

# FIREFLY: PHYlogeny-informed REpresentation learning to estimate PHyLogenetic dIstances

Meijun Gao<sup>1</sup>, Byungho Lee<sup>1,2</sup>, and Kevin J. Liu<sup>1,2,3</sup>

<sup>1</sup> Department of Computer Science and Engineering

<sup>2</sup> Ecology, Evolution, and Behavior Program

<sup>3</sup> Genetics and Genome Sciences Program

Michigan State University

East Lansing, MI, USA

`kjl@msu.edu`

**Abstract.** Phylogenetic distance estimation and distance-based phylogeny reconstruction are well-studied cornerstone topics in phylogenetics. Classical approaches for both utilize mathematical or probabilistic graphical models of biomolecular sequence evolution. But model violations can occur and model-based analysis can be impacted as a result.

Recent advances in statistical machine learning using deep neural networks provide an alternative in the form of representation learning. Newer applications of deep learning to phylogenetic distance estimation have followed. A number of challenges in this area remain, since state-of-the-art methods are often restricted to pairwise or subset-based analyses and retain other simplifying assumptions.

In fact, classical model-based methods and representation learning are orthogonal and, as we show, their combination can be greater than the sum of its parts. We bridge these different approaches by synthesizing mathematical and logical constraints with statistical machine learning – an approach from physics-informed machine learning (PIML). Our algorithmic solution takes the form of a Transformer-based framework for learning phylogeny-informed representations directly from MSAs, which we apply to the task of phylogenetic distance estimation. The result is FIREFLY, a computational framework for “PHYlogeny-informed REpresentation learning to estimate PHyLogenetic dIstances”).

We benchmarked FIREFLY’s performance against other state-of-the-art methods using simulated and empirical datasets. We found that FIREFLY improves both pairwise distance estimation accuracy and downstream phylogenetic inference compared with state-of-the-art methods. The gains are particularly pronounced under high indel rates and on estimated MSAs, where alignment errors and gap-induced uncertainty are most severe. Our results highlight the value of integrating phylogeny-based inductive bias into deep representation learning and suggest that MSA-level modeling offers a robust foundation for evolutionary inference under challenging conditions.

**Keywords:** Phylogenetic distance estimation, Multiple sequence alignment, Phylogeny-informed learning, PIML, Deep representation learning

## 1 Introduction

Phylogenetic inference aims to reconstruct evolutionary relationships among biological sequences and plays a central role in comparative genomics and molecular evolution. In a standard phylogenetic pipeline, homologous sequences are first assembled into a multiple sequence alignment (MSA), after which a phylogenetic tree is inferred under an explicit event-based model of sequence evolution using statistical or mathematical optimization [3, 23, 25]. These optimization-based methods are computationally intensive due to repeated optimization score calculation and the need to explore a tree space that grows super-exponentially with the number of taxa [3]. As a computationally efficient alternative, distance-based methods reconstruct phylogenies from estimated pairwise evolutionary distances [8, 11]. Despite their differences in formulation and computational cost, both optimization-based and distance-based approaches rely on simplifying assumptions about the evolutionary process, such as site independence and error-free input alignments. In practice, insertions and deletions are typically ignored or treated as missing data rather than explicitly modeled [24]. Consequently, phylogenetic inference can be sensitive to alignment uncertainty and upstream alignment errors, particularly when MSAs are noisy or approximately estimated [7, 22].

An alternative to these classical approaches has emerged thanks to recent advances in machine learning: an increasing number of studies have investigated the use of deep learning methods in phylogenetics, as reviewed in [14]. Several likelihood-free methods formulate topology inference as a classification problem over tree topologies [19, 27]. However, because the number of unrooted tree topologies grows rapidly with the number of taxa, these methods are restricted to quartets (trees with four leaves) and must be combined heuristically to infer larger trees. Benchmark studies have shown that, under challenging scenarios such as long branches or short alignments, these methods can underperform classical likelihood or distance-based approaches [26]. Other likelihood-free frameworks, including GAN-based methods, require retraining for each dataset and do not scale beyond a modest number of taxa [18]. Phyloformer [15] instead predicts pairwise evolutionary distances from MSAs using Transformer encoders and reconstructs trees via distance-based methods, enabling scalability to larger datasets. However, its reliance on deterministic pairwise signals makes it sensitive to noise in estimated MSAs, and its pairwise representations may discard shared evolutionary context captured at the MSA level [2, 4], particularly in gap-rich alignments.

To help address these limitations and explore new directions for applications of deep learning in phylogenetics, our work draws inspiration from recent advances in MSA-based representation learning for protein modeling, where deep networks are pretrained on large MSAs and transferred to downstream tasks [9, 16]. While such approaches primarily focus on relationships between sites, phylogenetic inference concerns relationships between sequences. In this study, we introduce a phylogeny-informed representation learning framework built on a DNA-MSA Transformer encoder augmented with a learnable phylo-

genetic attention bias. By operating on full MSAs, the model learns sequence-level representations that capture global relational structure across taxa, which cannot be recovered from pairwise-only features. We combine a pretrained transformer encoder with a supervised distance estimation module to infer pairwise evolutionary distances, which are subsequently used for downstream tree reconstruction. A key component of our framework is a learnable phylogenetic bias module that is integrated directly into the column-wise attention mechanism. Unlike fixed or hand-crafted phylogenetic priors, this bias is learned end-to-end from data and adapts naturally to alignment noise. It is activated during joint training with the distance estimation module, allowing the model to stabilize cross-sequence attention while remaining flexible. Through extensive simulation studies and empirical analyses, our framework consistently outperforms Phyloformer and achieves performance comparable to, or exceeding, state-of-the-art likelihood-based methods on both true and estimated MSAs. Importantly, it remains robust in challenging regimes with high gap content, where prior deep learning approaches and likelihood-based methods often degrade.

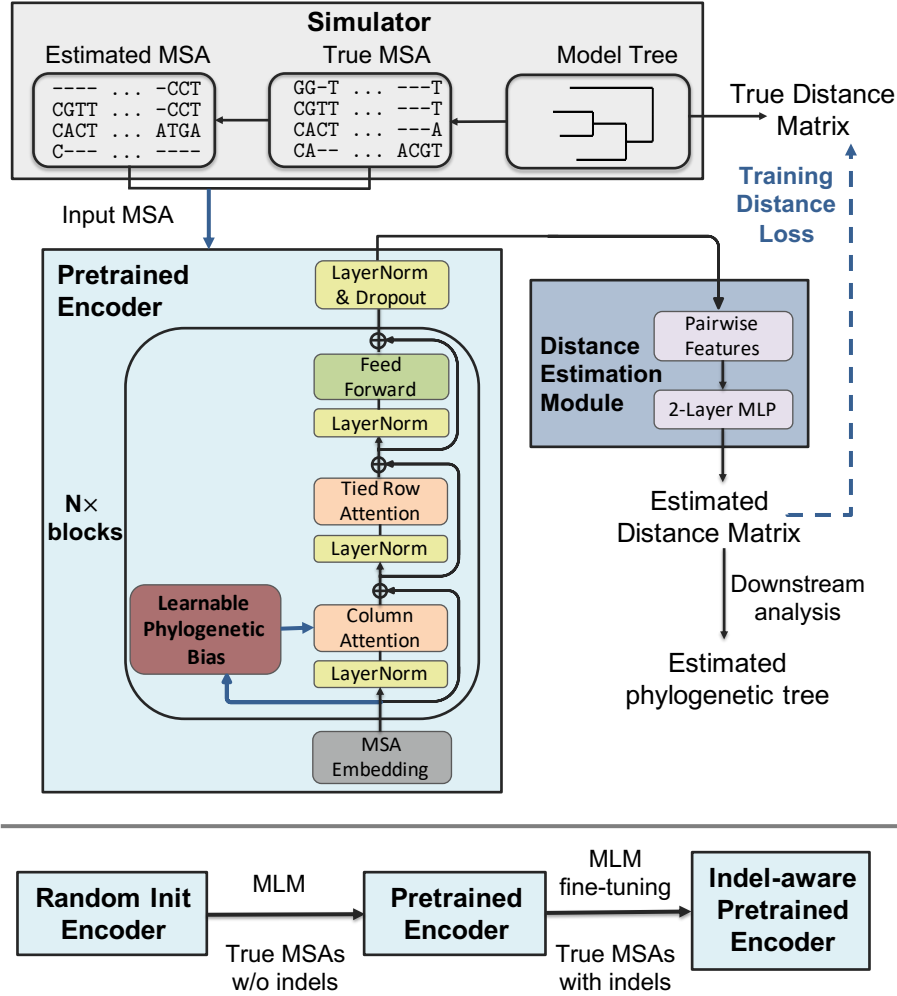
Our contributions can be summarized as follows:

- **MSA-level representation learning.** We introduce an MSA-level representation learning framework that encodes global evolutionary relationships among taxa, rather than relying on pairwise statistics derived from raw alignments.
- **Learnable phylogenetic inductive bias.** We propose a data-adaptive phylogenetic bias that guides attention toward sequences based on their evolutionary relationships, providing a flexible alternative to fixed evolutionary priors.
- **Robustness under realistic alignment noise.** By modeling full MSAs and learning adaptive phylogenetic structure, our approach substantially improved robustness in downstream phylogenetic inference under alignment errors and gap-induced uncertainty commonly encountered in estimated MSAs.

## 2 Method

*Problem formulation and framework overview.* Given a DNA multiple sequence alignment (MSA) with  $N$  taxa and alignment length  $L$ , our overall goal is to learn phylogeny-aware representations of MSAs that can support downstream phylogenetic inference tasks. In this work, we focus on pairwise evolutionary distance inference, followed by distance-based phylogenetic tree reconstruction.

Accurate phylogenetic representation learning requires modeling both site-wise dependencies along sequences and taxon-wise relationships across aligned sequences, while remaining robust to alignment noise. We therefore propose a Transformer-based framework that operates directly on full MSAs to learn sequence-level representations capturing global relational structure across taxa, and serves for downstream phylogenetic inferences. The framework consists of two stages. In the first stage, a shared Transformer encoder is pretrained using a



**Fig.1. Overview of the FIREFLY framework for downstream evolutionary pairwise distance inference and phylogenetic tree inference.** A simulator is used to generate phylogenetic trees together with true and estimated MSAs under specified evolutionary models. The framework consists of two stages. In the first stage, a Transformer encoder is pretrained using masked language modeling on true MSAs, first without indels and then fine-tuned on MSAs with indels to obtain an indel-aware encoder. The encoder architecture is based on the MSA Transformer [16], with a learnable phylogenetic bias module that is activated only during supervised distance learning. In the second stage, a distance estimation module is trained in a supervised manner using both true and estimated MSAs. Pairwise features are extracted from encoder hidden states and fed into a lightweight MLP to predict pairwise evolutionary distances, which are supervised using true distances derived from the model tree. The resulting pairwise evolutionary distances can serve as input for downstream phylogenetic tree reconstruction using distance-based inference methods.

masked language modeling (MLM) objective to learn general evolutionary patterns from MSAs. In the second stage, the pretrained encoder is jointly optimized with a supervised distance prediction module to infer a pairwise evolutionary distance  $D \in \mathbb{R}^{N \times N}$  from an input MSA, where each entry  $D_{ij}$  represents the evolutionary distance between taxa  $i$  and  $j$ . During joint supervised training of the encoder and the distance estimation module, a learnable phylogenetic bias module is activated to stabilize cross-sequence attention and guide representation learning under alignment noise. The inferred distance matrix is subsequently used for phylogenetic tree reconstruction via a distance-based method.

