nature portfolio

Peer Review File

PhyloTune: An Efficient Method to Accelerate Phylogenetic Updates Using a Pretrained DNA Language Model

Corresponding Author: Dr Wuqin Xu

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Deng et al. present a new method to improve the efficiency of phylogenetic sequence placement. It does so by first determining what clade a new sequence belongs to, then identifying parts of the sequences that are informative, and finally performing alignment and sequence reconstruction for the clade and the informative parts of the sequences. The first two steps are performed by a new method based on neural networks, while the alignment and phylogenetic reconstruction are performed by existing methods. The pipeline is tested on a database of plant sequences, and is found to perform better than other deep learning approaches, but not as well as standard sequence placement methods.

I found the manuscript clear overall, although I do have a few comments below on clarity in specific parts of the manuscript. I am not sure the title of the manuscript is clearly conveying what the method does. The term "Phylogenetic inference" to me really means the reconstruction of a phylogenetic tree from sequence data, but that's not what the method accelerates. The method accelerates a particular phylogenetic pipeline, where a reference taxonomy is available, some sequences/taxa need to be added to it, and only taxonomic identification seems to matter (e.g., the branch lengths do not seem to matter).

The method presented here, BERTPhylo, is more accurate than other machine learning approaches for taxonomic identification, but is less accurate than traditional methods. We don't know if the traditional methods are slower or faster than BERTPhylo.

The manuscript does not quantify how accurate the reconstructed phylogenys is. We are shown phylogenies obtained with high- or low-attention regions, but we don't know how the phylogeny updated by BERTPhylo compares to a phylogeny in which all sequences are analyzed at once. The authors could compute standard metrics like the Robinson-Foulds distance or the Kuhner-Felsenstein distance to compare the two trees, and they could also include standard approaches to place sequences in phylogenies.

The ability to identify regions of the sequences that are most informative seems one of the most interesting novelties in the manuscript. However, I had a few questions about the results. Are the high attention regions always the same across clades, in which case one could simply use them and ditch other parts of the sequences from the start? Are the high attention regions characterized by specific features: for instance do they evolve faster, do they have a higher or lower amino acid diversity per column? The authors did look at Fst and heterozygosity, but these statistics are mostly relevant for recent divergence, at the genus and species level.

I was wondering how representative is PlantSeqs of potential use cases of the method? In particular, I feel that Kraken is often used with microbial data. I think it would be useful to discuss how much we can extend the results obtained here on plants to other clades and other genes.

I was able to download the code and setup the conda environment. I did not try to reproduce the results in the paper with the command provided on github. Perhaps it would be useful to provide a smaller test set, possibly one that could be run on CPU, for users who do not have access to a good GPU but would still want to test the method.

Minor comments:

l66: "focusing on accelerating and parallelizing calculations have been developed, such as DupTree [21]" : I would not include DupTree in this list as it does not take sequence data as input, but gene trees containing events of gene duplication and loss.
l110: "ETE3 [43],": I find it strange to include ETE3 here, because it is a library for handling phylogenetic trees and, to a small extent, alignments.
l202: "This analysis is further corroborated by combining MMseqs2 with BERTPhylo, leading to additional enhancements across almost all metrics, notably a 5.04% increase in macro-F1 at the class rank.": it is not clear to me how MMseqs2 and BERTPhylo are combined, and so how this results in an improvement.
l243: "while in the topography constructed": I'm not sure what topography means here.
l273: "where heterozygosity serving as a measure": serves
l328: "Embyrophyta": Embryophyta
l350: "Drawing inspiration from the taxonomy we use should reflect shared ancestry (i.e. phylogeny)": weird sentence
l669: "an specific input sequence" : a specific
Fig. S6: "high-attention regions (right) versus low-attention regions (left).": it's the opposite.
Fig. S7: the legend does not explain what the grey stripes represent (in the legend of Fig. 7 it is explained). Fig. S7 would be easier to read if the x scales matched between the heatmap and the line plots, as in Fig. 7.

(Remarks on code availability)
I only checked that it installed properly, but did not test that it ran, as it requires a (big) GPU.

Reviewer #2

(Remarks to the Author)
The manuscript be Deng and co-authors presents a new method for integration of deep learning, in particular DNA language models (DLMs), into phylogeny reconstruction. This is a growing field of research that definitely warrants extensive researchers' attention. Despite the claim that is made in the abstract, introduction and a large part of the results section, I failed to find a detailed description of the procedure of integrating DLMs into the actual phylogeny reconstruction in the Methods section, see my comments below. On the other hand, the application of attention mechanisms to find important genomic markers caught my eye as a very elegant approach. Additionally, I have further critiques and suggestions that are also outlined below.

Major comments.

1. As mentioned, the Methods section does not contain a full description of the methods. From what I understood, the deep neural net is used merely to identify the correct taxon (at the level of genus, family, order or class), essentially solving a classification problem. To do the actual phylogenetic reconstruction, I imagine, one needs to apply some traditional sequence-based method, but the Methods section gives no details on this. If I am correct, the statement in the main part of the manuscript that BERTPhylo integrates DL into the phylogenetic reconstruction pipeline is overstated and misleading.

2. The manuscript lacks proper benchmarks. Essentially, all results are obtained for only one dataset from PlantSeqs. If the method is supposed to be generally applicable, as it is alluded to in the introduction, one needs a thorough benchmark on different phyla, as well as on simulated data where the ground truth is known.

3. Comparison with traditional methods for phylogenetic reconstruction (e.g. RAxML, BEAST, Mash) is missing. One needs to present the comparison of the quality of reconstruction, as well as of the needed compute resources.

4. Results of the taxonomic assignment with BERTPhylo (Table 1) are worse than with the traditional methods (in particular, Kraken2). Where is the advantage of using BERTPhylo? Additionally, in Table 1, I see that the genus-level assignment is predicted better that class-level, which is counter-intuitive to me – I would imagine that higher-level taxa are easier to predict. Could the authors comment on this?

5. How was split into the training/validation/test sets performed during the model training? What was the training goal?

Minor comments.

1. Lines 53-55: This makes an impression that people do classical alignment-based phylogenetic reconstruction on whole genomes, which is not the case anymore. Please consider rephrasing.

2. Lines 56-71: I find the distinction of methods into distance-based vs. character-based unusual. I am more used to classifying phylogenetic methods into alignment-based and alignment-free. All character-based methods, as defined by the authors, would fall into the alignment-based category, but distance-based methods are a mix of the both types.

3. Table 1: Elapsed times for all methods should be provided.

4. Lines 180-199: I strongly recommend using MCC (Matthew's correlation coefficient) as a measure of quality of classification, since this is the most unbiased measure.

5. Section 2.2: No details of application of DNABERT or BERTax or any of the traditional methods are given neither here nor

in the Methods section. This is particularly important for DNABERT, which is a DNA foundation model with many downstream tasks, so here I am completely at a loss how it was applied. For traditional tools, on the other hand, parameters may be extremely important: for example, for Kraken2, the choice of the database may play a crucial role in its performance.

6. Wording is strange at some places, e.g. I would recommend changing ‘equitably’ to ‘equally’ (line 259) or ‘heightened’ to ‘increased’ (line 298).

7. Section 4.4: What is sentence? Was some sliding window applied for longer sequences?

(Remarks on code availability)
The repository is there and it looks reasonable. There are installing and running instructions. However, the code itself is not commented. I have not tried to rerun it, but I think it will be possible, but not super straight-forward.

Reviewer #3

(Remarks to the Author)
Deng et al. proposed BERTPhylo, an approach for updating an existing phylogenetic backbone tree with new taxa using a BERT-based pretrained DNA language model. Their method identifies appropriate subtrees/clades to which the new sequences belong, and finds “valuable” regions within a sequence that they use for the update process. These contributions notably improve the speed of phylogenetic updates. While the paper presents an interesting and potentially valuable technique, I have several concerns regarding the evaluation of the proposed method:

1. The method involves identifying suitable clades for placing new sequences into an existing backbone tree. It is not clear how many new sequences are handled at a time. Does the method process sequences one by one, or can it handle multiple sequences simultaneously? More importantly, what happens when the input sequences belong to different taxonomic levels (e.g., genus, order, or family)? Can BERTPhylo assign multiple sequences to different clades? For example, in Figure 2, the three new sequences are all from the order Fabales. What if these sequences belong to different clades? An evaluation under such scenarios would strengthen the paper's contributions.

2. The authors compiled a dataset containing sequences of land plants to train and test their method. Given that phylogenomic datasets are often heterogeneous, a model trained on a specific dataset may not generalize well to others. It is essential to test BERTPhylo on unseen datasets from different lineages or taxonomic groups to assess its robustness and generalizability.

3. The novelty detection aspect of the model is innovative, but it may struggle with cases where sequence data is highly divergent from the training set. In such cases, the model could misclassify or fail to place taxa accurately within the phylogeny. Further exploration with examples of how the method performs with divergent sequences would be valuable.

4. The method identifies valuable regions in the sequences based on attention scores from the model. However, the decision-making process of the deep learning model, specifically how it identifies high-attention regions, may lack biological interpretability. This is a significant drawback for researchers seeking to understand the biological basis behind the model’s predictions. Can the authors provide more evidence or case studies demonstrating the biological relevance of these high-attention regions?

5. The authors mention that they divide sequences into three regions and then identify the high-attention regions. Does this imply that the attention regions are contiguous, corresponding to one of these three predefined divisions? If so, why must they be contiguous?

6. BERTPhylo depends heavily on the pretrained DNABERT model, and as a result, any limitations of DNABERT are likely to propagate into BERTPhylo. For instance, pretrained models like DNABERT typically truncate input sequences to a maximum length (e.g., 1024 bp). Is this also the case for BERTPhylo? Can the proposed method handle arbitrarily long sequences, and if not, how does it manage longer sequences?

7. May BERTPhylo’s reliance on existing taxonomic classification limit its ability to identify new evolutionary relationships? It would be useful for the authors to discuss the model's potential to discover clades or evolutionary patterns that do not fit within the taxonomic classification framework used in BERTPhylo.

8. Although BERTPhylo outperforms other deep learning based methods, it generally falls short when compared to traditional approaches like MMseq2 and BLAST. This is understandable, as methods like BLAST are more computationally intensive. However, the model conditions under which the comparisons were made (Tables S1 and S2) seem to be very easy as the precision, recall and F1-scores are close to 100%. I believe the evaluation would benefit from the inclusion of more challenging model conditions.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I have read the new version of the manuscript and the reply to our comments. The authors have adressed my comments satisfactorily. I only have some minor wording comments below. Congratulations to the authors for their work.

l24: "accelerate phylogenetic updates using a pretrained DNA language model": I don't think "phylogenetic update" is very clear; perhaps seomething like "add taxa to an existing phylogenetic tree using a pretrained language model"
l73: Another approach that may be worth citing in this paragraph is this: https://doi.org/10.1101/2024.06.17.599404
l194: "BERTPhylo process multiple sequences": processes
l263: "had a closer or equal RF distances to the full-sequence trees than those constructed using low-attention regions": this is not very well said: a distance is not "close" or "equal" to a tree, but it can be small.
l295: "the two nulcear markers": nuclear

(Remarks on code availability)
I have opened the google collab notebook, and the github repository, and have skimmed their contents. I have not seen anything problematic. I have not tried to execute any code.

Reviewer #2

(Remarks to the Author)
The authors have significantly improved the manuscript in response to my comments. In fact they have satisfactorily addressed all of them but one: No benchmarking on simulated data has been performed. I consider this a valuable exercise, since in this case the ground truth is known, and also the tree topology and clade distribution can be modelled freely.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I would like to thank the authors for their effort in revising the manuscript. My concerns were mostly about the generalizability of the pre-trained model and the evaluation/benchmarking of the proposed methods. Although the authors improved the manuscript with additional experiments, I still have some questions and concerns.

1. Regarding the evaluation of your model's generalizability, I was expecting to see the performance of the already trained model on unseen data. That means, the model trained on the PlantSeqs dataset should be evaluated on the new microbial data. Following this, you could fine-tune your model using the new microbial data and assess the performance again. This approach would allow you to evaluate how your pre-trained model performs on unseen data with and without fine-tuning. However, it is not clear whether you fine-tuned the model based on the microbial data or trained it from scratch using the microbial data.

The authors mentioned that they used two HLPs to adapt to the new dataset. How does your model perform with only one HLP layer (as you used for the PlantSeqs dataset)? This suggests that the model was not fine-tuned based on the microbial data; instead, a slightly different architecture with two HLP layers was trained for the microbial dataset. This is concerning, as it indicates that the architecture may lack the expected generalizability of a pre-trained model. Moreover, how would a user determine whether such adaptations in HLP layers are necessary for a new dataset? Pre-trained models are usually fine-tuned with new datasets while maintaining the same architecture.

I understand the challenge, as I mentioned in my original comment, that phylogenomic dataset are heterogeneous. But a systematic assessment of your proposed model in this regard is essential.

2. As for benchmarking, thank you for your clarification regarding my original Comment 8. However, I believe it is important to make direct comparisons to traditional phylogenetic inference methods (like RAxML) with an entire set of sequences. You have provided comparisons between the performance with full-length sequences and high/low-attention regions (Fig 6). However, it is equally important to compare the results when considering all the sequences together and estimating trees using a traditional approach (e.g., RAxML) versus your approach of updating the existing tree with a new sequence. While the updating approach makes the overall pipeline faster, readers need to understand its accuracy compared to traditional approaches when applied to the entire set of sequences.

3. I appreciate that the authors have changed the title of the manuscript by indicating that it “updates” an existing tree. However, I feel that the name “BERTPhylo” may not be appropriate and could be an overstatement. The authors are using DNABERT to identify the subtree to be updated and to identify high-attention regions in sequences. That does not make this method a BERT-like method. However, the name “BERTPhylo” suggests a BERT-like (eg., DNABERT, ProtBERT) model

for phylogenetic tree estimation, which is not the case for your method.

Minor comments:

In section 4.2, the authors added a new sentence, “PlantSeqs dataset is fine-tuned based on DNABERT.” But a dataset cannot be fine-tuned. Did you mean to say that DNABERT was fine-tuned based on PlantSeqs dataset?

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
The authors have satisfactorily addressed my comments.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I would like to thank the authors for their thorough revisions and for addressing my previous comments. The manuscript has been improved, and the new analyses provide important insights. Regarding the newly added Figure 2a, I have two suggestions to further strengthen the interpretation of the results.

Currently, Figure 2a shows the RF distances between the trees inferred using the proposed method (for both high-attention regions and full-length sequences) and the tree inferred from the complete set of sequences using RAxML. While this comparison is useful, a more informative approach would be to evaluate the methods directly against the true species tree, since this is a simulated dataset with a known ground truth. Otherwise, it is difficult to interpret the accuracy of the proposed method compared to RAxML. Specifically, I recommend revising Figure 2a to include the following: 1) RF distances between the trees obtained by your approach (high attention region) and the true tree, 2) RF distances between the trees obtained by your approach (full length) and the true tree, and 3) RF distances between the trees obtained from all sequences using RAxML and the true tree (this will be just a horizontal line as in Fig 2b since varying the number of sequences is not relevant here).
My second suggestion is to briefly discuss the RF distances of the high attention region approach. The authors observed that as the number of sequences increases, the RF distance between their method and RAxML increases (as I expected). But they reported the RF distances (0.007, 0.023 and 0.014) for only full-length sequences. For high attention region approach, the RF distances are even higher (more than 25% for 100 sequences). This is what I anticipated, and the authors should report and discuss this (but with respect to true trees as I suggested above).

(Remarks on code availability)

Rebuttal of BERTPhylo: An Efficient Method to Accelerate Phylogenetic Inference Using a pretrained DNA Language Model

Dear reviewers,

We sincerely appreciate your insightful and constructive feedback on our manuscript. We have carefully considered your suggestions and conducted additional experiments using the bacterial dataset from the *Bordetella* genus to further demonstrate the applicability of BERTPhylo across diverse datasets beyond plants. Additionally, we have revised the manuscript accordingly, with all changes highlighted in red. We hope the following point-to-point responses address your concerns and further clarify our contributions.

