Detection of Remote Homologous Sequences via Approximate Methods and ESM-2

Charis Marinakis
National and Kapodistrian University of Athens -
Department of Information and Telecommunications
February 2026

1. Introduction

The detection of remote homologous proteins (remote homologs) constitutes one of the central challenges in computational biology. Although homology reflects shared evolutionary ancestry, protein sequences may diverge substantially over time, resulting in low sequence identity (<30%), commonly referred to as the *Twilight Zone*. Within this identity range, traditional alignment-based methods such as BLAST lose sensitivity and fail to reliably detect distant homologs.

The development of Protein Language Models, such as ESM-2, offers a novel approach. By learning distributional and structural properties of proteins from large-scale sequence data, these models map sequences into a continuous vector space (embedding space). In this space, neighboring proteins are likely to be functionally or structurally related, even when their sequence identity is low.

This work experimentally evaluates various Approximate Nearest Neighbor (ANN) methods for fast and efficient search within this embedding space, enabling the detection of remote homologs that are not easily recognized by BLAST. The comparison includes both traditional ANN approaches (LSH, Hypercube, IVF-Flat, IVF-PQ) and more recent neural-based methods (Neural LSH combined with KaHIP or Louvain), examining their effectiveness in terms of accuracy, computational efficiency, and the biological relevance of retrieved neighboring proteins.

Detecting remote homologs is essential for functional annotation of newly discovered proteins, understanding evolutionary relationships, and supporting structure prediction efforts. As biological databases continue to grow rapidly, scalable computational approaches become increasingly necessary.

The goal of this study is to demonstrate that the combination of protein embeddings and ANN indexing structures can reveal hidden functional and structural homologies, providing an efficient framework for searching large-scale protein databases while overcoming the limitations of purely sequence-based methods such as BLAST.

2. Methodology

The experimental procedure was conducted using Google Colab, enabling GPU-accelerated computation and efficient large-scale embedding generation.

2.1 Data

The database used in this study was UniProtKB/Swiss-Prot (specifically the *swissprot_50k* subset), which consists of manually curated protein sequences accompanied by high-quality functional and structural annotations, including UniProt annotations, Pfam domains, and Gene Ontology (GO) terms.

The query set comprises 12 target proteins that are independent of the database (i.e., not included in the indexed collection), ensuring unbiased retrieval evaluation.

2.2 Generation of ESM-2 Embeddings

For each protein sequence, embeddings were computed using the ESM-2 model. Representations from layer 6 were extracted, and mean pooling was applied across the sequence length to produce fixed-dimensional vector representations.

This procedure ensures that proteins of varying sequence lengths are mapped into vectors of identical dimensionality, enabling efficient similarity search in the embedding space.

2.3 ANN Methods

The following Approximate Nearest Neighbor (ANN) methods were evaluated.

Classical ANN methods, implemented in C++, include:
- LSH (Locality Sensitive Hashing)
- Hypercube (Random Projection)
- IVFFlat (Inverted File with k-means clustering)
- IVFPQ (Inverted File with Product Quantization)

Neural ANN methods, implemented in Python, include:
- NLSH (Neural LSH) via KaHIP
- NLSH (Neural LSH) via Louvain method

2.4 BLAST Ground Truth

The results were evaluated against those returned by BLAST when querying the same database. The top BLAST hits were used as a reference set (ground truth) for computing Recall@N, which measures the proportion of relevant BLAST-retrieved proteins recovered among the top-N ANN results.

3. Experimental Comparison

3.1 Evaluation Metrics

Performance was evaluated based on the following metrics:

- **Recall@N**: The proportion of top-*N* hits returned by BLAST that are also retrieved by each ANN method.
- **QPS (Queries Per Second)**: The number of search queries processed per second, reflecting computational efficiency.
- **L2 distance in the embedding space**: Euclidean distance between protein embeddings, used as the similarity metric during retrieval.

3.2 ANN Hyperparameter Configuration

For each Approximate Nearest Neighbor (ANN) method, an extensive hyperparameter exploration (parameter tuning) was conducted to construct trade-off curves between accuracy (Recall@N) and computational performance (QPS).

The final configurations reported below correspond to operating points on the trade-off curve that achieve the highest possible recall under an acceptable response time, according to the experimental results.

