A More Informative and Reproducible Remote Homology Evaluation for Protein Language Models

Asher Moldwin[1], **Anowarul Kabir**[1], **Amarda Shehu**[1]
[1]George Mason University
amoldwin@gmu.edu, akabir4@gmu.edu, ashehu@gmu.edu

Recent studies exploring the abilities of transformer-based protein language models have highlighted their performance on the task of remote homology detection, but have not provided datasets or evaluation procedures geared toward properly measuring performance on this task. With the goal of obtaining more informative and reproducible results, we offer a detailed procedure for constructing datasets and evaluating remote homology detection performance in a way that allows detailed analyses to be performed that shed light on the remote homology detection performance throughout the "twilight zone" of low sequence similarity. Using the proposed procedures, we found that three state-of-the-art protein language models exhibit diminishing performance when the pairwise sequence similarity between the query sequence and other proteins is restricted to below 35% identity.

Introduction

Protein Language Models (PLMs)(Bepler and Berger 2021; Elnaggar et al. 2022a; Heinzinger et al. 2019), are transformer-based(Vaswani, Shazeer et al. 2017) models that are trained on large amounts of protein sequence data. In addition to effectively modelling the protein sequence, recent studies(Rives, Meier et al. 2021) have indicated that PLMs also implicitly model protein structure and function, making their sequence representations useful for downstream tasks such as predicting secondary structure (Elnaggar et al. 2022b), subcellular localization (Stärk et al. 2021; Elnaggar et al. 2022b), evolutionary relationships within protein families (Hie et al. 2022) and superfamily (Kabir and Shehu 2022) and family (Nambiar et al. 2020) membership. One task in particular that has been used as a "stress-test" of the degree to which PLMs can learn structural information is *remote homology detection*.

Remote homology detection refers to the task of identifying pairs of proteins that share similar structure but have a deceptively low level of similarity between their raw sequences. In particular, remote homology at the superfamily level refers to proteins in the same *superfamily* but different *families*, where superfamily membership indicates similar structural characteristics and family membership indicates a high degree of similarity in the protein sequence.

Several existing studies observe strong performance of PLMs on remote homology detection, but these studies seem to rely on simplified formulations of the remote homology problem to make this point. Specifically, the homology detection problem is often made easier by leaving high-similarity pairs of sequences in the evaluation dataset, making some homologs trivially identifiable from the sequence-level commonalities alone. The simplified formulation adopted by these studies also seems particularly inappropriate because contemporary structural biology literature frequently admits that the homology-detection problem only becomes truly challenging when considering sequences in the "twilight zone" of less than 35% identity(Rost 1999).

In addition, the existing studies do not provide detailed procedures for reproducing the relevant datasets or evaluations. This makes it difficult to replicate results and also necessitates starting from scratch when attempting to evaluate new models or modify the evaluation procedure in any way.

In this paper, we seek to introduce detailed and reproducible dataset construction procedures as well as a refined evaluation method aimed at providing more informative results regarding performance in the low sequence-identity setting.

Problem Formulation

A convenient definition of remote homology that has been adopted by many recent computational studies relies on the hierarchical protein classification system used to annotate proteins in the Structural Classification of Proteins (SCOP2) (Andreeva, Howorth et al. 2013; Andreeva, Kulesha et al. 2019) and SCOPe(Chandonia, Guan et al. 2021; Fox, Brenner, and Chandonia 2013) databases. In this system, if two proteins are believed to share a common ancestry due to common functional and structural features, they are said to belong to the same *superfamily*. Family membership, on the other hand, refers to proteins that share a high similarity in their raw sequence. As such, sequences sharing above 30% identity generally labeled as belonging to the same family. It should be noted that there appear to be exceptions to these criteria because the classification is based on identified clusters of similar proteins rather than describing all of the individual pairwise commonalities.

Basic Definition: Remote Homology

Following the definition from (Chen, Guo et al. 2016; Strodthoff, Wagner et al. 2020; Rives, Meier et al. 2021), 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.

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" (Rost 1999) in terms of sequence identity. The sequence identity is computed as a pairwise global alignment score. The extended equation is as follows:

(2)

$$\text{areRemoteHomologs}(\text{p}_i, p_j, th) = \begin{cases} 1, & \text{if } SF_i = SF_j \text{ and } F_i \neq F_j \\ & \text{and } identity(p_i, p_j) \leq th \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Homology Detection Using Sequence Representations

Following Rives et al.(Rives, Meier et al. 2021), we focus on representation-based detection of remote homology. This entails computing the cosine similarity between two vector-encoded protein representations to predict whether or not they are homologous. In this approach, sequence pairs with a higher cosine similarity between their vector-representations are taken to be more likely to be homologous than those with low cosine similarity.