Additional Experiments on the *Bordetella* Genus

BERTPhylo can leverage any pre-trained DNA language model to accelerate updates to a given phylogenetic tree. To evaluate the generalizability of BERTPhylo to non-plant clades and genes, we extended our experiments to microbial data, specifically focusing on the *Bordetella* genus. The dataset for this experiment was derived from the Nature Communications article, "A comprehensive resource for *Bordetella* genomic epidemiology and biodiversity studies" [1]. From this dataset, we selected 140 core genes with the highest coverage in all *Bordetella* species as the basis for our analysis. Below is a summary of the setup and findings:

1 Experiment Setup

We selected sequences from a subset of species to construct the initial phylogenetic tree, with these species considered as "known" and the others as "unknown." For new sequences (from both known and unknown species), BERTPhylo was used to identify the smallest taxonomic units and extract high-attention regions from the branching sequences. Existing tools were then applied to rebuild the clades based on these high-attention regions.

clade	species	#training	#validation	#test
clade1	B._bronchiseptica_I-1	151	19	50
	B._bronchiseptica_I-4	165	20	50
	B._bronchiseptica_II	465	56	50
	B._pertussis	116	14	50
	unknown (1)⋆	-	-	50
clade2	B._avium	96	11	50
	B._hinzii	129	16	50
	B._trematum	121	14	50
	Genomic_species_1	119	14	50
	Genomic_species_5	107	13	50
	unknown (2)⋆	-	-	50
clade3	B._bronchialis	112	13	50
	Genomic_species_2	116	14	50
	Genomic_species_7	112	13	50
	Genomic_species_9	121	14	50
	unknown (1)⋆	-	-	50
outgroup	unknown (4)⋆	-	-	50
		1,930	**231**	**850**

Updated Table S4. The taxonomic structure and statistics of *Bordetella* dataset used in experiment. unknown (X)⋆: X denotes the number of unknown species covered by the test sequences.

1.1 Taxonomic Structure and Training/Test Set Splitting

Out of the 39 species in the dataset, 13 were designated as "known species" for training and testing, while the remaining species were treated as "unknown" for testing. Following the phylogenetic hierarchy outlined in [1], we divided the 39 species into three clades and one outgroup. Each clade included both known and unknown species, while the outgroup contained only unknown species. For the known species, 50 sequences per species were randomly selected for the test set. The remaining sequences were split into training and validation sets, with 45% of sequences for training and 5% for validation, sampled from individual isolates. For the unknown species, 50 sequences were randomly selected from each clade and from the outgroup, resulting in a test set consisting of 850 sequences, including fully known, partially known (species known but clade unknown), and completely unknown sequences. The detailed taxonomic structure and dataset statistics are summarized in Table S4.

1.2 Model Architecture and Training Configuration

BERTPhylo was adapted for the *Bordetella* dataset using DNABERT-S as the pre-trained backbone. We appended a two-layer hierarchical linear probes (HLPs) to DNABERT-S [2] for identification of the smallest taxonomic unit. One layer performed

novelty detection and classification at the clade level, while the other layer operates at the species level. DNABERT-S overcomes the input length limitation of the original DNABERT model [3], enabling it to handle arbitrarily long sequences. We fine-tuned the model using the two-step training protocol described in the Section 4.3:

- **Phase 1**: The BERT backbone was frozen, and only the HLPs were trained for 10 epochs.
- **Phase 2**: All layers were unfrozen, and the entire model was fine-tuned for an additional 5 epochs.

The model was trained with a weighted cross-entropy loss to account for class imbalances.

2 Experimental Results

2.1 Identification of the Smallest Taxonomic Unit

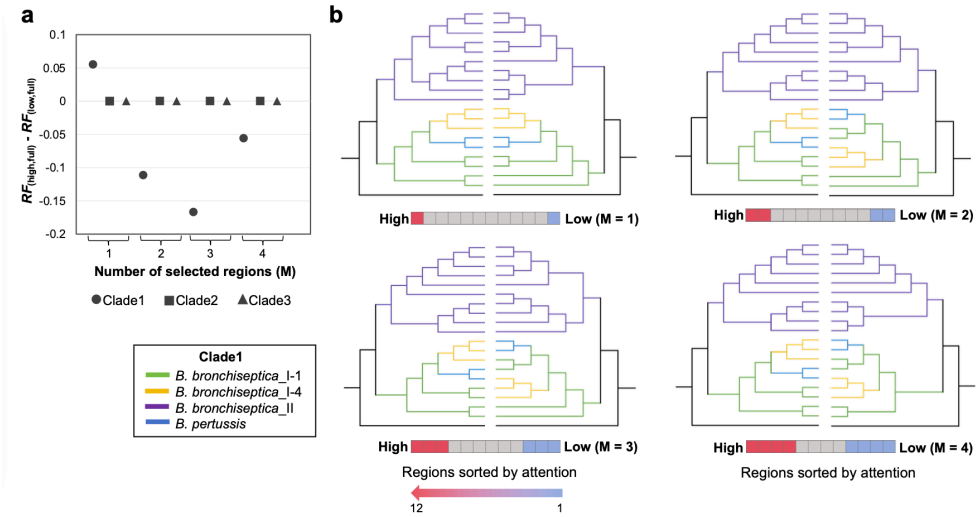
Identifying the smallest taxonomic unit of a new sequence involves two key tasks: novelty detection and taxonomic classification. Table S5 compares the performance of BERTPhylo and MMseqs2 on these two tasks. For a fair comparison, we used the *Bordetella* training set as the reference database for MMseqs2. BERTPhylo achieved an AUPR value of over 95% at both clade and species levels, demonstrating its superior ability to detect sequences from unknown taxa. In contrast, MMseqs is limited to identifying the most similar sequences and cannot ensure consistency across all taxonomic levels between the identified and query sequences, which significantly restricts its performance on test sets containing unknown taxa compared to BERTPhylo.

Method	**Rank**	**Novelty Detection** (↑)		**Taxonomic Classification** (↑)				
		AUROC	**AUPR**	**ACC**	**MCC**	**F1**	**P**	**R**
BERTPhylo	clade	89.73	99.28	88.38	83.30	87.49	89.42	87.84
	species	87.10	95.08	60.31	61.02	54.51	55.07	60.31
MMseqs2	clade	-	-	85.29	79.25	65.77	85.83	67.98
	species	-	-	47.18	45.44	49.81	48.92	57.29

Updated Table S5. The experimental results (%) of identifying the smallest taxonomic unit on the *Bordetella* test set. ACC, MCC, F1, P, R represent accuracy, Matthews correlation coefficient, precision and recall, respectively.

2.2 Phylogenetic Tree Construction Using High- vs. Low-Attention Regions

We next compare the phylogenies constructed from high- and low-attention regions. In this experiment, we divided the sequences into 12 regions, ranking each region by the average attention score from the final layer of the BERT model. The top M regions with the highest and lowest attention scores were selected to form the high-



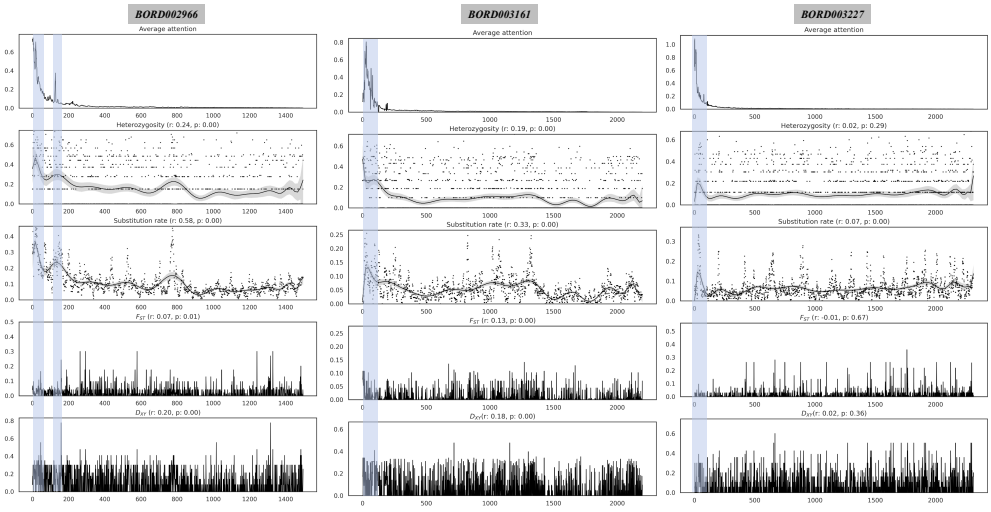
Updated Figure 8. Phylogenetic trees constructed from high-attention regions show greater similarity to full-length sequence trees compared to those built from low-attention regions. (**a**) Difference in Robinson-Foulds (RF) distance between trees constructed from high-attention and full-length sequences (RF(high, full)) versus those constructed from low-attention and full-length sequences (RF(low, full)). High- and low-attention regions are defined by dividing the sequence from the *Bordetella* dataset into 12 segments and selecting the top M $(1, 2, 3,$ and $4)$ regions with the highest and lowest attention scores, respectively. (**b**) Comparison of phylogenetic tree topologies of clade1 constructed from high- and low-attention regions at different M values.

and low-attention regions, respectively. We quantified the accuracy of the phylogenies constructed from high- and low-attention regions formed with various M values $(1, 2, 3,$ and $4)$ by calculating the Robinson-Foulds (RF) distance between these trees and the tree constructed using full-length sequences. As illustrated in Figure 8, the differences in RF distances are non-positive in most cases, indicating that trees constructed with high-attention regions are more similar to the full-length trees than those built with low-attention regions. This suggests that high-attention regions tend to provide more informative loci for tree reconstruction. In Clades 2 and 3, however, both high-attention and low-attention regions produced trees identical in topology to the full-length tree. This is likely due to the simplicity of these clades' tree structures, where even arbitrarily chosen regions contain sufficient phylogenetic signals to construct accurate topologies.

2.3 Analysis of Attention Regions

To further investigate the utility of attention weights for phylogenetic tree construction, we compared the average attention scores with heterozygosity, substitution rate, fixation index (F_{ST}), and absolute divergence (D_{XY}) values (Figure S11). Consistent

with findings from the PlantSeqs dataset, regions with higher attention scores tended to exhibit elevated levels of heterozygosity, substitution rate, F_{ST}, and D_{XY}. Pearson correlation analysis between average attention and these genetic metrics revealed positive correlations for most cases, with the highest correlation observed between average attention and substitution rate. We speculate that the high substitution rate may make these regions valuable for addressing systematic and evolutionary issues. Moreover, high-attention regions in both the *Bordetella* and PlantSeqs datasets were concentrated in the anterior, with those in *Bordetella* being closer to the 5' end. This may be related to the finding that the 5' region contains more variable sites, as suggested in [4].



Updated Figure S11. Average attention, heterozygosity, substitution rate, fixation index (F_{ST}), and absolute divergence (D_{XY}) curves for the *BORD002966*, *BORD003161*, and *BORD003227* genes in *Bordetella* dataset using BERTPhylo. The blue shaded area denotes the peak region of attention.

Although all genetic metric showed positive average correlation coefficients, some genes had p-value greater than 0.1 (Supplementary Table S6). This observation may stem from the *Bordetella* dataset, which contains a significantly larger number of genes (approximately 16 times more than PlantSeqs), but with fewer sequences per gene (fewer than 20 sequences per gene). This limitation likely affects the model's generalization capacity. Addressing the challenge of effectively applying pretrained LLMs to datasets with limited sample sizes but a high number of genes will be an important avenue for future research.

For Reviewer #1,

We would like to express our sincere gratitude to Reviewer #1 for taking the time to carefully read our manuscript and for providing such insightful and constructive feedback. We greatly appreciate the recognition of the novelty and potential impact of our method, particularly the identification of the most informative regions of sequences, which was highlighted as a key innovation. Below, we provide point-by-point responses to the reviewer's comments, which we hope will address the concerns raised and further clarify the contributions of our approach.

Comment 1

Deng et al. present a new method to improve the efficiency of phylogenetic sequence placement. It does so by first determining what clade a new sequence belongs to, then identifying parts of the sequences that are informative, and finally performing alignment and sequence reconstruction for the clade and the informative parts of the sequences. The first two steps are performed by a new method based on neural networks, while the alignment and phylogenetic reconstruction are performed by existing methods. The pipeline is tested on a database of plant sequences, and is found to perform better than other deep learning approaches, but not as well as standard sequence placement methods.

Response 1

Thank you for highlighting this important point.

As you correctly noted, the primary contributions of our method are the identification of the clade to which a new sequence belongs and the pinpointing of the most informative regions of sequences. **These capabilities are not present in traditional methods, such as MMseqs2 or Kraken2.** While these tools can identify the most similar sequence, they cannot confirm that the identified sequence and query sequence are consistent across all taxonomic levels. Table 1 in the original manuscript only reports the performance for sequences belonging to known taxa in PlantSeqs. **When sequences from unknown taxa are included, traditional methods exhibit significant performance drops, as shown in the Updated Table S1.**

To better highlight the strengths of BERTPhylo and the limitations of traditional methods, we have revised the manuscript (lines 137-152) to clarify the distinction between identifying the smallest taxonomic unit and traditional taxonomic classification. Additionally, we improved the description of smallest taxonomic unit results (lines 215-249) to reduce confusion between the two.

Comment 2

I found the manuscript clear overall, although I do have a few comments below on clarity in specific parts of the manuscript. I am not sure the title of the manuscript is clearly conveying what the method does. The term "Phylogenetic inference" to me really means the reconstruction of a phylogenetic tree from sequence data, but that's not what the method accelerates. The method accelerates a particular phylogenetic

	Macro Precision				Macro Recall				Macro F1			
Method	Class	Order	Family	Genus	Class	Order	Family	Genus	Class	Order	Family	Genus
MMseqs2	89.58	93.42	95.95	99.22	93.81	94.75	97.43	99.20	89.77	93.54	96.41	99.20
BLAST	89.16	93.55	95.96	99.24	90.40	94.09	97.16	99.20	87.60	93.24	96.25	99.21
Kraken2	93.79	94.71	95.66	99.13	89.01	92.47	95.87	97.78	90.17	93.10	95.44	98.42

(a) Performance on test sequences belonging to known taxa.

	Macro Precision				Macro Recall				Macro F1			
Method	Class	Order	Family	Genus	Class	Order	Family	Genus	Class	Order	Family	Genus
MMseqs2	78.92	88.06	84.81	74.80	78.28	92.17	95.98	98.24	69.40	87.58	87.92	82.46
BLAST	79.46	88.62	85.60	75.23	78.49	92.33	95.95	98.26	69.92	88.13	88.50	82.85
Kraken2	82.61	88.44	85.40	75.92	77.87	91.43	94.96	96.98	75.82	89.34	88.74	83.08

(b) Performance on all test sequences, including those from unknown taxa.

Updated Table S1. Performance comparison of traditional classification methods on the PlantSeqs test set. The training set of PlantSeqs serves as the reference database, and all metrics are calculated using macro-averaging. Results clearly highlight that classification performance significantly declines when test set contains unknown taxa.

pipeline, where a reference taxonomy is available, some sequences/taxa need to be added to it, and only taxonomic identification seems to matter (e.g., the branch lengths do not seem to matter).

Response 2

Thank you for your thoughtful feedback and for pointing out potential confusion regarding the manuscript title.

As you pointed out, BERTPhylo focuses on the process of updating an existing phylogenetic tree with newly collected sequences, rather than reconstructing a tree from scratch. To better reflect the focus and objectives of our method, we have revised the title to: "**BERTPhylo: An Efficient Method to Accelerate Phylogenetic Updates Using a Pretrained DNA Language Model**." We hope this updated title more clearly convey the scope and contributions of BERTPhylo.

Additionally, our current version of BERTPhylo primarily leverages taxonomic information to fine-tune the pretrained DNA model for the identification of the smallest taxonomic units and the extraction of high-interest regions. However, as we mentioned in discussion section, how to better utilize the modeling prowess of large language models (LLMs) to enhance phylogenetic tree construction is worth further investigation. Incorporating more information, such as branch lengths, to fine-tune the model is a direction worth exploring.

Comment 3

The method presented here, BERTPhylo, is more accurate than other machine learning approaches for taxonomic identification, but is less accurate than traditional methods. We don't know if the traditional methods are slower or faster than BERTPhylo.

Response 3

Thank you for raising this important point.

As discussed in Response 1, traditional methods cannot identify the smallest taxonomic unit as BERTPhylo does. These methods fail to ensure consistency across all taxonomic levels between the identified and query sequences, leading to a significant performance degradation when sequences from unknown taxa are included (Updated Table S1). Supplementary experiments on the *Bordetella* dataset demonstrated similar outcomes, where traditional methods underperformed compared to BERTPhylo in identifying the smallest taxonomic unit (Updated Table S5).

Regarding computational efficiency, BERTPhylo takes an average inference time of about 0.013 seconds per sequence using 8 A100 GPUs, compared to MMseqs2, which requires around 0.005 seconds per sequence. **Notably, BERTPhylo's primary contribution lies in accelerating the process of updating phylogenetic trees**. For instance, the time required for aligning and updating the subtree of the order Malpighiales takes only 0.33% (240 seconds) of the total time required by the standard pipeline (**Figure 3** in the manuscript). Furthermore, BERTPhylo's inference time is significantly lower than the time required for tree construction.

