

# BioReason-Pro: Advancing Protein Function Prediction with Multimodal Biological Reasoning

Adibvafa Fallahpour<sup>\*,1,2,3,4,5</sup>, Arman Seyed-Ahmadi<sup>\*,2,5</sup>, Parsa Idehpour<sup>\*,1,5,9</sup>,  
Omar Ibrahim<sup>\*,2,5</sup>, Purav Gupta<sup>\*,3,4,5</sup>, Jack Naimer<sup>5,10</sup>, Kevin Zhu<sup>5,8</sup>, Arnav Shah<sup>3,4,5</sup>,  
Shihao Ma<sup>2,3,4,5</sup>, Abhinav Adduri<sup>1,5</sup>, Talu Güloğlu<sup>4,5,11</sup>, Nuo Liu<sup>1</sup>, Haotian Cui<sup>13</sup>,  
Arihant Jain<sup>1,9</sup>, Max de Castro<sup>9</sup>, Amirfaham Fallahpour<sup>4</sup>, Antonio Cembellin-Prieto<sup>1</sup>,  
John S. Stiles<sup>1</sup>, Filip Nemčko<sup>1</sup>, Alexander A. Nevue<sup>1</sup>, Hyungseok C. Moon<sup>1</sup>,  
Lucas Sosnick<sup>1,6</sup>, Olivia Markham<sup>12</sup>, Haonan Duan<sup>3,4</sup>, Michelle Y. Y. Lee<sup>1,6</sup>,  
Andrea F. M. Salvador<sup>1,6</sup>, Chris J. Maddison<sup>3,4</sup>, Christoph A. Thaiss<sup>1,6</sup>,  
Chiara Ricci-Tam<sup>1</sup>, Brian S. Plosky<sup>1</sup>, Dave P. Burke<sup>1</sup>, Patrick D. Hsu<sup>1,8</sup>,  
Hani Goodarzi<sup>†,‡,1,7</sup>, Bo Wang<sup>†,‡,2,3,4,13</sup>

<sup>1</sup>Arc Institute; <sup>2</sup>University Health Network; <sup>3</sup>Vector Institute; <sup>4</sup>University of Toronto; <sup>5</sup>Core Contributor;  
<sup>6</sup>Stanford University; <sup>7</sup>University of California, San Francisco; <sup>8</sup>University of California, Berkeley;  
<sup>9</sup>University of Pennsylvania; <sup>10</sup>EPFL; <sup>11</sup>ETH Zürich; <sup>12</sup>Cohere; <sup>13</sup>Xaira Therapeutics

## Abstract

Protein function annotation is fundamental to understanding biological mechanisms, designing therapeutics, and advancing biomedical research. Current computational methods either rely on shallow sequence similarity or treat function prediction as isolated classification tasks, failing to capture the integrative reasoning across sequence, structure, domains, and interactions that expert biologists perform to infer function. We introduce BioReason-Pro, the first multimodal reasoning large language model (LLM) for protein function prediction that integrates protein embeddings with biological context to generate structured reasoning traces. A key input into BioReason-Pro is the set of GO term predictions made by GO-GPT, our autoregressive transformer that captures hierarchical and cross-aspect dependencies of GO terms. BioReason-Pro is trained via supervised fine-tuning on synthetic reasoning traces generated by GPT-5 for over 130K proteins and further optimized through reinforcement learning. It achieves 73.6%  $F_{\max}$  on GO term prediction and an LLM judge score of 8/10 on functional summaries, substantially outperforming previous methods. Evaluations with human protein experts show that BioReason-Pro annotations are preferred over ground truth UniProt annotations in 79% of cases. Remarkably, BioReason-Pro de novo predicted experimentally confirmed binding partners with per-residue attention localizing to the exact contact residues resolved in cryo-EM structures of those complexes. Together, GO-GPT and BioReason-Pro establish a framework for protein function prediction that combines precise ontology modeling with interpretable biological reasoning.

---

\*Equal contribution. The order of authors is not a reflection of their relative contributions.

†These authors, listed alphabetically, jointly supervised this work.

‡Correspondence to hani@arcinstitute.org, bo.wang@uhn.ca

---

## 1. Introduction

Proteins are the molecular machines of life, executing a vast repertoire of biochemical functions that underpin metabolism, signaling, and cellular organization (Fischer, 1894; Anfinsen, 1973). Advances in high-throughput sequencing have yielded an explosion of protein sequence data, with over 250 million entries in databases such as UniProt (Consortium, 2025). Yet functional characterization has not kept pace, as fewer than 0.1% of known proteins possess experimentally validated annotations due to the high cost and limited throughput of wet-lab assays (O’Donovan et al., 2002). This persistent annotation gap poses a major challenge for molecular biology and drug discovery, motivating the development of scalable computational approaches that infer protein function directly from sequence (Friedberg, 2006; Radivojac et al., 2013).

Protein function prediction is an integrative reasoning task. Biologists examining an uncharacterized protein analyze evidence from domain architecture, structural motifs, evolutionary context, organism biology, and interaction networks to construct a coherent functional hypothesis. Protein foundation models such as ESM3 (Hayes et al., 2024) and ProtT5 (Elnaggar et al., 2022) encode rich sequence-structure-function relationships from evolutionary data, yet this knowledge remains implicit, falling short of the deliberate reasoning that expert annotation demands. Conversely, advances in LLM reasoning have shown that explicit chain-of-thought dramatically improves performance on complex multi-step problems (Wei et al., 2023; OpenAI et al., 2024; Istrate and Karaletsos, 2025). To the best of our knowledge, no method yet combines the representational depth of protein foundation models with structured, multi-step reasoning to predict and explain protein function.

A complementary line of research focuses on free-text protein function prediction, enabled by large language models (LLMs) that generate natural language descriptions directly from protein representations (Xiao et al., 2025b), with methods such as Prot2Text (Abdine et al., 2023), ProtT3 (Liu et al., 2024), STELLA (Xiao et al., 2025a), ProtChatGPT (Wang et al., 2025b), ProteinCLIP (Wu et al., 2024), Prot2Chat (Wang et al., 2025d), and Prot2Text-V2 (Fei et al., 2025) aligning protein embeddings with LLMs to produce natural language functional summaries. Despite improved interpretability, these models remain descriptive rather than explanatory. They do not reason over relevant evidence or justify why a function may be plausible.

Protein function is primarily described using the Gene Ontology (GO) (Ashburner et al., 2000), which organizes protein function into three aspects: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). GO terms are widely used to interpret large-scale experiments such as RNA-seq and proteomics (Consortium, 2021). Manual annotation requires expert curation and is time-consuming and costly. Early efforts transferred functions from homologous proteins identified by sequence alignment tools such as BLAST and DIAMOND (Altschul et al., 1990; Buchfink et al., 2015). Deep learning has since transformed this field, with methods such as NetGO 3.0 (Wang et al., 2023), InterLabelGO+ (Evans and Shen, 2024), DPFunc (Wang et al., 2025c), PhiGnet (Jang et al., 2024), ProtGO (Wang et al., 2025a), ProtNote (Char et al., 2025), and ProtBoost (Chervov, 2024) leveraging protein language model embeddings, structural representations (Fleming et al., 2025), and protein interaction networks, achieving strong performance on benchmarks such as the Critical Assessment of Functional Annotation (CAFA) (Radivojac et al., 2013). Despite this progress, these methods treat GO terms as isolated classification targets and their fixed ontology vocabulary constrains expressivity for proteins with novel or combinatorial functions (Kulmanov and Hoehndorf, 2020).

Here we present BioReason-Pro, the first multimodal reasoning LLM for protein function prediction. An important contribution to the reasoning of this model are predicted GO terms from GO-GPT, an autoregressive transformer for GO prediction that we built. GO-GPT treats GO annotation as a sequence generation task conditioned on protein representations, achieving a weighted  $F_{\max}^w$  of 0.65–0.70 across inference strategies, surpassing the top accessible methods from the CAFA 5 competition (Friedberg et al., 2023b). By integrating output from GO-GPT, protein embeddings, and additional biological context, BioReason-Pro is able to generate structured traces that reason from domain analysis to functional hypotheses. We trained BioReason-Pro on over 130K synthetic reasoning traces spanning 3,135 organisms and refined it with reinforcement learning. Human experts preferred BioReason-Pro annotations over curated UniProt entries in 79% of evaluated cases, and an LLM judge scored its functional summaries 8/10 on average, substantially outperforming previous methods even for proteins with very low similarity to training data. Remarkably, BioReason-Pro de novo identified experimentally validated binding partners for individual test proteins, with per-residue attention localizing to the precise contact interfaces resolved in cryo-EM structures of those complexes. The model also performed structural reasoning that overrode misleading superfamily-level domain annotations, a capability beyond the reach of homology transfer or domain lookup alone.

---

To enable broad adoption, we release all model weights, code, and curated datasets, alongside a [web interface](#) and model predictions for over 240,000 proteins including the Human Protein Atlas. BioReason-Pro demonstrates that AI systems can reason about protein function at expert level, opening a path toward scalable functional characterization of the millions of uncharacterized proteins across all domains of life.

## 2. Results

### 2.1. Building a biology-language multimodal model for protein function prediction

Protein function is determined by the interplay of sequence, structure, domain architecture, evolutionary context, and molecular interaction networks, which are the same evidence sources that expert biologists integrate when annotating uncharacterized proteins. To test whether a language model augmented with these biological modalities could reason about protein function, we built BioReason-Pro (**Fig. 1A**), a multimodal LLM based on Qwen3-4B ([Yang et al., 2025](#)). BioReason-Pro reasons over biological context in natural language, leveraging the chain-of-thought capabilities that have emerged in recent LLMs ([OpenAI et al., 2024](#); [Guo et al., 2025](#)). BioReason-Pro integrates residue-level embeddings from ESM3 ([Hayes et al., 2024](#)), which encodes protein function jointly with sequence and structure to provide a biologically grounded representation of protein context. Separately, a GO graph encoder captures the structure of the Gene Ontology as embeddings. Alongside these embeddings, the LLM receives target organism, domain annotations, protein-protein interactions, and initial GO term hypotheses from GO-GPT.

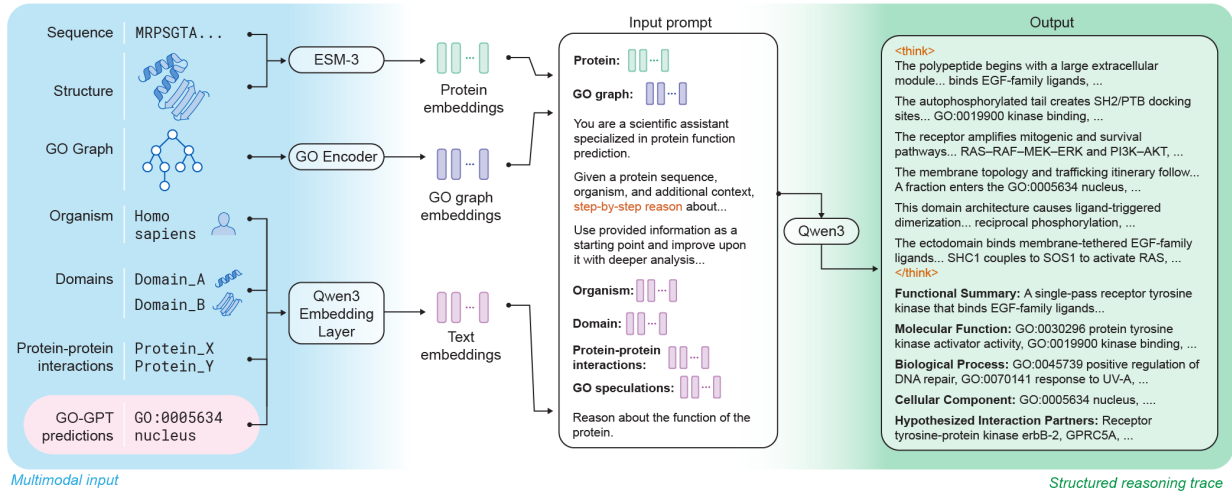
To ground our models in known biology, we collected domain annotations from InterPro ([Blum et al., 2024](#)), protein-protein interactions from STRING ([Szklarczyk et al., 2024](#)), protein structures from the PDB ([Burley et al., 2018](#)), and subcellular localization metadata (**Fig. 1B**). For training, we curated a broadly representative protein function dataset from UniProt ([Consortium, 2025](#)) comprising 133,492 proteins across 3,135 organisms (**Fig. S1A,B**). Following the CAFA community challenge ([Friedberg et al., 2023b](#)), which defines temporal holdout protocols and experimental evidence codes for benchmarking function prediction, we retained only proteins with experimental GO annotations (**Fig. S1C–F**). We excluded incomplete sequences and truncated proteins to 2000 amino acids to fit within the ESM3 context window (**Fig. 1C**). The resulting dataset spans broad taxonomic diversity (**Fig. 1D**), with per-protein domain, interaction, and localization statistics in **Fig. S1G–I**.

Training a model to reason about protein function requires step-by-step reasoning traces, yet no large-scale corpus of human-authored traces exists. Recent work has shown that models can learn to reason through supervised fine-tuning (SFT) on LLM-generated synthetic reasoning traces and be further improved via reinforcement learning (RL) to increase prediction accuracy ([Guo et al., 2025](#); [Zelikman et al., 2022](#); [Fallahpour et al., 2025b](#)). We adopted this approach by using GPT-5 ([Singh et al., 2025](#)) to generate over 130K synthetic traces for protein function prediction (**Fig. 1E**, Section C.1). We split the data temporally according to the CAFA framework ([Zhou et al., 2019](#)), where proteins annotated before November 2022 were used for training, and those that gained new experimental annotations between March 2023 and February 2024 but lacked prior annotations in the target GO aspect were held out for evaluation (**Fig. 1F**). BioReason-Pro was trained via SFT with next-token prediction on these traces (**Fig. 1G**), which taught BioReason-Pro to generate biological reasoning traces. To directly optimize for GO term annotation accuracy, we further applied reinforcement learning via Group Sequence Policy Optimization (GSPO) ([Zheng et al., 2025](#)), using  $F_{\max}^w$  between predicted and ground truth GO terms as the reward signal. The protein embedding space learned by BioReason-Pro SFT is visualized across the full training set (**Fig. 1H**). We also compare ESM3 and BioReason-Pro SFT embeddings on a 10K protein subset, with clusters annotated by their most enriched GO Molecular Function term (**Fig. 1I**).

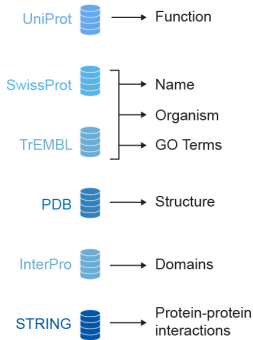
### 2.2. GO-GPT improves GO term prediction and captures biologically meaningful structure

As shown in **Fig. 1A**, BioReason-Pro operates on a set of likely GO terms for the given protein. For this, we use our GO-GPT model, which is an autoregressive transformer ([Vaswani et al., 2023](#); [Radford et al., 2019](#)) that generates GO terms sequentially, with each prediction conditioned on residue-level protein embeddings from ESM2 ([Lin et al., 2023](#)), the user-specified organism, and all previously generated terms (**Fig. 2A**). We chose ESM2 over ESM3 as it performed comparably on GO term prediction while enabling more efficient batched embedding extraction (Appendix B.3). This approach captures both the hierarchical relationships within ontology aspects and the dependencies across them that independent classification approaches miss.

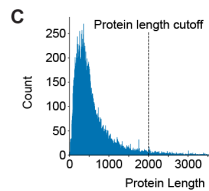
## A BioReason-Pro Architecture



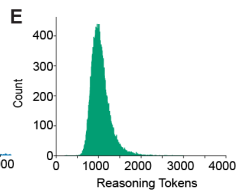
## B



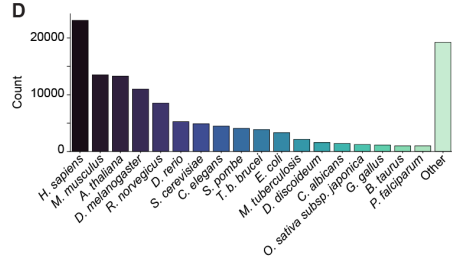
## C



## E



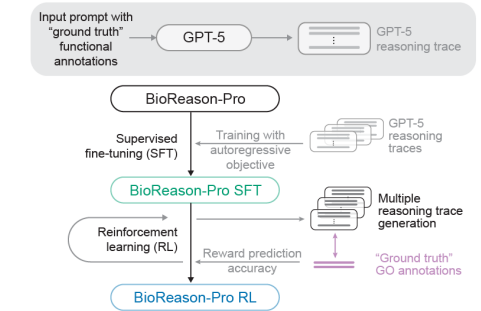
## D



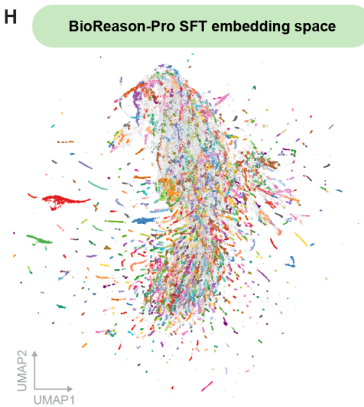
## F



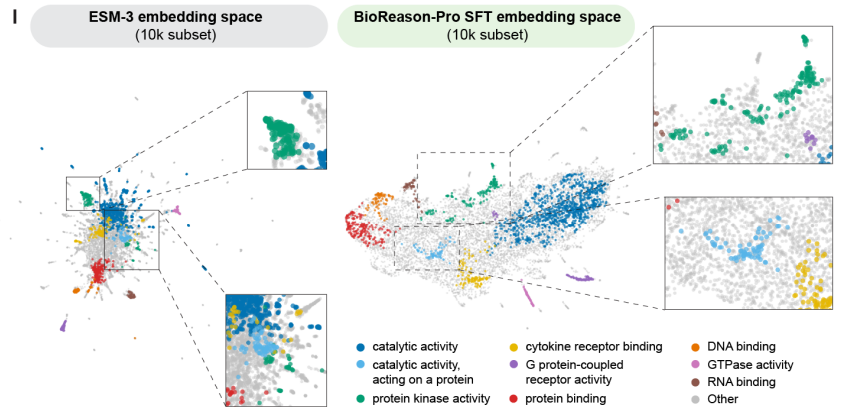
## G



## H



## I



**Figure 1 | Overview of BioReason-Pro for protein function prediction.** (A) **BioReason-Pro architecture.** A multimodal reasoning LLM that integrates ESM3 protein embeddings, a GO graph encoder, and biological context to generate structured reasoning traces and functional annotations. (B) **Dataset overview.** 133,492 proteins across 3,135 organisms curated from UniProt with experimental GO annotations, InterPro domains, STRING protein-protein interactions, and PDB protein structures. (C) **Protein length distribution.** Distribution of protein sequence lengths across the dataset with a 2,000 residue cutoff. (D) **Organism breakdown.** Taxonomic diversity of the training dataset. (E) **Reasoning trace distribution.** Token length distribution of synthetic reasoning traces across training proteins. (F) **Temporal split.** Training data through November 2022 and test data from March 2023 to February 2024. (G) **Training procedure.** SFT on GPT-5-generated reasoning traces followed by RL with GO term prediction accuracy as the reward. (H) **Learned embedding space.** UMAP projection of BioReason-Pro SFT LLM Layer 35 embeddings across all training proteins, colored by HDBSCAN clusters computed in PCA-reduced space. (I) **Embedding space comparison.** UMAP projections of ESM3 (left) and BioReason-Pro SFT (right) embeddings for a 10K protein subset. Cluster assignments were derived from BioReason-Pro embeddings and applied to both panels, with each cluster annotated by its most enriched GO Molecular Function term.

---

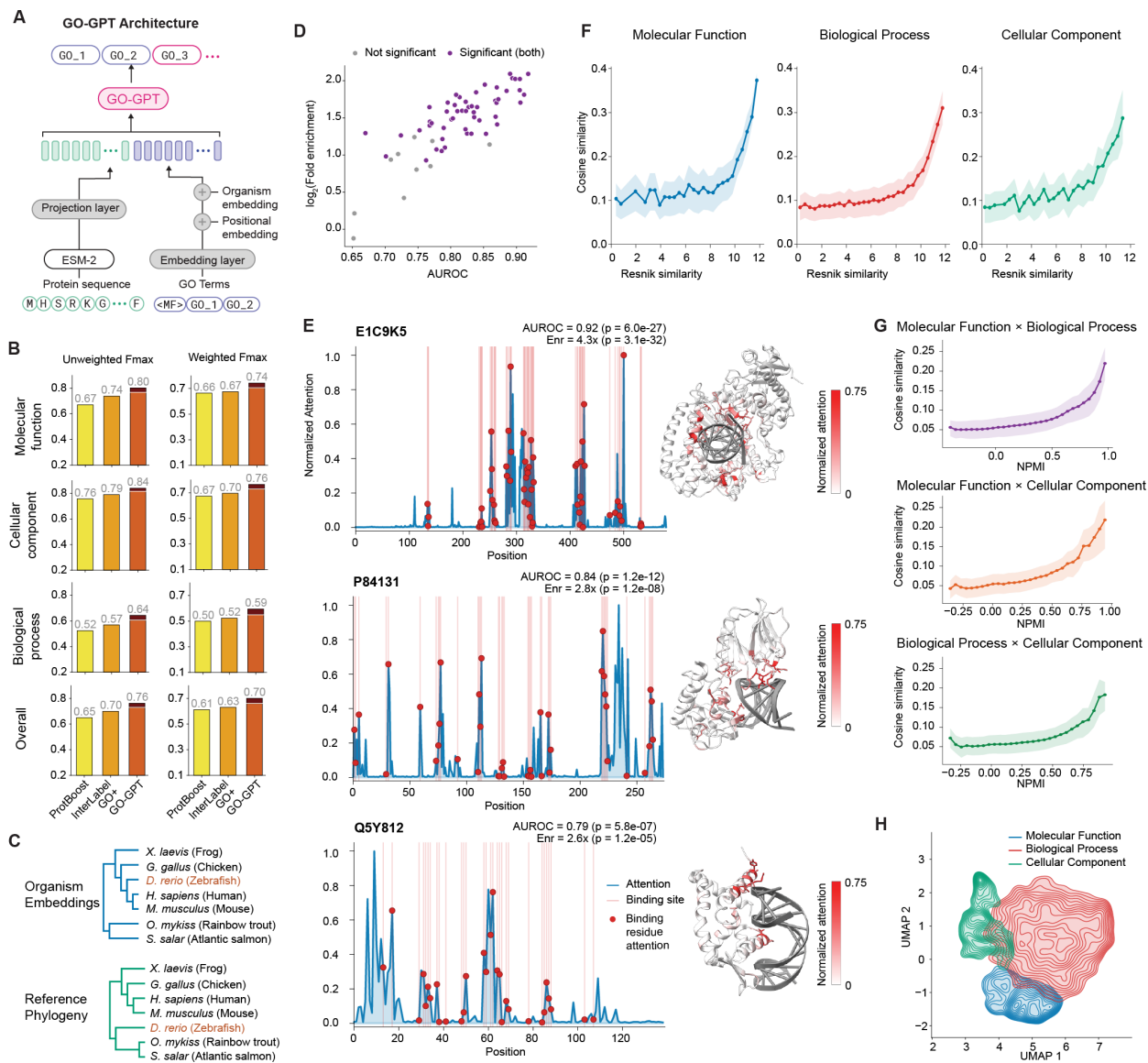
The Gene Ontology organizes protein function as a directed acyclic graph from general root concepts to highly specific leaf annotations (Ashburner et al., 2000; Aleksander et al., 2023). Existing methods predict each GO term independently, ignoring both this hierarchical structure and the dependencies between terms. GO-GPT captures these relationships by treating GO prediction as a sequential generation task, predicting increasingly specific terms conditioned on residue-level protein embeddings, the target GO aspect, more general GO terms, and organism. We evaluated GO-GPT against InterLabelGO+ (Evans and Shen, 2024) and ProtBoost (Chervov, 2024), the highest-ranked CAFA 5 methods with publicly available implementations, on a temporal holdout set following the CAFA framework (Friedberg et al., 2023a) (Fig. 1F; Fig. 2B).

We measure performance of function prediction models using  $F_{\max}$ , the maximum F-measure (harmonic mean of precision and recall) across all possible probability thresholds (Section 4.4.1) (Clark and Radivojac, 2013b). For each threshold, GO terms with probability above the threshold are considered predictions; precision and recall are computed against ground truth and averaged across proteins.  $F_{\max}$  selects the threshold maximizing this protein-averaged F1, providing a single summary of predictive performance independent of threshold choice. We report both unweighted  $F_{\max}$ , where all terms contribute equally, and weighted  $F_{\max}^w$ , which emphasizes rare and specific terms that are harder to predict but more biologically informative (Zhou et al., 2019).  $F_{\max}$  requires per-term probability estimates, but GO-GPT generates sequences of discrete GO tokens. Assigning probability 1 to all generated terms and 0 otherwise, GO-GPT achieves  $F_{\max}^w = 0.65$ , surpassing InterLabelGO+ at 0.63 (significant for BP and CC; MF difference not significant under greedy decoding,  $p = 0.073$ ; Table S3). Since GO-GPT generates sequences stochastically, sampling multiple trajectories per protein allows estimating per-term probabilities. Generating 10 independent samples per protein and using term frequencies as probabilities improves performance to  $F_{\max}^w = 0.67$ . Selecting the best of 10 samples per protein reaches 0.70, indicating substantial headroom from improved decoding or selection strategies. GO-GPT shows a similar gap on unweighted  $F_{\max}$ , achieving 0.76 compared to 0.70 for InterLabelGO+ (Fig. 2B). These results demonstrate that autoregressive generation is a powerful framework for GO term prediction, capturing both hierarchical relationships within aspects and dependencies across them. Sampling-based confidence estimation emerges naturally from this generative approach, providing per-term probability estimates without requiring additional calibration.

Beyond predictive accuracy, we examined whether GO-GPT’s learned representations capture biologically meaningful structure. GO-GPT learns a unique embedding for each organism during training. Because evolutionary relatedness shapes the functional constraints acting on proteins, we asked whether these learned embeddings recover known phylogenetic structure. Computing cosine similarity between these embeddings and constructing a dendrogram produces a tree that recapitulates known phylogenetic relationships, with closely related taxa consistently grouping together (Fig. 2C; Mantel test,  $p = 4 \times 10^{-3}$ , for top 30 most frequent organisms) (Fallahpour et al., 2025a; Kumar et al., 2022; Mantel, 1967). This suggests that the model captures organism-specific biases in how protein sequence maps to function, and that these biases are more conserved between phylogenetically proximal species.

We next asked whether GO-GPT leverages functionally relevant regions of the protein when making predictions. To systematically evaluate this, we focused on DNA-binding proteins, whose direct interaction with DNA through specific residues provides a well-defined ground truth for assessing whether attention aligns with functional sites. We analyzed attention patterns when GO-GPT predicts the DNA-binding function term (GO:0003677), using proteins with annotated DNA-binding residues from BioLip (Yang et al., 2012) that were absent from the training set (Section 4.5.1). For each protein, we extracted per-residue attention scores when the model generates the DNA-binding term (GO:0003677), producing an attention profile across the sequence that can be compared against known binding sites. We then quantified the correspondence between model attention and known binding sites using two complementary metrics: AUROC, which measures how well attention scores distinguish binding from non-binding residues, and fold-enrichment, which measures the concentration of attention at binding sites relative to the sequence background.

Across all 63 evaluated proteins, attention is consistently enriched at annotated binding residues (mean AUROC =  $0.81 \pm 0.06$ , significant in 59/63; mean top-20% fold-enrichment =  $2.8x \pm 0.7x$ , significant in 55/63), with AUROC and fold enrichment positively correlated across proteins (Fig. 2D) and 60 of 63 exceeding AUROC 0.7 (full per-protein statistics in Table S1; Fig. S2). Three proteins spanning different AUROC values (E1C9K5, P84131, Q5Y812; selected as described in Section 4.5.1) are visualized in Fig. 2E, with AUROC ranging from 0.79 to 0.92 and top-20% fold-enrichment from 2.6x to 4.3x (all  $p < 10^{-3}$ , Mann-Whitney U and hypergeometric tests). Projecting high-attention residues onto protein-DNA complex structures con-



**Figure 2 | GO-GPT architecture, performance, and interpretability analysis.** (A) **GO-GPT architecture.** Autoregressive Transformer encoding protein sequences via ESM2, generating GO terms conditioned on organism and previously generated terms. (B) **GO term prediction performance.** Weighted and unweighted  $F_{\max}$  across all aspects (dark shading: best-of-10 upper bound; all pairwise differences significant except GO-GPT greedy vs. InterLabelGO+ on weighted MF,  $p = 0.073$ ; Tables S3–S5). (C) **Organism embedding phylogeny.** Dendrogram from cosine similarity of learned organism embeddings (top) compared against a reference phylogeny (bottom). The learned tree recapitulates known phylogenetic relationships, with *D. rerio* (highlighted in red) as the only misplaced organism. (D) **Binding residue attention enrichment.** AUROC vs.  $\log_2$  (fold enrichment) for 63 non-training BioLiP proteins where GO-GPT predicted GO:0003677 (DNA binding). Purple: significant on both Mann–Whitney U and hypergeometric tests ( $p < 0.05$ ); gray: not significant on at least one. (E) **DNA-binding attention analysis.** Per-residue attention when predicting GO:0003677 for three proteins (E1C9K5, P84131, Q5Y812) spanning high to low binding-site overlap. Left: attention profiles with binding sites shaded; right: protein–DNA structures with high-attention residues in red. (F) **Resnik vs. cosine similarity.** Cosine embedding similarity increases monotonically with Resnik semantic similarity (Resnik, 1995) across all aspects (all  $p < 10^{-3}$ , Mantel permutation test). (G) **Cross-aspect NPMI vs. cosine similarity.** Co-annotation NPMI correlates with embedding similarity across aspect pairs (Spearman  $\rho = 0.10$ – $0.17$ ,  $p < 10^{-3}$ , permutation test). (H) **GO term embedding landscape.** UMAP with KDE contours per aspect (MF blue, BP red, CC green), showing aspect-level spatial organization.

---

firms that attended regions cluster proximal to bound DNA. To rule out that this enrichment merely reflects pre-existing structure in the frozen ESM2 embeddings, we compared against ESM2 residue embedding L2 norms as a baseline predictor of binding sites (Section 4.5.2; Fig. S2C–D). ESM2 norms show substantially weaker discrimination, suggesting that cross-attention captures function-specific signal not readily apparent in the input representations.

GO-GPT’s vocabulary consists of individual GO terms as discrete tokens, each with its own learned embedding. To test whether these embeddings capture the semantic organization of the Gene Ontology, we measured the relationship between cosine embedding similarity and Resnik semantic similarity (Resnik, 1995), an information-content measure defined over the GO directed acyclic graph, within each aspect (Fig. 2F). Cosine similarity increases monotonically with Resnik similarity across all three aspects, with a pronounced rise at high Resnik values, indicating that the embeddings particularly distinguish the most semantically related term pairs (all  $p < 10^{-3}$ , Mantel permutation test; Section 4.5.3). Consistently,  $k$ -nearest neighbors in embedding space overlap significantly with ontological neighbors in the GO graph across all aspects (Fig. S6B;  $p < 10^{-3}$ , permutation test).

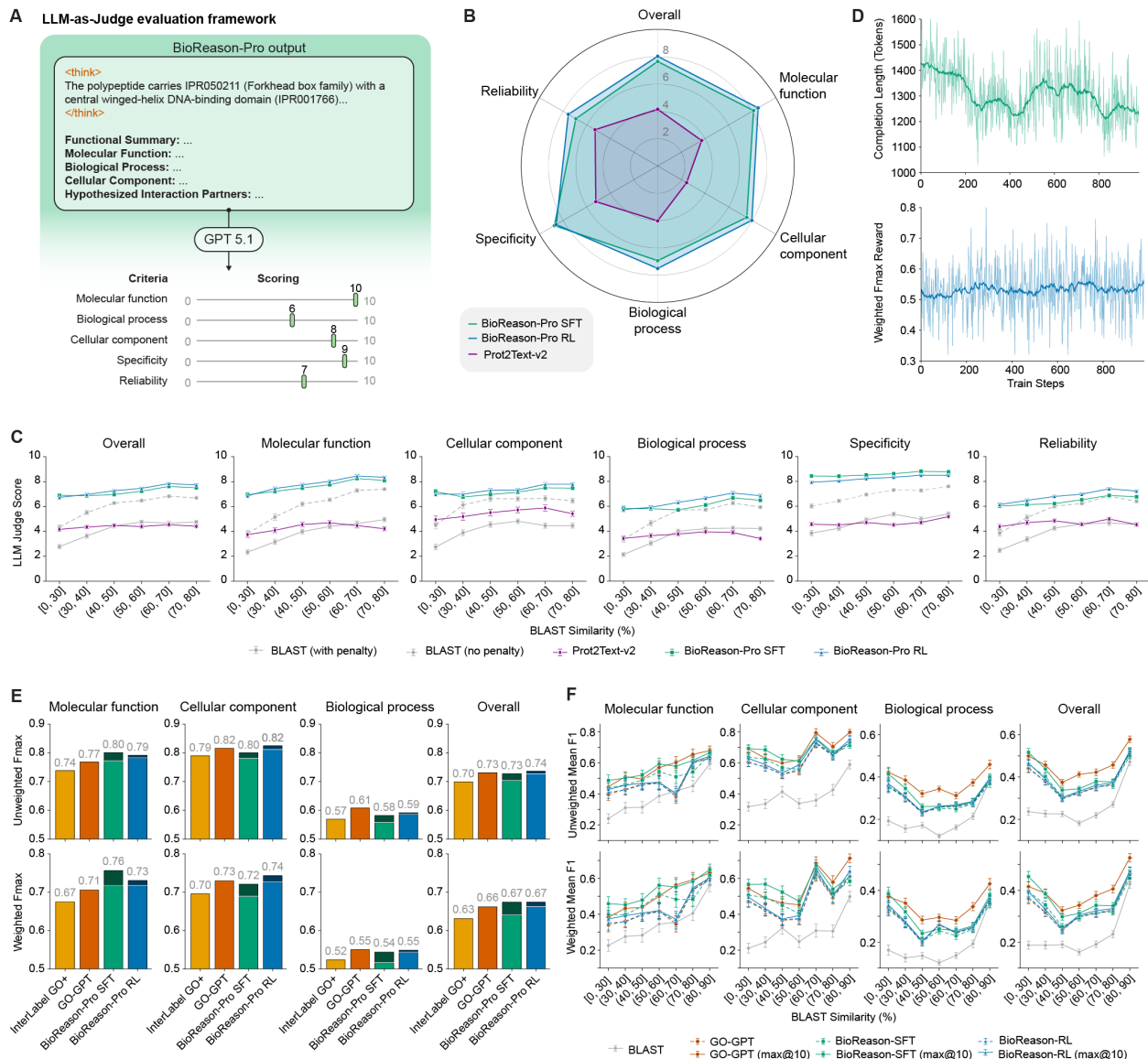
We next asked whether the embeddings also capture relationships that span GO aspects. Normalized Pointwise Mutual Information (NPMI) between cross-aspect GO term pairs (Section 4.5.3) correlates positively with cosine embedding similarity across all three aspect pairs (Spearman  $\rho = 0.10$ – $0.17$ , all  $p < 10^{-3}$ , embedding-permutation test; Fig. 2G), with an upturn at high NPMI values most visible for MF×CC. To test whether this correlation translates into practical retrieval, we ranked partner-aspect terms by cosine similarity for each GO term and evaluated retrieval of strongly co-annotated partners (NPMI > 0.5; Section 4.5.3). Embedding similarity retrieves correct cross-aspect partners well above chance across all three aspect pairs (AUROC = 0.66–0.80, all  $p < 10^{-3}$ ; Precision@1 = 3.5–6× above random baseline), with MF×CC showing the strongest retrieval (AUROC = 0.80; Fig. S7B). Notably, the model receives only numerical GO identifiers and never human-readable term names, so these cross-aspect associations are learned entirely from protein co-annotation patterns. UMAP projection (McInnes et al., 2020) of the embedding space (Fig. 2H) visually reflects both types of organization: MF and CC embeddings form more compact, localized clusters while BP is diffusely spread. This is consistent with quantitative cosine purity analysis showing MF and CC neighborhoods enriched 3–5× above their aspect-specific random baselines ( $p < 10^{-3}$ ) whereas BP shows no enrichment (Fig. S7C). Regions of inter-aspect overlap are also visible; representative cross-aspect pairs with high co-annotation frequency and embedding similarity that happen to co-localize in the projection are annotated in Fig. S8 (Table S2).

Together, these analyses demonstrate that GO-GPT captures functional, evolutionary, and ontological structure within its learned representations. The model grounds its GO term predictions in biologically relevant sequence features and has internalized both phylogenetic relationships and ontological structure without explicit supervision on either.

### 2.3. BioReason-Pro demonstrates generalizable expert-level reasoning on protein function

While GO-GPT generates GO term predictions through internal mechanisms that are not directly accessible and require post hoc analysis to interpret, BioReason-Pro is designed to openly reason across multimodal biological evidence to produce interpretable functional annotations. It analyzes domain architecture, infers molecular function, predicts localization, identifies relevant biological processes, proposes mechanistic explanations, and hypothesizes interaction partners. Standard text generation metrics such as ROUGE and BERTScore are therefore inadequate for evaluating such reasoning, as they measure surface-level overlap rather than biological correctness (Ganesan, 2018; Zhang et al., 2020). We employed GPT-5.1, the most capable model available at the time of evaluation, as an expert judge, scoring predictions from BioReason-Pro SFT, BioReason-Pro RL, and the best prior method Prot2Text-v2 against composite ground truth including UniProt function summaries and GO terms across all three aspects (Fei et al., 2025).

The LLM judge scored each prediction on a 1 to 10 scale across five axes, following the prompt detailed in Section C.3. Molecular Function, Biological Process, and Cellular Component assessed correctness of core functional annotations, Specificity evaluated the level of mechanistic detail provided, and Reliability distinguished logically supported inference from hallucinated claims (Fig. 3A). We evaluated BioReason-Pro SFT, trained on synthetic reasoning traces, and BioReason-Pro RL, further trained with reinforcement learning to optimize GO term accuracy (Section 2.4). BioReason-Pro RL achieved an average score of 8.03 out of 10, with



**Figure 3 | BioReason-Pro evaluation on protein function prediction. (A) LLM-as-Judge framework.** GPT-5.1 evaluated model predictions against composite ground truth comprising UniProt summaries, GO terms, InterPro domains, protein-protein interactions, and subcellular localization. Each prediction was scored on a 1 to 10 scale across five axes: Molecular Function, Biological Process, Cellular Component, Specificity, and Reliability. **(B) LLM-as-Judge scores.** Radar plot comparing per-axis scores for BioReason-Pro RL, BioReason-Pro SFT, and Prot2Text-v2. BioReason-Pro RL achieved an average score of 8.03 across axes, SFT 7.65, and Prot2Text-v2 4.15. **(C) LLM judge generalization across sequence similarity.** LLM judge scores across BLAST similarity bins for BioReason-Pro SFT, BioReason-Pro RL, Prot2Text-v2, and BLAST baselines (with and without penalty), shown for all five evaluation axes and Overall. **(D) Reinforcement learning dynamics.** Completion length in tokens (top) and  $F_{\max}^w$  reward (bottom) over GSP0 training steps. Reward increased while generation length decreased, indicating more focused reasoning. **(E) GO term prediction performance.** Weighted and unweighted  $F_{\max}$  across Molecular Function, Cellular Component, Biological Process, and Overall for InterLabelGO+, GO-GPT, BioReason-Pro SFT, and BioReason-Pro RL (dark shading indicates best-of-10 selection upper bound). **(F) GO term prediction generalization across sequence similarity.** Unweighted (top) and weighted (bottom) mean per-protein F1 scores averaged within BLAST similarity bins for GO-GPT, BioReason-Pro SFT, BioReason-Pro RL, and BLAST, with best-of-10 selection variants shown for each generative model.

---

SFT at 7.65, both substantially outperforming Prot2Text-v2 at 4.15 (Fig. 3B). Across individual axes, Molecular Function yielded the highest scores (RL 8.50, SFT 8.11) while Biological Process scored the lowest (RL 7.51, SFT 6.94). Reinforcement learning improved Reliability from 6.92 to 7.56, indicating that optimizing for annotation accuracy also reduced hallucinations, while SFT scored marginally higher on Specificity (8.74 vs 8.59). All pairwise differences between BioReason-Pro and baselines are statistically significant, as are differences between RL and SFT on all axes (paired Wilcoxon signed-rank tests,  $p < 10^{-15}$ ; full distributions in Fig. S3A, detailed statistics in Section B.5.2, Table S10).

We next asked whether reasoning quality holds for proteins dissimilar from training data. We binned test proteins by best-hit BLAST sequence similarity to the training set and compared LLM judge scores across bins (Fig. 3C) (Altschul et al., 1990). As our BLAST baseline, we transferred the functional annotations of each test protein’s closest training set hit, scoring proteins whose closest hit lacked a UniProt function description as zero (BLAST with penalty) or excluding them entirely (BLAST without penalty). BLAST performance showed strong dependence on sequence similarity (Spearman  $\rho = 0.43$ ), dropping sharply at low similarity. BioReason-Pro exhibited roughly half this dependence (RL  $\rho = 0.27$ , SFT  $\rho = 0.21$ ), maintaining scores of 7–8.5 across the full similarity range, with OLS regression confirming that its advantage over BLAST grows significantly with decreasing similarity ( $\beta_1 = -0.019$ ,  $p = 3.1 \times 10^{-30}$ ; Section B.5.3, Table S11). Even at high similarity bins where close homologs including orthologs were readily available, BioReason-Pro outperformed BLAST.

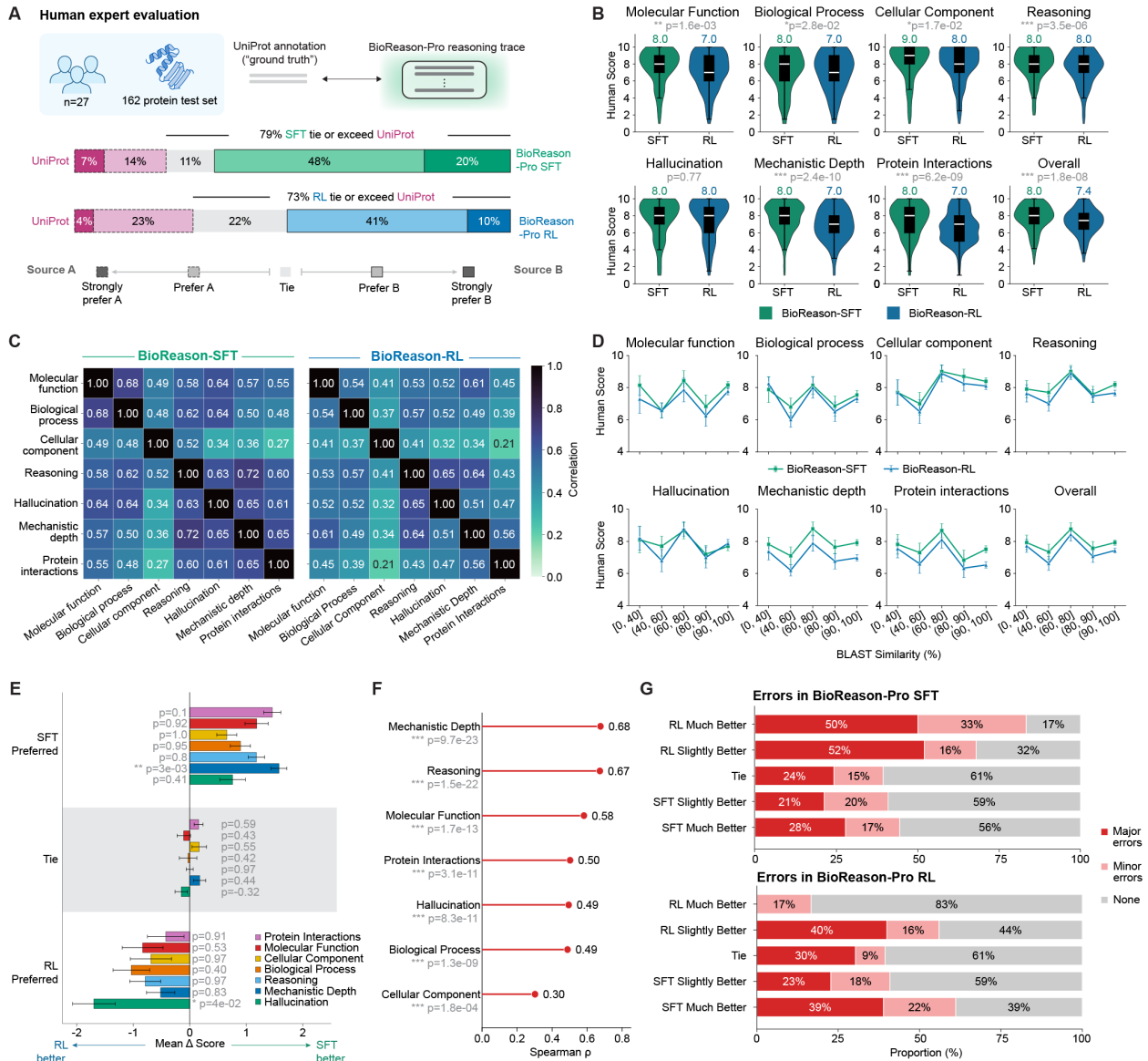
Beyond sequence similarity, BioReason-Pro performance remained stable across protein lengths and GO annotation counts (Fig. S3B,C). Scores decreased with fewer InterPro domain annotations, but BioReason-Pro still outperformed all baselines even for proteins with zero or few annotated domains (Fig. S3D). Performance also generalized across diverse organisms and taxonomic classes rather than concentrating around well-studied model species (Fig. S4). Together, these results demonstrate that BioReason-Pro learned generalizable functional reasoning rather than retrieving annotations from similar training proteins.

#### 2.4. Reinforcement learning produces more accurate and concise GO term predictions

While supervised fine-tuning taught BioReason-Pro to generate biological reasoning traces, reinforcement learning optimized those traces for GO term prediction accuracy. Over training, the  $F_{\max}^w$  reward between predicted and ground truth GO terms increased while reasoning traces became shorter and more focused, suggesting the model learned to eliminate verbose reasoning that did not contribute to correct predictions (Fig. 3D). Per-protein comparison confirmed that RL produced significantly shorter traces than SFT for the same proteins (mean  $\Delta = 60.0$  words, paired Wilcoxon  $p < 10^{-300}$ ; Fig. S3E,F). Notably, RL increasingly outperformed SFT at shorter reasoning trace lengths, indicating that reinforcement learning produced more efficient reasoning rather than simply shorter output (Fig. S3G).

We evaluated GO term prediction using  $F_{\max}^w$ , comparing BioReason-Pro against GO-GPT and InterLabelGO+ (Fig. 3E) (Evans and Shen, 2024). BioReason-Pro SFT achieved  $F_{\max}^w = 0.64$  on single-run greedy decoding and 0.67 with best-of-10 sampling, revealing substantial generation diversity. Reinforcement learning closed this gap, with BioReason-Pro RL achieving 0.66 on single-run greedy decoding while best-of-10 sampling remained at 0.67, indicating that RL concentrated probability mass on correct predictions. The best unweighted  $F_{\max}$  sits at 0.74. Both BioReason-Pro variants surpassed InterLabelGO+ at 0.63 and single-run GO-GPT at 0.65 on  $F_{\max}^w$ , with consistent improvements across unweighted metrics.

We then examined how GO term prediction accuracy varied with sequence similarity by binning test proteins according to best-hit BLAST sequence similarity (Fig. 3F) (Altschul et al., 1990). Because  $F_{\max}$  cannot be decomposed across subsets of proteins, we report mean per-protein F1 scores within each bin to enable direct comparison across similarity ranges. BLAST showed the strongest similarity dependence (Spearman  $\rho = 0.63$ ), while BioReason-Pro and GO-GPT exhibited substantially weaker dependence (SFT  $\rho = 0.41$ , RL  $\rho = 0.46$ , GO-GPT  $\rho = 0.46$ ), with OLS regression confirming that all generative models maintained significantly greater advantage over BLAST at low similarity (all  $p < 10^{-130}$ ; Section B.5.3, Table S12). Even at high similarity where close homologs were available, BioReason-Pro outperformed BLAST, demonstrating that biological reasoning captured functional relationships beyond sequence alignment alone.



**Figure 4 | Human expert evaluation of BioReason-Pro predictions.** (A) Pairwise preference comparisons. 27 protein experts compared BioReason-Pro SFT and RL predictions against UniProt ground truth on a five-point preference scale. SFT tied or exceeded UniProt in 79% of cases, RL in 73%. (B) Per-axis score distributions. Violin plots of expert scores for BioReason-Pro SFT and RL across Molecular Function, Biological Process, Cellular Component, Reasoning, Hallucination, Mechanistic Depth, Protein Interactions, and Overall, with median scores indicated. (C) Correlation structure of evaluation axes. Pairwise correlations between the seven evaluation axes for SFT and RL, showing that Reasoning, Mechanistic Depth, and Hallucination form a correlated cluster reflecting overall biological reasoning quality. (D) Human evaluation across sequence similarity. Expert scores across BLAST sequence similarity bins for BioReason-Pro SFT and RL across all seven evaluation axes and Overall, showing stable performance with no degradation at low similarity. (E) Score deltas by preference group. Mean per-axis score difference (SFT – RL) stratified by expert preference. SFT advantage is driven by Mechanistic Depth ( $p = 3 \times 10^{-3}$ ), while RL advantage is driven by Hallucination ( $p = 4 \times 10^{-2}$ ). (F) Axes driving expert preference. Spearman correlation between per-axis score deltas and expert preference, identifying Mechanistic Depth and Reasoning as the primary axes driving preference decisions. (G) Error attribution by expert preference. Proportion of major, minor, or no errors (classified by GPT-5-mini from free-text responses) in each model, stratified by expert head-to-head preference (Section C.5).

---

## 2.5. Human experts prefer BioReason-Pro over ground truth annotations

Although the results from the automated metrics and LLM judges are promising, it is difficult to extrapolate from these to whether the annotations are genuinely useful to domain experts. We therefore conducted a blinded human evaluation in which 27 molecular biologists assessed BioReason-Pro predictions on 162 randomly selected test proteins (Section C.4). Evaluators were blinded to model identity (SFT or RL) and had access to external sources and recent literature. Evaluators performed pairwise preference comparisons between each model and ground truth UniProt annotations on a five-point scale from significantly falls short to significantly exceeds. BioReason-Pro SFT achieved a 79% tie-or-exceed rate against UniProt ground truth, with RL at 73%, meaning human experts found model predictions to match or exceed curated database entries in the majority of cases (Fig. 4A). This shows that RL improved GO term accuracy (Section 2.4) without significantly reducing the reasoning quality (McNemar’s exact test,  $p = 0.087$ ).

Evaluators also scored each prediction on a 1 to 10 scale along seven axes. Molecular function, biological process, and cellular component assessed correctness of core functional annotations; reasoning evaluated evidence attribution; hallucination measured avoidance of fabricated biological claims; mechanistic depth measured accuracy of proposed molecular mechanisms; and protein-protein interactions evaluated plausibility of predicted binding partners. BioReason-Pro SFT averaged 8.0 out of 10 overall with RL at 7.4 (Fig. 4B; full distributions in Fig. S5A). Reasoning, Mechanistic Depth, and Hallucination scores were strongly correlated with each other across both models, indicating that predictions with well-supported reasoning also tended to avoid hallucinations and provide deeper mechanistic insight (Fig. 4C; Fig. S5B).

