PHYLA: TOWARDS A FOUNDATION MODEL FOR PHYLOGENETIC INFERENCE

Andrew Shen[1,2,*], **Yasha Ektefaie**[1,‡,*], **Lavik Jain**[3], **Maha Farhat**[1,4,‡], **Marinka Zitnik**[1,5,6,7,‡]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2]Department of Computer Science, Northwestern University, Evanston, IL, USA
[3]Harvard University, Cambridge, MA, USA
[4]Division of Pulmonary and Critical Care, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA
[5]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Allston, MA, USA
[6]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[7]Harvard Data Science Initiative, Cambridge, MA, USA

[*]Co-first authors
‡Corresponding authors: yasha_ektefaie@hms.harvard.edu, maha_farhat@hms.harvard.edu, marinka@hms.harvard.edu

ABSTRACT

Deep learning has made strides in modeling protein sequences but often struggles to generalize beyond its training distribution. Current models focus on learning individual sequences through masked language modeling, but effective protein sequence analysis demands the ability to reason across sequences, a critical step in phylogenetic analysis. Training biological foundation models explicitly for inter-sequence reasoning could enhance their generalizability and performance for phylogenetic inference and other tasks in computational biology. Here, we report an ongoing development of PHYLA, an architecture that operates on an explicit, higher-level semantic representation of phylogenetic trees. PHYLA employs a hybrid state-space transformer architecture and a novel tree loss function to achieve state-of-the-art performance on sequence reasoning benchmarks and phylogenetic tree reconstruction. To validate PHYLA's capabilities, we applied it to reconstruct the tree of life, where PHYLA accurately reclassified archaeal organisms, such as Lokiarchaeota, as more closely related to bacteria—aligning with recent phylogenetic insights. PHYLA represents a step toward molecular sequence reasoning, emphasizing structured reasoning over memorization and advancing protein sequence analysis and phylogenetic inference.

1 INTRODUCTION

Protein language models (PLM) use transformers with masked language or autoregressive self-supervision to model molecular sequences (Rives et al., 2021; Lin et al., 2022; Alley EC, 2019; Madani, 2023; Notin, 2022). PLMs have shown state-of-the-art performance across predictive (Meier et al., 2021; Rives et al., 2021; Rao et al., 2021; Elnaggar et al., 2021; Alley EC, 2019; Rao et al., 2020) and generative (Lin et al., 2022; Hayes et al., 2024; Madani, 2023; Ferruz, 2022) tasks. Despite the advantages of PLMs, they are limited in the context length of their inputs due to the quadratic nature of self-attention (Vaswani et al., 2017). State-space models have emerged as a way to increase the size of genomic context that can be integrated into a model (Gu & Dao, 2024; Poli et al., 2023; Sgarbossa et al., 2024). Hybrid architectures have emerged to combine the strength of transformers with state-space models (Nguyen et al., 2023; 2024). Despite the advances in model architecture and masked language modeling, self-supervised tasks used to train these models have remained largely the same. Recent studies have highlighted the shortcomings of masked language modeling from the lens of generalizability (Ektefaie et al., 2024), where learned representations are

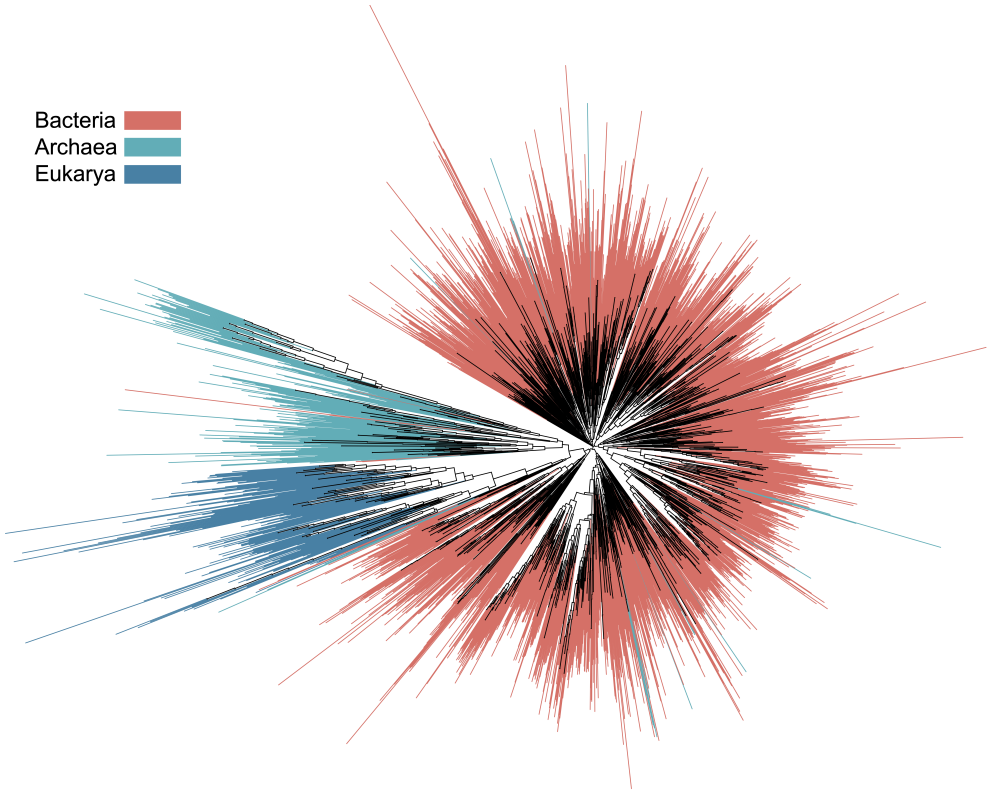

Figure 1: **PHYLA constructed phylogenetic tree of 3,084 ribosomal protein sequences among organisms in all the taxonomic domains of life.** Organisms classified as bacteria are labeled in red, archaea in light blue, and eukarya in dark blue.

heavily biased by training dataset composition (Ding & Steinhardt, 2024) or do not prove useful for many tasks (Li et al., 2024), leading to worse performance of generalist foundation models than their specialized counterparts in some cases.

PLMs are trained to model individual protein sequences but are not designed to reason across sequences. As such, these models excel at capturing *intra-sequence relationships* but are not explicitly trained to handle *inter-sequence relationships*. This limitation is rooted in the architecture and training paradigm of PLMs. *Understanding evolutionary relationships requires the identification of similarities and differences across sequences before modeling finer details at the amino acid level within each sequence.* PLMs implicitly learn some degree of inter-sequence relationships through exposure to large datasets, but they lack explicit architectural or training features designed to reason systematically across sequences. We hypothesize that models with explicit hierarchical reasoning capabilities, tailored to compare and integrate information across sequences, will perform better at tasks like phylogenetic inference and functional annotation.

Important biological insights can be discerned from reasoning across sequences, whether it is understanding a taxon's position in the tree of life (Hug, 2016), determining the impact of a protein variant (Meier et al., 2021; Brandes, 2023; et al., 2023), or annotating functions of poorly characterized proteins (Nguyen et al., 2024; Avsec, 2021; Zvyagin et al., 2022; Queen et al., 2024). PLMs using masked language tasks learn to implicitly compare sequences by observing many variations of the same sequence. However, these models tend to memorize variations rather than compare sequences, limiting their ability to generalize to unseen sequences or variants (Ektefaie et al., 2024). We hypothesize that a PLM explicitly trained to compare sequences can achieve better generalization than current models.

Assessing a model's ability to reason across molecular sequences is challenging because the specifics of sequence comparison vary with the biological application. Biologists use multiple sequence alignment algorithms to inform phylogenetic trees. Though these algorithms are foundational tools

for hypothesis generation (Chatzou et al., 2015), they are computationally intensive, with time and resource requirements that scale exponentially with sequence length and number (Katoh et al., 2002; Sievers et al., 2011). We propose that the ability to reconstruct phylogenetic trees from model-generated embeddings can serve as a proxy for evaluating sequence reasoning. Excelling at this task requires models to identify and prioritize differences between protein sequences and assess their impact on protein phylogeny. This demands the ability to reason across multiple sequences rather than modeling each sequence independently.

Present work. We report an ongoing development of PHYLA models, an approach that shifts away from processing at the token or sequence level and moves toward hierarchical reasoning in an abstract embedding space tailored for proteins. This embedding space captures relationships and insights independent of individual sequences, focusing instead on the underlying semantic and functional connections across proteins. PHYLA is a hybrid state-space and sparsified attention model trained on a novel tree loss and masked language loss. PHYLA models the reasoning process at a phylogenetic and functional level, rather than being confined to individual amino acid tokens or specific sequence alignments. PHYLA is explicitly trained to reconstruct the phylogenetic trees of protein sequences during training, in addition to predicting the identity of masked amino acids.