We hope this response clarifies the computational aspects, and we have revised the manuscript to highlight these points (lines 210-212).

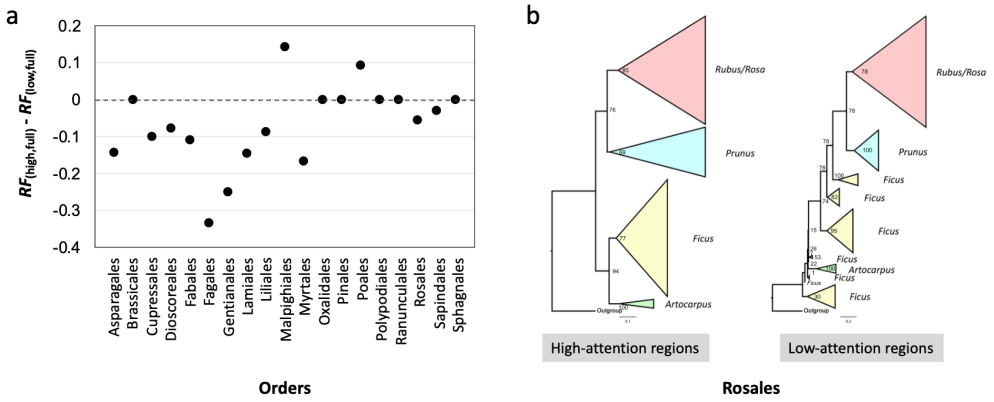
Comment 4

The manuscript does not quantify how accurate the reconstructed phylogenys is. We are shown phylogenies obtained with high- or low-attention regions, but we don't know how the phylogeny updated by BERTPhylo compares to a phylogeny in which all sequences are analyzed at once. The authors could compute standard metrics like the Robinson-Foulds distance or the Kuhner-Felsenstein distance to compare the two trees, and they could also include standard approaches to place sequences in phylogenies.

Response 4

Thank you for your valuable suggestion.

In response, we have quantified the accuracy of phylogenies reconstructed using high- and low-attention regions by calculating the Robinson-Foulds (RF) distance. We compared the RF distances between trees constructed using high- and low-attention regions and those built using the full sequence in the Plantseqs datase, with the comparison expressed as RF(high, full) - RF(low, full). Out of the 20 orders in the PlantSeqs dsataset, we report comparisons for 19, as the order Porellales with only one individual. **As shown in the Updated Figure 6a, nearly** 90% **of the trees constructed using high-attention regions had a closer or equal RF distances to the full-sequence trees than those constructed using low-attention regions.** This suggests that when there is a need to improve tree-building efficiency due to the length of input sequences, using high-attention regions as a substitute for full sequences is an excellent option.



Updated Figure 6. Phylogenetic trees constructed using high-attention regions perform better. (**a**) Difference in Robinson-Foulds (RF) distance between trees constructed from high-attention and full-length sequences (RF(high, full)) versus those constructed from low-attention and full-length sequences (RF(low, full)). (**b**) Phylogenetic trees constructed for the angiosperm order Rosales using high-attention (left) and low-attention regions (right) of sequences extracted by BERTPhylo, respectively.

Moreover, while the RF distance differences between the order Malpighiales and Poales suggest that trees constructed using low-attention regions are more similar to those based on full-length sequences, trees built with high-attention regions exhibit better clustering of individuals within the same genus (Supplementary Figure S6). Notably, regardless of whether high-attention, low-attention, or full-length regions were used, the resulting topologies of Malpighiales and Poales differed significantly from those reported in prior studies [5, 6]. This highlights the inherent gap between gene trees constructed from different genes—or even different segments of the same gene—and the species tree [7]. It also suggests that comparison with full-length sequences is only a reference to assess whether high-attention regions can replace them and does not fully reflect topology accuracy. For taxonomic groups with unstable topologies, careful consideration is needed in selecting appropriate markers or genetic regions to ensure reliable phylogenetic reconstruction [8].

We appreciate your suggestion and have incorporated this additional analysis into the manuscript (lines 258-283).

Comment 5

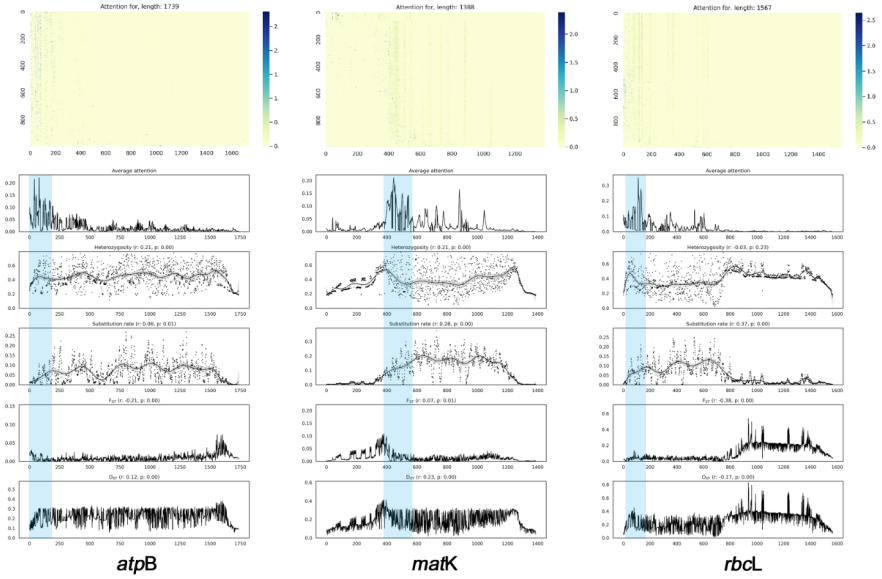
The ability to identify regions of the sequences that are most informative seems one of the most interesting novelties in the manuscript. However, I had a few questions about the results. Are the high attention regions always the same across clades, in which case one could simply use them and ditch other parts of the sequences from the start?

Response 5

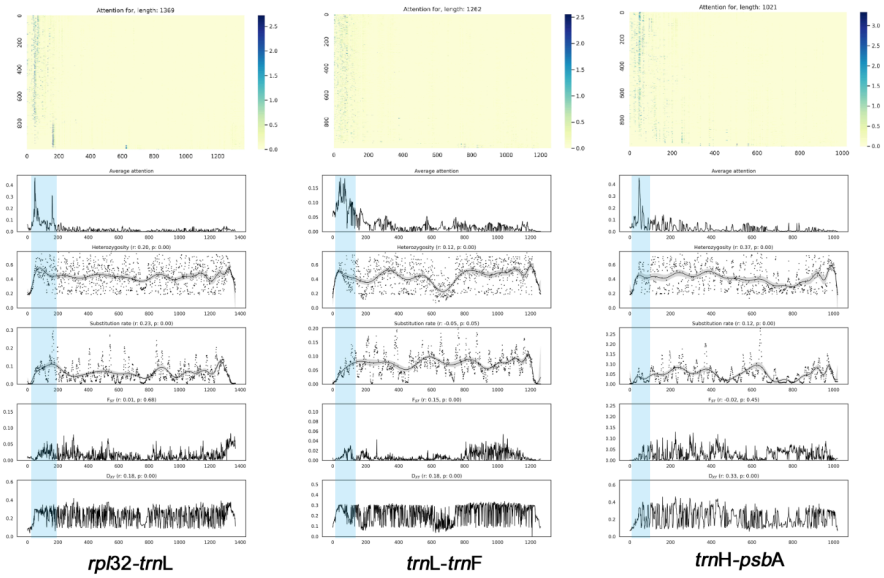
Thank you for this insightful question.

As shown in Updated Figures S7, S8, S9, while high-attention regions are typically concentrated in the 5' regions of the sequences, there are some differences across different molecular markers. For instance, the *mat*K and 5S rRNA markers exhibit high-attention regions concentrated in the middle of the sequence. Our supplementary experiments on the *Bordetella* dataset further revealed that, while high-attention regions tend to cluster at the 5' end, in some cases the most highly attended regions are located elsewhere in the sequence (Figure S11). For example, in the three clades we tested, the most highly attended regions are located within the first 1/12 of the sequence, but when considering the top three regions, 3.6%, 3.6%, and 18.7% of genes from each clade do not fall within the top 3/12 of the sequence. Specifically, the region with the third highest attention for *BORD003783* gene is located around 10/12 of the sequence.

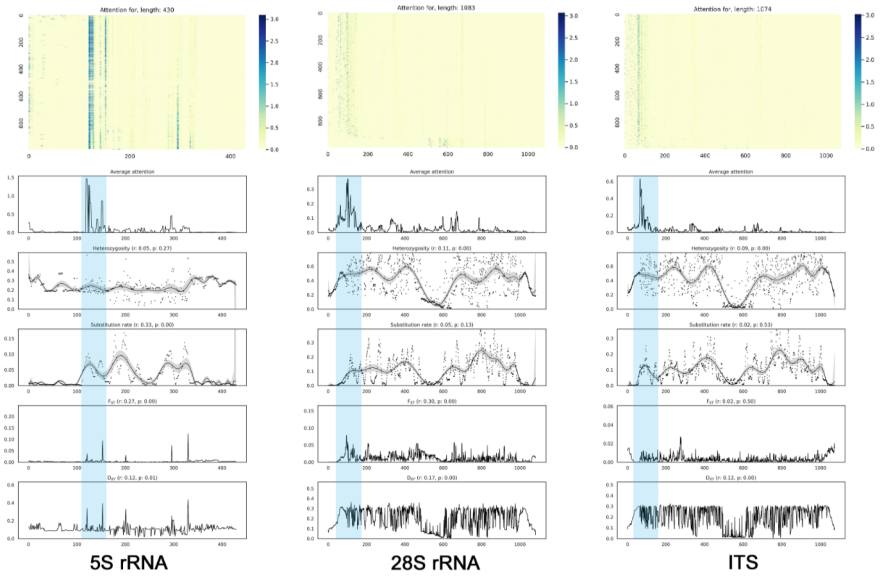
These findings indicates that while there are general patterns, high-attention regions are not identical across all clades and genes. Therefore, it would not be advisable to discard other parts of the sequence entirely from the outset.



Updated Figure S7. Attention heatmaps as well as average attention, heterozygosity, substitution rate, F_{ST}, and D_{XY} curves of three coding genes (*atp*B, *mat*K and *rbc*L). The blue shaded area denotes the peak region of attention.



Updated Figure S8. Attention heatmaps, average attention, heterozygosity, substitution rate, F_{ST}, and D_{XY} curves of three intergenic spacer regions (*rpl*32-*trn*L, *trn*L-*trn*F and *trn*H-*psb*A). The blue shaded area denotes the peak region of attention.



Updated Figure S9. Attention heatmaps, average attention, heterozygosity, substitution rate, F_{ST}, and D_{XY} curves of 5S rRNA, 28S rRNA and ITS. The blue shaded area denotes the peak region of attention.

Comment 6

Are the high attention regions characterized by specific features: for instance do they evolve faster, do they have a higher or lower amino acid diversity per column? The authors did look at F_{ST} and heterozygosity, but these statistics are mostly relevant for recent divergence, at the genus and species level.

Response 6

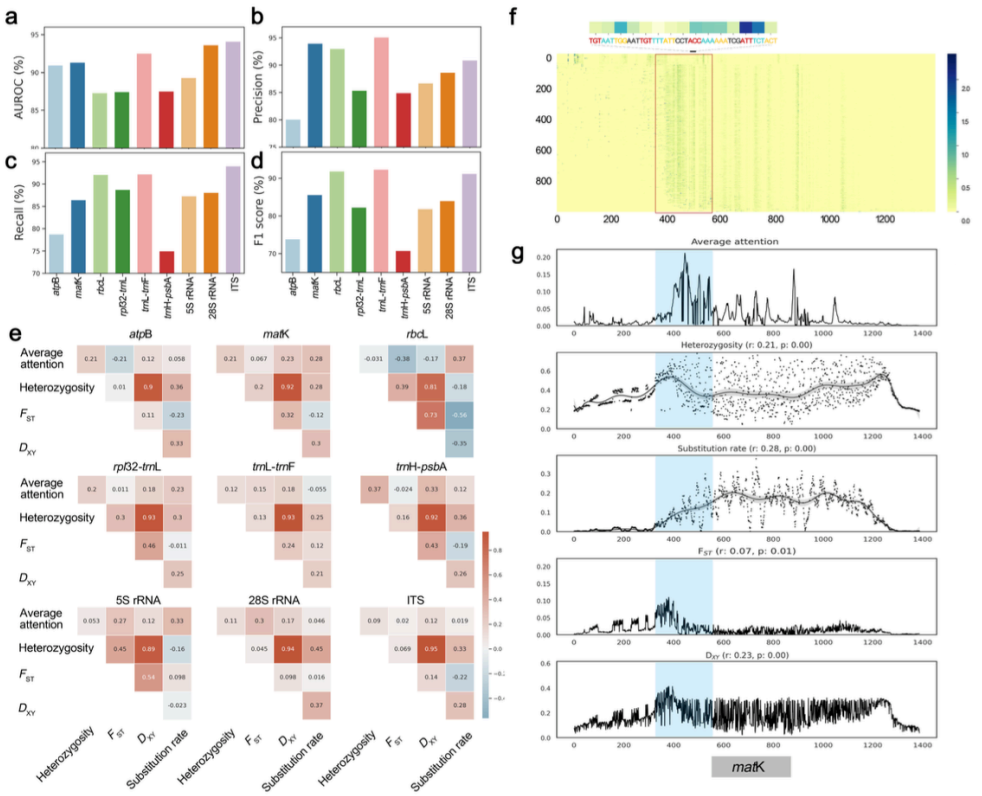
Thank you for your constructive suggestion. To further investigate the evolutionary features of high-attention regions, we expanded our analysis by incorporating two additional metrics:

- Absolute divergence (D_{XY}): Reflecting the more ancient divergence between sequences [9].
- Nucleotide substitution rate: Representing the evolutionary rate of different genetic regions [10].

We did not calculate amino acid diversity per column because the markers we selected are nucleotide sequences, and only three out of the nine sequences are protein-coding genes. The remaining sequences are non-coding and are not translated into proteins. Therefore, in the revised version, we focused on heterozygosity, nucleotide substitution rate, F_{ST}, and D_{XY} to assess the evolutionary characteristics of the regions. However, if you believe that calculating amino acid diversity is necessary, we are more than willing to perform this analysis for the three coding genes. Below are the brief conclusion for the analyses of D_{XY} and substitution rate.

As shown in the updated Figure 7 and Figures S7, S8, S9, we found that regions receiving heightened attention often show relatively high levels of D_{XY}, and substitution rate, which is consistent with the patterns observed for heterozygosity and F_{ST}. We also calculated the Pearson coefficient between the average attention and these genetic metrics, and found that most of correlations are positive, particularly the substitution rate, suggesting that the rapid evolutionary rate may make these regions valuable for evolutionary studies. However, in some markers, the correlation was not significant (with p-value greater than 0.1, Supplementary Table S7). **Therefore, we propose that the model's attention, which is not directly equivalent to other genetic metrics, can be considered an independent genetic parameter for reference in future research.** It highlights the regions the model deems most informative for downstream tasks, rather than directly reflecting traditional genetic variations. This makes attention a distinct and complementary measure that provides additional insight into sequence variability and evolutionary relevance.

We have updated the manuscript (lines 309-316) to include these additional findings and appreciate your suggestion to explore this aspect further.



Updated Figure 7. Visual analyses of BERTPhylo based on nine molecular markers of the PlantSeqs dataset. (**a**) Comparison of the average AUROC for four taxonomic ranks of different markers in novelty detection. | (**b-d**) Comparison of the average precision (**b**), recall (**c**) and F1-score (**d**) for four taxonomic ranks of different markers in taxonomic classification. | (**e**) Pearson correlation coefficient between average attention, heterozygosity, the fixation index (F_{ST}), absolute divergence (D_{XY}), and substitution rate based on nine molecular markers. | (**f**) Attention heatmap of 1,000 DNA sequences of the chloroplast marker *mat*K using BERTPhylo. The red box highlights the attention peak region for the majority of sequences, with an example of the corresponding DNA sequences displayed above it. | (**g**) Average attention, heterozygosity, substitution rate, F_{ST}, and D_{XY} curves of *mat*K. The blue shaded area denotes the peak region of attention.

Comment 7

I was wondering how representative is PlantSeqs of potential use cases of the method? In particular, I feel that Kraken is often used with microbial data. I think it would be useful to discuss how much we can extend the results obtained here on plants to other clades and other genes.

