scientific reports

Check for updates

OPEN

Analysis of phylogenetic signal in protein language model embeddings

Brendonas Stakauskas[1]✉ & Paweł Górecki[2]

Protein language models learn high-dimensional representations of amino acid sequences that capture structural, functional and evolutionary information without explicit modeling. In this study, we examine whether distances derived from such representations can be used for phylogenetic tree inference in a zero-shot setting. Using protein families from the PANTHER database and simulated datasets with controlled evolutionary parameters, we compare trees inferred from protein language model embedding distances to trees inferred using classical phylogenetic analysis techniques and to a transformer-based distance predictor trained under explicit evolutionary models. We show that in the zero-shot setting phylogenetic signal is largely lost when sequences are represented by one fixed-sized vector, resulting in poor recovery of tree topology and branch lengths. Accumulating distances across aligned residue-level embeddings substantially improves topological accuracy, particularly for MSA-aware models, and can even match the performance of models specifically trained to infer distances for tree inference. However, distances in protein language model embedding space do not reliably reproduce evolutionary branch lengths.

Phylogenetic inference aims to estimate the evolutionary relationships among biological sequences by inferring tree topologies and branch lengths from molecular data. This fundamental problem in computational biology seeks to infer the patterns of descent that connect contemporary species or genes through their common ancestors[1]. Classical approaches rely on multiple sequence alignments and explicit evolutionary models, using maximum likelihood or distance-based frameworks to evaluate competing phylogenetic hypotheses. Maximum likelihood methods, formalized by Felsenstein[2], evaluate the probability of observed sequence data given a particular tree topology and set of evolutionary parameters. Distance-based methods, most notably the neighbor-joining algorithm[3], infer trees by iteratively clustering taxa based on pairwise evolutionary distances, making them computationally efficient and particularly well-suited for large datasets where maximum likelihood approaches may be prohibitively slow.

These analyses are commonly applied to DNA or protein sequence data and depend on probabilistic models of sequence evolution that describe substitution processes. For nucleotide sequences, models range from simple equal-rate substitution schemes to complex general time-reversible models[4], while protein evolution models include empirical matrices like PAM[5] and BLOSUM[6], or more sophisticated approaches like WAG[7] and LG[8]. Modern phylogenetic inference incorporates rate heterogeneity across sites using gamma distributions[9] and model selection frameworks[10] to select appropriate evolutionary models. Widely-used phylogenetic software packages include RAxML[11], PhyML[12], and IQ-TREE[13], which implement optimized algorithms for efficient phylogenetic inference.

Recent advances in deep learning have led to the development of protein language models (PLMs) that adapt ideas from natural language processing to biological sequences. Trained in a self-supervised manner on large protein databases such as UniRef[14] and BFD[15], these models learn rich sequence representations that capture aspects of protein structure, function, and evolutionary patterns directly from amino acid sequences. The generic nature of such models allows their internal representations to be used for structure prediction, function annotation, and evolutionary analyses[16–19] without further fine-tuning. These representations are typically obtained by extracting the activations of intermediate model layers for each residue in a sequence, yielding high-dimensional embeddings that can be pooled for a sequence-level representation.

Such models are already being explored in a phylogenetic context. Lupo et al.[20] showed that protein language models trained on multiple sequence alignments encode phylogenetic signal in their internal representations,

[1]Institute of Data Science and Digital Technologies, Vilnius University, 08412 Vilnius, Lithuania. [2]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, 02-097 Warsaw, Poland. ✉email: brendonas.stakauskas@mif.stud.vu.lt

with simple combinations of attention features strongly correlating with pairwise sequence distances. However, this work does not perform phylogenetic tree inference and does not assess whether these representations can be used to recover tree topology or branch lengths. Yeung et al.[21] used dimensionality reduction and tree-based methods to build a visualizations of sequence embeddings, demonstrating that embedding spaces reflect patterns of sequence similarity. Their analysis focuses on clustering and visualization properties of embedding spaces, and does not aim to interpret embedding distances as evolutionary branch lengths or to benchmark against standard phylogenetic reconstruction methods. Nesterenko et al.[22] presented a neural network framework Phyloformer that learns phylogenetic distances directly from simulated multiple sequence alignments. While this approach is likelihood-free, it is not model-free, as the network is trained explicitly under specific evolutionary models and simulation settings.

In this work, we investigate whether distances in the embedding space of generic protein language models can be interpreted as distances in evolutionary space. Specifically, we examine whether such distances satisfy the properties required for phylogenetic inference in zero-shot settings, without task-specific training. Our goal is to characterize the conditions under which embedding-based distances can recover phylogenetic signal, as well as to identify their limitations for evolutionary analysis.

