervised instructionally following a normal pre-training period on independent protein and text corpora. To overcome the low quality of current datasets that are associated with an imbalance in annotation and poor instructional cues, the authors developed a knowledge graph-based instruction generation model. The system also automatically builds a high-quality, diverse and a balanced instruction dataset and this is essential in effectively aligning the two different domains. The framework helps in overcoming the biases that exist in raw data by converting the factual knowledge on the graph into explicit instructions so that the model is trained on a large range of tasks. The output is a model that performs state-of-the-art on both text-to-protein and protein-to-text generation tasks. The work is a breakthrough in terms of how proteins are treated as a collection of characters to the establishment of a complete, bilingual model. InstructProtein allows more intuitive and powerful applications by allowing LLMs to reason and manipulate proteins in a high-level conceptual language. This kind of matching helps to enable researchers to engage in a natural and in depth way with protein data and can increase the speed the hypothesis creation and in silico experimentation in protein design and functional annotation (Wang et al., 2023).

Abdine et al. (2023) [2] introduced Prot2Text, a multimodal protein function prediction model that generates free-text descriptions of protein function. Prot2Text's encoder uses a Relational Graph Convolutional Network (RGCN) to process 3D protein structure and a pretrained ESM-2 protein language model to encode the

amino-acid sequence, while a GPT-2 decoder generates the functional description. The model is trained on a large SwissProt dataset (256K entries) of proteins with annotated functions. Abdine et al. reported that combining structural and sequential modalities substantially improved performance even the RGCN encoder alone outperformed a vanilla Transformer encoder, and replacing the untrained Transformer with the pretrained ESM2-35M model yielded an additional 16 BLEU-point gain. The full Prot2Text model (RGCN+ESM2-35M) achieved BLEU-4 35.11 and ROUGE-L 48.33 on the test set, markedly higher than any unimodal baseline, was among the first to fuse graph-based structural data with large language models for protein annotation. The authors note limitations in their design: in particular, the RGCN encoder was not pretrained, so it may not fully capture complex structural patterns. They suggest that pretraining the graph encoder on related tasks could improve performance in future work.

Although Large Language Models (LLMs) have demonstrated potential in protein science, most studies have considered proteins as a one-dimensional sequence, ignoring the fact that the functionality of a protein is determined by its 3D structure. This sequence based view inhibits the ability of LLCMs to deal with problematic issues that need structure. Ma et al. (2025) [3] presented a solution mentioned as paradigm-shifting in ProtTeX which is a new approach towards structure in context Reasoning and editing the proteins with LLMs. ProtTeX is a new system that re-architectures the presentation of protein data to an LLM. The fundamental novelty of it is the tokenization of protein sequences, 3D structures and textual data into a single discrete space. Under this scheme, the joint training of the LLM can be performed in a straightforward Next-Token Prediction paradigm, during which the model is trained to understand the interaction between all the three modalities. With this approach, the LLM is able to do structure-in-context reasoning, viewing structural information as part of its reasoning process. The applied findings are significant: ProtTeX has a two-fold accuracy increase in predicting protein functions in comparison with expert models. In addition to prediction, it opens up strong generative abilities, such as high-quality conformational generation and text-promptable protein design. This work touches a new milestone shows that if the language of structure us taught to an LLM, a proper insight is achieved in-to the molecular world (Ma et al., 2025) [3].

More recent developments in protein science are utilizing Large Language Models (LLMs) to go beyond the simple analysis of sequences to more complex, multi-modal reasoning. The current research that is underway is aimed at adding more biological context to come up with stronger and more interpretable models. One of the difficulties lies in the symbolic language of amino acids being incompatible with the descriptive human language. This is dealt with by InstructProtein (Wang et al., 2023) which produces a model that has the ability to generate both ways, i.e., translation between protein sequence and its functional description by instructing through knowledge. Based on this, ProtTeX (Ma et al., 2025) [3] integrates the important aspect of 3D structure, tokenizing it with sequence and text into a common space. This can be used to think in a structure-in-context way which is vital to functionality. Such models are being developed using large datasets such as Mol-Instructions (Fang et al., 2024) which is a massively large, domain-specific dataset to train robust biomolecular LLMs. Moreover, an even bigger jump is to incorporate explicit reasoning into these

models. Likewise, ProLLM (Jin et al., 2024) [6] uses a special (specifically) Protein Chain of Thought to reason about protein-protein interactions by reasoning over the entire signaling pathway, which contains complex indirect interactions. These tools changed the view towards biology and genetics and inspired to use scientific discovery (Jin et al., 2024) [6].

Fallahpour et al. (2025) [17] proposed BioReason, a reasoning based multimodal framework for biology. BioReason tightly couples a DNA sequence encoder (a genomic foundation model) with an LLM, enabling the model to directly process genomic input and perform multi-step biological reasoning. Through combined supervised fine-tuning and reinforcement learning, BioReason learns to produce logical, domain-coherent deductions. In benchmark tasks (e.g. KEGG disease pathway and variant effect prediction), it markedly improved accuracy (KEGG pathway accuracy rose from 88% to 97%) and achieved on average 15% performance over strong single-modality baselines. Crucially, BioReason generates interpretable reasoning traces – step-by-step biological deductions that explain its predictions. This addresses a key limitation of "black-box" models by providing transparent, mechanistic explanations. Taken together, these studies illustrate the power of multimodal alignment and reasoning for protein function prediction: Prot2Text show that integrating sequence/structure data with LLMs yields richer annotations, while BioReason demonstrates that embedding explicit reasoning leads to more interpretable, accurate predictions. These insights motivate our proposed reasoning-based multimodal framework, which seeks to combine these strengths to further improve the explainability of protein function predictions.

Large Language Models (LLM) have proven their advanced reasoning skills in a variety of domains, the implementation of such systems on complex biological inference problems such as cell type annotation is emerging. The current models of single-cell-specific models do not usually have the sophisticated reasoning of large-scale LLMs. To fill this gap, Cao et al. (2025) [7] created CellReasoner, a small, reasoning-enhanced LLM which has been optimized entirely to be applied to single-cell type annotation. The authors present a small training plan that can effectively mobilize the reasoning abilities of a 7B-parameter LLM with an extraordinarily small dataset with solely 380 quality chain-of-thought exemplars. CellReasoner is an algorithm that is trained to provide a direct mapping between the profiles of cell-level gene expression to targets cell type labels, showing strong results in the zero-shot and few-shot learning setting. The big advantage of such method is that it is interpretable; the model produces expert-level, marker-by-marker reasoning, and it can have structured and verifiable annotations. It is a viable and potentially important solution to smart single-cell analysis and demonstrates that LLM-based thinking could be utilized efficiently to automate and optimize a complicated task in biological classification (Cao et al., 2025) [7].

Although Large Multimodal Models (LMMs) have been demonstrated as problem solvers with a remarkable ability, the aspect of reasoning with complex, domain-specific images, especially in academic research was not investigated sufficiently (Ma et al., 2025) [3]. Despite all their advantages, existing models tend to fail in tasks that entail not only the extraction of visual features but also multiple, rationale-

based inferences on the basis of the visual information represented in science graphs and schemes. This shortcoming points at a serious gap in the possibilities of the contemporary LMMs in addressing real-life academic and scientific problems. Ma et al. (2025) [14] have created SCI-Reason, the first dataset aimed at testing, as well as enhancing, the complex multimodal reasoning capabilities of LMMs with respect to academic fields. From PubMed, the dataset consists of more than 12,000 images and their corresponding question-answer pairs, and all of them are supplemented with a rich chain-of-thought explanatory text that helps the model to identify their meaning. An in-depth analysis of eight prominent LMMs in SCI-Reason indicated that they have major limitations in their ability to reason, where the highest-performing model was averaging out at 55.19 accuracy. More importantly, previous results of the analysis of errors showed that over fifty percent of the failures were associated with malfunctions of the multi-step inference chains and not with errors in basic visual perception. This finding underlines the need to use datasets like SCI-Reason to train and inspire the next generation of LMMs to advance to a higher and more advanced level of reasoning (Ma et al., 2025) [3].

Retrieval Augmented Generation (RAG) has become an effective tool to reduce problems such as hallucination and knowledge cutoff in Large Language Models (LLMs) ground their answers on retrieved data. But almost all of the existing RAG models are constructed to access information in one modality, usually text. It is a major weakness in real world, high-stakes disciplines such as the medical domain where the pertinent information commonly occurs in more than just one form, such as knowledge graphs, clinical notes (text), and complicated molecular structures. It is the capacity to retrieve, reason over, and synthesize such multi, domain-specific information that is critical to the generation of accurate and reliable responses, but development of such multi-modal RAG systems has been negated due to the absence of appropriate datasets. Sengupta et al. (2025) [15] have published a new multi-modal, question-answering dataset, BioMol-MQA on a complicated issue of polypharmacy to brid this critical gap. The dataset is structured as a multimodal knowledge graph which includes text and molecular structure to be retrieved, and a set of difficult questions which serve the purpose of evaluating whether an LLM can reason on the multimodal information. The benchmarks provided by the authors demonstrate that the current LLMs have much difficulty in answering these questions on their own, but their capabilities increase significantly when the background information is offered by the knowledge graph. This underscores the need to leverage the potential of LLMs to its maximum through powerful and multi-modal RAG frameworks to use them in high-stakes and domain-specific tasks and use them more effectively (Sengupta et al., 2025) [15].

ProteinReasoner [8] The author Liu et al. (2025) introduced a multimodal protein language model, which explicitly uses chain-of-thought reasoning. This is the work that is closest to our approach in terms of explicit reasoning. The most important innovation is to use evolutionary profiles using Multiple Sequence Alignments (MSAs) as a form of intermediate reasoning steps. The ProteinReasoner CoT architecture is an expert reasoning framework that uses evolutionary constraints to reason about protein design (inverse folding) and protein design-pertinent evolutionary profile (predicting mutational effects) using the protein structure as input and reasoning

to give output. It also significantly outperforms larger base models, showcasing that explicit reasoning has other advantages than merely being able to scale out model parameters. Also, the model presents in-context learning with protein optimization learning through experimental feedback to provide information on future designs. This feature allows a step-by-step enhancement of it through real-life testing.

Critical observation: ProteinReasoner is the work that is closest to our approach in terms of explicit reasoning. However, there are several key differences to be noted:

1. **Task focus:** ProteinReasoner targets protein design tasks such as, inverse folding and fitness prediction rather than protein function prediction.

2. **Reasoning content:** ProteinReasoner uses evolutionary profiles as intermediate reasoning steps, but our approach leverages bioinformatics tool outputs, including domain annotations, subcellular localization, and topology information, grounded in a curated template bank.

3. **Baseline comparison:** ProteinReasoner does not establish a controlled comparison between reasoning-based and non-reasoning variants on the same task, leaving open whether explicit reasoning itself (rather than other architectural choices) drives the observed improvements.

Our work addresses these gaps by focusing on protein function prediction, integrating diverse and biologically grounded bioinformatics evidence, and establishing explicit baseline comparisons to isolate the impact of reasoning.

2.2.1 Reasoning-Enhanced Biological Models

Several works outside protein function prediction demonstrate that explicit reasoning improves both predictive performance and interpretability across diverse biological domains. These studies provide strong motivation and precedent for our investigation of reasoning-based approaches in protein function annotation.

BioReason [17] Fallahpour et al. (2025) proposed a reasoning-based multimodal framework that tightly couples a DNA sequence encoder (genomic foundation model) with a large language model, enabling direct processing of genomic input with multi-step biological reasoning. Training combines supervised fine-tuning and reinforcement learning, explicitly teaching the model to produce logical, domain-coherent deductions. Importantly, reasoning is not treated as a post-hoc explanation but is integrated directly into the prediction pipeline, allowing the model to learn that improved reasoning leads to improved predictions. On benchmark tasks including KEGG disease pathway prediction and variant effect prediction, BioReason achieved substantial performance gains. KEGG pathway prediction accuracy increased from 88% to 97%, representing a 9 percentage point absolute improvement, while average performance improved by approximately 15% over strong single-modality baselines.

Key implication for our work: BioReason suggests that reasoning supervision enhances biological prediction accuracy, not just interpretability. Though BioReason uses genomic sequences to predict pathways at the level of prediction and does not

use its proteins to do the same instead of annotating their functions, its success is driving our hypothesis that reasoning power can also be used with proteins.

CellReasoner [7]: Cao et al. (2025) introduced a small reasoning-enhanced language model which is optimized for single-cell type prediction. The work has shown that reasoning abilities can be successfully applied to specialized biological problems with low supervision. Notably, a 7B-parameter model achieves strong reasoning performance using only 380 high-quality chain-of-thought exemplers, highlighting the data efficiency of reasoning-based learning. This approach maps gene expression profiles directly to cell type labels and performs strongly in both zero-shot and few-shot settings. Its main strength is in interpretability, certain expert-level, marker-by-marker reasoning is generated by the model, elucidating the use of certain gene expression patterns to predict cities. These systematic descriptions allow the professionals in the domain to judge the validity of prediction. These systematic descriptions allow the professionals in the domain to judge the validity of prediction.