Response 7

Thank you for raising this important question.

To explore the generalizability of BERTPhylo to non-plant clades and genes, we conducted supplementary experiments on a microbial dataset derived from the *Bordetella* genus. Compard with the PlantSeqs dataset, the *Bordetella* dataset presents a more challenging scenario, with approximately 16 times more genes but fewer sequences per gene (fewer than 20 sequences per gene). Additionally, the sequence similarity between the test and training sets is lower, with MMseqs achieving only 47.18% classification accuracy at the species level.

We evaluated and analyzed BERTPhylo in three aspects, including identification the smallest taxonomic units, comparison the phylogenies quality using high- and low-attention regions, and analysis of the relationship between average attention scores and genetic diversity metrics. Detailed information and results are provided in Additional Experiments on the *Bordetella* Genus. **Our results, as illustrated in Table S5 and Figure S11, demonstrate that BERTPhylo effectively identifies the smallest taxonomic units as well as key regions for efficient phylogeny reconstruction even under these more extreme conditions.** These findings highlight BERTPhylo's applicability to diverse datasets beyond plants.

We appreciate your suggestion and have added the supplementary experiments on the *Bordetella* genus into the manuscript (lines 351-381).

Comment 8

I was able to download the code and setup the conda environment. I did not try to reproduce the results in the paper with the command provided on github. Perhaps it would be useful to provide a smaller test set, possibly one that could be run on CPU, for users who do not have access to a good GPU but would still want to test the method.

Response 8

Thank you for your helpful feedback and suggestion. Given that BERTPhylo relies on a pretrained DNA language model that requires GPU resources for efficient execution, we acknowledge that users without access to GPUs may face challenges in reproducing our results. To address this, we provide a Colab link (https://colab.research.google.com/drive/1l2EdbAehS79a5GSy4PyGoVZv-nwjJ4Lt) that allows users to reproduce our results with the help of high-performance GPU resources provided by Google. This should enable a wider range of researchers to test the approach.

Minor comments

l66: "focusing on accelerating and parallelizing calculations have been developed, such as DupTree [21]" : I would not include DupTree in this list as it does not take sequence data as input, but gene trees containing events of gene duplication and loss.

Response: We have removed DupTree from the list of tools aimed at accelerating and parallelizing calculations.

l110: "ETE3 [43],": I find it strange to include ETE3 here, because it is a library for handling phylogenetic trees and, to a small extent, alignments.

Response: We have reconsidered this inclusion and removed ETE3 from the list.

l202: "This analysis is further corroborated by combining MMseqs2 with BERTPhylo, leading to additional enhancements across almost all metrics, notably a 5.04% increase in macro-F1 at the class rank.": it is not clear to me how MMseqs2 and BERTPhylo are combined, and so how this results in an improvement.

Response: we have revised the manuscript (lines 137-152) to clarify the distinction between BERTPhylo and MMseqs2, and also updated the description of smallest taxonomic unit results (lines 215-249) to reduce confusion between the two.

l243: "while in the topography constructed": I'm not sure what topography means here.

Response: This has been corrected to "topology".

l273: "where heterozygosity serving as a measure": serves

Response: We have corrected the typo, changing "serving" to "serves".

l328: "Embyrophyta": Embryophyta

Response: This typo has been corrected to "Embryophyta".

l350: "Drawing inspiration from the taxonomy we use should reflect shared ancestry (i.e. phylogeny)": weird sentence

Response: We have revised this sentence for clarity: "Since taxonomic classifications reflect shared ancestry, we leverage this principle to infer the smallest taxonomic unit of a sequence based on its position within the taxonomic hierarchy".

l669: "an specific input sequence" : a specific Response: We have corrected the typo, changing "an" to "a".

Fig. S6: "high-attention regions (right) versus low-attention regions (left).": it's the opposite.

Response: We have corrected the description in the figure legend to match the actual visualization: "high-attention regions (left) versus low-attention regions (right)".

Fig. S7: the legend does not explain what the grey stripes represent (in the legend of Fig. 7 it is explained). Fig. S7 would be easier to read if the x scales matched between the heatmap and the line plots, as in Fig. 7.

Response: We have updated the figure legend to include an explanation of the grey stripes, and adjusted the layout of the figure so that the x-axis scales of the heatmap and the line plots align for better readability. Additionally, we also added the curves of D_{XY} and substitution rate for each molecular marker.

For Reviewer #2,

We would like to sincerely thank Reviewer #2 for taking the time to carefully review our manuscript and provide valuable and constructive feedback. We greatly appreciate the reviewer's recognition of our approach to leveraging attention mechanisms for identifying important genomic markers, which was described as a very elegant approach. In response to the reviewer's critiques and suggestions, we provide point-by-point replies below. We hope that these responses and the corresponding revisions will address the raised concerns and further clarify the contributions of our method.

Comment 1

As mentioned, the Methods section does not contain a full description of the methods. From what I understood, the deep neural net is used merely to identify the correct taxon (at the level of genus, family, order or class), essentially solving a classification problem. To do the actual phylogenetic reconstruction, I imagine, one needs to apply some traditional sequence-based method, but the Methods section gives no details on this. If I am correct, the statement in the main part of the manuscript that BERTPhylo integrates DL into the phylogenetic reconstruction pipeline is overstated and misleading.

Response 1

Thank you for pointing out the need for a clearer and more complete description of our methodology. We have clarified the contributions of BERTPhylo and its integration into the phylogenetic reconstruction pipeline as follows:

1. Clarification of BERTPhylo's contributions and workflow

BERTPhylo utilizes pretrained DNA language models to **identify the smallest taxonomic unit** for new sequences and **extract high-attention regions** most useful for phylogenetic updating. Subsequently, existing tools such as MAFFT (for sequence alignment) and RAxML (for tree inference) are used to update the subtree topology based on these high-attention regions. The workflow of BERTPhylo is illustrated in Figure 1a of manuscript. This hybrid approach combines the efficiency and precision of deep learning in preprocessing with the reliability of traditional methods for phylogenetic reconstruction. Moreover, BERTPhylo represents the first attempt to leverage the attention of genomic large language models to accelerate phylogenetic updating and provides a feasible solution.

2. Distinction from classification tasks

Given a phylogenetic tree, a new sequence may belong to an unknown taxon at the genus rank but align with a known taxon at a higher rank. **Therefore, identifying the smallest taxonomic unit of a new sequence goes beyond standard classification; it requires determining the lowest rank at which the sequence can be classified into a known taxon.** Traditional methods, such as MMseqs, can identify the most similar sequences, but fail to ensure consistency across all taxonomic levels between the identified and query sequences. As illustrated in the Updated Table

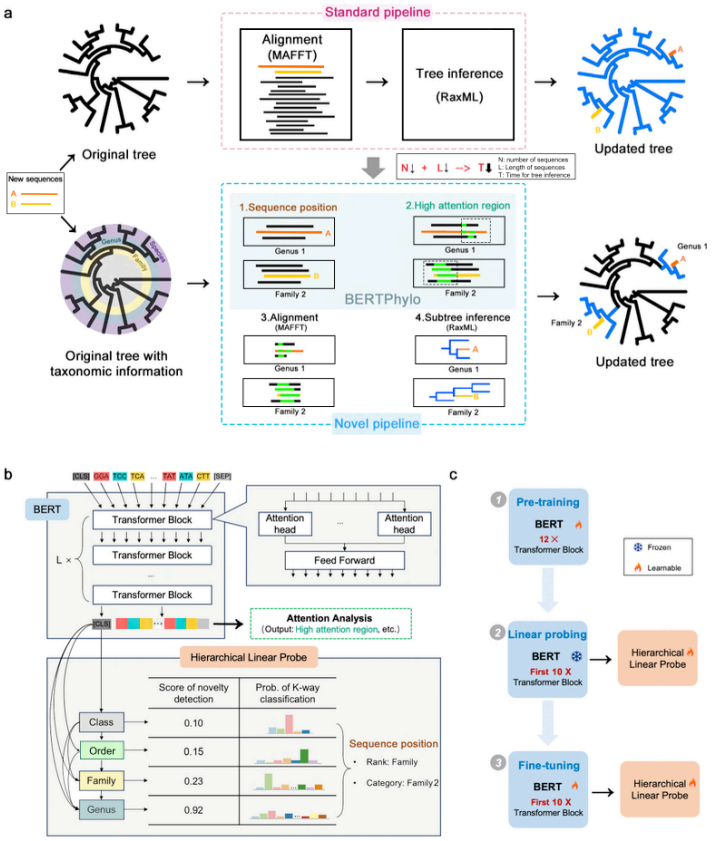


Figure 1. Pipeline explanation, model architecture and training protocol of BERTPhylo. (**a**) In contrast to standard pipelines, BERTPhylo prompts a novel pipeline which is tailored to enhance computational efficiency concerning both sequence number and length. Leveraging existing taxonomic classification systems, BERTPhylo narrows the input to alignment (e.g. MAFFT) and tree inference tools (e.g. RAxML) by restricting updates to designated subtrees and pinpointing potentially valuable regions in subtree sequences. | (**b**) The BERTPhylo model architecture adapted to the PlantSeqs dataset comprises a Transformer-based BERT model along with hierarchical linear probes (HLPs) across four taxonomic ranks (class, order, family, genus) to identify the taxonomic unit of new sequences and extract high-attention regions as potentially valuable regions. | (**c**) Training protocol of BERTPhylo. This utilizes a BERT framework pretrained on DNA sequences, starting with linear probing of randomly initialized HLPs via the first ten layers, followed by fine-tuning of the full network.

S1, traditional methods exhibit significant performance drops when sequences from unknown taxa are included.

Method	Macro Precision				Macro Recall				Macro F1			
	Class	Order	Family	Genus	Class	Order	Family	Genus	Class	Order	Family	Genus
MMseqs2	89.58	93.42	95.95	99.22	93.81	94.75	97.43	99.20	89.77	93.54	96.41	99.20
BLAST	89.16	93.55	95.96	99.24	90.40	94.09	97.16	99.20	87.60	93.24	96.25	99.21
Kraken2	93.79	94.71	95.66	99.13	89.01	92.47	95.87	97.78	90.17	93.10	95.44	98.42

(a) Performance on test sequences belonging to known taxa.

Method	Macro Precision				Macro Recall				Macro F1			
	Class	Order	Family	Genus	Class	Order	Family	Genus	Class	Order	Family	Genus
MMseqs2	78.92	88.06	84.81	74.80	78.28	92.17	95.98	98.24	69.40	87.58	87.92	82.46
BLAST	79.46	88.62	85.60	75.23	78.49	92.33	95.95	98.26	69.92	88.13	88.50	82.85
Kraken2	82.61	88.44	85.40	75.92	77.87	91.43	94.96	96.98	75.82	89.34	88.74	83.08

(b) Performance on all test sequences, including those from unknown taxa.

Updated Table S1. Performance comparison of traditional classification methods on the test set containing unknown taxa from the reference database. All metrics are calculated using macro-average methodology, highlighting a significant decline in classification results when unknown taxa are included.

To enhance clarity, we have moved the description of the BERTPhylo method from Section 4.1 to Section 2.1 and provided a more detailed explanation of BERTPhylo's methodology (lines 125-169). We hope that these changes, along with the example shown in Figure 1, will make the workflow clearer for readers.

Comment 2

The manuscript lacks proper benchmarks. Essentially, all results are obtained for only one dataset from PlantSeqs. If the method is supposed to be generally applicable, as it is alluded to in the introduction, one needs a thorough benchmark on different phyla, as well as on simulated data where the ground truth is known.

Response 2

Thank you for highlighting the need for broader benchmarking.

To explore the generalizability of BERTPhylo to non-plant clades and genes, we conducted supplementary experiments on a microbial dataset derived from the *Bordetella* genus. Compard with the PlantSeqs dataset, the *Bordetella* dataset presents a more challenging scenario, with approximately 16 times more genes but fewer sequences per gene (fewer than 20 sequences per gene). Additionally, the sequence similarity between the test and training sets is lower, with MMseqs achieving only 47.18% classification accuracy at the species level.

We evaluated and analyzed BERTPhylo in three aspects, including identification the smallest taxonomic units, comparison the phylogenies quality using high- and low-attention regions, and analysis of the relationship between average attention scores and genetic diversity metrics. Detailed information and results are provided in Additional Experiments on the *Bordetella* Genus. **Our results, as illustrated in**

Table S5 and Figure S11, demonstrate that BERTPhylo effectively identifies the smallest taxonomic units as well as key regions for efficient phylogeny reconstruction even under these more extreme conditions. These findings highlight BERTPhylo's applicability to diverse datasets beyond plants.

We appreciate your suggestion and have added the supplementary experiments on the *Bordetella* genus into the manuscript (lines 351-381).

Comment 3

Comparison with traditional methods for phylogenetic reconstruction (e.g. RAxML, BEAST, Mash) is missing. One needs to present the comparison of the quality of reconstruction, as well as of the needed compute resources.

Response 3

Thank you for raising this important point. We have addressed this issue as follows:

1. Comparison with traditional methods for reconstruction

The role of BERTPhylo is not to replace traditional phylogenetic methods but to enhance them by identifying clades and extracting high-attention regions. Although our experiments used RAxML for phylogenetic reconstruction, it can be used in conjunction with other tree-building methods.

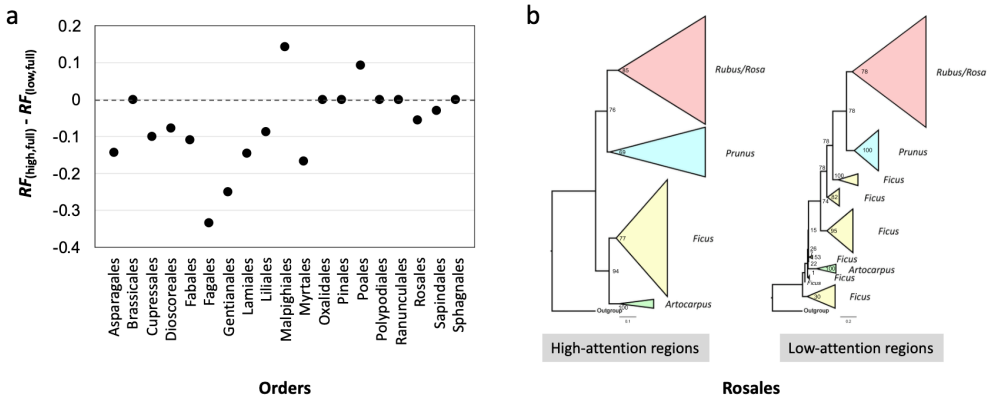
2. Reconstruction quality

In Section 2.4 of the manuscript, we have evaluated the quality of phylogenetic reconstruction, **demonstrating that high-attention regions yield superior phylogenetic resolution.** In the revised manuscript, we further quantified the quality of phylogenetic reconstruction using the Robinson-Foulds (RF) distance, comparing trees constructed with high-attention and low-attention regions to trees built with all sequences. **As shown in the Updated Figure 6a, nearly** 90% **of the trees constructed using high-attention regions had a closer or equal RF distances to the full-sequence trees than those constructed using low-attention regions.** This suggests that when there is a need to improve tree-building efficiency due to the length of input sequences, using high-attention regions as a substitute for full sequences is an excellent option.

3. Comparison of computational resources

In Figure 3 of the manuscript, we have provided a comparison of compute resource requirements between the standard pipeline and the BERTPhylo-based pipeline. The runtime of BERTPhylo depends on the size of the subtree and the length of the high-attention regions. For instance, when the updated sequence belongs to an unknown taxon of the order Malpighiales, and using high-attention regions reduced to one-third of the original length, aligning and updating the subtree takes only 4 minutes, **representing just** 0.33% **of the time required by the standard pipelines**.

We hope these clarifications addresse your concerns and highlight the efficiency benefits of BERTPhylo compared to standard pipeline.



Updated Figure 6. Phylogenetic trees constructed using high-attention regions perform better. (**a**) Difference in Robinson-Foulds (RF) distance between trees constructed from high-attention and full-length sequences (RF(high, full)) versus those constructed from low-attention and full-length sequences (RF(low, full)). (**b**) Phylogenetic trees constructed for the angiosperm order Rosales using high-attention (left) and low-attention regions (right) of sequences extracted by BERTPhylo, respectively.

Comment 4

Results of the taxonomic assignment with BERTPhylo (Table 1) are worse than with the traditional methods (in particular, Kraken2). Where is the advantage of using BERTPhylo? Additionally, in Table 1, I see that the genus-level assignment is predicted better that class-level, which is counter-intuitive to me – I would imagine that higher-level taxa are easier to predict. Could the authors comment on this?