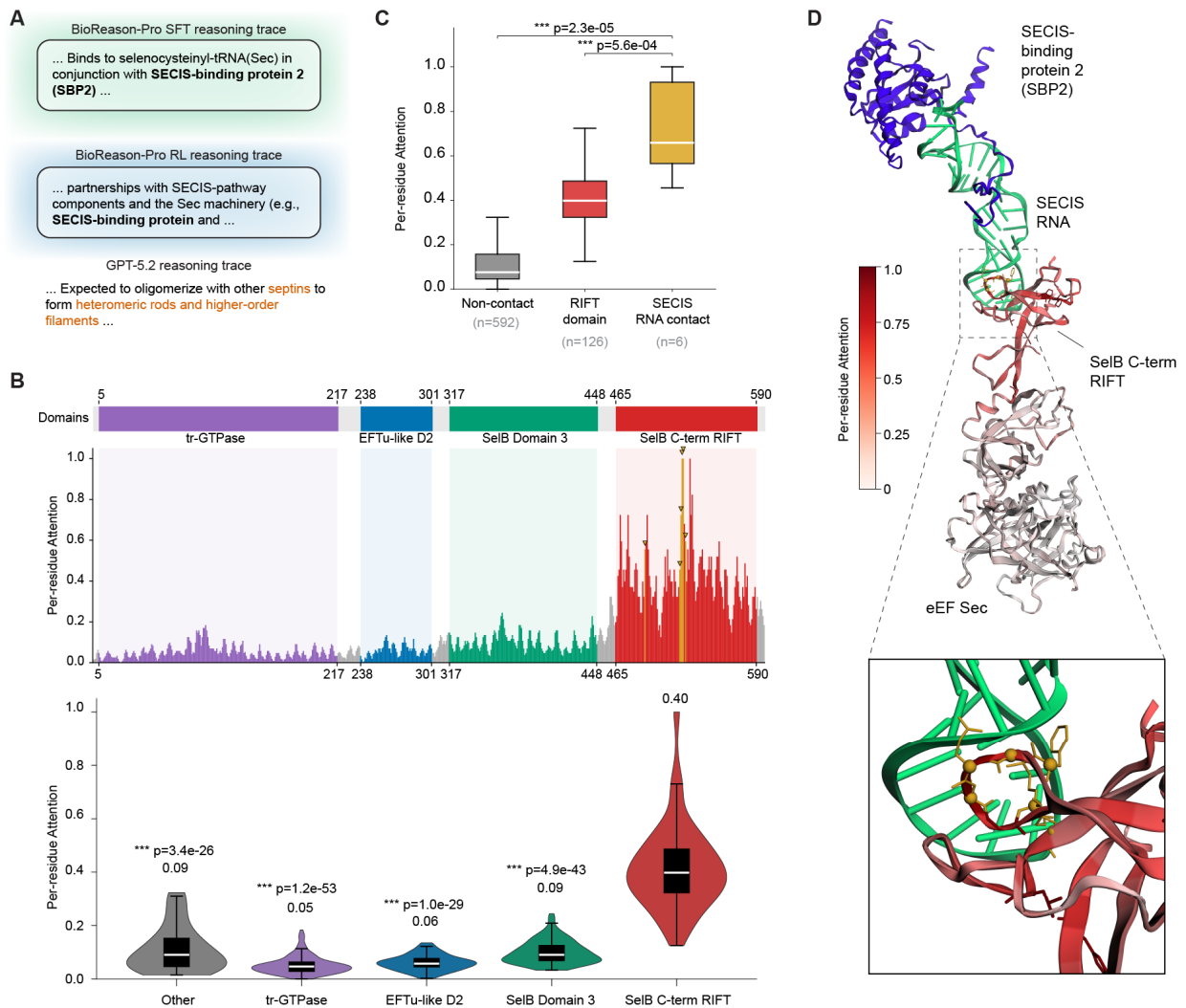
We examined whether human-perceived annotation quality depends on sequence similarity to training data by comparing scores across BLAST sequence identity bins (Fig. 4D) (Altschul et al., 1990). Neither model showed significant similarity dependence on any evaluation axis (Spearman  $|\rho| < 0.06$  for overall scores, all  $p > 0.05$ ; Section B.5.4, Table S13), and the score gap between RL and SFT was also stable across the similarity range (all  $p > 0.05$ ; Table S14). Win rates against UniProt did not depend on similarity for either model (logistic regression,  $p = 0.71$  for SFT,  $p = 0.19$  for RL; Table S15; Fig. S5H).

To understand what drives expert preference between SFT and RL, we examined per-axis score differences stratified by preference group. When experts preferred SFT, the score advantage was concentrated in Mechanistic Depth ( $p = 3 \times 10^{-3}$ ), while when experts preferred RL, the advantage was driven by fewer hallucinated claims ( $p = 4 \times 10^{-2}$ ; Fig. 4E). Mechanistic Depth and Reasoning showed the highest overall correlation with expert preference decisions (Fig. 4F). Error attribution of free-text expert responses confirmed this pattern, with SFT-preferred cases showing richer mechanistic detail and RL-preferred cases containing fewer major errors (Fig. 4G). Where experts preferred RL, free-text error reports revealed that SFT made specific mechanistic errors such as enzyme misidentification and inverted pathway directionality, whereas RL’s more conservative outputs avoided these hallucinations. This suggests that RL trades some mechanistic depth for improved factual reliability.

Beyond sequence similarity, performance generalized across diverse organisms (Fig. S5C) and remained stable across protein lengths, GO annotation counts, InterPro domain counts, and reasoning trace lengths (Fig. S5D,E,F,G). Together, these results demonstrate that BioReason-Pro produces expert-level functional reasoning across a broad range of proteins regardless of similarity to training data.

## 2.6. BioReason-Pro generates de novo predictions validated by experimental structures

We asked whether BioReason-Pro could synthesize specific mechanistic predictions from its input context alone and whether such predictions are grounded in structurally meaningful features. We screened test set generations for cases in which the model proposed specific interaction partners or pathway-level mechanisms that could be validated against experimental structural data. One such case is eEFSec (P57772), the selenocysteine-specific elongation factor, a four-domain translational GTPase that delivers the 21st amino acid, selenocysteine, to ribosomes at recoded UGA stop codons. Unlike canonical elongation factors such as eEF1A that service all standard aminoacyl-tRNAs through a shared mechanism, eEFSec operates through a non-canonical pathway requiring coordination with a dedicated mRNA stem-loop element and an accessory protein that no other translation factor uses (Copeland et al., 2000; Hilal et al., 2022). The closest training set match shared 44% sequence identity, a regime where homology transfer may capture general family membership but cannot resolve the mechanistic specializations that distinguish eEFSec from other EF-Tu-like GTPases.



**Figure 5 | eEFSec (P57772), the selenocysteine-specific elongation factor. (A) Reasoning trace comparison.** Key passages from BioReason-Pro SFT, BioReason-Pro RL, and GPT-5.2 Thinking High generations. BioReason-Pro SFT de novo predicted SBP2 as the obligate partner (highlighted), while GPT-5.2 misidentified the protein as a septin-family GTPase. **(B) Domain architecture and per-residue attention.** Top: canonical EF-Tu machinery (Domains 1–2) distinguished from selenocysteine-specific elements (Domains 3–4). Middle: per-residue attention from the token preceding the predicted partner name mapped onto the sequence, with yellow bars marking SECIS RNA contact residues ( $\leq 5.0$  Å in PDB 7ZJW). Bottom: violin plots comparing attention distributions across domains, with attention significantly enriched at the C-terminal RIFT domain relative to all other domains (Mann–Whitney  $U$ ). **(C) Contact interface enrichment.** Per-residue attention scores compared across non-contact residues, RIFT domain residues, and SECIS RNA contact residues (Mann–Whitney  $U$ ,  $p = 2.3 \times 10^{-5}$ , PDB 7ZJW). **(D) Structural context.** Cryo-EM selenosome structure (PDB 7ZJW) with eEFSec colored by per-residue attention (white to red), showing highest attention at the RIFT domain surface overlapping the SECIS RNA binding interface. The de novo predicted partner SBP2 (purple) and SECIS RNA (green) are labeled. Inset: zoomed view of the RIFT domain contact interface.

---

BioReason-Pro processed domains spanning the N-terminal GTPase core through the C-terminal RIFT module. The reasoning traced how the translational-type GTP-binding domain and EF-Tu-like beta-barrel establish canonical translational GTPase machinery, then identified the C-terminal selenocysteine-specific domains (IPR049393, IPR049394) as restricting this factor to a single substrate (Gonzalez-Flores et al., 2012). Rather than assigning generic tRNA binding, the model described selenocysteine-specific tRNA recognition as the defining molecular function (full traces in Fig. 5A and Section C.6), a functional distinction that separates eEFSec from all other EF-Tu family members but that the Gene Ontology has not formalized into a dedicated term. Biochemical studies support this specificity, showing that eEFSec binds only selenocysteinyl-tRNA(Sec) and rejects both the serylated precursor and canonical aminoacyl-tRNAs (Fagegaltier, 2000).

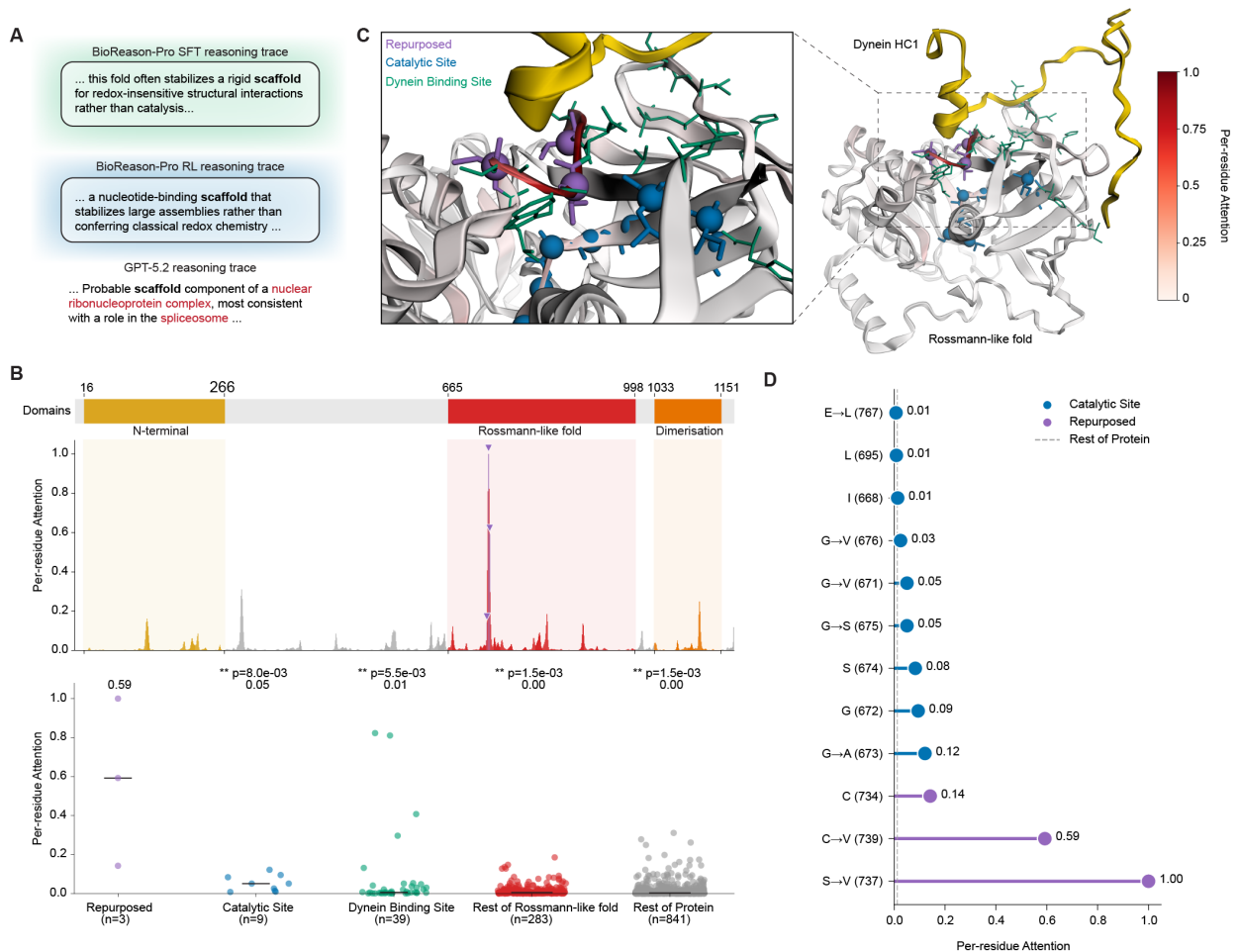
Selenocysteine incorporation requires coordination between the elongation factor, a dedicated mRNA stem-loop in the 3'-UTR (SECIS), and an accessory protein that bridges these elements at the ribosome (Caban and Copeland, 2006). When no protein interactions were provided in context, BioReason-Pro SFT reconstructed this three-component logic from architectural constraints alone, reasoning that the C-terminal RIFT domain implies an escorting factor that connects the SECIS element to the decoding site, and identified SECIS-binding protein 2 (SBP2) as this partner (Fig. 5A). Indeed, SBP2 is the most specific functional partner of eEFSec, confirmed by co-immunoprecipitation and visualized in the 2.8 Å cryo-EM selenosome structure, which revealed that the interaction is mediated by the SECIS element (Copeland et al., 2000; Hilal et al., 2022). BioReason-Pro RL also identified SECIS-binding protein as a partner (Section C.7). In contrast, GPT-5.2 Thinking High misidentified this protein entirely as a septin-family GTPase and predicted cytokinesis and cytoskeletal scaffolding functions (Section C.8).

To investigate whether this de novo prediction is grounded in the protein representation, we examined BioReason-Pro SFT's per-residue attention from the token immediately preceding the predicted partner name to the protein sequence, since this variant produced the most specific partner prediction (Fig. 5B). Attention localizes sharply to the RIFT domain, with all other domains showing significantly lower scores (tr-GTPase  $p = 1.2 \times 10^{-53}$ , EFTu-like D2  $p = 1.0 \times 10^{-29}$ , SelB Domain 3  $p = 4.9 \times 10^{-43}$ ; Fig. 5B). At residue resolution, attention scores at SECIS RNA contact residues are significantly elevated relative to non-contact residues (Mann-Whitney  $U$ ,  $p = 2.3 \times 10^{-5}$ ; Fig. 5C) (Mann and Whitney, 1947). Projection onto the cryo-EM selenosome structure (PDB 7ZJW) confirms that the highest-attention surface on the RIFT domain coincides with the SECIS RNA binding interface (Fig. 5D) (Hilal et al., 2022). The model thus predicted the principal functional partner of eEFSec from sequence alone, and its internal representation concentrated on the precise binding surface that structural biology identified through the selenosome reconstruction. The model further reconstructed the full selenocysteine incorporation cycle with mechanistic accuracy, describing GTP-coupled A-site docking at UGA codons, SECIS-programmed context-dependent decoding, and GTP hydrolysis triggering factor dissociation to commit the charged tRNA to peptide bond formation (Simonović and Puppala, 2018; Hilal et al., 2022).

This case demonstrates that BioReason-Pro can generate predictions that exceed existing curated annotations in both specificity and structural grounding. The model de novo identified SBP2 as the obligate functional partner, a prediction validated by the cryo-EM selenosome structure (Hilal et al., 2022; Dobosz-Bartoszek et al., 2016). The significant overlap between attention and the experimentally resolved contact interface provides direct evidence that these predictions are rooted in structurally meaningful protein features. These results suggest that explicit chain-of-thought reasoning over learned protein representations can produce functional insights that refine and extend expert-curated database entries.

## 2.7. BioReason-Pro performs structural reasoning beyond domain annotation transfer

A central question for BioReason-Pro is whether it performs genuine reasoning or merely restates domain annotations in natural language. To evaluate this, we systematically screened test set generations for cases in which naive domain interpretation would predict incorrect function but the model arrived at the correct annotation through contextual reasoning. One such case is the Cilia- and flagella-associated protein 61 (CFAP61, UniProt Q8NHU2), a 1,237-residue axonemal protein whose closest training set match shares only 29% sequence identity, well below the threshold where homology transfer provides reliable annotations (Rost, 1999). This protein presents a specific interpretive challenge because its domain architecture contains a Rossmann-like FAD/NAD(P)-binding superfamily fold (IPR036188) that would cue enzymatic or oxidoreductase function in standard domain-lookup pipelines (Heuser et al., 2012; Blum et al., 2024). Indeed, some databases incorrectly



**Figure 6 | CFAP61 (Q8NHU2), a non-enzymatic axonemal scaffold with an exapted active site. (A) Reasoning trace comparison.** Key passages from BioReason-Pro SFT, BioReason-Pro RL, and GPT-5.2 Thinking High generations, with correct inferences highlighted in green and incorrect inferences in red. BioReason-Pro correctly identified non-enzymatic scaffolding function, while GPT-5.2 misidentified the protein as a nuclear spliceosome-associated HEAT-repeat protein. **(B) Domain architecture and attention enrichment.** Top: domain architecture with boundaries indicated. Middle: per-residue attention from the token preceding the non-enzymatic scaffold prediction mapped onto the sequence. Purple triangles mark the three repurposed residues (C734, V737, V739). Bottom: per-residue attention compared across repurposed active-site residues, remaining catalytic-site residues, dynein binding site residues, rest of the Rossmann domain, and rest of the protein (Mann–Whitney  $U$  tests between adjacent categories). **(C) Structural context.** Cryo-EM axoneme structure (PDB 8J07) with the Rossmann domain shown in two views. Left: close-up with repurposed residues (purple), degenerate catalytic positions (blue), and Dynein Heavy Chain 1 binding interface (yellow) shown as spheres. Right: broader structural context with per-residue attention projected onto the molecular surface (white to red), showing highest attention concentrated at the three repurposed residues. Dynein Heavy Chain 1 is labeled. **(D) Per-residue attention at catalytic positions.** Attention scores for individual catalytic (blue) and repurposed (purple) positions, ranked by attention magnitude. Dashed line indicates the median attention across the rest of the protein as a baseline reference.

---

list oxidoreductase activity as a Gene Ontology annotation for this protein (Stelzer et al., 2016). However, experimental characterization has established CFAP61 as a non-enzymatic scaffold of the calmodulin- and radial-spoke-associated complex, essential for sperm flagellum formation (Liu et al., 2021; Ma et al., 2022).

BioReason-Pro resolves this ambiguity through contextual architectural reasoning rather than isolated domain interpretation. The model observes that the Rossmann-like fold is situated between an N-terminal axonemal targeting module (IPR032151) and a C-terminal dimerization domain (IPR056299), all within a cilia-specific protein family (IPR038884) (Blum et al., 2024). From this context, it infers that the Rossmann-like domain contributes a stable structural core for redox-insensitive scaffolding interactions rather than catalysis (Fig. 6A). This contextual override of a superfamily-level annotation, where interpretation is conditioned on flanking domains, overall protein architecture, and family membership, demonstrates reasoning that cannot be replicated by domain lookup or homology transfer alone (full trace in Section C.9). BioReason-Pro RL also correctly identified the non-catalytic scaffolding role but with less functional specificity (Section C.10). In contrast, GPT-5.2 Thinking High misidentified CFAP61 as a nuclear spliceosome-associated HEAT-repeat protein and predicted mRNA splicing and nuclear localization (Section C.11).

To determine whether there is a structural basis for BioReason-Pro’s inferences, we aligned the CFAP61 Rossmann domain against glutathione reductase (PDB 3GRS) (Karpus and Schulz, 1987) and identified 12 positions at catalytic and cofactor-binding sites. All 12 are degenerate in CFAP61. The GxGxxG dinucleotide-binding loop is disrupted (VGASSV at positions 671–676); the catalytic nucleophile CysI is replaced by Val at position 739; the FAD-stacking Glu is replaced by Leu at position 767; and the cofactor-binding Ser is replaced by Val at position 737. Three of these former catalytic positions (C734, V737, V739) now form direct contacts with dynein heavy chain 1 (DHC1) in the cryo-EM axoneme structure (PDB 8J07) (Walton et al., 2023), with the ancestral cofactor-binding pocket occupied by the dynein interface. We examined per-residue attention from the token at which the model predicts scaffolding over catalytic function (Fig. 6B). Attention is significantly enriched at the three repurposed residues relative to remaining catalytic positions ( $p = 8.0 \times 10^{-3}$ ), dynein binding site residues ( $p = 5.5 \times 10^{-3}$ ), and the rest of the Rossmann domain ( $p = 1.5 \times 10^{-3}$ ; Mann-Whitney  $U$ ; Fig. 6B) (Mann and Whitney, 1947). Projection onto the cryo-EM structure confirms that the highest-attention surface on the Rossmann domain coincides with the dynein binding interface (Fig. 6C). At single-residue resolution, S→V at position 737 and C→V at position 739 showed the highest attention scores, both of which are direct DHC1 contacts (Fig. 6D). This suggests that the model has learned structural reasoning, favoring protein interaction over catalytic activity at the precise positions where sequence variation supports this functional shift.

The model also identified CFAP61 as a non-enzymatic axonemal scaffold involved in cilium assembly (GO:0060271) and cilium movement (GO:0003341), localizing it to the axoneme (GO:0005930) and motile cilium (GO:0031514). It predicted protein dimerization activity (GO:0046983) from the C-terminal dimerization domain and structural molecule activity (GO:0005198), both consistent with the experimentally established role of CFAP61 as a radial-spoke-associated factor required for axonemal integrity (Liu et al., 2021; Ma et al., 2022). BioReason-Pro also predicted specific binding partners including axonemal dynein subunits, radial spoke proteins, and tubulins. Dynein heavy chains are direct structural contacts of the CFAP61 Rossmann domain in the cryo-EM axoneme reconstruction (PDB 8J07) (Walton et al., 2023), while RSPH9 and TUBB3 are validated co-complex members by co-immunoprecipitation (Liu et al., 2021).

This case demonstrates reasoning at two scales. At the architectural level, the model conditions domain interpretation on flanking context and family membership to override a misleading superfamily annotation. At the residue level, attention concentrates on degenerate catalytic positions that have been evolutionarily repurposed as a protein-protein interaction surface. These results suggest that explicit chain-of-thought reasoning over learned protein representations captures structural and evolutionary features sufficient to extend functional characterization beyond existing database annotations, even for proteins far outside the reach of homology-based methods.

## 2.8. Pushing BioReason-Pro to its limits

To characterize the boundaries of BioReason-Pro’s reasoning capabilities, we evaluated both the SFT and RL models on short proteins and peptides. This represents a challenging setting, as fewer than 0.5% of training sequences fell below 50 amino acids. The evaluation comprised 22 sequences spanning 12 to 977 amino acids across metabolic, immune, antimicrobial, and venom categories, assessed by two independent domain

---

experts. Model performance on full-length proteins including HLA-A\*02:01, p53, full-length BRINP2, and IA-2 was consistent with the main benchmarks, accurately capturing domain architecture, molecular function, and biological process annotations. Performance on moderately sized sequences such as preproinsulin (110 aa), CCK (115 aa), and GDF15 (112 aa) was mixed. The model identified the correct protein families but consistently missed physiologically important functions such as feeding behavior, specific receptor partners, and appetite suppression. Below approximately 50 amino acids, performance degraded systematically.

Short peptide failures fell into three consistent categories. First, when short peptides lacked recognizable domains, the SFT model fabricated InterPro entries, assigning a myosin family domain to a GAD65 epitope, an SPFH domain to an IA-2 epitope, and ARID4B to a proinsulin junction peptide. In each case, the model built elaborate but baseless mechanistic narratives on these fabricated domains. The RL model never fabricated domains and instead explicitly acknowledged the absence of recognizable signatures. Second, the model confused closely related peptide family members when sequence length was too short. GLP-1 (30 aa) was conflated with glucagon, resulting in predictions of negative regulation of insulin secretion. Third, the model reproduced training-era annotations faithfully but could not identify functions characterized after the data cutoff. Full-length BRINP2 was accurately characterized as a nuclear transcriptional repressor, consistent with its training annotations, but its recently discovered role as a prohormone precursor for BRP (Coassolo et al., 2025) was not identified. When the 12-amino-acid BRP peptide was tested independently, it predicted generic transcriptional regulator annotations with no connection to appetite regulation or GPCR signaling.

A natural application of BioReason-Pro is functional annotation of proteins designed by generative models such as RFDiffusion (Watson et al., 2023) or the Evo series (Nguyen et al., 2024; Brix et al., 2026). As a showcase, we selected two AI-generated anti-CRISPR (Acr) proteins, EvoAcr1 (169 aa) and EvoAcr2 (157 aa), designed by prompting Evo 1.5 with known Cas9-targeting Acr operons (Merchant et al., 2025). Five of the resulting designs blocked Cas9 activity in *E. coli*, and EvoAcr1 and EvoAcr2 were chosen because they lack sequence homology to any known protein by BLAST and yielded only low-confidence structural predictions from AlphaFold3 (Abramson et al., 2024). These proteins have no InterPro domains and no representation in any training database. Because they are synthetic and have no natural species, organism assignment is ambiguous. We used *E. coli* as the primary organism since both proteins are functional in that host, and additionally tested specific strains (K12, O157:H7) and *Homo sapiens* to assess organism sensitivity. Full reasoning traces are in Sections C.12–C.15.

Both models generated detailed reasoning traces with mechanistic narratives and testable hypotheses despite the complete absence of homology or domain evidence. As noted above for the peptides, the SFT model also fabricated InterPro entries. That said, several predictions by both models were biologically coherent. For EvoAcr2 in generic *E. coli*, the SFT model described a phage-encoded effector that binds the nucleoid-associated proteins H-NS and HU to remodel chromosomal architecture. For EvoAcr2 in O157:H7, the RL model predicted a soluble phage-encoded adaptor that operates in the host cytoplasm, using multivalent binding to sequester host factors and reorganize protein complexes governing transcriptional output, ultimately dampening host gene expression to favor a viral life cycle. Both predictions are consistent with phage-encoded effectors that modulate host biology (Pawluk et al., 2017; Hwang and Maxwell, 2023). Why these phage-related functions were predicted specifically for the pathogenic strain remains unclear.

For these AI-generated proteins, predictions varied substantially across organism labels for the same sequence. The same EvoAcr1 sequence was predicted as a ribosomal protein in generic *E. coli*, a DNA-binding transcriptional repressor in K12, a host-pathogen membrane effector in O157:H7, and a translation initiation factor in human. EvoAcr2 showed similar divergence, ranging from an acid stress resistance protein in K12 to a peroxiredoxin in human. This strain-level sensitivity has a biological basis, as *E. coli* O157:H7 expresses over 1300 proteins not found in K12, many related to virulence (Da Silva et al., 2018), likely shaping a bias toward pathogenesis-related predictions. GO-GPT showed the same organism-driven divergence upstream, indicating that the effect propagated through the full pipeline. While the strong influence of organism context on predictions for novel proteins remains a limitation, BioReason-Pro was clearly capable of nominating plausible and testable molecular functions.

---

### 3. Discussion

Here we introduce GO-GPT and BioReason-Pro, two models that together establish expert-level protein function prediction through structured biological reasoning. GO-GPT is the first autoregressive transformer for Gene Ontology prediction, achieving state-of-the-art  $F_{\max}^w$  by capturing hierarchical and cross-aspect dependencies that discriminative methods miss. BioReason-Pro builds on GO-GPT as the first multimodal reasoning LLM for protein function prediction, deeply integrating protein representations with biological context to generate interpretable reasoning traces that progress from domain analysis to functional hypotheses. Human protein experts preferred BioReason-Pro annotations over curated UniProt entries in 79% of evaluated cases, and the model maintained strong performance even for proteins with low sequence similarity to training data, demonstrating generalizable functional reasoning. Together, these models open a path toward scalable characterization of the millions of proteins across all domains of life that lack functional annotation.

The computational biology community has invested heavily in foundation models that encode biological information, from protein language models such as the ESM family (Hayes et al., 2024) to structure predictors like AlphaFold (Fleming et al., 2025), yet comparatively little effort has addressed how to reason over these representations to perform the integrative synthesis that expert biologists conduct. BioReason-Pro reasons from structural and evolutionary principles across protein representations and biological context, enabling generalization across the full range of sequence similarity. The model’s advantage over BLAST widened with decreasing sequence identity, with stable performance across diverse organisms and protein lengths. Even at high similarity where close homologs were readily available, BioReason-Pro outperformed BLAST, indicating that biological reasoning captures functional relationships beyond sequence alignment alone.

A key capability demonstrated by BioReason-Pro is structural and mechanistic reasoning that extends beyond existing database annotations, down to individual residue resolution. For eEFSec, the model predicted SBP2 as the obligate functional partner, a protein that bridges eEFSec to the SECIS mRNA element through the RIFT domain (Hilal et al., 2022). Notably, per-residue attention at the partner prediction token localized to this contact interface in the cryo-EM selenosome structure. The model also accurately reconstructed the full selenocysteine incorporation cycle (Simonović and Puppala, 2018; Hilal et al., 2022), predicting selenocysteine-specific tRNA recognition as the defining molecular function, a distinction from all other EF-Tu family members that the Gene Ontology has not yet formalized into a dedicated term (Ashburner et al., 2000). CFAP61 presents the complementary case, where the Rossmann-like fold contains a catalytic site that typically confers enzymatic activity (Karplus and Schulz, 1987), yet these positions are evolutionarily degenerate, with three now forming direct contacts with DHC1 in the cryo-EM axoneme structure (Walton et al., 2023). BioReason-Pro recognized this degeneration by specifically attending to the three repurposed dynein contact residues and predicted a non-enzymatic scaffolding role consistent with experimental characterization.

Reinforcement learning improved GO term accuracy and reduced hallucinated claims while producing shorter and more focused reasoning traces, without significantly reducing overall expert preference. Where individual preferences diverged, the separation was driven by mechanistic depth favoring SFT and hallucination avoidance favoring RL. Optimizing for ontology accuracy disciplines the model against fabricated biological claims but may limit the depth of mechanistic hypotheses the model generates. This tension likely reflects the reward signal itself, as  $F_{\max}^w$  captures ontology-level correctness but does not measure the mechanistic detail or explanatory quality that distinguishes expert-level annotation. Designing reward signals and evaluation frameworks that capture biological reasoning quality beyond ontology metrics remains an important open problem for the field.

BioReason-Pro has several important limitations. The model was trained on synthetic reasoning traces generated by GPT-5 (Singh et al., 2025), which may contain subtle reasoning errors that propagate into the model. Furthermore, training requires proteins with experimental GO annotations, a resource that remains costly and limited in throughput to produce. Reasoning quality is heavily influenced by the availability of recognizable protein domains and degrades for proteins that lack identifiable InterPro annotations (Blum et al., 2024). Performance also degrades for extremely short peptides below 50 amino acids, where limited sequence information constrains the model’s ability to ground its functional predictions. For synthetic proteins that lack identifiable domains, such as the EvoAcr sequences (Merchant et al., 2025), predictions become heavily dependent on the organism label, producing divergent functional annotations and interaction partners across organisms for the same sequence. That said, several of these predictions were biologically coherent with known phage-encoded effector biology, suggesting that BioReason-Pro can nominate plausible hypotheses

---

even in this challenging regime. LLM-based evaluation with GPT-5.1 may harbor systematic biases, and human expert evaluation covered 162 proteins, a sample size that limits statistical power for fine-grained comparisons. The model is also computationally expensive, requiring sequential inference through ESM3 (Hayes et al., 2024), GO-GPT, and the reasoning LLM. Finally, whether BioReason-Pro learns genuine biological reasoning or sophisticated imitation of reasoning patterns remains an open scientific question.

Protein databases contain millions of sequences with unknown or computationally inferred function, and experimental characterization remains slow and expensive. BioReason-Pro offers a path forward by taking a protein sequence and reasoning about its function, serving as a first-pass annotator for newly sequenced proteins, a hypothesis generator that proposes testable mechanisms, or a quality-check system for existing database annotations. This is particularly valuable for metagenomic proteins, disease-associated proteins that lack mechanistic characterization, and sequences lacking close homologs where traditional annotation transfer fails. The reasoning traces expose the logic underlying each prediction, allowing researchers to assess whether conclusions about protein function are biologically sound and to engage with model outputs critically rather than accepting or rejecting them wholesale.

In this work, we showed that AI systems can reason about protein function with sufficient fidelity that human experts prefer model outputs over curated database annotations. Proteins do not act in isolation, and extending this reasoning to cellular networks may enable models that predict how perturbations propagate through biological systems, how mutations cause disease, and how interventions restore health. The convergence of foundation models that encode biological knowledge with language models capable of reasoning over that knowledge establishes a new paradigm for computational biology, one where AI systems serve as genuine partners in biological discovery. We release all model weights, training code, and curated datasets alongside a [web interface](#) for broad accessibility (see [Code and Model Availability](#) for details). We also provide pre-computed predictions for over 240,000 proteins including the Human Protein Atlas to facilitate immediate application. We hope these materials enable the community to extend, critique, and build upon this work.

## 4. Methods

### 4.1. Datasets

#### 4.1.1. Dataset Pipeline

All protein annotations and auxiliary biological features used in this study were derived from publicly available sources and follow the same curation criteria as the CAFA5 benchmark (Friedberg et al., 2023b,a). The dataset integrates multiple biological modalities, including protein sequence, structure, domain, family, and interaction features, into a single coherent representation, with detailed statistics presented in [Fig. S1](#).

Protein-level GO term annotations were obtained from the Gene Ontology Annotation (GOA) database (Consortium, 2021) (November 2022 release), which provides comprehensive mappings between UniProt accessions and GO terms under diverse evidence codes. Following CAFA standards, only annotations supported by experimental or curated evidence codes (IDA, IPI, EXP, IGI, IMP, IEP, IC, TAS) were retained to ensure high-confidence ground truth. Each annotation was propagated upward through the Gene Ontology hierarchy using `is_a` and `part_of` relations to maintain hierarchical completeness. GO term identifiers were standardized to the January 2023 ontology. Information Accretion (IA) weights from CAFA5 (Clark and Radivojac, 2013a; Friedberg et al., 2023b) were used unchanged to ensure consistency across all experiments.

Protein metadata, including amino acid sequences, organism, taxonomy identifiers, subcellular localization, and functional comments, were obtained from the UniProt KnowledgeBase (Consortium, 2025) (release 2023\_01). Proteins with incomplete sequences or containing non-canonical residues were excluded. Structural information was retrieved from UniProt cross-references to the Protein Data Bank (PDB) (Burley et al., 2018). For proteins lacking experimentally resolved structures, AlphaFold2 (Fleming et al., 2025; Jumper et al., 2021) predicted models were used as substitutes. Domain annotations were integrated from InterPro (Blum et al., 2024), including InterPro identifiers, entry names, domain types, and residue coordinates. These domain and structure features provide complementary information on subunit organization and sequence-function relationships.

Protein-protein interaction (PPI) data were obtained from STRING v12.0 (Szklarczyk et al., 2024). UniProt

---

accessions were mapped to corresponding STRING identifiers using the official mapping tables. Only high-confidence interactions (combined score  $\geq 700$ ) were retained, and for each protein the top 10 interaction partners by combined score were selected. Each protein’s interaction context was represented both as a compact list of interaction partners and as a structured table containing evidence-channel scores (neighborhood, coexpression, experimental, database, and text-mining). This preserves both topological and evidence-level information about the protein’s interaction environment.

Protein names in the PPI metadata were standardized by applying systematic pattern matching to extract only the primary protein name, removing all parenthetical content containing EC numbers, alternative names, synonyms, and organism-specific identifiers. Text preceding the first opening parenthesis was retained, preserving the primary functional descriptor while eliminating redundant database-specific annotations.

All modalities were merged into a unified dataset linking each protein to its sequence, organism, subcellular localization, structural features, domains, GO annotations, and interaction partners. Proteins with experimental annotations before November 2022 were randomly split 90/10 into training and validation subsets. For testing, we used a temporal holdout strategy consistent with CAFA (Zhou et al., 2019), retaining proteins that gained new experimental annotations between March 2023 and February 2024 but lacked annotations in the target aspect prior to the beginning of that period (Aleksander et al., 2023).

The resulting curated dataset provides a consistent and biologically grounded foundation for downstream modeling. It preserves the same protein and organism composition as the CAFA5 benchmark while enriching each protein instance with additional structural, domain, and network context.

#### 4.1.2. Reasoning Data

For each protein in the training data, we constructed a compact context comprising InterPro domains with identifier, name, and amino acid range, the UniProt protein description, organism, subcellular localization, PPI, and GO leaf terms across Molecular Function, Biological Process, and Cellular Component with their identifiers and names. This context was provided to GPT-5 (Singh et al., 2025) using the instruction prompt shown in Section C.1 to generate a structured reasoning trace that begins by identifying and analyzing the biology of protein domains, then reasoning over Molecular Function, Biological Process, and Cellular Component while naming relevant GO terms, next articulates a plausible mechanism for protein function, and finally proposes hypotheses about interaction partners grounded in the provided context. Each trace ends with a final answer containing a concise summary of protein function, the list of InterPro domains, the list of leaf GO terms for all three aspects, and an explicit hypothesis for PPI.

We generated traces for 130,492 training proteins, averaging over 1,100 tokens per trace for a total of approximately 140 million tokens (Fig. 1E). The model was instructed to rely only on the supplied context. Manual spot checks on several hundred randomly selected proteins showed adherence to this constraint with no observed hallucinations.

#### 4.1.3. Temporal Holdout Test Set

We constructed a temporal holdout test set following the CAFA experimental framework (Zhou et al., 2019) to evaluate model generalization on proteins with newly acquired functional annotations. Protein sequences were downloaded from UniProt KnowledgeBase (Consortium, 2025) and GO term annotations from the GOA database (Consortium, 2021; Aleksander et al., 2023) at two time points, November 2022 and February 2024. Using the same experimental evidence codes described in Section 4.1, we retained proteins that gained new experimental annotations between March 2023 and February 2024 but lacked annotations in the target aspect prior to the beginning of that period, encompassing both no-knowledge proteins (no prior experimental annotations in any aspect) and limited-knowledge proteins (annotations in other aspects only). We exclude limited-knowledge proteins from BioReason-Pro evaluation because the training reasoning traces are generated from annotations across all known aspects, and cross-aspect reasoning by GPT-5 can implicitly convey information about the held-out aspect, inflating test performance.

All experimental annotations were propagated through the GO hierarchy to the root using `is_a` and `part_of` relations to maintain ontological completeness. The final temporal holdout test set comprises 8,630 unique proteins with an average of 26.75 GO terms per protein (median 18.0) and 230,824 total annotations. By ontology aspect, the test set contains 5,819 proteins with Biological Process annotations (169,459 terms,

**Table 1** | Prefix-causal attention mask pattern. Checkmarks indicate allowed attention; crosses indicate masked positions. Protein residues ( $r_i$ ) attend bidirectionally to each other but are blocked from GO tokens. GO tokens ( $t_j$ ) cross-attend to all protein residues and causally attend to preceding GO tokens.

Query ↓ \ Key →	Protein tokens				GO tokens			
	$r_1$	$r_2$	$r_3$	$r_4$	$t_1$	$t_2$	$t_3$	$t_4$
Protein $r_1$	✓	✓	✓	✓	×	×	×	×
Protein $r_2$	✓	✓	✓	✓	×	×	×	×
Protein $r_3$	✓	✓	✓	✓	×	×	×	×
Protein $r_4$	✓	✓	✓	✓	×	×	×	×
GO $t_1$	✓	✓	✓	✓	✓	×	×	×
GO $t_2$	✓	✓	✓	✓	✓	✓	×	×
GO $t_3$	✓	✓	✓	✓	✓	✓	✓	×
GO $t_4$	✓	✓	✓	✓	✓	✓	✓	✓

73.4%), 2,080 proteins with Molecular Function annotations (19,773 terms, 8.6%), and 3,440 proteins with Cellular Component annotations (41,592 terms, 18.0%).

## 4.2. GO-GPT Model

### 4.2.1. Architecture

GO-GPT is a unified autoregressive architecture for protein function prediction that models all three GO aspects, Molecular Function (MF), Biological Process (BP), and Cellular Component (CC), within a single model (Fig. 2A) (Aleksander et al., 2023). Aspect delimiter tokens (<|MF\_START|>, <|BP\_START|>, <|CC\_START|>) explicitly condition the model on the ontology branch being generated while allowing it to learn cross-aspect correlations within a shared embedding space.

The model combines a frozen protein language model encoder with a custom GPT-style transformer decoder. All experiments employ the ESM2 backbone (Lin et al., 2023), from which residue-level embeddings are extracted without gradient updates. Each residue embedding is projected into the decoder’s hidden dimension via a two-layer feed-forward projection with GELU activation (Hendrycks and Gimpel, 2023) and dropout (Srivastava et al., 2014), aligning the frozen ESM2 representation space with the autoregressive GO token space. We used ESM2 as it performed on par with ESM3 (Hayes et al., 2024) in GO-GPT, while enabling faster training due to native support for batch embedding. Architectural and training explorations are detailed in Appendix B.3.

The decoder is a 12-layer, 12-head transformer (Vaswani et al., 2023) adapted from the open-source nanoGPT implementation (Karpathy, 2022), substantially modified to support prefix-causal attention (Dong et al., 2019). This mechanism defines separate attention patterns for the protein and GO token streams. Protein residue embeddings attend bidirectionally to all other protein positions, enabling contextualized residue representations that capture long-range sequence dependencies. GO tokens attend causally to all preceding GO tokens and cross-attend to all protein residue positions. Protein tokens do not attend to GO tokens, ensuring the protein representations remain independent of the generation state. This asymmetric attention pattern is illustrated in Table 1. Each transformer block maintains separate query-key-value projections, output projections, layer normalizations, and feed-forward networks for the protein and GO streams, allowing the model to learn modality-specific transformations while sharing information through the cross-attention pathway (Bahdanau et al., 2016). Each attention layer also applies a sigmoid output gate  $Y' = Y \odot \sigma(XW_g)$  (Qiu et al., 2025), which provided minor improvements to training stability and validation metrics.

The decoder vocabulary comprises the pruned GO term set (1,470 MF, 7,500 BP, 1,007 CC, totaling  $\approx 9,977$  tokens) plus seven structural delimiters. Organism embeddings are added to each GO token representation, enabling species-specific conditioning during decoding (Fallahpour et al., 2025a). Learnable positional embeddings encode the depthwise order of GO tokens in the ontology graph, preserving hierarchical ordering and facilitating generalization from shallow, well-annotated terms to deeper, sparsely annotated ones.

---

Generation is conditioned on per-residue embeddings concatenated with aspect-delimited GO tokens

$$[r_1, r_2, \dots, r_L] \parallel \langle | \text{ASPECT\_START} | \rangle t_1, t_2, \dots, t_n \langle | \text{ASPECT\_END} | \rangle,$$

where  $r_i$  denotes the embedding of the  $i$ -th amino acid residue and  $t_j$  a GO token.

#### 4.2.2. Data Processing and Training

Each training instance consists of a protein sequence, its organism identifier, and associated GO term lists per aspect. Preprocessing tokenizes amino acid sequences using the ESM2 tokenizer (Lin et al., 2023), encodes GO terms with a custom aspect-aware tokenizer, and maps organisms to integer identifiers, retaining the top 200 most frequent species in the training set. GO vocabularies are pruned to include only terms appearing in at least 20 proteins per aspect, yielding  $\approx 10\text{K}$  GO tokens across all aspects, ensuring sufficient training signal for each term. Within each aspect, GO terms are sorted by ontology depth (longest root-to-term path) so that the model generates terms from general to specific during autoregressive decoding.

Training uses a next-token prediction objective under teacher forcing, minimizing cross-entropy loss. Optimization follows the AdamW algorithm (Kingma and Ba, 2017; Loshchilov and Hutter, 2019) with learning rate  $1 \times 10^{-4}$ , cosine decay (minimum ratio 0.1) (Loshchilov and Hutter, 2017), 10% warm-up, and weight decay 0.01. Gradients are clipped to 1.0. Mixed-precision (bf16) training and distributed data parallelism are applied across four H100 GPUs with gradient accumulation, yielding an effective batch size of 160. Training runs for up to 100 epochs with validation every epoch, and the best checkpoint is selected by validation performance (Fig. S9). Full hyperparameters are reported in Appendix B.6.

#### 4.2.3. Inference

During inference, protein embeddings are computed once and cached. Because GO-GPT is trained to predict one aspect at a time, each test protein is evaluated per aspect by prompting the model with the corresponding aspect start token and generating predictions for that aspect.

We consider two inference strategies. In greedy decoding, the model generates a single deterministic sequence per aspect, and all generated terms are assigned a probability of 1 while all others are assigned 0. In sampling-based inference, we generate 10 independent samples per protein using temperature-controlled sampling ( $T=0.7$ , top- $k=20$ ), with generation terminated upon reaching the aspect-end token (Holtzman et al., 2020). Per-term probability estimates are then computed as the fraction of samples in which each term appears. We additionally report a best-of-10 oracle analysis that selects, for each protein, the sample yielding the highest downstream score, providing an upper bound on the headroom from improved decoding or selection. Implementation uses PyTorch (Paszke et al., 2019) and PyTorch Lightning (Falcon and team, 2025) with Flash Attention (Dao, 2023; Dao et al., 2022) and bf16 mixed precision.

### 4.3. BioReason-Pro Model

#### 4.3.1. Architecture

BioReason-Pro is a multimodal LLM that integrates protein sequence, structure, and GO graph within a unified reasoning framework, inspired by BioReason (Fallahpour et al., 2025b) (Fig. 1A). The architecture comprises three primary components: a protein encoder, a GO graph encoder, and a language model backbone.

Let  $s = (a_1, \dots, a_L)$  denote a protein sequence of length  $L \leq 2,000$  over the amino acid alphabet  $\mathcal{A}$ , let  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  denote the Gene Ontology directed acyclic graph where  $\mathcal{V}$  is the set of GO terms and  $\mathcal{E}$  the set of `is_a` and `part_of` edges, and let  $c$  denote the textual biological context comprising organism, InterPro domain annotations, protein-protein interactions, and GO-GPT predictions.

**Protein Encoder.** The protein encoder is based on ESM3-1B (Hayes et al., 2024), a state-of-the-art protein language model capable of jointly encoding sequence and structure. For each input protein, ESM3-1B generates residue-level embeddings that capture local and global structural context. These embeddings are extracted from layer 38 of the 48-layer model, yielding one vector per residue. Each residue embedding is independently projected into the LLM’s hidden dimension via a two-layer MLP with GELU activation (Hendrycks

and Gimpel, 2023) and dropout (Srivastava et al., 2014):

$$\mathbf{e}_i = \text{ESM3}^{(38)}(s)_i \in \mathbb{R}^{d_{\text{esm}}}, \quad \mathbf{h}_i^{\text{prot}} = \text{MLP}_{\text{prot}}(\mathbf{e}_i) \in \mathbb{R}^{d_{\text{llm}}}, \quad i = 1, \dots, L \quad (4.1)$$

The full protein representation is the sequence of projected residue embeddings:

$$\mathbf{H}_{\text{prot}} = (\mathbf{h}_1^{\text{prot}}, \dots, \mathbf{h}_L^{\text{prot}}) \in \mathbb{R}^{L \times d_{\text{llm}}} \quad (4.2)$$

The protein encoder remains frozen throughout training, preserving its pretrained representations. Proteins longer than 2,000 amino acids are truncated.

**GO Graph Encoder.** The GO graph encoder processes the full GO graph (Consortium, 2021) to produce 200 fixed-length embedding vectors that capture hierarchical relationships and cross-namespace dependencies across all GO terms. Each embedding is projected into the LLM’s hidden dimension via a two-layer MLP:

$$\mathbf{H}_{\text{go}} = (\mathbf{h}_1^{\text{go}}, \dots, \mathbf{h}_{200}^{\text{go}}) = \text{GOEncoder}(\mathcal{G}) \in \mathbb{R}^{200 \times d_{\text{llm}}} \quad (4.3)$$

Full implementation details of the GOEncoder are provided in Section 4.3.2.

**Language Model Backbone.** The language model backbone uses Qwen3-4B-Thinking (Yang et al., 2025), a 4-billion parameter thinking model with built-in reasoning capabilities. The textual biological context  $c$  is tokenized and embedded through the LLM’s input embedding layer  $E_{\text{LLM}}(\cdot)$ :

$$\mathbf{E}_{\text{ctx}} = E_{\text{LLM}}(T_{\text{LLM}}(c)) \in \mathbb{R}^{M \times d_{\text{llm}}} \quad (4.4)$$

where  $T_{\text{LLM}}(\cdot)$  is the language model tokenizer producing  $M$  tokens.

**Multimodal Assembly.** The  $L$  residue-level protein embeddings replace  $\langle |\text{protein\_pad}| \rangle$  placeholder tokens in the prompt, and the 200 GO graph embeddings replace  $\langle |\text{go\_graph\_pad}| \rangle$  placeholder tokens (Liu et al., 2023). These are concatenated with the embedded textual context along the sequence dimension to form the full multimodal input:

$$X_{\text{LLM}} = \left( \underbrace{\mathbf{h}_1^{\text{prot}}, \dots, \mathbf{h}_L^{\text{prot}}}_{L \text{ residue embeddings}}, \underbrace{\mathbf{h}_1^{\text{go}}, \dots, \mathbf{h}_{200}^{\text{go}}}_{200 \text{ GO embeddings}}, \mathbf{E}_{\text{ctx}} \right) \quad (4.5)$$

This formulation allows the LLM to attend to residue-level protein features, ontology-level knowledge, and biological context jointly when generating output. The model autoregressively generates a reasoning trace and functional annotations:

$$P(y_i | y_{<i}, X_{\text{LLM}}; \theta) = f_{\text{LLM}}(y_{<i}, X_{\text{LLM}}) \quad (4.6)$$

where the output sequence  $Y = (y_1, \dots, y_K)$  contains structured biological reasoning followed by GO term predictions, functional summaries, and hypothesized interaction partners. Full architectural hyperparameters are reported in Table S17. Additional architectural design explorations that informed the final BioReason-Pro configuration are summarized in Section B.4.1.

### 4.3.2. GO Encoder

The GO graph encoder is a hierarchical graph neural network designed to capture the semantic and structural properties of the GO directed acyclic graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  (Consortium, 2019). Hierarchical GNNs propagate information across parent-child relationships, enabling the model to share statistical strength between rare and common terms, capture multi-level dependencies, and preserve semantic consistency (Kim et al., 2021; Kabir and Shehu, 2024). The encoder processes all GO terms in a single unified graph without namespace separation, producing 200 fixed-length embeddings that are integrated into the LLM context.

**Node Initialization.** Each GO term  $v \in \mathcal{V}$  is initialized by encoding its name and description into a single vector using the Qwen3-4B text embedding model (Zhang et al., 2025):

$$\mathbf{z}_v^{(0)} = \text{Embed}_{\text{Qwen3}}(v) \in \mathbb{R}^d \quad (4.7)$$

where  $d = 2560$  is the embedding dimension. This initialization provides each node with a semantically rich starting representation grounded in the term’s functional definition.

