

Multiple versus pairwise sequence alignments for protein phylogenetics using foundation models

Rohan Alibutud^{1,2} and Sudhir Kumar^{1,2,*}

¹Institute for Genomics and Evolutionary Medicine, Temple University, PA 19122, USA

²Department of Biology, Temple University, Philadelphia, PA 19122, USA

*Corresponding author (s.kumar@temple.edu)

Keywords: Molecular phylogenetics, Foundation models, Proteins, Multiple sequence alignment, ESM vs MSA transformers.

Abstract

Phylogenetic inference is a common task in molecular and evolutionary biology and has conventionally required a multiple sequence alignment (MSA), a statistical model of amino acid substitutions, and an optimality principle. Recently, global models of amino acid substitutions have been inferred from millions of MSAs using transformer-based deep learning, resulting in protein foundation models (pFMs), also known as protein language models (PLMs). Training pFMs on MSAs hypothetically enables them to encode residue dependencies and the phylogenetic structure of the MSA collection. In contrast, pFMs trained on individual sequences lack access to such phylogenetic structure. Here, we assess the phylogeny inference gains offered by the use of MSA for training pFMs by comparing the relative accuracies of phylogenies inferred using two types of pFMs: one trained on a large collection of MSAs (msat-pFM, [1]) and the other trained using a collection of single sequences (esm-pFM). For msat-pFM analysis, we inferred neighbor-joining trees using pairwise distances estimated directly from the sequence attention matrices. For esm-pFM [2], pairwise distances were obtained using the correlation of attentions of homologous residues, where pairwise sequence alignments (PSA) were used to establish residue homologies. Surprisingly, MSA phylogenies inferred using the msat-pFM were less accurate than esm-pFMs. This pattern was seen across datasets spanning both small and large numbers of species and proteins. Also, PSA phylogenies obtained using residue attentions from early ESM-PFM layers were much more accurate. These results suggest that the multiple sequence alignment step, which is obligatory to establish residue homologies across multiple sequences, may not add information when using evolutionary distances based on attentions in pFMs.

A. Introduction

The advent of deep learning in molecular biology has introduced a host of new methods for addressing existing tasks in molecular evolution and phylogenetics. In this article, we examine the potential application of protein foundation models (pFMs) for a specific task in molecular biology: phylogenetic inference. pFMs have been trained using millions of multiple sequence alignments (MSAs) by deep learning in a transformer architecture.[1–3] These pFMs are relevant to molecular phylogenetics for their ability to capture residue dependencies in homologous proteins.[4,5] This feature allows analyses conducted using pFMs to avoid the ubiquitous assumptions in traditional methods of independent evolution of residues within proteins.[6] The commonly used MSA Transformer model (referred to here as msat-pFM) is a network with over 100 million parameters and 144 attention heads, organized into 12 layers (**Fig. 1**).[1] The use of msat-pFM to analyze any MSA produces a collection of sequence attention matrices (SAMs), one for each residue, for every attention head. A SAM captures the model's weighted strength of association between sequences at a site. If there are N sequences in an MSA, then $144 \times L$ SAMs of size $N \times N$ are produced in an inference run using msat-pFM.

Lupo *et al.*[7] demonstrated that many SAMs are highly correlated with Hamming distances between sequences, a simple measure of evolutionary distance. They trained a linear regression model that could predict Hamming distances from SAMs across any MSA. Subsequently, Chen *et al.* 2025[8] have shown that neighbor-joining (NJ) trees produced from the inverse of layer-SAMs, which converts pairwise average attentions into a measure of dissimilarity, can yield phylogenies similar to those obtained with conventional evolutionary distances and phylogenetic methods. Therefore, msat-FM can be used to reconstruct phylogenetic relationships among sequences in an MSA, where msat-pFM replaces traditional substitution models of molecular evolution.

We investigate whether using an MSA-based pFM improves performance compared to an alternative method that relies instead on single sequences as the input. This method employs the ESM-2 model (referred to as esm-pFM; [2]), a model trained on millions of individual protein sequences, in order to derive attention-based distances. In this case, we used pairwise sequence alignments (PSAs) to establish residue homologies and then estimated pairwise distances based on the strength of correlation between the corresponding residue attentions. Not only are PSAs easier to generate, but they also avoid circularity as individual protein MSAs are assembled using heuristics that involve progressive alignment guided by an *ad hoc* phylogeny built from a matrix of pairwise distances between sequences.[9–12]

Here, we first present our new approach to estimate pairwise distances from PSAs, which uses residue attention matrices (RAMs) produced by esm-pFMs for individual sequences. We then compare the accuracy of the phylogenies produced using PSA (esm-pFM) and MSA (msat-pFM) approaches. In the following, we present analyses of increasingly difficult and

larger phylogenetic datasets, describe our new PSA approach, and compare results from MSA and PSA analyses.

B. Methods

Inference from esm-pFM Residue Attention Matrices

To test the impact of multiple sequence alignment on phylogenetic inference accuracy, we needed to infer trees using a model that did not rely on MSAs. Here, we introduce a new method that employs esm-pFM, a model that takes only single sequences as input. We used the 35-million-parameter checkpoint of the Evolutionary Scale Modeling (ESM-2) protein language model, retrieved from the fair-esm GitHub repository, as our example for esm-pFM. Checkpoints at 150M, 650M, and 3B parameters are also available, but we used the lower-parameter checkpoint to ensure that phylogenetic inference accuracy in esm-pFM was not due to using a larger model than msat-pFM, which has 100 million parameters.