We first describe the Transformer encoder architecture and the masked language model pretraining used in the first stage. We then introduce the learnable phylogenetic bias module and the evolutionary distance prediction module, followed by a description of the overall training procedure. An overview of the proposed framework is shown in Figure 1.

*Transformer encoder architecture.* Our encoder follows the MSA Transformer architecture proposed in [16], as illustrated in Figure 1. It consists of six stacked attention blocks, each equipped with eight attention heads. Each block adopts an axial attention design comprising tied row attention, column attention, and a feedforward layer with residual connections and layer normalization. Row attention models dependencies across alignment sites within each sequence, with parameters shared across all sequences via tied row attention for efficiency [16]. Column attention captures cross-sequence interactions at each alignment site, operating independently at each column of the MSA.

Since the ordering of taxa in an MSA is arbitrary, we do not include row-wise positional embeddings. Positional biases along the site dimension are also omitted in the current implementation. Additional architectural details are provided in the Supplementary Material.

*Transformer encoder pretraining with masked language modeling.* We represent a DNA multiple sequence alignment with  $N$  taxa and  $L$  aligned sites as a matrix,  $X \in \mathbb{R}^{N \times L}$ , where each entry corresponds to one of four nucleotides (A, C, G, T) or gap characters (“-”).

To mitigate the sparsity and high dimensionality of one-hot encoding and to preserve the column-wise homology structure of MSAs, we avoid subword tokenization (e.g., BPE) and k-mer encodings. Instead, each nucleotide or gap symbol is assigned a unique identifier from a fixed vocabulary and mapped to a learnable embedding via a shared embedding layer. This approach preserves site-level alignment while providing a compact and computationally efficient embedding.

We pretrain the Transformer encoder using a self-supervised masked language modeling (MLM) objective [16]: a subset of alignment entries is masked, and the model is trained to reconstruct the masked nucleotides from both row-wise (within-sequence) and column-wise (across-taxa) context. The training loss follows the standard cross-entropy formulation, with details provided in the Supplementary Material.

In this work, we adopt a similar column–row masking strategy in [16]. Columns are selected with probability  $P_{col}$ ; within each selected column, a fraction  $P_{row}$  of rows are masked while ensuring that at least  $K_{min}$  rows remain unmasked. This constraint prevents the model from receiving an insufficient observable signal within a column and stabilizes training. More details are provided in the Supplementary Materials.

*Learnable phylogenetic bias module.* The attention mechanism is largely data-driven and, by default, does not favor biologically structured interactions without additional inductive bias. In evolutionary MSAs, sequences are related through shared ancestry, which induces hierarchical and non-exchangeable dependencies across taxa. More closely related taxa are expected to exhibit stronger and more consistent correlations across sites than more distantly related ones. These relationships are structured by an underlying phylogenetic tree rather than being uniform across all sequence pairs. In contrast, column-wise attention infers cross-sequence similarity at each site directly from observed alignment patterns. Without additional constraints, this makes the attention scores sensitive to alignment noise, particularly in gap-rich or estimated MSAs.

More broadly, this idea of combining statistical deep learning with logical, mathematical and/or physical constraints has gained traction in the topic known as physics-informed machine learning (PIML) [13]. Rather than learning constraints and laws from scratch, incorporating basic structural information about a problem or domain under study can yield significant efficiencies. An intriguing corollary hypothesis suggests benefits in the other direction as well: it is possible that statistical representation learning can also augment mathematical and logical models in helpful ways and possibly also offset model mis-specification, although rigorous testing of this hypothesis would be needed for a given PIML application. Our goal is to implement this idea for the estimation task under study. To achieve this goal, we introduce a learnable phylogenetic bias module. This module encodes pairwise evolutionary relationships directly from learned taxa-level representations, without relying on external distance estimates, and integrates them into the column-wise attention mechanism to guide cross-sequence interactions.

For each taxon, we summarize its representation across all alignment sites by averaging the encoder hidden states along the site dimension. Given the encoder output  $\mathbf{H} \in \mathbb{R}^{N \times L \times d}$ , where  $d$  is the hidden dimension, this yields a taxa-level representation  $\mathbf{R} \in \mathbb{R}^{N \times d}$ . More expressive aggregation strategies are left for future work. We then apply  $M$  independent learned weight matrices to  $\mathbf{R}$ , each projecting it into a pair of query and key matrices, yielding  $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{M \times N \times d_h}$ , where  $d_h = d/M$  and  $M$  is the number of heads (multi-head attention). Using multiple independent transformations allows the model to capture different aspects of inter-sequence relationships simultaneously, analogous to how multiple sequence alignment tools consider different types of sequence similarity at once. For each head, we compute pairwise similarity scores between sequences as the

phylogenetic bias  $\mathbf{S}^{(m)} \in \mathbb{R}^{N \times N}$ ,

$$S_{ij}^{(m)} = \frac{1}{\sqrt{d_h}} \langle \mathbf{Q}_i^{(m)}, \mathbf{K}_j^{(m)} \rangle, \quad i, j \in 1, \dots, N. \quad (1)$$

where  $m$  indexes the attention heads, and  $\langle \cdot, \cdot \rangle$  denotes the dot product between two vectors. We enforce symmetry to reflect evolutionary relationships and remove self-interactions by zeroing the diagonal. The bias is then scaled by a set of learned head-specific positive coefficients  $\alpha \in \mathbb{R}^M$ , yielding  $\mathbf{B}^{(m)} = \alpha_m \mathbf{S}^{(m)}$  for each attention head. The resulting bias  $\mathbf{B} \in \mathbb{R}^{M \times N \times N}$  is added to the column-attention logits prior to softmax normalization. This encourages sequences with closer evolutionary relationships to receive higher attention weights during column-wise aggregation.

This design introduces phylogeny-informed structure into attention in a flexible and computationally efficient manner. We further include an oracle experiment using ground-truth evolutionary distances as a column-attention bias to validate the bias design; results are reported in the Supplementary Materials.

*Pairwise evolutionary distance prediction and phylogenetic tree inference.* To estimate pairwise evolutionary distances, we design a distance estimation module that operates on the encoder output  $\mathbf{H}$ . For each pair of taxa  $(i, j)$ , we construct symmetric pairwise features that capture complementary aspects of their relationship. Specifically, we use the element-wise absolute difference  $|\mathbf{H}_i - \mathbf{H}_j|$ , element-wise product  $\mathbf{H}_i \odot \mathbf{H}_j$ , and element-wise sum  $\mathbf{H}_i + \mathbf{H}_j$ . These features are concatenated and passed to a lightweight two-layer feedforward network to predict the pairwise evolutionary distance matrix. The prediction is supervised by a distance loss  $\mathcal{L}_{\text{dist}}$  that measures the discrepancy between predicted and true pairwise distances; the specific formulation is provided in Section 3 (“Performance evaluation criteria”). Additional details of the distance estimation module are provided in the Supplementary Materials.

Given the predicted pairwise distances from MSAs, phylogenetic trees can be reconstructed using distance-based methods such as FastME [11]. For this downstream analysis, we compare FIREFLY followed by FastME with a maximum-likelihood-based approach (FastTree) and with Phyloformer followed by FastME.

*Model training.* We adopt a staged training strategy for the encoder and the evolutionary distance estimation module. In the first stage, the encoder is pretrained using a masked language modeling (MLM) objective on 10,000 simulated true MSAs without indels, in order to learn fundamental sequence- and column-level representations. It is then further pretrained on 10,000 true MSAs containing indels under the same MLM objective to adapt the representations to alignment uncertainty.

In the second stage, the evolutionary distance estimation module is trained jointly with the pretrained encoder on 20,000 MSA–tree pairs, consisting of 10,000 true simulated MSAs and 10,000 corresponding estimated MSAs. During this stage, the parameters of the distance estimation module, the pretrained encoder, and the phylogenetic bias module are jointly optimized. The phylogenetic

bias module is activated only during supervised distance training. All training experiments use MSAs with a fixed number of taxa ( $N = 50$ ) and alignment length ( $L = 1024$ ).

The supervised training objective for evolutionary distance estimation module is defined as

$$\mathcal{L} = \mathcal{L}_{\text{dist}}^{\text{true}} + w_{\text{est}} \mathcal{L}_{\text{dist}}^{\text{est}},$$

where  $\mathcal{L}_{\text{dist}}^{\text{true}}$  and  $\mathcal{L}_{\text{dist}}^{\text{est}}$  denote the distance loss  $\mathcal{L}_{\text{dist}}$  computed on true and estimated MSAs, respectively; see Section 3 (“Performance evaluation criteria”) for more details. The weighting factor  $w_{\text{est}}$  is scheduled across training epochs: it is set to 0.5 for the first five epochs and then gradually increased to 1.0. This schedule allows the model to first emphasize cleaner evolutionary signals and progressively adapt to noise introduced by estimated alignments. Additional training details are provided in the Supplementary Materials.

### 3 Performance Benchmarking Study

*Simulation datasets.* Training and evaluation data are generated using a simulation-based pipeline designed to produce paired true and estimated MSAs, along with their corresponding phylogenetic trees, under controlled evolutionary scenarios. We first generate ultrametric phylogenies under a birth–death process, which are subsequently perturbed to obtain non-ultrametric trees following the procedure described in [20]. Each tree is then rescaled to match the empirical tree length distributions observed in public phylogenetic databases, consistent with the protocol used by Phyloformer [15].

Given each rescaled phylogeny, a true MSA is simulated using INDELible [5] under a finite-sites nucleotide substitution model with gamma-distributed among-site rate heterogeneity, incorporating insertion and deletion (indel) events. Substitution processes follow the general time-reversible (GTR) model [17], with base frequencies and substitution rate parameters derived from empirical rice datasets [6]. We adopt a medium gap-length distribution as described in [12]. An estimated MSA is then generated from the corresponding true MSA using a multiple sequence alignment algorithm.

All training datasets consist of MSAs with 50 taxa and an indel rate of 0.02, while datasets with 10 to 100 taxa and other indel rates (0.01 and 0.03) are used for evaluation. For all experiments, alignment lengths are truncated or padded to 1024 sites. Unless otherwise specified, all estimated MSAs used for both training and evaluation are generated using MAFFT [10], a widely used multiple sequence alignment method. Additional details on data simulation, parameter settings, and summary statistics of the datasets are provided in the Supplementary Materials.

*Empirical datasets.* We evaluate two downstream phylogenetic analysis tasks on three intronic RNA datasets – IGIC2, IGID, and IGIE – from the Comparative RNA Website (CRW) database [21]. The remaining intronic RNA datasets are excluded because their alignments exceed 6,000 sites, beyond the practical input



size supported by FIREFLY under current hardware constraints. Each dataset provides curated reference alignments and unaligned sequence data; maximum-likelihood reference trees from [21] are used in place of unknown ground-truth phylogenies. Estimated MSAs are constructed from the unaligned sequence files using MAFFT. For IGIE, which contains 249 taxa and 2,751 sites, the full alignment exceeds the processing limit and is randomly partitioned into five sub-alignments with their corresponding induced subtrees. For both downstream evaluation tasks, analyses are conducted on estimated alignments. Dataset summary statistics are provided in the Supplementary Materials.