A practical limitation of embedding-based phylogenetic analysis is that tree inference from embeddings is restricted to distance-based methods, such as neighbour joining, and cannot be directly integrated into likelihood-based or Bayesian frameworks that require explicit evolutionary models. In addition, the high dimensionality of embeddings require techniques such as mean-pooling be used to produce fixed-length vectors suitable for distance computation. Such a step can substantially reduce or distort phylogenetic signal present at the level of individual sites, motivating the development of alternative representation strategies for phylogenetic analysis.

In this study, we evaluate protein language models for phylogenetic inference in a zero-shot setting. We analyse whether distances in embedding space can be used as proxies for evolutionary distances. Using both empirical protein family data and simulated datasets we examine the effects of different embedding models, distance metrics and aggregation strategies. We show that embedding-derived distances can be used to recover phylogenetic topology under certain conditions, while also highlighting the limitations of these approaches in representing evolutionary distances.

Materials and methods

In this section, we describe the datasets, embedding models, and evaluation framework used to assess phylogenetic signal in protein language model embeddings.

Data acquisition

Empirical data

Protein families were obtained from the PANTHER 19 database. To ensure comparable input across embedding models, we included only the alignments with less than 512 positions. Gene families with a gap fraction below 10% were excluded. The remaining families were ranked by gap fraction, and the 500 families with the lowest gap content and at least 20 sequences each were selected. For these families, we downloaded the provided multiple sequence alignments (MSAs) and reference phylogenetic trees. PANTHER trees are inferred using GIGA software[23].

Simulated data

Our simulation procedure can be divided into three major phases: simulating species trees, simulating gene trees with gene duplication, gene loss and incomplete lineage sorting events, and simulating protein sequences. The parameter selection in all three phases is partially based on the simulation study from Molloy et al.[24].

First, we simulated species trees using SimPhy version 1.0.2 with a birth-death process conditioned on the number of extant species and tree age, using parameters from[24]: tree height = 1,800,000,337. 5 years, speciation rate = 1.8×10^{-9} events/year, and extinction rate = 0 events/year. The number of leaves in each species tree was set to either 12 or 20. By conditioning on a fixed tree age with all species sampled at the present, the resulting species trees are ultrametric, meaning all leaves are equidistant from the root in time.

Next, we simulated a single gene tree per species tree using SimPhy version 1.0.2[25], with each gene tree containing the same total number of leaves as its corresponding species tree. Following[24], we explored three duplication/loss (DL) rates $\{10^{-10}, 2 \cdot 10^{-10}, 5 \cdot 10^{-10}\}$ and two effective population sizes $\{10^7, 5 \cdot 10^7\}$, corresponding to low and medium levels of incomplete lineage sorting (ILS), respectively. These six parameter combinations produced a diverse collection of gene trees with varying evolutionary histories.

Gene duplication, loss, and ILS events resulted in incomplete taxon sampling and variable sequence copy numbers across simulated gene families. Consequently, some species were absent from the final gene trees while others had multiple sequences, reflecting realistic gene family evolution patterns. The lower DL and ILS parameters were calibrated using biological data from *Saccharomycetales* fungi[26], while higher parameters tested method performance under more challenging evolutionary scenarios.

Finally, for each gene tree, we simulated protein sequences using AliSim[27], implemented in IQ-TREE 2[13]. Sequences were generated under the WAG substitution model[7] with gamma-distributed rate heterogeneity across sites and insertions/deletions (WAG+G4+indels). The gamma shape parameter α was sampled from a lognormal distribution with log-mean -0.471 and log-standard deviation 0.349[24].

We incorporated indels using AliSim's "realistic" preset, which employs empirically-derived parameters[27,28]: insertion rate $= 0.03$ and deletion rate $= 0.09$ (relative to the substitution rate), yielding approximately 12 indels per 100 substitutions with a biologically realistic deletion-to-insertion ratio of 3:1. Indel lengths follow a power-law (Zipfian) distribution with exponent $a = 1.7$ and maximum length 50 amino acids, where approximately

60–70% of indels are single-character events. AliSim outputs true alignments with gap characters representing the evolutionary indels, and all sequences in the alignment have the same total length of 500 (including gaps).

In summary, for each set of parameters of duplication-loss rates, population sizes, and leaf-set sizes we simulated 100 species trees and 100 corresponding true gene trees with alignments. The true gene trees were used as references to assess the accuracy of phylogenetic inference from the simulated alignments in our pipeline.

Embedding models

Single-sequence protein language models included ProtT5 (Rostlab/prot_t5_xl_uniref50), ProtBERT (Rostlab/prot_bert)[29], E1[30], and ESM-C[31] (esmc_600m). For each model, embeddings were extracted for all residues in the input sequence. To obtain a single fixed-dimensional representation per sequence, residue-level embeddings were aggregated by mean pooling across sequence positions. For ESM-C, we additionally evaluated the CLS token as an alternative sequence-level representation.

For MSA-aware approaches we used MSA Pairformer[32] and Phyloformer (PF checkpoint). MSA Pairformer outputs embeddings with cross-sequence dependencies; mean pooling across residues was used to obtain sequence-level vectors. Phyloformer produces full distance matrices and therefore no pooling was applied.