As no MSAs are used, esm-pFM does not directly compare sequences and thus does not quantify similarities or differences between them. Instead, our approach calculates correlations of RAMs across species to generate a pairwise distance matrix (**Fig. 1**). The underlying expectation is that more closely related species will produce more similar RAMs (e.g., Chen et al.).[8] We use this pairwise distance matrix to infer a neighbor-joining tree for each RAM, yielding one tree for each layer/head combination. We assemble these layer/head RAM trees into a single consensus tree using ASTRAL.[13] Thus, from a set of unaligned sequences, we can infer a phylogenetic tree that requires only pairwise alignments between species and does not use MSAs.

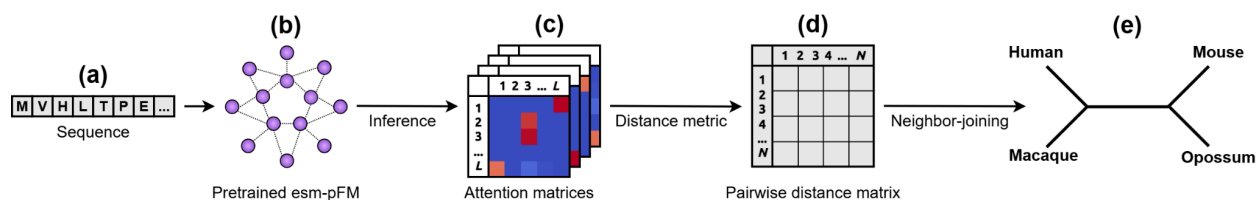


Fig. 1 | Phylogenetic inference using esm-pFM attention. (a) Two or more unaligned sequences are passed into the pretrained esm-pFM model (b). Model inference yields (c) N attention matrices of dimensions $L \times L$, where N is the number of sequences and L is the length of each sequence. Distance metrics are calculated for each layer and head's attention matrix $M_{l,h}$. Pairwise comparisons are made between the corresponding layer and head matrices of different species, producing a distance matrix (d) that represents the dissimilarity between the attentions of any two species. This distance matrix is passed to the neighbor-joining algorithm to infer a phylogenetic tree (e). This process yields a single tree for every attention layer/attention head combination. Multiple trees are combined using ASTRAL in order to yield a final consensus tree.

We passed the pairwise-aligned sequences into the model one at a time, performing a forward inference pass and extracting attention matrices for each head in each layer. We calculated Pearson's R for the esm-pFM attention matrix output for each pairwise relationship of two species in order to derive a correlation matrix for every network head. Once the correlation

between the attention matrices of a pair of species is known, a distance matrix is built by subtracting a normalized correlation value from 1 for each pairwise comparison, as described in the following equation, where v is the number of valid, non-gapped positions and L is the sequence length:

$$(1) \quad d_{a,b} = 1 - r \frac{v}{L}$$

This equation normalizes by the number of non-gapped positions in a given input sequence to account for gap tokens introduced by the Needleman-Wunsch pairwise alignment. After flattening the two-dimensional attention matrices a and b into two one-dimensional series such that $a = \{a_1, a_2, \dots, a_{mn}\}$ and $b = \{b_1, b_2, \dots, b_{mn}\}$, we can calculate how well the attention matrices of two sequences correlate with each other. Using Pearson's R to measure correlation yields a measure of similarity, but subtracting the absolute value of R from 1 converts the correlation into a measure of dissimilarity (distance, d). We can then use these pairwise dissimilarity metrics to populate a distance matrix that represents all species for which we wish to infer a tree. We then apply the neighbor-joining algorithm to this distance matrix to reconstruct a phylogeny for each layer-head combination. As the attention matrix of each head may capture part of the overall phylogeny, we combine information from all layer-head combinations by building a neighbor-joining tree from each head attention matrix and then deriving a consensus topology using quartet puzzling in ASTRAL.[13]

Inference from msat-pFM Sequence Attention Matrices

For comparison with esm-pFM, we introduce a method to build phylogenies of individual SAMs extracted from msat-pFM. This method differs from a previous approach (Chen *et al.* (2025)) in which SAMs were averaged together prior to inferring a single phylogenetic tree.[8] In our case, we generate a separate tree for the SAM of each head in every layer (a total of 144). Each layer/head SAM is then used to infer a separate phylogenetic tree. These separate trees are then fed into ASTRAL to derive a final consensus phylogeny. **Fig. 2** provides a step-by-step approach for estimating a phylogeny of sequences in an MSA using msat-pFM. This procedure is analogous to phylogenetic inference using an MSA and a substitution model, with msat-pFM in the place of a standard substitution model such as JTT.[14] Instead of describing evolutionary distance in terms of substitutions, msat-pFM estimates it via an attention mechanism.

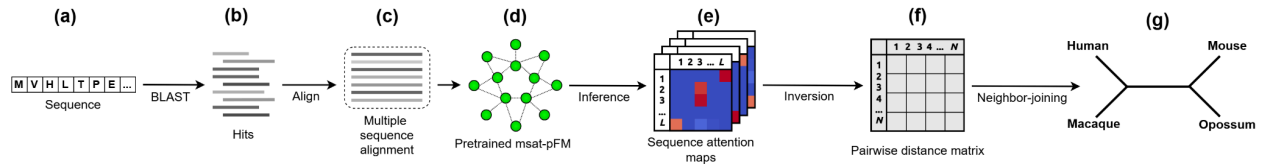


Fig. 2 | Phylogenetic inference using msat-pFM attention. (a) A target sequence must be used to compose a multiple sequence alignment alongside all matching BLAST hits (b). The resulting MSA (c) is then passed into the pretrained msat-pFM (d). Model inference then yields a set of 144 SAMs of dimensions $N \times N \times L$ (e), where N is the number of sequences in the input MSA and L is the sequence length. (f) Each SAM is averaged across L to derive a single SAM of dimensions $N \times N$, which is then inverted to create a distance matrix representing the predicted evolutionary distance between any two sequences. (g) The distance matrix is passed to the neighbor-joining algorithm to infer a phylogenetic tree. This process yields a single tree for every attention layer/attention head combination. Multiple trees are combined using ASTRAL in order to yield a final consensus tree.