*Performance evaluation criteria.* We evaluate pairwise evolutionary distance estimation using the mean absolute error (MAE) and mean relative error (MRE). Let  $d_{i,j}$  denote the true evolutionary distance between taxa  $i$  and  $j$ , and  $\hat{d}_{i,j}$  the corresponding predicted distance. The MAE and MRE are defined as

$$\text{MAE} = \frac{1}{|\mathcal{N}|} \sum_{(i,j) \in \mathcal{N}} |\hat{d}_{i,j} - d_{i,j}|, \quad \text{MRE} = \frac{1}{|\mathcal{N}|} \sum_{(i,j) \in \mathcal{N}} \frac{|\hat{d}_{i,j} - d_{i,j}|}{d_{i,j}},$$

where  $\mathcal{N} = (i, j) \mid 1 \leq i < j \leq N$  denotes the set of all unordered taxon pairs.

The distance loss  $\mathcal{L}_{\text{dist}}$  used for supervised training is defined as a weighted combination of these two evaluation criteria,

$$\mathcal{L}_{\text{dist}} = \text{MAE} + \lambda \text{MRE},$$

with  $\lambda = 0.3$  in all experiments.

For a reconstructed phylogenetic tree, the normalized Robinson–Foulds (RF) distance is used to evaluate the topological accuracy. In a phylogenetic tree, each branch induces a bipartition of the leaf set. Let  $A$  and  $B$  denote the sets of bipartitions in the true tree  $T$  and the estimated tree  $T^*$ , respectively. The normalized RF distance is defined as

$$\text{NRF}_{\text{norm}}(T, T^*) = (|A| + |B|)^{-1} (|A \cup B| - |A \cap B|),$$

which measures the fraction of discordant bipartitions between the two trees.

For a distance matrix induced by a phylogenetic tree, satisfaction of the four-point condition is expected, as it reflects the fundamental additivity property of tree-based evolutionary distances. To directly evaluate this structural property, we introduce the four-point residual as an additional assessment to quantify deviations from tree additivity. For any quartet of distinct taxa  $\{i, j, k, l\}$ , let  $d_T(\cdot, \cdot)$  denote the pairwise path length in tree  $T$ , and define  $S_1 = d_T(i, j) + d_T(k, l)$ ,  $S_2 = d_T(i, k) + d_T(j, l)$ , and  $S_3 = d_T(i, l) + d_T(j, k)$ . The four-point residual for this quartet is

$$\Delta_T(i, j, k, l) = \max\{S_1, S_2, S_3\} - \text{mid}\{S_1, S_2, S_3\},$$

where  $\text{mid}(\cdot)$  denotes the median. For an additive tree metric, the two largest sums are equal for every quartet, yielding  $\Delta_T = 0$ . Larger values indicate

stronger violations of tree additivity. To reduce computational cost, we randomly sample 2000 quartets and compute the four-point residual score by averaging  $\Delta_T(i, j, k, l)$  over the sampled quartets.

## 4 Results

*Performance on true MSAs with indels.* We evaluate FIREFLY on two downstream phylogenetic tasks using true MSAs with indels: pairwise evolutionary distance inference and phylogenetic tree reconstruction.

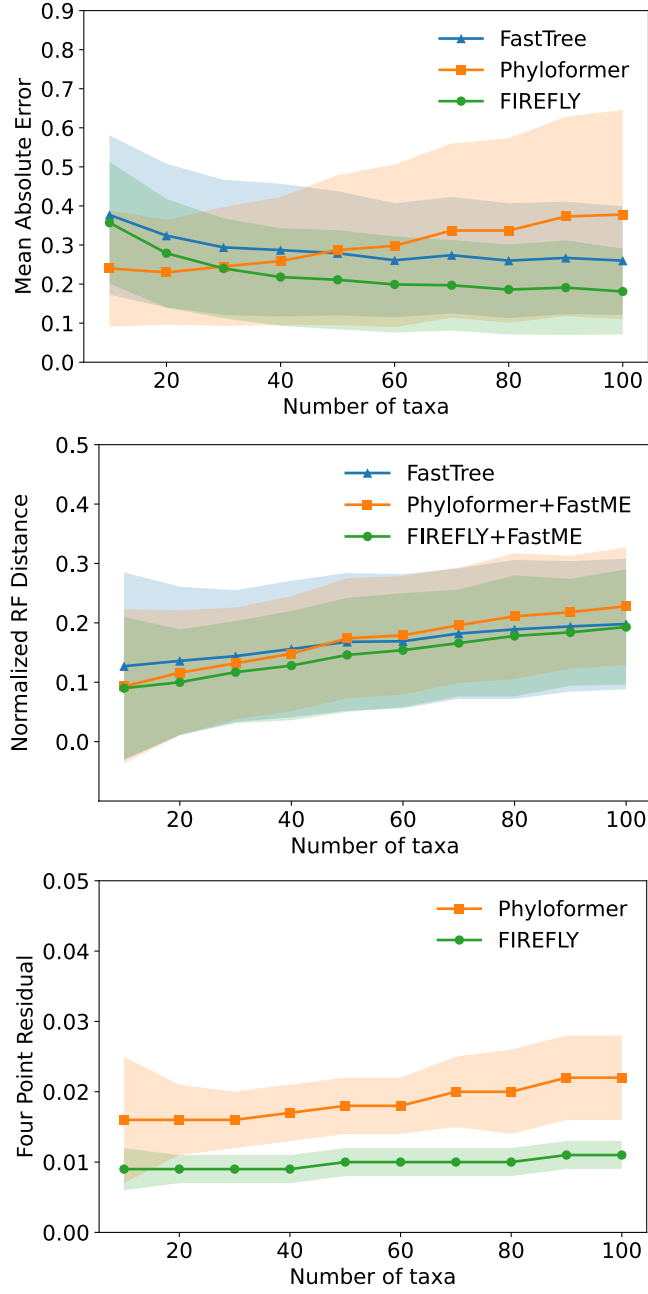
We first assess distance estimation accuracy using mean absolute error (MAE). As shown in Figure 2 (top right), FIREFLY outperforms Phyloformer across most taxa sizes, with the exception of small MSAs ( $N \leq 30$ ). The performance gap increases as the number of taxa grows. Compared with FastTree, FIREFLY achieves comparable or lower MAE and consistently shows reduced variance. We next evaluate phylogenetic tree reconstruction accuracy using the normalized Robinson–Foulds (NRF) distance. As shown in Figure 2 (top left), FIREFLY combined with FastME (FIREFLY+FastME) achieves comparable or lower NRF distances than both Phyloformer combined with FastME (Phyloformer+FastME) and FastTree across all taxa sizes. While NRF distance increases for all methods as the number of taxa grows, the relative trends differ: the advantage of FIREFLY+FastME over FastTree decreases with the number of taxa, whereas its advantage over Phyloformer+FastME increases.

Beyond topology and distance accuracy, we assess the additivity of the inferred distance matrices using the four-point residual. This analysis is restricted to Phyloformer and FIREFLY, since FastTree is not a distance-based method and distances derived from inferred trees are naturally additive. As shown in Figure 2 (bottom), FIREFLY consistently yields lower four-point residuals with smaller variance than Phyloformer across all taxa sizes, indicating closer adherence to tree additivity.

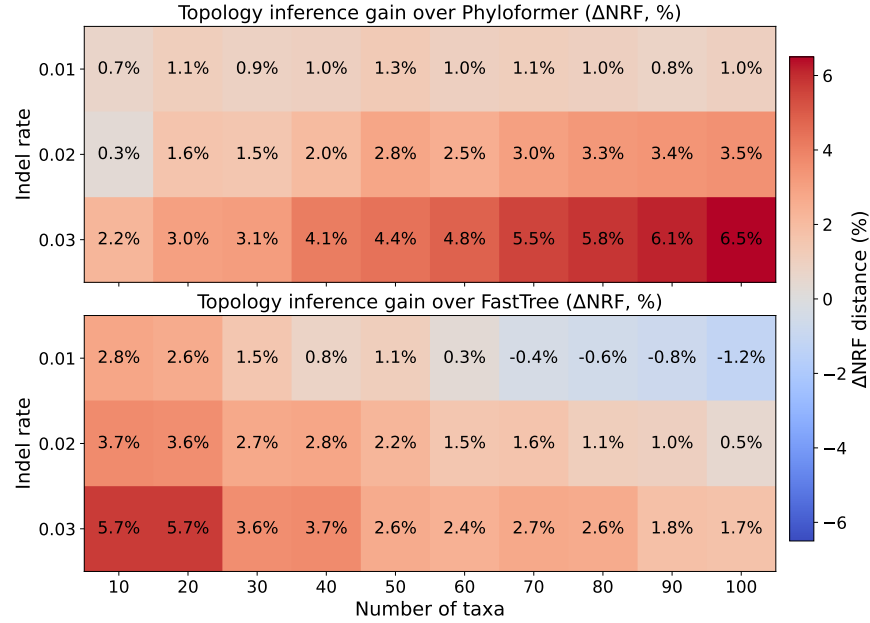
Overall, these results indicate that on true MSAs with indels, FIREFLY achieves competitive accuracy in pairwise distance estimation and phylogenetic tree topology inference, while exhibiting improved conformity to tree additivity in the predicted pairwise distances.

*Robustness under high indel rates.* To evaluate the robustness to insertion/deletion (indel) events, we compared the downstream phylogenetic tree topology inference performance of FIREFLY+FastME with Phyloformer+FastME and FastTree across increasing indel rates and varying numbers of taxa. Figure 3 reports the mean topology inference performance gain, measured as the difference in normalized Robinson–Foulds (NRF) distance, where positive values indicate improved accuracy of FIREFLY.

Across all experimental settings, FIREFLY+FastME consistently outperforms Phyloformer+FastME. At low indel rates ( $\rho = 0.01$ ), the gains are modest and relatively stable across taxa sizes, typically below 1.5%. As the indel rate increases, the advantage of FIREFLY and becomes more pronounced and grows



**Fig. 2. Performance on simulated true MSAs across different numbers of taxa.** Results are shown for three evaluation assessments: normalized Robinson-Foulds (NRF) distance, mean absolute error (MAE) of pairwise distances, and four-point residual. Curves (and shaded regions) report average (and standard deviation) across replicates.



