BEYOND STATIC ANNOTATION MATCHING: A TEMPORAL BENCHMARK FOR FREE-FORM PROTEIN FUNCTION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Advances in large language models have enabled protein models to generate detailed natural-language descriptions of protein function. However, evaluating such free-form predictions remains challenging, as existing benchmarks are primarily built around fixed labels or static annotations. We introduce ProLong-Bench, a temporal benchmark for free-form protein function prediction that evaluates models as protein annotations evolve over time. Across six representative Protein LLM systems, we observe substantial differences in performance across temporal settings. Models that perform well on existing functional annotations do not necessarily capture knowledge added later, while performance on newly added proteins shows a different pattern. These results highlight temporal evaluation as an important complement to conventional static benchmarking. We intend ProLong-Bench to provide a standardized and extensible evaluation framework for free-form protein function prediction, facilitating more comprehensive model comparison and supporting future progress in the field.

1 INTRODUCTION

Understanding protein function is fundamental to comparative genomics, experimental prioritization, variant interpretation, and protein engineering. Automated protein function prediction has traditionally focused on structured annotations such as Gene Ontology (GO) terms(Ashburner et al., 2000). Prior GO-specific methods range from sequence- and homology-based approaches (Altschul et al., 1990; 1997) to neural and protein-language-model predictors (Kulmanov & Hoehndorf, 2020; Wang et al., 2023; Rives et al., 2021), with more recent work incorporating literature and retrieval (You et al., 2018; Yan et al., 2024; 2025).

In recent years, rapid advances in protein language models and LLM-based systems have expanded the capabilities of computational protein modeling. Beyond conventional structured prediction, recent systems can generate free-form functional descriptions that capture substrates, reaction direction, binding partners, cellular roles, conditions, and exceptions. These systems reach free-form generation through several broad modeling paradigms, which we group into three families in Figure 1: *Modality-projection* methods that inject protein representations to language decoders, *modality-unification* methods that jointly model biological and textual modalities in a shared tokenizer, and *tool augmentation* methods that reason over retrieved evidence, domain-tool outputs, or other structured biological information to derive the final functional description (Fei et al., 2025; Liu et al., 2024; Pei et al., 2024; 2026; Luo et al., 2023; Fan et al., 2026; Fallahpour et al., 2026).

Despite this diversity in model design, evaluation has remained comparatively static. Existing protein-language benchmarks typically assess model outputs against fixed references, through settings such as protein captioning, question answering, and instruction following (Liu et al., 2024; Fang et al., 2024; Ye et al., 2025). While useful for measuring general protein-language capabilities, these benchmarks do not capture how model performance changes as functional annotations evolve over time. A concrete example from our benchmark illustrates this problem. For W1 accession F2R776 (Rifampicin monooxygenase), BioMatrix predicts that the protein is a monooxygenase that modifies rifampicin and inactivates a broad range of rifamycin antibiotics. The later Swiss-Prot annotation retains these functions but adds a new exception: "Does not act on rifabutin." BioMatrix does not mention this newly added exception (Appendix A.1). A static comparison against the final annotation

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.

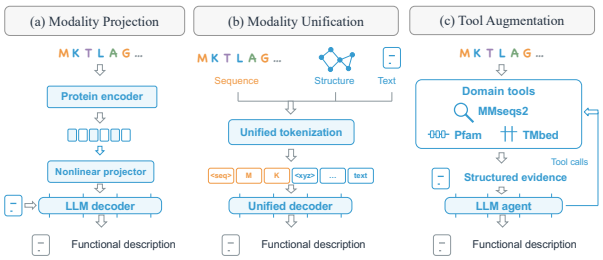


Figure 1: Integration paradigms for Protein LLMs.

mixes two distinct behaviors into a single score: recovery of established functional knowledge and coverage of information added later. On the other hand, temporal evaluation has been studied more extensively in structured protein function prediction, with benchmarks such as CAFA and LAFA assessing predictions against annotations that become available at later time points (Zhou et al., 2019; De Paolis Kaluza et al., 2026; Phan et al., 2026). However, these benchmarks focus on ontology-based outputs rather than detailed free-form functional descriptions. Recent work such as BioReason-Pro (Fallahpour et al., 2026) has extended temporal evaluation to free-form protein function generation. However, it neither characterizes how functional knowledge evolves over time, nor systematically compares distinct temporal groups to assess performance under different forms of temporal novelty.

To address this gap, we introduce ProLong-Bench, a temporal benchmark for free-form protein function prediction. ProLong-Bench separates longitudinal evaluation into three complementary settings: recovery of established functional knowledge for existing proteins, coverage of knowledge added through later annotation updates, and generalization to newly added proteins. This formulation distinguishes different forms of temporal change that are conflated by conventional static evaluation. Across six representative systems, we find that strong performance on established annotations does not necessarily translate into coverage of emerging knowledge, while newly added proteins exhibit a distinct performance pattern. These results show that protein novelty and annotation novelty represent different evaluation challenges, motivating temporal evaluation as an important complement to static protein-function benchmarking.

Our contributions are:

- **A longitudinal benchmark for free-form protein function prediction.** We introduce ProLong-Bench, which organizes temporal evaluation into three complementary regimes: established-knowledge recovery (Track A.0), emerging knowledge coverage (Track A.1), and generalization to newly added proteins (Track B).
- **A unified claim-level evaluation framework.** We evaluate all three regimes using atomic functional claims under a common support framework, with Track A.0/A.1 further enabling a controlled comparison of retained and newly added knowledge for the same protein sequences and model outputs.
- **A systematic evaluation across diverse Protein LLM systems.** We benchmark six representative systems spanning distinct modeling paradigms and complement the main evaluation with historical-information and robustness analyses to characterize differences in temporal functional knowledge.

2 Related Work

2.1 Protein LLMs for function understanding

Protein LLMs generate functional text through three broad paradigms. *Modality-projection* methods connect protein encoders to language models through learned adapters, often supplying protein representations as continuous soft prompts. ProtT3 and Prot2Text-V2 align sequence representations with language (Liu et al., 2024; Fei et al., 2025); ProteinChat uses structural inputs, while ProtChatGPT, ProteinGPT, Prot2Chat, and EvoLLa incorporate sequence and structure through projection or attention-based interfaces (Guo et al., 2025; Wang et al., 2025a; Xiao et al., 2025; Wang et al., 2025b; Zhou et al., 2026). *Modality-unification* methods, including BioMatrix, jointly model biological and textual modalities in shared token spaces (Pei et al., 2026; Luo et al., 2023).

Tool-augmented systems, such as PFUA and BioReason-Pro, incorporate domain tools, retrieved evidence, or structured biological reasoning (Fan et al., 2026; Fallahpour et al., 2026).

2.2 Benchmarks for protein function understanding

Protein-function benchmarks have traditionally focused on structured prediction, most prominently through Gene Ontology annotations. CAFA established a prospective evaluation setting in which predictions are made before later experimental annotations become available, and subsequent CAFA rounds and LAFA continue this longitudinal evaluation paradigm for ontology-based function prediction (Zhou et al., 2019; De Paolis Kaluza et al., 2026; Phan et al., 2026). PROBE (Çevrim et al., 2025) benchmarks protein representations on function-related tasks, including ontology-based prediction, supporting comparisons of representation baselines. ProLong-Bench instead evaluates temporal changes in free-form functional claims.

With the emergence of protein-language models, evaluation has expanded to protein captioning, question answering, retrieval, instruction following, and broader multi-task settings. ProtT3 evaluates protein captioning, protein QA, and protein–text retrieval (Liu et al., 2024); Mol-Instructions provides biomolecular instruction-following data (Fang et al., 2024); and ProteinBench evaluates protein foundation models across a broader range of prediction and generation tasks (Ye et al., 2025). These benchmarks largely rely on fixed references and therefore provide a static view of free-form functional prediction.

Recent work has begun to incorporate temporal splits into free-form protein-function generation. BioReason-Pro uses temporal evaluation as part of the assessment of a specific model under time-separated data (Fallahpour et al., 2026). In contrast, ProLong-Bench treats temporal change itself as the primary object of evaluation. It explicitly distinguishes retained functional knowledge from functional annotations added later for the same protein, while separately evaluating newly added proteins. This benchmark-centered formulation enables paired evaluation of retained and later-added functional content for the same protein sequence and frozen model output, thereby disentangling annotation novelty from protein novelty.

3 ProLong-Bench

3.1 Temporal regime construction

We use the Swiss-Prot FUNCTION field as the prediction target, as it is commonly used as a training target for protein function generation systems and is updated alongside changes in protein metadata, providing a free-form textual view of how the overall functional description of a protein evolves over time. Based on the documented training-data cutoffs of the evaluated methods, we construct data from three Swiss-Prot snapshots dated 2025-02-05, 2026-01-28, and 2026-09-03, defining two consecutive evaluation windows: **W1 (2025-02-05 → 2026-01-28) and W2 (2026-01-28 → 2026-09-03)**. W2 serves as the fully clean evaluation window, as it lies beyond the documented training-data cutoffs of all evaluated systems. Because W2 alone provides relatively limited coverage, we additionally include W1 to better represent the settings applicable to most models. W1 may contain a small amount of temporal overlap for some systems, but it provides broader coverage and also allows us to examine model behavior over a longer time span. All model-specific dates are summarized in Appendix C.

Within each window, ProLong-Bench organizes evaluation according to how proteins and their functional annotations change between the historical and target snapshots. **Track A** follows proteins that persist across the window, while **Track B** considers proteins that first appear in the target snapshot. For Track A, we further distinguish whether an evaluated functional statement was already represented historically: **A.0** contains retained statements, whereas **A.1** contains newly added statements. As shown in Figure 2a, this hierarchical construction yields three evaluation regimes: established-knowledge recovery (A.0), emerging knowledge coverage (A.1), and generalization to new-entry proteins (Track B).