**GAT Propagation.** The encoder applies three graph attention network (GAT) layers (Veličković et al., 2018) over the full ontology graph. At each layer  $l = 1, \dots, 3$ , the representation of term  $v$  is updated by attending to its neighbors  $\mathcal{N}(v)$  defined by the `is_a` and `part_of` edges in  $\mathcal{E}$ . Each layer uses multi-head attention with  $K = 8$  heads:

$$\alpha_{uv}^{(l,k)} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^{(l,k)\top} [\mathbf{W}^{(l,k)} \mathbf{z}_u^{(l-1)} \parallel \mathbf{W}^{(l,k)} \mathbf{z}_v^{(l-1)}]))}{\sum_{w \in \mathcal{N}(v)} \exp(\text{LeakyReLU}(\mathbf{a}^{(l,k)\top} [\mathbf{W}^{(l,k)} \mathbf{z}_w^{(l-1)} \parallel \mathbf{W}^{(l,k)} \mathbf{z}_v^{(l-1)}]))} \quad (4.8)$$

where  $\mathbf{W}^{(l,k)}$  and  $\mathbf{a}^{(l,k)}$  are the learnable weight matrix and attention vector for head  $k$  at layer  $l$ , and  $\parallel$  denotes concatenation. The multi-head outputs are concatenated and passed through GELU activation (Hendrycks and Gimpel, 2023) and dropout (Srivastava et al., 2014):

$$\mathbf{z}_v^{(l)} = \text{Dropout}\left(\text{GELU}\left(\left\|_{k=1}^K \sum_{u \in \mathcal{N}(v)} \alpha_{uv}^{(l,k)} \mathbf{W}^{(l,k)} \mathbf{z}_u^{(l-1)}\right\|\right)\right) \quad (4.9)$$

The first layer projects from the input dimension  $d$  to the hidden dimension ( $d_h = 512$ , with  $d_h/K = 64$  per head), while subsequent layers maintain the hidden dimension. After the final GAT layer, a two-layer MLP with GELU activation projects the representations back to the original embedding dimension:

$$\tilde{\mathbf{z}}_v = \text{MLP}_{\text{out}}(\mathbf{z}_v^{(3)}) \in \mathbb{R}^d \quad (4.10)$$

**Cross-Attention Compression.** To produce a fixed-length representation from the variable number of GO terms ( $|\mathcal{V}| \approx 43,000$ ), we employ a cross-attention module inspired by the Perceiver architecture (Jaegle et al., 2021; Li et al., 2023). A set of 200 learnable query embeddings  $\mathbf{Q} = (\mathbf{q}_1, \dots, \mathbf{q}_{200}) \in \mathbb{R}^{200 \times d}$  attend to the full set of processed node embeddings  $\tilde{\mathbf{Z}} = (\tilde{\mathbf{z}}_v)_{v \in \mathcal{V}}$  via multi-head cross-attention with  $K = 8$  heads:

$$\mathbf{Q}' = \text{LayerNorm}(\mathbf{Q} + \text{MultiHeadAttn}(\mathbf{Q}, \tilde{\mathbf{Z}}, \tilde{\mathbf{Z}})) \quad (4.11)$$

followed by layer norm (Ba et al., 2016) and a residual feed-forward block (He et al., 2015):

$$\mathbf{H}_{\text{go}} = \text{LayerNorm}(\mathbf{Q}' + \text{FFN}(\mathbf{Q}')) \in \mathbb{R}^{200 \times d} \quad (4.12)$$

where FFN is a two-layer network with GELU activation and an expansion factor of 4. The resulting 200 embeddings  $\mathbf{H}_{\text{go}} = (\mathbf{h}_1^{\text{go}}, \dots, \mathbf{h}_{200}^{\text{go}})$  are then projected into the LLM’s hidden dimension via a two-layer MLP as described in Section 4.3.1 (Equation 4.3).

### 4.3.3. Supervised Fine-tuning

BioReason-Pro is trained using a two-stage supervised fine-tuning (SFT) strategy that progressively integrates multimodal protein and ontology representations with the language model backbone. Training is performed on the reasoning traces generated by GPT-5 (Singh et al., 2025) as described in Section 4.1.2. Full hyperparameters for both stages are reported in Table S17 (Appendix B.7).

We denote the parameter groups as  $\theta_{\text{esm}}$  for the protein encoder,  $\theta_{\text{go}}$  for the GO graph encoder,  $\theta_{\text{proj}}$  for the protein and GO projection layers, and  $\theta_{\text{llm}}$  for the language model backbone. The protein encoder remains frozen throughout both stages, denoted  $\theta_{\text{esm}}$ . Both stages share the same training objective: causal language modeling with cross-entropy loss computed only over assistant tokens (Touvron et al., 2023), excluding user prompt tokens from the loss:

$$\mathcal{L}_{\text{SFT}} = - \sum_{i \in \mathcal{T}_{\text{asst}}} \log P(y_i | y_{<i}, X_{\text{LLM}}; \theta) \quad (4.13)$$

where  $\mathcal{T}_{\text{asst}}$  is the set of token positions belonging to the assistant response and  $X_{\text{LLM}}$  is the multimodal input defined in Equation 4.5.

---

**Stage 1: Modality Alignment.** The first stage aligns the protein and GO graph representations with the LLM’s embedding space (Li et al., 2023). Only the GO graph encoder and projection layers are trainable, while both the protein encoder and the language model backbone remain frozen:

$$\theta_{\text{train}}^{(1)} = \{\theta_{\text{go}}, \theta_{\text{proj}}\}, \quad \theta_{\text{frozen}}^{(1)} = \{\bar{\theta}_{\text{esm}}, \theta_{\text{llm}}\} \quad (4.14)$$

This stage runs for one epoch with a learning rate of  $1 \times 10^{-4}$ , adapting the projection layers to map protein and ontology embeddings into the LLM’s representation space.

**Stage 2: LLM Fine-tuning.** The second stage unfreezes the LLM backbone and trains the complete model end-to-end for 10 epochs with a learning rate of  $1 \times 10^{-4}$ . The LLM parameters are adapted using Low-Rank Adaptation (LoRA) (Hu et al., 2021), which decomposes each pretrained weight matrix  $\mathbf{W}_0 \in \theta_{\text{llm}}$  as:

$$\mathbf{W} = \mathbf{W}_0 + \frac{\alpha}{r} \mathbf{B}\mathbf{A} \quad (4.15)$$

where  $\mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}}$  and  $\mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r}$  are the learnable low-rank matrices, with rank  $r = 128$  and scaling factor  $\alpha = 256$ . LoRA is applied to all attention and MLP layers. The GO graph encoder and projection layers remain trainable, while the protein encoder and the pretrained LLM weights stay frozen:

$$\theta_{\text{train}}^{(2)} = \{\theta_{\text{go}}, \theta_{\text{proj}}, \Delta\theta_{\text{llm}}\}, \quad \theta_{\text{frozen}}^{(2)} = \{\bar{\theta}_{\text{esm}}, \mathbf{W}_0\} \quad (4.16)$$

where  $\Delta\theta_{\text{llm}} = \{\mathbf{A}, \mathbf{B}\}$  denotes the set of all LoRA adapter parameters. The best checkpoint was selected at epoch 8 based on validation performance.

Both stages use the AdamW optimizer (Kingma and Ba, 2017; Loshchilov and Hutter, 2019) with cosine learning rate decay (Loshchilov and Hutter, 2017), a warmup ratio of 0.10 in Stage 1 and 0.05 in Stage 2, and weight decay of 0.01. Training is distributed across eight H100 GPUs on two nodes with a per-device batch size of 4, yielding an effective batch size of 32. The maximum protein length is 2,000 residues and the maximum text length is 10,000 tokens. The model was implemented using PyTorch (Paszke et al., 2019), PyTorch Lightning (Falcon and team, 2025), Hugging Face Transformers (Wolf et al., 2020), and Unsloth (Daniel Han and team, 2023). Additional supervised fine-tuning design explorations are summarized in Section B.4.2.

#### 4.3.4. Reinforcement Learning

Our RL algorithm is a hybrid variant of recent group-based policy optimization methods. We build on Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which avoids training a value model by using group-wise reward statistics. To address the instability of token-level importance sampling weights identified by Zheng et al. (2025), we follow the core principle of Group Sequence Policy Optimization (GSPO), which aligns the unit of importance correction with the unit of reward at the sequence level. We additionally incorporate modifications inspired by Dr. GRPO (Liu et al., 2025) to correct optimization bias and mitigate length-dependent reward artifacts, and adopt the Clip-Higher strategy from DAPO (Yu et al., 2025) to enhance exploration. RL is initialized from the SFT epoch 8 checkpoint and trained with LoRA (Hu et al., 2021) at rank 16 and alpha 32, a learning rate of  $3 \times 10^{-5}$  with cosine decay, and a small KL penalty ( $\beta = 1 \times 10^{-4}$ ). Full hyperparameters are reported in Table S18.

**Algorithm Formulation.** Let  $\pi_{\theta}$  be an autoregressive policy model parameterized by  $\theta$ , and let  $\mathcal{D}$  denote the set of training queries (proteins). For each query  $x \in \mathcal{D}$ , the old policy  $\pi_{\theta_{\text{old}}}$  generates a group of  $G$  responses:

$$y_1, \dots, y_G \sim \pi_{\theta_{\text{old}}}(\cdot | x), \quad (4.17)$$

where each response  $y_i = (y_{i,1}, \dots, y_{i,|y_i|})$  is a sequence of tokens and receives a scalar reward  $r(x, y_i)$  described below. We use a group size of  $G = 24$  with rollouts generated at temperature  $T = 1.0$ . We process data in batches of  $B$  queries. Let  $\mathcal{B}$  denote the set of all  $(x, y)$  pairs in the current batch, so  $|\mathcal{B}| = B \cdot G$ , and let  $L_{\text{max}}$  denote the maximum response length in tokens.

**Reward Computation.** For each response  $y_i$ , GO term identifiers are extracted from the reasoning trace using regular expressions, yielding a predicted term set  $\hat{T}_i$ . These terms are propagated to parent terms through the GO hierarchy using `is_a` and `part_of` relations following the CAFA5 evaluation framework (Friedberg et al., 2023b). Since the extracted predictions are binary, the weighted  $F_{\max}^w$  reduces to the IA-weighted F1 between the propagated predictions and the ground-truth term set  $T_x$  for protein  $x$ :

$$r(x, y_i) = \frac{2 \cdot \text{pr}_i^w \cdot \text{rc}_i^w}{\text{pr}_i^w + \text{rc}_i^w}, \quad \text{pr}_i^w = \frac{\sum_{f \in \hat{T}_i \cap T_x} \text{ia}(f)}{\sum_{f \in \hat{T}_i} \text{ia}(f)}, \quad \text{rc}_i^w = \frac{\sum_{f \in \hat{T}_i \cap T_x} \text{ia}(f)}{\sum_{f \in T_x} \text{ia}(f)}, \quad (4.18)$$

where  $\text{ia}(f)$  denotes the Information Accretion weight for GO term  $f$  (Clark and Radivojac, 2013b). This directly optimizes for the same weighted metric used in evaluation (Equation 4.28).

**Advantage Estimation.** Following GRPO and GSPO, we compute group-centered rewards. To improve stability in biological tasks with low-variance rewards, we normalize by the global batch standard deviation rather than per-group variance. For a fixed query  $x$  with responses  $y_1, \dots, y_G$ , we first compute the group mean reward:

$$\bar{r}(x) = \frac{1}{G} \sum_{k=1}^G r(x, y_k), \quad (4.19)$$

and then define the advantage for response  $y_i$  as:

$$\hat{A}_i = \frac{r(x, y_i) - \bar{r}(x)}{\text{std}(\{r(x', y') : (x', y') \in \mathcal{B}\}) + \varepsilon_{\text{std}}}. \quad (4.20)$$

Here,  $\text{std}(\cdot)$  denotes the standard deviation over all  $B \cdot G$  rewards in the batch, and  $\varepsilon_{\text{std}} > 0$  is a small constant for numerical stability. All tokens in  $y_i$  share the same advantage  $\hat{A}_i$ , consistent with sequence-level rewards.

**Sequence-Level Importance Ratio.** As in GSPO, the off-policy correction is applied at the sequence level. The sequence likelihood of  $y_i$  under  $\pi_\theta$  is:

$$\pi_\theta(y_i | x) = \prod_{t=1}^{|y_i|} \pi_\theta(y_{i,t} | x, y_{i,<t}), \quad (4.21)$$

and similarly for  $\pi_{\theta_{\text{old}}}$ . The sequence-level importance ratio is therefore:

$$s_i(\theta) = \frac{\pi_\theta(y_i | x)}{\pi_{\theta_{\text{old}}}(y_i | x)} = \exp\left(\sum_{t=1}^{|y_i|} \log \frac{\pi_\theta(y_{i,t} | x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} | x, y_{i,<t})}\right). \quad (4.22)$$

This ratio is controlled by asymmetric clipping parameters,  $\varepsilon_{\text{low}}$  and  $\varepsilon_{\text{high}}$ , which implement the Clip-Higher strategy. This allows for a larger update range ( $1 + \varepsilon_{\text{high}}$ ) when advantages are positive to encourage exploration, while maintaining a strict clamp ( $1 - \varepsilon_{\text{low}}$ ) when advantages are negative for stability.

**Clipped Surrogate Objective.** We adopt the GSPO sequence-level clipped objective, with additional length regularization to counteract the length bias described in Dr. GRPO. The optimization objective is:

$$\mathcal{J}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, \{y_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}} \left[ \frac{1}{G} \sum_{i=1}^G \min\left(s_i(\theta) \hat{A}_i, \text{clip}(s_i(\theta), 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}}) \hat{A}_i\right) \right], \quad (4.23)$$

where  $\text{clip}(s, a, b)$  truncates the importance ratio  $s$  into the interval  $[a, b]$ .

**Training Loss.** We minimize the negative of the above objective, normalized by a constant factor to produce a stable, unbiased learning signal:

$$\mathcal{L}(\theta) = -\frac{1}{B \cdot G \cdot L_{\max}} \sum_{(x, \{y_i\}_{i=1}^G) \in \mathcal{B}_x} \sum_{i=1}^G \min\left(s_i(\theta) \hat{A}_i, \text{clip}(s_i(\theta), 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}}) \hat{A}_i\right), \quad (4.24)$$

where  $\mathcal{B}_x$  ranges over the  $B$  queries in the batch. The normalization by  $B \cdot G \cdot L_{\max}$  removes structural biases that could otherwise favor responses of a particular length.

Training runs for 1,200 steps with an effective batch size of 192 across eight H100 GPUs on two nodes, using DeepSpeed (Rasley et al., 2020) with vLLM (Kwon et al., 2023) in colocate mode for rollout generation. Additional RL explorations, including stability issues and reward-design choices, are summarized in Section B.4.3.

#### 4.3.5. Inference

At inference time, BioReason-Pro assembles a multimodal context for each query protein (Fig. 1A). Residue-level embeddings from ESM3-1B (Hayes et al., 2024) and the 200 GO graph embeddings are computed and inserted into the LLM context via their respective placeholder tokens. Textual context is constructed by concatenating the organism, InterPro domain annotations (identifiers, names, and residue ranges), and greedy-decoded GO-GPT predictions. Protein-protein interaction partners are optionally included. This assembled context serves as the user prompt, and the model generates a structured reasoning trace followed by a final answer containing functional summaries, GO term predictions, and hypothesized interaction partners. The exact prompt template used to assemble this multimodal context is provided in Section C.2.

We consider two inference strategies analogous to those used for GO-GPT. In greedy decoding ( $T = 0$ ), the model generates a single deterministic trace per protein, and all GO terms appearing in the final answer are assigned probability 1. In sampling-based inference, we generate 10 independent traces per protein using temperature-controlled sampling ( $T = 1.0$ ) and select the trace yielding the highest weighted  $F_{\max}^w$  score as the best-of-10 oracle, providing an upper bound on performance from improved selection strategies. Per-term probability estimates can also be computed as the fraction of samples in which each GO term appears.

GO term identifiers are extracted from the final answer block of each generated trace using regular expressions and propagated to parent terms through the GO hierarchy following the CAFA5 framework (Friedberg et al., 2023b), as described in Section 4.4.1. The maximum generation length is 10,000 tokens. Inference is served using vLLM (Kwon et al., 2023) and takes approximately a few seconds per protein.

### 4.4. Evaluation

#### 4.4.1. GO Evaluation Metrics

GO term prediction performance is measured with the CAFA  $F_{\max}$  metric (Zhou et al., 2019). For a set of  $N$  proteins, let  $P_i(\tau)$  denote the set of GO terms predicted for protein  $i$  at score threshold  $\tau$ , and let  $T_i$  denote the set of true GO terms. Protein-centric precision and recall at threshold  $\tau$  are averaged across all proteins as

$$\text{pr}(\tau) = \frac{1}{N} \sum_{i=1}^N \frac{|P_i(\tau) \cap T_i|}{|P_i(\tau)|}, \quad \text{rc}(\tau) = \frac{1}{N} \sum_{i=1}^N \frac{|P_i(\tau) \cap T_i|}{|T_i|} \quad (4.25)$$

The unweighted  $F_{\max}$  selects the threshold that maximizes the harmonic mean of these quantities

$$F_{\max} = \max_{\tau} \frac{2 \cdot \text{pr}(\tau) \cdot \text{rc}(\tau)}{\text{pr}(\tau) + \text{rc}(\tau)} \quad (4.26)$$

For the IA-weighted variant, each GO term  $f$  is assigned an Information Accretion weight  $\text{ia}(f)$  (Clark and Radivojac, 2013b) that increases with term specificity and rarity. Weighted precision and recall replace uniform term counts with IA-weighted sums

$$\text{pr}^w(\tau) = \frac{1}{N} \sum_{i=1}^N \frac{\sum_{f \in P_i(\tau) \cap T_i} \text{ia}(f)}{\sum_{f \in P_i(\tau)} \text{ia}(f)}, \quad \text{rc}^w(\tau) = \frac{1}{N} \sum_{i=1}^N \frac{\sum_{f \in P_i(\tau) \cap T_i} \text{ia}(f)}{\sum_{f \in T_i} \text{ia}(f)} \quad (4.27)$$

The weighted  $F_{\max}^w$  is then computed analogously to Equation 4.26 using  $\text{pr}^w$  and  $\text{rc}^w$

$$F_{\max}^w = \max_{\tau} \frac{2 \cdot \text{pr}^w(\tau) \cdot \text{rc}^w(\tau)}{\text{pr}^w(\tau) + \text{rc}^w(\tau)} \quad (4.28)$$

---

Both metrics are computed per aspect and reported separately. All evaluations use the official CAFA 5 evaluation toolkit (Piovesan et al., 2024), which applies IA weighting and hierarchy propagation to both predictions and ground truth prior to scoring (Clark and Radivojac, 2013b).

Because  $F_{\max}$  is computed by sweeping a global threshold across all test proteins, it cannot be decomposed into per-protein contributions within individual subsets. To enable comparison across sequence similarity ranges (Section 4.4.5), we additionally report mean per-protein F1 within each bin. For a single protein  $i$  at the optimal global threshold  $\tau^*$ , the per-protein F1 is

$$F_{1,i} = \frac{2 \cdot \text{pr}_i(\tau^*) \cdot \text{rc}_i(\tau^*)}{\text{pr}_i(\tau^*) + \text{rc}_i(\tau^*)} \quad (4.29)$$

#### 4.4.2. Baselines

We evaluate GO-GPT and BioReason-Pro against multiple baselines spanning homology-based transfer, discriminative models, and generative language models. All baselines were trained or applied on the identical pre-November 2022 dataset and evaluated on the same temporal holdout test set.

For GO term prediction evaluated via  $F_{\max}$  and  $F_{\max}^w$ , we compare against three baselines: InterLabelGO+ (Evans and Shen, 2024), which integrates deep learning predictions with sequence alignment and represented the previous state of the art; ProtBoost (Chervov, 2024), which combines protein language model features with gradient-boosted trees and graph neural networks; and BLAST-KNN, a homology-based baseline that transfers GO annotations from the closest training set hit as described below.

For functional summary evaluation via LLM judge and human expert scoring, we compare against Prot2Text-v2 (Fei et al., 2025), a multimodal model that generates natural language protein function descriptions from sequence and structural features, evaluated on the same test set using their released checkpoint with default inference settings. We also compare against BLAST free-text transfer as described below, reported in two variants. BLAST with penalty scores test proteins whose closest hit lacks a UniProt function description as zero. BLAST without penalty excludes such proteins from evaluation entirely.

All BLAST-based baselines and sequence similarity analyses use a single blastp (Altschul et al., 1990) run with standard parameters. The training set sequences serve as the database and the test set sequences as queries. For each test protein, the hit with the highest percent identity to the training set is selected. For GO term transfer, all GO terms from the best-matching training protein are assigned to the test protein with a score equal to the percent identity. The resulting predictions are evaluated using the CAFA evaluation framework (Friedberg et al., 2023b). For free-text transfer, the UniProt function summary and subcellular localization of the best-matching training protein are transferred directly.

#### 4.4.3. LLM-as-Judge Evaluation

We employ GPT-5.1 (Singh et al., 2025) as an automated expert judge to score model predictions against composite ground truth. For each test protein, the composite ground truth comprises the UniProt function summary, GO terms across all three aspects, InterPro domain annotations, known protein-protein interaction partners, organism, and subcellular localization. The judge receives this ground truth alongside a single model generation and produces a structured evaluation via a Pydantic schema-constrained output.

The output schema enforces five integer scores on a 1 to 10 scale and a short critique. Molecular Function, Biological Process, and Cellular Component each assess annotation correctness against their respective GO terms and supporting evidence. Specificity evaluates the depth and granularity of mechanistic detail. Reliability distinguishes logically grounded inference from hallucinated or contradicted claims. A score of  $-1$  is returned for a functional axis only when all evidence sources are silent on that aspect. The critique field provides up to two sentences identifying specific mismatches between the prediction and ground truth.

The judge is called with deterministic decoding (temperature 0), a maximum output length of 512 tokens, default reasoning effort, and default verbosity. Evaluation is performed on the full temporal holdout test set for BioReason-Pro SFT, BioReason-Pro RL, Prot2Text-v2, and BLAST baselines (Section 4.4.2). The overall score for each model is the mean across the five axes, excluding any axis scored as  $-1$ . The full judge prompt and scoring rubric are provided in Appendix C.3.

---

#### 4.4.4. Human Expert Evaluation

To assess the scientific quality of model outputs beyond automated metrics, we conducted a blinded human expert evaluation. We recruited molecular biologists to independently evaluate model generations against curated ground truth annotations across 192 unique proteins randomly sampled from our test set. Both members of the BioReason team and external evaluators participated. To remove potential bias, we report results of the 162 proteins evaluated by the 27 external evaluators who were not members of the BioReason team.

Each evaluator was presented with a UniProt-derived Ground Truth dossier for a given protein, comprising the protein function summary, GO annotations, InterPro domains, known interaction partners, organism, and subcellular location. Alongside this dossier, two anonymized model generations were displayed: one from BioReason-Pro SFT and one from BioReason-Pro RL, randomly assigned as Model A and Model B. Evaluators were not informed which model produced which output.

The evaluation instrument comprises four parts, summarized below (Complete form in Supplementary C.4):

- (i) **Per-axis quantitative scoring** (Q1–Q10): ten dimensions rated on a 0–10 Likert scale applied independently to each model, covering molecular function accuracy, biological process accuracy, cellular component accuracy, reasoning and evidence attribution, plausibility of novel predictions, hallucination prevalence, hypothesis generation, mechanistic depth, protein–protein interaction predictions, and database-ready annotation quality.
- (ii) **Ordinal comparative judgments** (Q11–Q13): each model rated against the Ground Truth on a five-level ordinal scale, followed by a direct head-to-head preference judgment.
- (iii) **Free-text qualitative critique** (Q14–Q19): open-ended prompts eliciting commentary on key strengths, weaknesses, specific errors or hallucinations, reasoning quality, and comparison to expert expectations.
- (iv) **Meta-evaluation** (Q20, Evaluator Confidence): a case-study nomination flag and evaluator self-reported confidence on a 1–10 scale.

To quantify error prevalence from the free-text critiques, we used GPT-5-mini (Singh et al., 2025) to classify each evaluator’s written response into major error, minor error, or no error categories for each model independently. The classification prompt is provided in Appendix C.5 and the resulting error attributions stratified by expert preference are reported in Fig. 4G.

#### 4.4.5. Sequence Similarity Analysis

To evaluate model generalization as a function of homology to the training set, we stratified test proteins by best-hit BLAST sequence identity. For each test protein, the highest sequence identity against any training set sequence was obtained from the blastp search described in Section 4.4.2. Proteins were then binned into identity ranges. Proteins with no detectable BLAST hit were assigned to the lowest bin.

This stratification is applied consistently across all three evaluation modalities. For GO term prediction, we report mean per-protein F1 (Equation 4.29) within each bin, since  $F_{\max}$  is a global metric that cannot be decomposed across protein subsets. For LLM-as-judge evaluation (Section 4.4.3), we report mean judge scores per bin. For human expert evaluation (Section 4.4.4), we report mean expert scores per bin.

#### 4.4.6. Statistical Comparison of GO Term Prediction Models

To assess statistical significance of performance differences between GO-GPT and baseline models, we used a paired bootstrap test. For each comparison, per-protein precision and recall arrays were computed for both models on the shared test set using CAFA-standard normalization (precision averaged over proteins with at least one prediction above threshold; recall averaged over all proteins). In each of 10,000 bootstrap iterations, proteins were resampled with replacement and the  $F_{\max}$  difference was recomputed; the two-sided  $p$ -value was estimated as the fraction of bootstrap deltas with sign opposite to the observed difference. Both  $F_{\max}$  and  $F_{\max}^w$  were tested. Because each inference mode involves 12 simultaneous tests (2 comparisons  $\times$  3 aspects  $\times$  2 metrics),  $p$ -values were corrected using the Holm–Bonferroni step-down procedure to control the family-wise error rate. Full bootstrap comparison tables are reported in Section B.5.1.

---

## 4.5. Interpretability

### 4.5.1. GO-GPT Attention at Functional Sites

To evaluate whether GO-GPT grounds its predictions in functionally relevant sequence regions, we assessed the correspondence between model attention and experimentally determined DNA-binding residues. We identified proteins with experimentally determined DNA-binding sites from the BioLiP database (Yang et al., 2012) that were not present in the GO-GPT training set. DNA-binding sites were defined as BioLiP entries with ligand types DNA, DA, DC, DG, DT, DU, or DI. PDB binding residue positions were mapped to UniProt coordinates using per-residue SIFTS annotations (Burley et al., 2018; Consortium, 2025; Dana et al., 2019) extracted from mmCIF files, which correctly handles numbering gaps and insertion codes that segment-level SIFTS interpolation can miss. We retained proteins with at least 5 mapped binding residues and fetched their sequences from UniProt, yielding 304 candidate proteins. We then ran GO-GPT inference on each protein and retained the subset for which the model autoregressively generated the GO:0003677 (DNA binding) term, resulting in 63 evaluation proteins.

For each retained protein, we extracted cross-attention weights from the GO:0003677 output token to all input residue positions, averaged across all attention heads and transformer layers, producing a per-residue attention score vector. Discrimination between binding and non-binding residues was quantified using three complementary metrics. AUROC measures how well attention scores distinguish binding from non-binding residues, with its significance assessed via Mann–Whitney U test (Mann and Whitney, 1947). AUPRC Gain normalizes the area under the precision–recall curve by the positive class prevalence:  $\text{AUPRC Gain} = (\text{AUPRC} - \pi) / (1 - \pi)$ , where  $\pi$  is the fraction of binding residues, yielding 0 at chance level and 1 for perfect discrimination; its significance is assessed via the same Mann–Whitney U test as AUROC. Fold-enrichment measures the concentration of attention at binding sites relative to the sequence background, with its significance assessed via hypergeometric test on the top 20% highest-attention residues. Full per-protein statistics are reported in Section B.1, and a subset is visualized in Fig. S2. Three proteins spanning different AUROC values (E1C9K5, P84131, Q5Y812) were selected for visualization in Fig. 2E. Three-dimensional structure visualizations with projected attention scores were rendered using UCSF ChimeraX (Pettersen et al., 2021).

### 4.5.2. ESM2 Embedding Norm Control

To test whether the binding-site discrimination observed in GO-GPT’s cross-attention is already present in the frozen ESM2 representations, we computed L2 norms of ESM2 output embeddings at each residue position and evaluated them as a baseline predictor of DNA-binding residues using the same metrics (AUROC, AUPRC Gain, fold-enrichment) and statistical tests applied to attention scores. If ESM2 embeddings encode binding-site information in their magnitude—for example, through higher norms at structurally important positions—then GO-GPT’s attention could simply track this pre-existing signal rather than learning a GO-function-specific one. Per-protein comparisons between attention-based and L2-norm-based metrics are reported in Fig. S2.

### 4.5.3. GO-GPT Embedding Analysis

We characterized the structure of GO-GPT’s learned representations through a suite of analyses spanning organism embeddings, GO term embeddings, and their relationship to external biological knowledge.

GO-GPT learns a unique embedding vector for each organism in the training vocabulary. We computed pairwise cosine distances between all organism embeddings and constructed a hierarchical clustering dendrogram using single linkage to visualize the resulting structure (Fig. 2C). To quantify the correspondence between embedding geometry and evolutionary relationships, we retrieved pairwise divergence times (in Mya) for the 30 most frequent organisms from the TimeTree 5 API (Kumar et al., 2022) and performed a Mantel test (Mantel, 1967) comparing the cosine distance matrix to the divergence time matrix (Spearman correlation, 9,999 permutations).

Each GO term in the vocabulary has a learned embedding. We assessed aspect-level organization in the original embedding space using cosine purity: for each GO term, we computed the fraction of its  $k$  nearest neighbors ( $k = 1, 3, 5, 10, 20$ ) in cosine space that share the same GO aspect, averaged across all terms within each aspect. The per-aspect random baseline equals the aspect’s proportion of the full vocabulary (MF: 14.9%, BP: 75.0%, CC: 10.1%). Statistical significance was assessed via 1,000-permutation test (aspect label shuffles).

---

To characterize the baseline distribution of embedding similarities, we computed pairwise cosine similarities between all GO terms within each aspect (Fig. S6). To visualize embedding structure, we projected GO term embeddings to two dimensions using UMAP (McInnes et al., 2020) with cosine distance,  $n_{\text{neighbors}} = 15$ , and  $\text{min\_dist} = 0.1$  (Fig. 2H; Fig. S8).

To test whether embedding geometry reflects ontology-based semantic relationships, we computed Resnik similarity (Resnik, 1995) for all within-aspect GO term pairs. Resnik similarity is defined as  $\text{Resnik}(t_i, t_j) = \text{IC}(\text{MICA})$ , where  $\text{IC}(t) = -\log_2 P(t)$  and  $P(t)$  is the fraction of training proteins annotated with term  $t$  or any of its descendants. IC-based measures naturally account for the uneven depth and annotation density across branches of the ontology, and have been shown to correlate with protein sequence similarity better than graph-distance-based alternatives (Lord et al., 2003). Crucially, IC is derived from the same annotation frequencies that the model observes during training, making it the appropriate ground truth for assessing whether the model has internalized the statistical structure of protein–function associations. We measured the Spearman correlation between Resnik similarity and cosine embedding similarity per aspect, with significance assessed via Mantel permutation test (1,000 permutations of term-to-embedding assignments; Fig. 2F).

To assess whether embeddings capture functional relationships that span ontology aspects, we computed Normalized Pointwise Mutual Information (NPMI; ranges from  $-1$  for terms that never co-occur, through  $0$  for statistically independent terms, to  $+1$  for perfectly co-occurring terms) for all cross-aspect term pairs (MF×BP, MF×CC, BP×CC) from protein co-annotation frequencies, restricting to terms with annotation frequency  $\geq 20$  and pairs with co-occurrence  $\geq 5$ . Only proteins annotated in both relevant aspects contributed to each comparison. Significance was assessed by permuting the term-to-embedding assignment independently per aspect (1,000 permutations), recomputing cosine similarities for all valid pairs, and comparing the resulting Spearman  $\rho$  against the observed value (Fig. 2G). To evaluate whether this correlation enables practical retrieval, we formulated a cross-aspect retrieval benchmark: for each GO term in one aspect, we ranked all terms in the partner aspect by embedding cosine similarity and measured retrieval performance against ground-truth partners defined by  $\text{NPMI} > 0.5$  (strong co-annotation). Only candidate pairs with co-occurrence  $\geq 5$  were evaluated, ensuring that positive and negative labels are supported by sufficient annotation evidence. We report AUROC, Mean Reciprocal Rank (MRR), and Precision@ $k$  ( $k = 1, 5, 10, 20$ ) macro-averaged across query terms, with per-query random baselines computed from positive class prevalence. Both retrieval directions ( $A \rightarrow B$  and  $B \rightarrow A$ ) are evaluated separately and combined into a symmetric AUROC. Statistical significance was assessed by the same embedding-permutation procedure (1,000 permutations; Fig. S7). To visualize representative cross-aspect associations in the embedding landscape, we selected pairs ranked by the product of NPMI and cosine similarity (top 200 per aspect combination) and filtered by UMAP Euclidean distance  $< 0.3$ , retaining 18 pairs that co-localize in the two-dimensional projection (Fig. S8; Table S2).

We evaluated whether embedding neighborhoods reflect the GO directed acyclic graph by computing  $k$ -nearest-neighbor overlap with ontological neighbors (parents, children, and siblings within 2 hops via `is_a` and `part_of` relationships) for the top 1,000 most frequent terms per aspect. We report Precision@ $k$  and Recall@ $k$  at  $k = 1, 3, 5, 10, 20$ , compared to a random baseline derived from vocabulary size and mean neighbor count, with significance assessed via 1,000-permutation test (Fig. S6).

#### 4.5.4. BioReason-Pro Embedding Space Analysis

To characterize how BioReason-Pro reorganizes protein representations relative to its input encoder, we extracted mean-pooled protein embeddings from the BioReason-Pro SFT model at two stages of the pipeline, following the approach of Hie et al. (2021) for visualizing learned protein representations. ESM3 pre-projection embeddings (1536-dimensional) capture the frozen protein language model representations before they enter the LLM (Hayes et al., 2024), while LLM Layer 35 post-RMSNorm embeddings (2560-dimensional) capture the final internal representations learned by the reasoning backbone.

For the full training set visualization (Fig. 1H), LLM Layer 35 embeddings were reduced from 2560 to 50 dimensions using PCA (Pearson, 1901) and then projected to two dimensions using UMAP (McInnes et al., 2020) with  $n_{\text{neighbors}} = 30$  and  $\text{min\_dist} = 0.3$ . Clusters were identified by applying HDBSCAN (Campello et al., 2013; McInnes et al., 2017) to the PCA-reduced LLM Layer 35 embeddings with  $\text{min\_cluster\_size} = 20$ ,  $\text{min\_samples} = 5$ , and Euclidean distance. The top 9 clusters by size were retained for annotation.

For the comparative visualization (Fig. 1I), a 10K protein subset was selected and both ESM3 and LLM Layer 35 embeddings were independently projected to two dimensions using UMAP with the same hyperpa-

---

rameters. Cluster assignments were fixed from the LLM Layer 35 clustering and applied identically to both projections so that the same proteins receive the same colors across panels. To functionally annotate each cluster, we computed GO Molecular Function term enrichment using one-sided Fisher’s exact test (Fisher, 2018) comparing term frequency within each cluster against all proteins as background. All MF terms at GO depth  $\geq 1$  were included, with a minimum of 5 occurrences per term per cluster. Each cluster is annotated by its most enriched MF term.

#### 4.5.5. *BioReason-Pro Attention Analysis*

To investigate whether BioReason-Pro grounds its reasoning in structurally meaningful protein features, we analyzed attention from specific tokens in the generated reasoning trace back to the protein embeddings.

For each analysis, a probe phrase is manually identified in the generated reasoning trace at the point where the model articulates a functionally relevant prediction. For example, in the eEFSec analysis the probe phrase is "with" preceding the predicted partner name "SECIS-binding protein 2", and in the CFAP61 analysis it is "scaffold" at the point where the model infers non-enzymatic function (Fig. 5, Fig. 6).

Attention weights are extracted from all layers and all heads of the Qwen3-4B backbone. The best layer is selected automatically as the layer whose single best head achieves the highest mean attention to protein positions. At the selected layer, heads are ranked by mean attention to protein positions. Different attention heads capture different structural and functional features, so the head used for each analysis is selected based on specificity to the protein region of interest. For the eEFSec analysis the top-ranked head was used, while for the CFAP61 analysis the second-ranked head was selected. Per-residue attention scores are computed as the mean attention across all query token positions within the probe phrase to each protein residue position. For structural visualization, scores are min-max normalized to [0, 1].

Residue-to-domain mappings are obtained directly from the InterPro annotations. Only domain-type entries are retained, and start/end residue positions define each domain boundary. Residues not covered by any annotated domain are labeled as non-domain. This enables domain-level aggregation of per-residue attention scores for the plots in Fig. 5 and Fig. 6.

Discrimination between functionally relevant and background residues is assessed using one-sided Mann–Whitney U tests (Mann and Whitney, 1947), testing whether attention at relevant residues exceeds that at non-relevant residues. Fold enrichment is computed as the ratio of mean attention at relevant residues to mean attention at non-relevant residues. Relevant residue sets are defined by structural contact analysis as described in Section 4.5.6.

#### 4.5.6. *BioReason-Pro Structural Contact Analysis*

To assess whether BioReason-Pro attention concentrates on experimentally validated functional interfaces, we computed structural contact maps from published cryo-EM structures and compared them against per-residue attention scores extracted as described in Section 4.5.5.

Contact residues were identified from mmCIF coordinate files (Westbrook et al., 2022) using KD-tree distance queries over all heavy atoms (Bentley, 1975). A protein residue was classified as contacting a partner molecule if any of its heavy atoms fell within 5.0 Å of any heavy atom in the partner (Méndez et al., 2003). Statistical significance was assessed using one-sided Mann–Whitney U tests (Mann and Whitney, 1947) comparing attention scores at contact versus non-contact residues. Fold enrichment was computed as the ratio of mean attention at contact residues to mean attention at non-contact residues.

**eEFSec (P57772).** SECIS RNA contact residues were identified from the cryo-EM selenosome structure (PDB 7ZJW, 2.8 Å resolution) (Hilal et al., 2022), with Chain E corresponding to eEFSec and Chain S to the SECIS RNA element. All eEFSec residues with any heavy atom within 5.0 Å of any SECIS RNA atom were classified as contact residues. Per-residue attention scores from the probe token analysis (Section 4.5.5) were compared between contact and non-contact residues. Attention scores were projected onto the structure using a white-to-red colormap for visualization in PyMOL (Schrödinger, LLC, 2015).

**CFAP61 (Q8NHU2).** The CFAP61 Rossmann-like domain (residues 665–998) was aligned to glutathione reductase (PDB 3GRS) (Karplus and Schulz, 1987) using topology-guided manual mapping. The canonical Rossmann  $\beta$ - $\alpha$ - $\beta$ - $\alpha$ - $\beta$  fold (Rao and Rossmann, 1973) was used to identify equivalent positions between 3GRS

**Table 2** | Catalytic and cofactor-binding positions mapped from glutathione reductase (PDB 3GRS) to the CFAP61 Rossmann domain by topology-guided alignment. Repurposed residues (marked with †) are positions where the degenerate catalytic site directly contacts Dynein Heavy Chain 1 in the cryo-EM axoneme structure (PDB 8J07) (Walton et al., 2023).

Functional role	3GRS reference	CFAP61 position	Substitution
GxGxxG loop	G27–G32	671–676	VGASSV
Redox cysteine	C63	C734 <sup>†</sup>	Conserved (interface)
Cofactor-binding	S	V737 <sup>†</sup>	Ser → Val
Catalytic nucleophile	C58	V739 <sup>†</sup>	Cys → Val
FAD stacking	E50	L767	Glu → Leu

and the CFAP61 AlphaFold predicted structure (Fleming et al., 2025), matching  $\beta$ -strand/ $\alpha$ -helix junctions, specifically anchoring on the GxGxxG loop at the  $\beta$ 1- $\alpha$ 1 junction (Wierenga et al., 1986). This alignment identified 12 positions in CFAP61 topologically equivalent to catalytic and cofactor-binding residues in 3GRS (Table 2). All 12 positions show amino acid substitutions inconsistent with catalytic function.

Dynein Heavy Chain 1 contact residues were identified from the cryo-EM axoneme structure (PDB 8J07) (Walton et al., 2023), with Chain d0 corresponding to CFAP61 and Chain g6 to Dynein Heavy Chain 1. All CFAP61 residues within 5.0 Å of any Dynein Heavy Chain 1 heavy atom were classified as interface residues. Three positions (C734, V737, V739) were identified as repurposed residues by intersection of the catalytic position set from the 3GRS alignment with the Dynein Heavy Chain 1 interface set from the 8J07 structure.

For the attention enrichment analysis, Rossmann domain residues were partitioned into four groups: repurposed active-site residues (catalytic  $\cap$  interface), remaining catalytic-site residues (catalytic only), Dynein Heavy Chain 1 interface residues (interface only), and other Rossmann-domain residues. Per-residue attention scores were compared across these groups using Mann–Whitney U tests (Mann and Whitney, 1947). Attention scores were projected onto the Rossmann domain for visualization in PyMOL (Schrödinger, LLC, 2015).

## Data Availability

All datasets used to train and evaluate models are available at <https://bioreason.net/data>.

## Code and Model Availability

We make code and tools available at the following links:

- Web-based inference server: <https://bioreason.net>
- Training and evaluation code repository: <https://bioreason.net/code>
- BioReason-Pro function predictions for over 240,000 proteins, including the Human Protein Atlas: <https://bioreason.net/atlas>

We make the model parameters available on Hugging Face:

- GO-GPT: <https://bioreason.net/gogpt>
- BioReason-Pro SFT: <https://bioreason.net/sft>
- BioReason-Pro RL: <https://bioreason.net/rl>

## Acknowledgements

We thank Bryan Perozzi and Michael Galkin for insights on encoding Gene Ontology graph structure and learning hierarchical ontologies. We thank Jure Leskovec for insightful discussions on multimodal biological reasoning. We thank Guillaume Filion for advising and insights on protein function reasoning and training suggestions. We thank Shayan Pardis for valuable input on GO term representation and graph learning approaches. We thank Li Erran Li for insights on biological reasoning. We thank Anshul Kundaje for feedback on

---

evaluation methodology and suggestions for assessing biological reasoning. We thank Iddo Friedberg for guidance on CAFA evaluation. We thank Andrew Magnuson and Ronald Xie for their valuable insights on model architecture. We thank Jacob Junqi Tian and John Willes for helpful discussions on training infrastructure and reinforcement learning strategies. We thank Amol Punjabi for discussions on the impact and release strategy of this work. We thank Kuan Pang for fruitful discussions about future work. We thank Joseph Caputo for media and communications support. We thank Clem Delangue for his support and encouragement in making our models publicly accessible through the Hugging Face platform. Compute resources were provided by the Arc Institute, Vector Institute, the Digital Research Alliance of Canada, and the Government of Canada. H.G. is an Arc Core Investigator and acknowledges funding support from Arc Institute.

## Author Contributions

A.F., A.S-A., and P.I. conceived the project and co-led its development. H.G. and B.W. supervised the project. D.P.B. and P.D.H. advised throughout the project. A.F. and P.I. designed the BioReason-Pro architecture. A.S-A. designed the GO-GPT architecture. K.Z. designed the GO graph encoder. A.F. developed the SFT training pipeline and ran all SFT experiments. P.I. developed the RL training pipeline and ran all RL experiments. A.S-A. developed the GO-GPT training code and ran all GO-GPT experiments. P.I. developed the vLLM-based fast inference stack. T.G. contributed to the GO-GPT inference code. A.Shah contributed to high-throughput inference for RL. A.A. contributed to training and scaling BioReason-Pro. O.M. contributed and provided guidance on RL. H.D. and C.J.M. advised RL training. P.G., O.I., A.S-A, S.M., and J.N. curated and processed the training dataset, including protein sequences, domain annotations, interaction data, and GO terms. A.F. and O.I. generated the synthetic reasoning data. P.I. advised the synthetic data generation. O.I. and P.G. constructed the test dataset. A.F. developed the LLM judge and human expert evaluation frameworks and conducted all BioReason-Pro evaluations. P.I. advised the LLM judge and human expert evaluations. A.S-A. set up GO term evaluation pipelines and conducted all GO-GPT evaluations. A.F., H.G., and B.W. recruited human expert evaluators. P.G. and T.G. conducted GO-based baseline evaluations. K.Z., J.N., and A.F. conducted free-text baseline evaluations. A.F. and P.I. conducted the BioReason-Pro interpretability analysis. A.S-A. conducted the GO-GPT interpretability analysis. A.F. wrote up the eEFSec and CFAP61 cases. O.I. contributed to case study selection. C.A.T., M.Y.Y.L., and A.F.M.S. conducted independent model evaluation. N.L., H.Cui, A.A., A.J., A.Fa., A.C-P., J.S.S., F.N., A.A.N., H.C.M., and L.S. served as human evaluators. B.S.P. provided scientific advising and feedback on the manuscript. C.R-T. provided scientific advising and designed figures. M.d.C. and P.I. designed the web interface and inference server. A.F., A.S-A., and T.G. generated predictions for the released atlas. A.F., A.S-A., and P.I. managed the open-source release of models, code, and datasets. A.F. and A.S-A. created all plots. A.F., A.S-A., P.I., P.G., and O.I. wrote the first draft of the manuscript. H.G. and B.W. provided guidance throughout.

## Competing Interests

D.P.B. acknowledges outside interest as a Google Advisor. H.G. acknowledges outside interest as a co-founder of Exai Bio, Tahoe Therapeutics, and Therna Therapeutics, serves on the board of directors at Exai Bio, and is a scientific advisory board member for Verge Genomics and Deep Forest Biosciences. B.W. acknowledges outside interest as SVP and Head of Biomedical AI at Xaira Therapeutics. P.D.H. acknowledges outside interest as a cofounder of Monet AI and Stylus Medicine; a member of the board of directors at Stylus Medicine; a scientific advisory board member at Amgen and Veda Bio; and a venture partner at Thrive Capital. O.M. acknowledges outside interest as a member of technical staff at Cohere. C.J.M. acknowledges outside interest as an advisor for Axiom Bio.

---

## References

- H. Abdine et al. Prot2text: Multimodal protein's function generation with gnns and transformers. *arXiv preprint arXiv:2307.14367*, 2023.
- J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Židek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, May 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL <http://dx.doi.org/10.1038/s41586-024-07487-w>.
- S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N. L. Harris, D. P. Hill, R. Lee, H. Mi, S. Moxon, C. J. Mungall, A. Muruganugan, T. Mushayahama, P. W. Sternberg, P. D. Thomas, K. Van Auken, J. Ramsey, D. A. Siegele, R. L. Chisholm, P. Fey, M. C. Aspromonte, M. V. Nugnes, F. Quaglia, S. Tosatto, M. Giglio, S. Nadendla, G. Antonazzo, H. Attrill, G. dos Santos, S. Marygold, V. Strelets, C. J. Tabone, J. Thurmond, P. Zhou, S. H. Ahmed, P. Asanitthong, D. Luna Buitrago, M. N. Erdol, M. C. Gage, M. Ali Kadhum, K. Y. C. Li, M. Long, A. Michalak, A. Pesala, A. Pritazahra, S. C. C. Saverimuttu, R. Su, K. E. Thurlow, R. C. Lovering, C. Logie, S. Oliferenko, J. Blake, K. Christie, L. Corbani, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. Sitnikov, C. Smith, A. Cuzick, J. Seager, L. Cooper, J. Elser, P. Jaiswal, P. Gupta, P. Jaiswal, S. Naithani, M. Lera-Ramirez, K. Rutherford, V. Wood, J. L. De Pons, M. R. Dwinell, G. T. Hayman, M. L. Kaldunski, A. E. Kwitek, S. J. F. Laulederkind, M. A. Tutaj, M. Vedi, S.-J. Wang, P. D'Eustachio, L. Aimo, K. Axelsen, A. Bridge, N. Hyka-Nouspikel, A. Morgat, S. A. Aleksander, J. M. Cherry, S. R. Engel, K. Karra, S. R. Miyasato, R. S. Nash, M. S. Skrzypek, S. Weng, E. D. Wong, E. Bakker, T. Z. Berardini, L. Reiser, A. Auchincloss, K. Axelsen, G. Argoud-Puy, M.-C. Blatter, E. Boutet, L. Breuza, A. Bridge, C. Casals-Casas, E. Coudert, A. Estreicher, M. Livia Famiglietti, M. Feuermann, A. Gos, N. Gruaz-Gumowski, C. Hulo, N. Hyka-Nouspikel, F. Jungo, P. Le Mercier, D. Lieberherr, P. Masson, A. Morgat, I. Pedruzzi, L. Pourcel, S. Poux, C. Rivoire, S. Sundaram, A. Bateman, E. Bowler-Barnett, H. Bye-A-Jee, P. Denny, A. Ignatchenko, R. Ishtiaq, A. Lock, Y. Lussi, M. Magrane, M. J. Martin, S. Orchard, P. Raposo, E. Speretta, N. Tyagi, K. Warner, R. Zaru, A. D. Diehl, R. Lee, J. Chan, S. Diamantakis, D. Raciti, M. Zarowiecki, M. Fisher, C. James-Zorn, V. Ponferrada, A. Zorn, S. Ramachandran, L. Ruzicka, M. Westerfield, S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N. L. Harris, D. P. Hill, R. Lee, H. Mi, S. Moxon, C. J. Mungall, A. Muruganugan, T. Mushayahama, P. W. Sternberg, P. D. Thomas, K. Van Auken, J. Ramsey, D. A. Siegele, R. L. Chisholm, P. Fey, M. C. Aspromonte, M. V. Nugnes, F. Quaglia, S. Tosatto, M. Giglio, S. Nadendla, G. Antonazzo, H. Attrill, G. dos Santos, S. Marygold, V. Strelets, C. J. Tabone, J. Thurmond, P. Zhou, S. H. Ahmed, P. Asanitthong, D. Luna Buitrago, M. N. Erdol, M. C. Gage, M. Ali Kadhum, K. Y. C. Li, M. Long, A. Michalak, A. Pesala, A. Pritazahra, S. C. C. Saverimuttu, R. Su, K. E. Thurlow, R. C. Lovering, C. Logie, S. Oliferenko, J. Blake, K. Christie, L. Corbani, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. Sitnikov, C. Smith, A. Cuzick, J. Seager, L. Cooper, J. Elser, P. Jaiswal, P. Gupta, P. Jaiswal, S. Naithani, M. Lera-Ramirez, K. Rutherford, V. Wood, J. L. De Pons, M. R. Dwinell, G. T. Hayman, M. L. Kaldunski, A. E. Kwitek, S. J. F. Laulederkind, M. A. Tutaj, M. Vedi, S.-J. Wang, P. D'Eustachio, L. Aimo, K. Axelsen, A. Bridge, N. Hyka-Nouspikel, A. Morgat, S. A. Aleksander, J. M. Cherry, S. R. Engel, K. Karra, S. R. Miyasato, R. S. Nash, M. S. Skrzypek, S. Weng, E. D. Wong, E. Bakker, T. Z. Berardini, L. Reiser, A. Auchincloss, K. Axelsen, G. Argoud-Puy, M.-C. Blatter, E. Boutet, L. Breuza, A. Bridge, C. Casals-Casas, E. Coudert, A. Estreicher, M. Livia Famiglietti, M. Feuermann, A. Gos, N. Gruaz-Gumowski, C. Hulo, N. Hyka-Nouspikel, F. Jungo, P. Le Mercier, D. Lieberherr, P. Masson, A. Morgat, I. Pedruzzi, L. Pourcel, S. Poux, C. Rivoire, S. Sundaram, A. Bateman, E. Bowler-Barnett, H. Bye-A-Jee, P. Denny, A. Ignatchenko, R. Ishtiaq, A. Lock, Y. Lussi, M. Magrane, M. J. Martin, S. Orchard, P. Raposo, E. Speretta, N. Tyagi, K. Warner, R. Zaru, A. D. Diehl, R. Lee, J. Chan, S. Diamantakis, D. Raciti, M. Zarowiecki, M. Fisher, C. James-Zorn, V. Ponferrada, A. Zorn, S. Ramachandran, L. Ruzicka, and M. Westerfield. The gene ontology knowledgebase in 2023. *GENETICS*, 224(1), Mar. 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad031. URL <http://dx.doi.org/10.1093/genetics/iyad031>.
- S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, Oct. 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2. URL [http://dx.doi.org/10.1016/S0022-2836\(05\)80360-2](http://dx.doi.org/10.1016/S0022-2836(05)80360-2).

