Fine-Tuning Protein Language Models Unlocks the Potential of Underrepresented Viral Proteomes

Rajan Sawhney[1], Barbra Ferrell[2, 5], Thibaut Dejean[1], Zachary Schreiber[2, 5], William Harrigan[3], Shawn W. Polson[2, 5], K. Eric Wommack[4, 5], and Mahdi Belcaid[1]

[1]**Department of Information and Computer Sciences, University of Hawai'i at Manoa**
[2]**Department of Computer and Information Sciences, University of Delaware**
[3]**Hawai'i Institute of Marine Biology, University of Hawai'i at Manoa**
[4]**Department of Plant and Soil Sciences, University of Delaware**
[5]**Delaware Biotechnology Institute, University of Delaware**

Corresponding author:
Mahdi Belcaid[1]

Email address: mahdi@hawaii.edu

ABSTRACT

Protein language models (pLMs) have revolutionized computational biology by generating rich protein vector representations—or embeddings—enabling major advancements in de novo protein design, structure prediction, variant effect analysis, and evolutionary studies. Despite these breakthroughs, current pLMs often exhibit biases against proteins from underrepresented species, with viral proteins being particularly affected—frequently referred to as the "dark matter" of the biological world due to their vast diversity and ubiquity, yet sparse representation in training datasets. Here, we show that fine-tuning pre-trained pLMs on viral protein sequences, using diverse learning frameworks and parameter-efficient strategies, significantly enhances representation quality and improves performance on downstream tasks. To support further research, we provide source code for fine-tuning pLMs and benchmarking embedding quality. By enabling more accurate modeling of viral proteins, our approach advances tools for understanding viral biology, combating emerging infectious diseases, and driving biotechnological innovation.

1 INTRODUCTION

Proteins are the molecular machinery of life, serving virtually every essential biological process. Understanding the structural and functional properties of proteins is fundamental to advancing biology and medicine. Artificial intelligence (AI) is profoundly enhancing our insights into protein mechanisms, enabling highly precise predictions of protein structures, interactions, and activities directly from sequence data. Deep learning models, such as AlphaFold (Senior et al., 2020; Jumper et al., 2021; Abramson et al., 2024), have achieved remarkable accuracy in structure prediction, revealing active sites, binding pockets, and interaction interfaces essential for understanding protein roles (Jumper et al., 2021). Furthermore, AI-driven tools accelerate the discovery of functional motifs and domains, facilitating experimental validation and the design of proteins with customized properties (Ferruz and Höcker, 2022). By augmenting empirical methods, AI is transforming proteomics, paving the way for advancements in drug discovery, biotechnology, and synthetic biology (Qureshi et al., 2023; Holzinger et al., 2023).

Central to these advancements are protein language models (pLMs), which apply principles of large language models to protein sequences. By learning the "grammar" of protein sequences, pLMs capture the complex evolutionary relationships, and functional properties of proteins as presented through structural information, all without explicit targeted training towards any singular goal. At the core of pLM breakthroughs are embeddings—vector representations of protein sequences learned by the model during training. Protein embeddings encapsulate rich, complex biological information (Lin et al., 2023),

thereby improving performance across bioinformatics applications such as sequence alignment (Harrigan et al., 2024; McWhite et al., 2023), function annotation (Flamholz et al., 2024; Hu et al., 2022b), enzyme classification (Teukam et al., 2024), and evolutionary analysis (Hie et al., 2024). Additionally, embeddings allow for large-scale analyses that surpass the computational limitations of traditional methods. For example, embedding-based protein similarity replicates evolutionary insights typically derived from computationally expensive phylogenetic approaches while efficiently processing extensive sequence datasets (Lupo et al., 2022; Tule et al., 2025).

Various machine learning architectures drive deep learning models of proteins, most notably convolutional neural networks (CNNs) (Zeng et al., 2016; Ragoza et al., 2017), recurrent neural networks/mLSTM (UniRep, Alley et al. (2019)), bi-LSTM (Zhang et al., 2022), shallow neural networks (using approaches like Word2Vec, Sharma et al. (2021)), and the currently prevalent Transformer-based architectures (Vaswani, 2017). Successful training of pLMs requires vast datasets of protein sequences. Transformers are pre-trained in an unsupervised fashion, with the aim of capturing intrinsic properties of protein sequences which addresses the scarcity of fully annotated protein data. For example, Transformer-based pLMs like ProtBert, ProtT5 (Elnaggar et al., 2021), and the ESM family (Rives et al., 2021; Lin et al., 2023) are trained using a Masked Language Modeling (MLM) objective (Devlin et al., 2019) where certain amino acids are masked and predicted based on surrounding context. Other models, such as ProGen (Madani et al., 2023; Nijkamp et al., 2023), ProtGPT (Ferruz et al., 2022), RITA (Hesslow et al., 2022) and DARK (Moffat et al., 2022) use autoregressive Causal Language Modeling approaches for predicting the next amino acid in the sequence (Bahdanau, 2014).

Unfortunately, the performance of pLMs and the quality of their generated embeddings varies considerably based on the similarity of the taxonomic levels represented in their training data (Ding and Steinhardt, 2024). For example, protein sequences from certain species are assigned higher pLM likelihoods, reflecting biases introduced by the imbalanced representation of species within training datasets (Ding and Steinhardt, 2024). This bias can, in turn, adversely affect downstream applications or render them suboptimal (Buolamwini and Gebru, 2018; Chen et al., 2018; Kleinberg et al., 2022; Shahbazi et al., 2023). Efforts to mitigate these issues through dataset re-weighting and diversity enhancement have shown promise in improving model performance (Zhao et al., 2017; Ryu et al., 2017; Yang et al., 2020; Gao et al., 2020; Rolf et al., 2021; Lee et al., 2022). However, despite these efforts and multi-tiered database sampling strategies (Elnaggar et al., 2021; Rives et al., 2021; Lin et al., 2023), the unbalanced species distribution in protein databases, like UniProt (The UniProt Consortium, 2024), often results in similar biases across pLMs. Microbes are particularly impacted by these biases, with many microbial species being underrepresented in large protein datasets (Ding and Steinhardt, 2024; Paez-Espino et al., 2016; Youle et al., 2012; Mahmoudabadi and Phillips, 2018). Bacteriophage proteins, such as those targeting ESKAPE pathogens (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa*, and *Enterobacter* spp.) (Lee et al., 2024), are frequently underrepresented or excluded due to limited annotation (Sanejouand, 2023). This lack of representation can limit the ability of pLMs to generalize across diverse microbial taxa, particularly viruses, which make up only a small fraction of public protein databases despite their ubiquity in biological systems (Moreno-Gallego and Reyes, 2021).

Fine-tuning—a process of further training pre-trained models on domain-specific datasets—can mitigate biases by refining embeddings to capture diverse sequences and context-specific features. Recent studies suggest this approach has been successful in natural language processing and proteomics (Ding and Steinhardt, 2024; Elnaggar et al., 2021; Rao et al., 2019; Gu et al., 2021; Brandes et al., 2022; Rives et al., 2021; Howard and Ruder, 2018; Devlin et al., 2019; Houlsby et al., 2019). However, none of these studies have investigated the impact on viruses. Fine-tuning for enhanced representation learning could potentially be important for viral proteins, where optimization might capture distinct patterns of viral proteins and enhance performance across diverse tasks.

A pLM can be fine-tuned by updating all model weights (full fine-tuning), or selectively updating specific parameters of the model. Full fine-tuning models with extremely large parameter counts have large computational demands (100's of GB of RAM), making it infeasible for most standard systems. For example, training the ESM2-15B model took 60 days using over 512 NVIDIA V100 GPUs (Lin et al., 2023). However, Low-Rank Adaptation (LoRA) (Hu et al., 2022a), a variant of Parameter Efficient Fine-Tuning (PEFT) (Mangrulkar et al., 2022), selectively trains specific components of a model, dramatically reducing the number of trainable parameters and computational requirements. Specifically, LoRA

decomposes the model weight matrices into a pair of smaller, low-rank matrices, reducing both memory and computational costs. LoRA allows fast, efficient adaptation without additional inference latency, making it ideal for large models. Using rank r, LoRA adjusts only a small subset of parameters during fine-tuning, mitigating catastrophic forgetting (French, 1999; Kirkpatrick et al., 2017) and alleviating the RAM burden as pLMs scale in size. A rank of 8, as demonstrated by Hu et al. (2022a), achieves competitive performance when fine-tuning while keeping computational costs low. Higher ranks may lead to better performance in some cases, but they also increase the computational overhead.

This research evaluates the impact of LoRA fine-tuning with three popular representation learning approaches—namely masked language modeling, classification, and contrastive learning—on improving the embedding quality for viral proteins. We compared pre-trained and LoRA-fine-tuned versions of three different and widely popular pLMs: ESM2-3B, ProtT5-XL, and ProGen2-Large. An additional layer with diverse unsupervised or self-supervised learning objectives was uniformly applied across all models, refining learned features to domain-specific applications (Devlin et al., 2019; Raffel et al., 2020). This standardization enabled fair performance comparisons, minimizing architectural biases (Radford et al., 2019; Elnaggar et al., 2021).

Our study demonstrates that LoRA fine-tuning for selectively updating pLM parameters using virus-domain specific data and diverse learning objectives enhances downstream bioinformatics performance. All model adaptation and evaluation code is available open-source at https://github.com/Hawaii-Bioinformatics/ViralFineTuning, offering a practical framework for groups with limited computational resources to leverage pLM adaptation.

2 METHODS

Data

Models were fine-tuned on viral protein sequences from the Virus Orthologous Groups Database (VOGDB) (http://vogdb.org) (Trgovec-Greif et al., 2024). VOGDB offers an opportunity for applying both unsupervised and supervised learning approaches by leveraging the orthologous group information provided within the same dataset, facilitating comparisons between pre-trained and fine-tuned models. The VOGDB sequence dataset consisted of 381,760 sequences distributed across 24,916 orthologous groups. The dataset was filtered to only include groups containing over 10 sequences, resulting in 345,261 sequences in 5,981 orthologous groups.

