A Comparative Analysis of Transformer-based Protein Language Models for Remote Homology Prediction

Anowarul Kabir*
Dept of Computer Science
George Mason University
akabir4@gmu.edu

Asher Moldwin†
Dept of Computer Science
George Mason University
amoldwin@gmu.edu

Amarda Shehu‡
Dept of Computer Science
George Mason University
amarda@gmu.edu

ABSTRACT

Protein language models based on the transformer architecture are increasingly shown to learn rich representations from protein sequences that improve performance on a variety of downstream protein prediction tasks. These tasks encompass a wide range of predictions, including prediction of secondary structure, subcellular localization, evolutionary relationships within protein families, as well as superfamily and family membership. There is recent evidence that such models also implicitly learn structural information. In this paper we put this to the test on a hallmark problem in computational biology, remote homology prediction. We employ a rigorous setting, where, by lowering sequence identity, we clarify whether the problem of remote homology prediction has been solved. Among various interesting findings, we report that current state-of-the-art, large models are still underperforming in the "twilight zone" of very low sequence identity.

CCS CONCEPTS

• **Computing methodologies** → **Machine Learning**; • **Applied computing** → **Molecular structural biology**; **Bioinformatics**.

KEYWORDS

remote homology, transformer, large language model.

ACM Reference Format:
Anowarul Kabir, Asher Moldwin, and Amarda Shehu. 2023. A Comparative Analysis of Transformer-based Protein Language Models for Remote Homology Prediction. In *14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '23), September 3–6, 2023, Houston, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3584371.3612942

1 INTRODUCTION

An explosion in the number of known protein sequences is allowing researchers to leverage the Transformer [29] architecture and build Protein Language Models (PLMs) [4, 11, 13]. PLMs are highly appealing due to their ability to learn task-agnostic representations of proteins. In particular, they provide an alternative framework to link protein sequence to function without relying on sequence alignments and similarity. Sequence representations learned via PLMs have been shown useful for various prediction tasks, including predicting secondary structure [11], subcellular localization [11, 26], evolutionary relationships within protein families [14], and superfamily [15] and family [20] membership.

Observations from recent studies indicate that PLMs, though trained exclusively on sequence data, learn structural information; work in [24] suggests that sequence-only PLMs indeed learn structural aspects. Scaling up to 15 billion parameters in ESM-2 (and training over 65 million unique sequences) yields representations that, harnessed through an equivariant NN, additionally predict tertiary structure (though not at AlphaFold2 accuracy) [17]. These reports are not entirely surprising; PLMs capture the well-understood selective pressures that have been exerted on protein sequences throughout millennia of evolution. These pressures originate from the functional requirements of proteins, which in-turn determine their structure by affecting the evolution of their underlying sequences. This ability to encode structure is perhaps also a major aspect of the utility of PLMs in downstream prediction tasks related to protein function, even if limited to superfamily prediction, function co-localization, Gene Ontology categorization [16, 18] and more.

We caution, however, that such performance, though seemingly impressive, may be somewhat exaggerated for various reasons. First, care has to be taken when constructing training datasets to remove sequence redundancy as well as to avoid data leakage, where proteins in the test data set may have high sequence identity with proteins in the training dataset. Second, structure and function are well preserved above 30% sequence identity [25]. Proteins with similar structure and function are also present below this level of identity but cannot be detected from sequence similarity alone [25]. It remains unclear how PLMs perform in this zone (which some authors have taken to referring to as the "twilight zone" [25]).

One challenging, hallmark problem in computational molecular biology, *remote homolog detection*, is a suitable stress test for how much a PLM has learned from sequence information alone, and whether indeed it can detect remote homologs in the twilight zone. It is worth noting that (protein) remote homology detection refers to the identification of proteins that are similar in structure but share low sequence identity; this is a working definition. The term remote homology was originally introduced to refer to proteins that share

*Equally-contributing Author
†Equally-contributing Author
‡Corresponding Author

BCB '23, September 3–6, 2023, Houston, TX, USA

ACM ISBN 979-8-4007-0126-9/23/09.
https://doi.org/10.1145/3584371.3612942

a superfamily[1] but not a family[2]. For the purpose of computational studies, this working definition lends itself to a gradated problem, where one lowers the sequence identity between proteins in the "test" dataset with the query/target protein, and determines whether proteins similar to the query can be detected. This is the setting for this paper, and it is in this setting, over decreasing levels of sequence identity, in which we evaluate pre-trained, transformer-based PLMs (over exclusively protein sequences) of various sizes for their ability to detect remote homologs.

Remote homology prediction is a particularly appropriate problem to determine whether a PLM pre-trained exclusively over protein sequences has also encoded/learned structure information. As one lowers sequence identity, it becomes increasingly difficult to identify homologous proteins based on sequence; remote homologs are those that retain their function (and structure) similarity at low levels of sequence identity. So, if a PLM allows identification of homologs at very low levels of sequence identity, then it has additionally encoded structure in its learned representations.

In this paper, we select powerful, representative, state-of-the-art transformer-based PLMs (trained exclusively over protein sequence data) and evaluate whether representations learned by them aid in remote homology detection/prediction. We employ a rigorous setting, where, by lowering sequence identity, we clarify whether the problem of remote homology prediction has been "solved." Indeed, in contrast to existing pre-prints and other reported findings that enthusiastically declare the problem solved (see Section 2), we show through a careful evaluation that these reports are highly exaggerated. The problem, particularly as one reaches the truly challenging setting of 30% or lower sequence identity, remains challenging for all current, SOTA PLMs, including large ones such as ESM2. This is one of the major findings of this paper.

An additional contribution of this paper is the presentation of metrics to objectively determine whether the distance between PLM-learned representations of proteins correlates with distance between corresponding sequences. This becomes particularly important after removing from consideration easy, high-sequence identity pairs. This analysis and others clarify and allow us to better understand the success and failure cases of PLMs for remote homology prediction. For instance, as we show here, we identify which protein domains are most and least amenable to remote-homology prediction based on PLM representation-similarity; we provide several visualizations to aid our understanding of whether useful structural information is easily-obtainable (or not) from PLM-learned representations of proteins.