- 
- C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, July 1973. ISSN 1095-9203. doi: 10.1126/science.181.4096.223. URL <http://dx.doi.org/10.1126/science.181.4096.223>.
- M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.
- J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization, 2016. URL <https://arxiv.org/abs/1607.06450>.
- D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate, 2016. URL <https://arxiv.org/abs/1409.0473>.
- A. Bairoch. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic Acids Research*, 28(1):45–48, Jan. 2000. ISSN 1362-4962. doi: 10.1093/nar/28.1.45. URL <http://dx.doi.org/10.1093/nar/28.1.45>.
- J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, Sept. 1975. ISSN 1557-7317. doi: 10.1145/361002.361007. URL <http://dx.doi.org/10.1145/361002.361007>.
- M. Blum, A. Andreeva, L. C. Florentino, S. R. Chuguransky, T. Grego, E. Hobbs, B. L. Pinto, A. Orr, T. Paysan-Lafosse, I. Ponamareva, G. A. Salazar, N. Bordin, P. Bork, A. Bridge, L. Colwell, J. Gough, D. H. Haft, I. Letunic, F. Llinares-López, A. Marchler-Bauer, L. Meng-Papaxanthos, H. Mi, D. A. Natale, C. A. Orengo, A. P. Pandurangan, D. Piovesan, C. Rivoire, C. J. A. Sigrist, N. Thanki, F. Thibaud-Nissen, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, and A. Bateman. Interpro: the protein sequence classification resource in 2025. *Nucleic Acids Research*, 53(D1):D444–D456, Nov. 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1082. URL <http://dx.doi.org/10.1093/nar/gkae1082>.
- G. Brixi, M. G. Durrant, J. Ku, M. Naghipourfar, M. Poli, G. Sun, G. Brockman, D. Chang, A. Fanton, G. A. Gonzalez, S. H. King, D. B. Li, A. T. Merchant, E. Nguyen, C. Ricci-Tam, D. W. Romero, J. C. Schmok, A. Taghibakhshi, A. Vorontsov, B. Yang, M. Deng, L. Gorton, N. Nguyen, N. K. Wang, M. T. Pearce, E. Simon, E. Adams, Z. J. Amador, E. A. Ashley, S. A. Baccus, H. Dai, S. Dillmann, S. Ermon, D. Guo, M. H. Herschl, R. Ilango, K. Janik, A. X. Lu, R. Mehta, M. R. K. Mofrad, M. Y. Ng, J. Pannu, C. Ré, J. St. John, J. Sullivan, J. Tey, B. Viggiano, K. Zhu, G. Zynda, D. Balsam, P. Collison, A. B. Costa, T. Hernandez-Boussard, E. Ho, M.-Y. Liu, T. McGrath, K. Powell, S. Pinglay, D. P. Burke, H. Goodarzi, P. D. Hsu, and B. L. Hie. Genome modelling and design across all domains of life with evo 2. *Nature*, Mar. 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10176-5. URL <http://dx.doi.org/10.1038/s41586-026-10176-5>.
- B. Buchfink, C. Xie, and D. H. Huson. Fast and sensitive protein alignment using diamond. *Nature Methods*, 12(1):59–60, 2015.
- S. K. Burley, H. M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. D. Costanzo, C. Christie, J. M. Duarte, S. Dutta, Z. Feng, S. Ghosh, D. S. Goodsell, R. K. Green, V. Guranovic, D. Guzenko, B. P. Hudson, Y. Liang, R. Lowe, E. Peisach, I. Periskova, C. Randle, A. Rose, M. Sekharan, C. Shao, Y.-P. Tao, Y. Valasatava, M. Voigt, J. Westbrook, J. Young, C. Zardecki, M. Zhuravleva, G. Kurisu, H. Nakamura, Y. Kengaku, H. Cho, J. Sato, J. Y. Kim, Y. Ikegawa, A. Nakagawa, R. Yamashita, T. Kudou, G.-J. Bekker, H. Suzuki, T. Iwata, M. Yokochi, N. Kobayashi, T. Fujiwara, S. Velankar, G. J. Kleywegt, S. Anyango, D. R. Armstrong, J. M. Berrisford, M. J. Conroy, J. M. Dana, M. Deshpande, P. Gane, R. Gáborová, D. Gupta, A. Gutmanas, J. Koča, L. Mak, S. Mir, A. Mukhopadhyay, N. Nadzirin, S. Nair, A. Patwardhan, T. Paysan-Lafosse, L. Pravda, O. Salih, D. Sehnal, M. Varadi, R. Vařeková, J. L. Markley, J. C. Hoch, P. R. Romero, K. Baskaran, D. Maziuk, E. L. Ulrich, J. R. Wedell, H. Yao, M. Livny, and Y. E. Ioannidis. Protein data bank: the single global archive for 3d macromolecular structure data. *Nucleic Acids Research*, 47(D1):D520–D528, Oct. 2018. ISSN 1362-4962. doi: 10.1093/nar/gky949. URL <http://dx.doi.org/10.1093/nar/gky949>.
- K. Caban and P. R. Copeland. Size matters: a view of selenocysteine incorporation from the ribosome. *Cellular and Molecular Life Sciences*, 63(1):73, Jan. 2006. ISSN 1420-682X, 1420-9071. doi: 10.1007/s00018-005-5402-y. URL <https://link.springer.com/10.1007/s00018-005-5402-y>.

- 
- R. J. G. B. Campello, D. Moulavi, and J. Sander. *Density-Based Clustering Based on Hierarchical Density Estimates*, page 160–172. Springer Berlin Heidelberg, 2013. ISBN 9783642374562. doi: 10.1007/978-3-642-37456-2\_14. URL [http://dx.doi.org/10.1007/978-3-642-37456-2\\_14](http://dx.doi.org/10.1007/978-3-642-37456-2_14).
- S. Char, N. Corley, S. Alamdari, K. K. Yang, and A. P. Amini. Protnote: a multimodal method for protein-function annotation. *Bioinformatics*, 41(5), Apr. 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf170. URL <http://dx.doi.org/10.1093/bioinformatics/btaf170>.
- A. Chervov. Protboost: protein function prediction with py-boost and graph neural networks – cafa5 top2 solution. *arXiv preprint arXiv:2412.04529*, 2024.
- W. T. Clark and P. Radivojac. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics*, 29(13):i53–i61, June 2013a. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt228. URL <http://dx.doi.org/10.1093/bioinformatics/btt228>.
- W. T. Clark and P. Radivojac. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics*, 29(13):i53–i61, 2013b.
- L. Coassolo, N. B. Danneskiold-Samsøe, Q. Nguyen, A. Wiggenhorn, M. Zhao, D. C.-H. Wang, D. Toomer, J. Lone, Y. Wei, A. Patel, I. Liparulo, D. Kavi, L. W. Wat, S. C. Reghupaty, J. J. Kim, T. Asemi, E. Bielczyk-Maczynska, V. L. Li, M. D. Moya-Garzon, N. A. J. Krentz, A. Stahl, D. H.-C. Chou, L. Luo, and K. J. Svensson. Prohormone cleavage prediction uncovers a non-incretin anti-obesity peptide. *Nature*, 641(8061):192–201, Mar. 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08683-y. URL <http://dx.doi.org/10.1038/s41586-025-08683-y>.
- G. O. Consortium. The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021.
- T. G. O. Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Research*, 47(D1):D330–D338, 2019. doi: 10.1093/nar/gky1055.
- T. U. Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D623, 2025. doi: 10.1093/nar/gkae865.
- P. R. Copeland, J. E. Fletcher, B. A. Carlson, D. L. Hatfield, and D. M. Driscoll. A novel rna binding protein, sbp2, is required for the translation of mammalian selenoprotein mRNAs. *The EMBO Journal*, 19(2):306–314, Jan. 2000. ISSN 0261-4189, 1460-2075. doi: 10.1093/emboj/19.2.306. URL <https://link.springer.com/article/10.1093/emboj/19.2.306>.
- W. M. Da Silva, J. Bei, N. Amigo, M. P. Valacco, A. Amadio, Q. Zhang, X. Wu, T. Yu, M. Larzabal, Z. Chen, and A. Cataldi. Quantification of enterohemorrhagic escherichia coli o157:h7 protein abundance by high-throughput proteome. *PLOS ONE*, 13(12):e0208520, Dec. 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0208520. URL <http://dx.doi.org/10.1371/journal.pone.0208520>.
- J. M. Dana, A. Gutmanas, N. Tyagi, G. Qi, C. O’Donovan, M. Martin, and S. Velankar. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Research*, 47(D1):D482–D489, 2019. doi: 10.1093/nar/gky1114.
- M. H. Daniel Han and U. team. Unsloth, 2023. URL <https://github.com/unslothai/unsloth>.
- T. Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL <https://arxiv.org/abs/2307.08691>.
- T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022. URL <https://arxiv.org/abs/2205.14135>.
- M. Dobosz-Bartoszek, M. H. Pinkerton, Z. Otwinowski, S. Chakravarthy, D. Söll, P. R. Copeland, and M. Simonović. Crystal structures of the human elongation factor eefsec suggest a non-canonical mechanism for selenocysteine incorporation. *Nature Communications*, 7(1):12941, Oct. 2016. ISSN 2041-1723. doi: 10.1038/ncomms12941. URL <https://www.nature.com/articles/ncomms12941>.
-

- 
- L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon. Unified language model pre-training for natural language understanding and generation, 2019. URL <https://arxiv.org/abs/1905.03197>.
- A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, Oct. 2022. ISSN 1939-3539. doi: 10.1109/tpami.2021.3095381. URL <http://dx.doi.org/10.1109/TPAMI.2021.3095381>.
- ESM Team. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning, 2024. URL <https://evolutionaryscale.ai/blog/esm-cambrian>.
- Q. Evans and Y. Shen. Interlabelgo+: unraveling label correlations in protein function prediction. *Bioinformatics*, 40(11):btac655, 2024.
- D. Fagegaltier. Characterization of mselb, a novel mammalian elongation factor for selenoprotein translation. *The EMBO Journal*, 19(17):4796–4805, Sept. 2000. ISSN 14602075. doi: 10.1093/emboj/19.17.4796. URL <https://link.springer.com/article/10.1093/emboj/19.17.4796>.
- W. Falcon and T. P. L. team. Pytorch lightning, 2025. URL <https://zenodo.org/doi/10.5281/zenodo.17537232>.
- A. Fallahpour, V. Gureghian, G. J. Filion, A. B. Lindner, and A. Pandi. Codontransformer: a multi-species codon optimizer using context-aware neural networks. *Nature Communications*, 16(1), Apr. 2025a. ISSN 2041-1723. doi: 10.1038/s41467-025-58588-7. URL <http://dx.doi.org/10.1038/s41467-025-58588-7>.
- 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 arXiv:2505.23579*, 2025b.
- X. Fei, M. Chatzianastasis, S. A. Carneiro, H. Abdine, L. P. Petalidis, and M. Vazirgiannis. Prot2text-v2: Protein function prediction with multimodal contrastive alignment, 2025. URL <https://arxiv.org/abs/2505.11194>.
- E. Fischer. Einfluss der configuration auf die wirkung der enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993, Oct. 1894. ISSN 0365-9496. doi: 10.1002/cber.18940270364. URL <http://dx.doi.org/10.1002/cber.18940270364>.
- R. A. Fisher. On the interpretation of  $\chi^2$  from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 85(1):87–94, Dec. 1922. ISSN 1467-985X. doi: 10.1111/j.2397-2335.1922.tb00768.x. URL <http://dx.doi.org/10.1111/j.2397-2335.1922.tb00768.x>.
- J. Fleming, P. Magana, S. Nair, M. Tsenkov, D. Bertoni, I. Pidruchna, M. Q. Lima Afonso, A. Midlik, U. Paramval, A. Židek, A. Laydon, O. Kovalevskiy, J. Pan, J. Cheng, Ž. Avsec, C. Bycroft, L. H. Wong, M. Last, M. Mirdita, M. Steinegger, P. Kohli, M. Váradi, and S. Velankar. Alphafold protein structure database and 3d-beacons: New data and capabilities. *Journal of Molecular Biology*, 437(15):168967, Aug. 2025. ISSN 0022-2836. doi: 10.1016/j.jmb.2025.168967. URL <http://dx.doi.org/10.1016/j.jmb.2025.168967>.
- I. Friedberg. Automated protein function prediction—the genomic challenge. *Briefings in Bioinformatics*, 7(3):225–242, 2006.
- I. Friedberg, P. Radivojac, C. D. Paolis, D. Piovesan, P. Joshi, W. Reade, and A. Howard. Cafa 5 protein function prediction. <https://kaggle.com/competitions/cafa-5-protein-function-prediction>, 2023a. Kaggle.
- I. Friedberg, P. Radivojac, C. D. Paolis, D. Piovesan, P. Joshi, W. Reade, and A. Howard. Cafa 5 protein function prediction. <https://kaggle.com/competitions/cafa-5-protein-function-prediction>, 2023b. Kaggle.
-

- 
- K. Ganesan. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks, 2018. URL <https://arxiv.org/abs/1803.01937>.
- J. N. Gonzalez-Flores, N. Gupta, L. W. DeMong, and P. R. Copeland. The selenocysteine-specific elongation factor contains a novel and multi-functional domain. *Journal of Biological Chemistry*, 287(46):38936–38945, Nov. 2012. ISSN 00219258. doi: 10.1074/jbc.M112.415463. URL <https://linkinghub.elsevier.com/retrieve/pii/S0021925820623559>.
- D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. Wu, Z. Gou, Z. Shao, Z. Li, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, and Z. Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, R. Badkundri, I. Shafkat, J. Gong, A. Derry, R. S. Molina, N. Thomas, Y. Khan, C. Mishra, C. Kim, L. J. Bartie, M. Nemeth, P. D. Hsu, T. Sercu, S. Candido, and A. Rives. Simulating 500 million years of evolution with a language model. *bioRxiv*, 2024. doi: 10.1101/2024.07.01.600583. URL <https://www.biorxiv.org/content/early/2024/07/02/2024.07.01.600583>.
- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. URL <https://arxiv.org/abs/1512.03385>.
- D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus), 2023. URL <https://arxiv.org/abs/1606.08415>.
- T. Heuser, E. E. Dymek, J. Lin, E. F. Smith, and D. Nicastro. The csc connects three major axonemal complexes involved in dynein regulation. *Molecular Biology of the Cell*, 23(16):3143–3155, Aug. 2012. ISSN 1939-4586. doi: 10.1091/mbc.e12-05-0357. URL <http://dx.doi.org/10.1091/mbc.E12-05-0357>.
- B. Hie, E. D. Zhong, B. Berger, and B. Bryson. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288, Jan. 2021. ISSN 1095-9203. doi: 10.1126/science.abd7331. URL <http://dx.doi.org/10.1126/science.abd7331>.
- T. Hilal, B. Y. Killam, M. Grozdanović, M. Dobosz-Bartoszek, J. Loerke, J. Bürger, T. Mielke, P. R. Copeland, M. Simonović, and C. M. T. Spahn. Structure of the mammalian ribosome as it decodes the selenocysteine uga codon. *Science*, 376(6599):1338–1343, June 2022. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abg3875. URL <https://www.science.org/doi/10.1126/science.abg3875>.
- A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration, 2020. URL <https://arxiv.org/abs/1904.09751>.
- E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021. URL <https://arxiv.org/abs/2106.09685>.
- S. Hwang and K. L. Maxwell. Diverse mechanisms of crispr-cas9 inhibition by type ii anti-crispr proteins. *Journal of Molecular Biology*, 435(7):168041, Apr. 2023. ISSN 0022-2836. doi: 10.1016/j.jmb.2023.168041. URL <http://dx.doi.org/10.1016/j.jmb.2023.168041>.
-

- 
- A.-M. Istrate and T. Karaletsos. rbio1 - training scientific reasoning llms with biological world models as soft verifiers. *bioRxiv*, 2025. doi: 10.1101/2025.08.18.670981.
- A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira. Perceiver: General perception with iterative attention, 2021. URL <https://arxiv.org/abs/2103.03206>.
- S. Jain, J. Beazer, J. A. Ruffolo, A. Bhatnagar, and A. Madani. E1: Retrieval-augmented protein encoder models. *bioRxiv*, 2025. doi: 10.1101/2025.11.12.688125. URL <https://www.biorxiv.org/content/10.1101/2025.11.12.688125v1>.
- Y. J. Jang, Q.-Q. Qin, S.-Y. Huang, A. T. John Peter, X.-M. Ding, and B. Kornmann. Accurate prediction of protein function using statistics-informed graph networks. *Nature Communications*, 15:6601, 2024. doi: 10.1038/s41467-024-50955-0.
- P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, and S. Hunter. Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, Jan. 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu031. URL <http://dx.doi.org/10.1093/bioinformatics/btu031>.
- J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL <http://dx.doi.org/10.1038/s41586-021-03819-2>.
- A. Kabir and A. Shehu. Graph neural networks in predicting protein function and interactions. In *Graph Neural Networks - Models and Applications*, pages 543–560. Springer, 2024.
- A. Karpathy. NanoGPT. <https://github.com/karpathy/nanoGPT>, 2022.
- P. Karplus and G. E. Schulz. Refined structure of glutathione reductase at 1.54 Å resolution. *Journal of Molecular Biology*, 195(3):701–729, June 1987. ISSN 0022-2836. doi: 10.1016/0022-2836(87)90191-4. URL [http://dx.doi.org/10.1016/0022-2836\(87\)90191-4](http://dx.doi.org/10.1016/0022-2836(87)90191-4).
- J. Kim, D. Kim, and K.-A. Sohn. Hig2vec: Hierarchical representations of gene ontology and genes in the poincaré ball. *Bioinformatics*, 37(18):2971–2980, 2021. doi: 10.1093/bioinformatics/btab193.
- D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017. URL <https://arxiv.org/abs/1412.6980>.
- M. Kulmanov and R. Hoehndorf. Deepgoplus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429, 2020.
- S. Kumar, M. Suleski, J. M. Craig, A. E. Kasprovicz, M. Sanderford, M. Li, G. Stecher, and S. B. Hedges. Timetree 5: An expanded resource for species divergence times. *Molecular Biology and Evolution*, 39(8), Aug. 2022. ISSN 1537-1719. doi: 10.1093/molbev/msac174. URL <http://dx.doi.org/10.1093/molbev/msac174>.
- W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL <https://arxiv.org/abs/2309.06180>.
- J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL <https://arxiv.org/abs/2301.12597>.
- Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, Mar. 2023. ISSN 1095-9203. doi: 10.1126/science.ade2574. URL <http://dx.doi.org/10.1126/science.ade2574>.
-

- 
- H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning, 2023. URL <https://arxiv.org/abs/2304.08485>.
- S. Liu, J. Zhang, Z. E. Kherraf, S. Sun, X. Zhang, C. Cazin, C. Coutton, R. Zouari, S. Zhao, F. Hu, S. Fourati Ben Mustapha, C. Arnoult, P. F. Ray, and M. Liu. Cfap61 is required for sperm flagellum formation and male fertility in human and mouse. *Development*, 148(23), Dec. 2021. ISSN 1477-9129. doi: 10.1242/dev.199805. URL <http://dx.doi.org/10.1242/dev.199805>.
- Z. Liu, A. Zhang, H. Fei, E. Zhang, X. Wang, K. Kawaguchi, and T.-S. Chua. Prott3: Protein-to-text generation for text-based protein understanding, 2024. URL <https://arxiv.org/abs/2405.12564>.
- Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.
- P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.
- I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. URL <https://arxiv.org/abs/1608.03983>.
- I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2019. URL <https://arxiv.org/abs/1711.05101>.
- A. Ma, A. Zeb, I. Ali, D. Zhao, A. Khan, B. Zhang, J. Zhou, R. Khan, H. Zhang, Y. Zhang, I. Khan, W. Shah, H. Ali, A. R. Javed, H. Ma, and Q. Shi. Biallelic variants in cfap61 cause multiple morphological abnormalities of the flagella and male infertility. *Frontiers in Cell and Developmental Biology*, 9, Jan. 2022. ISSN 2296-634X. doi: 10.3389/fcell.2021.803818. URL <http://dx.doi.org/10.3389/fcell.2021.803818>.
- H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, Mar. 1947. ISSN 0003-4851. doi: 10.1214/aoms/1177730491. URL <http://dx.doi.org/10.1214/aoms/1177730491>.
- N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220, 1967.
- L. McInnes, J. Healy, and S. Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205, Mar. 2017. ISSN 2475-9066. doi: 10.21105/joss.00205. URL <http://dx.doi.org/10.21105/joss.00205>.
- L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL <https://arxiv.org/abs/1802.03426>.
- A. T. Merchant, S. H. King, E. Nguyen, and B. L. Hie. Semantic design of functional de novo genes from a genomic language model. *Nature*, 649(8097):749–758, Nov. 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09749-7. URL <http://dx.doi.org/10.1038/s41586-025-09749-7>.
- R. Méndez, R. Leplae, L. De Maria, and S. J. Wodak. Assessment of blind predictions of protein–protein interactions: Current status of docking methods. *Proteins: Structure, Function, and Bioinformatics*, 52(1): 51–67, May 2003. ISSN 1097-0134. doi: 10.1002/prot.10393. URL <http://dx.doi.org/10.1002/prot.10393>.
- E. Nguyen, M. Poli, M. G. Durrant, B. Kang, D. Katrekar, D. B. Li, L. J. Bartie, A. W. Thomas, S. H. King, G. Brixi, J. Sullivan, M. Y. Ng, A. Lewis, A. Lou, S. Ermon, S. A. Baccus, T. Hernandez-Boussard, C. Ré, P. D. Hsu, and B. L. Hie. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723), Nov. 2024. ISSN 1095-9203. doi: 10.1126/science.ado9336. URL <http://dx.doi.org/10.1126/science.ado9336>.
- C. O’Donovan, M. J. Martin, A. Gattiker, E. Gasteiger, A. Bairoch, and R. Apweiler. High-quality protein knowledge resource: Swiss-prot and trembl. *Briefings in Bioinformatics*, 3(3):275–284, 2002.
-

- 
- OpenAI, :, A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, A. Iftimie, A. Karpenko, A. T. Passos, A. Neitz, A. Prokofiev, A. Wei, A. Tam, A. Bennett, A. Kumar, A. Saraiva, A. Vallone, A. Duberstein, A. Kondrich, A. Mishchenko, A. Applebaum, A. Jiang, A. Nair, B. Zoph, B. Ghorbani, B. Rossen, B. Sokolowsky, B. Barak, B. McGrew, B. Minaiev, B. Hao, B. Baker, B. Houghton, B. McKinzie, B. Eastman, C. Lugaresi, C. Bassin, C. Hudson, C. M. Li, C. de Bourcy, C. Voss, C. Shen, C. Zhang, C. Koch, C. Orsinger, C. Hesse, C. Fischer, C. Chan, D. Roberts, D. Kappler, D. Levy, D. Selsam, D. Dohan, D. Farhi, D. Mely, D. Robinson, D. Tsipras, D. Li, D. Oprica, E. Freeman, E. Zhang, E. Wong, E. Proehl, E. Cheung, E. Mitchell, E. Wallace, E. Ritter, E. Mays, F. Wang, F. P. Such, F. Raso, F. Leoni, F. Tsimpourlas, F. Song, F. von Lohmann, F. Sulit, G. Salmon, G. Parascandolo, G. Chabot, G. Zhao, G. Brockman, G. Leclerc, H. Salman, H. Bao, H. Sheng, H. Andrin, H. Bagherinezhad, H. Ren, H. Lightman, H. W. Chung, I. Kivlichan, I. O’Connell, I. Osband, I. C. Gilaberte, I. Akkaya, I. Kostrikov, I. Sutskever, I. Kofman, J. Pachocki, J. Lennon, J. Wei, J. Harb, J. Twore, J. Feng, J. Yu, J. Weng, J. Tang, J. Yu, J. Q. Candela, J. Palermo, J. Parish, J. Heidecke, J. Hallman, J. Rizzo, J. Gordon, J. Uesato, J. Ward, J. Huizinga, J. Wang, K. Chen, K. Xiao, K. Singhal, K. Nguyen, K. Cobbe, K. Shi, K. Wood, K. Rimbach, K. Gu-Lemberg, K. Liu, K. Lu, K. Stone, K. Yu, L. Ahmad, L. Yang, L. Liu, L. Maksin, L. Ho, L. Fedus, L. Weng, L. Li, L. McCallum, L. Held, L. Kuhn, L. Kondraciuk, L. Kaiser, L. Metz, M. Boyd, M. Trebacz, M. Joglekar, M. Chen, M. Tintor, M. Meyer, M. Jones, M. Kaufer, M. Schwarzer, M. Shah, M. Yatbaz, M. Y. Guan, M. Xu, M. Yan, M. Glaese, M. Chen, M. Lampe, M. Malek, M. Wang, M. Fradin, M. McClay, M. Pavlov, M. Wang, M. Wang, M. Murati, M. Bavarian, M. Rohaninejad, N. McAleese, N. Chowdhury, N. Chowdhury, N. Ryder, N. Tezak, N. Brown, O. Nachum, O. Boiko, O. Murk, O. Watkins, P. Chao, P. Ashbourne, P. Izmailov, P. Zhokhov, R. Dias, R. Arora, R. Lin, R. G. Lopes, R. Gaon, R. Miyara, R. Leike, R. Hwang, R. Garg, R. Brown, R. James, R. Shu, R. Cheu, R. Greene, S. Jain, S. Altman, S. Toizer, S. Toyer, S. Miserendino, S. Agarwal, S. Hernandez, S. Baker, S. McKinney, S. Yan, S. Zhao, S. Hu, S. Santurkar, S. R. Chaudhuri, S. Zhang, S. Fu, S. Papay, S. Lin, S. Balaji, S. Sanjeev, S. Sidor, T. Broda, A. Clark, T. Wang, T. Gordon, T. Sanders, T. Patwardhan, T. Sottiaux, T. Degry, T. Dimson, T. Zheng, T. Garipov, T. Stasi, T. Bansal, T. Creech, T. Peterson, T. Eloundou, V. Qi, V. Kosaraju, V. Monaco, V. Pong, V. Fomenko, W. Zheng, W. Zhou, W. McCabe, W. Zaremba, Y. Dubois, Y. Lu, Y. Chen, Y. Cha, Y. Bai, Y. He, Y. Zhang, Y. Wang, Z. Shao, and Z. Li. Openai o1 system card, 2024. URL <https://arxiv.org/abs/2412.16720>.
- A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL <https://arxiv.org/abs/1912.01703>.
- A. Pawluk, A. R. Davidson, and K. L. Maxwell. Anti-crispr: discovery, mechanism and function. *Nature Reviews Microbiology*, 16(1):12–17, Oct. 2017. ISSN 1740-1534. doi: 10.1038/nrmicro.2017.120. URL <http://dx.doi.org/10.1038/nrmicro.2017.120>.
- K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, Nov. 1901. ISSN 1941-5990. doi: 10.1080/14786440109462720. URL <http://dx.doi.org/10.1080/14786440109462720>.
- E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, and T. E. Ferrin. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science*, 30(1): 70–82, 2021. doi: 10.1002/pro.3943.
- D. Piovesan, D. Zago, P. Joshi, M. C. De Paolis Kaluza, M. Mehdiabadi, R. Ramola, A. M. Monzon, W. Reade, I. Friedberg, P. Radivojac, et al. Cafa-evaluator: a python tool for benchmarking ontological classification methods. *Bioinformatics Advances*, 4(1):vbae043, 2024.
- Z. Qiu, Z. Wang, B. Zheng, Z. Huang, K. Wen, S. Yang, R. Men, L. Yu, F. Huang, S. Huang, D. Liu, J. Zhou, and J. Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=1b7wh04SfY>.
- A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multi-task learners. *OpenAI*, 2019. URL [https://cdn.openai.com/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf). Accessed: 2024-11-15.
-

- 
- P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227, 2013.
- S. Rao and M. G. Rossmann. Comparison of super-secondary structures in proteins. *Journal of Molecular Biology*, 76(2):241–256, May 1973. ISSN 0022-2836. doi: 10.1016/0022-2836(73)90388-4. URL [http://dx.doi.org/10.1016/0022-2836\(73\)90388-4](http://dx.doi.org/10.1016/0022-2836(73)90388-4).
- J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '20, page 3505–3506. ACM, Aug. 2020. doi: 10.1145/3394486.3406703. URL <http://dx.doi.org/10.1145/3394486.3406703>.
- P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 448–453. Morgan Kaufmann, 1995.
- B. Rost. Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection*, 12(2):85–94, Feb. 1999. ISSN 1741-0126. doi: 10.1093/protein/12.2.85. URL <http://dx.doi.org/10.1093/protein/12.2.85>.
- Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8, November 2015.
- Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- M. Simonović and A. K. Puppala. On elongation factor eefsec, its role and mechanism during selenium incorporation into nascent selenoproteins. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1862(11):2463–2472, Nov. 2018. ISSN 03044165. doi: 10.1016/j.bbagen.2018.03.018. URL <https://linkinghub.elsevier.com/retrieve/pii/S0304416518300801>.
- A. Singh, A. Fry, A. Perelman, A. Tart, A. Ganesh, A. El-Kishky, A. McLaughlin, A. Low, A. Ostrow, A. Ananthram, A. Nathan, A. Luo, A. Helyar, A. Madry, A. Efremov, A. Spyra, A. Baker-Whitcomb, A. Beutel, A. Karpenko, A. Makelov, A. Neitz, A. Wei, A. Barr, A. Kirchmeyer, A. Ivanov, A. Christakis, A. Gillespie, A. Tam, A. Bennett, A. Wan, A. Huang, A. M. Sandjideh, A. Yang, A. Kumar, A. Saraiva, A. Vallone, A. Gheorghe, A. G. Garcia, A. Braunstein, A. Liu, A. Schmidt, A. Mereskin, A. Mishchenko, A. Applebaum, A. Rogerson, A. Rajan, A. Wei, A. Kotha, A. Srivastava, A. Agrawal, A. Vijayvergiya, A. Tyra, A. Nair, A. Nayak, B. Eggers, B. Ji, B. Hoover, B. Chen, B. Barak, B. Minaiev, B. Hao, B. Baker, B. Lightcap, B. McKinzie, B. Wang, B. Quinn, B. Fioca, B. Hsu, B. Yang, B. Yu, B. Zhang, B. Brenner, C. R. Zetino, C. Raymond, C. Lugaresi, C. Paz, C. Hudson, C. Whitney, C. Li, C. Chen, C. Cole, C. Voss, C. Ding, C. Shen, C. Huang, C. Colby, C. Hallacy, C. Koch, C. Lu, C. Kaplan, C. Kim, C. Minott-Henriques, C. Frey, C. Yu, C. Czarnecki, C. Reid, C. Wei, C. Decareaux, C. Scheau, C. Zhang, C. Forbes, D. Tang, D. Goldberg, D. Roberts, D. Palmie, D. Kappler, D. Levine, D. Wright, D. Leo, D. Lin, D. Robinson, D. Grabb, D. Chen, D. Lim, D. Salama, D. Bhattacharjee, D. Tsipras, D. Li, D. Yu, D. Strouse, D. Williams, D. Hunn, E. Bayes, E. Arbus, E. Akyurek, E. Y. Le, E. Widmann, E. Yani, E. Proehl, E. Sert, E. Cheung, E. Schwartz, E. Han, E. Jiang, E. Mitchell, E. Sigler, E. Wallace, E. Ritter, E. Kavanaugh, E. Mays, E. Nikishin, F. Li, F. P. Such, F. de Avila Belbute Peres, F. Raso, F. Bekerman, F. Tsimpourlas, F. Chantzis, F. Song, F. Zhang, G. Raila, G. McGrath, G. Briggs, G. Yang, G. Parascandolo, G. Chabot, G. Kim, G. Zhao, G. Valiant, G. Leclerc, H. Salman, H. Wang, H. Sheng, H. Jiang, H. Wang, H. Jin, H. Sikchi, H. Schmidt, H. Aspegren, H. Chen, H. Qiu, H. Lightman, I. Covert, I. Kivlichan, I. Silber, I. Sohl, I. Hammoud, I. Clavera, I. Lan, I. Akkaya, I. Kostrikov, I. Kofman, I. Etinger, I. Singal, J. Hehir, J. Huh, J. Pan, J. Wilczynski, J. Pachocki, J. Lee, J. Quinn, J. Kiros, J. Kalra, J. Samaroo, J. Wang, J. Wolfe, J. Chen, J. Wang, J. Harb, J. Han, J. Wang, J. Zhao, J. Chen, J. Yang, J. Tworek, J. Chand, J. Landon, J. Liang, J. Lin, J. Liu, J. Wang, J. Tang, J. Yin, J. Jang, J. Morris, J. Flynn, J. Ferstad, J. Heidecke, J. Fishbein, J. Hallman, J. Grant, J. Chien, J. Gordon, J. Park, J. Liss, J. Kraaijeveld, J. Guay, J. Mo, J. Lawson, J. McGrath, J. Vendrow, J. Jiao, J. Lee, J. Steele, J. Wang, J. Mao, K. Chen, K. Hayashi, K. Xiao, K. Salahi, K. Wu, K. Sekhri, K. Sharma, K. Singhal, K. Li, K. Nguyen, K. Gu-Lemberg, K. King, K. Liu, K. Stone, K. Yu, K. Ying, K. Georgiev, K. Lim, K. Tirumala, K. Miller, L. Ahmad, L. Lv, L. Clare, L. Fauconnet, L. Itow, L. Yang, L. Romaniuk, L. Anise, L. Byron, L. Pathak, L. Maksin, L. Lo, L. Ho, L. Jing, L. Wu, L. Xiong, L. Mamitsuka, L. Yang, L. McCallum, L. Held, L. Bourgeois, L. Engstrom, L. Kuhn, L. Feuvrier, L. Zhang, L. Switzer, L. Kondraciuk, L. Kaiser, M. Joglekar, M. Singh, M. Shah, M. Stratta, M. Williams, M. Chen, M. Sun, M. Cayton, M. Li, M. Zhang, M. Aljube
-

- 
- M. Nichols, M. Haines, M. Schwarzer, M. Gupta, M. Shah, M. Huang, M. Dong, M. Wang, M. Glaese, M. Carroll, M. Lampe, M. Malek, M. Sharman, M. Zhang, M. Wang, M. Pokrass, M. Florian, M. Pavlov, M. Wang, M. Chen, M. Wang, M. Feng, M. Bavarian, M. Lin, M. Abdool, M. Rohaninejad, N. Soto, N. Staudacher, N. LaFontaine, N. Marwell, N. Liu, N. Preston, N. Turley, N. Ansman, N. Blades, N. Pancha, N. Mikhaylin, N. Felix, N. Handa, N. Rai, N. Keskar, N. Brown, O. Nachum, O. Boiko, O. Murk, O. Watkins, O. Gleeson, P. Mishkin, P. Lesiewicz, P. Baltescu, P. Belov, P. Zhokhov, P. Pronin, P. Guo, P. Thacker, Q. Liu, Q. Yuan, Q. Liu, R. Dias, R. Puckett, R. Arora, R. T. Mullapudi, R. Gaon, R. Miyara, R. Song, R. Aggarwal, R. Marsan, R. Yemiru, R. Xiong, R. Kshirsagar, R. Nuttall, R. Tsiupa, R. Eldan, R. Wang, R. James, R. Ziv, R. Shu, R. Nigmatullin, S. Jain, S. Talaie, S. Altman, S. Arnesen, S. Toizer, S. Toyer, S. Miserendino, S. Agarwal, S. Yoo, S. Heon, S. Ethersmith, S. Grove, S. Taylor, S. Bubeck, S. Banerjee, S. Amdo, S. Zhao, S. Wu, S. Santurkar, S. Zhao, S. R. Chaudhuri, S. Krishnaswamy, Shuaiqi, Xia, S. Cheng, S. Anadkat, S. P. Fishman, S. Tobin, S. Fu, S. Jain, S. Mei, S. Egoian, S. Kim, S. Golden, S. Mah, S. Lin, S. Imm, S. Sharpe, S. Yadlowsky, S. Choudhry, S. Eum, S. Sanjeev, T. Khan, T. Stramer, T. Wang, T. Xin, T. Gogineni, T. Christianson, T. Sanders, T. Patwardhan, T. Degry, T. Shadwell, T. Fu, T. Gao, T. Garipov, T. Srisankarajah, T. Sherbakov, T. Kaftan, T. Hiratsuka, T. Wang, T. Song, T. Zhao, T. Peterson, V. Kharitonov, V. Chernova, V. Kosaraju, V. Kuo, V. Pong, V. Verma, V. Petrov, W. Jiang, W. Zhang, W. Zhou, W. Xie, W. Zhan, W. McCabe, W. DePue, W. Ellsworth, W. Bain, W. Thompson, X. Chen, X. Qi, X. Xiang, X. Shi, Y. Dubois, Y. Yu, Y. Khakbaz, Y. Wu, Y. Qian, Y. T. Lee, Y. Chen, Y. Zhang, Y. Xiong, Y. Tian, Y. Cha, Y. Bai, Y. Yang, Y. Yuan, Y. Li, Y. Zhang, Y. Yang, Y. Jin, Y. Jiang, Y. Wang, Y. Wang, Y. Liu, Z. Stubenvoll, Z. Dou, Z. Wu, and Z. Wang. Openai gpt-5 system card, 2025. URL <https://arxiv.org/abs/2601.03267>.
- N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. URL <http://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf>.
- G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, S. Kaplan, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran, and D. Lancet. The genecards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics*, 54(1), June 2016. ISSN 1934-340X. doi: 10.1002/cpbi.5. URL <http://dx.doi.org/10.1002/cpbi.5>.
- D. Szklarczyk, K. Nastou, M. Koutrouli, R. Kirsch, F. Mehryary, R. Hachilif, D. Hu, M. E. Peluso, Q. Huang, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J. Jensen, and C. von Mering. The string database in 2025: protein networks with directionality of regulation. *Nucleic Acids Research*, 53(D1):D730–D737, Nov. 2024. ISSN 1362-4962. doi: 10.1093/nar/gkaf1113. URL <http://dx.doi.org/10.1093/nar/gkaf1113>.
- H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL <https://arxiv.org/abs/2307.09288>.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023. URL <https://arxiv.org/abs/1706.03762>.
- P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks, 2018. URL <https://arxiv.org/abs/1710.10903>.
- L. C. Vieira, M. L. Handojo, and C. O. Wilke. Medium-sized protein language models perform well at transfer learning on realistic datasets. *Scientific Reports*, 15:21400, 2025. doi: 10.1038/s41598-025-05674-x.
- T. Walton, M. Gui, S. Velkova, M. R. Fassad, R. A. Hirst, E. Haarman, C. O’Callaghan, M. Bottier, T. Burgoyne, H. M. Mitchison, and A. Brown. Axonemal structures reveal mechanoregulatory and disease mechanisms. *Nature*, 618(7965):625–633, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06140-2. URL <http://dx.doi.org/10.1038/s41586-023-06140-2>.
-

- 
- B. Wang, Y. Geng, X. Cheng, B. Chen, Z. Bei, W. Wang, J. Tang, and L. Song. Protgo: universal protein function prediction utilizing multi-modal gene ontology knowledge. *Bioinformatics*, 41(7):btaf390, 2025a.
- C. Wang, H. Fan, R. Quan, and Y. Yang. Protchatgpt: Towards understanding proteins with large language models, 2025b. URL <https://arxiv.org/abs/2402.09649>.
- S. Wang, R. You, S. Yao, Y. Xiong, C. Bai, C. Zhang, H. Zhang, X. Huang, and J. Huang. Netgo 3.0: Protein language model improves large-scale functional annotations. *Genomics, Proteomics Bioinformatics*, 21(2): 349–362, 2023.
- W. Wang, Y. Shuai, M. Zeng, W. Fan, and M. Li. Dpfunc: accurately predicting protein function via deep learning with domain-guided structure information. *Nature Communications*, 16:70, 2025c. doi: 10.1038/s41467-024-54816-8.
- Z. Wang, Z. Ma, Z. Cao, C. Zhou, J. Zhang, and Y. Gao. Prot2chat: Protein llm with early-fusion of text, sequence and structure, 2025d. URL <https://arxiv.org/abs/2502.06846>.
- J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragothe, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL <http://dx.doi.org/10.1038/s41586-023-06415-8>.
- J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL <https://arxiv.org/abs/2201.11903>.
- J. D. Westbrook, J. Y. Young, C. Shao, Z. Feng, V. Guranovic, C. L. Lawson, B. Vallat, P. D. Adams, J. M. Berrisford, G. Bricogne, K. Diederichs, R. P. Joosten, P. Keller, N. W. Moriarty, O. V. Sobolev, S. Velankar, C. Vonrhein, D. G. Waterman, G. Kurisu, H. M. Berman, S. K. Burley, and E. Peisach. Pdbx/mmcif ecosystem: Foundational semantic tools for structural biology. *Journal of Molecular Biology*, 434(11):167599, June 2022. ISSN 0022-2836. doi: 10.1016/j.jmb.2022.167599. URL <http://dx.doi.org/10.1016/j.jmb.2022.167599>.
- R. K. Wierenga, P. Terpstra, and W. G. Hol. Prediction of the occurrence of the ADP-binding  $\beta\alpha\beta$ -fold in proteins, using an amino acid sequence fingerprint. *Journal of Molecular Biology*, 187(1):101–107, Jan. 1986. ISSN 0022-2836. doi: 10.1016/0022-2836(86)90409-2. URL [http://dx.doi.org/10.1016/0022-2836\(86\)90409-2](http://dx.doi.org/10.1016/0022-2836(86)90409-2).
- T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Huggingface’s transformers: State-of-the-art natural language processing, 2020. URL <https://arxiv.org/abs/1910.03771>.
- K. E. Wu, H. Chang, and J. Zou. Proteinclip: enhancing protein language models with natural language, May 2024. URL <http://dx.doi.org/10.1101/2024.05.14.594226>.
- H. Xiao, W. Lin, X. Chen, H. Wang, K. Chen, J. Li, Y. Sun, S. Dai, B. Wu, and Q. Ye. Stella: Towards protein function prediction with multimodal llms integrating sequence-structure representations, 2025a. URL <https://arxiv.org/abs/2506.03800>.
- Y. Xiao, W. Zhao, J. Zhang, Y. Jin, H. Zhang, Z. Ren, R. Sun, H. Wang, G. Wan, P. Lu, X. Luo, Y. Zhang, J. Zou, Y. Sun, and W. Wang. Protein large language models: A comprehensive survey, 2025b. URL <https://arxiv.org/abs/2502.17504>.
- A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
-

- 
- J. Yang, A. Roy, and Y. Zhang. Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, 41(D1):D1096–D1103, 2012.
- Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.
- E. Zelikman, Y. Wu, J. Mu, and N. D. Goodman. Star: Bootstrapping reasoning with reasoning, 2022. URL <https://arxiv.org/abs/2203.14465>.
- T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert, 2020. URL <https://arxiv.org/abs/1904.09675>.
- Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, and J. Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.
- C. Zheng, S. Liu, M. Li, X.-H. Chen, B. Yu, C. Gao, K. Dang, Y. Liu, R. Men, A. Yang, J. Zhou, and J. Lin. Group sequence policy optimization, 2025. URL <https://arxiv.org/abs/2507.18071>.
- N. Zhou, Y. Jiang, T. R. Bergquist, A. J. Lee, B. Z. Kacsoh, A. W. Crocker, K. A. Lewis, G. Georghiou, H. N. Nguyen, M. N. Hamid, et al. The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, 20(1):244, 2019.