3.3 ANN Methods and Final Configurations

3.3.1 LSH (Locality Sensitive Hashing)

LSH maps embeddings into hash buckets such that similar vectors are likely to collide, enabling fast approximate nearest neighbor search. The key parameters are the number of hash functions per table (*k*), the number of hash tables (*L*), and the bucket width (*w*), which together control hash granularity and redundancy. The selected configuration was:

- k = 5
- L = 6
- w = 1

3.3.2 Hypercube

The Hypercube method projects embeddings into a binary hypercube and searches for neighbors over hypercube vertices. The parameters are the number of projection dimensions (*kproj*), bucket width (*w*), the maximum number of candidates considered (*M*), and the number of vertices explored during search (*probes*). The final configuration was:

- kproj = 3
- w = 4
- M = 6000
- probes = 1

3.3.3 IVF-Flat

IVF-Flat clusters embeddings into centroids and restricts search to selected inverted lists. The parameters are the number of clusters (*kclusters*) and the number of clusters explored per query (*nprobe*). The final configuration was:

- kclusters = 750
- nprobe = 35

3.3.4 IVF-PQ

IVF-PQ extends IVF-Flat with Product Quantization, compressing embeddings for faster search while maintaining accuracy. The parameters include the number of clusters (*kclusters*), number of clusters searched per query (*nprobe*), the number of subquantizers (*M*), and bits per subvector (*b*). The selected configuration was:

- kclusters = 700
- nprobe = 35
- M = 64
- b = 8

3.3.5 Neural LSH

Neural LSH partitions the embedding space using graph-based clustering. Two construction strategies were evaluated:

KaHIP-based partitioning: Builds a k-NN graph with *k* neighbors per node, partitions the graph into *m* clusters with controlled imbalance (*imbalance*), and trains for *epochs* iterations.

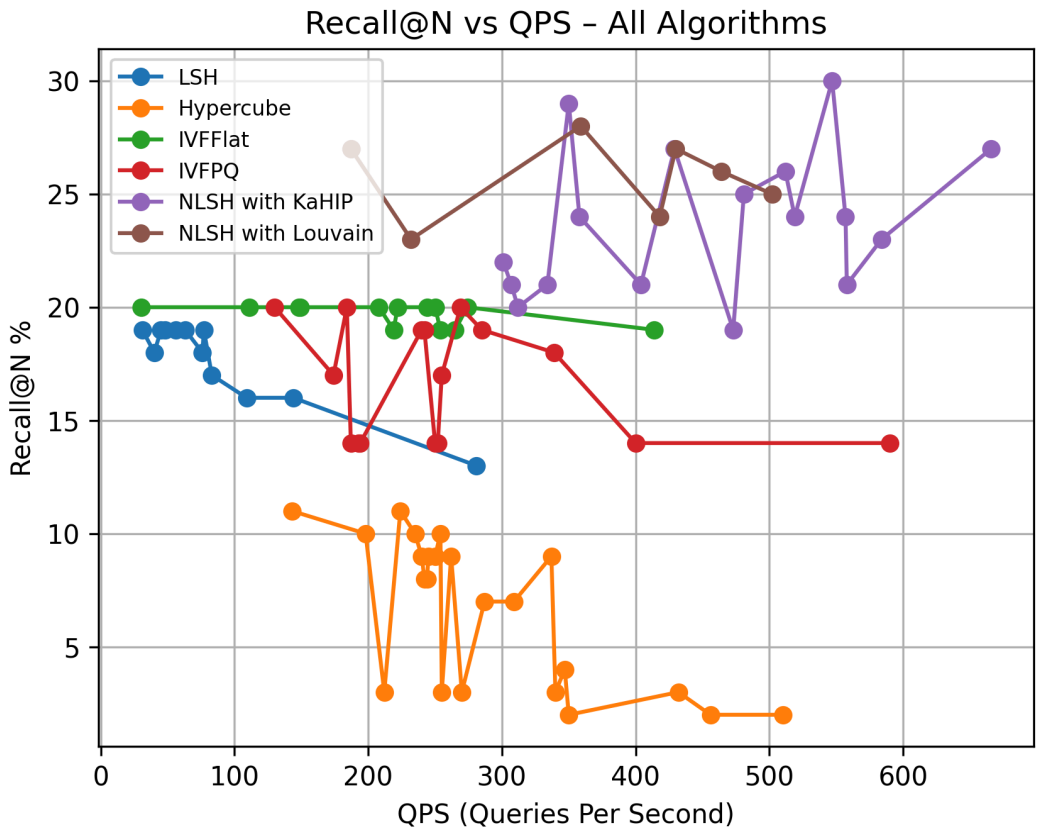
- k-NN graph: k = 30
- Number of partitions: m = 400
- Imbalance: 0.5
- Epochs: 20

