PROTOCOL: Late Interaction Retrieval for Protein Homolog Search

Gabrielle Cohn [*1] **Rohan Gumaste** [*1] **Minh Hoang** [*2] **Vihan Lakshman** [*1]

Abstract

Protein homology search underlies function annotation, structure prediction, and evolutionary analysis, but remains challenging in the "twilight zone," where global sequence similarity is weak and classical alignment methods lose sensitivity. Protein language models provide context-aware representations that could improve alignment sensitivity in this regime. However, prior protein embedding-based retrieval pipelines often pool these representations into a single vector, potentially obscuring local motifs, domains, or conserved residues that reveal remote homology. We introduce PROTOCOL, a model which represents proteins as sets of residue embeddings and uses ColBERT-style late interaction to test whether residue-level comparison improves homolog retrieval. PROTOCOL encodes proteins independently, keeps candidate representations precomputable, and scores candidates with MaxSim over residue embeddings. On SCOPe superfamily and Pfam clan benchmarks, PROTOCOL outperforms sequence-composition, alignment-based, pooled PLM, and trained single-vector baselines, supporting late interaction as an effective retrieval layer for remote homology search.

1. Introduction

Homologous proteins descend from a common ancestral sequence and often preserve related functions or structures. Detecting such relationships is central to computational biology, supporting function annotation, structure prediction, and evolutionary analysis. This task is especially difficult for remote homologs, whose sequences may diverge so substantially that direct sequence-level similarity becomes weak. In this "twilight zone," classical sequence-alignment methods can miss relationships that remain evident through conserved motifs, domains, or structural constraints (Altschul et al., 1990; Eddy, 2011; Steinegger & Söding, 2017).

Protein language models (PLMs) offer a promising sequence-only alternative because they produce contextual residue embeddings that encode structural and evolutionary signal (Lin et al., 2023b; Liu et al., 2024). However, many PLM-based retrieval pipelines pool these embeddings into a single protein vector and compare proteins by cosine similarity (Iovino et al., 2024). This is efficient, but can dilute local evidence such as a conserved motif, domain, or small set of structurally constrained residues that may be decisive for remote homology.

This pooling bottleneck motivates our central question: *does homolog retrieval improve when proteins are represented as sets of residue embeddings and compared through late interaction?* We hypothesize that residue-level scoring can preserve local evolutionary evidence while keeping database proteins independently encodable.

We test our hypothesis with PROTOCOL ("proteins" with "ColBERT") (see figure A.1), a late-interaction retrieval model for protein homology search. PROTOCOL adapts the ColBERT retrieval paradigm (Khattab & Zaharia, 2020) to protein sequences by representing each protein as residue-level PLM embeddings and comparing proteins through lightweight residue-level interactions[1].

We evaluate PROTOCOL on SCOPe superfamily and Pfam clan retrieval benchmarks using baselines chosen to isolate sequence composition, alignment sensitivity, PLM scale, contrastive fine-tuning, and late interaction. Across both settings, PROTOCOL performs best, supporting residue-level late interaction as an effective retrieval layer for remote homology search.

2. Related Work

Alignment-based homology search. Classical homology search relies on residue-level sequence comparison. BLAST (Altschul et al., 1990) performs fast pairwise local alignment, while profile-based methods such as PSI-BLAST and HMMER (Eddy, 2011) improve sensitivity using multi-

[*]Equal contribution [1]MIT [2]Princeton University. Correspondence to: Gabrielle Cohn <gcohn@mit.edu>, Rohan Gumaste <rgumaste@mit.edu>, Minh Hoang <minhhoang@princeton.edu>, Vihan Lakshman <vihan@mit.edu>.

Preprint. May 27, 2026.

[1]Our code is available at https://github.com/gabriellecohn/ProtoCol

ple sequence alignments. MMseqs2 (Steinegger & Söding, 2017) further improves the speed–sensitivity tradeoff for large-scale search. These methods remain strong baselines, but their reliance on detectable sequence similarity limits sensitivity for highly diverged homologs.

Structure-based search addresses some of these limitations by comparing proteins closer to their conserved three-dimensional form. Foldseek (van Kempen et al., 2024), for example, discretizes protein structures into residue-level alphabets and aligns the resulting strings. Such methods demonstrate the value of local matching beyond raw sequence identity, but require structural information to work. Our work instead asks whether sequence-only PLM representations can support an analogous retrieval primitive.