Key implication: From CellReasoner we got to know that reasoning-enhanced models can be both data-efficient and highly interpretable. Markers-by-marker logic is conceptually similar to our domain-wise reasoning framework (R1–R4) that implies that logical biological reasoning enhances predictive power and scientific application.

ProLLM [6]: Jin et al. (2024) developed Protein Chain-of-Thought (ProCoT) that prompts the prediction of protein-protein interaction (PPI). In contrast to classical PPI strategies that consider raw physical interactions as a result of sequence or structural similarity, ProCoT makes its inferences over whole signaling pathways to find indirect functional interactions that are mediated by biological networks. Proteins that are not physically bound by a biological system are common in much of biological systems; however, they can interact indirectly by signaling cascades. Such knowledge on pathways provided by ProCoT is explicitly represented in the form of reasoning chains, allowing the model to follow the interaction pathways within networks.

Key implication: ProLLM shows that learning biological knowledge as explicit reasoning chains rather to having models learn such relationships implicitly, has performance benefits. This supports our strategy of representing bioinformatics evidence as explicit lineages of reasoning.

SCI-Reason [14]: The first dataset to test complex multimodal reasoning in science scenarios was suggested by Ma et al. (2025). The first resolution was SCI-Reason, the first dataset to evaluate the complex multimodal reasoning. Even though the dataset is not protein-specific, it also puts light on the general issues with reasoning in science. SCI-Reason contains more than 12,000 images found in PubMed and annotated with question-answer and detailed chain-of-thought descriptions.Analysis of eight large multimodal models that have been popular showed that there are large reasoning constraints with the best of the models recording only 55.19% accuracy. Analysis of errors demonstrated that over 50% of failures were due to failures in multi-step inferences rather than misjudging visual perception, suggesting that models can frequently extract the correct features and they fail to reason over them effectively.

Key implication: These results illustrate that scientific reasoning is not obtained

by default as a result of prediction-based training and must be made by explicit supervision, which has motivated the integration of the organized reasoning in biological modeling.

Thought Graph [16]: Hsu et al. (2024) proposed a logic model in terms of a graph-form, highly structured thinking representation as opposed to a chain-of-thought process. As a gene set analysis, Thought Graph learns non-linear semantic relationships between biological processes that cannot be well modeled with sequential reasoning alone. Thought Graph outperformed typical Gene Set Enrichment Analysis on the gene set interpretation task by 40.28 points and baseline LLM methods by 5.38 points on the basis of cosine similarity to human annotations. These findings indicate that complex biological relationships can be well captured by structured representations of reasoning.

Key implication: Although our present work uses sequential reasoning traces (R1–R4), Thought Graph indicates that future efforts can use graph reasoning to make biologically inferences involving multiple types of evidence (graph-structured reasoning).

BioMol-MQA [15]: A multimodal question-answering dataset on polypharmacy and drug-drug interactions introduced by Sengupta et al. (2025) are aimed at answering questions. The dataset combines information on text and molecular structure into a multimodal biomedical knowledge graph, which is accompanied by questions, which involve a multi-step reasoning process using various types of evidence. As benchmarking findings demonstrate, large language models are weak at reasoning out of context but much stronger with knowledge graph grounding by means of retrieval-augmented generation.

Key implication: Reasoning which is based on structured biological knowledge plays a large role in improving the performance of the complex biomedical tasks. This underlies our strategy of basing reasoning traces on bioinformatics tool results and on template banks that have been curated instead of being based on parametric model knowledge alone.

2.3 Summary of Key Findings

New studies have shown significant advancements in the multimodal prediction of protein functions in various aspects, such as representation learning, multimodal fusion, generative modeling, interpretability (as well as uncertainty estimation).

Achievements in the Field:

- **Strong sequence representations:** Protein language models such as ESM-1b, ESM-2, ESM-3, and ProtT5 which are trained on millions of protein sequences, provide rich embeddings that capture evolutionary relationships, structural constraints and functional associations. These representations are successfully transferred to diverse downstream prediction tasks.
- **Benefits of multimodal fusion:** The inclusion of sequence data with structural data (e.g., contact maps and three-dimensional coordinates) and textual

annotations continues to have a positive effect on predictive performance. Models like CLEAN-Contact (+12.3% F1), STELLA (+10 BLEU-4 baselines unimodal) and Prot2Text (+16 BLEU when used with ESM-22 embeddings) show high multimodal learning improvements.

- **Natural language generation of function:** The analysis goes beyond discrete classification tags based on Prot2Text, STELLA and ProteinGPT that can generate coherent functional descriptions using protein inputs. These outputs are more functional and full of rich, detailed information.

- **Interpretability through detection models:** The residue-level attention maps displayed by detection-based architectures like ProtDETR highlight functional areas of proteins. While informative but this type of interpretability is different from explicit natural language reasoning that can be used in explaining biological logic.

- **Uncertainty quantification:** Conformal prediction algorithms (e.g. Protein-Vec+conformal-calibration) are also statistically based estimates of confidence which can be used to make principled abstinence in the event of unreliable predictions. This is one of the significant progressions towards credible AI in bioinformatics.

- **Reasoning benefits in related biological domains:** The previous research in genomics (BioReason), cell biology (CellReasoner), and protein interaction modeling (ProLLM) demonstrates that explicit reasoning oversight enhances predictive performance along with explanations.

Critical Research Gap: Despite these advances, there is no existing study systematically evaluates whether explicit reasoning traces improve protein function predictions or not. This gap could be manifests in several ways:

- Generative models like Prot2Text and STELLA have high quality generation but are not trained by a reasoning supervisor, so it is impossible to determine whether explicit reasoning would make predictions better or not.

- ProteinGPT offers interactive question-answer space, but it does not impose logical reasoning chains that are based on the use of molecular evidence in training.

- ProteinReasoner proposes an explicit chain-of-thought architecture, making it gain performance, but on protein design tasks (inverse folding and fitness prediction) not the annotation of functions. The reasoning are based on evolutionary profiles and cannot isolate the effect of reasoning by lack of control of baseline comparisons.

- BioReason and CellReasoner are strong evidence that reasoning enhances the accuracy of biological predictions, but work in other areas (genomics and cell biology), and the applicability of the benefits of reasoning to protein function prediction still remain uninvestigated.

- There is no controlled ablation experiment between identically trained models trained with and without reasoning supervision of protein function prediction. Consequently, the basic questions are still unanswered: does reasoning gives us us more accurate predictions, to what extent and what reasoning forms turn out to be the most effective?

Our Contribution Addressing This Gap: This study directly addresses the identified research gaps through the following contributions:

- **Reasoning dataset construction:** We introduce a pipeline that converts bioinformatics tool outputs, such as, Pfam domain annotations, DeepLoc subcellular localization, DeepTMHMM topology and SignalP signal peptides. By using these tools we introduced structured four-level reasoning traces (R1-R4), grounded in a curated template bank with biological knowledge.

- **Matched baseline establishment:** We have trained a non-reasoning baseline (V1) that predicts functional descriptions as directly from protein embeddings, which acts as a controlled baseline of reasoning augmented models.

- **Demonstration of reasoning improvements:** We show that reasoning-augmented generation (V7) is much better than non-reasoning generation (0.777 to 0.821, the improvement of BERT F1 by 5.7 percent). To the best of our understanding this is the first controlled demonstration that explicit reasoning performs better in predicting protein function predictions.

- **Systematic ablation study:** We test 16 variants of models with varying reasoning structure, order of output and input to understand major design principles:

  - Single integrated reasoning (R4) performs better than chains of exhaustive multi-step reasoning (R1-R4).
  - Reasoning should come before function prediction to have an effective chain-of-thought conditioning.
  - Embeddings of learned proteins do not need any further textual annotations.

- **Confidence-aware reasoning:** Inspired by BioReason and Protein-Vec with conformal prediction we incorporate calibrated confidence scores into reasoning traces, providing uncertainty estimates alongside functional predictions.

Remaining Challenges and Future Directions: Several challenges still remains despite the progress:

- **Limited generalization to remote homologs:** Prediction performance degrades substantially when the sequence similarity drops below 40%. While our homology-based clusters at $\leq 45\%$ identity, so further improvements are needed.

- **Multifunctional protein annotation:** Most existing models predict a single dominant function. Our reasoning framework could be extended to multi-function prediction by generating multiple reasoning chains.

- **Lack of experimental validation:** Predictions are evaluated computationally using held-out test sets; experimental validation remains future work.
- **Incomplete uncertainty calibration:** Although confidence scores are incorporated, calibration remains imperfect (Pearson $r = 0.33$), motivating exploration of advanced calibration techniques.
- **Computational cost:** State-of-the-art models require substantial resources. Our QLoRA-based training strategy enables experimentation on modest hardware, though large-scale training remains resource-intensive.
- **Absence of standardized reasoning benchmarks:** The field lacks dedicated benchmarks for evaluating reasoning quality independently of predictive accuracy. Our LLM-as-a-judge evaluation offers one solution, but community-wide benchmarks are needed.

Table 2.1: Summary of key findings from recent multimodal protein–learning papers

Ref.	**Study**	**Modalities, Datasets & Benchmarks**	**Core Method**	**Key Performance Highlight**	**Limitations**	**Explicit Reasoning?**
[17]	BioReason	DNA sequence + text; KEGG pathways	Genomic foundation model + LLM with SFT/RL	Pathway accuracy: 88% → 97% (+15% on average)	Focused on genomics; not protein-level	Yes
[2]	Prot2Text	Seq + structure; SwissProt 256K	RGCN + ESM-2 encoder with GPT-2 decoder	BLEU-4 = 35.1; ROUGE-L = 48.3	No reasoning traces; RGCN not pretrained	No
[1]	S-PLM	Sequence + contact map; SwissProt, CATH-S40, Kinase, CB513	ESM-2 + Swin-Transformer contrastive alignment; Struct-Adapter + LoRA	F_{max}=0.888 (EC); Acc = 87.48 % (SS)	Still needs structure at train-time; performance varies with tuning	No
[3]	ProtTeX	Sequence + 3D structures (PDB/AF)	Structure-aware tokenization + LLM decoding for editing	Better structure-aware predictions/editing	Engineering heavy; compute-intensive	Implicit
[4]	ProteinGPT	Seq + structure + abstracts; ProteinQA	Two-stage: modality alignment + instruction tuning (ESM-2, ESM-IF1, GVP-GNN)	BERTScore = 0.829; QA Acc = 80 %	Hallucinations; weak on novel structures; no citation links	Partial (Q/A format)

Continued on next page

Table 2.1 continued from previous page

Ref.	Study	Modalities, Datasets & Benchmarks	Core Method	Key Performance Highlight	Limitations	Explicit Reasoning?
[5]	ProtDETR	Sequence-only; SwissProt, New-392	DETR-style Transformer for residue–function detection	Recall = 0.608; F_1=0.594; PRG-AUC = 96.5	Low precision at $\leq 30\%$ identity; no 3-D; limited multimodal fusion	Attention maps
[6]	ProLLM	Sequence + PPI datasets	Chain-of-Thought (ProCoT) prompting/fine-tuning for PPI	Improved PPI accuracy/generalization vs baselines	Quality dependent on ProCoT prompts	Yes (pathway)
[7]	Cell-Reasoner	scRNA expression → cell labels; single-cell benchmarks	LLM fine-tuned with CoT exemplars for annotation	Strong zero/few-shot cell-type annotation	Sensitive to exemplar/domain shift	Yes
[15]	BioMol-MQA	Multimodal KG (drug/protein-/SMILES) QA	Multimodal QA benchmark (retrieval + reasoning)	Exposes LLM failure without multimodal retrievers	Focused on polypharmacy; retrieval challenge	Benchmark only
[8]	Protein-Reasoner	Seq + structure + evolutionary profiles; design benchmarks	Multimodal LLM with CoT-style intermediate reasoning	Gains in inverse-folding/design tasks (zero/few-shot)	Preprint stage; needs replication/wet-lab tests	Yes (evolutionary)
[9]	Protein-Vec+Conformal	Seq + structure; Pfam, UniProt	Risk-controlled retrieval via conformal prediction	39.6 % hit-rate on JCVI Syn3.0 @ FDR 0.1	Needs large calibration sets; not a standalone model	No
[18]	Instruct-Protein	Sequence ↔ text; KG-generated instruction corpus	Instruction-tune LLM on protein+KG instructions	Improved protein↔text generation vs baselines	Synthetic KG instructions; limited wet-lab validation	No
[11]	STELLA	Sequence + structure + text; OPI-Struc	ESM-3 encoder + LLaMA-3.1 decoder; multimodal instruction tuning	BLEU-4 = 43.0; ROUGE-L = 52.6; Acc = 88.85 %	No citation traceability; variable performance on new SwissProt; needs continual learning	No

Continued on next page

Table 2.1 continued from previous page

Ref.	Study	Modalities, Datasets & Benchmarks	Core Method	Key Performance Highlight	Limitations	Explicit Reasoning?
[19]	Structure-Aligned pLM	Seq + structure tokens; 129 k PDB, SaProt	Dual task: latent contrastive + token prediction	P@L/5 = 61.0; Fitness Sp = 0.951; EC F_{max}=0.861	Needs structure access; ↑ structure → ↑ perplexity; static tokens	No
[10]	CLEAN-Contact	Sequence + contact maps; SwissProt, AlphaFold2; New-392, Price-149	ResNet50 (contact maps) + ESM-2 embeddings; contrastive loss	F_1=0.566 (New-392), 0.525 (Price-149) — +16 % P, +18 % R vs CLEAN	2-D maps not full 3-D; hierarchy not modeled; rare ECs hard	No
[12]	ProtET	Seq + bio-text; 67 M pairs	CLIP-style contrastive pre-train + FiLM editing	F_{max}=0.883 (EC); stability ↑ 16.7 %	Needs heavy multimodal annotation; no kinetics	No
[14]	SCI-Reason	Multimodal academic QA (images+text) with CoT rationales	CoT-annotated multimodal benchmark release	Shows current LMMs struggle (low SOTA acc.)	Domain = academic images; annotator variance	Benchmark only
[16]	Thought Graph	Gene-set analysis benchmarks	Produce structured CoT ("thought graphs") + LLM reasoning	Better gene-set interpretation vs LLM/GSEA baselines	Evaluated mainly on gene-set tasks	Yes (graph)

Key observation from this table: Models achieving explicit reasoning (BioReason, CellReasoner, ProLLM, ProteinReasoner, Thought Graph) consistently demonstrate performance improvements in their respective domains. However, none of the protein function prediction models (Prot2Text, STELLA, ProteinGPT) incorporate explicit reasoning with controlled baseline comparisons. This gap—the absence of systematic reasoning evaluation for protein function annotation—is precisely what our work addresses.