Louvain-based partitioning: Builds a k-NN graph with *k* neighbors and detects communities with resolution parameter (*resolution*) over *epochs* iterations.

- k-NN graph: k = 20
- Resolution: 10
- Epochs: 20

During the search phase, the number of partitions explored per query (*T*) was set to 5.

The results of the hyperparameter exploration are summarized in the corresponding trade-off curve, illustrating the relationship between retrieval accuracy (Recall@N) and search speed (QPS).



The experimental results indicate that, the IVF-based methods (IVF-Flat and IVF-PQ) achieve higher recall compared to LSH and Hypercube, at the cost of reduced QPS, while, the Neural LSH approach provides the best overall trade-off, achieving competitive recall while maintaining high query throughput.

3.3 Results (N = 20)

Algorithm	Recall	QPS	tApprox	BuildTime
LSH	0.25	74.6	0.013	0.991445
Hypercube	0.14	194.2	0.005	0.125792
IVFFlat	0.27	180.4	0.006	289.897
IVFPQ	0.26	273.9	0.004	365.838
NLSH με KaHIP	0.35	362.7	0.003	72.427
NLSH με Louvain	0.33	410.8	0.002	137.205

4. Biological Evaluation

4.1 Definition of Remote Homolog

In this study, a pair of proteins is defined as a candidate remote homolog if it satisfies all of the following criteria:

1. BLAST sequence identity < 30%
2. Small L2 distance in the embedding space
3. At least one indication of functional or structural similarity (shared Pfam domain or common Gene Ontology (GO) terms)

This composite definition integrates sequence-based, embedding-based, and annotation-based evidence to identify biologically meaningful remote relationships.

4.2 Criteria for Result Characterization

The classification of retrieved pairs was performed using a combination of BLAST sequence identity, embedding distance (L2) and functional annotations from UniProt. For example, pairs with low sequence identity but shared Pfam domains were characterized as true positive remote homologs. Cases with small L2 distance but no shared Pfam domains or GO terms were characterized as potential false positives. This approach ensures that embedding-based similarity is interpreted within a biological context rather than purely geometrical proximity.

4.3 Επιλεγμένα Queries

The following queries were carefully selected as representative examples. For each query, the neighbors shown were consistently ranked within the top-5 across all evaluated ANN indices. This clearly demonstrates the ability of embeddings to detect remote homologs beyond sequence similarity, but also cases where embedding proximity may lead to biologically unsupported matches.

4.3.1 Remote homologs in Twilight Zone:

These cases exhibit BLAST identity < 30% but share Pfam domains and GO terms, demonstrating that embeddings can uncover hidden homologies beyond sequence similarity. These represent true positives that BLAST fails to rank highly.

Query protein: A0A001

Neighbor ID: sp|P9WQJ3|FATRP_MYCTU

BLAST Identity: **25.3%**

In BLAST Top-N?: **No**

Bio comment: True positive remote homolog (shared Pfam: PF00005)(shared GO terms)

Interpretation:
Despite the low sequence identity (<30%), shared Pfam domains and GO terms confirm functional and structural relatedness. The embedding-based proximity places this protein close in vector space, revealing hidden homology not captured by BLAST ranking.

4.3.2 False Positives

These neighbors fall within the Twilight Zone in terms of sequence identity but do not share Pfam domains or GO terms, suggesting that embedding proximity is not biologically supported.

Query Protein: A0A009I3Y5

Neighbor ID: sp|P24362|KRB2_VACCW

BLAST Identity: **26.8%**

In BLAST Top-N?: **No**

Bio comment: Embedding-based similarity without functional support (potential FP)

Interpretation:
The small L2 distance in embedding space may reflect shared motifs or model-specific learned features; however, the absence of shared annotations prevents confirmation of biological relatedness. This case is therefore considered a potential false positive.

