Composing Protein Evidence: A Benchmark for Fine-Grained Protein-Text Understanding

Anonymous ACL submission

Abstract

Protein-text models should not only recognize what a protein does, but also connect features that support reasoning. Existing evaluations often score final labels, local features, or protein-text matches separately, leaving unclear whether a model could predict composed features. We introduce PannotGround, a benchmark for systematic global-local protein evidence evaluation. PannotGround projects curated protein records into three aspects, then composes them into protein-to-text and text-to-protein tasks with biologically structured hard negatives. This design goes beyond plausible protein-text matching by requiring compositional features to be jointly consistent with the queried protein. Across protein language models, protein-text alignment models, protein-LLMs, and text LLMs, we find that single-level performance does not reliably transfer to compositional binding. Alignment models are strongest, but remain sensitive to local-evidence contrasts, and many errors are biologically related near misses rather than random failures. PannotGround provides a diagnostic benchmark for fine-grained protein-text understanding and exposes global-local evidence binding as a central bottleneck for current models.

1 Introduction

Protein understanding is not only label prediction. Functional annotations tie protein-level claims to sequence evidence: domains, motifs, catalytic residues, and cofactor-binding patterns. A model that predicts an enzyme class or binding function has answered only part of the biological question. For fine-grained protein-text understanding, the model should also recognize what local features accompany the annotation and where those features are in the protein.

Current evaluations stress pieces of this problem, but rarely the full binding problem. Protein benchmarks may score a final biological label, fine-grained tasks may score a local or residue-level feature, and protein-text tasks may score a caption or retrieval match (Notin et al., 2023; Tan et al., 2026; Liu et al., 2024). These tests are useful, but they do not systematically decompose whether a prediction matched the right global semantics, local feature types, and residue/span evidence within the same source protein. They also make it hard to tell whether an error is a biologically related near miss, a shortcut, or an unrelated match, a problem highlighted in compositional vision-language evaluation and recent protein-text analyses (Thrush et al., 2022; Hsieh et al., 2023; Wu et al., 2025; Rong et al., 2025).

Figure 1 shows the benchmark example. The source protein Q08209, the human calcineurin catalytic subunit, has a global phosphatase annotation, but its interpretation also depends on metal-binding sites, an active-site proton donor, substrate-interaction motifs, calmodulin-binding regions, autoinhibitory regions, and other local evidence. UniProt-derived function descriptions can mention several of these attributes. What is missing is a systematic evaluation that records which global and local components were matched, and whether they were matched to the same protein.

PannotGround makes this capability measurable. It projects curated protein records into synchronized attribute levels: P1 for global protein function, P2 for local feature types, and P3 for residue/span evidence. Single-level tasks test these levels independently. Compositional protein-to-text (P2T) and text-to-protein (T2P) tasks then test whether these attributes remain bound to one protein. Their hard negatives are biologically close: some share global function, some share local evidence, and some share partial overlap. This design exposes global-local shortcuts instead of hiding them in aggregate accuracy.

In this paper, our benchmark PannotGround

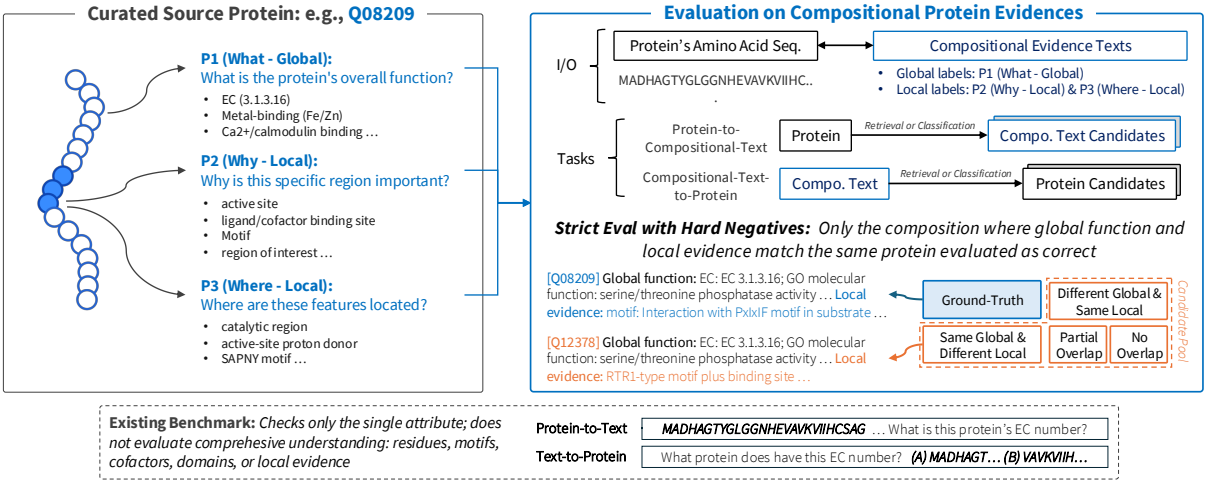


Figure 1: Overview of PannotGround using compositional protein evidence. Source proteins are curated by capturing both their global functions and local features (left). Based on this local-global information, PannotGround introduces a strict evaluation with hard negatives, such as candidates sharing global or local traits, to rigorously verify a model's comprehensive understanding of the protein (right), where existing benchmarks only evaluate single attributes (single-view), failing to assess this comprehensive understanding (bottom).

makes three contributions:

- We define *global-local evidence binding* as a core evaluation target for protein-text understanding, and instantiate it with PannotGround, a benchmark that synchronizes protein-level function, local feature identity, and residue/span evidence within the same curated source records.
- We turn this target into controlled P2T and T2P tasks with biologically structured hard negatives, enabling evaluation to distinguish exact matches from same-function, same-evidence, partial-overlap, and unrelated alternatives.
- We provide a cross-family diagnostic study showing that compositional binding is not predicted by single-level scores: PLM task heads can collapse under multi-field composition, protein-text alignment models are strongest but remain sensitive to local-evidence contrasts, and P2T/T2P errors reveal biologically structured near misses.

2 Related Work

Task-centric benchmarks and protein language models. Task-centric protein benchmarks define the dominant evaluation contract for protein understanding: given a protein sequence, variant, or representation, a model predicts a predefined biological endpoint and is compared across standardized splits and task-specific metrics. CAFA evaluates Gene Ontology function prediction; TAPE and PEER organize broad suites over structure, localization, interaction, and functional prediction; and ProteinGym focuses on mutation effects, fitness, and protein design (Zhou et al., 2019; Rao et al., 2019; Xu et al., 2022; Notin et al., 2023). Protein language models (PLMs) have become a standard model family evaluated under this contract because self-supervised training over large sequence collections learns transferable biochemical and evolutionary priors (Lin et al., 2023; Elnaggar et al., 2022; Brandes et al., 2022; Madani et al., 2023; Ferruz et al., 2022; Heinzinger et al., 2024; Elnaggar et al., 2023). This benchmark family is essential for measuring broad predictive competence, but its common evaluation contract remains task-centric: models are compared across separate endpoint tasks rather than deeply diagnosed for general understanding problem.

Fine-grained protein understanding and local evidence. Recent work increasingly recognizes that protein understanding cannot stop at global function labels. Biological resources such as InterPro and PROSITE organize families, domains, motifs, and functional sites, and benchmarks such as VenusX move evaluation below the whole-protein level by testing residue-, fragment-, and domain-level functional perception (Blum et al., 2025; Sigrist et al., 2026; Tan et al., 2026). In parallel, mechanistic and representation-interpretability

methods such as InterPLM, ProtSAE, and MotifAE suggest that dense PLM embeddings can be decomposed into human-interpretable or motif-like features associated with binding sites, structural motifs, and other local biological concepts (Simon and Zou, 2025; Liu et al., 2025; Hou et al., 2025). These works motivate the same premise as PannotGround: protein models should be evaluated on the sites, motifs, domains, and regions that provide evidence for a function, not only on the final label. The remaining gap is compositional. Fine-grained prediction or interpretability alone does not require a model to connect local evidence to a global function, a natural-language description, and the same source protein. PannotGround therefore treats local perception as a prerequisite and evaluates whether global semantics, local feature types, and explicit residue/span evidence can be synchronized and composed.

Compositional and multimodal reasoning benchmarks. Compositional multimodal benchmarks show that high aggregate accuracy can hide failures to bind entities, attributes, and evidence. CLEVR and GQA use controlled scene structure and question programs to test visual reasoning, while Winoground, CREPE, COLA, and SugarCrepe use contrastive image-text matching and hard negatives to expose failures in attribute-object binding (Johnson et al., 2017; Hudson and Manning, 2019; Thrush et al., 2022; Ma et al., 2023; Ray et al., 2023; Hsieh et al., 2023). ScienceQA, SEED-Bench, and MMMU further emphasize objective multimodal reasoning formats that combine perception with domain knowledge (Lu et al., 2022; Li et al., 2024; Yue et al., 2024). Protein-text resources similarly connect sequences, structures, phenotype descriptions, captions, and natural language through instruction following, QA, generation, and retrieval (Fang et al., 2024; PharMolix, 2025; Luo et al., 2024; Shen et al., 2024; Abdine et al., 2024; Liu et al., 2024; Xu et al., 2023; Su et al., 2025; Queen et al., 2024; Lv et al., 2025; Fallahpour et al., 2026).

3 Benchmark: PANNOTGROUND

Protein annotations are inherently multi-scale: a sequence can support global functions such as enzymatic activity or molecular binding, while the evidence for those functions often appears as local sites, domains, motifs, topology, or disordered regions. PannotGround turns this structure into a benchmark for fine-grained protein-text understanding. Each example is derived from a curated protein record $r_i = (x_i, A_i^g, A_i^\ell, s_i)$, where $x_i \in \Sigma^{L_i}$ is the protein, s_i is the split, A_i^g is a set of global annotation entries, and A_i^ℓ is a set of local annotation entries. A global entry is (n, y), with namespace n and normalized label y. A local entry is (t, d, S), with canonical feature type t, optional description d, and residue sites or spans S.

Table 1: Compact comparison with representative protein benchmarks. Coverage summarizes the biological task scope; composition indicates whether the benchmark diagnoses binding between global function, local evidence, and protein-text matching.

Benchmark	Scope	Grounding	Composition / diagnosis
Task-centric (Xu et al., 2022; Notin et al., 2023)	Broad	Partial	Separate endpoint tasks; no global-local evidence binding.
Fine-grained (Sigrist et al., 2026; Tan et al., 2026)	Local/deep	Yes	Local perception, but not protein-text global-local composition.
Protein-text (Fang et al., 2024; Shen et al., 2024; Queen et al., 2024)	Text-facing; Label-only	Partial	Often rewards plausible text or global label match without fixed evidence binding.
PannotGround	**Global function, local features, residue evidence**	**Yes**	**Binds different levels of tasks; evaluate diverse model types robustly.**

From the same record, PannotGround constructs three synchronized views. The key design principle is synchronization. They are not separately assembled tasks; they are projections of the same curated protein record. It requires us to distinguish three capabilities that single-label benchmarks often collapse: recognizing global protein function, recognizing local feature identity, and grounding local evidence to sequence coordinates. Figure 1 illustrates the distinction.

The compositional tasks test whether attributes from these levels are bound to the same protein. For eligible records, PannotGround selects global and local feature attributes and renders them as a protein-specific text description. P2T presents a protein sequence with $K = 8$ candidate captions and asks which caption matches the protein. T2P presents a compositional text query with $N = 64$ candidate proteins and asks which protein matches the query. Hard negatives are relation-aware: some preserve global while changing local evidence, some preserve local while changing global, some share only partial attributes, and some share no selected global or local signature. Controlled pro-

jections further isolate the source of evidence by varying attribute count or keeping only global or local components. Full task formats and split sizes are reported in Appendix Tables 7, 24, and 25.

PannotGround contains in-distribution and family-level out-of-distribution validation and test splits, with 282,021 unique protein records. P1 and P2 are record-level attribute sets, whereas P3 is query-conditioned; a single protein can therefore contribute multiple P3 rows. Detailed split-level counts are reported in Appendix Table 9. The benchmark is built from curated positive annotations: missing protein-level labels are not treated as complete negative evidence unless a specific evaluation view defines a closed label universe. Local spans support grounding evaluation, but PannotGround does not claim expert-curated causal links between every global label and every local span.

4 Evaluation

PannotGround evaluates two related questions. First, can a model recover the single-level tasks from protein input? Second, can it bind attributes from those levels into the same protein-specific text description? We therefore separate attribute-level evaluation from compositional retrieval, and we compare model families when they share the same answer space.

Table 2 summarizes the evaluation views used in the main results. PLM classification isolates sequence-encoder capacity in task-specific label spaces and is reported as supporting evidence in the appendix. MCQA converts single-level decisions into fixed candidate records shared across model families. P2T is an 8-way protein-to-text caption selection task, and T2P is a 64-way text-to-protein retrieval task. We use protein-LLM to refer to protein-conditioned generative language models.

Model-family labels in the result tables refer to evaluation interfaces. PLM denotes protein sequence encoders evaluated through classification heads, candidate-scoring heads, or frozen prototype similarity. PLM-TextAlign uses the same kind of frozen PLM protein encoder but maps its protein embeddings into a biomedical text-embedding space before scoring answer text; it is a post-hoc alignment probe, not a trained protein-text retrieval model. Alignment denotes protein-text models trained with contrastive protein-to-text and text-to-protein objectives, so they directly score protein-text compatibility. Protein-LLM denotes protein-conditioned generative language models evaluated by answer likelihood or direct response. Text LLM denotes language-only models that receive the benchmark prompt and answer choices as text.

In PLM classification, P1 and P2 are multi-label classification problems over global attributes and local feature types, while P3 is a localization problem over residues or spans conditioned on a local evidence query. These results measure how much each level's information is accessible from sequence encoder representations. They are useful encoder baselines, but they are not directly comparable to prompted LLM outputs because the prediction interface is different.

MCQA provides the shared cross-family interface for single-level tasks. Each example has a fixed candidate set before evaluation, but the candidate space is task-specific rather than uniformly 8-way: P1 and P3 are single-choice questions scored by accuracy, while P2 asks for the set of present local feature types and is scored by mean set F1. PLM 0-shot rows rank candidates by frozen-encoder prototype similarity; fine-tuned PLM rows use candidate-scoring heads over the same records. Text LLM and protein-LLM rows use prompting or likelihood scoring under the same candidate records. We also report an aggregate MCQA score as a compact summary, but single-level task remain the primary interpretation because they correspond to different biological capabilities.

P2T and T2P evaluate composition rather than isolated attribute recognition. In P2T, the query is a protein sequence and the candidate pool contains one gold caption plus seven hard-negative captions. A correct model must choose the caption whose P1/P2/P3 attributes all belong to the query protein. In T2P, the query is a compositional text description and the candidate pool contains 64 proteins. A correct model must rank the source protein above hard negatives that may share global function, local evidence, or partial attributes. Thus, P2T and T2P should be read as controlled retrieval/ranking tasks, not as open-ended generation tasks.

All prompts, thresholds, and decoding choices are selected on validation splits only. For prompt-based and likelihood-based systems, we additionally use no-sequence, sequence-shuffle, and choice-shuffle controls to test whether performance depends on protein input, residue order, or answer position. For P3, masked-evidence interventions test whether the score changes when the annotated evidence span is removed or isolated. Full metric

Table 2: Evaluation views used in the main results. P1/P2/P3 are biological evidence levels; P2T and T2P are task directions built from those levels.

View	Role in the paper	Prediction object or variants	Model families	Primary metrics
PLM classification	Appendix encoder-capacity baseline for P1/P2/P3	P1/P2 label sets; P3 residue or span evidence	PLM	Top-k F1; residue F1
Single-level MCQA	Main P1/P2/P3 comparison under a shared answer contract	One selected answer for P1/P3; selected feature-type set for P2	PLM, text LLMs, protein-LLMs	Accuracy; set F1
P2T multi-choice	Main compositional binding test	Rank $K = 8$ captions; Attr1/2/4, Global-only, Local-only, Full	PLM, PLM-TextAlign, bidirectional Alignment , text LLMs, protein-LLMs	Accuracy; relation diagnostics
T2P retrieval	Reverse compositional retrieval and failure analysis	Rank $N = 64$ proteins for a full compositional text query	Bidirectional contrastive models	R@1, R@5, MRR, mean rank; relation diagnostics

definitions and implementation details are given in Appendix C.

5 Results

We structure our evaluation to address three questions: (1) model performance on single-level tasks (P1 – 3), (2) the transferability of single-level competence to compositional tasks requiring the integration of global and local evidence, and (3) insights from failure patterns regarding fine-grained protein understanding. This section presents representative comparisons, while comprehensive results for single-level and compositional tasks are detailed in Appendix D and F, respectively.

5.1 Single-Level Task Performance

To evaluate model proficiency on single-level tasks, Table 3 presents results under a shared multi-choice framework, where each query contains a single ground truth among eight candidates. This standardized label space allows for a direct comparison across different model architectures.

The results indicate that performance across the three biological levels is highly variable. Zero-shot PLMs, which rank candidates by computing similarity scores between frozen protein embeddings and reference prototypes from labeled samples, demonstrate strong capabilities in global function prediction (P1) and localization (P3). Conversely, fine-tuned PLMs utilize task-specific heads, with the fine-tuned ESM-2 achieving the highest accuracy on local feature sets (P2).

In comparison, both Text LLMs and Protein-LLMs generally struggle under this rigid evaluation format, although ProLLaMA retains moderate performance in predicting local features. These discrepancies highlight that global function prediction, local feature identification, and localization require distinct modeling capabilities. For strictly single-level tasks, fine-tuning specialized PLMs remains the best approach. Furthermore, isolated PLM classification results within each subtask's native label space are detailed in Appendix Tables 14 and 15.

Table 3: Single-level multi-choice results. P1 and P3 use accuracy; P2 uses set F1. For PLM, 0-shot ranks candidates by frozen encoder prototypes, and FT uses fine-tuned candidate-scoring heads over the same multi-choice records. Split-specific results are reported in Appendix Tables 16 and 17.

Model	Setting	P1	P2	P3
PLM candidate scorers				
Ankh	0-shot	**.775**	.583	**.720**
Ankh	Fine-tuned	.498	.503	.312
ESM-2	0-shot	.760	.590	.460
ESM-2	Fine-tuned	.496	**.670**	.694
ProGen	0-shot	.675	.494	.265
ProGen	Fine-tuned	.648	.592	.301
ProstT5	0-shot	.765	.559	.674
ProstT5	Fine-tuned	.389	.579	.531
ProtBERT	0-shot	.710	.484	.455
ProtBERT	Fine-tuned	–	.623	.316
Protein-LLM				
BioReason-Pro-SFT	Prompt	.170	.013	.225
ProLLaMA	Likelihood	.230	.402	.210
Text LLM				
Llama-3.1-8B	Prompt	.370	.308	.189
Qwen2.5-7B	Prompt	.385	.281	.159
Qwen3-4B	Prompt	.345	.358	.169

5.2 Compositional Binding Is the Bottleneck

To comprehensively understand protein, models must process multiple biological levels. In our

Table 4: Controlled P2T accuracy for evidence hierarchy. Attr1 is a single-level, single-field reference: each example asks about only one P1, P2, or P3 evidence type rather than binding several fields into a caption. Full uses the original P1/P2/P3 compositional caption. Detailed split-specific rows are in Appendix Table 37. All values are percentages.

Model	Attr1	Global	Local	Full
Random	–	12.50	12.50	12.50
PLM				
Ankh	43.00	13.07	12.68	5.62
ESM-2	63.55	12.78	26.73	9.25
ProGen	49.80	13.50	19.82	7.15
ProstT5	51.35	12.53	12.93	5.92
PLM-TextAlign				
Ankh	24.12	14.88	30.30	24.18
ESM-2	23.87	26.80	40.75	36.40
ProGen	25.58	14.07	29.30	21.07
ProstT5	24.79	20.35	36.88	31.40
Text LLM				
Llama-3.1-8B	28.94	21.83	18.52	26.75
Qwen2.5-7B	27.53	11.85	10.50	13.57
Qwen3-4B	29.11	13.30	15.18	12.55
Alignment				
ProTrek-35M	53.05	73.15	43.23	57.35
ProTrek-650M	50.92	71.80	44.17	56.45
ProtST-ESM2	42.05	48.08	26.50	38.22
ProtST-ProtBERT	42.58	45.22	25.10	36.83
ProtT3	40.27	45.03	17.88	40.80
Pannot-FG	53.29	68.55	41.62	54.35
Protein-LLM				
ProLLaMA	28.08	23.38	18.90	14.22
BioReason-Pro-SFT	13.50	6.67	3.60	8.25

evaluation, a candidate text is not a simple label; it integrates global function (P1), local feature type (P2), and specific local evidence (P3). We investigate whether single-level proficiency transfers to this compositional caption selection and identify which evidence level models rely on.

As shown in Table 4, single-level performance does not reliably transfer to compositional grounding. Most model families, particularly task-specific PLMs, experience a severe performance collapse. For instance, the ESM-2 PLM achieves 63.55% on single-attribute recognition but drops to 9.25% on the Full P2T task. Alignment models perform better under this strict multi-choice framework, yet their capabilities reveal a stark imbalance.

The reliance on global versus local evidence varies heavily by architecture. Alignment models excel when global semantics are provided, with ProTrek-35M reaching 73.15% on Global-only tasks but dropping to 43.23% on Local-only tasks. Conversely, PLM-TextAlign models exhibit the opposite trend; ESM-2 TextAlign performs better on Local-only (40.75%) than Global-only (26.80%), suggesting its representations learned from masked language modeling align more naturally with local features. Ultimately, performance on the Full P2T task remains lower than on Global-only tasks across all architectures. This demonstrates that successfully binding global functional semantics with fine-grained local evidence remains the primary bottleneck for current protein-text models.

Table 5: Controlled P2T accuracy for field-count composition by model family. Attr1 is the same reference as in Table 4. Attr2 and Attr4 are 8-way P2T variants whose candidate captions display exactly two or four evidence fields, respectively. Values are family means over available model rows and the two held-out splits; detailed model-level rows are reported in Appendix Table 37. All values are percentages.

Family	Attr1	Attr2	Attr4
Random	–	12.50	12.50
PLM	51.93	8.33	8.30
PLM-TextAlign	24.52	11.16	11.86
Text LLM	28.53	17.80	14.26
Alignment	**47.03**	**40.76**	**36.76**
Protein-LLM	20.79	12.29	10.77

Table 5 investigates whether this performance degradation merely results from increased text length. By fixing the number of evidence fields, the Attr2 and Attr4 settings isolate the capacity for multi-field binding without the redundancy of full natural captions. These controlled evaluations confirm a significant transfer gap. While PLMs excel at single-level tasks, their performance drops near the 12.5% random baseline when required to integrate two or four fields. Text LLMs similarly degrade as condition constraints increase from Attr2 to Attr4. In contrast, alignment models demonstrate the most robust multi-field capability, consistently performing above the random baseline.

Importantly, Attr4 is not a direct proxy for the Full benchmark. Full captions often contain redundant biological cues, whereas Attr2 and Attr4 strictly limit evidence exposure. Ultimately, this demonstrates that the challenge of compositional understanding stems from the fundamental requirement to accurately bind corresponding global and local evidence to a specific protein, rather than simply processing longer text.

5.3 What the Failures Reveal

While §5.2 highlights the difficulty of compositional prediction, analyzing the failure patterns reveals whether these errors are arbitrary or biologically structured. To explore this, we utilize the complementary Text-to-Protein (T2P) retrieval task, where a compositional text query must iden-

Table 6: Full-compositional T2P 64-way retrieval. Random R@1 is 1.56%. R@1 and R@5 are percentages. MR is the mean rank. SG, SL, Part., and Other report the top-ranked wrong protein when R@1 fails, normalized over failures: Same-global wrong-local, Same-local wrong-global, Partial overlap, and Other.

Model	R@1	R@5	MRR	MR	SG	SL	Part.	Other
ProTrek-35M	9.67	37.80	.2463	10.59	36.0	26.8	32.4	4.8
ProTrek-650M	**11.13**	39.73	**.2622**	10.94	34.0	30.5	31.0	4.6
ProtST-ProtBERT	3.03	17.33	.1252	21.60	18.8	14.4	40.6	26.1
ProtST-ESM2	.93	7.03	.0670	34.36	9.1	9.1	14.5	67.3
ProtT3	3.20	21.60	.1394	19.56	19.8	13.4	34.6	32.2
Pannot-FG	8.90	**39.80**	.2463	**9.77**	33.6	25.4	34.9	6.1

tify the exact matching protein. Table 6 summarizes the 64-way full-compositional retrieval performance for alignment models, reporting averages across `val_id` and `test_id`. Detailed split-specific results are in Appendix Tables 32, 33, and 34.

The T2P results demonstrate that models tend to retrieve biologically related neighborhoods rather than exact protein matches. The highest R@1 remains at only 11.13%, indicating significant challenges in exact retrieval. However, the error distribution is highly non-random. For models like ProTrek and Pannot-FG, fewer than 7% of the top-ranked errors fall into the entirely unrelated "Other" category. Instead, the majority of incorrect retrievals share the same global function, local evidence, or exhibit partial overlap with the target query. Consequently, the T2P failure analysis corroborates the P2T findings: current models capture coarse protein-text compatibility but struggle to reliably bind the complete set of global and local evidence to the exact protein.

P2T diagnostics reveal the same structure from the caption-selection direction. The left panel of Figure 2 plots Global-only against Local-only P2T accuracy. Alignment models are global-semantics dominant: their Global-only score is much higher than their Local-only score. PLM-TextAlign shows the opposite tendency, with stronger Local-only than Global-only accuracy, suggesting that sequence encoders expose local feature evidence more directly than abstract global functions. Protein-LLM and text LLM rows remain closer to the random region.

The difficulty is also biologically structured rather than a generic caption length effect. The right panel of Figure 2 shows that local evidence is easier when the category is distinctive, such as transmembrane topology, active sites, or domain/motif/region evidence. It is harder for low-complexity/disorder/coiled-coil evidence, and for example, combining active-site and ligand-binding semantics. This pattern supports the biological motivation for PannotGround: fine-grained protein understanding requires connecting textual descriptions to the kinds of sequence evidence that biologists actually use.