Chapter 3

Requirements, Impacts and Constraints

Chapter 3 outlines how our scalable platform predicting protein functions was developed through the merging of protein sequence with functional annotations and reasoning datasets in contrastive and multi-modal learning models. Developed using PyTorch, LLM models and PLM models. The platform produces combined sequence-function embeddings, with reasoning outputs. Its goals include increased accessibility to the study of protein functions, facilitating hypothesis-driven research, decreasing dependence on expensive or time-intensive experiments, and maintaining ethical, reproducibility, and data quality standards. The project planning and risk management are outlined in the chapter and an economic assessment is given including the potential costs and time savings as compared to traditional experimental methods.

3.1 Final Specifications and Requirements

The final specifications of this research comprise the establishment of the multi-modal reasoning based protein function prediction framework with the integration of various sources of data, including UniProt sequences of proteins, Pfam, InterPro, ProSite, DeepLoc, and DeepTMHMM enriched annotations. The original 250K proteins dataset is first clustered with MMseq2 to filter out redundant data leaving a manageable set of 87K entries. Such entries are converted into a reasoning dataset comprising structured question-answer pairs, containing causal reasoning traces (R1-R4), to connect molecular evidence to predicted functions. The system uses a hybrid architecture based on a fine-tuning of Protein Language Models (PLMs) and Large Language Models (LLMs), fine-tuned in two stages on LLaMA-16 and LLaMA 3.1B embeddings, soft prompting, and weighted pooling strategies to improve the quality of predictions. The evaluation is performed by the means of accuracy, BLEU, ROUGE, and BERTScore, as well as ablation studies to understand the effect of architectural decisions. It is a framework that is strongly geared toward reproducibility and transparency and all the datasets, model checkpoints, and code are publicly accessible to facilitate future studies in protein functional prediction.

For local setup:

- OS: Linux Ubuntu

- CPU: AMD Ryzen 9 9950X
- GPU: RTX 5090 32GB
- RAM: 64GB DDR5
- SSD: 1 TB
- HDD: 1 TB

For cloud setup:

- Kaggle
- GPU: 2 T4 GPU (used in parallel)
- Google Colab
- Modal: High-performance AI infrastructure

The framework is scalable, interpretable, and offers a prediction platform that is versatile in predicting protein functions. It facilitates effective model training, and it can be customized to downstream applications like discovering functions with reasoning steps, which makes it appropriate across a broad spectrum of biological research applications.

3.2 Societal Impact

One of the main reasons that ProtReason was developed is to decrease the challenges of predicting interpretable protein functions. There are frequently advanced state-of-the-art models that have significant computational demands and expert knowledge to execute and interpret it. By using our explicit reasoning traces with predictions any researcher without core computational biological knowledge can test the validity of predictions against established domain knowledge.
The framework has a number of applications where interpretable function prediction can be of important value:

Drug Discovery: The role of drug target discovery is in the study of protein function that needs to be understood before making predictions about drug-drug interactions. Reasoning traces are mechanistic hypotheses that may be used to direct experiment validation, where might reduce the search space of identifying the target.
Enzyme Engineering: Explicitly predicting the functionality of enzymes by reasoning about domain contributions and structural properties can be used to rationally design biocatalysts to be used in industrial processes, such as biodegradation and biofuel production.
Functional Genomics: It involves the mass sequencing of protein sequences that do not have known functions. Confidence-calibrated reasoning on proteins produced by automated annotation allows prioritization of proteins to be experimentally characterized.

The structured reasoning trace format (R1 to R4) has an audit trail on every prediction through which a human can trace the model outputs. The outputs of ProtReason can be checked by biologists in terms of their biological plausibility as opposed to black-box classifiers which do not give any underlying reasons. Confidence scores are also used to get extra data on how well the item is predicted, it assists the users in setting the level of confidence they have with the model outputs.

3.3 Environmental Impact

ProtReason was designed with computational efficiency in mind.A number of architectural decisions save a lot of energy as compared to naive end-to-end implementations. The first component is that the ESM-2 encoder is frozen during training, which removes backpropagation using about 3 billion parameters. Computational cost gets reduced by 80% by our fully fine tuned architecture.

Secondly, aggressive quantization is used where we are using 8-bit precision for the encoder and 4-bit precision for the language model, respectively. This reduces considerably bandwidth needs on memory and allows a consumer grade GPUs to be trained instead of special data center processors. Also, QLoRA fine-tuning requires updating the language model with 0.25% of the parameters, which additionally minimizes the amount of memory and the number of floating point operations during training.

All of these efficiency measures allow training with NVIDIA T4 GPUs (70 W thermal design power) rather than NVIDIA A100 GPUs (400 W thermal design power), which has a six fold reduction in power usage for training time.

ProtReason not only saves computing resources but also enhances protein characterization workflows. Typically, experimental biochemical assays require a range of between $500 and $5,000 per protein and have a high cost in terms of laboratory needs, such as reagents, energy, and researcher time. ProtReason provided confidence score accompanied by reasoning traces which can high identical probability candidates and ignore unlikely matches. As a result it successfully reduces the total number of experimental phases that is required by any large scale protein characterization process.

To clear things up - our experimental validation is not being crossed out by computational predictions in high level work applications. Rather, ProtReason is a intermediate-scale filter that can be used to narrow experimental resources into the most promising applications and enhances the economic and environmental sustainability of protein discovery functional applications

3.4 Ethical Issues

The Research employs publicly available protein data only and is based on FAIR (Findable, Accessible, Interoperable, Reusable) principles and data does not imply any human or animal experimentation. Ethical issues are mostly relevant to the ultimate

utilization of proteins, and it is necessary to regulate the biosafety and environmental issues. The model promotes transparency through offering a rational method of each prediction through clear phases that enabled the form of human control and verification. No confidential, proprietary, and sensitive data and information are utilized. The interpretability and open measures of evaluation in the system make scientific research fair and accountable.

3.5 Standards

The study follows the established practices of computational biology, machine learning and AI modeling. Tabular representations containing protein sequences and annotations are done in standard tabular representations. Several data resources like UniProt, Pfam and InterPro meet international bioinformatics requirements. It is implemented using the best practices in model training, such as dataset normalization, controlled Cafa splits, and reproducible benchmarking. Our system provides support for the integration of many LLM libraries and pipelines of bioinformatics for future labwork and to ensure full transparency.

3.6 Project Management Plan

We carried our work in certain steps in a fixed iterative ways. It began with pre processing protein sequences which were annotated by Pfam, InterPro and AlphaFold tools. We organized the data in standardized way consisting verified labels and deterministic splits. Good base models are then trained to determine reference performance and computations costs.

Across many modalities our multi feature dataset benchmarking quantified accuracy and AUC. Interpretability research overcomes failures and success, and explainability methods justify model transparency. Cross-dataset tests: These tests are tested on a related dataset after being trained on a different dataset.

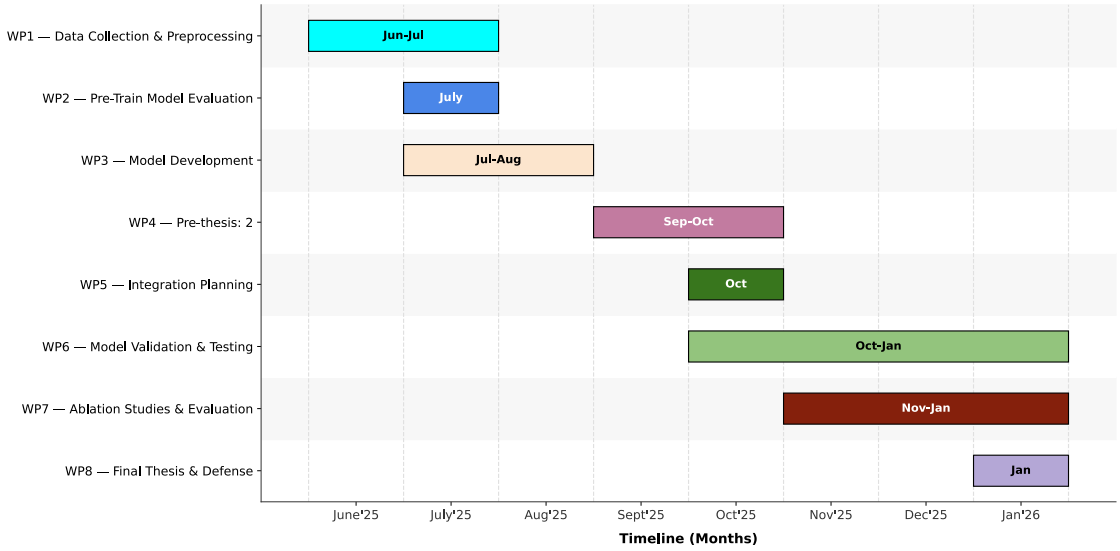


Figure 3.1: Project timeline using Gantt chart

The final stage consolidates results and narrative into the thesis and a reproducibility bundle. Decision gates after ablation and benchmarking ensure only the best configurations proceed to comprehensive testing. All stages are depicted in the project timeline Gantt chart.

3.7 Risk Management

We identified several risks and implemented mitigations:

Data Quality Risk: UniProt annotations may contain inconsistencies. Mitigation: Template-grounded reasoning generation with evidence whitelisting; validation module checking trace consistency.

Computational Resource Risk: Training may exceed available GPU memory. Mitigation: Quantization, gradient checkpointing, and cloud computing flexibility.

Negative Result Risk: Reasoning can be unsatisfactory on the basis of areas of reasoning. Mitigation: This would in itself be a useful finding; systematic ablation would determine what benefits would be offered.

3.8 Economic Analysis

The reasoning-based multimodal system proposed will aid in significantly lowering the research expenses by substituting the costly wet-tail validations with a computational prediction that may be used as a preliminary screening process. Licensing and costs of software licensing is reduced through the utilisation of open medical datasets and open-source software infrastructure such as PyTorch and Hugging Face.

Moreover, lightweight training plans allow efficient model construction with using small GU programs and have much fewer infrastructure and energy demands. In comparison to conventional experimental testing pipelines, the framework reduces cost overall, and allows affordable and rapid experimentation on small laboratories and academic institutions.

Our system encourages broader participation in data driven biological research by providing a low cost and low computation cost pathway for functional protein prediction with reasoning traces. This is how it breaks the chain of command for the need of expensive and high computational approaches.

Chapter 4

Proposed Methodology

The framework incorporates causal reasoning through advanced models such as GPT-OSS 20B and LLaMA, enabling the generation of reasoning traces to justify the predictions. Evaluation of the model is carried out using a range of metrics, with ablation studies to assess the impact of different architectural choices. This methodology offers a more comprehensive and interpretable approach to protein function prediction.

4.1 Methodology Overview

The main aim of the study is to come up with a framework of protein functions prediction based on reasoning. The usual approaches are based on sequence homology or the classification labels, which are restricted with novel or divergent proteins. The other solution that is suggested in this research is the integration of reasoning data to come up with more sensible predictions that are more correct.
The methodology starts with Prot2Text[2] dataset which contains 250K protein sequences. Pfam, DeepTMHMM, SignalP, and DeepLoc functional annotations are added to this dataset. Such heterogeneous data inputs bring up well-rounded representation of every protein and this is essential in precise function prediction.
In order to optimize the dataset to be used in training, the data is clustered with the MMseqs2, which reduced the size of the dataset by 250K to 87K entries. This kind of clustering can ensure a diverse and the same time a manageable dataset which prevents overfitting while protecting and keeping important biological information intact. For every protein, there is an evidence pack presents which consists of sequence data, domain annotations, and structural information. This evidence pack is put into GPT-OSS 20B that generates reasoning traces (R1 - R4) to provide justifications for the protein function predictions and it's confidence score is indicating the model's certainty correctly.
This model is trained in two stages. The initial step is the MSSP-128 model, which uses a Qformer-based projector based on LLaMA-16 embeddings. The weighted pooling strategies (mean, min, max and std) are used to process these embeddings in order to capture a rich set of features. At the second stage the model is refined with LLaMA 3.1 (8B), and soft prompting and reasoning-based outputs are implemented in improving the accuracy of the predictions made by the function.
The model is evaluated by a variety of measures, such as accuracy BLEU, ROUGE and BERTScore and LLM as a judge. There are also, ablation studies that are

conducted to determine the effect of various architectural decisions, including use of reasoning traces in differnt settings on model performance.