Our primary paired construction comes from Track A. We retain accessions that are present at both endpoints, have unchanged sequences, and receive updated FUNCTION annotations. Holding the sequence fixed also fixes the model prediction, allowing retained A.0 and newly added A.1 content to be evaluated against the same answer. This yields 291 proteins in W1 and 243 proteins in W2. Track B complements this paired setting with proteins that first appear in Swiss-Prot at the target snapshot.

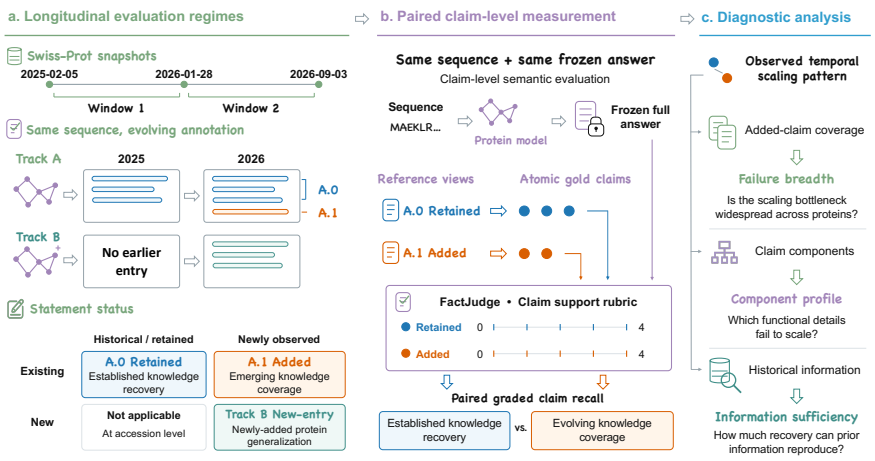


Figure 2: Overview of the ProLong-Bench temporal evaluation framework.

Their complete target FUNCTION annotations are evaluated using the same claim-support protocol, with dataset statistics summarized in Table 1.

Table 1: Dataset statistics by temporal regime and evaluation window.

	Track A.0		Track A.1		Track B	
Window	Proteins	Claims	Proteins	Claims	Proteins	Claims
W1	291	3,643	291	690	157	456
W2	243	4,434	243	573	90	244

3.2 Claim-level semantic measurement

Free-form functional descriptions can express the same biological fact in substantially different language, making whole-description similarity insufficient for temporal evaluation. We therefore evaluate whether model outputs cover the functional content represented by each reference, rather than how closely they reproduce its wording. As shown in Figure 2b, each reference view is decomposed into atomic functional claims, and FactJudge measures the degree to which the full model answer supports each claim on a common 0–4 scale.

This formulation places all three temporal regimes on the same semantic scale. In Track A, retained A.0 and newly added A.1 claims are evaluated against the same frozen model answer, enabling a paired comparison of coverage of established and emerging knowledge. Track B applies the same procedure to the complete target FUNCTION annotation of newly added proteins. We normalize each claim score and use protein-macro graded recall as the primary endpoint, giving each protein equal weight regardless of annotation length.

Beyond overall claim support, FactJudge records component-level diagnostics that characterize which aspects of a functional claim are recovered, including entities, relations, specificity, conditions, and certainty. These diagnostics are used to analyze the nature of emerging knowledge failures, while conventional lexical and whole-description metrics are reported separately for comparison. We further validate FactJudge against expert consensus, confirming that the LLM judge provides reliable claim-level semantic support scores. Full validation procedures, scoring prompts, and implementation details are provided in Appendix B.4 and Appendix B. We additionally apply the prediction-relevance judge to a high-risk subset selected for apparent prediction–reference mismatch. This targeted diagnostic distinguishes unrelated content, related extensions, and restatements within full model answers. Results show that explicit contradictions with reference claims are rare across the evaluated models, whereas missing support for emerging functional knowledge remains widespread, indicating that omission is the dominant failure mode under our reference-based evaluation(Appendix B.5). Because the reference annotation is not exhaustive, prediction-side content absent from the current reference cannot be reliably treated as incorrect. We therefore prioritize gold-side recall as the primary endpoint, asking whether models cover functional knowledge that is subsequently incorporated into Swiss-Prot rather than penalizing all reference-external content as false positives.

3.3 VALIDATION AND DIAGNOSTIC ANALYSES

The primary comparison first establishes how protein-function performance changes across temporal regimes, with the paired A.0/A.1 setting providing the clearest contrast between retained and later-added functional content. We then use a sequence of diagnostics to understand the resulting scaling pattern, as shown in Figure 2c.

First, we ask whether limited A.1 coverage reflects a small number of difficult cases or a broader failure across proteins. We therefore measure the prevalence of usable support for added claims at the protein level, characterizing the *failure breadth* of emerging knowledge coverage. Second, we ask which aspects of functional content fail to scale. FactJudge component scores decompose claim support into entities, relations, specificity, conditions, and certainty, providing a *component profile* of the missing information. Finally, strong recovery of retained annotations raises the question of how much can already be reproduced from information available before the target release. Historical annotation copying and leave-accession-out homology transfer probe this *information sufficiency*, while public-reference representation analysis characterizes its association with model behavior.

4 RESULTS AND ANALYSIS

To better understand how protein-function understanding models behave under temporal change, we organize our analysis around three research questions. RQ1 first establishes the overall performance landscape across retained knowledge, emerging knowledge, and newly added proteins, asking whether strong conventional static performance transfers uniformly across these regimes. Building on this landscape, RQ2 focuses on the novel side, using claim-level evaluation to compare annotation novelty with protein novelty and to identify which functional components are most often missing from later-added claims. RQ3 turns to the retained side, examining how closely retained-knowledge behavior tracks historically available information, through historical replay, training-reference distributions, and claim-level associations with public-reference representation.

4.1 RQ1: HOW DO SYSTEM STRENGTHS CHANGE ACROSS TEMPORAL KNOWLEDGE REGIMES?

We first establish the conventional static reference by evaluating complete target descriptions across both windows. Table 2 reports lexical overlap, embedding similarity, and deterministic fact matching against the full target annotation. **BioMatrix achieves the strongest agreement with complete target descriptions under conventional evaluation.** It has the highest overall scores and Track A Det-F1 in both W1 and W2, with overall Det-F1 increasing from 0.692 to 0.756. Prot2Text-V2 generally ranks second. In W2, BioMatrix reaches an overall Det-F1 of 0.756, compared with 0.249 for Prot2Text-V2, 0.142 for PFUA, and 0.061 for BioReason-Pro. This ordering is not fully universal even under conventional evaluation: Prot2Text-V2 leads Track B Det-F1 in W1 at 0.167, whereas BioMatrix leads in W2 at 0.155. These results provide the static reference against which we examine whether the same system strengths persist once functional content is separated by temporal status.

We then separate performance by temporal regime using protein-macro graded recall across A.0, A.1, and Track B in Figure 3. **Performance within each system changes sharply across temporal regimes: every evaluated system has lower recall on emerging A.1 knowledge than on retained A.0 knowledge in both windows.** BioMatrix provides the clearest example, with W2 recall falling from 0.829 on A.0 to 0.148 on A.1, corresponding to a paired A.1–A.0 contrast of -0.681. Because Track A evaluates the same frozen answer for the same protein sequence against retained and later-added claims, this contrast reflects a change in coverage with the temporal status of the target knowledge rather than a change in protein identity or generated output. Strong recovery of established functional content therefore does not imply comparable coverage of information added later to the same record.

Temporal decomposition changes not only performance within systems, but also their relative ordering. **Tool-augmentation systems show stronger coverage of the two novelty regimes despite weaker conventional and retained-side performance.** In W2, BioReason-Pro and PFUA reach A.1 recall of 0.211 and 0.210, respectively, compared with 0.148 for BioMatrix. On Track B, PFUA and BioReason-Pro reach 0.362 and 0.352, compared with 0.251 for BioMatrix, while BioMatrix remains strongest on retained A.0 knowledge. Thus, the system that best reproduces established or complete annotations need not be the system that best covers temporally novel functional information. The temporal regimes therefore reveal complementary system strengths that are obscured by a single conventional ranking.

Table 2: Conventional whole-description evaluation across the two windows. R-L and BERT denote ROUGE-L F1 and BioBERT BERTScore F1 on the overall cohort. Det-F1 denotes deterministic claim-matching F1 and is additionally reported for updated-entry Track A and newly-added-protein Track B. BioReason-Pro has $N = 457$ because five predictions are unavailable.

