COMPOSING PROTEIN EVIDENCE: A BENCHMARK FOR FINE-GRAINED PROTEIN-TEXT UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein-text understanding requires more than assigning functional labels: a biological claim is most informative when it is associated with various levels of evidence. Current protein understanding evaluations often fragment this relation, measuring global labels, local features, or protein–text matching individually without testing whether these elements remain bound to the same protein. We introduce PannotGround, a benchmark that turns curated protein annotations into a diagnostic test of evidence composition. PannotGround projects each protein record into synchronized levels and composes them into protein-to-text and text-to-protein tasks with biologically structured hard negatives. 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, yet remain sensitive to local-evidence contrasts, and many failures are biologically plausible near-misses. PannotGround establishes evidence composition as a key diagnostic axis for fine-grained protein-text models. The dataset and code are available on 🤗 and .

1 INTRODUCTION

Understanding proteins, the workhorse molecule of life, takes more than predicting their functional labels from sequences or structures. A model that predicts an enzyme class or binding function has answered only part of the biological question. Rather, informative functional annotations tie protein-level claims to sequence or structure evidence: domains, motifs, catalytic residues, and cofactor-binding patterns. Fine-grained protein-text understanding demands what local features accompany the annotation as potential evidence and where those features are in the protein.

Current evaluations of protein understanding stress pieces of this problem, but rarely the full evidence binding problem. Specifically, 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., 2025; 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 one example. The source protein Q08209, the human calcineurin catalytic subunit, has a global annotation of phosphatase, 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 across levels whether these attributes remain bound to one protein. Their hard negatives are biologically close: some share global function, some share local

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

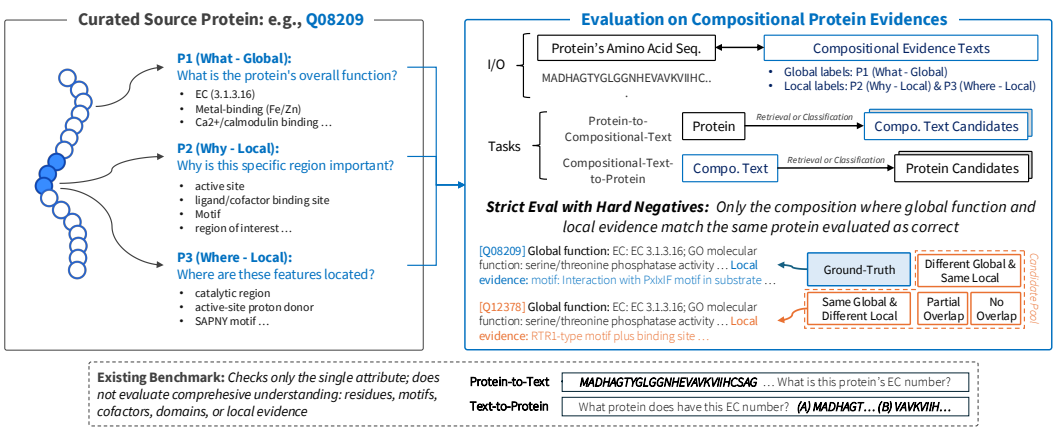


Figure 1: Overview of PannotGround using compositional protein evidence. Source proteins are curated by capturing both their global functions and local features (left). Based on this local-global compositional binding, PannotGround introduces a strict evaluation with hard negatives, such as candidates 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).

evidence, and some share partial overlap. This design exposes global-local shortcuts that used to be hidden in aggregate accuracy.

In this paper, our benchmark PannotGround makes three contributions:

- We define 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 (Friedberg et al., 2025; 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., 2021; 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 diagnosed for general protein understanding.

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., 2025). 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 & Zou, 2025; Liu et al., 2026; 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 & 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; Luo et al., 2026; Shen et al., 2024; Abdine et al., 2024; Liu et al., 2024; Xu et al., 2023; Su et al., 2024; Queen et al., 2025; Lv et al., 2025; Fallahpour et al., 2026).

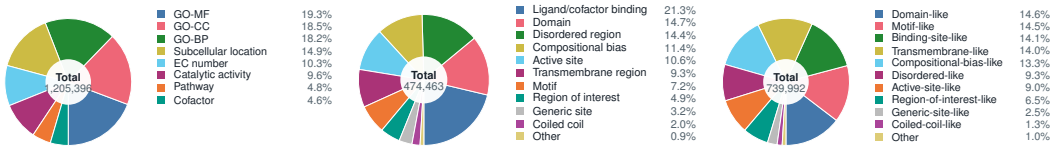
Table 1: Compact comparison with representative protein benchmarks. Coverage summarizes the biological task scope; composition indicates whether the benchmark diagnoses binding between global function, local evidence, and protein-text matching.

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., 2025)	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., 2025)	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; evaluates diverse model types robustly.**

3 BENCHMARK: PANNOTGROUND

Protein annotations are inherently multi-scale: a sequence can support global functions such as enzymatic activity, while the evidence for those functions often appears as local sites, domains, and motifs. PannotGround turns annotations into a comprehensive 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. A global entry is (n, y): namespace n (e.g., EC number, GO molecular function, cofactor) and normalized label y (e.g., "EC 3.1.3.16"). A local entry is (t, d, S): canonical feature type t (e.g., active site), optional description d (e.g., "metal-binding residue"), and residue sites or spans S (e.g., positions 199–201). Figure 1 illustrates a complete record for human calcineurin (Q08209).

Task levels. We create projections of the atomic protein record from UniProt (Consortium, 2025). For diagnostic purposes, we set three types of single-level tasks and the compositional cross-level task. The single-level results reveal whether the failure stems from an inability to recognize individual attributes. Its atomic record details are presented in Figure 2. *P1 (global function)* targets protein-



(a) P1 function subtasks. (b) P2 feature-type composition. (c) P3 feature-subtype localization tasks.

Figure 2: Subtask composition of PannotGround atomic record. P1 is protein-level semantic function annotation. P2 is local feature type. P3 is feature subtype localization task.

level semantic annotation: the model identifies the set of global labels within a given namespace (e.g., EC numbers or GO molecular function terms) given protein input. *P2 (local feature type)* targets feature-type recognition: the model recovers the set of canonical localized feature types present anywhere in the protein (e.g., active site, transmembrane region, domain). Local feature type identification is formulated as a multi-label prediction because multiple feature types typically co-occur. *P3 (residue-level evidence-subtype)* targets evidence localization: given a structured query specifying the feature family, canonical type, and optional description, the model identifies the corresponding residue span(s). This decomposition separates three capabilities that single-label benchmarks often collapse: semantic function recognition, local feature detection, and evidence localization.

Single-level MCQ. We constructed multi-choice single-level evaluation tasks from the task-view record to compare across different models (PLM, Protein-LLM, and Text LLM). We select an atomic annotation in the task-view record, set the ground truth answer as the gold answer, and sample negative labels in the same subtask space. After having the positive and negative choices for a question, we remove the false negative samples and shuffle the options. The number of multiple-choice options is set to 4, 8, and 5 for P1/2/3, and an example is shown in Appendix B.5.

Compositional MCQ. We construct each compositional instance from the annotations associated with a single protein, combining protein-level functional attributes, local feature types, and local feature descriptors. Localization is currently not included in the compositional tasks.

In the full protein-to-text (P2T) task, the input is a protein sequence and the candidates are eight compositional captions. The ground-truth caption is generated from all retained attributes of that protein from the task-view record, whereas the seven negative captions are drawn from biologically non-equivalent proteins in the same split. Each protein has a global label signature (G) and local label signature (L). The same-local/different-global negatives are selected from proteins indexed by the same (L) but a different (G); same-global/different-local negatives are vice-versa; both-different negatives are selected from the remaining proteins for which both signatures differ.

The text-to-protein (T2P) task reverses the matching direction but increases the retrieval space: a compositional caption is used as the query, and the model ranks 64 candidate protein sequences, only one of which is the protein from which the complete query composition was derived. The remaining candidates are sampled from the same split and include up to 15 proteins with the same (L) but different (G), up to 15 with the same (G) but different (L), 16 candidates sharing only a subset of the query attributes, and 16 candidates differing in both components, with non-equivalent random candidates used to fill any unoccupied positions.

Variant sets. We further construct controlled variants by projecting the full composition before candidate sampling. *Attr2* and *Attr4* retain exactly two and four annotation fields, respectively, and include at least one global attribute and one local attribute whenever both are available. *Global* retains global attributes and *local* retains local attributes. For these projected variants, negatives are ranked by similarity to the projected ground truth, prioritizing candidates that preserve one component or share the largest number of selected attributes. [1] The variants retain the original task cardinality of 8 P2T options and 64 T2P candidates. Together, these designs distinguish isolated attribute recognition from the more demanding requirement of compossed prediction.

[1] Candidates that are equivalent to the ground truth, or that duplicate an existing caption, composition, or biological signature, are excluded.

Table 2: PannotGround dataset statistics by split. Proteins count unique source proteins.

		Task-view records			Single-level tasks			Compositional tasks	
Split	Proteins	P1	P2	P3	P1	P2	P3	P2T	T2P
Train	127,619	119,614	100,015	350,177	—	—	—	—	—
Val-ID	22,587	21,834	16,951	56,439	50	50	49	1,000	1,000
Test-ID	22,014	21,255	16,940	57,780	100	100	98	2,000	2,000
Val-OOD	55,414	53,540	42,005	138,610	50	50	50	1,000	1,000
Test-OOD	54,387	52,346	42,523	136,986	100	100	98	2,000	2,000
Total	282,021	268,589	218,434	739,992	300	300	295	6,000	6,000

Data Split. To facilitate the record organization, we separate the atomic record into task-view records and split the protein into 5 splits. The proteins in SwissProt are clustered at 30% identity (C30) and we use C30 clusters to divide out-of-distribution (OOD) protein families. Within the remaining families, we use 90% identity to separate in-distribution (ID) splits (C90). The details of splits are in the appendix A.2.

The task-view stats in Table 2 highlight a large-scale resource for the field: 282,021 proteins across controlled ID and OOD splits, with >1.2M task-view records and 12K cross-level tasks. The demos in Appendix B.5 present the content in each task-view record. The canonical task-view records in P1/2 have multiple atomic annotations. A subset of 46,837 proteins are further identified with structures available from AlphaFold DB, with exact full-length sequence-structure mappings and associated confidence metadata. This structure-available subset preserves the original assignments of splits.

4 EVALUATION

PannotGround addresses two questions that existing benchmarks leave unresolved. First, can a model recognize individual biological attribute from a protein sequence? Second, when these evidence types are composed into a single protein-specific description, can the model bind them to the correct protein? These two questions motivate two evaluation stages: single-level tasks that isolate each evidence type independently, and compositional tasks that test whether isolated capabilities transfer to multi-attribute binding.