4.3.3 Pure embedding hits

These neighbors exhibit BLAST identity = 0% but small L2 distance, indicating strong embedding-based similarity without sequence evidence or confirmed annotations.

Query Protein: A0A009HN45

Neighbor ID: sp|A6QKD8|SECA2_STAAE

BLAST Identity: **0.0%**

In BLAST Top-N?: **No**

Bio comment: Strong embedding-based similarity (no sequence evidence)

Interpretation:
Despite zero sequence identity, proximity in embedding space suggests possible functional or structural similarity. However, in the absence of supporting annotations, this case may represent either a novel remote homolog or an embedding-only false positive.

4.3.4 Σαφείς BLAST homologs:

In these cases, the neighbor appears at the top of BLAST results, confirming established homology.

Query Protein: A0A009HLV9

Neighbor ID: sp|P28787|NTRC_PROHU

BLAST Identity: **62.5%**

In BLAST Top-N?: **Yes**

Bio comment: Clear homology (BLAST)

Interpretation:
The high sequence identity confirms homology via BLAST. Embedding-based proximity also places the protein nearby in vector space, demonstrating that embeddings successfully capture both well-known homologs and remote relationships.

5. Conclusions

The analysis of the experimental results demonstrates that Neural LSH (NLSH)-based methods provide the best balance between retrieval accuracy and computational efficiency. Specifically, they achieve high Recall while maintaining very low query latency and high QPS, offering the most favorable trade-off between speed and precision for remote homology search. Other methods, such as IVF-PQ and Hypercube, achieve faster execution in certain configurations but generally exhibit lower recall.

The use of protein embeddings generated by ESM-2 revealed remote homologs even in cases where BLAST identity is below 30%, i.e., within the Twilight Zone where traditional sequence-based comparison becomes unreliable. Examples such as A0A001 demonstrated that nearest neighbors in the embedding space share common Pfam domains and Gene Ontology (GO) terms, confirming structural and functional homology and uncovering biologically meaningful relationships that are not easily detected using BLAST alone.

The embeddings also successfully identified clear high-identity homologs, as observed in A0A009HLV9, indicating that the method preserves the ability to detect well-established homologous relationships while simultaneously identifying novel, remote ones. At the same time, cases such as A0A009I3Y5 highlighted neighbors with small L2 embedding distance but lacking biological annotation support. These instances are considered potential false positives and underscore both the limitations of embedding-based similarity and the necessity of biological validation.

Overall, the combined use of protein embeddings and ANN indexing structures enables fast and effective large-scale database search, revealing hidden functional and structural homologies beyond sequence similarity alone. Future improvements may arise from the adoption of more advanced similarity metrics instead of pure L2 distance, deeper integration of structural information or conserved motif features, hybrid scoring schemes that combine embedding similarity with alignment-based evidence, and the development of scalable dynamic pipelines capable of handling significantly larger protein databases. Such directions could further enhance the robustness, interpretability, and scalability of embedding-based remote homology detection frameworks.