4.2 Dataset Overview

This research uses two sets of dataset for training and evaluation. They are Prot2Text [2] dataset and our novel dataset generated for this study.

4.2.1 Prot2Text Dataset

Our model uses the Prot2Text [2] dataset as its main foundation. This is a publicly available dataset of 250,000 protein sequences in UniProt each having functional annotations and text descriptions attached to them. The annotations give hints into the biological functions, families, and domains of the proteins rendering it an important resource to the protein function prediction tasks.

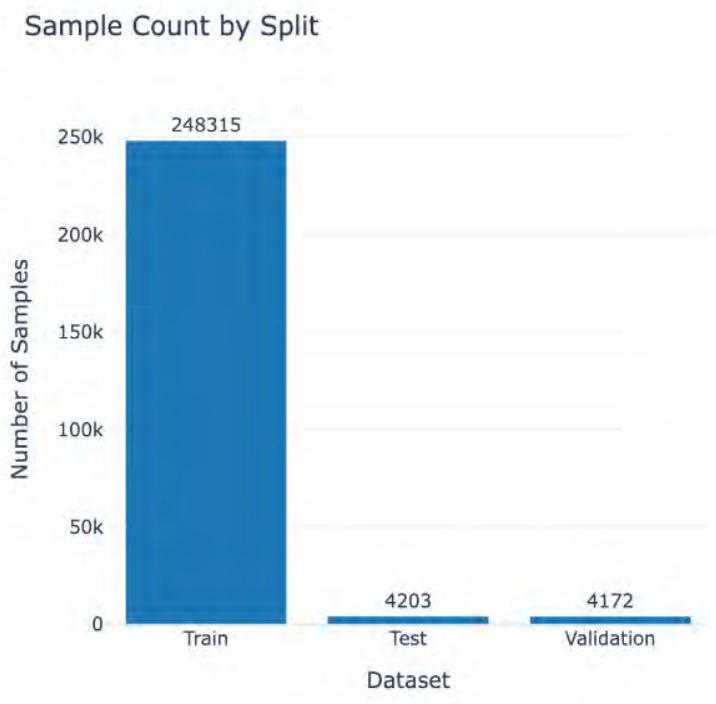


Figure 4.1: Prot2Text Samples

The dataset has been extensively utilized to predict functionalities of proteins and especially the broad coverage of various proteins has made it useful. Its drawbacks, however, are that it offers a wide range of annotations and does not provide much in-depth contextual information on protein structure and interactions.

4.2.2 Custom Dataset

We obtain our own custom 87k dataset that is optimized to work with our downstream task, together with Prot2Text retrive from uniprot, we used MMseqs2 clustering algorithm to cluster proteins using their sequence similarity. Sampling of this clusteres was then done.

Algorithm: Balanced Protein Subsampling

1: **Cluster-Based Initial Sampling:** For each cluster $c \in C$, select k members into subset S using a size-dependent quota:

- $k = 1$ (size ≤ 3), $k = 2$ (size ≤ 7), $k = 3$ (size ≤ 15), or $k = 4$ (size > 15).

2: **Targeted Feature Enrichment:** For each feature $f \in \{TM, SP\}$, if the current ratio $r_f < \text{Target}_f$:

- Calculate the deficit Δ_f needed to reach the target ratio.
- Sample Δ_f unique proteins possessing feature f from the remaining pool and add them to S.

By using this algorithm, constructs an unredundant protein set by sampling a representative set of sequence clusters and then injecting to obtain special biological ratio specifications.

Instead of setting a fixed number, we let the algorithm decide the dataset size. We ended up with 87,384 rows. We split our data based on a homology based split.

4.2.3 Reasoning Dataset Construction

Each protein entry in our dataset is associated with an evidence pack consisting of sequence cutouts, structural metadata, enrichment from uniprot and biological tool results. The evidence pack aggregates outputs from multiple computational tools: Pfam/InterPro for domain annotations, DeepTMHMM for transmembrane topology prediction, SignalP for signal peptide detection, and DeepLoc for subcellular localization prediction.We also provide our model with set of templetes for GPT-OSS 20B processes these inputs to produce structured reasoning traces (R1–R4). To ground the model, we provide a template bank of normalized literature tags (e.g., "7-transmembrane-helix topology defines GPCR architecture"), encouraging the LLM to connect raw evidence with known biological processes. Finally, a validator module compares the generated output against permissible evidence items to ensure consistency and reduce hallucinations.

Template Bank for Biologically Grounded Reasoning To guide the reasoning generation process and ensure biological accuracy, we construct a **template bank** normalized biological knowledge statements derived from established literature. These templates encode domain-specific knowledge that connects computational tool outputs to their functional implications, preventing the language model from generating biologically implausible explanations.

The template bank covers five categories of biological knowledge:

1. **Membrane Topology Templates (DeepTMHMM):** Describe the functional implications of predicted transmembrane architectures.

2. **Signal Peptide Templates (SignalP):** Explain targeting pathway implications of different signal peptide types.

3. **Subcellular Localization Templates (DeepLoc):** It connects predicted localization to biological processes that are occurring.

4. **Membrane Association Templates:** Differantiates between integral, peripheral and lipid-anchored membrane proteins.

5. **Targeting Signal Templates:** Report on the mechanisms of different targeting sequences.

To prevent hallucination and ensure traceability, we implement an **AllowedEvidenceItems** constraint system. For each protein, we programmatically extract a whitelist of permissible evidence citations. Then the reasoning model is advised to give references to items in this whitelist in the evidence arrays only, so that all its claims are based on real computational predictions and not on fabricated evidence.

Reasoning Trace Generation

We employ GPT-OSS-20B [20] to generate structured reasoning traces from the evidence packs. Manufacturing the process of generation employs a specially shaped system prompt that impose a number of quality restrictions:

1. **Mechanistic Explanation Requirement**: Not only do the reasoning traces assert consistency, each reasoning trace must explain *how* and *why* the evidence is related to the functionality. Circular phrases such as "consistent with the label" or "supports the annotation" are explicitly forbidden.

2. **Evidence Grounding**: All claims must reference specific evidence items from the AllowedEvidenceItems whitelist.

3. **Contradiction Acknowledgment**: When evidence conflicts with the functional annotation (e.g., cytoplasmic localization for a predicted secreted protein), the model must explicitly state the contradiction.

4. **Confidence Calibration**: Confidence scores must reflect evidence strength

Table 4.1: Representative examples from the literature-grounded template bank. Each template encodes domain knowledge connecting tool predictions to biological mechanisms.

Template ID	**Biological Knowledge Statement**
Membrane Topology (DeepTMHMM)	
DEEPTMHMM_TYPE_TM	DeepTMHMM predicts an α-helical transmembrane protein; membrane-spanning α-helices are typically $\sim$20 hydrophobic residues long, consistent with crossing the lipid bilayer core.
DEEPTMHMM_TMR_EQ_7	A seven-transmembrane-helix topology is the defining structural hallmark of G protein-coupled receptors (GPCRs) in eukaryotes.
DEEPTMHMM_TMR_EQ_12	A twelve-transmembrane-helix topology is characteristic of many secondary transporters, including members of the major facilitator superfamily (MFS).
Signal Peptides (SignalP)	
SIGNALP_SP	SignalP predicts a Sec/SPI signal peptide, which is typically cleaved by signal peptidase I during protein secretion.
SIGNALP_LIPO	SignalP predicts a lipoprotein signal peptide (Sec/SPII, LIPO) that leads to lipid anchoring after cleavage by signal peptidase II.
SIGNALP_TAT	SignalP predicts a Tat/SPI signal peptide, indicating export via the twin-arginine translocation (Tat) pathway.
Subcellular Localization (DeepLoc)	
DEEPLOC_LOC_CELL_MEMBRANE	DeepLoc predicts cell membrane localization; plasma membrane proteins frequently function in transport, signaling, or cell–environment interactions.
DEEPLOC_LOC_MITOCHONDRION	DeepLoc predicts mitochondrial localization; mitochondria are primary sites of cellular energy metabolism, including oxidative phosphorylation.
DEEPLOC_LOC_NUCLEUS	DeepLoc predicts nuclear localization; nuclear proteins are involved in DNA/RNA processing, transcriptional regulation, and genome maintenance.
Targeting Signals	
DEEPLOC_SIG_NUCLEAR_LOCALIZATION_SIGNAL	A nuclear localization signal (NLS) is a short peptide motif that mediates active import into the nucleus through importin-dependent pathways.
DEEPLOC_SIG_MITOCHONDRIAL_TRANSIT_PEPTIDE	Mitochondrial targeting sequences (transit peptides) are typically N-terminal, positively charged, and form amphipathic helices recognized by the mitochondrial import machinery.

Reasoning traces (R1–R4):

R1 (Functional Reasoning): Derives the molecular and catalytic function of a protein primarily from common domain annotations (e.g., Pfam/InterPro). This trace explains the predicted mechanism of action and biological role based on domain-level evidence with fuctional data.

R2 (Localization Reasoning): Describes the cellular or subcellular compartment in which the protein operates using localization models. This reasoning provides biological context for function and includes an associated confidence score reflecting prediction reliability.

R3 (Structural Reasoning): Describes the protein's structural architecture depending on domains and transmembrane topology, as inferred from sequence-based structure predictors. These features explain how the protein's architecture supports its predicted function.

R4 (Confidence Summary): Combines the confidence from functional, localization, and structural reasoning traces to provide an overall reliability assessment of the annotation. High confidence indicates consistency among multiple evidence sources.

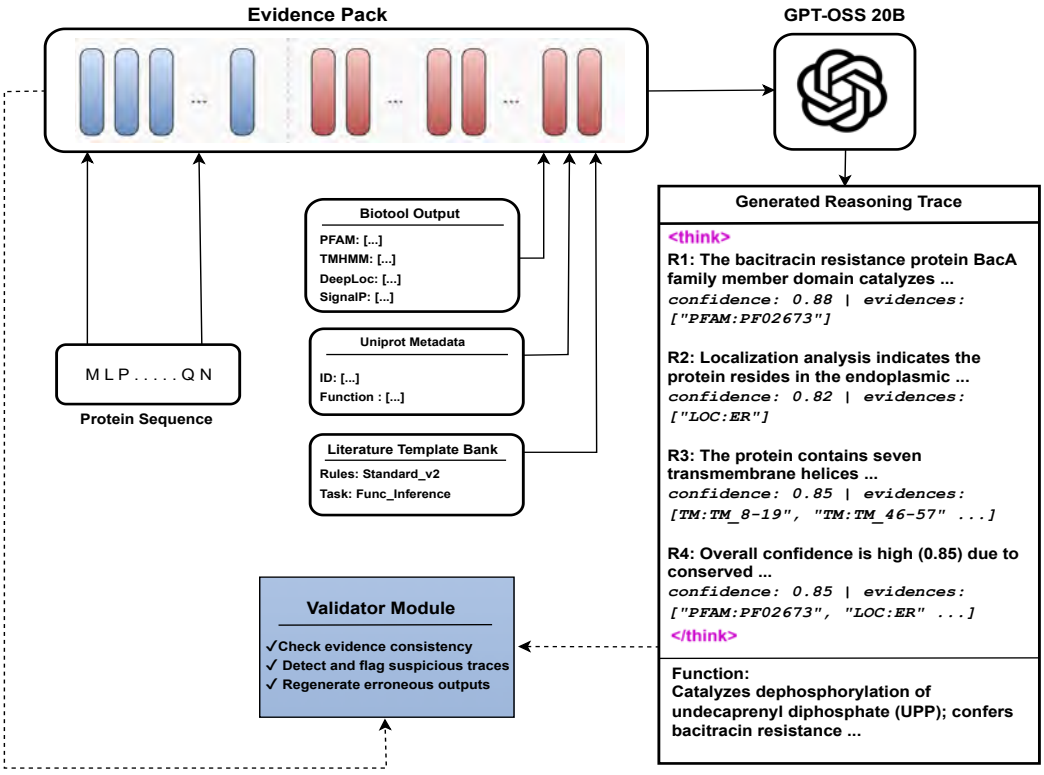


Figure 4.2: Protein Reasoning Trace Generation Pipeline

Table 4.3: Evidence Summary for BacA Family Protein

Component	Content
Input – (Evidence pack)	**Protein ID:** Q0SNQ3
	Protein sequence: MINILNAIILGIVQGITEFLPVSSS-GHLLLFRHFIL. . . (264)
	Organism: *Borrelia afzelii*
	Evidence sources: Pfam, DeepLoc, DeepTMHMM, SignalP, UniProt metadata
Reasoning (R1 – Functional Evidence)	**R1 Functional Anchor:** The bacitracin resistance protein BacA family member domain is known.
	Confidence: 0.88
	Evidence: PFAM:PF02673
Reasoning (R2 – Localization Evidence)	**R2 Localization Inference:** Localization analysis indicates the protein resides in the endoplasmic reticulum.
	Confidence: 0.82
	Evidence: LOC:ER
Reasoning (R3 – Structural Evidence)	**R3 Structural Reasoning:** The protein contains seven transmembrane helices.
	Confidence: 0.85
	Evidence: TM annotations
Reasoning (R4 – Evidence Synthesis & Validation)	**R4 Confidence Synthesis:** Overall confidence is high (0.85) due to conserved BacA domain and supporting localization and TM evidence.
	Confidence: 0.85
	Evidence: PFAM:PF0267, LOC:ER, TM annotation
Function	Catalyzes dephosphorylation of undecaprenyl diphosphate (UPP); confers bacitracin resistance

4.2.4 Statistics Table and Dataset Distribution

The following tables summarize statistics for the reasoning levels, including text length and confidence scores across reasoning categories (R1–R4). Additionally, metrics are broken down by dataset split (training, test, and validation) to show the average text length and confidence distribution within each subset.