Uniformed Evaluation Task. Answering these two questions across diverse model families introduces a design challenge. Protein language models (PLMs) encode sequences into embeddings and score candidates through learned heads. Protein-conditioned LLMs generate or rank text given a sequence input. Text-only LLMs reason over text sequence input. These architectures expose fundamentally different scoring interfaces, so no single evaluation format natively fits all of them. Our solution is to fix the answer space rather than the scoring mechanism. For single-level tasks, we map all biological targets to fixed multiple-choice candidate sets (MCQA): every model — regardless of whether it scores candidates by embedding similarity, conditional likelihood, or prompted generation — selects from the same options and is measured by the same metric. For compositional tasks, we construct protein-to-text (P2T) and text-to-protein (T2P) retrieval with biologically structured hard negatives. Therefore, errors are not only counted but diagnosed. As a complementary baseline, PLM classification with task-specific heads[2] measures what the encoder can recover in its native label space, independent of the shared MCQA interface.

Model Families. We evaluate 16 models across four model families that represent distinct approaches to protein understanding. (1) Protein-text *alignment* models (ProTrek-35M, ProTrek-650M, ProtST variants, and Pannot-FG[3]) jointly train protein and text encoders to produce compatible embeddings. The model scores candidates by direct protein-text cosine similarity and participates in all evaluation tracks, including T2P retrieval. (2) Protein language models (PLMs) (Ankh, ESM-2, ProGen, ProstT5, and ProtBERT) are unsupervised sequence encoders that do not natively produce text-compatible scores. On the structure-available subset, we additionally evaluate ESM-IF structure representations,

[2]Trained on their own subtask label space.

[3]A fine-grained trained model with details in appendix F

residue-wise concatenation of ESM-2 and ESM-IF embeddings, and SaProt representations from both amino-acid and Foldseek 3Di tokens. We define *PLM-TextAlign* as projecting their embeddings into the text space via Ridge regression on PubMedBERT caption embeddings and score candidates by cosine similarity in the projected space. (3) *Protein LLMs* (BioReason-Pro-SFT and ProLLaMA) accept a protein sequence as input and use their LLM decoder to rank candidates by normalized conditional log-likelihood of each candidate text. (4) *Text-only LLMs* (Llama-3.1-8B, Qwen2.5-7B, and Qwen3-4B) receive textual candidate descriptions with the protein sequence. They select answers via prompted generation or parsed responses, serving as controls that measure inference from textual semantics and parametric knowledge. All model families score over the same fixed candidate sets defined in Sec. 3, ensuring that performance differences reflect model capability rather than answer-space confounds.

Evaluation Metrics Single-level tasks are scored as follows. P1 and P3 each present a single correct answer among fixed candidates. Their performance is measured by accuracy (proportion of instances where the model's top-ranked candidate matches the gold answer). P2 requires recovering the complete set of feature types present in a protein. The ground-truth is a multi-label set and performance is measured by mean set F1. For each instance, the F1 score is computed between the predicted feature-type set and the gold feature-type set, and the metric is averaged across instances. Compositional tasks are scored analogously to their candidate-pool size. P2T accuracy measures the proportion of instances where the correct caption is ranked first among eight candidates (random level: 12.5%). T2P uses Recall@1 (proportion of instances where the correct protein is ranked first among 64 candidates; random level: 1.56%) and Mean Reciprocal Rank (MRR) as primary metrics. We report single-level results separately for P1, P2, and P3 rather than as an aggregate, because each corresponds to a distinct biological capability. Compositional results are further accompanied by relation-aware error diagnostics: when a model selects an incorrect candidate, we classify the error by the known relation between that candidate and the gold (same-global/wrong-local, same-local/wrong-global, partial overlap, or unrelated).

All hyperparameters, thresholds, and decoding configurations are selected on validation splits. Final results are reported on held-out ID and family-level OOD test splits.

5 RESULTS

We structure our evaluation to address three questions: (1) model performance on single-level tasks, (2) the transferability of single-level competence to compositional tasks requiring the integration of multiple evidences, 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 E respectively.

5.1 SINGLE-LEVEL TASK PERFORMANCE

To evaluate model performance on single-level tasks, we used curated multi-choice single-level task. Table 3 presents results under a shared multi-choice setup. Here we define PLM's *Proto* readout set up: PLM rank choices by computing similarity scores between frozen protein embeddings and reference prototypes from labeled samples. And *FT* readout means PLM use sub-task specifically finetuned MLP head to predict logit across the choices.

The results indicate that performance across the three biological levels is highly variable. PLMs with *Proto* demonstrate strong capabilities in global function prediction and localization. Conversely, PLMs with task-specific heads achieve the highest accuracy on local feature identification. On the other hand, both Text LLMs and Protein-LLMs generally struggle under this rigid evaluation format. These discrepancies highlight that global function prediction, local feature identification, and localization require distinct modeling capabilities. Moreover, isolated PLM classification results within each subtask's native label space are detailed in Appendix (Tables 16 and 17).

5.2 COMPOSITIONAL MATCHING REMAINS CHALLENGING

To comprehensively understand protein, models need ability to predict multiple biological attributes. We analyze the challenge for compositional protein attribute understanding. In our evaluation, a

Table 3: Single-level multi-choice results on test set. We present accuracy for P1&P3 and set F1 for P2. Results are averaged among ID and OOD sets. Separate ID&OOD results are reported in Appendix (Table 18).

(a) PLM candidate scorers

Model	Readout	P1	P2	P3
Ankh	Proto	**.775**	.583	**.720**
	FT	.740	.503	–
ESM-2	Proto	.760	.590	.460
	FT	.730	**.670**	–
ProGen	Proto	.675	.494	.265
	FT	.725	.592	–
ProstT5	Proto	.765	.559	.674
	FT	.700	.579	–
ProtBERT	Proto	.710	.484	.455
	FT	.715	.623	–

(b) Protein-LLM

Model	Readout	P1	P2	P3
BioReason-Pro-SFT	Prompt	.170	.013	.225
ProLLaMA	Likeli.	.230	.402	.210

(c) Text LLM

Model	Readout	P1	P2	P3
Llama-3.1-8B	Prompt	.370	.308	.189
Qwen2.5-7B	Prompt	.385	.281	.159
Qwen3-4B	Prompt	.345	.358	.169

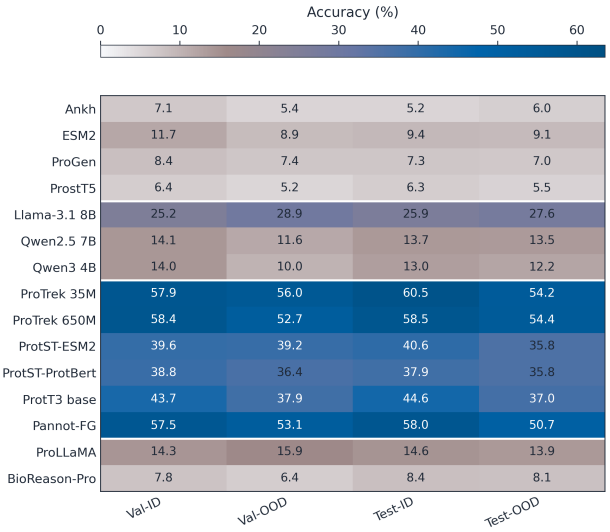


Figure 3: Compositional P2T accuracy across splits.

Table 4: Controlled P2T accuracy on variant compositional sets.

Model	Attr1	Global	Local	Full
Random	–	12.50	12.50	12.50
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

compositional text comprises attributes across previously defined tasks. The experiments are built upon the compositional retrieval task of PannotGround in Sec. 3. We found that single-level prediction proficiency is hard to transfer to compositional caption selection in different levels.

Figure 3 presents the compositional performance of models across split sets. We find *Alignment* models are currently the most effective at selecting the full compositional caption. The strongest Full results come from ProTrek and Pannot-FG, all above 54%. *PLM-TextAlign* mapping from frozen protein embeddings is weaker, reaching 36.40% at best. Both text-only and protein LLMs remain lower still. Thus, the main-table comparison shows that solving the full caption task requires more than isolated evidence recognition.

Table 4 reviews models' behavior on different set of compositional prediction. *Alignment* models are strongest when the decision can be made mostly from global function: the best such model reaches 73.15% on the global task, but drops to 43.23% on the local task. The *PLM-TextAlign* shows a different pattern: for every completed encoder, local task is easier than

Table 5: Controlled P2T accuracy for attribute-count composition by model family. Attr1 is the same reference as in Table 4. 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

global, suggesting that local evidence is easier to align. Even when a model can annotate global labels or local features in isolation, predicting both to the same protein-specific caption is still a hard problem.

Table 5 investigates the performance degradation resulted from increased attribute number. *Attr2* and *Attr4* are 8-way P2T variant dataset whose candidate captions display exactly two or four evidence fields respectively. Values in the table are family means over available model rows and the two held-out splits. 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 surpassing the random baseline. *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. The whole results are reported in Appendix Table 29. The results demonstrate that the challenge of compositional understanding stems from the fundamental requirement to accurately predict corresponding global and local evidence to a specific protein.

Table 6: P2T accuracy (%) on matched structure-available questions in the test set using the PLM-TextAlign setup.

Model	ID	OOD
Sequence-only baselines		
Ankh	16.01	14.77
ProstT5	22.83	22.78
ESM-2	26.28	27.43
Structure-aware models		
ESM-IF	24.01	26.58
ESM-2+ESM-IF	**28.39**	**29.96**
SaProt	21.99	21.52

We further evaluate the structural impact on the structure-available subset, retaining the original ID/OOD assignments. Table 6 compares ESM-2, ESM-IF, their concatenated representations, and SaProt on identical questions using the PLM-TextAlign setup. ESM-2+ESM-IF reaches 29.17% mean accuracy, compared with 26.86% for matched sequence-only ESM-2, whereas ESM-IF and SaProt reach 25.30% and 21.75%. Thus, the fusion of protein sequence and structure improves caption selection in this setting, but the performance drop compared to single-level evaluations remains significant.

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. We present the complementary Text-to-Protein (T2P) retrieval task, where a compositional text query needs to identify the matching protein. Table 7 summarizes the 64-way full-compositional retrieval performance for alignment models on `test_id`. Detailed split-specific results are in Appendix Tables 24, 25, and 26.