The rest of this paper is organized as follows. We first relate some definitions, preliminaries, and necessary details about existing PLMs in Section 2. Section 3 relates our analysis setting, and the metrics utilized. Section 4 reports our findings, and Section 5 concludes the paper.

[1]A protein superfamily is the largest grouping (clade) of proteins for which common ancestry can be inferred.

[2]A protein family is a group of proteins with a common evolutionary origin, reflected by their related functions and similarities in sequence or structure.

2 RELATED WORK AND BACKGROUND

2.1 Protein Classification and Homology

Currently, the most commonly used definition of remote homology in computational studies is based on the hierarchical classification system for proteins provided in the SCOP2 [2, 3] and SCOPe [5, 12] databases [6]. These databases divide protein sequences into "domains" in levels of *classes*, *folds*, *superfamilies*, *families*, *protein regions*, and *protein types*. Generally, the criteria for family membership are related to sequence-level similarities; SCOP's documentation indicates that all sequences sharing sequence identity above 30% are grouped in the same family.

However, this appears to be a simplification of the actual criteria, as the analysis in [6] is based on similarity-based sequence clustering rather than all-to-all alignment and comparison of all protein sequence pairs in the database. Using this system, proteins belonging to the same superfamily are referred to as superfamily-level homologs [27]. Proteins in the same superfamily but in different families are considered remote homologs at the superfamily level [6, 22, 27].

2.2 Protein Language Models

Several iterations of PLMs have been developed since the advent of the transformer architecture. In particular, in this paper we employ three publicly available, pre-trained, SOTA PLMs to obtain representations for our analysis:

(1) **ESM-1** is the Evolutionary-Scale Modeling PLM [23]. ESM has been trained on 250 million protein sequences (a total of 86 billion amino acids) on masked-language-modelling tasks. While there are several lighter-weight ESM-1 variants, we utilize the ESM-1b variant with 33-layers and 650 M parameters.
(2) **ESM-2** is a more recent update to the ESM-1 architecture and was trained with variations spanning from 8 M to 15 B parameters [17]. For consistency, we used the 33-layers, 650 M parameter version.
(3) **ProtTrans T5** [10] is another, more recent PLM with self-supervised training, based on the original T5 model [21] for natural language processing. Specifically, ProtTrans-T5 is a 3 B parameter encoder-decoder model, and it was trained on a denoising task where 15% of the amino acids in the input were randomly masked.

All three of these models employ masked-token prediction as their training objective.

3 METHODS

3.1 Classic Definition: Remote Homology

In this study, we utilize the Structural Classification of Proteins SCOP2 [2, 3] database (latest update: 29 June 2022), containing 5,936 families and 2,816 superfamilies. SCOP2 defines family as a group of closely related proteins with clear evidence for their evolutionary origin and superfamily as a group that brings together more distantly-related protein domains. The similarity among proteins in a superfamily is frequently limited to common structural features that, along with a conserved architecture of active or binding sites

or similar modes of oligomerization, suggest a probable common evolutionary ancestry.

Following the definition from [6, 22, 27], we first define that a pair of proteins, p_i and p_j, are remote homologs if they belong to the same superfamily but different families, as follows:

$$areRemoteHomologs(p_i, p_j) = \begin{cases} 1, & \text{if } SF_i = SF_j \text{ and } F_i \neq F_j \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where SF_i and F_i define the superfamily and family label annotation of the i-th protein.

3.2 Hardened Definition: Remote Homology

We harden the above definition to accommodate the sequence identity threshold and focus on the truly hard cases; that is, no pair of remote homologs will share sequence identity more than a predefined threshold. This threshold will ensure that this pair falls into the "twilight zone" [25] in terms of sequence identity. The sequence identity is computed as a pairwise global alignment score. The extended equation is as follows:

$$areRemoteHomologs(p_i, p_j, th) = \begin{cases} 1, & \text{if } SF_i = SF_j \text{ and } F_i \neq F_j \\ & \text{and } identity(p_i, p_j) \leq th \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

We report experiments and results considering both of the above equations which allows us to truly gauge the performance of various PLMs as the problem becomes harder (lower sequence identity). We observe in our study that there is a high number of sequence pairs in different families with above 30% sequence identity than the sequence pairs belonging to the same family domain in the SCOP2 database (Figures not shown). This reinforces our hypothesis that extra filtering may be required if we want to identify the nontrivial remote-homologs without high sequence-level similarity, to test PLMs with. Fig. 1 shows the pairwise sequence identity distribution in the SCOP2 [2, 3] database.

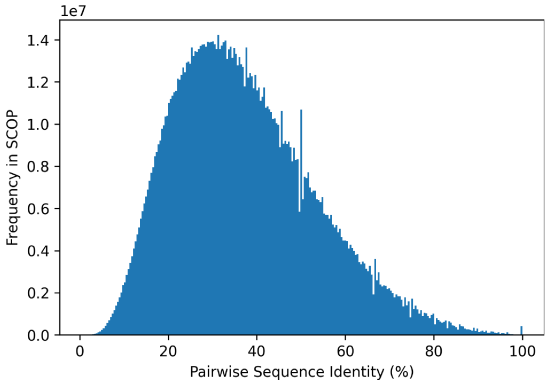

Figure 1: Histogram showing the sequence identity distribution of sequence pairs from the SCOP2 database.

3.3 Learned Amino-acid Level and Protein-level Representations

For a protein, $1 \leq i \leq N$, in the SCOP2 database, each defined by its sequence of l_i amino acids, we obtain a corresponding representation $s_i \in \mathbb{R}^{l_i \times D}$ from a PLM transformer; in this representation, each amino acid of a protein is mapped into $\mathbb{R}^D$.