Table 4.5: Length statistics across reasoning levels

Reasoning Level	Avg Length	Median Length	Min Length	Max Length	Std Dev
R1 (Functional Domain)	123.23	122.0	28	283	25.91
R2 (Localization)	132.68	127.0	51	305	30.49
R3 (Structure)	132.98	121.0	45	525	41.05
R4 (Confidence)	159.96	155.0	77	330	32.43

Table 4.6: Confidence statistics across reasoning levels

Reasoning Level	Avg Confidence	Median	Min	Max	Std Dev
R1 (Functional Domain)	0.8121	0.85	0.00	0.92	0.1561
R2 (Localization)	0.6719	0.75	0.20	0.92	0.1815
R3 (Structure)	0.7713	0.80	0.10	0.95	0.1284
R4 (Confidence)	0.7142	0.78	0.15	0.93	0.1548

Table 4.7: Metrics by dataset split (train/test/validation)

Split	Count	Avg Text Len	Avg Confidence
Train	81195	139.04	0.74
Test	3065	138.31	0.73
Eval	3124	138.10	0.74

4.3 Architecture Design

In order to overcome the modality gap between high-dimensional protein sequence embeddings and the semantic space of Large Language Models (LLMs), we introduce the ProtReason Architecture. The system is pretrained at two phases (1) Representation Alignment through Multi-Scale Perceiver Pooling (MSPP-128) network and (2) Generative Fine-tuning on LLaMA 3.1 8B with Soft Prompting and Low-Rank Adaptation (LoRA).

4.3.1 Stage 1: Multi-Scale Perceiver Pooling (MSPP-128)

The primary challenge in protein-to-text generation is compressing variable-length protein sequences ($L \approx 50$ to 2000+ residues) into a fixed-length, semantically rich representation compatible with LLM context windows. We introduce MSPP-128, a specialized Q-Former architecture that extracts hierarchical features ranging from local active sites to global tertiary structures into a fixed set of $K = 128$ prefix tokens.

A. Input Representation and Pre-projection: Let a protein sequence S of length L be processed by a ESM-2 (3B) encoder. The output is a sequence of hidden states $H_{\text{esm}} \in \mathbb{R}^{L \times d_{\text{esm}}}$, where $d_{\text{esm}} = 2560$. To align dimensions for the Perceiver, we apply a learnable multi-layer perceptron (MLP) pre-projection with layer normalization:

$$H = \text{LayerNorm}(\text{GELU}(\text{LayerNorm}(H_{\text{esm}})W_1 + b_1)\, W_2 + b_2)$$

where $H \in \mathbb{R}^{L \times d_{\text{model}}}$ and $d_{\text{model}} = 2048$. The input layer normalization stabilizes the ESM-2 outputs, while the output layer normalization facilitates training of the subsequent attention layers.

B. Multi-Scale Feature Extraction: Protein function is encoded at multiple scales. Such as, single residues (active sites), motifs (secondary structure) and domains (tertiary structure). To capture this, we generate three views of the input embeddings using distinct pooling kernels:

- **Fine-Grained View (H_{fine}):** The full-resolution sequence (H).
- **Mid-Level View (H_{mid}):** Mean pooling with kernel size $k = 8$.
- **Coarse-Grained View (H_{coarse}):** Mean pooling with kernel size $k = 32$.

For a pooling window size k, the pooled representation h'_i for the i-th chunk is calculated as

$$h'_i = \frac{1}{|M_i|} \sum_{j \in M_i} h_j$$

where M_i denotes the set of valid (non-padded) indices in the i-th window. This hierarchical input allows the attention mechanism to attend to domain-level features in H_{coarse} while retaining residue-level precision from H_{fine}.

C. Perceiver Resampler with Gated Cross-Attention: The core of MSPP-128 is a Perceiver Resampler with $N = 4$ layers and $n_{\text{heads}} = 16$ attention heads. We initialize $K = 128$ learnable latent queries $\mathbf{Q} \in \mathbb{R}^{K \times d_{\text{model}}}$, augmented with learnable positional encodings $\mathbf{P}_{\text{pos}}$. Each layer l performs iterative cross-attention to the multi-scale features followed by self-attention.

To improve training stability, we employ gated cross-attention, where the contribution of each attention block is modulated by a learnable scalar gate parameter α, initialized to zero. For a given input view $\mathbf{V} \in \{\mathbf{H}_{\text{fine}}, \mathbf{H}_{\text{mid}}, \mathbf{H}_{\text{coarse}}\}$, the gated cross-attention update is defined as

$$\mathbf{Q}^{(l+1)} = \mathbf{Q}^{(l)} + \alpha \cdot \text{CrossAttn}\big(\mathbf{Q}^{(l)},\ \mathbf{V},\ \mathbf{V}\big)\,, \tag{4.1}$$

where CrossAttn$(\cdot)$ denotes standard multi-head attention with the latent queries as queries and the multi-scale protein features as keys and values. The gate parameter α is learned during training, allowing the model to gradually incorporate cross-modal information while avoiding early optimization instability.

$$Q' = Q + \tanh(\alpha) \cdot \text{MultiHeadAttn}(Q, V, V)$$
$$Q_{\text{out}} = Q' + \tanh(\beta) \cdot \text{FFN}(\text{LayerNorm}(Q'))$$

where β is an additional learnable gate parameter. By initializing $\alpha, \beta = 0$, the hidden layer acts as an identity function at the start of training, allowing the model to progressively learn to extract information from the protein encoder without destabilizing gradients.

D. Loss Function (Hard Negative Mining & Asymmetric Weighting:) To prevent mode collapse and encourage diverse token utilization, we incorporate auxiliary regularizers:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{NCE}} + \lambda_2 \mathcal{L}_{\text{diversity}} + \lambda_3 \mathcal{L}_{\text{consistency}}$$

Here, $\mathcal{L}_{\text{diversity}}$ penalizes cosine similarity between the K distinct queries so they attend to different protein regions, and $\mathcal{L}_{\text{consistency}}$ enforces that statistical views (mean, max, std) of the protein embeddings maintain consistent geometric relationships with the teacher embeddings. After epoch 4, we introduce small regularization loss like vicreg to help the model generalize better.

4.3.2 Stage 2: Generative Fine-Tuning with Soft Prompting

In the second stage, the pre-trained MSPP-128 projector is frozen, and we fine-tune the LLaMA 3.1 8B model to generate functional descriptions and reasoning traces. We utilize soft prompting combined with QLoRA (Quantized Low-Rank Adaptation) to achieve high performance with efficient memory usage.

A. Soft Prompt Injection: Unlike standard LLM fine-tuning which uses discrete text tokens, we inject the continuous embeddings P generated by MSPP-128 directly into the LLM's embedding space. Let the input protein be x_{prot} and the target text sequence (reasoning trace + function) be $Y = \{y_1, y_2, \ldots, y_T\}$. The input sequence to the LLM is constructed as a concatenation of the soft prefix P and the token embeddings of the text instructions E_{inst}:

$$\text{Input}_{\text{LLM}} = [P_1, P_2, \ldots, P_{128}, E_{\text{inst}}, E_{\text{BOS}}].$$

This effectively prompts the LLM with the structural context of the protein, allowing it to attend to specific protein regions (encoded in the 128 tokens) when generating text.

B. QLoRA Implementation (4-bit Quantization): To fine-tune the 8-billion parameter LLaMA model efficiently, we employ 4-bit quantization on the base model weights W_0. The weights are stored using the NormalFloat4 (NF4) data type. During the forward pass, the quantized weights are dequantized to BF16. We inject trainable low-rank adaptation matrices A and B into the attention and feed-forward layers:

$$W_{\text{new}} = \text{dequant}(W_0) + \frac{\alpha}{r} BA.$$

Here, $B \in \mathbb{R}^{d_{\text{in}} \times r}$ and $A \in \mathbb{R}^{r \times d_{\text{out}}}$ are low-rank matrices with rank $r = 16$, while W_0 remains frozen. Only A and B are updated via gradient descent.

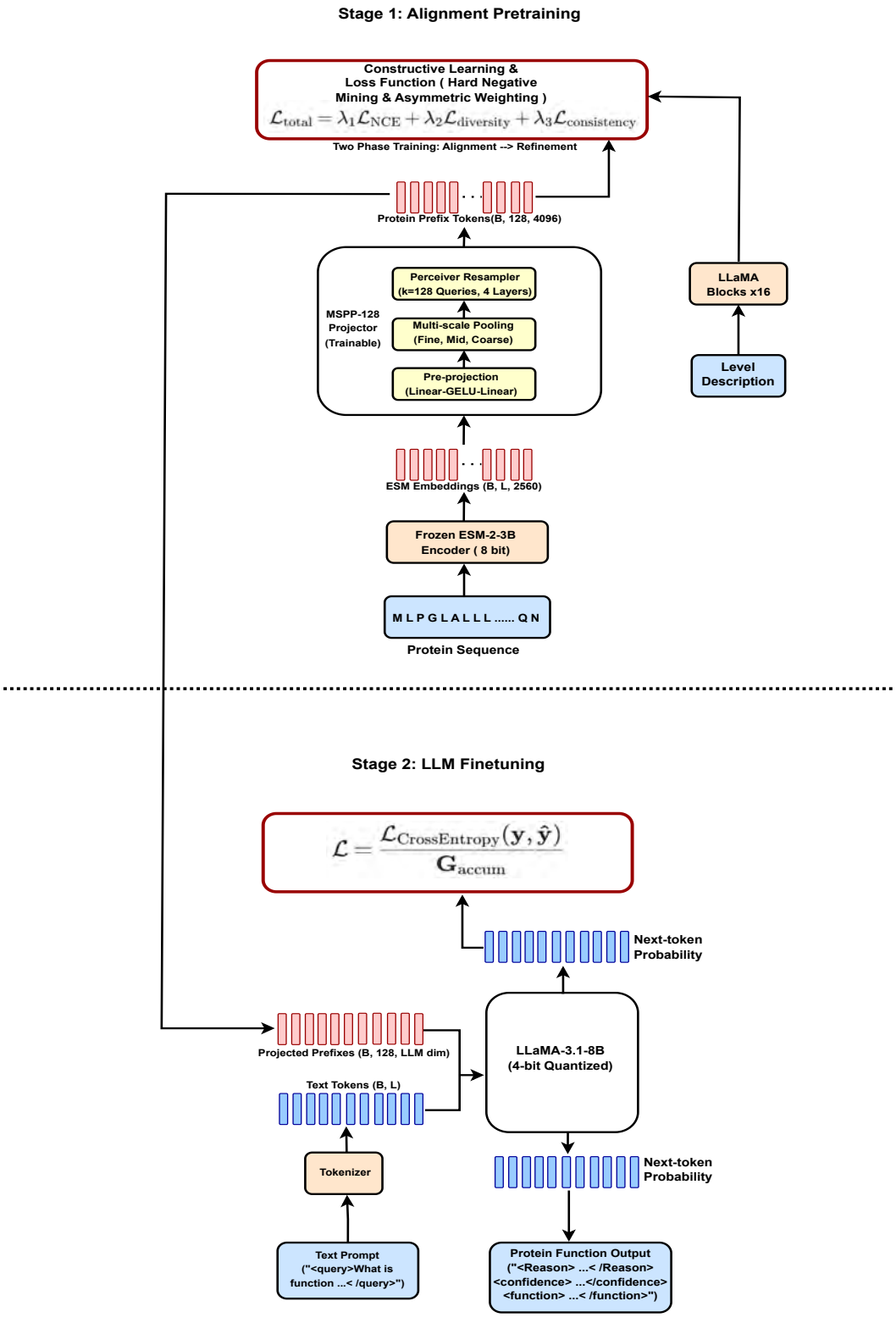


Figure 4.3: Architecture Diagram of proposed ProtReason

C. Causal Language Modeling Objective: The model is trained to maximize the likelihood of the reasoning trace (R) and the functional description (Y) given the protein prefix P. The loss function is the standard causal language modeling (CLM) loss:

$$\mathcal{L} = \frac{\mathcal{L}\text{CrossEntropy}(\mathbf{y}, \hat{\mathbf{y}})}{G\text{accum}}$$

4.4 Implementation of Selected Design

Stage 1 uses a contrastive learning framework with a batch size of 96, and gradient accumulation steps are used to maintain an effective batch size of 768. Stage 2 fine-tunes the LoRA-augmented Causal LLM with a batch size of 128 and a learning rate of 2e-4. The learning rate scheduler is employed with a warm-up ratio of 0.03 to stabilize the fine-tuning process.