The previously observed correlation between attention value and Hamming distance supports the idea that SAMs effectively measure evolutionary relatedness, with more closely related species having higher attention values.[7] Thus, SAMs must be inverted to produce a distance matrix that instead describes how evolutionarily divergent two species are. The neighbor-joining algorithm can then be applied to this distance matrix in order to yield a phylogeny. In order to compensate for the variance of attention across different heads, we calculate a normalized attention distance by subtracting elementwise attention weights from the maximum attention weight of each matrix, as described in Equations 2 and 3:

$$(2) \quad A_{a,b} = \text{mean}(A_{a,b,n})$$

$$(3) \quad d_{a,b} = \max(A_{a,b}) - A_{a,b}$$

Where $A_{a,b}$ is the attention value for the pair of species a and b and n is the number of residues in the protein. We average all n residue matrices to obtain a single mean matrix of pairwise species attentions. Chen *et al.* (2025) used a similar method to build neighbor-joining trees from transformed attention matrices (Q-matrix). The equation for the Q-matrix is thus:

$$(4) \quad Q_{a,b} = -((n-2)A_{a,b} - \sum_n A_{a,n} - \sum_n A_{b,n})$$

Where $Q_{a,b}$ is the Q-distance matrix for species a and b, $\sum_n A_{a,n}$ is the sum of the attention matrices for species a along all residues n, and $\sum_n A_{b,n}$ is the sum of the attention matrices for species b along all residues n.

OrthoMaM Data

To obtain a well-curated set of protein sequences, we used the Orthologous Mammalian Markers (OrthoMaM) database.[15] We sampled a set of proteins that were shorter than 1024 residues (the context limit of msat-pFM). From this set, we selected three species subsets (the 4-mammal, 4-ape, and 34-species sets) and then selected proteins with representative

sequences for all three (*Supplemental Table S1*). For the msat-pFM analyses, we left these alignments as is, but for the esm-pFM analyses, we removed the gap characters (“-”) for each species, yielding a set of unaligned sequences. We also used the OrthoMaM SuperMatrix tree as a reference “ground truth” tree, as it was assembled from a much larger pool of protein sequences than were used in any analyses in this paper.

C. Results

C.1. Accuracy of MSA-based pFM phylogenetic inference

To test the performance of the SAM phylogenetic inference approach, we passed a sample protein alignment from the Orthologous Mammalian Markers (OrthoMaM) database [15] into the model. We selected transferrin (TF) as our example protein, with its alignment containing four distantly related mammalian species: human (*Homo sapiens*), macaque (*Macaca fascicularis*), rat (*Rattus norvegicus*), and gray short-tailed opossum (*Monodelphis domestica*) (**Fig. 3a**).

We inferred a phylogenetic tree from SAM-based distance matrices, yielding one phylogeny for each attention layer-head combination. Of the 144 phylogenies produced, 79 correctly clustered human and macaque sequences. **Fig. 3b** shows the grid arranged by layers and heads within each layer, in which there appears to be no specific pattern to the layers or heads that return the correct tree. The majority-rule consensus phylogeny of the transferrin protein yielded the correct species tree.

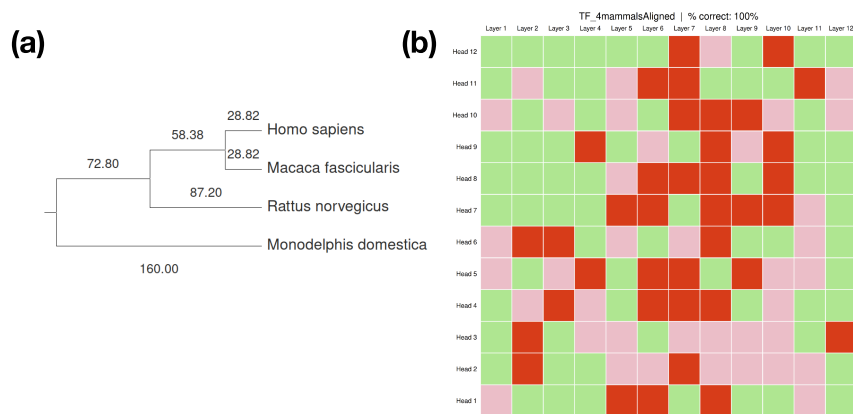


Fig. 3 | msat-pFM SAM phylogenies inferred from transferrin. (a) The expected phylogeny of transferrin with 4 mammal species, with divergence times from TimeTree 5. **(b)** A map of layer and head combinations that yielded correct phylogeny (green fill) with human-macaque grouping. Pink and red cells mark incorrect topologies, in which humans were adjacent to rat and opossum, respectively.

To assess the generality of the observations made using transferrin protein, we repeated the msat-pFM phylogenetic analyses on 19 additional OrthoMaM protein alignments for the same set of 4 mammal species (see *Methods*). We first assessed which of these proteins was faster evolving vs. more conserved, using the sum of branch lengths for the corresponding Poisson-model neighbor-joining tree as a point of comparison (**Fig. 4a**). The proportion of

attention heads producing the correct topology ranged from 44.4% to 81.25% (**Fig. 4b**). Generally, more highly conserved proteins performed worse than other proteins in these datasets, as the faster-evolving proteins returned more accurate msat-pFM trees. This is likely because a greater number of evolutionary substitutions means a greater signal for the model. On average, 58% of attention heads produced correct topology over all proteins, with the correct topology being the most common for every protein. With these results establishing a baseline for an MSA-based foundation model, we proceeded to a pairwise alignment-based esm-pFM analysis.