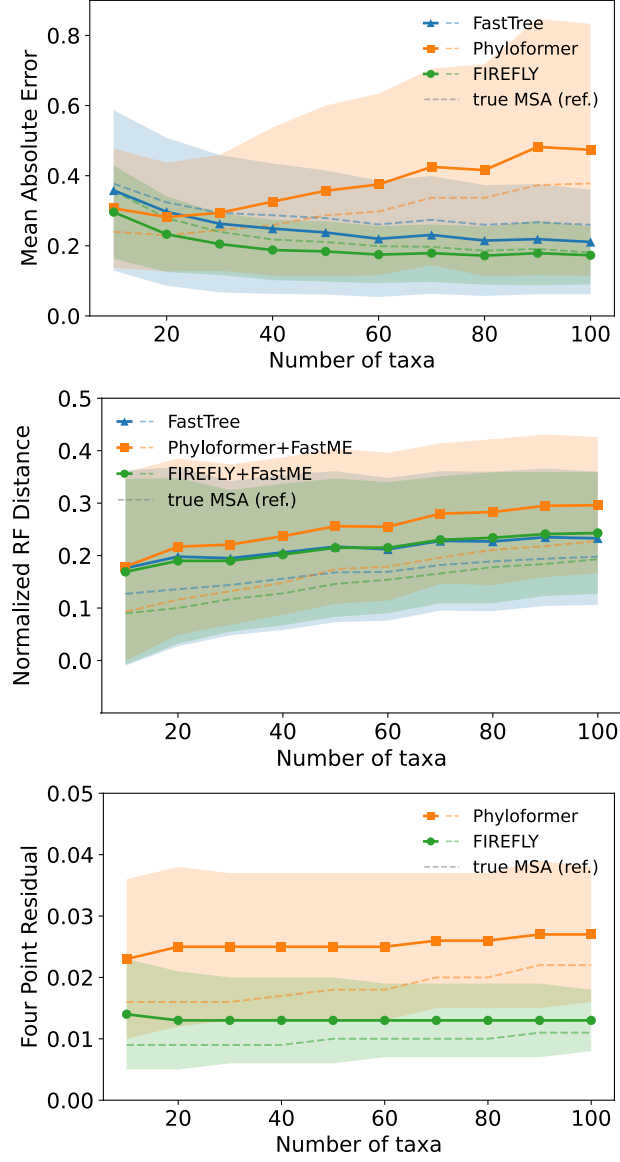
**Fig. 3. Phylogenetic tree topology inference gain of FIREFLY+FastME over two baselines on simulated true MSAs with different indel rates.** Heatmaps report the mean performance gain of FIREFLY in phylogenetic topology inference, measured by the difference in normalized Robinson–Foulds (NRF) distance across varying indel rates and numbers of taxa. Positive values (red) indicate improved topology inference by FIREFLY relative to the corresponding baseline. The upper panel compares FIREFLY+FastME with Phyloformer+FastME, while the lower panel compares FIREFLY+FastME with FastTree.

with the number of taxa. At a high indel rate ( $\rho = 0.03$ ) and  $N = 100$  taxa, the performance gain exceeds 6%. Compared with FastTree, a similar trend is observed with respect to indel rates: the topology inference gain increases as the indel rate rises. However, a different pattern emerges regarding the number of taxa. In this case, the gain decreases as the number of taxa increases. For example, at  $\rho = 0.03$  and small taxa sizes ( $N = 10, 20$ ), the gain exceeds 5%, while it is smaller for larger trees. This difference is expected. Both Phyloformer+FastME and FIREFLY+FastME rely on two-step, distance-based tree construction. In contrast, FastTree performs direct likelihood-based optimization over the tree space and is specifically designed for large numbers of taxa. As tree size increases, the advantage of improved distance estimation is partially offset by the limitations of distance-based tree reconstruction.

Overall, the increasing performance gain with higher indel rates indicates that FIREFLY is particularly effective at mitigating the detrimental effects of frequent indel events. We attribute this robustness to its full-MSA representation learning strategy, which better leverages gap patterns and alignment structure under high indel conditions.

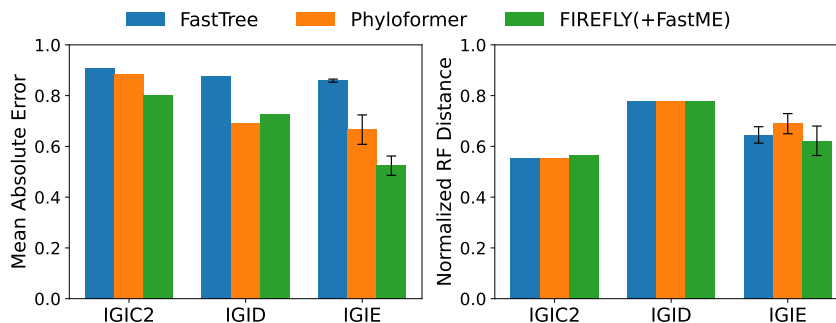
*Robustness to alignment errors on estimated MSAs with indels.* Most existing deep learning-based phylogenetic analysis methods are evaluated on simulated true MSAs, where sequences are perfectly aligned. While this setting is useful for theoretical experimentation, it does not reflect typical real-world analyses, where MSAs are estimated and inevitably contain alignment errors. Such alignment errors can distort evolutionary signals and propagate into downstream phylogenetic analysis, as shown in prior studies [1, 7]. Therefore, assessing robustness to alignment errors is critical for evaluating practical applicability. For this purpose, we evaluated FIREFLY on estimated MSAs containing indels.

As shown in Fig. 4, the impact of alignment errors varies across methods and evaluation metrics. For NRF distance and four-point residual, all methods show performance degradation on estimated MSAs compared to true MSAs. For MAE, Phyloformer exhibits pronounced degradation on estimated MSAs, reflecting strong sensitivity to alignment noise, while FIREFLY and FastTree show slight improvement, suggesting greater robustness in distance estimation. For phylogenetic tree reconstruction, the degraded distance estimates of Phyloformer lead to reduced topology accuracy. As a result, the performance gap between Phyloformer+FastME and FIREFLY+FastME increases on estimated MSAs compared with true MSAs. For example, at  $N = 100$  taxa, the NRF difference increases from 3% to 5%. In contrast, the performance gap between FIREFLY+FastME and FastTree diminishes, and the two methods achieve comparable accuracy across taxa sizes. This trend likely reflects the increased impact of alignment errors on two-step distance-based pipelines. Most notably, alignment errors lead to substantially larger four-point residual values and increased variance for both methods. However, the relative advantage of FIREFLY over Phyloformer also becomes more pronounced under estimated MSAs, indicating improved robustness of FIREFLY to alignment-induced noise.



**Fig. 4. Performance comparison on estimated MSAs across different numbers of taxa.** Results are shown for the same assessments and layout as in Fig. 2, but evaluated using estimated MASs from the simulated datasets.

Overall, these results indicate that the advantages of FIREFLY extend beyond idealized settings and are most evident under realistic conditions with estimated MSAs. In the supplementary material, we further observe consistent trends across estimated MSAs generated by alternative alignment methods.

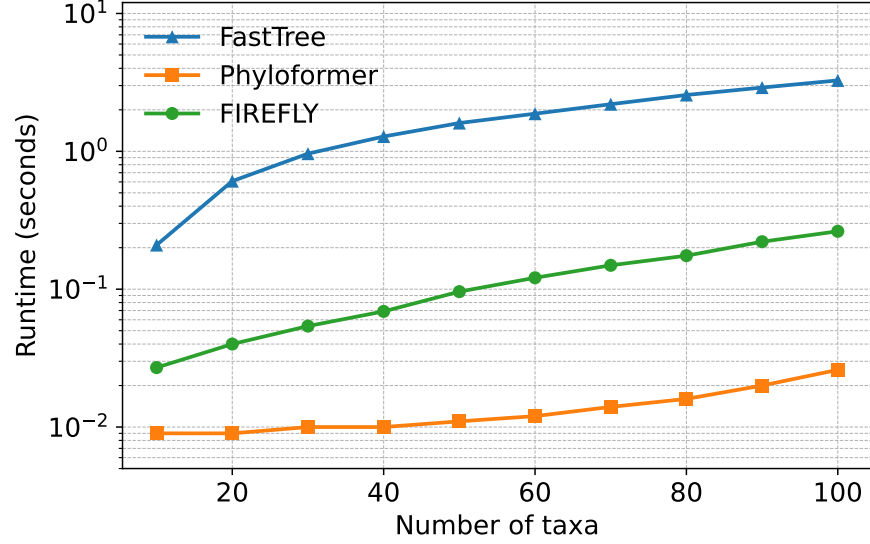


**Fig. 5. Performance comparison on CRW intronic RNA datasets.** The left panel shows mean absolute error (MAE) of pairwise distance prediction, and the right panel shows the normalized Robinson–Foulds distance (NRFD) of downstream tree inference, evaluated on estimated alignments. Error bars denote standard deviation across IGIE partitions.

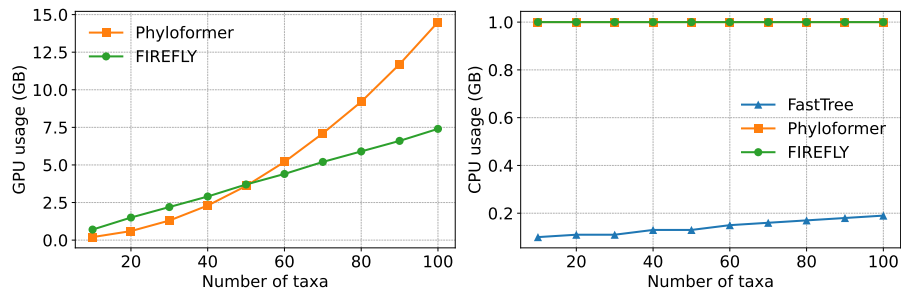
*Computational runtime and memory usage.* Similar to other deep learning-based approaches, FIREFLY enables substantially faster inference after training than full maximum-likelihood or Bayesian methods. We assessed computational efficiency by measuring runtime as a function of the number of taxa in Figure 6. Phyloformer achieves the fastest runtime across all settings, while FastTree incurs substantially higher computational cost due to explicit likelihood evaluation and tree search procedures. FIREFLY introduces a moderate runtime overhead relative to Phyloformer, reflecting the additional cost of learning representations from the full MSA rather than from pairwise-averaged inputs, but remains one to two orders of magnitude faster than FastTree. Importantly, FIREFLY maintains sub-second runtimes even for trees with 100 taxa, demonstrating favorable scaling and practical applicability.

As shown in Fig. 7, GPU memory increases as the number of taxa grows for both deep learning models. The increase is steeper for Phyloformer than for FIREFLY. In contrast, CPU memory remains stable across taxa sizes for both models, since most computations are carried out on the GPU rather than the CPU. FastTree only uses CPU memory, which stays low and increases slightly with taxa.

*Ablation Study on the Learned Phylogenetic Bias Module.* We conducted ablation experiments to evaluate the contribution of the learned phylogenetic bias module.



**Fig. 6. Comparison of downstream phylogenetic tree inference runtime across different methods.** Execution time for tree reconstruction pipelines evaluated on simulated MSAs with 1024 sites. For deep learning-based approaches, the runtime includes distance prediction followed by FastME tree inference, excluding model initialization and weight loading.



**Fig. 7. Computational memory usage across different numbers of taxa.** Left: GPU memory usage for deep learning models (Phyloformer and FIREFLY). Right: CPU memory usage for all methods.



Results show that the bias term provides increasing benefit as the number of taxa and indel rate grow, yielding up to 37% reduction in distance estimation error. Furthermore, different attention heads learn complementary correlation patterns with the true evolutionary distances, suggesting that the multi-head formulation captures diverse aspects of inter-sequence relationships. Details are presented in Supplementary Appendix subsections 3.2 and 3.3.

*Performance on empirical datasets.* We compare the performance of FastTree, Phyloformer, and FIREFLY on three CRW intronic RNA datasets using estimated alignments. Figure 5 reports the mean absolute error (MAE) for pairwise distance prediction and the normalized Robinson–Foulds (NRFD) distance for downstream phylogenetic tree inference. Error bars are shown only for IGIE, where multiple partitions are available; the other datasets consist of a single instance.

Across datasets, FIREFLY achieves the lowest MAE in most cases, indicating more accurate pairwise distance prediction, except for IGID, where Phyloformer performs slightly better. The performance gap is particularly pronounced on IGIE, which remains more challenging despite partitioning due to its larger taxon sampling and higher gappiness compared to the other datasets. For downstream tree inference, all methods yield comparable NRFD values on IGIC2 and IGID. In contrast, on the more challenging IGIE dataset, FIREFLY produces trees with lower NRFD, particularly compared to Phyloformer, indicating improved robustness under increased alignment noise and larger taxon sampling.