Response 4

Thank you for your valuable comments. We provide the following clarifications:

1. Clarification of Table 1 performance and the advantages of BERTPhylo

As noted in Response 1, the identification of the smallest taxonomic unit accomplished by BERTPhylo is not a standard classification task, it involves determining the lowest rank at which the sequence can be classified into a known taxon. **Traditional methods lack the ability to confirm whether the identified sequence and the query sequence are consistent at the taxonomic level, so they cannot complete this task.**

Table 1 in the manuscript only reports the performance for sequences belonging to known taxa in the PlantSeqs dataset. However, when sequences from unknown taxa are included, traditional methods exhibit significant performance drops, as shown in the Updated Table S1. Supplementary experiments on the *Bordetella* dataset demonstrated similar outcomes, where traditional methods underperformed compared to BERTPhylo in identifying the smallest taxonomic unit (Updated Table S5).

2. Explanation for genus-level assignment outperforming class-level assignment

The seemingly counterintuitive result that genus-level predictions outperform class-level predictions arises from differences in the test sequence distributions across taxonomic ranks. Table 1 reports performance exclusively for sequences belonging to known taxonomic units in the PlantSeqs dataset. A sequence may belong to a known class but not to a known genus, leading to a varying number of test sequences across taxonomic ranks.

Specifically, as shown in Figure 9 of manuscript, the PlantSeqs dataset includes five types of sequences:

① Known genus,
② Unknown genus but known family,
③ Unknown genus and family but known order,
④ Unknown genus, family, and order but known class, and
⑤ Completely unknown taxa in all ranks.

This variation causes the proportion of in-distribution (ID) versus out-of-distribution (OOD) sequences to change across taxonomic levels. As detailed in Table 2 of manuscript, coarser taxonomic levels have more ID test sequences from known taxonomic units (e.g., $14,700$ for class vs. $10,000$ for genus). Because Table 1 reports performance solely for ID sequences, coarse taxonomic levels may show seemingly poorer performance due to the higher proportion of ID test sequences.

Notably, Table 1 highlights the challenge of correctly classifying partially known sequences (e.g., sequences known at the family level but unknown at the genus level). This challenge underscores the added value of BERTPhylo in identifying the most informative taxonomic level rather than merely finding the closest match.

We appreciate your suggestion and have revised the manuscript (lines 137-152) to better highlight the strengths of BERTPhylo and the limitations of traditional methods. Additionally, we improved the description of smallest taxonomic unit results (lines 215-249) to reduce confusion.

Comment 5

How was split into the training/validation/test sets performed during the model training? What was the training goal?

Response 5

Thank you for your question.

1. Training/Validation/Test Set Split

The data split process is explained in **Section 4.1**, with Figures 8a and 8b illustrating how the training, validation, and test sets were created. It is important to note that only sequences with known genus were used for training and testing, and

other types of sequences were only used in the testing phase. The number of sequences across different taxonomic ranks for each set is also provided in Table 2.

2. Training Goal

The training goal of fine-tuning was to enable the model to correctly classify each sequence at every taxonomic level. In classification tasks, cross-entropy loss is commonly used to quantify the difference between the predicted and actual probability distributions. We fine-tuned the model by minimizing this difference during training. To address class imbalance, we employed a weighted cross-entropy loss. The detailed formula and training protocol are provided in **Section 4.3** of manuscript.

We hope these clarifications address your concerns regarding the experimental datasets and model training.

Minor comments

1. Lines 53-55: This makes an impression that people do classical alignment-based phylogenetic reconstruction on whole genomes, which is not the case anymore. Please consider rephrasing.

Response: We have removed "up to entire genomes" to avoid confusion.

2. Lines 56-71: I find the distinction of methods into distance-based vs. character-based unusual. I am more used to classifying phylogenetic methods into alignment-based and alignment-free. All character-based methods, as defined by the authors, would fall into the alignment-based category, but distance-based methods are a mix of the both types.

Response: We divided methods into distance-based and character-based methods, following references [11, 12]. To avoid confusion, we have added relevant citations when introducing these classifications.

3. Table 1: Elapsed times for all methods should be provided.

Response: Thank you for your valuable suggestion. The comparison of computational efficiency is discussed in lines 200-212. The average inference time per sequence for BERTPhylo is about 0.013 seconds(s), while MMseqs, BLAST, and Kraken2 takes about 0.005, 0,008, 0.01 s. Notably, BERTPhylo's primary contribution lies in accelerating the process of updating phylogenetic trees. For instance, the time required for aligning and updating the subtree of the order Malpighiales takes only 0.33% (240 s) of the total time required for the standard pipeline (Figure 3 in the manuscript). Furthermore, BERTPhylo's inference time is significantly lower than the time required for tree construction.

4. Lines 180-199: I strongly recommend using MCC (Matthew's correlation coefficient) as a measure of quality of classification, since this is the most unbiased measure.

Response: Thank you for your valuable suggestion. We have added MCC to classification-related evaluations, including the Updated Table 1 for the PlantSeqs

dataset, and Table S5 for the supplementary experiments. BERTPhylo shows better MCC performance than the baseline on both datasets.

5. Section 2.2: No details of application of DNABERT or BERTax or any of the traditional methods are given neither here nor in the Methods section. This is particularly important for DNABERT, which is a DNA foundation model with many downstream tasks, so here I am completely at a loss how it was applied. For traditional tools, on the other hand, parameters may be extremely important: for example, for Kraken2, the choice of the database may play a crucial role in its performance.

Response: Thank you for your valuable suggestions. We describe how to apply DNABERT, BERTax, and traditional methods to the PlantSeqs dataset in lines 186-189 of the original manuscript: "For a meaningful comparison, traditional methods use the training set of PlantSeqs as reference database, while deep learning-based methods leverage their respective open source pretrained models to fine-tune randomly initialized hierarchical linear probes (HLPs) on the same dataset." Lines 224-226 of the revised manuscript also describe how the baseline DNABERT was adapted to PlantSeqs dataset.

6. Wording is strange at some places, e.g. I would recommend changing 'equitably' to 'equally' (line 259) or 'heightened' to 'increased' (line 298).

Response: Thank you for your suggestion, we have changed "equitably" to "equally" and "heightened" to "increased".

7. Section 4.4: What is sentence? Was some sliding window applied for longer sequences?

Response: Thank you for your question. We have changed "sentence" to "sequence" to keep the context consistent and avoid confusion. Additionally, we used a 3-mer with sliding window 3 on the PlantSeq dataset. Due to the input restrictions of the pretrained model DNABERT, PlantSeqs filtered out sequences longer than 1,530 bp (lines 438-441). However, it is important to note that the framework of BERTPhylo is designed to be adaptable to any BERT-based large language model. This flexibility helps address the limitations of specific pretrained models, such as sequence length restrictions. For example, in the newly added experiment, we used DNABERT-S, a pretrained model that removes the sequence length restriction, to fine-tune the Bordetella dataset. With DNABERT-S, BERTPhylo is capable of handling arbitrarily long sequences, demonstrating its ability to adapt to more advanced DNA models.

For Reviewer #3,

We would like to express our sincere gratitude to Reviewer #3 for taking the time to carefully read our manuscript and for providing such insightful and constructive feedback. We greatly appreciate to the reviewer for assessing our method as interesting and potentially valuable. Below, we provide point-by-point responses to the reviewer's comments, which we hope will address the concerns raised and further clarify the contributions of our approach.

Comment 1

The method involves identifying suitable clades for placing new sequences into an existing backbone tree. It is not clear how many new sequences are handled at a time. Does the method process sequences one by one, or can it handle multiple sequences simultaneously? More importantly, what happens when the input sequences belong to different taxonomic levels (e.g., genus, order, or family)? Can BERTPhylo assign multiple sequences to different clades? For example, in Figure 2, the three new sequences are all from the order Fabales. What if these sequences belong to different clades? An evaluation under such scenarios would strengthen the paper's contributions.

Response 1

Thank you for your thoughtful questions. BERTPhylo can process multiple sequences simultaneously, with no theoretical upper limit. However, for efficiency, we cap the number of sequences processed at a time to 16. These sequences can originate from the same or different taxonomic levels. BERTPhylo outputs sequence sets based on whether their clades overlap:

- **Overlapping clades:** A unified sequence set is generated to reconstruct the largest evolutionary clade.
- **Non-overlapping clades:** Separate sequence sets are generated for each clade to reconstruct individual clades.

For example, in Figure 2, BERTPhylo processes five sequences and outputs two sequence sets:

- **Fabales clade sequence set (Figure 2i):** Includes the high-attention regions of the three new Fabales sequences and all original Fabales sequences.
- **Primulina clade sequence set (Figure 2ii):**, Includes the high-attention regions of the two new *Primulina* sequences and all original *Primulina* sequences.

If three new sequences in Figure 2i belong to different orders, BERTPhylo operates as follows:

- **No overlap with the *Primulina* clade:** Four sequence sets will be generated, one for the *Primulina* clade and one for each of the three new sequences.
- **Overlap with the *Primulina* clade:** Three sequence sets will be generated, among which the set of the overlapping order containing both the new sequence belonging to the order level and the two newly added sequences in Figure 2ii.

We have clarified in the revised manuscript how BERTPhylo handles multiple sequences at different taxonomic levels or clades (lines 194-200). Thank you again for your insightful feedback, which has helped us better explain the method's capabilities!

Comment 2

The authors compiled a dataset containing sequences of land plants to train and test their method. Given that phylogenomic datasets are often heterogeneous, a model trained on a specific dataset may not generalize well to others. It is essential to test BERTPhylo on unseen datasets from different lineages or taxonomic groups to assess its robustness and generalizability.

Response 2

Thank you for this constructive question. We address your concerns about the robustness and generalizability of BERTPhylo under the following two scenarios:

1. Testing on unseen datasets from different lineages or taxonomic groups

A key advantage of BERTPhylo is its ability to detect sequences that do not belong to known taxonomic units through novelty detection scores. For test sequences from lineages or taxonomic groups distinct from the training data, BERTPhylo is likely to classify them as out-of-distribution (OOD) at the coarsest levels. These sequences can be treated as outgroups to update the existing phylogenetic tree. We have explored this scenario in the PlantSeqs dataset, where sequences at all taxonomic ranks belong to unknown taxa. **As shown in the Table 1a and Figure 4 of revised manuscript, BERTPhylo effectively identifies these test sequences that differ significantly from the training data.**

2. Generalizability to non-plant datasets

The framework proposed by BERTPhylo, which leverages pretrained large language models (LLMs) to accelerate phylogenetic tree updates, is not limited to land plant lineages. For non-plant datasets, the pretrained model can be fine-tuned using the training strategy outlined by BERTPhylo, enabling the identification of taxonomic units and extraction of high-attention regions. **Supplementary experiments on the microbial dataset derived from the *Bordetella* genus demonstrate how BERTPhylo can be applied to non-plant clades** (details in Additional Experiments on the *Bordetella* Genus), which highlights BERTPhylo's generalizability across diverse datasets and lineages.

It is also important to note that the performance of pretrained LLMs on downstream tasks depends on the representational capacity of the LLM, as well as the quantity and complexity of data used for the downstream task.

Comment 3

The novelty detection aspect of the model is innovative, but it may struggle with cases where sequence data is highly divergent from the training set. In such cases, the model could misclassify or fail to place taxa accurately within the phylogeny. Further

exploration with examples of how the method performs with divergent sequences would be valuable.

Response 3

Thank you for acknowledging the innovative aspects of our novelty detection approach. We address your concerns regarding highly divergent test sequences under the following two scenarios.

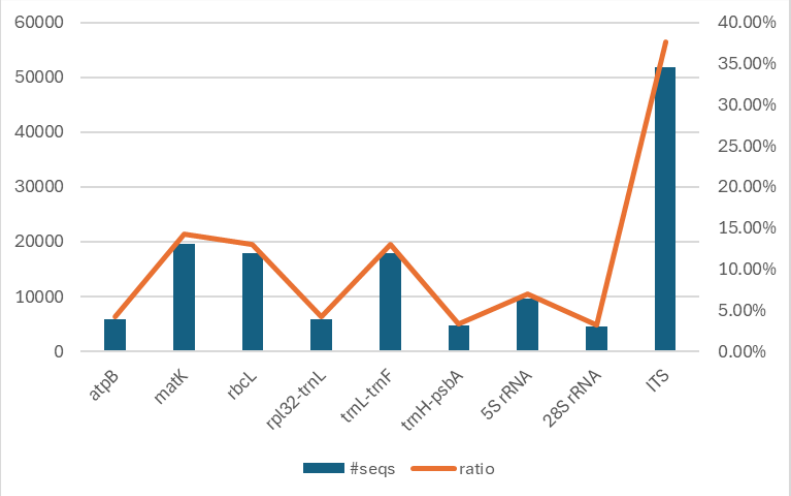
1. Test sequence belongs to an unknown taxon.

As noted in Response 2, one of BERTPhylo's key advantages over traditional classification methods is its ability to avoid forcing a test sequence into an incorrect taxon. Instead, BERTPhylo provides a novelty detection score for each test sequence at every taxonomic level. A high score suggests that the sequence likely belongs to an unknown taxon at that level, rather than erroneously assigning it to a known taxon.

For instance, the PlantSeqs dataset includes sequences from unknown taxa across all taxonomic ranks (Figure 9a). These sequences differ significantly from the training data due to their greater phylogenetic distance. However, as demonstrated in Table 1 and Figure 4 of revised manuscript, BERTPhylo achieves higher accuracy in detecting novelty even for highly divergent sequences, reinforcing its effectiveness in this context.

2. Test sequence belongs to a known taxon but derives from a molecular marker with limited training data coverage.

The figure below shows the proportion of training sequences corresponding to different molecular markers in the PlantSeqs dataset. Combined with the novelty detection and taxonomic classification performance for different molecular markers (Figure 7a,b,c,d), it is evident that markers like *atp*B, *rpl*32-*trn*L, and *trn*H-*psb*A, which have lower representation in the training data, exhibit comparatively weaker



performance. Conversely, 28S rRNA, despite having the lowest representation, performs robustly in both tasks. This may be due to attributed to its genomic proximity to ITS, a highly represented nuclear marker, allowing the model to effectively capture nuclear marker characteristics.

Indeed, when the model encounters sequences from the same taxon as the training data but with high divergence, the performance would be affected. Expanding training dataset to include a broader range of sequences can further improve the model's generalization and robustness in such cases.

Comment 4

The method identifies valuable regions in the sequences based on attention scores from the model. However, the decision-making process of the deep learning model, specifically how it identifies high-attention regions, may lack biological interpretability. This is a significant drawback for researchers seeking to understand the biological basis behind the model's predictions. Can the authors provide more evidence or case studies demonstrating the biological relevance of these high-attention regions?

Response 4

Thank you for emphasizing the importance of biological interpretability. We fully agree that understanding and validating the high-attention regions identified by BERTPhylo is critical. To address this concern, we performed an analysis of these regions in relation to heterozygosity and F_{ST}, as detailed in Section 2.4 and 2.5 of the manuscript. **Our findings indicate that regions receiving high attention generally exhibit elevated levels of heterozygosity and $\boldsymbol{F_{\mathrm{ST}}}$.**

Building on this foundation and incorporating Reviewer #1's suggestions, we expanded our analysis by incorporating two additional metrics:

- Absolute divergence (D_{XY}): Reflecting the more ancient divergence between sequences [9].
- Nucleotide substitution rate: Representing the evolutionary rate of different genetic regions [10].