In the case of PLMs, protein representations can be obtained by extracting their final-layer embeddings from the transformer model.

PLMs for Remote Homology Prediction

In the context of PLMs, our interest in their remote homology detection performance is twofold: (1) More accurate tools for remote homology prediction are needed in structural biology research, where the still-imperfect HHblits(Remmert et al. 2011) algorithm is currently the state of the art method for homology detection, and (2) observing PLM performance on remote homology detection can shed light on the degree to which PLMs implicitly learn about protein structure despite being trained only on raw sequences.

For both of these goals, the ability to gain a detailed understanding of PLM performance for different sequence identity thresholds is critical. For the purposes of this study, we will consider protein representations from 3 state-of-the-art PLMs: (1) ESM-1b(Rives, Meier et al. 2021), (2)ESM-2(Verkuil et al. 2022), and (3)ProtTrans-T5(Elnaggar et al. 2022a).

Evaluation Datasets

In this section, we describe (1) the dataset construction pipeline used by Rives et al. to evaluate PLM remote homology detection and (2) our own enhanced dataset construction and evaluation procedures.

Procedure from Rives et al.

The primary study of PLM performance on remote homology detection is included in the paper by Rives et al. introducing and analyzing the ESM1-b model.(Rives, Meier et al. 2021) Here we reproduce their dataset construction pipeline to the best of our ability, while noting that many details in their description were unclear and no dataset was made public for this task. The authors also did not respond to our request for access to their evaluation dataset. Also note that their evaluation involved calculation of area under the receiver operating characteristic (AUROC) and HIT-10 metrics, but it is not clear whether these values are meant to be macro-averaged across all queries or calculated for the dataset as a whole. Due to time and resource constraints, we take a random sample of 1000 query sequences to calculate performance metrics on this dataset, and macro-average the results across all queries.

Our attempt to reproduce their dataset construction pipeline is shown in Figure 1. Note that our attempt to reconstruct their dataset had over twice as many unique remote homolog pairs at the superfamily level when compared with the number given in their original evaluation.

In addition to reproducing their original experiments, we also evaluate PLMs on datasets generated using the same procedure but using different sequence-identity threshold subsets provided by the ASTRAL compendium (Dai et al. 2019).

Proposed Procedure

Unlike Rives et al., we choose to use the SCOP2(Andreeva, Howorth et al. 2013; Andreeva, Kulesha et al. 2019) database instead of SCOPe due to our observation that the superfamily annotations in SCOP2 tend to be more reliable.

Our experimental setup is designed with the goal of accessibility and reproducibility. As such, we opt to construct our datasets 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 sub-sequences. 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* (Rives, Meier et al. 2021) that reports performance on $92,944$ pairs of remote-homologs from SCOPe.

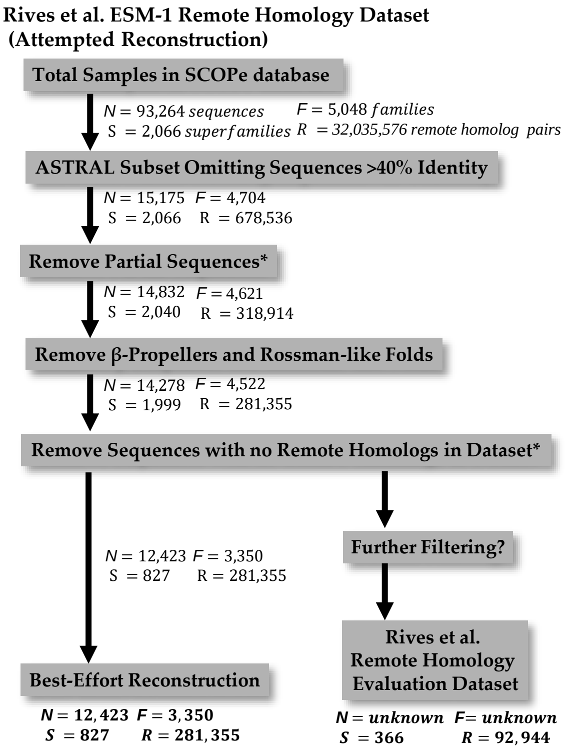
Rives et al. ESM-1 Remote Homology Dataset
(Attempted Reconstruction)
Total Samples in SCOPe database
N = 93,264 sequences F = 5,048 families
S = 2,066 superfamilies R = 32,035,576 remote homolog pairs
ASTRAL Subset Omitting Sequences >40% Identity
N = 15,175 F = 4,704
S = 2,066 R = 678,536
Remove Partial Sequences*
N = 14,832 F = 4,621
S = 2,040 R = 318,914
Remove β-Propellers and Rossman-like Folds
N = 14,278 F = 4,522
S = 1,999 R = 281,355
Remove Sequences with no Remote Homologs in Dataset*
N = 12,423 F = 3,350
S = 827 R = 281,355
Further Filtering?
Best-Effort Reconstruction
N = 12,423 F = 3,350
S = 827 R = 281,355
Rives et al.
Remote Homology
Evaluation Dataset
N = unknown F= unknown
S = 366 R = 92,944