To evaluate PHYLA's generated trees, we introduce a sequence reasoning benchmark designed to assess the ability of PLMs to reconstruct phylogenetic trees and perform phylogenetic inference based on these trees. We find that PLMs exhibit significant limitations in their ability to reason across sequences. PHYLA achieves stronger performance in phylogenetic tree reconstruction and functional prediction despite having significantly fewer parameters (291M parameters) compared to larger models. On the OpenFold small benchmark, PHYLA achieves a normalized Robinson-Foulds (Robinson & Foulds, 1981) (normRF) metric of 0.8187, outperforming much larger models like ESM3 (1.4B parameters) and Evo (7B parameters). On the ProteinGym benchmark, PHYLA achieves a Spearman's rank correlation of 0.696, which is competitive with models like ESM2 (3B parameters) and better than several other methods. We use PHYLA to reconstruct the tree of life from ribosomal protein sequences of eukarya, archaea, and bacteria (Figure 1). Our analysis suggests that PHYLA captures relationships among subspecies of bacteria and archaea, differing from the current tree of life and aligning more closely with the functional characteristics of these subspecies. This motivates future development of PHYLA as part of a new generation of sequence models aimed at advancing biological sequence reasoning.

2 RELATED WORK

Protein Language Models (PLMs). State-of-the-art protein language models include transformer-based models such as ESM2 (Lin et al., 2022) and ProGen (Madani, 2023) that are trained using masked or autoregressive language modeling. These models learn to model the language of proteins by learning the co-occurrence of amino acid residues within a diverse training set. Other PLMs, such as ESM3 (Hayes et al., 2024), model additional data modalities. ESM3 considers structural and functional information in addition to the background amino acid sequences. These models have demonstrated good performance on intra-sequence reasoning from sequence modeling pre-training tasks but have not explicitly been trained on inter-sequence reasoning between different sequences in the training set.

Alternatives to self-attention. Self-attention is the backbone of the transformer but suffers from quadratic scaling with sequence length, making modeling longer protein sequences difficult (Vaswani et al. (2017)). The Mamba state-space architecture has been proposed as an alternative backbone architecture for sequence-based foundation models. The architecture builds upon the S4 class of structured state-space models (Gu et al. (2022)) by adding a selection mechanism and a hardware-aware parallel algorithm. These advances allow Mamba to model long sequences efficiently. Beyond Mamba, other approaches use similar ideas to extend context length, including Hyena (Poli et al., 2023) and xLSTM (Beck et al., 2024).

Bioinformatics approaches to phylogenetic analysis. Traditional tree reconstruction methods for a set of input protein sequences consist of generating a multiple sequence alignment (MSA) using one of many alignment algorithms. The MAFFT and Clustal Omega alignment algorithms are popular choices for efficient and accurate MSA generation (Katoh et al. (2002); Sievers et al. (2011)). These alignment algorithms align the input sequences by matching the location of the most conserved

amino acids within the sequences. After generating the MSA, a phylogenetic tree is reconstructed using a tree reconstruction algorithm, like FastTree and IQTree (Price et al. (2010); Nguyen et al. (2014)). These algorithms infer the structure of the phylogenetic tree with and without parametric models and usually with various heuristics to generate the most likely phylogenetic tree topology. The primary limitation of tree reconstruction is runtime inefficiency as tree sizes grow.

3 PHYLA APPROACH

We design and implement PHYLA, which explicitly operates at two levels of abstraction: sequence tokens and phylogenetic trees. We define a phylogenetic tree as a higher-order abstraction that encapsulates evolutionary and functional relationships among sequences. This approach stands in sharp contrast to current PLMs, which are primarily sequence-centric and token-based, lacking the capacity to perform structured, multi-level reasoning across datasets. PHYLA introduces a hierarchical architecture that enables reasoning in this abstract embedding space, setting a new standard for protein modeling and phylogenetic inference.

This paper aims to provide proof of concept for this high-level vision of an alternative architecture to current best practices in protein language modeling.

3.1 OVERVIEW OF PHYLA MODEL ARCHITECTURE

Given a set of protein sequences S, the goal is to construct a phylogenetic tree T of S. To address this problem, we propose a hybrid state-space transformer model, PHYLA. During training, a phylogenetic tree T is sampled, where T consists of N sequences S. Each sequence is tokenized into a stream of 22 tokens, corresponding to 20 standard amino acids, a mask token, and a pad token. The input to PHYLA is S with a $[CLS]$ token concatenated in front of each tokenized sequence, $s \in S$: $\{[CLS]s_1 \parallel [CLS]s_2 \parallel [CLS]s_3, ..., [CLS]s_n\}$ and the output is a phylogenetic tree which is then compared to the sampled tree to calculate the loss. The size and number of trees considered in each training step are determined at each training step by an adaptive batch size sampler.

The architecture of PHYLA comprises of a sequence of blocks, each containing 16 Mamba layers (Gu & Dao (2024)) followed by a sparsified self-attention layer. The sparsified self-attention employs an attention mask M:

$$M_{ij} = \begin{cases} 1, & \text{if the } j\text{-th token is within the } i\text{-th sequence,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

This architecture incorporates inductive biases tailored to sequence comparison. Specifically, Mamba layers facilitate inter-sequence comparisons, capturing relationships between different sequences, and sparsified attention layers apply self-attention between the CLS token of each input protein sequence and its sequence positions to perform intra-sequence comparisons. PHYLA is trained on 13,696 phylogenetic trees from OpenProteinSet (Ahdritz et al., 2023) with 40 sequence blocks, and the current model release has 291 million model parameters.

3.2 ADAPTIVE BATCH SIZING

We employ an adaptive batch sizing approach to efficiently utilize GPU memory and avoid overfitting to a specific tree topology. We determine the largest subtree $t \in T$ at every training step that can fit within the available GPU memory. Next, we randomly sample a subtree size n such that $5 \leq n \leq |t|$, where $|t|$ is the number of sequences in t. Finally, we identify how many subtrees of the sampled size $|t|$ can be accommodated within the GPU memory. If the model encounters an out-of-memory (OOM) error during this process, the subtrees are resampled with both the subtree size and the number of subtrees halved. Details are given in Appendix A.1.

3.3 PHYLA LOSS FUNCTION

Phyla's loss function is a combination of a masked language loss (MLM) and tree loss (TREE):

$$L_{\text{PHYLA}} = L_{\text{TREE}} + L_{\text{MLM}}. \tag{2}$$

To compute the tree loss, we first normalize the distance matrix of the sampled tree, D, by dividing each element D_{ij} by the maximum value of its corresponding row i:

$$D'_{ij} = \frac{D_{ij}}{\max(D_{i1}, D_{i2}, ..., D_{iN})} \tag{3}$$

Next, we compute the pairwise distances between embeddings of CLS tokens of the sequences to create a predicted distance matrix P. We then row-normalize P in the same way as D. L_{TREE} is the L1 loss between the row-normalized distance matrices:

$$L_{\text{TREE}} = \sum_{i}^{n} \sum_{j}^{n} |D'_{ij} - P'_{ij}| \tag{4}$$

Lastly, for the MLM loss, we mask 15% of the input sequence and have the model predict the identity of the masked sequences as described previously (Devlin et al., 2019).

4 EXPERIMENTS

Datasets. We evaluate the ability to reconstruct trees using a held-out subset of the OpenProteinSet (Ahdritz et al., 2024) comprising 119 trees. The trees are stratified into three categories based on the number of sequences in the tree: "Openfold Small" (0 to 1,000 sequences, 45 trees), "Openfold Medium" (1,000 to 2,000 sequences, 45 trees), and "Openfold Large" (2,000+ sequences, 29 trees). We also evaluate the ability to predict functional labels using 83 datasets from the ProteinGym (Notin et al., 2023a) benchmark. The 83 datasets were chosen based on which would fit on a single 80GB H100 GPU during inference.

Baselines. We consider two protein language models, one genomic foundation model, six models from the ProteinGym benchmark, and two traditional tree reconstruction methods. The protein language models include ESM2 and ESM3 (Lin et al. (2022); Hayes et al. (2024)). The genomic foundation model is Evo (Nguyen et al. (2024)). The 6 ProteinGym benchmarks include ProteinNPT, MSA Transformer, ESM-1v, Tranception, TranceptEVE, and DeepSequence (Notin et al. (2023a;b); Rao et al. (2021); Meier et al. (2021); Notin et al. (2022); Riesselman (2018); Notin (2022)). The two traditional tree reconstruction methods include MAFFT + FastTree and Clustal + FastTree (Katoh et al. (2002); Price et al. (2010); Sievers et al. (2011)).

Evaluation setup. We consider two evaluation settings. **Tree reconstruction**: This setting evaluates the model's ability to reconstruct phylogenetic trees given solely the original sequences. We evaluate tree reconstruction by comparing the predicted tree to the reference tree using the Robinson-Foulds metric (Robinson & Foulds (1981)). **Functional prediction**: This setting evaluates the model's ability to predict functional labels given solely the original sequences. We assess functional prediction by training a linear probe classifier on the generated embeddings. We also consider a case study of **reconstructing the tree of life using ribosomal protein sequences** to demonstrate a potential biological use case for PHYLA.