Phylogenetic tree inference from embeddings and MSAs

All phylogenetic trees were reconstructed from pairwise distance matrices using FastME 2.0 with the Neighbour Joining[3] algorithm. Although negative branch lengths can occur when distance matrices violate the additive tree metric assumption, this was rare in our datasets, and such cases were excluded from downstream analyses.

For trees inferred from multiple sequence alignments (NJ-MSA trees), pairwise distances were computed using BLOSUM62-based amino acid substitution scores[6]. For trees inferred from sequence embeddings (NJ-embedding trees), distances were calculated using cosine and Euclidean distances.

Aggregating distances

Sequence-level embeddings summarize information across alignment positions, which may limit direct access to site-specific variation when used for phylogenetic inference. As an alternative, we consider distance construction strategies that operate on residue-level embeddings aligned to a MSA. This formulation allows distances to be aggregated from individual alignment positions or from subsets of positions, providing a flexible framework for evaluating different aggregation schemes.

Within this framework, we evaluate three classes of distance aggregation strategies: sliding-window aggregation, in which site embeddings are pooled over contiguous alignment regions to incorporate local context; sorted coverage aggregation, where distances are constructed from subsets built in an information-quantity–based manner; and weighted subset consensus, which aggregates distances from randomly sampled subsets of alignment positions and infers trees from the resulting distances.

For this approach, we focused on ESM-C and MSA Pairformer. Results were compared to those obtained with Phyloformer using a published model checkpoint from the authors' GitHub repository. As Phyloformer provides only full distance matrices, window-based aggregation could not be applied. In contrast, MSA Pairformer enables per-residue embedding extraction and was therefore subjected to the same distance accumulation procedures. The applied parameters are summarized in Table 1.

Sliding-window aggregation

To incorporate local context, a sliding-window strategy was applied. Embeddings were average-pooled within overlapping windows of adjacent alignment positions, and distance matrices were computed for each windowed subset. Window-based and site-wise (single-position) distance matrices were treated uniformly in downstream analyses. Window size and overlap were varied to control the extent of contextual aggregation, with window size 1 corresponding to site-wise aggregation and a single full-length window corresponding to the previous method of sequence-level embeddings.

Category	Values	Method
Models	ESM-C, MSA Pairformer	All
Distance metrics	Cosine, Euclidean	All
Sliding window sizes	10, 15, 20, 25	Sliding window
Window overlap sizes	1, 2, 3	Sliding window
Special window cases	Window size 1 and sequence length, overlap 0	Sliding window
Site weighting schemes	Unit, unique residue count, Shannon entropy	Sorted coverage, Weighted subset consensus
Sorted coverage thresholds (%)	80, 85, 90, 95	Sorted coverage
Coverage selection direction	Most informative, least informative	Sorted coverage
Distance normalization	Weight sum, site count	Sorted coverage, Weighted subset consensus
Random subset coverage (%)	30, 50, 70, 90	Weighted subset consensus
Subset replicates per setting	10, 30, 50, 70	Weighted subset consensus

Table 1. Summary of parameters used across distance aggregation experiments.

Sorted coverage

Residue-level embeddings were extracted for each alignment position in the multiple sequence alignment, similar to the sliding-window approach with window size one. Each alignment position was assigned a weight using three alternative schemes: a uniform weight, the number of unique residues observed at the site, and Shannon entropy computed over the residue distribution.

Sites were ranked by weight to reflect site informativeness. Distance matrices were constructed by aggregating site-wise distances in rank order until a predefined coverage threshold was reached, using either descending or ascending weight order. Distance normalization was performed either by the sum of site weights or by the number of contributing sites.

Weighted subsets consensus

Using the same weighting framework, we generated random subsets of alignment sites. For each replicate, sites were sampled uniformly at random until the accumulated site weights reached a predefined coverage threshold. This procedure was repeated to generate multiple subsets per configuration. For each subset, distance matrices were computed and phylogenetic trees were inferred independently. The resulting trees were then combined using majority-rule consensus with a threshold of 50% and extended majority-rule consensus as implemented in IQ-TREE2[13].

Measures to compare unrooted phylogenetic trees

Tree similarity was quantified using the normalized forms of the following measures: classical Robinson–Foulds distance (RF)[33], branch score distance (KF)[34], Nye similarity[35], and shared phylogenetic information (SPI)[36]. These metrics capture both topological differences and discrepancies in branch lengths between trees. Normalized similarity measures (SPI and Nye) were converted to distances by computing 1 minus similarity, so that all metrics a consistent interpretation, with 0 indicating identical trees and larger values indicating greater dissimilarity. In addition, all metric values, except KF, are in the range [0,1].

