LOST IN SEQUENCE, FOUND IN TEXT: RETRIEVING FUNCTIONALLY RELATED PROTEINS UNDER JOINT SEQUENCE AND STRUCTURE SIMILARITY CONTROLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Convergent evolution can give rise to proteins with similar functions despite dissimilar sequences and structures, motivating functional retrieval beyond conventional similarity search. Yet the ability of existing methods to recover such functional relationships remains insufficiently evaluated under joint controls on sequence and structural similarity. We introduce FURETON, a FUnctional RETrieval benchmark for this setting built from 89,116 reviewed UniProtKB/Swiss-Prot proteins with AlphaFold-predicted structures and selected Gene Ontology (GO) annotations. Designed to assess functional retrieval beyond conventional similarity, the benchmark controls both sequence and structural similarity between query–target pairs and defines relevance through informative shared GO ancestors. We evaluate five families of methods in zero-shot settings: conventional sequence-, structure-, and signature-based search, protein sequence language models, protein structure-aware language models, protein–text contrastive models, and protein-to-text generative models. We additionally adapt protein language models by training projection heads with a contrastive objective while keeping the pretrained backbones frozen. This contrastive adaptation improves retrieval across Molecular Function, Biological Process, Cellular Component, and the overall task, suggesting that it can make pretrained protein representations more effective for functional retrieval. Nevertheless, the best adapted model achieves 4.65% MAP in overall task, compared with 7.65% for retrieval using protein-to-text generative models. These results highlight functional text as a promising representation for retrieving functionally related proteins under joint sequence and structural similarity controls. Code, data, and evaluation protocols are available at `https://anonymous.4open.science/r/FURETON-EA8D`.

1 INTRODUCTION

Proteins that share a common ancestor often retain similar sequences, structures, and functions, supporting functional annotation through sequence and structure searches with tools such as BLAST (Altschul et al., 1990), MMseqs2 (Steinegger & Söding, 2017), and Foldseek (Van Kempen et al., 2024). However, proteins can also evolve similar functions independently rather than inherit them from a common ancestor, a process known as convergent evolution (Stern, 2013). For example, enzymes can independently evolve to catalyze the same reaction despite having dissimilar sequences and overall structures (Omelchenko et al., 2010). Motivated by such functional similarities, we therefore study *similarity-controlled functional protein retrieval* as a complement to conventional similarity search: given a query, a method ranks proteins from a common reference collection by functional relevance, with sequence and structural similarity jointly controlled between the query and each candidate (see Figure 1).

Existing benchmarks cover a broad range of protein prediction and retrieval tasks (Zhou et al., 2019; Rao et al., 2019; Xu et al., 2022; Dallago et al., 2021; Notin et al., 2023; Yang et al., 2024; Tan et al., 2026; Su et al., 2025), but the similarity-controlled retrieval setting defined above remains insufficiently evaluated. Addressing this gap is important because modern protein deep learning models have shown promise in recovering functional relationships missed by conventional similarity search (Qian et al., 2025). Protein sequence language models such as ESM-2 (Lin et al., 2023)

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.

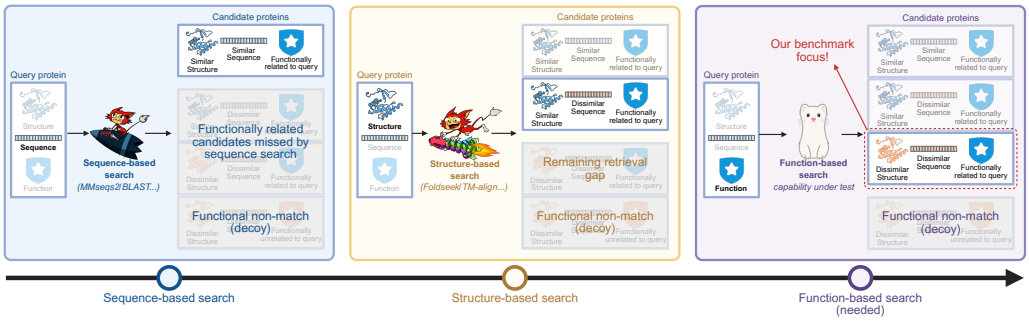

Figure 1: From conventional search to functional protein retrieval. Sequence search retrieves proteins with similar sequences, and structure search extends retrieval to structurally similar proteins. FURETON targets the remaining regime: whether methods can retrieve proteins that are functionally related to the query but dissimilar in both sequence and structure.

and ESM-C (Candido et al., 2026) learn from large sequence collections, while protein structure-aware language models such as SaProt (Su et al., 2024) incorporate structural information. Protein–text contrastive models such as ProtST (Xu et al., 2023), ProteinDT (Liu et al., 2025) and ProTrek (Su et al., 2025) align protein representations with functional descriptions. Protein-to-text generative models such as Prot2Text-V2 (Fei et al., 2026), InstructBioMol (Zhuang et al., 2025), and Evolla (Zhou et al., 2025) instead generate functional descriptions that can serve as textual representations of proteins for retrieval.

We therefore introduce FURETON, a FUnctional RETrieval benchmark built from 89,116 reviewed UniProtKB/Swiss-Prot proteins (The UniProt Consortium, 2023) with AlphaFold-predicted structures (Varadi et al., 2024) and Gene Ontology (GO) annotations retained under a selected evidence-code policy (Ashburner et al., 2000). We construct a joint similarity graph by connecting protein pairs when a sequence or structure search reports similarity above the corresponding threshold. Each connected component forms a cluster, so neither search reports similarity above its threshold for proteins in different clusters. We assign each cluster as a whole to the training, validation, or test split. Protein pairs from different clusters are considered relevant when its GO annotations share a sufficiently informative ancestor (Yu et al., 2010; Resnik, 1999). These annotations define relevance but are not provided as retrieval inputs. This construction yields 3.6 million positive pairs across molecular function, biological process, and cellular component. To evaluate retrieval of these relations, methods rank a common collection of nearly 89,000 proteins for each held-out query (see Table 1).

Under a unified full-corpus evaluation protocol with mean average precision (MAP) as the primary metric (Schütze et al., 2008), we evaluate five families of methods in zero-shot settings: conventional sequence-, structure-, and signature-based search, protein sequence language models, protein structure-aware language models, protein–text contrastive models, and protein-to-text generative models. To test whether contrastive adaptation can further improve retrieval, we adapt protein language models by training only projection heads while keeping the pretrained backbones frozen.

We make the following contributions:

- We introduce FURETON, a dedicated benchmark for functional retrieval combining GO-defined relevance with joint controls on sequence and structural similarity between queries and candidates. Under a unified full-corpus protocol, we systematically evaluate five representative method families for their ability to retrieve functionally related proteins.
- We show that protein-to-text generative models achieve higher overall MAP and Recall@100 than all other zero-shot models and even contrastively adapted protein language models in our primary comparison. These results suggest that generated functional text may serve as a useful representation for functional protein retrieval.
- We find that contrastive adaptation improves average retrieval performance across queries for all eight protein language models. However, query-level analysis shows that many queries benefit from adaptation, but a small subset shows no improvement with any model.

These findings indicate that gains of contrastive adaptation are consistent across models but not universal across queries.

2 Related Work

Similarity-based annotation transfers evidence from sequence, profile, domain, or structural neighbors. BLAST, MMseqs2, HMMER, and HHsearch search sequence or profile space; Pfam, InterPro, and InterProScan organize conserved signatures; and Foldseek extends scalable search to structural neighbors (Altschul et al., 1990; Steinegger & Söding, 2017; Eddy, 1998; Söding, 2005; Mistry et al., 2021; Paysan-Lafosse et al., 2023; Jones et al., 2014; Van Kempen et al., 2024). FURETON complements similarity-based annotation by evaluating functional retrieval under joint sequence and structural similarity controls.

Protein models show promise for functional retrieval beyond conventional sequence and structure searches, but differ in the information they learn from and how they represent proteins. ESM-2 and ESM-C encode sequence information, while SaProt further incorporates structural tokens (Lin et al., 2023; Candido et al., 2026; Su et al., 2024). ProtST, ProteinDT, and ProTrek align proteins with functional language, while Prot2Text-V2, InstructBioMol, and Evolla generate descriptions that can serve as retrieval keys (Xu et al., 2023; Liu et al., 2025; Su et al., 2025; Fei et al., 2026; Zhuang et al., 2025; Zhou et al., 2025). We evaluate the functional retrieval performance of these models on FURETON.

Existing benchmarks evaluate a range of protein prediction and retrieval tasks. Prediction benchmarks include CAFA for GO term prediction, TAPE and PEER for diverse downstream prediction tasks, and FLIP and ProteinGym for fitness prediction (Zhou et al., 2019; Rao et al., 2019; Xu et al., 2022; Dallago et al., 2021; Notin et al., 2023). Retrieval-related benchmarks include CARE for reaction-to-enzyme retrieval, VenusX for assessing functional similarity between sampled protein pairs, and ProTrek for GO-overlap retrieval (Yang et al., 2024; Tan et al., 2026; Su et al., 2025). However, none of these retrieval-related benchmarks combines a common protein collection with joint controls on sequence and structural similarity. Table 3 provides a detailed comparison of the evaluation protocols.

3 Benchmark Construction

FURETON is constructed in four stages. Figure 2 summarizes corpus curation, joint similarity controls, component-disjoint splitting, and GO-based relevance construction. Formal definitions and implementation details are provided in Appendix A.

3.1 Protein Corpus and Annotation Curation

We construct the corpus from reviewed UniProtKB/Swiss-Prot entries (Release `2026_02`), whose records are manually curated (The UniProt Consortium, 2023). GO annotations are obtained from UniProt-GOA (Huntley et al., 2015). To reduce noise from automatically inferred annotations, we retain only experimentally supported or manually traceable GO annotations. These annotations are used only to define functional relevance and are not provided as retrieval inputs. Each benchmark protein is represented by its canonical amino-acid sequence and a matched AlphaFold-predicted structure.

Starting from $575{,}503$ reviewed entries, we retain proteins between 40 and 2,000 residues, require a matching predicted structure in the AlphaFold Protein Structure Database (Varadi et al., 2024), and require at least one retained GO annotation in Molecular Function, Biological Process, or Cellular Component. These filters reduce the corpus as

$$575{,}503 \rightarrow 562{,}905 \rightarrow 538{,}314 \rightarrow 89{,}116,$$

yielding the final protein pool $\mathcal{P}$ with $|\mathcal{P}| = 89{,}116$ (Figure 2a). Details of GO evidence filtering and corpus construction are provided in Appendix A.3.

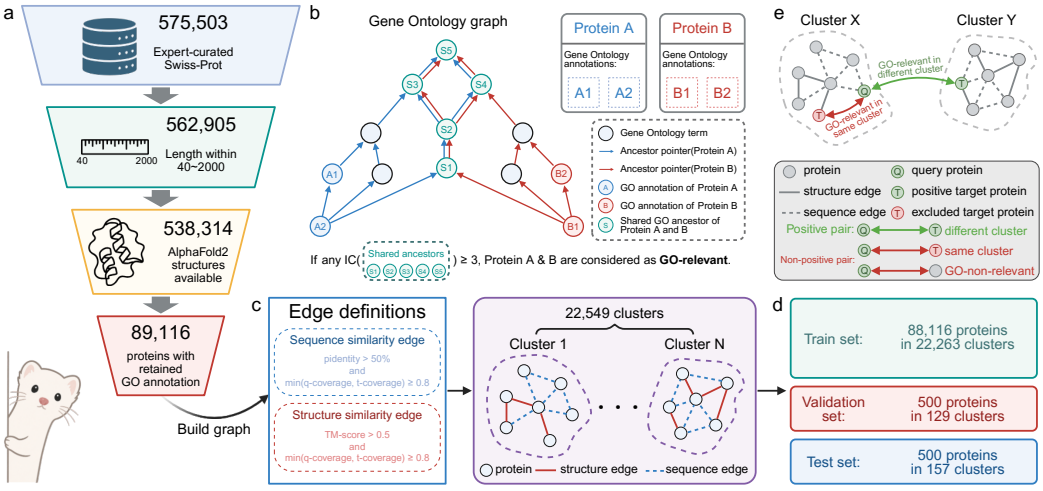

Figure 2: FURETON construction. **(a)** Expert-curated Swiss-Prot proteins are filtered by sequence length, AlphaFold structure availability, and retained GO annotation. **(b)** Direct annotations are propagated through `is_a` links, and a pair is relevant when its most informative shared ancestor has IC ≥ 3. **(c)** MMseqs2 and Foldseek matches define a joint similarity graph. Between different components, neither search reports a match meeting its criterion: sequence identity $> 50\%$ for MMseqs2 or alignment-normalized TM-score > 0.5 for Foldseek, both requiring alignment coverage ≥ 0.8 of each protein. **(d)** Each connected component is assigned intact to exactly one of the training, validation, or test sets. **(e)** Only GO-relevant protein pairs from different clusters are considered positive. Same-cluster pairs are excluded even if GO-relevant, while cross-cluster pairs that do not meet the GO relevance criterion are non-positive. Only positive pairs are stored.