Given a learned $s_i \in \mathbb{R}^{l_i \times D}$, we obtain protein-level representation $p_i \in \mathbb{R}^{1 \times D}$ by taking the average of the learned amino-acid-level features over the sequence length as in:

$$p_i = \frac{1}{l_i} \sum_{j=1}^{l_i} s_{ij} \quad (3)$$

3.4 Comparing PLM-learned Representations of Proteins

Following the methodology of Rives et al.[22], we adopt cosine similarity between a pair of protein representations as our similarity metric. Specifically, for each pair of sequences in SCOP2, we compute the representation similarity as follows:

$$sim(p_i, p_j) = \frac{p_i \cdot p_j}{\|p_i\| \|p_j\|} \quad (4)$$

3.5 Comparing Sequences of Proteins

To enable our analysis of the embedding representations of remote homologs in PLMs, we compute pairwise sequence alignments and identity scores for each of the $2 \times \binom{N}{2}$ pairs of sequences in the SCOP2 database. To compute these, we used Biopython's[7] pairwise alignment tool with default parameters.

3.6 From Representation Similarity to Prediction of Remote Homology

We employ several metrics and forms of analysis to evaluate whether structural commonalities between pairs of sequences are reflected in their embeddings.

3.6.1 Query-based Analysis. Using each sequence's PLM-learned embedding as a query (q_i), we exclude all other sequences from the same family (F_i) from the corpus of sequences (C) that will be queried. In our case, C refers to the set of all N sequences in SCOP2. We then exclude from C all sequences sharing a sequence identity above a given threshold th with the query sequence. The remaining query-sequence pairs are denoted as $\{(q_i, s) | q_i \in C, s \in C_i\}$, where $C_i = C \setminus (F_i \cup \{s \in C : \text{identity}(q_i, s) > th\})$.

For evaluating the performance, we consider the ground truth to be (i.e., the sequences are true homologs) if a sequence in the test dataset is from the same superfamily as the query and false otherwise, in accordance with Equation 2.

We then compare the pairwise embedding similarities and ground truths across all queries to obtain the following metrics:

(1) Area Under Reciever Operating Characteristic Curve (AUROC) [8]. We also report DeLong variances in the AUROC[9].
(2) Area Under Precision-Recall Curve (AUPRC) [8].
(3) Hit-10 [19] is the percentage of queries for which a true homolog was in the top-10 sequences with the most similar embeddings.

3.6.2 Clustering Analysis. We perform k-means clustering on the embeddings of sequences from the most-successfully-predicted and least-successfully-predicted superfamilies from each PLM based on the AUC (see above). We evaluate the quality of the resulting clusters and their agreement with the ground truth (i.e., whether sequences from the same superfamily are likely to be clustered together).

4 RESULTS & ANALYSIS

4.1 Experimental Setup

Our experimental setup is designed with the goal of accessibility and reproducibility.

4.1.1 Data Preprocessing. We opt to perform our analysis using all sequences in SCOP2 with minimal preprocessing or filtering. One exception to this is the removal of sequences where multiple spans were indicated within the same sequence, due to the ambiguity this creates when assessing the domains of the sequence and subsequences. We remove 506 such sequences compared with the total of 36, 900 sequences provided in SCOP2 database. Consequently, we have 2, 260, 440 remote-homolog pairs at the superfamily level. Note that we analyze significantly more remote-homologs (24 times) compared to Rives *et. al* [22] that reports performance on 92, 944 pairs of remote-homologs from SCOPe due to heavy filtration.

4.1.2 Sequence-Identity Thresholding. We compute the performance metrics using all protein sequences as individual queries. The thresholds we choose vary from 10% to 100% sequence identity with 5% increment. To compute AUROC, AUPRC, and HIT-10, we do not perform any sub-sampling or averaging of the protein sequences but instead choose to calculate all query-vs-ground-truth pairs and compute the metrics once over all samples, for each value of the sequence-identity threshold. This has the advantage of providing robust and reliable metrics, but this strategy also weights our results in favor of the larger superfamilies when compared with the strategy of sampling a single query from each superfamily. So, to provide a more fine-grain analysis at the superfamily level, we also report the same metrics for individual superfamilies from "hard" and "soft" domains, that is, difficult-to-predict and easy-to-predict superfamilies for each PLM.

4.2 Performance in the Twilight Zone

Figures 2, 3, and 4 show AUROC, AUPRC, and HIT-10 respectively for all three PLMs at varying levels of the sequence identity threshold. These metrics are also reported in numerical form in Tables 1, 2, and 3. For reference, "random" is added as a random baseline model; in it, the distribution of ground-truths is unchanged, but random numbers are used for embedding similarities. In Table 1 the DeLong variance is reported below the AUROC scores. These results appear to confirm that PLMs still struggle to identify remote homologs in the "twilight zone" [1, 25] from the sequence alone. We observe AUROC dropping sharply when the sequence identity threshold is lowered below 40%, indicating that above this threshold the problem is much easier.

We also note much lower performance than Rives *et. al* [24] at remote homology, even with no filtering (see "AUROC (Eq. 1)" in Table 1, or th=100% in Figure 2). Because the dataset used in their study of remote-homology prediction is not publicly available, we can only speculate as to the lower performance observed here. It is possible that it is due to the differences in the filtration applied to the dataset mentioned in Section 4.1.1, or differences in methodology for computing embeddings or calculating metrics that go beyond the details listed in their paper.