	W1						W2				
	Overall			Track A	Track B		Overall			Track A	Track B
Model	R-L	BERT	Det-F1	Det-F1	Det-F1	N	R-L	BERT	Det-F1	Det-F1	Det-F1
Modal-projection systems											
Prot2Text-V2	0.3308	0.788	0.316	0.336	**0.167**	462	0.3000	0.773	0.249	0.260	0.091
EvoLLa	0.0825	0.656	0.092	0.106	0.031	462	0.0927	0.658	0.106	0.116	0.034
BioT5+	0.1055	0.679	0.009	0.010	0.000	462	0.1063	0.668	0.007	0.007	0.000
Modal-unification systems											
BioMatrix	**0.5124**	**0.850**	**0.692**	**0.757**	0.134	462	**0.5763**	**0.870**	**0.756**	**0.788**	**0.155**
Tool-augmentation systems											
PFUA	0.1350	0.704	0.159	0.177	0.042	462	0.1327	0.708	0.142	0.149	0.042
BioReason-Pro	0.1127	0.711	0.069	0.077	0.018	457	0.1110	0.707	0.061	0.062	0.056

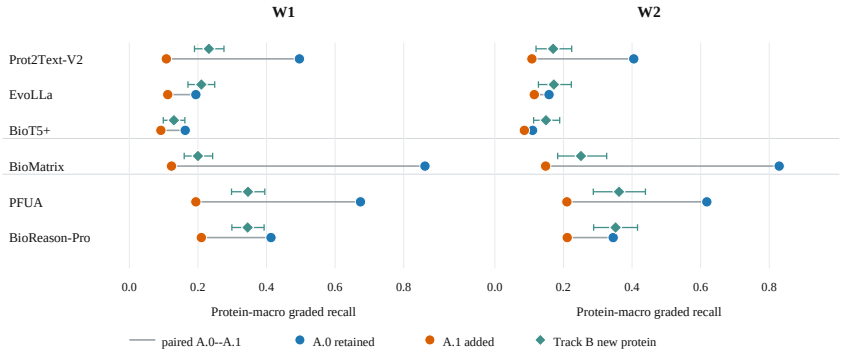


Figure 3: Protein-function performance across A.0, A.1, and Track B in W1 and W2. Track B whiskers denote 95% cluster-bootstrap intervals.

Take-Away for RQ1. System strengths are regime-dependent: BioMatrix leads on conventional and retained A.0 performance, while tool-augmentation systems are stronger on emerging A.1 knowledge and Track B new proteins, showing that the relative advantages of systems shift with the temporal status of the target knowledge.

4.2 RQ2: How do models behave across different forms of temporal novelty?

Table 3: Claim-level coverage across Track A.0, Track A.1, and Track B in W2. FJ-R denotes protein-macro graded FactJudge recall. Miss is the fraction of proteins with no claim receiving usable support ($s \geq 2$).

	A.0	A.1		Track B		Track A paired
Model	FJ-R ↑	FJ-R ↑	Miss ↓	FJ-R ↑	Miss ↓	A.1−A.0 [95% CI]
Prot2Text-V2	0.405	0.108	0.893	0.170	0.700	-0.297 [-0.367, -0.234]
EvoLLa	0.158	0.115	0.922	0.172	0.722	-0.043 [-0.072, -0.018]
BioT5+	0.110	0.086	0.905	0.149	0.733	-0.024 [-0.044, -0.005]
BioMatrix	**0.829**	0.148	0.802	0.251	0.589	-0.681 [-0.733, -0.625]
PFUA	0.618	0.210	0.724	**0.362**	**0.411**	-0.408 [-0.470, -0.348]
BioReason-Pro	0.345	**0.211**	**0.723**	0.352	**0.411**	-0.134 [-0.179, -0.093]

RQ1 reveals substantial performance gaps once evaluation moves from retained to temporally novel functional content. Specifically, the relative strength of tool-augmented systems changes across regimes. This reversal highlights a limitation of reference-matching evaluation for free-form prediction: systems that do not generate Swiss-Prot-style descriptions may be penalized for differences in output form even when they recover relevant functional content. Whole-description and reference-matching metrics therefore do not directly identify which biological facts are supported



Figure 4: Component-level breakdown of emerging-knowledge coverage in W2. Cells show complete-match rates for applicable A.1 claim components.

and may partly reflect alignment with the reference style rather than functional coverage itself. **We therefore use FactJudge to evaluate atomic reference claims against the complete free-form answer, placing A.0, A.1, and Track B on a common semantic scale without requiring structured outputs.** The paired A.0/A.1 setting provides an internal anchor, since the same frozen answer is evaluated against retained and later-added claims. It also reveals a notable difference from whole-description evaluation: on W2 A.0, PFUA reaches 0.618 FJ-R compared with 0.405 for Prot2Text-V2, reversing their ordering under conventional reference-matching metrics and aligning with their relative behavior on A.1 and Track B. This suggests that surface-form agreement can under- or overstate functional coverage when systems differ in generation style or objective. We report protein-level Miss as the fraction of proteins with no usable claim support ($s \geq 2$). This threshold achieves precision 0.909, recall 1.000, and F1 0.952 against expert consensus (Appendix B.4).

Annotation novelty is consistently more difficult than protein novelty at the claim level. All six systems have lower graded recall and higher Miss on A.1 than on Track B. Across systems, A.1 Miss ranges from 72.3% to 92.2%, compared with 41.1% to 73.3% on Track B. For BioMatrix, recall increases from 0.148 on A.1 to 0.251 on Track B, while Miss falls from 80.2% to 58.9%; PFUA shows the same contrast, with recall of 0.210 versus 0.362 and Miss of 72.4% versus 41.1%. The difference therefore extends beyond average coverage: usable support is absent for a substantially larger fraction of proteins under annotation novelty. Although the regimes involve different proteins and reference claims, the consistency of this contrast shows that newly added proteins do not substitute for evaluating newly added functional knowledge within existing records. Additional robustness analyses confirm that this conclusion persists when reweighting A.1 and Track B to account for annotation density and unequal sequence-family representation. Detailed results are provided in Appendix E.

A.1 failures further exhibit a structured component-level pattern rather than a uniform loss of information. Figure 4 shows that certainty is completely matched for 0.71–0.83 of applicable claims in four systems, whereas entity/object and relation/direction matches remain low across all six systems; specificity is only partially recovered, and conditions are rarely matched completely. Models therefore often preserve coarse modality or confidence while missing the biological participants, relations, specificity, and contextual qualifiers that make later-added claims informative. Claim-level evaluation thus reveals both the broader difficulty of annotation novelty relative to protein novelty and the functional dimensions in which that deficit concentrates. Explicit contradictions with reference claims are uncommon in the gold-support judgments. Together with the low A.1 coverage and component-level omissions, this indicates that missing or insufficiently specified functional information is the main observed failure in reference-claim coverage.

Take-Away for RQ2. Claim-level evaluation provides a less surface-sensitive view of temporal novelty: later-added A.1 knowledge is less broadly recovered than Track B newly-added-protein knowledge, and A.1 failures are concentrated in biological entities, relations, specificity, and conditions.

4.3 RQ3: How closely does retained behavior reflect historically available information?

RQ1 and RQ2 show a large gap between retained A.0 knowledge and later-added A.1 knowledge. This raises a simple question: how much of the strong retained-side performance is related to information

Table 4: Historical-information sufficiency across A.0 and A.1.

	W1			W2		
Method	A.0	A.1	Δ	A.0	A.1	Δ
Historical baselines						
Old annotation copy	**0.986**	**0.083**	−0.903	**0.991**	**0.109**	−0.882
Homology transfer	0.866	0.077	−0.789	0.881	0.105	−0.776
Model outputs						
BioMatrix	0.861	0.077	−0.784	0.837	0.101	−0.737
PFUA	0.240	0.062	−0.178	0.238	0.062	−0.177
EvoLLa	0.114	0.054	−0.061	0.116	0.051	−0.065



Figure 5: BioMatrix test distributions and whole-description outcomes relative to the training reference.

that was already available historically? We first examine how much of the retained-side agreement can be reproduced by historical copying and homology transfer, then assess how model outcomes vary with historical representation at the protein and claim levels. Table 4 compares model outputs with old-annotation copying and leave-accession-out homology transfer on the same A.0 and A.1 references using TF–IDF cosine similarity.

Historically available information can reproduce much of the retained-side lexical signal, but not emerging knowledge. In W2, simply copying old annotations yields an A.0 similarity of 0.991 (0.109 on A.1). Homology transfer provides a realistic baseline, recovering an A.0 similarity of 0.881 (0.105 on A.1). BioMatrix closely mirrors this historical baseline (0.837 on A.0, 0.101 on A.1), a pattern consistently observed in W1 as well. Our paired evaluation confirms that this alignment is not driven by coverage differences, with BioMatrix offering no lexical advantage over homology transfer (difference of -0.054 on A.0; Appendix Table 18). Overall, these reference-agreement comparisons demonstrate that retained lexical content is readily recoverable from historical sources, whereas emerging functional knowledge is not.

BioMatrix's whole-description failures also vary with the distribution of its training reference. To understand the mechanics behind BioMatrix's strong reliance on historical patterns, we examine its predictions relative to its public, task-aligned training reference. Figure 5 locates the W1 and W2 cohorts (630 and 462 proteins) relative to this reference in sequence and functional-description spaces. This visualization reveals two key patterns. On one hand, zero and partial deterministic Fact F1 outcomes in W2 are visibly concentrated outside densely represented training regions, confirming that model failures correlate with sparse reference coverage in sequence or description space. On the other hand, the distinct distributional footprints between the W1 and W2 cohorts highlight the temporal shifts in annotation data over time, directly underscoring the necessity of time-aware evaluations. The scatter plots thus support a distribution-level association with historical coverage, while the claim-level analysis below directly measures the corresponding recall differences.

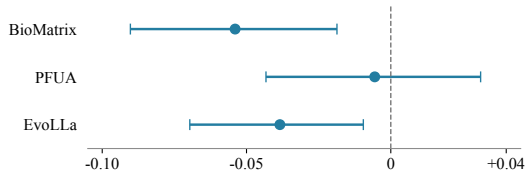


Figure 6: W2 A.0 graded-recall differences between historically unseen and represented claims. Points show unseen-minus-represented contrasts; bars denote 95% sequence-cluster bootstrap intervals.