Table 7: Full-compositional T2P 64-way retrieval on `test_id`. Random R@1 is 1.56%. R@1 and R@5 are percentages. MR is the mean rank, where lower is better. SG, SL, Part., and Other report the top-ranked wrong protein when R@1 fails, normalized over test-set 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	9.40	39.05	.2471	10.15	36.4	27.3	31.5	4.7
ProTrek-650M	**11.05**	**40.55**	**.2655**	10.64	35.0	28.9	30.8	5.3
ProtST-ProtBERT	3.30	17.85	.1283	22.10	20.6	15.0	37.6	26.7
ProtST-ESM2	.55	4.25	.0503	38.45	8.2	6.4	13.1	72.3
ProtT3	3.55	22.10	.1427	19.64	20.6	12.5	31.5	35.4
Pannot-FG	8.10	39.75	.2406	**9.72**	33.8	24.2	35.5	6.4

The T2P results demonstrate that models are able to retrieve biologically related neighbors of the matching protein. The R@5 for T2P is as high as 40% for several models. 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. Consequently, the T2P failure analysis corroborates the P2T findings: current models capture coarse protein-text compatibility but struggle to reliably compose the complete set of multi-scale evidence.

The P2T outcome breakdown provides a complementary view. Figure 4 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. 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.

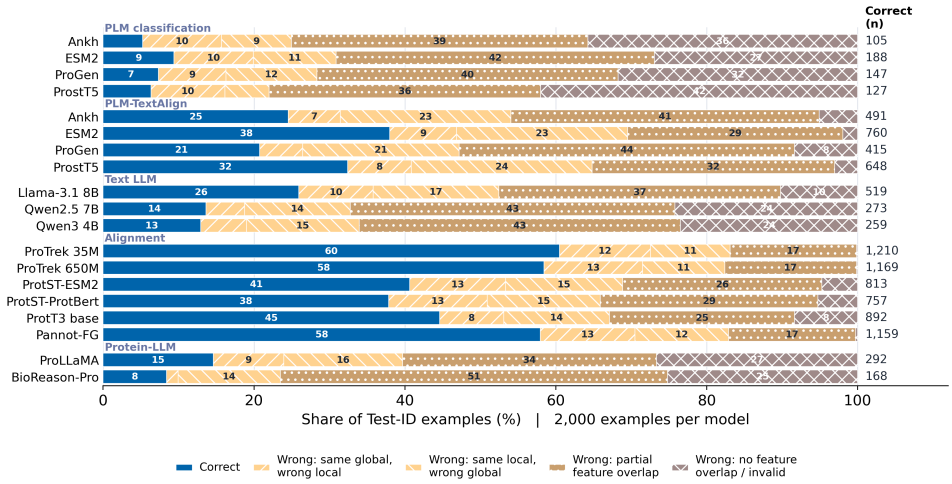


Figure 4: Structured P2T outcomes on `test_id`. The left rail groups rows by model family. Each stacked bar partitions all predictions into correct answers, relation-preserving wrong answers, partial-overlap wrong answers, and no-feature-overlap or invalid answers. The right column reports correct counts out of 2,000 examples.

Local-evidence discrimination varies across biological categories (Figure 5). Transmembrane, active-site, and domain/motif/region examples score above their corresponding model/split baselines. In contrast, low-complexity, disorder, coiled coil, and combined active-site/ligand-binding examples score below these baselines. These descriptive differences identify categories in which local evidence is harder to distinguish when global annotations are shared.

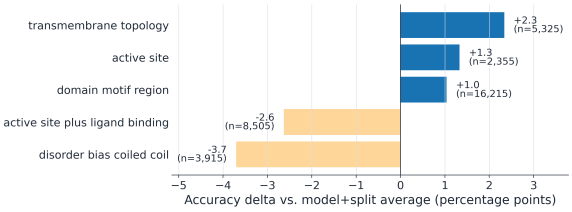


Figure 5: Local-evidence discrimination. Bars show P2T accuracy deviations from the corresponding model/split means on the same-global/different-local subset in percentage.

Taken together, these diagnostics distinguish exact matching, retained annotation overlap, and evidence-specific difficulty. They show where aggregate P2T/T2P scores conceal differences in model behavior and provide concrete targets for improving local-evidence discrimination.

6 Conclusion

The large-scale and systematic PannotGround frames three aspects that are often conflated in the evaluation of protein natural language understanding. It creates compositional tasks using structured hard negatives to evaluate whether global and local annotations are matched to the same protein. Across the evaluated model families, strong single-level performance coexists with substantial performance drop in compositional matching. *Alignment* models lead the P2T evaluation, while error analyses reveal incomplete annotation matches and differences across local-evidence categories. These findings motivate evaluating annotation recognition, compositional matching, and error structure separately. PannotGround provides a common experiment playground for developing protein natural language understanding models that reliably associate functional descriptions with specific local evidence.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data set, implementing methods, assisting with translation, cleaning and reformating dataset, supporting qualitative and thematic data analysis, interpreting results. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments. Additionally, we used generative AI tools for sourcing/searching for information and creation of artifacts. We have reviewed AI-assisted work. The LLM-generated code snippets and technical artifacts were independently verified, executed, and unit-tested. The synthetic dataset are checked with ground truth from Uniprot. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of machine learning — 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 benchmark study.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide an anonymized supplementary artifact containing the benchmark data, evaluation code, split metadata, and paper-facing results. The accompanying appendix and supplementary materials describe the dataset construction, task definitions, evaluation protocols, and implementation details. All reported experiments use fixed data splits and deterministic evaluation procedures, enabling independent verification of the presented results.

REFERENCES

Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. Prot2Text: Multimodal protein's function generation with GNNs and transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 10757–10765, 2024. doi: 10.1609/aaai.v38i10.28948.

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, Aron Marchler-Bauer, Laetitia Meng-Papaxanthos, Huaiyu Mi, Darren A. Natale, Christine A. Orengo, Arun P. Pandurangan, Damiano Piovesan, Catherine Rivoire, Christian J. A. Sigrist, Narmada Thanki, Françoise Thibaud-Nissen, Paul D. Thomas, Silvio C. E. Tosatto, Cathy H. Wu, and Alex Bateman. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research*, 53(D1):D444–D456, jan 2025. doi: 10.1093/nar/gkae1082. URL https://doi.org/10.1093/nar/gkae1082.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022. doi: 10.1093/bioinformatics/btac020.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic acids research*, 53(D1):D609–D617, 2025. URL https://doi.org/10.1093/nar/gkae1010.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing, 2021. URL https://arxiv.org/abs/2007.06225.

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

Adibvafa Fallahpour, Arman Seyed-Ahmadi, Parsa Idehpour, Omar Ibrahim, Benedict M.H. Choi, Purav Gupta, Jack Naimer, Kevin Zhu, Abhinav Adduri, Arnav Shah, Shihao Ma, Talu Güloglu, Nuo Liu, Haotian Cui, Arihant Jain, Max de Castro, Amirfaham Fallahpour, Antonio Cembellin-Prieto, John S. Stiles, Filip Nemčko, Alexander A. Nevue, Hyungseok C. Moon, Lucas Sosnick, Olivia Markham, Haonan Duan, Michelle Y. Y. Lee, Andrea F. M. Salvador, Chris J. Maddison, Christoph A. Thaiss, Chiara Ricci-Tam, Brian S. Plosky, Dave P. Burke, Patrick D. Hsu, Hani Goodarzi, and Bo Wang. BioReason-Pro: Advancing protein function prediction with multimodal biological reasoning. *bioRxiv*, 2026. doi: 10.64898/2026.03.19.712954. URL https://www.biorxiv.org/content/early/2026/03/20/2026.03.19.712954.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-Instructions: A large-scale biomolecular instruction dataset for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Tlsdsb6l9n.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348, 2022. doi: 10.1038/s41467-022-32007-7.

Iddo Friedberg, Predrag Radivojac, Paul D Thomas, An Phan, M. Clara De Paolis Kaluza, Damiano Piovesan, Parnal Joshi, Chris Mungall, Martyna Plomecka, Walter Reade, and María Cruz. CAFA 6 protein function prediction. https://kaggle.com/competitions/cafa-6-protein-function-prediction, 2025. Kaggle.

Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 2024. doi: 10.1093/nargab/lqae150.

Chao Hou, Di Liu, and Yufeng Shen. MotifAE reveals functional sequence patterns from protein language model: Unsupervised discovery and interpretability analysis. *bioRxiv*, 2025. doi: 10.1101/2025.11.04.686576. URL https://www.biorxiv.org/content/early/2025/12/04/2025.11.04.686576.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=Jsc7WSCZd4.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702, 2019. doi: 10.1109/CVPR.2019.00686.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–1997, 2017. doi: 10.1109/CVPR.2017.215.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308. IEEE, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://doi.org/10.1126/science.ade2574.

Xiangyu Liu, Haodi Lei, Yi Liu, Yang Liu, and Wei Hu. Protsae: Disentangling and interpreting protein language models via semantically-guided sparse autoencoders, 2026. URL https://arxiv.org/abs/2509.05309.

Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. ProtT3: Protein-to-text generation for text-based protein understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 5949–5966, 2024. doi: 10.18653/v1/2024.acl-long.324. URL https://aclanthology.org/2024.acl-long.324/.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521. Curran Associates, Inc., 2022. doi: 10.52202/068431-0182. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/11332b6b6cf4485b84afadb1352d3a9a-Paper-Conference.pdf.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Massimo Hong, Yushuai Wu, Mu Qiao, and Zaiqing Nie. BioMedGPT: An open multimodal large language model for biomedicine. *IEEE Journal of Biomedical and Health Informatics*, 30(2):981–992, 2026. doi: 10.1109/JBHI.2024.3505955.

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. ProLLaMA: A protein large language model for multi-task protein language processing, 2025. URL https://arxiv.org/abs/2402.16445.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally?, 2023. URL https://arxiv.org/abs/2212.07796.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023. doi: 10.1038/s41587-022-01618-2. URL https://doi.org/10.1038/s41587-022-01618-2.

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. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379. Curran Associates, Inc., 2023. doi: 10.52202/075280-2810. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/cac723e5ff29f65e3fcbb0739ae91bee-Paper-Datasets_and_Benchmarks.pdf.

Owen Queen, Yepeng Huang, Robert Calef, Valentina Giunchiglia, Tianlong Chen, George Dasoulas, LeAnn Tai, Gianmarco Abbadessa, Owain Howell, Michelle M. Li, Yasha Ektefaie, Ayush Noori, Ildiko Farkas, Joseph Brown, Tom Cobley, Karin Hrovatin, Tom Hartvigsen, Fabian J. Theis, Bradley L. Pentelute, James Zou, Vikram Khurana, David Owen, Richard Nicholas, Manolis Kellis, and Marinka Zitnik. ProCyon: A multimodal foundation model for protein phenotypes. *bioRxiv*, 2025. doi: 10.1101/2024.12.10.627665. URL https://www.biorxiv.org/content/early/2025/11/03/2024.12.10.627665.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with TAPE. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/37f65c068b7723cd7809ee2d31d7861c-Paper.pdf.