The filtered VOGDB sequence dataset was divided into training and testing subsets in a 90:10 ratio (313,873 training sequences and 31,388 test sequences) using a stratified sampling strategy; all orthologous groups were proportionately represented in the training and test subsets.

For model evaluation, we used four datasets: (1) the VOGDB test subset, generated by selecting 673 VOGs, each containing at least 10 sequences, from the filtered VOGDB sequence dataset. (2) A multiple sequence alignment (MSA) dataset from VOGDB, generated by selecting a subset of 300 VOGs with each VOG's MSA containing 60 to 200 alignments. (3) HeliBase, a dataset of 174 manually curated reference DNA helicase sequences identified through a combination of historical and recent literature, published metagenomes, and biochemical characterizations (https://github.com/ud-veil/HeliBase) (Brosh Jr and Matson, 2020; Singleton et al., 2007; Nasko et al., 2018; Keown, 2024) . (4) The UvsW helicase dataset, containing 47 sequences sourced from Swiss-Prot (The UniProt Consortium, 2024).

Model Training

Our study used three popular pre-trained pLMs: ESM2-3B, ProtT5-XL and ProGen2-Large (Table 1). For each, we chose a variant with approximately 3 billion parameters, balancing performance and computational resource requirements, making the model appropriate for a wide range of research environments.

The original pLMs were fine-tuned using three approaches: (1) masked language modeling (MLM), (2) classification with cross-entropy loss (CLS), and (3) contrastive learning with cosine mean squared error loss (CON) (Fig. 1).

1. Masked Language Modeling: Transformer models, such as ESM and ProtT5, leverage MLM (introduced in BERT, Devlin et al. (2019)) for predicting masked tokens within sequences, and capturing bidirectional context to learn relationships essential for tasks like function prediction and structure analysis. MLM is optimized using cross-entropy loss:

Model	Number of parameters	Max. sequence length	Encoded layers	Embedding size	Database
ESM2-3B	3B	1,024	36	2,560	Uniref50/90
ProtT5-XL	3B	512	24	1,024	BFD100/UniRef50
ProGen2-Large	2.7B	1,024	32	2,560	UniRef90

Table 1. Protein language models applied in the study.

$$Loss = -\sum_{i=1}^{N} \log(P(true_token_i | context))$$

Where N is the number of masked tokens, and $P(true_token_i|context)$ is the probability assigned to the true token at the masked position i. Following the approach described in Devlin et. al (2019), we masked 15% of tokens in each protein sequence at random, with masking probability of 80% for ESM2 and 100% for ProtT5 and ProGen2. We computed the loss between the predicted tokens and original tokens using the cross-entropy loss function.

2. Classification with cross-entropy loss: Classification-based self-supervised methods are often used due to the availability of large datasets that group proteins into families or orthologous groups. The training was designed to predict an input protein's most likely orthologous group. The model was trained using the cross-entropy loss function, which measured the difference between the predicted probability distribution and the true orthologous group label.

$$Loss = -\sum_{i=1}^{N} \log(P(true_group_i | x_i))$$

Where N is the number of samples, $true_group_i$ is the true label, and $P(true_group_i|x_i)$ is the predicted probability of the true group given input x_i.

3. Contrastive Learning: Contrastive learning is another powerful representation learning method that trains models to embed similar protein sequences closer together in feature space while positioning dissimilar sequences further apart. Siamese networks, a contrastive learning architecture, can be particularly effective for tasks comparing or ranking sequences, such as protein sequence alignment or similarity search. We fine-tuned models with Siamese networks optimized using a hybrid loss function combining cosine similarity and mean squared error (MSE).

$$Loss = \alpha \cdot CosineLoss + (1-\alpha) \cdot \frac{1}{n} \sum_{i=1}^{n} (A_i - B_i)^2$$

Where A and B are the two vectors being compared, and α is a weight factor balancing the cosine and MSE loss. The Siamese network was trained on triplets of protein sequences: an anchor, a positive (a sequence from the same orthologous group), and a negative (a sequence from a different orthologous group) according to the method described in Schroff et al. 2015. We used the cosine MSE loss for computing the similarity between the embeddings. Specifically, the cosine similarity was calculated for the anchor–positive and anchor–negative pairs, and the MSE was applied to encourage anchor–positive similarity to be close to 1 and anchor–negative similarity to approach 0.

These methods can also be integrated. For example, XLNet (Yang et al., 2019) combined autoregressive objectives while adhering to the contrastive learning framework. Through these representation learning methods, the model captures complex biological relationships, performing optimally on downstream tasks relevant to viral protein function and mutation analysis.

Training and Implementation Specifications

All models were implemented in PyTorch (v.2.4.0), using HuggingFace implementations of ESM2 (facebook/esm2_t36_3B_UR50D) and ProtT5 (Rostlab/prot_t5_xl_uniref50) from the transformers package (v.4.42.4), ProGen2 from its github repository (https://github.com/enijkamp/progen2), and LoRA from the peft package (v.0.10.0) (https://pypi.org/project/peft). Models were trained using an NVIDIA A6000 graphics card with 48GB memory and an NVIDIA A100 with 80GB memory. Model weights were

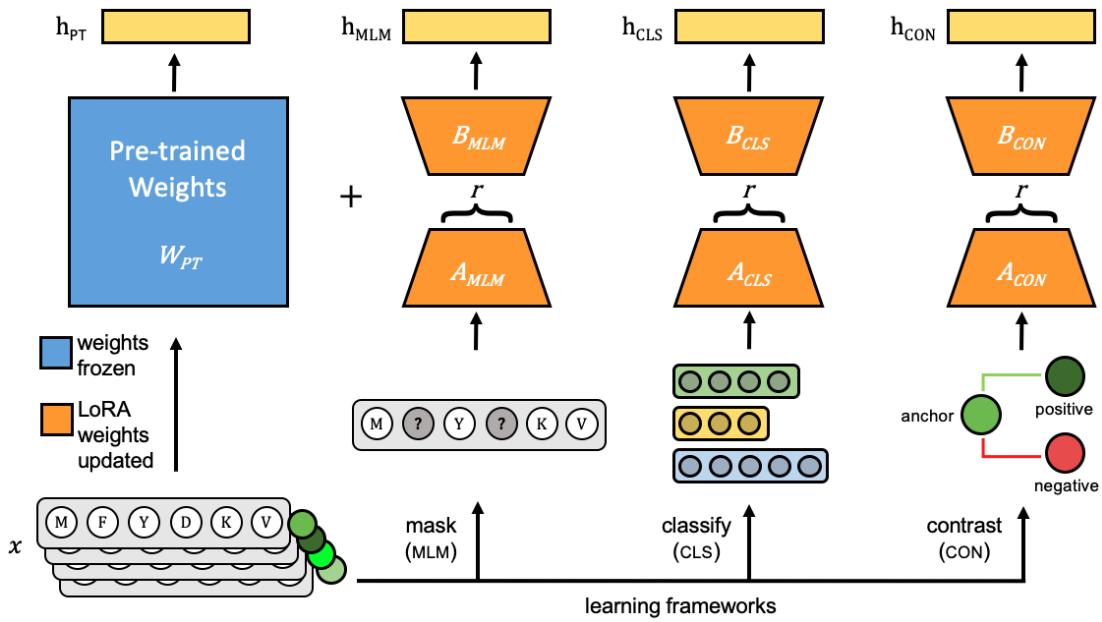


Figure 1. Each pre-trained protein language model is fine-tuned on viral protein sequences (x) using Parameter-Efficient Fine-tuning (PEFT)'s LoRA method. During fine-tuning, pre-trained weights of the pre-trained model (W_{PT}), shown in blue, are frozen and only weights for LoRA-trainable parameters, shown in orange, are updated using one of three learning frameworks: Masked Language Modeling (MLM), classification with cross-entropy loss (CLS), and contrastive learning (CON). During inference, h_{PT} is the output of the pre-trained model, and the frozen pre-trained weights and LoRA-learned weights are added together to generate model outputs indicated as h_{MLM}, h_{CLS} and h_{CON}.

optimized via backpropagation using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of $1e-9$ for epochs ranging from 2 to 10. We used LoRA $r = 8$ to fine-tune the entire updated architecture (pLM and PEFT layers). LoRA matrices were added to query, key, and value weights (*Wq, Wk*, and *Wv*, respectively) and learned through the variant's newly incorporated representation learning frameworks. Model parallelism was used for ESM2-3B given its memory demands with parameters assigned to designated GPUs. Table 2 provides model training specifications and their average loss values.

Protein embeddings

Protein embeddings were generated by tokenizing each residue in the protein sequence and using the weights from the model's last hidden layer to extract a vector for each token. Each fine-tuned model's LoRA-learned weights are added to the pre-trained weights for extracting residue vector representations. The dimensionality of these vectors was defined by the model's embedding size (Table 1). Padding and special tokens were removed for matching the number of vectors to the residues. Sequence-level embeddings, or pooled embeddings, were obtained by averaging the amino acid embeddings across the sequence. All test sequences were embedded prior to evaluation, except for vector clustering alignments (vcMSA, McWhite et al. (2023)) alignments, where embeddings were generated on the fly.

Model evaluation

Our model evaluation presumed that improvements in protein embedding quality will be reflected in both statistical evaluations and downstream tasks. Specifically, we assessed the embeddings generated from each pre-trained and fine-tuned model using pairwise comparisons, clustering, and alignment-based approaches to determine whether fine-tuning improved the overall quality of the embeddings.