At the claim level, historical representation is associated with retained recall for BioMatrix and EvoLLa, but shows little association for PFUA. To investigate this effect at a fine-grained level, we partition W2 retained (A.0) claims into two groups (those with a qualifying match in the public historical reference and those without) and compute the recall difference (unseen minus represented) across three models spanning different architectures and performance tiers: BioMatrix, EvoLLa, and PFUA (Figure 6). The results reveal a clear divergence in model behavior based on their generation paradigms. For the sequence-conditioned generators, BioMatrix and EvoLLa exhibit recall contrasts of -0.054 and -0.038, respectively. This substantial performance drop on unseen claims confirms that their high retained-side recall is heavily supported by repeating historically represented knowledge. In contrast, PFUA shows a near-zero recall difference (-0.006), demonstrating that its predictions are largely insensitive to representation in this static historical reference. This distinction corroborates that tool-augmentation systems operate under a fundamentally different mechanism, where retained coverage is decoupled from static reference recovery.

Take-Away for RQ3. Strong retained-side performance can largely reflect historically available information, but the strength of this relationship differs across system designs. Modal-projection/modal-unification systems are more closely tied to historical representation, whereas the tool-augmentation system is less sensitive to it.

5 Discussion and Conclusion

Protein-function capability does not scale uniformly across temporal regimes. Performance on established annotations, later-curated functional content, and newly added proteins reflects distinct aspects of model capability. In particular, the A.1 gap is broad rather than concentrated in a small number of difficult cases, and spans multiple dimensions of functional content. Thus, static annotation agreement provides an incomplete view of protein-function prediction, motivating temporal evaluation as a complementary axis for assessing Protein LLMs.

Strong retained recovery can often be supported by historically available information. High agreement with retained annotations should be interpreted relative to information already available before the target release. Historical-information diagnostics show that substantial retained-side recovery can be reproduced or approximated from prior annotations and related proteins. Evaluation should therefore separate retained from newly added functional content, since aggregate scores may be dominated by historically available information and obscure failures on later-curated knowledge. Consequently, strong static or retained performance does not by itself establish capability beyond historically represented knowledge.

The temporal gap suggests concrete directions for model improvement. The broad A.1 deficit suggests that improving static sequence-to-description generation alone may be insufficient. Future systems may benefit from dynamically updated retrieval over recent literature and curated databases, combined with evidence-grounded reasoning that preserves biological entities, relations, specificity, and conditions. ProLong-Bench provides a framework for testing whether such interventions specifically improve later-curated knowledge coverage.

Limitations and future work. First, ProLong-Bench currently focuses on the Swiss-Prot FUNCTION field, while protein function is represented across a broader set of curated annotations. Future extensions could incorporate additional functional fields to provide a more comprehensive view of free-form protein-function prediction. Second, the evaluated systems do not exhaust the rapidly growing space of Protein LLMs. Future work should include a broader set of models to provide a more comprehensive evaluation of existing Protein LLMs.

AI USE STATEMENT

AI assistants were used to revise the manuscript and write analysis and plotting code. DeepSeek Flash was used for reference decomposition and rubric-based scoring, as specified in the evaluation protocol. Post-hoc expert validation labels were supplied through collaborative expert consensus. AI-generated analyses are checked against the reported results; the limitations of the judge and its post-hoc validation are reported explicitly.

ETHICS STATEMENT

The study evaluates descriptions of publicly curated proteins. Benchmark scores measure agreement with a supplied reference and are not experimental confirmation of function, safety, or therapeutic utility. Misinterpreting reference-unverified text as false could penalize valid scientific hypotheses; conversely, fluent text can contain unsupported functional claims. We therefore separate coverage from relevance and retain uncertainty.

REFERENCES

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. doi: 10.1016/S0022-2836(05)80360-2. URL https://doi.org/10.1016/S0022-2836(05)80360-2.

Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

Elif Çevrim, Melih Gökay Yiğit, Erva Ulusoy, Ardan Yılmaz, and Tunca Doğan. A benchmarking platform for assessing protein language models on function-related prediction tasks. In *Protein Function Prediction: Methods and Protocols*, volume 2947 of *Methods in Molecular Biology*, pp. 241–268. Humana, 2025. doi: 10.1007/978-1-0716-4662-5_14. URL https://doi.org/10.1007/978-1-0716-4662-5_14.

M. Clara De Paolis Kaluza, Rashika Ramola, Parnal Joshi, Damiano Piovesan, Walter Reade, Sandra Orchard, Maria J. Martin, Alex Ignatchenko, Burkhard Rost, Christine A. Orengo, et al. Advances in protein function prediction from the fifth CAFA challenge. *bioRxiv*, 2026. doi: 10.64898/2026.04.27.716980. URL https://www.biorxiv.org/content/10.64898/2026.04.27.716980.

Adibvafa Fallahpour, Arman Seyed-Ahmadi, Parsa Idehpour, Omar Ibrahim, Purav Gupta, Jack Naimer, Kevin Zhu, Arnav Shah, Shihao Ma, Abhinav Adduri, et al. 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/10.64898/2026.03.19.712954v1. Version 1.

Chuanliu Fan, Zicheng Ma, Huanran Meng, Aijia Zhang, Wenjie Du, Jun Zhang, Ziqiang Cao, and Guohong Fu. Interleaved tool-call reasoning for protein function understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12977–12995, 2026. doi: 10.18653/v1/2026.acl-long.592. URL https://aclanthology.org/2026.acl-long.592/.

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

Xiao Fei, Michail Chatzianastasis, Sarah Almeida Carneiro, Hadi Abdine, Lawrence P. Petalidis, and Michalis Vazirgiannis. Prot2Text-V2: Protein function prediction with multimodal contrastive alignment. *arXiv preprint arXiv:2505.11194*, 2025. URL https://arxiv.org/abs/2505.11194.

Han Guo, Mingjia Huo, Xingyi Cheng, Digvijay Singh, Hamidreza Rahmani, Shen Li, Philipp Gerlof, Trey Ideker, Danielle A. Grotjahn, Elizabeth Villa, Le Song, and Pengtao Xie. Multi-modal large language model enables protein function prediction. In *ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences*, 2025.

Maxat Kulmanov and Robert Hoehndorf. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429, 2020. doi: 10.1093/bioinformatics/btz595. URL https://doi.org/10.1093/bioinformatics/btz595.

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 (Volume 1: Long Papers)*, pp. 5949–5966, 2024. doi: 10.18653/v1/2024.acl-long.324. URL https://aclanthology.org/2024.acl-long.324/.

Yizhen Luo, Jiahua Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. BioMedGPT: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023. URL https://arxiv.org/abs/2308.09442.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. BioT5+: Towards generalized biological understanding with IUPAC integration and multi-task tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. URL https://aclanthology.org/2024.findings-acl.71/.

Qizhi Pei, Zhimeng Zhou, Yi Duan, Yiyang Zhao, Wei Li, Han Guo, Liang He, Chengping Li, Chang-Yu Hsieh, Conghui He, Rui Yan, and Lijun Wu. BioMatrix: Towards a comprehensive biological foundation model spanning the modality matrix of sequences, structures, and language. *arXiv preprint arXiv:2606.22138*, 2026. URL https://arxiv.org/abs/2606.22138.

An Phan, Yanli Wang, Frimpong Boadu, Maxat Kulmanov, Robert Hoehndorf, Jianlin Cheng, Predrag Radivojac, and Iddo Friedberg. LAFA: A framework for reproducible longitudinal assessment of protein function annotation models. *arXiv preprint arXiv:2604.20782v1*, 2026. URL https://arxiv.org/abs/2604.20782v1.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL https://doi.org/10.1073/pnas.2016239118.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 2017. doi: 10.1038/nbt.3988. URL https://www.nature.com/articles/nbt.3988.

Chao Wang, Hehe Fan, Ruijie Quan, Lina Yao, and Yi Yang. ProtChatGPT: Towards understanding proteins with hybrid representation and large language models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1076–1086. Association for Computing Machinery, 2025a. doi: 10.1145/3726302.3730064.

Shaojun Wang, Ronghui You, Yunjia Liu, Yi Xiong, and Shanfeng Zhu. NetGO 3.0: A protein language model improves large-scale functional annotations. *Genomics, Proteomics & Bioinformatics*, 21(2):349–358, 2023.

Zhicong Wang, Zicheng Ma, Ziqiang Cao, Changlong Zhou, Jun Zhang, and Yiqin Gao. Prot2Chat: Protein large language model with early fusion of text, sequence, and structure. *Bioinformatics*, 41 (8):btaf396, 2025b. doi: 10.1093/bioinformatics/btaf396.

Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. ProteinGPT: Multimodal llm for protein property prediction and structure understanding. In *ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences*, 2025.

Huiying Yan, Shaojun Wang, Hancheng Liu, Hiroshi Mamitsuka, and Shanfeng Zhu. GORetriever: reranking protein-description-based GO candidates by literature-driven deep information retrieval for protein function annotation. *Bioinformatics*, 40(Supplement_2):ii53–ii61, 2024. doi: 10.1093/bioinformatics/btae401. URL https://doi.org/10.1093/bioinformatics/btae401.

Huiying Yan, Hancheng Liu, Shaojun Wang, and Shanfeng Zhu. GOAnnotator: accurate protein function annotation using automatically retrieved literature. *Bioinformatics*, 41(Supplement_1): i410–i419, 2025. URL https://academic.oup.com/bioinformatics/article/41/Supplement_1/i410/8199377.

Fei Ye, Zaixiang Zheng, Dongyu Xue, Yuning Shen, Lihao Wang, Yiming Ma, Yan Wang, Xinyou Wang, Xiangxin Zhou, and Quanquan Gu. ProteinBench: A holistic evaluation of protein foundation models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BksqWM8737.

Ronghui You, Xiaodi Huang, and Shanfeng Zhu. DeepText2GO: Improving large-scale protein function prediction with deep semantic text representation. *Methods*, 145:82–90, 2018. ISSN 1046-2023.