As shown in the Updated Figure 7 and Figures S7, S8, S9, we found that regions receiving heightened attention often show relatively high levels of $\boldsymbol{D_{\mathrm{XY}}}$, and substitution rate, which is consistent with the patterns observed for heterozygosity and $\boldsymbol{F_{\mathrm{ST}}}$. We also calculated the Pearson coefficient between the average attention and these genetic metrics, and found that most of correlations are positive, particularly the substitution rate, suggesting that the rapid evolutionary rate may make these regions valuable for evolutionary studies. However, in some markers, the correlation was not significant (with p-value greater than 0.1). **Therefore, we propose that the model's attention, which is not directly equivalent to other genetic metrics, can be considered an independent genetic parameter for reference in future research.** It highlights the regions the model deems most informative for downstream tasks, rather than directly reflecting traditional genetic

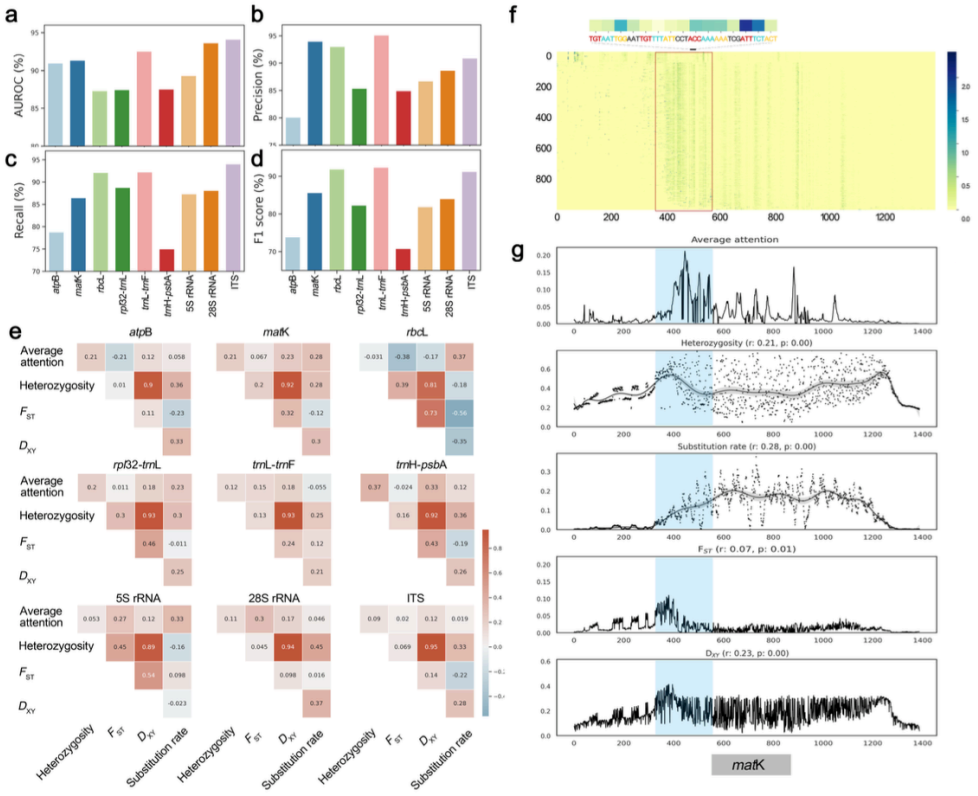


Figure 7. Visual analyses of BERTPhylo based on nine molecular markers. (**a**) Comparison of the average AUROC for four taxonomic ranks of different markers in novelty detection. | (**b-d**) Comparison of the average precision (**b**), recall (**c**) and F1-score (**d**) for four taxonomic ranks of different markers in taxonomic classification. | (**e**) Pearson correlation coefficient between average attention, heterozygosity, the fixation index (F_{ST}), absolute divergence (D_{XY}), and substitution rate based on nine molecular markers. | (**f**) Attention heatmap of 1,000 DNA sequences of the chloroplast marker *mat*K using BERTPhylo. The red box highlights the attention peak region for the majority of sequences, with an example of the corresponding DNA sequences displayed above it. | (**g**) Average attention, heterozygosity, substitution rate, F_{ST}, and D_{XY} curves of *mat*K. The blue shaded area denotes the peak region of attention.

variations. This makes attention a distinct and complementary measure that provides additional insight into sequence variability and evolutionary relevance.

Comment 5

The authors mention that they divide sequences into three regions and then identify the high-attention regions. Does this imply that the attention regions are contiguous,

corresponding to one of these three predefined divisions? If so, why must they be contiguous?

Response 5

Thank you for your thoughtful question.

To clarify, the division of sequences into three regions is a specific choice for the PlantSeqs dataset. **The high-attention regions identified by BERTPhylo are not necessarily contiguous or confined to any single predefined divison.** Instead, the following points highlight the flexibility of this approach:

1. Non-contiguity of high-attention regions:

High attention regions can consist of multiple small segments and do not need to be continuous within a given division. For example, the sequence can be divided into K regions, and the top M (<K) regions with highest attention scores can be selected as high-attention regions. In the supplementary experiment, we tested the effect of dividing sequences into 12 regions and selecting different numbers of high-attention regions (Figure S11).

2. Selection of the number of regions:

The choice of dividing sequences into K regions and selecting M high-attention regions is not arbitrary. For the PlantSeqs dataset, we chose to split the sequences into three regions for two main reasons:

- High-attention regions in most sequences were concentrated (Figure 7 and Figure S7, S8, S9).
- The shortest sequence in the PlantSeqs dataset is about 300 bp, so dividing it into three regions ensures that high-attention regions are of sufficient length for validation.

For longer sequences or when model attention is more distributed, we recommend using a larger number of divisions to better capture high-attention regions.

We hope this explanation clarifies the role of the sequence divisions and the flexibility in identifying high-attention regions. Thank you again for your insightful questions, we have expanded this explanation in the revised manuscript (lines 167-169).

Comment 6

BERTPhylo depends heavily on the pretrained DNABERT model, and as a result, any limitations of DNABERT are likely to propagate into BERTPhylo. For instance, pretrained models like DNABERT typically truncate input sequences to a maximum length (e.g., 1024 bp). Is this also the case for BERTPhylo? Can the proposed method handle arbitrarily long sequences, and if not, how does it manage longer sequences?

Response 6

Thank you for your insightful question. You are correct that **BERTPhylo, as it leverages DNA pretrained models, inherits certain limitations of the underlying models.** For instance, when we used DNABERT with a 3-mer tokenization to fine-tune PlantSeqs, the maximum input sequence length was limited to $1,530$ bp. Consequently, PlantSeqs filterd out sequences longer than $1,530$ bp to accommodate the input length limit of the pretrained model (lines 438-441).

However, the BERTPhylo framework is designed to be highly adaptable and compatible with any BERT-based model. This flexibility addresses the limitations of specific pretrained models, such as sequence length restrictions. For example, in the newly added experiment, we used DNABERT-S, a pretrained model without sequence length limitations, to fine-tune the *Bordetella* dataset. By leveraging DNABERT-S, BERTPhylo is capable of handling arbitrarily long sequences, demonstrating its ability to adapt to more advanced pretrained LLMs.

This adaptability underscores BERTPhylo's potential to evolve in tandem with advancements in DNA pretrained models, ensuring its utility across diverse datasets and scenarios. We appreciate your thoughtful question and hope our response effectively addresses your concerns.

Comment 7

May BERTPhylo's reliance on existing taxonomic classification limit its ability to identify new evolutionary relationships? It would be useful for the authors to discuss the model's potential to discover clades or evolutionary patterns that do not fit within the taxonomic classification framework used in BERTPhylo.

Response 7

Thank you for raising this thought-provoking question.

BERTPhylo's reliance on existing taxonomic classifications as a framework for determining the appropriate clades for new sequences may indeed limit its ability to independently identify novel evolutionary relationships outside the predefined taxonomy. To address this concern, we highlight the following points:

1. Potential for identifying inconsistencies or anomalies:

Although BERTPhylo is grounded in existing taxonomic classifications, it has the capability to detect inconsistencies or anomalies during the classification process. For instance, when a sequence exhibits a high novelty score at a particular taxonomic level, two scenarios are possible:

- **Unknown Taxon:** The sequence may belong to an entirely unknown taxon, reflecting high distributional uncertainty.
- **Known Taxon with Significant Divergence:** Alternatively, the sequence may belong to a known taxon but exhibit significant divergence from existing sequences, indicating high data uncertainty.

The latter scenario may highlight new lineages or evolutionary patterns not captured in the current taxonomic framework. BERTPhylo's ability to flag such anomalies provides opportunities for further exploration and phylogenomic analysis to uncover new relationships.

2. Focus on high-attention regions:

BERTPhylo identifies regions of high phylogenetic importance through its attention mechanism. By analyzing these high-attention regions, researchers can explore whether they reveal patterns that deviate from the expected taxonomic relationships. This analysis could shed light on novel clades or evolutionary paths that challenge the current taxonomy.

3. Adapting to expanded taxonomic frameworks:

The BERTPhylo framework is inherently flexible and can adapt to revised or expanded taxonomic classifications. For instance, as new evolutionary relationships are identified and incorporated into the backbone phylogenetic tree, BERTPhylo can be re-applied to refine and expand phylogenetic analyses, ensuring its relevance as classifications evolve.

4. Future extensions:

While the current version focuses on efficiently updating existing phylogenies, future iterations of BERTPhylo could include unsupervised approaches for clustering sequences. Such approaches would allow the model to identify entirely new evolutionary relationships without relying on predefined taxonomic classifications, enhancing its ability to discover novel patterns.

In summary, while BERTPhylo currently depends on existing taxonomies, it serves as a valuable tool for detecting anomalies, refining phylogenies, and even contribute to the discovery of novel evolutionary relationships in collaboration with other methodologies. We appreciate your insightful suggestion and expand the discussion of this point in the revised manuscript (lines 409-414).

Comment 8

Although BERTPhylo outperforms other deep learning based methods, it generally falls short when compared to traditional approaches like MMseq2 and BLAST. This is understandable, as methods like BLAST are more computationally intensive. However, the model conditions under which the comparisons were made (Tables S1 and S2) seem to be very easy as the precision, recall and F1-scores are close to 100%. I believe the evaluation would benefit from the inclusion of more challenging model conditions.

Response 8

Thank you for your thoughtful feedback.

We would like to clarify that the identification of the smallest taxonomic unit by BERTPhylo is not a conventional classification task. Instead, it involves determining

the lowest rank at which the sequence can be classified into a known taxon. **Traditional methods like MMseqs2 or BLAST lack the capability to verify whether the identified sequence and the query sequence are consistent at each taxonomic level.** Therefore, they cannot accomplish this task.

To further explore the generalizability of BERTPhylo on more challenging datasets, we conducted supplementary experiments on the *Bordetella* genus. Compard with the PlantSeqs dataset, the *Bordetella* dataset contains approximately 16 times more genes but fewer sequences per gene (fewer than 20 sequences per gene). Additionally, the sequence similarity sbetween the test and training sets is lower, with MMseqs achieving only 47.18% classification accuracy at the species level. Detailed information and results are provided in the Section of Additional Experiments on the *Bordetella* Genus. **Our results, as illustrated in Table S5 and Figure S11, demonstrate that BERTPhylo effectively identifies the smallest taxonomic units as well as key regions for efficient phylogeny reconstruction even under these more extreme conditions.** These findings highlight BERTPhylo's applicability to diverse datasets beyond plants.

We have incorporated these supplementary experiments on the *Bordetella* into the manuscript (lines 351-381). Thank you for your suggestion, which has strengthened our work!

References

[1] Bridel, S. *et al.* A comprehensive resource for bordetella genomic epidemiology and biodiversity studies. *Nature Communications* **13**, 3807 (2022).

[2] Zhou, Z. *et al.* Dnabert-s: Learning species-aware dna embedding with genome foundation models. *ArXiv* (2024).

[3] Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

[4] Hilu, K. W. & Liang, g. The matk gene: sequence variation and application in plant systematics. *American journal of botany* **84**, 830–839 (1997).

[5] TCH, C. Malpighiales phylogeny poster (2023).

[6] Wang, H., Wu, Z., Li, T. & Zhao, J. Phylogenomics resolves the backbone of poales and identifies signals of hybridization and polyploidy. *Molecular Phylogenetics and Evolution* **200**, 108184 (2024).

[7] Szöllősi, G. J., Tannier, E., Daubin, V. & Boussau, B. The inference of gene trees with species trees. *Systematic biology* **64**, e42–e62 (2015).

[8] Li, Y.-X. *et al.* Phylogenomics of bivalvia using ultraconserved elements reveal new topologies for pteriomorphia and imparidentia. *Systematic Biology* syae052 (2024).

902 [9] Nagylaki, T. Fixation indices in subdivided populations. *Genetics* **148**, 1325–
903 1332 (1998).

904 [10] Nei, M. *Molecular evolutionary genetics* (Columbia university press, 1987).

905 [11] Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nature*
906 *Reviews Genetics* **13**, 303–314 (2012).

907 [12] Kapli, P., Yang, Z. & Telford, M. J. Phylogenetic tree building in the genomic
908 age. *Nature Reviews Genetics* **21**, 428–444 (2020).

Response to referees for PhyloTune: An Efficient Method to Accelerate Phylogenetic Updates Using a Pretrained DNA Language Model

Dear all reviewers,

We sincerely appreciate your thorough evaluation of our revised manuscript and the insightful, constructive feedback provided. We are grateful for your recognition of the improvements made.

To address concerns regarding benchmarking and generalizability, we have conducted additional experiments on simulated datasets to validate the feasibility and robustness of our approach across varying sequence quantities. Furthermore, we have revised the manuscript accordingly to better articulate the contributions of our method. All modifications are highlighted in red for clarity.

We hope that our point-by-point responses below adequately address your concerns and further clarify the significance of our work. Thank you again for your time and consideration.

For Reviewer #1,

We would like to express our sincere gratitude to Reviewer #1 for the meticulous evaluation and constructive feedback on our revised manuscript and response. The reviewer's positive comments are deeply encouraging. We have implemented all suggested wording improvements with careful attention, as outlined below.

Comment 1

l24: "accelerate phylogenetic updates using a pretrained DNA language model": I don't think "phylogenetic update" is very clear; perhaps seomething like "add taxa to an existing phylogenetic tree using a pretrained language model".

Response 1

We appreciate the reviewer's valuable suggestion regarding terminological clarity. Following this recommendation, we have revised the sentence to:

> "accelerate the integration of novel taxa into an existing phylogenetic tree using a pretrained DNA language model".

Comment 2

l73: Another approach that may be worth citing in this paragraph is this: https://doi.org/10.1101/2024.06.17.599404

Response 2

We are grateful to the reviewer for highlighting this significant work. After reading the method presented in the cited study, we have incorporated the reference in lines 78–80. This approach outputs evolutionary distances between all sequence pairs on the tree, thereby qualifying as a distance-based deep learning method.

Comment 3

l194: "BERTPhylo process multiple sequences": processes

Response 3

We thank the reviewer for identifying this typographical error. The text has been corrected from "process" to "processes".

Comment 4

l263: "had a closer or equal RF distances to the full-sequence trees than those constructed using low-attention regions": this is not very well said: a distance is not "close" or "equal" to a tree, but it can be small.

Response 4

Thank you for pointing out this imprecision. We have revised the sentence to "smaller or equal RF distances", which more accurately conveys the intended meaning.

Comment 5

l295: "the two nulcear markers": nuclear

Response 5

We have corrected the typographical error by changing "nulcear" to "nuclear".

We hope that these revisions address all of the reviewer's remaining concerns and further enhance the clarity and precision of our manuscript. Thank you again for your time and insightful suggestions.

For Reviewer #2,

Comment

The authors have significantly improved the manuscript in response to my comments. In fact they have satisfactorily addressed all of them but one: No benchmarking on simulated data has been performed. I consider this a valuable exercise, since in this case the ground truth is known, and also the tree topology and clade distribution can be modelled freely.

Response

We sincerely appreciate the reviewer's positive evaluation of our manuscript improvements and thank you for suggesting the inclusion of simulated-data benchmarking. This recommendation has provided a valuable opportunity to evaluate our method under controlled conditions with a known ground truth, thereby further demonstrating its robustness and efficacy. Below, we detail our experimental setup, evaluation metrics, and key findings.

1. Experimental Setup

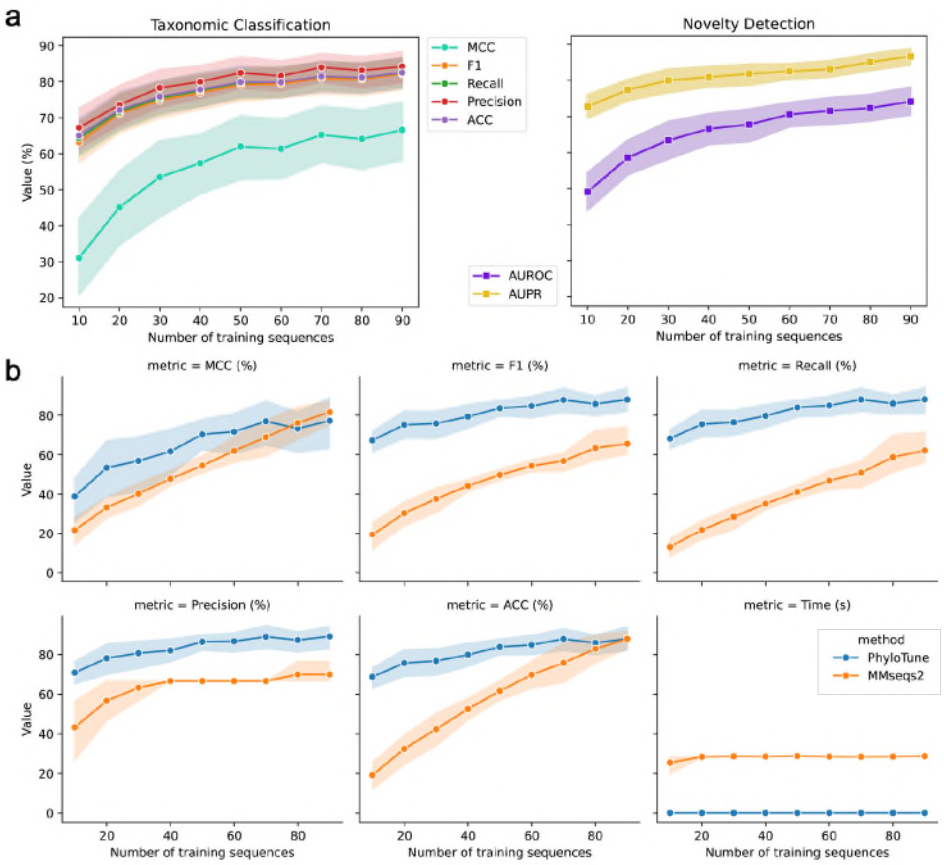
We developed a custom Python script to generate random rooted binary trees for a specified number of species (n). Sequences of $3,000$ bp were then simulated along these trees using Seq-Gen v1.3.5 [1].