## 5 Discussion

Across experiments on both true and estimated MSAs, as well as under varying indel rates, FIREFLY consistently improves pairwise distance estimation and phylogenetic tree reconstruction. These gains are most pronounced for MSAs with larger numbers of taxa and higher levels of noise, arising from alignment errors and higher indel events.

Comparisons between true and estimated MSAs highlight the practical value of this design. On estimated MSAs, where alignment errors introduce noise and weaken column-wise evolutionary signals, FIREFLY shows a clearer advantage over Phyloformer and achieves performance comparable to maximum likelihood estimation. This robustness is particularly important in realistic phylogenetic settings, where true alignments are unavailable, and estimation error is unavoidable. The improvement over Phyloformer underscores the importance of full-MSA representation learning for phylogenetic analysis. By modeling evolutionary structure at the alignment level, FIREFLY avoids the systematic biases associated with purely pairwise learning strategies, which have been shown to affect phylogenetic comparison and construction [2, 4]. These systematic biases become more severe in the presence of alignment errors.

Across simulations with increasing indel rates, FIREFLY shows progressively larger performance gains. Higher indel rates reflect more complex evolutionary

processes and lead to alignments with a larger fraction of gap-related uncertainty. In such settings, traditional likelihood-based phylogenetic methods typically treat gaps as missing data, limiting their ability to exploit evolutionary signals carried by indel patterns. In contrast, FIREFLY learns representations directly from full MSAs and integrates both sequence content and gap structure through row- and column-wise context. By comparison, Phyloformer relies on pairwise sequence representations, which compress the alignment and lose indel-related structure. Together, these factors explain why FIREFLY substantially outperforms both FastTree and Phyloformer in high-indel regimes, highlighting the benefit of MSA-level representation learning.

More broadly, this work highlights the value of integrating modern representation learning with rules and constraints based on first principles. By incorporating phylogenetically motivated inductive bias directly into the model architecture, FIREFLY moves beyond purely data-driven approaches and better aligns learned representations with evolutionary structure.

## 6 Conclusion

In this study, we introduced FIREFLY, a Transformer-based framework that learns phylogeny-informed representations directly from multiple sequence alignments and supports downstream phylogenetic analyses. By integrating a learnable phylogenetic bias into MSA representation learning, FIREFLY improves both evolutionary distance estimation and downstream phylogenetic tree inference, particularly in settings characterized by high indel rates and alignment uncertainty. Using simulated and empirical datasets, we conduct extensive performance benchmarking and ablation experiments, and we demonstrate that explicitly modeling phylogenetic structure within deep representations leads to more robust and accurate inference. These findings underscore the potential of combining modern representation learning with classical evolutionary knowledge to advance phylogenetic analysis in complex evolutionary regimes.

We conclude with thoughts on future research directions. We hypothesize that FIREFLY’s phylogeny-informed representation learning can be applied not only to phylogenetic distance estimation and tree reconstruction, but also to other tasks in computational phylogenetics. More generally, FIREFLY provides a case study of an important insight from physics-informed machine learning (PIML): logic and theory synergizes with purely statistical representation learning. This insight can be leveraged throughout computational biology and bioinformatics and beyond.

*Data availability.* Data and scripts used in this study are publicly available under an open copleyleft license at <https://gitlab.msu.edu/liulab/firefly-study-data-and-scripts>.

*Acknowledgments.* This research has been supported in part by the National Science Foundation (DBI-2144121, DBI-2214038, and CCF-1714417 to KJL). All computational experiments and analyses were performed on the MSU High Performance Computing Center, which is part of the MSU Institute for Cyber-Enabled Research.

*Disclosure of Interests.* The authors have no competing interests to declare.

## References

1. Ashkenazy, H., Sela, I., Levy Karin, E., Landan, G., Pupko, T.: Multiple sequence alignment averaging improves phylogeny reconstruction. *Systematic Biology* **68**(1), 117–130 (2019)
2. Dunn, C.W., Zapata, F., Munro, C., Siebert, S., Hejnol, A.: Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proceedings of the National Academy of Sciences* **115**(3), E409–E417 (2018)
3. Felsenstein, J.: Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**(6), 368–376 (1981)
4. Felsenstein, J.: Phylogenies and the comparative method. *The American Naturalist* **125**(1), 1–15 (1985)
5. Fletcher, W., Yang, Z.: INDELible: a flexible simulator of biological sequence evolution. *Molecular Biology and Evolution* **26**(8), 1879–1888 (2009)
6. Gao, M., Liu, K.J.: Statistical analysis of GC-biased gene conversion and recombination hotspots in eukaryotic genomes: a phylogenetic hidden Markov model-based approach. In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. pp. 1–24 (2021)
7. Gao, M., Wang, W., Liu, K.J.: The impact of gene sequence alignment and gene tree estimation error on summary-based species network estimation. In: *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. pp. 1–17 (2022)
8. Gascuel, O.: BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* **14**(7), 685–695 (1997)
9. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al.: Highly accurate protein structure prediction with AlphaFold. *Nature* **596**(7873), 583–589 (2021)
10. Katoh, K., Standley, D.M.: MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**(4), 772–780 (2013)
11. Lefort, V., Desper, R., Gascuel, O.: FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular Biology and Evolution* **32**(10), 2798–2800 (2015)
12. Liu, K., Warnow, T.J., Holder, M.T., Nelesen, S.M., Yu, J., Stamatakis, A.P., Linder, C.R.: SATE-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology* **61**(1), 90 (2012)
13. Meng, C., Griesemer, S., Cao, D., Seo, S., Liu, Y.: When physics meets machine learning: A survey of physics-informed machine learning. *Machine Learning for Computational Science and Engineering* **1**(1), 20 (2025)
14. Mo, Y.K., Hahn, M.W., Smith, M.L.: Applications of machine learning in phylogenetics. *Molecular Phylogenetics and Evolution* **196**, 108066 (2024)
15. Nesterenko, L., Blassel, L., Veber, P., Boussau, B., Jacob, L.: Phyloformer: fast, accurate, and versatile phylogenetic reconstruction with deep neural networks. *Molecular Biology and Evolution* **42**(4), msaf051 (2025)
16. Rao, R.M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., Rives, A.: MSA transformer. In: *International Conference on Machine Learning*. pp. 8844–8856. PMLR (2021)

17. Rodriguez, F., Oliver, J., Marin, A., Medina, J.: The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology* **142**, 485–501 (1990)
18. Smith, M.L., Hahn, M.W.: Phylogenetic inference using generative adversarial networks. *Bioinformatics* **39**(9), btad543 (2023)
19. Suvorov, A., Hochuli, J., Schrider, D.R.: Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Systematic Biology* **69**(2), 221–233 (2020)
20. Szöllösi, G.J., Höhna, S., Williams, T.A., Schrempf, D., Daubin, V., Boussau, B.: Relative time constraints improve molecular dating. *Systematic Biology* **71**(4), 797–809 (2022)
21. Wang, W., Hejasebazzi, A., Zheng, J., Liu, K.J.: Build a better bootstrap and the RAWR shall beat a random path to your door: phylogenetic support estimation revisited. *Bioinformatics* **37**(Supplement\_1), i111–i119 (2021)
22. Warnow, T.: Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. *PLoS Currents* **4**, RRN1308 (2012)
23. Warnow, T.: *Computational phylogenetics: an introduction to designing methods for phylogeny estimation*. Cambridge University Press (2018)
24. Wong, K.M., Suchard, M.A., Huelsenbeck, J.P.: Alignment uncertainty and genomic analysis. *Science* **319**(5862), 473–476 (2008)
25. Yang, Z., Rannala, B.: Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Molecular Biology and Evolution* **14**(7), 717–724 (1997)
26. Zaharias, P., Grosshauser, M., Warnow, T.: Re-evaluating deep neural networks for phylogeny estimation: the issue of taxon sampling. *Journal of Computational Biology* **29**(1), 74–89 (2022)
27. Zou, Z., Zhang, H., Guan, Y., Zhang, J.: Deep residual neural networks resolve quartet molecular phylogenies. *Molecular Biology and Evolution* **37**(5), 1495–1507 (2020)

# Supplementary Appendix – FIREFLY: PHYlogeny-informed REpresentation learning to estimate PHyLogenetic dIstances

Meijun Gao<sup>1</sup>, Byungho Lee<sup>1,2</sup>, and Kevin J. Liu<sup>1,2,3</sup>

<sup>1</sup> Department of Computer Science and Engineering

<sup>2</sup> Ecology, Evolution, and Behavior Program

<sup>3</sup> Genetics and Genome Sciences Program

Michigan State University

East Lansing, MI, USA

kjl@msu.edu

## 1 Supplementary Methods

### 1.1 The procedures for data simulation

In this study, we follow exactly the simulation procedure described in Phyloformer [7] to generate model gene trees. Specifically, an ultrametric tree is first simulated under a birth–death process, after which branch lengths are individually rescaled to deviate from ultrametric. The tree is then rescaled to match a target diameter sampled from empirical data. Finally terminal branches shorter than a predefined minimum are extended to satisfy the lower bound constraint.

DNA sequence alignments are simulated under a finite-sites nucleotide substitution model with gamma-distributed among-site rate heterogeneity along the model trees using INDELible [1].

For site-rate heterogeneity, we assume a gamma distribution with no invariant sites ( $\text{pinv} = 0$ ). For the gamma shape parameter  $\alpha$ , we follow exactly the sampling procedure described in Phyloformer [7], where  $\alpha$  values are drawn from the empirical distribution inferred from HOGENOM alignments, with additional Gaussian perturbation and a minimum threshold of 0.05. Substitution processes follow the general time-reversible (GTR) model [8], with base frequencies and substitution rate parameters derived from empirical rice datasets [2]. When insertion and deletion (indel) events are included, we adopt the medium gap-length distribution described in [5]. The indel rate is set to 0.01, 0.02, or 0.03 depending on the experimental setting.

### 1.2 Transformer encoder architecture

We adopt an encoder-only Transformer architecture, as illustrated in Figure 1. The encoder consists of six stacked attention blocks, each with eight attention heads and an embedding dimension of  $d = 256$ .

Input MSAs are encoded using an index-based tokenization scheme with a fixed vocabulary  $\{\text{A, C, G, T, -, PAD, MASK}\}$ . Each token is mapped to a learnable embedding vector, resulting in an input tensor of shape  $N \times L \times d$ , where  $N$  is the number of taxa and  $L$  the alignment length. Within each block, column attention is first applied independently to each alignment column to model interactions across sequences at a given site. This is followed by row attention applied along each sequence to capture dependencies across alignment positions. A position-wise feedforward network with hidden dimension  $d_f = 4d$  and ReLU activation is then applied. We include one LayerNorm normalization layer before and one residual connection after each of the row-wise, column-wise attention and feedforward layers. After six attention blocks and a final layer normalization with dropout, the resulting MSA representation  $\mathbf{H}$  is passed to the distance estimation module.