To control for dataset complexity, we adjusted all scores relative to the NJ-MSA baseline computed from BLOSUM62 distances. For each dataset, distances were computed for all methods and, for every metric, the corresponding NJ-MSA value was subtracted according to $DIST^{i}_{\text{relative}} = DIST^{i}_{\text{raw}} - DIST^{i}_{\text{NJ-MSA}}$, i is a dataset number. This per-dataset adjustment normalizes results against a common reference, ensuring that comparisons reflect methodological differences rather than inherent dataset difficulty. Under this convention, negative values indicate performance better than the NJ-MSA baseline, whereas positive values indicate worse performance.

For comparison purposes branch lengths were normalized by the total sum of all branch weights present in a tree.

Score errors (where applicable) are reported as a standard deviation.

Results

Sequence-level embeddings lack phylogenetic signal

We first analyse the results of the comparative study for the PANTHER dataset. A summary is depicted in Fig. 1 and Table 2. We observe that using sequence-level embeddings, either by averaging residue embeddings or by relying on a special sequence token, results in poor performance with respect to both topology and branch-length

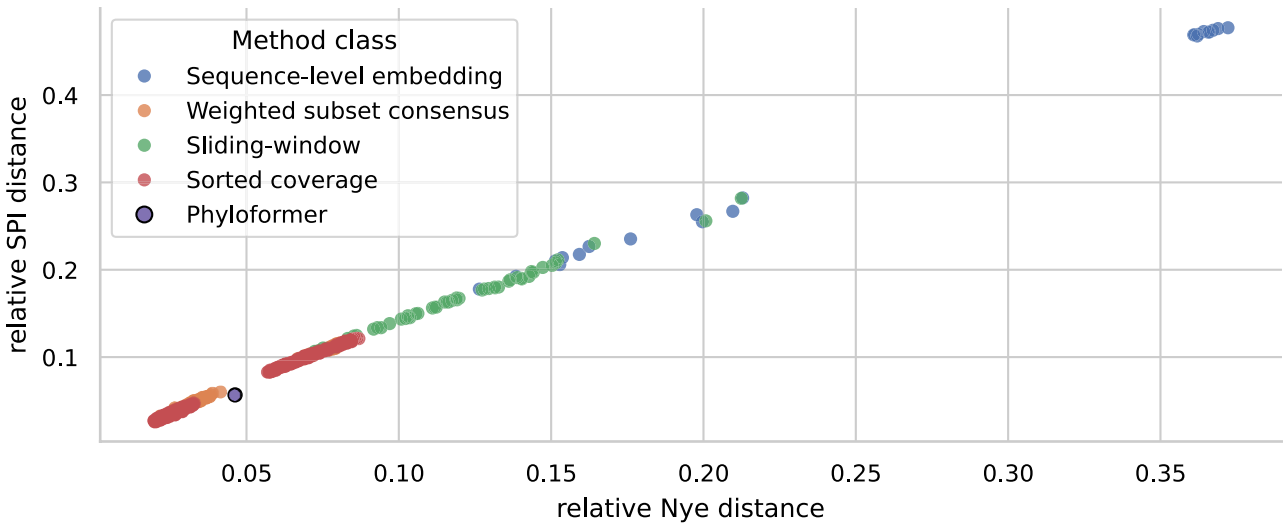

Fig. 1. Scatter plot of results for the PANTHER dataset. Each point represents a parameter configuration, coloured by experiment type. Scores are shown relative to the NJ-MSA baseline and averaged across 500 datasets. Phyloformer result is highlighted as a reference point. Points in the lower-left region indicate better performance.

Model	Metric used	Score
MSA Pairformer	Euclidean	0.23 ± 0.16
MSA Pairformer	Cosine	0.29 ± 0.17
ProtT5	Euclidean	0.34 ± 0.20
ESMC-CLS	Euclidean	0.35 ± 0.20
E1	Euclidean	0.36 ± 0.19

Table 2. Top five results of tree inference using a single vector sequence representation on the PANTHER dataset. Results are ranked by normalized Robinson–Foulds distance, relative to the NJ-MSA baseline.