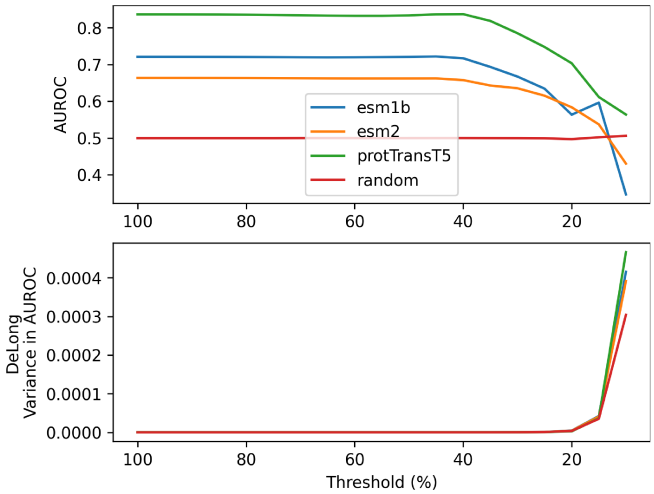

Figure 2: AUROC and DeLong variance embedding similarity as a predictor of homology for embeddings from all three PLMs, as the filtering sequence identity threshold is decreased from 100% to 10%. A threshold of 100% indicates no filtering beyond removal of sequences in the same family as the query, following Eq. 1. PLM embeddings of each sequence from the sequences in SCOP2 are used as queries.

Table 1: AUROC Comparison. DeLong variances are shown below the AUROC score.

PLMs	AUROC (Eq. 1)	AUROC (Eq. 2)			
		th=40%	th=30%	th=20%	th=10%
ESM1b	0.721± 4.31E-08	0.717± 6.02E-08	0.667± 2.86E-07	0.563± 4.51E-06	0.346± 4.15E-04
ESM2	0.663± 4.32E-08	0.658± 6.04E-08	0.635± 2.66E-07	0.584± 4.16E-06	0.430± 3.92E-04
ProtTrans-T5	0.836± 2.21E-08	0.837± 3.01E-08	0.785± 1.73E-07	0.703± 2.69E-06	0.564± 4.66E-04

Because remote homolog pairs are exceedingly rare when compared with the number of possible sequence pairs in SCOP2, this created a significant class-imbalance in the ground truth, calculating the AUROC scores. Thus, DeLong variances are also provided to give a measure of the reliability of the provided AUROC scores, especially as the threshold is lowered and the number of positive-ground-truth examples becomes even lower. In addition, we observe the similar trend of decreasing performance in AUPRC scores, indicating that these results are not simply an artifact of the worsening class imbalance as the threshold is lowered. The random baselines shown in Figures 2, 3, and 4 also confirm that the changing ground truth distribution for different values of the threshold are not to blame for the decrease in performance.

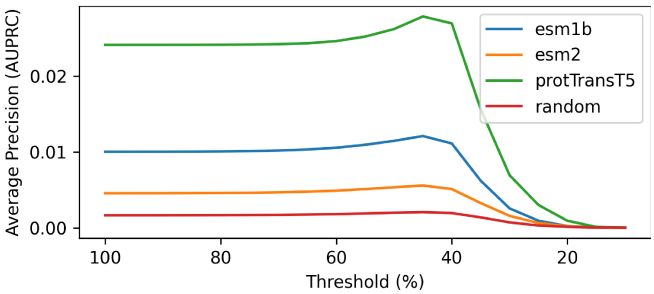

Figure 3: AUPRC of embedding similarity as a predictor of homology, as filtering threshold is decreased from 100% to 10%, from all three PLMs.

Table 2: AUPRC Comparison.

PLMs	AUPRC (Eq. 1)	AUPRC (Eq. 2)			
		th=40%	th=30%	th=20%	th=10%
ESM1b	1.003e-02	1.111e-02	2.563e-03	2.370e-04	3.186e-05
ESM2	4.559e-03	5.111e-03	1.582e-03	2.067e-04	3.519e-05
Prottrans-T5	2.411e-02	2.692e-02	6.927e-03	9.341e-04	5.901e-05

The Hit-10 scores show a similar trend regarding model performance in the "twilight zone", but with the difference that ESM-1b now outperforms ProtTransT5 and ESM-2 on this metric. Because this metric is calculated at the query level and then averaged over all queries, this may indicate that there are some classes of query where ESM-1b can identify the remote homologs at least to some degree, but the other two models completely fail to assign a high "top-10" rank to the true homologs.

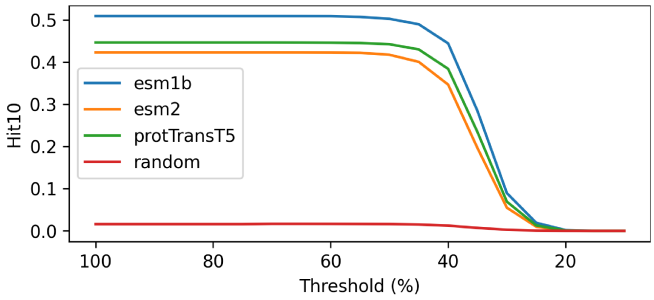

Figure 4: HIT-10 of embedding similarity as a predictor of remote-homology, as filtering threshold is decreased from 100% to 10% for all three PLMs.

Table 3: HIT-10 comparison.

Methods	HIT-10 (Eq. 1)	HIT-10 (Eq. 2)			
		th=40%	th=30%	th=20%	th=10%
ESM1b	1.003e-02	1.111e-02	2.563e-03	2.370e-04	3.186e-05
ESM2	4.559e-03	5.111e-03	1.582e-03	2.067e-04	3.519e-05
Prottrans-T5	2.411e-02	2.692e-02	6.927e-03	9.341e-04	5.901e-05

4.3 Protein Domain Analysis for Remote Homology Prediction

In addition to calculating AUROC across all queries in the SCOP2 database, we also calculate the same metrics separately for query sequences coming from each superfamily in SCOP2. To identify the "hard" and "soft" (i.e., difficult-to-predict and easy-to-predict) superfamilies for each PLM, we start with the 150 superfamilies with the highest number of remote homologs in SCOP2, and identify the 10 superfamilies with the highest AUC and the 10 with the lowest AUC when attempting to predict homologs based on PLM embeddings, when using queries from that superfamily.