3.2 Joint Similarity Controls and Component-Disjoint Splits

Sequence-only clustering can leave structurally similar proteins across splits, whereas structure-only clustering can leave sequence-similar proteins. To prevent highly similar proteins from appearing across different splits, we jointly consider sequence and structural similarity by constructing a similarity graph in which proteins are nodes and an undirected edge is added when a pair satisfies either a sequence or structure match criterion (Figure 2c).

Sequence and structural matches are identified separately using MMseqs2 (Steinegger & Söding, 2017) and Foldseek (Van Kempen et al., 2024), respectively. For sequence similarity, we perform an all-against-all MMseqs2 search over the corpus and define a match when the reported sequence identity exceeds 50% with at least 80% coverage of both proteins. The corresponding structural one is defined by an all-against-all Foldseek search, where a match requires an alignment-normalized TM-score above 0.5 and at least 80% coverage of both proteins. We construct an undirected similarity graph by connecting protein pairs that satisfy either the sequence or structural match criterion.

The connected components of this graph define the benchmark's similarity clusters, yielding $22{,}549$ clusters. Each cluster is assigned entirely to one split, ensuring that no proteins linked through sequence or structural similarity are separated across splits. The resulting partition contains $88{,}116$ training proteins in $22{,}263$ clusters, 500 validation proteins in 129 clusters, and 500 test proteins in 157 clusters (Figure 2d).

The training split is used for supervised adaptation, the validation split for checkpoint selection, and the test split as the fixed query set for final evaluation. Complete search parameters, formal graph definitions, and the component assignment algorithm are provided in Appendix A.4.

3.3 GO-Based Functional Relevance and Pair Construction

We define functional relevance separately for the three GO aspects. For each protein and aspect, GO terms directly annotated to the protein are expanded by GO ancestor closure, including the annotated

Table 1: Benchmark composition and retrieval statistics. Pair counts report relevant records under each task. For the ALL task, a pair is relevant if it meets the relevance criterion in at least one GO aspect. Effective queries have at least one relevant candidate in the test gallery.

Statistic	MF	BP	CC	ALL
Proteins	62,125	60,308	64,024	89,116
Training pairs	600,102	2,607,396	414,204	3,517,524
Validation pairs	6,379	29,440	4,296	38,832
Test pairs	8,978	27,079	4,202	38,866
Test queries with pairs	190	270	74	364
Median relevant per query	44	73	49.5	73

terms and all terms reachable through `is_a` relations. We quantify the specificity of each GO term t using information content (IC), computed from its frequency in our benchmark corpus (Resnik, 1999; Lord et al., 2002),

$$\mathrm{IC}(t) = -\log_{10}\left(\frac{f(t)}{N}\right), \qquad N = |\mathcal{P}| = 89{,}116, \tag{1}$$

where $f(t)$ is the number of proteins in the corpus whose propagated annotations contain t.

Let $\bar{T}_i^a$ denote the propagated annotation set of protein i in aspect $a \in \{\mathrm{MF}, \mathrm{BP}, \mathrm{CC}\}$. Following Resnik's most-informative-common-ancestor criterion (Resnik, 1999; Lord et al., 2002; Yu et al., 2010), we define

$$S_{\mathrm{GO}}^a(i,j) = \max_{t \in \bar{T}_i^a \cap \bar{T}_j^a} \mathrm{IC}(t), \qquad y_{ij}^a = \mathbf{1}[S_{\mathrm{GO}}^a(i,j) \geq 3]\,. \tag{2}$$

The similarity is set to zero when the intersection is empty. We define four retrieval tasks, $\mathcal{T} = \{\mathrm{MF}, \mathrm{BP}, \mathrm{CC}, \mathrm{ALL}\}$. The first three correspond to Molecular Function, Biological Process, and Cellular Component, respectively, while ALL considers a pair relevant if the threshold is met in at least one aspect. The formal definition of the ancestor closure and relevance criteria is provided in Appendix A.5.

We form training pairs from functionally relevant proteins in different similarity clusters within the training split. Validation queries are paired with all training proteins, while test queries are paired with all training and validation proteins. We store only relevant pairs and treat candidates that do not meet the corresponding GO relevance criterion as non-relevant during evaluation. Queries without a relevant candidate under a given criterion are excluded from that criterion's metrics.

In total, we obtain 3,595,222 relevant protein pairs for training, validation, and testing, counting each pair only once even if it is relevant under multiple GO aspects. Table 1 summarizes the related statistics. Exact set definitions are provided in Appendix A.6.

4 Evaluation Protocol

We evaluate each method by how well it ranks functionally relevant proteins in the benchmark's retrieval tasks. Further scoring, training, and metric details are provided in Appendix B.

4.1 Retrieval Methods

Methods are grouped by how they represent proteins and compute query–candidate similarity. Table 4 summarizes the three method families considered in our evaluation.

Conventional sequence and structure search. MMseqs2 (Steinegger & Söding, 2017) and Foldseek (Van Kempen et al., 2024) provide direct sequence- and structure-similarity baselines, respectively. Returned MMseqs2 hits are ranked by aligned sequence identity, whereas returned Foldseek hits are ranked by alignment TM-score. Both methods are evaluated without benchmark-specific training (details in Appendix B.1).

Protein-embedding retrieval. We evaluate three model families: protein sequence language models, protein structure-aware language models, and protein–text contrastive models. The sequence models include ESM-2 (Lin et al., 2023) and ESM-C (Candido et al., 2026), while the structure-aware model is SaProt (Su et al., 2024). The contrastive models include ProtST (Xu et al., 2023), ProTrek (Su et al., 2025), and ProteinDT (Liu et al., 2025). We compute query and candidate embeddings with each model, using only the protein branch for contrastive models. Candidates are ranked by cosine similarity between these embeddings.

Text-mediated retrieval. Text-mediated retrieval uses functional descriptions from three sources: protein-to-text generative models, conventional signature-based annotation, and expert-curated protein records. The generative models include Evolla (Zhou et al., 2025), Prot2Text-V2 (Fei et al., 2026), and InstructBioMol (Zhuang et al., 2025). Signature-based descriptions are assembled from InterProScan annotations (Jones et al., 2014), while the curated reference is assembled from Swiss-Prot protein-name, comment, and keyword fields. To avoid directly exposing the relevance labels, we exclude GO identifiers, mapped GO terms, and fields used to define relevance from the InterProScan and Swiss-Prot descriptions. All descriptions are embedded with the same fixed text encoder (see Appendix C for the text-encoder ablation). Detailed scoring and text-processing procedures are provided in Appendix B.1.

4.2 Zero-Shot Evaluation and Contrastive Adaptation

We evaluate all methods in a zero-shot setting and selected protein language models after contrastive adaptation.

Zero-shot evaluation. We evaluate all methods without using the FURETON training data for adaptation.

Contrastive adaptation. We apply contrastive adaptation to selected protein language models using the FURETON training pairs and a CLIP-style contrastive objective (Radford et al., 2021). For each backbone f_m, we freeze its parameters and train a projection head p_ω together with a learnable temperature $\tau > 0$. The projection head is shared between the query and candidate branches.

Let $\mathcal{B} = \{(q_i, c_i)\}_{i=1}^{N} \subset \mathcal{R}_{\text{train}}^{\text{ALL}}$ be a batch of positive, cross-component training pairs. Given a native backbone embedding z_v^m, we define

$$\bar{z}_v^m = \frac{p_\omega(z_v^m)}{\|p_\omega(z_v^m)\|_2}, \qquad \ell_{ij} = \frac{(\bar{z}_{q_i}^m)^\top \bar{z}_{c_j}^m}{\tau}. \tag{3}$$

The projection head and temperature are optimized with the symmetric InfoNCE objective (Oord et al., 2018)

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N} \sum_{i=1}^{N} \left[\log \frac{\exp(\ell_{ii})}{\sum_{j=1}^{N} \exp(\ell_{ij})} + \log \frac{\exp(\ell_{ii})}{\sum_{j=1}^{N} \exp(\ell_{ji})} \right]. \tag{4}$$

All off-diagonal pairings are treated as in-batch negatives. Sampling and optimization details are given in Appendix B.2.

After each training epoch, the 500 validation queries are ranked against the fixed $88{,}116$-protein validation candidates. We retain the checkpoint with the highest validation MAP under the ALL task. The selected checkpoint is then evaluated once on the test queries against the $88{,}616$-protein test gallery, without retraining on the validation proteins. The complete training and model-selection workflow is shown in Appendix B.2.

4.3 Evaluation Metrics

For each task $t \in \mathcal{T}$, we compute average precision (AP), Recall@K, and AUROC for each query and report their averages, giving each query equal weight. For the ALL task, metrics are computed using the combined relevance criterion in Eq. (21), rather than by averaging the MF, BP, and CC results. MAP is our primary metric and evaluates the ranking of all relevant candidates. Recall@K,

for $K \in \{1, 5, 10, 100\}$, measures the fraction of relevant candidates retrieved within the top K results. For methods that score the complete candidate collection, AUROC measures global positive–negative discrimination, but higher AUROC does not necessarily indicate better retrieval among the top-ranked candidates. Formal metric definitions are provided in Appendix B.3.

5 Experimental results

5.1 Functional text leads retrieval, while metrics reveal different strengths

The metrics capture complementary aspects of retrieval: MAP, our primary metric, summarizes ranking quality across all relevant candidates; Recall@K measures their recovery within a fixed retrieval budget; and AUROC measures global positive–negative discrimination. We therefore assess retrieval using both MAP and Recall@K, with AUROC providing a complementary view.

Under joint sequence and structure controls, protein-to-text generative models achieved the strongest MAP and Recall@100 for ALL task (Figure 3; Table 2). On the ALL task, Evolla led MAP at 7.65%, followed by Prot2Text-V2 (7.25%) and InstructBioMol (6.62%), exceeding the best adapted protein language model (ESM2-35M, 4.65%) and protein–text contrastive model (ProtST-ESM2, 5.71%). Retrieval depth also affected the ordering: Evolla led learned Recall@1/5/10, whereas InstructBioMol led Recall@100 (0.1359), followed by Prot2Text-V2 (0.1351) and Evolla (0.1313) (Appendix E).

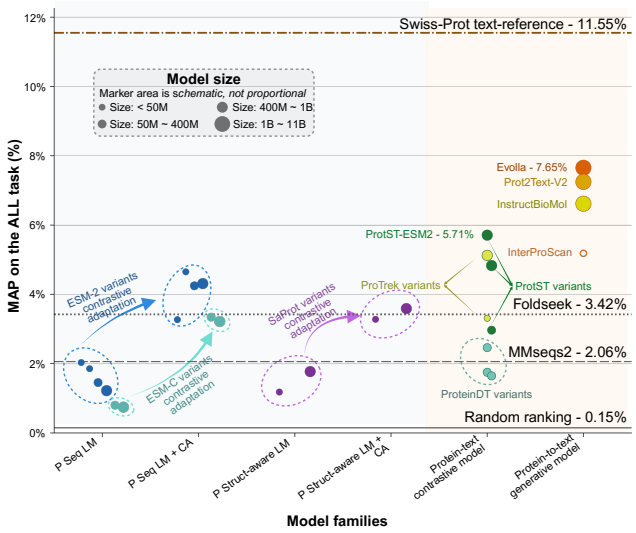

Figure 3: MAP on the ALL task across five method families, including zero-shot and contrastively adapted protein language models. Matched zero-shot and adapted protein language models are connected within their model family. MMseqs2, Foldseek, random ranking, and the Swiss-Prot text-reference are shown as horizontal references. Marker size indicates only a coarse parameter-size bin and is not proportional to parameter count. Abbreviations: P Seq LM, protein sequence language model; P Struct-aware LM, protein structure-aware language model; CA, contrastive adaptation.