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan Plummer, Ranjay Krishna, and Kate Saenko. COLA: A benchmark for compositional text-to-image retrieval. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46433–46445. Curran Associates, Inc., 2023. doi: 10.52202/075280-2014. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/917cd410aa55b61594fa2a6f6e5a9e94-Paper-Datasets_and_Benchmarks.pdf.

Dingyi Rong, Zijian Chen, Qi Jia, Kaiwei Zhang, Haotian Lu, Guangtao Zhai, and Ning Liu. LiveProteinBench: A contamination-free benchmark for assessing models' specialized capabilities in protein science, 2025. URL https://arxiv.org/abs/2512.22257.

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

Christian J A Sigrist, Béatrice A Cuche, Edouard de Castro, Elisabeth Coudert, Nicole Redaschi, and Alan Bridge. The PROSITE database for protein families, domains, and sites. *Nucleic Acids Research*, 54(D1):D451–D458, 01 2026. ISSN 1362-4962. doi: 10.1093/nar/gkaf1188. URL https://doi.org/10.1093/nar/gkaf1188.

Elana Simon and James Zou. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nature methods*, 22(10):2107–2117, 2025. doi: 10.1038/s41592-025-02836-7.

Jin Su, Xibin Zhou, Xuting Zhang, and Fajie Yuan. ProTrek: Navigating the protein universe through tri-modal contrastive learning. *bioRxiv*, 2024. doi: 10.1101/2024.05.30.596740. URL https://www.biorxiv.org/content/early/2024/09/11/2024.05.30.596740.

Yang Tan, Wenrui Gou, Bozitao Zhong, Liang Hong, Huiqun Yu, and Bingxin Zhou. VenusX: Unlocking fine-grained functional understanding of proteins, 2025. URL https://arxiv.org/abs/2505.11812.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238, 2022. doi: 10.1109/CVPR52688.2022.00517.

Juntong Wu, Zijing Liu, He Cao, Hao Li, Bin Feng, Zishan Shu, Ke Yu, Li Yuan, and Yu Li. Rethinking text-based protein understanding: Retrieval or LLM?, 2025. URL https://arxiv.org/abs/2505.20354.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. PEER: A comprehensive and multi-task benchmark for protein sequence understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 35156–35173. Curran Associates, Inc., 2022. doi: 10.52202/068431-2548. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/e467582d42d9c13fa9603df16f31de6d-Paper-Datasets_and_Benchmarks.pdf.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. ProtST: Multi-modality learning of protein sequences and biomedical texts. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38749–38767. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/xu23t.html.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024. doi: 10.1109/CVPR52733.2024.00913.

A Data Split

A.1 Data Source

PannotGround is constructed from reviewed UniProtKB/Swiss-Prot entries in release `2025_03`. Each source record provides a protein sequence and its retained functional, site, and region annotations. Splits are assigned at the protein-accession level before task construction. All P1/P2/P3 records and compositional examples derived from a source protein inherit its split.

A.2 Homology-Based Splitting

Let $C_{30}(p)$ and $C_{90}(p)$ denote the MMseqs2 cluster assignments of protein p at nominal sequence-identity levels of 30% and 90% respectively. [4] C30 clusters define the benchmark's operational family groups; C90 clusters control near-duplicate sharing among training and ID splits.

Split construction

1. Randomly partition C30 clusters into a retained pool and two held-out cluster sets for OOD validation and OOD test.
2. Assign each protein in a held-out C30 cluster to that cluster's designated OOD split. Leave proteins in retained C30 clusters unassigned.
3. Within the retained pool, partition C90 clusters into Train, $\mathrm{Val}_{\mathrm{ID}}$, and $\mathrm{Test}_{\mathrm{ID}}$, keeping each C90 group intact.

Using random seed 13, we partitioned the 128,569 C30 groups represented in the benchmark universe into an 80% retained pool and two 10% held-out pools: 102,855 retained C30 groups and 12,857 groups each for OOD validation and OOD test. All proteins belonging to a held-out C30 group were assigned to the corresponding OOD split. Within the retained C30 pool, the globally computed C90 assignments were restricted to retained proteins and partitioned in a 90/5/5 ratio into Train, Val-ID, and Test-ID. This assigned 242,981 C90 groups to Train and 13,499 groups each to Val-ID and Test-ID.

The resulting accession sets are disjoint. No C30 identifier is shared between the retained pool and either OOD split. No C90 identifier is shared among Train, Val-ID, and Test-ID. After training-set subsampling, 77.7% of Val-ID proteins and 77.3% of Test-ID proteins have a C30 identifier represented in $\mathrm{Train_condensed}$.

These guarantees concern assigned cluster identifiers. Because the source snapshot was not explicitly deduplicated by normalized amino-acid sequence, the current release audit also identifies a small number of cross-split exact sequence collisions; we report and release these records separately rather than treating MMseqs2 cluster membership as an exact-sequence equivalence relation.

A.3 Training-Set Subsampling

Before family-balanced sampling, we retained training proteins between 80 and 1,000 amino acids, reducing the full training partition from 404,867 to 371,170 proteins. For each C30 group f, define

$$\mathcal{P}_f^{\mathrm{eligible}} = \{p \in \mathrm{Train} : C_{30}(p) = f,\ 80 \le |p| \le 1000\}.$$

Using random seed 13, we sampled at most two proteins uniformly without replacement from every nonempty eligible C30 group:

$$\mathrm{Train}_{\mathrm{condensed}} = \bigcup_{f:\,|\mathcal{P}_f^{\mathrm{eligible}}|>0} \mathrm{Sample}_{\min(2,|\mathcal{P}_f^{\mathrm{eligible}}|)}(\mathcal{P}_f^{\mathrm{eligible}}). \tag{1}$$

This retains all 86,945 eligible training C30 groups and produces 127,619 training proteins. No subsequent global-size cap was applied. Validation and test partitions were not subsampled

[4] The MMseqs2 version is 13.45111. C30 are clustered with `linclust --min-seq-id 0.3 -c 0.8 --cov-mode 1` and C90 with `linclust --min-seq-id 0.9 -c 0.9 --cov-mode 1`.

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.

Table 8: 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

B.2 Record in SwissProt

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.

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 9).
- **Sites:** Individual residues or short spans associated with specific functional roles such as catalysis or ligand binding (4 feature types; see Table 9).

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

Table 9: 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.

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.

B.3 Data Source and Field Selection Rationale

We source all annotations from UniProtKB/Swiss-Prot[5], 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.

[5]We use UniProtKB release 2025_03. All entries are restricted to the reviewed (Swiss-Prot) partition to ensure manual curation quality.

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.

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.

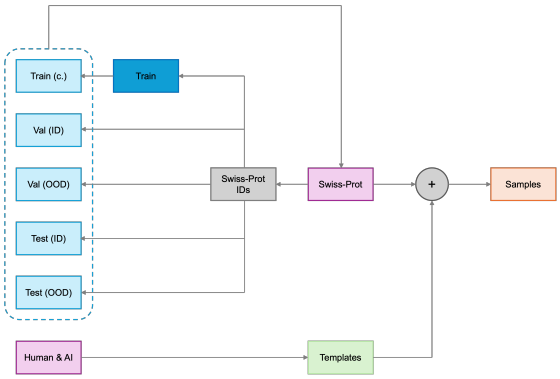


Table 10: End-to-end dataset construction pipeline, from accession-level splitting to template-based sample instantiation.

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

All native task-view record 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. Their raw format is presented in record as follows

P1 Record


```
{
  "task_id": "PGV0:P1:P12345",
  "record_id": "P12345",
  "split": "test_id",
  "task_family": "P1_protein_labels",
  "labels": {
    "go_mf": [
      {
        "label_id": "GO:0005524",
        "label_text": "ATP binding"
      }
    ],
    "ec": [
      {
        "label_id": "2.7.11.1"
      }
    ]
  }
}
```


P2 Record


```
{
  "task_id": "PGV0:P2:P12345",
  "record_id": "P12345",
  "split": "test_id",
  "task_family": "P2_feature_typing",
  "feature_types_present": [
    "active_site",
    "ligand_or_cofactor_binding_site",
    "domain",
    "transmembrane_region"
  ]
}
```


P3 Record


```
{
  "task_id": "PGV0:P3:P12345:active_site:proton_donor",
  "record_id": "P12345",
  "split": "test_id",
  "task_family": "P3_grounding",
  "query": {
    "canonical_type": "active_site",
    "feature_detail": "Proton donor"
  },
  "target_spans": [
    {
      "start": 155,
      "end": 155
    }
  ]
}
```


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. The space cardinarity is shown in table 11 Missing labels should be treated as unobserved in the dataset construction sense; they are treated as negatives only inside an explicitly closed evaluation

Table 11: P1 label-vocabulary sizes by namespace.

Namespace	# Labels	Namespace	# Labels
GO-BP	16,269	Catalytic activity	29,277
GO-MF	7,643	Cofactor	6,787
GO-CC	2,740	Subcellular location	30,085
EC	5,540	Pathway	9,479

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 assuming exactly one site.