Protein language models for retrieval. Large protein language models such as ProtTrans (Elnaggar et al., 2021) and ESM (Lin et al., 2023b) learn contextual residue embeddings from unaligned sequence databases. These embeddings have been used for retrieval via mean-pooling into a single protein-level vector and ranking candidates by cosine similarity (Schütze et al., 2022; Iovino et al., 2024). This bi-encoder-style setup supports efficient nearest-neighbor search, analogous to dense retrieval methods in NLP such as Sentence-BERT (Reimers & Gurevych, 2019) and Dense Passage Retrieval (Karpukhin et al., 2020), but discards the token-level structure of PLM embeddings. Methods such as PLMAlign and pLM-BLAST (Liu et al., 2024; Kaminski et al., 2023) point to the utility of residue-level PLM embeddings for remote homology detection by using PLM representations in alignment-oriented comparison.

Late-interaction retrieval. Late-interaction models occupy a middle ground between efficient global-vector retrieval and expensive pairwise interaction models. ColBERT (Khattab & Zaharia, 2020) represents each text passage as token embeddings and scores a query–document pair by matching each query token to its most similar document token using MaxSim. This preserves fine-grained matching while allowing document representations to be precomputed. Following the introduction of ColBERT, a number of works have further developed this notion of late interaction retrieval (Santhanam et al., 2022b;a; Formal et al., 2024; Faysse et al., 2025; Lee et al., 2023; Lin et al., 2023a; Chaffin & Sourty, 2025; Dhulipala et al., 2024; Engels et al., 2023). All of these prior works focus on traditional information retrieval settings. PROTOCOL adapts this idea to proteins: residues replace tokens, candidate proteins replace documents, and MaxSim provides a residue-level similarity score for homolog retrieval. In contrast to alignment-oriented PLM methods, PROTOCOL uses late interaction as a learned retrieval mechanism rather than computing an explicit alignment path.

3. Methodology

This section defines the components of PROTOCOL: a residue-level PLM encoder, MaxSim scoring, and a contrastive objective based on weak homology labels. Together, these choices test whether homology retrieval benefits from preserving residue embeddings through the scoring layer rather than compressing each protein before comparison. Dataset construction, baselines, and evaluation protocol are described in Section 4.

Late-interaction encoder. We instantiate ColBERT-style late interaction over a protein language model. Let $x = (x_1, \ldots, x_T)$ be a protein sequence, and let $h_t \in \mathbb{R}^H$ denote the contextual residue embedding produced at position t by an ESM-2 backbone f_θ (Lin et al., 2023b). We attach a linear projection $W \in \mathbb{R}^{D \times H}$ followed by L2 normalization:

$$e_t = \frac{W h_t}{\|W h_t\|_2} \in \mathbb{R}^D, \tag{1}$$

with $D = 128$ in all experiments. Unless otherwise stated, the backbone is ESM-2 35M ($H = 480$, 12 layers). To keep the trainable footprint modest, we freeze the embedding layer and lower transformer blocks and fine-tune only the final three transformer layers, the post-stack LayerNorm, and W. This yields roughly 8.4M trainable parameters out of 33.6M.

MaxSim scoring. A protein is represented by its variable-length set of L2-normalized residue embeddings $E = \{e_1, \ldots, e_T\}$. MaxSim operationalizes residue-level retrieval by allowing each query residue to contribute its strongest match anywhere in the candidate protein. Given query embeddings E^q and candidate embeddings E^d, we score the pair using the asymmetric MaxSim operator of Khattab & Zaharia (2020),

$$\mathrm{MaxSim}(E^q, E^d) = \sum_{i=1}^{T_q} \max_{j \in [T_d]} \langle e_i^q, e_j^d \rangle. \tag{2}$$

Because embeddings are L2-normalized, each inner product is a cosine similarity. Padding positions are masked in both the inner maximum and outer sum.