Finally, wrong answers have a different semantic quality. Figure 3 decomposes P2T outcomes on `test_id` at the model level, with the left rail grouping rows by model family. PLM classification rows have small correct segments and large no-feature-overlap errors, showing that independent task heads often fail to recover any biologically matching caption component. Text-aligned PLM rows move substantial mass from no-overlap errors into correct or relation-preserving choices, especially for ESM2 and ProstT5. Alignment models are strongest: ProTrek and Pannot-FG get more than half of the examples correct, and their remaining errors are dominated by captions that still preserve some global or local evidence. Protein-LLMs do not show the same binding behavior. Thus, the benchmark reveals not only whether a model is wrong, but whether its wrong answer is a biologically meaningful near miss, a partial match, or a protein-unrelated choice.

Overall, PannotGround separates three phenomena that are often conflated in protein-language evaluation. Sequence encoders can recover useful global and local signals; compositional P2T/T2P tests whether those signals are bound to the same protein; and the diagnostic views reveal whether failures are biologically meaningful near misses or shortcuts. Current protein-text alignment models are the strongest compositional systems, but local evidence binding remains the main bottleneck for fine-grained protein understanding.

6 Limitations and Future Work

- **KG semantics:** structured semantic validation via knowledge graphs is promising.
- **Annotation sparsity:** Swiss-Prot evidence is incomplete; a large portion of proteins are under-explored; methods should handle missing labels.

Impact/Ethics Statement

This paper presents work whose goal is to advance the field of language modeling through the study of simple algorithmic tasks inspired by creativity. There are many potential societal consequences of

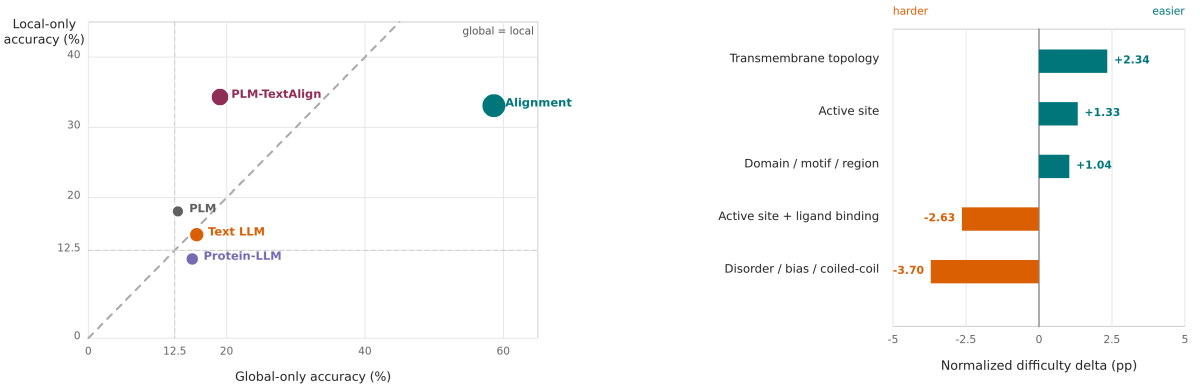


Figure 2: Evidence-source and biology-conditioned diagnostics. Left: Global-only versus Local-only P2T accuracy; the dashed reference is equal global/local performance, and point size tracks Full P2T accuracy. Right: local-evidence families that make the same-global local-evidence discrimination easier or harder after model/split normalization. The vertical zero line is the normalized model/split baseline; positive bars are easier than that baseline and negative bars are harder, measured in percentage points.

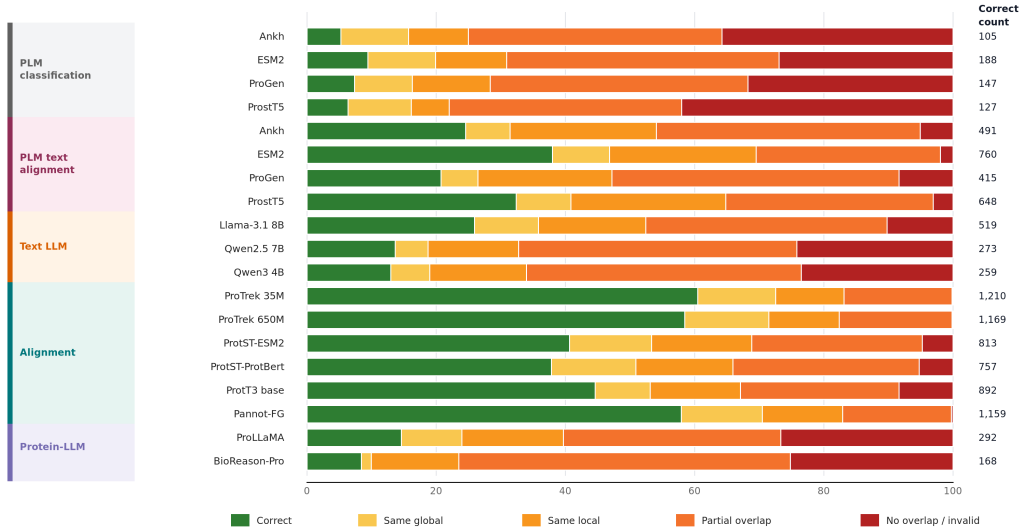


Figure 3: Structured P2T outcomes on `test_id`. The left rail groups rows by model family. Each stacked bar partitions all predictions into correct answers, relation-preserving wrong answers, partial-overlap wrong answers, and no-feature-overlap or invalid answers. The right column reports correct counts out of 2,000 examples.

language modeling — particularly if such methods were applied to sensitive domains such as biology or biotechnology. As such, while misuse of AI in biological contexts is an important concern in general, we do not believe any specific biosecurity or biological misuse risks arise from this focused, abstract, algorithmic study. AI tools were used to assist with writing clarity and code debugging. All scientific content was developed and verified by the authors.

Licenses

PannotGround is constructed from curated UniProtKB/Swiss-Prot records. UniProtKB/Swiss-Prot and UniRef resources are distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Gene Ontology, Rhea, and ChEBI resources used for functional labels, reaction identifiers, and ligand/cofactor normalization are also distributed under CC BY 4.0. If applicable, AlphaFold DB structures are licensed under CC BY 4.0, and RCSB PDB / wwPDB structure data are distributed under the CC0 1.0 Public Domain Dedication. We preserve attribution to all original resources and cite the corresponding database papers.

For model evaluation, we use publicly available pretrained models and follow their respec-

tive licenses: ESM-2 and PubMedBERT/BiomedBERT are MIT-licensed; ProTrek checkpoints are MIT-licensed; Qwen2.5/Qwen3 and BioReason-Pro-SFT are Apache-2.0 licensed; Llama-3.1 follows the Llama 3.1 Community License; ProLLaMA follows the Llama 2 license; Ankh follows CC BY-NC-SA 4.0; ProtT3 follows CC BY-NC 4.0 for models/code; and VenusX, when used for pilot comparison, follows CC BY-NC-ND 4.0. Our released derived artifacts will preserve the attribution requirements and license restrictions of the upstream resources.

References

2025. Uniprot: the universal protein knowledgebase in 2025. *Nucleic acids research*, 53(D1):D609–D617.

Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. 2024. Prot2text: Multimodal protein's function generation with gnns and transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10757–10765.

Matthias Blum, Antonina Andreeva, Laise Cavalcanti Florentino, Sara Rocio Chuguransky, Tiago Grego, Emma Hobbs, Beatriz Lazaro Pinto, Ailsa Orr, Typhaine Paysan-Lafosse, Irina Ponamareva, Gustavo A. Salazar, Nicola Bordin, Peer Bork, Alan Bridge, Lucy Colwell, Julian Gough, Daniel H. Haft, Ivica Letunic, Felipe Llinares-López, and 15 others. 2025. Interpro: the protein sequence classification resource in 2025. *Nucleic Acids Research*, 53(D1):D444–D456.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. 2022. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110.

Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. 2023. Ankh: Optimized protein language model unlocks general-purpose modelling. *Preprint*, arXiv:2301.06568.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, and 1 others. 2022. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127.

Adibvafa Fallahpour and 1 others. 2026. BioReason-Pro: Advancing protein function prediction with multimodal biological reasoning. *bioRxiv*.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2024. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *Preprint*, arXiv:2306.08018.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. 2022. Protgpt2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348.

Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. 2024. Bilingual language model for protein sequence and structure. *Nature Methods*, 21:2288–2298.

Yuheng Hou and 1 others. 2025. MotifAE reveals functional sequence patterns from protein language model: Unsupervised discovery and interpretability analysis. *Preprint*, bioRxiv:2025.11.04.686576.

Cheng-Yu Hsieh, Jieyu Zhang, Ziyang Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, volume 36.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2024. SEED-Bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Zhiheng Liu and 1 others. 2025. ProtSAE: Disentangling and interpreting protein language models via semantically-guided sparse autoencoders. *Preprint*, arXiv:2509.05309.

Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024. ProtT3: Protein-to-text generation for text-based protein understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 5949–5966.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Massimo Hong, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2024. BioMedGPT: Open multimodal generative pre-trained transformer for biomedicine. *IEEE Journal of Biomedical and Health Informatics*.

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. 2025. ProLLaMA: A protein large language model for multi-task protein language processing. *Preprint*, arXiv:2402.16445.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. CREPE: Can vision-language foundation models reason compositionally? *Preprint*, arXiv:2212.07796.

Ali Madani, Ben Krause, Ethan R. Greene, and 1 others. 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41:1099–1106.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. 2023. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, volume 36.

PharMolix. 2025. UniProtQA. Hugging Face dataset.

Owen Queen and 1 others. 2024. ProCyon: A multimodal foundation model for protein phenotypes. *bioRxiv*.

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

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A. Plummer, Ranjay Krishna, and Kate Saenko. 2023. COLA: A benchmark for compositional text-to-image retrieval. In *Advances in Neural Information Processing Systems*, volume 36.

Dingyi Rong, Zijian Chen, Qi Jia, Kaiwei Zhang, Haotian Lu, Guangtao Zhai, and Ning Liu. 2025. LiveProteinBench: A contamination-free benchmark for assessing models' specialized capabilities in protein science. *Preprint*, arXiv:2512.22257.

Yiqing Shen, Zan Chen, Michail Mamalakis, Luhan He, Haiyang Xia, Tianbin Li, Yanzhou Su, Junjun He, and Yu Guang Wang. 2024. A fine-tuning dataset and benchmark for large language models for protein understanding. *Preprint*, arXiv:2406.05540.

Christian J. A. Sigrist, Bertrand A. Cuche, Emmanuel de Castro, Etienne Coudert, Nicolas Redaschi, and Alan Bridge. 2026. The prosite database for protein families, domains, and sites. *Nucleic Acids Research*, 54(D1):D451–D458.

Evan Simon and James Zou. 2025. Interplm: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22:2107–2117.

Jin Su and 1 others. 2025. A trimodal protein language model enables advanced protein searches. *Nature Biotechnology*.

Yang Tan, Wenrui Gou, Bozitao Zhong, Liang Hong, Huiqun Yu, and Bingxin Zhou. 2025. Venusx: Unlocking fine-grained functional understanding of proteins. *Preprint*, arXiv:2505.11812.

Yang Tan, Wenrui Gou, Bozitao Zhong, Huiqun Yu, Liang Hong, and Bingxin Zhou. 2026. VenusX: Unlocking fine-grained functional understanding of proteins. In *International Conference on Learning Representations*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Juntong Wu, Zijing Liu, He Cao, Hao Li, Bin Feng, Zishan Shu, Ke Yu, Li Yuan, and Yu Li. 2025. Rethinking text-based protein understanding: Retrieval or LLM? *Preprint*, arXiv:2505.20354.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. 2023. Protst: Multi-modality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, pages 38749–38767. PMLR.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Naihui Zhou and 1 others. 2019. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20(1):244.

A Split Definitions

Our dataset is curated from UniProtKB, a comprehensive repository of protein sequences and annotations. UniProtKB is divided into two components: Swiss-Prot, which contains manually reviewed and curated protein entries, and TrEMBL, which contains automatically annotated entries. In this work, we use Swiss-Prot exclusively as the source of proteins for training(both pre-training and post-training) and evaluation. Swiss-Prot entries are manually curated by domain experts and are associated with higher-quality functional annotations, including experimentally supported evidence for catalytic activity, binding sites, motifs, and domain organization. In contrast, TrEMBL contains a much larger number of sequences, but many annotations are inferred computationally and may be noisy or inconsistent.

We choose Swiss-Prot for three main reasons. First, our objectives involve fine-grained and evidence-grounded reasoning, which requires reliable residue-level supervision. Second, automatically annotated databases often propagate systematic errors, which can confound both training and evaluation. Third, restricting to curated proteins simplifies interpretation and error analysis, allowing model behavior to be analyzed in terms of known biochemical mechanisms.

A.1 Principles for Dataset Splitting

Protein datasets violate standard i.i.d. assumptions in two distinct ways.

First, a single protein may generate multiple training examples (e.g., multiple prompts or residue-level labels). If such examples are split independently, information can leak across splits. Therefore, all splits must be defined *at the protein level*.

Second, protein sequences exhibit strong homology structure. Many proteins are nearly identical, while others share only remote evolutionary relationships. Naïve random splitting leads to near-duplicate leakage, resulting in artificially inflated performance.

We therefore adopt the following principles:

Principle 1 (Protein atomicity). All examples derived from the same protein accession must belong to the same split.

Principle 2 (Hierarchical homology control). We distinguish between two regimes of sequence similarity: (i) near-duplicate similarity (approximately 90% sequence identity), and (ii) family-level similarity (approximately 30% sequence identity). These represent fundamentally different generalization challenges and must be controlled separately.

Principle 3 (Train once, evaluate many). A single training set is constructed, while multiple evaluation sets are defined with increasing levels of distribution shift.

A.2 Hierarchical Homology-Based Splitting

Let

$$\mathcal{P} = \{p_1, \ldots, p_N\}$$

be the set of curated Swiss-Prot proteins.

We leverage MMseqs2 to implement the following two clustering methods:

- $C_{30}(p)$ assigns protein p to a family-level cluster (proteins with approximately 30% identity in one cluster).
- $C_{90}(p)$ assigns protein p to a near-duplicate cluster (proteins with approximately 90% identity in one cluster).

The goal is to construct five disjoint protein ID sets: Train, Val_{ID}, Test_{ID}, Val_{OOD}, Test_{OOD}.

Step 1: Split family-level clusters. Let

$$\mathcal{F}_{30} = \{C_{30}(p) \mid p \in \mathcal{P}\}$$

be the set of unique family clusters. We randomly partition:

$$\mathcal{F}_{30} = \mathcal{F}_{30}^{\text{train}} \cup \mathcal{F}_{30}^{\text{val}} \cup \mathcal{F}_{30}^{\text{test}}.$$

Step 2: Assign OOD-family proteins. For each protein $p \in \mathcal{P}$:

$$\text{split}(p) = \begin{cases} \text{Val}_{\text{OOD}}, & C_{30}(p) \in \mathcal{F}_{30}^{\text{val}}, \\ \text{Test}_{\text{OOD}}, & C_{30}(p) \in \mathcal{F}_{30}^{\text{test}}, \\ \text{unassigned}, & C_{30}(p) \in \mathcal{F}_{30}^{\text{train}}. \end{cases}$$

Define:

$$\mathcal{P}_{\text{train-fam}} = \{p \in \mathcal{P} \mid C_{30}(p) \in \mathcal{F}_{30}^{\text{train}}\}.$$

Step 3: Split near-duplicate clusters inside training families. Let:

$$\mathcal{F}_{90} = \{C_{90}(p) \mid p \in \mathcal{P}_{\text{train-fam}}\}.$$

We randomly partition:

$$\mathcal{F}_{90} = \mathcal{F}_{90}^{\text{train}} \cup \mathcal{F}_{90}^{\text{val}} \cup \mathcal{F}_{90}^{\text{test}}.$$

Step 4: Assign ID and training proteins. For each protein $p \in \mathcal{P}_{\text{train-fam}}$:

$$\text{split}(p) = \begin{cases} \text{Val}_{\text{ID}}, & C_{90}(p) \in \mathcal{F}_{90}^{\text{val}}, \\ \text{Test}_{\text{ID}}, & C_{90}(p) \in \mathcal{F}_{90}^{\text{test}}, \\ \text{Train}, & C_{90}(p) \in \mathcal{F}_{90}^{\text{train}}. \end{cases}$$

This construction ensures: (i) each protein is assigned to exactly one split, (ii) no family-level cluster (C_{30}) is shared between the training split and the OOD validation or test splits, (iii) no near-duplicate cluster (C_{90}) spans the Train, Val_{ID}, and Test_{ID} splits, and (iv) and the resulting ID and OOD splits correspond to distinct generalization regimes—respectively, within-family generalization and cross-family generalization.

A.3 Condensed Training Set via Family-Balanced Sampling

The full training split contains several hundred thousand proteins, which is computationally expensive for large-scale pretraining. Moreover, the distribution of protein families is highly imbalanced, with a small number of families dominating the dataset.

To construct a more compact yet diverse training set, we apply *within-train sampling*, without modifying any evaluation splits.

Let:

$$\text{Train} = \bigcup_{f \in \mathcal{F}_{30}^{\text{train}}} \mathcal{P}_f,$$

where $\mathcal{P}_f$ denotes the set of training proteins in family f.

We define the condensed training set as:

$$\text{Train}_{\text{condensed}} = \bigcup_{f \in \mathcal{F}_{30}^{\text{train}}} \text{Sample}_k(\mathcal{P}_f),$$

where $\text{Sample}_k(\cdot)$ selects at most k proteins uniformly at random from each family.

Optionally, a global target size N can be enforced by further subsampling across families.

This sampling scheme preserves broad family-level diversity while preventing overrepresentation of large families. In contrast, random subsampling tends to concentrate on well-studied protein families and discard rare functional classes.

Importantly, sampling is applied *only within the training split*. All validation and test splits remain unchanged, and therefore all homology guarantees of the original split protocol are preserved.

B Dataset Construction

PannotGround is constructed to evaluate protein annotation as biological perception. Rather than treating each source annotation field as a separate benchmark task, we organize the data into three perception tasks that ask progressively more localized questions: global semantic annotation (P1), localized feature-type recognition (P2), and residue/span evidence localization (P3). This organization matches the evaluation protocol in Appendix C: the same biological targets can be evaluated either by native PLM readouts or by candidate-selection interfaces for LLMs, protein-LLMs, and converted PLM scores.

B.1 Task Definitions and Biological Motivation

Protein annotation spans multiple biological scales. Some annotations describe the entire protein, such as molecular function, enzyme class, catalytic activity, cofactors, pathway participation, or subcellular localization. Other annotations describe localized sequence evidence, such as domains, motifs, membrane segments, binding sites, active sites, and disordered regions. We turn these complementary sources into three perception tasks:

- **P1: semantic function prediction.** What are the protein-level functions?
- **P2: feature-type recognition.** What fine-grained features exist in the protein?
- **P3: residue evidence localization.** Where are those fine-grained fetures?

The previous source categories, such as GO, EC, catalytic activity, regions, and sites, are therefore not separate benchmark tasks in the current design. They are annotation families used to instantiate P1, P2, and P3. This task-centered grouping makes the benchmark easier to interpret: each result table answers one aspect of perception ability, while subtasks expose which annotation families or feature subtypes contributed to that aspect.

B.2 Content in tasks

Protein understanding requires inferring biologically meaningful information across multiple levels of abstraction—from global functional classification to fine-grained residue-level annotations. We define eight tasks organized into two complementary levels of granularity, collectively capturing the hierarchical nature of protein biology.

Table 7: Current PannotGround perception tasks and their source annotation families. Source fields are grouped by biological target rather than reported as independent benchmark tasks.

Task	Biological perception target	Source annotation families	Gold object	Main evaluation use
P1	Global semantic annotation supported by the whole sequence	GO biological process, GO molecular function, GO cellular component, EC number, catalytic activity, cofactor, subcellular location, pathway	Protein-level label set by namespace	Semantic label prediction
P2	Presence of localized biological feature types	Curated site and region annotations, canonicalized into a shared feature-type vocabulary	Set of feature types present in the protein	Feature-type recognition
P3	Residue or span evidence for a queried localized feature	Curated site and region annotations with valid sequence coordinates	One or more residue spans matching the query	Evidence localization

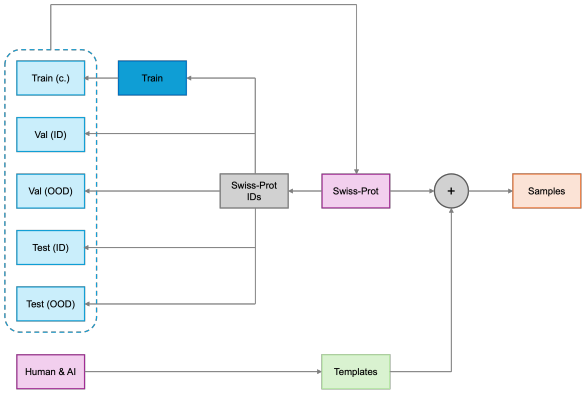


Figure 4: End-to-end dataset construction pipeline, from accession-level splitting to template-based sample instantiation.

Protein-level tasks. These tasks assign global annotations defined over entire protein sequences:

- **Gene Ontology (GO):** Controlled vocabulary terms from the Gene Ontology Consortium, partitioned into three sub-ontologies: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC).
- **EC Number:** Enzyme Commission classification providing a four-level hierarchical encoding of enzymatic catalytic function.
- **Catalytic Activity:** The specific biochemical reaction(s) catalyzed by the protein.
- **Cofactor:** Metal ions or small organic molecules required for enzymatic activity.
- **Subcellular Location:** Cellular compartments in which the protein resides, including membrane topology and orientation when available.
- **Pathway:** Biological pathways in which the protein participates, with hierarchical pathway organization and step-level roles.

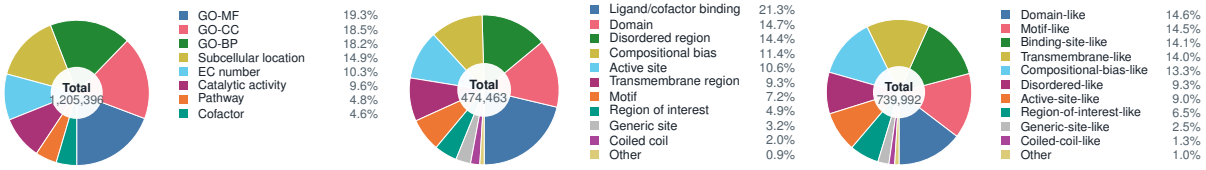
Residue-level tasks. These tasks provide localized annotations tied to specific sequence positions or segments:

- **Regions:** Contiguous sequence segments defining structural, topological, or functional domains (13 feature types; see Table 8).
- **Sites:** Individual residues or short spans associated with specific functional roles such as catalysis or ligand binding (4 feature types; see Table 8).

These eight tasks were selected to satisfy three criteria simultaneously: (i) they collectively span the major axes of protein biology (function, localization, interaction partners, and local sequence features); (ii) each task admits a learnable mapping from the protein sequence itself, rather than depending on extrinsic context; and (iii) they exhibit minimal semantic redundancy with one another while remaining complementary.

Figure 5 reports the subtask composition used in the benchmark appendix. For P1, the denominator is the number of positive protein-namespace pairs, because one protein-level P1 record may contain annotations from multiple namespaces. For P2, the denominator is the number of positive protein-feature pairs; although the task is encoded as a single multi-label feature-set prediction problem, per-feature scores are the natural subtask analysis. For P3, the denominator is the number of query-conditioned grounding rows, grouped by the frozen evidence-subtype map.

The largest P1 namespace subtasks are GO-MF, GO-CC, GO-BP, and subcellular location. The largest P2 feature-type subtasks are ligand/cofactor binding, domain, disordered region, compositional bias, active site, transmembrane region, and mo-



(a) P1 namespace subtasks. (b) P2 feature-type positives. (c) P3 evidence-subtype tasks.

Figure 5: Subtask composition of PannotGround. P1 is counted as positive protein-namespace pairs. P2 is decomposed by positive canonical feature type inside the multi-label feature-set task. P3 is counted by query-conditioned grounding rows using the frozen evidence-subtype map.

tif. The largest P3 evidence subtypes are domain-like, motif-like, binding-site-like, transmembrane-like, compositional-bias-like, disordered-like, and active-site-like. This distribution reflects the biological annotation density of the curated source tables and should be considered when interpreting macro-averaged or per-subtask results.

B.3 Data Source and Field Selection Rationale

We source all annotations from UniProtKB/Swiss-Prot[1], which provides expert-curated protein records in structured JSON format. Each record contains three categories of information:

1. **Entry-level metadata** (accession identifiers, versioning, curation status)—encodes database bookkeeping with no biological signal.
2. **Biological metadata** (source organism, gene names, taxonomic lineage)—reflects protein provenance. While biologically relevant, these attributes primarily indicate *where* a protein originates rather than *what it does*. Using them as prediction targets would encourage taxonomic memorization over functional reasoning, since homologous proteins across species often share identical functions.
3. **Functional semantic annotations** (cross-references, comments, features, protein description)—directly describe intrinsic, sequence-encoded biochemical properties.

All tasks are derived exclusively from the third category. Within this category, we further apply the following selection principles:

- **Semantic control:** We retain only annotation types expressed in controlled or semi-controlled vocabularies, ensuring consistent label semantics across proteins.
- **Intrinsic learnability:** We include only annotations that reflect properties encoded in the protein sequence itself, excluding those contingent on experimental conditions, organism-specific phenotypes, or post-translational processing states.
- **Non-redundancy:** We exclude annotation types that aggregate information across multiple tasks. For instance, the `FUNCTION` comment type provides free-text summaries that often recapitulate catalytic activity, pathway, and GO information, risking shortcut learning. Similarly, the `keywords` field offers coarse-grained descriptors that substantially overlap with GO terms.

Table 8 provides the complete mapping between tasks and their UniProtKB source fields.

B.4 Data Construction Pipeline

The construction pipeline has four stages.

Protein normalization and split assignment. Each protein is represented by a stable record identifier, a sequence, sequence length, and C30/C90 cluster metadata. Each record is assigned to exactly one split. Task-view rows refer to the canonical protein table through `record_id` or `sequence_ref`, so model runners can join in the sequence only when needed.

Annotation normalization. Protein-level annotations are normalized within their source namespace. GO annotations are separated into biological process, molecular function, and cellular component namespaces; EC annotations are split into individual labels; and catalytic activity, cofactor, subcellular localization, and pathway annotations are retained as namespace-specific labels after text normalization. Local site and region annotations are mapped to a canonical feature-type vocabulary.