Notably, the better-performing superfamilies tended to have fewer included sequences on average, indicating that these may be the superfamilies with more refined and restrictive definitions than the larger superfamilies shown in Table 4 and 5. Another explanation is that the inflated AUROC scores may be caused by the increased class imbalance for queries from the smaller superfamilies. However, the PRC column indicates that generally the bottom-10 superfamilies tended to also have lower PRC. To a lesser degree, this also holds for the Hit-10 scores, despite the fact that even many the top-10 superfamilies had a hit-10 score of zero.

Table 4 shows the superfamilies with the highest AUROC when using embeddings from the ProtTransT5 model (the best-performing PLM, judging by its AUROC in Table 1) to predict remote homologs, and Table 5 shows the superfamilies where the AUC was lowest. Similar tables, giving the hard and soft domains for the other two PLMs, are provided in the supplement. Note that the AUC scores used here are using the sequence-identity filtering threshold of 30%.

4.4 Visual Analysis of Hard and Soft sets

To visualize how well the "hard" and "soft" domains are separated in the representational space of the PLMs, we perform a T-SNE [28] dimensionality reduction to view the embeddings from these superfamilies in a two dimensional plot. The T-SNE transformation is fit using all sequences in the SCOP2 database. Note that superfamily-based filtering is only applied later, when producing the visualizations. In Figure 5, 6 and 7, we report the top-5 "soft" domains in the top panel and the bottom-5 "hard" domains in the bottom panel for ESM1b, ESM2 and ProtTransT5, respectively. Subjectively, in all cases this appears to show cleaner and more defined clusters when considering the "soft" domains, relative to the "hard". This indicates some level of agreement between the distances between sequences in our T-SNE projection, and the cosine similarity between pairs of sequences that we used to define remote homologs in the high dimensional protein embedding space.

4.5 Distribution of Pairwise Embedding Similarity

To better understand the significance of a given similarity level between two sequences in the representational space of the PLMS, we visualize the distribution of embedding similarities across all sequence pairs in SCOP2 in all three PLMs in Figure 8. All three PLMs show unimodal distributions of embedding similarities. However, the distributions for both ESM models is skewed heavily toward higher similarities between embeddings. Interestingly, the distribution for ProtTransT5 embedding similarities is almost identical to

Table 4: Superfamilies with highest AUROC(@th=0.3) in ProtTransT5, selected from a list of 150 superfamilies with the most remote homologs in the SCOP database. Note that the maximum possible number of sequence pairs that can be remote homlogs is actually higher than the number of sequences in the superfamily. This is because the number of pairs is $2\times\binom{|SF|}{2}$. Also note that the sequence counts reported in the "Num Seqs" and "Num RHs" columns are prior to applying the threshold.

Super family	Description	ROC	PRC	HIT-10	Num Seqs	Num RHs
3000845	Furanosidase-like	1.00	0.01	0.00	51	1870
3000255	omega toxin-like	1.00	0.58	0.22	67	1980
3002200	NADH-quinone oxidoreductase subunit 11-like	1.00	0.59	0.00	44	1322
3000050	Thioredoxin reductase-like	1.00	0.33	0.02	44	1236
3001604	beta-lactamase/ transpeptidase-like	1.00	0.29	0.02	211	4070
3000069	DEATH domain	1.00	0.53	0.13	61	2580
3000309	Scorpion toxin-like	0.99	0.77	0.38	129	11884
3000224	Porins	0.99	0.44	0.29	52	1916
3001790	Thiamin diphosphate-binding fold (THDP-binding)	0.99	0.17	0.19	59	2920
3000884	Ribosomal protein L16p/L10e	0.98	0.02	0.00	54	1450

Table 5: Superfamilies with lowest AUROC(@th=0.3) in ProtTransT5.

Super family	Description	ROC	PRC	HIT-10	Num Seqs	Num RHs
3000110	RNA-binding domain RBD	0.47	0.00	0.03	236	11876
3000570	UBC-like	0.52	0.00	0.04	85	1994
3001694	WD40 repeat-like	0.54	0.00	0.00	72	3204
3001397	SIS domain/ Ribosomal protein S2-like	0.57	0.00	0.00	101	5824
3000053	S13-like H2TH domain	0.59	0.00	0.01	84	3098
3000098	Nudix	0.60	0.00	0.09	66	1916
3001284	S15/NS1 RNA-binding domain	0.61	0.00	0.00	55	1152
3000066	Protein kinase-like (PK-like)	0.64	0.00	0.00	413	21446
3001808	Cysteine proteinases	0.65	0.01	0.23	166	22180
3001593	Adenine nucleotide alpha hydrolases-like	0.66	0.00	0.00	57	2314

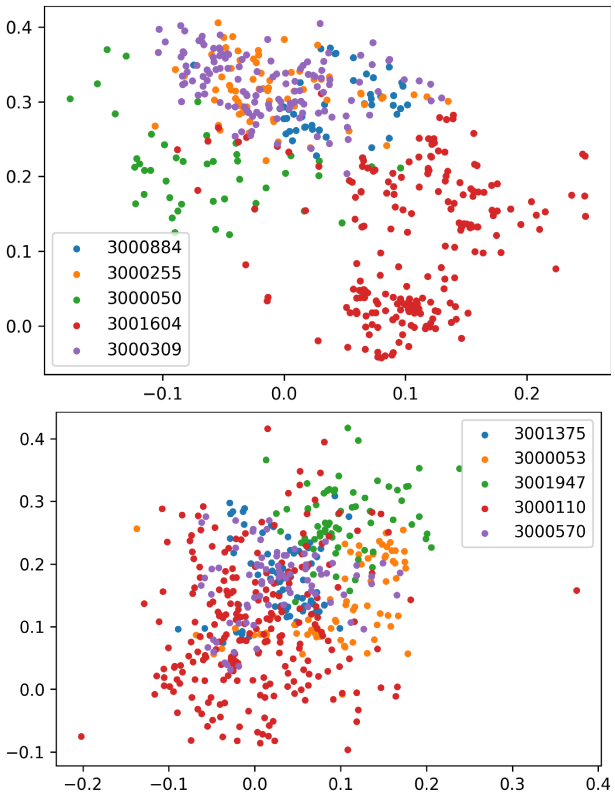

Figure 5: T-SNE plot of ESM-1b embeddings for sequences from superfamilies that had the (top panel) top-5 AUROC and (bottom panel) bottom-5 AUROC shown in Tables 7 and 8.

the distribution of pairwise sequence identities in SCOP2 shown in Figure 1. However, while this may seem to indicate that the sequence information is retained by the model, it may actually be a coincidence: Figure 9 shows little correlation between pairwise sequence identity and pairwise embedding similarity in the ProtTransT5 model.