AUROC yielded a different ordering: adapted ESM2-650M achieved the highest learned AUROC in the primary comparison (0.8239), while its MAP and Recall@100 remained below all three protein-to-text generative models. This advantage illustrates that stronger global discrimination need not yield better ranking of relevant candidates or greater recovery within a limited retrieval budget.

We further evaluated two other text sources: conventional signature-based descriptions from InterProScan and expert-curated protein records from Swiss-Prot. Under the same text encoder, InterProScan descriptions reached 5.18% MAP and 0.1035 Recall@100, whereas the Swiss-Prot reference reached 11.55% MAP and 0.1715 Recall@100. Swiss-Prot provides expert-curated functional

Table 2: MAP, Recall@100, and AUROC of representative methods on the ALL task, including zero-shot and contrastively adapted protein language models. In the Setting column, "contrastive adapt." denotes contrastive adaptation. For ESM-C and SaProt, we show the highest-MAP adapted variant in each series. For ESM-2, we show the best adapted variant for each metric. All adapted variants are shown with their zero-shot counterparts. For each protein–text contrastive model series, we include the highest-MAP variant. Underlining marks the best displayed result within each family; bold marks the best result excluding the shaded Swiss-Prot reference. Full results are provided in Table 11.

Family	Method	Setting	MAP	R@100	AUROC
Control	random	–	0.0015	0.0011	0.5055
Conventional search	MMseqs2	–	0.0206	0.0368	0.5457
	Foldseek	–	0.0342	0.0679	0.5637
	InterProScan	–	0.0518	0.1035	0.7075
Protein sequence language model	ESM2-35M	zero-shot	0.0185	0.0480	0.6416
	ESM2-35M	contrastive adapt.	0.0465	0.0735	0.7791
	ESM2-150M	zero-shot	0.0146	0.0363	0.6396
	ESM2-150M	contrastive adapt.	0.0425	0.0771	0.8097
	ESM2-650M	zero-shot	0.0122	0.0295	0.6163
	ESM2-650M	contrastive adapt.	0.0431	0.0722	**0.8239**
	ESMC-300M	zero-shot	0.0080	0.0231	0.6408
	ESMC-300M	contrastive adapt.	0.0334	0.0598	0.8148
Protein structure-aware language model	SaProt-650M	zero-shot	0.0177	0.0382	0.6784
	SaProt-650M	contrastive adapt.	0.0359	0.0569	0.8129
Protein–text contrastive model	ProTrek-650M	zero-shot	0.0512	0.0803	0.7778
	ProtST-esm2	zero-shot	0.0571	0.0975	0.8165
	ProteinDT-InfoNCE	zero-shot	0.0246	0.0658	0.7497
Protein-to-text generative model	Evolla-10B-DPO	zero-shot	**0.0765**	0.1313	0.8018
	Prot2Text-V2	zero-shot	0.0725	0.1351	0.7459
	InstructBioMol	zero-shot	0.0662	**0.1359**	0.7454
Expert-curated reference	Swiss-Prot	–	0.1155	0.1715	0.8156

descriptions, indicating substantial room for improving functional text representations for protein retrieval. GO-aspect results appear in Appendix D.

5.2 Contrastive adaptation improves retrieval, but gains remain query-selective

Contrastive adaptation improved MAP, AUROC, and Recall@100 for every matched backbone across MF, BP, CC, and ALL: all paired comparisons across eight protein language models, four tasks, and three metrics improved (Figure 4; Table 12). ESMC illustrates the gains from weak zero-shot representations: ESMC-300M increased from 0.0080 to 0.0334 MAP and from 0.6408 to 0.8148 AUROC, while ESMC-600M increased from 0.0074 to 0.0322 MAP and from 0.6446 to 0.8177 AUROC. These results suggest that frozen representations can contain functional signal that is poorly expressed in their native embedding space and that contrastive adaptation can induce these latent signals to become more accessible for GO-defined retrieval.

Retrieval gains were consistent across models but uneven across queries (Appendix E.1). Roughly seven in ten queries gained AP on at least five of the eight backbones, while a small minority showed no gain on any backbone. A small number of query–backbone pairs contributed about two-thirds of the total positive AP gain, indicating that adaptation benefits were shared across many queries but not uniformly distributed.

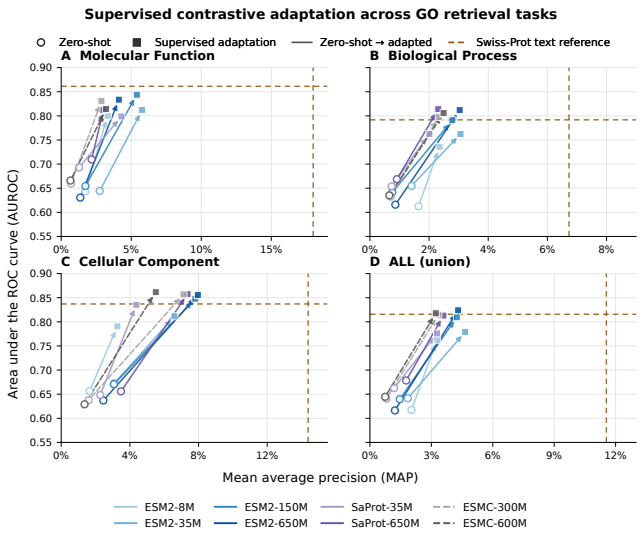

Figure 4: Zero-shot (open circles) and contrastively adapted (filled squares) results for eight backbones across MF, BP, CC, and ALL. Axes show MAP (horizontal) and AUROC (vertical); lines connect each backbone's paired results. Brown dashed lines mark Swiss-Prot text-reference MAP (vertical) and AUROC (horizontal). MAP ranges vary by task; the AUROC range is shared.

6 Discussion and Conclusion

Functional similarity does not have a single universally accepted definition. FURETON adopts a GO-based definition of relevance as a reproducible starting point for evaluating functional retrieval under joint sequence and structural similarity controls. To extend evaluation beyond this GO-based formulation, future benchmarks could address different biological questions by defining relevance through, for example, shared pathway membership, disease-associated relationships, or reaction similarity.

Beyond the question of how functional similarity is defined, the concept of protein function itself varies across biological contexts. Enzymologists may focus on catalytic properties such as reaction specificity or kinetic parameters, whereas disease-oriented studies may emphasize molecular interactions, pathway involvement, or disease associations. These differences extend to protein function prediction, where models trained with different data and methods may perform well on specific tasks without necessarily generalizing to others. This motivates the search for a common representation that can capture different types of functional information and support generalization across tasks.

A potential basis for such a representation is natural language, which researchers across disciplines already use to describe protein functions and communicate their findings. This shared use suggests that text could serve as a common medium for representing diverse functional information in retrieval.

Our results provide initial evidence supporting this direction: protein-to-text generative models outperformed other learned representations under the FURETON setting. However, the higher performance of the expert-curated Swiss-Prot reference suggests that current generated descriptions do not yet fully capture the functional information available in curated records. Even this reference achieved only 11.55% MAP, leaving substantial room for improvement on the benchmark. This score does not establish an inherent limit of functional text, as the Swiss-Prot descriptions were used without optimization for this retrieval task. Future work could investigate whether protein-to-text models can improve retrieval by generating more precise descriptions of functional properties and relationships, such as catalytic activities, interaction partners, or pathway roles.

More broadly, functional protein retrieval remains a challenging problem. We hope FURETON will provide a common benchmark for the community to compare approaches, identify their limitations, and develop more effective retrieval methods.

AI USE STATEMENT

We used ChatGPT to assist with language editing, manuscript organization, and discussions of result interpretation and limitations.

The authors take responsibility for the final manuscript and research artifacts.

REPRODUCIBILITY STATEMENT

Appendix A details corpus curation, similarity-controlled splitting, and GO-based relevance construction. Appendix B specifies retrieval scoring, generation settings, contrastive adaptation, checkpoint selection, and evaluation metrics. The anonymous repository is available at `https://anonymous.4open.science/r/FURETON-EA8D`.

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.

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, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z Wu, Jin Sub Lee, Elise S Bruguera, Jehan A Keval, et al. Language modeling materializes a world model of protein biology. *bioRxiv*, pp. 2026–06, 2026.

Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K. Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://openreview.net/forum?id=p2dMLEwL8tF.

Sean R. Eddy. Profile hidden markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763, 1998.

Evidence and Conclusion Ontology developers. Default mappings from GO evidence codes to ECO identifiers, 2026. URL http://purl.obolibrary.org/obo/eco/gaf-eco-mapping.txt. Permanent official mapping; accessed 2026-09-15.

Xiao Fei, Michail Chatzianastasis, Sarah Carneiro, Hadi Abdine, Lawrence Petalidis, and Michalis Vazirgiannis. Prot2text-v2: Protein function prediction with multimodal contrastive alignment. *Advances in Neural Information Processing Systems*, 38:26205–26232, 2026.

Gameselo. STS-multilingual-mpnet-base-v2. Hugging Face model repository, 2024. URL https://huggingface.co/Gameselo/STS-multilingual-mpnet-base-v2. Model ID: Gameselo/STS-multilingual-mpnet-base-v2; revision: 449f917af30f590fc31f9ffb226c94f21a2f47b8;.

Rachael P Huntley, Tony Sawford, Prudence Mutowo-Meullenet, Aleksandra Shypitsyna, Carlos Bonilla, Maria J Martin, and Claire O'Donovan. The goa database: gene ontology annotation updates for 2015. *Nucleic acids research*, 43(D1):D1057–D1063, 2015.

infgrad. Jasper-Token-Compression-600M. Hugging Face model repository, 2025. URL https://huggingface.co/infgrad/Jasper-Token-Compression-600M. Model ID: infgrad/Jasper-Token-Compression-600M; revision: e1373068c35fba2a37568da4eb4105377e17eae9;.

Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, et al. Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 2014.

Kingsoft-LLM. QZhou-Embedding. Hugging Face model repository, 2025. URL https://huggingface.co/Kingsoft-LLM/QZhou-Embedding. Model ID: Kingsoft-LLM/QZhou-Embedding; revision: 9b2eccf023325e047ffd3fe4aa34f135ab48a3d2;.

Xianming Li and Jing Li. Aoe: Angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1825–1839, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, et al. A text-guided protein design framework. *Nature Machine Intelligence*, 7(4):580–591, 2025.

Phillip W Lord, Robert D Stevens, Andy Brass, and Carole A Goble. Semantic similarity measures as tools for exploring the gene ontology. In *Biocomputing 2003*, pp. 601–612. World Scientific, 2002.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1):D412–D419, 2021.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in neural information processing systems*, 36: 64331–64379, 2023.

Marina V Omelchenko, Michael Y Galperin, Yuri I Wolf, and Eugene V Koonin. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biology direct*, 5(1):31, 2010.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, et al. Interpro in 2022. *Nucleic acids research*, 51(D1):D418–D427, 2023.

Hui Qian, Yuxuan Wang, Xibin Zhou, Tao Gu, Hui Wang, Hao Lyu, Zhikai Li, Xiuxu Li, Huan Zhou, Chengchen Guo, et al. Esm-ezy: a deep learning strategy for the mining of novel multicopper oxidases with superior properties. *Nature Communications*, 16(1):3274, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

Sentence-Transformers. all-MiniLM-L6-v2. Hugging Face model repository, 2021. URL https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Model ID: sentence-transformers/all-MiniLM-L6-v2; revision: 1110a243fdf4706b3f48f1d95db1a4f5529b4d41;.