### 1.3 Pairwise distance estimation module

Let the the encoder output hidden states be  $\mathbf{H} \in \mathbb{R}^{N \times L \times d}$ , where  $N$  the number of taxa,  $L$  the alignment length, and  $d$  the embedding dimension. For each taxa pair  $(i, j)$ , we construct a symmetric pairwise feature:

$$\mathbf{F}_{ij} = \left[ |\mathbf{H}_i - \mathbf{H}_j|, \mathbf{H}_i \odot \mathbf{H}_j, \mathbf{H}_i + \mathbf{H}_j \right] \in \mathbb{R}^{L \times 3d}, \quad (1)$$

where  $\mathbf{H}_i \in \mathbb{R}^{L \times d}$  denotes the embedding of taxon  $i$ . This results in a feature tensor  $\mathbf{F} \in \mathbb{R}^{N \times N \times L \times 3d}$ .

To aggregate information across alignment sites, we apply a learnable pooling operation over the site dimension  $L$  of  $\mathbf{F}$ , where site-wise weights are computed using a learnable scoring function. The weighted features are summed across sites to obtain a pooled pair representation for each taxa pair. The pooled representations are then mapped to scalar scores via a projection head. To reflect the absence of self-distances, diagonal entries are fixed to zero. The outputs are transformed using a softplus function to ensure non-negative values, yielding the final distance matrix  $\mathbf{D} \in \mathbb{R}^{N \times N}$ .

Both the scoring function and the projection head consist of LayerNorm followed by a two-layer feedforward network with GELU activation and dropout.

#### 1.4 Model training details

*Pretraining via masked language modeling.* The Transformer encoder is first pretrained on 10K simulated MSAs without indel events, and is then fine-tuned on an additional 10K simulated MSAs with indel events, both using a masked language modeling (MLM) objective that trains the model to reconstruct randomly masked tokens from their surrounding context. This encourages the encoder to capture both local sequence patterns and long-range evolutionary dependencies across taxa without requiring labeled data. We employ a structured row-column masking strategy. Specifically,  $P_{token}$  denotes the fraction of all tokens to be masked, and  $P_{row}$  denotes the fraction of rows selected for masking. During the first 5% of training steps,  $P_{token}$  and  $P_{row}$  are linearly increased from 12% to 15% and from 15% to 30%, respectively, and are then kept fixed for the remainder of training. The column masking probability  $P_{col}$  is derived implicitly as

$$P_{col} = \frac{P_{token}}{P_{row}}.$$

To ensure sufficient unmasked context for learning, we enforce that at least  $K_{\min}$  rows remain unmasked in each MSA, where  $K_{\min}$  is set to 25% of the total number of taxa.

The MLM pretraining objective minimizes the cross-entropy loss between the predicted and true nucleotide tokens at masked positions:

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{|\mathcal{M}|} \sum_{(i,j) \in \mathcal{M}} \log p_{\theta}(x_{ij} | \tilde{\mathbf{X}}),$$

where  $\mathcal{M}$  denotes the set of masked positions,  $x_{ij}$  is the true token at row  $i$  and column  $j$ , and  $\tilde{\mathbf{X}}$  is the masked input MSA.

*Supervised distance estimation module training.* Subsequently, the distance estimation module is trained jointly with the pretrained encoder on 20K MSA-tree pairs, consisting of 10K true simulated MSAs and 10K corresponding estimated MSAs. True pairwise distances are used as labels, and the training objective combines distance losses  $\mathcal{L}_{\text{dist}}$  on both data types:

$$\mathcal{L} = \mathcal{L}_{\text{dist}}^{\text{true}} + w_{\text{est}} \mathcal{L}_{\text{dist}}^{\text{est}},$$

where  $\mathcal{L}_{\text{dist}}^{\text{true}}$  and  $\mathcal{L}_{\text{dist}}^{\text{est}}$  denote the distance loss  $\mathcal{L}_{\text{dist}}$  computed on true and estimated MSAs respectively, and  $w_{\text{est}}$  is a weighting coefficient.

All models are trained using the AdamW optimizer [6]. Training is performed for a fixed maximum number of epochs with early stopping, where training is terminated if the validation loss does not improve. The checkpoint with the lowest validation loss on a held-out validation set is selected. We adopt a linear learning rate schedule with 10% warmup steps to the target learning rate, followed by linear decay for the remaining training steps. All experiments are conducted on a single NVIDIA H200 GPU with 141 GB of VRAM. Training configurations, wall-clock training time, and the peak GPU memory usage for all models are summarized in Table S1.

**Supplementary Table S1. Training configurations and resource usage for all models.** The table reports dataset sizes, batch sizes, numbers of training epochs, target learning rates, wall-clock training time, and peak GPU memory usage for each model.

Model name	Batch size	Dataset size	Indels model	Epochs	GPUs	Target Learning rate	Training time (wall-clock)	Peak GPU memory
Base pretrained encoder	4	10k	without	20	1×H200	$5 \times 10^{-4}$	32 hours	44GB
Pretrained encoder	4	10k	with	20	1×H200	$1 \times 10^{-3}$	30 hours	47GB
Pretrained encoder & Distance estimation module	4	20k	with	30	1×H200	$1 \times 10^{-3}$	106 hours	132GB

## 2 Supplementary Materials

### 2.1 Simulation datasets

Tables S2 and S4 report the summary statistics of the simulated test datasets under different indel rates, along with the corresponding estimated MSAs produced by the various alignment methods. We additionally report the alignment errors of the estimated MSAs produced by the different alignment methods in Table S3.

**Supplementary Table S2. Summary statistics of simulated testing datasets with 0.02 indel rate across varying numbers of taxa.** Average normalized Hamming distance (“ANHD”) and the percentage of indels (“Gap.”) are reported for both simulated true and estimated MSAs, while the number of MSA sites (“Len.”) is reported for estimated MSAs only. All values are averaged over 1000 test replicates.

$N$	True MSA		MAFFT MSA			ClustalW MSA			Clustal Omega MSA		
	ANHD	Gap.	Len.	ANHD	Gap.	Len.	ANHD	Gap.	Len.	ANHD	Gap.
10	0.398	0.323	856.2	0.411	0.205	775.1	0.427	0.114	797.2	0.433	0.142
20	0.394	0.386	825.5	0.411	0.260	729.1	0.436	0.150	746.4	0.443	0.175
30	0.383	0.419	813.8	0.401	0.292	707.9	0.430	0.174	722.1	0.438	0.196
40	0.376	0.448	798.6	0.393	0.319	684.8	0.427	0.191	696.7	0.434	0.212
50	0.367	0.470	785.8	0.384	0.339	667.8	0.419	0.206	677.1	0.428	0.225
60	0.358	0.479	788.5	0.372	0.353	665.4	0.409	0.216	671.0	0.417	0.231
70	0.366	0.522	758.6	0.381	0.386	623.4	0.422	0.233	631.0	0.429	0.252
80	0.345	0.512	770.2	0.358	0.383	641.1	0.397	0.241	642.6	0.407	0.252
90	0.354	0.542	750.2	0.367	0.409	611.6	0.408	0.255	613.2	0.417	0.267
100	0.341	0.543	755.3	0.353	0.413	616.6	0.395	0.261	614.9	0.405	0.269

### 2.2 Empirical datasets

Table S5 summarizes the statistics of the three CRW intronic RNA datasets for both the reference and estimated MSAs. For the IGIE partitions dataset, statistics are computed on partitioned alignments after removing columns that contain only gaps. Reported values are averaged across all partitions.

## 3 Supplementary Results

### 3.1 Ground-Truth Bias as an Upper Bound

As an upper bound sanity check, we first replace the learnable phylogenetic bias with the ground-truth pairwise evolutionary distances and use them as the column-attention bias  $S$  during both training and

**Supplementary Table S3. Alignment estimation error on testing datasets with 0.02 indel rate measured by sum-of-pairs false positives (SPFP) and false negatives (SPFN).** SPFP quantifies the proportion of nucleotide homologies present in the estimated alignment but absent from the true alignment, SPFN measures the proportion of true nucleotide homologies missing from the estimated alignment. Average SPFP and SPFN values are reported for each MSA method.

<i>N</i>	MAFFT MSA		ClustalW MSA		Clustal Omega MSA	
	SPFP	SPFN	SPFP	SPFN	SPFP	SPFN
10	0.425	0.403	0.481	0.445	0.459	0.430
20	0.374	0.358	0.453	0.425	0.439	0.420
30	0.327	0.315	0.423	0.400	0.411	0.398
40	0.306	0.298	0.411	0.392	0.398	0.390
50	0.286	0.279	0.396	0.379	0.388	0.382
60	0.257	0.253	0.372	0.360	0.363	0.362
70	0.267	0.264	0.394	0.382	0.380	0.380
80	0.228	0.226	0.350	0.342	0.344	0.347
90	0.238	0.238	0.364	0.357	0.358	0.363
100	0.215	0.215	0.344	0.340	0.337	0.344

**Supplementary Table S4. Summary statistics of simulated testing datasets with other indel rates across varying numbers of taxa.** Average normalized Hamming distance (“ANHD”) and the percentage of indels (“Gap.”) are reported for both simulated true and estimated MSAs, while the number of MSA sites (“Len.”) is reported for estimated MSAs only. All values are averaged over 1000 test replicates.

<i>N</i>	indel rate=0.01					indel rate=0.03				
	True MSA		MAFFT MSA			True MSA		MAFFT MSA		
	ANHD	Gap.	Len.	ANHD	Gap.	ANHD	Gap.	Len.	ANHD	Gap.
10	0.406	0.213	949.2	0.409	0.157	0.409	0.412	773.4	0.430	0.242
20	0.408	0.267	928.8	0.415	0.202	0.400	0.475	744.6	0.424	0.309
30	0.393	0.297	916.8	0.402	0.227	0.371	0.485	749.9	0.396	0.329
40	0.386	0.321	905.1	0.394	0.246	0.372	0.526	724.1	0.397	0.366
50	0.398	0.358	890.1	0.406	0.277	0.358	0.540	720.9	0.382	0.384
60	0.391	0.377	881.4	0.399	0.293	0.355	0.565	706.5	0.377	0.408
70	0.350	0.357	891.9	0.356	0.278	0.353	0.585	692.4	0.375	0.428
80	0.352	0.377	886.5	0.357	0.297	0.361	0.613	674.2	0.382	0.453
90	0.355	0.403	872.2	0.361	0.316	0.352	0.620	676.3	0.372	0.464
100	0.347	0.410	869.5	0.352	0.323	0.340	0.624	674.6	0.359	0.470