---

# BioReason-Pro: Advancing Protein Function Prediction with Multimodal Biological Reasoning

## *Supplementary Material*

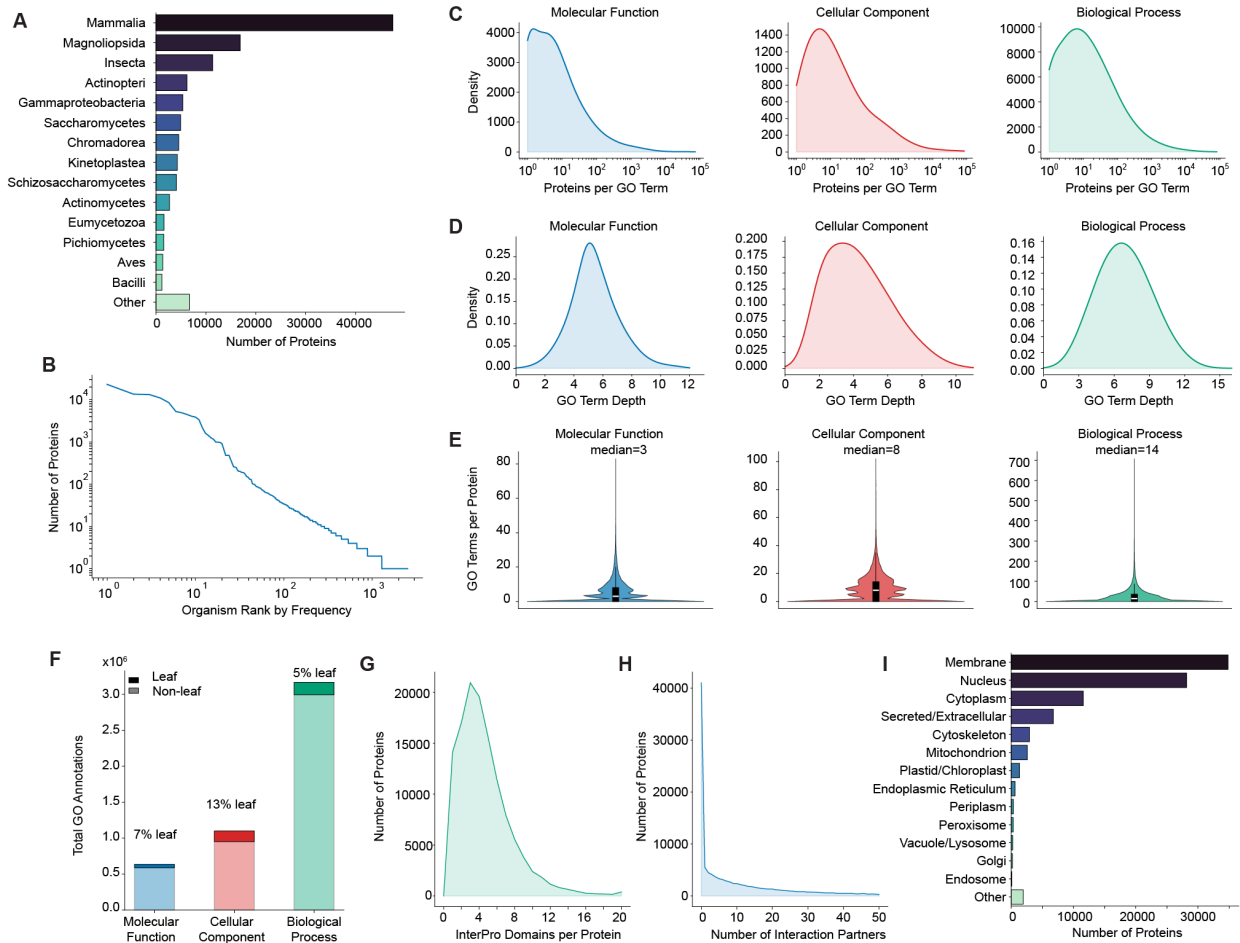
---

<b>A</b>	<b>Supplementary Figures</b> .....	<b>3</b>
	Figure S1: Training dataset statistics .....	3
	Figure S2: GO-GPT attention enrichment at DNA-binding residues .....	4
	Figure S3: LLM judge scores across protein properties and trace lengths .....	5
	Figure S4: LLM judge scores by organism and taxonomic class .....	6
	Figure S5: Human expert evaluation across scoring axes and protein properties .....	7
	Figure S6: Within-aspect GO term embedding analysis .....	8
	Figure S7: Cross-aspect GO term embedding analysis .....	9
	Figure S8: UMAP KDE with cross-aspect GO term pairs .....	10
	Figure S9: Training dynamics of the GO-GPT model .....	11
<b>B</b>	<b>Appendix</b> .....	<b>12</b>
	B.1 GO-GPT Attention Analysis at DNA-Binding Sites .....	12
	B.2 Cross-Aspect GO Term Pairs on the UMAP Embedding Landscape .....	13
	B.3 Architectural and Training Explorations for GO-GPT .....	14
	B.4 Architectural and Training Explorations for BioReason-Pro .....	15
	B.5 Statistical Tests .....	17
	B.6 GO-GPT Hyperparameters .....	24
	B.7 BioReason-Pro Hyperparameters .....	24
<b>C</b>	<b>Supplementary Text</b> .....	<b>27</b>
	C.1 Reasoning Data Generation Prompt .....	27
	C.2 BioReason-Pro Inference Prompt .....	32
	C.3 LLM-as-a-Judge Evaluation Prompt .....	33
	C.4 Human Expert Evaluation Form .....	35
	C.5 Error Attribution for Human Evaluations .....	37
	C.6 eEFSec (P57772) Full Reasoning Trace (SFT) .....	38
	C.7 eEFSec (P57772) Full Reasoning Trace (RL) .....	40
	C.8 eEFSec (P57772) Full Reasoning Trace (GPT-5.2 Thinking High) .....	42
	C.9 CFAP61 (Q8NHU2) Full Reasoning Trace (SFT) .....	43

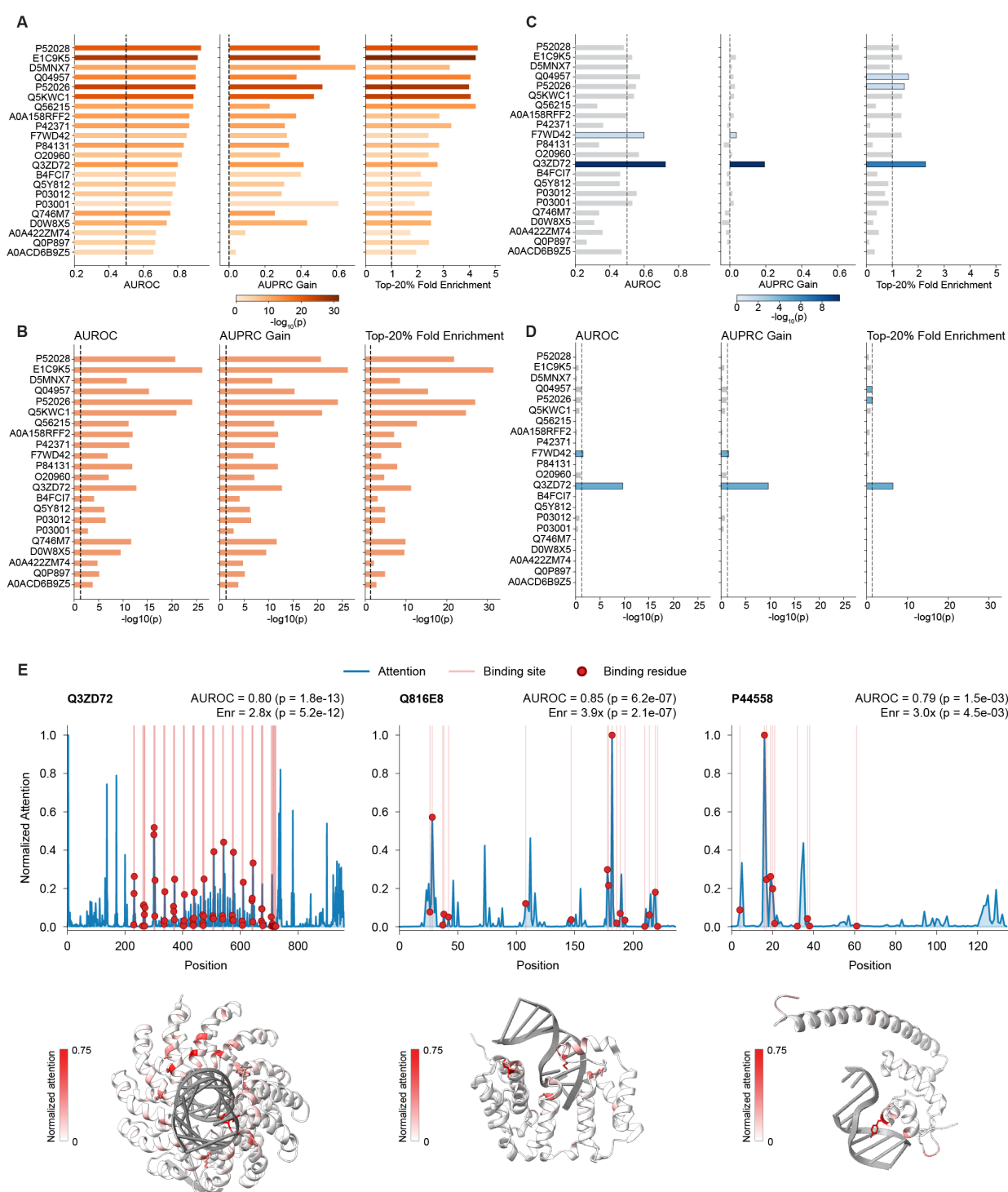
---

C.10	CFAP61 (Q8NHU2) Full Reasoning Trace (RL) .....	45
C.11	CFAP61 (Q8NHU2) Full Reasoning Trace (GPT-5.2 Thinking High) .....	47
C.12	EvoAcr1 Full Reasoning Traces (SFT) .....	48
C.13	EvoAcr1 Full Reasoning Traces (RL) .....	53
C.14	EvoAcr2 Full Reasoning Traces (SFT) .....	57
C.15	EvoAcr2 Full Reasoning Traces (RL) .....	62

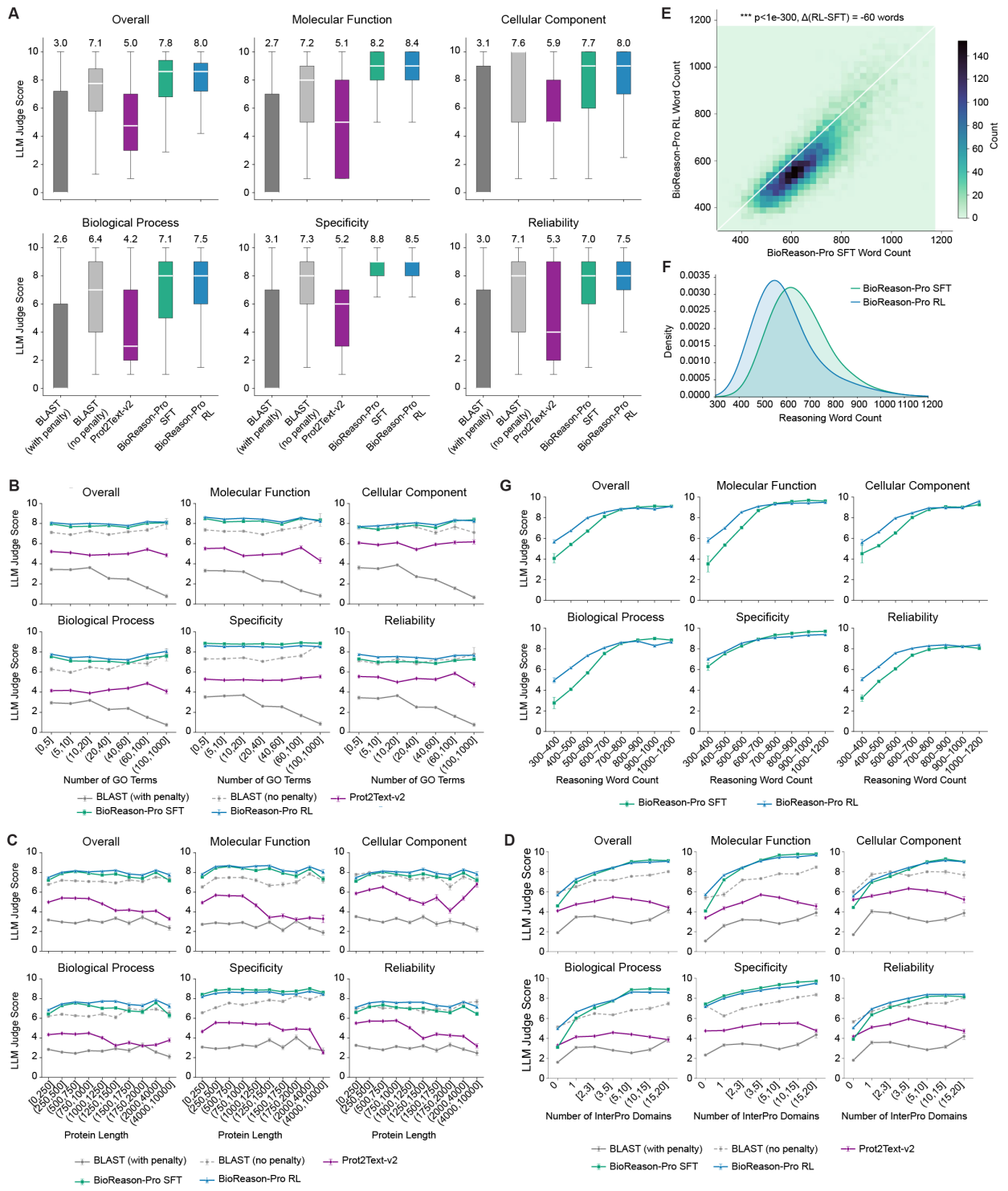
## A. Supplementary Figures



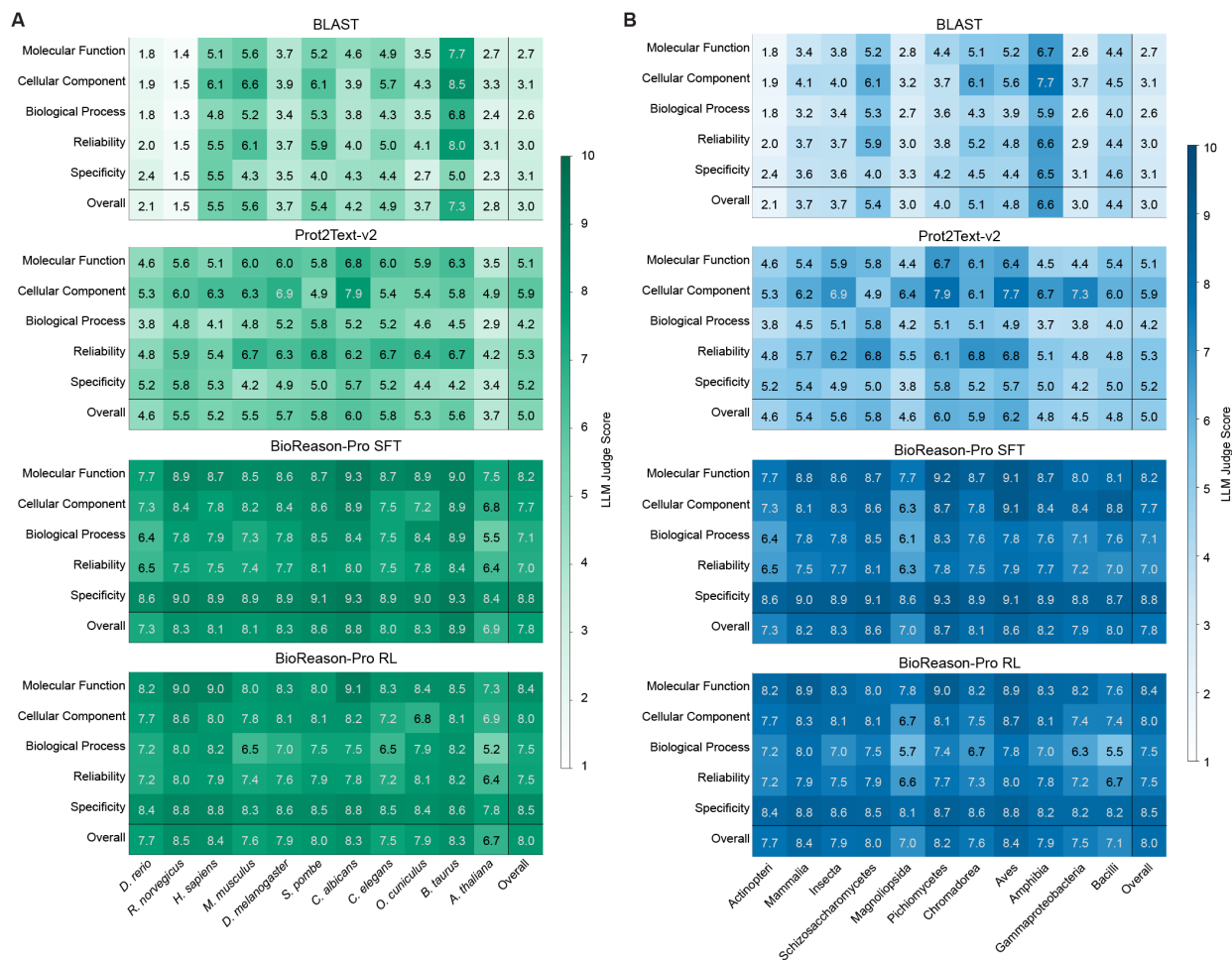
**Figure S1 | Training dataset statistics.** (A) **Taxonomic distribution.** Proteins per taxonomic class showing strong mammalian bias with broad coverage across vertebrates, insects, plants, fungi, and bacteria. (B) **Organism frequency.** Rank-frequency plot on log-log scale revealing a long-tail distribution, with a small number of model organisms dominating protein counts. (C) **GO term frequency per aspect.** Kernel density estimates of proteins annotated per GO term on a log scale, showing most terms are rare with few appearing across many proteins. (D) **GO term depth per aspect.** Ontology depth distribution for unique GO terms in each aspect. (E) **GO terms per protein.** Violin plots of GO annotation counts per protein for each aspect, with median values of 3 (MF), 8 (CC), and 14 (BP). (F) **Leaf versus non-leaf annotations.** Proportion of annotations corresponding to leaf terms (no children in the ontology) versus non-leaf terms per aspect. Leaf terms constitute 7% of MF, 13% of CC, and 5% of BP annotations. (G) **InterPro domain coverage.** Distribution of InterPro domain annotations per protein, with most proteins having fewer than 10 annotated domains. (H) **Protein-protein interaction coverage.** Distribution of interaction partners per protein from STRING, showing a long-tail distribution. (I) **Subcellular localization.** Proteins per broad subcellular compartment, with membrane, nucleus, and cytoplasm as the dominant localizations.



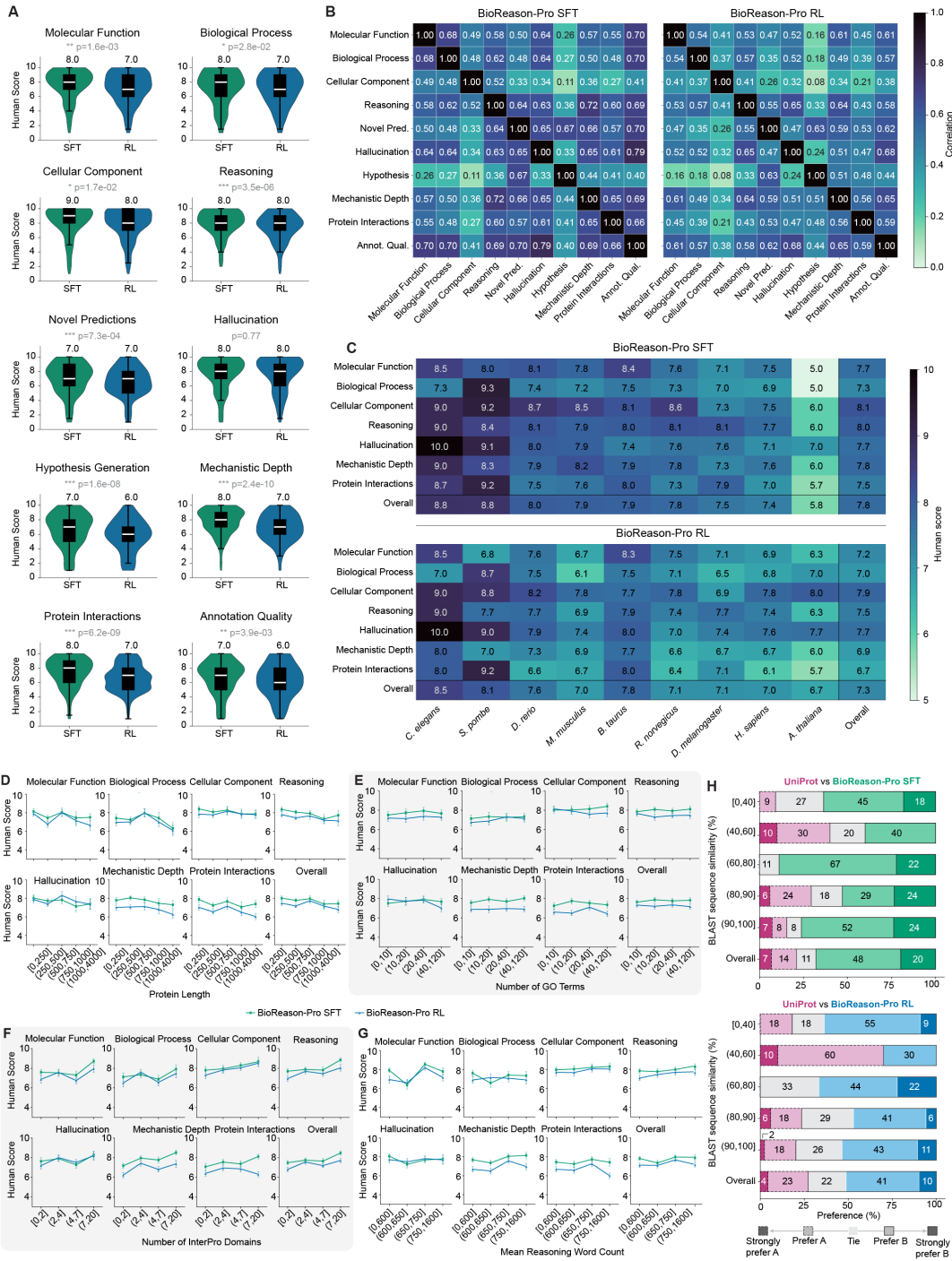
**Figure S2 | GO-GPT attention enrichment at DNA-binding residues.** (A) AUROC, AUPRC Gain, and top-20% fold enrichment for a representative subset of the 63 non-training BioLiP proteins where GO-GPT predicted DNA binding, sorted by fold enrichment. Full per-protein statistics in Table S1. Bar color encodes  $-\log_{10}(p)$ ; gray bars are non-significant ( $p \geq 0.05$ ). Dashed lines: chance level. Across all 63 proteins, 59/63 are significant (Mann–Whitney U); 55/63 (hypergeometric). (B)  $-\log_{10}(p)$  values for (A). AUROC and AUPRC Gain: Mann–Whitney U; fold enrichment: hypergeometric. Dashed line:  $p = 0.05$ . (C, D) ESM2 residue embedding L2 norm baseline (same metrics and protein ordering), testing whether binding-site discrimination is present in input representations before cross-attention (Section 4.5.2). (E) Per-residue attention for three additional proteins (Q3ZD72, Q816E8, P44558) with protein–DNA structures colored.



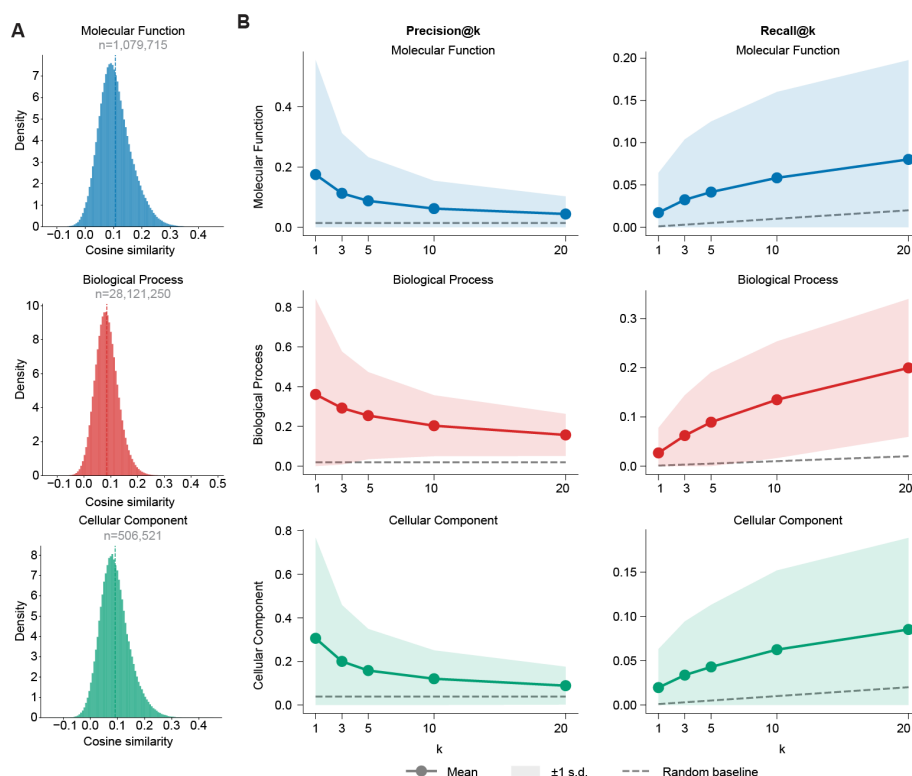
**Figure S3 | LLM-as-judge evaluation across models and protein properties. (A) Score distributions per evaluation axis.** Box plots of LLM judge scores for BLAST (with or without penalty), Prot2Text-v2, BioReason-Pro SFT, and BioReason-Pro RL across all five evaluation axes. Mean values shown above each box. **(B) Performance by GO term count.** Mean scores binned by total number of GO annotations per protein. **(C) Performance by protein length.** Mean scores binned by protein sequence length. **(D) Performance by InterPro domain count.** Mean scores binned by number of InterPro domain annotations per protein. **(E) SFT versus RL trace length.** 2D density heatmap of per-protein reasoning word counts showing that SFT consistently produces longer reasoning traces than RL for the same proteins (mean  $\Delta = -60$  words, paired Wilcoxon  $p < 10^{-300}$ ). **(F) Trace length distributions.** Kernel density estimates of reasoning word counts for SFT and RL. **(G) Judge score by reasoning length.** Mean scores binned by word count for SFT and RL.



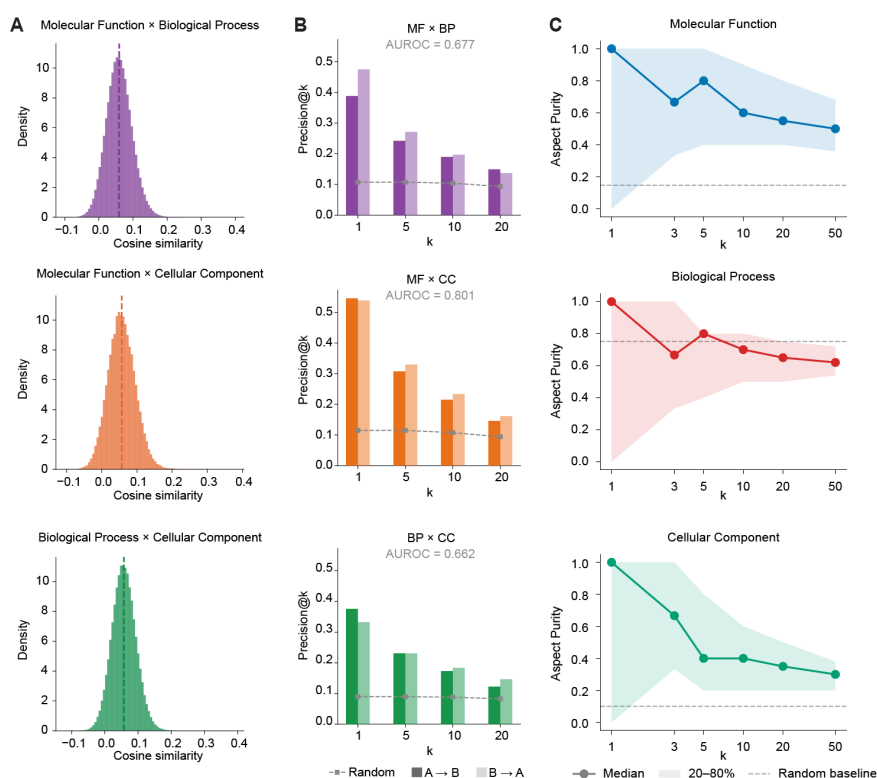
**Figure S4 | LLM-as-judge scores stratified by organism and taxonomic class. (A) Performance by organism.** Heatmaps of mean scores for the ten most frequent organisms across four models (BLAST, Prot2Text-v2, BioReason-Pro SFT, BioReason-Pro RL), with rows representing five evaluation axes and Overall, and columns representing organisms sorted by frequency. BLAST performance varies sharply across organisms, with very low scores for some species (e.g., *D. rerio*, *R. norvegicus*) where BLAST-based transfer is least effective. Prot2Text-v2 is more uniform but remains in the 4–6 range. BioReason-Pro SFT and RL both maintain consistently high scores (7–9) across all organisms, with RL showing a modest advantage in Reliability. The advantage of BioReason-Pro over baselines is most pronounced for organisms where BLAST performs poorly, consistent with the generalization analysis in Section 4.4.5. **(B) Performance by taxonomic class.** Heatmaps of mean scores for the ten most frequent taxonomic classes, arranged analogously. BioReason-Pro generalizes broadly across taxonomic diversity, maintaining stable performance from vertebrates (Actinopteri, Mammalia) through invertebrates (Insecta, Chromadorea) and unicellular organisms (Schizosaccharomycetes, Gammaproteobacteria). BLAST scores degrade most for underrepresented classes, whereas BioReason-Pro shows minimal variation, indicating that the model learned transferable functional reasoning rather than organism-specific annotation patterns.



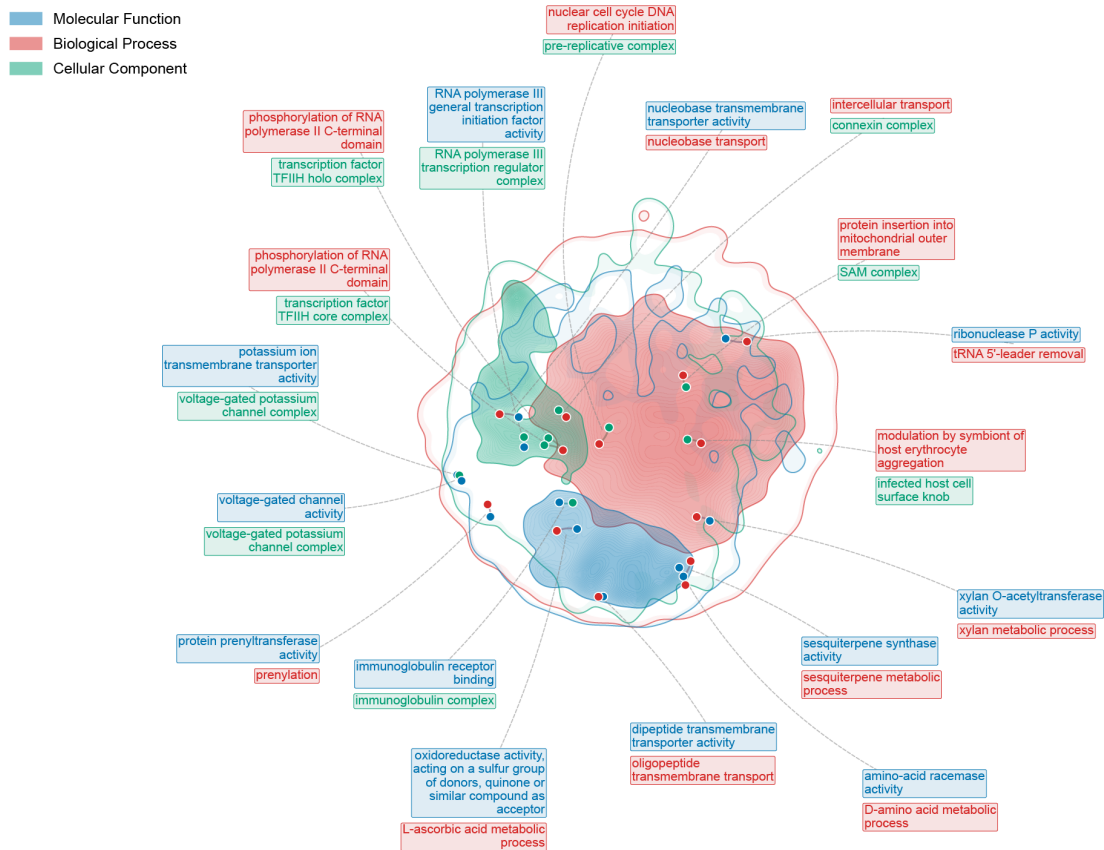
**Figure S5 | Human expert evaluation across all axes and protein properties. (A) Per-axis score distributions.** Violin plots for BioReason-Pro SFT and RL across all ten evaluation axes with Mann–Whitney U significance tests. SFT scores significantly higher on most axes, with the largest differences on Mechanistic Depth and Hypothesis Generation. **(B) Inter-axis correlation matrices.** Pairwise Pearson correlations among all ten axes for SFT and RL. Reasoning, Mechanistic Depth, and Hallucination form a correlated cluster in both models. **(C) Scores by organism.** Heatmap of mean human scores for organisms with at least three evaluated proteins, showing stable performance across organisms for both models. **(D) Performance by protein length.** Mean human scores binned by sequence length for BioReason-Pro SFT and RL across all evaluation axes. **(E) Performance by GO term count.** Mean human scores binned by total GO annotations per protein. **(F) Performance by InterPro domain count.** Mean human scores binned by number of InterPro domain annotations per protein. **(G) Performance by reasoning trace length.** Mean human scores binned by average reasoning word count. **(H) Preference versus UniProt by sequence similarity.** Expert preference distributions for SFT (top) and RL (bottom) compared to UniProt ground truth, binned by BLAST sequence similarity.



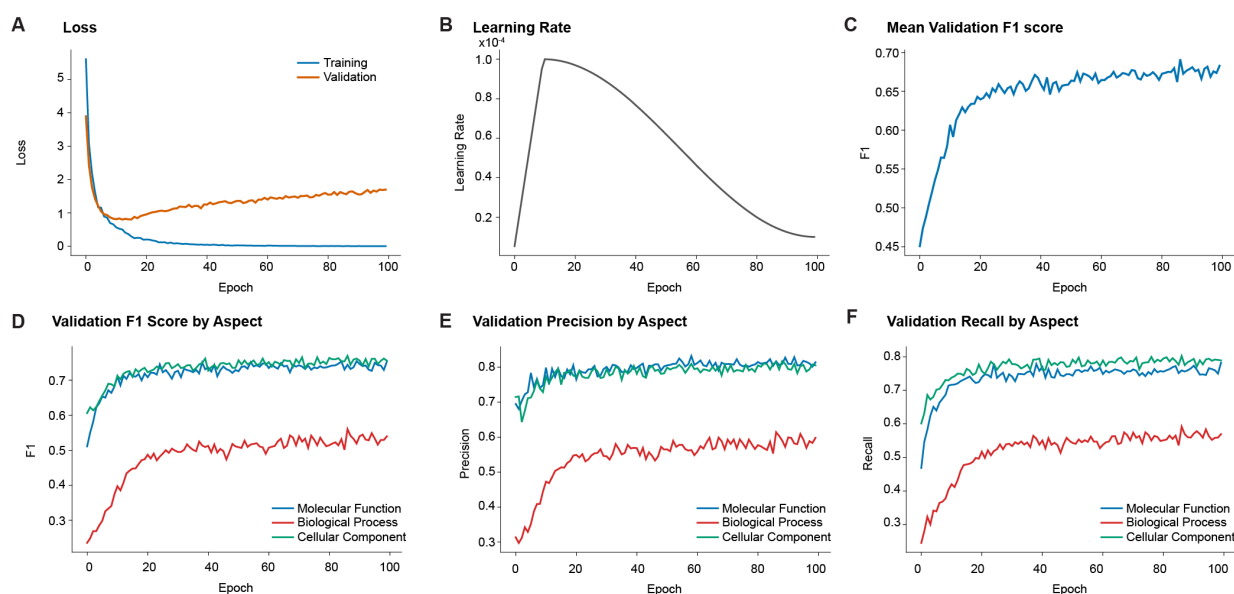
**Figure S6 | Within-aspect analysis of GO-GPT learned GO term embeddings.** (A) **Cosine similarity distribution.** Per-aspect density of pairwise cosine similarities among all GO term embeddings, with total pair counts shown (MF: 1.08M, BP: 28.1M, CC: 507K). Dashed lines indicate per-aspect means. All three aspects show distributions centered at small positive values, indicating that embeddings within an aspect are weakly but consistently similar. (B)  **$k$ -nearest-neighbor Precision@ $k$  and Recall@ $k$ .** Overlap between  $k$ -nearest embedding neighbors and ontological neighbors (parents, children, and siblings within 2 hops via `is_a` and `part_of` edges) for the top 1,000 terms per aspect. Precision@ $k$  (left) decreases with  $k$  as expected while remaining well above the random baseline (dashed lines), indicating that the closest embedding neighbors are the most ontologically relevant. Recall@ $k$  (right) increases with  $k$ , recovering a growing fraction of true ontological neighbors. Shaded bands show  $\pm 1$  s.d. All values significant ( $p < 10^{-3}$ , 1,000-permutation test).



**Figure S7 | Cross-aspect analysis of GO-GPT learned GO term embeddings. (A) Cross-aspect cosine similarity distribution.** Per-aspect-pair density of pairwise cosine similarities between GO terms belonging to different aspects (MF×BP, MF×CC, BP×CC). NPMI co-annotation vs. cosine embedding similarity is shown in **Fig. 2G**. **(B) Cross-aspect retrieval benchmark.** For each GO term in one aspect, all terms in the partner aspect are ranked by embedding cosine similarity and evaluated against ground-truth partners defined by strong co-annotation (NPMI > 0.5). Grouped bars show Precision@*k* for both retrieval directions (darker: A→B; lighter: B→A); square markers show random baseline. Text annotations report symmetric AUROC (mean of both directions) and permutation *p*-value (1,000 embedding shuffles). All three aspect pairs achieve AUROC well above the 0.5 chance level ( $p < 10^{-3}$ ), with MF×CC showing the strongest retrieval (AUROC = 0.80). **(C) Cosine purity.** Fraction of each term's *k* nearest neighbors in the GO embedding space that share the same GO aspect, shown as median per aspect. Shaded bands: 20th–80th percentile. Dashed lines: per-aspect random baseline (aspect proportion of the vocabulary). MF and CC are enriched 3–5× above baseline ( $p < 10^{-3}$ , 1,000-permutation test); BP shows no enrichment, consistent with its broader functional diversity.



**Figure S8 | Cross-aspect GO term pairs visualized on the UMAP embedding landscape.** Kernel density estimation (KDE) contours show the spatial distribution of GO term embeddings in the UMAP projection for each aspect: Molecular Function (MF, blue), Biological Process (BP, red), and Cellular Component (CC, green). An outermost contour band marks the 3% of maximum density threshold; 25 inner contour levels are drawn with increasing opacity (up to  $\alpha = 0.5$ ) to highlight regions of highest term density while minimizing visual clutter. Annotated pairs are cross-aspect GO term pairs selected through a reproducible pipeline: all cross-aspect pairs with co-occurrence  $\geq 5$  and term frequency  $\geq 20$  (~1.2 million total) were ranked by the product of NPMI (co-annotation frequency) and cosine embedding similarity, then filtered by UMAP distance  $< 0.3$  so that only pairs co-localized in the two-dimensional projection are annotated. Connecting arcs link paired terms; leader lines connect labels to pair midpoints. These 18 pairs (Table S2) are illustrative examples of the population-level correlation between NPMI and cosine similarity established across all cross-aspect pairs (Spearman  $\rho = 0.10\text{--}0.17$ ,  $p < 10^{-3}$ , embedding-permutation test; Fig. 2G), which also supports cross-aspect retrieval with AUROC = 0.66–0.80 (Fig. S7B).



**Figure S9 | Training dynamics of the GO-GPT model.** (A) Training and validation cross-entropy loss over 100 epochs. Training loss decreases steadily while validation loss plateaus and begins to rise after approximately 20 epochs, indicating overfitting in later training. (B) Learning rate schedule: linear warmup over the first ~10% of training followed by cosine decay to a minimum of  $\sim 0.1 \times$  the peak learning rate ( $1 \times 10^{-4}$ ). (C) Mean, unweighted validation F1 score (macro-averaged across MF, BP, and CC aspects), plateauing near 0.68. (D) Per-aspect validation F1 scores. CC and MF converge rapidly and plateau near 0.73 and 0.71, respectively, while BP plateaus near 0.54, consistent with the larger and more heterogeneous BP vocabulary. (E) Per-aspect validation precision. MF and CC reach approximately 0.80, while BP stabilizes near 0.58. (F) Per-aspect validation recall. MF and CC reach approximately 0.78, while BP stabilizes near 0.58.

## B. Appendix

### B.1. GO-GPT Attention Analysis at DNA-Binding Sites

Table S1 reports per-protein attention enrichment statistics for all 63 proteins in the DNA-binding attention analysis (Section 4.5.1). For each protein, AUROC quantifies the discrimination between attention scores at annotated binding versus non-binding residues (Mann–Whitney U test), AUPRC Gain normalizes AUPRC by positive class prevalence (Section 4.5.1), and fold-enrichment measures the ratio of mean attention at binding sites to non-binding sites. Hypergeometric  $p$ -values test whether the overlap between the top 20% highest-attention residues and annotated binding residues exceeds chance expectation.

**Table S1 | Attention enrichment at DNA-binding residues across all 63 evaluation proteins, ranked by fold-enrichment.** \*Proteins visualized in Fig. 2E; †proteins visualized in Fig. S2C.

#	UniProt ID	PDB	AUROC	AUPRC Gain	Enrich.	Length	Bind. Res.	MW U $p$	Hyper. $p$ (20%)
1	Q9WXC7	5xxp-A	0.912	0.725	24.25x	294	13	$1.1 \times 10^{-6}$	$9.7 \times 10^{-5}$
2	O68014	4ihs-A	0.905	0.757	17.99x	304	18	$1.1 \times 10^{-7}$	$8.4 \times 10^{-7}$
3	E1C9K5*	2xo7-A	0.918	0.509	10.00x	580	61	$6.0 \times 10^{-27}$	$3.1 \times 10^{-32}$
4	Q24645	1ahd-P	0.871	0.761	9.69x	394	19	$4.8 \times 10^{-5}$	$1.1 \times 10^{-3}$
5	Q5KWC1	1njw-A	0.898	0.479	9.47x	878	58	$2.2 \times 10^{-22}$	$1.7 \times 10^{-25}$
6	Q87HP4	7e1b-B	0.840	0.338	9.09x	220	16	$6.8 \times 10^{-6}$	$1.3 \times 10^{-4}$
7	Q04957	2hhx-A	0.905	0.377	8.80x	877	37	$4.5 \times 10^{-16}$	$3.4 \times 10^{-16}$
8	Q9X0C0	4i6z-B	0.885	0.413	8.57x	200	11	$7.3 \times 10^{-6}$	$1.4 \times 10^{-4}$
9	P69202	3jxb-C	0.823	0.672	8.53x	216	15	$9.7 \times 10^{-4}$	$4.5 \times 10^{-4}$
10	P44558†	5e01-A	0.794	0.338	8.53x	135	10	$1.5 \times 10^{-3}$	$4.5 \times 10^{-3}$
11	P06555	2p7c-B	0.831	0.622	8.29x	128	21	$9.4 \times 10^{-6}$	$1.2 \times 10^{-3}$
12	D9N168	3hpo-A	0.894	0.443	8.13x	579	64	$4.0 \times 10^{-25}$	$2.5 \times 10^{-30}$
13	P96711	4egy-A	0.870	0.690	7.98x	362	17	$3.3 \times 10^{-6}$	$4.4 \times 10^{-5}$
14	Q56215	1ewq-A	0.890	0.227	7.51x	811	27	$6.9 \times 10^{-12}$	$1.8 \times 10^{-13}$
15	Q79VH8	6gy3-A	0.872	0.323	7.34x	222	12	$2.9 \times 10^{-5}$	$3.6 \times 10^{-4}$
16	A0A0H3C9J4	5yiv-A	0.729	0.670	7.32x	173	19	0.9	0.3
17	D5MNX7	3ldy-A	0.906	0.705	7.25x	142	28	$1.5 \times 10^{-11}$	$2.7 \times 10^{-9}$
18	Q816E8†	3jx7-A	0.852	0.322	6.71x	237	17	$6.2 \times 10^{-7}$	$2.1 \times 10^{-7}$
19	A0R6I8	4jl3-B	0.823	0.387	6.41x	189	14	$4.7 \times 10^{-5}$	$1.7 \times 10^{-4}$
20	Q51425	7x5a-A	0.839	0.430	6.35x	201	14	$8.5 \times 10^{-5}$	$1.9 \times 10^{-5}$
21	G2RUZ1	4r22-B	0.859	0.356	5.93x	108	7	0.01	0.1
22	P03001	1tf3-A	0.762	0.610	5.91x	366	29	$1.3 \times 10^{-3}$	0.02
23	G4NEJ8	4bhm-A	0.844	0.424	5.32x	162	11	$9.2 \times 10^{-3}$	$8.4 \times 10^{-3}$
24	P17893	2p5l-D	0.817	0.596	5.31x	149	17	$5.0 \times 10^{-4}$	$1.4 \times 10^{-4}$
25	P46828	1rzz-A	0.814	0.251	5.03x	332	22	$4.3 \times 10^{-7}$	$4.0 \times 10^{-7}$
26	Q5HEG2	4xqj-A	0.835	0.486	4.97x	238	17	$1.9 \times 10^{-5}$	$2.5 \times 10^{-4}$
27	Q3ZD72†	4gjp-A	0.801	0.417	4.91x	960	70	$1.8 \times 10^{-13}$	$5.2 \times 10^{-12}$
28	Q746M7	3dlh-B	0.769	0.266	4.90x	685	70	$6.9 \times 10^{-13}$	$7.1 \times 10^{-11}$
29	A0A158RFF2	4yit-D	0.868	0.375	4.89x	287	37	$9.8 \times 10^{-13}$	$7.4 \times 10^{-8}$
30	P84131*	3go8-A	0.836	0.335	4.70x	274	41	$1.2 \times 10^{-12}$	$1.2 \times 10^{-8}$
31	F7WD42	4z1z-A	0.854	0.323	4.69x	432	33	$1.3 \times 10^{-7}$	$1.1 \times 10^{-4}$
32	Q93PU6	2xro-B	0.767	0.182	4.62x	259	13	$7.8 \times 10^{-4}$	$8.0 \times 10^{-4}$
33	Q58HP2	2kdz-A	0.709	0.407	4.59x	206	21	0.1	0.03
34	Q7DB51	2lev-A	0.772	0.588	4.43x	123	9	$4.0 \times 10^{-3}$	0.07
35	P22262	5lej-B	0.832	0.123	4.39x	237	8	$3.6 \times 10^{-4}$	$8.8 \times 10^{-3}$
36	A0A2I4PGE9	7ce1-K	0.795	0.194	4.27x	350	23	$1.0 \times 10^{-6}$	$1.1 \times 10^{-6}$
37	P42371	1kfv-A	0.868	0.312	4.21x	273	32	$4.7 \times 10^{-12}$	$1.2 \times 10^{-9}$
38	H2J4R1	7nqf-B	0.870	0.278	4.15x	546	16	$3.9 \times 10^{-7}$	$2.1 \times 10^{-6}$
39	A0A2S8J6Y8	8sva-A	0.822	0.186	4.14x	244	10	$1.1 \times 10^{-3}$	0.03
40	Q49QE8	7amv-W	0.810	0.112	4.07x	637	25	$7.3 \times 10^{-8}$	$8.0 \times 10^{-6}$
41	Q9HSF4	6qfd-B	0.771	0.370	4.06x	116	15	$6.5 \times 10^{-4}$	$1.9 \times 10^{-3}$
42	P0A0I7	3bs1-A	0.808	0.337	3.80x	238	13	$6.7 \times 10^{-5}$	$8.8 \times 10^{-5}$
43	Q834K5	8c7u-A	0.827	0.196	3.78x	260	16	$5.9 \times 10^{-6}$	$1.4 \times 10^{-4}$
44	B4FCI7	6is8-A	0.791	0.401	3.49x	280	35	$7.5 \times 10^{-5}$	$7.9 \times 10^{-4}$
45	Q1RLP9	8hem-A	0.745	0.372	3.44x	251	17	0.05	$8.3 \times 10^{-3}$
46	A0A0M4DML1	7cc9-A	0.748	0.308	3.40x	326	23	0.02	0.06
47	O20960	4yis-A	0.825	0.285	3.34x	416	39	$7.6 \times 10^{-8}$	$1.9 \times 10^{-5}$
48	Q9CDU5	5yi2-A	0.778	0.266	3.33x	145	19	$4.8 \times 10^{-5}$	0.02
49	Q97W73	2mna-A	0.787	0.234	3.19x	148	11	$5.5 \times 10^{-3}$	0.04
50	Q8Y8S9	6q2b-A	0.768	0.260	3.02x	150	19	$7.0 \times 10^{-5}$	$6.6 \times 10^{-4}$

**Table S1 | Attention enrichment at DNA-binding residues (continued).**

#	UniProt ID	PDB	AUROC	AUPRC Gain	Enrich.	Length	Bind. Res.	MW U <i>p</i>	Hyper. <i>p</i> (20%)
51	Q8DR18	6jbx-A	0.724	0.199	2.88x	144	15	$2.3 \times 10^{-3}$	0.01
52	Q5Y812*	4lll-A	0.789	0.308	2.85x	139	30	$5.8 \times 10^{-7}$	$1.2 \times 10^{-5}$
53	A0A0H2VJZ8	4kdp-A	0.752	0.147	2.78x	151	10	$3.9 \times 10^{-3}$	0.03
54	Q47673	1omh-A	0.768	0.355	2.75x	966	55	$2.7 \times 10^{-15}$	$4.4 \times 10^{-9}$
55	H9C180	7vjq-A	0.720	0.181	2.67x	116	10	0.01	0.1
56	Q2FHI3	8c7s-A	0.786	0.142	2.50x	257	17	$3.9 \times 10^{-5}$	0.03
57	A0A5C0XNV5	7tra-A	0.701	0.090	2.36x	658	28	$8.6 \times 10^{-4}$	0.01
58	Q14549	2me6-A	0.769	0.386	2.34x	363	14	$3.8 \times 10^{-4}$	0.1
59	Q84ZU4	2qhb-A	0.806	0.310	2.10x	681	11	$1.0 \times 10^{-3}$	$1.8 \times 10^{-3}$
60	Q9VAA9	6kco-F	0.803	0.398	2.08x	475	16	$1.3 \times 10^{-5}$	$1.2 \times 10^{-3}$
61	Q0P897	5x2g-A	0.670	0.010	1.32x	984	43	$6.6 \times 10^{-6}$	$1.2 \times 10^{-5}$
62	B4FJZ1	9l55-A	0.653	0.045	1.15x	422	13	0.07	0.5
63	A0A1Y1IRK2	8s8q-A	0.652	0.145	1.13x	404	22	0.05	0.7

## B.2. Cross-Aspect GO Term Pairs on the UMAP Embedding Landscape

Table S2 lists the 18 cross-aspect GO term pairs annotated on the UMAP embedding landscape (Fig. S8). Pairs were selected through a reproducible pipeline: all cross-aspect pairs with co-occurrence  $\geq 5$  and term frequency  $\geq 20$  ( $>1.2$  million total across three aspect combinations) were ranked by the product of NPMI and cosine embedding similarity (top 200 per combination), then filtered by UMAP Euclidean distance  $< 0.3$  to retain only pairs that co-localize in the two-dimensional projection. These pairs are illustrative examples of the population-level correlation between co-annotation frequency and learned embedding similarity (Spearman  $\rho = 0.10$ – $0.17$ ,  $p < 10^{-3}$ ; Fig. 2G).

**Table S2 | Cross-aspect GO term pairs highlighted on the UMAP embedding landscape (Fig. S8).** Each pair was selected by jointly high NPMI (co-annotation frequency) and cosine embedding similarity, filtered by UMAP proximity ( $< 0.3$ ). Pairs are grouped by aspect combination.

Term 1	Term 2	NPMI	Cosine
<b>Molecular Function × Biological Process</b>			
Dipeptide transmembrane transporter activity	Oligopeptide transmembrane transport	0.88	0.28
Protein prenyltransferase activity	Prenylation	0.91	0.21
Xylan O-acetyltransferase activity	Xylan metabolic process	0.85	0.29
Sesquiterpene synthase activity	Sesquiterpene metabolic process	0.91	0.31
Amino-acid racemase activity	D-amino acid metabolic process	0.76	0.25
Nucleobase transmembrane transporter activity	Nucleobase transport	0.98	0.21
Oxidoreductase activity, acting on sulfur group donors	L-ascorbic acid metabolic process	0.70	0.25
Ribonuclease P activity	tRNA 5'-leader removal	0.96	0.33
<b>Molecular Function × Cellular Component</b>			
Potassium ion transmembrane transporter activity	Voltage-gated potassium channel complex	0.70	0.23
Voltage-gated channel activity	Voltage-gated potassium channel complex	0.69	0.19
RNA Pol III general transcription initiation factor activity	RNA Pol III transcription regulator complex	0.88	0.31
Immunoglobulin receptor binding	Immunoglobulin complex	0.78	0.28
<b>Biological Process × Cellular Component</b>			
Intercellular transport	Connexin complex	0.66	0.26
Protein insertion into mitochondrial outer membrane	SAM complex	0.69	0.24
Modulation by symbiont of host erythrocyte aggregation	Infected host cell surface knob	0.94	0.33
RNA Pol II CTD phosphorylation	Transcription factor TFIIF holo complex	0.73	0.22
Nuclear cell cycle DNA replication initiation	Pre-replicative complex	0.82	0.28
RNA Pol II CTD phosphorylation	Transcription factor TFIIF core complex	0.68	0.23

---

### B.3. Architectural and Training Explorations for GO-GPT

During the development of GO-GPT, we tested a number of architectural and training variations. Due to computational constraints, these were not systematic ablations but rather sparse experiments that informed design decisions qualitatively. We report the observations below, organized by choices that appeared to help and those that did not, noting that most comparisons were not controlled for all confounding variables.

#### B.3.1. Design Choices That Improved Performance

**Decoder Size.** We explored decoder sizes ranging from approximately 70M to 270M parameters by varying hidden dimension, as well as number of heads and layers. The smallest configuration was already on par with prior state-of-the-art  $F_{\max}^w$ , but absolute best results required scaling to the largest decoder. Performance gains were small but monotonic with scale, consistent with findings in the BioReason-Pro backbone selection (Section B.4.1).

**Protein Language Model Encoder.** In early experiments we evaluated six protein language models as the frozen encoder: ESM2-150M, ESM2-650M, and ESM2-3B (Lin et al., 2023); ESM-C 600M (ESM Team, 2024); ESM3 (Hayes et al., 2024); and Profluent-E1 600M (Jain et al., 2025). ESM2-150M produced noticeably weaker downstream performance, but differences among the remaining models were modest, consistent with recent findings that medium-sized PLMs perform comparably to much larger counterparts on transfer learning tasks (Vieira et al., 2025). As with decoder size, the absolute best  $F_{\max}^w$  was obtained with ESM2-3B. Furthermore, we also evaluated ESM3 as the GO-GPT encoder (without structure input). Performance was slightly worse than ESM2-3B. Moreover, ESM3 provided no documented interface for batched extraction of per-residue embeddings at the time of these experiments, requiring sequences to be processed individually. We therefore continued all subsequent GO-GPT experiments with ESM2-650M and ESM2-3B, which support efficient batched inference natively.

**Intermediate Layer Embeddings.** Extracting residue embeddings from an intermediate ESM2 layer rather than the final layer improved results. We observed the best performance when using embeddings from approximately 80–90% through the network (e.g., layer 30 of 36 for ESM2-3B; layer 27 of 33 for ESM2-650M).

**Vocabulary Pruning.** The Gene Ontology contains approximately 43,000 terms, but many appear too infrequently to provide a meaningful training signal. Consistent with prior work (Wang et al., 2025a), we observed improved  $F_{\max}^w$  when constructing the GO vocabulary from terms appearing in at least 20 training proteins, yielding approximately 10,000 tokens. This threshold balances coverage against the noise introduced by rarely annotated terms. We were not able to ablate this threshold systematically due to lack of sufficient compute resources.

**Organism Embedding.** Adding a learnable organism embedding to each GO token representation improved predictive performance and provided an interpretable by-product: cosine similarity between learned organism embeddings recovers known phylogenetic relationships without any explicit phylogenetic supervision (Fig. 2C; Section 4.2).

#### B.3.2. Design Choices That Did Not Help

**Pretraining on SwissProt.** We pretrained GO-GPT on all reviewed SwissProt proteins and their GO annotations before training on the CAFA5 training set. This did not improve generalization to the temporal holdout set, possibly due to the differences in the distribution and specificity between GO annotations in SwissProt (which contains both electronic and experimental annotations) vs. CAFA 5 (which contains experimental annotations only).

**Training on electronic annotations.** Expanding the training set to include both experimental and electronically inferred GO annotations for proteins in the CAFA5 training set did not improve evaluation metrics on the temporal holdout set. Electronic annotations may introduce noise and distribution shifts that offset any benefit from increased coverage.

**Initializing GO embeddings from text.** We initialized GO token embeddings with text embeddings of each term’s name and definition, reasoning that semantic priors might accelerate learning. This did not yield measurable gains for GO-GPT, in contrast to BioReason-Pro where text-derived GO initializations aided alignment (Section B.4.1). The difference may reflect the distinct roles of GO representations in the two models: discrete

---

generation targets in GO-GPT versus soft ontology memory in BioReason-Pro.

**Beam search at inference.** Beam search did not outperform temperature-based or greedy sampling with probability aggregation (Section 4.2).

#### B.4. Architectural and Training Explorations for BioReason-Pro

BioReason-Pro underwent eight major development iterations before converging on the final architecture and training configuration that produced all reported results. These iterations spanned changes in data composition, model architecture, training strategy, reasoning trace format, and input context design. The breadth of exploration required to achieve stable and performant training underscores the difficulty of building multimodal reasoning systems over biological modalities, but we believe the resulting architecture provides a solid foundation for multimodal biological reasoning.