4.6 Clustering Analysis

To better quantify how well-defined and separated the superfamilies from the "hard" and "soft" domains are in the representational space of PLMs, we provide a clustering analysis. Table 6 shows the performance of k-means clustering on the "hard" and "soft" domains using embeddings from each of the PLMs. Note that for each PLM, we use its own "hard" and "soft" domains based on the AUROC of that PLM's embeddings at predicting the remote homologs in those domains. Predictably, this unsupervised clustering is more successful at differentiating the "soft" superfamilies from each other than it is at differentiating the "hard" superfamilies. These results serve to bolster the results achieved using pairwise cosine similarity, indicating that this distinction between "hard" and "soft" domains holds, even at the cluster (rather than just sequence-pair) level.

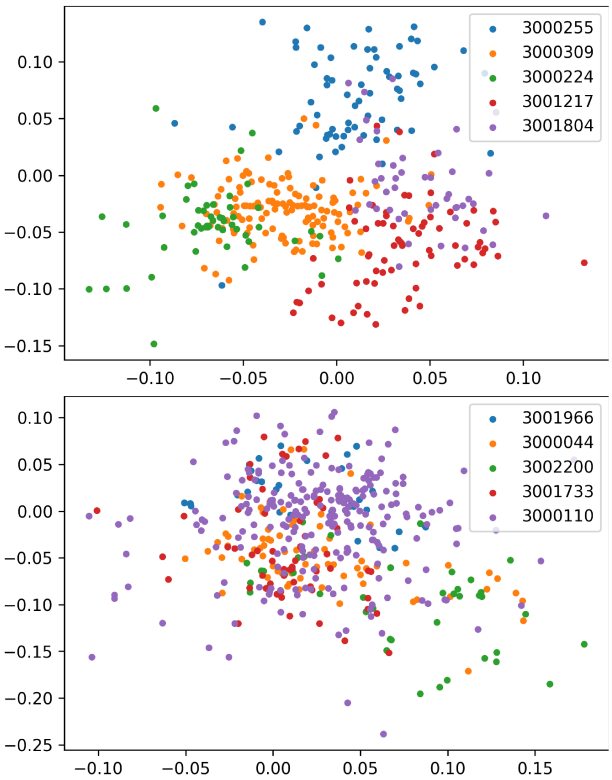

Figure 6: T-SNE plot of ESM-2 embeddings for sequences from superfamilies that had the (top panel) top-5 AUROC and (bottom panel) bottom-5 AUROC shown in Tables 9 and 10.

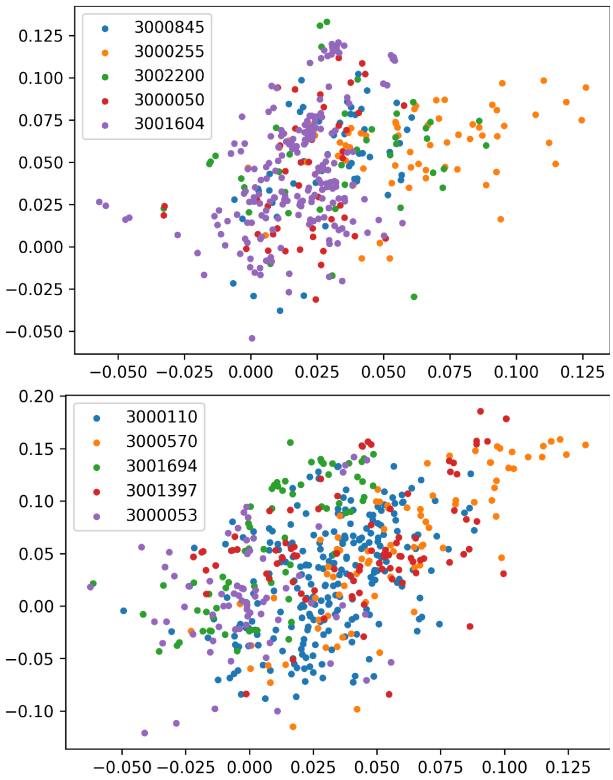

Figure 7: Top: T-SNE plot of ProtTransT5 embeddings for sequences from superfamilies that had the (top panel) top-5 AUROC and (bottom) bottom-5 AUROC shown in Tables 4 and 5.

Table 6: Cluster separability and accuracy metrics for K-means applied to embeddings for all sequences coming from the top-10 "hard" and "soft" domains (i.e., superfamilies) for each model, where the superfamily label is taken as the ground truth.

	esm1b	esm2	protTransT5
Top 10 AUC	("Soft" Domains)		
Silhouette	0.210186	0.210186	0.210186
ARI	0.716840	0.716840	0.716840
NMI	0.868765	0.868765	0.868765
Bottom 10 AUC	("Hard" Domains)		
Silhouette	0.120892	0.120892	0.120892
ARI	0.479750	0.479750	0.479750
NMI	0.595362	0.595362	0.595362

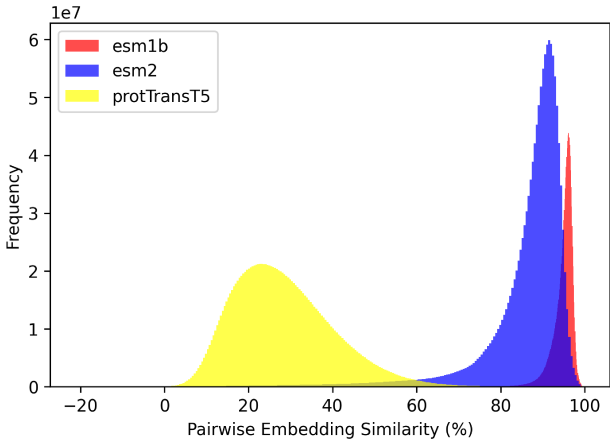

Figure 8: Histogram showing pairwise embedding similarities (cosine) for each model, using all pairwise comparisons between sequences from the SCOP2 database.