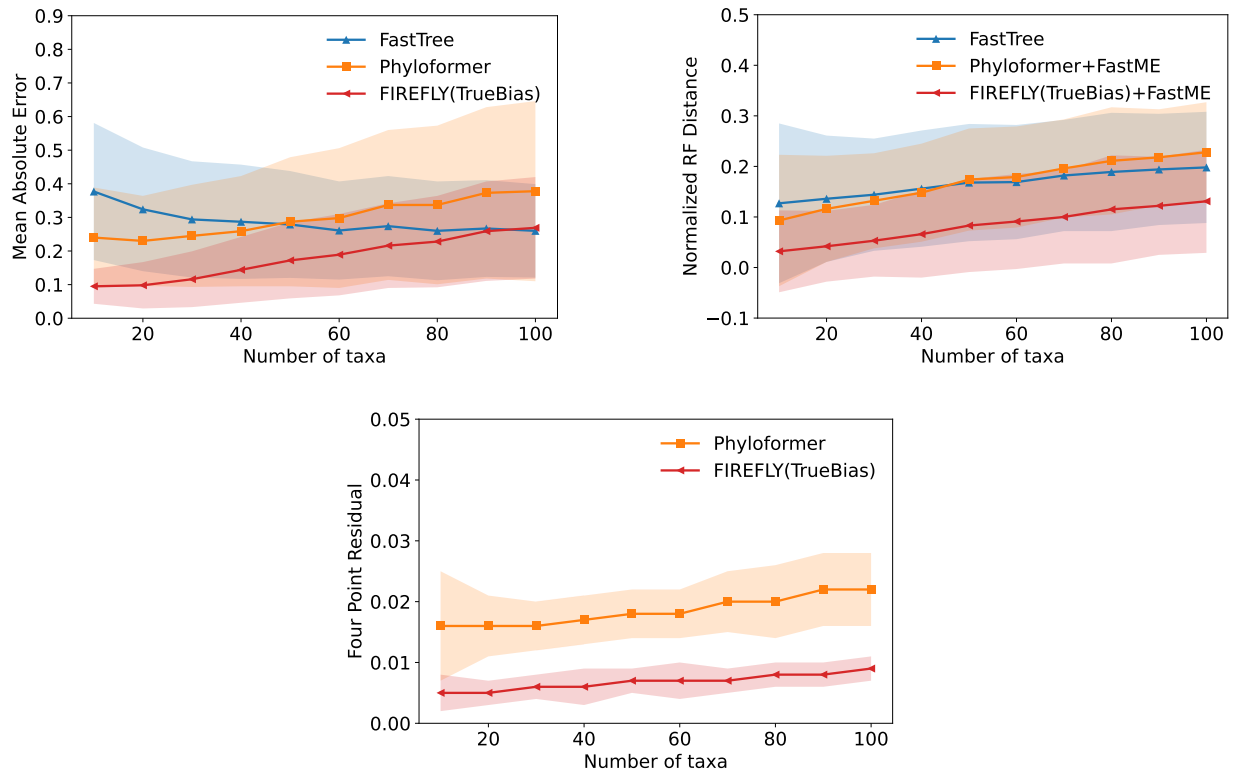
**Supplementary Table S5. Summary statistics of the CRW intronic RNA datasets.** The number of MSA sites (“Len.”), average normalized Hamming distance (“ANHD”) and the percentage of indels (“Gap.”) are reported for both reference and MAFFT estimated MSAs.

Dataset	Number of taxa	Reference MSA			MAFFT estimated MSA		
		Len.	ANHD	Gap.	Len.	ANHD	Gap.
IGIC2	32	4243	0.511	0.700	3530	0.353	0.640
IGID	21	5061	0.587	0.729	3032	0.522	0.548
IGIE	249	2751	0.430	0.839	2819	0.353	0.640
IGIE partitions	49.8	1569	0.432	0.715	1663	0.405	0.732



evaluation. This setting is not intended as a realistic protocol, since the ground-truth distances are unavailable in practice. Instead, it provides an upper bound and a mechanistic validation of our design choice – that injecting accurate phylogenetic distance into column attention can improve representation learning and downstream inference.

Figure S1 reports the resulting performance on pairwise distance prediction and downstream tree reconstruction, along with FastTree and Phyloformer as reference baselines. As shown in the figure, FIREFLY with the true bias consistently achieves lower mean absolute error (MAE) in pairwise distance prediction across all evaluated numbers of taxa. Relative to FastTree, the performance gap gradually decreases as the number of taxa increases, whereas the improvement over Phyloformer remains comparatively stable. These gains in distance estimation further translate into downstream phylogenetic inference. Trees reconstructed from the true-biased FIREFLY estimated distances exhibit a marked and stable improvement in topology accuracy, with normalized RF distances reduced by approximately 10% across all range of taxa. Overall, these results indicate that injecting the accurate phylogenetic relationship into column attention can effectively guide representation learning and improve both distance prediction and tree reconstruction, thereby validating the design principle underlying the proposed bias module.



**Supplementary Figure S1. Oracle experiment using ground-truth pairwise distances as column-attention bias.** Performance comparison on simulated true MSAs when the true pairwise distances are provided as the column-attention bias during both training and evaluation of FIREFLY. Results are shown for the same metrics and layout as in Fig. 2.

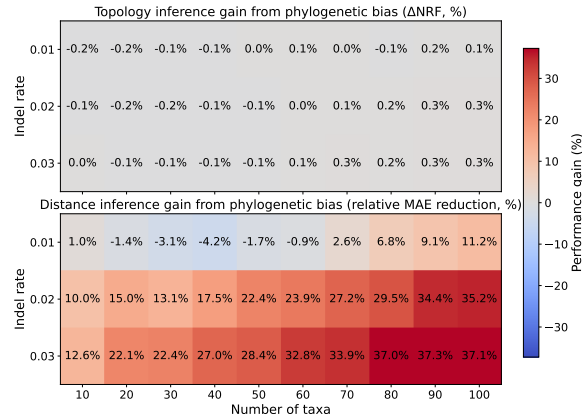
### 3.2 Learned Phylogenetic Bias: Ablation and Correlation Analysis

We conducted an ablation study on estimated MSAs to assess the contribution of the learnable phylogenetic bias module by comparing the full FIREFLY model with a variant without this module. Performance was evaluated on pairwise distance prediction and downstream phylogenetic tree inference. Supplementary

Figure S2 reports the performance gain of the full FIREFLY module relative to the variant without the phylogenetic bias. For tree topology inference (top panel), the effect of the phylogenetic bias is modest. Across all indel rates and taxa sizes, the absolute difference in normalized Robinson–Foulds (NRF) distance remains within 0.3%, indicating comparable topology inference performance with or without the bias module. In contrast, the impact on pairwise distance prediction is substantial (bottom panel). Incorporating the phylogenetic bias leads to pronounced reductions in MAE, particularly at moderate to high indel rates. The relative MAE reduction increases with both indel rate and number of taxa, exceeding 35% at high indel rates and large taxa sizes. Overall, the learnable phylogenetic bias has a stronger impact on distance estimation accuracy under alignment noise than on downstream topology inference.

We assessed the learned phylogenetic bias by computing Pearson correlations between the last-layer bias scores of FIREFLY and true evolutionary distances over estimated MSAs with 50 taxa (Supplementary Fig. S3). Approximately half of the attention heads display clear negative correlations, consistent with higher bias assigned to closely related taxa. In particular, head 7 shows strong and stable phylogenetic signals, with a median correlation of  $-0.75$ . Averaging correlations across heads yields a consistently negative and more concentrated distribution, indicating a robust encoding of global phylogenetic structure. Heads with weaker or positive correlations likely capture complementary, non-phylogenetic information.

The ablation study examines the role of the learnable phylogenetic bias in FIREFLY. Introducing this bias improves pairwise distance prediction accuracy, particularly on more challenging datasets. Correlation analysis further shows that the learned phylogenetic bias – both in specific attention heads and when averaged across heads – exhibits significant Pearson correlation with true evolutionary distances. This indicates that the bias guides the learned representations toward a phylogeny-based organization. In contrast, the phylogenetic bias has little effect on downstream distance-based tree reconstruction. One possible explanation is that topology inference is a discrete process that depends mainly on the relative ordering of pairwise distances rather than their absolute values. Once estimated distances fall within the same topological equivalence region, further reductions in distance error do not change the topology output by FastME. Overall, these results suggest that the primary contribution of the phylogenetic bias lies in improving distance prediction and global consistency, rather than resolving ambiguous phylogenetic relationships that would lead to alternative tree topologies.

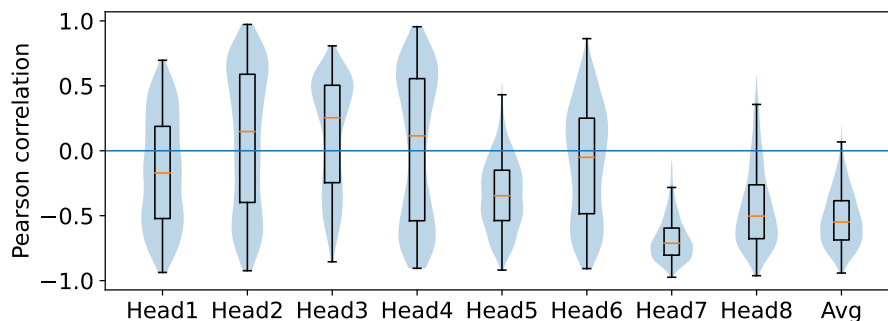


**Supplementary Figure S2. Ablation study of the learnable phylogenetic bias on estimated MSAs.** Heatmaps report the mean performance gain of the full FIREFLY model over a variant without the phylogenetic bias module. The upper panel shows the improvement in downstream topology inference, quantified by the difference in normalized Robinson–Foulds (NRF) distance. The lower panel shows the improvement in distance inference, measured as the relative reduction in mean absolute error (MAE), defined as  $(MAE_{nobias} - MAE_{full}) / MAE_{nobias}$ . Positive values indicate improved performance of FIREFLY relative to the no-bias variant. Results are shown across varying indel rates and numbers of taxa.

### 3.3 Correlation Between Learned Phylogenetic Bias and True Evolutionary Distances

To examine whether the learnable phylogenetic bias captures meaningful evolutionary information, we analyze its relationship with the ground-truth pairwise evolutionary distances. Specifically, we compute the Pearson correlation between the learned bias values and the true distances across attention heads. This analysis serves as an interpretability check, assessing whether the learnable bias module organizes representations in a manner consistent with underlying evolutionary relationships.

Figure S3 displays the distribution of Pearson correlations between the learned phylogenetic bias and true pairwise evolutionary distances across attention heads. Several heads exhibit clear negative correlations, most notably heads 7 ( $-0.75$ ) and 8 ( $-0.5$ ). This negative correlation indicates that taxa pairs with smaller evolutionary distances tend to receive larger attention bias weights. Averaging correlations across heads further stabilizes this relationship, suggesting that the phylogenetic bias organizes the learned representations in a globally consistent and evolution-informative manner. In addition, some attention heads show positive correlations with evolutionary distance. This diversity suggests that different heads might capture complementary, non-phylogenetic structure.