Experiment 1: Pairwise comparison-based experiments

We hypothesized that fine-tuning results in high quality embeddings that lead to more accurate and effective detection of similarity when compared with the pre-trained model. This hypothesis was tested using three pairwise comparisons: (1) sequence similarity across proteins within the same orthologous

Model	Storage size (MB)	Training with Adam (GB)	Trained parameters (LoRA)	Training time (hrs)	Train loss	Test loss	Hardware (GPU)
ESM2-3B$_{MLM}$	51	42.34	4.42M	127	0.048	0.051	1 Nvidia A6000
ESM2-3B$_{CLS}$	52	42.34	4.42M	146	6.424	8.731	1 Nvidia A6000
ESM2-3B$_{CON}$	62	51.48	4.99M	150	0.063	0.085	2 Nvidia A6000
ProtT5-XL$_{MLM}$	97	24.60	8.85M	48	0.011	0.124	4 Nvidia A100
ProtT5-XL$_{CLS}$	97	34.50	8.85M	136	5.283	8.624	1 Nvidia A6000
ProtT5-XL$_{CON}$	48	39.40	4.09M	150	0.117	0.127	1 Nvidia A6000
ProGen2-Large$_{MLM}$	46	22.40	3.93M	100	0.862	0.874	1 Nvidia A6000
ProGen2-Large$_{CLS}$	46	23.10	3.93M	100	5.321	8.813	1 Nvidia A6000
ProGen2-Large$_{CON}$	48	25.40	4.09M	120	0.864	0.961	1 Nvidia A6000

appended subscripts indicate the model's training type: masked language modeling (MLM), classification (CLS), or contrastive (CON).

Table 2. Fine-tuning specifications for the protein language models.

group; (2) similarity at conserved positions in MSAs; and (3) similarity at non-homologous positions in MSAs. All three experiments utilized cosine similarity, a robust metric for comparing embeddings based on vector orientation, which avoided biases related to sequence length and composition.

1.1 Pairwise sequence comparisons with pooled embeddings

To evaluate whether fine-tuning improved the quality of pooled embeddings, we computed pairwise cosine similarity between the 10 members of each of 673 VOGs in the VOGDB test subset (30,285 within-family pairwise comparisons).

1.2 Similarity within Conserved MSA Columns

To investigate conserved and non-conserved MSA positions, from the VOGDB MSA dataset, we selected 300 VOGs, with each VOG containing between 60 and 200 alignments. Conserved positions were defined as those with fewer than 5% unknown residues (denoted as "X" according to IUPAC standards), consisting of at most four different amino acid types, and with at least 70% of the residues in the column composed of the most abundant amino acid. Each MSA could contain multiple conserved columns. For each conserved column, we calculated the pairwise amino acids embedding cosine similarity. This analysis resulted in 1,145,044 pairwise comparisons across 193 conserved columns, covering 5,323 sequences from 51 VOGs.

1.3 Similarity Across Random Sites

The ability of fine-tuned models to discriminate between non-homologous positions across sequences, was evaluated using an iterative sampling method for sequences within an MSA. For each test iteration, embeddings of amino acid pairs derived from distinct positions and sequences were compared, ensuring that these amino acids did not originate from conserved columns and were separated by at least 20 positions. More formally, for each iteration, a random position p_i is selected for each sequence s_i and p_j for each sequence s_j, such that $p_i \neq p_j$ and $|p_i - p_j| > 20$ for all $i \neq j$. This sampling methodology ensured that embeddings were compared between non-homologous, well-separated positions across distinct sequences. This procedure yielded a total of 1,049,148 pairwise cosine similarity comparisons across 19,071 sequences from 196 VOGs.

Experiment 2. Clustering-based experiments

An iterative clustering analysis using pooled sequence embeddings was used for determining whether fine-tuning produces more distinct and biologically meaningful clusters compared to pre-trained models. Specifically, for each of the 1000 iterations per model variant (pre-trained and fine-tuned), 500 VOGs were randomly sampled from the VOGDB test subset including all their corresponding sequences (at least 10 per group). Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm (Campello et al., 2013; McInnes et al., 2017) was applied for clustering pooled embeddings

of sequences from these selected groups. Cluster quality was assessed by computing their silhouette scores, providing a quantitative measure of distinctiveness and cohesion. The highest silhouette score achieved across all iterations for each model served as a metric for evaluating the model's ability to discern biologically meaningful distinctions among viral proteins.

HDBSCAN parameters for each model suite were optimized individually to attain the best silhouette scores. For the ESM2-3B and ProtT5-XL models, the parameters were set to min_cluster_size=10, min_samples=5, epsilon=1.0, and alpha=1.0. In contrast, ProGen2-Large required a min_cluster_size of 20 to avoid consistent negative silhouette scores. The silhouette score was computed as:

$$S(i) = \frac{(b(i) - a(i))}{max(a(i), b(i))}$$

where a(i) is the mean intra-cluster distance and b(i) is the mean nearest-cluster distance. The overall silhouette score, computed as the mean of S(i) across all points, ranges from -1 to 1, with higher values indicating superior clustering.

Further, the ability of fine-tuned models to improve protein family and function identification was assessed through clustering of diverse DNA helicase sequences within HeliBase. Given the limitations of traditional homology-based methods in precise family-level functional annotation (Flamholz et al., 2024), helicases served as an effective candidate for evaluating model's efficacy in achieving precise family-level assignments.

Helicase embeddings derived from pre-trained and fine-tuned models were converted to two-dimensional vectors using Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP, Sainburg et al. (2021)). HDBSCAN (min_samples=5, min_cluster_size=5) was used for clustering UMAP representations. To compare the performance of embeddings with conventional phylogenetic approaches, a phylogenetic tree was constructed alongside Pfam domain analysis of all helicases. Using InterProScan (version 5.67-990) (Jones et al., 2014) and the Pfam database, significant domain hits, with start and stop regions were obtained. Full length helicase sequences were aligned with MAFFT (–auto mode, version 7) (Katoh and Standley, 2013), and an approximate maximum-likelihood tree was produced using FastTree (version 2.1) (Price et al., 2010) (Price et al., 2010) and annotated with domain information in Iroki (Moore et al., 2020).

Experiment 3: Alignment-based experiments

These experiments evaluated whether fine-tuning improved the quality of protein embeddings specifically in the context of alignment-based methods, including both pairwise soft alignment (Harrigan et al., 2024) and vector-clustering MSA (vcMSA, McWhite et al. (2023)) approaches. It was expected that fine-tuned embeddings would yield longer soft alignments of homologous residues and improved vcMSA consistency as compared with pre-trained models. Such an outcome would demonstrate a model's enhanced ability for capturing biologically meaningful relationships within and between protein sequences.

3.1 Soft Alignments

Pairwise soft alignments were generated using amino acid-level embeddings for sequence pairs belonging to the same VOG, derived from the VOGDB test subset (resulting in a total of 30,285 pairwise soft alignments) (Harrigan et al., 2024). Resulting pairwise alignments were filtered to retain only those exhibiting path length (longest homologous residue path) exceeding 18 for sequences shorter than 1024 residues for all models. We used the length of the soft alignment path as a metric for comparing the quality of embeddings produced by each model, with the expectation that fine-tuned embeddings would result in longer paths of protein sequences exhibiting remote homology.

3.2 Multiple Sequence Alignments

MSAs were generated using the vcMSA algorithm (McWhite et al., 2023) on amino acid-level embeddings of sequences from 100 randomly sampled VOGs from the VOGDB test subset. A maximum of 150 sequences were randomly selected from these VOGs and were shuffled, increasing alignment difficulty and simulating realistic alignment challenges. The vcMSA algorithm, which originally used the pre-trained ProtT5 model for generating amino acid-level embeddings, was modified to incorporate all models developed in this study. Fine-tuning should enhance the accuracy of vcMSA alignments by refining the contextual embeddings used for sequence clustering and ordering, thus yielding alignments with greater

biological relevance compared to those derived from pre-trained embeddings. The generated vcMSAs were evaluated using T-Coffee's (Notredame et al., 2000) Transitive Consistency Score (TCS) (Chang et al., 2014), a metric that measures the reliability and consistency of aligned sequences across various alternative alignments.

Further, as a focused analysis of UvsW helicase proteins from bacteriophages, we evaluated the quality of the vcMSA using Mutual Information (MI) and occupancy, comparing against conventional alignment methods (EMBOSS Needle pairwise alignments and Clustal Omega MSAs, Madeira et al. (2024)). MI quantifies the dependency between positions in an MSA, indicating co-evolutionary relationships, while occupancy measures the proportion of sequences with residues (non-gaps) at a position, reflecting conservation and alignment completeness.

3 RESULTS

Overall, pLMs ESM2-3B, ProtT5-XL, and ProGen2-Large benefited from fine-tuning with viral sequences based on statistical analyses and the outcome of downstream tasks performed with sequence embeddings. Fine-tuning improved the overall embedding quality based on: (1) pairwise comparisons evaluating pooled embeddings and sequence similarity within conserved or non-homologous MSA positions; (2) clustering; and (3) alignments.

Experiment 1: Pairwise comparison-based experiments

Fine-tuning should improve embedding quality, resulting in increased cosine similarity between similar proteins or amino acids, and decreased cosine similarity between non-homologous proteins or amino acids. Evaluation of model performance was based on three pairwise embedding assessments: (1) pooled embedding similarity for proteins within the same VOG (30,285 comparisons across 673 groups in the VOGDB test subset); (2) amino acid-level embedding similarity at conserved MSA positions (1,145,044 comparisons across 193 conserved sites over 5,323 alignments from 51 orthologous groups); and (3) amino acid-level embedding similarity at non-homologous MSA positions (1,049,148 comparisons across 19,071 alignments from 196 orthologous groups).