Task P1 Record Spec: Protein Label Recognition


```
{
  "task_id": "RBV0:P1:A0A009IHW8",
  "record_id": "A0A009IHW8",
  "split": "train_condensed",
  "task": "P1_protein_label_recognition",
  "input": {
    "instruction": "Predict protein-level biological
        labels for the requested namespaces.",
    "query_namespaces": [
      "catalytic_activity",
      "ec",
      "go_bp",
      "go_mf"
    ],
    "sequence_ref": "A0A009IHW8"
  },
  "gold": {
    "labels_by_namespace": {
      "ec": [
        {
          "label_id": "3.2.2.-",
          "label_text": "3.2.2.-",
          "normalized_label": "3.2.2.-",
          "quality_tier": "medium"
        }
      ],
      "go_mf": [...]
    }
  },
  "metrics": [
    "micro_f1",
    "macro_f1",
    "label_ranking_average_precision",
    "auprc"
  ]
}
```

Task P2 Record Spec: Feature Type Recognition


```
{
  "task_id": "RBV0:P2:A0A009IHW8",
  "record_id": "A0A009IHW8",
  "split": "train_condensed",
  "task": "P2_feature_type_recognition",
  "input": {
    "instruction": "Predict which curated site/region
        feature types are present.",
    "sequence_ref": "A0A009IHW8"
  },
  "gold": {
    "feature_families_present": [
      "region",
      "site"
    ],
    "feature_family_counts": {
      "region": 2,
      "site": 6
    },
    "feature_type_counts": {
      "active_site": 1,
      "coiled_coil_region": 1,
      "domain": 1,
      "ligand_or_cofactor_binding_site": 4,
      "site": 1
    },
    "feature_types_present": [
      "active_site",
      "coiled_coil_region",
      "domain",
      "ligand_or_cofactor_binding_site",
      "site"
    ]
  },
  "metrics": [
    "micro_f1",
    "macro_f1",
    "exact_set_match"
  ]
}
```


Task P3 Record Spec: Evidence Grounding


```
{
  "task_id": "RBV0:P3:000011962ca80292",
  "record_id": "Q568Z0",
  "split": "train_condensed",
  "task": "P3_grounding",
  "input": {
    "instruction": "Locate all residue positions or sequence
        regions matching disordered_region.",
    "query": {
      "canonical_type": "disordered_region",
      "feature_detail": "Disordered",
      "feature_family": "region",
      "source_namespace": "regions"
    },
    "sequence_ref": "Q568Z0"
  },
  "gold": {
    "spans": [
      {
        "annotation_id": "RBV0:91f8f8d9d705206d",
        "start": 1,
        "end": 59
      }
    ]
  },
  "metrics": [
    "span_exact_match",
    "span_iou",
    "site_precision_recall_f1"
  ]
}
```


Using multi-choice examples. 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. Their examples are referred as follows:

Example 1: P1 Semantic Function Prediction

Source: `pannotground_llm_qa_test_id.jsonl`
Question ID: `PGLLM:test_id:P1:00001:O15381`
Protein: `O15381` (sequence length = 856)
Task: `P1:go_mf` — **Answer type:** Single choice

Question:

```
Which GO molecular-function annotation is curated for this
protein?
```

Options:

Option	Candidate answer
A	alpha-L-rhamnosidase activity (GO:0030596)
B	ATP binding (GO:0005524)
C	lipoprotein lipase activator activity (GO:0060230)
D	dTDP-6-deoxy-L-talose 4-dehydrogenase activity (GO:0047003)

Gold answer: `B`

Note: This example represents P1 because the model must infer a protein-level semantic function label from the sequence.

Example 2: P2 Feature-Type Recognition

Question ID: `PGLLM:test_id:P2:00006:P70289`
Protein: `P70289` (sequence length = 1705)
Task: `P2:feature_type_set` — **Answer type:** Multi choice

Question:

```
Select all curated site/region feature types present in this
protein.
```

Options:

Option	Candidate feature type
A	`active_site`
B	`compositional_bias_region`
C	`domain`
D	`ligand_or_cofactor_binding_site`
E	`disordered_region`
F	`lipidation_site`
G	`coiled_coil_region`
H	`transmembrane_region`

Gold answers: `A, C, D, E, H`

Gold feature types:

```
active_site
domain
ligand_or_cofactor_binding_site
disordered_region
transmembrane_region
```

Note: This example represents P2 because it is not asking for the biological function itself, but for which local annotation feature categories are present in the protein.

Example 3: P3 Local Evidence Localization

Question ID: `PGLLM:test_id:P3:00010:Q5H0H0`
Protein: `Q5H0H0` (sequence length = 589)
Task: `P3:transmembrane_like` — **Answer type:** Single choice

Question:

`Which candidate residue span or region corresponds to the localized evidence?`

Evidence query:
`transmembrane_region / Helical`

Protein-level context:
ATP binding; ABC-type oligopeptide transporter activity; plasma membrane; cell inner membrane; multi-pass membrane protein.

Options:

Option	Candidate local evidence
A	Residue span 226-246; local sequence window 218-254: `HQEVKIYGAQQTEMERYGALADRNLRLAMKVESTRGI`
B	Residue span 458-478; local sequence window 450-486: `DASQLERAILGANAMEFVAQLPEGLQSHVGAKGGRLS`
C	Residue span 29-49; local sequence window 21-57: `VAFAKPYRLLLVAALIAALIEAAGTTGFLALMKPITD`
D	Residue span 461-481; local sequence window 453-489: `QLERAILGANAMEFVAQLPEGLQSHVGAKGGRLSGGQ`
E	Residue span 506-526; local sequence window 498-534: `AMLKDAPILILDEATAALDNESERLVQDALHKLMPDR`

Gold answer: `C` — **Gold span:** residues `29-49`

Note: This example represents P3 because the label/query is already given, and the task is to localize the supporting residue region. Here the correct answer is a hydrophobic helical transmembrane segment.

B.6 Dataset Statistics

Structure-available subset. For structure-based evaluation, we retain proteins with AlphaFold DB structures, exact full-length sequence–structure mappings, and associated confidence metadata. The resulting `main_afdb_exact` subset contains 46,837 proteins (16.61% of the benchmark) and preserves all original split assignments. Table 12 reports coverage by split. Because structure availability differs across splits, sequence–structure comparisons use matched evaluation subsets.

Table 12: Protein coverage of the structure-available subset. Coverage is measured relative to the protein count in each original split. These counts describe proteins, rather than task-specific evaluation questions.

Split	All proteins	Structure subset	Coverage (%)
Train	127,619	10,735	8.41
Val-ID	22,587	11,708	51.84
Test-ID	22,014	11,235	51.04
Val-OOD	55,414	6,735	12.15
Test-OOD	54,387	6,424	11.81
Total	282,021	46,837	16.61

Multi-choice counts. The 6,000 P2T questions cover 5,905 unique gold proteins. Sequence length has mean 419.05, median 358, and 90th percentile 734 residues. Gold answer positions A–H account for 11.7%, 12.3%, 12.6%, 12.0%, 12.4%, 12.9%, 12.9%, and 13.2% of questions, respectively.

Table 13: Mean T2P query size and component counts. Global atoms, local types, and evidence descriptors are different annotation units.

Split	Characters	Words	Global atoms	Local types	Evidence
Test-ID	414.52	49.54	2.92	2.37	3.46
Test-OOD	444.35	52.67	3.42	2.39	3.42

Table 14: Candidate relations in the full compositional release. T2P entries are mean counts per 64-candidate query; P2T entries are exact counts across 6,000 eight-choice questions.

Relation to gold	T2P mean / query	P2T choices
Gold	1.00	6,000
Same global / different local	7.33	7,664
Same local / different global	10.21	7,420
Partial attribute overlap	16.00	—
Both different	16.00	11,997
Random non-equivalent	13.46	14,919
Total	64.00	48,000

T2P and P2T results are reported separately by split. P2T additionally reports both diagnostic subsets; relation-conditioned errors accompany aggregate scores as defined in Appendix C.

B.7 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.

C Evaluation Protocol and Metric Definitions

This appendix defines the scoring rules for individual attribute recognition and compositional protein–text matching in PannotGround. It also specifies the diagnostic analyses used to assess sequence dependence, evidence sensitivity, and the structure of compositional errors.

C.1 Evaluation Overview and Common Conventions

P1, P2, and P3 denote biological attribute levels: protein-level annotations, local feature types, and query-specific residue or span evidence, respectively. We evaluate these targets through candidate-based recognition and native protein language model (PLM) readouts. Compositional protein-to-text (P2T) and text-to-protein (T2P) tasks evaluate matching between proteins and descriptions assembled from their global annotations and local descriptors. Table 15 summarizes the prediction objects and primary metrics.

Let u_{ic} be the score of candidate $c \in C_i$ for example i. Single-choice predictions use $\hat{y}_i = \arg\max_{c \in C_i} u_{ic}$, or the parsed final answer for generation-based models.

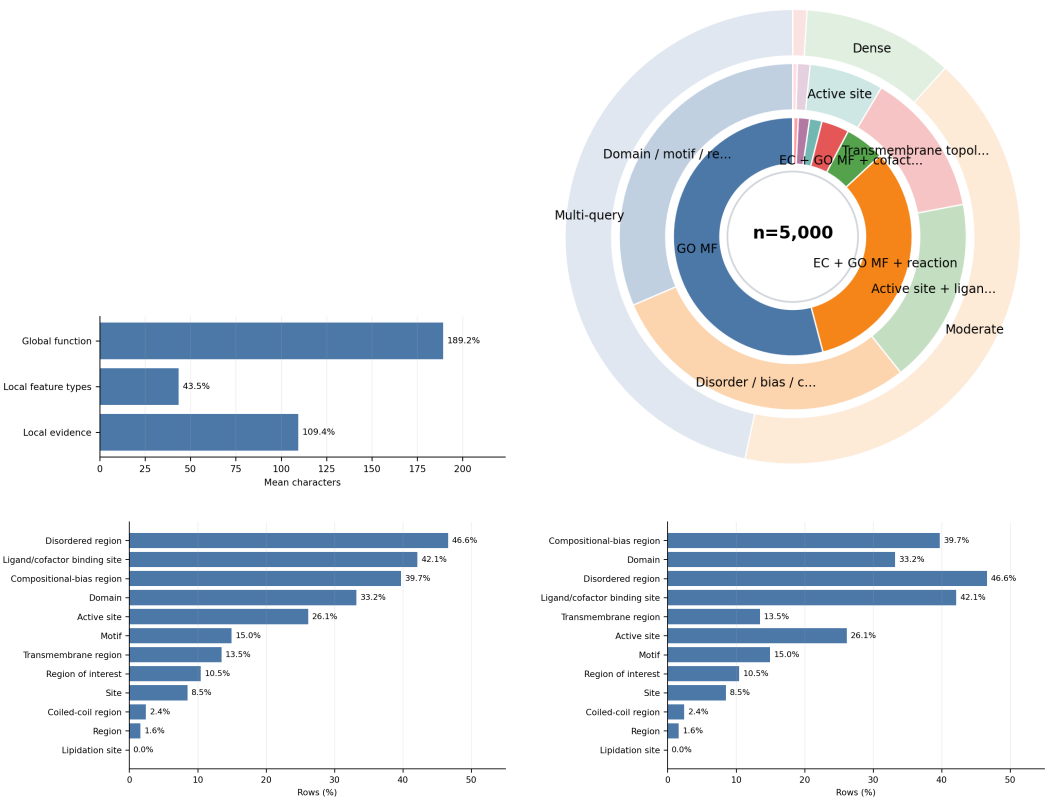


Figure 6: T2P compositional query structure. Top left: query text length by global, local-feature, and local-evidence component. Top right: joint query-composition hierarchy. Bottom left: local feature-type prevalence. Bottom right: local evidence-query prevalence.