4.1 PHYLA CAN REASON OVER PROTEIN SEQUENCES

Experimental setup. To assess the ability of PHYLA to reason over sequences, we assess PHYLA's ability to reconstruct phylogenetic trees on the "Openfold Small", "Openfold Medium", and "Openfold Large" datasets. We use the metric of Robinson-Foulds distance, or "RF", whereby a larger RF value is equivalent to a larger distance between predicted and reference tree, and can be interpreted as a lower quality predicted tree. The RF metric is not invariant to tree size, so we compute the normalized RF, or "normRF", to directly compare the tree reconstruction performance between trees of different sizes. We utilize the ETE3 Toolkit implementation of RF and normRF distance (Jaime Huerta-Cepas & Bork (2016)). We compare the performance of PHYLA against state-of-the-art PLMs (ESM2, ESM3) and genomic foundation models (Evo) (Lin et al. (2022); Hayes et al. (2024); Nguyen et al. (2024)). Table 1 shows the normRF performance of PHYLA and benchmark models on the three stratifications of the OpenProteinSet.

Results. PHYLA achieves the best performance on the Openfold Small evaluation set, beating benchmark models with 2 to 24 times more parameters (Table 1). Although PHYLA does not outperform the benchmarks in the Openfold Medium and Openfold Large evaluation sets, these results suggest a trend in all models worsening their performance as tree size increases.

Table 1: **Tree reconstruction performance.** Average normRF metric values (lower value indicates better performance) across all datasets within the Openfold Small (0 to 1,000 sequences), Medium (1,000 to 2,000 sequences), and Large (2,000+ sequences) evaluation sets for PHYLA vs. ESM models vs. Evo.

Model	**Openfold Small** normRF ↓	**Openfold Medium** normRF↓	**Openfold Large** normRF ↓
ESM2 (650M)	0.8735	0.9084	0.9292
ESM2 (3B)	0.8391	**0.8609**	**0.8859**
ESM3 (1.4B)	0.9070	0.9297	0.9387
Evo (7B)	0.9877	0.9949	0.9963
PHYLA (291M)	***0.8187***	*0.8980*	*0.9357*

4.2 PHYLA TREES ENCODE PROTEIN FUNCTIONAL INFORMATION

Experimental setup. To evaluate the expressivity of the learned embeddings from PHYLA, we train a linear probe on predicting functional labels from the embeddings. We utilize the 83 datasets from the ProteinGym (Notin et al. (2023a)) benchmark as our evaluation set. Table 2 shows the average Spearman correlation metric for linear probe performance on the 83 ProteinGym evaluation datasets.

Results. PHYLA ranks among the top 4 models out of 15 evaluated (Table 2) on the Linear Probe metric, despite having significantly fewer parameters and being trained on a smaller dataset. In contrast, Evo performs the worst among all models on this metric, which aligns with expectations given that Evo was trained on prokaryotic genomes, whereas ProteinGym comprises human protein sequences (Nguyen et al., 2024).

Table 2: **Functional prediction performance.** Average Spearman correlation coefficient values (higher values indicate better performance) averaged across 83 datasets within the ProteinGym evaluation sets using Linear Probe calculation for PHYLA vs. ESM models vs. Evo vs. ProteinGym benchmarks, *: pulled from pre-computed ProteinGym benchmark. Note that Evo model is trained on millions of microbial genomes and thus is not expected to generalize well to human protein sequences in ProteinGym.

Model	**ProteinGym** Spearman correlation ↑
ESM2 (650M)	**0.7754**
ProteinNPT*	0.7081
ESM2 (3B)	0.7044
PHYLA (291M)	*0.6962*
MSA Transformer Embeddings*	0.6944
ESM-1v Embeddings*	0.6482
Tranception Embeddings*	0.6239
TranceptEVE + One-Hot Encodings*	0.4839
MSA Transformer + One-Hot Encodings*	0.4738
Tranception + One-Hot Encodings*	0.4672
DeepSequence + One-Hot Encodings*	0.4591
ESM-1v + One-Hot Encodings*	0.4415
ESM3 (1.4B)	0.2743
One-Hot Encodings*	0.2725
Evo (7B)	-0.0044

4.3 RUNTIME COMPARISON

Experimental setup. To evaluate the efficiency of PHYLA compared to the benchmark models on embedding generation, we calculate the runtime required to generate sequence embeddings, generate a predicted distance matrix, and run the neighbor-joining algorithm to construct the predicted tree (scikit-bio development team, 2020). Table 3 shows the average runtime in seconds for tree reconstruction across the three stratifications of the OpenProteinSet and the ProteinGym evalution set for PHYLA and benchmark models. We also include the runtime for a larger Phyla model "Phyla (660M)" with 660M parameters for a more fair comparison to the benchmark models, the smallest of which had 650M parameters.

We also compared tree reconstruction of PHYLA with phylogenetic tree reconstruction methods. We calculate the runtime required to construct a multiple sequence alignment (MSA) from each dataset using the state-of-the-art aligners MAFFT and Clustal Omega, and then construct a tree from the MSA using a FastTree efficient tree construction method (Katoh et al., 2002; Price et al., 2010; Sievers et al., 2011). Table 4 shows the average runtime in seconds for tree reconstruction across the three stratifications of the OpenProteinSet for PHYLA and the benchmark methods.

Results. PHYLA generates embeddings much faster than the benchmark models across all three stratifications of the OpenProteinSet and the ProteinGym evaluation set (Table 3). The larger "Phyla (660M)" model generates embeddings faster than all benchmarks. In addition, PHYLA outperforms phylogenetic tree reconstruction methods across all stratifications of the OpenProteinSet, particularly on the Openfold Large evaluation set of trees larger than 2,000 sequences (Table 4). We see that the runtime for tree reconstruction increases as a function of tree size, but PHYLA's absolute runtime is still significantly faster than the other methods at all scales.

Table 3: **Runtime analyses.** Average runtime in seconds (lower is better) averaged across all datasets within the Openfold Small (0 to 1,000 sequences), Medium (1,000 to 2,000 sequences), Large (2,000+ sequences), and ProteinGym evaluation sets for Phyla vs. ESM models vs. Evo.

Model	Openfold Small Seconds ↓	Openfold Medium Seconds ↓	Openfold Large Seconds ↓	ProteinGym Seconds ↓
ESM2 (650M)	12.65	58.17	280.15	78.49
ESM2 (3B)	37.96	138.22	509.97	93.14
ESM3 (1.4B)	18.38	70.18	251.55	110.73
Evo (7B)	37.02	126.47	477.72	78.38
PHYLA (291M)	***2.08***	***23.38***	***179.00***	***69.74***
PHYLA (660M)	***3.22***	***26.59***	***216.17***	***70.52***

Table 4: **Runtime analyses.** Average runtime in seconds (lower is better) averaged across all datasets within the Openfold Small (0 to 1,000 sequences), Medium (1,000 to 2,000 sequences), Large (2,000+ sequences), and ProteinGym evaluation sets for PHYLA vs. traditional benchmarks that involve performing multiple-sequence alignment followed by running a phylogenetic tree construction algorithm on the aligned sequences.

Model	Openfold Small Seconds ↓	Openfold Medium Seconds ↓	Openfold Large Seconds ↓
MAFFT + FastTree	38.20	190.02	695.67
Clustal + FastTree	26.71	354.53	1594.88
PHYLA (291M)	***2.08***	***23.38***	***179.00***

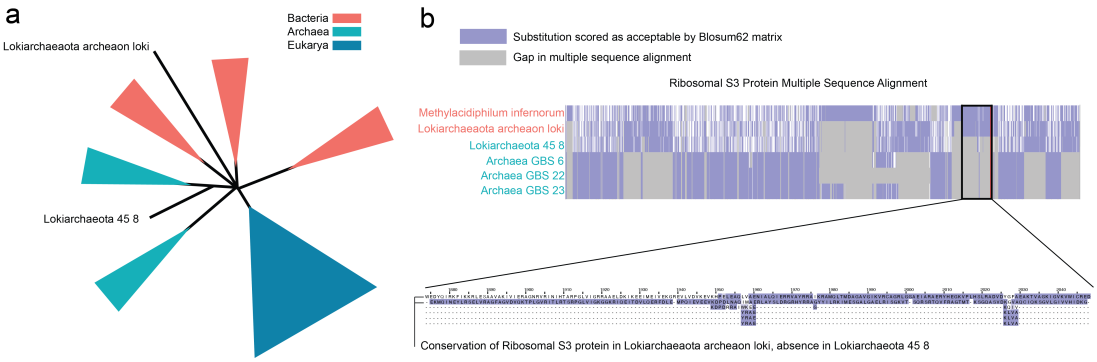

Figure 2: **PHYLA created a new placement of Lokiarchaea.** a. Lokiarchaeaota archeaon loki (L-A) was placed among bacterial neighbors while Lokiarchaeota 45 8 (L-45) was placed among archaeal neighbors. b. Analysis of the multiple sequence alignment revealed that L-A placed with bacteria retained a conserved S3 ribosomal protein, aligning with its bacterial neighbors. In contrast, the L-45 placed with archaea exhibited a deletion of the S3 ribosomal protein, aligning with its archaeal neighbors.