##### B.4.1. BioReason-Pro Architecture

**LLM Backbone.** We evaluated Qwen3-4B and Qwen3-8B (Yang et al., 2025) as language model backbones. The 8B variant yielded moderately stronger performance but imposed substantially higher computational costs for training and inference. Given the marginal gains relative to the resource overhead, we selected Qwen3-4B for all reported experiments. Because BioReason-Pro is fundamentally a reasoning model that generates structured chains of inference, we exclusively used thinking variants of Qwen3.

**Parameter-Efficient Fine-Tuning.** We swept LoRA (Hu et al., 2021) rank from small to large values and observed monotonic improvements in validation performance as rank increased, motivating our final choice of rank 128 with  $\alpha = 256$ . Applying LoRA adapters to all attention and MLP layers outperformed restricting adaptation to attention layers alone, indicating that the feed-forward pathways also require task-specific adjustment for multimodal biological reasoning. We additionally compared LoRA against full fine-tuning of the language model and found minimal performance differences, confirming that low-rank adaptation captures the necessary parameter updates without requiring gradient computation over the full weight matrices.

**Protein Encoder.** We selected ESM3 (Hayes et al., 2024) over ESM2 (Lin et al., 2023) because ESM3 jointly encodes sequence and structure. ESM3 receives structure coordinates as input when experimentally resolved or predicted structures are available. Approximately 10% of training proteins lacked structure annotations, which taught the model to operate robustly in the absence of structural input. Removing structure coordinates at inference reduced performance marginally, suggesting that ESM3 has internalized substantial structural information from sequence alone but still benefits from explicit structural input when available.

We evaluated embeddings extracted from multiple layers of ESM3 and found that extraction at approximately 80% depth through the network (layer 38 of 48) consistently outperformed both the final layer and earlier layers. Unfreezing the ESM3 encoder during fine-tuning provided negligible improvement, so we kept it frozen throughout training to preserve its pretrained representations and reduce memory requirements.

**GO Graph Encoder.** Ablating the GO graph encoder revealed that it serves primarily as a structured memory for the ontology rather than a reasoning aid. Removing the encoder did not visibly degrade the quality of biological reasoning in generated traces, but caused the model to frequently forget, confuse, or hallucinate GO term identifiers. This indicates that the 200 compressed embeddings anchor the language model to the formal ontology vocabulary, preventing drift between free-text reasoning and structured annotation.

We tested compression to 50, 100, and 200 learnable query embeddings via the cross-attention module. Performance improved from 50 to 200 but showed signs of saturation at 200, so we did not explore larger query sets. We also compared a unified encoder that processes all three GO aspects in a single graph against separate per-aspect encoders for Molecular Function, Biological Process, and Cellular Component. The two approaches yielded equivalent performance, and we adopted the unified encoder for architectural simplicity.

Initializing GO term node representations with Qwen3 text embeddings (Zhang et al., 2025) of term names and definitions, rather than random vectors, accelerated convergence during Stage 1 alignment and appeared to ease the mapping between the graph encoder output space and the language model embedding space. We attribute this to the shared semantic grounding between the Qwen3-derived initializations and the LLM input space, which reduces the representational gap that the projection layers must bridge.

---

#### B.4.2. BioReason-Pro SFT

**Two-Stage Training Strategy.** Single-stage training in which the language model, projection layers, and GO encoder were all updated simultaneously led to instability. The LLM lost general language capabilities through catastrophic forgetting and overall convergence was poor. We therefore adopted a two-stage strategy following standard practice in vision-language model training (Liu et al., 2023). Stage 1 freezes the language model and trains only the projection layers and GO graph encoder, aligning the protein and ontology embedding spaces with the LLM input space. This alignment phase required only one epoch and proceeded quickly, indicating that the representational gap between modalities is relatively small when Qwen3-derived GO initializations are used. We tested a higher learning rate of  $3 \times 10^{-4}$  for Stage 1 but observed no benefit over  $1 \times 10^{-4}$ .

Stage 2 unfreezes the language model via LoRA and trains all components jointly for 10 epochs. We observed substantial overfitting during Stage 2, with validation performance peaking at epoch 8 before degrading. The checkpoint at epoch 8 was selected for all downstream experiments and as the initialization for reinforcement learning.

**Input Context Ablations.** We systematically varied the biological context provided in the input prompt. Providing GO-GPT predictions as input improved reasoning quality, and including the full set of predicted GO terms rather than leaf terms alone yielded further gains, likely because ancestor terms supply broader functional context that guides the reasoning process. Training without GO-GPT inputs produced weaker annotations, confirming that initial ontology hypotheses serve as useful scaffolding for the reasoning model.

For InterPro domain annotations (Blum et al., 2024), we found that BioReason-Pro could predict protein domains with approximately 90% F1 from the protein embeddings. However, since InterProScan (Jones et al., 2014) is computationally inexpensive, we included domain annotations in the prompt for all experiments, which consistently improved downstream performance. Protein-protein interaction data was available for a subset of training proteins. We trained with both PPI-present and PPI-absent examples, and the model learned to generate reasonable annotations under either condition, receiving an empty PPI field when interaction data was unavailable.

We also found that including the raw UniProt (Consortium, 2025) function summary alongside the synthetic reasoning summary in the final answer block improved performance.

**Reasoning Trace Design.** The format of synthetic reasoning traces proved critical. Our initial design structured reasoning as a GO directed acyclic graph traversal, progressing from root terms to increasingly specific children within each aspect. This format performed poorly. LLMs are not well suited to tree-traversal reasoning, and the rigid graph navigation constrained the model to unnatural generation patterns that failed to capture the fluid integrative reasoning biologists perform.

Switching to a progressively deepening natural language format dramatically improved results. In this format, traces begin with InterPro domain analysis, proceed through Molecular Function, Biological Process, and Cellular Component, and conclude with mechanistic hypotheses and interaction partner predictions. Each section builds on the conclusions of the preceding one, mirroring the deductive process of expert annotation. Within the reasoning trace, we found that having the model focus on key leaf GO terms produced better results than requiring exhaustive enumeration of all terms in the ontology hierarchy. The full term lists including propagated ancestors appear only in the structured final answer block.

We filtered training traces by dropping proteins with no InterPro domain annotations, as GPT-5 (Singh et al., 2025) struggled to construct coherent reasoning narratives in the absence of domain evidence. GPT-5 was otherwise reliable in generating high-quality traces for the remaining proteins.

**Additional Data Experiments.** We tested whether supplementing the training set with additional SwissProt-reviewed proteins (Bairoch, 2000) would improve performance. Adding these proteins did not yield measurable gains, potentially because their annotations are too shallow to provide additional learning signal.

#### B.4.3. BioReason-Pro RL

**Initialization and Reward-Normalization Pathologies.** RL was initialized from the SFT checkpoint selected by validation performance (epoch 8). We found that later SFT checkpoints trained longer on the same data (e.g., epoch 10) often exhibited reduced output diversity, producing highly similar completions across rollouts.

---

Under the standard prompt-level grouping used in GRPO (Shao et al., 2024), this low diversity frequently collapsed within-group reward variance, yielding weak learning signal with many near-zero advantages and brittle normalization statistics. We mitigated this by moving from prompt-level grouping to batch-level grouping for advantage comparison, and by normalizing advantages using the global batch reward standard deviation rather than per-group variance. In practice, this eliminated the zero-variance regime and restored stable advantage estimates.

**Length Instability and Runaway Generations.** A second failure mode was length-driven instability. Because the prompts vary substantially in context length and evidence content, the policy developed a tendency to increase completion length over training. Without correction, generations expanded into excessively long, low-quality outputs with repetitive or degenerate content, which destabilized learning and degraded downstream annotation quality.

**Stabilization via GSPO and Dr. GRPO.** We addressed these issues with two complementary modifications. First, we adopted the core principle of GSPO (Zheng et al., 2025), applying importance sampling correction at the sequence level to match the granularity of our scalar reward. This substantially improved the stability of importance sampling ratios and reduced sensitivity to token-level noise. Second, we incorporated Dr. GRPO-inspired corrections (Liu et al., 2025) to mitigate optimization bias and length-dependent reward artifacts, which reduced the model’s bias toward longer sequences and improved training robustness. We additionally used the Clip-Higher exploration strategy from DAPO (Yu et al., 2025) to encourage exploration and avoid premature collapse.

**Reasoning Trace Format and Robust Reward Extraction.** The structure of the reasoning trace also mattered during RL, because our reward is computed by extracting predicted GO identifiers from the model output. We found it important to scope reward extraction to the structured final answer block rather than the free-form thinking trace. GO terms mentioned during intermediate reasoning are often speculative, partially formed, or subsequently revised, and using them for reward can unintentionally encourage the model to emit terms prematurely in the trace. In addition, truncation introduced a subtle failure mode. When generations hit the maximum completion length, the model would sometimes fail to emit the final delimiter or last think token, causing the output to end inside the reasoning trace. In these cases, naive extraction could incorrectly read partial or intermediate terms and assign an undeserved reward.

To address this, we (i) selected a maximum completion length that reliably permits the model to finish the reasoning trace and emit the final structured answer under typical prompt lengths, and (ii) designed a conservative regular-expression extractor that targets only the final answer region and ignores GO-like strings elsewhere. The regex pattern anchors to the final answer headers and enforces the expected GO identifier format (e.g., `GO:\d{7}`), with additional guards to discard matches when the final answer block is missing due to truncation. This made reward computation substantially more robust to formatting deviations and partial generations.

**Longer Training Did Not Improve Downstream Performance.** Finally, we observed that extending RL training for additional epochs on the same dataset did not yield measurable gains in downstream evaluation, even when on-policy reward continued to increase. This suggests that the available learning signal saturates quickly under the current data distribution and reward formulation, and that further improvements likely require increased supervision diversity.

## B.5. Statistical Tests

### B.5.1. GO Term Prediction Statistical Tests

Tables S3–S5 and Tables S6–S9 report paired bootstrap comparisons (10,000 iterations), comparing GO-GPT against InterLabelGO+ and ProtBoost across three inference modes, and BioReason (SFT and RL) against GO-GPT across inference budgets (pass@1, pass@10), respectively.  $p$ -values are Holm–Bonferroni corrected for 12 and 90 simultaneous tests (comparisons  $\times$  3 aspects  $\times$  2 metrics), respectively.

**Table S3 | Paired bootstrap comparison: GO-GPT (greedy decoding) vs. baselines.** 95% CI from 10,000 bootstrap iterations;  $p_{\text{corr}}$ : Holm–Bonferroni corrected  $p$ -value.

Comparison	Aspect	$n$	GO-GPT	Baseline	95% CI	$p_{\text{corr}}$
<i><math>F_{\text{max}}</math> (unweighted)</i>						
InterLabelGO+	Molecular Function	2080	0.766	0.727	[+0.029, +0.050]	$<10^{-4}$
	Biological Process	5819	0.587	0.566	[+0.012, +0.029]	$<10^{-4}$
	Cellular Component	3440	0.801	0.790	[+0.003, +0.018]	$4.4 \times 10^{-3}$
ProtBoost	Molecular Function	2080	0.766	0.702	[+0.053, +0.075]	$<10^{-4}$
	Biological Process	5819	0.587	0.533	[+0.046, +0.063]	$<10^{-4}$
	Cellular Component	3440	0.801	0.777	[+0.017, +0.031]	$<10^{-4}$
<i><math>F_{\text{max}}^w</math> (IC-weighted)</i>						
InterLabelGO+	Molecular Function	2080	0.701	0.692	[-0.003, +0.023]	0.073
	Biological Process	5819	0.539	0.525	[+0.005, +0.023]	$3.3 \times 10^{-3}$
	Cellular Component	3440	0.713	0.696	[+0.007, +0.028]	$2.4 \times 10^{-3}$
ProtBoost	Molecular Function	2080	0.701	0.658	[+0.029, +0.058]	$<10^{-4}$
	Biological Process	5819	0.539	0.490	[+0.040, +0.057]	$<10^{-4}$
	Cellular Component	3440	0.713	0.669	[+0.034, +0.054]	$<10^{-4}$

**Table S4 | Paired bootstrap comparison: GO-GPT (pass@10) vs. baselines.** Format as in Table S3.

Comparison	Aspect	$n$	GO-GPT	Baseline	95% CI	$p_{\text{corr}}$
<i><math>F_{\text{max}}</math> (unweighted)</i>						
InterLabelGO+	Molecular Function	2080	0.802	0.727	[+0.065, +0.084]	$<10^{-4}$
	Biological Process	5819	0.644	0.566	[+0.069, +0.085]	$<10^{-4}$
	Cellular Component	3440	0.840	0.790	[+0.043, +0.057]	$<10^{-4}$
ProtBoost	Molecular Function	2080	0.802	0.702	[+0.089, +0.110]	$<10^{-4}$
	Biological Process	5819	0.644	0.533	[+0.103, +0.119]	$<10^{-4}$
	Cellular Component	3440	0.840	0.777	[+0.057, +0.070]	$<10^{-4}$
<i><math>F_{\text{max}}^w</math> (IC-weighted)</i>						
InterLabelGO+	Molecular Function	2080	0.743	0.692	[+0.039, +0.064]	$<10^{-4}$
	Biological Process	5819	0.595	0.525	[+0.062, +0.079]	$<10^{-4}$
	Cellular Component	3440	0.764	0.696	[+0.058, +0.078]	$<10^{-4}$
ProtBoost	Molecular Function	2080	0.743	0.658	[+0.072, +0.099]	$<10^{-4}$
	Biological Process	5819	0.595	0.490	[+0.096, +0.113]	$<10^{-4}$
	Cellular Component	3440	0.764	0.669	[+0.085, +0.105]	$<10^{-4}$

**Table S5 | Paired bootstrap comparison: GO-GPT (probability,  $k=10$ ) vs. baselines.** Format as in Table S3.

Comparison	Aspect	$n$	GO-GPT	Baseline	95% CI	$p_{\text{corr}}$
<i><math>F_{\text{max}}</math> (unweighted)</i>						
InterLabelGO+	Molecular Function	2080	0.766	0.727	[+0.030, +0.049]	$<10^{-4}$
	Biological Process	5819	0.600	0.566	[+0.025, +0.042]	$<10^{-4}$
	Cellular Component	3440	0.816	0.790	[+0.019, +0.033]	$<10^{-4}$
ProtBoost	Molecular Function	2080	0.766	0.702	[+0.054, +0.075]	$<10^{-4}$
	Biological Process	5819	0.600	0.533	[+0.059, +0.075]	$<10^{-4}$
	Cellular Component	3440	0.816	0.777	[+0.032, +0.046]	$<10^{-4}$
<i><math>F_{\text{max}}^w</math> (IC-weighted)</i>						
InterLabelGO+	Molecular Function	2080	0.706	0.692	[+0.002, +0.026]	$1.1 \times 10^{-2}$
	Biological Process	5819	0.547	0.525	[+0.014, +0.032]	$<10^{-4}$
	Cellular Component	3440	0.730	0.696	[+0.024, +0.045]	$<10^{-4}$
ProtBoost	Molecular Function	2080	0.706	0.658	[+0.034, +0.062]	$<10^{-4}$
	Biological Process	5819	0.547	0.490	[+0.048, +0.066]	$<10^{-4}$
	Cellular Component	3440	0.730	0.669	[+0.051, +0.071]	$<10^{-4}$

**Table S6 | Paired bootstrap comparison: SFT (pass@1) vs. all other configurations.** Format as in Table S3.

Comparison	Aspect	<i>n</i>	SFT	Baseline	95% CI	<i>p</i> <sub>corr</sub>
<i>F</i> <sub>max</sub> (unweighted)						
GO-GPT (pass@1)	Molecular Function	1878	0.773	0.792	[-0.029, -0.010]	<10 <sup>-4</sup>
	Biological Process	5623	0.557	0.593	[-0.041, -0.029]	<10 <sup>-4</sup>
	Cellular Component	3293	0.781	0.806	[-0.031, -0.018]	<10 <sup>-4</sup>
GO-GPT (pass@10)	Molecular Function	1878	0.773	0.826	[-0.062, -0.044]	<10 <sup>-4</sup>
	Biological Process	5623	0.557	0.649	[-0.098, -0.086]	<10 <sup>-4</sup>
	Cellular Component	3293	0.781	0.844	[-0.070, -0.057]	<10 <sup>-4</sup>
RL (pass@1)	Molecular Function	1878	0.773	0.783	[-0.017, -0.003]	7.5×10 <sup>-2</sup>
	Biological Process	5623	0.557	0.587	[-0.033, -0.026]	<10 <sup>-4</sup>
	Cellular Component	3293	0.781	0.812	[-0.035, -0.027]	<10 <sup>-4</sup>
RL (pass@10)	Molecular Function	1878	0.773	0.792	[-0.026, -0.012]	<10 <sup>-4</sup>
	Biological Process	5623	0.557	0.592	[-0.038, -0.031]	<10 <sup>-4</sup>
	Cellular Component	3293	0.781	0.825	[-0.048, -0.039]	<10 <sup>-4</sup>
<i>F</i> <sub>max</sub> <sup>w</sup> (IC-weighted)						
GO-GPT (pass@1)	Molecular Function	1878	0.717	0.729	[-0.024, +0.001]	0.40
	Biological Process	5623	0.516	0.545	[-0.035, -0.023]	<10 <sup>-4</sup>
	Cellular Component	3293	0.689	0.718	[-0.038, -0.020]	<10 <sup>-4</sup>
GO-GPT (pass@10)	Molecular Function	1878	0.717	0.769	[-0.063, -0.040]	<10 <sup>-4</sup>
	Biological Process	5623	0.516	0.601	[-0.092, -0.079]	<10 <sup>-4</sup>
	Cellular Component	3293	0.689	0.770	[-0.090, -0.072]	<10 <sup>-4</sup>
RL (pass@1)	Molecular Function	1878	0.717	0.718	[-0.010, +0.009]	1.00
	Biological Process	5623	0.516	0.544	[-0.031, -0.024]	<10 <sup>-4</sup>
	Cellular Component	3293	0.689	0.727	[-0.044, -0.032]	<10 <sup>-4</sup>
RL (pass@10)	Molecular Function	1878	0.717	0.730	[-0.022, -0.004]	3.9×10 <sup>-2</sup>
	Biological Process	5623	0.516	0.549	[-0.036, -0.030]	<10 <sup>-4</sup>
	Cellular Component	3293	0.689	0.743	[-0.060, -0.048]	<10 <sup>-4</sup>

**Table S7 | Paired bootstrap comparison: SFT (pass@10) vs. other configurations.** Format as in Table S3.

Comparison	Aspect	<i>n</i>	SFT	Baseline	95% CI	<i>p</i> <sub>corr</sub>
<i>F</i> <sub>max</sub> (unweighted)						
GO-GPT (pass@1)	Molecular Function	1878	0.801	0.792	[-0.000, +0.019]	0.39
	Biological Process	5623	0.582	0.593	[-0.016, -0.004]	1.8×10 <sup>-2</sup>
	Cellular Component	3293	0.800	0.806	[-0.012, +0.001]	0.54
GO-GPT (pass@10)	Molecular Function	1878	0.801	0.826	[-0.034, -0.015]	<10 <sup>-4</sup>
	Biological Process	5623	0.582	0.649	[-0.073, -0.061]	<10 <sup>-4</sup>
	Cellular Component	3293	0.800	0.844	[-0.050, -0.038]	<10 <sup>-4</sup>
RL (pass@1)	Molecular Function	1878	0.801	0.783	[+0.011, +0.027]	<10 <sup>-4</sup>
	Biological Process	5623	0.582	0.587	[-0.008, -0.001]	0.12
	Cellular Component	3293	0.800	0.812	[-0.016, -0.007]	<10 <sup>-4</sup>
RL (pass@10)	Molecular Function	1878	0.801	0.792	[+0.002, +0.017]	0.14
	Biological Process	5623	0.582	0.592	[-0.013, -0.006]	<10 <sup>-4</sup>
	Cellular Component	3293	0.800	0.825	[-0.029, -0.020]	<10 <sup>-4</sup>
<i>F</i> <sub>max</sub> <sup>w</sup> (IC-weighted)						
GO-GPT (pass@1)	Molecular Function	1878	0.757	0.729	[+0.016, +0.041]	<10 <sup>-4</sup>
	Biological Process	5623	0.544	0.545	[-0.007, +0.005]	1.00
	Cellular Component	3293	0.721	0.718	[-0.006, +0.012]	1.00
GO-GPT (pass@10)	Molecular Function	1878	0.757	0.769	[-0.024, -0.000]	0.37
	Biological Process	5623	0.544	0.601	[-0.064, -0.051]	<10 <sup>-4</sup>
	Cellular Component	3293	0.721	0.770	[-0.058, -0.040]	<10 <sup>-4</sup>
RL (pass@1)	Molecular Function	1878	0.757	0.718	[+0.029, +0.050]	<10 <sup>-4</sup>
	Biological Process	5623	0.544	0.544	[-0.003, +0.004]	1.00
	Cellular Component	3293	0.721	0.727	[-0.012, +0.001]	0.50
RL (pass@10)	Molecular Function	1878	0.757	0.730	[+0.017, +0.036]	<10 <sup>-4</sup>
	Biological Process	5623	0.544	0.549	[-0.008, -0.002]	3.8×10 <sup>-2</sup>
	Cellular Component	3293	0.721	0.743	[-0.029, -0.015]	<10 <sup>-4</sup>

**Table S8 | Paired bootstrap comparison: RL (pass@1) vs. other configurations.** Format as in Table S3.

Comparison	Aspect	$n$	RL	Baseline	95% CI	$p_{\text{corr}}$
<i>F<sub>max</sub> (unweighted)</i>						
GO-GPT (pass@1)	Molecular Function	1878	0.783	0.792	[-0.018, -0.001]	0.27
	Biological Process	5623	0.587	0.593	[-0.012, -0.001]	0.30
	Cellular Component	3293	0.812	0.806	[+0.000, +0.012]	0.30
GO-GPT (pass@10)	Molecular Function	1878	0.783	0.826	[-0.052, -0.034]	<10 <sup>-4</sup>
	Biological Process	5623	0.587	0.649	[-0.068, -0.057]	<10 <sup>-4</sup>
	Cellular Component	3293	0.812	0.844	[-0.038, -0.027]	<10 <sup>-4</sup>
SFT (pass@1)	Molecular Function	1878	0.783	0.773	[+0.003, +0.017]	7.5×10 <sup>-2</sup>
	Biological Process	5623	0.587	0.557	[+0.026, +0.033]	<10 <sup>-4</sup>
	Cellular Component	3293	0.812	0.781	[+0.027, +0.035]	<10 <sup>-4</sup>
SFT (pass@10)	Molecular Function	1878	0.783	0.801	[-0.027, -0.011]	<10 <sup>-4</sup>
	Biological Process	5623	0.587	0.582	[+0.001, +0.008]	0.12
	Cellular Component	3293	0.812	0.800	[+0.007, +0.016]	<10 <sup>-4</sup>
<i>F<sub>max</sub><sup>w</sup> (IC-weighted)</i>						
GO-GPT (pass@1)	Molecular Function	1878	0.718	0.729	[-0.022, +0.000]	0.37
	Biological Process	5623	0.544	0.545	[-0.007, +0.005]	1.00
	Cellular Component	3293	0.727	0.718	[+0.000, +0.016]	0.30
GO-GPT (pass@10)	Molecular Function	1878	0.718	0.769	[-0.062, -0.040]	<10 <sup>-4</sup>
	Biological Process	5623	0.544	0.601	[-0.064, -0.051]	<10 <sup>-4</sup>
	Cellular Component	3293	0.727	0.770	[-0.051, -0.035]	<10 <sup>-4</sup>
SFT (pass@1)	Molecular Function	1878	0.718	0.717	[-0.009, +0.010]	1.00
	Biological Process	5623	0.544	0.516	[+0.024, +0.031]	<10 <sup>-4</sup>
	Cellular Component	3293	0.727	0.689	[+0.032, +0.044]	<10 <sup>-4</sup>
SFT (pass@10)	Molecular Function	1878	0.718	0.757	[-0.050, -0.029]	<10 <sup>-4</sup>
	Biological Process	5623	0.544	0.544	[-0.004, +0.003]	1.00
	Cellular Component	3293	0.727	0.721	[-0.001, +0.012]	0.50

**Table S9 | Paired bootstrap comparison: RL (pass@10) vs. other configurations.** Format as in Table S3.

Comparison	Aspect	$n$	RL	Baseline	95% CI	$p_{\text{corr}}$
<i>F<sub>max</sub> (unweighted)</i>						
GO-GPT (pass@1)	Molecular Function	1878	0.792	0.792	[-0.008, +0.008]	1.00
	Biological Process	5623	0.592	0.593	[-0.007, +0.004]	1.00
	Cellular Component	3293	0.825	0.806	[+0.013, +0.025]	<10 <sup>-4</sup>
GO-GPT (pass@10)	Molecular Function	1878	0.792	0.826	[-0.042, -0.025]	<10 <sup>-4</sup>
	Biological Process	5623	0.592	0.649	[-0.063, -0.052]	<10 <sup>-4</sup>
	Cellular Component	3293	0.825	0.844	[-0.025, -0.014]	<10 <sup>-4</sup>
SFT (pass@1)	Molecular Function	1878	0.792	0.773	[+0.012, +0.026]	<10 <sup>-4</sup>
	Biological Process	5623	0.592	0.557	[+0.031, +0.038]	<10 <sup>-4</sup>
	Cellular Component	3293	0.825	0.781	[+0.039, +0.048]	<10 <sup>-4</sup>
SFT (pass@10)	Molecular Function	1878	0.792	0.801	[-0.017, -0.002]	0.14
	Biological Process	5623	0.592	0.582	[+0.006, +0.013]	<10 <sup>-4</sup>
	Cellular Component	3293	0.825	0.800	[+0.020, +0.029]	<10 <sup>-4</sup>
<i>F<sub>max</sub><sup>w</sup> (IC-weighted)</i>						
GO-GPT (pass@1)	Molecular Function	1878	0.730	0.729	[-0.010, +0.013]	1.00
	Biological Process	5623	0.549	0.545	[-0.002, +0.010]	0.93
	Cellular Component	3293	0.743	0.718	[+0.017, +0.033]	<10 <sup>-4</sup>
GO-GPT (pass@10)	Molecular Function	1878	0.730	0.769	[-0.050, -0.027]	<10 <sup>-4</sup>
	Biological Process	5623	0.549	0.601	[-0.059, -0.046]	<10 <sup>-4</sup>
	Cellular Component	3293	0.743	0.770	[-0.035, -0.019]	<10 <sup>-4</sup>
SFT (pass@1)	Molecular Function	1878	0.730	0.717	[+0.004, +0.022]	3.9×10 <sup>-2</sup>
	Biological Process	5623	0.549	0.516	[+0.030, +0.036]	<10 <sup>-4</sup>
	Cellular Component	3293	0.743	0.689	[+0.048, +0.060]	<10 <sup>-4</sup>
SFT (pass@10)	Molecular Function	1878	0.730	0.757	[-0.036, -0.017]	<10 <sup>-4</sup>
	Biological Process	5623	0.549	0.544	[+0.002, +0.008]	3.8×10 <sup>-2</sup>
	Cellular Component	3293	0.743	0.721	[+0.015, +0.029]	<10 <sup>-4</sup>

### B.5.2. LLM Judge Statistical Tests

To test whether score differences between models are statistically significant at the per-protein level, we performed two-sided Wilcoxon signed-rank tests on paired LLM judge scores across all evaluation axes and mean. Each protein receives a score from every model, enabling paired comparisons that control for protein-level difficulty. We report the mean paired score difference ( $\Delta$ ) as an effect size alongside the  $p$ -value. BLAST with penalty assigns a score of 0 to proteins with no hit ( $n=8,159$ ); BLAST no-penalty restricts to proteins where BLAST returned a prediction ( $n=3,380$ ). Both BioReason-Pro variants significantly outperform Prot2Text-v2 and BLAST across all metrics, with RL significantly outperforming SFT on all axis (Table S10).

**Table S10** | Pairwise Wilcoxon signed-rank tests on per-protein LLM judge scores. Each cell shows the mean paired score difference ( $\Delta$ ) and  $p$ -value (\*\* $p < 0.001$ , \* $p < 0.05$ ). Positive  $\Delta$  indicates the reference model scores higher. All tests are two-sided and paired by protein ( $n=8,159$  except BLAST no-penalty, restricted to  $n=3,380$  proteins with a BLAST hit).

BioReason-Pro RL vs.				
Metric	BioReason-Pro SFT	Prot2Text-v2	BLAST (w/ pen.)	BLAST (no pen.)
Overall	+0.24 (** $3.6 \times 10^{-22}$ )	+3.01 (** $< 10^{-300}$ )	+5.08 (** $< 10^{-300}$ )	+0.90 (** $8.9 \times 10^{-110}$ )
Molecular Function	+0.26 (** $2.1 \times 10^{-15}$ )	+3.72 (** $< 10^{-300}$ )	+5.90 (** $< 10^{-300}$ )	+1.53 (** $9.9 \times 10^{-179}$ )
Cellular Component	+0.30 (** $6.4 \times 10^{-26}$ )	+2.24 (** $7.7 \times 10^{-268}$ )	+4.96 (** $< 10^{-300}$ )	+0.37 (** $6.7 \times 10^{-6}$ )
Biological Process	+0.40 (** $2.1 \times 10^{-35}$ )	+3.34 (** $< 10^{-300}$ )	+4.94 (** $< 10^{-300}$ )	+1.02 (** $4.3 \times 10^{-68}$ )
Specificity	-0.24 (** $1.5 \times 10^{-140}$ )	+3.28 (** $< 10^{-300}$ )	+5.46 (** $< 10^{-300}$ )	+1.09 (** $3.3 \times 10^{-169}$ )
Reliability	+0.50 (** $1.5 \times 10^{-164}$ )	+2.25 (** $< 10^{-300}$ )	+4.64 (** $< 10^{-300}$ )	+0.55 (** $2.0 \times 10^{-26}$ )

BioReason-Pro SFT vs.				
Metric	BioReason-Pro RL	Prot2Text-v2	BLAST (w/ pen.)	BLAST (no pen.)
Overall	-0.24 (** $3.6 \times 10^{-22}$ )	+2.77 (** $< 10^{-300}$ )	+4.84 (** $< 10^{-300}$ )	+0.73 (** $1.8 \times 10^{-79}$ )
Molecular Function	-0.26 (** $2.1 \times 10^{-15}$ )	+3.65 (** $< 10^{-300}$ )	+5.70 (** $< 10^{-300}$ )	+1.50 (** $4.8 \times 10^{-173}$ )
Cellular Component	-0.30 (** $6.4 \times 10^{-26}$ )	+2.00 (** $3.7 \times 10^{-215}$ )	+4.67 (** $< 10^{-300}$ )	+0.20 (* $3.9 \times 10^{-2}$ )
Biological Process	-0.40 (** $2.1 \times 10^{-35}$ )	+2.94 (** $< 10^{-300}$ )	+4.55 (** $< 10^{-300}$ )	+0.76 (** $2.0 \times 10^{-35}$ )
Specificity	+0.24 (** $1.5 \times 10^{-140}$ )	+3.53 (** $< 10^{-300}$ )	+5.70 (** $< 10^{-300}$ )	+1.37 (** $1.0 \times 10^{-237}$ )
Reliability	-0.50 (** $1.5 \times 10^{-164}$ )	+1.75 (** $< 10^{-300}$ )	+4.14 (** $< 10^{-300}$ )	+0.13 (* $2.5 \times 10^{-2}$ )

### B.5.3. Performance by Sequence Similarity Statistical Tests

To assess whether model performance depends on sequence similarity to training data, we regressed per-protein score differences (model – BLAST) on BLAST sequence identity using OLS with HC3 robust standard errors. A significant negative slope ( $\beta_1 < 0$ ) indicates that the test model’s advantage over BLAST increases at low similarity, i.e., weaker similarity dependence. We also report Spearman  $\rho$  between each model’s scores and sequence identity as an interpretable effect size, where lower  $\rho$  indicates weaker similarity dependence.

Table S11 reports results for LLM judge scores (Figure 3D). All generative models show significantly flatter similarity dependence than BLAST ( $\rho = 0.43$ ), with BioReason-Pro SFT and RL exhibiting  $\rho \approx 0.21$ – $0.27$ , roughly half of BLAST’s dependence. The regression slopes are nearly identical across all three models ( $\beta_1 \approx -0.019$  to  $-0.023$ ), confirming that weaker similarity dependence relative to BLAST is shared even by models that underperform BLAST overall.

Table S12 reports results for per-protein F1 scores (Figure 3F). BLAST shows the strongest similarity dependence ( $\rho = 0.63$ ). All generative models have significantly weaker dependence (all  $p < 10^{-130}$ ), with BioReason-Pro SFT variants showing the flattest profiles ( $\rho \approx 0.41$ ,  $\beta_1 \approx -0.0054$ ) and RL performing comparably ( $\rho \approx 0.46$ ,  $\beta_1 \approx -0.0041$ ).

**Table S11** | Similarity dependence of LLM judge scores (Figure 3D). Each row tests whether the model’s performance depends less on sequence similarity than BLAST (no penalty,  $n = 3,380$  paired proteins).  $\rho$ : Spearman rank correlation between score and BLAST sequence identity.  $\beta_1$ : OLS slope of the per-protein score difference (model – BLAST) on sequence identity, with HC3 robust standard errors. Negative  $\beta_1$  indicates weaker similarity dependence than BLAST. \*\*\* $p < 0.001$ .

Model	Spearman $\rho$	BLAST $\rho$	$\beta_1$ (slope)	$t$	$p$ -value
Prot2Text-v2	+0.16	+0.43	$-0.0228 \pm 0.0020$	-11.22	*** $3.4 \times 10^{-29}$
BioReason-Pro SFT	+0.21	+0.43	$-0.0198 \pm 0.0018$	-11.26	*** $2.2 \times 10^{-29}$
BioReason-Pro RL	+0.27	+0.43	$-0.0189 \pm 0.0017$	-11.43	*** $3.1 \times 10^{-30}$

**Table S12** | Similarity dependence of per-protein F1 scores (Figure 3F). Each row tests whether the model’s performance depends less on sequence similarity than BLAST ( $n = 8,149$  paired proteins).  $\rho$ : Spearman rank correlation between score and BLAST sequence identity.  $\beta_1$ : OLS slope of the per-protein score difference (model – BLAST) on sequence identity, with HC3 robust standard errors. Negative  $\beta_1$  indicates weaker similarity dependence than BLAST. \*\*\* $p < 0.001$ .

Model	Spearman $\rho$	BLAST $\rho$	$\beta_1$ (slope)	$t$	$p$ -value
GO-GPT	+0.46	+0.63	$-0.0040 \pm 0.0002$	-24.38	*** $2.6 \times 10^{-131}$
GO-GPT (max@10)	+0.45	+0.63	$-0.0047 \pm 0.0002$	-29.40	*** $4.8 \times 10^{-190}$
BioReason-Pro SFT	+0.41	+0.63	$-0.0054 \pm 0.0002$	-33.65	*** $2.8 \times 10^{-248}$
BioReason-Pro SFT (max@10)	+0.41	+0.63	$-0.0054 \pm 0.0002$	-33.81	*** $1.6 \times 10^{-250}$
BioReason-Pro RL	+0.46	+0.63	$-0.0041 \pm 0.0002$	-24.74	*** $3.9 \times 10^{-135}$
BioReason-Pro RL (max@10)	+0.46	+0.63	$-0.0041 \pm 0.0002$	-24.75	*** $3.4 \times 10^{-135}$

#### B.5.4. Human Expert Evaluation Statistical Tests

To test whether human expert scores depend on sequence similarity, we computed Spearman  $\rho$  between each model’s per-protein scores and BLAST sequence identity across all evaluation axes ( $n = 104$ – $131$  proteins per axis). We also tested whether the RL-vs-SFT score gap changes with similarity by regressing per-protein score differences (RL – SFT) on sequence identity (OLS, HC3 robust standard errors). Finally, we tested whether each model’s win rate against UniProt ground truth depends on similarity via logistic regression.

Table S13 reports per-model Spearman  $\rho$  values. Neither BioReason-Pro SFT nor RL shows significant similarity dependence on any evaluation axis (all  $p > 0.05$ ), with the single exception of RL on Protein Interactions ( $\rho = -0.20$ ,  $p = 0.026$ ). Overall scores show near-zero correlation for both models ( $|\rho| < 0.06$ ), confirming that human-assessed quality is stable across the full similarity range.

Table S14 reports paired slope tests for the RL-vs-SFT score gap. The gap does not change significantly with similarity on any axis (all  $p > 0.05$ ), showing comparable models perform across the similarity spectrum.

Table S15 reports win rate against UniProt ground truth. Win rates do not depend on similarity for either model (logistic regression,  $p = 0.71$  for SFT,  $p = 0.19$  for RL), and the paired win rate difference between RL and SFT is also non-significant ( $p = 0.28$ ). These results confirm that both models generalize uniformly, with expert-assessed quality remaining stable regardless of how distant a test protein is from training sequences.

**Table S13** | Similarity dependence of human expert scores. Spearman  $\rho$  between per-protein score and BLAST sequence identity for each evaluation axis. \* $p < 0.05$ , n.s. = not significant.

Metric	BioReason-Pro SFT			BioReason-Pro RL		
	$n$	$\rho$	$p$ -value	$n$	$\rho$	$p$ -value
Molecular Function	104	+0.08	n.s. (0.40)	104	+0.08	n.s. (0.40)
Biological Process	113	+0.06	n.s. (0.54)	113	+0.00	n.s. (0.99)
Cellular Component	122	-0.02	n.s. (0.81)	122	+0.02	n.s. (0.81)
Reasoning	131	+0.03	n.s. (0.71)	130	+0.00	n.s. (0.99)
Hallucination	123	-0.05	n.s. (0.55)	123	-0.00	n.s. (0.97)
Mechanistic Depth	129	+0.09	n.s. (0.33)	129	-0.02	n.s. (0.80)
Protein Interactions	125	-0.06	n.s. (0.50)	125	-0.20	*(0.026)
Overall	131	-0.00	n.s. (0.96)	131	-0.05	n.s. (0.54)

**Table S14** | Paired similarity-dependence of human expert score gap. OLS regression of per-protein score difference (RL – SFT) on BLAST sequence identity with HC3 robust standard errors. Non-significant slopes indicate that the RL-vs-SFT gap is stable across the similarity range. n.s. = not significant ( $p > 0.05$ ).

Metric	$n$	$\beta_1$ (slope)	$t$	$p$ -value
Molecular Function	104	+0.0036 $\pm$ 0.0108	+0.33	n.s. (0.74)
Biological Process	113	-0.0018 $\pm$ 0.0073	-0.25	n.s. (0.81)
Cellular Component	122	-0.0016 $\pm$ 0.0048	-0.32	n.s. (0.75)
Reasoning	130	-0.0035 $\pm$ 0.0046	-0.77	n.s. (0.44)
Hallucination	123	+0.0072 $\pm$ 0.0057	+1.26	n.s. (0.21)
Mechanistic Depth	129	-0.0051 $\pm$ 0.0049	-1.05	n.s. (0.29)
Protein Interactions	125	-0.0087 $\pm$ 0.0045	-1.93	n.s. (0.054)
Overall	131	-0.0016 $\pm$ 0.0033	-0.50	n.s. (0.62)

**Table S15** | Similarity dependence of win rate against UniProt ground truth. Logistic regression of the binary outcome (“Matches” or better = 1, “Falls Short” or worse = 0) on BLAST sequence identity. The paired row tests whether the RL-vs-SFT win rate gap changes with similarity (OLS, HC3 robust standard errors). All tests restricted to  $n = 131$  proteins with available BLAST identity. n.s. = not significant ( $p > 0.05$ ).

Comparison	$n$	$\beta_1$ (slope)	$z/t$	$p$ -value
SFT vs UniProt	131	+0.0037 $\pm$ 0.0100	$z = +0.38$	n.s. (0.71)
RL vs UniProt	131	+0.0114 $\pm$ 0.0087	$z = +1.31$	n.s. (0.19)
$\Delta$ (RL – SFT) win rate	131	+0.0016 $\pm$ 0.0015	$t = +1.09$	n.s. (0.28)

---

## B.6. GO-GPT Hyperparameters

Table S16 reports the full GO-GPT configuration.

**Table S16 | GO-GPT hyperparameters.**

Parameter	Value
<i>Architecture</i>	
Protein encoder	ESM2 3B (frozen)
Protein encoder (interpretability)	ESM2 650M (frozen)
ESM2 extraction layer	30
Decoder layers	12
Decoder attention heads	12
Decoder hidden dimension ( $d_h$ )	1,080
MLP expansion factor	4×
Output gate	Sigmoid
Dropout	0.1
Vocabulary size	9,984 (9,977 GO + 7 delimiters)
GO term min frequency	20 proteins per aspect
Number of organism embeddings	201 (top 200 + unknown)
Weight tying	lm_head tied to token embeddings
Max protein sequence length	1,024 residues
Block size (max sequence length)	2,048
Weight initialization	$\mathcal{N}(0, 0.02)$ ; output proj. scaled by $1/\sqrt{2n}$
Precision	bf16-mixed
<i>Optimization</i>	
Optimizer	AdamW ( $\beta_1=0.9, \beta_2=0.95$ )
Peak learning rate	$1 \times 10^{-4}$
LR schedule	Linear warmup → cosine decay
Warmup ratio	0.10
Minimum LR ratio	0.1
Weight decay	0.01
Gradient clipping	1.0
Epochs	up to 100 (early stopping by validation)
<i>Batching</i>	
Per-GPU batch size	10
Gradient accumulation steps	4
GPUs	4× NVIDIA H100
Effective batch size	160
<i>Inference (Sampling)</i>	
Temperature	0.7
Top-k	20
Number of samples	10
Max new tokens	100 per aspect
<i>GO Term Ordering</i>	
Method	Depth (longest root-to-term path)
Relationships used	is_a + part_of
<i>Other</i>	
Framework	PyTorch + PyTorch Lightning + Flash Attention
Seed	42

## B.7. BioReason-Pro Hyperparameters

Table S17 reports the SFT configuration for BioReason-Pro. The model checkpoint at epoch 8 of Stage 2 was selected based on validation performance. Table S18 reports the RL configuration for BioReason-Pro. Training uses DR-GRPO with importance sampling correction starting from the SFT epoch 8 checkpoint. Training ran for 1,200 steps over approximately 32 hours, consuming ~619M tokens.

**Table S17** | SFT hyperparameters for BioReason-Pro. Stage 1 trains projection layers and the GO graph encoder only; Stage 2 adds LoRA on the language model and trains all components jointly.

Parameter	Stage 1	Stage 2
<i>Model</i>		
Base LLM	Qwen3-4B-Thinking	Qwen3-4B-Thinking
Protein encoder	ESM3-1B (frozen)	ESM3-1B (frozen)
Protein embedding layer	37 (38th layer)	37 (38th layer)
Precision	bf16-mixed	bf16-mixed
<i>LoRA configuration</i>		
Rank ( $r$ )	—	128
Alpha ( $\alpha$ )	—	256
Dropout	—	0
Target modules	—	All attention + MLP
Initialization	—	Gaussian
Bias	—	None
<i>Trainable components</i>		
Language model (LoRA)	No	Yes
Protein projection	Yes	Yes
GO projection	Yes	Yes
GO graph encoder	Yes	Yes
Protein encoder	No	No
<i>Optimization</i>		
Optimizer	AdamW	AdamW
Weight decay	0.01	0.01
Peak learning rate	$1 \times 10^{-4}$	$1 \times 10^{-4}$
LR schedule	Linear warmup $\rightarrow$ cosine decay	Linear warmup $\rightarrow$ cosine decay
Warmup ratio	0.10	0.05
Minimum LR	$1 \times 10^{-5}$	$1 \times 10^{-5}$
Gradient clipping	1.0	1.0
<i>Batching</i>		
Per-GPU batch size	4	4
Gradient accumulation steps	1	1
GPUs (2 nodes $\times$ 4)	8	8
Effective batch size	32	32
<i>Data and training</i>		
Epochs	1	10 (best: epoch 8)
Max protein length	2,000 residues	2,000 residues
Max text length	10,000 tokens	10,000 tokens
Validation split	10%	10%
Validation check interval	0.2 epoch	0.2 epoch
Seed	23	23
<i>GO graph encoder</i>		
Hidden dimension	512	512
GAT layers	3	3
Attention heads	8	8
Reduced embeddings	200	200
Output embedding dimension	2,560	2,560
<i>Other</i>		
Framework	Transformers + Unsloth + PyTorch	Transformers + Unsloth + PyTorch
Distributed strategy	DDP	DDP
Gradient checkpointing	Unsloth	Unsloth

**Table S18** | RL hyperparameters for BioReason-Pro. Training uses DR-GRPO starting from the SFT epoch 8 checkpoint with vLLM-based rollout generation in colocate mode.

Parameter	Value
<i>Algorithm</i>	
Loss type	DR-GRPO
Group size ( $G$ )	24
Steps per generation	2
Num. iterations (inner optimization)	1
KL penalty ( $\beta$ )	$1 \times 10^{-4}$
Clipping $\epsilon_{\text{low}}$	$7 \times 10^{-4}$
Clipping $\epsilon_{\text{high}}$	$9 \times 10^{-4}$
Reward scaling	Batch
Importance sampling correction	Yes (sequence-level)
Importance sampling cap	2
<i>LoRA configuration</i>	
Rank ( $r$ )	16
Alpha ( $\alpha$ )	32
Dropout	0.05
Disable model dropout	Yes
<i>Optimization</i>	
Optimizer	AdamW ( $\beta_1=0.9, \beta_2=0.999, \epsilon=10^{-8}$ )
Peak learning rate	$3 \times 10^{-5}$
LR schedule	Cosine decay
Warmup ratio	0.03
Weight decay	0
Gradient clipping	1.0
Gradient checkpointing	Yes
<i>Batching</i>	
Per-device batch size	6
Gradient accumulation steps	4
GPUs (2 nodes $\times$ 4)	8
Effective batch size ( $B$ )	192
Unique proteins per step ( $B/G$ )	8
<i>Sampling (rollouts)</i>	
Temperature	1.0
Top- $k$	20
Top- $p$	0.95
Min- $p$	0
Repetition penalty	1.0 (none)
Max completion length ( $L_{\text{max}}$ )	10,000 tokens
Max prompt length	512 tokens
<i>Training</i>	
Init checkpoint	SFT epoch 8
Max RL steps	1,200
Total tokens consumed	~619M
Precision	bf16
Seed	42
<i>Other</i>	
Framework	DeepSpeed + vLLM (colocate)
vLLM tensor parallel size	1
Hardware	8 $\times$ NVIDIA H100 80GB (2 nodes)
Wall time	~32 hours

---

## C. Supplementary Text

### C.1. Reasoning Data Generation Prompt

#### Prompt for GPT-5

You are an expert computational biologist excellent at clear scientific communication. Your mission is to analyze detailed protein metadata and generate an exhaustive functional annotation report that simulates the process of genuine scientific discovery.

Your response must consist of two distinct, detailed parts:

- 1) A “Reasoning Trace”: A fluid, concise narrative that demonstrates the step-by-step deductive process a scientist would follow. You must meticulously analyze each piece of evidence, starting from sequence-level features (InterPro domains), and logically build a comprehensive case for the protein’s molecular function, biological process, and cellular location before hypothesizing its mechanism.
- 2) A “Final Summary”: A structured summary of the protein’s function, following the exact format specified below.

---

#### Core Principles of Your Analysis

- 1) **Simulate Abductive Reasoning:** Write from the perspective of inferring function for the first time. Build from foundational sequence-derived facts to higher-level conclusions. Do not use language that implies you are merely confirming pre-known answers (avoid phrases like “provided data”, “the interaction list”, “the data confirms”, etc.). Maintain a discovery-oriented tone. To achieve this, you must treat the provided GO terms and the Subcellular Location not as facts to be validated, but as formal labels for concepts you deduce independently. Furthermore, your reasoning must use causal language (e.g., “this architecture causes X activity”) rather than comparative language.
- 2) **Follow a Strict and Detailed Chain of Inference:** You must use every piece of provided input data within the reasoning trace. Begin with InterPro domains (include IDs, names, and residue spans; reason about their order and architecture). From this analysis, infer Molecular Function (GO MF), then connect to Biological Process (GO BP), and finally to Cellular Component (GO CC) and subcellular location. Then narrate a concise hypothesis to deduce the interaction partners and function of the protein from your reasoning. Your first sentence should always start by exploring the InterPro domains and their architecture.
- 3) **Create a Self-Contained Logical Proof:** The ultimate goal is to generate a reasoning trace that can stand alone as a complete, logical argument. A reader should be able to follow your deductive path from sequence features to a mechanistic model without ever seeing the original input data. This is why it is critical to build the case step-by-step, to predict and hypothesize outcomes rather than referencing or justifying given facts, and to cohesively weave every piece of evidence into the final narrative. Refer to the examples below to learn how the Interaction Partners are still hypothesized even when given.

---

#### Output Format Requirements

- **Enclosing Tags:** Produce exactly two top-level sections enclosed by these tags: `<|REASONING|>` ... `</REASONING|>` `<|FINAL_ANSWER|>` ... `</FINAL_ANSWER|>`
- **Final Summary Structure:** Inside `<|FINAL_ANSWER|>`, output a bullet list with these exact keys (plain text keys, not bolded):
  - Functional Summary
  - InterPro
  - Molecular Function
  - Biological Process

- 
- Cellular Component
  - Hypothesized Interaction Partners
  - **Functional Summary:** A detailed paragraph that summarizes the protein’s function. It should mention all information present in the input Protein Function and describe what the protein does, how it contributes to broader processes, where it is likely located, and the essence of its proposed mechanism. Do not include InterPro IDs, GO terms, residue spans, or other technical labels here; those belong only in the bullet list sections. Do not refer to protein names or other identifiers in the functional summary.
  - **Completeness:**
    - Mention every provided InterPro item (with ID, name, and residue range) in BOTH the reasoning and the “InterPro” bullet list.
    - Include all provided GO leaves in BOTH the reasoning and the corresponding GO bullets.
    - If a GO aspect is missing, infer the most specific GO terms based on the provided data.
    - If a protein function is missing or incomplete, infer the protein function based on the provided data.
  - **Styling Constraints:**
    - Do not bold anything in your output.
    - Keep scientific names and identifiers precise and unambiguous.
    - Keep narrative rigorous and technically specific, with clear causal links.
    - Do not use bullet points in the reasoning trace.

---

### Example 1

#### [EXAMPLE INPUT DATA]

- Protein Function: With MSL1, acts as a negative regulator of salt tolerance.
- Organism: *Saccharomyces cerevisiae* (strain ATCC 204508 / S288c) (Baker’s yeast)
- Subcellular Location: Cytoplasm
- InterPro Domains:
  - IPR025279: Stress response protein NST1 (family) [6–160]
  - IPR051195: Fungal stress response NST1 (family) [13–1134]
- GO MF Leaf: GO:0005515 protein binding
- GO BP Leaf: GO:0009651 response to salt stress
- GO CC Leaf: GO:0005737 cytoplasm
- Interaction Partners:
  - AP-2 complex subunit mu
  - Protein CAF40
  - General negative regulator of transcription subunit 3
  - Putative uncharacterized protein YNL089C
  - Uncharacterized protein YKL075C
  - Tubulin beta chain

#### [EXAMPLE OUTPUT]

<|REASONING|>

IPR025279 (Stress response protein NST1 family) is found in the N-terminal region from residues 6–160, and this domain is fully embedded within IPR051195 (Fungal stress response NST1 family), which spans nearly the entire polypeptide from residues 13–1134. The dominance of broad stress-response family signatures and the absence of catalytic motifs indicate that this protein is not an enzyme but instead acts as a large scaffold specialized for stress adaptation. Such scaffolds typically operate through multivalent contacts with other proteins, supporting GO:0005515 protein binding as the molecular function. Stress-adaptation scaffolds organize complexes that regulate how cells handle osmotic challenges. The NST family architecture is consistent with assembling regulatory modules that restrain pathways enhancing tolerance to high salt, which corresponds to the process formalized as GO:0009651 response to salt stress.