5 CONCLUSION

Through our rigorous experiments where we carefully controlled the difficulty of the setting for remote homology prediction, we have gained valuable insights into the current state of PLMs in identifying remote homology and capturing structural features of protein sequences. Our main set of results largely conflicts with the analogous analyses performed by other research groups investigating their own state-of-the-art PLMs. In summary, remote homology prediction remains difficult for PLMs where it matters; that is, as sequence identity gets lower.

By conducting analyses in the challenging "twilight zone" and excluding numerous trivial samples from the dataset used to evaluate remote homology prediction metrics, we have shed light on the behavior of PLMs under difficult conditions. We have examined specific superfamilies where PLMs effectively capture remote homologs as well as cases where they exhibit poor performance,

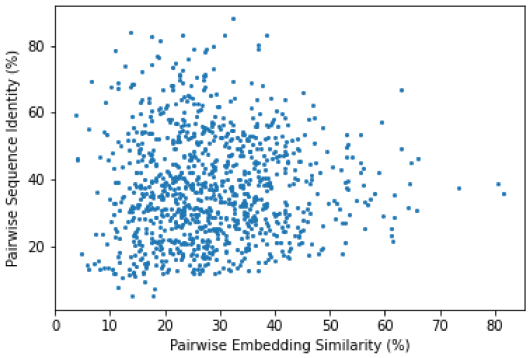

Figure 9: Scatterplot of embedding similarity vs Sequence identity for 1000 randomly-sampled pairs of sequences in the SCOP2 database. Embeddings shown are from ProtTransT5.

offering valuable insights for improving future PLMs and even facilitating the development of novel protein-modeling approaches beyond the traditional PLM paradigm.

In addition, our thorough analysis includes visualizations of various aspects of PLM representations that provide further understanding of their successes and failures. These visualizations complement our main conclusion and offer useful insights into the factors contributing to PLMs' performance.

We also uncovered important details regarding the distribution of protein domains and pairwise sequence identities in the SCOP database that supplement their original documentation and provide missing information regarding the presence of many sequences officially categorized as being in different families, that in reality share a high sequence identity.

In future work, we plan to leverage these findings to inform our exploration of different training regimes and model architectures. Rather than relying on sequence-level similarity, we aim to focus on performance in the "twilight zone" using a new benchmark dataset. Furthermore, we aspire to incorporate more biological knowledge to explain the successes and failures of existing PLMs through further analysis.

We believe that our work will be valuable to researchers dedicated to advancing protein structure models. The datasets, code, and analyses presented here are available at: github.com/amoldwin/plm-remote-homolog-analysis.

6 ACKNOWLEDGMENTS

Computations were run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University, VA (URL: http://orc.gmu.edu).

REFERENCES

[1] Guillermin Agüero-Chapin, Deborah Galpert, et al. 2020. Graph Theory-Based Sequence Descriptors as Remote Homology Predictors. *Biomolecules* 10, 1 (2020).
[2] Antonina Andreeva, Dave Howorth, et al. 2013. SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Research* 42, D1 (2013), D310–D314.
[3] Antonina Andreeva, Eugene Kulesha, et al. 2019. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research* 48, D1 (2019), D376–D382.
[4] T. Bepler and B. Berger. 2021. Learning the protein language: Evolution, structure, and function. *Cell Syst* 12, 6 (2021), 654–669.e3.
[5] John-Marc Chandonia, Lindsey Guan, et al. 2021. SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Research* 50, D1 (2021), D553–D559.
[6] Junjie Chen, Mingyue Guo, et al. 2016. A comprehensive review and comparison of different computational methods for protein remote homology detection. *Briefings in Bioinformatics* 19, 2 (Nov. 2016), 231–244.
[7] P. A. Cock, T. Antao, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioniformatics* 25 (2009), 1422–1423.
[8] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*. 233–240.
[9] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. 1988. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 44, 3 (1988), 837–845.
[10] Ahmed Elnaggar, Michael Heinzinger, et al. 2020. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. *bioRxiv* (2020).
[11] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, et al. 2022. ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Trans Patern Anal Mach Intell* 44, 10 (2022), 7112–7127.
[12] Naomi K. Fox, Steven E. Brenner, and John-Marc Chandonia. 2013. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research* 42, D1 (2013), D304–D309.
[13] M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, et al. 2019. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* 20, 723 (2019), 1–17.
[14] B. Hie, L. Yang, Kim. K. K., and P. S. Kim. 2022. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Syst* 13, 4 (2022), 274–285.e6.
[15] Anowarul Kabir and Amarda Shehu. 2022. Transformer Neural Networks Attending to Both Sequence and Structure for Protein Prediction Tasks. In *Intl Conf on Knowledge Graphs (ICKG)*. IEEE. https://arxiv.org/abs/2206.11057 accepted.
[16] K. L. Kabir, B. Ma, R. Nussinov, and A. Shehu. 2022. Fewer Dimensions, More Structures for Improved Discrete Models of Dynamics of Free versus Antigen-Bound Antibody. *Biomolecules* 21, 7 (2022), 1011.
[17] Zeming Lin, Halil Akin, et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* (2022).
[18] M. Littmann, M. Heinzinger, C. Dallago, T. Olenyi, and B. Rost. 2021. Embeddings from deep learning transfer GO annotations beyond homology. *Sci Rep* 11, 1 (2021), 1160.
[19] Jianzhu Ma, Sheng Wang, et al. 2014. MRFalign: Protein Homology Detection through Alignment of Markov Random Fields. In *Research in Computational Molecular Biology*, Roded Sharan (Ed.). Springer International Publishing, Cham, 173–174.
[20] Ananthan Nambiar, Simon Liu, Mark Hopkins, Maeve Heflin, Sergei Maslov, et al. 2020. Transforming the Language of Life: Transformer Neural Networks for Protein Prediction Tasks. In *Intl Conf on Bioinformatics, Computational Biology, and Health Informatics (BCB)*. ACM, 1–8.
[21] Colin Raffel, Noam Shazeer, et al. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *CoRR* abs/1910.10683 (2019).
[22] Alexander Rives, Joshua Meier, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* 118, 15 (April 2021).
[23] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA* 118, 15 (2021), e2016239118.
[24] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA* 118, 15 (2021), e2016239118.
[25] B. Rost. 1999. Twilight zone of protein sequence alignments. *Protein Eng* 12, 2 (1999), 85–94.
[26] H. Stärk, C. Dallago, M. Heinzinger, and B. Rost. 2021. Light attention predicts protein location from the language of life. *Bioinformatics Adv* 1, 1 (2021), vbab035.
[27] Nils Strodthoff, Patrick Wagner, et al. 2020. UDSMProt: universal deep sequence models for protein classification. *Bioinformatics* 36, 8 (2020), 2401–2409.
[28] L. J. P. van der Maaten and G. E. Hinton. 2008. Visualizing High-Dimensional Data Using t-SNE. *J Mach Learn Res* 9 (2008), 2579–2605.
[29] Ashish Vaswani, Noam Shazeer, et al. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