Contrastive training. Training shapes the embedding space so homologs receive high late-interaction scores and in-batch non-homologs receive lower scores. Each training pair consists of an anchor protein a and a positive protein p sampled from the same superfamily for SCOPe (Chandonia et al., 2022) or from the same clan for Pfam (Mistry et al., 2021). For a batch of B pairs, we form $S \in \mathbb{R}^{B \times B}$ with $S_{ij} = \mathrm{MaxSim}(E^{a_i}, E^{p_j})$, treat off-diagonal entries as in-batch negatives, and minimize the symmetric InfoNCE objective

$$\mathcal{L} = \tfrac{1}{2}\left[\mathrm{CE}(S/\tau, y) + \mathrm{CE}(S^\top/\tau, y)\right], \tag{3}$$

where $y_i = i$ and τ is a temperature. We do not filter accidental positive collisions among in-batch negatives.

Implementation details. Sequences are tokenized with the ESM-2 tokenizer and truncated to $T \leq 256$ residues. We optimize with AdamW using weight decay 0.01 for three epochs at batch size 16. The learning rate follows a OneCycleLR schedule with peak learning rate 2×10^{-5} and 10% warmup. Training uses fp16 autocast on a single GPU; gradients are unscaled and clipped to global norm 1.0 before each step. We set $\tau = 1$ throughout.

4. Experiments

4.1. Datasets and Retrieval Protocol

We evaluate homolog retrieval in two complementary settings. SCOPe provides a hierarchical structural classification of protein domains; we use superfamily labels as evidence of shared ancestry among potentially remote homologs (Chandonia et al., 2022). Pfam groups protein sequences into families using sequence alignments and profile HMMs; we use clan labels, which group related families, as a broader test of remote homology retrieval.

For each dataset, we construct train and test splits over evolutionary groups and train a separate PROTOCOL model on that dataset's training split. At evaluation time, each protein is used as a query against the remaining proteins in the corresponding test database, excluding self-matches. Retrieved proteins are relevant if they share the query's held-out evolutionary group: superfamily for SCOPe and Pfam clan for Pfam. Group-disjoint train/test splits ensure that evaluation measures generalization to unseen homologous groups rather than memorization of training labels.

4.2. Compared Methods

We compare against baselines designed to distinguish the contribution of late-interaction scoring from other sources of retrieval signal: sequence composition, alignment sensitivity, PLM scale, contrastive fine-tuning, and pretrained residue similarity without task-specific adaptation. The trained PROTOCOL model is described in Section 3.

MinHash Jaccard. We compute MinHash-approximated Jaccard similarity over amino acid 5-mers. Each sequence is decomposed into overlapping 5-mers, and a MinHash signature with 256 permutations is computed using `datasketch`. Candidates are ranked by the fraction of matching hash values between signatures.

MMseqs2. We evaluate MMseqs2 (Steinegger & Söding, 2017) as a strong alignment-based sequence retrieval baseline. All test sequences are searched against the full test set using `mmseqs easy-search` with sensitivity 7.5. Hits are ranked by e-value in ascending order.

Method	cR@1	cR@10	cR@100	Lat. (ms)
MinHash	0.3172	0.1932	0.1567	**4.23**
MMseqs2	0.8578	0.6177	0.3421	1068.73
ESM-2 650M pool	0.8773	0.6664	0.5127	49.58
ESM-2 35M uni	0.6819	0.6777	0.7141	20.31
PROTOCOL-F	0.8825	0.8093	0.7422	24.33
PROTOCOL	**0.9460**	**0.8947**	**0.8796**	24.10

Table 1. SCOPe superfamily retrieval performance ($n_{\text{test}} = 2314$). Scores are capped recall; latency is in ms. MMseqs2 uses sensitivity 7.5; MinHash uses 5-mers; PROTOCOL-F is frozen.

Mean-pooled ESM-2 650M. To assess the importance of encoder scale, we embed each protein using frozen ESM-2 650M (`facebook/esm2_t33_650M_UR50D`), which is substantially larger than the ESM-2 35M backbone used by PROTOCOL. Final-layer residue embeddings are mean-pooled, L2-normalized, and ranked by cosine similarity.

Uni-vector ESM-2 35M. This is the direct ablation of late interaction. It uses the same ESM-2 35M backbone and contrastive objective as PROTOCOL, but mean-pools residue embeddings into one L2-normalized protein vector and retrieves by cosine similarity. This tests whether gains come from residue-level scoring rather than fine-tuning alone.

Frozen PROTOCOL. To isolate task-specific optimization, we evaluate a frozen PROTOCOL variant with the same ESM-2 35M backbone, 128-dimensional projection, and MaxSim scoring function as the trained model, but with all parameters left at their initial values.