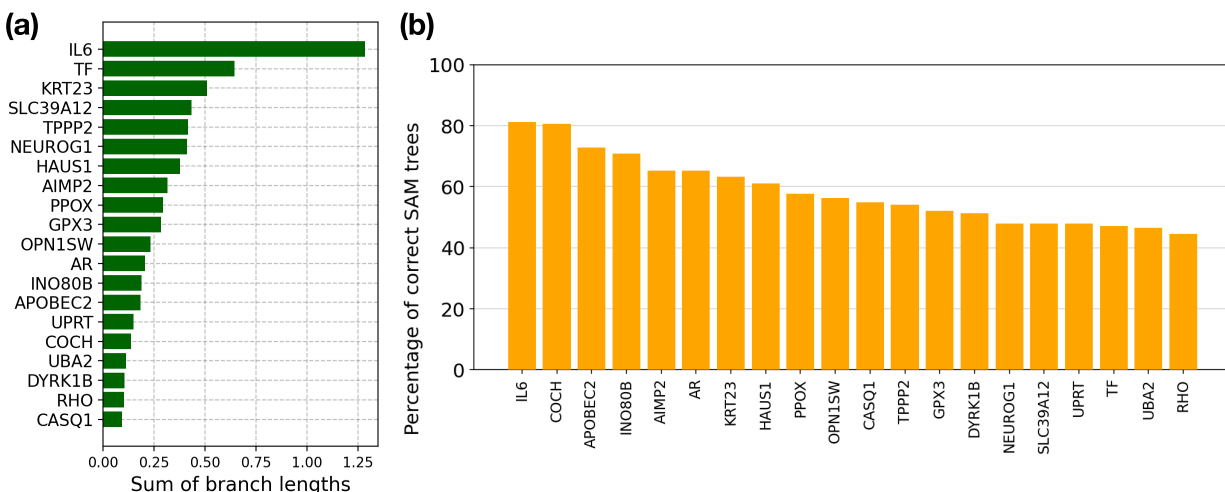


Fig. 4 | Accuracy of SAM phylogenetic inference. (a) The sum of branch lengths of each neighbor-joining tree inferred using a Poisson model, with fast-evolving proteins having longer total branch lengths and more conserved proteins having shorter total branch lengths. **(b)** Bar graph representing the percentage of correctly inferred trees from all 144 layer/head combinations of the SAMs extracted from the msat-pFM analysis of each protein.

C.2 Accuracy of PSA-based phylogenetic inference

To test the alternative approach to pFM phylogenetic inference, we applied our method based on esm-pFM RAMs to the same set of four species (human/macaque/rat/opossum). The length of the TF sequences varied across the 4 example species, with the opossum TF sequence being the shortest (628 residues) and the macaque being the longest (695 residues). With this approach, global site homologies are no longer required; instead, pairwise alignments are used to compare RAMs. Thus, homology is established separately for each pair of species, rather than being derived globally across the entire alignment.

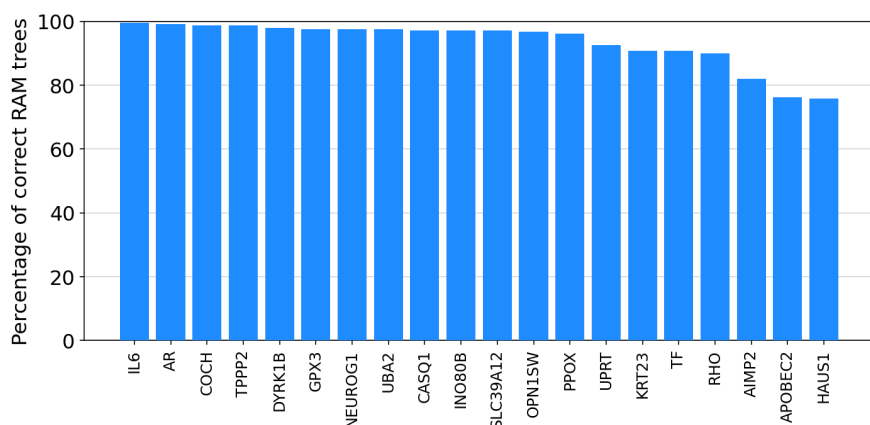


Fig. 5 | Accuracy of PSA (esm-pFM) analysis. Each bar represents the percentage of correctly inferred trees from all 240 layer/head combinations of each protein.

The checkpoint of esm-pFM used contained 12 layers, each with 20 attention heads. This yielded a total of 240 RAMs and, thus, 240 inferred phylogenies. Of the 240 RAM phylogenies, 218 were correct (**Fig. 5**). Therefore, 90.8% of the layer-heads produced the correct phylogeny, as opposed to 54.8% for msat-pFM. This surprising result was confirmed in the analysis of 19 additional proteins above, with esm-pFM achieving 88.1% accuracy across all proteins, as compared to 58% for msat-pFM.

C.3 Accuracy for closely related species

We hypothesized that the contribution of MSAs to greater phylogenetic inference accuracy might be more apparent when species are more closely related than in the prior example, and the time elapsed on in the internal branch of the tree is shorter. For this reason, we tested MSA and PSA accuracies for a set of protein alignments composed of four sequences corresponding to the great apes: human, chimpanzee, gorilla, and orangutan (**Fig. 6**). In this case, the length of the lineage leading to the common ancestor of humans and chimpanzees is very short (~2.20 Ma), which could make it harder for PSA relative to MSA, because residue homologies are only established between sequences rather than across all four sequences.