Overall, fine-tuned versions of the three models outperformed their original pre-trained model, though the scale of improvement depended on the training framework. Notably, fine-tuning of the ESM2-3B pLM using the contrastive model demonstrated the most significant improvement. The pre-trained ESM2-3B model yielded consistently high cosine similarities across all tests (Fig. 2), which aligns with a known limitation wherein embeddings using this pLM yield a narrow distribution of high cosine similarity scores, regardless of the evolutionary or functional relationships between the pairs of proteins compared (Tran et al., 2023). Fine-tuning broadened this distribution, particularly in distinguishing conserved from non-conserved sites.

As illustrated in Fig. 2A, the pre-trained ESM2-3B model exhibited a median cosine similarity of 0.99 for pooled embeddings, even though the average pairwise protein similarity (global alignment using BLOSUM62 scoring matrix) within individual protein families was only 0.51 (see Supplementary Material Section 1.2). The narrow cosine similarity distribution observed in pre-trained ESM2-3B broadens significantly after fine-tuning. The impact of fine-tuning was particularly evident when comparing the cosine similarity of amino acid embeddings at conserved versus random sites (Figs. 2B and 2C). Fine-tuning ESM2-3B using the contrastive model exhibited the largest cosine similarity difference between conserved and non-conserved sites (0.14). This large difference reflected better performance in distinguishing conserved from non-conserved sites. All other models and the pre-trained model showed little difference between conserved and non-conserved sites.

Fine-tuning of the ProtT5-XL pLM resulted in only marginal improvements (Fig. 2, middle column). Across all pairwise comparisons, the changes in cosine similarity scores for all fine-tuned ProtT5-XL models were minimal compared to the pre-trained model.

Similarly, fine-tuning of Progen2-Large did little to improve performance according to cosine similarity. When looking at non-conserved regions (Fig. 2C), the median scores remained high, suggesting that fine-tuning did not enhance the model's ability to differentiate between homologous and non-homologous amino acids.

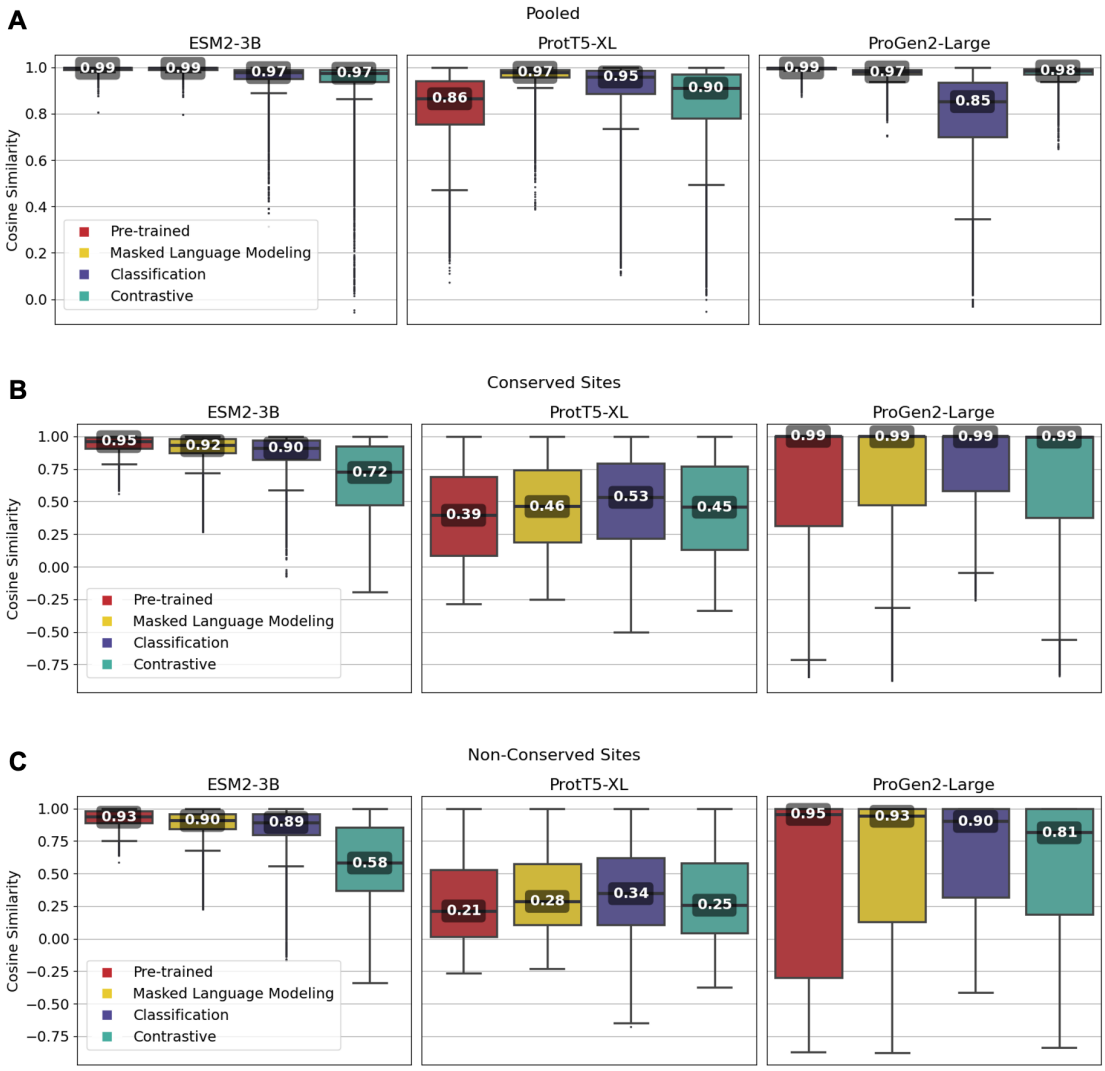


Figure 2. Pairwise cosine similarity distribution and median values for **(A)** pooled sequence embeddings of sequences within an orthologous group, **(B)** conserved sites within multiple sequence alignments, and **(C)** non-conserved sites within multiple sequence alignments.

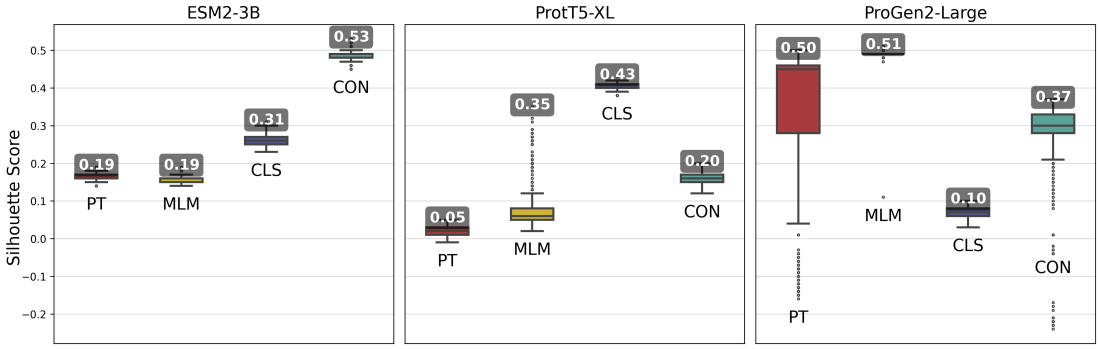
Experiment 2 : Clustering-based experiments

Pooled embeddings of sequences from randomly selected 500 VOGs from the VOGDB test subset were clustered using HDBSCAN, and cluster quality was evaluated with silhouette scores for assessment of whether fine-tuning improved formation of biologically meaningful clusters. This process was repeated over 1000 iterations, with the maximum scores used to compare each models' capability for capturing meaningful distinctions among viral proteins.

Fine-tuned ESM2-3B and ProtT5-XL models achieved the same or higher silhouette scores (Fig. 3) compared with their pre-trained pLMs. Notably, contrastive fine-tuning of ESM2-3B showed a score approximately three times higher than that of its pre-trained model (Fig. 3). Among the ProtT5-XL fine-tuning models, the classification approach yielded the highest silhouette score, but still underperformed compared to contrastive fine-tuning of ESM2-3B.

Fine-tuning approaches with the ProGen2-Large model produced widely varying changes in silhouette scores. By and large, fine-tuning approaches performed worse than the pre-trained model, indicating that additional training led to limited gains. Interestingly, selecting cluster size of minimum size 10, while optimal for the ESM2-3B and ProtT5-XL models, consistently resulted in negative silhouette scores for ProGen2-Large embeddings. Thus, increasing the minimum cluster size to 20 was necessary for this

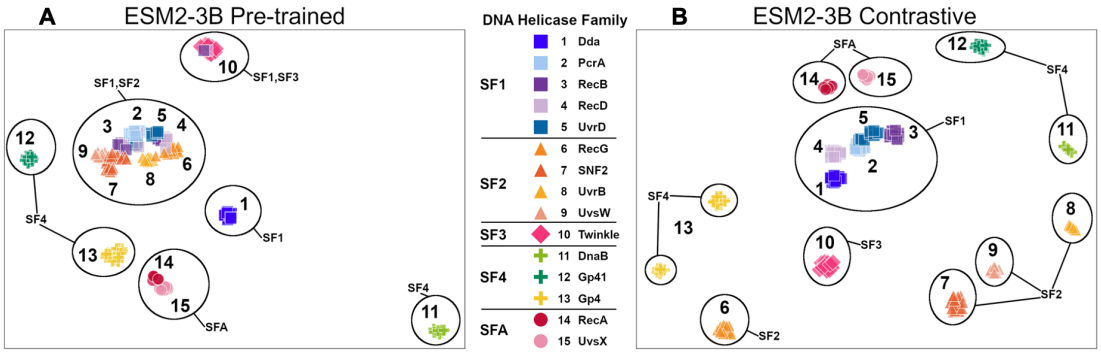
pLM.



ESM2-3B and ProtT5-XL results with min_cluster_size=10; Progen2-Large results with min_cluster_size=20; pre-trained (PT), masked language model (MLM), classification (CLS), or contrastive (CON) indicate the model's training type.