4.4.1 Stage 1: Representation Alignment

Stage 1 of the model focuses on representation alignment, where a frozen ESM2-3B encoder is employed along with a multi-scale Perceiver Resampler to encode protein features into 128 prefix tokens. To ensure robust mapping of protein information to text, the model is trained using assymetric InfoNCE loss and hard negative mining. The model also uses token-level contrastive loss to directly optimize prefix tokens rather than pooled statistics.

Hyperparameters for Stage 1:

- **Batch size**: 96
- **Effective batch size**: 768
- **Learning rate**: 2e-4
- **Number of epochs**: 10
- **Gradient clipping**: 1.0
- **Warmup ratio**: 0.06
- **Contrastive temperature**: 0.12
- **Curriculum stages**: (256, 512, 1024) varying max length throughout training
- **Hard negative weight**: 0.5
- **Trainable Parameters:** $\sim$857.7 M

4.4.2 Stage 2: Structured Output Generation and Fine-Tuning

In stage 2, The model uses the compatible embeddings produced in Stage 1 by injecting them into a LoRA-augmented Causal LLM as soft prefixes. The fine-tuning step conditions the model to produce structured outputs like reasoning traces, confidence scores and functional descriptions with a custom XML prompt schema. This is the step that combines the reasoning based output with the learned embeddings so that the model can conduct an in-depth and understandable analysis of protein functions.

Hyperparameters for Stage 2:

- **Batch size**: 4
- **Gradient Acum**: 8
- **Effective Batch size**: 32
- **Learning rate**: 2e-4
- **Number of epochs**: 10
- **LoRA scaling factor**: 16
- **LoRA dropout**: 0.05
- **Warm-up ratio**: 0.03
- **Max Sequence Length**: 512 tokens

Figure 4.4: Training Graph

Chapter 5

Result Analysis

This chapter presents the experimental evaluation of ProtReason. We first describe the evaluation methodology, including alignment assessment and generation metrics. We then present the systematic ablation study across 16 model variants, identifying key design principles. Finally, we compare ProtReason against established baselines and analyze reasoning quality and confidence calibration.

5.1 Performance Evaluation

This chapter revolves around the benchmark evaluation in which the protein sequence embeddings are combined with text descriptions in order to enhance protein function prediction. The chapter has employed retrieval-type metrics including protein-to-text (P2T) and text-to-protein (T2P) accuracy and recall to determine alignment between protein and text embeddings. Assessment of the projector on Prot2Text dataset , demonstrates the usefulness of the projector in matching protein sequences with functional descriptions. The findings give ideas on its strong and weak aspects, which can give a better perspective of its prospects in predicting protein functions. Recent studies have shown that the multi-modal alignment between textual annotations and sequence data is important in the evaluation of the data, which highly improves the accuracy and interpretability of the predictions. The method uses contrastive learning to inject alignment so that protein sequences have some sense of connection to descriptions of their functions. The stage-based training approach with MSPP-128 as the feature extractor and LLaMA 3.1 as the fine-tuning model shows how reasoning-based feature embedding can be used to achieve a higher accuracy of retrieval and better understandability of the predictions made by the functions. It is also highlighted in the findings that additional model optimization is necessary to provide the ability to generalize to a wide variety of protein functions to introduce more scalable and interpretable solutions in protein function annotation.

5.1.1 Projector Alignment (Retrieval-Style Evaluation)

We evaluated alignment on benchmark test sets (Prot2Text). Although MSPP-128 is not explicitly trained as a retrieval model, we use a retrieval-style protocol to quantify how well protein and text embeddings are aligned. Concretely, for protein-to-text (P2T), we embed each protein and rank all candidate text descriptions by cosine similarity; accuracy (Acc) reports top-1 retrieval, and Recall@20 (R@20) reports whether the correct description appears among the top 20 results. We perform the analogous evaluation for text-to-protein (T2P) retrieval. We report both in-batch results (where candidates are restricted to the current batch) and full test-set results (where each query is ranked against all items in the test set), to highlight the gap between easy and hard retrieval settings.

5.1.2 Mathematical Formulation of the Alignment Loss:

The output of the Perceiver is projected to the LLM dimension $d_{\text{llm}} = 4096$, yielding the prefix tokens $P \in \mathbb{R}^{128\times4096}$. Stage 1 training minimizes a composite loss designed to align these tokens with the text embeddings of a frozen teacher model (LLaMA). We employ a symmetric InfoNCE loss with hard negative mining. Let p be the pooled protein representation and t be the corresponding text embedding. The contrastive loss for a batch size B is

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{B}\sum_{i=1}^{B}\left(\log\frac{e^{\text{sim}(p_i,t_i)/\tau}}{\sum_{j=1}^{B} w_{ij}\cdot e^{\text{sim}(p_i,t_j)/\tau}}\right)$$

where $\text{sim}(u, v)$ is cosine similarity, τ is a learnable temperature parameter clamped to $[0.07, 0.5]$, and w_{ij} is the hard negative weight (with $w_{ij} > 1$ for semantically close negatives).

To prevent mode collapse and encourage diverse token utilization, we incorporate auxiliary regularizers:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{NCE}} + \lambda_2\mathcal{L}_{\text{diversity}} + \lambda_3\mathcal{L}_{\text{consistency}}$$

Here, $\mathcal{L}_{\text{diversity}}$ penalizes cosine similarity between the K distinct queries so they attend to different protein regions, and $\mathcal{L}_{\text{consistency}}$ enforces that statistical views (mean, max, std) of the protein embeddings maintain consistent geometric relationships with the teacher embeddings. After epoch 4, we introduce small regularization loss like vicreg to help the model generalize better.

5.2 Analysis of Design Solutions

5.2.1 Evaluation Metrics

To measure the performance of the model, we applied a wide variety of automatic measures to assess the quality of the generated functional annotations in a comprehensive manner. We seelected each measure in order to reflect different feature of the output, that ranged from strict agreement to the semantic coherence.

1. **BLEU:**
    - It uses n-gram precision to calculate local sequence similarity.
    - It helps in proving that correct local wording and ordering are protected by predicted descriptions.
2. **ROUGE-Based Metrics (ROUGE-1, ROUGE-2, ROUGE-L):**
    - **ROUGE-1:** It refers to unigram overlap (basic content coverage).
    - **ROUGE-2:** It refers to bigram overlap (local phrase consistency).
    - **ROUGE-L:** It refers to the longest common subsequence (LCS) similarity (structural/ordering consistency).
3. **BERT-Based Precision, Recall, and F1:**
    - It measures semantic similarity aacurately using contextualized representations (for example - BioBERT style embeddings).
    - It evaluates meaning level alignment even when exact wording can be different.
    - It supports interpretability by emphasizing biologically relevant descriptions carefully.
4. **Mean Absolute Error (MAE):**
    - It quantifies the average level of the discrepancy among the model's predicted confidence scores and it's own actual predictive accuracy.
    - It is the primary indicator of calibration quality - less MAE values tells that predicted confidence levels closely align with true correctness probabilities.
5. **Pearson Correlation Coefficient (r):**
    - It can measure the strength and direction of the linear relationship among predicted confidence scores and observed prediction accuracy.
    - A high positive value signals a strong linear association, which implies that increasing confidence corresponds proportionally to a higher likelihood of correct predictions.
6. **Spearman's Rank Correlation (ρ):**
    - It evaluates the single relationship between confidence scores and accuracy based on their ordering rather than raw values.
    - It is useful for checking relative calibration, which ensures that high confidence predictions generally can correspond to high accuracy ranks even when the relationship is not strictly linear.
7. **Bias:**
    - It usually captures the systematic tendency of our model to overestimate or underestimate its predictive reliability across the whole dataset.
    - All positive bias values can indicate overconfidence, where the values that are close to zero reflect a better calibrated and unbiased estimation of uncertainty.

Summary of what is measured

- **Lexical overlap:** BLEU, ROUGE.
- **Structural similarity:** ROUGE (importantly ROUGE-L).
- **Semantic coherence:** BERT-based Precision/Recall/F1.
- **Calibration Reliability:** MAE, Bias.
- **Statistical Association:** Pearson r, Spearman ρ.

Together, these measures offer a strict and multifaceted evaluation framework that captures both lexical accuracy and semantic alignment, which is essential for benchmarking interpretable and reasoning-based protein function prediction models.

5.2.2 Ablation Study

We have tested 16 different variants with 10k training samples based on inputs and outputs.

Table 5.1: Input–Output configuration for model variants (V1–V16)

Variant	**Input (included data)**	**Output (format/content)**
V1	k=128 projected embedding	Function only (`<function>`)
V2	Compact biotools annotations, k=128 projected embedding	Function only (`<function>`)
V3	Compact biotools annotations, k=128 projected embedding	R4 reasoning, evidence, confidence, function
V4	Compact biotools annotations, k=128 projected embedding	Reasoning R1–R4 with confidence, function
V5	Compact biotools annotations, k=128 projected embedding	Reasoning R1–R4 with evidence and confidence, function
V6	Compact biotools annotations, k=128 projected embedding	Function then reasoning R1–R4 with evidence and confidence
V7	k=128 projected embedding	R4 reasoning, confidence and function
V8	k=128 projected embedding	Function, R4 reasoning, evidence and confidence
V9	k=128 projected embedding	R4 reasoning, confidence, evidence, function
V10	Structured biotools annotations, k=128 projected embedding	R4 reasoning, confidence, function
V11	Structured biotools annotations, k=128 projected embedding	Reasoning R1–R4 with confidences, function
V12	Structured biotools annotations, k=128 projected embedding	Reasoning R1–R3 with confidences, function
V13	Structured biotools annotations, k=128 projected embedding	Reasoning R1–R3, function, average confidence
V14	k=128 projected embedding	Reasoning R1–R4 with confidences, function
V15	k=128 projected embedding	Reasoning R1–R3 with confidences, function
V16	k=128 projected embedding	Reasoning R1–R4 (no confidences), function

The table provides a summary of the input-output settings that will be applied to the sixteen model variants (V1-V16) that are specified by the `build_input_output`

function. Each variant works on a fixed embedded k = 128 as the embedded input representation, substituting raw text queries as the main input representation. Variants vary in whether this embedding is applied on its own or in combination with the application of compact or structured biotools annotations. Previous variants (V1-V2) are modeled as bottom-of-the-line baselines, and they only generate the predicted function without carrying out any inference in between. Later variants (V3- V6, V9-V13) gradually add explicit reasoning elements, such as multi-step reasoning traces (R1-R4), evidence, confidence scores and (in some instances) average-confidence estimates. These elements occur in different orders and with different presence in different variants, which means that it is possible to analyze the influence of reasoning structure and supervision on model behavior in a controlled way. Subsequent variants (V14-V16) factor out the influence of reasoning depth and confidence supervision, by fixing the inputs to the projection to the embedded alone, and varying the number of reasoning steps and the presence or absence of confidence scores. In general, the table creates a systematic space of ablation of input structure, reasoning granularity, evidence inclusion and confidence modeling, and allows a specific comparison of representational and supervision decisions between variants.

Table 5.2: Evaluation Results for Variants 1–16 with Rankings

Variant	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BERT Prec	BERT Rec	BERT F1
V1(Baseline)	0.1471	0.2835	0.1597	0.2391	0.7637	0.7934	0.7767
V2	0.0335	0.1603	0.0910	0.1447	0.7314	0.6856	0.7044
V3	0.0403	0.1944	0.1204	0.1728	0.6834	0.7604	0.7154
V4	0.0330	0.2344	0.1308	0.2060	0.7385	0.7650	0.7486
V5	0.0648	0.2675	0.1473	0.2337	0.7728	0.7680	0.7672
V6	0.0257	0.0740	0.0220	0.0554	0.6571	0.6614	0.6539
V7	**0.2083**	**0.3563**	**0.2251**	**0.3197**	**0.8252**	**0.8188**	**0.8210**
V8	0.0305	0.0937	0.0527	0.0771	0.5305	0.6765	0.5918
V9	0.0352	0.2301	0.1413	0.2090	0.7106	0.7610	0.7311
V10	0.0456	0.1679	0.0925	0.1447	0.6610	0.7499	0.7001
V11	0.0144	0.1182	0.0592	0.1034	0.6375	0.7126	0.6704
V12	0.0370	0.1192	0.0641	0.1040	0.6826	0.6542	0.6639
V13	0.0408	0.2044	0.1295	0.1824	0.6897	0.7036	0.6922
V14	0.0575	0.2164	0.1150	0.1866	0.7573	0.7395	0.7456
V15	0.1257	0.2321	0.1352	0.2069	0.7723	0.7300	0.7486
V16	0.1296	0.2889	0.1550	0.2537	0.8058	0.7880	0.7952

5.3 Final Design Adjustments

Based on the ablation results, V7 is selected as the final model configuration:

- **Input:** k=128 projected protein embedding only, without textual annotations.
- **Output:** Single integrated reasoning (R4) with a confidence score, followed by the function description.