4.4 ABLATION ANALYSES

Experimental setup. To understand the effect of the sequence reasoning loss, we trained PHYLA with only masked language modeling loss (PHYLA-MLM). We evaluated (PHYLA-MLM) on tree reconstruction using the three stratifications of the OpenProteinSet and also on functional prediction using the Linear Probe metric.

Results. As shown in Table 5, we found PHYLA-MLM consistently performed worse than PHYLA on tree reconstruction and functional prediction.

Table 5: **Tree loss ablation performance.** Average normRF metric values (lower is better) across all datasets within the Openfold Small (0 to 1,000 sequences), Medium (1,000 to 2,000 sequences), and Large (2,000+ sequences) evaluation sets and average Spearman rank correlation (higher is better) averaged across 83 datasets within the ProteinGym evaluation set for PHYLA vs. PHYLA-MLM.

Model	Openfold Small normRF ↓	Openfold Medium normRF ↓	Openfold Large normRF ↓	ProteinGym Spearman ↑
PHYLA-MLM	0.9306	0.9663	0.9711	0.6174
PHYLA	***0.8187***	***0.8980***	***0.9357***	**0.6962**

4.5 USING PHYLA TO CONSTRUCT A PHYLOGENETIC TREE ACROSS 3,084 ORGANISMS

PHYLA demonstrates promising performance in sequence reasoning. To showcase its capabilities, we applied PHYLA to the task of phylogenetic tree construction. The tree of life is a fundamental framework in biology, delineating evolutionary relationships between organisms and serving as an indicator of relative phenotypic traits. Current approaches to constructing the tree of life typically rely on multiple sequence alignments of ribosomal proteins (Hug et al., 2016). We used PHYLA to analyze a set of 3,084 phylogenetic sequences, successfully reconstructing the tree of life in just 16 hours, compared to the 3,840 hours required by traditional methods (Hug et al., 2016).

As shown in Figure 1, PHYLA accurately places sequences within their respective domains in the tree of life. PHYLA identifies overlap between certain archaeal isolates and bacteria, a result consistent with current phylogenetic reasoning. Lokiarchaeota, an archaeal lineage clustered with bacteria, is known to have a mosaic genome with over 30% of its genome derived from bacteria (Levasseur et al., 2017). Within this genus, Phyla placed Lokiarchaeaota archaeon loki (L-A) paraphyletic to bacteria while Lokiarchaeota 45 8 (L-45) is paraphyletic to archaea (Figure 2a). Examination

of the multiple sequence alignment of L-45 and L-A with their immediate phylgenetic neighbors, revealed that L-45 harbors a deletion of the S3 ribosomal protein while L-A retains this protein (Figure 2b). The S3 deletion has been noted in previous studies of Lokiarchaea genomes Da Cunha et al. (2017). Biologically, these differences may relate to adaptation to extreme environments. L-45 was isolated from the bottom of the Arctic Ocean, while L-A was isolated from the Horonobe Underground Research Laboratory (URL) in Japan. In fact, L-A's neighbor, Methylacidiphilum infernorum, is an acidophilic methanotroph originally isolated from a geothermal area in New Zealand Hou et al. (2008). This environment shares similarities with the conditions in the URL, where extensive methane metabolism has been observed Amano et al. (2024). This highlights PHYLA's ability to discover potentially biologically meaningful evolutionary relationships.

5 CONCLUSION

Molecular sequence reasoning presents unique challenges, requiring models to represent individual sequences while reasoning across multiple sequences at varying levels of abstraction. Here, we report an ongoing development of PHYLA, a hybrid state-space and transformer model that operates at two levels of abstraction: sequence tokens and phylogenetic trees. By defining phylogenetic trees as higher-order abstractions that encapsulate evolutionary and functional relationships among sequences, PHYLA can overcome limitations of current protein language models, which are primarily sequence-centric and token-based. This hierarchical architecture enables structured, multi-level reasoning and sets a new benchmark for molecular sequence modeling and phylogenetic inference.

Preliminary results show that PHYLA achieves competitive or state-of-the-art performance in reconstructing phylogenetic trees, outperforming traditional multiple sequence alignment algorithms and existing machine learning approaches in runtime efficiency. Using PHYLA, we reconstructed the tree of life, revealing a phylogeny that aligns with established biological reasoning. These results motivate future development of PHYLA to establish a foundational model for molecular sequence reasoning and more efficient and insightful phylogenetic analysis.

ACKNOWLEDGMENTS

Y.E. is supported by grant T32 HG002295 from the National Human Genome Research Institute and the NSDEG fellowship. Y.E. and M.Z. gratefully acknowledge the support of NIH R01-HD108794, NSF CAREER 2339524, US DoD FA8702-15-D-0001, ARPA-H BDF program, awards from Chan Zuckerberg Initiative, Bill & Melinda Gates Foundation INV-079038, Amazon Faculty Research, Google Research Scholar Program, AstraZeneca Research, Roche Alliance with Distinguished Scientists, Sanofi iDEA-iTECH, Pfizer Research, John and Virginia Kaneb Fellowship at Harvard Medical School, Biswas Computational Biology Initiative in partnership with the Milken Institute, Harvard Medical School Dean's Innovation Fund for the Use of Artificial Intelligence, Harvard Data Science Initiative, and Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

COMPETING INTERESTS

The authors declare no competing interests.

CODE AVAILABILITY STATEMENT

Code to run Phyla can be found in the project github.

DATA AVAILABILITY STATEMENT

Data to run Phyla can be found in the project github specifically in the "dataset_info" folder.

REFERENCES

Gustaf Ahdritz, Nazim Bouatta, Sachin Kadyan, Lukas Jarosch, Daniel Berenberg, Ian Fisk, Andrew M. Watkins, Stephen Ra, Richard Bonneau, and Mohammed AlQuraishi. Openproteinset:

Training data for structural biology at scale, 2023. URL https://arxiv.org/abs/2308.05326.

Gustaf Ahdritz, Nazim Bouatta, Sachin Kadyan, Lukas Jarosch, Dan Berenberg, Ian Fisk, Andrew Watkins, Stephen Ra, Richard Bonneau, and Mohammed AlQuraishi. Openproteinset: Training data for structural biology at scale. *Advances in Neural Information Processing Systems*, 36, 2024.

Biswas S AlQuraishi M Church GM Alley EC, Khimulya G. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods*, 2019. doi: 10.1038/s41592-019-0598-1.

Yusuke Amano, Rishabh Sachdeva, Daniel Gittins, Karthik Anantharaman, Shuwei Lei, Laura E. Valentin-Alvarado, Samuel Diamond, Hiroshi Beppu, Teruki Iwatsuki, Ayako Mochizuki, Keiko Miyakawa, Eri Ishii, Hiroshi Murakami, Adam L. Jaffe, Cindy Castelle, Adi Lavy, Yohei Suzuki, and Jillian F. Banfield. Diverse microbiome functions, limited temporal variation and substantial genomic conservation within sedimentary and granite rock deep underground research laboratories. *Environmental Microbiome*, 19(1):105, Dec 2024. doi: 10.1186/s40793-024-00649-3.

Agarwal V. Visentin D. et al. Avsec, Ž. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 2021. doi: 10.1038/s41592-021-01252-x. URL https://www.nature.com/articles/s41592-021-01252-x.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory, 2024. URL https://arxiv.org/abs/2405.04517.

Goldman G. Wang C.H. et al. Brandes, N. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 2023. doi: 10.1038/s41588-023-01465-0. URL https://www.nature.com/articles/s41588-023-01465-0.

Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, and Cedric Notredame. Multiple sequence alignment modeling: methods and applications. *Briefings in Bioinformatics*, 17(6):1009–1023, 11 2015. ISSN 1467-5463. doi: 10.1093/bib/bbv099. URL https://doi.org/10.1093/bib/bbv099.

Violette Da Cunha, Morgan Gaia, Daniele Gadelle, Arshan Nasir, and Patrick Forterre. Lokiarchaea are close relatives of euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.*, 13(6):e1006810, June 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Frances Ding and Jacob Steinhardt. Protein language models are biased by unequal sequence sampling across the tree of life. *bioRxiv*, 2024. doi: 10.1101/2024.03.07.584001. URL https://www.biorxiv.org/content/early/2024/03/12/2024.03.07.584001.

Yasha Ektefaie, Andrew Shen, Daria Bykova, Maximillian Marin, Marinka Zitnik, and Maha Farhat. Evaluating generalizability of artificial intelligence models for molecular datasets. *bioRxiv*, 2024. doi: 10.1101/2024.02.25.581982. URL https://www.biorxiv.org/content/early/2024/02/28/2024.02.25.581982.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing, 2021. URL https://arxiv.org/abs/2007.06225.

J. Cheng et al. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 2023. doi: 10.1126/science.adg7492. URL https://www.science.org/doi/10.1126/science.adg7492.