Johannes Söding. Protein homology detection by hmm–hmm comparison. *Bioinformatics*, 21(7): 951–960, 2005.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

David L Stern. The genetic causes of convergent evolution. *Nature Reviews Genetics*, 14(11): 751–764, 2013.

Jin Su, Yan He, Shiyang You, Shiyu Jiang, Xibin Zhou, Xuting Zhang, Yuxuan Wang, Xining Su, Igor Tolstoy, Xing Chang, et al. A trimodal protein language model enables advanced protein searches. *Nature Biotechnology*, pp. 1–7, 2025.

Jin Su et al. SaProt: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations (ICLR)*, 2024.

Yang Tan, Wenrui Gou, Bozitao Zhong, Huiqun Yu, Bingxin Zhou, et al. Venusx: Unlocking fine-grained functional understanding of proteins. In *International Conference on Learning Representations*, volume 2026, pp. 44475–44507, 2026.

The Gene Ontology Consortium. Introduction to GO annotations, 2026a. URL `https://geneontology.org/docs/go-annotations/`. Official documentation; accessed 2026-09-15.

The Gene Ontology Consortium. Guide to GO evidence codes, 2026b. URL `https://geneontology.org/docs/guide-go-evidence-codes/`. Official documentation; accessed 2026-09-15.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic acids research*, 51(D1):D523–D531, 2023.

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, et al. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic acids research*, 52(D1):D368–D375, 2024.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

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. *Advances in Neural Information Processing Systems*, 35:35156–35173, 2022.

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

Jason Yang, Ariane Mora, Shengchao Liu, Bruce J Wittmann, Anima Anandkumar, Frances H Arnold, and Yisong Yue. Care: a benchmark suite for the classification and retrieval of enzymes. *Advances in Neural Information Processing Systems*, 37:3094–3121, 2024.

Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, 26(7):976–978, 2010.

Peng Yu, En Xu, Bin Chen, Haibiao Chen, and Yinfei Xu. QZhou-Embedding Technical Report, 2025. URL `https://arxiv.org/abs/2508.21632`.

Dun Zhang, Ziyang Zeng, Yudong Zhou, and Shuyang Lu. Jasper-token-compression-600m technical report. *arXiv preprint arXiv:2511.14405*, 2025.

Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.

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(1):244, 2019.

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*, pp. 2025–01, 2025.

Xiang Zhuang, Keyan Ding, et al. Advancing biomolecular understanding and design following human instructions. *Nature Machine Intelligence*, 7:1154–1167, 2025. doi: 10.1038/s42256-025-01064-0.

Table 3: Protocol-level comparison with representative protein benchmarks and retrieval evaluations. "Common gallery" denotes a benchmark-defined candidate collection shared across queries; "full ranking" denotes scoring the entire defined gallery rather than sampled pairs.

Benchmark	Query/target	Relevance	Common gallery	Full ranking	Sequence control	Structure control	Supervision	Main metrics
CAFA	Protein → GO	GO labels	No	No	Temporal/sequence	No	Supervised	$F_{\max}$, $S_{\min}$
TAPE / PEER	Task-specific	Task labels	No	No	Varies	No	Supervised	Task-specific
FLIP / ProteinGym	Variant → fitness	Experimental fitness	No	No	Mutation/split	No	Zero-shot/Supervised	Spearman's ρ
CARE	Reaction → enzyme	Reaction match	Restricted	Yes	No	No	Supervised	Top-1 accuracy
VenusX	Protein pair	InterPro family	No	No	Identity clusters	No	Zero-shot	AUC
ProTrek	Protein → protein	GO overlap	Yes	Yes	Yes	No	Zero-shot	Accumulated TP
FURETON	**Protein → protein**	**GO-based**	**Yes**	**Yes**	**Yes**	**Yes**	**Zero-shot/Supervised**	**MAP, Recall@k, AUROC**

Table 4: Overview of evaluated retrieval methods. Adapted protein encoders use cosine similarity in the learned projection space described in Section 4.2.

Family	Methods	Retrieval representation	Ranking score
Pairwise search	MMseqs2	sequence alignments	aligned sequence identity
	Foldseek	structure alignments	alignment TM-score
Protein embedding	ESM-2 / ESM-C	sequence representation	cosine similarity
	SaProt	sequence–3Di representation	cosine similarity
	ProtST / ProTrek / ProteinDT	pretrained protein branch	cosine similarity
Text-mediated	Evolla-10B-DPO / Prot2Text-V2 / InstructBioMol	generated description	text-embedding cosine
	InterProScan	annotation-derived text	text-embedding cosine
	Swiss-Prot reference	curated annotation text	text-embedding cosine

A Additional Benchmark Construction Details

A.1 Protocol Context and Evaluated-Method Registry

Table 3 expands the protocol-level comparison summarized in the related-work discussion. Table 4 records the representation and ranking score used by each evaluated family.

A.2 Formal Retrieval Task Definition

Each protein $i \in \mathcal{P}$ is associated with an amino-acid sequence x_i, a predicted structure X_i, and retained GO annotations A_i^a for $a \in \mathcal{A} = \{\mathrm{MF}, \mathrm{BP}, \mathrm{CC}\}$. The annotations are used to define relevance judgments but are not directly provided as retrieval inputs.

We define

$$\mathcal{T} = \{\mathrm{MF}, \mathrm{BP}, \mathrm{CC}, \mathrm{ALL}\}.$$

Let $\mathcal{Q}_\tau$ and $\mathcal{C}_\tau$ denote the query and candidate sets at evaluation stage $\tau \in \{\mathrm{val}, \mathrm{test}\}$, as defined in Appendix A.6.

Given a retrieval method m, let $s_m(q, c)$ denote its score for query q and candidate c. For methods that represent proteins as vectors, let u_i^m denote the method-specific input for protein i, which may comprise its sequence and structure. We write

$$z_i^m = f_m(u_i^m), \qquad s_m(q, c) = g_m(z_q^m, z_c^m), \tag{5}$$

where f_m is the representation function and g_m is the corresponding pairwise scoring function.

The induced full-collection ranking is

$$\pi_{q,\tau}^m = \mathrm{argsort}^{\downarrow}_{c \in \mathcal{C}_\tau} s_m(q, c). \tag{6}$$

The same candidate collection is used across the four GO tasks at a given evaluation stage. The tasks differ through their relevance judgments y_{qc}^a, rather than through the candidate set.

A.3 Corpus and GO Annotation Processing

The reviewed-accession universe, length filter, and annotation text were obtained from UniProtKB/Swiss-Prot release `2026_02` with 575,503 reviewed entries (The UniProt Consortium, 2023). AlphaFold database structures (Varadi et al., 2024) were matched by UniProt accession.

Table 5: Exact GO evidence whitelist and final retained annotation counts.

GO code	ECO identifier	Evidence name	Classification	Final counts
EXP	ECO:0000269	Inferred from Experiment	experimental	48,695
IDA	ECO:0000314	Inferred from Direct Assay	experimental	301,159
HDA	ECO:0007005	Inferred from High Throughput Direct Assay	high-throughput experimental	47,252
IPI	ECO:0000353	Inferred from Physical Interaction	experimental	361,065
HTP	ECO:0006056	Inferred from High Throughput Experiment	high-throughput experimental	1,209
IGI	ECO:0000316	Inferred from Genetic Interaction	experimental	32,299
HGI	ECO:0007003	Inferred from High Throughput Genetic Interaction	high-throughput experimental	65
IEP	ECO:0000270	Inferred from Expression Pattern	experimental	18,200
HEP	ECO:0007007	Inferred from High Throughput Expression Pattern	high-throughput experimental	568
IMP	ECO:0000315	Inferred from Mutant Phenotype	experimental	167,990
HMP	ECO:0007001	Inferred from High Throughput Mutant Phenotype	high-throughput experimental	2,006
TAS	ECO:0000304	Traceable Author Statement	author statement	130,306
IC	ECO:0000305	Inferred by Curator	curator inference	5,079

Table 6: Corpus filtering and retained direct GO annotation coverage.

Stage/aspect	Proteins	Coverage	Denominator
Swiss-Prot reviewed	575,503	100.00%	Swiss-Prot reviewed
Length 40–2,000 aa	562,905	97.81%	Swiss-Prot reviewed
AlphaFold structure available	538,314	93.54%	Swiss-Prot reviewed
Retained GO annotation available	89,116	15.48%	Swiss-Prot reviewed
MF	62,125	69.71%	final benchmark pool
BP	60,308	67.67%	final benchmark pool
CC	64,024	71.84%	final benchmark pool

GO annotations carry evidence codes for experimental, phylogenetic, computational, author-statement, curator-statement, and automatically generated support (The Gene Ontology Consortium, 2026b;a). The experimental family includes direct and high-throughput subtypes. The benchmark retains those experimental codes together with traceable author statements (TAS) and curator inferences (IC). It excludes electronic inference (IEA), phylogenetic and computational inference, non-traceable statements, and all other codes. This policy emphasizes evidence that is directly measured or manually traceable and reduces reliance on automatically or computationally transferred annotations, including codes explicitly based on sequence or structural similarity. Table 5 gives the exact GO-code to ECO-identifier whitelist (Evidence and Conclusion Ontology developers, 2026). The evidence code IC is distinct from the information-content quantity used in the relevance definition.

The filters were applied in the following order:

1. Apply the 40–2,000-residue eligibility filter to the release `2026_02` canonical sequences, reducing the corpus from $575{,}503$ to $562{,}905$ proteins.
2. Require an AlphaFold Database structure matched by UniProt accession, leaving $538{,}314$ proteins.
3. Require at least one retained GO annotation in Molecular Function, Biological Process, or Cellular Component, leaving the final $89{,}116$ proteins.

The distribution after each step and aspect is in Table 6.

A.4 Formal Similarity Graph and Split Construction

We queried every corpus member against the same corpus database with MMseqs2 (Steinegger & Söding, 2017) and Foldseek (Van Kempen et al., 2024). These are database-against-itself searches, but the hit caps and prefilters do not enumerate every protein pair. Software versions and complete command-line parameters are given below. The preserved MMseqs2 run log records commit `17b688d21dda57fc5f5b7286ecba7ec003d4717f`:

```
mmseqs createdb full_pool.fasta poolDB
mmseqs search poolDB poolDB searchDB search_tmp \
  --min-seq-id 0.0 -e 10 --max-seqs 100000 -s 7.5
```

```
mmseqs convertalis poolDB poolDB searchDB seq_results_all.m8 \
  --format-output query,target,pident,alnlen,evalue,bits
```

Foldseek version (commit): `70cea935733e0f70e3e10a0572b45ba6e21bfa2d`. The command used was as follows:

```
foldseek createdb pdb_pool poolDB --chain-name-mode 0
foldseek search poolDB poolDB alignDB tmp -a 1 -e 1 \
  --max-seqs 100000 -c 0.3 --cov-mode 1 --threads 96
foldseek convertalis poolDB poolDB alignDB struct_results_all.m8 \
  --format-output query,target,alntmscore,evalue,bits,qcov,\
tcov,alnlen
```

For a returned directional record $i \to j$, let $I_{\mathrm{seq}}(i,j) \in [0,1]$ denote the MMseqs2-reported default estimated percent identity divided by 100, and let $T_{\mathrm{str}}(i,j) \in [0,1]$ denote the Foldseek alignment-normalized TM-score (Zhang & Skolnick, 2005). For MMseqs2, the graph builder computes query and target coverage from the exported alignment length and the two FASTA sequence lengths. For Foldseek, it uses the exported `qcov` and `tcov` fields. We write these values as $q^r_{\mathrm{cov}}(i,j)$ and $t^r_{\mathrm{cov}}(i,j)$ for modality $r \in \{\mathrm{seq}, \mathrm{str}\}$. If no directional record $i \to j$ was exported for modality r, we define $\widetilde{m}_r(i,j) = 0$.