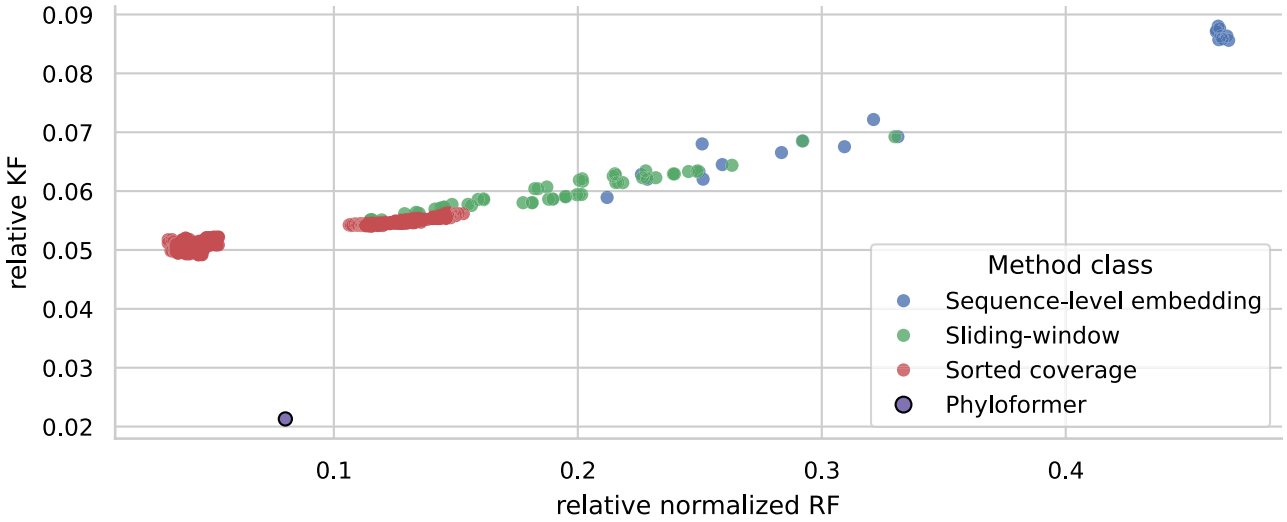

Fig. 2. Same as Fig. 1, but with scores focusing on branch lengths. Consensus methods often infer multifurcating trees, thus they are not included here.

Model	Metric used	Window size	Overlap	Score
MSA Pairformer	Cosine	1	0	0.04 ± 0.11
MSA Pairformer	Euclidean	1	0	0.04 ± 0.11
MSA Pairformer	Euclidean	10	2	0.12 ± 0.14
MSA Pairformer	Euclidean	10	1	0.12 ± 0.14
MSA Pairformer	Euclidean	10	3	0.12 ± 0.14

Table 3. Top five results of tree inference using windowed aggregation on the PANTHER dataset. Results are ranked by normalized Robinson–Foulds distance relative to the NJ-MSA baseline.

preservation. While the relative normalized Robinson–Foulds distance for the PANTHER dataset obtained with Phyloformer was 0.08 ± 0.15 as indicated in Fig. 2, none of the protein or MSA language models approached this level of accuracy (see Table 2).

Reducing each sequence to a single vector consistently fails to preserve phylogenetic structure. Because evolutionary signal is distributed unevenly across alignment sites, full sequence averaging effectively removes informative variation, resulting in trees that poorly match the alignment-based baseline.

Recovering phylogenetic structure from residue-level embeddings

In contrast to sequence-level representations, retaining residue-level information from aligned positions substantially improves phylogenetic inference. Windowed aggregation experiments showed that the strongest signal is obtained in the limiting case where each alignment position is treated independently (window size one). As shown in Table 3, MSA Pairformer consistently outperforms ESM-C, while window overlap has little influence on overall performance.

The effect of window size shows a clear monotonic trend. Site-wise distance accumulation (window size one) yields the best agreement with reference trees, with a relative normalized Robinson–Foulds distance of 0.09 ± 0.15. Increasing the window size progressively degrades topological accuracy, reaching 0.20 ± 0.17 for

Model	Metric used	Weights	Coverage	Weight priority	Score
MSA Pairformer	Euclidean	Shannon	80	High first	0.03 ± 0.11
MSA Pairformer	Euclidean	Shannon	90	High first	0.03 ± 0.11
MSA Pairformer	Cosine	Shannon	90	High first	0.03 ± 0.11
MSA Pairformer	Euclidean	Shannon	80	High first	0.03 ± 0.11
MSA Pairformer	Cosine	Shannon	95	High first	0.03 ± 0.12

Table 4. Top five results of tree inference using the sorted coverage approach on the PANTHER dataset. Results are ranked by normalized Robinson–Foulds distance relative to the NJ-MSA baseline.