Deduplication and coordinate validation. Repeated protein-level labels are deduplicated within

[1]We use UniProtKB release 2024_05. All entries are restricted to the reviewed (Swiss-Prot) partition to ensure manual curation quality.

Task	Source Field	Filter Key	Granularity	Samples/Entry
Gene Ontology	`uniProtKBCrossReferences`	`database="GO"`	Protein	≤ 3
EC Number	`proteinDescription`	—	Protein	≥ 1
Catalytic Activity	`comments`	`commentType="CATALYTIC ACTIVITY"`	Protein	≥ 1
Cofactor	`comments`	`commentType="COFACTOR"`	Protein	≥ 1
Subcellular Location	`comments`	`commentType="SUBCELLULAR LOCATION"`	Protein	≥ 1
Pathway	`comments`	`commentType="PATHWAY"`	Protein	≥ 1
Regions	`features`	`type` $\in \mathcal{R}$	Residue	≥ 1
Sites	`features`	`type` $\in \mathcal{S}$	Residue	≥ 1

Table 8: Mapping between tasks and UniProtKB source fields. $\mathcal{R}$ = {Topological domain, Transmembrane, Intramembrane, Domain, Repeat, Calcium binding, Zinc finger, DNA binding, Nucleotide binding, Region, Coiled coil, Motif, Compositional bias}. $\mathcal{S}$ = {Active site, Metal binding, Binding site, Site}. The "Samples/Entry" column indicates how many samples a single protein can yield for a given task.

protein and namespace. For local annotations, the benchmark retains a row for grounding only when the start and end coordinates are numeric, ordered, and lie inside the protein sequence. Repeated local annotations with the same protein, source family, canonical type, and feature detail are merged into one P3 query with all matching gold spans.

Task-view instantiation. P1 groups normalized protein-level labels by namespace. P2 converts valid site/region annotations for each protein into a set of canonical feature types. P3 creates query-conditioned localization records in which the input query specifies the feature family, canonical type, and feature detail, and the gold answer is the corresponding span set. This construction separates global semantic recognition, local feature recognition, and residue-level grounding while preserving the connection between feature existence and feature coordinates.

B.5 General usage

Overall, PannotGround contains 282,021 unique proteins and 3,666,013 deduplicated curated annotations. The constructed task views contain 268,589 P1 records, 218,434 P2 records, and 739,992 P3 localization records. Residue-level localization contributes the largest number of records because a single protein can contain multiple queried local features and multiple spans per query. This imbalance is intentional and reflects the biological multiplicity of localized annotations; evaluation tables should therefore report P1, P2, and P3 separately rather than relying only on an aggregate score. Details are included in the Table 9.

All native task views are compact records that reference a protein sequence by identifier. A model runner should first load `canonical/proteins.jsonl.gz` as a map from `record_id` to sequence, then join the appropriate task view on `record_id` or `sequence_ref`. This design avoids duplicating long sequences across P1, P2, and P3.

Using P1. P1 is a multi-label semantic prediction task over namespaces. For native PLM evaluation, a model can use a pooled sequence representation with namespace-specific label heads. The `gold.labels_by_namespace` field contains the positive labels for each namespace present in the record. Missing labels should be treated as unobserved in the dataset construction sense; they are treated as negatives only inside an explicitly closed evaluation universe. For compact main-table reporting, a P1-core setting can focus on structured namespaces such as EC and GO, while P1-all includes all available global namespaces.

Using P2. P2 is a multi-label feature-type recognition task. The input is the full protein sequence, and the target is the set of canonical local feature types present anywhere in the protein. P2 is appropriate for pooled or sequence-level readouts because the gold object removes coordinate requirements while still testing whether the model detects local biological feature classes. Per-feature metrics should be reported because the feature distribution is imbalanced.

Using P3. P3 is a query-conditioned grounding task. The input includes the protein sequence and a structured query containing the feature family, canonical type, and feature detail. The target is the set of matching spans. A native PLM baseline can condition a token or span head on the query and score residues or candidate spans. If multiple spans match the same query in one protein, the record contains all matching spans; models should therefore support multi-span targets rather than as-

Table 9: PannotGround v0 task-view statistics by split. P1, P2, and P3 counts refer to constructed perception-task records, not raw source rows.

Split	Proteins	P1 records	P2 records	P3 records
`train_condensed`	127,619	119,614	100,015	350,177
`val_id`	22,587	21,834	16,951	56,439
`test_id`	22,014	21,255	16,940	57,780
`val_ood_family`	55,414	53,540	42,005	138,610
`test_ood_family`	54,387	52,346	42,523	136,986
Total	282,021	268,589	218,434	739,992

Table 10: Curation summary for the v0 source annotations. Quality tiers reflect how directly an annotation supports sequence-based perception in the current benchmark.

Item	Count	Interpretation
Deduplicated annotations	3,666,013	Normalized source annotations retained in the canonical audit
High-quality annotations	1,478,169	Valid localized site/region annotations with non-generic feature types
Medium-quality annotations	1,680,917	Structured protein-level labels, primarily GO and EC-derived labels
Weak annotations	506,925	Usable but less structured labels or generic localized feature types
Excluded annotations	2	Localized annotations rejected because coordinates are invalid

suming exactly one site.

Using candidate-selection views. For LLMs, protein-LLMs, and cross-family comparisons, the rendered QA views provide fixed candidate choices and matched controls. P1 and P3 are rendered as single-choice tasks, while P2 is rendered as a multi-answer feature-set task. These views are useful when comparing model families with different native interfaces because every model is scored against the same answer space.

C Evaluation Protocol and Metric Definitions

PannotGround evaluates two related capabilities. The first is attribute-level recognition: whether a model can recover global protein attributes (P1), local feature-type attributes (P2), and residue- or span-grounded local evidence (P3). The second is compositional protein-text matching: whether attributes from these levels are bound to the same source protein in protein-to-text (P2T) and text-to-protein (T2P) retrieval. We therefore separate PLM classification, MCQA, and compositional retrieval. PLM classification measures what a sequence encoder can recover through task-specific heads. MCQA maps P1/P2/P3 decisions to fixed candidate sets so that text LLMs, protein-LLMs, and PLM adapters share the same answer space. P2T and T2P then test whether the recognized attributes compose into a protein-specific text description.

In addition to the primary scores, we report controls and interventions that test whether a score is grounded in the protein input. Input controls measure dependence on sequence content, residue order, and answer order; masked-evidence interventions test whether P3 decisions depend on the annotated residue evidence.

The appendix is organized as follows. We first define the evaluation tracks and their comparability scope. We then give metric definitions for PLM classification, MCQA, and compositional retrieval. The final sections define grounding interventions, input controls, aggregate scores, and interface diagnostics.

C.1 Evaluation Tracks

Let x_i denote the protein input for example i, and let $a_i \in \{\mathrm{P1}, \mathrm{P2}, \mathrm{P3}\}$ denote the attribute level. The attribute-level tracks are

$$\mathrm{P1\text{-}cls},\ \mathrm{P2\text{-}cls},\ \mathrm{P3\text{-}cls} \quad \text{for PLM classification,}$$

$$\mathrm{P1\text{-}mcqa},\ \mathrm{P2\text{-}mcqa},\ \mathrm{P3\text{-}mcqa} \quad \text{for MCQA.}$$

The compositional tracks are P2T and T2P retrieval.

In Table 12, PLM adapter rows denote PLM scores converted to the same fixed candidate choices used for LLMs and protein-LLMs.

C.2 Reporting Scope

The tracks answer different benchmark questions and should be reported as separate result families.

Table 11: PannotGround attribute levels. Each level defines one biological target; evaluation tracks are listed in Table 12.

Level	Biological target	Prediction target
P1	Global semantic function prediction: identify the protein-level annotation supported by the sequence.	Protein-level annotation label
P2	Feature-type recognition: identify which localized feature types are present in the protein.	Set of localized feature types
P3	Residue evidence localization: select the residue or span evidence supporting a queried feature.	Residue set, span set, or candidate span

Table 12: Evaluation tracks and primary metrics. P1/P2/P3 are attribute levels; P2T and T2P are compositional retrieval directions.

Track	Model family	Evaluation record	Prediction object	Primary metric
P1-cls	PLM encoder	Sequence with annotation label universe	Scores over annotation labels	Top-k-by-gold micro-F1; mean Jaccard secondary
P1-mcqa	LLM, protein-LLM, or PLM adapter	Candidate annotation question	One selected annotation	Accuracy; MRR when rankings are available
P2-cls	PLM encoder	Sequence with feature-type label set	Scores over feature-type labels	Top-k-by-gold micro-F1; mean Jaccard secondary
P2-mcqa	LLM, protein-LLM, or PLM adapter	Candidate feature-type question	Selected feature-type set	Mean set F1; Jaccard and exact match secondary
P3-cls	PLM encoder	Sequence with queried feature and residue/span target	Residue or span scores	Top-k-by-gold residue F1; top-k mean best IoU secondary
P3-mcqa	LLM, protein-LLM, or PLM adapter	Candidate-span question	One selected span	Candidate-span accuracy; MRR when rankings are available
P2T	Protein-text or LLM-style scorer	Protein query with 8 candidate captions	Ranked captions	Accuracy / R@1; relation diagnostics
T2P	Protein-text scorer	Text query with candidate proteins	Ranked proteins	R@1, R@5, MRR, mean rank

- **PLM classification results (P1-cls, P2-cls, P3-cls)** evaluate how much annotation or residue information is accessible from a sequence encoder through a task-specific classification head. These are the primary encoder-only attribute-level baselines.
- **MCQA results (P1-mcqa, P2-mcqa, P3-mcqa)** evaluate whether a model selects the correct candidate answer for the same biological target. Text-only LLMs and protein-LLMs are evaluated only in this track. PLMs can also be included in this track when 0-shot scores or fine-tuned scores are converted into candidate choices.
- **Compositional retrieval results (P2T, T2P)** evaluate whether global and local attributes are bound to the same protein in a text description. These results should be interpreted separately from isolated P1/P2/P3 recognition.

Thus, cross-family P1/P2/P3 claims should be made within the MCQA track. PLM classification metrics provide complementary evidence about encoder capacity, but they are not the same measurement as MCQA accuracy or set F1.

C.3 Label and MCQA Candidate Assumptions

PannotGround is built from curated positive annotations. For P1-cls and P2-cls, a closed label universe is defined within each reported namespace or feature-type set, and labels absent from the gold set are treated as negative for that specific evaluation. This convention is needed to compute multi-label F1 and ranking metrics, but it should be interpreted as a benchmark evaluation assumption rather than proof that unobserved annotations are biologically false.

For P1-mcqa and P3-mcqa, each example has one gold candidate. For P2-mcqa, each example has a nonempty gold candidate set. Candidate sets are fixed before model evaluation; models are scored only on the final selected candidate or candidate set. Candidate ranking metrics are reported only for models that expose scores for all candidates.

The same candidate set is used for all model

families within a reported example, so differences between text-only LLMs, protein-LLMs, and PLM adapters are not caused by different answer spaces. MCQA results are reported with input controls when available, including no-sequence, sequence-shuffle, and choice-order controls, to distinguish biological signal from candidate priors or option-order artifacts. For single-choice candidate tasks, candidate-set size is reported as a diagnostic; the corresponding random choice baseline is

$$\text{Chance} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C_i|}.$$

C.4 PLM Classification Metrics

PLM classification evaluations use task-specific fine-tuned heads or 0-shot scores over encoder representations. They are reported as encoder-only attribute-level metrics.

P1-cls and P2-cls: top-k label metrics. For example i, let $\mathcal{L}$ be the task label universe, $Y_i \subseteq \mathcal{L}$ the gold label set, and $s_{i\ell}$ the model score for label ℓ. We set $k_i = |Y_i|$ and define the top-k prediction as

$$\widehat{Y}_i^k = \text{TopK}_{\ell\in\mathcal{L}}(s_{i\ell}, k_i). \tag{1}$$

Aggregated true positives, false positives, and false negatives are

$$\text{TP}_k = \sum_i |\widehat{Y}_i^k \cap Y_i|, \tag{2}$$

$$\text{FP}_k = \sum_i |\widehat{Y}_i^k \setminus Y_i|, \tag{3}$$

$$\text{FN}_k = \sum_i |Y_i \setminus \widehat{Y}_i^k|. \tag{4}$$

The primary P1-cls and P2-cls metric is top-k-by-gold micro-F1:

$$\text{TopK-MicroF1} = \frac{2\text{TP}_k}{2\text{TP}_k + \text{FP}_k + \text{FN}_k}. \tag{5}$$

The secondary metric is top-k-by-gold mean Jaccard:

$$\text{TopK-MeanJaccard} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\widehat{Y}_i^k \cap Y_i|}{|\widehat{Y}_i^k \cup Y_i|}. \tag{6}$$

Exact set match under the same top-k rule is

$$\text{TopK-Exact} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}[\widehat{Y}_i^k = Y_i]. \tag{7}$$

When thresholded diagnostics are reported, the prediction is

$$\widehat{Y}_i(\tau) = \{\ell \in \mathcal{L} : s_{i\ell} \geq \tau\}, \tag{8}$$

where τ is selected on validation data only. Micro-F1 and macro-F1 at τ are diagnostic because sparse annotation namespaces can make a single global threshold unstable. Ranking diagnostics such as AUPRC and label-ranking average precision may also be reported, but they are not the main table metric.

P3-cls: top-k residue and span metrics. For example i, let $G_i \subseteq \{1, \ldots, L_i\}$ be the gold residue set and r_{ij} the model score for residue j. We set $k_i = |G_i|$ and define

$$\widehat{G}_i^k = \text{TopK}_{j\in\{1,\ldots,L_i\}}(r_{ij}, k_i). \tag{9}$$

Aggregated residue counts are

$$\text{TP}_k^{\text{res}} = \sum_i |\widehat{G}_i^k \cap G_i|, \tag{10}$$

$$\text{FP}_k^{\text{res}} = \sum_i |\widehat{G}_i^k \setminus G_i|, \tag{11}$$

$$\text{FN}_k^{\text{res}} = \sum_i |G_i \setminus \widehat{G}_i^k|. \tag{12}$$

The primary P3-cls metric is top-k-by-gold residue F1:

$$\text{TopK-ResidueF1} = \frac{2\text{TP}_k^{\text{res}}}{2\text{TP}_k^{\text{res}} + \text{FP}_k^{\text{res}} + \text{FN}_k^{\text{res}}}. \tag{13}$$

For span diagnostics, let S_i be the gold span set and $\widehat{S}_i$ the predicted span set obtained from the predicted residue mask. For spans $a = [a_1, a_2]$ and $b = [b_1, b_2]$,

$$\text{IoU}(a, b) = \frac{|a \cap b|}{|a \cup b|}. \tag{14}$$

The top-k mean best IoU used for P3-cls is

$$\text{MeanBestIoU} = \frac{1}{\sum_i |S_i|}\sum_i \sum_{g\in S_i} \max_{\hat{s}\in\widehat{S}_i} \text{IoU}(g, \hat{s}), \tag{15}$$

with the maximum defined as zero when no span is predicted.

C.5 MCQA Metrics

MCQA evaluations score the final answer over a fixed candidate set. For example i, let

$$e_i = (x_i, q_i, C_i, y_i, s_i, r_i),$$

where q_i is the query, C_i is the candidate set, y_i is the gold answer or gold answer set, s_i is the split, and r_i is the control condition.

Rationale for MCQA. MCQA is used for LLM and protein-LLM evaluation because open-ended protein annotation introduces synonym, formatting, and ontology-name normalization ambiguities that can dominate the score. A fixed candidate set keeps the biological question unchanged while making the output space explicit: the model must recognize the annotation, feature type, or evidence span that is best supported by the protein input. This also gives text-only LLMs, protein-LLMs, and PLMs converted to candidate scores the same prediction object and denominator. Thus, MCQA measures controlled recognition and grounding among plausible alternatives; it should not be interpreted as an open-ended annotation-discovery metric.

MCQA scoring adapters. MCQA rows may be produced by different model interfaces, but the scored object is always the same final answer. For models that directly produce a candidate answer, $\hat{y}_i$ or $\widehat{Y}_i$ is parsed from the final response. For models that expose candidate scores, the single-choice prediction for P1-mcqa and P3-mcqa is

$$\hat{y}_i = \arg\max_{c \in C_i} u_{ic}. \tag{16}$$

For likelihood-based protein-LLMs, u_{ic} may be a length-normalized conditional log-likelihood,

$$u_{ic} = \frac{1}{|c|} \sum_{m=1}^{|c|} \log p_\theta(c_m \mid x_i, q_i, c_{<m}). \tag{17}$$

For P2-mcqa likelihood scoring, each candidate c can be reduced to a binary decision score,

$$d_{ic} = \log p_\theta(\text{yes} \mid x_i, q_i, c) - \log p_\theta(\text{no} \mid x_i, q_i, c), \tag{18}$$

and the predicted set is

$$\widehat{Y}_i(\tau) = \{c \in C_i : d_{ic} \geq \tau\}. \tag{19}$$

The threshold τ is selected on validation data and then fixed for all test splits.

Two PLM MCQA adapters are used in cross-family comparisons. **PLM 0-shot** maps frozen PLM representations to candidate scores without supervised training on the MCQA examples; in our PLM rows, candidates are ranked by similarity to label or evidence reference vectors computed from training embeddings. **PLM fine-tune** trains a lightweight candidate-scoring head on the training split and then applies the same MCQA metric as the other model families. In these rows, "adapter" refers only to the scoring interface that maps PLM outputs to fixed candidates; the reported prediction object and denominator are the same as the corresponding LLM and protein-LLM MCQA rows.

P1-mcqa. P1-mcqa selects one annotation candidate. If the model assigns scores u_{ic} to candidates $c \in C_i$, the prediction is

$$\hat{y}_i = \arg\max_{c \in C_i} u_{ic}. \tag{20}$$

The primary metric is accuracy:

$$\mathrm{Acc}_{\mathrm{P1}} = \frac{1}{N_{\mathrm{P1}}} \sum_{i=1}^{N_{\mathrm{P1}}} \mathbb{1}[\hat{y}_i = y_i]. \tag{21}$$

When a complete candidate ranking is available, we report

$$\mathrm{MRR}_{\mathrm{P1}} = \frac{1}{N_{\mathrm{P1}}} \sum_{i=1}^{N_{\mathrm{P1}}} \frac{1}{\mathrm{rank}_i(y_i)}. \tag{22}$$

P2-mcqa. P2-mcqa predicts a subset of feature-type candidates. Let $Y_i \subseteq C_i$ be the gold feature set and $\widehat{Y}_i \subseteq C_i$ the prediction. Per-example precision, recall, set F1, and Jaccard are

$$P_i = \frac{|\widehat{Y}_i \cap Y_i|}{|\widehat{Y}_i|}, \tag{23}$$

$$R_i = \frac{|\widehat{Y}_i \cap Y_i|}{|Y_i|}, \tag{24}$$

$$F_i = \frac{2P_iR_i}{P_i + R_i}, \tag{25}$$

$$J_i = \frac{|\widehat{Y}_i \cap Y_i|}{|\widehat{Y}_i \cup Y_i|}. \tag{26}$$

Undefined precision, recall, or F1 values are set to zero. The primary metric is mean set F1:

$$\mathrm{SetF1}_{\mathrm{P2}} = \frac{1}{N_{\mathrm{P2}}} \sum_{i=1}^{N_{\mathrm{P2}}} F_i. \tag{27}$$

Secondary metrics are mean Jaccard

$$\mathrm{MeanJaccard}_{\mathrm{P2}} = \frac{1}{N_{\mathrm{P2}}} \sum_{i=1}^{N_{\mathrm{P2}}} J_i \tag{28}$$

and exact set match

$$\mathrm{Exact}_{\mathrm{P2}} = \frac{1}{N_{\mathrm{P2}}} \sum_{i=1}^{N_{\mathrm{P2}}} \mathbb{1}[\widehat{Y}_i = Y_i]. \tag{29}$$

For likelihood-based protein-LLMs, the selected set may be formed by thresholding candidate-wise yes/no scores. The threshold is calibrated on validation data and then fixed for test evaluation.

P3-mcqa. P3-mcqa selects one evidence-span candidate. With candidate scores u_{ic},

$$\hat{y}_i = \arg\max_{c \in C_i} u_{ic}. \tag{30}$$

The primary metric is candidate-span accuracy:

$$\mathrm{Acc}_{\mathrm{P3}} = \frac{1}{N_{\mathrm{P3}}} \sum_{i=1}^{N_{\mathrm{P3}}} \mathbb{1}[\hat{y}_i = y_i]. \tag{31}$$

MRR is reported when the full candidate ranking is available.

C.6 Compositional Retrieval Metrics

P2T and T2P evaluate composition rather than isolated P1/P2/P3 recognition. Each example is built from a source protein and a compositional description whose global and local attributes should refer to that same protein. The candidate pools are constructed with relation-aware hard negatives, so the selected wrong answer can be interpreted by its relation to the gold example.

P2T retrieval. In P2T, the query is a protein sequence x_i and the candidate set C_i contains $K = 8$ text descriptions. Exactly one description $y_i \in C_i$ matches the source protein. A scoring model assigns u_{ic} to each caption $c \in C_i$, and the predicted caption is

$$\hat{y}_i = \arg\max_{c \in C_i} u_{ic}. \tag{32}$$

The primary metric is 8-way accuracy, equivalently R@1:

$$\mathrm{Acc}_{\mathrm{P2T}} = \frac{1}{N_{\mathrm{P2T}}} \sum_{i=1}^{N_{\mathrm{P2T}}} \mathbb{1}[\hat{y}_i = y_i]. \tag{33}$$

Prompted models that return an answer option are scored by the parsed final choice. Likelihood-based systems and protein-text alignment systems are scored by ranking all candidate captions. The random baseline is $1/K = 12.5\%$.

For wrong P2T predictions, we report the known relation $\rho_i(\hat{y}_i)$ between the selected caption and the gold caption. Relations distinguish errors that preserve global function but change local evidence, preserve local evidence but change global function, preserve only partial attributes, or share no selected global/local signature. These diagnostics are not separate task scores; they explain whether an error is a biologically close near miss or an unrelated shortcut.

T2P retrieval. In T2P, the query is a compositional text description q_i and the candidate set C_i contains proteins. The controlled setting uses $|C_i| = 64$, and full-pool retrieval ranks the source protein against the canonical protein pool. A model assigns scores u_{ic} to candidate proteins and induces a rank $\mathrm{rank}_i(y_i)$ for the gold protein y_i. Recall at k is

$$\mathrm{R@}k = \frac{1}{N_{\mathrm{T2P}}} \sum_{i=1}^{N_{\mathrm{T2P}}} \mathbb{1}[\mathrm{rank}_i(y_i) \le k], \tag{34}$$

and mean reciprocal rank is

$$\mathrm{MRR} = \frac{1}{N_{\mathrm{T2P}}} \sum_{i=1}^{N_{\mathrm{T2P}}} \frac{1}{\mathrm{rank}_i(y_i)}. \tag{35}$$

We also report mean rank,

$$\mathrm{MeanRank} = \frac{1}{N_{\mathrm{T2P}}} \sum_{i=1}^{N_{\mathrm{T2P}}} \mathrm{rank}_i(y_i). \tag{36}$$

For controlled 64-way T2P, the random R@1 baseline is $1/64 = 1.56\%$. For full-pool T2P, the random baseline is the reciprocal of the pool size. When R@1 fails, the top-ranked wrong protein is assigned its known relation-aware error type when available.

Controlled compositional variants. Attr2, Attr4, Global-only, and Local-only are data projections, not new metrics. Attr2 and Attr4 fix the number of displayed attributes. Global-only keeps only global P1 information, and Local-only keeps only local P2/P3 information. They use the same P2T or T2P scoring rules as the full compositional setting and are reported to separate attribute-count effects from global/local evidence effects.

C.7 Masked-Evidence Intervention Tests

P3 candidate-span accuracy measures whether the model ranks the correct evidence span above decoys. The masked-evidence intervention tests ask a different question: whether the model's score for the queried biological statement depends on the annotated evidence span. These tests are therefore reported as grounding diagnostics, not as replacements for P1/P2/P3 primary metrics.

For example i, let g_i denote the gold evidence span, and let $f_\theta(x_i, q_i, g_i)$ be the model score for

the queried biological statement under a specified sequence mask. We compute three scores:

$$s_{\text{full}}(i) = f_\theta(x_i, q_i, g_i), \tag{37}$$

$$s_{\text{bg}}(i) = f_\theta(x_i \setminus g_i, q_i, g_i), \tag{38}$$

$$s_{\text{local}}(i) = f_\theta(x_i[g_i], q_i, g_i). \tag{39}$$

Here $x_i \setminus g_i$ denotes the background sequence with the annotated evidence span removed or masked, and $x_i[g_i]$ denotes the local sequence view that keeps the annotated span while masking the remaining residues. The two intervention scores are

$$\text{Nec}(i) = s_{\text{full}}(i) - s_{\text{bg}}(i), \tag{40}$$

$$\text{Suf}(i) = s_{\text{local}}(i) - s_{\text{bg}}(i). \tag{41}$$

$\text{Nec}(i) > 0$ indicates that removing the annotated evidence weakens the model's support for the queried statement. $\text{Suf}(i) > 0$ indicates that the local evidence span alone provides more support than the background without that span. A high necessity value with a lower sufficiency value should be interpreted as evidence that the annotated span is important but not fully self-contained; surrounding sequence context may still be required.

For a group G, such as a split or site subtype, we report the mean intervention scores

$$\overline{\text{Nec}}_G = \frac{1}{|G|} \sum_{i \in G} \text{Nec}(i), \tag{42}$$

$$\overline{\text{Suf}}_G = \frac{1}{|G|} \sum_{i \in G} \text{Suf}(i). \tag{43}$$

Because these quantities are score differences, they should be compared within the same scoring model and evaluation protocol.

C.8 Input Controls and Aggregate Scores

All reported tables distinguish in-distribution (ID) and family-level out-of-distribution (OOD) test splits. For MCQA evaluations, we additionally report input controls when available. Let $M(r)$ denote the primary score under control condition r. The main condition is $r = \text{Full}$, where the original protein input and candidate set are provided. The NoSeq control removes the protein sequence. The SeqShuffle control preserves amino-acid composition but shuffles residue order. The ChoiceShuffle control permutes answer-choice order. The sequence-use and order-use gaps are

$$\text{SeqGap} = M(\text{Full}) - M(\text{NoSeq}), \tag{44}$$

$$\text{OrderGap} = M(\text{Full}) - M(\text{SeqShuffle}). \tag{45}$$

SeqGap measures the gain from providing protein input. OrderGap measures the gain from preserving amino-acid order rather than only composition. Choice shuffle is used to check option-order sensitivity.