- **Rationale:** Provides the best trade-off between predictive performance (BERT-F1 = 0.8210) and interpretability.

5.4 Comparison Analysis

Table 5.3: Retrieval-based alignment performance of MSPP-128 on the SwissProt test dataset under different in-batch sizes and full test evaluation.

Evaluation Setting	**P2T Acc (%)**	**P2T R@20 (%)**	**T2P Acc (%)**	**T2P R@20 (%)**
In-batch (bs = 64)	43.22	87.76	44.77	79.22
In-batch (bs = 128)	37.07	75.64	38.77	70.94
In-batch (bs = 256)	30.87	64.37	33.21	63.45
In-batch (bs = 512)	24.73	56.22	27.53	56.89
In-batch (bs = 1024)	19.35	49.74	22.53	50.65
Full test set	6.18	29.03	8.53	32.27

We tested and assessed our model on two standard sets(benchmark) of data to understand the quality of protein text representations alignment. In spite of the fact that MSPP-128 is not actually constructed as a retrieval model, retrieval-type evaluation represents a robustly valid proxy in measuring the consistency of cross-modal embedding. In this direction, we obtain retrieval scores between protein embeddings and corresponding textual descriptions and this enables us to experimentally measure the faithfulness of alignments. The measures of protein-to-text (P2T) and text-to-protein (T2P) retrieval are employed to test whether semantically matched protein with text pairs are placed in closer proximity in the common embedding than divergent pairs in the common embedding. This evaluation plan makes it possible to assess the quality of alignment without regard to downstream generative performance, which justifies the quality of the representation alignment stage.

Table 5.5: Retrieval-based alignment performance of MSPP-128 on the Prot2Text test dataset under different in-batch sizes and full test evaluation.

Evaluation Setting	**P2T Acc (%)**	**P2T R@20 (%)**	**T2P Acc (%)**	**T2P R@20 (%)**
In-batch (bs = 64)	41.80	88.27	50.62	86.49
In-batch (bs = 128)	34.84	78.78	44.56	78.17
In-batch (bs = 256)	28.61	67.85	38.57	71.00
In-batch (bs = 512)	22.80	57.50	32.71	64.36
In-batch (bs = 1024)	17.38	49.19	26.93	57.59
Full test set	7.92	34.21	15.25	43.73

The most notable aspect of the ProtReason is that it can produce structured rea soning results that give precise reasons as to why predicted protein functions have been made. Although the organizational approach that traditional models are based on concerns sequence homology or classification in nomenclature, ProtReason manages to combine both biological sequence data and language-based reasoning so that it only takes data related to the structural bioinformatics fields and turns into predictive computational modeling. Through the combination of Asymmetric InfoNCE loss and Hard Negative Mining in Stage 1 of the training, ProtReason acquires a powerful correlation between protein sequences and textual representations. Stage 2 adds reasoning traces (R1-R4), which are applied by the LoRA-augmented Causal LLM, to enable ProtReason to produce detailed functional descriptions, confidence scores, and causal explanations that do not only assist in making predictions more accurately but also increase predictability.

Table 5.7: Evaluation Results for Related Models

Model	B-2	B-4	R-1	R-2	R-L	BERT Score
GPT-4o-mini (0-shot)	2.65	0.31	10.81	1.25	7.73	81.44
GPT-4o-mini (1-shot)	5.86	1.24	15.38	2.66	11.58	83.21
Claude-3.5-Sonnet(0-shot)	5.47	0.90	13.91	1.92	10.22	82.98
Claude-3.5-Sonnet (1-shot)	6.62	1.54	16.02	3.09	12.14	83.58
P2T-GPT	33.30	26.06	38.14	23.72	35.47	88.23
Vanilla-Transformer	17.48	15.75	27.80	19.44	26.07	82.31
ESM2GPT	35.49	32.11	47.46	39.18	45.31	89.97
RGCN2Transformer	31.64	27.97	42.43	34.91	40.72	90.58
RGCNxTransformer	33.01	30.39	45.75	37.38	43.63	89.04
ESM2-35M	-	32.11	47.46	39.18	45.31	43.21
RGCN	-	21.63	36.20	28.01	34.40	78.91
RGCN + ESM2-35M	-	30.39	45.75	37.38	43.63	82.51
RGCN × Vanilla-Transformer	-	27.97	42.43	34.91	40.72	81.12
SEQ2LLaMA	24.97	21.44	33.99	25.86	32.18	87.13
RGCN2LLaMA	25.96	22.46	39.36	31.34	37.60	88.31
RGCNxESM2LLaMA	35.84	31.84	50.87	43.28	48.98	90.92
Prot2Text	36.36	32.45	50.49	42.60	48.33	90.56
ProtReason	**29.96**	**26.29**	**40.91**	**30.87**	**38.31**	**89.45**

The combination of tracing reasoning is another important capability that identifies ProtReason as operational effects of the current approaches since it delivers transparent, understandable reasoning behind every prediction. Comparing ProtReason with a baseline model like Prot2Text, we can show that ProtReason is much better in terms of several metrics, among them, BLEU, ROUGE, and BERTScore. The Basis-Expainable feature of ProtReason is of special use in drugs discovery and functional genomics because it is important to know the explanation of the predictions.

Moreover, using textual sequence embeddings, ProtReason is always more accurate and better generalized than their predecessors. The findings reveal that ProtReason performs better in those tasks necessitating predictive accuracy and explainability of the models, and thus, it is a state-of-the-art option to analyze protein functions.

Table 5.8: Reasoning (Lexical) Performance Metrics

Metric	Value
BLEU-2	47.99
BLEU-4	36.78
ROUGE-1	68.05
ROUGE-2	47.66
ROUGE-L	64.01

The performance metrics of the reasoning phase are described in this table with the special attention to the lexical similarity measures, i.e. BLEU and ROUGE. These

values give a solid evaluation of the quality of generation of the values in terms of precision and recall coupled with standard benchmarking.

Table 5.9: Confidence Calibration Metrics

Metric	**Value**
MAE	0.1219
Pearson r	0.3333
Spearman ρ	0.2668
Bias	0.0870

The following table has summarized the most important calibration measures such as Mean Absolute Error (MAE) and correlation coefficients. These are measures of the fidelity of the scores of model confidence which are considered a major gauge of predictive reliability of the system.

Table 5.10: Confidence Mean Statistics

Metric	**Mean**	**± p-value**
Expected Mean	0.7022	± 0.1579
Predicted Mean	0.7892	± 0.1089

The above table gives a statistical breakdown of the confidence means pointing towards the deviation of the Expected and Predicted values. This information is also necessary in the diagnosis of aggregate calibration errors, which indicate the overall inclination of the model to overconfidence or underconfidence.

5.5 Summary of Results

5.5.1 LLM as a Judge

The table below will be a performance assessment of Large Language Models (LLMs) as evaluators of predictions of biological functions. The data is specifically the comparison of two models in a number of qualitative measures over a 1-5 scale.

Table 5.11: Function Prediction Quality (1–5)

Model	**Semantic**	**Biochemical**	**Completeness**	**Specificity**	**Aggregate**
LLaMA-3.3-70B	**3.22**	**4.60**	2.52	3.44	**3.44**
GPT-OSS-120B	2.80	3.80	**2.62**	**3.61**	3.21

In this case, Strong biochemical score imply that the model result is supportive of strong biochemical properties.

5.5.2 Reasoning & Confidence Analysis

The next analysis will evaluate the quality of the reasoning of the models on four important dimensions, including scientific validity, evidentiary support, hallucination avoidance, and coherence. On a 1-5 scale, performance is measured, and it is necessary to point out trade-offs between the test of logical accuracy and readability as a text.

Table 5.12: Reasoning Quality (1–5)

Model	Scientific	Support	No Hallucination	Coherence	Aggregate
LLaMA-3.3-70B	**3.92**	**3.76**	**4.36**	3.82	**3.96**
GPT-OSS-120B	3.26	3.23	2.20	**4.66**	3.34

The total score of LLaMA-3.3-70B is 3.96, which can be significantly explained by the fact that it is better scientifically grounded and has an enormous benefit of not hallucinating (4.36). Interestingly, the GPT-OSS-120B implies that its results are much more offshoot (4.66), but since it fails to reliably prevent hallucinations, it pulls its overall quality to 3.34.

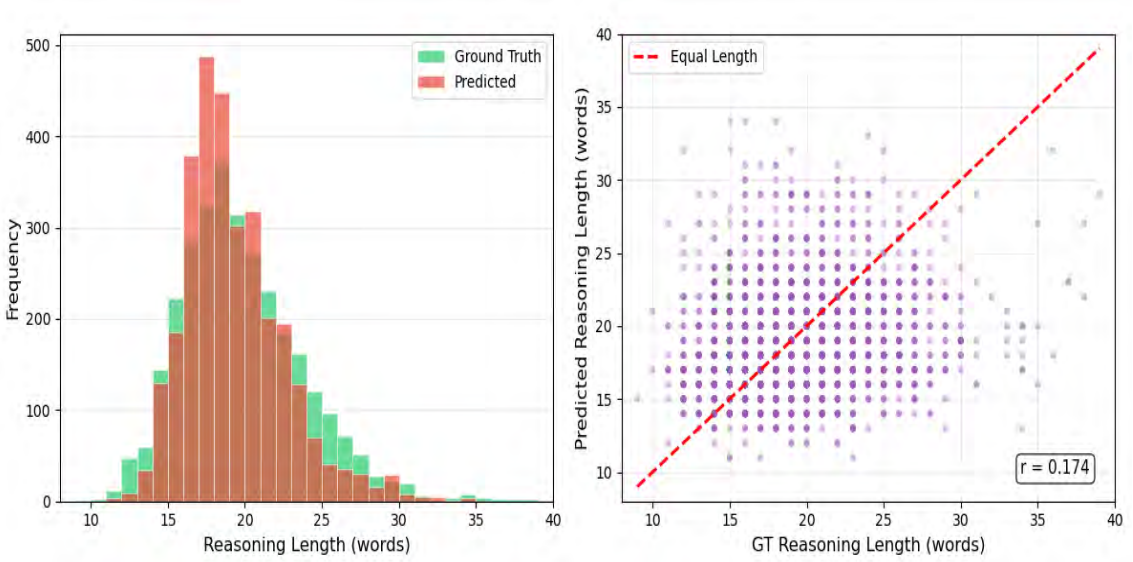


Figure 5.1: Comparison of ground truth and predicted reasoning lengths.

In order to find out the extent to which the models estimate their own knowledge, this section includes the comparative analysis of confidence calibration. Other important indicators are how to deal with uncertainty and correspondence between confidence and accuracy.

Table 5.13: Confidence Calibration (1–5)

Model	Calibrated	Uncertainty Acknowledgment	Accuracy	Aggregate
LLaMA-3.3-70B	2.78	3.95	4.57	3.36
GPT-OSS-120B	**3.02**	**4.03**	4.57	**3.52**

The aggregate score of 3.52 was higher than that of 3.36 in the category in which GPT-OSS-120B was engaged. The results of both models imply the same high performance of raw accuracy (4.57).

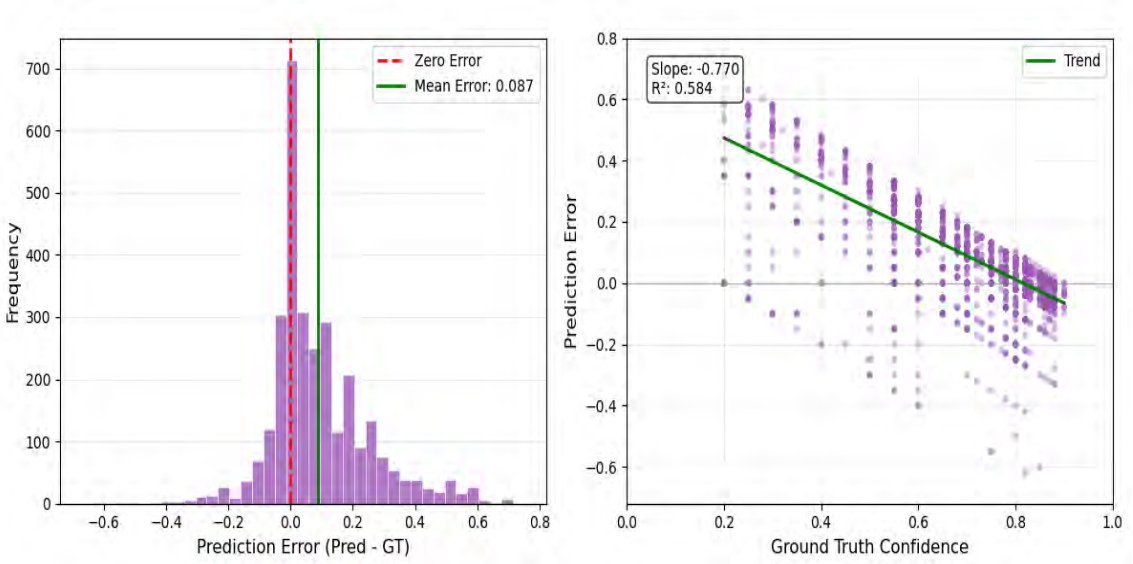


Figure 5.2: Error Analysis and Confidence Correlation.