The directional sequence-match indicator is

$$\widetilde{m}_{\mathrm{seq}}(i,j) = \mathbf{1}[I_{\mathrm{seq}}(i,j) > 0.5 \ \wedge \ \min\{q^{\mathrm{seq}}_{\mathrm{cov}}(i,j), t^{\mathrm{seq}}_{\mathrm{cov}}(i,j)\} \geq 0.8]\,. \tag{7}$$

The directional structure-match indicator is

$$\widetilde{m}_{\mathrm{str}}(i,j) = \mathbf{1}\big[T_{\mathrm{str}}(i,j) > 0.5 \ \wedge \ \min\big\{q^{\mathrm{str}}_{\mathrm{cov}}(i,j), t^{\mathrm{str}}_{\mathrm{cov}}(i,j)\big\} \geq 0.8\big]\,. \tag{8}$$

For either modality, a search match is present if a qualifying record was returned in at least one direction:

$$m_r(i,j) = \max\{\widetilde{m}_r(i,j), \widetilde{m}_r(j,i)\}\,, \qquad r \in \{\mathrm{seq}, \mathrm{str}\}. \tag{9}$$

We construct the undirected joint similarity graph

$$G_{\mathrm{sim}} = \big(\mathcal{P}, E_{\mathrm{sim}}\big), \qquad \{i,j\} \in E_{\mathrm{sim}} \iff m_{\mathrm{seq}}(i,j) = 1 \ \vee \ m_{\mathrm{str}}(i,j) = 1. \tag{10}$$

The connected components

$$\mathcal{K}_{\mathrm{sim}} = \mathrm{CC}(G_{\mathrm{sim}}) \tag{11}$$

define the similarity clusters. This construction yields 22,549 components. Component sizes are summarized in Figure 5 and Table 7.

Let $C(i) \in \mathcal{K}_{\mathrm{sim}}$ denote the component containing protein i, and let

$$\rho : \mathcal{K}_{\mathrm{sim}} \to \{\mathrm{train}, \mathrm{val}, \mathrm{test}\}$$

denote the component-to-split assignment. The protein subsets are

$$\mathcal{P}_s = \big\{i \in \mathcal{P} : \rho\big(C(i)\big) = s\big\}\,, \qquad s \in \{\mathrm{train}, \mathrm{val}, \mathrm{test}\}. \tag{12}$$

Relative to the records returned by these two searches, a pair satisfying either match criterion is connected by an edge and therefore belongs to the same connected component. Consequently, proteins in different splits satisfy

$$i \in \mathcal{P}_s, \quad j \in \mathcal{P}_{s'}, \quad s \neq s' \Longrightarrow m_{\mathrm{seq}}(i,j) = m_{\mathrm{str}}(i,j) = 0. \tag{13}$$

For the reported split, the component identifiers were read in their stored order, shuffled once with Python's `random.Random(42)`, and traversed once. At each step the complete current component was assigned to test if the current test count was below 500, otherwise to validation if its count was below 500, and otherwise to training. This produced

$$\begin{aligned} |\mathcal{P}_{\mathrm{train}}| &= 88{,}116 && \text{proteins in } 22{,}263 \text{ components}, \\ |\mathcal{P}_{\mathrm{val}}| &= 500 && \text{proteins in } 129 \text{ components}, \\ |\mathcal{P}_{\mathrm{test}}| &= 500 && \text{proteins in } 157 \text{ components}. \end{aligned} \tag{14}$$

The exact counts of 500 proteins in both validation and test arose by chance: because components were assigned atomically, the splitting procedure only ensured that each split was filled until its size reached or exceeded 500, rather than guaranteeing exactly 500 proteins.

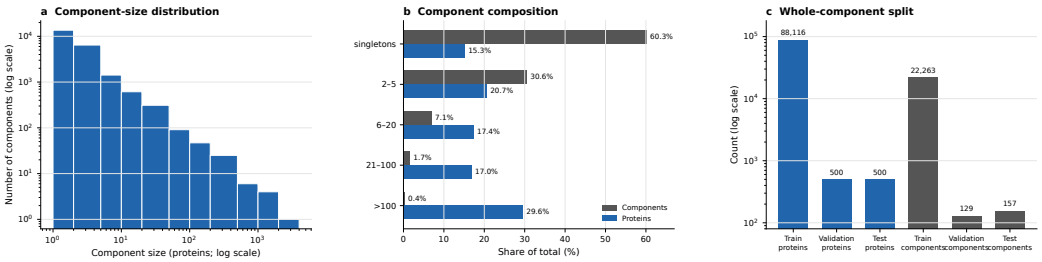

Figure 5: Similarity-graph and split diagnostics. Panels report the component-size distribution, cluster and protein mass across size bins, and the whole-component train/validation/test allocation. The final graph contains 22,549 components, including 13,592 singletons; the largest component contains 3,182 proteins.

Table 7: Component-size distribution and protein coverage.

Component size	Components	Components (%)	Proteins	Proteins (%)
1	13,592	60.28	13,592	15.25
2–5	6,905	30.62	18,448	20.70
6–20	1,594	7.07	15,540	17.44
21–100	377	1.67	15,120	16.97
> 100	81	0.36	26,416	29.64

A.5 Formal GO Relevance Definition

Let

$$\mathcal{A} = \{\mathrm{MF}, \mathrm{BP}, \mathrm{CC}\}$$

denote the three GO aspects, and let T_i^a denote the set of GO terms directly assigned to protein i in aspect $a \in \mathcal{A}$.

For a GO term t in aspect a, define its ancestor closure as

$$\mathrm{Anc}_a^*(t) = \{t\} \cup \mathrm{Anc}_a(t), \tag{15}$$

where $\mathrm{Anc}_a(t)$ is the set of ancestors reachable through `is_a` edges. The propagated annotation set is

$$\bar{T}_i^a = \bigcup_{u \in T_i^a} \mathrm{Anc}_a^*(u). \tag{16}$$

Only `is_a` edges are used. Because these relations do not cross GO aspects, $\bar{T}_i^a$ contains only terms from aspect a.

For a term t belonging to aspect a, its propagated corpus frequency is

$$f(t) = \left|\left\{i \in \mathcal{P} : t \in \bar{T}_i^a\right\}\right|. \tag{17}$$

Its information content is

$$\mathrm{IC}(t) = -\log_{10}\left(\frac{f(t)}{N}\right), \qquad N = |\mathcal{P}| = 89{,}116. \tag{18}$$

Illustrative term frequencies and information-content values are given in Table 8.

For each aspect, pairwise functional similarity is the maximum information content among the shared propagated terms:

$$S_{\mathrm{GO}}^a(i,j) = \max_{t \in \bar{T}_i^a \cap \bar{T}_j^a} \mathrm{IC}(t). \tag{19}$$

We set $S_{\mathrm{GO}}^a(i,j) = 0$ when the intersection is empty. The aspect-specific relevance label is

$$y_{ij}^a = \mathbb{1}[S_{\mathrm{GO}}^a(i,j) \geq \lambda], \qquad \lambda = 3. \tag{20}$$

Table 8: Randomly sampled GO terms from each aspect and information-content bin.

Aspect	IC bin	GO identifier	Term	$f(t)$	IC
MF	[0,1)	GO:0016740	transferase activity	9,238	0.984
MF	[1,2)	GO:0046872	metal ion binding	2,622	1.531
MF	[2,3)	GO:0051087	protein-folding chaperone binding	199	2.651
MF	[3,4)	GO:0017176	phosphatidylinositol N-acetylglucosaminyltransferase activity	10	3.950
MF	[4,max]	GO:0010354	homogentisate prenyltransferase activity	4	4.348
BP	[0,1)	GO:0048519	negative regulation of biological process	10,993	0.909
BP	[1,2)	GO:0051174	regulation of phosphorus metabolic process	1,061	1.924
BP	[2,3)	GO:0022600	digestive system process	125	2.853
BP	[3,4)	GO:0016445	somatic diversification of immunoglobulins	49	3.260
BP	[4,max]	GO:1901551	negative regulation of endothelial cell development	3	4.473
CC	[0,1)	GO:0016020	membrane	20,175	0.645
CC	[1,2)	GO:0005576	extracellular region	7,626	1.068
CC	[2,3)	GO:0005856	cytoskeleton	872	2.009
CC	[3,4)	GO:0071203	WASH complex	28	3.503
CC	[4,max]	GO:0043509	activin A complex	1	4.950

The combined similarity and relevance label are

$$S^{\text{all}}_{\text{GO}}(i,j) = \max_{a \in \mathcal{A}} S^{a}_{\text{GO}}(i,j), \qquad y^{\text{all}}_{ij} = \mathbb{1}\left[S^{\text{all}}_{\text{GO}}(i,j) \geq \lambda\right]. \tag{21}$$

Equivalently, a pair is relevant under ALL if it is relevant under at least one aspect. The aspect-specific labels are non-exclusive. A negative label means that the available annotations do not establish sufficiently specific shared GO evidence; it should not be interpreted as proof that the two proteins have different functions.

A.6 Formal Retrieval Relations

Let

$$a \in \{\text{MF}, \text{BP}, \text{CC}, \text{all}\}$$

index the four relevance criteria. The training relation contains unordered relevant pairs from distinct training components:

$$\mathcal{R}^{a}_{\text{train}} = \left\{(i,j) : i,j \in \mathcal{P}_{\text{train}},\ i < j,\ C(i) \neq C(j),\ y^{a}_{ij} = 1\right\}. \tag{22}$$

The validation query set and candidate gallery are

$$\mathcal{Q}_{\text{val}} = \mathcal{P}_{\text{val}}, \qquad \mathcal{C}_{\text{val}} = \mathcal{P}_{\text{train}}, \tag{23}$$

and the corresponding relevant relation is

$$\mathcal{R}^{a}_{\text{val}} = \left\{(q,c) : q \in \mathcal{Q}_{\text{val}},\ c \in \mathcal{C}_{\text{val}},\ y^{a}_{qc} = 1\right\}. \tag{24}$$

The test query set and candidate gallery are

$$\mathcal{Q}_{\text{test}} = \mathcal{P}_{\text{test}}, \qquad \mathcal{C}_{\text{test}} = \mathcal{P}_{\text{train}} \cup \mathcal{P}_{\text{val}}, \tag{25}$$

and the test relation is

$$\mathcal{R}^{a}_{\text{test}} = \left\{(q,c) : q \in \mathcal{Q}_{\text{test}},\ c \in \mathcal{C}_{\text{test}},\ y^{a}_{qc} = 1\right\}. \tag{26}$$

Component-level splitting implies that every validation or test query–candidate pair satisfies

$$\forall \tau \in \{\text{val}, \text{test}\}, \quad \forall (q,c) \in \mathcal{Q}_{\tau} \times \mathcal{C}_{\tau}, \qquad m_{\text{seq}}(q,c) = m_{\text{str}}(q,c) = 0. \tag{27}$$

No explicit negative relations are stored. For a query q and relevance criterion a, the relevant candidate set is

$$\mathcal{R}_q^a = \left\{ c \in \mathcal{C}_\tau : y_{qc}^a = 1 \right\}. \tag{28}$$

All other candidates are treated as non-relevant under criterion a during evaluation. Queries for which $\mathcal{R}_q^a = \varnothing$ are excluded from metrics for that criterion.

The validation gallery contains $88{,}116$ candidates, and the test gallery contains $88{,}616$ candidates. The combined ALL relations contain $3{,}517{,}524$ training pairs, $38{,}832$ validation pairs, and $38{,}866$ test pairs, for a total of $3{,}595{,}222$ unique relevant pair records.

B Additional Evaluation Protocol Details

B.1 Method-Specific Scoring Functions

For an evaluation stage, let $(\mathcal{Q}, \mathcal{C})$ denote the corresponding query set and candidate collection defined in the benchmark-construction appendix. Each method assigns a score $s_m(q, c)$ to a query–candidate pair. Candidates are sorted by decreasing score, and all score ties are resolved using a fixed, method-independent candidate order.