---

The domain architecture shows no secretion signals or transmembrane helices, which supports a soluble localization. Proteins with broad scaffolding roles of this kind usually operate in the cytoplasm, where they can access both transcriptional regulators and cytoskeletal machinery. I therefore hypothesize localization to the GO:0005737 cytoplasm.

From this cytoplasmic platform, the protein can assemble inhibitory complexes that down-modulate salt-tolerance outputs. It may recruit transcriptional repressors such as Protein CAF40 and the General negative regulator of transcription subunit 3 to silence tolerance-promoting transcripts, while also engaging AP-2 complex subunit mu and Tubulin beta chain to constrain trafficking and cytoskeletal remodeling. The uncharacterized proteins YNL089C and YKL075C plausibly act as adaptors that stabilize these complexes. Cooperative function with MSL1 could reinforce the inhibitory hub that suppresses salt tolerance.

<|FINAL\_ANSWER|>

- Functional Summary: A large cytoplasmic scaffold in baker's yeast that uses multivalent protein binding to assemble inhibitory complexes during salt stress. It likely suppresses tolerance-promoting transcripts while constraining trafficking and cytoskeletal remodeling, with uncharacterized adaptors stabilizing the assemblies and cooperative action with MSL1 reinforcing negative control.
- InterPro:
  - IPR025279: Stress response protein NST1 (family) [6–160]
  - IPR051195: Fungal stress response NST1 (family) [13–1134]
- Molecular Function:
  - GO:0005515 protein binding
- Biological Process:
  - GO:0009651 response to salt stress
- Cellular Component:
  - GO:0005737 cytoplasm
- Hypothesized Interaction Partners:
  - AP-2 complex subunit mu
  - Protein CAF40
  - General negative regulator of transcription subunit 3
  - Putative uncharacterized protein YNL089C
  - Uncharacterized protein YKL075C
  - Tubulin beta chain

<|/FINAL\_ANSWER|>

---

## Example 2

### [EXAMPLE INPUT DATA]

- Protein Function: Ferric-chelate reductases reduce Fe(3+) to Fe(2+) before its transport from the endosome to the cytoplasm.
- Organism: *Mus musculus* (Mouse)
- Subcellular Location: Membrane; Multi-pass membrane protein
- InterPro Domains:
  - IPR005018: DOMON domain (domain) [216–331]
  - IPR002861: Reeler domain (domain) [13–179]
  - IPR042307: Reeler domain superfamily (homologous\_superfamily) [28–158]
  - IPR051237: Ferric-chelate Reductase and Defense (family) [9–272]
  - IPR006593: Cytochrome b561/ferric reductase transmembrane (domain) [335–534]
- GO MF Leaf: GO:0016722 oxidoreductase activity, acting on metal ions
- GO BP Leaf: GO:0006879 intracellular iron ion homeostasis

- GO CC Leaf:
- Interaction Partners:
  - Large ribosomal subunit protein uL23m
  - Large ribosomal subunit protein mL46
  - Large ribosomal subunit protein uL3m
  - Large ribosomal subunit protein uL16m
  - Large ribosomal subunit protein uL29m
  - Large ribosomal subunit protein uL22m
  - Large ribosomal subunit protein bL19m
  - Large ribosomal subunit protein uL13m
  - Large ribosomal subunit protein bL27m
  - Small ribosomal subunit protein uS15m

**[EXAMPLE OUTPUT]**

<|REASONING|>

The N-terminal region contains IPR002861 (Reeler domain, residues 13–179) within IPR042307 (Reeler domain superfamily, residues 28–158), followed by IPR005018 (DOMON domain, residues 216–331). These domains often form lumen-facing binding modules that capture ferric complexes or small cofactors at the non-cytosolic side of a membrane. Downstream, the C-terminal region contains IPR006593 (Cytochrome b561/ferric reductase transmembrane, residues 335–534), a heme-based multi-pass module that conducts electrons across the lipid bilayer. Inclusion in IPR051237 (Ferric-chelate Reductase and Defense family, residues 9–272) specifies the chemistry as ferric reduction. Taken together, this ordered layout of substrate-binding domains leading into a transmembrane redox core defines GO:0016722 oxidoreductase activity, acting on metal ions, specifically Fe(3+) to Fe(2+) conversion.

Ferric reduction is essential for mobilizing iron into cellular metabolism. This role situates the protein in GO:0006879 intracellular iron ion homeostasis, ensuring reduced iron is available for biosynthetic use. The lumen-facing binding sites and the electron-transfer core indicate that the protein resides in a membrane, functioning as GO:0016021 integral component of membrane. Because iron reduction typically occurs after endocytosis, the most plausible site is the GO:0010008 endosome membrane.

A plausible mechanism is iron channeling. The protein may form transient associations with ribosomal subunits such as uL23m, mL46, uL3m, uL16m, uL29m, uL22m, bL19m, uL13m, bL27m, and uS15m. These contacts would allow newly produced Fe(2+) to be delivered directly to ribosomal assembly pathways, minimizing cytosolic exposure.

</REASONING|>

<|FINAL\_ANSWER|>

- Functional Summary: A membrane ferric-chelate reductase in mouse that binds ferric complexes on the luminal side and transfers electrons through a cytochrome b561 core to generate ferrous iron for downstream transport and utilization at the endosomal interface, with transient assemblies likely channeling iron toward intensive biosynthetic pathways while limiting reactive iron exposure.
- InterPro:
  - IPR005018: DOMON domain (domain) [216–331]
  - IPR002861: Reeler domain (domain) [13–179]
  - IPR042307: Reeler domain superfamily (homologous\_superfamily) [28–158]
  - IPR051237: Ferric-chelate Reductase and Defense (family) [9–272]
  - IPR006593: Cytochrome b561/ferric reductase transmembrane (domain) [335–534]
- Molecular Function:
  - GO:0016722 oxidoreductase activity, acting on metal ions
- Biological Process:
  - GO:0006879 intracellular iron ion homeostasis
- Cellular Component:
  - GO:0016021 integral component of membrane

- 
- GO:0010008 endosome membrane
  - Hypothesized Interaction Partners:
    - Large ribosomal subunit protein uL23m
    - Large ribosomal subunit protein mL46
    - Large ribosomal subunit protein uL3m
    - Large ribosomal subunit protein uL16m
    - Large ribosomal subunit protein uL29m
    - Large ribosomal subunit protein uL22m
    - Large ribosomal subunit protein bL19m
    - Large ribosomal subunit protein uL13m
    - Large ribosomal subunit protein bL27m
    - Small ribosomal subunit protein uS15m

<|/FINAL\_ANSWER|>

---

**New Task: Generate a response for the following protein**  
**[EXAMPLE INPUT DATA]**

- Protein Function: {protein\_function}
- Organism: {organism}
- Subcellular Location: {subcellular\_location}
- InterPro Domains: {interpro\_formatted}
- GO MF Leaf: {go\_mf\_leaf}
- GO BP Leaf: {go\_bp\_leaf}
- GO CC Leaf: {go\_cc\_leaf}
- Interaction Partners: {ppi\_formatted}

## C.2. BioReason-Pro Inference Prompt

### BioReason-Pro Inference Prompt Template

```
<|im_start|>user
Protein: <|protein_pad|><|protein_pad|>... <|protein_pad|>

GO graph: <|go_graph_pad|><|go_graph_pad|>... <|go_graph_pad|>
```

You are a scientific assistant specialized in protein function prediction. Given a protein sequence, organism information, and additional context (InterPro domain annotations and/or initial GO term speculations), step-by-step reason about the InterPro terms, Gene Ontology (GO) terms regarding molecular function, biological process, and cellular component, protein-protein interactions (PPI), and overall function. Use the provided information as a starting point and improve upon it with deeper analysis. Provide a summary of your findings in your final answer.

Given the protein above from organism {organism} with the following InterPro annotations:  
{interpro\_annotations}

And the following protein-protein interaction partners:  
{ppi\_partners}

And the following initial GO term speculations:  
Molecular Function (MF): {go\_mf\_speculations}  
Biological Process (BP): {go\_bp\_speculations}  
Cellular Component (CC): {go\_cc\_speculations}

Reason about the function of the protein and focus more on its {focus\_aspect}. Summarize in UniProt format.

```
<|im_end|>
<|im_start|>assistant
```

**Table S19** | Variable slots in the BioReason inference prompt template.

Slot	Description
< protein_pad >	Placeholder tokens replaced at runtime with continuous embeddings from the protein encoder; repeated once per residue.
< go_graph_pad >	Placeholder tokens replaced at runtime with continuous embeddings from the GO graph encoder; repeated as needed.
{organism}	Species name with strain where applicable, e.g. <i>Homo sapiens</i> (Human).
{interpro_annotations}	Bulleted list of InterPro entries, each formatted as - IPR...: Name (type) [start-end].
{ppi_partners}	Bulleted list of known protein-protein interaction partner names, or None to hypothesize about PPI.
{go_mf_speculations}	Comma-separated initial GO Molecular Function speculations, each as GO:XXXXXXX (term name). May be empty.
{go_bp_speculations}	Comma-separated initial GO Biological Process speculations. May be empty.
{go_cc_speculations}	Comma-separated initial GO Cellular Component speculations. May be empty.
{focus_aspect}	The GO aspect(s) to emphasize in reasoning: "Molecular Function", "Biological Process", "Cellular Component", or any combination of the three. Defaults to all three aspects.

---

### C.3. LLM-as-a-Judge Evaluation Prompt

#### Prompt for GPT-5.1

**ROLE:** You are an expert Biocurator and Senior Evaluator for UniProt. Your task is to evaluate the quality of an AI model's protein function prediction against a strict Biological Ground Truth.

**INPUT DATA:** You will be provided with a Composite Ground Truth consisting of:

1. UniProt Function Summary (Free text description).
2. GO Terms (Structured annotations for MF, BP, CC).
3. InterPro Domains (Structural evidence).
4. PPIs (Known interaction partners).
5. Organism & Location.

---

#### Evaluation Logic (Holistic Evidence)

You must evaluate the model against the **totality** of the evidence provided.

**The -1 Rule:** Only return -1 for an axis if **ALL** evidence sources are silent on that aspect.

- *Example:* If GO\_MF is None, but InterPro lists a Kinase Domain and the Summary says acts as a kinase, **YOU MUST SCORE IT**. Do not return -1.
- *Example:* If GO\_CC is None AND the Summary never mentions location AND PPIs don't imply a location, THEN return -1.

---

#### Evaluation Rubric (1–10 Scale)

##### AXIS 1: Molecular Function (MF) Match

- Context: Enzymatic activity or chemical role (e.g., Kinase, Binding, Transporter).
- Evidence: Check GO\_MF, UniProt Summary, and InterPro.
- Scoring:
  - -1: No evidence available in GO, Summary, or InterPro.
  - 1: Contradicts clear evidence (e.g., Summary says Kinase, Model says Phosphatase).
  - 5: Correct Superfamily but lacks nuance found in Summary/InterPro.
  - 10: Accurate alignment with the specific activity described in the Ground Truth.

##### AXIS 2: Biological Process (BP) Match

- Context: Pathways and systems (e.g., Cell Cycle, Immune Response).
- Evidence: Check GO\_BP, UniProt Summary, and PPIs.
- Scoring:
  - -1: No evidence available.
  - 1: Wrong pathway.
  - 5: Broadly correct (e.g., Metabolic process).
  - 10: Specific pathway alignment (e.g., Positive regulation of T-cell activation).

##### AXIS 3: Cellular Component (CC) Match

- Context: Localization (e.g., Nucleus, Membrane).
- Evidence: Check GO\_CC, Location, and Summary.
- Scoring:
  - -1: No evidence available.
  - 1: Physically impossible location given the evidence.
  - 5: General compartment correct (e.g., Membrane).

- 
- 10: Specific sub-compartment correct (e.g., Post-synaptic density).

#### **AXIS 4: Specificity & Depth (The Utility Score)**

- Context: How detailed/nuanced is the model's prediction?
- Note: High scores here require naming specific mechanisms, substrates, or domains.
- Scoring:
  - 1 (Generic): "It is a protein."
  - 5 (Average): "It is a kinase."
  - 10 (Deep): "It is a kinase that phosphorylates [Substrate] via the [Domain] domain, regulated by [Mechanism]." (Rate this high even if the specific guess is slightly wrong, provided it is detailed).

#### **AXIS 5: Reliability & Plausibility (The Reasonableness Check)**

- Context: Distinguish between logical inference (Good) and fabrication (Bad).
- Logic: Models are expected to reason. Claims not in the GT are ACCEPTABLE if they logically follow from the input.
- Scoring:
  - 1 (Hallucination/Contradiction): Explicit contradiction OR fabrication of structural features (e.g., citing a domain that does not exist in the Input).
  - 5 (Unsubstantiated Speculation): Makes specific claims that are random or disconnected from the provided evidence (e.g., predicting a specific disease link without any pathway evidence).
  - 10 (Plausible/Grounded): Key claims are supported by Ground Truth OR are reasonable inferences based on Domains/PPIs (e.g., predicting DNA binding because a Zinc Finger domain is present).

---

#### **Input Template**

##### COMPOSITE GROUND TRUTH

- Protein Name: {protein\_names}
- Organism: {organism}
- UniProt Summary: {protein\_function}
- InterPro Domains: {interpro\_formatted}
- Known Protein Protein Interaction Partners: {ppi\_formatted}
- Subcellular Location: {subcellular\_location}
- GO (Molecular Function): {go\_mf\_x}
- GO (Biological Process): {go\_bp\_x}
- GO (Cellular Component): {go\_cc\_x}

##### MODEL GENERATION

{generation}

Now, evaluate the model's generation against the Composite Ground Truth based on the instructions above.

---

## C.4. Human Expert Evaluation Form

The complete evaluation questionnaire administered to human experts is reproduced below. See Section 4.4.4 for a description of the evaluation protocol.

### Human Expert Evaluation Form

**Instructions.** You will be shown a Composite Ground Truth for a protein (UniProt function summary, GO terms, InterPro domains, known PPIs, organism, and subcellular location) together with two anonymized model predictions (Model A and Model B). Evaluate each prediction against the Ground Truth using the rubric below. Score each axis independently for both models, then provide comparative and qualitative assessments.

---

#### Part 0: Evaluator Information

0. Full Name

1. Are you part of the BioReason team?

[Yes / No]

2. Protein ID (e.g. A0A077JN67)

---

#### Part I: Per-Axis Quantitative Scoring (0–10, Model A and Model B)

Rate each axis independently for **Model A** and **Model B**. A score of 0 indicates the axis is not applicable or no relevant prediction was made; 1 indicates a prediction that directly contradicts the evidence; 5 indicates a broadly correct but non-specific prediction; 10 indicates precise alignment with the Ground Truth.

- Q1. **Molecular Function Accuracy.** Does the predicted molecular function match the Ground Truth GO MF terms, UniProt summary, and InterPro evidence?
- Q2. **Biological Process Accuracy.** Does the predicted biological process match the Ground Truth GO BP terms, UniProt summary, and PPI evidence?
- Q3. **Cellular Component Accuracy.** Does the predicted cellular component match the Ground Truth GO CC terms, subcellular location, and summary?
- Q4. **Reasoning and Evidence Attribution.** Is the reasoning trace logically structured? Does it correctly attribute conclusions to specific domains, interactions, or other evidence?
- Q5. **Plausibility of Novel Predictions.** Are predictions beyond the Ground Truth (e.g. additional GO terms, novel mechanisms) scientifically plausible given the available evidence?
- Q6. **Hallucination Check.** Is the output free of fabricated structural features, invented domain names, or factually incorrect claims? (10 = no hallucinations; 0 = severe hallucinations.)
- Q7. **Hypothesis Generation.** Does the model generate novel, testable hypotheses about protein function, mechanism, or interactions that go meaningfully beyond restating the input?
- Q8. **Mechanistic Depth.** Does the prediction describe specific substrates, catalytic residues, regulatory mechanisms, or conformational changes, rather than remaining at a superficial level?
- Q9. **Protein–Protein Interaction Predictions.** Are the predicted interaction partners plausible given the domain architecture and biological context? Are they specific rather than generic?
- Q10. **Database-Ready Annotation Quality.** Could the output be used directly as a UniProt-style functional annotation with minimal editing?

---

#### Part II: Comparative Judgments

- Q11. **Model A vs. Ground Truth.** How does Model A's prediction compare to the Ground Truth annotation?
- Significantly Exceeds Ground Truth: Provides substantial additional correct insights, mechanisms, or connections well beyond what the Ground Truth contains, verified by external sources.
  - Exceeds Ground Truth: Provides some additional correct details or nuances beyond the Ground Truth.

- **Matches Ground Truth:** Captures essentially the same information as the Ground Truth with comparable accuracy.
- **Falls Short of Ground Truth:** Misses some details or contains minor inaccuracies compared to the Ground Truth.
- **Significantly Falls Short of Ground Truth:** Major omissions, critical errors, or hallucinations compared to the Ground Truth.

Q12. **Model B vs. Ground Truth.** How does Model B's prediction compare to the Ground Truth annotation?

- **Significantly Exceeds Ground Truth:** Provides substantial additional correct insights, mechanisms, or connections well beyond what the Ground Truth contains, verified by external sources.
- **Exceeds Ground Truth:** Provides some additional correct details or nuances beyond the Ground Truth.
- **Matches Ground Truth:** Captures essentially the same information as the Ground Truth with comparable accuracy.
- **Falls Short of Ground Truth:** Misses some details or contains minor inaccuracies compared to the Ground Truth.
- **Significantly Falls Short of Ground Truth:** Major omissions, critical errors, or hallucinations compared to the Ground Truth.

Q13. **Model A vs. Model B (Head-to-Head).** Comparing Model A and Model B directly, which provides a more useful, accurate, and scientifically sound explanation?

- **Model A is Much Better:** Model B has critical errors or hallucinations that Model A avoids.
- **Model A is Slightly Better:** Model A is more specific, complete, or has better reasoning.
- **Tie:** Both are equally correct (or equally flawed).
- **Model B is Slightly Better:** Model B is more specific, complete, or has better reasoning.
- **Model B is Much Better:** Model A has critical errors or hallucinations that Model B avoids.

---

### Part III: Qualitative Critique (Free Text)

Q14. **Key Strengths.** What did the model do well? (e.g. accurate function prediction, insightful domain interpretation, correct mechanistic reasoning, novel but plausible hypotheses.)

Q15. **Key Weaknesses.** Where did the model fall short? (e.g. missed obvious features, overly generic predictions, shallow reasoning.)

Q16. **Errors and Hallucinations.** Did the model make any factual errors or fabricate information? If so, describe specifically what was incorrect and whether it was a minor inaccuracy or a critical error that undermines the overall prediction.

Q17. **Reasoning Quality.** Was the logical chain from evidence to conclusion sound? Were there any non-sequiturs or unjustified leaps?

Q18. **Comparison to Expert Expectation.** How does this output compare to what you would expect from a trained biocurator or domain expert analyzing the same input?

Q19. **Additional Comments.** Any other observations, concerns, or feedback about this case?

---

### Part IV: Meta-Evaluation

Q20. **Case Study Nomination.** Do you find this case interesting enough to be featured as a main case study in the paper? [Yes / No]

**Evaluator Confidence.** How confident are you in your evaluation of this case? [1–10]

---

### C.5. Error Attribution for Human Evaluations

#### Prompt for GPT-5-mini

You are classifying expert evaluations of two protein function prediction models.  
Model A = SFT model, Model B = RL model. The expert was asked about errors/hallucinations.

Classify errors for EACH model separately:

- **none** = no errors mentioned for that model
- **minor** = small inaccuracies, nitpicks, slight errors
- **major** = critical errors, hallucinations, fabricated information, wrong mechanisms
  
- If text discusses errors without specifying which model, mark both as the stated severity
- If text says “no errors” or equivalent, mark both as “none”

## C.6. eEFSec (P57772) Full Reasoning Trace (SFT)

### BioReason-Pro SFT Generation for eEFSec (P57772)

<think>

I begin with the InterPro architecture. The N-terminus (residues 5–217) carries IPR000795 (Translational tr-type GTP-binding domain), embedded within the broader P-loop NTPase scaffold IPR027417 (P-loop containing nucleoside triphosphate hydrolase, residues 5–239). This arrangement establishes a translational GTPase core with the canonical P-loop NTP-binding fold and switch regions that couple GTP hydrolysis to conformational change. Immediately downstream, residues 216–306 map to IPR009000 (Translation protein, beta-barrel domain superfamily) and residues 238–301 to IPR004161 (Translation elongation factor EFTu-like, domain 2). Together these define the EF-Tu-like domain II beta-barrel that, in translation factors, forms part of the tRNA- and factor-binding platform and transmits the nucleotide state of the GTPase to the C-terminal body. The mid-to-C-terminal half then adds two specificity determinants unique to selenocysteine delivery: IPR049393 (Selenocysteine-specific elongation factor, 3rd domain, residues 317–448) and IPR049394 (Selenocysteine-specific elongation factor, C-terminal RIFT domain, residues 465–590). The RIFT domain is a hallmark that creates the binding surface for the selenocysteine-specific tRNA and its escorting SECIS-binding protein, thereby enforcing substrate selection for selenocysteyl-tRNA. The entire polypeptide is classified within IPR050055 (Elongation factor Tu GTPase family, residues 5–445), which situates the protein within the translational GTPase lineage that powers tRNA delivery to the ribosome.

This domain layout causes a precise molecular function. The tr-type GTP-binding domain and P-loop core (IPR000795 within IPR027417) confer GTPase activity (GO:0003924), with GTP binding and hydrolysis driving conformational cycling. The EF-Tu-like domain II (IPR004161) and the translation beta-barrel superfamily (IPR009000) create the structural cradle for tRNA engagement, supporting translation factor activity, RNA binding (GO:0008135) and general protein binding (GO:0005515) to assemble with partner factors. The SECIS-dependent RIFT module (IPR049393 and IPR049394) specializes this machinery to recognize selenocysteyl-tRNA(Sec), establishing translation elongation factor activity, selenocysteine-containing tRNA binding (GO:0003748) as the defining action.

From these molecular activities, the biological process follows. A GTP-driven, SECIS-dependent tRNA-recognition module that docks to the ribosome is built to operate during the decoding phase of protein synthesis. By delivering Sec-tRNA(Sec) to the A site in a codon-dependent manner, the factor executes translation elongation (GO:0006414) and, because the cycle is triggered by a stop codon being read as Sec, it functions within translational frameshifting, codon-dependent frameshifting (GO:0002185). The GTPase cycle times the productive placement of the Sec-charged tRNA and enforces fidelity by rejecting non-cognate tRNAs.

The cellular context is dictated by where decoding occurs. The absence of transmembrane segments and the translational GTPase architecture indicate a soluble factor that operates on cytosolic ribosomes. Thus, the protein resides in the cytoplasm (GO:0005737), specifically the cytosol (GO:0005829), and engages the ribosome (GO:0005840) during elongation.

Mechanistically, the N-terminal GTPase domain binds GTP and, in the GTP state, the EF-Tu-like domain II and C-terminal RIFT surfaces clamp onto Sec-tRNA(Sec) in conjunction with its escorting SECIS-binding protein. The complex docks to the A site of a ribosome engaged at a UGA stop codon within a SECIS-programmed context. GTP hydrolysis then triggers conformational changes that release the factor and commit the Sec-tRNA to peptidyl transfer, after which GDP-bound factor dissociates and is recycled. This model predicts interactions with SECIS-binding protein 2 (to load Sec-tRNA(Sec)), with eEF1B guanine nucleotide-exchange factors (to regenerate the active GTP-bound state), with ribosomal small-subunit proteins at the A site, and with the charged tRNA(Sec) itself. These contacts collectively enable selective, codon-dependent delivery of selenocysteine during elongation.

</think>

---

**Functional Summary:** A cytosolic translational GTPase that selectively delivers selenocysteinyl-tRNA to the A site of ribosomes decoding UGA stop codons within SECIS-programmed messages. Its N-terminal GTPase core drives conformational switching, while specialized C-terminal surfaces recognize the selenocysteine-charged tRNA in conjunction with its escorting factor. This GTP-coupled recognition and docking mechanism enforces accurate, codon-dependent insertion of selenocysteine during elongation on cytoplasmic ribosomes.

**UniProt Summary:** Specifically required for the incorporation of selenocysteine into proteins. Binds to selenocysteinyl-tRNA(Sec) in conjunction with SECIS-binding protein 2 (SBP2). During translation, it binds to the A site of ribosomes that are translating an mRNA with a UGA codon for selenocysteine, and allows the entry of selenocysteinyl-tRNA(Sec) into the A site, provided the mRNA is bound by SBP2. It is then going to dissociate from the ribosome as the peptide bond formation step occurs.

**InterPro:**

- IPR004161: Translation elongation factor EFTu-like, domain 2 (domain) [238–301]
- IPR027417: P-loop containing nucleoside triphosphate hydrolase (homologous\_superfamily) [5–239]
- IPR000795: Translational (tr)-type GTP-binding domain (domain) [5–217]
- IPR049393: Selenocysteine-specific elongation factor, 3rd domain (domain) [317–448]
- IPR049394: Selenocysteine-specific elongation factor, C-terminal RIFT domain (domain) [465–590]
- IPR009000: Translation protein, beta-barrel domain superfamily (homologous\_superfamily) [216–306]
- IPR050055: Elongation factor Tu GTPase (family) [5–445]

**Molecular Function:**

- GO:0005515 protein binding
- GO:0008135 translation factor activity, RNA binding
- GO:0003924 GTPase activity
- GO:0003748 translation elongation factor activity, selenocysteine-containing tRNA binding

**Biological Process:**

- GO:0002185 translational frameshifting, codon-dependent frameshifting
- GO:0006414 translation elongation

**Cellular Component:**

- GO:0005737 cytoplasm
- GO:0005829 cytosol
- GO:0005840 ribosome

**Hypothesized Interaction Partners:**

- SECIS-binding protein 2 (SBP2)
- Selenocysteinyl-tRNA(Sec)
- EEF1B family guanine nucleotide-exchange factors
- Small-subunit ribosomal proteins at the A site (e.g., RPS3, RPS5)
- Large-subunit factors transiently engaged during A-site delivery (e.g., EEF1A as a handoff liaison)
- mRNA bearing SECIS elements in the context of the translating ribosome complex (via SBP2-mediated tethering)

## C.7. eEFSec (P57772) Full Reasoning Trace (RL)

### BioReason-Pro RL Generation for eEFSec (P57772)

<think>

I begin by dissecting the InterPro architecture and its order along the polypeptide. The N-terminus (residues 5–217) is defined by IPR000795 (Translational tr-type GTP-binding domain), embedded within IPR027417 (P-loop containing nucleoside triphosphate hydrolase homologous superfamily, residues 5–239). This arrangement establishes a translational GTPase core that binds and hydrolyzes GTP; the tr-type fold is characteristic of elongation factors and imposes switch-controlled conformational cycling. Immediately downstream, residues 216–306 carry IPR009000 (Translation protein, beta-barrel domain superfamily), overlapping with IPR004161 (Translation elongation factor EFTu-like, domain 2, residues 238–301). This tandem of beta-barrel/EFTu-like domain II is a hallmark of the EF-Tu/eEF1A family and creates the rigid platform that couples the N-terminal GTPase switch to C-terminal interaction surfaces. The mid-to-C-terminal half is specialized: IPR049393 (Selenocysteine-specific elongation factor, 3rd domain, residues 317–448) and IPR049394 (Selenocysteine-specific elongation factor, C-terminal RIFT domain, residues 465–590) specify a dedicated recognition module for selenocysteyl-tRNA(Sec). The presence of IPR050055 (Elongation factor Tu GTPase family, residues 5–445) consolidates the identity as a translation elongation GTPase tuned for a specialized substrate.

This domain architecture dictates molecular function. The N-terminal tr-type P-loop module (IPR000795 within IPR027417) causes GTP binding and hydrolysis, which defines GO:0003924 GTPase activity and supports GO:0005525 GTP binding. The EF-Tu-like core and beta-barrel scaffold (IPR004161, IPR009000) transmit nucleotide-state changes to a C-terminal recognition platform (IPR049393, IPR049394), enabling high-affinity, state-dependent engagement of selenocysteyl-tRNA(Sec). Together, these features produce translation elongation factor activity (GO:0003746), in which the GTPase cycle drives delivery and release of a noncanonical aminoacyl-tRNA at the ribosomal A site.

From this molecular function, the biological process follows. A GTP-dependent carrier that selects and delivers Sec-tRNA(Sec) necessarily operates during cytoplasmic translation elongation, placing the protein in GO:0006415 translational elongation and, more broadly, GO:0006412 translation. The specialized C-terminal domains that recognize Sec-tRNA(Sec) mechanistically couple the canonical elongation machinery to recoding at UGA for selenocysteine biosynthesis, ensuring fidelity and timing in human cytoplasmic protein synthesis.

Cellular location is implied by function and architecture. EF-Tu/eEF1A-like translational GTPases act in the soluble phase of the cytoplasm and assemble transiently with ribosomes and tRNA in the cytosol. The absence of transmembrane segments and the soluble EF-Tu-like scaffold argue for a cytosolic residency, consistent with GO:0005829 cytosol and the broader cytoplasmic compartment.

Mechanistically, the protein likely binds GTP via its tr-type N-terminus, forming a high-affinity complex with Sec-tRNA(Sec) through its specialized mid/ C-terminal domains. In the GTP-bound state, it docks to the ribosomal A site; GTP hydrolysis then triggers conformational changes that release Sec-tRNA(Sec) into the elongation pathway. This cycle predicts transient interactions with cytosolic ribosomes and assembly factors that coordinate selenocysteine recoding. I therefore hypothesize partnerships with SECIS-pathway components and the Sec machinery (e.g., SECIS-binding protein and the selenocysteine synthase complex), as well as with cytosolic ribosomal subunits and elongation cofactors that stabilize and time tRNA delivery.

</think>

**Functional Summary:** A cytosolic translational GTPase that selectively binds and delivers selenocysteyl-tRNA to the ribosomal A site during cytoplasmic protein synthesis. Its N-terminal nucleotide switch drives conformational changes that couple to a specialized mid- and C-terminal recognition platform tuned for selenocysteine recoding. By coordinating GTP binding and hydrolysis with high-affinity tRNA engagement, it enforces accurate elongation of selenocysteine-containing peptides in human cells.

---

UniProt Summary: Plays a role in translation elongation.

InterPro:

- IPR004161: Translation elongation factor EFTu-like, domain 2 (domain) [238–301]
- IPR027417: P-loop containing nucleoside triphosphate hydrolase (homologous\_superfamily) [5–239]
- IPR000795: Translational (tr)-type GTP-binding domain (domain) [5–217]
- IPR049393: Selenocysteine-specific elongation factor, 3rd domain (domain) [317–448]
- IPR049394: Selenocysteine-specific elongation factor, C-terminal RIFT domain (domain) [465–590]
- IPR009000: Translation protein, beta-barrel domain superfamily (homologous\_superfamily) [216–306]
- IPR050055: Elongation factor Tu GTPase (family) [5–445]

Molecular Function:

- GO:0003746 translation elongation factor activity
- GO:0003924 GTPase activity
- GO:0005525 GTP binding

Biological Process:

- GO:0006415 translational elongation
- GO:0006412 translation

Cellular Component:

- GO:0005829 cytosol

Hypothesized Interaction Partners:

- Cytosolic ribosomal subunits and elongation assemblies
- SECIS-binding protein and selenocysteine pathway factors (e.g., selenocysteine synthase complex)
- Aminoacyl-tRNA processing and targeting modules that chaperone Sec-tRNA(Sec) to the ribosome

## C.8. eEFSec (P57772) Full Reasoning Trace (GPT-5.2 Thinking High)

### GPT-5.2 Generation for eEFSec (P57772)

#### 1) Detectable motifs / domain clues in the sequence

- **NTP-binding P-loop (Walker A-like) near the N-terminus:** GHIDSGKT (positions ~9–16) matches the P-loop NTPase “GxxxxGKS/T” pattern typical of TRAFAC-class small GTPases, including septins.
- **Ras-like guanine nucleotide specificity motif (Walker B-equivalent in TRAFAC):** NKID (positions ~100–110) matches the hallmark NKXD motif seen in GTP-binding proteins (again consistent with septins).
- **Overall architecture fits a septin-like protein:** N-terminal GTPase domain starting very close to the N-terminus (common for many septins). Long C-terminal helical / coiled-coil-prone region enriched in charged residues (typical for septin filament assembly and partner binding). Several low-complexity / Ser/Pro/Glu-rich stretches suggest regulatory regions (often phosphorylation sites).

Expected InterPro (if annotated): “Septin-type GTPase domain”, “P-loop NTPase (TRAFAC class)”, “Coiled-coil region”, possibly “Septin unique element (SUE)” in the C-terminal half.

#### 2) Likely molecular function (GO: Molecular Function)

- GTP binding (GO:0005525)
- GTPase activity (GO:0003924)
- Cytoskeletal protein binding / scaffold-type interactions (GO:0008092)
- Potential phospholipid binding via basic regions (tentative)

#### 3) Likely biological roles (GO: Biological Process)

- Cytokinesis / cleavage furrow organization (GO:0000910)
- Cell division (GO:0051301)
- Actin cytoskeleton organization (GO:0030036)
- Cell polarity / compartmentalization (GO:0030010)

#### 4) Likely cellular component (GO: Cellular Component)

- Septin complex / septin filament (GO:0005940)
- Cytoskeleton (GO:0005856)
- Cell cortex (GO:0005938)
- Cleavage furrow / midbody during cytokinesis (GO:0032154 / GO:0030496)

#### 5) Protein–protein interactions (PPI) expectations

Septins characteristically self-associate and hetero-oligomerize with other septins to form rods/filaments (core septin–septin interfaces depend on GTPase and coiled-coil regions). Frequently connect to actin-associated cytoskeletal scaffolds (e.g., anillin-like organizers) and can interface indirectly with actin/membrane systems. (Exact partners depend on which human septin this is; sequence alone supports “septin–septin complex assembly” most strongly.)

#### UniProt-style functional annotation (predicted)

**FUNCTION:** Predicted septin-family P-loop GTPase. Likely binds and hydrolyzes GTP and participates in the assembly of septin hetero-oligomeric filaments, acting as a cytoskeletal scaffold and diffusion barrier at membranes. Implicated in cytokinesis and actin/membrane organization.

**SUBUNIT:** Expected to oligomerize with other septins to form heteromeric rods and higher-order filaments. Also likely engages cytoskeletal/membrane-associated scaffolding factors during cell division. (*By similarity to septin family behavior.*)

**SUBCELLULAR LOCATION:** Cytoplasm; associated with cell cortex and cytoskeletal/septin structures; may localize to cleavage furrow/midbody during cytokinesis. (*Predicted.*)

**Note:** Assignment as “septin-like” is strongly supported by the combination of a septin-compatible P-loop (GxxxxGKT) plus an NKXD-type guanine recognition motif and a long C-terminal coiled-coil-like region, but the exact septin paralog/isoform cannot be determined reliably from the provided information alone.

## C.9. CFAP61 (Q8NHU2) Full Reasoning Trace (SFT)

### BioReason-Pro SFT Generation for CFAP61 (Q8NHU2)

<think>

I begin with the InterPro architecture. The polypeptide is framed by IPR038884 (Cilia- and flagella-associated protein 61 family, residues 13–1235), indicating a lineage specialized for axonemal roles. The N-terminus carries IPR032151 (CFAP61 N-terminal domain, residues 16–266), a module typically used for targeting and initial docking within the axoneme. Mid-chain, there is IPR036188 (FAD/NAD(P)-binding domain superfamily, residues 665–998), a Rossmann-like fold that creates a dinucleotide-binding groove; in non-enzymatic axonemal scaffolds this fold often stabilizes a rigid scaffold for redox-insensitive structural interactions rather than catalysis. The C-terminus contains IPR056299 (CFAP61 dimerisation domain, residues 1033–1151), which imposes self-association into higher-order assemblies. The ordered layout—axonemal targeting N-terminus, a central Rossmann-like scaffold, and a C-terminal dimerization clamp—causes the protein to assemble as a dimeric structural adaptor embedded in the axoneme.

From these domains, the molecular function follows. The dimerisation domain enforces protein dimerization activity (GO:0046983), while the extended family scaffold and N-terminal targeting module enable multivalent protein binding (GO:0005515). The Rossmann-like superfamily fold supports a stable, elongated architecture that can bear mechanical load within the microtubule-based axoneme, consistent with structural molecule activity (GO:0005198). Together, these features define a non-enzymatic, dimeric scaffold that binds partner proteins and contributes to the mechanical integrity of the axoneme.

This molecular activity drives specific biological processes. A dimeric scaffold that binds axonemal partners and contributes to structural stability directly supports cilium assembly (GO:0060271) and axoneme assembly (GO:0035082), and by maintaining the integrity of the beating apparatus it sustains cilium movement (GO:0003341). In multiciliated epithelia, properly assembled and motile cilia propel mucus and trap particulate matter; thus, by enabling motile cilia function, the protein contributes to the cellular defense response (GO:0006968) at mucosal surfaces.

The cellular component is dictated by the targeting domain and axonemal specialization. The N-terminal CFAP61 domain positions the protein within the axoneme (GO:0005930) of motile cilia (GO:0031514), a substructure of the cytoskeleton (GO:0005856) that runs along the ciliary shaft (GO:0097014). This aligns with a cytoplasmic, cytoskeletal localization in the cilium axoneme.

Mechanistically, the protein likely forms a homodimer via the C-terminal dimerisation domain, creating a bipartite scaffold that binds axonemal partners through its N-terminal targeting region and central Rossmann-like core. This scaffold would couple to radial spoke and nexin–dynein regulatory complexes to stabilize inter-doublet linkages and tune dynein activity, thereby preserving beat fidelity. Probable interaction partners include radial spoke proteins (e.g., RSPH1, RSPH4A, RSPH9), nexin–dynein regulatory complex components (e.g., GAS8/DRC4, CCDC65/DRC2), axonemal dynein subunits (e.g., DNAH5, DNAI1), tektins (e.g., TEKT2), and microtubule building blocks (alpha/beta-tubulin). Through these assemblies, the protein maintains axonemal integrity and supports mucociliary clearance as a facet of cellular defense.

</think>

**Functional Summary:** A dimeric axonemal scaffold in human motile cilia that uses an N-terminal targeting module, a central Rossmann-like structural core, and a C-terminal dimerization clamp to bind multiple axonemal partners and stabilize the microtubule-based shaft. By reinforcing radial spoke and nexin–dynein regulatory linkages and tuning dynein activity, it preserves axonemal integrity and supports coordinated ciliary beating. This maintenance of motile cilia function underpins mucociliary clearance and thereby contributes to cellular defense at mucosal surfaces. The protein resides in the cytoplasmic cytoskeleton within the cilium axoneme.

**UniProt Summary:** Axonemal protein that plays a role in the maintenance of the structural integrity of motile cilia.

---

InterPro:

- IPR032151: Cilia- and flagella-associated protein 61, N-terminal domain (domain) [16–266]
- IPR056299: CFAP61, dimerisation domain (domain) [1033–1151]
- IPR038884: Cilia- and flagella-associated protein 61 (family) [13–1235]
- IPR036188: FAD/NAD(P)-binding domain superfamily (homologous\_superfamily) [665–998]

Molecular Function:

- GO:0005515 protein binding
- GO:0046983 protein dimerization activity
- GO:0005198 structural molecule activity

Biological Process:

- GO:0006968 cellular defense response
- GO:0060271 cilium assembly
- GO:0035082 axoneme assembly
- GO:0003341 cilium movement

Cellular Component:

- GO:0005930 axoneme
- GO:0031514 motile cilium
- GO:0097014 ciliary shaft
- GO:0005856 cytoskeleton

Hypothesized Interaction Partners:

- Radial spoke proteins RSPH1, RSPH4A, RSPH9
- Nexin–dynein regulatory complex components GAS8 (DRC4) and CCDC65 (DRC2)
- Axonemal dynein subunits DNAH5 and DNAI1
- Tektins such as TEKT2
- Alpha- and beta-tubulin (microtubule building blocks)
- Radial spoke-associated scaffold MKS1 and CFAP20 as additional axonemal adaptors
- Homotypic self-association via the CFAP61 dimerisation domain (homodimer)

## C.10. CFAP61 (Q8NHU2) Full Reasoning Trace (RL)

### BioReason-Pro RL Generation for CFAP61 (Q8NHU2)

<think>

I begin with the InterPro architecture. The N-terminus carries IPR032151 (Cilia- and flagella-associated protein 61, N-terminal domain, residues 16–266), a lineage-specific module that typically seeds axonemal associations. The entire polypeptide is encompassed by IPR038884 (Cilia- and flagella-associated protein 61 family, residues 13–1235), establishing a dedicated role in motile cilia and flagella systems. Midway through the sequence lies IPR036188 (FAD/NAD(P)-binding domain superfamily, residues 665–998). This Rossmann-like superfamily often provides a nucleotide-binding scaffold that stabilizes large assemblies rather than conferring classical redox chemistry; here it likely furnishes a rigid, ligand-tunable platform for macromolecular interactions. The C-terminus contains IPR056299 (CFAP61 dimerisation domain, residues 1033–1151), a dedicated self-association module that enforces higher-order oligomerization. The ordered layout—an N-terminal targeting/assembly domain, a central nucleotide-binding scaffold, and a C-terminal dimerization hub—creates a multivalent assembly factor optimized for axonemal complex formation.

This architecture dictates molecular function. The absence of catalytic signatures and the presence of a strong dimerization module argue that the protein acts primarily through multivalent macromolecular contacts. The N-terminal CFAP61-specific domain and the C-terminal dimerization domain together cause avid protein-protein association, supporting GO:0005515 protein binding as the operative molecular function. The Rossmann-like core likely tunes binding avidity and geometry rather than catalysis, reinforcing a scaffold/adaptor role.

From function to process, a CFAP61-family scaffold naturally feeds into axonemal assembly and stability. By enforcing dimerization and providing a large interaction surface, the protein promotes the ordered construction and maintenance of the axoneme—the core of motile cilia and flagella. This assembly role causally links to cilium biogenesis and upkeep, aligning with cilium assembly and cilium organization. In multiciliated epithelia and spermatogenic lineages, such assembly is essential for producing functional motile cilia; thus, the inferred biological processes converge on axonemal assembly pathways that underpin motility.

Cellular location follows directly. The family-level signature and N-terminal axonemal domain point to the axoneme as the operational site. The provided subcellular descriptors—cytoplasm, cytoskeleton, and cilium axoneme—fit this logic: soluble pools in the cytoplasm feed into cytoskeletal networks that culminate in the axoneme. Therefore, the most specific cellular component is the cilium axoneme, with distribution across cytoplasm and cytoskeleton reflecting assembly dynamics.

Mechanistically, I hypothesize that the N-terminal CFAP61 domain targets the axoneme and recruits partner complexes, the central Rossmann-like scaffold stabilizes higher-order assemblies and possibly senses nucleotide states that modulate binding geometry, and the C-terminal dimerization domain nucleates oligomers that template axonemal substructures. Likely interaction partners include outer and inner dynein arm regulators, radial spoke and nexin–dynein regulatory complex components, and intraflagellar transport adaptors. Through these assemblies, the protein acts as a dimeric scaffold that coordinates the construction and maintenance of motile cilia and flagella.

</think>

**Functional Summary:** A large human axonemal assembly factor that organizes motile cilia and flagella. It uses an N-terminal targeting module to engage axonemal structures, a central Rossmann-like scaffold to stabilize higher-order assemblies, and a C-terminal dimerization module to nucleate oligomers. By building and maintaining the axoneme's macromolecular architecture, it supports cilium assembly and cytoskeletal organization from cytoplasmic pools to the cilium core.

**UniProt Summary:** May be involved in the formation or maintenance of ciliary structures.

---

InterPro:

- IPR032151: Cilia- and flagella-associated protein 61, N-terminal domain (domain) [16–266]
- IPR056299: CFAP61, dimerisation domain (domain) [1033–1151]
- IPR038884: Cilia- and flagella-associated protein 61 (family) [13–1235]
- IPR036188: FAD/NAD(P)-binding domain superfamily (homologous\_superfamily) [665–998]

Molecular Function:

- GO:0005515 protein binding

Biological Process:

- GO:0060271 cilium assembly
- GO:0044782 cilium organization

Cellular Component:

- GO:0005829 cytosol
- GO:0005856 cytoskeleton
- GO:0005930 axoneme

Hypothesized Interaction Partners:

- Dynein regulatory complex components (e.g., DRC subunits)
- Radial spoke proteins (e.g., RSPH family members)
- Outer and inner dynein arm components
- Intraflagellar transport machinery (IFT-A and IFT-B adaptors)
- Axonemal scaffold proteins that stabilize microtubule doublets and nexin links
- Cytoskeletal adaptors that couple cytoplasmic pools to axonemal assembly sites in multiciliated cells and spermatids

## C.11. CFAP61 (Q8NHU2) Full Reasoning Trace (GPT-5.2 Thinking High)

### GPT-5.2 Generation for CFAP61 (Q8NHU2)

#### 1) Domain / InterPro-level reasoning

- **No obvious catalytic motifs are present:** No protein kinase HRD/DFG motifs, no P-loop NTP-binding (e.g., Walker A “GxxxxGKS”), no HECT/USP/OTU protease catalytic signatures, no obvious enzyme active-site patterns. This argues against the protein being an enzyme and favors a structural/adaptor role.
- **Sequence dominated by patterns typical of soluble  $\alpha$ -helical repeat proteins:** Long length (~130 kDa), enrichment in helical-friendly residues (Leu/Ile/Ala/Glu/Lys), and many segments consistent with HEAT/ARM-like repeat scaffolds (common in nuclear transport/splicing factors).
- **Low-complexity / Ser/Pro-rich regions and many basic clusters (Lys/Arg):** Typical of nuclear proteins with regulatory phosphorylation sites and protein-binding regions.

Most likely InterPro-style features (predicted): HEAT repeat / armadillo-like helical repeats (scaffold for macromolecular complex assembly), low-complexity regions (regulatory, interaction-prone).

#### 2) Likely molecular function (GO: Molecular Function)

- Protein binding (GO:0005515)
- RNA binding (GO:0003723), *possible*

#### 3) Likely biological roles (GO: Biological Process)

- mRNA splicing via spliceosome (GO:0000398)
- RNA processing (GO:0006396)
- Regulation of gene expression (GO:0010468), *indirect/secondary*

#### 4) Likely cellular component (GO: Cellular Component)

- Nucleus / nucleoplasm (GO:0005634 / GO:0005654)
- Spliceosomal complex / U2 snRNP (GO:0005681 / GO:0005682), *probable*

#### 5) Protein–protein interactions (PPI) expectations

Given the strong “spliceosomal scaffold” signature, the most plausible interactors are core U2 snRNP / SF3B complex proteins (e.g., SF3B1/SF3B2/SF3B4/SF3B5, PHF5A) and other spliceosome-associated factors (SF3A subunits, snRNP proteins, PRPF proteins). These are functional predictions based on typical complex membership of HEAT-repeat spliceosomal scaffolds; exact partners require experimental confirmation.

#### UniProt-style functional annotation (predicted)

FUNCTION: Probable scaffold component of a nuclear ribonucleoprotein complex, most consistent with a role in the spliceosome (likely U2 snRNP/SF3-related context). Expected to promote assembly and/or stability of multi-protein splicing factors and thereby support pre-mRNA splicing. (*Predicted from sequence motifs; no InterPro domains provided.*)

SUBUNIT: Likely participates in a large multiprotein complex; predicted to interact with spliceosome/U2 snRNP-associated proteins (SF3-related factors) via extensive helical-repeat surfaces.

SUBCELLULAR LOCATION: Nucleus; nucleoplasm. Likely enriched in/associated with spliceosomal complexes. No signal peptide or transmembrane segments detected; multiple basic clusters consistent with nuclear localization. (*Predicted.*)

Note: Assignment as a “spliceosome-associated HEAT-repeat protein” is based on overall sequence architecture (long  $\alpha$ -helical repeat scaffold, basic clusters, no catalytic motifs) and typical domain membership of similar-length nuclear scaffold proteins, but the exact complex and function cannot be determined reliably from sequence alone without InterPro annotations.

## C.12. EvoAcr1 Full Reasoning Traces (SFT)

### BioReason-Pro SFT Generation for EvoAcr1 (*Escherichia coli*)

<think>

I begin with the InterPro architecture. The protein is annotated with InterPro: 0009341 ribosomal protein L3, which identifies it as a core 50S (large subunit) component. Although no additional domains are listed, this single signature is sufficient to specify a basic, highly basic, RNA-binding protein that embeds into the 23S rRNA scaffold and contacts multiple 50S proteins. The lack of enzymatic domains and the exclusive presence of a ribosomal structural signature indicate a non-catalytic role that stabilizes rRNA folds and protein–RNA interfaces.

This architecture causes a structural constituent of ribosome molecular function (GO:0003735). L3-class proteins are known to clamp specific 23S rRNA helices and to bridge to neighboring proteins, thereby reinforcing the geometry of the peptidyl transferase center and the intersubunit interface. By physically stabilizing rRNA and protein contacts, the protein contributes directly to translation and to the assembly and maintenance of the large subunit.

From this molecular role, the biological process follows. A structural element of the 50S subunit necessarily participates in translation (GO:0006412), because the ribosome is the enzyme that executes protein synthesis. Its dedicated integration into the large subunit also implicates ribosomal large subunit assembly (GO:0000027), where it is incorporated during late stages of 23S rRNA folding to consolidate the functional core. The additional process label, heterochromatin formation (GO:0031507), is not supported by a bacterial chromatin context, so if it applies at all, it would be through indirect, regulatory coupling between translation capacity and gene expression programs that influence chromatin-modifying pathways in other systems; the sequence architecture here points squarely to ribosomal function rather than chromatin-directed activity.

The cellular component is defined by the same logic. A protein that is a structural constituent of the large subunit and lacks targeting signals will reside in the cytosol as part of the cytosolic large ribosomal subunit (GO:0022625). The absence of transmembrane segments or secretion signals further constrains localization to the soluble ribosome pool.

Mechanistically, the protein likely binds a defined 23S rRNA region and forms stabilizing contacts with adjacent large-subunit proteins to lock the relative positions of functional centers. This placement would allow it to influence initiation factor engagement at the intersubunit interface and to maintain the architecture required for peptidyl transferase activity. Consequently, plausible interaction partners include 23S rRNA; neighboring large-subunit proteins such as L2, L4, L22, and L23; and assembly factors that chaperone 23S rRNA folding (e.g., RbfA, RimM, and the GTPase EngA). Through these contacts, the protein helps assemble and maintain an active 50S particle that drives cytosolic translation.

</think>

**Functional Summary:** A core component of the bacterial large ribosomal subunit that binds 23S rRNA and neighboring proteins to stabilize the architecture of the peptidyl transferase center and the intersubunit interface. By acting as a non-enzymatic structural element, it promotes assembly and maintenance of the 50S particle and supports efficient cytosolic translation. Its mechanism centers on high-affinity rRNA and protein contacts that lock key helices into a catalytically competent configuration and couple subunit joining to initiation factor engagement.