Figure 3. Silhouette scores of 1000 iterations of embedding clusters annotated with the highest score across all iterations. Each iteration used Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to generate clusters of sequence pooled embeddings from 500 randomly selected VOGs from the VOGDB test subset.

The ESM2-3B pLM, fine-tuned using the contrastive approach, was chosen for further performance evaluation based on observations of the cosine similarity (Fig. 2) and silhouette score (Fig. 3) tests. Using the contrastive ESM2-3B model we assessed whether fine-tuning improved meaningful cluster formation for well known proteins having extensive family and functional domain annotations. Comparison of HDBSCAN cluster distributions from the pre-trained (Fig. 4A) and contrastive fine-tuned ESM2-3B pLMs (Fig. 4B) for DNA helicase sequences within HeliBase demonstrated dramatic improvements in groupings according to helicase families. Member proteins within helicase families SF1 and SF2 were clearly separated using embeddings generated with the contrastive fine-tuned ESM2-3B whereas embeddings generated with the pre-trained ESM2-3B mostly overlapped for these families. Interestingly, the contrastive ESM2-3B differentiated two subgroups within the Gp4 family (cluster 13 in Fig. 4B) based on the strict presence or absence of a Pfam PF08273 zinc-binding domain, a distinction missed in more classical phylogenetic tree analysis (yellow clade, Fig. 4C).



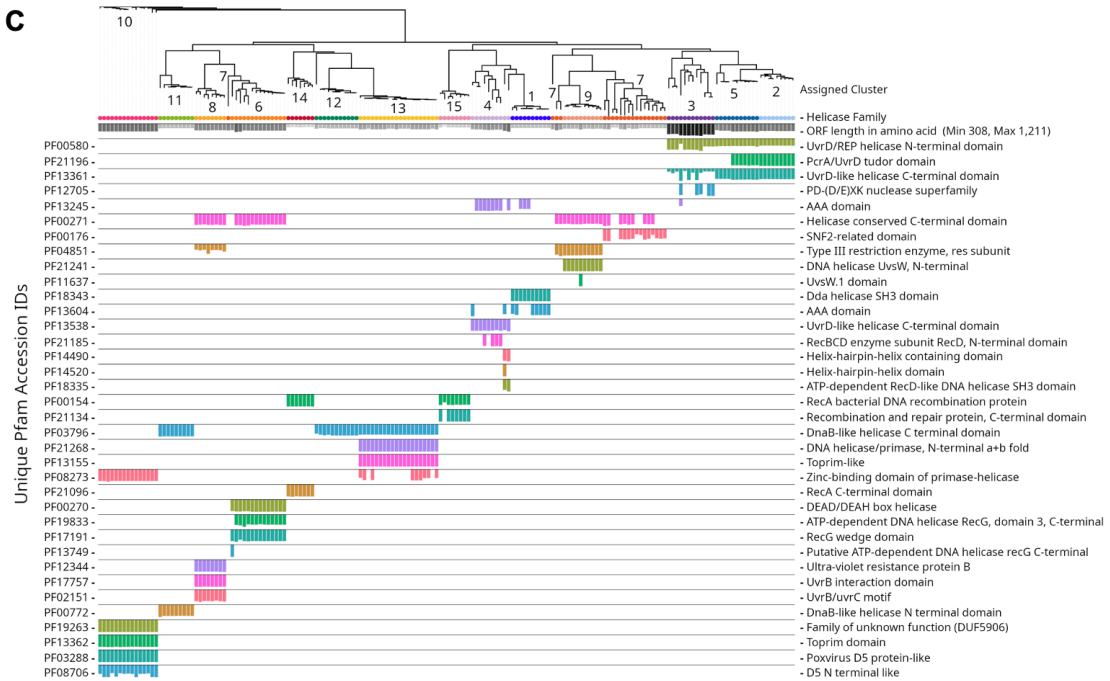


Figure 4. Family-level clusters of 174 DNA helicase sequences using embeddings from **(A)** pre-trained and **(B)** contrastive ESM2-3B models. Applying Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to (A) ESM2-3B pre-trained embeddings led to limited separation among helicase families, while (B) contrastive ESM2-3B embeddings led to clearer distinctions. **(C)** FastTree approximate maximum-likelihood tree of full length helicase sequences, along with their domain composition identified through InterProScan. Clades in the tree are annotated with HDBSCAN cluster numbers, and leaf nodes are colored to correspond to their respective helicase families from (A) and (B) using Iroki. Bars below the tree indicate relative domain coverage, labeled by Pfam identifier (left) and domain annotation (right).

Experiment 3: Alignment-based experiments

We evaluated whether fine-tuning improves protein embeddings for alignment-based tasks, including soft alignments and vcMSAs. Fine-tuned models were hypothesized to improve the identification of homologous residues, yielding longer alignments and enhancing vcMSA quality, thereby reflecting an improved ability to capture biologically meaningful evolutionary relationships.

3.1 Soft Alignments

Pairwise soft alignments were generated from amino acid-level embeddings for 30,285 sequence pairs belonging to the same VOG in the VOGDB test subset (Harrigan et al., 2024). Only those pairwise alignments with a path length longer than 18 for sequences shorter than 1024 residues were retained for model assessment. Fine-tuned models were expected to produce longer paths of homologous residues.

ESM2-3B fine-tuned using the contrastive approach outperformed the pre-trained and other fine-tuned models, achieving a mean homologous residue path length of 173 (Fig. 5). However, mean values alone offer only a partial view of the improvements. Among 27,046 alignments, 1,665 saw the residue path length double, indicating that contrastive fine-tuning substantially enhanced the quality of alignments generated using ESM2-3B embeddings. Contrastive fine-tuning also improved the ProtT5-XL and ProGen2-Large models, yet neither of these improved models demonstrated as substantial an improvement of path length as the contrastive ESM2-3B model.

Soft alignment using two Megamimivirinae putative ankyrin repeat protein sequences: AFX93168.1 from Megavirus courdo11 *(Megavirus; Megavirus chilense)* and AHA45685.1 from Hirudovirus strain Sangsue *(Mimivirus)* illustrate improvements in sequence alignment based on contrastive fine-tuning of ESM2-3B. The contrastive fine-tuned ESM2-3B model identified a homologous residue path of length

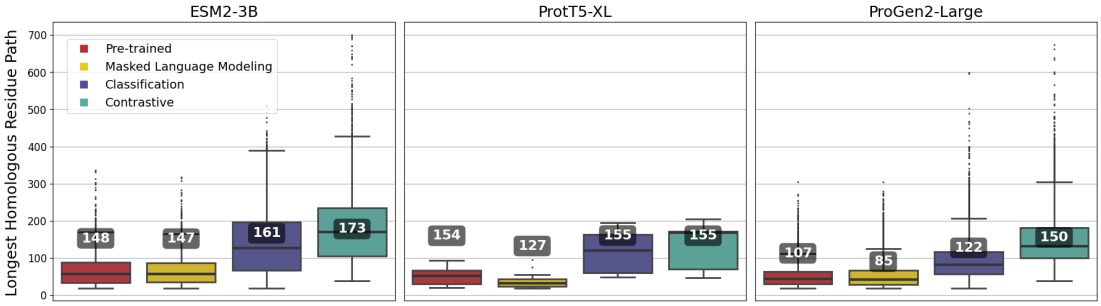


Figure 5. Longest homologous residue paths identified from pairwise soft alignments where the contrastive variant's identified lengths exceed twice the length identified by the pre-trained model. Numbers correspond to the overall mean of path lengths across all pairwise alignments.

174 for this sequence pair (Fig. 6B), significantly outperforming the pre-trained model, which yielded a path length of 71 (Fig. 6A). A BLASTP pairwise sequence alignment produced a highly fragmented result comprising 43 high-scoring pairs with an E-value of 0.089 (Fig. 6C).

3.2 Multiple Sequence Alignments

Vector-clustering MSAs were generated using amino-acid level embeddings of sequences from 100 randomly sampled VOGs from the VOGDB test subset (McWhite et al., 2023), and evaluated with T-Coffee's Transitive Consistency Score (TCS) (Notredame et al., 2000). Fine-tuned models were expected to improve alignment reliability, yielding vcMSAs with higher TCS.

Fine-tuning using the contrastive approach outperformed most other approaches for all of the pLMs, and among the pLMs, ESM2-3B produced the highest-quality vcMSAs with a mean TCS of 484 (Fig. 7).

The impact of pLM fine-tuning on vcMSAs was examined using embeddings for 47 bacteriophage proteins (UvsW helicase), with lengths ranging from 268 to 1019 residues produced by ESM2-3B fine-tuned using the contrastive approach. The vcMSA generated from embeddings using pre-trained ESM2-3B yielded a TCS of 939, an average Mutual Information (MI) of 0.16, and an average occupancy 0.51 (Fig. 8A). In contrast, alignments using embeddings from the contrastive fine-tuned ESM2-3B model resulted in a TCS of 972, average MI of 0.22, and occupancy of 0.59 (Fig. 8B). Most notably, pronounced gaps occurred in alignments from the pre-trained model (Fig. 8A). For instance, alanine residues (A) at alignment positions 28–31 and 34–37 remained unaligned due to vcMSA's assessment that their embeddings generated from the pre-trained ESM2-3B model were too distant to indicate homology. Similarly, the arginine (R) and methionine (M) at alignment positions 39–41 remained unaligned, despite multiple instances of each at these positions. Conversely, alignments produced using embeddings from contrastive fine-tuned ESM2-3B produced no gaps in these positions (Fig. 8B). This result agreed with HMM models associated with the protein family Pfam PF00271, which have co-occurrence probabilities of 0.086 and 0.068, respectively, for these positions. Additionally, the contrastive fine-tuned ESM2-3B model accurately aligned residues R and M at alignment position 28, which exhibit HMM probabilities of 0.040 and 0.011, respectively. These alignments were consistent with results obtained from EMBOSS Needle pairwise alignments and Clustal Omega MSAs (data not shown), further substantiating the reliability of vcMSA generated from embeddings obtained using the the contrastive fine-tuned ESM2-3B model.