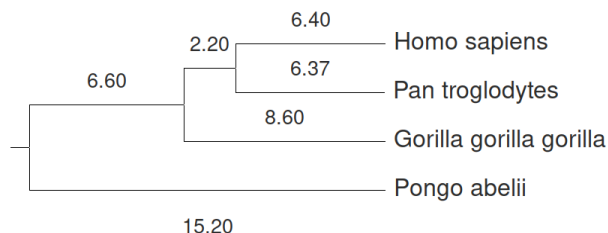


Fig. 6 | Expected phylogeny of transferrin. The expected phylogeny of transferrin with 4 mammal species with times from TimeTree 5.

Analysis of 20 great ape protein alignments yielded a range of accuracies for MSA and PSA (**Fig. 7**). On average, esm-pFM outperformed msat-pFM, although with a wider range of outcomes. The trees inferred using esm-pFM (RAM-based distances) ranged from 0% correct (three proteins) to 100% correct (5 proteins), with a mean of 57.0% and a median of 61.3%. By contrast, msat-pFM had a minimum accuracy of 10.4% in one protein to a maximum of 65.2%, with a mean of 36.4% and a median of 31.6%. Overall, msat-pFM recovered the correct consensus phylogeny for only 8 of 20 proteins for this more challenging set of species, whereas 13 of 20 proteins could recover the correct phylogeny when using esm-pFM. This indicates a better overall performance of the PSA phylogenetic inference approach.

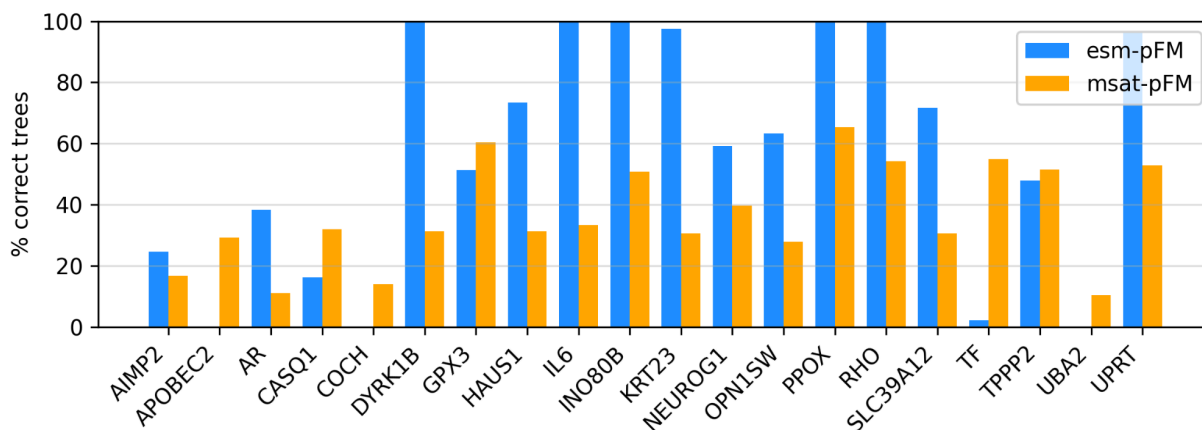


Fig. 7 | Accuracy of pFM phylogenetic inference for 20 proteins from 4 great ape species. Each protein had 240 esm-pFM RAMs and 144 msat-pFM SAMs, with phylogenetic inference accuracy expressed as the percentage of layer-head combinations that yielded a tree with the correct phylogeny. Any phylogeny that placed the human sequence and chimpanzee sequence adjacent to each other was deemed correct.

C.4 Accuracy of pFM phylogenetic inference for a larger tree

The four-taxa phylogeny is the simplest case, and MSAs may be more useful for datasets with larger numbers of sequences. To investigate this, we expanded each of the 20 OrthoMaM protein sequence alignments we used in the previous analyses from 4 taxa to 34, encompassing a range of placental mammal species, with the gray short-tailed opossum as the outgroup. This tree contains both deep and shallow divergence times, with the deepest at ~160 Ma, at the divergence point between marsupials and placental mammals, and the shallowest at 0.45 Ma, between cow (*Bos taurus*) and zebu (*Bos indicus*).[16]

We reconstructed MSA phylogenies for each attention head and each protein, yielding 144 phylogenies per protein. The accuracy of each phylogeny was quantified using the Robinson-Foulds distance[17] from the OrthoMaM reference tree, which is derived from the entire OrthoMaM database (rather than just the 20 proteins in our sample), allowing it to serve as a proxy for a ground-truth tree.[15] We use this to calculate a percentage metric, PC, where

the percentage of clades inferred correctly for a protein is $PC = [1 - (RF/2) / (N - 3)] * 100$, where RF is the Robinson-Foulds distance.

Across 20 proteins, PCs for individual msat-pFM layer-head phylogenies ranged from 0% to 71.0%, with a mean of 16.7%. Then, a majority-rule consensus was derived from 144 phylogenies for each protein, and a PC-consensus was estimated. These consensus phylogenies were overall more accurate, with PC ranging from 6.5% to 54.8% and a mean of 35.0%. Then, a global 20-protein consensus was generated, with a PC of 61.3%.