For MCQA summary tables, one primary score is first computed per attribute level:

$$M_{\text{P1}} = \text{Acc}_{\text{P1}}, \quad M_{\text{P2}} = \text{SetF1}_{\text{P2}}, \quad M_{\text{P3}} = \text{Acc}_{\text{P3}}.$$

The aggregate score is an example-count-weighted average:

$$\text{Full} = \frac{N_{\text{P1}} M_{\text{P1}} + N_{\text{P2}} M_{\text{P2}} + N_{\text{P3}} M_{\text{P3}}}{N_{\text{P1}} + N_{\text{P2}} + N_{\text{P3}}}. \tag{46}$$

The same aggregation is used for control conditions. Because the aggregate can hide task-specific behavior, main result tables should report the P1, P2, and P3 scores next to the aggregate.

C.9 Interface Diagnostics

For prompt-following or generation-based systems, format compliance is reported separately from the biological score. Let v_i, b_i, and z_i denote whether example i has a valid answer, an empty prediction, or a runtime or parser error. We report

$$\text{ValidAnswerRate} = \frac{1}{N} \sum_i v_i, \tag{47}$$

$$\text{EmptyPredictionRate} = \frac{1}{N} \sum_i b_i, \tag{48}$$

$$\text{ErrorRate} = \frac{1}{N} \sum_i z_i. \tag{49}$$

These diagnostics should not be averaged into the biological score. They explain whether a low MCQA score reflects wrong biological choices, answer-format failure, or runtime failure. Invalid, empty, or errored MCQA outputs remain in the task denominator and receive zero credit for the corresponding task metric.

C.10 Comparability

PLM classification rows, MCQA rows, and compositional retrieval rows answer related but distinct questions. P1-cls, P2-cls, and P3-cls measure how well an encoder head can recover the target labels or residues. P1-mcqa, P2-mcqa, and P3-mcqa measure whether a model can select the correct answer from a fixed candidate set. P2T and T2P measure whether global and local attributes are bound to the same protein in a compositional retrieval setting.

Cross-family P1/P2/P3 claims should therefore be made within the MCQA track, while PLM classification metrics should be reported as encoder-only baselines. Architecture-specific diagnostic analyses, such as Q-former soft-query geometry or evidence-generation prompt probes, are useful for explaining model behavior but should not be merged into the main P1/P2/P3 benchmark score unless they are evaluated with the same examples, splits, controls, and metrics.

D Individual Task Results

This appendix reports individual P1/P2/P3 results in the same attribute-level order used throughout the benchmark: P1 global semantic annotation, P2 feature-type recognition, and P3 residue evidence localization. We separate PLM classification from MCQA because they answer different questions. PLM classification measures what a protein encoder can recover from sequence representations with task-specific heads. MCQA measures whether models can choose the correct annotation, feature type, or evidence span from a fixed candidate set. Diagnostic results are then grouped by the question they answer: input controls test sequence dependence, masked-evidence interventions test residue grounding, and model-specific analyses are reported as diagnostics rather than as a primary benchmark leaderboard.

Reading this appendix. Table 13 summarizes the result families before the detailed tables. The central distinction is between attribute recovery from sequence encoders, controlled answer selection under MCQA, and diagnostic analyses that explain when scores depend on protein input or model interface behavior.

D.1 Attribute-Level Results

We first report the two main individual-result families. PLM classification tests how much P1/P2/P3 information is recoverable from sequence encoders under task-specific heads. MCQA then evaluates the same biological targets under a shared fixed-answer interface, which is the appropriate cross-family comparison for text LLMs, protein-LLMs, and PLM adapters.

PLM classification. PLM classification experiments use 8,192 training examples and 2,048 examples per test split for the controlled-pilot setting. Each table cell reports ID/OOD.

The PLM classification tables should be read as encoder-capacity baselines. They show whether sequence representations expose global labels, local feature-type labels, and residue-level evidence, but they do not test whether a model can follow the same MCQA answer contract used for LLM-style systems.

MCQA. MCQA tables use the same P1/P2/P3 biological targets but score the final candidate answer. The aggregate "Full" score is the task-count weighted average of P1, P2, and P3 primary metrics. P1 and P3 use accuracy; P2 uses set F1. These rows use the compact MCQA evaluation view; some PLM fine-tune rows cover only the available heads in that run.

The "0-shot" PLM interface is a frozen-encoder before-fine-tuning baseline: candidates are ranked by similarity to label or evidence reference vectors computed from labeled training protein or span embeddings. It does not train a P1/P2/P3 head, does not update PLM weights, and does not compare protein embeddings to text embeddings of label names or answer options. The "Fine-tune" PLM interface is the MCQA-bridge setting: a fine-tuned PLM head scores the exact same candidate records used for MCQA, using candidate metadata such as label ids, feature types, or spans rather than natural-language option text. For P1/P2, the head receives a mean-pooled non-special-token protein embedding. For P3, the scorer receives per-token residue embeddings concatenated with the queried feature embedding, and a candidate span is scored by the mean residue logit over that span. In these fine-tune runs, the head is trained jointly with the PLM tuning parameters: ESM-2, ProtBERT, and Ankh use full PLM fine-tuning, whereas ProstT5 and ProGen use PLM LoRA adapters plus the head. Thus, the cross-family comparison is metric-aligned and answer-space-aligned, but not interface-identical; PLMs score candidates with encoder-derived heads, while LLM and protein-LLM rows use prompting or likelihood scoring.

Together, the ID and OOD MCQA tables show the effect of the evaluation interface. PLM adapters can score the fixed candidates with encoder-derived signals, whereas text and protein LLMs must use prompt or likelihood interfaces. The rows are therefore comparable by answer space and metric, but not by internal scoring mechanism.

Table 13: How to read the P1/P2/P3 individual-result appendix. This table maps each attribute-level result family to its evaluation track, reported quantity, and comparable claim. Primary comparisons should be made within a result family and evaluation track.

Result family	Tracks	Models	Reported quantity	Main score/diagnostic	Comparable claim
PLM classification	P1-cls, P2-cls, P3-cls	PLMs only	Label scores or residue/span scores from task-specific heads	Top-k micro-F1 or top-k residue F1	Encoder attribute capacity
MCQA	P1-mcqa, P2-mcqa, P3-mcqa	PLM adapters, text LLMs, protein-LLMs	One selected option or selected option set	Accuracy for P1/P3, set F1 for P2	Controlled recognition and grounding
Input-control diagnostics	P1-mcqa, P2-mcqa, P3-mcqa	LLMs and protein-LLMs	Same MCQA examples under altered input	SeqGap and OrderGap relative to Full	Sequence and order dependence
Masked-evidence intervention	P3-mcqa	Models with span scores	Full, background, and local evidence scores	Necessity and sufficiency	Residue evidence dependence
Architecture diagnostics	Model-specific	Q-former models	Soft-query or generation behavior	Geometry or behavior metrics	Mechanistic interpretation, not leaderboard ranking

Table 14: P1-cls global semantic annotation results for PLM classification. The metric is top-k-by-gold micro-F1.

Model	EC	GO-BP	GO-MF	GO-CC	Loc.	Cat.	Path.	Cof.
Random	.004/.003	.001/.001	.001/.001	.002/.001	.000/.000	.000/.000	.000/.000	.000/.000
ESM-2	.453/.443	.035/.037	.028/.027	.055/.039	.118/.133	.016/.012	.002/.005	.014/.012
ProtBERT	.030/.023	.033/.029	.027/.030	.041/.032	.020/.019	.016/.012	.002/.005	.010/.007
Ankh	.028/.028	.028/.024	.050/.055	.054/.047	.023/.023	.011/.009	.002/.005	.008/.007
ProstT5	.160/.159	.095/.091	.158/.154	.209/.203	.039/.046	.019/.019	.001/.003	.006/.008
ProGen	.027/.037	.033/.036	.032/.035	.071/.056	.018/.018	.012/.009	.001/.004	.009/.003

D.2 Input and Grounding Diagnostics

The next tables are diagnostics rather than additional leaderboards. Input controls ask whether MCQA scores depend on protein sequence information. Masked-evidence interventions focus on P3 and ask whether the annotated local span contributes to the model score.

Input controls. Input controls are reported for generation- or likelihood-based candidate systems. The no-sequence condition tests whether a model can answer from candidate priors alone. The sequence-shuffle condition preserves amino-acid composition while removing residue order. The choice-shuffle condition tests option-order sensitivity.

The input-control results separate useful sequence dependence from answer-prior effects. Positive SeqGap values indicate that a model benefits from seeing the protein sequence, while near-zero or negative gaps indicate that the score can be reproduced without meaningful sequence evidence. OrderGap similarly tests whether residue order adds signal beyond amino-acid composition.

Masked-evidence interventions. Masked-evidence interventions are reported for residue-grounded P3 examples. They are not additional leaderboard metrics. Instead, they test whether the model score for the queried biological statement changes when the annotated evidence span is removed or isolated. Necessity is $s_{\text{full}} - s_{\text{bg}}$; sufficiency is $s_{\text{local}} - s_{\text{bg}}$, as defined in Appendix C.7. Positive necessity indicates that the annotated span contributes to the full-sequence score. Positive sufficiency indicates that the local evidence span alone carries more signal than the background without that span.

The aggregate site results show a consistent pattern. Candidate-span ranking degrades modestly from ID to OOD, while necessity and sufficiency remain close across splits. This means the family-shift effect is clearer in localization ranking than in the masked-evidence score differences. At subtype level, active-site and binding-site examples carry most of the sample volume and the clearest localization signal. Metal-binding rows have very small sample counts and should be treated as diagnostics rather than stable subtype conclusions.

Table 15: P2-cls and P3-cls PLM classification results. P2 primary is top-k-by-gold micro-F1 and secondary is mean Jaccard. P3 primary is top-k-by-gold residue F1 and secondary is mean best IoU. Each cell reports ID/OOD.

Model	P2 primary	P2 secondary	P3 primary	P3 secondary
Random	.198/.189	.105/.103	.391/.396	.054/.055
ESM-2	.729/.725	.675/.666	.612/.614	.405/.418
ProtBERT	.690/.695	.623/.627	.475/.465	.180/.179
Ankh	.587/.591	.506/.509	.461/.463	.146/.145
ProstT5	.567/.563	.484/.488	.505/.495	.194/.204
ProGen	.485/.468	.357/.340	.464/.474	.113/.112

Table 16: MCQA results on the ID test split. PLM rows use candidate scores from 0-shot frozen encoders or fine-tuned heads; LLM and protein-LLM rows use prompted MCQA or likelihood scoring.

Model	Family	Interface	Full	P1	P2	P3
Ankh 0-shot	PLM	0-shot	.681	.780	.559	.704
Ankh fine-tune	PLM	Fine-tune	.440	.532	.485	.337
ESM-2 0-shot	PLM	0-shot	.601	.770	.592	.439
ESM-2 fine-tune	PLM	Fine-tune	.621	.484	.654	.674
ProGen 0-shot	PLM	0-shot	.467	.690	.493	.214
ProGen fine-tune	PLM	Fine-tune	.506	.629	.596	.337
ProstT5 0-shot	PLM	0-shot	.678	.770	.581	.684
ProstT5 fine-tune	PLM	Fine-tune	.507	.387	.579	.510
ProtBERT 0-shot	PLM	0-shot	.567	.730	.490	.480
ProtBERT fine-tune	PLM	Fine-tune	.464	–	.608	.316
BioReason-Pro-SFT	Protein-LLM	Protein prompt	.138	.170	.022	.225
ProLLaMA	Protein-adapted LM	Sequence-text likelihood	.262	.150	.401	.235
Llama-3.1-8B	Text LLM	Text prompt	.277	.340	.296	.194
Qwen2.5-7B	Text LLM	Text prompt	.255	.330	.291	.143
Qwen3-4B	Text LLM	Text prompt	.285	.330	.339	.184

D.3 Model and Subtask Diagnostics

The final group collects analyses that help interpret individual P1/P2/P3 results but should not be merged into the main benchmark score. The comparison rules clarify which model-specific rows are protocol-matched; the Q-former tables diagnose representation behavior; and the subtask breakdown shows how a single aggregate MCQA number decomposes across namespaces and local evidence types.

Comparison rules. Architecture and objective ablations are reported only when the compared rows share the same benchmark examples, split protocol, input controls, and metric. This restriction prevents mixing task-comparable baselines with architecture-matched ablations. Under this rule, encoder-only PLM rows are benchmark reference points, while Q-former or projector variants are controlled ablations only when their P1/P2/P3 scores are computed under the same MCQA or PLM classification evaluation view.

Q-former diagnostics. Q-former analyses are useful for explaining why fine-grained grounding remains difficult, but they are not attribute-level scores by themselves. We therefore report them as diagnostics. The main observed pattern is that protein-conditioned soft queries stay substantially inside the protein-derived subspace, while the query set becomes highly homogeneous in deeper layers and captures only a compressed slice of token-level geometry.

The Q-former diagnostics support a narrow interpretation: the module is protein-conditioned, but its soft-query bottleneck can compress token-level evidence. The Stage2 generation probes show that open-ended evidence text is not yet reliable enough to serve as the main benchmark output. For this reason, the paper should use Q-former material to interpret model behavior, while the main empirical claim remains anchored in P1/P2/P3 attribute-level scores and masked-evidence interventions.

Subtask breakdown. Table 23 gives a representative subtask breakdown for the ESM-2 fine-tune MCQA interface on the ID full condition. The two score columns follow the task-level metric definitions: for P1 and P3, Primary is accuracy and Secondary is MRR; for P2, Primary is set F1 and Secondary is Jaccard.

The breakdown shows that the aggregate MCQA score hides substantial subtask variation. For this ESM-2 fine-tune run, P1 performance is stronger on GO-CC and GO-MF than on EC or GO-BP, P2 feature-type recognition is moderately strong,

Table 17: MCQA results on the family-level OOD test split. Metrics and interfaces match Table 16.

Model	Family	Interface	Full	P1	P2	P3
Ankh 0-shot	PLM	0-shot	.703	.770	.606	.735
Ankh fine-tune	PLM	Fine-tune	.420	.464	.521	.286
ESM-2 0-shot	PLM	0-shot	.607	.750	.588	.480
ESM-2 fine-tune	PLM	Fine-tune	.650	.507	.686	.714
ProGen 0-shot	PLM	0-shot	.491	.660	.494	.316
ProGen fine-tune	PLM	Fine-tune	.490	.667	.588	.265
ProstT5 0-shot	PLM	0-shot	.653	.760	.536	.663
ProstT5 fine-tune	PLM	Fine-tune	.520	.391	.578	.551
ProtBERT 0-shot	PLM	0-shot	.533	.690	.477	.429
ProtBERT fine-tune	PLM	Fine-tune	.479	–	.638	.316
BioReason-Pro-SFT	Protein-LLM	Protein prompt	.132	.170	.003	.225
ProLLaMA	Protein-adapted LM	Sequence-text likelihood	.300	.310	.403	.184
Llama-3.1-8B	Text LLM	Text prompt	.302	.400	.319	.184
Qwen2.5-7B	Text LLM	Text prompt	.295	.440	.270	.174
Qwen3-4B	Text LLM	Text prompt	.297	.360	.376	.153

Table 18: Input-control results for LLM and protein-LLM MCQA systems. SeqGap is Full minus NoSeq; OrderGap is Full minus sequence shuffle.

Model	Split	Full	NoSeq	SeqShuf	ChoiceShuf	SeqGap	OrderGap
BioReason-Pro-SFT	ID	.138	.134	.128	.149	.004	.011
ProLLaMA	ID	.262	.134	.245	.247	.128	.018
Llama-3.1-8B	ID	.277	.158	.254	.295	.119	.023
Qwen2.5-7B	ID	.255	.275	.266	.295	-.020	-.011
Qwen3-4B	ID	.285	.289	.225	.267	-.004	.060
BioReason-Pro-SFT	OOD	.132	.114	.117	.132	.018	.014
ProLLaMA	OOD	.300	.154	.312	.284	.145	-.013
Llama-3.1-8B	OOD	.302	.178	.253	.313	.124	.049
Qwen2.5-7B	OOD	.295	.294	.285	.287	.001	.010
Qwen3-4B	OOD	.297	.321	.265	.316	-.023	.032

and P3 grounding depends heavily on evidence subtype. Common or sequence-constrained local evidence such as binding-site, disordered-region, and transmembrane-like examples is easier than domain-, motif-, or other-region-like examples in this sample. Rows with very small n, such as active-site-like, motif-like, and other-region-like, should be treated as diagnostic rather than stable subtype conclusions. The table therefore motivates reporting P1/P2/P3 and subtype breakdowns instead of relying only on the aggregate MCQA score.

Reporting notes. The PLM classification tables and MCQA tables should not be collapsed into a single leaderboard. PLM classification rows evaluate sequence-encoder capacity, whereas MCQA rows evaluate controlled answer selection under a shared MCQA interface. Full MCQA scores should also be interpreted together with the input-control table: a high Full score with a small SeqGap may indicate reliance on label priors or option wording rather than protein-sequence evidence.

E Compositional Data Construction

This appendix describes the compositional PannotGround datasets used to test whether a model can bind global function (P1), local feature types (P2), and local evidence descriptions (P3) within a single example. The construction intentionally separates two complementary directions. The text-to-protein (T2P) retrieval task asks whether a model can retrieve the sequence whose global/local/evidence composition matches a compositional text query. The protein-to-text (P2T) caption-selection task asks whether a model can select the caption whose global and local statements jointly match a given sequence. Both tasks are built from the same canonical global/local/evidence substrate, but they stress different failure modes: T2P emphasizes retrieval among many proteins, whereas P2T emphasizes distinguishing hard textual alternatives for the same protein.

E.1 Task Format

Table 24 summarizes the two compositional task families. In both cases, the global-function namespace set is restricted to the core annotation sources `ec`, `go_mf`, `catalytic_activity`, and `cofactor`. P3 descriptions are constructed without residue-coordinate requirements in the current compositional release, so the benchmark evaluates binding between global semantics and local evidence categories rather than coordinate-precise span recovery.

Table 19: P3 site-subtype localization and masked-evidence intervention results. MC-Acc, MC-MRR, and MC-Rec@5 evaluate candidate-span ranking. Necessity and sufficiency are score-difference diagnostics and should be interpreted within the same scoring model.

Split	Site subtype	n	MC-Acc	MC-MRR	MC-Rec@5	Necessity	Sufficiency
ID	Active site	7,675	38.33	.589	88.05	.527	.148
ID	Binding site	50,378	32.46	.536	84.87	.566	.204
ID	Generic site	2,566	21.63	.423	68.28	.517	.038
ID	Metal binding	28	35.71	.594	89.29	.383	-.041
ID	All sites	60,647	32.75	.538	84.58	.559	.189
OOD	Active site	19,596	34.05	.553	85.90	.540	.126
OOD	Binding site	124,874	31.67	.529	84.35	.568	.209
OOD	Generic site	4,159	20.65	.430	73.77	.498	.065
OOD	Metal binding	25	24.00	.495	72.00	.671	.166
OOD	All sites	148,654	31.67	.530	84.25	.562	.194

Table 20: Recommended interpretation of model-comparison rows. Rows should be promoted to the main ablation table only when the rightmost column is satisfied.

Comparison type	What it tests	Proper placement	Same-protocol requirement
PLM classification	Encoder attribute capacity without an LLM decoder	PLM classification result table	P1-cls/P2-cls/P3-cls metrics
PLM MCQA score	Whether encoder scores transfer to fixed candidate decisions	MCQA table	P1-mcqa/P2-mcqa/P3-mcqa metrics
Linear projector protein-LLM	Protein-to-language interface under the current benchmark protocol	Main MCQA table	Same examples and controls
Q-former matching variant	Protein-text alignment module or local objective effect	Controlled ablation or appendix	Matched checkpoint protocol
Stage2 evidence generation	Whether the model can generate usable evidence text	Appendix diagnostic	Separate behavior metrics

E.2 Splits and Scale

Table 25 gives the released split sizes. The P2T dataset is exactly balanced across the two diagnostic subtasks in every split: same-local/different-global, where local evidence is controlled and global function must be distinguished, and same-global/different-local, where global function is controlled and local evidence must be distinguished. This balance is important because an aggregate P2T accuracy can otherwise conflate global-function discrimination with local-evidence discrimination.

The T2P release contains 5,000 retrieval queries. The P2T release contains 6,000 caption-selection questions over 5,905 unique gold proteins. The P2T mean sequence length is 419.05 amino acids, with a median of 358 amino acids and a 90th percentile of 734 amino acids. Thus, the P2T task contains realistic Swiss-Prot-scale proteins rather than short synthetic snippets.

E.3 Controlled Compositional Ablations

In addition to the full compositional view, we construct four controlled projections of the same gold P1/P2/P3 source annotations. These variants are used to separate two questions that are entangled in the full benchmark: whether performance changes as the number of displayed attributes increases, and whether a model relies more on global function or local evidence. Table 26 defines the projections. The P2T variants preserve the 8-way caption-selection format and contain the same split sizes as the full P2T release: 1,000 `val_id`, 1,000 `val_ood_family`, 2,000 `test_id`, and 2,000 `test_ood_family` rows. The T2P variants preserve the 64-way retrieval format and contain 1,000 `val_id`, 2,000 `test_id`, and 2,000 `test_ood_family` rows.

All controlled-variant attributes were provenance-audited as subsets of the corresponding source protein's full P1/P2/P3 composition. The audit therefore checks biological faithfulness to the current PannotGround gold annotations; it does not independently re-curate UniProt biology. For the one-sided `p1_only` and `local_only` ablations, inherited hard-negative relation labels should be treated as implementation metadata rather than literal biological relation labels, because one side of the global/local signature is intentionally empty. The Attr1 columns reported in the results appendix are different: they are individual P1, P2, and

Table 21: Summary of Q-former soft-query diagnostics across Stage2 checkpoints, layers, and sampled tasks. These metrics describe representation geometry and should not be averaged with P1/P2/P3 attribute-level scores.

Diagnostic	Mean	Interpretation
Mean pairwise query cosine	.908	High query homogeneity
Mean distance to query centroid	.105	Low query diversity at deeper layers
PC1 variance ratio	.884	Query states are close to low-rank geometry
Query-in-protein-span ratio	.679	Soft queries remain protein-conditioned
Protein-in-query-span ratio	.100	Query states cover a narrow part of token geometry
Mean principal cosine	.677	Moderate subspace alignment
Protein-induced drift norm	10.949	Protein input changes query states substantially
Protein-induced drift parallel ratio	.624	Much of the drift lies in the protein-derived span

Table 22: Stage2 evidence-generation probe summary. These prompt variants test whether a protein-conditioned generator can emit a reliable evidence block from sequence alone. They are diagnostic attempts, not main PannotGround benchmark rows.

Prompt variant	BLEU-2	BLEU-4	METEOR	ROUGE-1	Behavioral read
Strict schema	8.36	0.00	9.77	2.19	Mostly repeats rule text rather than producing evidence fields
Family first	8.73	0.00	10.01	4.43	Similar rule copying or degenerate filler
Minimal	8.48	0.00	14.18	4.38	Produces content, but often the wrong annotation type
Soft prompt	11.88	5.75	16.47	6.94	Best lexical overlap, but still unreliable and format-unstable

P3 task references, not generated compositional projections.

E.4 T2P Query Composition

The T2P queries are comparatively long because they concatenate global function, local feature-type, and local-evidence descriptions. Table 27 summarizes the query statistics. The mean query contains 51.1 words, 3.16 global atoms, 2.39 local feature types, and 3.43 local evidence queries. The `test_ood_family` split is slightly more compositionally dense than `test_id`, especially in global atoms, which should be considered when interpreting OOD-family retrieval results.

The query text is explicitly factored into three components. The average global function component is 189.21 characters (22.71 words), the average local feature-type component is 43.48 characters (5.15 words), and the average local evidence-query component is 109.36 characters (13.25 words). This confirms that T2P is not a keyword-only retrieval task: most examples require matching a multi-part description whose strongest lexical evidence may be split across global and local clauses.

The T2P candidate pool is therefore not a standard open-set retrieval pool. It is deliberately contrastive: many incorrect proteins preserve either the global component, the local evidence component, or a partial attribute overlap. A model that retrieves near-miss candidates may be biologically closer than one that selects no-overlap negatives, so error relations should be reported together with R@1 or MRR.

E.5 P2T Caption-Selection Composition

The P2T dataset uses a fixed 8-way choice format. Each row contains one gold caption and seven distractors whose relation to the gold composition is known. The two diagnostic subtasks create complementary controlled comparisons. In same-local/different-global rows, all plausible captions share local evidence, so the model must recover the correct global function. In same-global/different-local rows, global function is controlled, so the model must recover the correct local evidence. This design makes it possible to ask whether a model is globally competent, locally competent, or genuinely compositionally binding the two.

Gold answer positions are close to uniform: A–H contain 11.7%, 12.3%, 12.6%, 12.0%, 12.4%, 12.9%, 12.9%, and 13.2% of gold labels, respectively. This does not prevent position bias in generative models, but it does make such bias diagnosable rather than an artifact of the gold labels.

Table 23: Representative MCQA subtask cases for ESM-2 fine-tuned heads on the ID full condition. Because the prediction object differs by attribute level, n is the number of evaluated MCQA examples in each subtask, and the score columns use task-specific metrics: for P1 and P3, Primary score is accuracy and Secondary score is MRR; for P2, Primary score is set F1 and Secondary score is Jaccard.

Task	Subtask	n	Primary score	Secondary score
P1	EC	6	.333	.569
P1	GO-BP	18	.333	.579
P1	GO-CC	18	.611	.764
P1	GO-MF	20	.550	.729
P2	Feature-type set	100	.654	.777
P3	Active-site-like	2	1.000	1.000
P3	Binding-site-like	39	.718	.830
P3	Compositional-bias-like	6	.667	.806
P3	Disordered-like	9	.778	.822
P3	Domain-like	10	.400	.608
P3	Generic-site-like	3	.667	.833
P3	Motif-like	3	.000	.344
P3	Other region	1	.000	.250
P3	Region-of-interest-like	2	.500	.625
P3	Transmembrane-like	23	.783	.854

Table 24: Compositional task formats.