Schmidt S. Höcker B. Ferruz, N. Protgpt2 is a deep unsupervised language model for protein design. *Nature Communications*, 2022. doi: 10.1038/s41467-022-32007-7. URL https://www.nature.com/articles/s41467-022-32007-7.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces, 2022. URL https://arxiv.org/abs/2111.00396.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *bioRxiv*, 2024. doi: 10.1101/2024.07.01.600583. URL https://www.biorxiv.org/content/early/2024/07/02/2024.07.01.600583.

Shaobin Hou, Kira S Makarova, Jimmy H W Saw, Pavel Senin, Benjamin V Ly, Zhemin Zhou, Yan Ren, Jianmei Wang, Michael Y Galperin, Marina V Omelchenko, Yuri I Wolf, Natalya Yutin, Eugene V Koonin, Matthew B Stott, Bruce W Mountain, Michelle A Crowe, Angela V Smirnova, Peter F Dunfield, Lu Feng, Lei Wang, and Maqsudul Alam. Complete genome sequence of the extremely acidophilic methanotroph isolate v4, methylacidiphilum infernorum, a representative of the bacterial phylum verrucomicrobia. *Biol. Direct*, 3:26, July 2008.

Baker B. Anantharaman K. et al. Hug, L. A new view of the tree of life. *Nature Microbiology*, 2016. doi: 10.1038/nmicrobiol.2016.48. URL https://www.nature.com/articles/nmicrobiol201648.

Laura A Hug, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, Cindy J Castelle, Cristina N Butterfield, Alex W Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A Relman, Kari M Finstad, Ronald Amundson, Brian C Thomas, and Jillian F Banfield. A new view of the tree of life. *Nat. Microbiol.*, 1:16048, April 2016.

Francois Serra Jaime Huerta-Cepas and Peer Bork. Ete 3: Reconstruction, analysis and visualization of phylogenomic data. *Mol Biol Evol*, 2016. doi: 10.1093/molbev/msw046.

Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30 (14):3059–3066, 07 2002. ISSN 0305-1048. doi: 10.1093/nar/gkf436. URL https://doi.org/10.1093/nar/gkf436.

Anthony Levasseur, Vicky Merhej, Emeline Baptiste, Vikas Sharma, Pierre Pontarotti, and Didier Raoult. The rhizome of lokiarchaeota illustrates the mosaicity of archaeal genomes. *Genome Biol. Evol.*, 9(10):2635–2639, October 2017.

Francesca-Zhoufan Li, Ava P. Amini, Yisong Yue, Kevin K. Yang, and Alex X. Lu. Feature reuse and scaling: Understanding transfer learning with protein language models. *bioRxiv*, 2024. doi: 10.1101/2024.02.05.578959. URL https://www.biorxiv.org/content/early/2024/02/14/2024.02.05.578959.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Krause B. Greene E.R. et al. Madani, A. Large language models generate functional protein sequences across diverse families. *Nat Biotechnol 41, 1099–1106*, 2023. doi: 10.1038/s41587-022-01618-2.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021. doi: 10.1101/2021.07.09.450648. URL https://www.biorxiv.org/content/early/2021/07/10/2021.07.09.450648.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen Baccus. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 43177–43201. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/86ab6927ee4ae9bde4247793c46797c7-Paper-Conference.pdf.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Armin W. Thomas, Brian Kang, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aman Patel, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with evo. *bioRxiv*, 2024. doi: 10.1101/2024.02.27.582234. URL https://www.biorxiv.org/content/early/2024/03/06/2024.02.27.582234.

Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, 11 2014. ISSN 0737-4038. doi: 10.1093/molbev/msu300. URL https://doi.org/10.1093/molbev/msu300.

Dias M. Frazer J. Marchena-Hurtado J. Gomez A. Marks D.S. Gal Y. Notin, P. Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval. *ICML*, 2022.

Pascal Notin, Lood Van Niekerk, Aaron W Kollasch, Daniel Ritter, Yarin Gal, and Debora S. Marks. Trancepteve: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. *bioRxiv*, 2022. doi: 10.1101/2022.12.07.519495. URL https://www.biorxiv.org/content/early/2022/12/10/2022.12.07.519495.

Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Rose Orenbuch, Yarin Gal, and Debora S. Marks. Proteingym: Large-scale benchmarks for protein design and fitness prediction. *bioRxiv*, 2023a. doi: 10.1101/2023.12.07.570727. URL https://www.biorxiv.org/content/early/2023/12/08/2023.12.07.570727.

Pascal Notin, Ruben Weitzman, Debora S. Marks, and Yarin Gal. Proteinnpt: Improving protein property prediction and design with non-parametric transformers. *bioRxiv*, 2023b. doi: 10.1101/2023.12.06.570473. URL https://www.biorxiv.org/content/early/2023/12/07/2023.12.06.570473.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Re. Hyena hierarchy: Towards larger convolutional language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28043–28078. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/poli23a.html.

Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. Fasttree 2 – approximately maximum-likelihood trees for large alignments. *PLOS ONE*, 5(3):1–10, 03 2010. doi: 10.1371/journal.pone.0009490. URL https://doi.org/10.1371/journal.pone.0009490.

Owen Queen, Yepeng Huang, Robert Calef, Valentina Giunchiglia, Tianlong Chen, George Dasoulas, LeAnn Tai, Yasha Ektefaie, Ayush Noori, Joseph Brown, et al. Procyon: A multimodal foundation model for protein phenotypes. *bioRxiv*, pp. 2024–12, 2024.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *bioRxiv*, 2020. doi: 10.1101/2020.12.15.422761. URL https://www.biorxiv.org/content/early/2020/12/15/2020.12.15.422761.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8844–8856. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/rao21a.html.

Ingraham J.B. Marks D.S. Riesselman, A.J. Deep generative models of genetic variation capture the effects of mutations. *Nature Method*, 2018. doi: 10.1038/s41592-018-0138-4. URL https://www.nature.com/articles/s41592-018-0138-4.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.

D.F. Robinson and L.R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147, 1981. ISSN 0025-5564. doi: https://doi.org/10.1016/0025-5564(81)90043-2. URL https://www.sciencedirect.com/science/article/pii/0025556481900432.

The scikit-bio development team. scikit-bio: A bioinformatics library for data scientists, students, and developers, 2020. URL http://scikit-bio.org.

Damiano Sgarbossa, Cyril Malbranke, and Anne-Florence Bitbol. Protmamba: a homology-aware but alignment-free protein state space model. *bioRxiv*, 2024. doi: 10.1101/2024.05.24.595730. URL https://www.biorxiv.org/content/early/2024/05/25/2024.05.24.595730.

Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D Thompson, and Desmond G Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 7(1):539, 2011. doi: https://doi.org/10.1038/msb.2011.75. URL https://www.embopress.org/doi/abs/10.1038/msb.2011.75.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Maxim Zvyagin, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, Carla M. Mann, Michael Irvin, J. Gregory Pauloski, Logan Ward, Valerie Hayot, Murali Emani, Sam Foreman, Zhen Xie, Diangen Lin, Maulik Shukla, Weili Nie, Josh Romero, Christian Dallago, Arash Vahdat, Chaowei Xiao, Thomas Gibbs, Ian Foster, James J. Davis, Michael E. Papka, Thomas Brettin, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath, and Arvind Ramanathan. Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics. *bioRxiv*, 2022. doi: 10.1101/2022.10.10.511571. URL https://www.biorxiv.org/content/early/2022/10/11/2022.10.10.511571.

A APPENDIX

A.1 ADAPTIVE BATCH SIZE ALGORITHM

We empirically determined that our model could process inputs of length 213,350 on a 32 GB GPU and 302,350 on a 48 GB GPU. For untested GPU memory sizes, we used a linear model to estimate the maximum input length. Given the length of the longest sequence in a phylogenetic tree, we calculated the largest tree that can fit within the available GPU memory. To mitigate overfitting, we randomly sampled a tree size between 5 and the maximum permissible tree size. From the sampled tree size, we determined the number of trees we could sample.

If an out-of-memory (OOM) error occurred during training, the model resampled with both the tree size and the number of trees halved.

A.2 PHYLOGENETIC TREE RECONSTRUCTION

See Figure S1 and Figure S2.

A.3 FUNCTIONAL PREDICTION

See Figure S3 and Figure S4.

A.4 RUNTIME

See Figure S5 and Figure S6.

A.5 MODEL SIZE

PHYLA has 291M model parameters, which is less than 650M parameters of ESM2, 1.4B parameters of ESM3, 3B parameters or large ESM2, and 7B parameters of Evo model. We also train on 4 GPUs for 6 days, which is less than the 512 GPUs that ESM2 trains on for 8 days. We also train on a much smaller training set of $\sim$13,000 sequences compared to the $\sim$50,000,000 sequences that ESM2 trains on.

A.6 ABLATIONS

Tree loss ablation results in worse performance on tree reconstruction and functional prediction (for linear probe performance). See Figure S7 and Figure S8.