1.1 Smallest Taxonomic Unit Identification

To assess smallest taxonomic unit identification, species (excluding the outgroup) were divided into two taxonomic units based on the topology of a randomly generated tree ($n = 100$). The outgroup was expanded into a branch of five species, designated as an unknown taxonomic unit for novelty detection. For sequences belonging to known taxonomic units, we performed stratified 10-fold cross-validation to generate 10 sets of training and testing datasets. In addition to the original training set size, we subsampled $10, 20, 30, 40, \cdots, 80$ sequences to study the impact of varying training sample sizes. Consistent with our experiments on the microbial *Bordetella* dataset, we used DNABERT-S [2]—capable of processing sequences of arbitrary lengths—as our backbone model. To minimize sampling bias, three independent sequence matrices were generated with Seq-Gen, resulting in a total of 30 experimental replicates for each training sample size.

1.2 Phylogenetic Tree Construction from Attention Regions

For a comparative analysis, we evaluated trees constructed from high-attention and low-attention regions across simulated datasets with $n = 20, 40, 60, 80, 100$ sequences. Each simulated sequence was divided into K segments (with K values of $2, 3$, and 4). We then selected regions based on the highest and lowest attention scores obtained from the last layer of the BERT module, respectively. For each size of the species tree, 10 trees were randomly generated for validation.

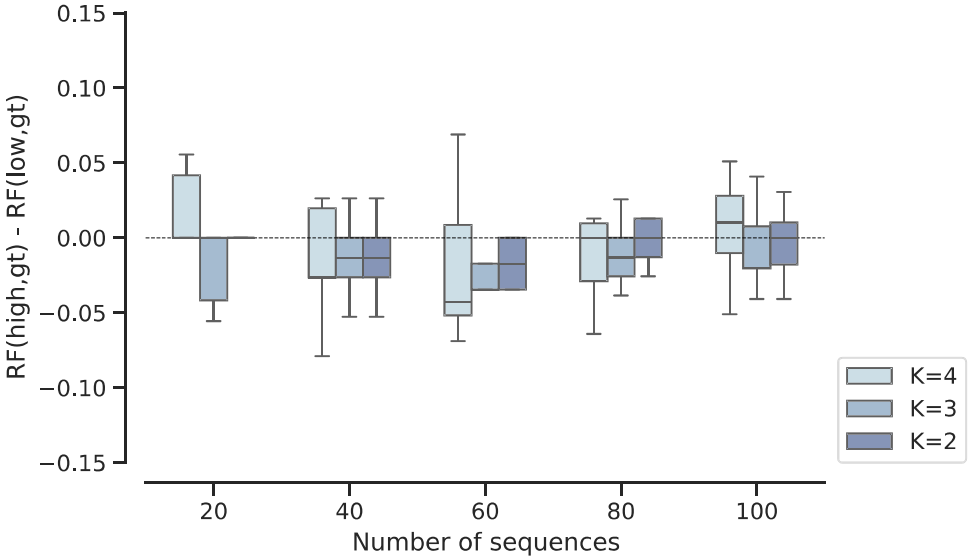


Updated Figure 4. Our method (renamed as PhyloTune) improves with increasing training data and outperforms MMseqs2 when the reference database is limited. (**a**) Performance (%) of PhyloTune in identifying the smallest taxonomic unit on simulated datasets with varying training sequences. (Left) Taxonomic Classification: Matthews correlation coefficient (MCC), macro F1 score, macro recall, macro precision, and accuracy (ACC) for known taxa. (Right) Novelty Detection: AUPR and AUROC for unknown taxa. | (**b**) Comparison of taxonomic classification between PhyloTune and MMseqs2, which uses the training data as a reference database. Metrics include MCC, macro F1 score, macro recall, macro precision, ACC, and computational time (seconds).

2. Experimental Results

2.1 Identification of the Smallest Taxonomic Unit

Our findings indicate that both taxonomic classification and novelty detection performance improve steadily as the number of training samples increases (Figure 4a). For example, using 90 training samples resulted in a 35.5% improvement in the Matthews



Updated Figure 6. Difference in Robinson-Foulds (RF) distance between the tree constructed from high-attention regions and the ground truth tree (RF(high, gt)) and the RF distance between the tree constructed from low-attention regions and the ground truth tree (RF(low, gt)) at different numbers of sequences and lengths of attention regions (1/K of the full length) on the simulated datasets.

Correlation Coefficient (MCC) and a 25.1% increase in the AUROC compared to using only 10 training samples. These findings highlight the challenges associated with fine-tuning pretrained large language models like DNABERT-S on limited data, as the scarcity of training samples can lead to overfitting and hinder effective learning of distribution features.

Furthermore, when compared to MMseqs2 [3]—which uses training samples as its reference database—our method exhibits a clear advantage in classifying sequences belonging to known taxonomic units. Almost all evaluation metrics across all experimental settings demonstrate superior performance relative to MMseqs2, with the advantage being particularly pronounced in scenarios with fewer training samples (see Figure 4b). This observation suggests that deep learning approaches are better equipped to handle complex sequences that are less likely to be present in conventional reference databases, such as sequences involving long-range homology [4]. Moreover, MMseqs2 is inherently limited in its capacity to identify unknown taxonomic units because it lacks the mechanism to assess consistency between the query sequences and the sequences it identifies across different taxonomic levels.

2.2 Resolution of Phylogenetic Trees Using High-Attention Regions

We evaluated the accuracy of phylogenetic trees by calculating the normalized Robinson-Foulds (RF) distance between the trees constructed from attention-based

regions and the ground truth trees. In most cases, trees built from high-attention regions more closely resembled the ground truth trees than those built from low-attention regions. This trend was particularly pronounced under the $K = 3$ setting, where high-attention regions consistently outperformed low-attention regions across species trees of all sizes (n = $20, 40, 60, 80$, and 100; see Figure 6). However, when only a quarter of the sequence length was extracted as attention regions ($K = 4$), the accuracy of species trees of sizes 20 and 100 built from high-attention regions dropped significantly. These results underscore a trade-off between tree resolution and sequence length and indicate that while the attention mechanism offers valuable insights into the assessment of the informativeness of different positions in a sequence, determining the optimal fragment extraction strategy—both in terms of region selection and fragment length—remains a promising area for further research.

Conclusion

Our comprehensive benchmarking on simulated datasets demonstrates that our method effectively identifies the smallest taxonomic units and constructs accurate phylogenetic trees even under challenging few-shot settings. The ability to extract high-attention fragments not only enhances tree construction efficiency but also maintains high accuracy. We hope these additional results address the reviewer's concern and further demonstrate the value of our approach. The additional experiments have been added to the revised manuscript (lines 228-244, 277-290, 443-455). We deeply appreciate your thoughtful feedback and constructive suggestions.

For Reviewer #3,

We sincerely thank Reviewer #3 for taking the time to carefully read our revised manuscript and responses and for providing insightful and constructive feedback. Below, we provide point-by-point responses to the your comments, aiming to address the concerns regarding the generalizability and benchmarking of our method while further clarifying the contributions of our approach.

Comment 1

Regarding the evaluation of your model's generalizability, I was expecting to see the performance of the already trained model on unseen data. That means, the model trained on the PlantSeqs dataset should be evaluated on the new microbial data. Following this, you could fine-tune your model using the new microbial data and assess the performance again. This approach would allow you to evaluate how your pretrained model performs on unseen data with and without fine-tuning. However, it is not clear whether you fine-tuned the model based on the microbial data or trained it from scratch using the microbial data.

The authors mentioned that they used two HLPs to adapt to the new dataset. How does your model perform with only one HLP layer (as you used for the PlantSeqs dataset)? This suggests that the model was not fine-tuned based on the microbial data; instead, a slightly different architecture with two HLP layers was trained for the microbial dataset. This is concerning, as it indicates that the architecture may lack the expected generalizability of a pretrained model. Moreover, how would a user determine whether such adaptations in HLP layers are necessary for a new dataset? pretrained models are usually fine-tuned with new datasets while maintaining the same architecture.

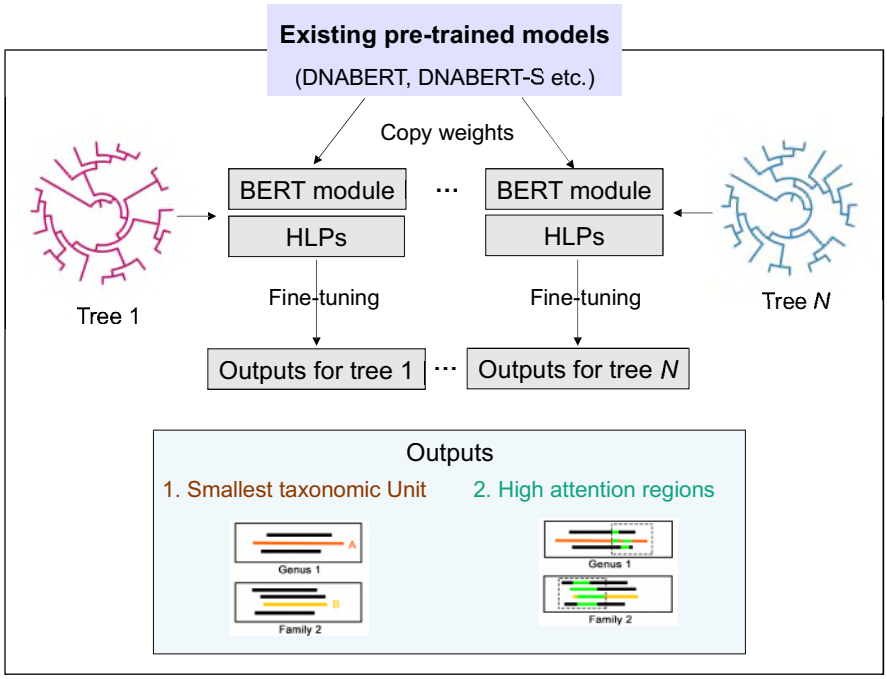
I understand the challenge, as I mentioned in my original comment, that phylogenomic dataset are heterogeneous. But a systematic assessment of your proposed model in this regard is essential.

Response 1

We sincerely appreciate reviewer's thoughtful comments and the opportunity to clarify our approach. We respond to your concerns about the generalizability of our method from the following points:

1. Clarification of Our Method

First, we would like to emphasize that our work does not propose a universal pretrained model; rather, it introduces a general fine-tuning framework for rapidly updating a given phylogenetic tree by leveraging existing DNA pretrained models (e.g., DNABERT [5], DNABERT-S [2]). Specifically, our method utilizes hierarchical linear probes (HLPs) tailored to the target phylogenetic tree. These HLPs are then connected to a pretrained model and fine-tuned to accurately identify the smallest taxonomic unit of the query sequence on the specified tree and to extract high-attention regions for all sequences within that branch (see Figure 1b in the revised manuscript).



Updated Figure 1b. Overview of PhyloTune. For a given phylogenetic tree requiring updates, hierarchical linear probes (HLPs) are specifically designed to align with its taxonomic hierarchy. These probes are fine-tuned on a pre-trained DNA model to accurately classify query sequences at the smallest taxonomic unit within the specified tree while extracting high-attention regions from all sequences within the corresponding clade.

Consequently, the same query sequence can be mapped to different taxonomic hierarchies when evaluated on models fine-tuned with different datasets (e.g., plant versus microbial), reflecting its position on distinct phylogenetic trees. For instance, a microbial test sample may be recognized as an unknown taxonomic unit at the class rank in the model fine-tuned using the plant dataset, yet classified into a known taxonomic unit in the model fine-tuned using the microbial dataset. It is important to note that all test data in our experiments remained unseen during training.

2. Design of HLPs

HLPs are designed to identify the smallest taxonomic unit of a sequence in a given phylogenetic tree, so the number of levels and the output dimensions at each level depend on the taxonomic hierarchy of the phylogenetic tree to be updated. For example, the PlantSeqs dataset (renamed Plant for consistency) contains four taxonomic ranks (class, order, family, and genus), so we employ four HLP layers—each

corresponding to a specific level for novelty detection and classification. The *Bordetella* dataset has only two levels, hence it uses two HLP layers. A single HLP layer would indicate that if a query sequence is identified as belonging to an unknown taxonomic unit, the entire tree must be updated. In contrast, multi-level HLPs can pinpoint the taxonomic unit at various hierarchical levels. This flexible design ensures that our method can accommodate phylogenetic trees with diverse topologies without resorting to a "one-size-fits-all" architecture.

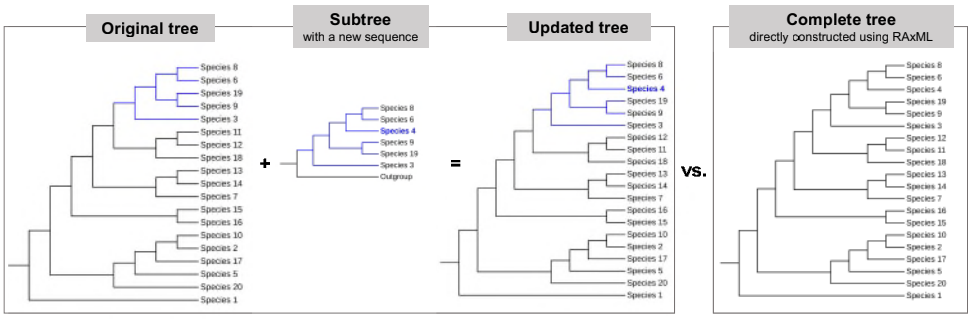
3. Generalizability of Our Method

To further investigate the generalizability of our method, we expanded our evaluation beyond plant (focusing on Embryophyta) and microbial (focusing on *Bordetella* genus) datasets by including assessments on simulated datasets. These simulations cover multiple replicates across different sequence counts ($20, 40, 60, 80$, and 100), with detailed experimental setups and results provided in the revised manuscript (lines 228-244, 277-290, 443-455). Comprehensive benchmarking on these datasets demonstrates that our method can effectively identify the smallest classification unit and construct accurate phylogenetic trees even under challenging few-shot settings. In particular, compared to MMseqs2—which relies on a reference database built from training samples—our method shows a clear advantage in classifying sequences from known taxonomic units, especially when training data are sparse. Furthermore, the ability to extract high-attention regions not only improves the efficiency of tree updates but also maintains high accuracy.

To avoid confusion and further clarify our contribution, we have renamed "BERT-Phylo" to "PhyloTune" in the revised manuscript, added a schematic diagram of our method in Figure 1b, and detailed the design of HLPs for different datasets in Section 4.2 (lines 500-517). We hope that these additional experiments and clarifications address the reviewers' concerns regarding generalizability and benchmarking. We greatly appreciate your constructive feedback.

Comment 2

As for benchmarking, thank you for your clarification regarding my original Comment 8. However, I believe it is important to make direct comparisons to traditional phylogenetic inference methods (like RAxML) with an entire set of sequences. You have provided comparisons between the performance with full-length sequences and high/low-attention regions (Fig 6). However, it is equally important to compare the results when considering all the sequences together and estimating trees using a traditional approach (e.g., RAxML) versus your approach of updating the existing tree with a new sequence. While the updating approach makes the overall pipeline faster, readers need to understand its accuracy compared to traditional approaches when applied to the entire set of sequences.