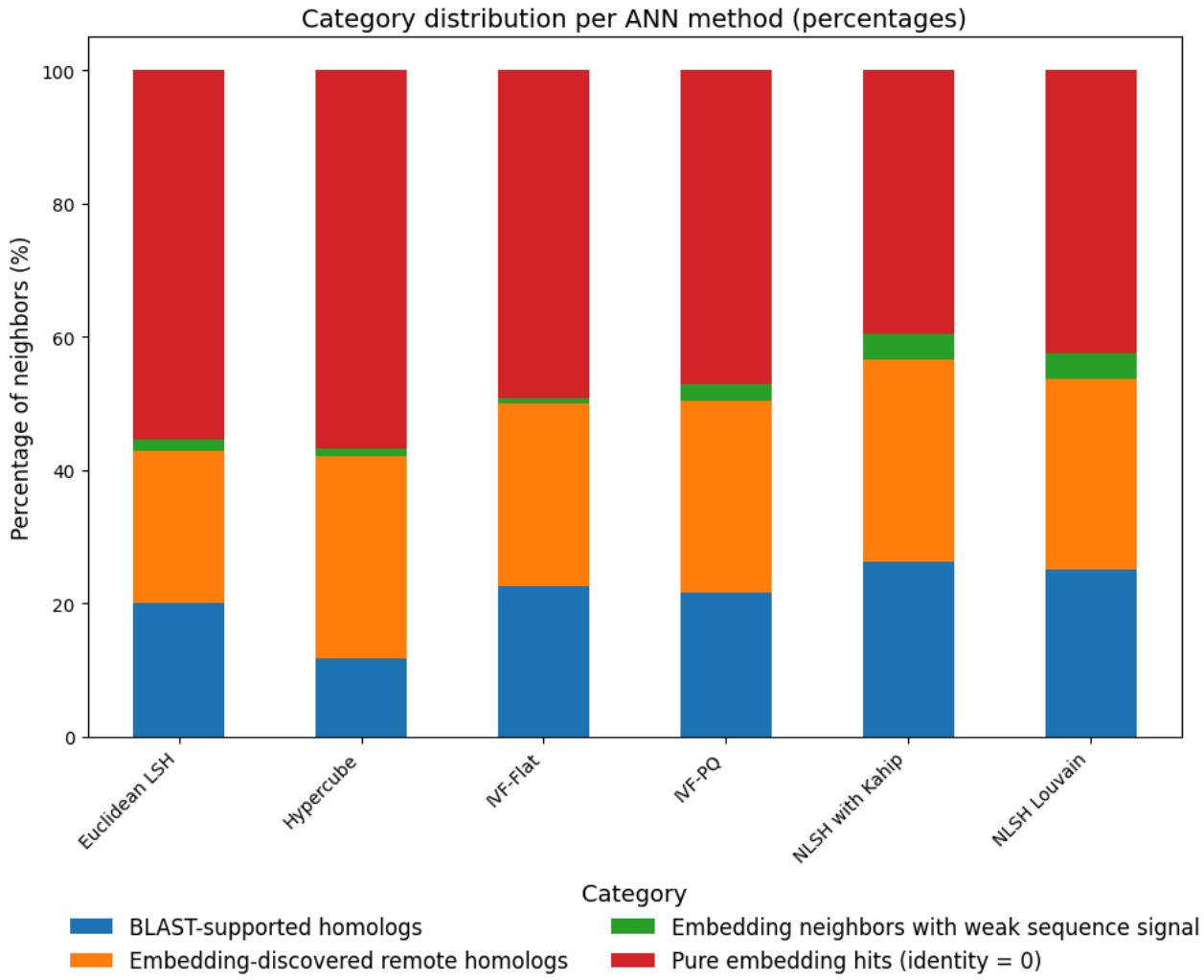
Beyond performance metrics, this study highlights a broader methodological shift in computational biology: similarity is no longer restricted to explicit sequence alignment but can be inferred through high-dimensional representations learned from large-scale protein databases. Models such as ESM-2 implicitly capture evolutionary, structural, and functional constraints embedded in sequence data, enabling the detection of biologically meaningful relationships even when classical alignment tools such as BLAST fail due to low sequence conservation. This suggests that embedding-based approaches provide a complementary representation of protein space that can extend the boundaries of homology detection and functional annotation in modern bioinformatics workflows.

6. Visualization

6.1 Category distribution per ANN method

The bar chart presents the percentage distribution of neighbor categories for each ANN method. The categories include BLAST-supported homologs, embedding-discovered remote homologs, embedding neighbors with weak sequence signal, and pure embedding hits (identity = 0).

The figure clearly shows that embedding-based methods identify a significant proportion of remote homologs that are not highly ranked by BLAST. At the same time, neighbors lacking sequence or functional evidence appear in limited proportions. The chart provides a direct comparison of each ANN method's ability to retrieve biologically meaningful neighboring proteins and highlights differences in how effectively each approach captures remote evolutionary relationships.



6.2 Recall vs Biorelevance

This plot examines the relationship between the classical performance metric Recall@N and biological relevance, defined based on functional annotations such as Pfam domains and GO terms. It illustrates that even in cases where BLAST fails to detect homologs (resulting in low recall), embedding-based methods are still capable of retrieving proteins that are biologically related in terms of function or structure.

The visualization demonstrates that ANN-based embeddings can uncover hidden homologs beyond sequence similarity alone, reinforcing the idea that embedding proximity may reflect deeper biological relationships that are not captured by traditional alignment-based methods.