4 DISCUSSION

Our results demonstrate that fine-tuning substantially enhances the performance of pLMs on tasks critical to understanding viral function. Notably, the contrastive fine-tuned ESM2-3B model consistently outperformed its pre-trained counterpart across pairwise sequence embedding comparisons, clustering, and alignment-based evaluations.

In pairwise embedding assessments, the broadening of cosine similarity distributions, and the improved discrimination between homologous and non-homologous sites suggests that fine-tuning allows the pLM to better represent subtle differences in sequence context. This is in line with previous work indicating that domain-specific fine-tuning can mitigate biases arising from imbalanced training datasets and enhance

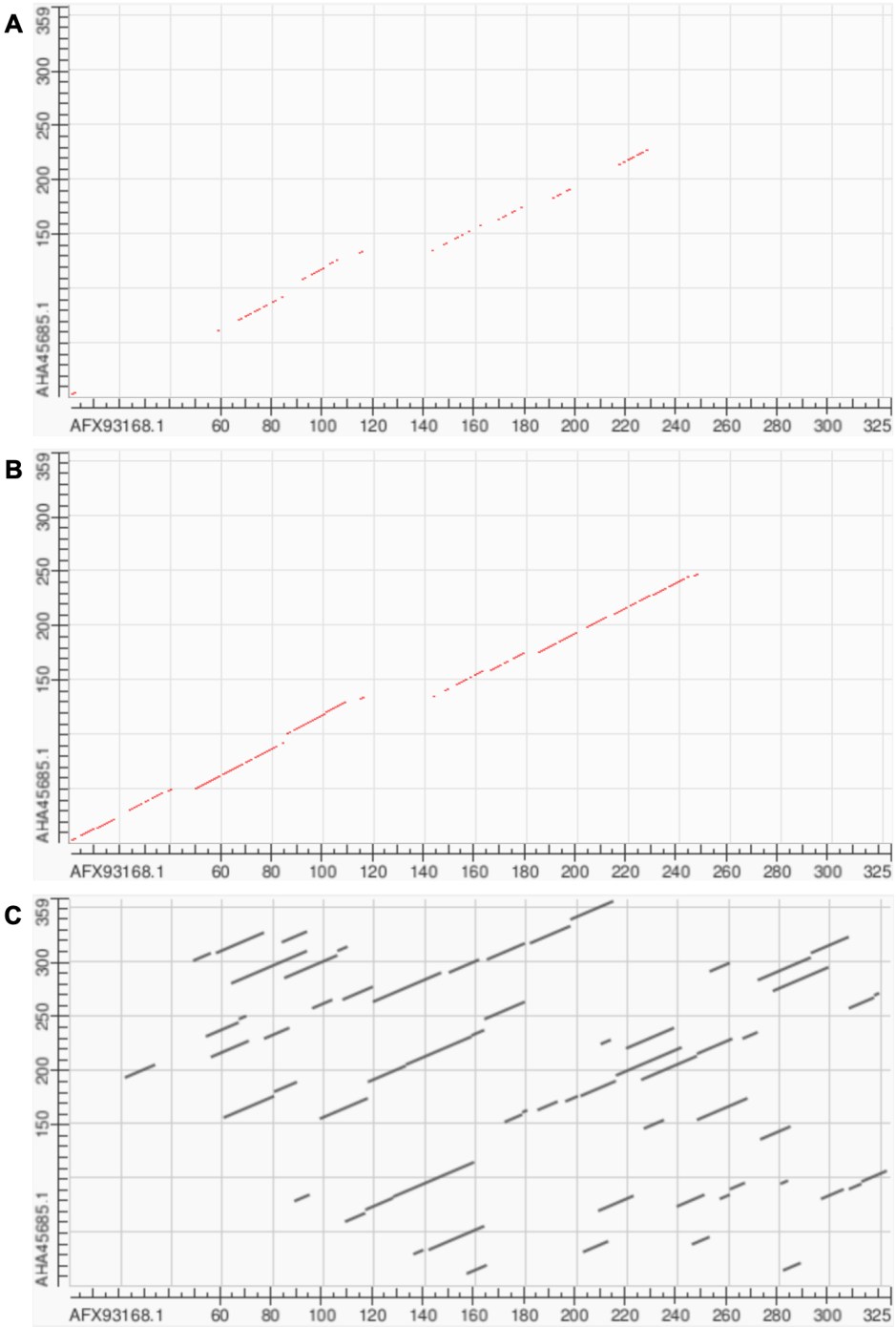


Figure 6. Soft alignment for protein sequences of AFX93168.1 from Megavirus courdo11 *(Megavirus)* and AHA45685.1 from Hirudovirus strain Sangsue *(Mimivirus)*. **(A)** ESM2-3B pre-trained embeddings identify the longest soft alignment path of length 71, whereas **(B)** ESM2-3B embeddings fine-tuned using the contrastive framework identify the longest path of length 174. **(C)** Traditional BLASTP homology search generated multiple fragmented regions but was unable to find the longest path.

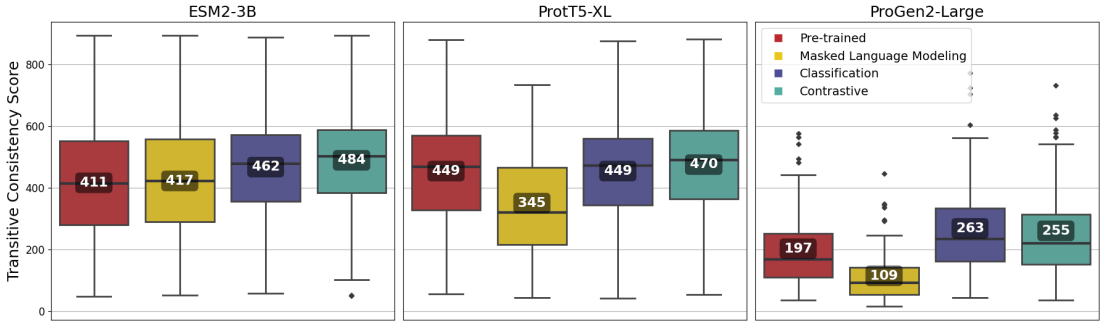


Figure 7. Transitive Consistency Scores of 100 vector-clustering MSAs (vcMSAs) annotated with overall mean values. Each vcMSA contained a maximum of 150 sequences with sequence lengths ranging from 62 to 3050 amino-acids.

the model's sensitivity to biologically relevant variations. Such improvements are particularly crucial for viral proteins, which have historically been underrepresented in the training data for many pLMs.

Clustering experiments further reinforced these findings. The high silhouette scores achieved by the contrastive fine-tuned ESM2-3B illustrate its enhanced ability to group proteins into biologically coherent clusters. For instance, the refined separation of helicase families—where distinct functional subgroups were more clearly delineated compared to pre-trained embeddings—underscores the potential of fine-tuning to preserve and even enhance functionally relevant features within protein families.

Alignment-based assessments provide additional evidence of the benefits conferred by fine-tuning. The contrastively fine-tuned model not only produced longer homologous residue paths in soft alignments but also generated MSAs with higher TCS and improved MI metrics. These improvements suggest that fine-tuned embeddings can capture evolutionary signals more effectively, which is essential for accurate functional annotation and comparative sequence analysis.

Interestingly, while ESM2-3B and, to a moderate extent, ProtT5-XL exhibited clear improvements following fine-tuning, the ProGen2-Large model showed limited gains. This divergence indicates that the efficacy of fine-tuning may depend on the underlying model architecture and its pre-training regime.

Beyond the immediate performance improvements, our study underscores the broader utility of parameter-efficient fine-tuning approaches like LoRA. By dramatically reducing computational requirements, LoRA makes it feasible for research groups with limited resources to adapt large-scale pLMs for specialized applications, such as the study of viral proteins. These advances offer innovative and unprecedented opportunities for understanding viral function and ecology.

5 CONCLUSION

Twelve protein language model variants were evaluated for their ability to capture viral protein biology. In these evaluations, the pre-trained pLMs were compared with versions fine-tuned for viral sequences using one of three learning frameworks. Our application of PEFT revealed clear benefits: Using viral proteins to selectively update pLM weights (initially learned through mlm or autoregressive Causal Language Modeling) with classification or contrastive learning frameworks resulted in performance improvements across diverse tasks. Our fine-tuned models demonstrated improved clustering capabilities, with both contrastive ESM2-3B and ProtT5-XL variants producing clearer distinctions among viral protein orthologous groups, families, and domains, indicating that fine-tuning enhances the models' ability to differentiate distinct protein groups despite low sequence similarity, thus supporting biologically meaningful hierarchies.

Additionally, the fine-tuned models produced better quality alignments, with both contrastive and classification fine-tuning frameworks significantly improving global alignment accuracy and MSA fidelity. These improvements underscore the fine-tuned pLMs' refined ability to capture conserved motifs and infer evolutionary relationships with greater precision. The results were particularly significant for contrastive fine-tuned ESM2-3B, which consistently outperformed all other models and fine-tuned variants. This model showed robust performance across numerous downstream tasks.