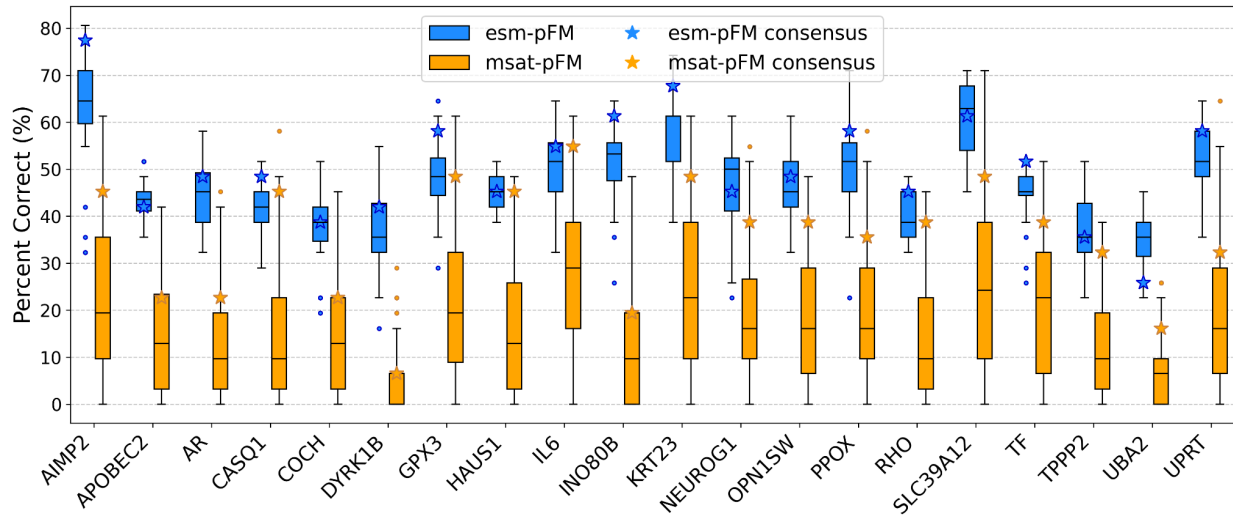


Fig. 8 | Percentage of correctly (PC) inferred clades across all attention head trees for esm-pFM vs msat-pFM. $PC = [1 - (RF/2) / (N - 3)] * 100$, where RF is the Robinson-Foulds distance, and N is the count of species in the tree. Bars are for individual layer-heads, and the PC of the consensus of head trees is denoted by a star.

PSA analyses performed better, with PC values across 20 proteins ranging from 0% to 83.9% and a mean of 38.6%. The PC of esm-pFM consensus trees ranged from 25.8% to 77.4% per protein for the 34 taxa datasets, with a mean of 50.7% (**Fig. 8**). The PC for the consensus esm-pFM tree derived from all 20 proteins combined was 74.2%. For these 34 taxa datasets, msat-pFM performed worse than esm-pFM. For 19 out of 20 proteins, msat-pFM PC was lower than the corresponding esm-pFM PC. These results show that transformer models trained on individual sequences can outperform those trained on MSAs for datasets with many species.

C.5 Concentration of phylogenetic signals in individual heads

Lupo *et al.* (2022) analyzed SAMs for protein family alignments and reported greater phylogenetic signal in certain layers and heads. Their approach was to train a single logistic model on the means of the SAMs of one set of alignments, and then assess the ability to predict Hamming distances in the remainder of the alignments. They observed larger regression coefficients in the early layers of msat-pFM, with the highest correlation with

Hamming distance consistently occurring at head 5 of msat-pFM. They interpreted this result as a concentration of phylogenetic signal at an early stage of the model.

We tested whether this pattern was also evident in phylogenetic inference results. Because Lupo *et al.* did not build phylogenies and lacked ground-truth trees, they could only use Hamming distance as a proxy for evolutionary distance. In contrast, our approach directly measures phylogenetic signals by inferring phylogenies. Reasoning that the effects of individual protein sequences would be averaged out in a larger sample, we used the trees from the 34-species set to maximize the available data. We calculated the mean PC for each layer-head of both models across all 20 proteins to obtain an average PC for each layer-head (**Fig. 10**). This contrasts with the previous approach, which calculated PC for each *protein* across the layers and heads.[8]

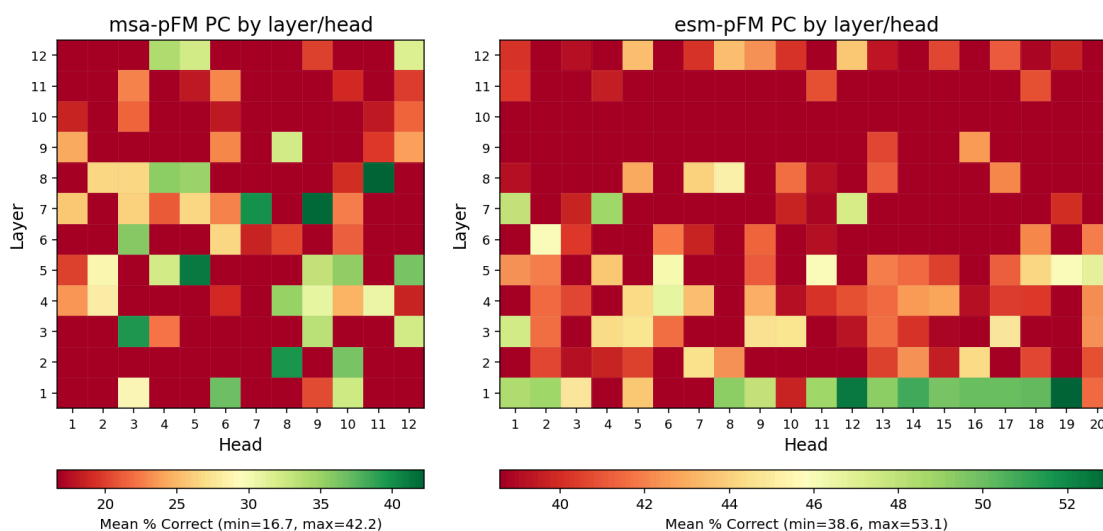


Fig. 10 | Heatmaps of average PC across msat-pFM and esm-pFM layerheads. The mean PC is reported for each layer-head combination across the 20 proteins. Green denotes a higher PC or more accurate phylogenetic inference. Both heatmaps are scaled to the minimum and maximum PC for each model.