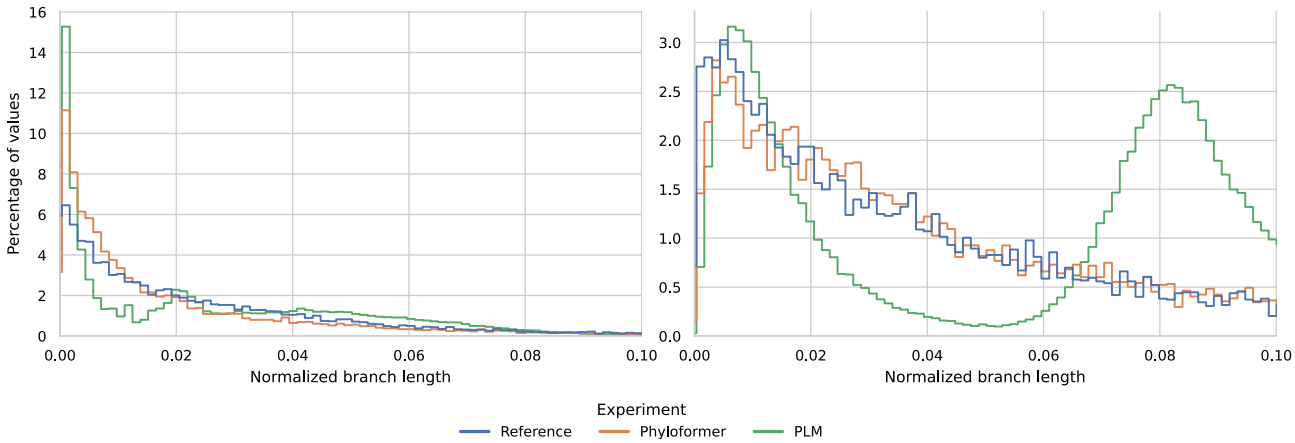

Fig. 3. Histogram of normalized branch lengths across different methods of inference. Left shows empirical data and right—simulated. PLM line groups all trees inferred from sequence-level embedding distance matrixes.

a window size of 25, while aggregation across the full sequence performs worst with score of 0.28 ± 0.19 . This behaviour mirrors that observed for sequence-level embeddings, where aggregation across many sites reduces phylogenetic signal.

These results indicate that phylogenetic signal arises from accumulating distances across alignment positions rather than from aggregating embeddings into a single sequence representation. Moreover, the use of all alignment sites is not required to obtain improved trees, as some positions contribute little or no informative signal. Figure 1 shows that restricting distance accumulation to subsets of alignment sites can yield better performance than window-based aggregation.

The observed separation between result clusters is largely explained by model choice, with MSA Pairformer consistently outperforming ESM-C across comparable settings (Table 4).

We additionally evaluated weighted subset consensus trees constructed from multiple randomly sampled site subsets. Across configurations, this approach yielded results comparable to those obtained with sorted coverage aggregation, without consistently reducing relative Robinson–Foulds (RF) distances. Consistent with earlier experiments, MSA Pairformer produced lower relative RF distances than ESM-C across all evaluated consensus settings. The mean relative RF distance for ESM-C was 0.13 ± 0.16 , compared to 0.04 ± 0.12 for MSA Pairformer.

A limitation of the consensus approach is the frequent occurrence of multifurcations in the inferred trees. This complicates direct comparison with fully bifurcating reference trees.

Embedding space distances do not represent evolutionary distances

Although it is possible to obtain results without fine-tuning a model, a major drawback of using pre-trained models without fine-tuning is the difficulty of reproducing the branch lengths. Phyloformer is better at preserving branch lengths, as shown in Fig. 2.

Branch-length difference is clearly seen in Fig. 3. There are more short branches than in reference or Phyloformer trees. While reference and Phyloformer branch-length distributions are smooth and unimodal PLM branch lengths have a way steeper drop-off followed by increase of branch lengths. This pattern suggests that distances in embedding space do not map uniformly to evolutionary divergence. This behaviour is observable in trees inferred by using weighted/windowed representation techniques as well.

Simulated data

The findings presented in previous sections are supported by results from experiments with simulated datasets. As shown in Fig. 4, embedding-based methods recover tree topology with accuracy comparable to Phyloformer,

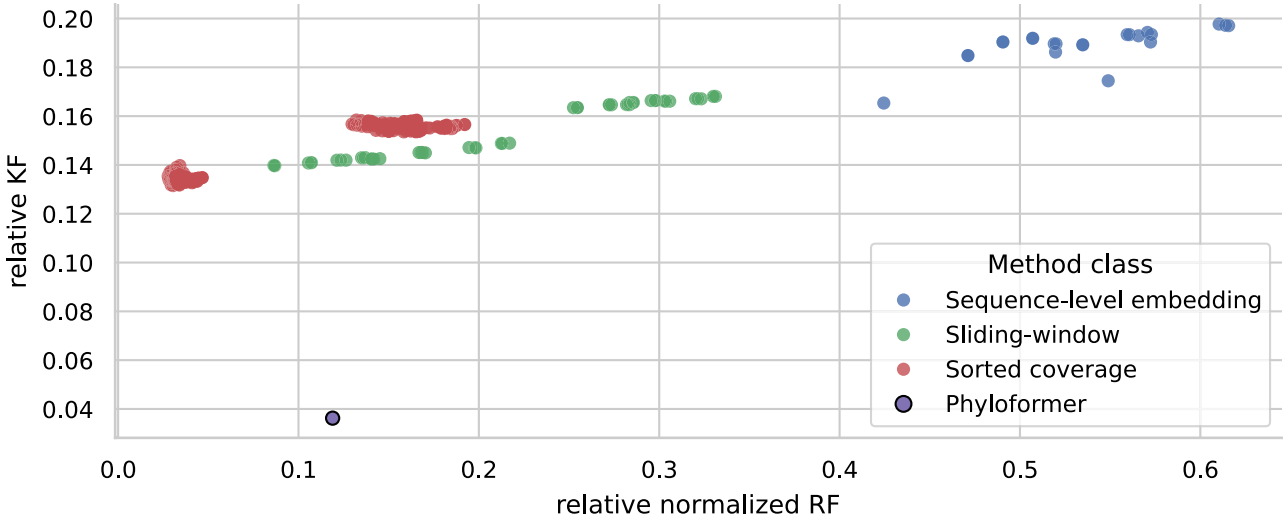

Fig. 4. Same as Fig. 1 but with simulated dataset. Scoring methods focuses both on topology (x-axis) and branch lengths (y-axis).