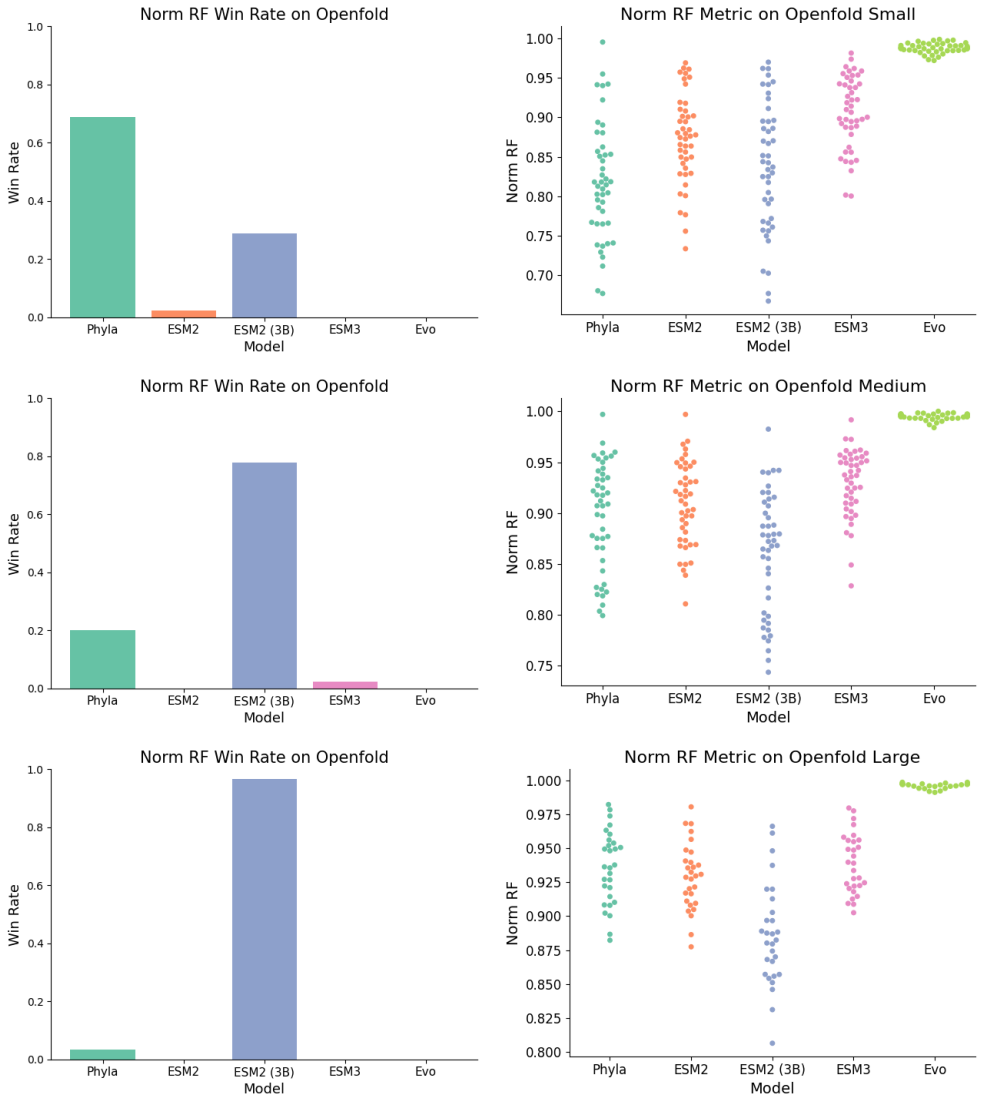

Figure S1: (a) Win rate (higher value is better) of normRF metric values for the Openfold Small (0 to 1000 sequences) evaluation set for Phyla vs ESM models vs Evo. (b) Swarm plot of normRF metric (lower is better) values for all datasets in the Openfold Small (0 to 1000 sequences) evaluation set for Phyla vs ESM models vs Evo. (c) Win rate (higher value is better) of normRF metric values for the Openfold Medium (1000 to 2000 sequences) evaluation set for Phyla vs ESM models vs Evo. (d) Swarm plot of normRF metric (lower is better) values for all datasets in the Openfold Medium (1000 to 2000 sequences) evaluation set for Phyla vs ESM models vs Evo. (e) Win rate (higher value is better) of normRF metric values for the Openfold Large (2000+ sequences) evaluation set for Phyla vs ESM models vs Evo. (f) Swarm plot of normRF metric (lower is better) values for all datasets in the Openfold Large (2000+ sequences) evaluation set for Phyla vs ESM models vs Evo.

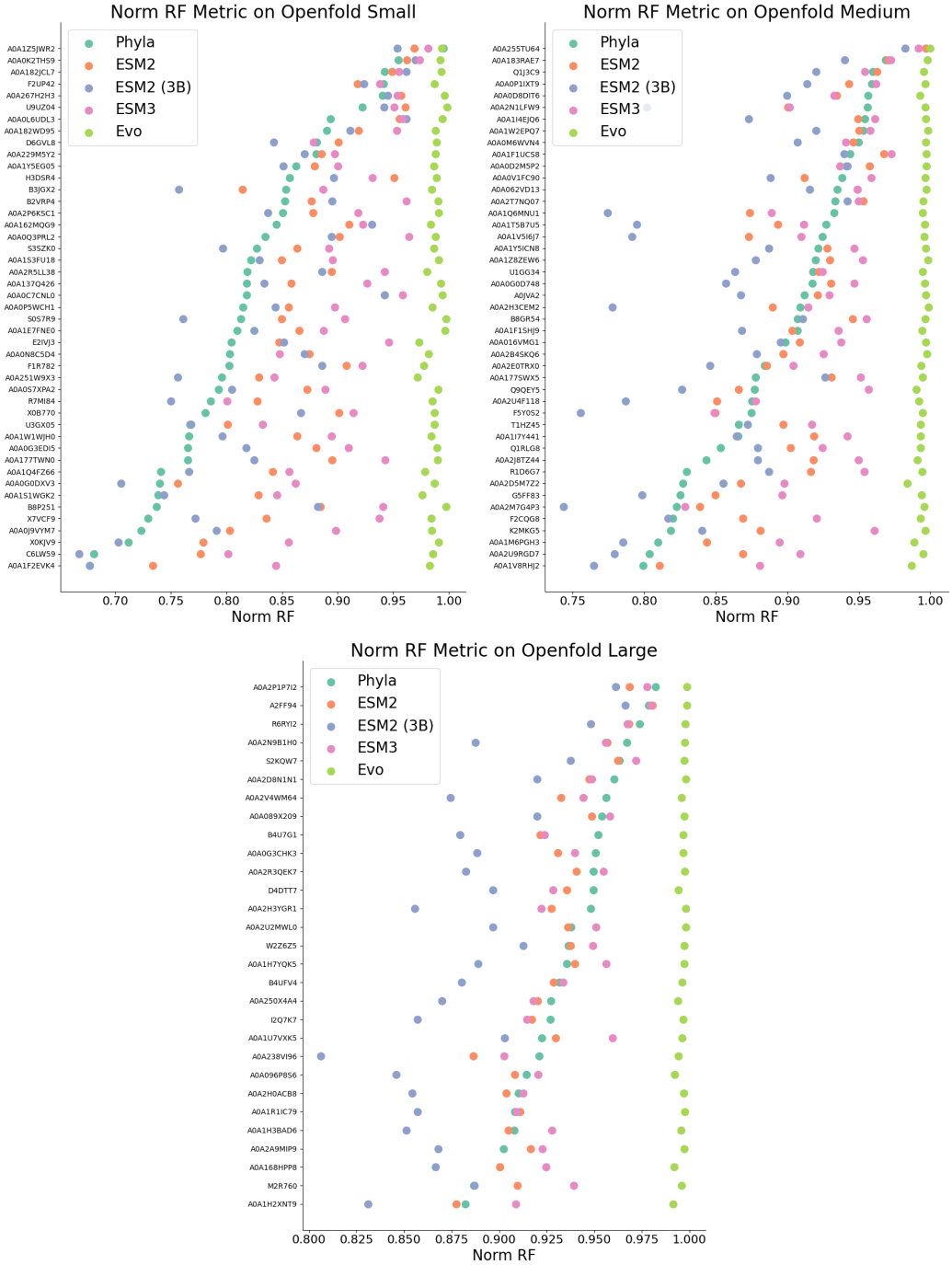

Figure S2: (a) Norm RF metric (lower, i.e. to the left, is better) values for all datasets in the Openfold Small (0 to 1000 sequences) evaluation set for Phyla vs ESM models vs Evo. (b) Norm RF metric (lower, i.e. to the left, is better) values for all datasets in the Openfold Medium (1000 to 2000 sequences) evaluation set for Phyla vs ESM models vs Evo. (c) Norm RF metric (lower, i.e. to the left, is better) values for all datasets in the Openfold Large (2000+ sequences) evaluation set for Phyla vs ESM models vs Evo.