To detect for the presence of a pattern of stronger phylogenetic signal in specific layer-head combinations, we established a null hypothesis that the mean PC of a given layer-head would be equal to the mean PC of all layer-heads as a whole. Given that the amount of phylogenetic signal varies across proteins (due to conservation, sequence length, or selective pressures), we normalized PC scores within each protein. Thus, we are only comparing PCs between layers and heads of the same protein. We applied Wilcoxon signed-rank tests for each layer, adjusting for the risk of false positives using the Benjamini-Hochberg correction, because the layers are sequential in the model and pass information to one another, so they are not independent. Considering each protein as a replicate, we also tested if there was a directional pattern in which either model increased or decreased in mean PC across successive layers, using Spearman correlation coefficients.

In the msat-pFM analysis, some intermediate layers (4, 5, 7, and 8) showed a significantly higher accuracy than average ($P < 10^{-2}$, $P < 10^{-6}$, $P < 10^{-5}$, $P < 10^{-3}$, respectively). But, the difference in accuracy was not particularly large, with a median change in PC ranging from +1.9 in layer 84 to +6.7 in layer 5. Lupo *et al.* also previously identified layer 5 as containing more phylogenetic information than the other layers, albeit without directly inferring any trees. However, there was no significant trend in PC accuracy as the model progressed from layers 1 through 12 ($\rho = 0.01$, $P = 0.28$).

In contrast, layer 1 in esm-pFM analysis produced phylogenies with rather high accuracy ($P < 10^{-6}$), with modestly higher accuracies seen for layers 4 and 5 ($P < 10^{-3}$, $P < 10^{-2}$, respectively). In addition, there was a significant trend of decreasing phylogenetic accuracy across layers, as the Spearman's ρ for PC over the successive layers of the model was -0.465 ($P < 10^{-3}$). This means that earliest layers capture bulk of the phylogenetic information, which is the reason why the accuracy of layer 1 of esm-pFM is higher than consensus across all the layers (**Fig. 11**).

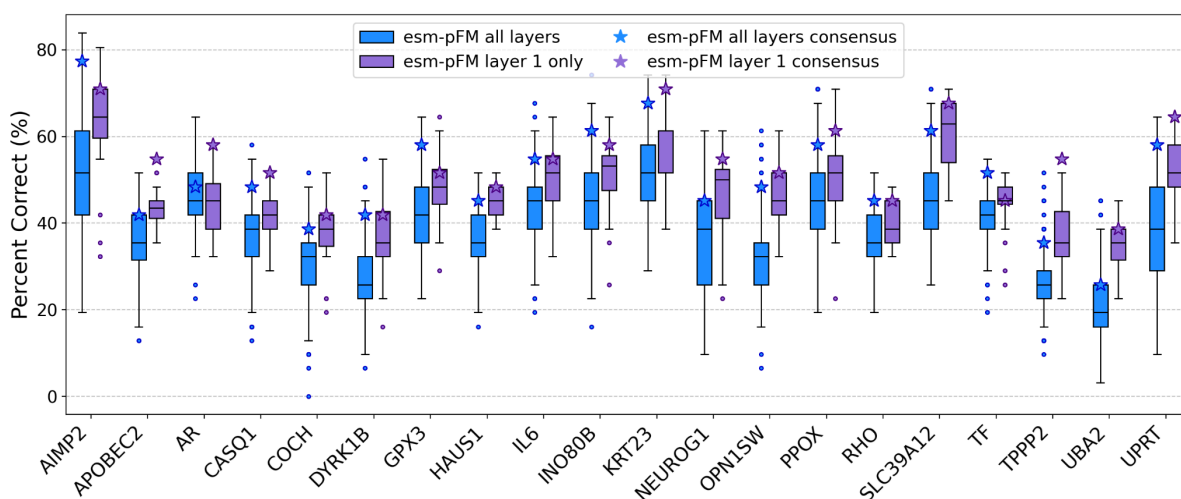


Fig. 11 | Percentage of correctly (PC) inferred clades across all attention head trees when using all layers vs. only layer 1 of esm-pFM. PC was determined by $PC = [1 - (RF/2) / (N - 3)] * 100$, where RF is the Robinson-Foulds distance, and N is the count of species in the tree. Bars are for individual layer-heads, and the PC of the consensus of head trees is denoted by a star.

D. Discussion

D.1 pFMs as an analog of substitution models

We initially explored the ability of pFMs to extract phylogenetic signal from protein sequences, building on prior investigations by Lupo *et al.* (2022) and Chen *et al.* (2025).[7,8] In particular, we sought to identify the contribution of multiple sequence alignment to the performance of pFMs for phylogenetic inference. We compared two approaches: one based on MSA-dependent SAMs and one based on single-sequence-trained RAMs. In classical

phylogenetics, evolutionary distances (and therefore tree topologies) are inferred using sequence differences and a stochastic model of sequence evolution.[9]

Under the SAM/RAM tree-building approaches discussed in this paper, pFMs play an analogous role: rather than modeling substitution probabilities, they capture evolutionary relatedness based on the similarity of residue interdependencies revealed by attention scores. Phylogenetic signal is extracted from the attention matrices, which themselves contain weights derived from the inherent functional, structural, and statistical properties of the individual sequences or alignments on which they were trained.