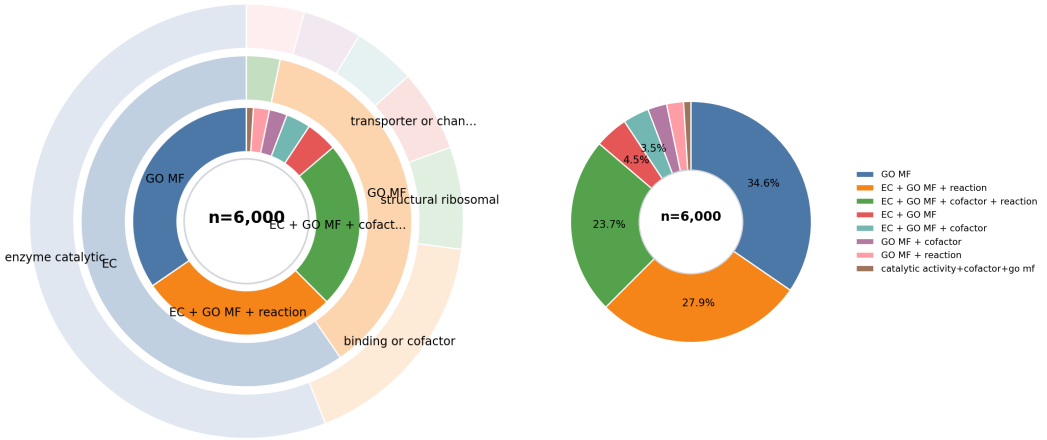


Figure 7: P2T global-composition visualizations. Left: global composition hierarchy, including primary global identity and rendered annotation bundles. Right: global annotation-bundle distribution.

PLM readouts. **Proto** uses similarities to reference vectors built from labeled training embeddings with a frozen encoder. **FT** maps task-head outputs to candidate labels or feature types; a P3 span receives the mean residue logit over that span. Encoder updates differ by model and belong to

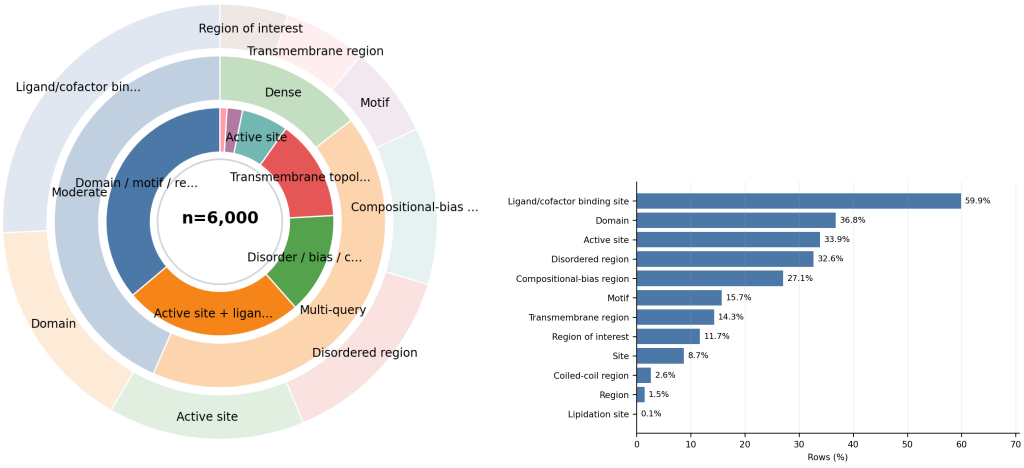


Figure 8: P2T local-composition visualizations. Top left: P2/P3 local-composition hierarchy. Top right: P2 feature presence.

Table 15: Evaluation settings and primary metrics. MCQA denotes multiple-choice question answering. Native PLM metrics use the gold number of positive labels or residues to determine the prediction cardinality. Secondary metrics and diagnostic analyses are defined in the subsequent subsections.

Task	Prediction object	Primary metric
Single-level candidate-based recognition		
P1-mcqa	One annotation candidate	Accuracy
P2-mcqa	A set of feature-type candidates	Mean set F1
P3-mcqa	One evidence-span candidate	Accuracy
Compositional matching		
P2T	One caption selected from eight candidates	Accuracy (R@1)
T2P	A ranking of candidate proteins	R@1 and MRR
Native PLM readouts		
P1-cls	A set of protein-level annotation labels	Top-k-by-gold micro-F1
P2-cls	A set of local feature-type labels	Top-k-by-gold micro-F1
P3-cls	A set of residues matching the evidence query	Top-k-by-gold residue micro-F1

the training configuration. The factorized P2T bridge averages its available global, local-type, and local-evidence component scores.

Protein–text similarity. Alignment models score protein–caption compatibility. PLM-TextAlign uses cosine similarity after a ridge map from frozen protein embeddings to PubMedBERT caption embeddings. Task-specific rerankers retain their own compatibility score.

Language-model scoring. Likelihood interfaces score candidate tokens by

$$u_{ic} = \frac{1}{|c|} \sum_{m=1}^{|c|} \log p_\theta(c_m \mid x_i, q_i, c_{<m}), \tag{2}$$

where x_i is the protein input, q_i the task prompt, and $|c|$ the candidate token count. For P2, generation returns a candidate set; the likelihood interface uses

$$d_{ic} = \log p_\theta(\text{yes} \mid x_i, q_i, c) - \log p_\theta(\text{no} \mid x_i, q_i, c), \qquad \widehat{Y}_i = \{c \in C_i : d_{ic} \geq \tau\}. \tag{3}$$

Thresholds and other calibrated settings are selected on validation data and fixed for test evaluation. Prompted interfaces are scored from parsed answers.

C.2 Task Metrics

Single-choice accuracy and retrieval. For N examples, let y_i be the gold candidate and r_i its one-based rank. P1-mcqa, P3-mcqa, and P2T share

$$\mathrm{Acc} = \frac{1}{N}\sum_i \mathbf{1}[\hat{y}_i = y_i]. \tag{4}$$

For complete rankings,

$$\mathrm{R@}k = \frac{1}{N}\sum_i \mathbf{1}[r_i \le k], \qquad \mathrm{MRR} = \frac{1}{N}\sum_i \frac{1}{r_i}, \qquad \mathrm{MeanRank} = \frac{1}{N}\sum_i r_i. \tag{5}$$

Thus R@1 equals accuracy. T2P uses either 64 candidates or the full pool of 282,021 proteins; these settings are reported separately. Uniform random R@1 is $N^{-1}\sum_i |C_i|^{-1}$: 12.5% for P2T and 1.56% for 64-way T2P.

Set metrics. For nonempty gold set Y_i, prediction $\widehat{Y}_i$, and $a_i = |\widehat{Y}_i \cap Y_i|$, define

$$F_i = \frac{2a_i}{|\widehat{Y}_i| + |Y_i|}, \qquad J_i = \frac{a_i}{|\widehat{Y}_i \cup Y_i|}, \qquad E_i = \mathbf{1}[\widehat{Y}_i = Y_i]. \tag{6}$$

P2-mcqa reports $\mathrm{SetF1} = N^{-1}\sum_i F_i$ as primary, with mean Jaccard $N^{-1}\sum_i J_i$ and exact match $N^{-1}\sum_i E_i$ as secondary metrics. Empty predictions receive zero on all three.

Native PLM labels and residues. Use Y_i for a gold label set (P1/P2) or residue set (P3), and let $\widehat{Y}_i^k$ contain the $k_i = |Y_i|$ highest-scoring elements. Since predicted and gold cardinalities match,

$$\mathrm{TopK\text{-}MicroF1} = \frac{\sum_i |\widehat{Y}_i^k \cap Y_i|}{\sum_i |Y_i|}. \tag{7}$$

This is an oracle-cardinality measure: the number of positives is supplied by the gold target. Mean Jaccard and exact match use Eq. 6 with $\widehat{Y}_i^k$.

For P3, convert predicted residues into maximal contiguous spans. With inclusive residue coordinates, $\mathrm{IoU}(g,h) = |g \cap h|/|g \cup h|$. For gold spans S_i and predicted spans $\widehat{S}_i$,

$$\mathrm{MeanBestIoU} = \frac{\sum_i \sum_{g \in S_i} \max_{h \in \widehat{S}_i} \mathrm{IoU}(g,h)}{\sum_i |S_i|}. \tag{8}$$

An empty prediction has zero overlap. Matching is gold-directed; extra spans are not separately penalized by this diagnostic.

D Individual Task Results

This appendix reports individual P1/P2/P3 results in the same attribute-level order used throughout the benchmark.

D.1 Single-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 randomly sampled 8,192 training examples from training set and 2,048 examples per test split for the controlled-pilot setting. Each table cell in table 16 and 17reports ID/OOD.

The PLM classification tables would be regard as encoder-capacity baselines. They show whether sequence representations expose global labels, local feature-type labels, and residue-level evidence.

Table 16: P1-cls global semantic annotation performance (top-k-by-gold Micro-F1). Each cell reports ID/OOD scores across different annotation categories.

	EC & Gene Ontologies				Specific Attributes			
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

Table 17: PLM performance on P2 (feature-type classification) and P3 (residue localization) tracks. Each cell reports ID/OOD scores.

	P2 Classification		P3 Localization	
Model	Micro-F1	Jaccard	Residue F1	IoU
Random	0.198 / 0.189	0.105 / 0.103	0.391 / 0.396	0.054 / 0.055
ESM-2	**0.729 / 0.725**	**0.675 / 0.666**	**0.612 / 0.614**	**0.405 / 0.418**
ProtBERT	0.690 / 0.695	0.623 / 0.627	0.475 / 0.465	0.180 / 0.179
Ankh	0.587 / 0.591	0.506 / 0.509	0.461 / 0.463	0.146 / 0.145
ProstT5	0.567 / 0.563	0.484 / 0.488	0.505 / 0.495	0.194 / 0.204
ProGen	0.485 / 0.468	0.357 / 0.340	0.464 / 0.474	0.113 / 0.112

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. The results are shown in table 18. The result for the structure sets are reported in table 19

Table 18: MCQA results on the ID and family-level OOD test splits, reported as **ID / OOD**. PLM rows use candidate scores from Proto frozen encoders or fine-tuned heads; LLM and protein-LLM rows use prompted MCQA or likelihood scoring.

Model	Family	Readout	Full	P1	P2	P3
Ankh	PLM	Proto	.681 / .703	.780 / .770	.559 / .606	.704 / .735
		Fine-tune	– / –	.730 / .750	.485 / .521	– / –
ESM-2	PLM	Proto	.601 / .607	.770 / .750	.592 / .588	.439 / .480
		Fine-tune	– / –	.740 / .720	.654 / .686	– / –
ProGen	PLM	Proto	.467 / .491	.690 / .660	.493 / .494	.214 / .316
		Fine-tune	– / –	.720 / .730	.596 / .588	– / –
ProstT5	PLM	Proto	.678 / .653	.770 / .760	.581 / .536	.684 / .663
		Fine-tune	– / –	.720 / .680	.579 / .578	– / –
ProtBERT	PLM	Proto	.567 / .533	.730 / .690	.490 / .477	.480 / .429
		Fine-tune	– / –	.710 / .720	.608 / .638	– / –
BioReason-Pro-SFT	Protein-LLM	Protein prompt	.138 / .132	.170 / .170	.022 / .003	.225 / .225
ProLLaMA	Protein-LLM	Sequence-text likelihood	.262 / .300	.150 / .310	.401 / .403	.235 / .184
Llama-3.1-8B	Text LLM	Text prompt	.277 / .302	.340 / .400	.296 / .319	.194 / .184
Qwen2.5-7B	Text LLM	Text prompt	.255 / .295	.330 / .440	.291 / .270	.143 / .174
Qwen3-4B	Text LLM	Text prompt	.285 / .297	.330 / .360	.339 / .376	.184 / .153

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.