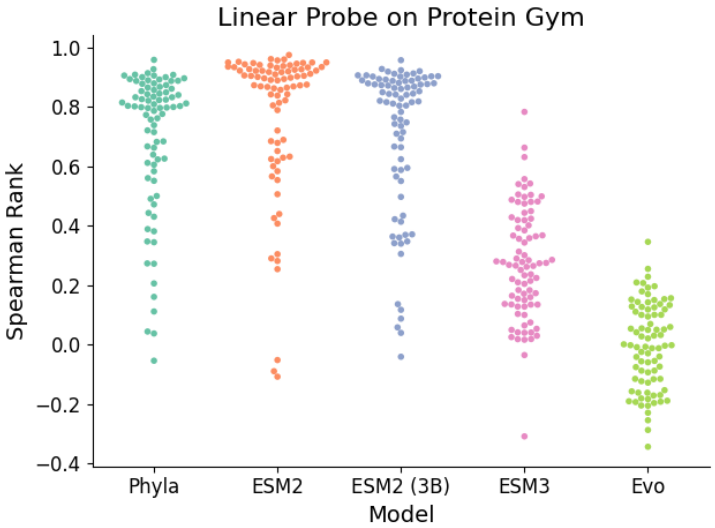

Figure S3: (a) Swarm plot of Spearman correlation (higher is better) values using linear probe calculation method for 83 datasets in the Protein Gym evaluation set for Phyla vs ESM models vs Evo.

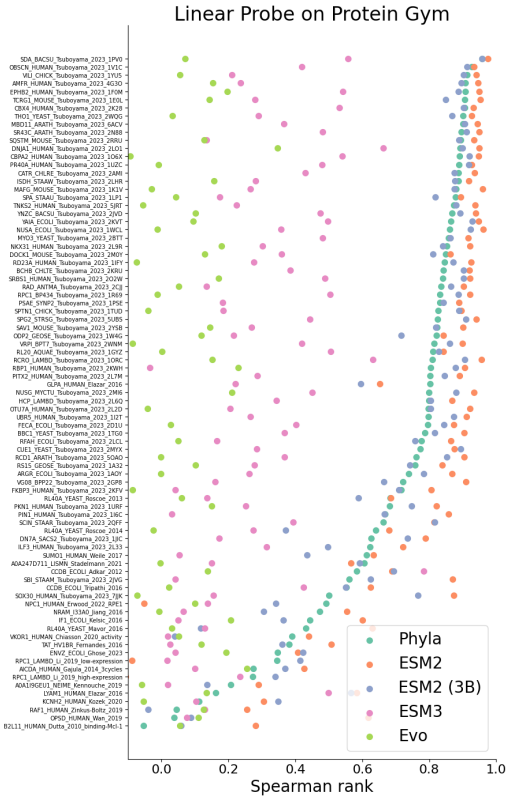

Figure S4: (a) Spearman correlation metric (higher, i.e. to the right, is better) values using linear probe calculation method for 83 datasets in the Protein Gym evaluation set for Phyla vs ESM models vs Evo.

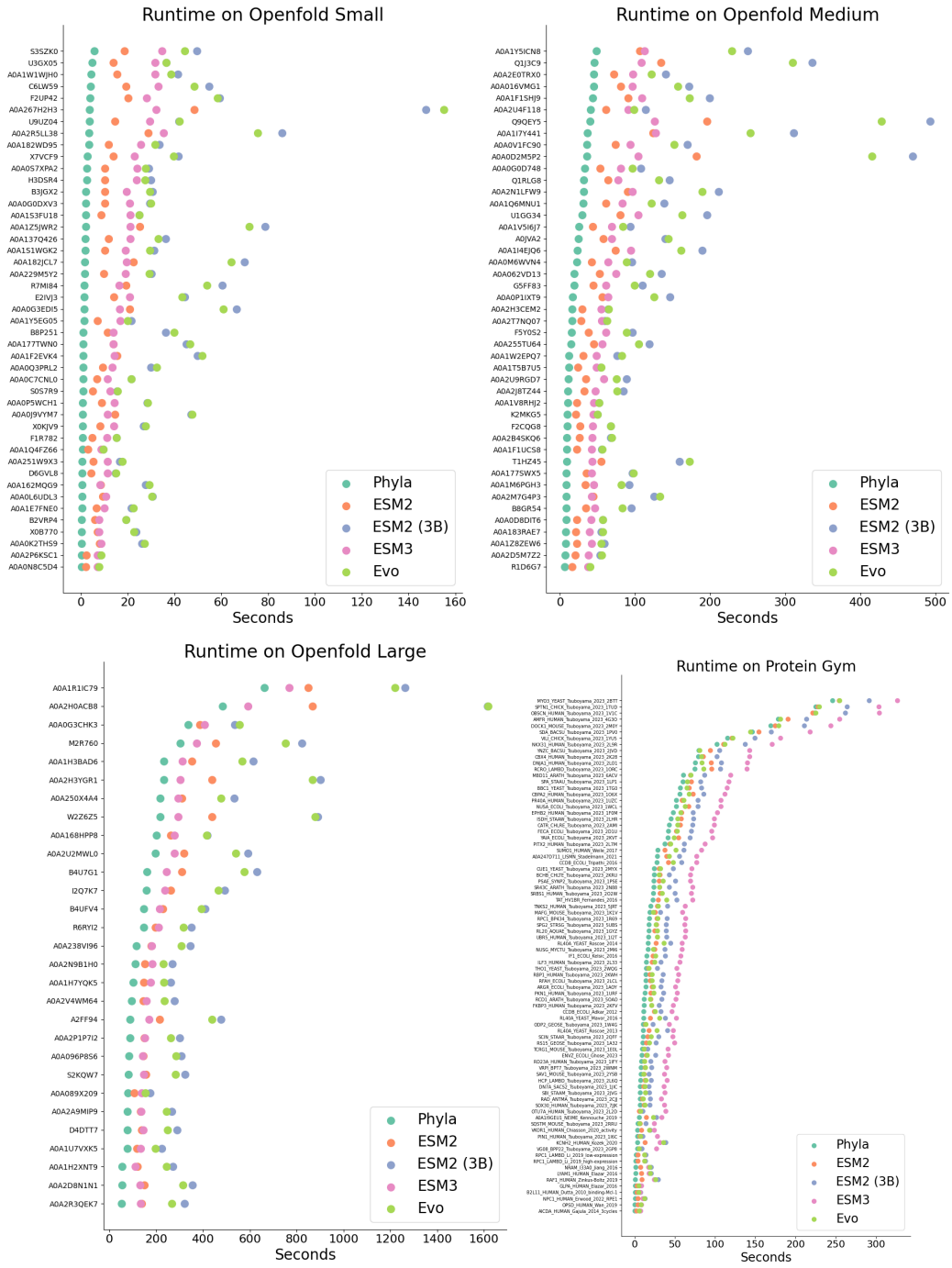

Figure S5: (a) Runtime metric (lower, i.e. to the left, is better) in seconds for all datasets in the Openfold Small (0 to 1000 sequences) evaluation set for Phyla vs ESM models vs Evo. (b) Runtime metric (lower, i.e. to the left, is better) in seconds for all datasets in the Openfold Medium (1000 to 2000 sequences) evaluation set for Phyla vs ESM models vs Evo. (c) Runtime metric (lower, i.e. to the left, is better) in seconds for all datasets in the Openfold Large (2000+ sequences) evaluation set for Phyla vs traditional method (in green) vs ESM models. (d) Runtime metric (lower, i.e. to the left, is better) in seconds for 83 datasets in the Protein Gym evaluation set for Phyla vs ESM models vs Evo.