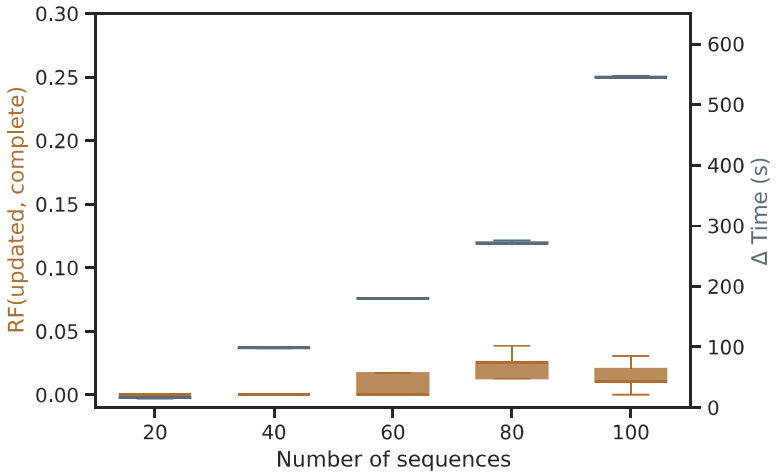
Updated Figure S1. Schematic diagram comparing the updated tree obtained by updating only a subtree of the original tree (marked in blue) and the complete tree reconstructed using all sequences.

Response 2

We appreciate this constructive comment and fully agree on the importance of comparing our subtree updating approach with traditional complete-tree reconstruction methods to assess its feasibility and limitations.

To address this, we conducted comparative experiments on simulated datasets. Specifically, we randomly generated phylogenetic trees of n species ($n = 20, 40, 60, 80, 100$) and simulated their sequence alignments using Seq-Gen v1.3.5 [1]. From each tree, we randomly selected a clade with at least four leaves as the "subtree" and designated one sequence within this subtree as the "new" sequence to be added. The remaining $n - 1$ sequences were considered the "original" sequences used to construct the "original" tree, while the tree reconstructed using all n sequences was termed the "complete" tree. We used RAxML to construct the original tree, subtree, and complete tree, then updated the original tree with the subtree to obtain the "updated" tree (see Figure S1). To quantify the differences between the updated and complete trees, we used the normalized Robinson-Foulds (RF) distance. For each n, we conducted five replicates with non-overlapping clades.

Our results show that for smaller datasets ($n = 20, 40$), the updated trees were topologically identical to the complete trees (Figure 2a, RF(updated, complete)). As the number of sequences increased, minor differences emerged. Specifically, for $n = 60, 80$, and 100, the average normalized RF distances were $0.007, 0.023$, and 0.014, respectively. This outcome is expected when adding new sequences, especially in more complex topologies [6, 7]. However, it is noteworthy that as n increases, the subtree-based approach achieves exponential gains in computational efficiency compared to complete-tree reconstruction (Figure 2a, $\Delta Time$). Thus, while focusing on subtree updates may overlook some global topological changes introduced by new sequences, it remains a widely used strategy to balance computational efficiency and accuracy, as demonstrated by methods such as pplacerDC [8] and SCAMPP [9]. Notably, the well-known APG phylogeny of angiosperms [10] was also constructed by iteratively



Updated Figure 2a. Robinson-Foulds (RF) distance (RF(updated, complete)) and tree-building time differences ($\Delta Time = Time_{\text{complete tree}} - Time_{\text{updated tree}}$) between the updated tree obtained by updating only the subtree and the complete tree built using all sequences on simulated datasets.

"connecting" subtrees. Phylogenetic inference is influenced by available data and computational constraints, making it challenging to obtain an absolutely "correct" result. However, compared to reconstructing an entire tree from scratch using all sequences, updating an existing tree through subtree reconstruction offers a balanced trade-off between efficiency and accuracy.

We have incorporated the above comparisons in the revised manuscript (lines 162-183) to illustrate both the feasibility and limitations of our subtree updating approach as an alternative to full-tree reconstruction. Once again, we appreciate your valuable suggestions.

Comment 3

I appreciate that the authors have changed the title of the manuscript by indicating that it "updates" an existing tree. However, I feel that the name "BERTPhylo" may not be appropriate and could be an overstatement. The authors are using DNABERT to identify the subtree to be updated and to identify high-attention regions in sequences. That does not make this method a BERT-like method. However, the name "BERTPhylo" suggests a BERT-like (eg., DNABERT, ProtBERT) model for phylogenetic tree estimation, which is not the case for your method.

Response 3

Thank you for this valuable feedback. We understand the concern that "BERTPhylo" in the original title might suggest we are providing a new BERT-like pretrained model specifically for phylogenetic tree updating. As mentioned in Response 1, our work

instead presents a general approach that leverages existing DNA pretrained models to accelerate the update of a given phylogenetic tree, rather than introducing a universal pretrained model. To better reflect the focus and contributions of our work, we have changed the term "BERTPhylo" to "PhyloTune" in the revised manuscript. We hope this adjustment prevents any confusion and more accurately conveys our approach and its impact. Thank you again for your constructive suggestion.

Minor comment

In section 4.2, the authors added a new sentence, "PlantSeqs dataset is fine-tuned based on DNABERT." But a dataset cannot be fine-tuned. Did you mean to say that DNABERT was fine-tuned based on PlantSeqs dataset?

Response

Thank you for pointing out this imprecision. We have revised the sentence to "we fine-tuned DNABERT for the Plant dataset", which more accurately conveys the original meaning.

References

[1] Rambaut, A. & Grass, N. C. Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Bioinformatics* **13**, 235–238 (1997).

[2] Zhou, Z. *et al.* Dnabert-s: Learning species-aware dna embedding with genome foundation models. *ArXiv* (2024).

[3] Hauser, M. *MMseqs: ultra fast and sensitive clustering and search of large protein sequence databases*. Ph.D. thesis, LMU, Munich (2014).

[4] Mock, F., Kretschmer, F., Kriese, A., Böcker, S. & Marz, M. Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. *Proceedings of the National Academy of Sciences, USA* **119**, e2122636119 (2022).

[5] Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

[6] Nabhan, A. R. & Sarkar, I. N. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in bioinformatics* **13**, 122–134 (2012).

[7] Bernot, J. P. *et al.* Major revisions in pancrustacean phylogeny and evidence of sensitivity to taxon sampling. *Molecular Biology and Evolution* **40**, msad175 (2023).

[8] Koning, E., Phillips, M. & Warnow, T. *ppiacerdc: a new scalable phylogenetic placement method*, 1–9 (2021).

[9] Wedell, E., Cai, Y. & Warnow, T. Scampp: scaling alignment-based phylogenetic placement to large trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **20**, 1417–1430 (2022).

[10] Group, A. P. *et al.* An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: Apg iv. *Botanical journal of the Linnean Society* **181**, 1–20 (2016).

Response to referees for PhyloTune: An Efficient Method to Accelerate Phylogenetic Updates Using a Pretrained DNA Language Model

Dear all reviewers,

We would like to express our sincere gratitude to the editors and reviewers for their continued time, effort, and constructive feedback throughout the review process. We are especially thankful for the encouraging comments from Reviewers #2 and #3, which have greatly contributed to improving the clarity and scientific rigor of our manuscript. Below we address the final set of comments from Reviewer #3 in detail.

For Reviewer #3,

Comment 1

Currently, Figure 2a shows the RF distances between the trees inferred using the proposed method (for both high-attention regions and full-length sequences) and the tree inferred from the complete set of sequences using RAxML. While this comparison is useful, a more informative approach would be to evaluate the methods directly against the true species tree, since this is a simulated dataset with a known ground truth. Otherwise, it is difficult to interpret the accuracy of the proposed method compared to RAxML. Specifically, I recommend revising Figure 2a to include the following: 1) RF distances between the trees obtained by your approach (high attention region) and the true tree, 2) RF distances between the trees obtained by your approach (full length) and the true tree, and 3) RF distances between the trees obtained from all sequences using RAxML and the true tree (this will be just a horizontal line as in Fig 2b since varying the number of sequences is not relevant here).

Response 1

We thank the reviewer for this insightful suggestion to improve the interpretability of the feasibility of updating existing trees through targeted subtree reconstruction. We have revised Figure 2 accordingly. Specifically, the updated Figure 2b now clearly illustrates the Robinson-Foulds (RF) distances between:

- Trees updated using high-attention regions (high tree) and the ground-truth tree (RF(high, gt));
- Trees updated using full-length sequences (full tree) and the ground-truth tree (RF(full, gt));
- Trees reconstructed from complete sequence datasets using RAxML (complete tree) and the ground-truth tree (RF(complete, gt)).

Additionally, the corresponding tree reconstruction times are shown in the updated Figure 2b, enabling a clearer joint analysis of accuracy and computational efficiency.

To clarify the experimental setup: the x-axis in updated Figure 2b,c represents the number of sequences in each randomly generated ground-truth tree. Each such tree yields a corresponding complete tree reconstructed using RAxML. For each ground-truth tree, we select five random, non-overlapping subtrees (each with ≥ 4 leaves), and use either full-length sequences or high-attention regions to reconstruct them. These updated subtrees are then merged with the rest branches to form the full tree and high tree, respectively (Figure 2a).

As shown in Figure 2b, both update strategies yield trees with correct topologies for smaller numbers of sequences (n = 20, 40). As tree complexity increases, topological discrepancies emerge. For n = 60, 80, and 100, the average normalized RF distances for full trees is 0.007, 0.046, and 0.027, respectively; for high trees, it is 0.021, 0.054, and 0.031. Notably, even complete trees reconstructed from complete sequences set show non-trivial discrepancies from the ground truth in more complex topologies (e.g., RF = 0.038 and 0.020 for n = 60 and 80, respectively), which is consistent with known challenges in reconstructing complex topologies [1, 2].

Importantly, as shown in Figure 2c, the time required for complete tree reconstruction grows exponentially with the number of sequences. In contrast, our subtree update strategy, which only rebuilds a small portion of the tree, makes update time relatively insensitive to the total number of sequences, significantly saving computational cost.

As discussed in the main text (now clarified in lines 180–185), this trade-off between speed and accuracy aligns with findings in prior work on subtree placement strategies such as pplacerDC [3] and SCAMPP [4], and mirrors the pragmatic design of large-scale phylogenies like the APG system for angiosperms [5], which are constructed by iteratively connecting subtrees. Taken together, these results underscore that subtree-based updates provide a practical and scalable solution that balances accuracy with efficiency in large phylogenetic analyses.

Comment 2

My second suggestion is to briefly discuss the RF distances of the high attention region approach. The authors observed that as the number of sequences increases, the RF distance between their method and RAxML increases (as I expected). But they reported the RF distances (0.007, 0.023 and 0.014) for only full-length sequences. For high attention region approach, the RF distances are even higher (more than 25% for 100 sequences). This is what I anticipated, and the authors should report and discuss this (but with respect to true trees as I suggested above).

Response 2

We appreciate the reviewer's valuable suggestion and fully agree that the RF distances of the high-attention region approach warrant explicit discussion. As expected, the RF distances for trees updated using high-attention regions are slightly higher than those based on full-length sequences, particularly as tree complexity increases (Figure 2b). For sequence counts of 60, 80, and 100, the average differences between RF(high, gt) and RF(full, gt) are 0.014, 0.008, and 0.004, respectively. Although these differences are relatively small, they illustrate the expected trade-off between reduced sequence length and topological accuracy.

On the other hand, the high-attention region approach further reduces computational cost, with observed time reductions of 30.3%, 30.1%, and 14.3% compared to updating with full-length sequences in the same settings (Figure 2c). This highlights the practical benefit of our method, which offers significant efficiency gains with only a modest compromise in accuracy.

We have revised Figure 2 and incorporated the above discussion into the manuscript (lines 167–182), highlighting that while our method delivers computational efficiency, it does involve a trade-off in topological accuracy under certain conditions. We believe this clarification strengthens the interpretation of our findings and provides a balanced perspective on the strengths and limitations of our approach. We are grateful for the reviewer's insightful and constructive suggestions, which have led to significant improvements in the manuscript. We hope that the current version fully addresses all remaining concerns.

References

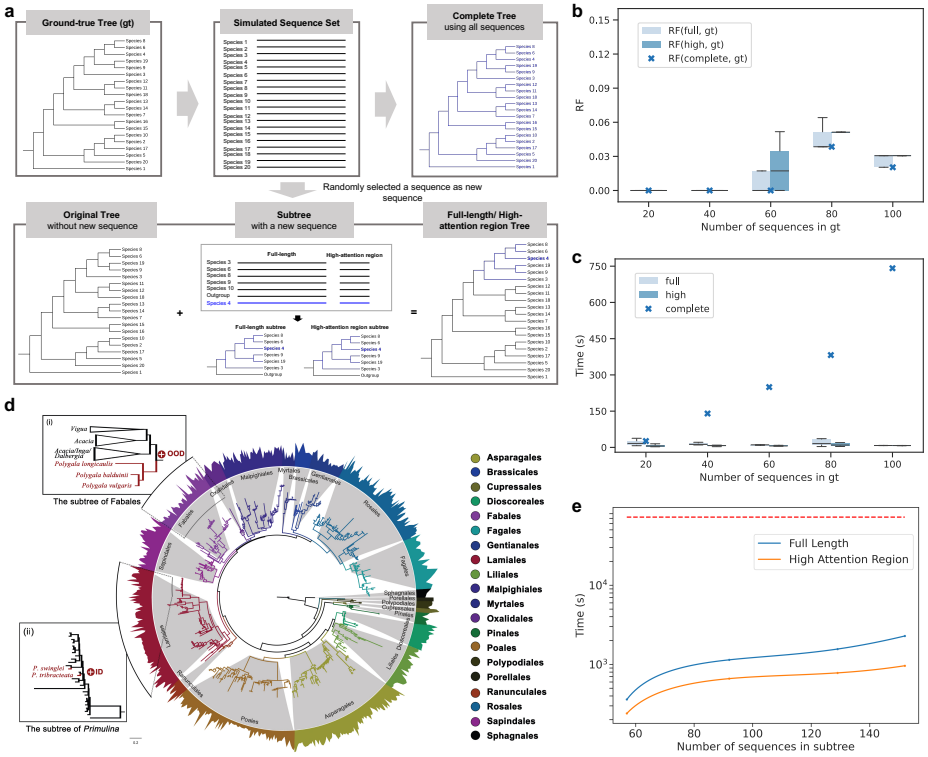
[1] Nabhan, A. R. & Sarkar, I. N. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Brief. Bioinform.* **13**, 122–134 (2012).

[2] Bernot, J. P. *et al.* Major revisions in pancrustacean phylogeny and evidence of sensitivity to taxon sampling. *Mol. Biol. Evol.* **40**, msad175 (2023).

[3] Koning, E., Phillips, M. & Warnow, T. pplacerDC: a new scalable phylogenetic placement method. *Proc. ACM Int. Conf. Bioinf. Comput. Biol. Health Inf.* 1–9 (2021).

[4] Wedell, E., Cai, Y. & Warnow, T. SCAMPP: scaling alignment-based phylogenetic placement to large trees. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **20**, 1417–1430 (2022).

[5] Group, A. P. *et al.* An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **181**, 1–20 (2016).



Updated Figure 2. Performance comparison of phylogenetic tree updating methods. (**a**) Schematic overview of tree reconstruction strategies. The original tree is built from sequences simulated on the ground-truth tree (gt) with one sequence removed as the new sequence. Updates are performed using: all sequences (complete tree), full-length sequences of a target subtree (full-length tree), or high-attention regions of the subtree (high-attention region tree). Using the example of the addition of species 4, the updated parts of the three trees using RAxML are highlighted in blue. | (**b,c**) Robinson-Foulds (RF) distance and construction time compare updated trees to gt. Each box plot (n=5 independent experiments) shows the median, interquartile range (25th to 75th percentile), and whiskers to minima/maxima within 1.5 times IQR. | (**d**) Example of updating the phylogenetic tree using PhyloTune. The original tree consisted of 677 species of 20 orders from Embryophyta. The tree was built using RAxML, with organisms colored based on order. The scale represents the normalized fraction of total branch length. The rugged bars at the outer circle represents the normalized length of input DNA sequences. *i*) Update of out-of-distribution (OOD) sequences: the three newly added sequences belong to the order Fabales, but do not belong to any families or genera in the original tree, so only the subtree of Fabales is updated. *ii*) Update of in-distribution (ID) sequences: the two newly added sequences belong to the genus *Primulina*, so the subtree of *Primulina* is updated. | (**e**) Time comparison for the example tree. Blue and orange curves show subtree reconstruction times using full-length sequences and high-attention regions (one-third of the full length), respectively. The red dotted line indicates the time needed to update the tree using all sequences (about 20.1 hours). Source data are provided as a Source Data file.