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.

Table 19: Native P1/P2/P3 performance on matched structure-available evaluation subsets. All models within each task are evaluated on the same examples. Scores are arithmetic means of Test-ID and Test-OOD results. Metrics follow Appendix C. Bold indicates the highest score in each column.

Model	Input	P1	P2	P3
ESM-2	Sequence	0.1467	**0.6729**	**0.5618**
ESM-IF	Structure	0.1175	0.5143	0.4893
ESM-2+ESM-IF	Sequence + structure	**0.1879**	0.6151	0.5386
SaProt	Sequence + structure	0.1664	0.5782	0.4961

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.

Table 20: 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

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 $s_{\text{local}} - s_{\text{bg}}$. 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 in table 21. 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.

E Compositional Evaluation Results

This appendix reports the compositional evaluation of PannotGround. 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 21: 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.

		Candidate Ranking			Intervention Diagnostics	
Site Subtype	n	MC-Acc	MC-MRR	MC-Rec@5	Necessity	Sufficiency
In-Distribution (ID)						
Active site	7,675	38.33	0.589	88.05	0.527	0.148
Binding site	50,378	32.46	0.536	84.87	0.566	0.204
Generic site	2,566	21.63	0.423	68.28	0.517	0.038
Metal binding	28	35.71	0.594	89.29	0.383	-0.041
All sites	60,647	32.75	0.538	84.58	0.559	0.189
Out-of-Distribution (OOD)						
Active site	19,596	34.05	0.553	85.90	0.540	0.126
Binding site	124,874	31.67	0.529	84.35	0.568	0.209
Generic site	4,159	20.65	0.430	73.77	0.498	0.065
Metal binding	25	24.00	0.495	72.00	0.671	0.166
All sites	148,654	31.67	0.530	84.25	0.562	0.194

table. Finally, we decompose P2T behavior into diagnostic axes: global versus local discrimination, wrong-answer relation structure, shortcut behavior, and biology-conditioned difficulty.

E.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 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 22 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.

E.2 P2T Scoring Example

Table 23 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

Table 22: 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`, `prost t5`	Factorized caption scoring from independent annotation heads	Not evaluated
PLM-TextAlign	`ankh`, `esm2`, `progen`, `prost t5`	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

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.

Table 23: 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

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

Table 24: T2P 64-way compositional retrieval results on **Val** (1,000 rows) and **Test** (2,000 rows). The last four columns break down top-1 failure cases: **Same Global** (same global, wrong local), **Same Local** (same local, wrong global), **Partial** match, and **Other**.

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

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.

E.3 T2P Retrieval

Table 24 reports the original full-compositional 64-way T2P alignment evaluations on `val_id` and `test_id`. Tables 25 and 26 add the newly completed controlled-variant and full-pool retrieval runs.

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 25 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 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.

E.4 P2T Comparative Accuracy

Table 27 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.

Table 25: Controlled T2P 64-way retrieval results. Scores report percentage R@1 and R@5 across `test_id` (ID) and `test_ood_family` (OOD) splits.

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

Table 26: Full-pool T2P retrieval on original full-compositional queries. The gold protein is ranked against the canonical pool of 282,021 proteins (random R@1 is .00035%).

	ID (`test_id`)					OOD (`test_ood`)				
Model	R@1	R@10	R@50	MRR	Mean rank	R@1	R@10	R@50	MRR	Mean rank
`protrek_35m`	**.60**	**5.25**	**20.60**	**.0258**	**4,169.84**	.65	6.00	23.10	.0291	3,869.87
`protrek_650m`	.20	4.85	20.15	.0225	5,281.47	**.70**	**7.25**	**24.45**	**.0317**	**3,738.37**
`protst_esm2`	.10	.25	.75	.0017	51,104.57	.00	.15	.85	.0009	48,237.14
`protst_protbert`	.00	.15	.75	.0008	52,412.30	.00	.00	.90	.0008	54,859.38

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 below random.

Table 27: P2T held-out 8-way compositional caption-selection accuracy (%). Underlined with gray background indicates the best performance, and underlined indicates the second-best performance per column (excluding baseline). SL/DG: same-local/different-global; SG/DL: same-global/different-local.

	ID (test_id)			OOD (test_ood_family)		
Model	Overall	SL/DG	SG/DL	Overall	SL/DG	SG/DL
Random Baseline						
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
esm2	9.40	7.90	10.90	9.10	7.90	10.30
progen	7.35	6.30	8.40	6.95	6.10	7.80
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
esm2	38.00	37.10	38.90	34.80	33.70	35.90
progen	20.75	18.50	23.00	21.40	18.30	24.50
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
qwen25_7b	13.65	13.10	14.20	13.50	13.00	14.00
qwen3_4b	12.95	12.90	13.00	12.15	12.00	12.30
Alignment Models						
protrek_35m	60.50	62.80	58.20	54.20	56.30	52.10
protrek_650m	58.45	62.30	54.60	54.45	55.00	53.90
protst_esm2	40.65	43.90	37.40	35.80	39.50	32.10
protst_protbert	37.85	39.60	36.10	35.80	39.80	31.80
prott3_base	44.60	48.90	40.30	37.00	39.70	34.30
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
bioreason_pro_sft	8.40	8.30	8.50	8.10	9.40	6.80

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 28: 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

Table 28 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

Table 29: Controlled P2T accuracy by model. Underlined with gray background indicates the best performance, and underlined indicates the second-best performance per column (excluding baseline). Attr1 is an individual-task reference; Full is the original full P1/P2/P3 compositional P2T benchmark.

	Attr1		Attr2		Attr4		Global-only		Local-only		Full	
Model	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
Random Baseline												
analytic_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
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
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
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
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
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
prostt5	24.98	24.60	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
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
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
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
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
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
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
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
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

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 four attributes.

Table 29 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, binding 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 23 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.

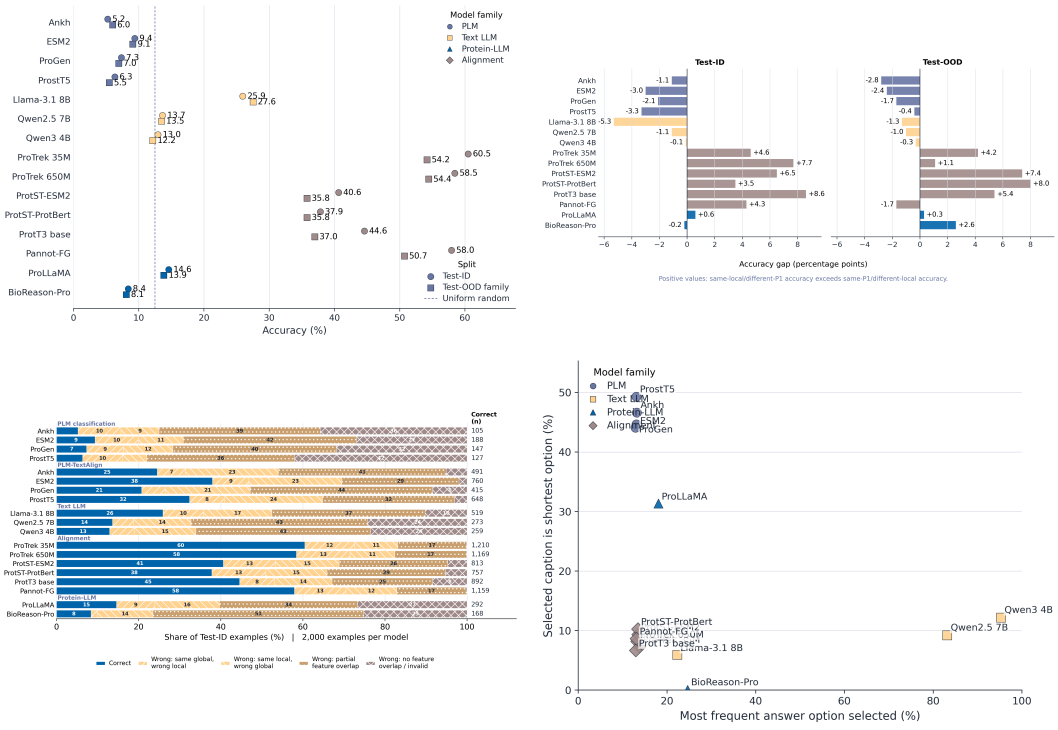


Figure 9: Core P2T compositional behavior visualizations. Top left: held-out accuracy. Top right: difference between global-function and local-evidence diagnostics. Bottom left: relation type of wrong answers on test-id. Bottom right: answer-position and caption-length shortcuts.

E.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 30 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 27. 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

Table 30: 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 31: Aggregated P2T behavior taxonomy over the completed validation and test splits. Accuracies and behavior rates are percentages. Values are percentages except the behavior label. Local adv. is Same-global task minus Same-local task, reported in percentage points; its negative values indicate that local-evidence discrimination under fixed P1 is harder. Hard-neg. wrong is computed only among incorrect predictions: it is the percentage of wrong answers whose selected caption preserves either Same-global or Same-local. Underlined with gray background indicates the best performance, and underlined indicates the second-best performance per column.

Model	Acc	Same-local	Same-global	Local adv.	Hard-neg wrong	Behavior Label
PLM						
ankh	5.58	4.64	6.52	1.88	20.86	Length-biased factorized scorer
esm2	9.18	7.92	10.44	2.52	24.51	Length-biased factorized scorer
progen	7.20	6.32	8.08	1.76	22.59	Length-biased factorized scorer
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
esm2	36.02	34.65	37.40	2.75	51.04	Ridge-aligned caption matcher
progen	21.46	18.35	24.58	6.23	34.21	Ridge-aligned caption matcher
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
qwen25_7b	13.18	12.96	13.40	0.44	24.58	Position-biased generator
qwen3_4b	12.04	12.12	11.96	-0.16	26.40	Position-biased generator
Alignment Models						
protrek_35m	57.08	58.88	55.28	-3.60	56.90	Strong partial-compositional matcher
protrek_650m	55.70	57.76	53.64	-4.12	54.76	Balanced partial-compositional matcher
protst_esm2	38.42	41.96	34.88	-7.08	48.62	Balanced partial-compositional matcher
protst_protbert	36.74	39.92	33.56	-6.36	46.29	Balanced partial-compositional matcher
prott3_base	44.60	43.20	37.24	-5.96	44.40	Balanced partial-compositional matcher
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
bioreason_pro_sft	7.88	8.48	7.28	-1.20	17.63	Local-first partial matcher

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-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 31 reports the same diagnostic view at the model level. 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.