Naihui Zhou, Yuxiang Jiang, Timothy R. Bergquist, Alexandra J. Lee, Balint Z. Kacsoh, Alex W. Crocker, Kimberley A. Lewis, George Georghiou, Huy N. Nguyen, Md Nafiz Hamid, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20:244, 2019. doi: 10.1186/s13059-019-1835-8. URL https://doi.org/10.1186/s13059-019-1835-8.

Xibin Zhou, Chenchen Han, Yingqi Zhang, Huan Du, Jiayuan Tian, Jin Su, Renju Liu, Kai Zhuang, Shiyu Jiang, Anthony Gitter, et al. Decoding the molecular language of proteins with Evolla. *bioRxiv*, 2026. doi: 10.1101/2025.01.05.630192. URL https://www.biorxiv.org/content/10.1101/2025.01.05.630192v4. Version 4.

A Data construction and evidence

Each sample is a protein accession observed over one evaluation window, with sequence and FUNCTION annotations from both endpoints. Track membership depends on whether the accession is present at the baseline snapshot. Snapshot dates define annotation availability; model release dates alone do not establish training cutoffs. Qualifying additions require manually supported evidence, including experimental evidence, author statements, curator inference, and manually curated propagation. Following UniProt's documentation, we selected evidence codes corresponding to manually curated annotations: ECO:0000269, ECO:0000303, ECO:0000304, ECO:0000305, ECO:0000250, ECO:0000255, ECO:0000312, and ECO:0007744. Automatically asserted evidence does not qualify an annotation change for inclusion, but evaluation against the full target description retains the complete FUNCTION text.

A.1 Illustrative annotation update

For W1 accession F2R776, the BioMatrix answer is: "Monooxygenase that can modify rifampicin, thereby inactivating its antibiotic activity. Inactivates a broad range of rifamycin antibiotics." The target reference adds "Does not act on rifabutin." The independent reference decomposition yields three retained claims (monooxygenase activity, rifampicin inactivation, and broad rifamycin inactivation), each assigned support 4, and one added claim (the rifabutin exception), assigned support 0.

B Evaluation protocol

B.1 Conventional evaluation

Lexical metrics remove ECO/PubMed evidence citations, join split hyphenated words, lowercase text, and apply the same tokenization to predictions and references. Exact match requires complete equality of the resulting token sequences. BLEU-2 and BLEU-4 use sentence-level modified n-gram precision with add-one smoothing and the standard brevity penalty before macro averaging. ROUGE-1/2 use multiset n-gram-overlap F1, and ROUGE-L uses token-LCS F1. BERTScore uses `dmis-lab/biobert-base-cased-v1.2`, layer 12, without baseline rescaling. Deterministic facts are extracted using sentence, coordinated-predicate, and object-list rules. The fixed matcher combines embedding similarity with lexical/entity anchors, type compatibility, and relation/negation checks. Its support threshold is 0.78, and the separate contradiction threshold is 0.72. One-to-one assignment defines the matched count. Deterministic matching is used as a reference-mismatch diagnostic, separate from the FactJudge endpoint.

The full conventional tables are shown in Tables 5 and 6. Table 5 reports lexical machine metrics for the W1 overall cohort and the W2 whole-description cohort used by Table 2. Lexical metrics are not separately aggregated for Track A and Track B. BERTScore columns in Table 6 are protein-macro; Fact columns pool claim counts. Zero-hit denotes no matched deterministic reference fact. Thus zero-hit in this table differs from zero *usable rubric support* in Table 3.

Table 5: Whole-description lexical metrics across W1 and W2.

Win.	Model	N	Exact	BLEU-2	BLEU-4	R-1	R-2	R-L
W1	Prot2Text-V2	630	0.0000	0.2184	0.1978	0.3497	0.2688	0.3308
W1	EvoLLa	630	0.0000	0.0411	0.0142	0.1215	0.0259	0.0825
W1	BioT5+	630	0.0000	0.0577	0.0310	0.1470	0.0239	0.1055
W1	BioMatrix	630	0.0000	**0.3934**	**0.3718**	**0.5267**	**0.4623**	**0.5124**
W1	PFUA	630	0.0000	0.0941	0.0491	0.2047	0.0649	0.1350
W1	BioReason-Pro	630	0.0000	0.0567	0.0275	0.1723	0.0215	0.1127
W2	Prot2Text-V2	462	0.0043	0.1994	0.1818	0.3230	0.2441	0.3000
W2	EvoLLa	462	0.0000	0.0498	0.0165	0.1454	0.0302	0.0927
W2	BioT5+	462	0.0000	0.0516	0.0270	0.1491	0.0260	0.1063
W2	BioMatrix	462	**0.0108**	**0.4710**	**0.4513**	**0.5956**	**0.5374**	**0.5763**
W2	PFUA	462	0.0000	0.0914	0.0455	0.2126	0.0635	0.1327
W2	BioReason-Pro	457	0.0000	0.0513	0.0243	0.1714	0.0207	0.1110

Table 6: Conventional whole-description metrics at native coverage.

Model	Win.	N	BERT P/R/F1	Fact P/R/F1	Claims/prot.	Zero-hit
BioReason-Pro	W1	630	0.704 / 0.721 / 0.711	0.089 / 0.057 / 0.069	4.135	0.737
Prot2Text-V2	W1	630	0.829 / 0.756 / 0.788	0.555 / 0.221 / 0.316	2.590	0.359
PFUA	W1	630	0.677 / 0.741 / 0.704	0.144 / 0.179 / 0.159	8.122	0.583
EvoLLa	W1	630	0.636 / 0.680 / 0.656	0.065 / 0.161 / 0.092	16.206	0.537
BioMatrix	W1	630	**0.877 / 0.829 / 0.850**	**0.787 / 0.617 / 0.692**	5.098	**0.270**
BioT5+	W1	630	0.693 / 0.669 / 0.679	0.026 / 0.005 / 0.009	1.359	0.965
BioReason-Pro	W2	457	0.710 / 0.709 / 0.707	0.101 / 0.044 / 0.061	3.928	0.698
Prot2Text-V2	W2	462	0.817 / 0.739 / 0.773	0.479 / 0.168 / 0.249	3.141	0.379
PFUA	W2	462	0.685 / 0.740 / 0.708	0.150 / 0.135 / 0.142	8.000	0.528
EvoLLa	W2	462	0.646 / 0.673 / 0.658	0.081 / 0.153 / 0.106	16.818	0.481
BioMatrix	W2	462	**0.894 / 0.851 / 0.870**	**0.844 / 0.685 / 0.756**	7.251	**0.199**
BioT5+	W2	462	0.689 / 0.653 / 0.668	0.027 / 0.004 / 0.007	1.260	0.965

B.2 Reference decomposition and scoring

The decomposer preserves entities, relation direction, negation, uncertainty, conditions, and biological context. It splits independent predicates while keeping a complete biochemical reaction with its substrates and products together. Necessary context is repeated to make a proposition interpretable; no new biological information may be inferred. The protein is an implicit subject, and generic introductory prefixes are removed. Retained and added reference views are processed independently.

Table 7 summarizes reference-claim support. Component scores constrain overall support: a mismatched entity or relation caps it at one, and an explicit contradiction forces zero. Component scores remain diagnostic rather than independently validated endpoints. Prediction relevance instead uses a three-level scale: unrelated/incompatible, related extension, or redundant paraphrase.

Table 7: FactJudge gold-support rubric.

Score	Meaning
4	All essential entities, objects, relations, directions, and conditions are stated or faithfully paraphrased.
3	The core fact is supported, with a noncritical qualifier omitted or slightly weaker certainty.
2	Partial support, with an important detail, object, mechanism, or condition missing.
1	Only the broad topic or function class is related.
0	No relevant support, or an explicit contradiction.

B.3 FactJudge prompts

Component	Scored output
Gold-support judge	Overall support on a 0–4 scale, plus diagnostic component scores for entity/object, relation/direction, specificity, conditions, and certainty.
Prediction-relevance judge	Prediction-side relevance on a 0–2 scale, reported separately from gold-claim recall.
Gold-claim decomposition	Atomic reference claims preserving entities, relations, direction, negation, uncertainty, conditions, and biological context.

Table 8: FactJudge scoring components.