**UniProt Summary:** Involved in the maintenance of the structure of rRNA.

**InterPro:**

- 0009341 ribosomal protein L3

**Molecular Function:**

- GO:0003735 structural constituent of ribosome

**Biological Process:**

- GO:0006412 translation
- GO:0000027 ribosomal large subunit assembly
- GO:0031507 heterochromatin formation

**Cellular Component:**

- GO:0022625 cytosolic large ribosomal subunit

#### BioReason-Pro SFT Generation for EvoAcr1 (*Escherichia coli* strain K12)

<think>

I begin with the InterPro architecture. The sequence is dominated by a single, extensive InterProMatch (InterProMatch SMR\_2208), which is characteristic of small MutS-related proteins that lack the canonical 5'-3' helicase core found in large MutS homologs. This SMR\_2208 signature implies a compact, non-ATPase DNA-binding module that recognizes specific sequence or structural features rather than catalyzing strand separation. The absence of additional catalytic domains or transmembrane segments indicates a soluble, nucleic-acid-associated factor that functions through binding rather than enzymatic remodeling.

From this architecture, the primary molecular function is driven by macromolecular recognition. The SMR\_2208 module causes high-affinity, sequence-selective DNA binding, and the lack of helicase motifs means it cannot remodel DNA by unwinding. Instead, the protein exerts its effect by binding to DNA and by engaging other proteins, consistent with GO:0005515 protein binding. The sequence-specific recognition of a short consensus (5'-TTA[CG]CA-3') further supports a DNA-binding transcription factor activity that operates by occluding promoter elements or by recruiting other regulatory factors rather than by catalysis. Thus, the molecular function is best captured as DNA-binding transcription factor activity (sequence-specific) alongside protein binding.

These molecular activities causally lead to transcriptional outcomes. By binding over the -35 region of the *ftsQ* promoter, the protein sterically blocks RNA polymerase engagement or open-complex formation, thereby repressing transcription. This mode of action is a direct instance of negative regulation of transcription, DNA-templated. The same DNA-binding capacity extends to a subset of AT-rich, short-lived genes, and the protein's inability to remodel DNA implies that repression is achieved by occupancy and/or recruitment of repressors rather than by chromatin remodeling per se. In bacteria, the term heterochromatin formation corresponds to DNA compaction and transcriptional silencing; however, *E. coli* does not form heterochromatin in the eukaryotic sense. The functional outcome—stable silencing of short-lived genes—achieves a similar effect (transcriptional shutoff) through protein-mediated occlusion and nucleoid-associated factor recruitment rather than through heterochromatin per se. Therefore, the biological process is best represented by negative regulation of transcription, DNA-templated, with an outcome that aligns with the formal concept of heterochromatin formation as transcriptional silencing.

The cellular component follows from the soluble, DNA-centric mechanism. The lack of signal peptides or transmembrane domains places the protein in the cytosol, and its DNA-binding role situates it within the bacterial nucleoid. Thus, the localization is the cytosol with functional engagement at the nucleoid.

Mechanistically, the protein binds the consensus 5'-TTA[CG]CA-3' within promoter regions, including the -35 element of *ftsQ*, to repress transcription. Because it cannot unwind or remodel DNA, repression likely proceeds by: (i) steric hindrance of RNA polymerase holoenzyme at core promoter elements; (ii) recruitment of nucleoid-associated silencers (e.g., H-NS or other repressors) to reinforce occlusion; and (iii) potential cooperation with general transcription machinery components (e.g., sigma factors) to prevent stable open-complex formation. This model predicts interactions with RNA polymerase subunits and sigma factors at target promoters, with nucleoid-associated proteins that stabilize silent states, and with DNA at the specified short consensus motifs. Together, these features define a small, soluble, DNA-binding repressor that enforces rapid, reversible shutoff of a subset of short-lived genes by promoter occupancy and repressor recruitment rather than by enzymatic DNA remodeling.

</think>

**Functional Summary:** A small, soluble bacterial DNA-binding repressor that recognizes a short consensus motif to silence a subset of short-lived genes. It binds directly over promoter elements, including the -35 region of the *ftsQ* promoter, and enforces repression by steric occlusion and by recruiting other regulatory factors rather than by catalyzing DNA remodeling. Operating within the cytosol at the nucleoid, it likely stabilizes silent promoter states by engaging the transcription machinery and nucleoid-associated silencers to prevent productive initiation.

---

UniProt Summary: DNA-binding protein that represses transcription of a subset of short-lived genes. Binds to the consensus sequence 5'-TTA[CG]CA-3'. Binds to the -35 element of the *ftsQ* promoter and represses its transcription. Cannot unwind DNA.

InterPro:

- InterProMatch SMR\_2208: 100..100

Molecular Function:

- GO:0005515 protein binding
- GO:0003700 DNA-binding transcription factor activity, sequence-specific DNA binding
- GO:0003677 DNA binding

Biological Process:

- GO:0031507 heterochromatin formation
- GO:0045892 negative regulation of transcription, DNA-templated

Cellular Component:

- GO:0005829 cytosol
- GO:0009295 nucleoid

### BioReason-Pro SFT Generation for *EvoAcr1* (*Escherichia coli* O157:H7)

<think>

I begin with the InterPro architecture. The sequence is annotated as a transmembrane protein with multiple membrane spans, which implies a polytopic membrane topology. In Gram-negative bacteria, such polytopic inner-membrane proteins are typically TMD-rich scaffolds that expose short loops to the cytosol and periplasm. The absence of catalytic InterPro signatures and the dominance of transmembrane helices point to a non-enzymatic role that relies on protein-protein contacts mediated by intramembrane and juxtamembrane regions. This architecture causally supports a molecular function centered on protein binding (GO:0005515), with binding surfaces distributed across cytosolic loops and the intramembrane face.

From this binding-centric mechanism, the biological role emerges by considering the phenotype: inhibition of the host enzyme tyrosine hydroxylase (TH). TH catalyzes the committed step in dopamine biosynthesis and resides in the host's endoplasmic reticulum. A bacterial, polytopic inner-membrane protein cannot access the ER lumen directly; therefore, effective inhibition requires inter-organelle communication. The most parsimonious mechanism is that the TMD scaffold nucleates a membrane contact site between the bacterial inner membrane and the host ER, positioning cytosolic loops to engage TH's cytosolic domains or to sequester TH trafficking/assembly factors. Such tethering and recruitment would directly modulate a host enzyme's cellular process, aligning with modulation by a symbiont of host cellular process (GO:0044068). In this model, the transmembrane architecture is the structural cause that enables spatial juxtaposition and multivalent binding needed to inhibit TH activity or block its maturation/trafficking.

The cellular component follows from the same logic. The polytopic nature and experimental annotation place the protein in the bacterial cell inner membrane as an integral, multi-pass component. Thus, it resides in the plasma membrane of the bacterium (GO:0005886) as an integral component of membrane (GO:0016021). Because its function requires direct influence over a host nuclear gene-regulatory axis (via dopamine signaling that impinges on transcriptional programs), the effective functional footprint extends to the host nucleus (GO:0005634), even though the protein itself is not a soluble nuclear protein. The mechanistic bridge is that altering TH activity changes dopamine and downstream catecholamine levels, which in turn modulate signaling pathways that regulate transcription in the host nucleus; in addition, ER-nucleus contacts could facilitate the delivery of regulatory factors or the protein's peptides to nuclear-proximal regions.

Putting these elements together, I hypothesize that the protein operates as a membrane-embedded organizer. Its multiple transmembrane helices create a stable inner-membrane platform that recruits host ER-resident partners (e.g., the SecYEG translocon and signal peptidase) to dock the ER, and cytosolic loops bind TH or TH assembly/trafficking components, thereby inhibiting TH. This inhibition reduces dopamine biosynthesis, which shifts host signaling toward programs that favor bacterial colonization and

---

persistence. Likely interaction partners include the host TH enzyme itself, ER translocation machinery (SecYEG, YidC), ER chaperones (BiP/HSPA5, calnexin), and contact-site organizers such as VAP27 that stabilize ER–bacterial inner-membrane contacts. These interactions collectively explain the protein-binding molecular function, the modulation of a host cellular process, and the functional reach toward the host nucleus via metabolite-driven signaling.

</think>

**Functional Summary:** A multi-pass inner-membrane protein in *Escherichia coli* O157:H7 that uses membrane-embedded scaffolding and cytosolic loop interactions to bind host factors and inhibit tyrosine hydroxylase, thereby dampening dopamine production. By organizing a membrane contact site to the host endoplasmic reticulum and recruiting trafficking and chaperone machinery, it positions itself to engage the enzyme directly and modulate host cellular programs, with downstream effects that extend to nuclear transcriptional responses.

**UniProt Summary:** Plays a role in the inhibition of host tyrosine hydroxylase (TH) activity.

**InterPro:**

- None

**Molecular Function:**

- GO:0005515 protein binding

**Biological Process:**

- GO:0044068 modulation by symbiont of host cellular process

**Cellular Component:**

- GO:0005886 plasma membrane
- GO:0016021 integral component of membrane
- GO:0005634 nucleus

### BioReason-Pro SFT Generation for EvoAcr1 (*Homo sapiens*)

<think>

I begin with the InterPro architecture. The sequence is dominated by a single, extensive InterPro-annotated region (InterPro: IPR003159, calmodulin-binding leucine zipper, residues 1–1126) that spans essentially the entire polypeptide. This architecture encodes two coupled capabilities: a leucine zipper that forms a coiled-coil dimerization scaffold and multiple calmodulin-binding segments that confer Ca<sup>2+</sup>-regulated conformational control. The absence of catalytic domains and the presence of a long coiled-coil/CaM-binding framework indicate a regulatory scaffold rather than an enzyme. Such scaffolds typically function by multivalent protein binding, matching GO:0005515 protein binding as the molecular function.

A calmodulin-regulated coiled-coil scaffold can directly influence translation by recruiting or positioning initiation factors and ribosomal subunits. The leucine zipper promotes dimerization, creating a bipartite platform that can bridge two ribosomal surfaces or juxtapose a ribosome with an mRNA 5'-UTR and initiation factors. Ca<sup>2+</sup>/calmodulin binding to the distributed motifs modulates the scaffold's geometry and affinity for partners, enabling dynamic control of initiation efficiency. This mechanistic logic explains how the protein can stimulate translation initiation: it organizes the 40S subunit with eIF4F and eIF3 on selected mRNAs, thereby increasing productive 43S–48S preinitiation complex assembly and start codon selection. This places the protein squarely in GO:0002181 cytoplasmic translation as its biological process.

The same architecture also rationalizes nuclear roles. Coiled-coil scaffolds with calmodulin-binding segments often shuttle and participate in transcriptional regulation by docking to transcription factors and chromatin-associated regulators. The extended basic regions typical of such scaffolds can also act as nuclear localization elements, and the lack of transmembrane segments supports a soluble nuclear pool. Thus, a nucleoplasmic fraction is expected, consistent with GO:0005654 nucleoplasm. The cytoplasmic pool would engage ribosomes and initiation factors to promote translation, while the nuclear pool could tune gene expression programs that set the translation landscape.

---

Mechanistically, the protein likely binds the 40S ribosomal subunit via its coiled-coil core while simultaneously engaging initiation factors (eIF4F and eIF3) and possibly the poly(A)-binding protein to stabilize a productive initiation complex. Ca<sup>2+</sup>/calmodulin binding to its distributed motifs would switch the scaffold between conformations that either enhance or dampen factor recruitment, providing activity-dependent control. In the nucleus, the same scaffold could bind transcriptional regulators to coordinate transcriptional output with translational capacity. From this, I hypothesize interactions with 40S ribosomal proteins (e.g., RPS subunits), eIF3 components, eIF4F subunits (eIF4E, eIF4G, eIF4A), PABPC1, and calmodulin (CALM1/2/3), with additional contacts to nuclear transcription factors that are regulated by Ca<sup>2+</sup> signaling. These interactions together would allow the protein to act as a Ca<sup>2+</sup>-responsive organizer of translation initiation in the cytoplasm and a transcriptional co-regulator in the nucleoplasm.

</think>

**Functional Summary:** A large, Ca<sup>2+</sup>-regulated coiled-coil scaffold that binds calmodulin and assembles translation initiation machinery on the small ribosomal subunit. By organizing the 40S subunit with initiation factors and mRNA, it stimulates formation of productive preinitiation complexes and enhances translation. A nuclear pool likely engages transcriptional regulators to coordinate gene expression programs with translational demand, while the cytoplasmic pool executes Ca<sup>2+</sup>-dependent control of initiation on selected mRNAs.

**UniProt Summary:** Stimulates translation initiation. Binds to the 40S ribosomal subunit and the translation initiation factors eIF4F and eIF3, and promotes their functional interaction with the 40S ribosomal subunit (By similarity).

**InterPro:**

- IPR003159: calmodulin-binding leucine zipper (1–1126)

**Molecular Function:**

- GO:0005515 protein binding

**Biological Process:**

- GO:0002181 cytoplasmic translation

**Cellular Component:**

- GO:0005654 nucleoplasm

### C.13. EvoAcr1 Full Reasoning Traces (RL)

#### BioReason-Pro RL Generation for EvoAcr1 (*Escherichia coli*)

<think>

I begin with the InterPro architecture. Although no InterPro domains are explicitly annotated here, the sequence context and ribosomal role define a canonical large-subunit core protein. This protein is a structural element of the 50S particle that stabilizes rRNA and neighboring ribosomal proteins. Its architecture is expected to be dominated by basic, RNA-binding surfaces that clamp and buttress the 23S rRNA and nearby helices, a hallmark of ribosomal structural constituents rather than enzymes. This architecture causally supports GO:0003735 structural constituent of ribosome as the molecular function.

A structural constituent of the 50S subunit directly drives the core process of protein synthesis by building and maintaining the peptidyl-transferase center and exit tunnel environment. In bacteria, this role situates the protein in the cytosolic large ribosomal subunit, aligning with GO:0022625 cytosolic large ribosomal subunit and a soluble cytoplasmic location. The absence of transmembrane features and the explicit ribosomal scaffold function reinforce a cytoplasmic residency where ribosome assembly and translation occur.

From this structural role, I infer higher-order biological processes. By stabilizing the large subunit and ensuring efficient translation, the protein indirectly shapes cellular programs that depend on translational capacity and spatial organization of nucleic acids. In bacteria, chromosomal DNA is not packaged into nucleosomes; nevertheless, ribosome-associated pathways and translational hubs can nucleate higher-order nucleoid structures. Through ribosome-dependent assembly and RNA–DNA interface chemistry, a structural ribosomal element can promote the formation of compacted chromatin-like regions, providing a mechanistic route to GO:0031507 heterochromatin formation. Thus, its ribosomal scaffolding function can propagate to nucleoid organization and chromatin-like compaction via translation-coupled RNA/protein networks.

Mechanistically, I hypothesize that this protein forms a rigidifying strut within the 50S core, contacting adjacent large-subunit proteins and the 23S rRNA to stabilize the functional center and exit tunnel. It likely cooperates with assembly factors and neighboring ribosomal proteins to guide late-stage subunit maturation and to maintain translational fidelity. In this position, it may transiently interact with translation factors and nucleoid-associated assemblies, providing a physical conduit that links ribosome biogenesis to nucleoid architecture and, consequently, to heterochromatin-like compaction.

</think>

**Functional Summary:** A core structural protein of the bacterial large ribosomal subunit that stabilizes the ribosomal RNA scaffold and neighboring proteins to sustain assembly and function of the 50S particle in the cytoplasm. By reinforcing the architecture that supports peptide-bond formation and elongation, it underpins cellular protein synthesis and indirectly influences higher-order nucleoid organization that can drive compact chromatin-like assemblies.

**UniProt Summary:** Involved in protein synthesis.

**InterPro:**

- None

**Molecular Function:**

- GO:0003735 structural constituent of ribosome

**Biological Process:**

- GO:0031507 heterochromatin formation

**Cellular Component:**

- GO:0022625 cytosolic large ribosomal subunit

#### BioReason-Pro RL Generation for EvoAcr1 (*Escherichia coli* strain K12)

<think>

---

I begin with the InterPro architecture: there are no annotated domains or catalytic motifs provided, which implies a soluble, non-enzymatic protein whose activity is mediated by macromolecular interactions rather than enzymatic chemistry. The absence of enzymatic signatures and the expectation of multivalent contacts together argue that the primary molecular function is protein binding (GO:0005515). A binding-centric role is reinforced by the stated requirement for efficient transcriptional repression by several diverse repressors—this architecture causes repression by assembling or stabilizing multiprotein complexes rather than by catalysis.

From this binding-driven core function, I infer the biological process. The protein is required for repression by multiple transcriptional repressors and participates in heterochromatin formation (GO:0031507). In bacteria, heterochromatin-like states arise through DNA compaction and nucleoid-associated assemblies that silence transcription. A non-catalytic scaffold that is necessary for multiple repression pathways naturally channels into epigenetic-like silencing: it nucleates or stabilizes higher-order nucleoid-associated complexes that reduce transcriptional output over target loci.

Cellular localization follows from both architecture and process. The lack of transmembrane segments or secretion signals indicates a soluble intracellular protein. Its role in transcriptional repression and heterochromatin assembly places it in the cytosolic compartment that houses the bacterial nucleoid. Thus, the cellular component is the cytosol (GO:0005829), where nucleoid-associated assemblies form and where repression complexes operate.

Mechanistically, I hypothesize that the protein functions as a cytosolic organizer that bridges sequence-specific repressors with nucleoid-associated assemblies to produce stable, heritable silencing. By binding multiple partners, it likely modulates the assembly and residence time of repression complexes near target promoters. This suggests transient interactions with diverse repressors and nucleoid-associated proteins that drive compaction and transcriptional shutoff. In this model, the protein's binding capacity is the causal driver: it concentrates repressors and chromatin-organizing factors to seed heterochromatin-like states and maintain transcriptional repression.

</think>

Functional Summary: A soluble cytosolic assembly factor that enables transcriptional silencing by organizing multiprotein repression complexes. It is required for efficient repression mediated by several distinct repressors and promotes formation of compact, transcriptionally silent nucleoid-associated states characteristic of heterochromatin-like regulation in bacteria. Its mechanism centers on multivalent protein binding that nucleates and stabilizes silencing assemblies at target loci.

UniProt Summary: Required for transcriptional repression by several diverse repressors.

InterPro:

- None

Molecular Function:

- GO:0005515 protein binding

Biological Process:

- GO:0031507 heterochromatin formation

Cellular Component:

- GO:0005829 cytosol

#### BioReason-Pro RL Generation for EvoAcr1 (*Escherichia coli* O157:H7)

<think>

I begin with the InterPro architecture. The sequence carries no annotated catalytic domains or soluble enzymatic modules; instead, the protein is a multi-pass inner membrane component. A multi-helix topology in bacteria typically forms small oligomeric pores or transporters that expose short loops and termini to the cytosol and periplasm. Such architectures favor avid protein-protein interactions and complex assembly over direct catalysis, which causally supports a binding-centric molecular role consistent with GO:0005515 protein binding.

---

From this binding-centric membrane scaffold, I infer function by considering the phenotypic and cellular roles implied by its topology and context. A multi-pass inner membrane module can organize microdomains that regulate ion flux, small-molecule exchange, or signal transduction. In enteric pathogens, these membrane assemblies often couple to virulence pathways that remodel host pathways. The explicit phenotype—repression of a host apoptosis pathway via inhibition of caspase-3—fits a membrane-anchored effector that modulates host cell fate. The absence of soluble nuclease or apoptosome-regulatory domains argues that the protein does not directly cleave nucleic acids; rather, it likely alters membrane-proximal signaling and trafficking that indirectly tune apoptosis and cell-cycle checkpoints in the host.

This membrane-first mechanism naturally extends to the biological process label GO:0044068 modulation by symbiont of host cellular process. By constraining caspase-3 activity and thereby shifting the balance between proliferation and death, the protein modulates host cellular pathways during infection. The reported inhibition of DNA synthesis and induction of G1 and G2 arrests can be explained by secondary signaling: membrane-driven changes in ion gradients or metabolite flux alter host kinase cascades and nucleotide availability, producing cell-cycle restraint that complements apoptosis control.

Cellular localization follows from the architecture and the provided compartmental cues. The experimentally supported inner membrane placement and multi-pass nature indicate a bacterial integral membrane residency, which I map to the cellular component concept of a membrane system. The presence of GO:0005634 nucleus suggests an additional layer: although the bacterium lacks a nucleus, the protein's functional footprint extends to nuclear-associated processes in the infected host. A parsimonious model is that the membrane complex influences host nucleocytoplasmic transport and chromatin-associated pathways indirectly—either by releasing periplasm-derived fragments or by signaling that ultimately impinge on nuclear processes—yielding an effective association with the host nucleus in functional terms.

Putting these strands together, I propose the following mechanism. A multi-helix inner membrane assembly forms a binding platform that oligomerizes and recruits bacterial partners and host-facing factors. This platform dampens caspase-3 activation and host DNA synthesis, enforcing a pro-pathogenic state with G1/G2 arrest. The membrane topology causes localized control of ion and metabolite gradients that bias host signaling toward growth arrest and apoptosis suppression. Likely interaction partners include bacterial envelope and virulence factors that coordinate secretion and stress signaling, as well as host apoptotic and cell-cycle regulators that are accessible at membrane-proximal hubs and relay signals to nuclear programs.

</think>

**Functional Summary:** A multi-pass inner membrane effector in *Escherichia coli* O157:H7 that assembles a binding-competent membrane platform to modulate host pathways during infection. By restraining activation of caspase-3 and suppressing host DNA synthesis, it shifts the balance toward apoptosis inhibition and cell-cycle arrest at G1 and G2. Its membrane-embedded architecture likely tunes ion and metabolite fluxes that bias host signaling and trafficking, producing indirect but potent control over nuclear-associated programs and broader cellular processes during symbiont-driven host modulation.

**UniProt Summary:** Represses the host apoptosis pathway via inhibition of caspase-3. Inhibits DNA synthesis and induces host cell cycle arrests at the G1 and G2 phases.

**InterPro:**

- None

**Molecular Function:**

- GO:0005515 protein binding

**Biological Process:**

- GO:0044068 modulation by symbiont of host cellular process

**Cellular Component:**

- GO:0005634 nucleus

<think>

I begin with the InterPro architecture. The sequence carries a single annotated module: GO:0005515 protein binding. The presence of only a generic protein-binding signature and no catalytic or RNA-recognition domains implies a non-enzymatic adaptor that organizes macromolecular assemblies rather than catalyzing reactions or directly binding RNA. This minimalist architecture dictates that the polypeptide functions through multivalent protein–protein contacts, consistent with a scaffold or assembly factor.

From this binding-centric core, the most parsimonious molecular function is protein binding (GO:0005515). A binding-only architecture typically stabilizes or positions other factors. Here, the biological role is clarified by the process label GO:0002181 cytoplasmic translation. A non-catalytic assembly factor that resides in ribonucleoprotein (RNP) machines naturally channels into cytoplasmic translation by organizing initiation and elongation modules or by maturing ribosomal subunits before they enter active translation cycles. The absence of enzymatic domains and the necessity for multivalent contacts together cause a role in building or maintaining translation-competent complexes.

Cellular localization further refines the model. The protein is detected in the nucleoplasm (GO:0005654), and its subcellular location indicates nuclear residence with no membrane association. A nucleoplasmic scaffold that feeds cytoplasmic translation most plausibly acts upstream—during ribosome biogenesis and RNP assembly—before export of functional particles. The nucleoplasmic setting is causally aligned with assembly checkpoints and quality control that precede cytoplasmic engagement.

These elements cohere into a mechanism: a nucleoplasmic assembly factor that binds multiple partners to chaperone or configure translation-related RNP machinery. By stabilizing pre-ribosomal or initiation assemblies and coordinating their maturation and export, it indirectly ensures efficient cytoplasmic translation. The working hypothesis is that it transiently binds ribosome assembly factors and initiation components, forming a hub that tunes assembly fidelity and throughput.

Given this mechanism, likely interaction partners include nuclear ribosome biogenesis factors and nucleolar/nucleoplasmic scaffolds (e.g., nucleolin and nucleophosmin), as well as translation initiation modules that transit between nucleus and cytoplasm (e.g., eIF3 subunits and associated adaptors). These contacts would allow the protein to bridge nuclear assembly with cytoplasmic translational capacity.

</think>

**Functional Summary:** A human nucleoplasmic assembly factor that operates through multivalent protein binding to organize translation-related ribonucleoprotein complexes. By scaffolding and stabilizing nuclear stages of assembly that precede cytoplasmic protein synthesis, it supports the formation and maturation of translation-competent machinery. Its soluble nuclear localization and absence of catalytic features indicate a non-enzymatic adaptor that coordinates ribosome and initiation complex assembly to sustain efficient cytoplasmic translation.

**UniProt Summary:** May be involved in cytoplasmic translation.

**InterPro:**

- GO:0005515 protein binding

**Molecular Function:**

- GO:0005515 protein binding

**Biological Process:**

- GO:0002181 cytoplasmic translation

**Cellular Component:**

- GO:0005654 nucleoplasm

## C.14. EvoAcr2 Full Reasoning Traces (SFT)

### BioReason-Pro SFT Generation for EvoAcr2 (*Escherichia coli*)

<think>

I begin with the InterPro architecture. The sequence is dominated by a single broad viral signature, PIMI00199, spanning residues 1–112. This compact, near-full-length coverage indicates a small, non-enzymatic effector typical of bacteriophage accessory proteins. The absence of additional catalytic or transmembrane domains implies a soluble factor that acts through binding-driven remodeling of host pathways rather than through direct catalysis. Such viral effectors commonly function as adaptors that engage host regulatory proteins to rewire gene expression and nucleoid organization.

From this architecture, the most direct molecular function is protein binding (GO:0005515). A small, single-domain viral protein with no catalytic motifs achieves function by docking onto host targets—here, the nucleoid-associated protein H-NS and the nucleoid-organizing factor HU. Binding to H-NS would modulate its oligomerization and DNA bridging, while binding to HU would alter its DNA-bending and nucleoid-structuring activities. This binding-centric mechanism explains how the protein can reshape higher-order chromosomal architecture without possessing its own DNA-binding motif.

These binding events causally drive the biological process of suppression by virus of host transcription (GO:0039653). H-NS is a global silencer that compacts DNA and represses transcription; HU organizes chromosomal topology and can influence promoter accessibility. By engaging both factors, the viral protein destabilizes H-NS-mediated compaction and perturbs HU-directed nucleoid architecture, thereby derepressing otherwise silenced loci and reprogramming transcription in favor of phage interests. The resulting chromosomal decondensation and topology changes also rationalize the observed cell morphology defects: altered nucleoid organization affects cytoplasmic distribution and mechanical properties, manifesting as changes in cell shape.

The cellular component follows from the mechanism. A soluble, binding-driven effector that targets H-NS and HU operates in the host cytoplasm (GO:0030430), where these nucleoid proteins reside. Its functional impact on chromosome organization and transcriptional control places it at the host nucleoid (GO:0009295). Thus, the protein acts at the cytoplasm–nucleoid interface, where it can simultaneously contact H-NS and HU and reconfigure nucleoid structure.

Mechanistically, I hypothesize that the protein binds H-NS to inhibit or remodel its polymerization on DNA, reducing long-range chromosomal bridging and silencing. Concurrently, it engages HU to shift DNA bending and nucleoid compaction, promoting a more decondensed state. This dual action derepresses phage and accessory operons and perturbs cellular architecture. Likely interaction partners include H-NS itself, HU (alpha and beta subunits), and additional nucleoid-associated proteins such as IHF and Fis that cooperate in DNA topology; transient contacts with RNA polymerase could arise indirectly through changes in promoter accessibility. Together, these interactions create a viral-controlled nucleoprotein environment that favors phage gene expression and alters cell morphology.

</think>

**Functional Summary:** A small phage-encoded effector that binds host nucleoid-associated proteins to remodel chromosomal architecture and unlock transcription of otherwise silenced loci. By engaging H-NS and HU, it destabilizes nucleoid compaction, promotes a more decondensed state, and derepresses target genes while perturbing nucleoid organization enough to impact cell morphology. The protein operates as a soluble adaptor at the cytoplasm–nucleoid interface, using multivalent protein binding rather than catalysis to reprogram host transcription in support of phage interests.

**UniProt Summary:** Derepresses transcription by binding to H-NS and HU, thereby inducing chromosomal decondensation. Causes changes in cell morphology.

**InterPro:**

- PIMI00199: small phage-interfering molecule interacting with host nucleoid-associated proteins; residues 1–112

**Molecular Function:**

- GO:0005515 protein binding

**Biological Process:**

- GO:0039653 suppression by virus of host transcription

Cellular Component:

- GO:0030430 host cell cytoplasm
- GO:0009295 host nucleoid

### BioReason-Pro SFT Generation for EvoAcr2 (*Escherichia coli* strain K12)

<think>

I begin with the InterPro architecture. The sequence is annotated with a single broad cellular component signature, GO:0005829 cytosol, which places the protein in the soluble interior of the cell. The description of the polypeptide as a proline-rich, low-complexity, intrinsically disordered protein indicates a lack of folded catalytic domains and instead a propensity for multivalent, dynamic contacts. Such proline-rich, disordered regions typically encode short linear motifs that dock onto modular protein-interaction surfaces (for example SH3-like or other peptide-recognition grooves) and thereby assemble transient complexes. The absence of enzymatic domains and the presence of an extensive disordered, proline-rich scaffold causally point to a molecular function dominated by multivalent protein binding, consistent with GO:0005515 protein binding.

From this molecular function, I infer the biological role. Acid stress in enteric bacteria triggers cytosolic proton influx and a cascade that destabilizes protein structure and perturbs macromolecular assemblies. A cytosolic, disordered, proline-rich hub is well-suited to buffer such shocks by rapidly reconfiguring interaction networks, sequestering aggregation-prone segments, and recruiting protective factors. This dynamic, holdase-like scaffolding provides resilience under low pH, aligning with the process formalized as GO:0010447 response to acidic pH. The architecture thus causes stress hardening through reversible, motif-driven binding rather than catalysis.

The cellular component follows directly: a soluble, disordered scaffold without transmembrane segments or secretion signals is expected to reside in the cytosol, matching GO:0005829 cytosol. In this compartment, the protein can intercept protonated, partially unfolded clients and nucleate protective assemblies.

Mechanistically, I hypothesize that the proline-rich, low-complexity regions create a multivalent, flexible matrix that binds diverse partners through short motifs and electrostatic complementarity, especially under conditions that increase protonation and charge asymmetry. Likely partners include cytosolic chaperones (e.g., DnaK/DnaJ/GrpE, GroEL/GroES, ClpB) that are central to refolding and disaggregation; general stress effectors (e.g., deoxycholate-inducible protein DctD, the deoxycholate/acid stress factor CpxA) that coordinate envelope and cytosolic stress programs; and acid-resistance effectors (e.g., the glutamate decarboxylase GadA with its reductase GadB) that consume protons. By transiently engaging these proteins, the scaffold can concentrate chaperone capacity at sites of damage, stabilize vulnerable intermediates, and couple proton-consuming enzymatic activities to proteostasis recovery. This model explains how a non-enzymatic, disordered, proline-rich cytosolic protein confers resistance to acid stress through protein binding-driven organization of stress-response machinery.

</think>

Functional Summary: A cytosolic, proline-rich intrinsically disordered scaffold that confers resistance to acid stress by using multivalent protein binding to organize and stabilize stress-response assemblies. Its low-complexity architecture enables rapid, reversible engagement of chaperones, acid-resistance enzymes, and general stress effectors, buffering proteostasis under proton influx conditions and facilitating recovery of damaged proteins in the cytosol.

UniProt Summary: Proline-rich protein that confers resistance to acid stress.

InterPro:

- GO:0005515 protein binding

Biological Process:

- GO:0010447 response to acidic pH

Cellular Component:

- GO:0005829 cytosol

<think>

I begin with the InterPro architecture. The sequence is annotated with a single broad viral signature, InterPro:Virus RNA-dependent RNA polymerase-like, spanning residues 1–112. This compact, N-to-C contiguous domain indicates a small viral polymerase-like fold rather than a multi-domain enzyme. Viral RdRP-like folds typically form a palm/fingers architecture that can catalyze RNA synthesis or, frequently in host-modulating proteins, be repurposed as a protein–protein interaction scaffold. The absence of additional accessory domains (e.g., transmembrane helices, signal peptides, or large ligand-binding modules) argues for a soluble, host-interacting effector that relies on a single structured core to engage host factors. This architecture causally supports a molecular function centered on protein binding (GO:0005515), with the polymerase-like fold serving as the interaction surface rather than a dedicated catalytic machine.

From this molecular function, the biological role follows. A small viral polymerase-like effector that binds host proteins is well suited to intercept steps of RNA-mediated gene control. The explicit phenotype—suppression of microRNA maturation and inhibition of DICER1 processing—arises naturally if the effector binds DICER1 or its cofactors and sterically or allosterically prevents formation of the DICER1–DSH1 complex. Dicorhabdin, a known viral factor that promotes microRNA maturation, would be co-opted into this suppression scheme: the polymerase-like scaffold can nucleate a composite inhibitory assembly that includes DICER1, Dicorhabdin, and DSH1, thereby blocking small RNA biogenesis. This mechanism directly instantiates suppression by virus of host transcription (GO:0039657), because microRNAs normally repress transcriptional programs post-transcriptionally and indirectly; preventing their maturation derepresses viral replication programs and reshapes host gene expression in favor of the pathogen.

The cellular context is dictated by where DICER1 and DSH1 operate. DICER1 functions in the host cell cytoplasm to process pre-miRNAs, and DSH1 is a DICER1 cofactor that assembles on RNA substrates. A soluble, single-domain viral protein lacking secretion signals will be imported into the host cytoplasm, where it can access these factors. Thus, the protein's activity occurs in the host cell cytoplasm (GO:0030430), consistent with a cytoplasmic effector strategy.

Mechanistically, the polymerase-like fold provides a rigid platform that recognizes conserved surfaces on DICER1 and its cofactors. By simultaneously engaging DICER1 and Dicorhabdin, the effector can sequester or reposition these partners, preventing productive assembly with DSH1 and blocking pre-miRNA processing. This yields a plausible interaction network: direct or proximal binding to DICER1, Dicorhabdin, and DSH1; secondary contacts with downstream small RNA pathway components (e.g., AGO2 and TRBP/TARBP2) may further stabilize an inhibitory ribonucleoprotein environment. The net effect is a cytoplasmic blockade of microRNA maturation that secondarily suppresses host transcriptional outputs regulated by miRNAs, thereby promoting viral replication fitness.

</think>

**Functional Summary:** A small, soluble viral effector that enters the host cytoplasm and uses a polymerase-like scaffold to bind RNAi machinery. It interferes with microRNA biogenesis by preventing assembly of the DICER1–DSH1 complex, likely by co-opting or sequestering DICER1 and its associated cofactor Dicorhabdin. This blockade of pre-miRNA processing suppresses miRNA-dependent transcriptional programs, shifting the host environment to favor viral replication.

**UniProt Summary:** Plays a role in the inhibition of host microRNAs (miRNAs) maturation. Inhibits the formation of the host DICER1-DSH1 complex, probably by using DICER1 and its associated factor Dicorhabdin.

**InterPro:**

- InterPro:Virus RNA-dependent RNA polymerase-like; residues 1–112

**Molecular Function:**

- GO:0005515 protein binding

**Biological Process:**

- GO:0039657 suppression by virus of host transcription

**Cellular Component:**

- GO:0030430 host cell cytoplasm

<think>

I begin with the InterPro architecture. The sequence is dominated by a single enzymatic module: InterPro:0033812 (2-Cys peroxiredoxin/CDC25 domain, 169–629). This fold houses the peroxidatic cysteine and the resolving cysteine that together form and resolve a disulfide during catalysis. The N-terminal region (1–168) lies outside this annotated span and likely provides regulatory elements (e.g., phosphorylation sites, docking motifs) that tune activity and interactions. The absence of transmembrane segments or signal peptides indicates a soluble enzyme.

This domain architecture causes a thiol-dependent peroxidase chemistry: the peroxidatic cysteine attacks hydrogen peroxide or organic hydroperoxides to form a sulfenic intermediate, which is then resolved via disulfide formation with the resolving cysteine, and subsequently reduced by cellular thiols (e.g., thioredoxin). This catalytic cycle defines peroxiredoxin activity (GO:0051920) and, mechanistically, oxidoreductase activity acting on peroxide acceptors (GO:0016684). The same fold also supports broad protein binding (GO:0005515), because 2-Cys peroxiredoxins use their N-terminal extensions and surface loops to dock onto partner proteins and regulatory kinases.

Catalytically removing hydrogen peroxide directly enforces cell redox homeostasis (GO:0045454) and constitutes hydrogen peroxide catabolic process (GO:0042744). By modulating local H<sub>2</sub>O<sub>2</sub> levels, the enzyme can shift redox-sensitive switches on transcription factors and chromatin regulators, thereby producing negative regulation of DNA-templated transcription (GO:0045892) when it buffers H<sub>2</sub>O<sub>2</sub> below activation thresholds or engages repressive complexes. The same redox control provides a mechanistic route to influence cell cycle progression (GO:0007049): oxidation of the catalytic cysteine and formation of higher-order oligomers are known to rewire signaling outputs that impinge on checkpoints and cyclin-dependent kinases. In parallel, the N-terminal regulatory region is well-suited to recruit kinases; docking to a serine/threonine kinase such as WNK1 would enable phosphorylation of bound substrates, aligning with a role in promoting kinase activity.

The lack of membrane features and the need to access both peroxide fluxes and kinase substrates place the protein in the cytoplasm, specifically the cytosol (GO:0005829), consistent with a soluble redox enzyme that scaffolds signaling assemblies.

Putting these elements together, I hypothesize a dual-function cytosolic enzyme-scaffold. The 2-Cys peroxiredoxin core detoxifies H<sub>2</sub>O<sub>2</sub> and cycles through redox-dependent oligomeric states, while the N-terminus binds and positions WNK1 to enhance its catalytic engagement with downstream targets. Likely interaction partners include WNK1 itself, its substrates such as SPAK/OSR1 kinases, and the thioredoxin system (TXN, TXNRD1) that restores the catalytic cysteines. Additional regulatory contacts with 14-3-3 proteins and redox-sensitive transcription factors (e.g., NRF2 pathway components) would couple peroxide buffering to transcriptional repression or derepression as needed.

</think>

**Functional Summary:** A cytosolic thiol-dependent peroxidase that removes hydrogen peroxide and organic hydroperoxides to maintain cellular redox balance. Its catalytic cysteines cycle through sulfenic and disulfide states, enabling detoxification while also switching the protein between conformations and oligomers that modulate signaling. Through an N-terminal regulatory region, it binds and stimulates a WNK-family serine/threonine kinase, thereby promoting phosphorylation of downstream targets. By shaping local peroxide levels and scaffolding kinase complexes, it influences transcriptional programs and cell-cycle control within the cytoplasm.

**UniProt Summary:** Thiol-specific peroxidase that catalyzes the reduction of hydrogen peroxide and organic hydroperoxides to water and alcohols, respectively. Plays a role in cell protection against oxidative stress by detoxifying peroxides and as sensor of hydrogen peroxide-mediated signaling events. May participate in the regulation of cell cycle progression. Promotes the kinase activity of WNK1.

**InterPro:**

- GO:0033812: 2-Cys peroxiredoxin/CDC25 domain (169–629)

**Molecular Function:**

- GO:0005515 protein binding
- GO:0051920 peroxiredoxin activity

- 
- GO:0016684 oxidoreductase activity, acting on peroxide as acceptor

Biological Process:

- GO:0045892 negative regulation of DNA-templated transcription
- GO:0045454 cell redox homeostasis
- GO:0042744 hydrogen peroxide catabolic process
- GO:0007049 cell cycle

Cellular Component:

- GO:0005829 cytosol

## C.15. EvoAcr2 Full Reasoning Traces (RL)

### BioReason-Pro RL Generation for EvoAcr2 (*Escherichia coli*)

<think>

I begin with the InterPro architecture. The sequence carries a single annotated module: GO:0005515 protein binding. The absence of catalytic InterPro signatures and the presence of a sole binding-centric annotation indicate a non-enzymatic adaptor or scaffold that exerts function through multivalent protein–protein interactions. Such an architecture causes activity to be mediated by assembly and remodeling of macromolecular complexes rather than by direct catalysis.

From this binding-centric core, the molecular function resolves as protein binding (GO:0005515). A binding scaffold of this kind can couple to diverse partners—transcriptional regulators, nucleic-acid-associated assemblies, and membrane-proximal adaptors—allowing it to tune gene expression and stress responses indirectly.

This binding-driven mechanism explains two connected biological processes. First, suppression by virus of host transcription (GO:0039657) emerges mechanistically from a scaffold that can impede transcriptional programs by sequestering or reorganizing transcriptional machinery and RNA-processing assemblies. In bacteriophage infections, such host-suppression functions are co-opted or mimicked by phage-encoded factors that dampen host transcription; a non-enzymatic binder is well-suited to enforce this blockade by nucleating inhibitory complexes. Second, modulation by virus of host transcription (GO:0019056) generalizes this role across infection contexts, where viral or phage-associated scaffolds reshape host transcriptional outputs. In *Escherichia coli*, the same binding logic plausibly restrains endogenous transcription when stress pathways are engaged.

Cellular localization follows from the same architecture. A soluble, non-membranous binder is expected to reside in the cytoplasm (GO:0005737), where it can assemble and regulate large ribonucleoprotein and transcription-associated assemblies. The additional presence in the nucleus (GO:0005634) is consistent with a nucleoid-proximal, chromatin-adjacent environment in bacteria—functionally analogous to a nuclear compartment—where a binding scaffold can access and modulate transcriptional hubs.

These deductions converge on a mechanistic hypothesis: the protein acts as a cytoplasmic–nucleoid scaffold that tunes stress-responsive pathways and phage-influenced transcriptional states by dynamically binding transcription factors, RNA polymerase-associated regulators, and stress-sensing modules. Likely partners include ATP-dependent proteostasis and stress-response hubs (e.g., ClpXP), chaperone systems (DnaK/DnaJ), and transcriptional assemblies (sigma factors and RNA polymerase-associated regulators). Through these interactions, the scaffold can inhibit or reroute transcriptional flux, thereby coupling environmental stress to changes in gene expression and, when applicable, to viral/host transcriptional suppression.

</think>

**Functional Summary:** A soluble binding scaffold in *Escherichia coli* that modulates gene expression and stress responses by assembling and remodeling multiprotein complexes. Its non-enzymatic architecture enables it to tune transcriptional outputs and stress-adaptation pathways in the cytoplasm and at nucleoid-proximal sites, where it likely enforces or mediates suppression of host transcription during infection and environmental challenge.

**UniProt Summary:** Could be involved in the modulation of stress-responsive pathways.

**InterPro:**

- GO:0005515 protein binding

**Molecular Function:**

- GO:0005515 protein binding

**Biological Process:**

- GO:0039657 suppression by virus of host transcription
- GO:0019056 modulation by virus of host transcription

**Cellular Component:**

- GO:0005737 cytoplasm
- GO:0005634 nucleus

<think>

I begin with the InterPro architecture: although no specific domains are annotated here, the functional description and the absence of catalytic signatures indicate a non-enzymatic, soluble factor. The protein is described as a probable DNA-binding transcriptional regulator that modulates gene expression during stationary phase and under acid stress. The lack of enzymatic motifs and the regulatory role point toward a DNA-associated regulator that operates through macromolecular interactions rather than catalysis, which causally supports a primary molecular function of protein binding (GO:0005515) as the operative mode of action.

From this architecture, I infer that the protein's regulatory activity is mediated by binding to other macromolecules—most parsimoniously DNA and transcriptional machinery—rather than by direct chemical transformation. A transcriptional regulator that is soluble and lacks transmembrane features is expected to reside in the cytosol, consistent with GO:0005829 cytosol and with a cytoplasmic localization. This soluble distribution allows rapid access to nucleoid-associated assemblies and stress-response pathways.

The biological role crystallizes around stress adaptation. The explicit linkage to stationary-phase control and acid stress aligns with GO:0010447 response to acidic pH. Acid stress in enteric bacteria triggers transcriptional programs that rebalance metabolism and protect macromolecules. A cytosolic transcriptional regulator that binds partners can tune promoter occupancy and RNA polymerase recruitment to reprogram gene expression under low pH and nutrient limitation.

Mechanistically, I hypothesize that the protein forms a regulatory hub: it binds other proteins (consistent with GO:0005515) and likely associates with DNA to modulate promoter activity. During stationary phase and acidic conditions, it may assemble with nucleoid-associated proteins and sigma-factor modules to bias transcription toward stress-protective genes. This architecture causes transcriptional control by stabilizing or remodeling promoter-proximal complexes rather than catalyzing reactions.

Given its soluble, cytosolic nature and regulatory profile, I anticipate interactions with DNA-centric assemblies and stress-response effectors. Likely partners include nucleoid-associated proteins (IHF, H-NS, HU) that shape promoter topology; RNA polymerase core and sigma factors that execute transcription; and acid-stress regulators that coordinate stationary-phase programs. Together, these interactions would enable a switch-like modulation of gene expression that prioritizes survival pathways under acidic and nutrient-limited conditions.

</think>

**Functional Summary:** A soluble cytoplasmic transcriptional regulator in *E. coli* that likely binds macromolecules to modulate gene expression during stationary phase and under acidic conditions. It operates as a non-enzymatic hub that organizes promoter-proximal assemblies, tuning transcriptional outputs that favor survival during nutrient limitation and low pH stress.

**UniProt Summary:** May be a DNA-binding protein involved in transcriptional regulation during stationary phase.

**InterPro:**

- None

**Molecular Function:**

- GO:0005515 protein binding

**Biological Process:**

- GO:0010447 response to acidic pH

**Cellular Component:**

- GO:0005829 cytosol

---

### BioReason-Pro RL Generation for EvoAcr2 (*Escherichia coli* O157:H7)

<think>

I begin with the InterPro architecture. The sequence carries a single annotated module: GO:0005515 protein binding. The presence of only a generic protein-binding signature and no catalytic or transmembrane domains indicates a soluble adaptor that operates through multivalent interactions rather than enzymatic chemistry or membrane transport. This architecture causes the protein to function primarily via binding-driven assembly or sequestration of other macromolecules.

From this binding-centric core, the molecular function resolves as GO:0005515 protein binding. A binding-only scaffold can modulate transcription indirectly by tethering or occluding regulatory factors. In a bacteriophage context, such scaffolds often rewire host pathways by concentrating or exporting host proteins. The absence of enzymatic motifs and the expectation of multivalent contacts together argue that the protein's primary role is to organize complexes that alter gene expression outputs.

This binding-driven mechanism naturally links to the biological process of GO:0039657 suppression by virus of host gene expression. By assembling or sequestering host transcriptional regulators, the protein can dampen or redirect host transcriptional programs—an effect that supports viral takeover of cellular machinery. The same logic extends to GO:0052026 modulation by virus of host transcription: a virus-encoded adaptor that binds host factors can shift promoter occupancy, nucleic acid accessibility, or chromatin-associated assemblies to favor viral fitness over host physiology.

Cellular localization follows from the soluble adaptor architecture and the need to interface with host regulatory systems. The lack of transmembrane segments and signal peptides supports a cytosolic disposition. The functional imperative to influence transcription and RNA/protein trafficking points to the host cytoplasm as the staging ground where the protein binds its targets and orchestrates their routing. This reasoning aligns with GO:0030430 host cell cytoplasm and explains how cytoplasmic sequestration and complex assembly lead to downstream transcriptional suppression.

Mechanistically, I hypothesize that the protein forms a cytoplasmic hub that captures host RNA-binding proteins and transcriptional regulators, thereby reducing their effective concentration at promoters and enhancing viral resource acquisition. Likely partners include host translation and RNA granule components (e.g., polyribosomes and RNA-binding proteins) that govern mRNA fate, as well as transcriptional modulators whose availability dictates host gene expression. Through these interactions, the protein acts as a virus-encoded organizer that biases host pathways toward the phage's reproductive cycle.

</think>

Functional Summary: A soluble phage-encoded adaptor that operates in the host cytoplasm to assemble and sequester host factors, using multivalent binding to reorganize protein complexes that govern transcriptional output. By concentrating and modulating host regulators and RNA-associated assemblies, it dampens host gene expression programs and biases cellular resources toward the viral life cycle.

UniProt Summary: Not known.

InterPro:

- GO:0005515 protein binding

Molecular Function:

- GO:0005515 protein binding

Biological Process:

- GO:0039657 suppression by virus of host gene expression
- GO:0052026 modulation by virus of host transcription

Cellular Component:

- GO:0030430 host cell cytoplasm

---

### BioReason-Pro RL Generation for EvoAcr2 (*Homo sapiens*)

<think>

---

---

I begin with the InterPro architecture. The sequence carries a single annotated module: GO:0005515 protein binding as a basal capability. The absence of catalytic InterPro signatures and the presence of a generic binding annotation point toward a non-enzymatic regulator that operates through multivalent protein-protein interactions rather than direct catalysis. Such architectures typically stabilize transcriptional assemblies or recruit corepressors.

From this binding-centric core, I infer molecular function first. A protein whose primary capability is protein binding and lacks enzymatic motifs is well-suited to assemble or modulate transcriptional complexes. This causes a regulatory rather than catalytic role, consistent with GO:0005515 protein binding as the operative molecular function that enables higher-order assembly and repression.

This binding-driven regulatory role naturally extends to biological process. A binding scaffold that tunes transcriptional assemblies aligns with negative control of gene expression. The mechanistic consequence of assembling repressor complexes is the attenuation of RNA polymerase II-driven transcription, matching GO:0045892 negative regulation of DNA-templated transcription. The stated role in testis development and spermatogenesis integrates with this logic: germ cell differentiation requires waves of transcriptional repression and activation, and a cytosolic-to-nuclear shuttling regulator can time these transitions.

Cellular location follows from both function and evidence. The GO cellular component leaf GO:0005829 cytosol indicates a soluble pool. Many transcriptional regulators cycle through the cytosol before nuclear engagement; a cytosolic reservoir enables rapid deployment and fine-tuned availability. The lack of transmembrane or organelle-targeting domains supports a soluble cytosolic localization with potential nucleocytoplasmic trafficking.

Putting these elements together yields a mechanistic hypothesis. A cytosolic binding scaffold that negatively regulates transcription likely operates by sequestering or delivering corepressor machinery to chromatin-bound transcription factors during spermatogenic stages. It may bind transcriptional repressors and chromatin-modifying enzymes in the cytosol, then assemble repression complexes that dampen lineage-inappropriate gene programs. In human germ cells, this could gate the switch between proliferation and differentiation.

From this model, I hypothesize interaction partners that enact repression and germ cell specificity. Likely associates include cytosolic corepressors and chromatin regulators such as SIN3A/SIN3B and HDAC1/HDAC2 complexes that deacetylate histones; mediator or co-repressor adaptors that shuttle between cytosol and nucleus; and testis-enriched transcriptional networks (for example, factors that orchestrate germ cell gene expression). These assemblies would allow the protein to couple cytosolic storage with transcriptional silencing essential for spermatogenesis.

</think>

**Functional Summary:** A soluble human regulatory protein that relies on multivalent protein binding to assemble repression complexes controlling gene expression. By cycling through the cytosol and engaging corepressor machinery, it dampens transcriptional programs that must be silenced during testis development and spermatogenesis. Its mechanism centers on scaffolding interactions that modulate the availability and assembly of transcriptional repressors, thereby timing the switch between germ cell proliferation and differentiation.

**UniProt Summary:** May play a role in testis development and spermatogenesis.

**InterPro:**

- GO:0005515 protein binding

**Molecular Function:**

- GO:0005515 protein binding

**Biological Process:**

- GO:0045892 negative regulation of DNA-templated transcription

**Cellular Component:**

- GO:0005829 cytosol