	Task family	Input	Candidate set	Random baseline
T2P	Text-to-protein compositional retrieval	Compositional text query containing global-function, local-feature, and local-evidence statements	64 protein candidates per query, including hard negatives that preserve global function, local evidence, or partial attributes	R@1 = 1/64 = 1.56%
P2T	Protein-to-text compositional caption selection	Protein sequence and an 8-way caption-choice prompt	One gold caption plus seven contrastive captions per question	Accuracy = 1/8 = 12.5%

E.6 Visual Summary of Data Composition

Figure 6 summarizes the text-side structure of the T2P retrieval queries. The component-length panel shows that the global clause carries the largest textual mass, while the P3 evidence clause is still a substantial part of the query. The sunburst and feature-presence panels show that query composition is not dominated by one local evidence family: domain and motif evidence, disordered and compositional-bias regions, ligand/cofactor binding sites, active sites, and transmembrane topology all contribute meaningfully to the retrieval challenge.

Figures 7 and 8 give the corresponding P2T gold-caption composition. The global panels show that the caption-selection benchmark is centered on molecular-function and enzyme-related global semantics, but not restricted to a single identity source. The local panels show that local evidence is distributed across binding sites, domains, active sites, disordered/low-complexity regions, motifs, and transmembrane topology. This mixture is important because P2T hard negatives can preserve global function while altering local evidence, or preserve local evidence while altering global function.

E.7 Recommended Use

For benchmark reporting, T2P and P2T should not be collapsed into a single score. T2P measures whether a text-conditioned protein retriever can rank the unique fully matching sequence above proteins that preserve only part of the composition. P2T measures whether a sequence-conditioned model can bind the correct global caption to the correct local evidence under a controlled multiple-choice format. In both directions, ID and OOD-family splits should be reported separately. For P2T, the two diagnostic subtasks should also be reported separately because they identify whether remaining errors are driven by global-function confusion or local-evidence confusion.

Table 25: Compositional split sizes. T2P rows are retrieval queries; P2T rows are 8-way caption-selection questions.

Direction	Split	Rows	Candidates / choices	same-local / different-global	same-global / different-local
T2P	`val_id`	1,000	64 candidates	–	–
T2P	`test_id`	2,000	64 candidates	–	–
T2P	`test_ood_family`	2,000	64 candidates	–	–
P2T	`val_id`	1,000	8 choices	500	500
P2T	`val_ood_family`	1,000	8 choices	500	500
P2T	`test_id`	2,000	8 choices	1,000	1,000
P2T	`test_ood_family`	2,000	8 choices	1,000	1,000

Table 26: Controlled compositional ablation variants. Each variant is a projection of the same source composition rather than a separately curated annotation set.

Variant	Projection rule	Diagnostic comparison	Released task views
Full	Keep all available P1 global atoms, P2 local feature types, and P3 local evidence descriptors for the selected protein.	Original compositional benchmark with naturally varying attribute count and redundant biological evidence.	`retrieval_v1`; `core_v3_distinguishable`
Attr2	Keep exactly two attributes: one P1 global atom and one P3 local evidence descriptor.	Minimal controlled composition; tests whether a model can bind one global cue to one local evidence cue.	`retrieval_attr2`; `core_v3_attr2`
Attr4	Keep exactly four attributes. The exact P1/P2/P3 shape is data-driven, but the total attribute count is fixed.	Attribute-count ablation against Attr2 while retaining global/local composition.	`retrieval_attr4`; `core_v3_attr4`
Global-only	Keep all available P1 global attributes and remove P2/P3 local attributes.	Tests whether global function alone is sufficient for retrieval or caption selection.	`retrieval_p1_only`; `core_v3_p1_only`
Local-only	Keep all available P2/P3 local attributes and remove P1 global attributes.	Tests whether local feature and evidence descriptions alone are sufficient.	`retrieval_local_only`; `core_v3_local_only`

F Compositional Evaluation Results

This appendix reports the compositional evaluation of PannotGround along two directions introduced in Appendix E. The text-to-protein (T2P) task is a 64-way retrieval problem evaluated with R@1, R@5, MRR, and mean rank. The protein-to-text (P2T) task is an 8-way caption-selection problem evaluated with accuracy and relation-aware diagnostic statistics. The random baselines are 1.56% R@1 for T2P and 12.5% accuracy for P2T.

The presentation below is organized by evaluation question rather than by model family, but the model interfaces are still explicit. First, we define the evaluation surfaces and the role of each family. Second, we report T2P retrieval. Third, we compare all P2T model families in a single accuracy table. Finally, we decompose P2T behavior into diagnostic axes: global versus local discrimination, wrong-answer relation structure, shortcut behavior, and biology-conditioned difficulty.

F.1 Evaluation Protocol and Coverage

All P2T systems are evaluated on the same 8-way caption-selection instances: the input protein is paired with one gold caption and seven contrastive captions whose global/local/evidence relation to the gold answer is known. The primary metric is accuracy, and the diagnostic metrics ask whether a wrong selection preserves the global-function component, the local-evidence component, both partially, or neither. This protocol is important because a model can be above random yet fail to bind global function and local evidence, or be low-accuracy but still choose biologically close hard negatives.

The PLM bridge results should be read as a set of concrete factorized PLM baselines, not as a single illustrative demo. Four protein encoders are evaluated: `ankh`, `esm2`, `progen`, and `prostt5`. For each encoder, independent annotation-style evidence is converted into an 8-way caption score. These baselines test whether marginal recognition of individual protein properties is sufficient for compositional

Table 27: T2P query statistics from the retrieval-v1 text-component statistics.

Split	Rows	Mean chars	Mean words	Mean global atoms	Mean local types	Mean evidence queries
All	5,000	429.05	51.10	3.16	2.39	3.43
`val_id`	1,000	427.53	51.10	3.13	2.46	3.40
`test_id`	2,000	414.52	49.54	2.92	2.37	3.46
`test_ood_family`	2,000	444.35	52.67	3.42	2.39	3.42

Table 28: T2P hidden candidate relation mix over all 5,000 queries.

Candidate relation	Mean per query	Share
`both_wrong`	16.00	25.40%
`partial_attribute_overlap`	16.00	25.40%
`random_wrong`	13.46	21.37%
same-local/wrong-global	10.21	16.20%
same-global/wrong-local	7.33	11.64%

Table 29: P2T 8-way choice-relation mix.

Choice relation	Choices	Mean / question	Share
`gold`	6,000	1.00	12.5%
same-global/wrong-local	7,664	1.28	16.0%
same-local/wrong-global	7,420	1.24	15.5%
`both_wrong`	11,997	2.00	25.0%
`random_wrong`	14,919	2.49	31.1%

caption selection. The answer is negative in the current release: all four PLM bridges fall below the 12.5% random baseline on both held-out splits, and their behavior diagnostics show strong caption-length bias.

We also report a separate PLM-to-text alignment interface, denoted PLM-TextAlign. This interface does not use independent annotation heads. Instead, a frozen PLM protein embedding is mapped into a PubMedBERT text space with a ridge regressor and choices are ranked by protein-caption cosine compatibility. It therefore tests a different hypothesis from the PLM bridge: whether a sequence embedding can be linearly aligned to the text space used by the captions.

Table 30 summarizes the evaluated model families and interfaces. The P2T benchmark covers all model families in the current compositional result set: factorized PLM bridges, PLM-TextAlign, text LLMs, protein-LLMs, and protein-text alignment models. The T2P retrieval benchmark is currently reported for models that expose a retrieval-compatible protein-text scoring interface. Thus, the absence of a T2P row for a family should be read as missing coverage in the current release, not as a negative result.

F.2 Worked P2T Scoring Example

To make the evaluation rule concrete, Table 31 traces one real held-out instance: `PGCOMP:test_id:00001:P0DJD3`. The row belongs to `test_id` and to the same-local/different-global subtask. The gold answer is choice H. Its global labels include RNA binding (GO:0003723) and mRNA binding (GO:0003729); its local P2 feature types are compositional-bias region, disordered region, and domain; and its P3 evidence mentions basic/acidic residues, low complexity, polar residues, disorder, and an RRM domain. The candidate set is diagnostic by construction: choices F and G preserve the same local evidence but change the global function, while choices A and D preserve part of the global signal but change local evidence.

This example illustrates what the aggregate P2T metrics count. For a factorized PLM bridge, the current runs use equal component weights, so $s(c) = (s_{\text{global}}(c) + s_{\text{local}}(c) + s_{\text{evidence}}(c))/3$ over finite component scores. In the `progen` row, choice F obtains $(-3.116 - 0.464 - 2.123)/3 = -1.901$, whereas the gold choice H obtains $(-3.806 - 0.464 - 2.123)/3 = -2.131$. F and H have identical local/evidence scores in this instance; the error is therefore caused by the marginal global score ranking the same-local/wrong-global caption above the gold caption. Alignment models instead rank captions by protein-text compatibility, protein-LLMs rank choices by conditional likelihood, and prompted text LLMs are evaluated from the parsed generated answer. In all cases, the primary row-level score is binary accuracy, and the relation-aware diagnostics are read from the selected choice's known distractor type. The T2P retrieval evaluation is analogous but uses a 64-way protein ranking: R@1 is one only when the gold protein is ranked first, and the top-ranked wrong candidate supplies the relation-aware retrieval error type.

F.3 T2P Retrieval

Table 32 reports the original full-compositional 64-way T2P alignment evaluations on `val_id` and `te`

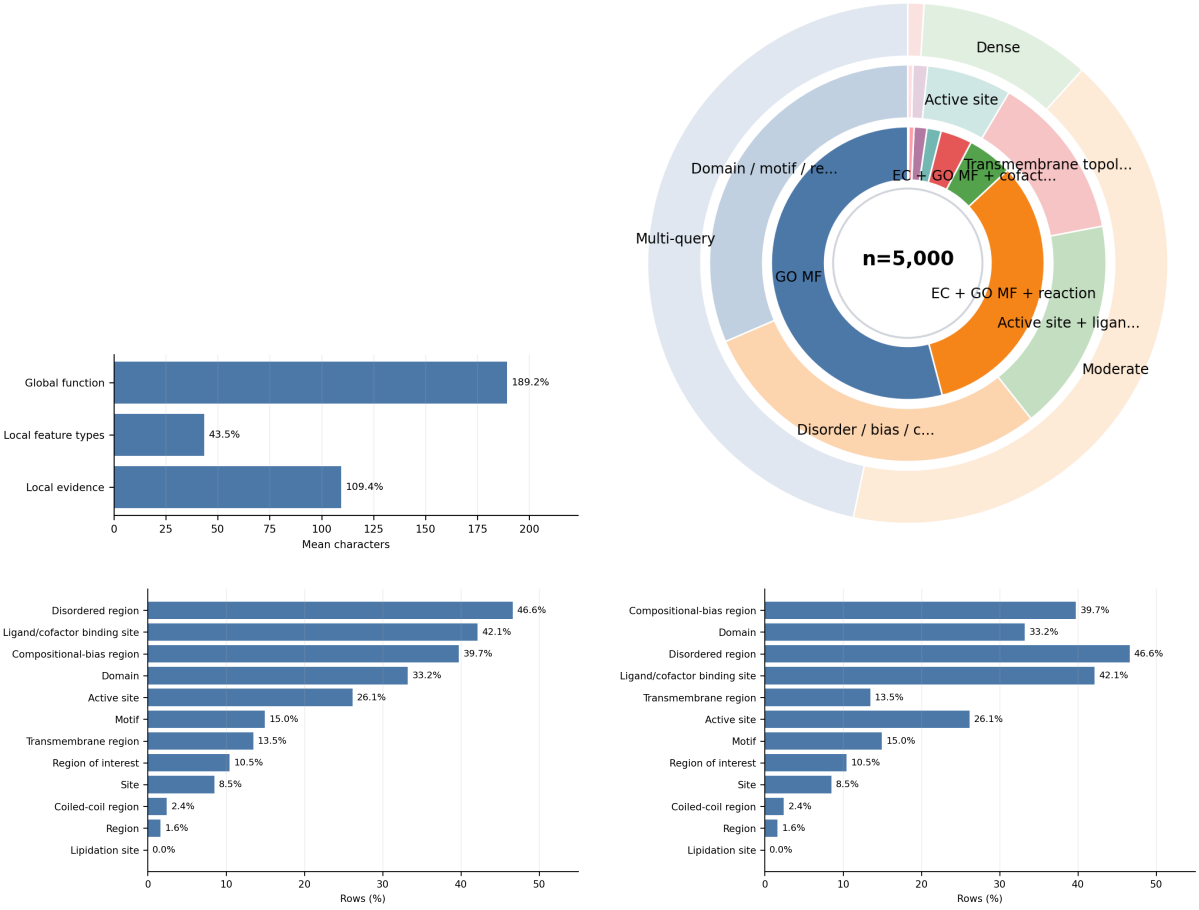


Figure 6: T2P compositional query structure. Top left: query text length by global, local-feature, and local-evidence component. Top right: joint query-composition hierarchy. Bottom left: local feature-type prevalence. Bottom right: local evidence-query prevalence.

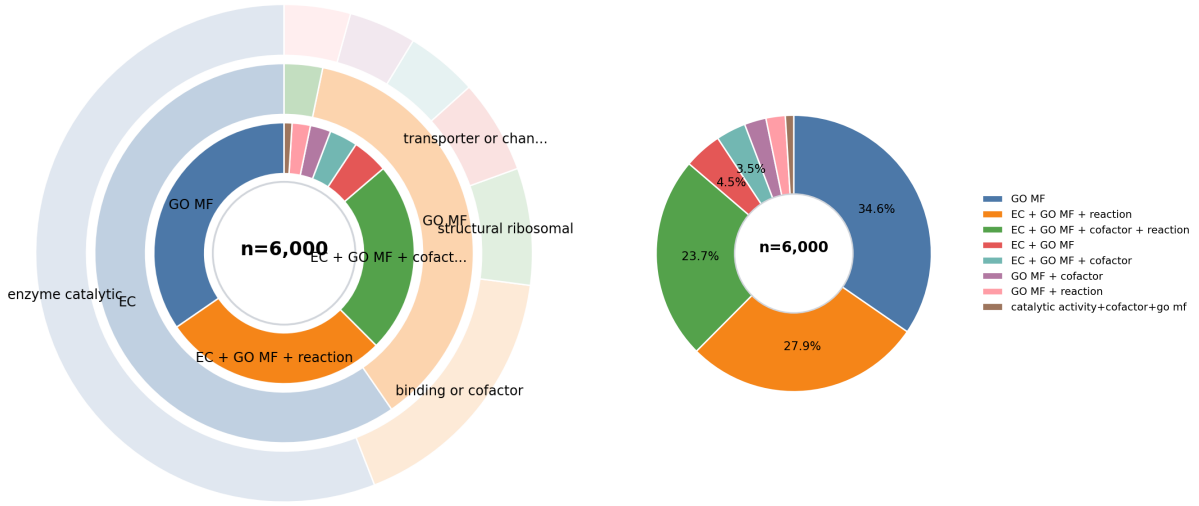


Figure 7: P2T global-composition visualizations. Left: global composition hierarchy, including primary global identity and rendered annotation bundles. Right: global annotation-bundle distribution.

`st_id`. Tables 33 and 34 add the newly completed controlled-variant and full-pool retrieval runs.

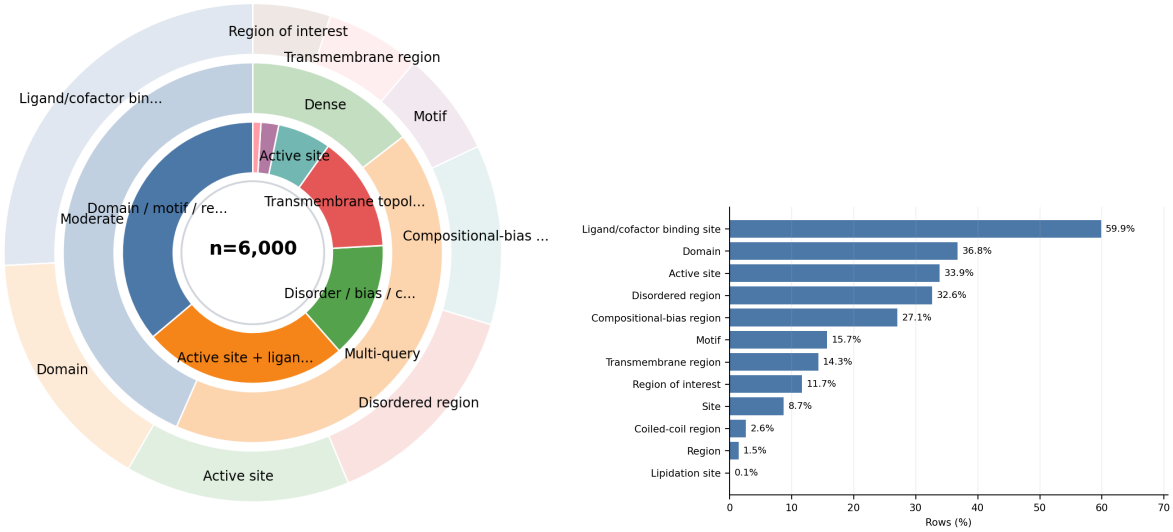


Figure 8: P2T local-composition visualizations. Top left: P2/P3 local-composition hierarchy. Top right: P2 feature presence.

Table 30: Compositional evaluation coverage. P2T rows are evaluated on `val_id`, `val_ood_family`, `test_id`, and `test_ood_family`. T2P rows are evaluated on the completed retrieval splits available in the current result directory.

Family	Models	P2T evaluation interface	T2P coverage
PLM bridge	`ankh`, `esm2`, `progen`, `prostt5`	Factorized caption scoring from independent annotation heads	Not evaluated
PLM-TextAlign	`ankh`, `esm2`, `progen`, `prostt5`	Ridge-mapped PLM embedding scored against caption text embeddings	Not evaluated
Text LLM	`llama31_8b`, `qwen25_7b`, `qwen3_4b`	Prompted 8-way answer selection	Not evaluated
Protein-LLM	`prollama`, `bioreason_pro_sft`	Choice-likelihood or direct answer selection	Not evaluated
Protein-text alignment	`protrek_35m`, `protrek_650m`, `protst_esm2`, `protst_protbert`, `prott3_base`, `pannot_fg`	Protein-caption compatibility scoring	Completed for available retrieval-compatible models

The completed T2P results show that compositional retrieval remains difficult even for protein-text alignment models. On `test_id`, `protrek_650m` obtains the best R@1 (11.05%) and R@5 (40.55%), while `pannot_fg` obtains the best mean rank (9.72) but a lower R@1 (8.10%). This means that some systems can move the gold protein into a plausible neighborhood without consistently ranking it first. The error columns further indicate that many top-ranked failures are not arbitrary negatives: they preserve global function, local evidence, or partial global/local attributes. T2P should therefore be reported with both retrieval metrics and relation-aware error analysis.

Controlled T2P variants. Table 33 reports the completed 64-way retrieval runs for the four controlled T2P projections. These use the same query rows and hard-negative interface as the full T2P benchmark, but vary the number or source of query attributes.

The controlled T2P variants answer a different question from the original full T2P table. Attr2 and Attr4 isolate the effect of attribute count under a fixed 64-way hard-negative interface. Global-only asks whether functional labels alone retrieve the protein; local-only asks whether local sequence evidence is sufficiently discriminative. The strongest systems show the same broad pattern as P2T: global function is the easiest retrieval signal, but local evidence still carries nontrivial signal for ProTrek models.

Full-pool retrieval is much harder than constructed 64-way retrieval. ProTrek still places the gold protein in the top 50 for roughly 20–24% of

Table 31: Worked P2T scoring example for PGCOMP:test_id:00001:P0DJD3. For rows with stored scalar choice scores, higher scores are better. Accuracy is one only when the selected answer equals the gold answer H; otherwise the known relation of the selected distractor becomes the error type.

Family / model	Scoring interface	Pred.	Selected score / output	Gold score / rank	Evaluation outcome
PLM bridge / progen	Mean of factorized global, local-feature, and local-evidence scores	F	F: -1.901	H: -2.131 / 4	Incorrect; same-local/wrong-global
PLM bridge / esm2	Mean of factorized global, local-feature, and local-evidence scores	D	D: -1.377	H: -1.472 / 4	Incorrect; same-global/wrong-local
Alignment / protrek_35m	Protein-caption embedding compatibility	F	F: 0.208	H: 0.189 / 3	Incorrect; same-local/wrong-global
Alignment / pannot_fg	Q-Former caption reranking by compatibility score	H	H: 0.813	H: 0.813 / 1	Correct
Text LLM / llama31_8b	Prompted JSON answer generation and parsing	F	F generated	H not selected	Incorrect; same-local/wrong-global
Protein-LLM / prollama	Normalized conditional log-likelihood of each answer choice	F	F: -4.619	H: -5.009 / 2	Incorrect; same-local/wrong-global
Protein-LLM / bioreason_pro_sft	Normalized conditional log-likelihood of each answer choice	A	A: -0.210	H: -0.438 / 5	Incorrect; same-global/wrong-local

Table 32: T2P 64-way compositional retrieval results. The last four columns describe the top-ranked wrong answer when R@1 fails.

Model	Split	Rows	R@1	R@5	MRR	Mean rank	Same global wrong local	Same local wrong global	Partial	Other
protrek_35m	val_id	1,000	10.20%	35.30%	0.2447	11.48	316	231	307	44
protrek_35m	test_id	2,000	9.40%	39.05%	0.2471	10.15	660	495	571	86
protrek_650m	val_id	1,000	11.30%	38.10%	0.2555	11.53	283	298	278	28
protrek_650m	test_id	2,000	**11.05%**	**40.55%**	**0.2655**	10.64	623	514	548	94
protst_protbert	val_id	1,000	2.50%	16.30%	0.1191	20.60	149	129	454	243
protst_protbert	test_id	2,000	3.30%	17.85%	0.1283	22.10	399	291	728	516
protst_esm2	val_id	1,000	1.70%	12.60%	0.1004	26.17	105	144	172	562
protst_esm2	test_id	2,000	0.55%	4.25%	0.0503	38.45	164	127	260	1,438
prott3_stage1_base	val_id	1,000	2.50%	20.60%	0.1327	19.41	178	147	399	251
prott3_stage1_base	test_id	2,000	3.55%	22.10%	0.1427	19.64	398	241	607	683
pannot_fg	val_id	1,000	10.50%	**39.90%**	**0.2577**	**9.87**	297	248	301	49
pannot_fg	test_id	2,000	8.10%	39.75%	0.2406	**9.72**	622	445	653	118

queries, but R@1 falls below 1%. This supports treating constructed 64-way T2P as a controlled hard-negative diagnostic and full-pool T2P as the realistic large-candidate retrieval setting.

F.4 P2T Comparative Accuracy

Table 35 is the central P2T comparison. It reports all model families in a single table on the two held-out splits most relevant for paper-level comparison. The two diagnostic columns decompose each split into same-local/different-global rows, where local evidence is controlled and global function must be distinguished, and same-global/different-local rows, where global function is controlled and local evidence must be distinguished.

The main full-compositional P2T result is a clear family-level separation. Alignment models are the strongest group, followed by PLM-TextAlign and then text LLMs. The strongest systems are protrek_35m, protrek_650m, and pannot_fg; all remain substantially above random on the OOD-family split. PLM-TextAlign is much stronger than the factorized PLM bridge: esm2 PLM-TextAlign reaches 38.00% on test_id, compared with 9.40% for the esm2 PLM bridge. This means that the failure of the PLM bridge should not be read as a failure of all PLM-derived sequence representations. It is specifically a failure of independent marginal annotation scores to bind into an 8-way caption decision. The ridge-aligned text interface recovers a substantial amount of caption-level compatibility signal, especially when the full caption contains many redundant global and local cues. Text LLMs are weaker but not uniformly random: llama31_8b reaches 25.95% on test_id and 27.55% on test_ood_family, whereas the two Qwen models remain closer to the random baseline. Protein-LLMs are mixed: prollama is slightly above random, whereas bioreason_pro_sft is

Table 33: Controlled T2P 64-way retrieval results. ID and OOD columns report R@1/R@5 on `test_id` and `test_ood_family`, respectively. All R values are percentages.

Variant	Model	ID R@1	ID R@5	OOD R@1	OOD R@5
Attr2	protrek_35m	29.15	67.00	22.15	58.75
Attr2	protrek_650m	**31.15**	**70.80**	22.85	61.05
Attr2	protst_esm2	8.45	29.15	6.95	27.70
Attr2	protst_protbert	8.10	28.60	6.30	25.05
Attr2	prott3_stage1_base	7.55	25.45	6.60	23.10
Attr2	pannot_fg	23.35	65.05	16.00	55.60
Attr4	protrek_35m	30.15	62.45	28.75	57.75
Attr4	protrek_650m	**30.75**	**63.40**	**30.85**	**60.90**
Attr4	protst_esm2	8.10	27.20	8.40	26.45
Attr4	protst_protbert	8.00	26.05	6.95	25.50
Attr4	prott3_stage1_base	8.90	27.70	9.65	27.65
Attr4	pannot_fg	21.90	57.45	24.95	54.90
Global-only	protrek_35m	**56.30**	78.20	53.45	77.05
Global-only	protrek_650m	55.60	80.00	**54.40**	77.55
Global-only	protst_esm2	11.95	36.80	13.35	38.70
Global-only	protst_protbert	10.45	34.05	10.60	32.50
Global-only	prott3_stage1_base	16.55	43.25	17.90	44.50
Global-only	pannot_fg	53.80	**82.25**	48.65	**80.05**
Local-only	protrek_35m	21.60	43.95	**22.30**	42.70
Local-only	protrek_650m	**22.50**	**46.35**	21.85	**45.65**
Local-only	protst_esm2	2.90	10.60	2.65	11.30
Local-only	protst_protbert	2.90	8.55	2.90	10.50
Local-only	prott3_stage1_base	3.50	13.50	4.15	12.90
Local-only	pannot_fg	16.60	45.85	16.65	44.20

Table 34: Full-pool T2P retrieval on the original full-compositional queries. The gold protein is ranked against the canonical pool of 282,021 proteins. Random R@1 is 0.00035%.

Model	test_id					test_ood_family				
	R@1	R@10	R@50	MRR	Mean rank	R@1	R@10	R@50	MRR	Mean rank
protrek_35m	0.60	5.25	20.60	0.0258	4169.84	0.65	6.00	23.10	0.0291	3869.87
protrek_650m	0.20	4.85	20.15	0.0225	5281.47	**0.70**	**7.25**	**24.45**	**0.0317**	**3738.37**
protst_esm2	0.10	0.25	0.75	0.0017	51104.57	0.00	0.15	0.85	0.0009	48237.14
protst_protbert	0.00	0.15	0.75	0.0008	52412.30	0.00	0.00	0.90	0.0008	54859.38

below random.