Figure 1: Our attempt to reconstruct the remote homology evaluation dataset used by Rives et al. Asterisks denote filtering steps that were not explicitly mentioned but which we executed in an attempt to bring the number of samples closer to those reported by Rives et al.

Computing Sequence Similarity

To enable our analysis of the PLM representations of remote homologs, we compute pairwise sequence alignments and identity scores for each of the $2 \times \binom{N}{2}$ pairs of sequences in the SCOP2 database. (Unlike the above ASTRAL subsets available for SCOPe, we needed to compute these manually for SCOP2.) To compute these, we used Biopython's(Cock, Antao et al. 2009) pairwise alignment tool with default parameters.

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 subsampling 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.

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 .

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) (Davis and Goadrich 2006). We also report DeLong variances in the AUROC(DeLong, DeLong, and Clarke-Pearson 1988).
2. Hit-10 (Ma, Wang et al. 2014) is the percentage of queries for which a true homolog was in the top-10 sequences with the most similar embeddings.

Comparative Analysis using Both Procedures

Using our enhanced dataset construction and evaluation procedure to measure Hit-10 and AUC, we are able to observe diminished performance beginning at 35% sequence similarity. These metrics are shown in Figure 3.

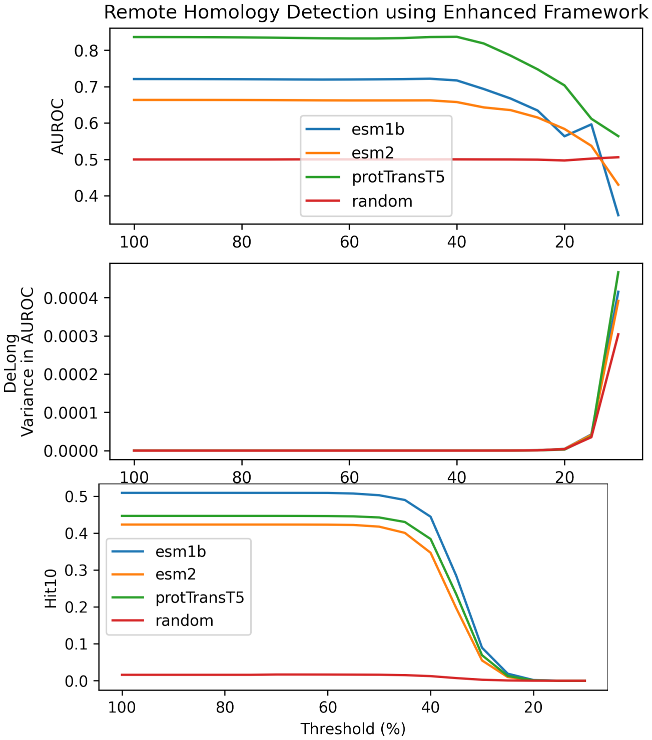

Figure 3: AUROC, DeLong variance in AUC, and Hit-10 score using embedding similarity as a predictor of homology for embeddings from all three PLMs, as the filtering sequence identity threshold is decreased from 100% to 10%. A threshold of 100% indicates no filtering beyond removal of sequences in the same family as the query, following Eq. 1. PLM embeddings of each sequence from the sequences in SCOP2 are used as queries.

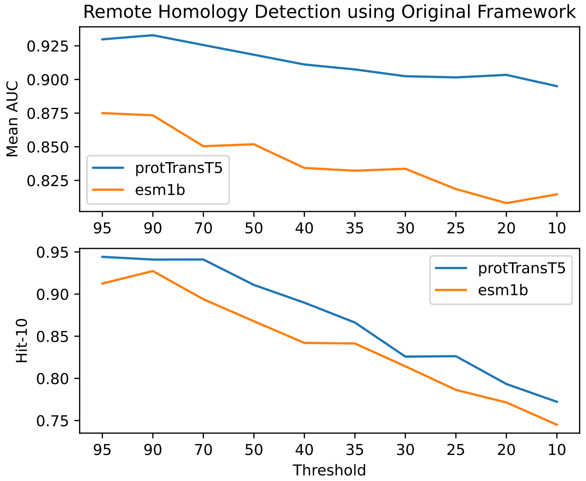

Figure 2: Remote homology Detection Performance when using the attempted reconstruction of the dataset and evaluation in Rives et al. with the addition of multiple sequence-identity thresholds.