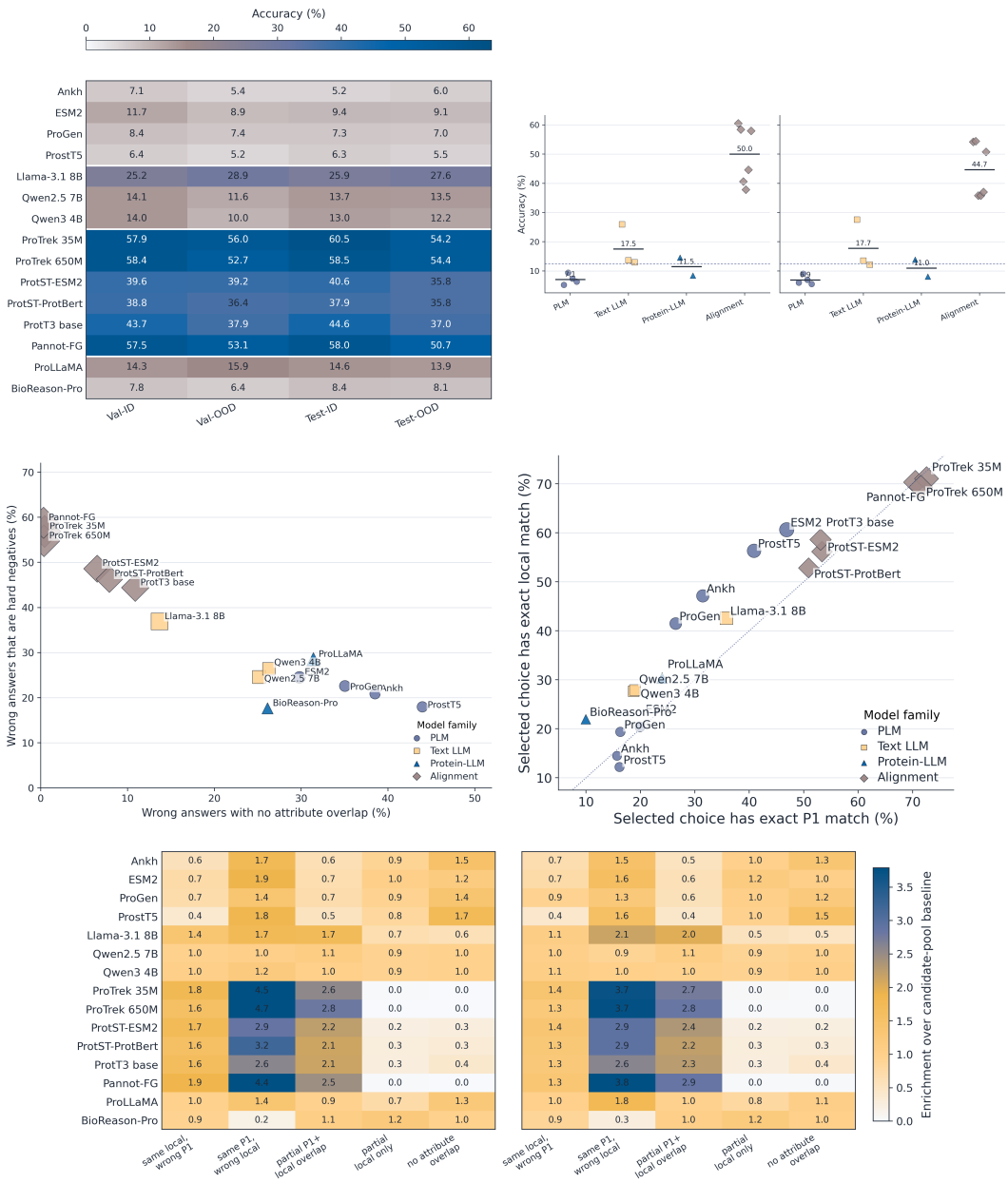


Figure 10: P2T model diagnostics beyond aggregate accuracy. Top left: model-by-split accuracy. Top right: held-out family-level separation. Middle left: error quality measured by hard-negative and no-overlap wrong selections. Middle right: whether selected captions preserve exact global function and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

Moreover, we have the results more structure-subset P2T task. The results are shown as table 32.

E.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 33 summarizes the most interpretable groups. Positive deltas indicate contexts easier than the corresponding model/split average; negative deltas indicate harder contexts.

Table 32: Controlled P2T accuracy (%) using PLM-TextAlign. Each value is the unweighted arithmetic mean of Test-ID and Test-OOD accuracies. The full-set ESM-2 row is included for reference; bold marks the highest score among the structure-subset rows. Candidate sets can differ between variants.

Model	Attr2	Attr4	Global-only	Local-only	Full
Full evaluation sets (reference)					
ESM-2	13.95	14.83	26.80	40.75	36.40
Structure-available subsets					
ESM-2	8.90	11.31	16.63	36.99	26.86
ESM-IF	7.63	9.76	14.01	37.11	25.30
ESM-2+ESM-IF	**9.17**	**12.19**	**19.61**	**39.79**	**29.17**

Table 33: Biological difficulty patterns for P2T local-evidence discrimination. Deltas are percentage points after model/split normalization.

Category	N	Delta	Interpretation
Global Namespace			
`ec+go_mf`	1,965	+7.29	Enzyme identity plus GO-MF can provide a strong global anchor.
`go_mf`	13,170	+1.66	GO-MF-only examples are modestly easier than average.
`catalytic_activity+cofactor+ec+go_mf`	8,790	-2.52	Dense enzyme/cofactor descriptions make local binding harder.
`cofactor+ec+go_mf`	1,050	-5.61	Cofactor-enriched enzyme contexts are difficult.
`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.
nucleic-acid binding or regulation	10,290	+1.07	Regulatory/DNA-binding contexts are mildly easier.
enzyme catalysis	23,970	-0.94	Enzyme contexts are slightly harder after normalization.
small-molecule or ion binding	21,360	-1.11	Binding/cofactor semantics increase ambiguity.
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.
active site	2,355	+1.33	Active-site-only cases are easier than average.
domain/motif/region	16,215	+1.04	Domain-like local evidence is mildly easier.
active site plus ligand binding	8,505	-2.63	Combining active-site and ligand-binding semantics is harder.
disorder/bias/coiled-coil	3,915	-3.70	Low-complexity and disorder evidence is the hardest local family.

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 evidence. 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.

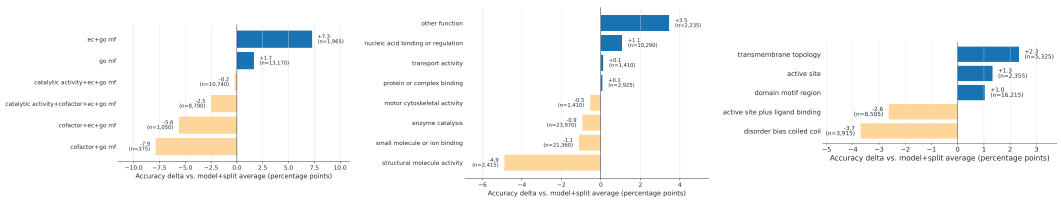


Figure 11: Biology-conditioned P2T difficulty diagnostics. Left: global namespace groups. Center: global function-slim groups. Right: local-evidence families.

F Fine-grained training for Pannot-FG

Model architecture. Pannot-FG is a protein–text model built upon Q-Former and LLM decoder. Learnable query tokens in the Q-former cross-attend to protein embeddings, collecting protein information into a compact set of representations.

Base training. Following the BLIP-2 stage-1 design, training combines three objectives: protein–text contrastive learning (PTC), which aligns paired representations using the maximum query–text cosine similarity; protein–text matching (PTM), which classifies matched and mismatched pairs; and language modeling (LM), which predicts annotation text conditioned on the protein queries. The base objective is

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{PTC}} + \lambda_{\text{PTM}}\mathcal{L}_{\text{PTM}} + \lambda_{\text{LM}}\mathcal{L}_{\text{LM}}. \tag{9}$$

Fine-grained supervision. The Pannot-FG checkpoint evaluated in our tables jointly uses two forms of residue-level supervision in addition to the base objective. *Local supervision* restricts the protein features supplied to the Q-Former to an annotated residue set and applies matching and language-modeling losses with its associated description. This encourages the model to associate local protein features with annotation text. *Masked-residue supervision* instead masks the annotated positions in the protein embedding sequence and applies a margin loss that favors the original protein–text pair over the masked pair. This encourages sensitivity to the annotated positions; it does not establish their biological necessity. The combined objective is

$$\begin{aligned}\mathcal{L}_{\text{FG}} = \mathcal{L}_{\text{base}} &+ \lambda_{\text{loc}}\big(\mathcal{L}_{\text{local-PTM}} + \mathcal{L}_{\text{local-LM}}\big) \\ &+ \lambda_{\text{cf}}\mathcal{L}_{\text{cf}},\end{aligned} \tag{10}$$

where $\mathcal{L}_{\text{cf}}$ is the masked-residue margin loss and λ_{loc} and λ_{cf} weight the two fine-grained objectives. We follow the model choice and hyperparameter as ProtT3. Our checkpoints uses ProtT3's stage1 checkpoint as base model, fine-grained supervised Q-former as the continual learning stage for stage1, and followed by the stage2 fine-tuning with LLM decoder. The dataset used for training is the task-view record based instruction set.

G Limitations

PannotGround is constructed from curated protein annotations and therefore reflects the coverage and granularity of the source database. Proteins with richer annotations are more likely to support multi-attribute evaluation, while sparsely annotated proteins are less well represented. Structure-based experiments further use a subset satisfying structure-availability and sequence-mapping requirements. Although comparisons within this subset use matched proteins, the findings may not generalize uniformly to other protein families or annotation regimes.

Our evaluation uses structured task views and template-based descriptions to enable controlled comparisons across models. This design covers a limited range of linguistic variation and does not directly assess open-ended biological explanation. Although P3 evaluates residue-level localization, the compositional tasks assess joint matching of global and local annotation descriptors without explicit residue coordinates. Extending the benchmark to paraphrased descriptions and position-aware compositional queries would broaden its coverage of protein–language understanding.