Gold-support judge system message

```
Score atomic reference claims against one protein prediction.

Use only the prediction as evidence. Do not use outside knowledge.
The omitted subject of each atomic reference claim is the same protein described by the
↪  prediction.

Overall support score:
- 4: Complete support. All essential entities, objects, relations, directions, and
↪  conditions are stated exactly or by faithful paraphrase.
```

Gold-support judge system message (continued)

```
- 3: The core fact is supported, but one non-critical qualifier is omitted or certainty
↪  is slightly weaker.
- 2: Partial support. The central function is related and partly correct, but an
↪  important detail, object, mechanism, or condition is missing.
- 1: Only the broad topic or function class is related; the specific claim is not
↪  supported.
- 0: No relevant support, or the prediction explicitly contradicts the claim.

Contradiction score:
- 0: No contradiction.
- 1: Potential incompatibility, but no explicit opposite assertion.
- 2: Explicit negation, opposite direction/effect, or mutually exclusive assertion about
↪  the same entities and relation.

Dimension scores use 0 for wrong/missing, 1 for partial/broader, and 2 for complete
↪  match.
Use null only when a dimension is not applicable. Specificity is null when the relevant
↪  entity or relation is absent or mismatched. Conditions are contextual qualifiers
↪  such as cell type, biological state, infection context, stimulus, or developmental
↪  stage; reaction participants, products, and mechanisms are not conditions.

Rules:
1. Score every supplied claim exactly once and preserve each claim_id exactly.
2. A different function or object is not a contradiction unless the prediction
↪  explicitly rules out the reference claim.
3. An uncertain prediction (for example, "may") does not fully support a definite claim.
↪  A definite prediction may support an otherwise identical uncertain claim.
4. Extra statements in the prediction do not affect support for a reference claim.
5. If contradiction_score is 2, support_score must be 0.
6. If entity_and_object or relation_and_direction is 0, support_score cannot exceed 1.
7. A support_score of 4 requires entity_and_object=2, relation_and_direction=2,
↪  contradiction_score=0, and every other applicable dimension=2.
8. Return integers and null only. Do not return labels, evidence, rationales, or other
↪  fields.
9. Return one JSON object and no other text.

Return this schema:
{
  "judgments": [
    {
      "claim_id": "copied exactly from the input",
      "support_score": 0,
      "contradiction_score": 0,
      "dimension_scores": {
        "entity_and_object": 0,
        "relation_and_direction": 0,
        "specificity": null,
        "conditions": null,
        "certainty": null
      }
    }
  ]
}
```

Gold-support judge user message template

```
REFERENCE CLAIMS:
<JSON array of {"claim_id":"...","claim":"..."} objects>

PREDICTION:
<complete prediction text>
```

Prediction-relevance judge system message

```
Score how each functional claim in one protein prediction relates to the supplied
↪  reference. Score relevance, not whether the claim is proven true.

Use the reference and general biological knowledge only to decide the relation.
Split the prediction into atomic functional claims. Ignore length, structure-only
↪  statements, discourse, and bare identity that conveys no function. Keep repeated
↪  functional claims as separate entries.

Scores:
- 2: The same functional fact or a faithful paraphrase; it adds no new information.
- 1: Related but not the same: it stays within the same specific function and adds a
↪  possible mechanism, condition, substrate, or consequence.
```

Prediction-relevance judge system message (continued)

```
- 0: No meaningful relation, or a different/incompatible core function, substrate, or
↪  process. Broad overlap such as transport, enzyme, homeostasis, stress, signaling, or
↪  membrane is not enough.

Return one JSON object and no other text:
{"claims":[{"claim":"...","relevance_score":0}]}
```

Prediction-relevance judge user message template

```
PREDICTION:
<complete prediction text>

REFERENCE:
<complete reference FUNCTION text>
```

`relevance_score=0` is a reference-relative mismatch signal, not independent evidence that the prediction claim is biologically false.

Gold-claim decomposition system message

```
You split protein descriptions into atomic biological claims.

Rules:
1. Each claim must contain one independently verifiable biological proposition.
2. Preserve the original meaning, including entities, relations, direction, negation,
↪  uncertainty, conditions, and biological context.
3. Do not add, infer, correct, generalize, or remove information.
4. Split independent predicates joined by conjunctions.
5. Keep a complete biochemical reaction, including its substrates and products, as one
↪  claim.
6. The protein is the implicit subject. Start each claim directly with the biological
↪  predicate. Never add generic prefixes such as "The described protein", "This
↪  protein", or "The protein".
7. Do not output vague or elliptical claims. If a claim depends on earlier text, repeat
↪  only the necessary entity, reaction, or object from the input.
8. Ignore headings, formatting, citations, and non-biological filler.
9. Return one JSON object with a "claims" array. Use an empty array if there is no
↪  biological claim.
```

Gold-claim decomposition user message template

```
Split the following protein description into atomic claims.

Description:
<reference FUNCTION description>

Return this JSON object:
{"claims": [{"claim": "..."}, {"claim": "..."}]}
```

B.4 POST-HOC EXPERT VALIDATION

We validate FactJudge post hoc against collaborative expert consensus; expert labels were not used to select or tune the evaluator, rubric, thresholds, or primary endpoint. The 50 reviewed descriptions comprise 25 A.0 and 25 A.1 cases: 17 from W1, 17 from an update interval ending 2026-06-30, and 16 from the interval ending 2026-09-03. Sampling spans score strata and difficult cases. Support and relevance agreement are evaluated on 50 and 99 items, respectively (Table 9).

Table 9: Agreement between FactJudge and expert consensus.

Endpoint	N	Exact agreement	Quadratic weighted κ
Prediction relevance	99	0.747	0.734
Gold support	50	0.780	0.944

For gold support, thresholding at $s \geq 2$ gives precision 0.909, recall 1.000, and F1 0.952; at $s \geq 3$, precision is 0.929, recall 1.000, and F1 0.963. On this subset, judge scores average 0.20 rubric points above consensus (0.05 graded-recall units). Reference decomposition is complete for 49 of 50 descriptions; 43 of 50 satisfy both completeness and atomicity, yielding an 86% joint pass rate.

B.5 Prediction-side relevance on a high-risk subset

Gold-side recall measures whether an answer covers the reference claims, but does not characterize additional content in the answer. We therefore apply the prediction-relevance judge to an intentionally enriched, nonrandom subset of outputs with apparent reference mismatch. This pilot contains 139 model–interval–protein output cases from BioMatrix, EvoLLa, and PFUA, yielding 2,097 atomic prediction claims. Cases are output-level observations and need not correspond to distinct proteins.

Selection uses four surface-risk features: prediction-to-reference word-count ratio; excess prediction-sentence count over reference-claim count, floored at zero; the fraction of prediction sentences whose maximum TF–IDF cosine similarity to a reference claim is below 0.15; and the mean of one minus this maximum similarity. TF–IDF uses word unigrams and bigrams with sublinear term frequency. Each feature is converted to a percentile rank within its model and source update interval, and all eligible outputs with a mean percentile of at least 0.80 are selected, without an additional sampling cap.

The judge decomposes each selected full answer into atomic functional claims and assigns the three relevance levels defined in Appendix B.3: 0 denotes unrelated or incompatible functional content, 1 denotes a related extension, and 2 denotes the same functional fact or a faithful paraphrase. Table 10 reports the full ordinal distribution as claim-micro proportions within each model's selected outputs; no binary relevance endpoint is introduced, and these proportions are not combined with gold-support recall.

Table 10: Prediction-relevance distributions in the high-risk pilot. Percentages pool prediction claims within each model's selected output cases.

Model	Cases	Claims	Score 0 (%)	Score 1 (%)	Score 2 (%)
BioMatrix	14	110	31.8	36.4	31.8
EvoLLa	83	1,343	86.7	12.2	1.0
PFUA	42	644	16.1	51.2	32.6

The selected outputs exhibit different relevance profiles. For PFUA, 51.2% of claims are scored as related extensions and 32.6% as restatements, whereas 86.7% of EvoLLa's claims receive score 0. BioMatrix is more evenly distributed, with 31.8%, 36.4%, and 31.8% at scores 0, 1, and 2, respectively. Apparent surface mismatch therefore includes both reference-related elaboration and unrelated functional content; it cannot be equated with semantic irrelevance. Because this subset is deliberately enriched for apparent prediction–reference mismatches, the proportions are stress-test diagnostics rather than model-wide relevance rates. Score 0 is a reference-relative functional mismatch signal, not independently verified biological falsity, and a related extension is not necessarily a validated additional fact.

C Training-data cutoffs across models

The evaluated systems differ in the temporal boundaries of their training data and inference-time resources. Table 11 summarizes the reported temporal information for each system. We distinguish documented annotation snapshots from model-publication dates, which are not treated as corpus cutoffs. All reported training data or inference-time resources predate the benchmark constructed in this work.

Table 11: Reported training-data cutoffs and temporal information for the evaluated systems. Publication dates are not treated as corpus cutoffs.

System	Reported temporal information
Prot2Text-V2	Swiss-Prot release 2025_02 (Fei et al., 2025).
EvoLLa	Swiss-Prot release 2023_03 for training; 2024_05 for the reported newly-added-protein test set (Zhou et al., 2026).
BioT5+	Original study publicly available in 2024-02; instruction data include Mol-Instructions (Pei et al., 2024).
BioMatrix	Swiss-Prot and TrEMBL release 2026_01 (Pei et al., 2026).
PFUA	Reported Swiss-Prot retrieval database released in 2025_03 (Fan et al., 2026).
BioReason-Pro	Swiss-Prot release 2022_10 (Fallahpour et al., 2026).

D MULTI-GRANULAR FACTJUDGE PROFILES

FactJudge is a reporting hierarchy rather than a single-score evaluator. Overall support on the 0–4 scale remains the primary scalar endpoint, and the judge output also permits ordinal support distributions and component profiles. Tables 12 and 13 report these diagnostics by model, window, and retained/added view. Null component fields indicate that a dimension is not applicable to the claim and are excluded from component denominators rather than counted as missing.