Pairwise sequence and structure search. For a pairwise search method m, let $\mathcal{H}_m(q) \subseteq \mathcal{C}$ denote the candidates returned for query q. The score is

$$s_m(q,c) = \begin{cases} \mathrm{pident}_{\max}(q,c), & m = \mathrm{MMseqs2}, \quad c \in \mathcal{H}_m(q), \\ \mathrm{TM}_{\max}(q,c), & m = \mathrm{Foldseek}, \quad c \in \mathcal{H}_m(q), \\ -\infty, & c \notin \mathcal{H}_m(q). \end{cases} \tag{29}$$

Here, $\mathrm{pident}_{\max}(q,c)$ and $\mathrm{TM}_{\max}(q,c)$ denote the maximum sequence identity and TM-score, respectively, when multiple alignments are returned for the same pair. Returned candidates are ranked by decreasing score. Candidates not returned by the search method form a deterministic accession-ordered tail, ensuring that every query has a complete ranking over the same gallery.

Protein-embedding retrieval. Let

$$z_i^m = f_m(x_i, X_i) \tag{30}$$

denote the mean-pooled representation produced by method m for protein i, where x_i is the amino-acid sequence and X_i denotes an optional structural representation, such as Foldseek 3Di states. The native retrieval score is cosine similarity:

$$s_m(q,c) = \frac{(z_q^m)^\top z_c^m}{\|z_q^m\|_2 \|z_c^m\|_2}. \tag{31}$$

For ProtST, ProTrek, and ProteinDT, z_i^m is obtained only from the protein branch of the pretrained model. For the evaluated ProtST, ProTrek, and ProteinDT configurations, only the sequence-input protein branch is used at retrieval time, without paired text or a separate structure input. SaProt additionally receives structure-derived 3Di tokens during inference.

Text-mediated retrieval. Each text-mediated method m provides a description d_i^m for protein i, obtained through generation, sequence annotation, or database lookup. A common text encoder e_ψ produces

$$h_i^m = e_\psi(d_i^m), \tag{32}$$

and query–candidate pairs are scored by

$$s_m(q,c) = \frac{(h_q^m)^\top h_c^m}{\|h_q^m\|_2 \|h_c^m\|_2}. \tag{33}$$

For the deterministic generation pipelines, decoding is capped at `max_new_tokens=256`. Generated descriptions are subsequently tokenized by `all-MiniLM-L6-v2` and truncated to its maximum input length of 256 word pieces. These limits use different tokenizers.

The generation settings are source-specific. Evolla uses the chat prompt with system message `You are an AI expert that can answer any questions about protein.` and user message `What is the function of this protein?`. Its input contains the amino-acid sequence and the Foldseek 3Di sequence. The saved Evolla processor configuration truncates each protein stream to `protein_max_length=1024` residues (the underlying tokenizer uses 1,026 positions including special tokens) and truncates the chat prompt to `text_max_length=512`; before tokenization, the generation wrapper truncates a longer 3Di stream to the amino-acid length or pads it with the literal hash character. Evolla generation uses `do_sample=False` and `max_new_tokens=256`. Prot2Text-V2 uses the system message `You are a scientific assistant specialized in protein function predictions. Given the sequence embeddings and other information of a protein, describe its function clearly and concisely in professional language.` and a user message consisting of `Sequence embeddings:` followed by one placeholder token per ESM token. Its wrapper keeps the first 1,024 amino-acid residues (plus the ESM special tokens), and uses `num_beams=1`, `do_sample=False`, and `max_new_tokens=256`.

InstructBioMol uses the instruction `What is the function of this protein?`; its ESM-2 and SaProt streams are tokenized with truncation at 512 tokens, while the Llama input is truncated at the configured 450-token length. Generation uses greedy decoding with `num_beams=1`, `do_sample=False`, and `max_new_tokens=256`.

Descriptions from all sources are embedded with the same `all-MiniLM-L6-v2` text encoder (Sentence-Transformers, 2021; Wang et al., 2020; Reimers & Gurevych, 2019). The InterProScan description is assembled from sequence-derived annotation outputs. The Swiss-Prot reference concatenates the recommended protein name, function description, and keywords. GO identifiers, mapped GO terms, and other fields used directly to construct benchmark relevance labels are excluded from both annotation-derived text representations.

B.2 Frozen-Backbone Supervised Adaptation Workflow

For each reported backbone, we run the same fixed-backbone procedure independently. The ALL training relation contains $3{,}517{,}524$ stored, unordered cross-component relations. In each epoch, a seeded sampler repeatedly draws an eligible training protein uniformly and then draws one of its relevant partners uniformly. This is sampling with replacement across batches and epochs: draws are fresh each epoch and may repeat across batches; within a batch, a rejected draw is resampled until every query and candidate protein is distinct. The stored relation count and the number of sampled training instances are separate but numerically similar quantities.

For every reported backbone, we train for 20 epochs with batch size 256 and $3{,}500{,}000$ sampled pair instances per epoch. Here, batch size counts positive pair instances rather than individual proteins: each full mini-batch contains 256 query–candidate pairs and, because the sampler enforces endpoint uniqueness, 512 distinct protein endpoints. Its query–candidate score matrix is therefore 256×256; each query has one diagonal positive and 255 off-diagonal in-batch negatives, and the symmetric candidate-to-query term has the same count. Across 20 epochs this is $70{,}000{,}000$ sampled pair instances. The same projection head is applied to both elements of each pair. It maps the backbone-specific pooled width d_m to 512 dimensions, applies GELU and layer normalization, and then applies a linear mapping to 128 dimensions followed by L_2 normalization. For a mini-batch $\mathcal{B} = \{(q_i, c_i)\}_{i=1}^{N}$, with $N = 256$ for a full batch, a backbone embedding z_v^m gives

$$\bar{z}_v^m = \frac{p_\omega(z_v^m)}{\|p_\omega(z_v^m)\|_2}, \qquad \ell_{ij} = \frac{(\bar{z}_{q_i}^m)^\top \bar{z}_{c_j}^m}{\tau}. \tag{34}$$

The implementation parameterizes the inverse temperature as $s = \exp(\texttt{log_temp}) = 1/\tau$, so the same logits are computed by multiplying the dot products by s. The projection head and logit-scale parameter minimize symmetric InfoNCE,

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N} \sum_{i=1}^{N} \left[\log \frac{\exp(\ell_{ii})}{\sum_{j=1}^{N} \exp(\ell_{ij})} + \log \frac{\exp(\ell_{ii})}{\sum_{j=1}^{N} \exp(\ell_{ji})} \right]. \tag{35}$$

The endpoint-uniqueness constraint applies only to protein identities and does not verify whether off-diagonal query–candidate pairs are relevant under the ALL relation. We do not apply multi-

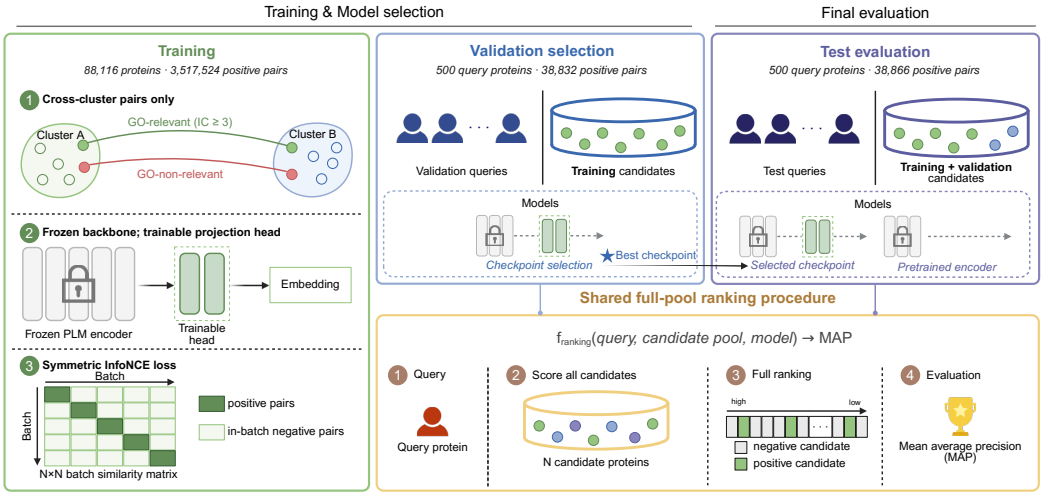

Figure 6: Supervised adaptation, checkpoint selection, and final evaluation. The 3,517,524 stored ALL training relations form the source pool for 3,500,000 freshly sampled positive pair instances per epoch. These pairs train a query–candidate-shared projection head and temperature above a frozen protein encoder. After each epoch, checkpoints are evaluated against the same fixed validation collection of 88,116 training proteins. The checkpoint with the highest validation ALL-MAP is selected and applied once to the fixed test retrieval task without retraining on validation proteins.

positive masking: every off-diagonal pair is treated as an in-batch negative, including any additional relevant pair that may occur off diagonal. Consequently, such pairs may act as false negatives. No explicit negative set is stored.

Backbone parameters remain frozen and receive no gradients. We use the precomputed pooled representations for the reported final runs. The adapted ESM-2, ESM-C, and SaProt runs use the implemented masked mean over all non-padding tokens, including special tokens. AdamW (Loshchilov & Hutter, 2019) uses a head learning rate of 10^{-3} and weight decay 10^{-4}; the stored logit-scale parameter, whose exponential is the inverse temperature, uses a learning rate of 10^{-4}. It is initialized so that the equivalent temperature is $\tau = 0.07$. The cosine schedule is stepped after every optimizer update over all 20 epochs, and gradients are clipped to norm 1.0. The sampler uses seed 42 plus the zero-based epoch index. After every epoch, the model ranks validation queries against the training collection; the checkpoint with the highest validation ALL-MAP is selected and evaluated once on held-out test queries against the fixed train-plus-validation collection.

Using the normalized projected representations from Eq. (34), the adapted retrieval score is

$$s_m^{\text{adapt}}(q, c) = (\bar{z}_q^m)^\top \bar{z}_c^m. \tag{36}$$

Because the representations are normalized, this score is equivalent to cosine similarity in the learned projection space. The temperature is used during contrastive optimization but is not required to order candidates during retrieval.

B.3 Formal Metric Definitions

Fix an evaluation stage, retrieval method, and relevance criterion, and omit the corresponding superscripts. Let $c_k = \pi_q(k)$ denote the candidate at rank k, and let $\mathcal{R}_q$ denote the relevant candidate set for query q. The set of queries with at least one relevant collection candidate is

$$\mathcal{Q}' = \{q \in \mathcal{Q} : |\mathcal{R}_q| > 0\}. \tag{37}$$

Mean average precision. Precision at rank k, average precision for query q, and mean average precision are defined as

$$\begin{aligned} \mathrm{P}_q@k &= \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}\left[c_j \in \mathcal{R}_q\right], \\ \mathrm{AP}(q) &= \frac{1}{|\mathcal{R}_q|}\sum_{k=1}^{|\mathcal{C}|} \mathrm{P}_q@k\, \mathbb{1}\left[c_k \in \mathcal{R}_q\right], \\ \mathrm{MAP} &= \frac{1}{|\mathcal{Q}'|}\sum_{q\in\mathcal{Q}'} \mathrm{AP}(q). \end{aligned} \tag{38}$$

Recall at K. Recall@K is macro-averaged across eligible queries:

$$\mathrm{Recall@}K = \frac{1}{|\mathcal{Q}'|}\sum_{q\in\mathcal{Q}'} \frac{|\mathcal{R}_q \cap \{c_1,\ldots,c_K\}|}{|\mathcal{R}_q|}, \qquad K \in \{1,5,10,100\}. \tag{39}$$

AUROC. Let $\mathrm{AUROC}(q)$ denote the standard binary AUROC computed from the scores and relevance labels of all candidates in $\mathcal{C}$ for query q. Define

$$\mathcal{Q}'' = \{q \in \mathcal{Q}' : |\mathcal{C} \setminus \mathcal{R}_q| > 0\}. \tag{40}$$

The reported query-level macro-AUROC is

$$\mathrm{AUROC} = \frac{1}{|\mathcal{Q}''|}\sum_{q\in\mathcal{Q}''} \mathrm{AUROC}(q). \tag{41}$$

All metrics are computed independently for MF, BP, CC, and ALL. In particular, the ALL results use the combined relevance labels directly and are not averages of the three aspect-specific metric values.

