IQ-NET: A Deep Learning Approach for Fast and Accurate Phylogenetic Inference from Real Alignments

Chen Yang[1*], Zixin Zhuang[1], Piyumal Demotte[1], Cuong Cao Dang[2], Le Sy Vinh[2], Bui Quang Minh[1], Nhan Ly-Trong[1]

[1]School of Computing, Australian National University, Canberra, 2600, ACT, Australia
[2]Faculty of Information Technology, University of Engineering and Technology, Vietnam National University, Hanoi, 144 Xuan Thuy, Cau Giay, 10000 Hanoi, Vietnam

*Corresponding author. E-mail: chen.yang1@anu.edu.au

Abstract

Phylogenetic inference is fundamental to modern biology, with applications spanning evolutionary biology, epidemiology, and comparative genomics. While maximum likelihood and Bayesian methods remain the gold standard due to their statistical rigor, they rely on simplifying evolutionary assumptions and are computationally intensive. Existing machine learning approaches offer speed advantages, but face several limitations: exclusive reliance on simulated training data, inadequate handling of gaps, focus primarily on topology rather than complete tree reconstruction, and sensitivity to input sequence order. Here, we introduce IQ-NET (Intelligent Quartet NETwork), a machine learning framework that addresses these limitations through training exclusively on real datasets, simultaneous inference of topology and branch lengths from gapped alignments without substitution model assumptions, and robustness to the order of input sequences. IQ-NET outperforms existing machine learning methods and achieves both higher accuracy and a 24-fold speedup over the IQ-TREE software. We also demonstrate IQ-NET's utility in species tree reconstruction by integrating it with ASTRAL.

Keywords: Phylogenetic inference, Machine learning, Quartet analysis, Empirical data training

1 Introduction

Phylogenetic inference reconstructs evolutionary relationships from molecular sequence data, serving as a cornerstone of modern biology with applications

spanning evolutionary biology, epidemiology, ecology, and comparative genomics (Felsenstein, 2004; Grenfell et al., 2004; Delsuc et al., 2005). These methods have elucidated species origins across diverse taxa and proven essential for understanding evolutionary history and contemporary challenges, as exemplified in tracking SARS-CoV-2 emergence and variants during the COVID-19 pandemic (Li et al., 2020; Hodcroft et al., 2021; Attwood et al., 2022; Turakhia et al., 2021; Gómez-Carballa et al., 2020).

Maximum likelihood and Bayesian inference are the gold standard for phylogenetic inference due to their statistical rigor and accuracy, as implemented in software such as IQ-TREE, RAxML, PHYML, MrBayes, and BEAST (Nguyen et al., 2015; Stamatakis, 2014; Guindon and Gascuel, 2003; Ronquist and Huelsenbeck, 2003; Drummond and Rambaut, 2007). However, these methods rely on substitution models with simplifying assumptions - stationarity, reversibility, and homogeneity (Felsenstein, 2004; Yang and Rannala, 2012; Jermiin et al., 2016) - that may fail to present complex evolutionary processes such as incomplete lineage sorting, hybridization, and recombination. Moreover, they are computationally intensive (Izquierdo-Carrasco and Stamatakis, 2011; Stamatakis, 2006), taking days or even months to analyze large datasets.

Machine learning offers promising alternatives for phylogenetic inference with potential for rapid analysis once trained. However, existing approaches (Suvorov et al., 2020; Zou et al., 2020; Wang et al., 2023; Smith and Hahn, 2023; Suvorov and Schrider, 2024; Kulikov et al., 2024; Nesterenko et al., 2025) still face several limitations. These methods rely exclusively on simulated training data, potentially limiting generalization to empirical datasets as demonstrated by Zhu et al. (2025). Many methods (Kulikov et al., 2024; Zou et al., 2020) either ignore or fail to handle gaps, despite their prevalence in real alignments. Several studies (Suvorov et al., 2020; Zou et al., 2020; Wang et al., 2023) focus primarily on topology inference, with only one (Suvorov and Schrider, 2024) addressing branch length estimation. Moreover, most models use neural network architectures originally designed for image recognition, making their predictions sensitive to the order of input sequences.