SUPPLEMENT

This provides more analysis on "hard" and "soft" domains for ESM1b and ESM2.

Table 7: Superfamilies with highest AUC(@th=0.3) in ESM1b.

Super family	Description	ROC	PRC	HIT-10	Num Seqs	Num RHs
3000884	Ribosomal protein L16p/L10e	1.00	0.70	0.33	54	1450
3000255	omega toxin-like	0.99	0.34	0.22	67	1980
3000050	Thioredoxin reductase-like	0.99	0.00	0.00	44	1236
3001604	beta-lactamase/ transpeptidase-like	0.99	0.27	0.02	211	4070
3000309	Scorpion toxin-like	0.99	0.70	0.43	129	11884
3000020	4Fe-4S ferredoxins	0.98	0.51	0.25	63	2856
3000224	Porins	0.98	0.04	0.29	52	1916
3000113	2Fe-2S ferredoxin-like	0.98	0.01	0.14	49	1140
3000845	Furanosidase-like	0.98	0.03	0.04	51	1870
3000545	Snake toxin-like	0.98	0.42	0.11	66	1402

Table 8: Superfamilies with lowest AUC(@th=0.3) in ESM1b.

Super family	Description	ROC	PRC	HIT-10	Num Seqs	Num RHs
3001375	Isocitrate/ Isopropylmalate dehydrogenase-like	0.17	0.00	0.00	74	2722
3000053	S13-like H2TH domain	0.33	0.00	0.04	84	3098
3001947	FMN-dependent nitroreductase-like	0.34	0.00	0.00	70	1162
3000110	RNA-binding domain RBD	0.40	0.00	0.06	236	11876
3000570	UBC-like	0.40	0.00	0.09	85	1994
3001059	Acid proteases	0.48	0.02	0.18	146	10592
3000143	Ribonuclease H-like	0.48	0.00	0.03	145	17824
3000738	Bet v1-like	0.51	0.00	0.10	78	4842
3000098	Nudix	0.52	0.00	0.03	66	1916
3001261	C-type lectin-like	0.54	0.00	0.38	132	4820

Table 9: Superfamilies with highest AUC(@th=0.3) in ESM2.

Super family	Description	PRC	HIT-10	ROC	Num Seqs	Num RHs
3000255	omega toxin-like	0.97	0.15	0.07	67	1980
3000309	Scorpion toxin-like	0.96	0.60	0.37	129	11884
3000224	Porins	0.94	0.16	0.27	52	1916
3001217	Flavoproteins	0.93	0.01	0.10	70	3992
3001804	Trimeric LpxA-like enzymes	0.93	0.05	0.42	45	1560
3001375	Isocitrate/ Isopropylmalate dehydrogenase-like	0.91	0.00	0.00	74	2722
3000433	50S Ribosomal protein L14-like	0.90	0.01	0.34	86	2346
3001604	beta-lactamase/ transpeptidase-like	0.90	0.03	0.02	211	4070
3002736	FimA/ Mfa2-like	0.89	0.00	0.07	42	1336
3000091	GHKL (Gyrase Hsp90 Histidine Kinase MutL) do...	0.89	0.00	0.00	56	2196

Table 10: Superfamilies with lowest AUC(@th=0.3) in ESM2.

Super family	Description	PRC	HIT-10	ROC	Num Seqs	Num RHs
3001966	L28p-like	0.29	0.00	0.00	57	1620
3000044	D-2-hydroxyacid dehydrogenase-like	0.40	0.00	0.00	72	1626
3002200	NADH-quinone oxidoreductase subunit 11-like	0.41	0.00	0.07	44	1322
3001733	YjgF-like	0.42	0.00	0.00	58	1530
3000110	RNA-binding domain RBD	0.44	0.00	0.02	236	11876
3000473	G protein-coupled receptor-like	0.45	0.00	0.00	130	9018
3000264	Calponin-homology domain CH-domain	0.46	0.00	0.00	49	1664
3001261	C-type lectin-like	0.48	0.00	0.03	132	4820
3000098	Nudix	0.49	0.00	0.03	66	1916
3001694	WD40 repeat-like	0.50	0.00	0.00	72	3204