C Text-Embedding Encoder Choice and Sensitivity

For the primary comparison of text-mediated methods, we embed generated and annotation-derived descriptions with the same frozen `all-MiniLM-L6-v2` encoder (Sentence-Transformers, 2021; Wang et al., 2020; Reimers & Gurevych, 2019) and rank proteins by cosine similarity (Appendix B.1). This fixed downstream encoder reduces variation attributable to the embedding step when comparing descriptions from Evolla, Prot2Text-V2, InstructBioMol, InterProScan, and the Swiss-Prot reference. It is a common evaluation setting, not a claim that `all-MiniLM-L6-v2` is the optimal encoder for every text source.

We assessed sensitivity to this choice using four text encoders applied to the same Evolla-10B-DPO description artifact. The four runs use the same test queries and candidate collection; only the downstream text-embedding configuration changes. Table 9 reports their overall (ALL) retrieval scores. The complete MF, BP, CC, and ALL results appear in Table 10.

Table 9: Text-encoder sensitivity for the same Evolla-10B-DPO descriptions on the overall test task. Parameter counts refer to the frozen downstream text encoder; the Evolla-10B-DPO generator is fixed across rows. Counts are rounded nominal values. The all-MiniLM configuration is used in the primary cross-method comparison; the others are sensitivity runs.

Text encoder	Parameters	MAP	AUROC	R@100
all-MiniLM-L6-v2 (main)	22.7M	0.0765	0.8018	0.1313
Jasper-Token-Compression-600M	0.6B	0.0810	0.7931	0.1308
QZhou-Embedding	7B	0.0679	0.7361	0.1169
STS-multilingual-mpnet-base-v2	0.3B	0.0535	0.7170	0.0989

Jasper yields the highest overall MAP in this four-head comparison, whereas all-MiniLM gives the highest AUROC and R@100. All four Evolla configurations exceed the best contrastively adapted protein-only encoder's overall MAP (0.0465), but the magnitude depends on the text head.

Table 10: Aspect-specific text-encoder sensitivity for the same Evolla-10B-DPO descriptions. Each entry reports MAP, AUROC, and R@100 for the indicated GO aspect. Parameter counts are rounded nominal counts for the frozen downstream text encoder; the generator and evaluation collection are held fixed across rows.

Text encoder	Parameters	MF			BP			CC		
		MAP	AUROC	R@100	MAP	AUROC	R@100	MAP	AUROC	R@100
all-MiniLM-L6-v2 (main)	22.7M	0.1233	0.8645	0.2055	0.0405	0.7671	0.0973	0.0859	0.8069	0.1762
Jasper-Token-Compression-600M (infgrad, 2025; Zhang et al., 2025)	0.6B	0.1267	0.8534	0.2129	0.0424	0.7600	0.0947	0.0794	0.8009	0.1679
QZhou-Embedding (Kingsoft-LLM, 2025; Yu et al., 2025)	7B	0.1116	0.7934	0.1896	0.0321	0.7016	0.0806	0.0671	0.7540	0.1534
STS-multilingual-mpnet-base-v2 (Gameselo, 2024; Li & Li, 2024)	0.3B	0.0913	0.7930	0.1800	0.0220	0.6798	0.0554	0.0684	0.7333	0.1427

D Additional Results by GO Aspect

The advantage of the strongest text-mediated methods in MAP is present in all three GO aspects, although the magnitude and the ordering of other method families vary (Figure 7). The annotation-rich Swiss-Prot reference obtains MF/BP/CC MAP values of 0.1799, 0.0673, and 0.1438. Among the learned methods in the primary comparison (Table 11), Evolla with the controlled `all-MiniLM-L6-v2` text head (Sentence-Transformers, 2021; Wang et al., 2020; Reimers & Gurevych, 2019) has the highest MAP in each aspect, at 0.1233, 0.0405, and 0.0859, respectively.

AUROC yields a different view. ProtST-ESM2 reaches 0.8656, 0.7928, and 0.8456 for MF, BP, and CC, while adapted SaProt-650M reaches 0.8139 in BP and adapted ESMC-600M reaches 0.8615 in CC. Their corresponding MAP values remain below Evolla. Alignment methods are also aspect dependent: MMseqs2 MAP is 0.0608/0.0095/0.0162 across MF/BP/CC, whereas Foldseek obtains 0.0600/0.0266/0.0259. Thus, the best text-mediated MAP result is consistent across aspects, but a single complete family ordering is not.

The standard complete-coverage effective-query counts are 190, 270, and 74 for MF, BP, and CC. Differences between aspects should therefore not be read as intrinsic task difficulty: prevalence, ontology structure, annotation density, and the macro-average denominator all differ. Full numerical results are reported in Table 11; paired adaptation deltas are reported in Table 12.

E Retrieval-Depth Results

Recall increases with retrieval depth by definition, but the relative ordering of text-mediated methods changes (Figure 8). Swiss-Prot Recall@1/5/10/100 is 0.0118/0.0424/0.0674/0.1715. Evolla leads the selected learned methods at the first three depths (0.0082/0.0285/0.0456), whereas at depth 100 InstructBioMol (0.1359) and Prot2Text-V2 (0.1351) are slightly above Evolla (0.1313). This pattern shows that performance at a single cutoff does not determine the entire retrieval profile.

ProtST-ESM2 provides a complementary contrast: its overall AUROC of 0.8165 exceeds Evolla's 0.8018, but its MAP and Recall@100 are lower (0.0571 and 0.0975 versus 0.0765 and 0.1313).

E.1 Query-Level Adaptation Audit

The query-level adaptation audit uses the same fixed test split and the ALL relevance relation as the headline results. The test split contains 500 nominal query proteins, of which 364 have at least one relevant candidate and therefore contribute to the complete-coverage ALL metrics.

For each query q and backbone m, we define the per-query AP change as

$$\Delta AP_{m,q} = AP_{m,q}^{\text{adapted}} - AP_{m,q}^{\text{zero-shot}}. \tag{42}$$

Positive values indicate that adaptation improves that query's average precision, and negative values indicate a regression. Panel A of Table 13 groups each common query by the number of the eight backbones for which $\Delta AP_{m,q} > 0$; it measures whether a gain recurs across backbones for the same query. The first row of panel B measures how frequently a positive AP change occurs. Its second row measures gain concentration by sorting positive ΔAP values and computing the share contributed by the largest 10% of them. The audit shows that positive changes are common but not uniform, and that the largest changes are concentrated in a small subset of query–backbone pairs.

Table 11: Final-run results for the retained in-scope configurations under joint sequence and structure similarity controls, grouped by GO aspect. MAP is the primary metric and is listed first. Values are macro-averaged over effective queries; all metrics are higher-is-better. Bold denotes the best non-reference score in each column, and underlining denotes the best score within each method family. Shaded rows are annotation-based references (InterProScan and Swiss-Prot) and are excluded from highlighting.

Family	Method	Setting	Input	MF						BP						CC						Overall					
				MAP	AUROC	R@1	R@5	R@10	R@100	MAP	AUROC	R@1	R@5	R@10	R@100	MAP	AUROC	R@1	R@5	R@10	R@100	MAP	AUROC	R@1	R@5	R@10	R@100
Control	random	–	–	0.0006	0.507	0.000	0.000	0.000	0.001	0.0015	0.502	0.000	0.000	0.000	0.001	0.0009	0.506	0.000	0.000	0.000	0.003	0.0015	0.505	0.000	0.000	0.000	0.001
Alignment	MMseqs2	–	seq	0.0608	0.593	0.008	0.035	0.064	0.096	0.0095	0.514	0.002	0.007	0.010	0.020	0.0162	0.557	0.005	0.012	0.020	0.032	0.0206	0.546	0.002	0.012	0.023	0.037
	Foldseek	–	struct	0.0600	0.616	0.010	0.040	0.062	0.124	0.0266	0.533	0.004	0.016	0.027	0.058	0.0259	0.564	0.008	0.023	0.032	0.047	0.0342	0.564	0.003	0.020	0.037	0.068
Sequence	ESM2-8M	ZS	seq	0.0171	0.643	0.002	0.007	0.012	0.076	0.0164	0.612	0.003	0.010	0.014	0.043	0.0165	0.657	0.004	0.012	0.015	0.044	0.0203	0.617	0.003	0.010	0.016	0.053
	ESM2-35M	ZS	seq	0.0275	0.644	0.009	0.021	0.027	0.073	0.0141	0.655	0.003	0.009	0.011	0.038	0.0309	0.672	0.009	0.019	0.022	0.054	0.0185	0.642	0.004	0.009	0.014	0.048
	ESM2-150M	ZS	seq	0.0173	0.655	0.002	0.009	0.016	0.059	0.0076	0.641	0.001	0.005	0.007	0.018	0.0307	0.671	0.009	0.019	0.024	0.060	0.0146	0.640	0.002	0.006	0.010	0.036
	ESM2-650M	ZS	seq	0.0136	0.631	0.005	0.008	0.010	0.041	0.0086	0.616	0.002	0.006	0.007	0.020	0.0245	0.637	0.007	0.017	0.024	0.055	0.0122	0.616	0.003	0.006	0.008	0.030
	ESMC-300M	ZS	seq	0.0070	0.660	0.001	0.005	0.007	0.033	0.0070	0.633	0.002	0.004	0.005	0.015	0.0160	0.638	0.009	0.014	0.017	0.035	0.0080	0.641	0.002	0.004	0.005	0.023
	ESMC-600M	ZS	seq	0.0066	0.666	0.001	0.003	0.006	0.033	0.0066	0.635	0.001	0.003	0.005	0.018	0.0136	0.629	0.004	0.014	0.017	0.034	0.0074	0.645	0.001	0.003	0.005	0.023
	ESM2-8M-FT	FT	seq	0.0332	0.800	0.002	0.009	0.014	0.091	0.0235	0.736	0.003	0.012	0.017	0.054	0.0327	0.790	0.004	0.010	0.017	0.080	0.0327	0.763	0.003	0.012	0.017	0.063
	ESM2-35M-FT	FT	seq	0.0578	0.812	0.009	0.025	0.033	0.096	0.0306	0.762	0.002	0.009	0.015	0.055	0.0660	0.811	0.006	0.020	0.033	0.113	0.0465	0.779	0.004	0.012	0.019	0.073
	ESM2-150M-FT	FT	seq	0.0540	0.844	0.007	0.018	0.028	0.103	0.0279	0.791	0.003	0.008	0.012	0.065	0.0781	0.848	0.011	0.026	0.041	0.141	0.0425	0.810	0.003	0.009	0.014	0.077
	ESM2-650M-FT	FT	seq	0.0411	0.834	0.002	0.015	0.026	0.086	0.0303	0.812	0.003	0.009	0.016	0.064	0.0796	0.856	0.010	0.027	0.043	0.147	0.0431	**0.824**	0.002	0.010	0.018	0.072
	ESMC-300M-FT	FT	seq	0.0287	0.831	0.002	0.009	0.013	0.064	0.0233	0.798	0.002	0.008	0.014	0.056	0.0712	0.857	0.014	0.031	0.042	0.142	0.0334	0.815	0.002	0.009	0.015	0.060
	ESMC-600M-FT	FT	seq	0.0320	0.814	0.002	0.010	0.013	0.065	0.0249	0.806	0.002	0.006	0.010	0.053	0.0551	**0.861**	0.009	0.023	0.033	0.124	0.0322	0.818	0.002	0.007	0.011	0.059
Sequence+ structure	SaProt-35M	ZS	seq+3Di	0.0128	0.693	0.003	0.009	0.011	0.037	0.0074	0.654	0.002	0.003	0.004	0.018	0.0227	0.648	0.007	0.013	0.019	0.053	0.0118	0.662	0.002	0.005	0.008	0.024
	SaProt-650M	ZS	seq+3Di	0.0217	0.710	0.005	0.013	0.020	0.054	0.0091	0.669	0.002	0.005	0.007	0.022	0.0348	0.656	0.010	0.022	0.029	0.070	0.0177	0.678	0.003	0.008	0.012	0.038
	SaProt-35M-FT	FT	seq+3Di	0.0430	0.799	0.004	0.016	0.029	0.078	0.0201	0.763	0.001	0.005	0.009	0.045	0.0438	0.835	0.004	0.009	0.019	0.088	0.0328	0.776	0.002	0.007	0.013	0.056
	SaProt-650M-FT	FT	seq+3Di	0.0301	0.812	0.005	0.013	0.017	0.064	0.0231	**0.814**	0.002	0.005	0.008	0.044	0.0736	0.858	0.007	0.019	0.028	0.139	0.0359	0.813	0.002	0.007	0.012	0.057
Protein–text CLIP	ProTrek-35M	ZS	seq	0.0553	0.766	0.012	0.033	0.048	0.108	0.0151	0.717	0.003	0.009	0.013	0.035	0.0432	0.735	0.010	0.028	0.041	0.097	0.0331	0.732	0.005	0.016	0.026	0.064
	ProTrek-650M	ZS	seq	0.0785	0.799	0.009	0.036	0.063	0.125	0.0263	0.764	0.005	0.016	0.020	0.053	0.0580	0.836	0.015	0.039	0.053	0.121	0.0512	0.778	0.007	0.023	0.037	0.080
	ProtST-esm1b	ZS	seq	0.0753	0.832	0.013	0.039	0.055	0.144	0.0322	0.789	0.006	0.015	0.024	0.083	0.0589	0.829	0.019	0.034	0.054	0.124	0.0483	0.803	0.006	0.019	0.032	0.101
	ProtST-esm2	ZS	seq	0.0829	**0.866**	0.011	0.041	0.066	0.140	0.0323	0.793	**0.006**	**0.017**	0.026	0.069	0.0739	0.846	0.018	**0.042**	**0.063**	0.143	0.0571	0.816	0.006	0.024	0.041	0.097
	ProtST-protbert	ZS	seq	0.0459	0.806	0.010	0.026	0.035	0.103	0.0200	0.761	0.005	0.012	0.015	0.049	0.0633	0.804	**0.019**	0.040	0.049	0.121	0.0297	0.770	0.005	0.014	0.019	0.068
	ProteinDT-InfoNCE	ZS	seq	0.0381	0.817	0.008	0.014	0.025	0.095	0.0175	0.720	0.004	0.009	0.015	0.056	0.0391	0.796	0.007	0.020	0.029	0.081	0.0246	0.750	0.003	0.009	0.017	0.066
	ProteinDT-EBM-NCE-1	ZS	seq	0.0288	0.798	0.012	0.015	0.018	0.056	0.0132	0.712	0.003	0.006	0.007	0.045	0.0180	0.770	0.004	0.010	0.017	0.049	0.0175	0.736	0.004	0.007	0.009	0.043
	ProteinDT-EBM-NCE-2	ZS	seq	0.0180	0.814	0.002	0.008	0.015	0.047	0.0123	0.703	0.003	0.009	0.012	0.034	0.0233	0.770	0.003	0.011	0.020	0.068	0.0165	0.739	0.002	0.007	0.011	0.037
Function text	Evolla-10B-DPO	ZS	text	**0.1233**	0.865	**0.018**	**0.059**	**0.091**	0.206	**0.0405**	0.767	0.005	0.017	**0.029**	**0.097**	**0.0859**	0.807	0.011	0.034	0.060	**0.176**	**0.0765**	0.802	**0.008**	**0.028**	**0.046**	0.131
	Prot2Text-V2	ZS	text	0.1126	0.809	0.005	0.045	0.075	**0.214**	0.0297	0.720	0.001	0.013	0.023	0.082	0.0541	0.741	0.004	0.016	0.031	0.113	0.0725	0.746	0.003	0.024	0.042	0.135
	InstructBioMol	ZS	text	0.0983	0.824	0.007	0.041	0.066	0.213	0.0311	0.719	0.002	0.014	0.022	0.083	0.0695	0.719	0.011	0.028	0.042	0.139	0.0662	0.745	0.004	0.021	0.036	**0.136**
Annotation	InterProScan	–	text	0.0959	0.809	0.016	0.052	0.070	0.190	0.0372	0.681	0.004	0.020	0.033	0.091	0.0978	0.739	0.019	0.060	0.084	0.165	0.0518	0.707	0.005	0.022	0.036	0.104
	Swiss-Prot	–	text	0.1799	0.861	0.029	0.094	0.132	0.264	0.0673	0.792	0.007	0.028	0.043	0.124	0.1438	0.837	0.026	0.067	0.092	0.228	0.1155	0.816	0.012	0.042	0.067	0.172