D.2 Model mismatch and causes of msat-pFM underperformance

Our finding that msat-pFM underperforms esm-pFM for phylogeny inference is initially counterintuitive. Previous research has indicated that msat-pFM may outperform esm-pFM for specific tasks in protein structure prediction, particularly for secondary structure prediction.[18] Hu *et al.* (2022) explored the effect of “evolution-awareness” via the use of multiple sequence alignment used in training and found that msat-pFM performed better than esm-pFM in terms of Precision@L, a metric for accurate secondary structure prediction. In this analysis, ESM-1b model performed slightly worse than MSA Transformer (0.703 vs. 0.748).[18] Evoformer, another PLM trained using multiple sequence alignments, performed slightly better than both (Precision@L = 0.785). These findings suggested that training on MSAs would allow a pFM to incorporate the information from positional homology in making predictions, which could improve its ability to infer evolutionary relationships. However, we found that msat-pFM performs worse across the board than esm-pFM, i.e., training pFMs using MSAs does not improve performance of pFMs for phylogeny inference.

One potential explanation is that this is due to model mismatch, as our interest was in inferring species phylogenies in which orthologous sequence alignments are used, but msat-pFM was trained using UniRef50 clusters, which are diverse protein family alignments replete with paralogous proteins.[1] Thus, one future direction would be to train an msat-pFM using large collections of orthologous sequences from across the tree of life, which would likely capture more neutral evolutionary patterns of residue dependencies, as compared to MSAs that contain many many paralogous proteins. Another possible explanation is the violation of the assumption that position-wise interdependencies between residues are shared across all sequences in an alignment.[1] In reality, these residue interdependencies may change more dramatically than currently assumed due to evolving epistatic relationships across time.

D.3 pFMs vs. classical phylogenetic methods

We found that the accuracy of pFM phylogenies is lower than for NJ tree inferred using standard evolutionary distances, as the 20-protein consensus tree achieved a PC of 77.4 for pFM as compared to 83.9 for standard NJ tree using p-distance. This could be due to the fact that attention based distances are likely to have much greater variance as compared to

classical amino acid substitution models, as the number of model parameters different by multiple orders of magnitude (a few to millions).

E. Conclusions

In this study, we evaluated the impact of multiple sequence alignment in pFM-based phylogenetic inference. To this end, we introduced methods for inferring phylogenies from attention maps extracted from two example models. After directly comparing these two methods, we found that the MSA-based model consistently produced less accurate phylogenies than the single-sequence model. We demonstrated that this remained true across multiple proteins, species sets, and for the consensus topologies derived across attention layers and heads. Beyond implications for the design and development of future pFMs, our findings suggest that directly extracting evolutionary relationships from individual sequence data using a deep learning model may be suitable. This can be attractive, as no MSAs, or substitution models are required. Furthermore, pFMs produce phylogenies without explicit assumptions of stationarity, rate homogeneity, and reversibility of evolutionary processes.[9] As newer pFMs are developed with a greater understanding of how evolutionary information can be encoded in foundation models, we hope to see a corresponding increase in their accuracy and applicability to long-standing questions in phylogenetics, as our findings demonstrate that one long-held requirement of classical phylogenetics (multiple sequence alignment) may be unnecessary or even detrimental when extracting evolutionary information using deep learning.

F. Funding

This work was supported by a research grant to S.K. (R35GM139540-05). This research includes calculations carried out on HPC resources supported in part by the National Science Foundation through a major research instrumentation grant 1625061. We are grateful for the technical assistance and scientific advice provided by Dr. Erfan Mowlai, Dr. John Allard, and Dr. Sudip Sharma.

G. Acknowledgments

We are grateful for the technical assistance and scientific advice provided by M. Erfan Mowlai, John Allard, and Sudip Sharma.

H. Competing interests statement

The authors declare that they have no competing interests.

I. Data availability

The supplemental information, datasets, and scripts are available at <https://github.com/RohanAlibutud/PLM-Phylo.git>. The foundation models referenced are available at <https://github.com/facebookresearch/esm>.

I. Bibliography

1. Rao R, Liu J, Verkuil R, et al. MSA Transformer. *Proceedings of Machine Learning Research* 2021; 8844–8856
2. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023; 379:1123–1130
3. Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* 2021; 118:e2016239118
4. Tule S, Foley G, Bodén M. Do protein language models learn phylogeny? *Brief. Bioinform.* 2024; 26:
5. Sgarbossa D, Malbranke C, Bitbol A-F. ProtMamba: a homology-aware but alignment-free protein state space model. *Bioinformatics* 2025; 41:
6. Kumar S, Filipski A. Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome Res.* 2007; 17:127–135
7. Lupo U, Sgarbossa D, Bitbol A-F. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *Nat. Commun.* 2022; 13:6298
8. Chen R, Foley G, Bodén M. Learning the language of phylogeny with MSA Transformer. *Cell Syst.* 2025; 101445
9. Nei M, Kumar S. *Molecular Evolution and Phylogenetics.* 2000;
10. Katoh K, Misawa K, Kuma K-I, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 2002; 30:3059–3066
11. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. *Bioinformatics* 2007; 23:2947–2948
12. Eddy S, HMMER development team. *HMMER User's Guide, Version 3.4.* 2023;
13. Zhang C, Rabiee M, Sayyari E, et al. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 2018; 19:
14. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 1992; 8:275–282
15. Allio R, Delsuc F, Belkhir K, et al. OrthoMaM v12: a database of curated single-copy ortholog alignments and trees to study mammalian evolutionary genomics. *Nucleic Acids Res.* 2024; 52:D529–D535
16. Kumar S, Suleski M, Craig JM, et al. TimeTree 5: An expanded resource for species divergence times. *Mol. Biol. Evol.* 2022; 39:
17. Robinson DF, Foulds LR. Comparison of phylogenetic trees. *Math. Biosci.* 1981; 53:131–147
18. Hu M, Yuan F, Yang KK, et al. Exploring evolution-aware & -free protein language models as protein function predictors. *arXiv [q-bio.QM]* 2022;