By contrast, we can see that these weak points are hidden when using our reconstruction of the original dataset and evaluation procedure, whose results are shown for ESM-1b and ProtTransT5 in Figure 2. The reason for the drastic performance differences between the two datasets are unclear, and further analysis must be done to determine whether it stems from the dataset or evaluation procedures. A few potential sources for the discrepancy could be: (1) A difference in alignment algorithm and scoring function when computing identity scores between ASTRAL and biopython, (2) the use of macro-averaged AUC rather than calculating the AUC across all ground truths and predictions at once as we did in our enhanced procedure, (3) our use of random sampling when selecting queries for the reconstructed dataset rather than using every available sequence as a query, (4) drastic unexplained differences between the distributions of sequences included in the SCOP2 and SCOPe databases.

These results highlight the degree to which dataset selection and evaluation procedures affect performance in remote homology detection. In follow-up work, we will investigate each of these possibilities to determine the exact causes for the performance discrepancies and what they can tell us about the specific success and failure cases of PLMs in detecting remote homology.

Data Availability

The code to download and reproduce the datasets in this paper can be found at: https://github.com/amoldwin/plm-remote-homology-evaluation

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

Andreeva, A.; Howorth, D.; et al. 2013. SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Research*, 42(D1): D310–D314.

Andreeva, A.; Kulesha, E.; 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): D376–D382.

Bepler, T.; and Berger, B. 2021. Learning the protein language: Evolution, structure, and function. *Cell Syst*, 12(6): 654–669.e3.

Chandonia, J.-M.; Guan, L.; 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): D553–D559.

Chen, J.; Guo, M.; et al. 2016. A comprehensive review and comparison of different computational methods for protein remote homology detection. *Briefings in Bioinformatics*, 19(2): 231–244.

Cock, P. A.; Antao, T.; et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioniformatics*, 25: 1422–1423.

Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q. V.; and Salakhutdinov, R. 2019. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context.

Davis, J.; and Goadrich, M. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240.

DeLong, E. R.; DeLong, D. M.; and Clarke-Pearson, D. L. 1988. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3): 837–845.

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. 2022a. 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): 7112–7127.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; et al. 2022b. 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): 7112–7127.

Fox, N. K.; Brenner, S. E.; and Chandonia, J.-M. 2013. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, 42(D1): D304–D309.

Heinzinger, M.; Elnaggar, A.; Wang, Y.; Dallago, C.; Nechaev, D.; Matthes, F.; and Rost, B. 2019. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20(723): 1–17.

Hie, B.; Yang, L.; K., K. K.; and Kim, P. S. 2022. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Syst*, 13(4): 274–285.e6.

Kabir, A.; and Shehu, A. 2022. Transformer Neural Networks Attending to Both Sequence and Structure for Protein Prediction Tasks. In *Intl Conf on Knowledge Graphs (ICKG)*. IEEE. Accepted.

Ma, J.; Wang, S.; et al. 2014. MRFalign: Protein Homology Detection through Alignment of Markov Random Fields. In Sharan, R., ed., *Research in Computational Molecular Biology*, 173–174. Cham: Springer International Publishing. ISBN 978-3-319-05269-4.

Nambiar, A.; Liu, S.; Hopkins, M.; Heflin, M.; Maslov, S.; 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)*, 1–8. ACM.

Remmert, M.; Biegert, A.; Hauser, A.; and Söding, J. 2011. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, 9: 173–5.

Rives, A.; Meier, J.; 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).

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng*, 12(2): 85–94.

Stärk, H.; Dallago, C.; Heinzinger, M.; and Rost, B. 2021. Light attention predicts protein location from the language of life. *Bioinformatics Adv*, 1(1): vbab035.

Strodthoff, N.; Wagner, P.; et al. 2020. UDSMProt: universal deep sequence models for protein classification. *Bioinformatics*, 36(8): 2401–2409.

Vaswani, A.; Shazeer, N.; et al. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Verkuil, R.; Kabeli, O.; Du, Y.; Wicky, B. I. M.; Milles, L. F.; Dauparas, J.; Baker, D.; Ovchinnikov, S.; Sercu, T.; and Rives, A. 2022. Language models generalize beyond natural proteins. *bioRxiv*.