4.3. Evaluation Metric

We evaluate retrieval using the capped recall@k metric,

$$\text{cRecall@}k(q) = \frac{\text{hits@}k(q)}{\min(k, N_q)}, \tag{4}$$

where $\text{hits@}k(q)$ is the number of top-k retrieved proteins that are real homologs of q, and N_q is the number of other proteins in that group. Unlike standard recall@k, capped recall@k normalizes by the maximum number of relevant proteins that can appear in the top k, so a ranking whose top-k entries are all relevant receives a score of 1 regardless of group size (Ji et al., 2025; Chen et al., 2023). We computed capped recall by treating every protein in the test set as a query and retrieving from the remaining set of examples.

4.4. Results

PROTOCOL outperforms other retrieval baselines. Across both benchmarks, trained PROTOCOL achieves the strongest performance at every cutoff. On SCOPe, it improves over the next best baseline by 6.87 points at cR@1, 21.07 at cR@10, and 16.55 at cR@100. These gains are largest at deeper retrieval cutoffs, suggesting that ranking

Method	cR@1	cR@10	cR@100	Lat. (ms)
MinHash	0.4957	0.1684	0.1104	**3.54**
MMseqs2	0.8833	0.3627	0.1539	1050.74
ESM-2 650M pool	0.8903	0.6048	0.4324	49.86
ESM-2 35M uni	0.7897	0.6657	0.6378	19.14
PROTOCOL-F	0.8147	0.5881	0.4533	24.90
PROTOCOL	**0.9377**	**0.7630**	**0.7065**	26.49

Table 2. Pfam clan retrieval performance ($n_{\text{test}} = 3000$). Metrics and parameter settings are identical to Table 1.

many homologs requires evidence beyond a single global representation or sequence-level similarity comparison.

The Uni-vector ESM-2 35M baseline is the key ablation. It uses the same backbone and supervision as PROTOCOL, but removes residue-level scoring. Because this comparison holds the encoder and training signal fixed, PROTOCOL's consistent gains indicate that late interaction adds value beyond contrastive fine-tuning alone.

Frozen PROTOCOL further separates the contribution of architecture from task-specific training. Its competitive SCOPe performance suggests that pretrained ESM-2 residue embeddings already encode useful local similarity structure. However, the substantial gains of the trained model indicate that weak homology supervision sharpens this structure further — as evidenced by the block-diagonal similarity maps in Fig. 1, where the trained model produces coherent high-similarity blocks that align with secondary structure boundaries, suggesting that contrastive fine-tuning encourages the embeddings to organize similarity along structurally meaningful lines rather than purely sequential ones.

The Pfam benchmark shows the same overall pattern, with a sharper distinction between nearest-neighbor and deeper retrieval. MMseqs2 and frozen mean-pooled ESM-2 650M remain competitive at cR@1, but drop more substantially at cR@10 and cR@100. In contrast, trained PROTOCOL improves over trained Uni-vector ESM-2 35M by 14.80, 9.73, and 6.87 points at cR@1, cR@10, and cR@100, respectively. This supports the benefit of residue-level late interaction under a different protein-family taxonomy.

PROTOCOL embeddings reflect structural organization. We investigate whether the retrieval mechanism learned by PROTOCOL is tied to the biological structure of protein sequences. We obtain PDB structure files for all SCOPe domain sequences in the test set and annotate each domain with per-residue secondary structure labels using the pyDSSP package. The DSSP algorithm assigns each residue one of three coarse-grained secondary structure states: α-helix (orange), β-strand (blue), or coil/loop (gray). The annotations for an exemplary query and its top-ranked true positive retrieval are overlaid along both axes of their pairwise residue embedding similarity matrix (Fig. 1).

Transitions between secondary structure elements, particularly between β-strands and coil/loop regions, correspond to visible discontinuities in the similarity map. At these boundaries, similarity drops sharply, producing a block-diagonal appearance that corresponds to shared secondary structure elements. For example, the β-strand-dominated region spanning residues 35-135 on the query and residues 25–130 on the top-hit forms a large, coherent diagonal block. This finding provides evidence that PROTOCOL's learned embeddings implicitly encode structural organization beyond raw sequence identity.