The full table alone should not be interpreted as a monotonic "more-attributes-is-harder" result. The full captions often contain more than four attributes, but they also contain more redundant biological cues. Controlled variants are needed to separate attribute-count difficulty from cue-density effects.

Table 36 adds Attr1 where the disjoint P1/P2/P3 evaluation gives a meaningful individual-task reference. For PLMs, Attr1 uses the QA-bridge rerun on the same LLM-QA Attr1 candidate files: the PLM reads the sequence and scores candidate metadata with the corresponding P1/P2/P3 task head, but it does not read the option text. For text LLMs and protein-LLMs, Attr1 uses the previous MCQA-style choice evaluation; for PLM-TextAlign/alignment systems, it uses calibrated similarity scoring on the same LLM-QA Attr1 files. The PLM-TextAlign family mean excludes `prostt5`, whose Attr1 similarity jobs failed during sequence tokenization. Attr1 should therefore be read as an individual-task reference, not as the same answer space as the 8-way compositional P2T columns. The PLM contrast is especially informative: the task-head aggregate is about 52%, but the same encoders fall near or below random when those marginal signals are factorized into compositional caption selection. For MCQA-style models, the drop is clear for text LLMs and `prollama`; `bioreason_pro_sft` is already weak on the disjoint aggregate, so its controlled compositional scores mainly indicate continued low choice reliability. Among the compositional columns, the table shows two distinct effects. First, Attr4 is generally harder than Attr2 for models that perform meaningful caption selection, especially text LLMs and alignment models. This is the cleanest evidence for an attribute-count effect, because Attr2 and Attr4 use fixed-size projected captions. Second, the full benchmark is not simply harder than Attr4. Full captions can be easier for alignment models and PLM-TextAlign because they provide extra redundant evidence. For example, a full caption can state both a global activity and several local sequence features; matching any one highly distinctive component can help the model choose the right option. In contrast, Attr2 removes most of that redundancy and Attr4 fixes the amount of evidence to exactly

Table 35: P2T held-out 8-way compositional caption-selection accuracy. SL/DG is same-local/different-global; SG/DL is same-global/different-local.

Family	Model	Test-ID	SL/DG	SG/DL	Test-OOD	SL/DG	SG/DL
Random	analytic_uniform_random	12.50	12.50	12.50	12.50	12.50	12.50
PLM	ankh	5.25	4.70	5.80	6.00	4.60	7.40
PLM	esm2	9.40	7.90	10.90	9.10	7.90	10.30
PLM	progen	7.35	6.30	8.40	6.95	6.10	7.80
PLM	prostt5	6.35	4.70	8.00	5.50	5.30	5.70
PLM-TextAlign	ankh	24.55	22.20	26.90	23.80	21.50	26.10
PLM-TextAlign	esm2	38.00	37.10	38.90	34.80	33.70	35.90
PLM-TextAlign	progen	20.75	18.50	23.00	21.40	18.30	24.50
PLM-TextAlign	prostt5	32.40	30.60	34.20	30.40	28.10	32.70
Text LLM	llama31_8b	25.95	23.30	28.60	27.55	26.90	28.20
Text LLM	qwen25_7b	13.65	13.10	14.20	13.50	13.00	14.00
Text LLM	qwen3_4b	12.95	12.90	13.00	12.15	12.00	12.30
Alignment	protrek_35m	**60.50**	**62.80**	**58.20**	54.20	**56.30**	52.10
Alignment	protrek_650m	58.45	62.30	54.60	**54.45**	55.00	**53.90**
Alignment	protst_esm2	40.65	43.90	37.40	35.80	39.50	32.10
Alignment	protst_protbert	37.85	39.60	36.10	35.80	39.80	31.80
Alignment	prott3_base	44.60	48.90	40.30	37.00	39.70	34.30
Alignment	pannot_fg	57.95	60.10	55.80	50.75	49.90	51.60
Protein-LLM	prollama	14.60	14.90	14.30	13.85	14.00	13.70
Protein-LLM	bioreason_pro_sft	8.40	8.30	8.50	8.10	9.40	6.80

Table 36: Controlled P2T family-level mean accuracy over test_id and test_ood_family. Attr1 is the rerun disjoint P1/P2/P3 aggregate for comparable individual-task evaluations: PLM task-head QA-bridge scoring on the same Attr1 candidate files, MCQA-style choice for text LLMs/protein-LLMs, and calibrated similarity scoring for PLM-TextAlign/alignment rows. Attr2 and Attr4 isolate the number of displayed attributes; Global-only and Local-only isolate the source of the evidence.

Family	Attr1	Attr2	Attr4	Global-only	Local-only	Full
Random	–	12.50	12.50	12.50	12.50	12.50
PLM	51.93	8.33	8.30	12.97	18.04	6.99
PLM-TextAlign	24.52	11.16	11.86	19.02	34.31	28.26
Text LLM	28.53	17.80	14.26	15.66	14.73	17.63
Alignment	47.03	40.76	36.76	58.64	33.08	47.33
Protein-LLM	20.79	12.29	10.77	15.03	11.25	11.24

four attributes.

Table 37 should be read as a compositional variant comparison, with Attr1 as an individual-task reference rather than the same 8-way caption-choice problem. For PLMs, the Attr1 values are task-head QA-bridge decisions rerun on the same LLM-QA Attr1 candidate files. For text LLMs and protein-LLMs, they are MCQA-style disjoint choices; for PLM-TextAlign/alignment rows, they are calibrated similarity decisions on the same LLM-QA Attr1 files. The remaining "–" entries either do not have an applicable Attr1 similarity protocol or, for prostt5 PLM-TextAlign, failed because empty sequence inputs reached the ProstT5 tokenizer. The PLM rows therefore show a sharp interface gap: the encoders can choose among individual P1/P2/P3 candidates with their task heads, but those marginal signals do not compose into reliable caption selection. The visible drop from Attr1 to Attr2/Attr4 for text LLMs and prollama similarly shows that single-task choice accuracy does not by itself imply compositional caption selection. For bioreason_pro_sft, both Attr1 and the compositional variants are weak, so the table gives less evidence of a composition-specific degradation. The Full column restores the original full P1/P2/P3 caption, so it tests whether the model can exploit all available redundant global and local evidence.

The global-only/local-only ablation reveals that different model families are using different biological signals. Alignment models prefer global P1 labels: their mean accuracy is 58.64% on Global-only but 33.08% on Local-only. This suggests that these protein-text compatibility models are especially strong at matching sequence-level representations to global functional descriptions. PLM-TextAlign shows the opposite pattern: it is 34.31% on Local-only, 28.26% on Full, and 19.02% on Global-only. Raw PLM bridges also improve on Local-only. This is consistent with local descriptors such as domains, transmembrane segments, bind-

Table 37: Controlled P2T held-out accuracy by model. Each variant has two columns: `test_id` and `test_ood_family`. Attr1 is an individual-task reference: PLM rows use the rerun task-head QA-bridge scoring on the same LLM-QA Attr1 candidate files, text LLM/protein-LLM rows use MCQA-style choice, and PLM-TextAlign/alignment rows use calibrated similarity scoring on the same LLM-QA Attr1 files. Full is the original full P1/P2/P3 compositional P2T benchmark.

Family	Model	Attr1		Attr2		Attr4		Global-only		Local-only		Full	
		ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
Random	analytic_uniform_random	–	–	12.50	12.50	12.50	12.50	12.50	12.50	12.50	12.50	12.50	12.50
PLM	ankh	44.03	41.98	7.55	7.45	6.55	7.55	12.55	13.60	13.05	12.30	5.25	6.00
PLM	esm2	62.09	65.00	8.10	7.85	7.95	8.95	11.50	14.05	26.50	26.95	9.40	9.10
PLM	progen	50.62	48.98	5.15	4.65	5.55	7.45	12.75	14.25	20.40	19.25	7.35	6.95
PLM	prostt5	50.73	51.98	12.95	12.90	10.65	11.75	11.75	13.30	12.05	13.80	6.35	5.50
PLM-TextAlign	ankh	25.36	22.88	9.75	9.75	9.55	10.30	15.45	14.30	29.40	31.20	24.55	23.80
PLM-TextAlign	esm2	23.99	23.75	14.20	13.70	13.35	16.30	28.30	25.30	39.10	42.40	38.00	34.80
PLM-TextAlign	progen	26.69	24.46	8.40	8.35	9.05	9.80	14.75	13.40	28.65	29.95	20.75	21.40
PLM-TextAlign	prostt5	–	–	13.05	12.05	11.95	14.60	20.15	20.55	36.00	37.75	32.40	30.40
Text LLM	llama31_8b	27.71	30.17	22.40	23.20	17.85	16.25	21.55	22.10	17.40	19.65	25.95	27.55
Text LLM	qwen25_7b	25.52	29.54	17.60	18.75	13.00	12.30	11.20	12.50	10.10	10.90	13.65	13.50
Text LLM	qwen3_4b	28.50	29.72	11.75	13.10	13.20	12.95	12.45	14.15	14.25	16.10	12.95	12.15
Alignment	protrek_35m	53.68	52.43	53.15	51.30	45.50	47.75	75.45	70.85	42.95	43.50	60.50	54.20
Alignment	protrek_650m	51.72	50.12	53.65	51.40	42.70	45.90	73.40	70.20	43.20	45.15	58.45	54.45
Alignment	protst_esm2	44.32	39.78	35.55	29.95	29.15	31.50	48.10	48.05	26.15	26.85	40.65	35.80
Alignment	protst_protbert	41.71	43.44	32.35	29.80	26.05	31.70	45.80	44.65	25.15	25.05	37.85	35.80
Alignment	prott3_base	40.80	39.75	27.05	28.00	25.70	27.95	44.95	45.10	17.35	18.40	44.60	37.00
Alignment	pannot_fg	50.67	55.91	49.95	46.95	42.70	44.50	70.35	66.75	42.15	41.10	57.95	50.75
Protein-LLM	prollama	26.20	29.96	11.65	11.75	12.90	11.10	21.85	24.90	18.60	19.20	14.60	13.85
Protein-LLM	bioreason_pro_sft	13.82	13.18	12.95	12.80	10.15	8.95	6.35	7.00	3.95	3.25	8.40	8.10

ing sites, and disorder being more directly encoded in sequence representations than abstract global function labels. Text LLMs do not show a large global/local preference; their best setting is still the full caption, which supplies richer language context.

The worked instance in Table 31 gives a concrete interpretation. The gold protein is described by global RNA/mRNA binding labels and by local evidence such as compositional-bias regions, disorder, and an RRM domain. A Global-only projection asks whether the model can choose the caption whose functional labels match the protein after local evidence has been removed. A Local-only projection asks whether the model can choose the caption from local evidence alone, after RNA/mRNA binding has been removed. The results above imply that alignment models are better at the first problem, whereas PLM-TextAlign and raw PLM bridges extract more signal from the second. Attr2 and Attr4 ask a third question: how much performance remains when the caption is restricted to exactly two or exactly four attributes. The drop from Attr2 to Attr4 for alignment and text LLM models shows that more controlled attributes increase the compositional burden, but the recovery on Full shows that redundant biological cues can offset that burden.

F.5 Cross-Family P2T Diagnostics

Accuracy alone does not determine whether a model is compositionally meaningful. A model can be low-accuracy because it chooses biologically plausible hard negatives, or because it collapses to no-overlap captions, a fixed answer position, or the shortest option. Table 38 therefore aggregates the held-out behavior by family. The rightmost diagnostic columns are computed on `test_id`.

The diagnostic summary sharpens the interpretation of Table 35. Alignment models are not only more accurate; when they fail, their wrong answers are usually near the correct composition. Their no-overlap error rate is only 5.48%. PLM-TextAlign has a similar near-miss profile: it remains less accurate than the alignment models, but it rarely collapses to no-overlap captions and has almost no shortest-caption shortcut. In contrast, PLM bridges combine low accuracy with a high no-

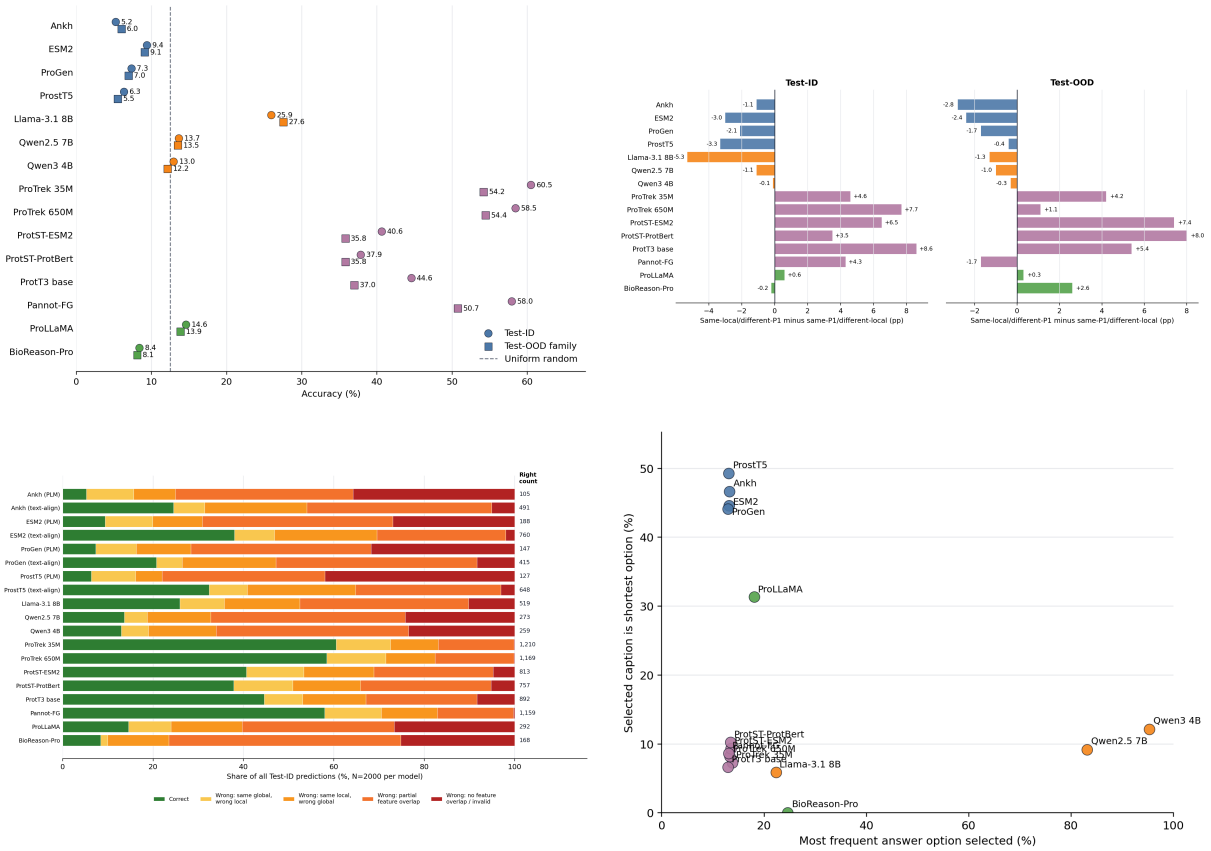


Figure 9: Core P2T compositional behavior visualizations. Top left: held-out accuracy. Top right: difference between global-function and local-evidence diagnostics. Bottom left: relation type of wrong answers on test-id. Bottom right: answer-position and caption-length shortcuts.

overlap rate (36.65%) and a strong shortest-caption bias (45.12%), even though the gold caption is shortest for only approximately 5% of `test_id` rows. Text LLMs occupy an intermediate regime: `llama31_8b` uses enough textual semantics to rise above random, but the Qwen models exhibit severe answer-position concentration. Protein-LLMs do not yet show robust compositional binding in this setting.

Table 39 reports the same diagnostic view at the model level. The taxonomy is not a new metric; it is an interpretive layer over accuracy, wrong-answer composition, answer-position entropy, and caption-length bias. The local-advantage column is defined as same-global/different-local accuracy minus same-local/different-global accuracy. Negative values mean that local-evidence discrimination under a fixed global function is harder than global-function discrimination under fixed local evidence.

The taxonomy shows that the strongest alignment models still have negative local advantage: they are generally better at distinguishing global function when local evidence is held fixed than at distinguishing local evidence when global function is held fixed. `pannot_fg` has the smallest negative local advantage among the alignment systems, consistent with fine-grained supervision improving local evidence binding. This diagnostic is also useful for weaker families: PLM bridges have mildly positive local advantage, but their absolute accuracies are far below random, so the apparent direction of advantage should not be interpreted as successful local reasoning.

F.6 Biology-Conditioned Difficulty

The deep behavior analysis normalizes difficulty by model and split, then asks which biological contexts make fixed-global local-evidence discrimination easier or harder. Table 40 summarizes the most interpretable groups. Positive deltas indicate contexts easier than the corresponding model/split average; negative deltas indicate harder contexts.

These patterns indicate that P2T difficulty is not a generic caption-length effect. Local evidence is easiest when the local category is distinctive, such as transmembrane topology or active-site ev-

Table 38: P2T family-level held-out behavior summary. Hard/partial wrong denotes wrong selections that preserve biologically related global or local attributes.

Family	Test-ID	Test-OOD	Hard / partial wrong	No-overlap wrong	Shortest chosen
Alignment	50.00	44.67	94.52	5.48	9.18
PLM-TextAlign	28.92	27.60	93.76	6.24	3.18
Text LLM	17.52	17.73	77.07	22.22	9.60
Protein-LLM	11.50	10.98	70.64	29.36	15.55
PLM	7.09	6.89	63.35	36.65	45.12

Table 39: Aggregated P2T behavior taxonomy over the completed validation and test splits. Accuracies and behavior rates are percentages.

Family	Model	Acc	Same-local task	Same-global task	Local adv.	Hard-neg. wrong	Behavior label
PLM	ankh	5.58	4.64	6.52	1.88	20.86	length-biased factorized scorer
PLM	esm2	9.18	7.92	10.44	2.52	24.51	length-biased factorized scorer
PLM	progen	7.20	6.32	8.08	1.76	22.59	length-biased factorized scorer
PLM	prostt5	5.78	5.00	6.56	1.56	18.00	length-biased factorized scorer
PLM-TextAlign	ankh	24.54	21.88	27.20	5.33	38.94	ridge-aligned caption matcher
PLM-TextAlign	esm2	36.02	34.65	37.40	2.75	51.04	ridge-aligned caption matcher
PLM-TextAlign	progen	21.46	18.35	24.58	6.23	34.21	ridge-aligned caption matcher
PLM-TextAlign	prostt5	30.65	28.08	33.23	5.15	48.46	ridge-aligned caption matcher
Text LLM	llama31_8b	27.18	25.36	29.00	3.64	36.89	mixed partial matcher
Text LLM	qwen25_7b	13.18	12.96	13.40	0.44	24.58	position-biased generator
Text LLM	qwen3_4b	12.04	12.12	11.96	-0.16	26.40	position-biased generator
Alignment	protrek_35m	57.08	58.88	55.28	-3.60	56.90	strong partial-compositional matcher
Alignment	protrek_650m	55.70	57.76	53.64	-4.12	54.76	balanced partial-compositional matcher
Alignment	protst_esm2	38.42	41.96	34.88	-7.08	48.62	balanced partial-compositional matcher
Alignment	protst_protbert	36.74	39.92	33.56	-6.36	46.29	balanced partial-compositional matcher
Alignment	prott3_base	40.22	43.20	37.24	-5.96	44.40	balanced partial-compositional matcher
Alignment	pannot_fg	54.10	54.96	53.24	-1.72	58.78	strong partial-compositional matcher
Protein-LLM	prollama	14.56	15.00	14.12	-0.88	28.51	mixed partial matcher
Protein-LLM	bioreason_pro_sft	7.88	8.48	7.28	-1.20	17.63	local-first partial matcher

idence. It is hardest for low-complexity/disorder evidence and for cases that combine active-site and ligand-binding semantics. On the global side, enzyme/cofactor-rich contexts increase ambiguity, whereas EC plus GO-MF without dense cofactor language gives models a stronger global anchor.

F.7 Version and Interpretation

The current P2T compositional benchmark is the `core_v3_distinguishable` version. Relative to the completed matched Core-v2 rows, Core-v3 is modestly harder overall: PLM models decrease by 0.98 percentage points, protein-LLMs by 1.63 points, and text LLMs by 0.30 points. The change is largest for the same-global/different-local diagnostic, supporting the intended role of Core-v3 as a sharper local-evidence distinguishability split.

Overall, the compositional results support three conclusions. First, protein-text alignment models are currently the strongest candidates for global/local/evidence binding, especially in P2T, but even they remain sensitive to local-evidence contrasts. Second, generative LLMs can sometimes exploit textual semantic priors, yet answer-distribution diagnostics are necessary because position bias can inflate or obscure apparent accuracy. Third, independent PLM task heads do not compose into reliable caption-level decisions: the four PLM encoders all show below-random held-out accuracy, high no-overlap error rates, and strong shortest-caption bias.

G Worked Biological Example for PannotGround Composition

This appendix gives one source-grounded example that can be used directly for schematic figures explaining PannotGround. The example is `Q08209`, the human calcineurin catalytic subunit alpha (`PP2BA_HUMAN`, gene `PPP3CA`). It is useful because the same Swiss-Prot entry contains a recognizable global function, multiple curated protein-level labels, and many local sequence features. The example therefore shows why a label-only protein benchmark is incomplete: the correct answer is not merely "serine/threonine phosphatase", but the binding of that global function to the right local evidence pattern.

G.1 Exact Source Protein Record

The canonical PannotGround protein record is:

```
record_id: Q08209
accession_id: Q08209
header: sp|Q08209|PP2BA_HUMAN Protein
```

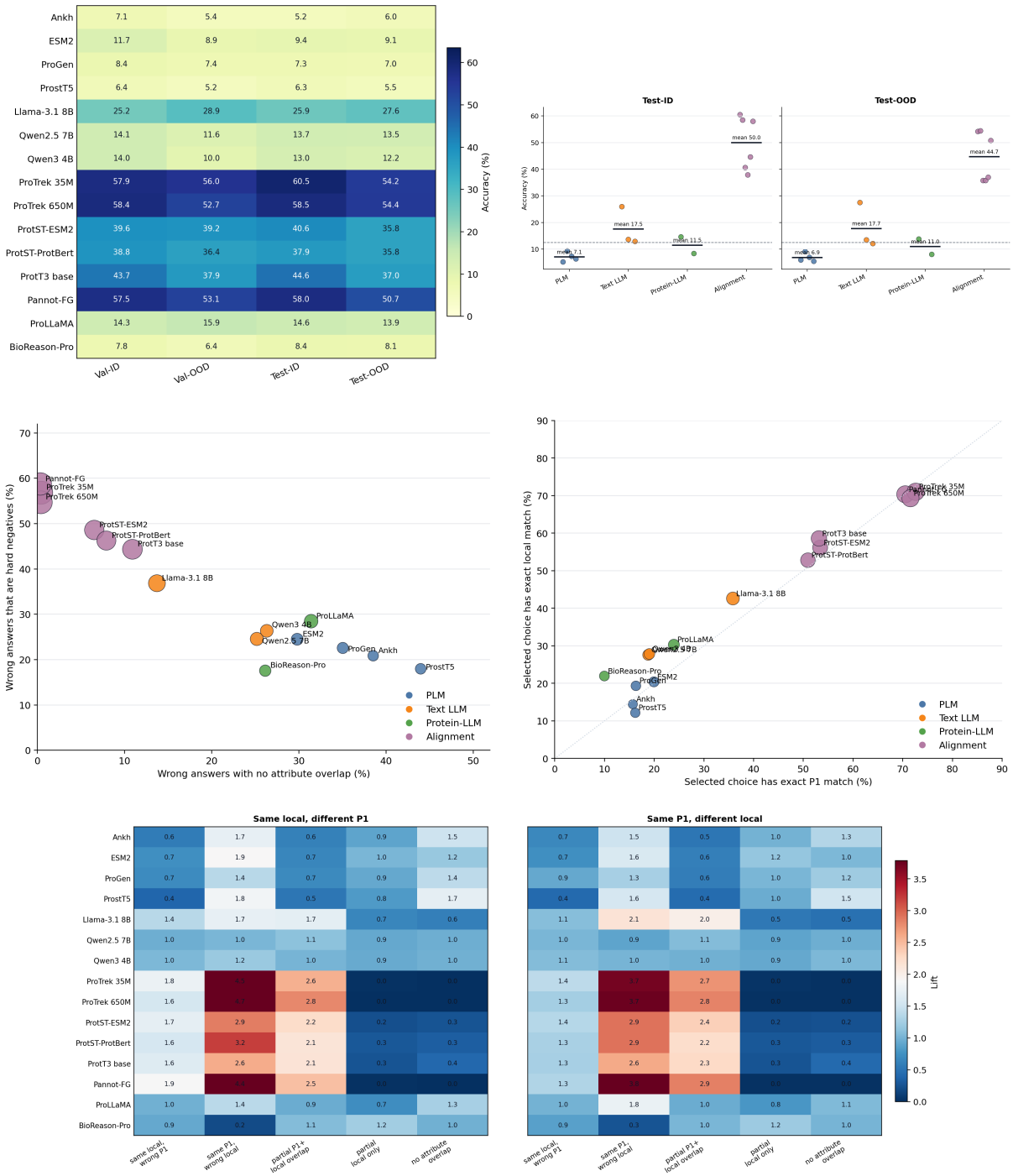


Figure 10: P2T model diagnostics beyond aggregate accuracy. Top left: model-by-split accuracy. Top right: held-out family-level separation. Middle left: error quality measured by hard-negative and no-overlap wrong selections. Middle right: whether selected captions preserve exact global function and exact local evidence. Bottom: relation lift after correcting for candidate-pool availability.

phosphatase 3 catalytic subunit alpha
OS=Homo sapiens OX=9606 GN=PPP3CA
PE=1 SV=1
organism: Homo sapiens
taxon: 9606
gene: PPP3CA
sequence_length: 521
split: test_id
cluster.c30_id: O42773
cluster.c90_id: P48452

The corresponding sequence is:

MSEPKAIDPKLSTTDRVVKAVPFPPSHRLTAKEVFDNDGKPRVDIL
KAHLMKEGRLEESVALRIITEGASILRQEKNLLDIDAPVTVCGDIH
GQFFDLMKLFEVGGSPANTRYLFLGDYVDRGYFSIECVLYLWALKI
LYPKTLFLLRGNHECRHLTEYFTFKQECKIKYSERVYDACMDAFDC
LPLAALMNQQFLCVHGGLSPEINTLDDIRKLDRFKEPPAYGPMCDI
LWSDPLEDFGNEKTQEHFTHNTVRGCSYFYSYPAVCEFLQHNNLLS
ILRAHEAQDAGYRMYRKSQTTGFPSLITIFSAPNYLDVYNNKAAVL
KYENNVMNIRQFNCSPHPYWLPNFMDVFTWSLPFVGEKVTEMLVNV
LNICSDDELGSEEDGFDGATAAARKEVIRNKIRAIGKMARVFSVLR
EESESVLTLKGLTPTGMLPSGVLSGGKQTLQSATVEAIEADEAIKG

Table 40: Biological difficulty patterns for P2T local-evidence discrimination. Deltas are percentage points after model/split normalization.