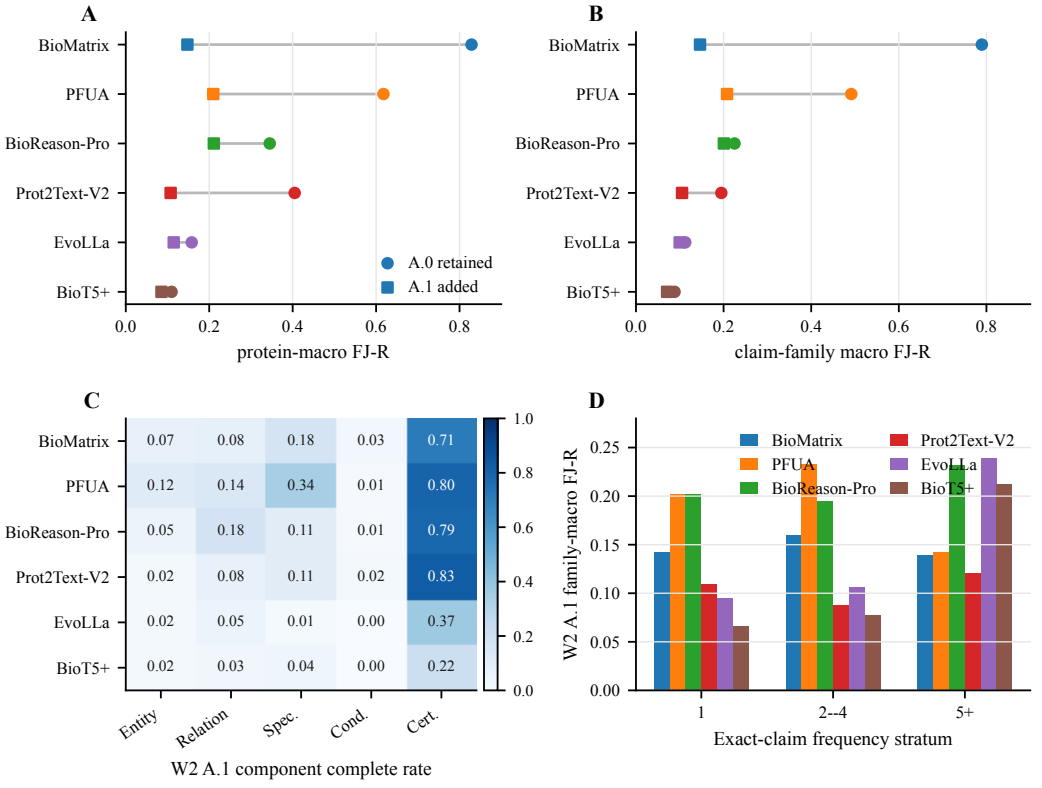


Figure 7: Multi-granular analysis of retained and emerging functional knowledge in W2.

Table 12: Ordinal claim-support distributions across models, windows, and A.0/A.1 views.

Model	Window	View	Claims	s_0	s_1	s_2	s_3	s_4	Family macro
Prot2Text-V2	W1	A.0	3643	0.589	0.134	0.053	0.011	0.214	0.288
EvoLLa	W1	A.0	3643	0.617	0.271	0.071	0.015	0.027	0.141
BioT5+	W1	A.0	3643	0.706	0.203	0.062	0.022	0.007	0.108
BioMatrix	W1	A.0	3643	0.105	0.057	0.017	0.012	0.809	**0.826**
PFUA	W1	A.0	3643	0.226	0.196	0.117	0.087	0.374	0.554
BioReason-Pro	W1	A.0	3643	0.278	0.372	0.208	0.057	0.085	0.328
Prot2Text-V2	W1	A.1	690	0.709	0.201	0.075	0.007	0.007	0.101
EvoLLa	W1	A.1	690	0.655	0.284	0.046	0.003	0.012	0.110
BioT5+	W1	A.1	690	0.772	0.165	0.045	0.014	0.003	0.081
BioMatrix	W1	A.1	690	0.667	0.236	0.077	0.009	0.012	0.118
PFUA	W1	A.1	690	0.555	0.283	0.100	0.016	0.046	0.176
BioReason-Pro	W1	A.1	690	0.484	0.354	0.119	0.010	0.033	**0.188**
Prot2Text-V2	W2	A.0	4434	0.683	0.151	0.040	0.009	0.117	0.195
EvoLLa	W2	A.0	4434	0.641	0.288	0.044	0.009	0.018	0.113
BioT5+	W2	A.0	4434	0.706	0.239	0.043	0.009	0.003	0.088
BioMatrix	W2	A.0	4434	0.140	0.050	0.021	0.010	0.779	**0.790**
PFUA	W2	A.0	4434	0.283	0.183	0.130	0.092	0.312	0.492
BioReason-Pro	W2	A.0	4431	0.436	0.359	0.123	0.030	0.053	0.225
Prot2Text-V2	W2	A.1	573	0.667	0.279	0.038	0.007	0.009	0.105
EvoLLa	W2	A.1	573	0.606	0.353	0.040	0.000	0.002	0.100
BioT5+	W2	A.1	573	0.729	0.218	0.049	0.003	0.000	0.071
BioMatrix	W2	A.1	573	0.578	0.312	0.075	0.012	0.023	0.146
PFUA	W2	A.1	573	0.515	0.305	0.082	0.019	0.079	**0.208**
BioReason-Pro	W2	A.1	571	0.431	0.403	0.126	0.011	0.030	0.201

Table 13: Claim-component profiles across models, windows, and A.0/A.1 views.

Model	Window	View	Ent.	Rel.	Spec.	Cond.	Cert.
Prot2Text-V2	W1	A.0	0.236/0.113/0.651	0.265/0.086/0.649	0.679/0.249/0.072	0.345/0.128/0.527	0.885/0.087/0.027
EvoLLa	W1	A.0	0.056/0.152/0.792	0.081/0.187/0.732	0.177/0.495/0.328	0.053/0.139/0.807	0.528/0.434/0.038
BioT5+	W1	A.0	0.045/0.156/0.799	0.068/0.129/0.803	0.117/0.543/0.339	0.025/0.164/0.811	0.204/0.772/0.024
BioMatrix	W1	A.0	0.825/0.044/0.131	0.834/0.027/0.139	0.956/0.032/0.012	0.894/0.022/0.083	0.972/0.023/0.005
PFUA	W1	A.0	0.516/0.207/0.277	0.493/0.164/0.343	0.641/0.285/0.075	0.480/0.246/0.274	0.920/0.069/0.011
BioReason-Pro	W1	A.0	0.191/0.337/0.472	0.273/0.310/0.417	0.203/0.561/0.236	0.153/0.255/0.592	0.836/0.133/0.031
Prot2Text-V2	W1	A.1	0.029/0.135/0.836	0.075/0.133/0.791	0.100/0.689/0.211	0.110/0.034/0.856	0.742/0.168/0.090
EvoLLa	W1	A.1	0.023/0.132/0.845	0.068/0.168/0.764	0.132/0.407/0.462	0.034/0.121/0.845	0.328/0.560/0.112
BioT5+	W1	A.1	0.022/0.086/0.893	0.048/0.106/0.846	0.092/0.434/0.474	0.000/0.057/0.943	0.261/0.697/0.042
BioMatrix	W1	A.1	0.042/0.145/0.813	0.086/0.129/0.786	0.133/0.478/0.389	0.009/0.052/0.939	0.721/0.164/0.115
PFUA	W1	A.1	0.088/0.235/0.677	0.125/0.167/0.709	0.271/0.571/0.157	0.124/0.176/0.699	0.843/0.092/0.065
BioReason-Pro	W1	A.1	0.070/0.196/0.735	0.149/0.213/0.638	0.175/0.565/0.260	0.061/0.061/0.878	0.854/0.111/0.035
Prot2Text-V2	W2	A.0	0.141/0.107/0.752	0.193/0.092/0.715	0.489/0.274/0.237	0.310/0.143/0.548	0.892/0.094/0.014
EvoLLa	W2	A.0	0.042/0.127/0.831	0.079/0.200/0.721	0.152/0.521/0.327	0.028/0.061/0.910	0.536/0.375/0.089
BioT5+	W2	A.0	0.029/0.098/0.873	0.040/0.173/0.786	0.077/0.552/0.371	0.008/0.017/0.975	0.252/0.696/0.052
BioMatrix	W2	A.0	0.796/0.037/0.167	0.797/0.035/0.168	0.959/0.033/0.009	0.910/0.023/0.067	0.983/0.015/0.002
PFUA	W2	A.0	0.492/0.164/0.345	0.422/0.198/0.380	0.593/0.338/0.069	0.370/0.200/0.430	0.915/0.080/0.004
BioReason-Pro	W2	A.0	0.111/0.223/0.666	0.200/0.247/0.554	0.194/0.556/0.250	0.080/0.121/0.799	0.851/0.131/0.017
Prot2Text-V2	W2	A.1	0.019/0.141/0.839	0.075/0.178/0.747	0.113/0.400/0.487	0.016/0.048/0.937	0.826/0.129/0.045
EvoLLa	W2	A.1	0.021/0.134/0.845	0.052/0.250/0.698	0.014/0.486/0.500	0.000/0.018/0.982	0.374/0.511/0.115
BioT5+	W2	A.1	0.016/0.106/0.878	0.033/0.175/0.792	0.039/0.513/0.447	0.000/0.000/1.000	0.216/0.741/0.043
BioMatrix	W2	A.1	0.072/0.152/0.777	0.084/0.216/0.700	0.179/0.537/0.284	0.029/0.129/0.843	0.708/0.179/0.113
PFUA	W2	A.1	0.119/0.190/0.691	0.143/0.223/0.634	0.345/0.448/0.207	0.011/0.125/0.864	0.799/0.113/0.088
BioReason-Pro	W2	A.1	0.051/0.270/0.680	0.180/0.270/0.550	0.113/0.506/0.381	0.009/0.064/0.927	0.792/0.154/0.054

E AGGREGATION ROBUSTNESS

Protein-macro recall is the primary estimand because ProLong-Bench evaluates functional coverage at the protein level. We report two sensitivity estimands for W2 A.0, A.1, and Track B. Claim-micro recall gives each reference claim occurrence equal weight and therefore tests sensitivity to annotation density. Sequence-cluster-macro recall first averages within each unified sequence cluster and then averages clusters, testing whether large homologous families dominate the result. Neither sensitivity estimand is treated as more correct than protein-macro recall.

Table 14 lists the grouping counts used for aggregation. W2 exact groups are exact judge-claim text groups in the scoring input. For the paired Track A cohort, the unified sequence-cluster count is 135 for all systems except BioReason-Pro, which has 134 because one prediction is unavailable.