5.6 Discussions

Results of this paper prove that ProtReason performs better than other protein function prediction benchmarks because it uses reasoning traces, which give it an opportunity to make predictions that are both more accurate and understandable. ProtReason involves the two-stage training which match protein sequences with textual descriptions during Stage 1, and during Stage 2, it produces structured output, for example functional descriptions and confidence scores with LoRA augmented Causal LLMs. Our model is much better off than current standard existing benchmark, as it scores better in terms of BLEU, ROUGE, and BERTScore, but it also offers clear and causal reasoning explanations as to why that is the case. These findings point to the combination of predictive accuracy and explainability as two important characteristics of the model, which make it suitable to apply it in the provides of applications that need high performance and interpretability.

Chapter 6

Conclusion

6.1 Summary of Findings

To conclude, ProtReason is a major improvement on protein functionality prediction, as it uses reasoning traces to improve model predictions in accuracy and interpretability. ProtReason plugs the gap between textual knowledge and protein sequences unlike traditional models that largely depend on sequence homology or an existing set of classification labels, which offer justification of each prediction in a way that is clear and not just organized. It trains its Causal LLM on two stages: initially matches the representations of proteins and then uses the results to produce structured predictions enhanced with the support of QLoRA. ProtReason is a better performer in functional prediction and model explainability. When compared to other models, such as Prot2Text, ProtReason performs better in terms of essential metrics, such as BLEU, ROUGE, and BERTScore, which proves that the model is better in use. This increased precision coupled with the capacity to produce interpretable arguments makes ProtReason a great tool to be used in modern biology, for instance, drug discovery and functional genomics and any other area where it is important to have an idea of the reasoning behind predictions. A combination of offering actionable information and high predictive accuracy makes ProtReason a cutting-edge model in the analysis of protein functionality, which has even greater potential of future development and application.

6.2 Contributions to the Field

6.2.1 Scientific Trust and Interpretability

Earlier sequence based predictors of protein functions are black boxes that lack transparency, biological validation and calibrated confidence. Conversely, our explanations based on reasoning are multimodal, which gives step-by-step explanations about all the predictions. The model generates reasoning traces (R1-R4) giving a traceable history of inference, which gives an account of the biological explanations of each prediction. This coincides with the increased requirement of explainable

AI in biological research whereby accuracy should be encompassed with scientific plausibility. The framework by making reasoning transparent allows biologists to clarify the origin of the predictions to their biological context to enable further validation and the enhancement of understanding of the protein functioning.

6.2.2 Hypothesis Generation and Discovery

The system also hastens the transformation of raw information into actionable information by considering the prediction of functionalities of proteins an interactive reasoning process. Complex functional questions can be posed by researchers in natural language and grounded and stepwise explanations provide mechanistic connections, unexpected homologs, or additional pathway connections. In addition to standard predictive models, this functionality helps generate and discover hypotheses based on synthesizing protein sequences and textual knowledge.

6.2.3 Availability and Democratization

Exposing complicated workflows via a conversational interface, the framework reduces technical complexity normally linked with the state-of-the-art bioinformatics tools. The system makes protein analysis accessible to a wider audience, such as wet-lab teams and cross-disciplinary teams by making access to it democratic. This will facilitate more effective cooperation and may save time and resources that were utilized in the course of experimental validation.

6.2.4 Integration of Sequence and Textual Knowledge

The result of this fusion is better in protein representations and predictions of low-homology or poorly characterized proteins. In particular, the integration of biological text information, e.g., Pfam, DeepLoc and InterPro as well as the sequence data can be used to make more confident predictions that are not limited to similarity in sequences alone, which further develop functional genomics of uncharacterized proteins.

6.2.5 Implications for Biological Data Science Applications

Our framework supports scalable and interpretable framework of protein annotation that is useful in functional genomics. In drug discovery, the system may assist in target discovery and variant analysis by helping to understand the relationship of the predicted protein functions with disease biology. In protein engineering, it can be used to inform the design of desired sequences with functionality on sequence evidence and functional annotations. In general, the strategy increases the reproducibility and influence throughout biotechnology research pipelines.

Enhancements

1. This has involved the references to the reasoning trace with confidence and step by step description into the text in a more expressive manner to represent the fundamental innovations of the framework.

2. Emphasis on biological context, explainability and democratization of advanced tools to be on par with the focus on accessibility, transparency and wider adoption.

3. Increased enhancement on multimodal fusion and its contribution to the extension of the framework usage to low-homology proteins or proteins with undetermined functions.

6.3 Recommendations for Future Work

6.3.1 Ablation of Scalability and Modality Projector

The future directions of the research should undertake a systematic evaluation of the scalability of the multimodal projector through training on more and larger numbers of epochs and biological data. Performance on generalizing to previously unseen protein families and to full scale genomic or meta-genomic sets of data requires extensive ablation experiments to explore the effects of dataset size, training time, and projector architecture on performance. Such analysis will assist in understanding the saturation levels of performance and ensure that the increase in size does not impact the interpretability. An intensely optimized projector would enable exploration of large molecular libraries, which would be used in target discovery and designing proteins.

6.3.2 Domain-Specific Language Model Pretraining

Another promising avenue to enhancing interpretability and functional specificity is to pretrain a base language model on large biological corpora, and then multimodally fine-tune it. More biological concepts (pathways, protein families, functional terminology) are more likely to be encoded in domain-adapted models, leading to more human-coherent reasoning and biologically-reasoned output. Assessment of the effects of such pretraining on functional accuracy and inner representations would improve confidence in the model predictions and would bolster the use of such models in functional annotation and protein engineering.

6.3.3 Evidence-Based Rationality and Timely Design

Another way is to develop domain-focused prompting strategies to clearly correlate model reasoning with molecular evidence. Causal and traceable reasoning can be

promoted with the help of structuring prompts to include biochemical rules, structural motifs, or constraints of experiments. Implementation of evidence-based reasoning would minimize hallucinations, enhance scientific credibility, and enable researchers to gain a better understanding of the reasoning behind predictions, which would boost confidence in downstream biological conjecture.

6.3.4 Combination of Structural and Graph-Based Modalities

Explicit structural and graph-based representations are necessary in the multimodal framework to enhance functional resolution, especially of low-homology proteins or novel proteins. The model could be improved with the inclusion of residue-level contact graphs, 3-dimensional structural embeddings, or even molecular interaction graphs in combination with sequence data. The anticipated results of this multimodal fusion include improved functional prediction and a wider range of applications in drug discovery, protein engineering, and rational therapeutic design.

Bibliography

[1] D. Wang, M. Pourmirzaei, U. L. Abbas *et al.*, "S-plm: Structure-aware protein language model via contrastive learning between sequence and structure," *bioRxiv*, 2023. [Online]. Available: https://doi.org/10.1101/2023.08.06.552203

[2] H. Abdine, M. Chatzianastasis, C. Bouyioukos, and M. Vazirgiannis, "Prot2text: Multimodal protein's function generation with gnns and transformers," *arXiv preprint*, vol. arXiv:2307.14367, 2023, arXiv:2307.14367. [Online]. Available: https://arxiv.org/abs/2307.14367

[3] Z. Ma, C. Fan, Z. Wang *et al.*, "Protex: Structure-in-context reasoning and editing of proteins with large language models," *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/abs/2503.08179

[4] Y. Xiao, E. Sun, Y. Jin, Q. Wang, and W. Wang, "Proteingpt: Multimodal llm for protein property prediction and structure understanding," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2408.11363

[5] S. Awasthi, A. Singh, A. Gupta, R. Sharma, P. Kumar, M. Patel, and S. Mehta, "Protdetr: Protein function prediction with transformer-based detection," *arXiv preprint*, vol. arXiv:2501.05644, 2025, arXiv:2501.05644. [Online]. Available: https://arxiv.org/abs/2501.05644

[6] M. Jin, H. Xue, Z. Wang *et al.*, "Prollm: Protein chain-of-thoughts enhanced llm for protein-protein interaction prediction," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2405.06649

[7] G. Cao, Y. Shen, J. Wu *et al.*, "Cellreasoner: A reasoning-enhanced large language model for cell type annotation," *bioRxiv*, 2025. [Online]. Available: https://www.biorxiv.org/content/10.1101/2025.05.20.655112v1

[8] C. Li *et al.*, "Proteinreasoner: A multi-modal protein language model with chain-of-thought reasoning for efficient protein design," *bioRxiv*, 2025. [Online]. Available: https://www.biorxiv.org/content/10.1101/2025.07.21.665832v1

[9] R. S. Boger, S. Chithrananda, A. N. Angelopoulos *et al.*, "Functional protein mining with conformal guarantees," *Nature Communications*, vol. 16, p. 85, 2025. [Online]. Available: https://doi.org/10.1038/s41467-024-55676-y

[10] Y. Yang, A. Jerger, S. Feng *et al.*, "Clean-contact: Contrastive learning-enabled enzyme functional annotation prediction with structural inference," *bioRxiv*, 2024. [Online]. Available: https://doi.org/10.1101/2024.05.14.594148

[11] H. Xiao, W. Lin, X. Chen *et al.*, "Stella: Towards protein function prediction with multimodal llms integrating sequence–structure representations," *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/abs/2506.03800

[12] M. Yin, H. Zhou, Y. Zhu, M. Lin, Y. Wu, J. Wu, H. Xu, C.-Y. Hsieh, T. Hou, J. Chen, and J. Wu, "Multi-modal clip-informed protein editing," *arXiv preprint*, vol. arXiv:2407.19296, 2024, arXiv:2407.19296. [Online]. Available: https://arxiv.org/abs/2407.19296

[13] K. Van der Weg, E. Merdivan, M. Piraud *et al.*, "Topec: Prediction of enzyme commission classes by 3d graph neural networks and localized 3d protein descriptor," *Nature Communications*, vol. 16, p. 2737, 2025. [Online]. Available: https://doi.org/10.1038/s41467-025-57324-5

[14] C. Ma, H. E., J. Ding, J. Zhang, Z. Ma, Q. Huang, B. Gao, L. Chen, and M. Song, "Sci-reason: A dataset with chain-of-thought rationales for complex multimodal reasoning in academic areas," *arXiv preprint*, vol. arXiv:2504.06637, 2025, arXiv:2504.06637. [Online]. Available: https://arxiv.org/abs/2504.06637

[15] S. Sengupta, S. Yang, P. Kwong Yu *et al.*, "Biomol-mqa: A multi-modal question answering dataset for llm reasoning over bio-molecular interactions," *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/abs/2506.05766

[16] C.-Y. Hsu, K. Cox, J. Xu *et al.*, "Thought graph: Generating thought process for biological reasoning," in *Proceedings of the Web Conference (WWW) Companion*, 2024, pp. 537–540. [Online]. Available: https://doi.org/10.1145/3589335.3651572

[17] A. Fallahpour, A. Magnuson, P. Gupta, S. Ma, J. Naimer, A. Shah, H. Duan, O. Ibrahim, H. Goodarzi, C. J. Maddison, and B. Wang, "Bioreason: Incentivizing multimodal biological reasoning within a dna-llm model," *arXiv preprint*, vol. arXiv:2505.23579, 2025, arXiv:2505.23579. [Online]. Available: https://arxiv.org/abs/2505.23579

[18] Z. Wang, Q. Zhang, K. Ding, M. Qin, X. Zhuang, X. Li, and H. Chen, "Instructprotein: Aligning human and protein language via knowledge instruction," *arXiv preprint*, vol. arXiv:2310.03269, 2023, arXiv:2310.03269. [Online]. Available: https://arxiv.org/abs/2310.03269

[19] Y. Chen, Z. Li, X. Wang, J. Zhang, and L. Wang, "Structure-aligned protein language model," *arXiv preprint*, vol. arXiv:2505.16896, 2025, arXiv:2505.16896. [Online]. Available: https://arxiv.org/abs/2505.16896

[20] D. Kumar, D. Yadav, and Y. Patel, "Gpt-oss-20b: A comprehensive deployment-centric analysis of openai's open-weight mixture of experts model," *arXiv preprint*, vol. arXiv:2508.16700, 2025, arXiv:2508.16700. [Online]. Available: https://arxiv.org/abs/2508.16700

[21] X. Liu, Z. Zhang, Y. Wang, Y. Lu, J. Tang, and J. Leskovec, "Mol-instructions: A benchmark for instruction-finetuned molecular language models," *arXiv preprint*, vol. arXiv:2306.08018, 2023, arXiv:2306.08018. [Online]. Available: https://arxiv.org/abs/2306.08018

[22] J. Wang, Z. Yang, C. Chen *et al.*, "Mpek: A multitask deep learning framework for enzymatic reaction kinetic parameter prediction," *Briefings in Bioinformatics*, vol. 25, no. 5, p. bbae387, 2024. [Online]. Available: https://academic.oup.com/bib/article/25/5/bbae387/7731495

[23] K. A. Sajeevan, A. Osinuga *et al.*, "Robust prediction of enzyme variant kinetics with realkcat," *bioRxiv*, 2025. [Online]. Available: https://www.biorxiv.org/content/10.1101/2025.02.10.637555v1

[24] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.