Grouping	Category	N	Delta	Interpretation
Global namespace	`ec+go_mf`	1,965	+7.29	Enzyme identity plus GO-MF can provide a strong global anchor.
Global namespace	`go_mf`	13,170	+1.66	GO-MF-only examples are modestly easier than average.
Global namespace	`catalytic_activity+cofactor+ec+go_mf`	8,790	-2.52	Dense enzyme/cofactor descriptions make local binding harder.
Global namespace	`cofactor+ec+go_mf`	1,050	-5.61	Cofactor-enriched enzyme contexts are difficult.
Global namespace	`cofactor+go_mf`	375	-7.87	Cofactor language without EC support is the hardest namespace group.
Global function-slim	other function	2,235	+3.46	Less enzyme-like functions are easier on average.
Global function-slim	nucleic-acid binding or regulation	10,290	+1.07	Regulatory/DNA-binding contexts are mildly easier.
Global function-slim	enzyme catalysis	23,970	-0.94	Enzyme contexts are slightly harder after normalization.
Global function-slim	small-molecule or ion binding	21,360	-1.11	Binding/cofactor semantics increase ambiguity.
Global function-slim	structural molecule activity	2,415	-4.88	Structural/ribosomal contexts are difficult for local evidence binding.
Local evidence	transmembrane topology	5,325	+2.34	Topological evidence is relatively distinctive.
Local evidence	active site	2,355	+1.33	Active-site-only cases are easier than average.
Local evidence	domain/motif/region	16,215	+1.04	Domain-like local evidence is mildly easier.
Local evidence	active site plus ligand binding	8,505	-2.63	Combining active-site and ligand-binding semantics is harder.
Local evidence	disorder/bias/coiled-coil	3,915	-3.70	Low-complexity and disorder evidence is the hardest local family.

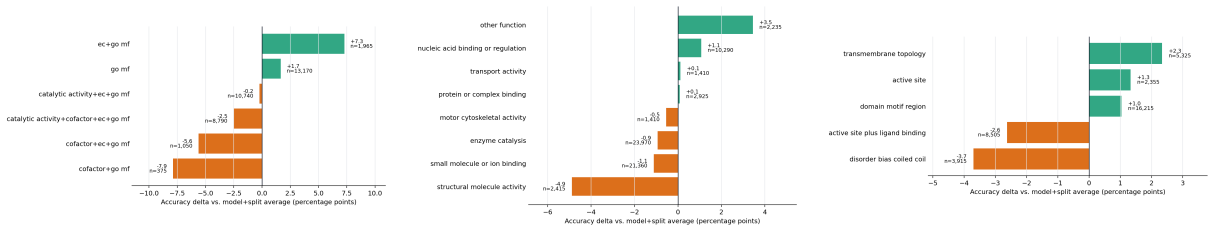


Figure 11: Biology-conditioned P2T difficulty diagnostics. Left: global namespace groups. Center: global function-slim groups. Right: local-evidence families.

```
FSPQHKITSFEEAKGLDRINERMPPRRDAMPSDANLNSINKALTSE
TNGTDSNGSNSSNIQ
```

The source biology is: calcium-dependent, calmodulin-stimulated protein phosphatase activity; dephosphorylation of phosphoserine and phosphothreonine protein substrates; Fe(3+) and Zn(2+) cofactors; and participation in calcineurin-NFAT signaling, calcium signaling, T-cell activation, neuronal signaling, kidney function, and muscle-related cellular contexts.

G.2 Source Annotations Projected into P1, P2, and P3

For `Q08209`, the canonical annotation file contains 90 source annotations: 44 GO biological-process labels, 13 GO cellular-component labels, 7 GO molecular-function labels, 11 region annotations, 9 site annotations, 2 catalytic-activity labels, 2 co-

factor labels, 1 EC label, and 1 subcellular location label. The figure should not print all 90 annotations. It should show that these source annotations are projected into three benchmark views: protein-level global labels (P1), local feature-type recognition (P2), and residue/span grounding (P3).

The most informative P1 global labels are:

- EC: `3.1.3.16`.
- Catalytic activity: O-phospho-L-seryl-[protein] + H2O = L-seryl-[protein] + phosphate.
- Catalytic activity: O-phospho-L-threonyl-[protein] + H2O = L-threonyl-[protein] + phosphate.
- Cofactors: Fe(3+) and Zn(2+), each with one ion per subunit.
- GO molecular function: protein serine/threonine phosphatase activity (GO:0004722), calcium ion binding (GO:0005509), calmodulin binding (GO:0005516), calmodulin-dependent protein phosphatase activity (GO:0033192), enzyme binding (GO:0019899), protein dimerization activity (GO:0046983), and ATPase binding (GO:0051117).

The local evidence used for P2/P3 is shown in Table 41. These are the coordinates that should be drawn on the protein track in a framework figure.

The released task-view rows for this protein are:

- P1: `PGV0:P1:Q08209`. The model predicts global labels from the namespaces `catalytic_activity`, `cofactor`, `ec`, `go_bp`, `go_cc`, `go_mf`, and `subcellular_location`.
- P2: `PGV0:P2:Q08209`. The gold feature types are `active_site`, `compositional_bias_region`, `disordered_region`, `ligand_or_cofactor_binding_site`, `motif`, `region_of_interest`, and `site`.
- P3: fourteen grounding queries locate the feature details in Table 41, such as `region_of_interest:Catalytic` at 56–340, `active_site:Protondonor` at 151, and `motif:SAPNYmotif` at 307–311.

G.3 Exact P2T Composition Row

The primary worked P2T row is:

```
question_id: PGCOMP:test_id:01026:Q08209
record_id: Q08209
task_family: P123_compositional_caption_selection
input_control: full
subtask: same_p1_different_local
sequence_length: 521
gold_answer: H
choice_count: 8
gold composition hash: 8e7b50a7745e84e4
```

The gold composition keeps a compact global-function signature and a coordinate-free local signature:

- Global identity: `ec:3.1.3.16`.
- Stable global labels: Fe(3+) cofactor, Zn(2+) cofactor, EC 3.1.3.16, GO:0004722, GO:0005509, GO:0005516, GO:0019899, GO:0033192, GO:0046983, and GO:0051117.
- Local feature types: active site, compositional-bias region, disordered region, ligand/cofactor binding site, motif, region of interest, and site.
- Local evidence descriptors: basic/acidic residues, low complexity, polar residues, disorder, PxIxIF substrate-interaction motif, SAPNY motif, autoinhibitory domain, autoinhibitory segment, calcineurin B binding, calmodulin-binding, catalytic region, proton donor, binding site, and PxVP substrate-interaction site.

The exact gold caption rendered in the benchmark is:

> Global function: EC: EC 3.1.3.16; GO molecular function: protein serine/threonine phosphatase activity (GO:0004722); calcium ion binding (GO:0005509); calmodulin binding (GO:0005516); +4 more; catalytic activity: O-phospho-L-seryl-[protein] + H2O = L-seryl-[protein] + phosphate; The enzyme mediates the catalytic process described as: O-phospho-L-threonyl-[protein] + H2O = L-threonyl-[protein] +...; cofactor: The enzyme functions with Fe(3+) as cofactors, and binds 1 Fe(3+) ion per subunit; This enzyme utilizes Zn(2+) for catalysis, and binds 1 zinc ion per subunit. Local feature types: active site; compositional bias region; disordered region; ligand/cofactor binding site; motif; region of interest; +1 more. Local evidence queries without coordinates: compositional bias region: Basic and acidic residues; compositional bias region: Low complexity; compositional bias region: Polar residues; disordered region: Disordered; motif: Interaction with PxIxIF motif in substrate; motif: SAPNY motif; +8 more.

Table 41: Curated local source annotations for Q08209. These rows instantiate P2 feature-type recognition and P3 residue/span grounding.

Annotation ID	Canonical type	Feature detail	Coordinate
PGV0:cd53e241cea360c2	region of interest	Catalytic	56–340
PGV0:adf2757536175a60	ligand/cofactor binding site	binding site	90
PGV0:cee8dd78fac425e5	ligand/cofactor binding site	binding site	92
PGV0:8abbd8de76266053	ligand/cofactor binding site	binding site	118
PGV0:fcaaedcbed7cc230	ligand/cofactor binding site	binding site	118
PGV0:acab5205dc23cee6	ligand/cofactor binding site	binding site	150
PGV0:5a65c8cb17044cd0	active site	Proton donor	151
PGV0:337aa2b4747982e2	ligand/cofactor binding site	binding site	199
PGV0:715a9fa17b26c15b	ligand/cofactor binding site	binding site	281
PGV0:aa0719ab2d14e065	motif	SAPNY motif	307–311
PGV0:339bb3b992a03017	motif / region of interest	Interaction with PxIxIF motif in substrate	327–336
PGV0:6c1f55183bb2cfdf	region of interest	Calcineurin B binding	341–369
PGV0:4a377a0436de6192	site	Interaction with PxVP motif in substrate	352
PGV0:1acb7481d7d67d28	region of interest	Calmodulin-binding	392–406
PGV0:780bfdcace3fe020	region of interest	Autoinhibitory segment	407–414
PGV0:b5b52298e1454b68	region of interest	Autoinhibitory domain	465–487
PGV0:b2ac0bcabf7cb540	compositional-bias region	Basic and acidic residues	475–490
PGV0:43e6fa9c004846df	disordered region	Disordered	475–521
PGV0:0dd1669445bb2c4e	compositional-bias region	Polar residues	494–503
PGV0:c67a752a39b7fd0a	compositional-bias region	Low complexity	504–521

Table 42 lists the exact candidate proteins in this row and gives a diagnostic interpretation of each option. The stored row-level distractor type records how the candidate was constructed, whereas the diagnostic relation describes its relationship to the Q08209 gold composition after comparing the global identity, full local signature, and overlapping attributes. In the rendered captions, statements introduced by "cofactor:" belong to the global-function component. In contrast, "ligand/cofactor binding site" is a local feature type, because it refers to residue-level binding-site evidence rather than to the protein-level cofactor annotation itself.

This row contains one exact gold choice, two same-primary-global/wrong-local choices, several partial-overlap choices, and one no-overlap choice. It therefore demonstrates why the compositional benchmark is stricter than a label-only evaluation. A model that recognizes only the primary EC identity could assign high confidence to choices A or C, since both retain an EC 3.1.3.16 phosphatase identity. These choices are nevertheless incorrect: their local evidence does not match the Q08209 source record. Choice A further clarifies the global/local separation. Its Mg/Mn statement is a global cofactor annotation, while its local component is the PPM-type-phosphatase, disorder, compositional-bias, and binding-site evidence pattern.

G.4 Specific Candidate for the Opposite Relation Direction

The Q08209 gold row is a same_p1_different_local question, so its candidate set is designed to emphasize same-global/wrong-local errors. The opposite hard error, same-local/wrong-global, appears in the companion held-out row PGCOMP:test_id:00005:Q8H4S6. This companion row should be used if the figure needs an exact candidate for the other direction.

In that row, the gold choice is E for protein Q8H4S6:

> Global function: EC: EC 3.1.3.16; GO molecular function: protein serine/threonine phosphatase activity (GO:0004722); metal ion binding (GO:0046872); catalytic activity: O-phospho-L-threonyl-[protein] + H2O = L-threonyl-[protein] + phosphate; This enzyme performs the catalytic reaction represented as: O-phospho-L-seryl-[protein] + H2O = L-seryl-[protein] + p...; cofactor: This enzyme makes use of Mg(2+) and Mn(2+) for catalysis, and binds 2 magnesium or manganese ions per subunit. Local feature types: disordered region; domain; ligand/cofactor binding site. Local evidence queries without coordinates: disordered region: Disordered; domain: PPM-type phosphatase; ligand/cofactor binding site: binding site.

The exact same-local/wrong-global candidate is choice G, protein B4G653:

> Global function: GO molecular function: protein serine/threonine phosphatase activity (GO:0004722); According to the information, this protein is responsible for metal ion binding (GO:0046872); catalytic activity: O-phospho-L-seryl-[protein] + H2O = L-seryl-[protein] + phosphate; O-phospho-L-threonyl-[protein] + H2O = L-threonyl-[protein] + phosphate; cofactor: Mg(2+) and Mn(2+) serve as vital cofactors for the enzyme. Local feature types: disordered region; domain; ligand/cofactor binding site. Local evidence queries without coordinates: disordered region: Disordered; domain: PPM-type phosphatase; ligand/cofactor binding site: binding site.

Table 42: Exact P2T choices for PGCOMP:test_id:01026:Q08209.

Choice	Protein	Stored type	Diagnostic relation	What the candidate says
A	Q8N819	same_p1_wrong_local	Same primary global, wrong local	EC 3.1.3.16 serine/threonine phosphatase. Its Mg/Mn cofactor statement is a global cofactor label, while its local evidence is a PPM-type phosphatase/disorder/binding-site pattern rather than the calcineurin local pattern.
B	P81373	both_wrong	Partial local overlap	Electron-transfer, metal-ion, and 2Fe-2S cluster binding protein with a 2Fe-2S ferredoxin-type domain and a generic ligand/cofactor binding site. It shares only generic binding-site evidence with the gold row.
C	Q12378	same_p1_wrong_local	Same primary global, wrong local	EC 3.1.3.16 phosphatase-like global identity, but the local evidence is an RTR1-type motif plus binding site, not the calcineurin catalytic, SAPNY, PxIxIF, calcineurin B, calmodulin-binding, and autoinhibitory pattern.
D	Q9FIU7	random_wrong	Partial local overlap	Beta-glucosidase (EC 3.2.1.21). The global function is wrong, but it still shares local descriptors such as active site, proton donor, and ligand/cofactor binding site.
E	Q6BZD9	random_wrong	Partial local overlap	DNA helicase (EC 5.6.2.3) with ATPase/helicase labels and 4Fe-4S cofactor. It shares motif and ligand/cofactor-binding feature categories, but the global function and motif details are wrong.
F	Q9C7S1	random_wrong	No feature overlap	Actin-binding FH2-domain protein. It does not share the gold global identity or the selected local evidence descriptors.
G	P17157	both_wrong	Partial local overlap	Cyclin-dependent protein kinase (EC 2.7.11.22). The global function is a kinase rather than a phosphatase, but the candidate partially overlaps through active-site, ligand/cofactor-binding-site, and generic site categories.
H	Q08209	gold	Correct	The only choice that binds EC 3.1.3.16 calcineurin-like global function to the Q08209 local evidence pattern: catalytic region, Fe/Zn binding sites, proton donor, SAPNY, PxIxIF, PxVP, calcineurin B binding, calmodulin-binding, autoinhibition, disorder, and compositional-bias regions.

The local side is exactly the same as the gold row: disordered_region, domain, and ligand_or_cofactor_binding_site, with P3 evidence descriptors Disordered, PPM-typephosphatase, and bindingsite. The global side differs because the full P1 identity and stable global composition do not match the gold row. This is the exact example to draw for "same local evidence, wrong global function".

G.5 Figure Guidance from This Example

For a general framework figure, draw the Q08209 sequence as a 521-residue protein track. Above the track, show global labels: EC 3.1.3.16, serine/threonine phosphatase activity, calcium/-calmodulin binding, Fe(3+) and Zn(2+) cofactors, and calcineurin-NFAT/T-cell signaling. On the track, draw the local evidence from Table 41. Then show three benchmark outputs: global label prediction, local feature-type prediction, and residue/span grounding.

For the composition-contribution figure, contrast two evaluation styles. A label-only benchmark asks whether the sequence has a global label such as "EC 3.1.3.16 phosphatase". PannotGround asks whether the model binds that global function to the correct local evidence. The figure should explicitly show candidate swaps:

- Correct: H:Q08209.
- Same global, wrong local: A:Q8N819 and C:Q12378.
- Partial feature overlap: B:P81373, D:Q9FIU7, E:Q6BZD9, and G:P17157.
- No feature overlap: F:Q9C7S1.
- Same local, wrong global: use the companion row PGCOMP:test_id:00005:Q8H4S6, choice G:B4G653.

This is the main contribution illustrated by the example: the benchmark can distinguish a model that recognizes a global phosphatase label from a model that correctly binds that label to catalytic, metal-binding, substrate-motif, calmodulin-binding, autoinhibitory, and disordered local evidence in the same protein.

H Detailed Fine-grained Pre-training algorithm for Protein–Text Alignment

H.1 Common Notation

Let a protein sequence be denoted by x of length L, and its residue-level embeddings produced by a pretrained protein language model (PLM, e.g., ESM) be

$$H = f_P(x) \in \mathbb{R}^{L \times d_P}, \quad H_i \in \mathbb{R}^{d_P}$$

where $\mathbb{H}_{\square}$ denote the embedding of residue i.

A textual description y (e.g., functional annotation) is encoded as

$$h_T = f_T(y)_{\texttt{[CLS]}} \in \mathbb{R}^{d_T},$$

where the [CLS] token representation is used.

The Q-former(F_Q) contains two parts, they are protein cross-modality module($F_{Q,prot}$) and text encoding module($F_{Q,text}$). In the protein cross-modality module, soft queries(learnable embeddings) are processed through self-attention layer and then apply their attention on protein embeddings through a cross attention layer. After the protein embeddings are gathered in the cross attention module, the embeddings would be processed by the feed-forword network. While for the text encoding module, the architecture is simple text encoder, i.e. PubMedBert.

The Q-Former consists of N_q learnable query tokens $Q \in \mathbb{R}^{N_q \times d_Q}$ that cross-attend to the protein encoder outputs H:

$$Z = F_{Q,prot}(Q, H) \in \mathbb{R}^{N_q \times d_Q}.$$

Or if the attention of protein and text is bidirectional we could write Z as:

$$Z = F_Q(Q, H, y) \in \mathbb{R}^{N_q \times d_Q}. \tag{50}$$

If the attention is causal from protein to text, we could write:

$$y = F_{Q,causal}(Q, H, y_{1:T-1}) \tag{51}$$

H.2 Baseline: BLIP2 Stage–1 Objectives

Let a protein sequence be denoted by x of length L, and a textual description by y. The protein encoder (PLM) produces contextual residue embeddings

$$H = f_P(x) \in \mathbb{R}^{L \times d_P}, \tag{52}$$

while the text encoder (e.g., PubMedBERT) outputs a sequence of token representations, from which we take the [CLS] token as

$$h_T = F_{Q,text}(y)_{\texttt{[CLS]}} \in \mathbb{R}^{d_T}. \tag{53}$$

The Q-Former employs N_q learnable query tokens $Q \in \mathbb{R}^{N_q \times d_Q}$ and attends to the protein embeddings:

$$Z = F_{Q,prot}(Q, H) \in \mathbb{R}^{N_q \times d_Q}. \tag{54}$$

Or if the attention on the text is not blocked (bidirectional) we could write Z as:

$$Z = F_Q(Q, H, y) \in \mathbb{R}^{N_q \times d_Q}. \tag{55}$$

Protein–Text Contrastive (PTC) loss. For each protein–text pair (x_i, y_i), we first individually calculate $Z = F_{Q,prot}(Q, H) \in R^{N_q d_Q}$ and h_T, then compute the cosine similarity as the maximum query–text similarity:

$$\mathcal{L}_{\text{PTC}} = \frac{1}{2B} \sum_{i=1}^{B} \Big[-\log \frac{\exp(s_{ii}/\tau)}{\sum_j \exp(s_{ij}/\tau)} - \log \frac{\exp(s_{ii}/\tau)}{\sum_j \exp(s_{ji}/\tau)} \Big] \tag{56}$$

where τ is a temperature parameter.

Protein–Text Matching (PTM). In the matching scenario, the attention is bidirectional between protein and text modalities. Given the query outputs Z and text features, a binary classifier distinguishes matched and mismatched pairs:

$$\mathcal{L}_{\text{PTM}} = \text{CE}(\,\text{MLP}_{\text{clf}}(F_Q(Q, H, y)),\, \text{label} \in \{0, 1\})\,. \tag{57}$$

Language Modeling (LM). In the language modeling, there is causal attention from protein to text. When the Q-Former is used to condition a text decoder, the LM loss is

$$\mathcal{L}_{\text{LM}} = \mathcal{L}_{NTP}\big(F_{Q,causal}(Q, H, y_{1:T-1}),\, y_{2:T}\big). \tag{58}$$

Baseline objective.

$$\boxed{\mathcal{L}_{\text{BLIP2-base}} = \mathcal{L}_{\text{PTC}} + \lambda_{\text{PTM}}\mathcal{L}_{\text{PTM}} + \lambda_{\text{LM}}\mathcal{L}_{\text{LM}}.} \tag{59}$$

H.3 Strategy 1: Local Contrastive Learning (ESM-only Pooling)

Let $R \subseteq \{1, \ldots, L\}$ be annotated residues and $\tilde{R}$ be annotated the rest of annotated residues. We pool the PLM embeddings and normalize:

$$H_R = \mathcal{M}_{\tilde{R}}(H) \tag{60}$$

If you want only the mean pooling of the annotated embeddings, you could set

$$H_R = \text{norm}\Big(\frac{1}{|R|}\sum_{i \in R} H_i\Big) \tag{61}$$

Under this local setup, we could have Q The local matching loss is

$$\boxed{\mathcal{L}_{\text{local-ptm}} = \mathcal{L}_{\text{PTM}}(F_Q(Q, H_R, y)).} \tag{62}$$

$$\boxed{\mathcal{L}_{\text{local-lm}} = \mathcal{L}_{\text{lm}}(F_{Q,causal}(Q, H_R, y)).} \quad (63)$$

Total objective:

$$\mathcal{L}^{(1)} = \mathcal{L}_{\text{BLIP2-base}} + \lambda_{\text{loc}}(\mathcal{L}_{\text{local-ptm}} + \mathcal{L}_{\text{local-lm}}). \quad (64)$$

H.4 Strategy 2: Masked-Residue Counterfactual Negatives

To capture residue necessity, the annotated residues R are masked:

$$\widetilde{H} = \mathcal{M}_R(H), \quad (65)$$

$$z_P^+ = F_{Q,prot}(Q, H), \quad z_P^- = F_{Q,prot}(Q, \widetilde{H}). \quad (66)$$

A margin-based counterfactual loss is defined as

$$\boxed{\mathcal{L}_{\text{cf}} = \mathbb{E}\big[\max\big(0, m + s(z_P^-, h_T) - s(z_P^+, h_T)\big)\big].} \quad (67)$$

The total loss is

$$\mathcal{L}^{(2)} = \mathcal{L}_{\text{BLIP2-base}} + \lambda_{\text{cf}}\mathcal{L}_{\text{cf}}. \quad (68)$$

I Appendix: Pretraining Experiments Details

I.1 Metrics

We evaluate models along two complementary dimensions: global protein–text alignment and fine-grained evidence localization. While global retrieval metrics quantify semantic correspondence at the sequence level, localization metrics explicitly measure whether predictions are grounded in correct biochemical evidence.

I.1.1 Global Protein–Text Retrieval

We first evaluate global alignment between protein and text representations using bidirectional retrieval metrics. Let $\mathcal{P}$ denote a set of protein sequences and $\mathcal{T}$ their corresponding textual descriptions. Both modalities are encoded using a shared Q-Former architecture into a joint embedding space.

We consider two retrieval directions:

- **Protein-to-Text (P2T):** given a protein embedding, retrieve its corresponding text description.
- **Text-to-Protein (T2P):** given a text embedding, retrieve its corresponding protein.

For each direction, we report results under two settings:

- **Fullset retrieval**, where each query is ranked against all candidates in the evaluation split.
- **In-batch retrieval**, where candidates are restricted to the current mini-batch.

We report the following metrics:

- **Top-1 Accuracy (Acc):** the fraction of queries for which the correct match is ranked first.
- **Recall@20 (Rec@20):** the fraction of queries for which the correct match appears among the top 20 retrieved candidates.

These metrics primarily measure global semantic alignment between protein and text representations, but do not capture whether predictions rely on correct localized biochemical evidence.

I.1.2 Fine-Grained Evidence Localization (Local MC)

To explicitly evaluate evidence grounding, we introduce a *local multi-choice* (loc_mc) localization task that probes whether the model can identify the correct protein location associated with a given textual evidence description.

Task formulation. For each annotated protein–text pair, we construct a candidate set of protein locations

$$\mathcal{L} = \{\ell_1, \ell_2, \ldots, \ell_K\},$$

where $K = 8$ by default. Each candidate ℓ_k corresponds to a plausible biochemical location on the protein, which may represent an individual residue, a contiguous residue span, or a predefined functional region.

Exactly one candidate corresponds to the ground-truth annotated evidence, while the remaining candidates are randomly sampled distractor locations. Under this setup, random guessing yields an expected accuracy of $1/K$.

Scoring and selection. Each candidate location ℓ_k is encoded using the protein-side tower, while the corresponding textual evidence is encoded using the text-side tower. Localization is performed by computing a similarity score between the text embedding and each candidate location embedding.

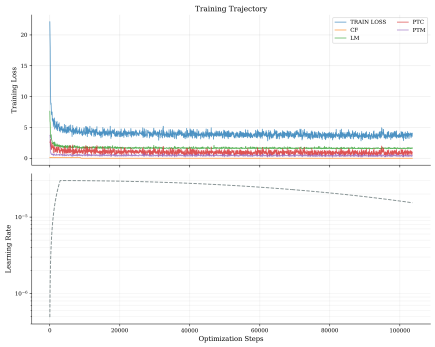


Figure 12: Behavior of training loss during stage 1: Q-Former training

Candidates are ranked according to these similarity scores, and the highest-scoring location is selected.

Localization metrics are computed per annotation and averaged across all annotations associated with a protein.