but consistently underperform in preserving branch lengths, as reflected by higher branch score distances. In the simulated setting, as can be seen in Fig. 3, this branch-length distortion is more pronounced, and normalized branch-length distributions inferred from embedding distances exhibit a clear multimodal structure driven by an increased proportion of short branches. This indicates that the limitations of embedding-space distances in representing evolutionary divergence persist even under controlled evolutionary simulations.

Discussion and conclusion

In this study, we evaluated protein language models for phylogenetic inference in a zero-shot setting. Using both empirical protein family data and simulated datasets, we examined whether distances constructed from protein language model representations can be interpreted as evolutionary distances. We found that phylogenetic topology can be recovered when distances are accumulated across aligned residue positions, whereas trees inferred from sequence-level representations fail to preserve branch lengths.

For both empirical and simulated data, branch lengths were not preserved in trees inferred from PLM embedding distances. Normalized branch-length histograms reveal the emergence of an additional peak for embedding-based trees that is absent in the reference data. In contrast, the histograms of reference trees and Phyloformer-inferred trees are similar in shape, differing primarily in the relative frequency of very short branches. This effect is even more pronounced in the simulated datasets, where normalized branch-length distributions become distinctly multimodal, suggesting that embedding-space distances may not scale linearly with evolutionary divergence.

For comparison with Phyloformer, we used a single set of pretrained weights provided by the authors; alternative checkpoints were not evaluated, and our results therefore reflect the performance of this specific published configuration. Across the evaluated protein language models, MSA-aware architectures consistently outperformed single-sequence models. In particular, MSA Pairformer showed improved recovery of phylogenetic structure compared to ESM-C, indicating that incorporating site information across sequences is important for extracting evolutionary signal. In contrast, sequence-level representations obtained by mean pooling residue embeddings consistently failed to recover phylogenetic structure. Improved performance was observed only when distances were accumulated across aligned positions, suggesting that mean pooling does not preserve phylogenetic signal and that alternative representations or aggregation strategies are required. Even simple site-based distance construction strategies yielded substantial improvements, suggesting further study direction.

The best-performing strategies in this study still rely on multiple sequence alignments, which raises the question of practical advantage over classical phylogenetic methods. Our aim was not to propose an alternative pipeline, but to find whether the phylogenetic signal resides in pretrained protein language model representations. Signal was present at the residue level but is lost through sequence-level pooling. This signal may be valuable in contexts where vectorial representations of evolutionary relationships are needed. This also serves as a caution against naive use of distances in embedding space for phylogenetics. Finding ways to recover this residue-level signal without relying on multiple sequence alignments is a promising direction for future work.

Our conclusions are limited to the subset of protein language models evaluated in this study, and may not generalize to all available architectures. Furthermore, all models were used in a zero-shot setting without task-specific adaptation. It remains unclear whether the observed limitations are inherent to protein language model representations or arise from the zero-shot setting. Phyloformer, which was specifically trained to predict evolutionary distances, showed better branch-length recovery. This suggests that task-specific training may be beneficial, however, neither fine-tuning the PLMs themselves nor training dedicated models on top of their embeddings was explored in this study.

Future directions include evaluating diverse empirical datasets and testing more complex simulation scenarios, particularly those with more indel patterns. Increasing sample sizes may improve the precision of our estimates, though we expect the core findings to remain unchanged. We note that our analysis relies on distance-based methods, particularly neighbor-joining, which are less statistically sophisticated than maximum likelihood approaches used in phylogenetics. However, it remains unclear how to effectively apply likelihood-based methods in the embedding space setting, where explicit evolutionary models and probabilistic frameworks are absent. Another important extension would be analyzing multifurcating trees, which frequently occur in phylogenetic studies when insufficient signal exists to resolve all relationships. This would provide a more complete picture of protein language models' capabilities for evolutionary inference.

Data availability

Pipeline is available at GitHub: https://github.com/brendonsa/phyloembed. Inferred trees are uploaded to Zeno do. DOI https://doi.org/10.5281/zenodo.18098552.