We experimented with various learning frameworks, highlighting the benefits of diversifying learning

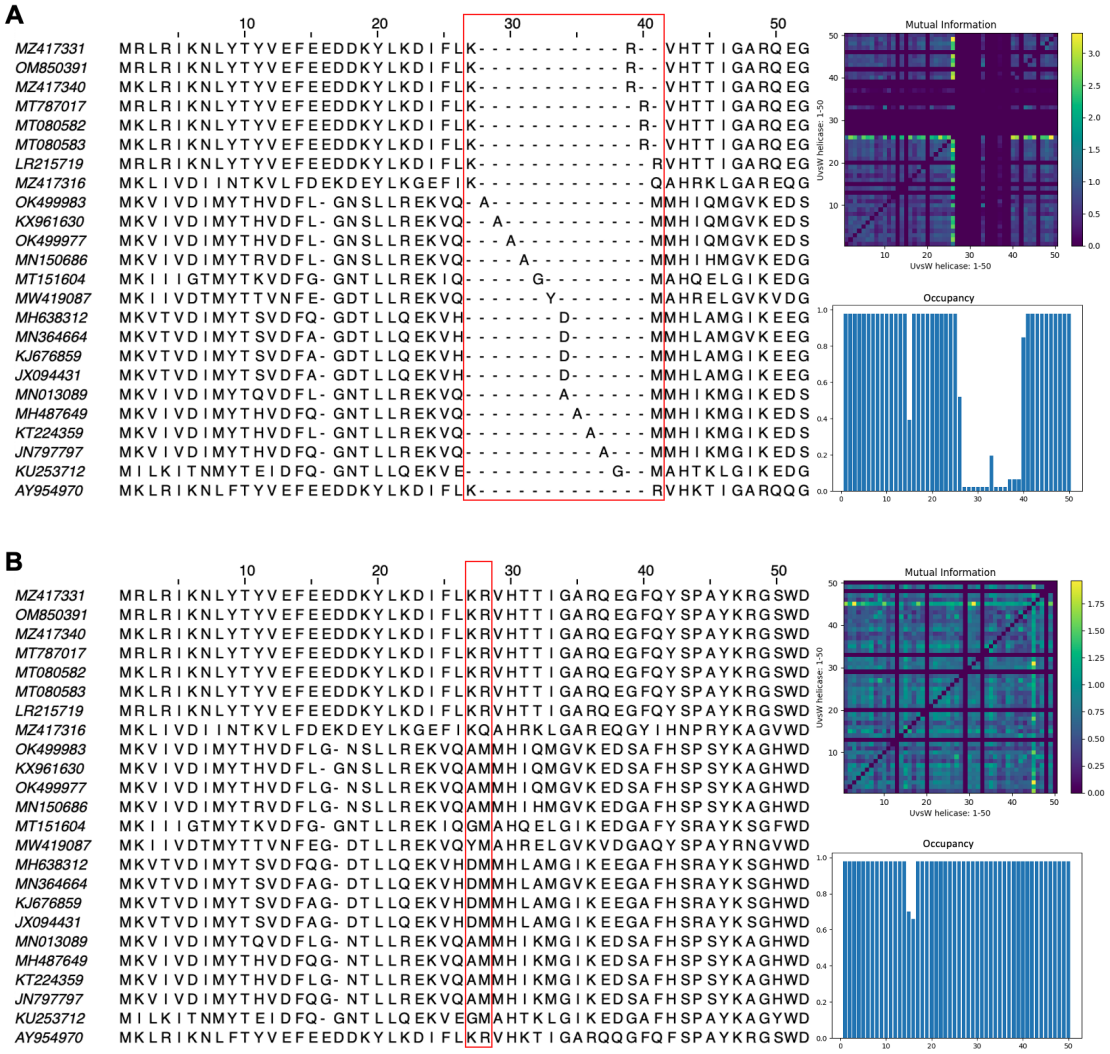


Figure 8. Subset of the first 50 positions in the vector-clustering MSA of UvsW helicase bacteriophage proteins using **(A)** pre-trained ESM2-3B and **(B)** contrastive fine-tuned ESM2-3B. The pre-trained model aligned sequences with lower mean Mutual Information (MI) (0.16) (top right), occupancy (0.51) (bottom right), and T-Coffee's Transitive Consistency Score (TCS, 939). The contrastive fine-tuned ESM2-3B model produced an alignment of greater quality with higher mean MI (0.22), occupancy (0.59), and TCS (972).

methodologies in fine-tuning pLMs. This diversity in training techniques proved advantageous, as it enriched model representations and enabled superior performance across clustering and alignment tasks. Our findings emphasize that PEFT methods can substantially reduce the computational and memory footprint of training, making these techniques accessible for large models without sacrificing efficiency. In sum, these findings pave the way for more accessible, high-performance tools in bioinformatics, advancing our ability to analyze and interpret complex biological data at scale.

6 ACKNOWLEDGMENTS

This work was supported by National Science Foundation grant numbers 1736030 and 2025567. Access to high performance computational resources, storage, and expertise were provided by the University of Delaware Bioinformatics Data Science Core Facility (RRID:SCR_017696) with support from Delaware INBRE (NIH P20GM103446), NIH Shared Instrumentation Grant (NIH S10OD028725), the Delaware Biotechnology Institute, and the State of Delaware.

REFERENCES

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500.

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322.

Bahdanau, D. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bakan, A., Meireles, L. M., and Bahar, I. (2011). Prody: protein dynamics inferred from theory and experiments. *Bioinformatics*, 27(11):1575–1577.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110.

Brosh Jr, R. M. and Matson, S. W. (2020). History of dna helicases. *Genes*, 11(3):255.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Chang, J.-M., Di Tommaso, P., and Notredame, C. (2014). TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Molecular biology and evolution*, 31(6):1625–1637.

Chen, I., Johansson, F. D., and Sontag, D. (2018). Why is my classifier discriminatory? *Advances in neural information processing systems*, 31.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ding, F. and Steinhardt, J. (2024). Protein language models are biased by unequal sequence sampling across the tree of life. *BioRxiv*, pages 2024–03.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2021). Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127.

Ferruz, N. and Höcker, B. (2022). Controllable protein design with language models. *Nature Machine Intelligence*, 4(6):521–532.

Ferruz, N., Schmidt, S., and Höcker, B. (2022). Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348.

Flamholz, Z. N., Li, C., and Kelly, L. (2024). Improving viral annotation with artificial intelligence. *Mbio*, 15(10):e03206–23.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. (2020). The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Harrigan, W. L., Ferrell, B. D., Wommack, K. E., Polson, S. W., Schreiber, Z. D., and Belcaid, M. (2024). Improvements in viral gene annotation using large language models and soft alignments. *BMC bioinformatics*, 25(1):165.

Hesslow, D., Zanichelli, N., Notin, P., Poli, I., and Marks, D. (2022). Rita: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*.

Hie, B. L., Shanker, V. R., Xu, D., Bruun, T. U., Weidenbacher, P. A., Tang, S., Wu, W., Pak, J. E., and Kim, P. S. (2024). Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 42(2):275–283.

Holzinger, A., Keiblinger, K., Holub, P., Zatloukal, K., and Müller, H. (2023). Ai for life: Trends in artificial intelligence for biotechnology. *New biotechnology*, 74:16–24.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Annual Meeting of the Association for Computational Linguistics*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022a). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hu, M., Yuan, F., Yang, K., Ju, F., Su, J., Wang, H., Yang, F., and Ding, Q. (2022b). Exploring evolution-aware &-free protein language models as protein function predictors. *Advances in Neural Information Processing Systems*, 35:38873–38884.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589.

Katoh, K. and Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780.

Keown, R. A. (2024). *From Single Amino Acid to Environmental Ecology and Back: The Saga of DNA Polymerase I*. PhD thesis, University of Delaware.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Kleinberg, G., Diaz, M. J., Batchu, S., and Lucke-Wold, B. (2022). Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare. *Journal of biomed research*, 3(1):42–47.

Lee, J., Hunter, B., and Shim, H. (2024). A pangenome analysis of eskape bacteriophages: the underrepresentation may impact machine learning models. *Frontiers in Molecular Biosciences*, 11:1395450.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. (2022). Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Lupo, U., Sgarbossa, D., and Bitbol, A.-F. (2022). Protein language models trained on multiple sequence

alignments learn phylogenetic relationships. *Nature Communications*, 13(1):6298.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos Jr, J. L., Xiong, C., Sun, Z. Z., Socher, R., et al. (2023). Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106.

Madeira, F., Madhusoodanan, N., Lee, J., Eusebi, A., Niewielska, A., Tivey, A. R., Lopez, R., and Butcher, S. (2024). The embl-ebi job dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Research*, page gkae241.

Mahmoudabadi, G. and Phillips, R. (2018). A comprehensive and quantitative exploration of thousands of viral genomes. *elife*, 7:e31955.

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., and Bossan, B. (2022). Peft: State-of-the-art parameter-efficient fine-tuning methods. *URL: https://github. com/huggingface/peft*.

McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11).

McWhite, C. D., Armour-Garb, I., and Singh, M. (2023). Leveraging protein language models for accurate multiple sequence alignments. *Genome Research*, 33(7):1145–1153.

Moffat, L., Kandathil, S. M., and Jones, D. T. (2022). Design in the dark: learning deep generative models for de novo protein design. *bioRxiv*, pages 2022–01.

Moore, R. M., Harrison, A. O., McAllister, S. M., Polson, S. W., and Wommack, K. E. (2020). Iroki: automatic customization and visualization of phylogenetic trees. *PeerJ*, 8:e8584.

Moreno-Gallego, J. L. and Reyes, A. (2021). Informative regions in viral genomes. *Viruses*, 13(6):1164.

Nasko, D. J., Chopyk, J., Sakowski, E. G., Ferrell, B. D., Polson, S. W., and Wommack, K. E. (2018). Family a dna polymerase phylogeny uncovers diversity and replication gene organization in the virioplankton. *Frontiers in microbiology*, 9:3053.

Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., and Madani, A. (2023). Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217.

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., and Kyrpides, N. C. (2016). Uncovering earth's virome. *Nature*, 536(7617):425–430.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490.