Metrics. We report three complementary localization metrics:

- **MC Accuracy ($\mathbf{Acc_{loc}}$):** the fraction of instances for which the ground-truth location is ranked first.
- **MC Mean Reciprocal Rank ($\mathbf{MRR_{loc}}$):**

$$\mathrm{MRR}_{\mathrm{loc}} = \mathbb{E}\left[\frac{1}{\mathrm{rank}(\ell^*)}\right],$$

  where ℓ^* denotes the ground-truth location.
- **MC Recall@5 ($\mathbf{Rec@5_{loc}}$):** the fraction of instances for which the ground-truth location appears among the top 5 ranked candidates.

While global retrieval metrics assess whether protein and text representations are aligned at the sequence level, strong retrieval performance does not guarantee that predictions depend on correct localized evidence. The proposed localization metrics explicitly evaluate *where* the model grounds its reasoning on the protein sequence. Improvements in loc_mc metrics therefore reflect enhanced evidence-dependent reasoning, rather than reliance on global correlations or memorization.

I.2 Training Behavior

We could notice from the figure 12 that the loss is stably decreasing with descending learning rate. The learning rate is largest learning rate is $3e-5$ and it has warm-up period for the first 3000 steps.

I.3 Strategy Ablation Results

Table 43 studies the effect of weighting the local contrastive objective (λ_{local}) and the counterfactual masking objective (λ_{cf}) during fine-grained pretraining. When both objectives are removed, localization performance degrades substantially, indicating that global supervision alone fails to induce representations that reliably encode causal evidence at the residue or span level. Introducing either objective in isolation improves localization quality, suggesting that each objective imposes a complementary inductive bias. Specifically, increasing λ_{local} primarily improves top-ranked localization accuracy and mean reciprocal rank, reflecting sharper alignment between evidence representations and textual descriptions. In contrast, increasing λ_{cf} consistently improves recall-oriented metrics, indicating enhanced coverage of alternative plausible evidence under perturbations. Combining both objectives yields the strongest overall localization performance, while preserving protein-to-text retrieval accuracy on held-out test data, demonstrating that fine-grained supervision enhances evidence grounding without degrading global semantic alignment.

I.4 Hyperparameter ablation

Table 44 reports an ablation study of fine-grained pretraining strategies across multiple datasets with varying sequence diversity and homology levels. Strategy 1 (local contrast) explicitly aligns localized protein evidence with textual descriptions, while Strategy 2 (counterfactual masking) enforces causal dependence on evidence by contrasting masked and unmasked variants. When neither strategy is applied, localization performance degrades consistently across all datasets, indicating that global supervision alone does not induce robust evidence-aware representations. Enabling either strategy in isolation improves localization metrics, with Strategy 1 favoring precision-oriented measures such as Accloc and MRRloc, and Strategy 2 improving recall-oriented behavior. Combining both strategies (code 12) yields the strongest and most consistent improvements, particularly on datasets with higher redundancy or structural similarity, while preserving retrieval performance. These results suggest that the two objectives impose complementary inductive biases that generalize across data regimes.

Run	λ_{local}	λ_{cf}	$\text{Acc}_{\text{loc}} \uparrow$	$\text{MRR}_{\text{loc}} \uparrow$	$\text{Rec@5}_{\text{loc}} \uparrow$	P2T Acc_{test}	P2T $\text{Rec@20}_{\text{test}}$
0	0.25	0.25	37.14	0.5015	60.41	15.94	80.08
1	0.50	0.25	37.96	0.5047	65.31	15.54	79.68
2	0.25	0.50	38.78	0.5115	64.49	15.94	80.48
3	0.50	0.50	37.55	0.5113	66.12	15.94	79.28
4	0.75	0.50	**42.45**	**0.5350**	67.76	15.54	78.88
5	0.50	0.75	41.35	0.5240	67.25	15.94	78.88
6	0.75	0.75	40.00	0.5299	**68.16**	15.94	80.08
abl	0.25	0.00	35.51	0.4552	48.98	15.54	80.88
abl	0.00	0.25	31.43	0.4462	58.37	15.94	80.88
abl	0.00	0.00	30.20	0.4190	46.53	**16.33**	80.08

Table 43: Ablation of fine-grained objective weights. Acc_{loc}, MRR_{loc}, and $\text{Rec@5}_{\text{loc}}$ measure evidence localization quality on the validation set. Protein-to-text (P2T) retrieval metrics are reported on the test set.

Strat.	P2T Acc	P2T Rec@20	T2P Acc	T2P Rec@20	InB P2T Acc	InB T2P Acc	Acc_{loc}	MRR_{loc}	$\text{Rec@5}_{\text{loc}}$
act_X									
12	1.36	20.68	1.59	22.73	3.07	3.07	44.09	0.588	76.25
1	1.36	21.25	1.59	22.73	3.07	3.07	42.16	0.546	65.00
2	1.48	20.57	1.59	22.61	3.07	3.07	43.30	0.558	67.05
0	1.25	21.59	1.48	22.84	3.07	3.07	41.48	0.544	68.41
act_mf50									
12	33.87	96.24	36.02	96.77	60.22	61.29	58.70	0.722	91.85
1	34.95	96.77	34.95	96.77	60.75	60.75	52.72	0.676	90.22
2	35.48	96.77	34.95	96.77	61.83	61.29	52.17	0.658	85.87
0	34.95	96.77	36.02	96.77	61.29	60.75	46.74	0.620	80.98
act_mp50									
12	21.04	80.60	21.04	80.87	50.27	51.09	59.83	0.723	91.29
1	21.31	81.15	21.58	81.15	51.09	51.09	49.16	0.633	82.87
2	21.58	80.60	21.31	80.87	50.27	51.09	56.18	0.696	89.04
0	21.31	81.15	21.58	81.15	51.09	51.09	49.16	0.633	82.87
act_mp90									
12	9.67	77.36	9.78	77.80	54.84	55.38	56.98	0.691	84.59
1	21.58	80.60	21.31	80.87	50.27	51.09	56.18	0.696	89.04
2	9.34	77.03	10.11	77.69	54.18	55.27	55.88	0.676	83.48
0	9.34	77.25	10.00	77.91	54.29	55.38	47.89	0.622	81.26

Table 44: Dataset-wise ablation of fine-grained pretraining strategies. Strategy codes indicate which objectives are enabled: 0 = none, 1 = local contrast only, 2 = counterfactual masking only, 12 = both. When enabled, each strategy uses a default weight $\lambda = 0.25$.

Task	Split	Total Loss ↓	LM ↓	PTC ↓	PTM ↓
Cofactor	ID	4.148	1.788	1.636	0.725
	OOD	4.183	1.821	1.612	0.750
EC number	ID	2.843	1.542	0.683	0.618
	OOD	2.869	1.543	0.688	0.638
Gene Ontology (MF)	ID	2.834	1.648	0.761	0.425
	OOD	2.882	1.656	0.777	0.449
Catalytic activity	ID	3.621	1.592	0.936	1.094
	OOD	3.583	1.623	0.926	1.035
Sites (binding/catalytic)	ID	5.881	1.527	2.974	0.628
	OOD	5.964	1.536	3.038	0.633
Regions (motifs/domains)	ID	–	–	–	–
	OOD	–	–	–	–

Table 45: Validation loss decomposition for Stage-1 pretraining. OOD-family splits exhibit systematically higher loss for functionally demanding tasks, indicating reduced robustness under family shift.

Task	Split	MC Acc ↑	MC MRR ↑	MC Rec@5 ↑
Sites	ID	31.72	0.5235	84.76
	OOD	30.91	0.5092	83.48

Table 46: Multi-choice localization performance of Stage-1 models. Only the Sites task shows non-zero localization in the provided logs.

I.5 large scale fine-grained pretraining

The table 45 presents the loss performance for different dataset.

The table46 presents the localization multi-choice performance on the two residue-level tasks.

The table 47 presents the baseline retrieval performance on our PANNOTGROUND's val and test dataset. Here the metrics are not distinguished between id and ood dataset.

The table 48 present the re-rank test retrieval performance of the baseline model.

The table 49 present the different loss for the baseline model. We notice that the diffenrence between ID and OOD is not significant.

The table 50 presents the localization metrics for the baseline model.

J Fine-grained ablations

J.1 Setup

We evaluate fine-grained protein function annotation under a BLIP2-style protein–text architecture. Our experiments are designed to (i) assess the ben-

Table 47: Test retrieval performance of the Stage-1 pretrained model. Since ID and OOD splits share identical retrieval metrics, results are reported without split.

	Full-set				In-batch			
Task	P2T Acc	P2T R@20	T2P Acc	T2P R@20	P2T Acc	P2T R@20	T2P Acc	T2P R@20
Cofactor	0.27	2.54	0.23	1.62	6.09	52.03	4.73	47.21
EC	0.29	3.90	0.21	2.38	12.63	59.40	8.96	55.32
Catalytic activity	1.32	12.95	0.50	5.24	24.89	76.32	15.48	70.50
Gene ontology	3.17	16.81	1.00	7.58	28.91	68.92	19.36	64.43
Sites	0.07	0.42	0.01	0.19	1.97	31.65	1.74	31.61
Regions	0.37	2.09	0.12	0.91	5.73	38.60	3.65	37.28

Table 48: Re-ranked test retrieval performance of the Stage-1 pretrained model. Retrieval metrics are identical across ID and OOD splits and therefore reported jointly.

	Full-set				In-batch			
Task	P2T Acc	P2T R@20	T2P Acc	T2P R@20	P2T Acc	P2T R@20	T2P Acc	T2P R@20
Cofactor	0.23	3.16	0.08	1.46	7.30	58.18	4.14	48.75
EC	0.27	3.69	0.15	2.34	13.82	67.62	7.56	55.69
Catalytic activity	1.48	14.37	0.40	4.11	29.00	87.20	12.80	73.25
Gene ontology	3.12	17.22	0.51	4.70	30.94	77.16	12.66	59.01
Sites	0.10	0.50	0.01	0.23	2.17	31.73	1.80	31.69
Regions	0.45	2.38	0.09	0.78	6.90	43.78	3.80	38.75

efit of PANNOTGROUND, which provides multi-granular biochemical evidence and supports both ID and OOD-family evaluation, and (ii) quantify the effectiveness of our **fine-grained pretraining objectives**. We used A100 GPUs on computing clusters for the experiments. More details about the experiment are attached in Appendix H.

J.2 Pilot Study: Validating Fine-Grained Objectives

Prior to large-scale pre-training on our proprietary datasets, we conduct a pilot study on the **VenusX** benchmark (Tan et al., 2025) to validate the efficacy of our proposed objective functions. This preliminary investigation serves to isolate the performance gains attributable to the *Local Contrastive* (LC) and *Counterfactual* (CF) strategies in a controlled setting, ensuring the architectural choices are optimized for evidence-grounded reasoning before scaling.

Quantitative Analysis. As summarized in Table 51, the integration of both strategies in the **Full (LC+CF)** framework consistently yields superior performance across all benchmarks. In the cross-family (`Res_Act_X`) split, our full model achieves a **7.84%** improvement in `Loc_mc_Rec5` over the baseline, demonstrating robust generalization to novel protein families.

The synergy between LC and CF becomes particularly evident in the mixed-family splits. In the high-similarity `MP90` setting, the **Full** model outperforms the baseline by **9.09%** in `Loc_mc_Acc`. These results indicate that the **LC** objective enhances localized precision, while the **CF** objective effectively mitigates rote sequence memorization. These findings provide a rigorous empirical foundation for our subsequent large-scale pre-training phase.

Faithfulness The ablation results in Table 51 provide evidence for the faithfulness of our pretraining strategies. While local contrastive learning (LC) improves discriminability of annotated regions, its effect is inconsistent across splits, particularly under more challenging distribution shifts. In contrast, counterfactual masking (CF) yields consistent improvements, indicating that explicitly enforcing dependence on annotated regions is critical for reliable localization. The full strategy combining LC and CF achieves the strongest performance across all benchmarks, suggesting that faithful reasoning requires both recognizing relevant evidence and being causally dependent on it. Notably, the largest gains are observed on MF and MP splits, where global sequence correlations are weaker, further supporting the role of evidence grounding in robust generalization.

Table 49: Validation loss decomposition for the Stage-1 pretrained model. Each task is reported with two subrows corresponding to ID and OOD-family splits.

Task	Split	Val Loss	CF	LM	Local LM	Local PTM	PTC	PTM
Cofactor	ID	13.77	0.00	6.28	0.00	0.00	3.60	3.88
	OOD	13.79	0.00	6.26	0.00	0.00	3.67	3.86
EC	ID	14.20	0.00	7.02	0.00	0.00	3.35	3.83
	OOD	14.17	0.00	7.03	0.00	0.00	3.31	3.83
Catalytic activity	ID	12.85	0.00	6.47	0.00	0.00	2.79	3.59
	OOD	12.80	0.00	6.47	0.00	0.00	2.77	3.57
Gene ontology	ID	14.10	0.00	7.53	0.00	0.00	2.85	3.72
	OOD	14.07	0.00	7.53	0.00	0.00	2.83	3.71
Sites	ID	25.85	0.10	8.24	8.19	10.75	3.96	4.05
	OOD	25.83	0.10	8.24	8.18	10.72	3.95	4.06
Regions	ID	25.42	0.11	8.26	8.05	10.25	4.10	3.79
	OOD	25.40	0.11	8.27	8.05	10.22	4.09	3.79

Table 50: Localization multiple-choice performance on the validation set. This evaluation is only applicable to fine-grained localization tasks.

Task	MC Acc (%)	MC MRR	MC Rec@5 (%)
Sites	8.63	0.285	51.02
Regions	29.51	0.472	69.50

J.3 Large Scale Cross-Modal Pretraining

To extend fine-grained training to a large-scale setting, we apply PANNOTGROUND to cross-modal pretraining using a Q-Former–based Stage-1 framework that aligns protein representations with textual annotations. We adopt a curriculum that first emphasizes global protein–text pairs to establish coarse functional semantics, followed by fine-grained annotations to refine localized and attribute-level alignment. All annotation types are unified under a single pretraining objective, with coefficient values selected based on the ablation study in Table 43.

We first evaluate the effect of Stage-1 pretraining on fine-grained localization using the multi-choice localization (Loc_mc) task. As shown in Table 52, fine-grained pretraining leads to a substantial improvement in localization sensitivity compared to the baseline model. With eight candidate locations per query, the baseline performs close to random guessing, whereas the Stage-1 model achieves approximately 30% accuracy, indicating that cross-modal pretraining successfully induces non-trivial site-level grounding between textual descriptions and protein regions.

Beyond localization, Stage-1 pretraining consistently improves cross-modal retrieval performance.

	Strategy	**Acc ↑**	**Mrr ↑**	**Rec5 ↑**
X	BL	41.48	0.544	68.41
	BL + LC	42.16	0.546	65.00
	BL + CF	43.30	0.558	67.05
	BL+LC+CF	**44.09**	**0.588**	**76.25**
MF50	BL	46.74	0.620	80.98
	BL + LC	52.72	0.676	90.22
	BL + CF	52.17	0.658	85.87
	BL+LC+CF	**58.70**	**0.722**	**91.85**
MP50	BL	49.16	0.633	82.87
	BL + LC	46.91	0.629	84.27
	BL + CF	56.18	0.696	89.04
	BL+LC+CF	**59.83**	**0.723**	**91.29**
MP90	BL	47.89	0.622	81.26
	BL + LC	45.90	0.613	84.37
	BL + CF	55.88	0.676	83.48
	BL+LC+CF	**56.98**	**0.691**	**84.59**

Table 51: Ablation results of pre-training strategies across four VenusX_Res_Act benchmarks. We evaluate the impact of Local Contrastive (LC) and Counterfactual (CF) strategies compared to baseline (BL). The metrics (Acc, Mrr, and Rec5) are for localization multiple choice as defined in sec **??**

As shown in Table 53, in-batch Recall@20 exceeds 90% across most tasks and approaches saturation for enzyme-centric annotations such as EC numbers. While in-batch top-1 accuracy remains moderate, this behavior indicates that fine-grained pretraining effectively organizes protein–text representations into semantically coherent neighborhoods, even when precise ranking among highly similar candidates remains challenging.

Full-set retrieval remains challenging, with low absolute accuracy but non-trivial Recall@20, reflecting the scale and ambiguity of the candidate space. Notably, EC number prediction exhibits

Task	Split	Acc ↑	Mrr ↑	Rec5 ↑
FG Stage1	ID	31.72	0.5235	84.76
	OOD	30.91	0.5092	83.48
Baseline	mix	8.63	0.285	51.02

Table 52: Multi-choice localization performance of Stage-1 models compared with baseline models. Here we use sites tasks for the evaluation.

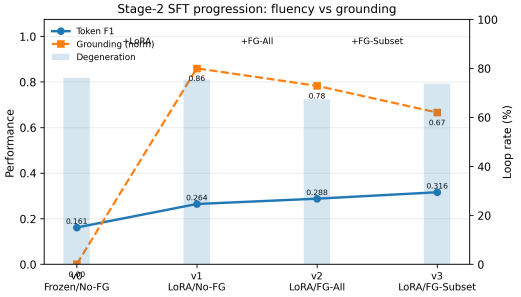


Figure 13: Decoder adaptation improves fluency, fine-grained pretraining improves robustness, but grounding remains the bottleneck.

the strongest performance, consistent with conserved catalytic motifs and standardized functional descriptions, whereas attributes such as cofactors show lower accuracy due to their heterogeneous and weakly localized nature. Together, these results indicate that Stage-1 cross-modal pretraining effectively structures the global embedding space but is insufficient on its own for precise decision-making or evidence-faithful prediction, motivating subsequent post-training stages.

J.4 Training with LLM Decoder

In Stage-2, we further fine-tune the pretrained model using task-specific instruction–response pairs derived from SwissProt-Evidence. The training aims to align the Q-former's soft embeddings with LLM decoder. It substantially improves fluency and task recognition across all annotation types: the model reliably adopts task-appropriate language (e.g., enzymatic descriptions for EC and catalytic tasks, biological process narratives for GO). Mechanistic evaluation reveals that the model predominantly produces abstract functional narratives.

We analyze four Stage-2 configurations that differ in decoder adaptation and Stage-1 initialization. Comparing v0 (frozen LLM) and v1 (LoRA-enabled) reveals that enabling decoder adaptation yields the largest performance gain, with a substantial increase in token-level F1 across tasks. This indicates that Stage-2 SFT primarily addresses a decoder alignment bottleneck, allowing the model to produce task-appropriate language and structured outputs. Transitioning from v1 to v2, where the model is initialized from a fine-grained Stage-1 checkpoint trained on all tasks, further improves performance consistency and reduces degeneration, suggesting that Stage-1 enhances the semantic quality of Q-Former representations. However, restricting Stage-1 pretraining to a subset of tasks (v3) introduces representation bias, leading to uneven task performance and reduced stability.

Notably, across all configurations, improvements are concentrated in linguistic fidelity rather than mechanistic grounding. While token-level F1 increases substantially from v0 to v3, grounding metrics remain comparatively low and largely flat, indicating that the model learns to produce fluent, task-aligned descriptions without reliably identifying the underlying biochemical evidence. This highlights a fundamental limitation of Stage-2 SFT: it aligns how the model expresses knowledge.

J.5 Additional Analysis of Stage-2 SFT Variants

To better understand the behavior of Stage-2 supervised fine-tuning (SFT), we compare four variants that differ in decoder adaptation and Stage-1 initialization. Specifically, **v0** (*Frozen/No-FG*) fine-tunes the Q-Former while keeping the LLM decoder frozen and does not use a fine-grained Stage-1 checkpoint; **v1** (*LoRA/No-FG*) additionally enables LoRA adaptation on the decoder; **v2** (*LoRA/FG-All*) initializes from a fine-grained Stage-1 checkpoint trained on all tasks; and **v3** (*LoRA/FG-Subset*) initializes from a Stage-1 checkpoint trained on only a subset of tasks.

Figure 14(a) shows that the largest improvement occurs from **v0** to **v1**, where the only major change is enabling LoRA on the LLM decoder. This suggests that, without decoder adaptation, Stage-2 SFT is primarily limited by a *decoder alignment bottleneck*: the Q-Former can be optimized, but a frozen decoder cannot fully translate its representations into task-appropriate outputs. Once LoRA is enabled, the model becomes substantially better at producing fluent, structured, and task-aware responses.

Comparing **v1** and **v2** isolates the effect of fine-grained Stage-1 pretraining under the same LoRA-based Stage-2 setup. Initializing from a

Task	Baseline Stage-1				Fine-grained Stage-1 (ID)				Fine-grained Stage-1 (OOD)			
	InAcc	InR20	FullAcc	FullR20	InAcc	InR20	FullAcc	FullR20	InAcc	InR20	FullAcc	FullR20
Cofactor	6.09	52.03	0.27	2.54	34.74	94.37	1.70	23.64	33.43	94.21	1.70	22.85
EC	12.63	59.40	0.29	3.90	69.70	99.57	3.30	45.57	68.36	99.57	3.27	45.01
GO	28.91	68.92	3.17	16.81	68.43	99.36	7.88	48.00	68.24	99.27	7.88	48.36
CA	24.89	76.32	1.32	12.95	58.75	98.89	4.32	47.82	58.12	98.35	4.42	47.45
Sites	1.97	31.65	0.07	0.42	12.32	56.87	0.85	6.41	12.73	56.12	0.85	6.41

(a) Protein → Text retrieval (p2t)

Task	Baseline Stage-1				Fine-grained Stage-1 (ID)				Fine-grained Stage-1 (OOD)			
	InAcc	InR20	FullAcc	FullR20	InAcc	InR20	FullAcc	FullR20	InAcc	InR20	FullAcc	FullR20
Cofactor	4.73	47.21	0.23	1.62	32.98	94.59	1.55	19.58	32.24	93.25	1.55	19.22
EC	8.96	55.32	0.21	2.38	66.78	99.47	2.24	37.59	66.40	99.47	2.46	37.23
GO	19.36	64.43	1.00	7.58	66.27	98.93	5.38	43.89	66.42	98.79	5.41	43.26
CA	15.48	70.50	0.50	5.24	57.97	98.74	3.90	45.90	57.13	98.74	3.85	45.90
Sites	1.74	31.61	0.01	0.19	12.70	54.68	0.47	4.93	11.47	54.02	0.47	4.93

(b) Text → Protein retrieval (t2p)

Table 53: Stage-1 retrieval performance comparing vanilla Q-Former baseline vs SwissProt-Evidence. Top: protein-to-text (p2t). Bottom: text-to-protein (t2p). In-batch metrics measure global semantic alignment; full-set metrics test large-pool retrieval.

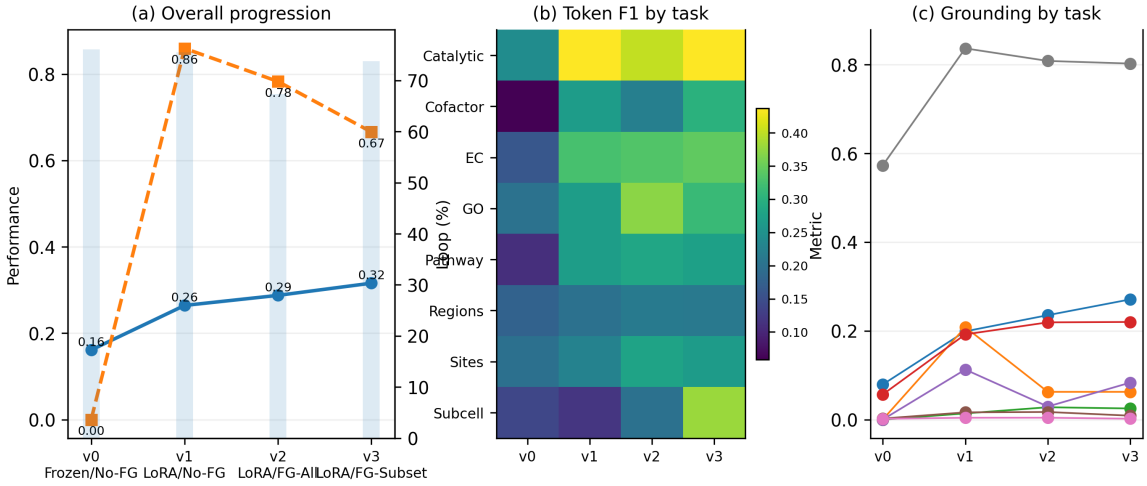


Figure 14: Additional analysis of Stage-2 SFT variants. **(a)** Overall progression across training configurations. Enabling decoder LoRA from v0 to v1 yields the largest gain in token-level performance, indicating that decoder adaptation is the primary bottleneck when the LLM is frozen. Moving from v1 to v2, i.e., introducing a fine-grained Stage-1 checkpoint trained on all tasks, further improves overall performance and reduces degeneration. Replacing this checkpoint with a subset-task Stage-1 initialization (v3) leads to slightly higher fluency but weaker robustness. **(b)** Token-level F1 by task. Improvements are broad but uneven, with stronger gains on semantic tasks such as catalytic activity, EC description, and gene ontology, while structurally grounded tasks remain comparatively difficult. **(c)** Task-specific grounding metrics. Despite clear improvements in fluency, grounding remains weak and highly task-dependent across all variants, revealing a persistent gap between task-conditioned language generation and mechanistic evidence localization.

fine-grained Stage-1 checkpoint trained on all tasks improves performance consistency and lowers degeneration, indicating that Stage-1 contributes a stronger and more stable representation prior for the Q-Former. In contrast, **v3**, which uses a Stage-1 checkpoint trained on only a subset of tasks, shows a less favorable trade-off: it retains strong fluency but exhibits weaker robustness, consistent with a more biased or less transferable intermediate representation.

The task-level heatmap in Figure 14(b) further shows that these gains are not uniform. Semantic tasks, such as catalytic activity and gene ontology, benefit most from decoder adaptation and improved initialization, whereas the structurally grounded tasks remain challenging. This observation is reinforced by Figure 14(c), where task-specific grounding metrics remain relatively low

and unstable across variants. In particular, improvements in token-level F1 do not translate proportionally into gains in EC tuple recovery or residue-level localization.

Overall, these results suggest a three-stage interpretation of the Stage-2 variants: **v0**→**v1** mainly reflects the benefit of decoder adaptation; **v1**→**v2** reflects the value of a stronger fine-grained Stage-1 prior; and **v2**→**v3** highlights the cost of reducing Stage-1 task coverage. Across all configurations, however, the dominant remaining limitation is unchanged: Stage-2 SFT improves *how the model expresses function* more than *how the model grounds function in biochemical evidence*.