This table contains 32 retained final-run configurations. The test collection contains 88,616 proteins. All displayed function-text and annotation-reference runs use `all-MiniLM-L6-v2` for text embedding. Setting: ZS = zero-shot and FT = contrastively adapted. Input: seq = sequence, struct = structure, seq+3Di = sequence with 3Di tokens, and text = function text.

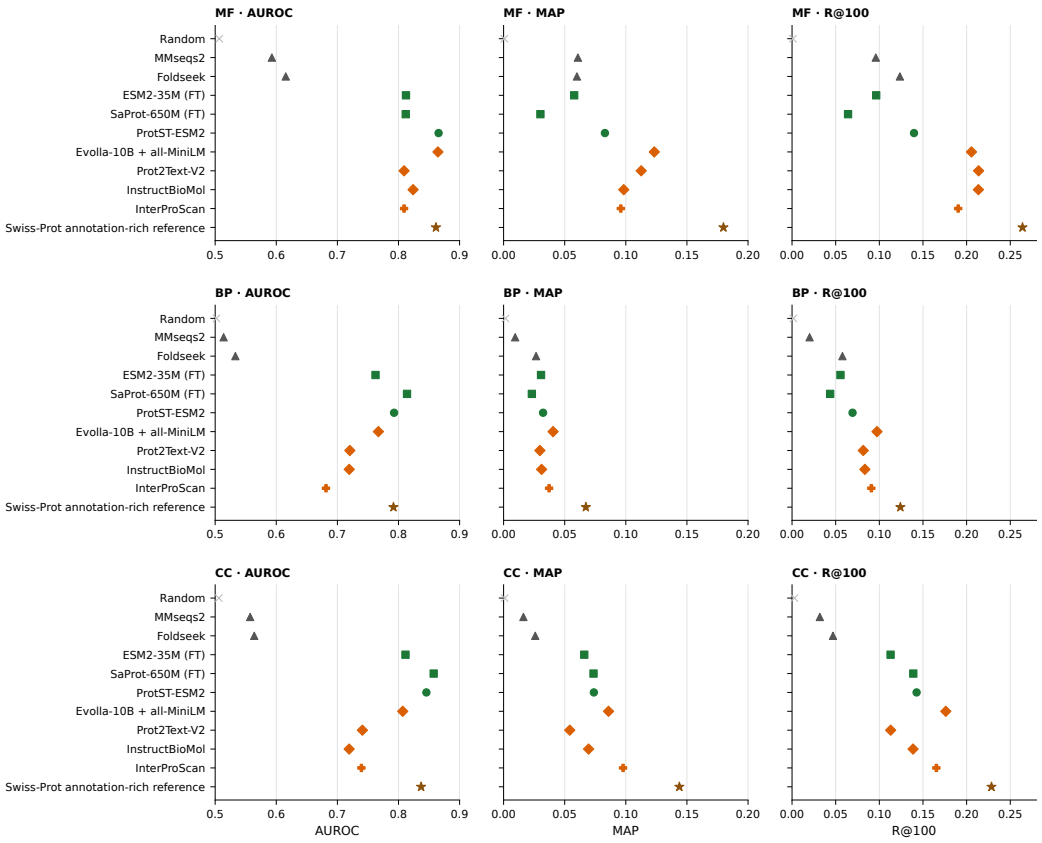

Figure 7: Aspect-specific AUROC, MAP, and Recall@100 for representative methods. Each column shares its metric scale across MF, BP, and CC.

Table 12: Paired changes in MAP, AUROC, and R@100 after supervised contrastive adaptation (adapted minus zero-shot), in percentage points. Positive values denote improvement.

	MF			BP			CC			ALL		
Backbone	ΔMAP	ΔAUROC	ΔR@100	ΔMAP	ΔAUROC	ΔR@100	ΔMAP	ΔAUROC	ΔR@100	ΔMAP	ΔAUROC	ΔR@100
ESM2-8M	+1.61	+15.69	+1.55	+0.71	+12.34	+1.09	+1.62	+13.40	+3.56	+1.24	+14.57	+1.04
ESM2-35M	+3.02	+16.79	+2.36	+1.66	+10.76	+1.75	+3.51	+13.90	+5.89	+2.80	+13.75	+2.55
ESM2-150M	+3.67	+18.93	+4.42	+2.03	+14.93	+4.72	+4.74	+17.73	+8.10	+2.79	+17.01	+4.08
ESM2-650M	+2.75	+20.30	+4.48	+2.17	+19.61	+4.38	+5.51	+21.88	+9.21	+3.09	+20.76	+4.26
SaProt-35M	+3.01	+10.56	+4.11	+1.28	+10.87	+2.75	+2.10	+18.70	+3.52	+2.10	+11.36	+3.19
SaProt-650M	+0.84	+10.22	+1.01	+1.40	+14.52	+2.22	+3.88	+20.17	+6.91	+1.82	+13.45	+1.87
ESMC-300M	+2.17	+17.14	+3.11	+1.63	+16.48	+4.08	+5.52	+21.91	+10.76	+2.54	+17.40	+3.67
ESMC-600M	+2.54	+14.83	+3.17	+1.83	+17.09	+3.55	+4.15	+23.25	+8.98	+2.47	+17.31	+3.52

Table 13: Query-level audit of AP changes after contrastive adaptation. Panel A groups the 364 queries by the number of backbones with positive per-query AP change. Panel B uses the same 2,912 query–backbone pairs; the top-10% share is computed only among positive AP changes.

Panel A: cross-backbone consistency

Positive backbones per query	Queries	Share
0/8	19/364	5.2%
1–4/8	87/364	23.9%
5–7/8	162/364	44.5%
8/8	96/364	26.4%

Panel B: prevalence and concentration of positive AP changes

Statistic	Value	
Positive per-query AP changes	2,009/2,912	69.0%
Share from largest 10% of positive changes	66.8%	

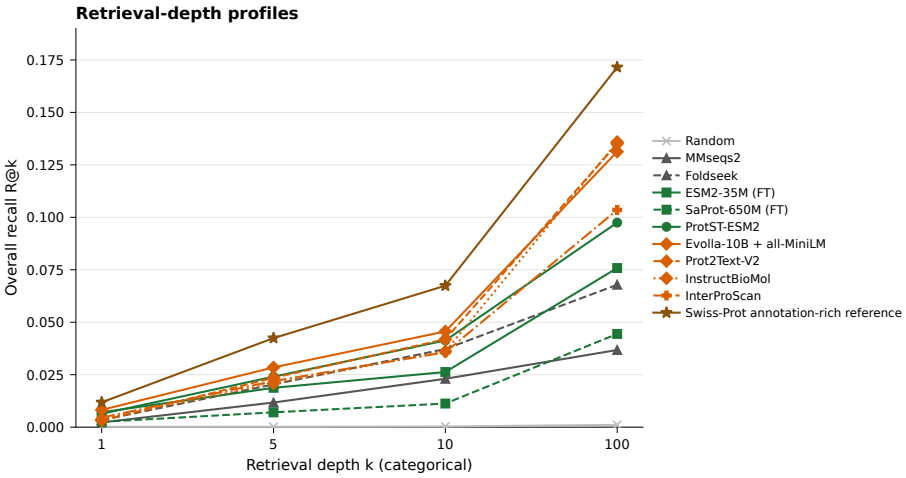

Figure 8: Overall recall at retrieval depths 1, 5, 10, and 100 for the representative method set. The horizontal positions are categorical and connect the four evaluated cutoffs without interpolation. Values are fixed-split point estimates.