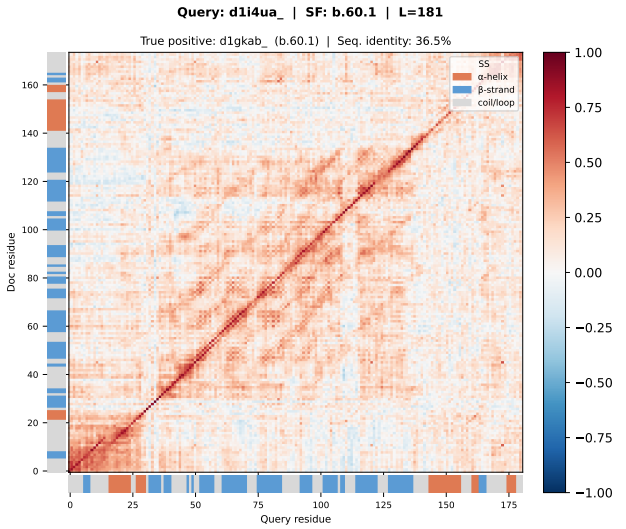

Figure 1. **ColBERT attention maps for true positive pair.** We visualize the residue-level similarity matrix between a representative query and its highest-ranked true positive match. Secondary structure annotations are shown along each axis. The similarity map exhibits block diagonal structure that coincides with secondary structure boundaries, indicating that PROTOCOL indeed learns meaningful structural organization patterns to facilitate retrieval.

5. Conclusion

Taken together, these comparisons show that representing proteins as sets of residue embeddings and comparing them through late interaction improves homolog retrieval. The strongest controlled evidence comes from the trained Uni-vector comparison, which holds the backbone and supervision fixed while replacing residue-level MaxSim scoring with a single global vector. PROTOCOL's consistent gains show that the benefit is not only due to contrastive fine-tuning, but also to preserving and comparing local residue-level evidence. The substantial gains of the trained model over the frozen variant indicate that weak homology supervision sharpens understanding of secondary structure, which is likely responsible for the improved performance.

In future, we hope to build on these preliminary results and, in particular, investigate how to scale PROTOCOL to perform efficient late-interaction search over orders-of-magnitude larger protein databases.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

Chaffin, A. and Sourty, R. Pylate: Flexible training and retrieval for late interaction models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 6334–6339, 2025.

Chandonia, J.-M., Guan, L., Lin, S., Yu, C., Fox, N. K., and Brenner, S. E. SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Research*, 50(D1):D553–D559, 2022.

Chen, C., Yang, C.-W., Lin, C.-Y., and Kao, H.-Y. Breaking boundaries in retrieval systems: Unsupervised domain adaptation with denoise-finetuning. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1630–1642, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.110. URL https://aclanthology.org/2023.findings-emnlp.110/.

Dhulipala, L., Hadian, M., Jayaram, R., Lee, J., and Mirrokni, V. Muvera: Multi-vector retrieval via fixed dimensional encoding. *Advances in Neural Information Processing Systems*, 37:101042–101073, 2024.

Eddy, S. R. Accelerated profile HMM searches. *PLOS Computational Biology*, 7(10):e1002195, 2011.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2021.

Engels, J., Coleman, B., Lakshman, V., and Shrivastava, A. Dessert: an efficient algorithm for vector set search with vector set queries. *Advances in Neural Information Processing Systems*, 36:67972–67992, 2023.

Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., and Colombo, P. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 61424–61449, 2025.

Formal, T., Clinchant, S., Déjean, H., and Lassance, C. Splate: Sparse late interaction retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2635–2640, 2024.

Iovino, B. G., Tang, H., and Ye, Y. Protein domain embeddings for fast and accurate similarity search. *Genome Research*, 34:1434 – 1444, 2024. doi: 10.1101/gr.279127.124.

Ji, X., Glenn, P., Parameswaran, A. G., and Hulsebos, M. Target: Benchmarking table retrieval for generative tasks, 2025. URL https://arxiv.org/abs/2505.11545.

Kaminski, K., Ludwiczak, J., Pawlicki, K., Alva, V., and Dunin-Horkawicz, S. plm-blast: distant homology detection based on direct comparison of sequence representations from protein language models. *Bioinformatics*, 39(10):btad579, 10 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad579. URL https://doi.org/10.1093/bioinformatics/btad579.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6784, 2020.

Khattab, O. and Zaharia, M. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020.