Table 14: Grouping statistics for claim-level aggregation.

Window	View	Occurrences	Exact groups	Seq. clusters	Cells
W1	A.0	3,643	2,707	210	2,788
W1	A.1	690	535	210	576
W2	A.0	4,434	3,044	135	3,125
W2	A.1	573	394	135	417

Table 15: Robustness of W2 A.0/A.1 results to aggregation choice.

Model	Estimand	A.0 FJ-R [95% CI]	A.1 FJ-R [95% CI]	Gap [95% CI]
BioMatrix	Claim-micro	0.810 [0.727, 0.870]	0.147 [0.116, 0.180]	-0.662 [-0.731, -0.574]
BioMatrix	Protein-macro	0.829 [0.776, 0.876]	0.148 [0.118, 0.179]	-0.681 [-0.733, -0.625]
BioMatrix	Sequence-cluster-macro	0.820 [0.766, 0.870]	0.149 [0.122, 0.175]	-0.671 [-0.722, -0.618]
PFUA	Claim-micro	0.492 [0.409, 0.580]	0.210 [0.161, 0.271]	-0.281 [-0.374, -0.196]
PFUA	Protein-macro	0.618 [0.561, 0.671]	0.210 [0.169, 0.253]	-0.408 [-0.470, -0.348]
PFUA	Sequence-cluster-macro	0.630 [0.579, 0.680]	0.215 [0.176, 0.255]	-0.415 [-0.469, -0.359]
EvoLLa	Claim-micro	0.119 [0.094, 0.150]	0.110 [0.086, 0.134]	-0.009 [-0.041, 0.016]
EvoLLa	Protein-macro	0.158 [0.130, 0.187]	0.115 [0.092, 0.138]	-0.043 [-0.072, -0.018]
EvoLLa	Sequence-cluster-macro	0.155 [0.129, 0.183]	0.109 [0.086, 0.134]	-0.046 [-0.072, -0.019]
BioReason-Pro	Claim-micro	0.226 [0.190, 0.274]	0.201 [0.173, 0.233]	-0.025 [-0.073, 0.011]
BioReason-Pro	Protein-macro	0.345 [0.307, 0.386]	0.211 [0.182, 0.242]	-0.134 [-0.179, -0.093]
BioReason-Pro	Sequence-cluster-macro	0.372 [0.335, 0.411]	0.223 [0.192, 0.255]	-0.149 [-0.193, -0.106]
Prot2Text-V2	Claim-micro	0.182 [0.136, 0.254]	0.103 [0.079, 0.132]	-0.079 [-0.145, -0.038]
Prot2Text-V2	Protein-macro	0.405 [0.335, 0.481]	0.108 [0.083, 0.135]	-0.297 [-0.367, -0.234]
Prot2Text-V2	Sequence-cluster-macro	0.464 [0.398, 0.533]	0.112 [0.088, 0.137]	-0.352 [-0.414, -0.290]
BioT5+	Claim-micro	0.091 [0.072, 0.111]	0.082 [0.059, 0.105]	-0.010 [-0.030, 0.011]
BioT5+	Protein-macro	0.110 [0.089, 0.131]	0.086 [0.064, 0.109]	-0.024 [-0.044, -0.005]
BioT5+	Sequence-cluster-macro	0.108 [0.088, 0.130]	0.082 [0.061, 0.104]	-0.027 [-0.046, -0.009]

The main findings are robust to aggregation choice. Added-claim recall is stable across protein-macro, claim-micro, and sequence-cluster-macro weighting: for example, BioMatrix remains near 0.15, PFUA and BioReason-Pro remain near 0.21, and the modal-projection systems remain near the low end of the scale. Sequence-cluster-macro estimates are close to protein-macro estimates for most systems, so low A.1 coverage is not explained by a few large homologous families. Retained recall is more sensitive to claim-micro weighting for some systems, especially PFUA and Prot2Text-V2, showing that close retained-side ordering can depend on the estimand. Protein-macro remains primary because the benchmark target is protein-level functional coverage rather than annotation volume.

Table 16: Robustness of W2 Track B results to aggregation choice. FJ-R is graded FactJudge recall; Miss denotes absence of usable claim support, reported with protein and sequence-cluster weighting.

	FJ-R ↑			Miss ↓	
Model	Claim-micro	Protein-macro	Cluster-macro	Protein	Cluster
BioMatrix	0.262	0.251	0.248	<u>0.589</u>	0.593
PFUA	**0.369**	**0.362**	**0.375**	**0.411**	**0.389**
EvoLLa	0.164	0.172	0.180	0.722	0.704
BioReason-Pro	<u>0.358</u>	<u>0.352</u>	<u>0.360</u>	**0.411**	<u>0.395</u>
Prot2Text-V2	0.181	0.170	0.171	0.700	0.698
BioT5+	0.140	0.149	0.147	0.733	0.728

Track B extends this robustness check to the comparison between temporal novelty regimes (Table 16). All six models have higher Track B recall than A.1 recall under each of the three aggregation schemes. The point-estimate differences range from 0.054 to 0.159 under claim-micro weighting, from 0.057 to 0.152 under protein-macro weighting, and from 0.059 to 0.160 under sequence-cluster-macro

weighting. For example, PFUA's cluster-macro recall is 0.375 on Track B versus 0.215 on A.1, while BioMatrix's is 0.248 versus 0.149. Giving each sequence cluster equal weight prevents large homologous families from dominating the aggregate, yet Track B recall remains higher for every model. Together with the claim-micro comparison, this shows that the A.1–Track B recall gap is not solely driven by annotation density or overrepresentation of large homologous families. Within Track B, switching from equal protein weights to equal sequence-cluster weights changes Miss by at most 0.022 across models; the failure-breadth estimates therefore also remain similar after reweighting.

F HISTORICAL-REFERENCE REPRESENTATION ANALYSIS

We label claims against a common Swiss-Prot 2025-02-05 historical reference. A best word unigram/bigram TF–IDF cosine similarity of at least 0.70 identifies a historically represented function; claims without a qualifying match are labeled unseen. Represented claims include matches within the target's historical sequence cluster (pair seen) and matches elsewhere (function seen, new pair). These labels indicate whether similar functional information is represented in the historical public reference.

Retrieval uses sublinear term frequency, English stop words, and minimum document frequency two. All systems use the same historical reference and labeling procedure. For Figure 6, claim scores are first averaged within each protein and representation category, then macro-averaged over proteins in that category. The reported contrast is unseen minus represented graded recall, with 95% confidence intervals from sequence-cluster bootstrap. The W2 A.0 cohort contains 243 proteins in 135 sequence clusters; 211 proteins contribute unseen claims and 241 contribute represented claims. The category means therefore need not involve identical protein populations.

Table 17: W2 A.0 graded-recall contrasts between historically unseen and represented claims. Intervals are 95% sequence-cluster bootstrap confidence intervals.

Model	Difference [95% CI]	Unseen proteins	Represented proteins
BioMatrix	-0.054 $[-0.090, -0.019]$	211	241
PFUA	-0.006 $[-0.043, +0.031]$	211	241
EvoLLa	-0.038 $[-0.070, -0.010]$	211	241

Training-reference distribution diagnostic. The main-text distribution figure compares the two test cohorts with BioMatrix's public task-aligned training reference. W1 uses 630 whole-description proteins, and W2 uses the 462-protein whole-description cohort used by Table 2. Training-anchored sequence clusters use MMseqs2 identity at least 30% and bidirectional coverage at least 80%; description clusters use connected components of training-fitted TF–IDF cosine at least 0.70. Test-only clusters do not merge existing training clusters. Layout uses sequence 3-mer or description word TF–IDF, SVD to 50 dimensions, normalization, and t-SNE with perplexity 40. Joint layout fitting serves visualization only; novelty labels are fixed from training-reference matches. Sequence novelty is one minus the identity of the highest-bit-score qualifying training hit, with exact sequence matches overriding retrieval and no qualifying hit assigned one. Thus it is a censored alignment score, not an unrestricted nearest-neighbor distance. Description novelty is one minus maximum cosine over unique training descriptions. The public SFT reference need not cover every source used to train the model.

G HISTORICAL TRANSFER AND LEXICAL SENSITIVITY

Old-description copying supplies the baseline FUNCTION text. Historical transfer uses the shared 2025-02-05 reference and excludes the same accession. MMseqs2 (Steinegger & Söding, 2017) candidates require at least 30% identity and 80% query and target coverage; candidates are ranked by decreasing bit score, decreasing identity, then increasing accession. It can still retrieve an identical sequence belonging to another accession. In the historical transfer diagnostic, transfer is available for 267/291 W1 strict proteins and 231/243 W2 proteins. Table 4 averages over available transfers. Table 18 reports paired differences for 267 W1 and 217 W2 proteins; its W2 estimates describe this smaller analysis subset, not all 231 transfer-covered proteins.

Table 18: BioMatrix versus historical transfer on the shared covered subset.

Window	View	N	SEQ30	BioMatrix − transfer [95% CI]
W1	A.0	267	186	-0.002 [-0.049, 0.044]
W1	A.1	267	186	-0.002 [-0.007, 0.002]
W2	A.0	217	128	-0.054 [-0.103, -0.006]
W2	A.1	217	128	-0.004 [-0.011, 0.003]

The distinction between recall and precision-sensitive lexical metrics is substantive. On W1 A.1, EvoLLa minus transfer has ROUGE-L recall difference 0.095 [0.067, 0.125], but ROUGE-L F1 difference -0.052 $[-0.061, -0.044]$. The same outputs therefore support opposite lexical comparisons. This motivates graded reference-claim coverage, rather than interpreting one lexical ordering as a factual ranking.