Qureshi, R., Irfan, M., Gondal, T. M., Khan, S., Wu, J., Hadi, M. U., Heymach, J., Le, X., Yan, H., and Alam, T. (2023). Ai in drug discovery and its clinical relevance. *Heliyon*, 9(7).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. (2017). Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling*, 57(4):942–957.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. (2019). Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, 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):e2016239118.

Rolf, E., Worledge, T. T., Recht, B., and Jordan, M. (2021). Representation matters: Assessing the importance of subgroup allocations in training data. In *International Conference on Machine Learning*, pages 9040–9051. PMLR.

Ryu, H. J., Adam, H., and Mitchell, M. (2017). Inclusivefacenet: Improving face attribute detection with race and gender diversity. *arXiv preprint arXiv:1712.00193*.

Sainburg, T., McInnes, L., and Gentner, T. Q. (2021). Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881–2907.

Sanejouand, Y.-H. (2023). On the unknown proteins of eukaryotic proteomes. *Journal of molecular evolution*, 91(4):492–501.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710.

Shahbazi, N., Lin, Y., Asudeh, A., and Jagadish, H. (2023). Representation bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys*, 55(13s):1–39.

Sharma, R., Shrivastava, S., Kumar Singh, S., Kumar, A., Saxena, S., and Kumar Singh, R. (2021). Deep-abppred: identifying antibacterial peptides in protein sequences using bidirectional lstm with word2vec. *Briefings in Bioinformatics*, 22(5):bbab065.

Singleton, M. R., Dillingham, M. S., and Wigley, D. B. (2007). Structure and mechanism of helicases and nucleic acid translocases. *Annu. Rev. Biochem.*, 76(1):23–50.

Teukam, Y. G. N., Dassi, L. K., Manica, M., Probst, D., Schwaller, P., and Laino, T. (2024). Language models can identify enzymatic binding sites in protein sequences. *Computational and structural biotechnology journal*, 23:1929–1937.

The UniProt Consortium (2024). Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617.

Trgovec-Greif, L., Hellinger, H.-J., Mainguy, J., Pfundner, A., Frishman, D., Kiening, M., Webster, N. S., Laffy, P. W., Feichtinger, M., and Rattei, T. (2024). Vogdb—database of virus orthologous groups. *Viruses*, 16(8):1191.

Tule, S., Foley, G., and Bodén, M. (2025). Do protein language models learn phylogeny? *Briefings in Bioinformatics*, 26(1):bbaf047.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., and Russakovsky, O. (2020). Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 547–558.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Neural Information Processing Systems*.

Youle, M., Haynes, M., and Rohwer, F. (2012). Scratching the surface of biology's dark matter. *Viruses: essential agents of life*, pages 61–81.

Zeng, H., Edwards, M. D., Liu, G., and Gifford, D. K. (2016). Convolutional neural network architectures for predicting dna–protein binding. *Bioinformatics*, 32(12):i121–i127.

Zhang, S., Krieger, J. M., Zhang, Y., Kaya, C., Kaynak, B., Mikulska-Ruminska, K., Doruker, P., Li, H., and Bahar, I. (2021). Prody 2.0: increased scale and scope after 10 years of protein dynamics modelling with python. *Bioinformatics*, 37(20):3657–3659.

Zhang, Y., Zhu, G., Li, K., Li, F., Huang, L., Duan, M., and Zhou, F. (2022). Hlab: learning the bilstm features from the protbert-encoded proteins for the class i hla-peptide binding prediction. *Briefings in Bioinformatics*, 23(5):bbac173.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

SUPPLEMENTARY MATERIAL

Experiment 1: Pairwise comparison-based experiments

1.1 ECDF of pairwise cosine similarity distribution

Provided below is the empirical cumulative distribution function (ECDF) plot of the cosine similarity distribution (Fig. S1). This accompanies the cosine similarity distribution box plot shown in Fig. 2.

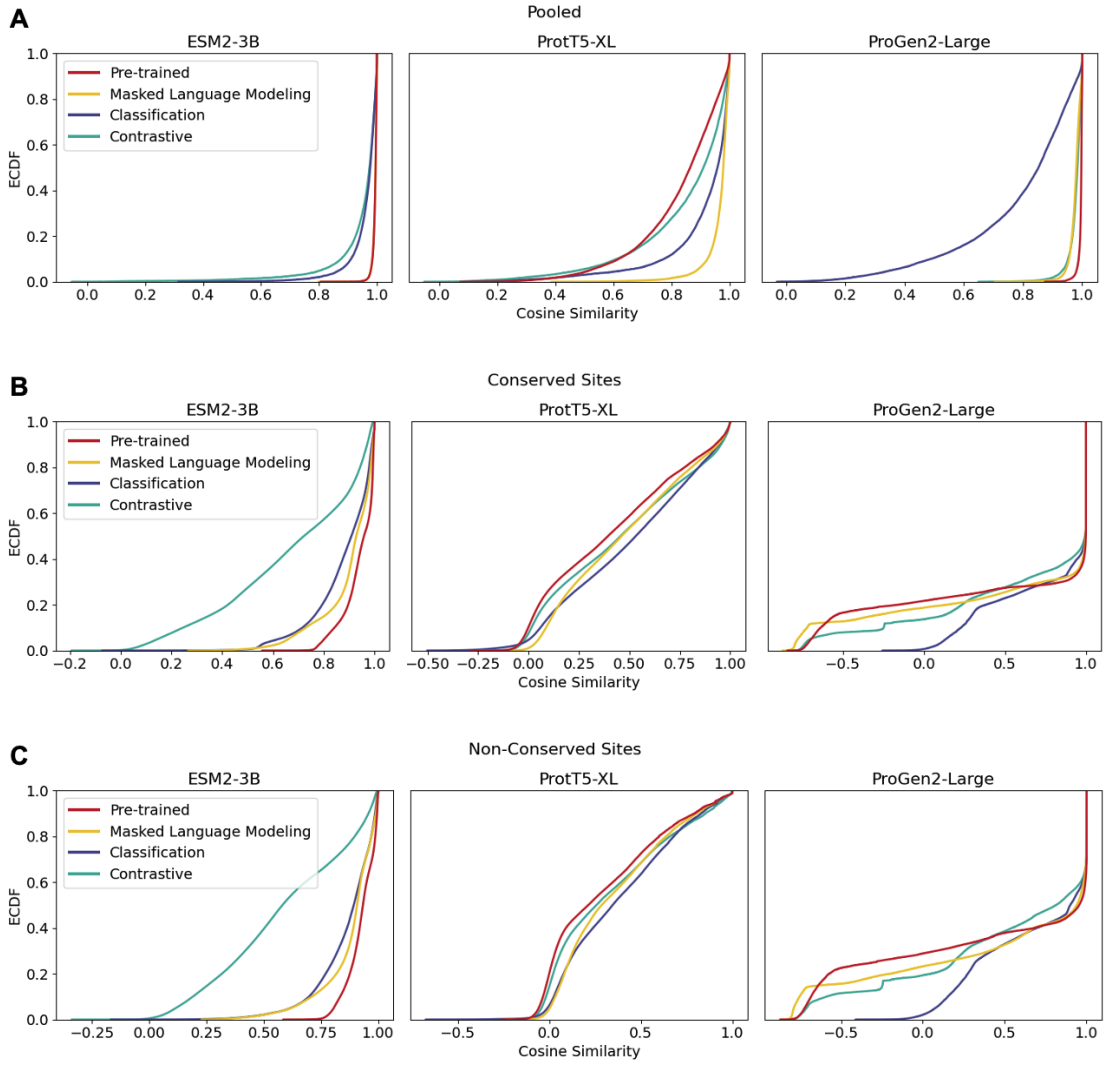


Figure S1. ECDF of pairwise cosine similarity values for (A) pooled sequence embeddings of sequences within an orthologous group, (B) conserved sites within multiple sequence alignments, and (C) non-conserved sites within multiple sequence alignments.

1.2 Average pairwise protein similarity using global sequence alignments with the BLOSUM62 scoring matrix

We observed that pre-trained ESM2-3B yields inflated cosine similarity scores for pooled sequence embeddings when compared against the average pairwise protein similarity computed using global sequence alignments with the BLOSUM62 scoring matrix. For each protein Viral Orthologous Group (VOG), we aligned ten randomly selected sequences (same sequence used to compute pairwise sequence cosine similarities from the VOG test dataset) using the Biopython pairwise2 module. The alignment scores were normalized by the maximum sequence length for each pair to obtain a similarity value between 0 and 1. Mean similarity scores were computed per VOG and averaged across all groups to

assess sequence redundancy and conservation within the dataset. This approach resulted in an average pairwise protein similarity score of 0.51, which is considerably low compared to the elevated cosine similarity score of 0.99 obtained via pre-trained ESM2-3B embeddings.

Experiment 2: Clustering-based experiments

2.1 DNA helicase domain analysis and phylogenetic tree

Pfam domain calculation

Each of the 174 DNA helicase amino acid sequences were analyzed using InterProScan Tool version 5.67-99.0 with default parameters, querying against the Pfam database. The output file generated consists of all significant domain hits for each open reading frame (ORF) ID, along with the start and stop regions of where the domains are aligned to the reference sequence.

Phylogentic tree

The DNA helicase amino acid sequences were aligned using MAFFT –auto and their alignments were analyzed using FastTree (version 2.1) with the default parameters to generate an approximate maximum-likelihood tree newick file of their protein alignments. The newick file along with a file containing contextual domain metadata was then loaded into the Iroki tree visualization web tool to generate the final tree.

Experiment 3: Alignment-based experiments

3.1 Quality of vcMSA for bacteriophage proteins

Mutual information and occupancy for vcMSA generated for UvsW helicase proteins from bacteriophages was derived using ProDy version 2.4.1 (Bakan et al., 2011; Zhang et al., 2021). Mutual information was normalized using minimum entropy to account for sequence variability and to highlight relevant residue correlations.