Lee, J., Dai, Z., Duddu, S. M. K., Lei, T., Naim, I., Chang, M.-W., and Zhao, V. Rethinking the role of token retrieval in multi-vector retrieval. *Advances in Neural Information Processing Systems*, 36:15384–15405, 2023.

Lin, W., Chen, J., Mei, J., Coca, A., and Byrne, B. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36:22820–22840, 2023a.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023b.

Liu, W., Wang, Z., You, R., Xie, C., Wei, H., Xiong, Y., Yang, J., and Zhu, S. Plmsearch: Protein language model powers accurate and fast sequence search for remote homology. *Nature Communications*, 15, 2024. doi: 10.1038/s41467-024-46808-5.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., and

Bateman, A. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 01 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa913. URL https://doi.org/10.1093/nar/gkaa913.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 3982–3992, 2019.

Santhanam, K., Khattab, O., Potts, C., and Zaharia, M. Plaid: an efficient engine for late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 1747–1756, 2022a.

Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., and Zaharia, M. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734, 2022b.

Schütze, K., Heinzinger, M., Steinegger, M., and Rost, B. Nearest neighbor search on embeddings rapidly identifies distant protein relations. *Frontiers in Bioinformatics*, Volume 2 - 2022, 2022. ISSN 2673-7647. doi: 10.3389/fbinf.2022.1033775. URL https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2022.1033775.

Steinegger, M. and Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., and Steinegger, M. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2):243–246, 2024.

A. Appendix

A.1. Framework Overview

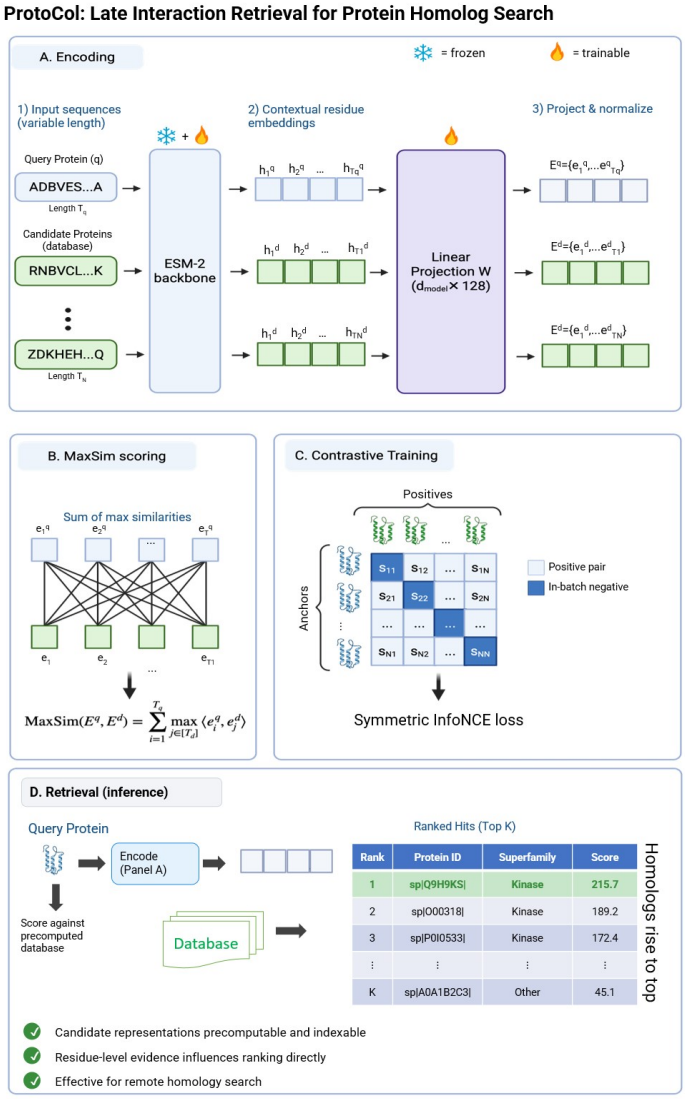

Figure 2. Overview of the ProtoCoL framework for protein homolog retrieval. Variable-length query and candidate protein sequences are encoded with a frozen ESM-2 backbone, projected into residue-level embeddings, and compared using MaxSim scoring. The projection layer is trained with a symmetric contrastive objective, enabling retrieval using precomputed candidate representations.