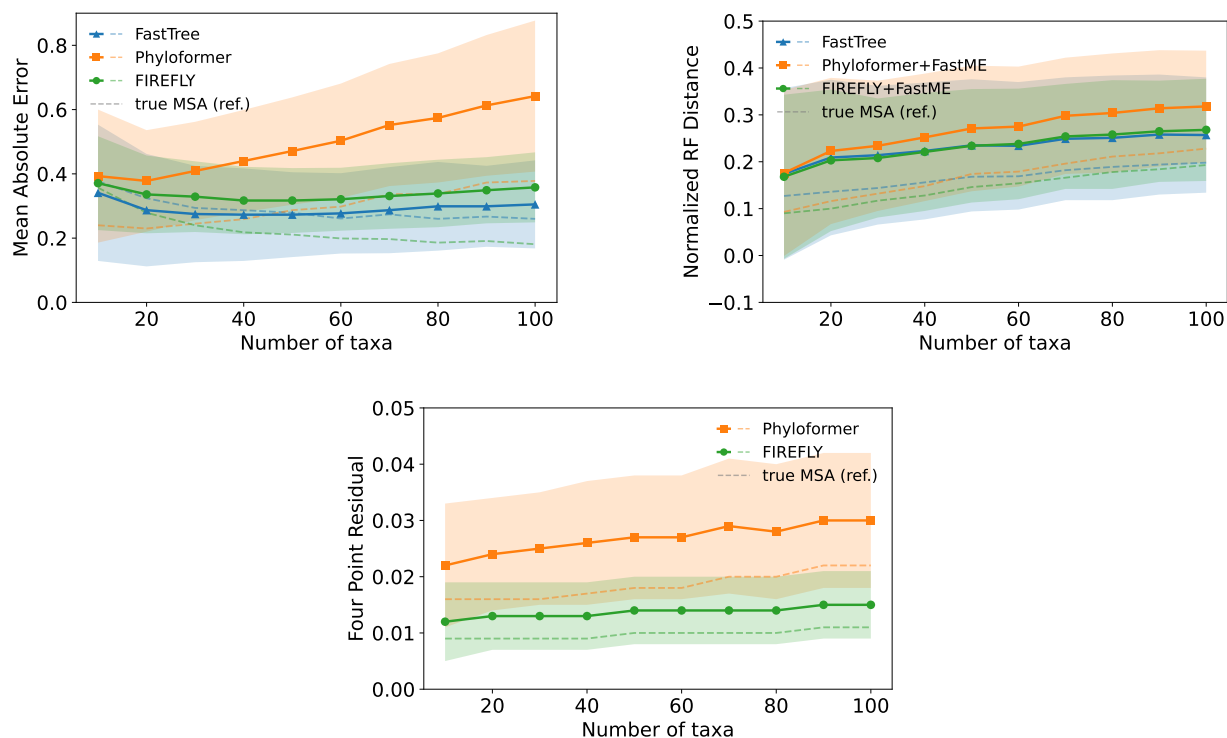
**Supplementary Figure S3. Pearson correlation between learned phylogenetic bias and evolutionary distance.** Violin plots show Pearson correlations between the learned phylogenetic bias from each attention head in the final layer of FIREFLY and true pairwise evolutionary distances across estimated MSAs with 50 taxa. The embedded box plots indicate the median, interquartile range, and whiskers.

### 3.4 Additional Experiments on Robustness to Alignment Methods

For both FIREFLY training and the primary evaluation of estimated MSAs, we use MAFFT as the default alignment method. To assess the generalization of FIREFLY to unseen alignment procedures, we additionally evaluate its performance on estimated MSAs generated using alternative methods, including ClustalW [3] and Clustal Omega [9, 10]. Figures S4 and S5 display the comparison results. We observe consistent performance trends across these alignment methods, comparable to those obtained on MAFFT-estimated MSAs. It suggests that FIREFLY exhibits consistent and robust behavior across different estimated MSAs generated by distinct alignment methods.

### 3.5 Additional Experiments with Classical Model-based Distance Corrections

For comparison purposes, we also utilized corrected distances under classical sequence evolution models to obtain a distance matrix and then perform distance-based phylogenetic estimation. The Tamura-Nei (TN93) model of nucleotide substitution [11] was used for the former; as in the rest of the study, FastME’s [4] implementation of neighbor-joining was used for the latter. The resulting pipeline was run on true MSAs. The accuracy of the corrected distance estimates, the topological error of the resulting distance-based tree estimates, and the four-point residual are shown in the top, middle, and bottom panels of Supplementary Figure S6, respectively. As classical distance corrections paired with neighbor-joining underperformed all

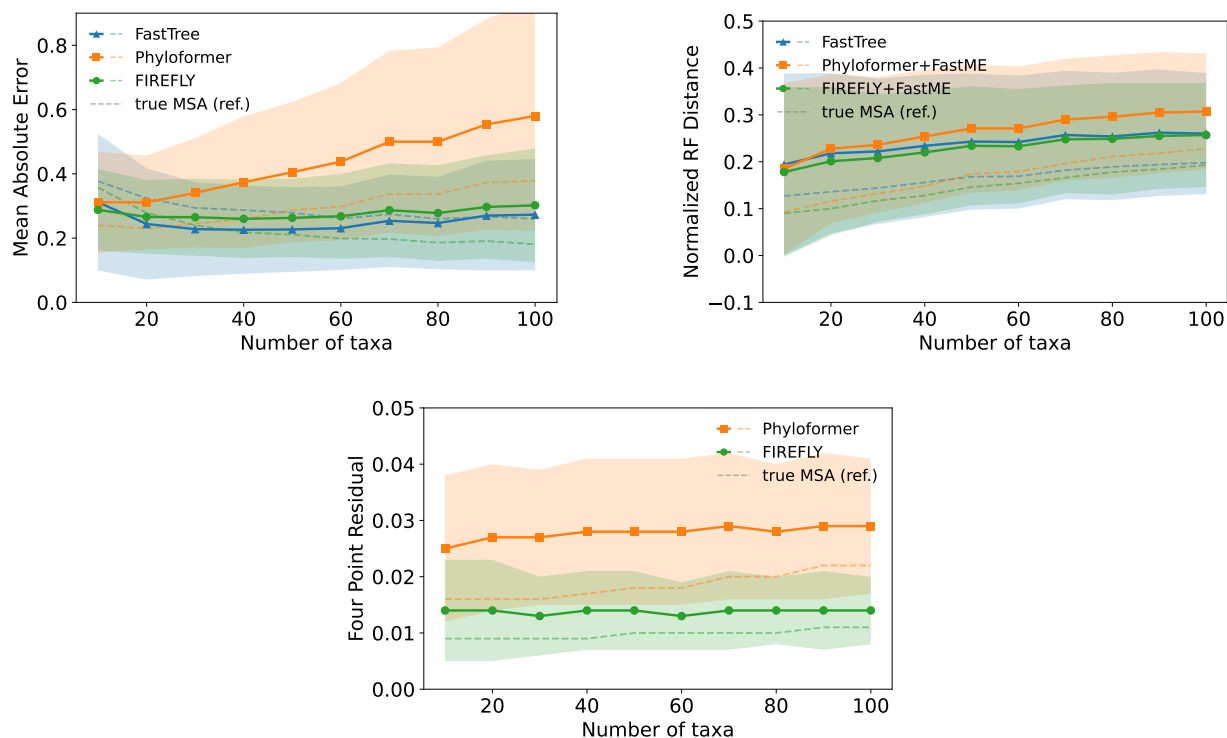


**Supplementary Figure S4. Performance comparison on Clustal Omega-estimated MSAs across different numbers of taxa.** Results are shown for the same metrics and layout as in Fig. 2.

other methods under study, the rest of the study focuses on the latter rather than the former. We attribute the performance discrepancy to two sources of model mis-specification. First, the nucleotide substitution model used for distance correction is less complex than that used for simulation. Second, none of the classical substitution models in this study account for sequence insertion and deletion processes, as is typical in the state of the art [12].

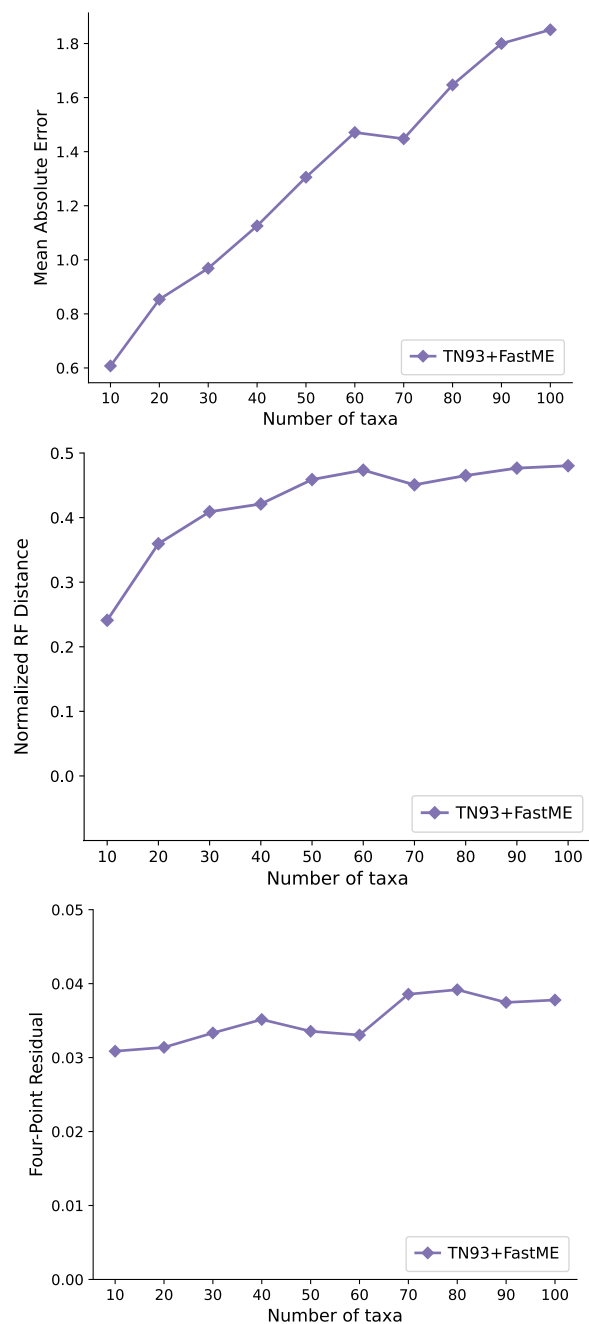
## References

1. Fletcher, W., Yang, Z.: INDELible: a flexible simulator of biological sequence evolution. *Molecular Biology and Evolution* **26**(8), 1879–1888 (2009)
2. Gao, M., Liu, K.J.: Statistical analysis of GC-biased gene conversion and recombination hotspots in eukaryotic genomes: a phylogenetic hidden Markov model-based approach. In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. pp. 1–24 (2021)
3. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al.: Clustal W and Clustal X version 2.0. *Bioinformatics* **23**(21), 2947–2948 (2007)
4. Lefort, V., Desper, R., Gascuel, O.: FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular Biology and Evolution* **32**(10), 2798–2800 (2015)
5. Liu, K., Warnow, T.J., Holder, M.T., Nelesen, S.M., Yu, J., Stamatakis, A.P., Linder, C.R.: SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology* **61**(1), 90 (2012)
6. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017)
7. Nesterenko, L., Blassel, L., Veber, P., Boussau, B., Jacob, L.: Phyloformer: fast, accurate, and versatile phylogenetic reconstruction with deep neural networks. *Molecular Biology and Evolution* **42**(4), msaf051 (2025)
8. Rodriguez, F., Oliver, J., Marin, A., Medina, J.: The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology* **142**, 485–501 (1990)
9. Sievers, F., Higgins, D.G.: Clustal Omega. *Current Protocols in Bioinformatics* **48**(1), 3–13 (2014)



**Supplementary Figure S5. Performance comparison on ClustalW-estimated MSAs across different numbers of taxa.** Results are shown for the same metrics and layout as in Fig. 2.

10. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al.: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**(1), 539 (2011)
11. Tamura, K., Nei, M.: Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Molecular biology and evolution* **10**(3), 512–526 (1993)
12. Warnow, T.: Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. *PLoS Currents* **4**, RRN1308 (2012)



**Supplementary Figure S6. Performance of distance-based tree estimation using corrected distances under a classical substitution model.** Distances were corrected under the Tamura-Nei (TN93) model [11], and distance-based tree estimation was performed using FastME's neighbor-joining implementation [4]. Results are shown for three evaluation assessments: mean absolute error (MAE) of pairwise distance estimates, normalized Robinson-Foulds (NRF) distance of estimated tree topologies versus ground truth, and four-point residual. Curves report average across replicates for model conditions with varying numbers of taxa – from 10 to 100.