Received: 31 December 2025; Accepted: 9 June 2026
Published online: 17 June 2026

References

1. Felsenstein, J. *Inferring Phylogenies* (Sinauer Associates, 2004).
2. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376. https://doi.org/10.1007/BF01734359 (1981).
3. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454 (1987).
4. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **17**, 57–86 (1986).
5. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* 5, 345–352 (National Biomedical Research Foundation, 1978).
6. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10915–10919. https://doi.org/10.1073/pnas.89.22.10915 (1992).
7. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851 (2001).
8. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320. https://doi.org/10.1093/molbev/msn067 (2008).
9. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314. https://doi.org/10.1007/BF00160154 (1994).
10. Posada, D. jModelTest: phylogenetic model averaging. *Mol. Biol. Evol.* **25**, 1253–1256. https://doi.org/10.1093/molbev/msn083 (2008).
11. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033 (2014).
12. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321. https://doi.org/10.1093/sysbio/syq010 (2010).
13. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534. https://doi.org/10.1093/molbev/msaa015 (2020).
14. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288. https://doi.org/10.1093/bioinformatics/btm098 (2007).
15. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542. https://doi.org/10.1038/s41467-018-04964-5 (2018).
16. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* **118**, e2016239118. https://doi.org/10.1073/pnas.2016239118 (2021).
17. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130. https://doi.org/10.1126/science.ade2574 (2023).
18. Rao, R. et al. Evaluating protein transfer learning with tape. *Adv. Neural Inf. Process. Syst.* **32**, 9689–9701 (2019).
19. Hie, B., Zhong, E. D., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. *Science* **371**, 284–288. https://doi.org/10.1126/science.abd7331 (2021).
20. Lupo, U., Sgarbossa, D. & Bitbol, A.-F. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *Nat. Commun.* **13**, 6298. https://doi.org/10.1038/s41467-022-34032-y (2022).
21. Yeung, W. et al. Tree visualizations of protein sequence embedding space enable improved functional clustering of diverse protein superfamilies. *Brief. Bioinform.* **24**, bbac619. https://doi.org/10.1093/bib/bbac619 (2023).
22. Nesterenko, L., Blassel, L., Veber, P., Boussau, B. & Jacob, L. Phyloformer: fast, accurate, and versatile phylogenetic reconstruction with deep neural networks. *Mol. Biol. Evol.* **42**, msaf051. https://doi.org/10.1093/molbev/msaf051 (2025).
23. Thomas, P. D. GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinf.* **11**, 312. https://doi.org/10.1186/1471-2105-11-312 (2010).
24. Molloy, E. K. & Warnow, T. TreeShrink: Fast and Accurate Detection of Outlier Long Branches in Collections of Phylogenetic Trees. *BMC Genomics* **21**, 332. https://doi.org/10.1186/s12864-020-6620-x (2020).
25. Mallo, D., De Oliveira Martins, L. & Posada, D. SimPhy: phylogenomic simulation of gene, locus, and species trees. *Syst. Biol.* **65**, 334–344. https://doi.org/10.1093/sysbio/syv082 (2016).
26. Rasmussen, M. D. & Kellis, M. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.* **22**, 755–765. https://doi.org/10.1101/gr.123901.111 (2012).
27. Ly-Trong, N., Naser-Khdour, S., Lanfear, R. & Minh, B. Q. AliSim: a fast and versatile phylogenetic sequence simulator for the genomic era. *Mol. Biol. Evol.* **39**, msac092. https://doi.org/10.1093/molbev/msac092 (2022).
28. Benner, S. A., Cohen, M. A. & Gonnet, G. H. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol.* **229**, 1065–1082. https://doi.org/10.1006/jmbi.1993.1105 (1993).
29. Elnaggar, A. et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381 (2022).
30. Jain, S., Beazer, J., Ruffolo, J. A., Bhatnagar, A. & Madani, A. E1: retrieval-augmented protein encoder models. *bioRxiv*. arXiv:2025.11126 (2025).
31. ESM Team. *ESM Cambrian: Revealing the Mysteries of Proteins with Unsupervised Learning* (2024).

32. Akiyama, Y., Zhang, Z., Mirdita, M., Steinegger, M. & Ovchinnikov, S. Scaling down protein language modeling with MSA Pairformer. *bioRxiv* https://doi.org/10.1101/2025.08.02.668173. https://www.biorxiv.org/content/early/2025/08/03/2025.08.02.668173.full.pdf (2025).
33. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147. https://doi.org/10.1016/0025-5564(81)90043-2 (1981).
34. Kuhner, M. K. & Felsenstein, J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11**, 459–468 (1994).
35. Nye, T. M. W., Lio, P. & Gilks, W. R. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics* **22**, 117–119. https://doi.org/10.1093/bioinformatics/bti720 (2006).
36. Smith, M. R. Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. *Bioinformatics* **36**, 5007–5013. https://doi.org/10.1093/bioinformatics/btaa614 (2020).

Acknowledgements
The research was supported by Excellence Initiative – Research University (2020–2026) at University of Warsaw. The research was partially funded by the National Science Center grant 2023/51/B/ST6/02792.

Author contributions
B.S. built the experimental pipeline and conducted the experiments, P.G. built data simulation pipeline. Both authors wrote and reviewed the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to B.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.