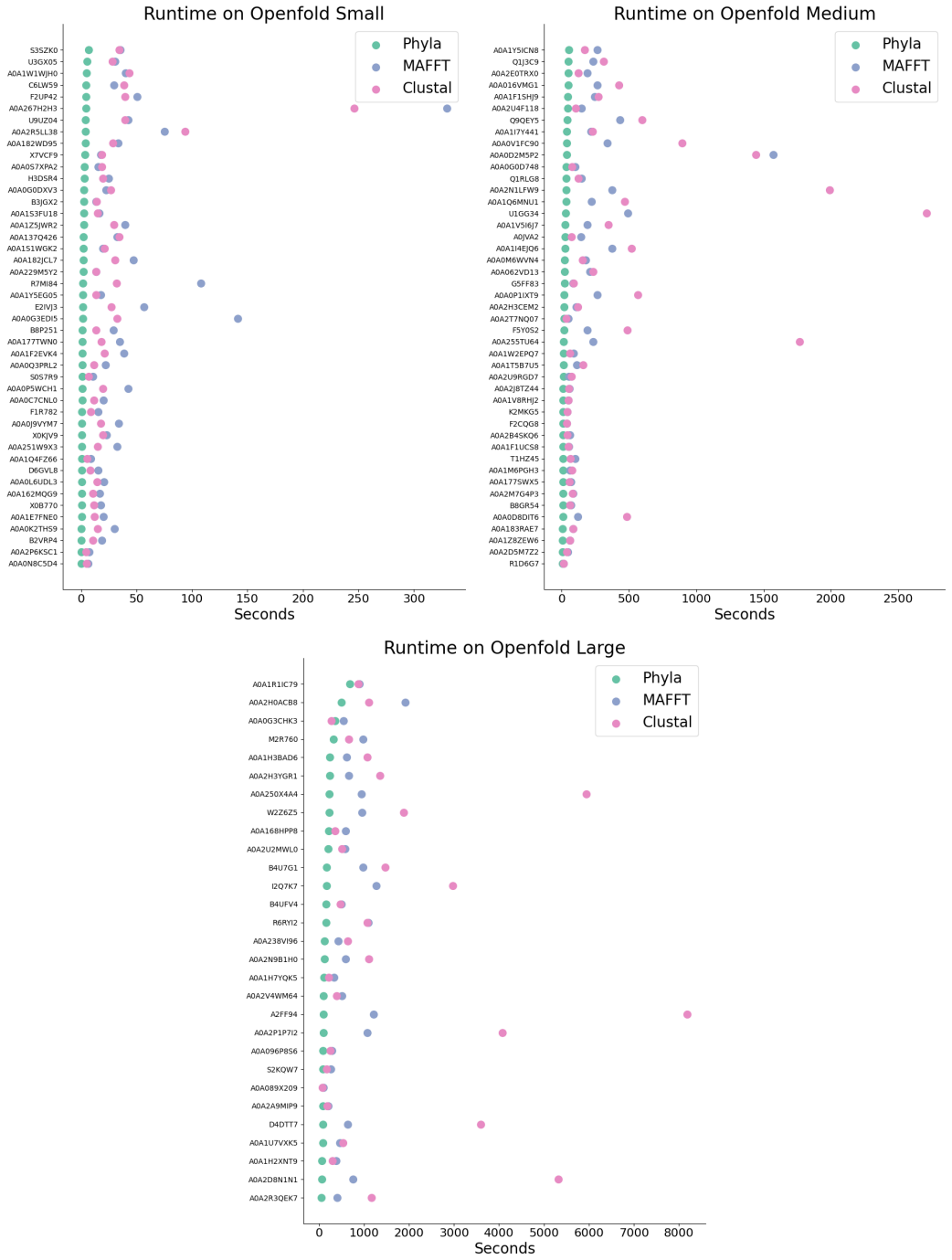

Figure S6: (a) Runtime metric (lower, i.e. to the left, is better) in seconds for all datasets in the Openfold Small (0 to 1000 sequences) evaluation set for Phyla vs traditional methods. (b) Runtime metric (lower, i.e. to the left, is better) in seconds for all datasets in the Openfold Medium (1000 to 2000 sequences) evaluation set for Phyla vs traditional methods. (c) Runtime metric (lower, i.e. to the left, is better) in seconds for all datasets in the Openfold Large (2000+ sequences) evaluation set for for Phyla vs traditional methods. *Note*: "traditional method" consists of taking sequences, aligning those sequences to create an MSA, and then using that MSA to construct a tree.

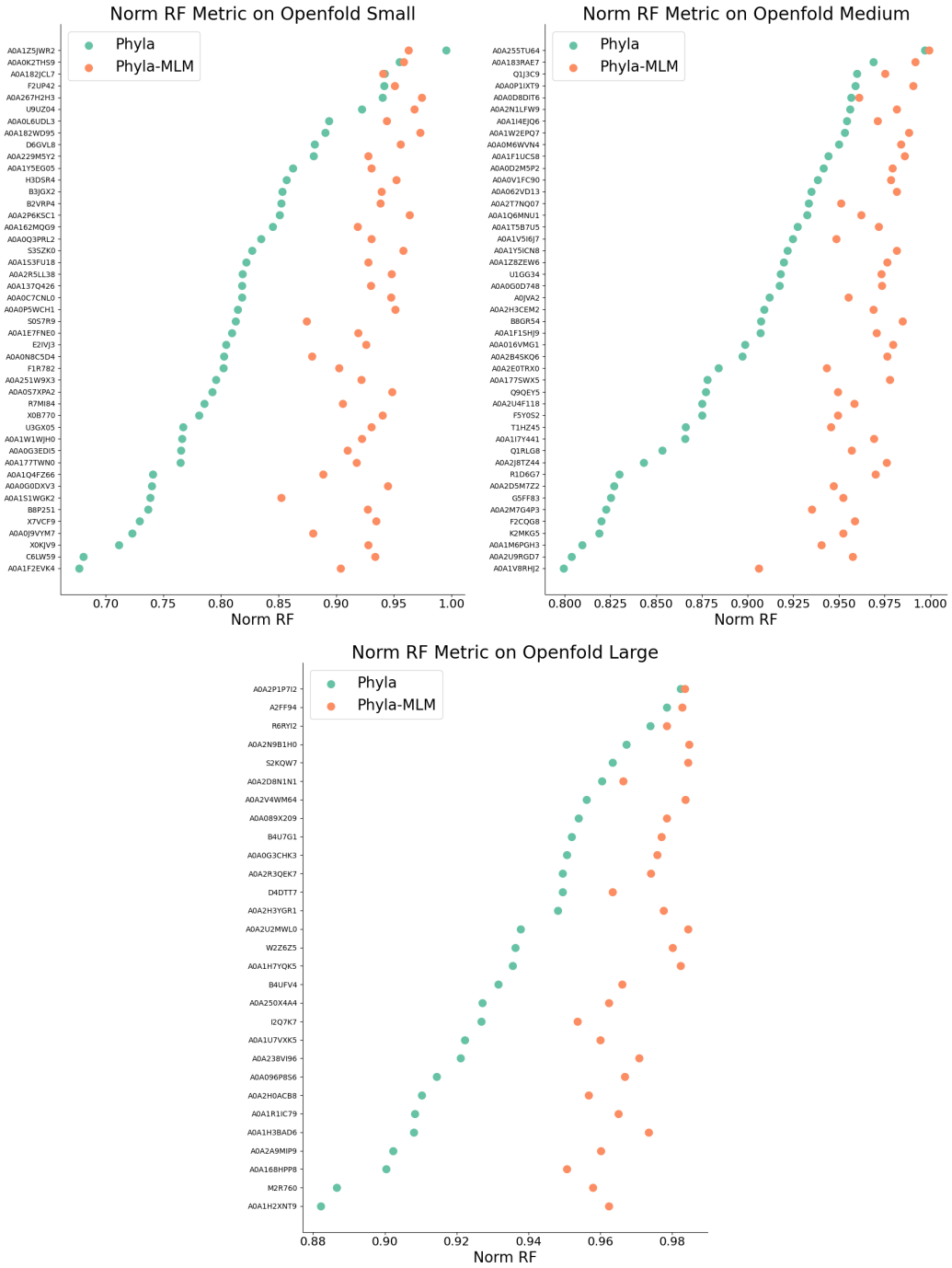

Figure S7: (a) Norm RF metric (lower, i.e. to the left, is better) values for all datasets in the Openfold Small (0 to 1000 sequences) evaluation set for Phyla vs Phyla-MLM. (b) Norm RF metric (lower, i.e. to the left, is better) values for all datasets in the Openfold Medium (1000 to 2000 sequences) evaluation set for Phyla vs Phyla-MLM. (c) Norm RF metric (lower, i.e. to the left, is better) values for all datasets in the Openfold Large (2000+ sequences) evaluation set for Phyla vs Phyla-MLM.

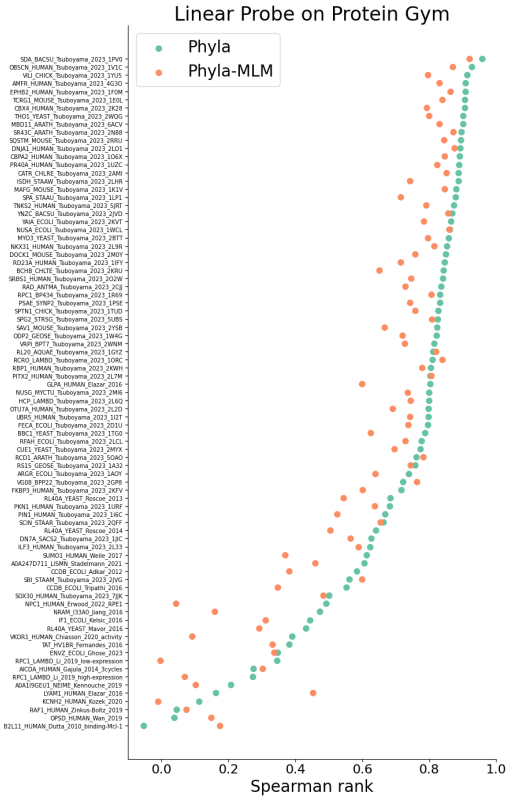

Figure S8: (a) Spearman correlation metric (higher, i.e. to the right, is better) using linear probe calculation method values for 83 datasets in the Protein Gym evaluation set for Phyla vs Phyla-MLM.