To address these limitations, we introduce IQ-NET (Intelligent Quartet NETwork), a machine learning framework for complete phylogenetic reconstruction of four-taxon trees directly from multiple sequence alignments. IQ-NET advances the field through several key innovations: (1) exclusive training and testing on real datasets with generalization validated on independent TreeBase dataset, (2) simultaneous inference of both topology and branch lengths from gapped alignments without assuming any substitution model, (3) inherent permutation invariance for consistent predictions regardless of sequence order, and (4) superior performance compared to existing machine learning methods, achieving higher accuracy than IQ-TREE with a 24-fold speedup that reduces runtime from 2.67 hours to 6.7

minutes when reconstructing over 50,000 quartet trees. We further demonstrate IQ-NET's utility for species tree reconstruction by integrating it with ASTRAL (Mirarab et al., 2014).

2 Materials and Methods

IQ-NET contains two key components: (1) a Tree topology classifier and (2) a Branch length regressor. This section outlines the development of these components through four main steps, detailed below.

2.1 Data Generation

To address the limited generalisation of machine learning models trained on simulated data, as recently highlighted by Zhu et al. (2025), we trained and tested IQ-NET on real data from the EvoNAPS database (`https://github.com/Cibiv/EvoNAPS`)and conducted extensive tests on the independent TreeBASE dataset (Piel et al., 2002).

The Empirical EvoNAPS Database

The EvoNAPS database comprises empirical alignments, their corresponding phylogenetic trees, and model parameters inferred with IQ-TREE (v2.2.0.5). Alignments were collected from published resources, including BenchmarkAlignments (`https://github.com/roblanf/BenchmarkAlignments/`), PANDIT (Whelan et al., 2006), OrthoMaM (Scornavacca et al., 2019), and TreeBASE (Piel et al., 2002). EvoNAPS includes both DNA and protein alignments and represents a diverse range of species, from microbes and fungi to plants, birds, turtles, and mammals.

The database contains 48,706 DNA alignments. The number of taxa per alignment ranges from 4 to 2,957, and sequence lengths range from 12 to 127,813 sites. Then we removed 9,998 alignments derived from TreeBASE, reserving TreeBASE exclusively as an independent dataset for evaluating model generalisation. Since EvoNAPS includes two versions (v10c and v12a) from the OrthoMaM database, we also removed all 14,509 alignments from v10c to avoid duplication with v12a.

Data Cleaning

To ensure data quality, we applied two filtering criteria. First, we excluded trees containing branches longer than 9 substitutions per site, as these often indicate problematic sequence alignments where nucleotides at the same site are not homologous (Mai and Mirarab, 2018). Second, we excluded alignments shorter than 200 sites, as these typically lack sufficient evolutionary signal for reliable phylogenetic inference and may generate erroneous phylogenies that introduce noise during model training. Additionally, all gap representations in the alignments ('-', 'N', and '.') were standardised to '-' for consistency.

Sampling quartet trees and alignments

The remaining phylogenetic trees were divided into independent training, validation, and testing sets with a ratio of 80:10:10. From each tree with N taxa, we randomly sampled $\max(1, \min(30, \frac{N}{4} - 1))$ quartet subtrees along with their corresponding sub-alignments. The factor $\frac{N}{4}$ represents the maximum number of unique 4-taxon sets that can be extracted from N taxa, helping to reduce the chance of repetition. To avoid oversampling from large trees and enhance dataset diversity, we limited the number of subtrees extracted per original tree to 30. The final dataset comprised 415,303 training, 51,401 validation, and 51,241 testing samples.

Feature Extraction

To encode the input alignments, we extracted site pattern frequencies as features, following the approach proposed in Leuchtenberger et al. (2020). This encoding approach is widely adopted in maximum likelihood-based phylogenetic methods, such as IQ-TREE, RAxML, and PhyML. Specifically, DNA sequences are represented using five characters: the four nucleotides (A, C, G, T) and the gap ('-'). For an alignment of four taxa, this results in $5^4 = 625$ possible site patterns. We counted the occurrences of each site pattern in the alignment and normalised them by the alignment length, ensuring that the resulting frequency vector sums up to one.

2.2 Model Design and Implementation

Let $[S_A, S_B, S_C, S_D]$ denote the input sequences representing the four taxa A, B, C, and D; and ψ denote the data encoder.

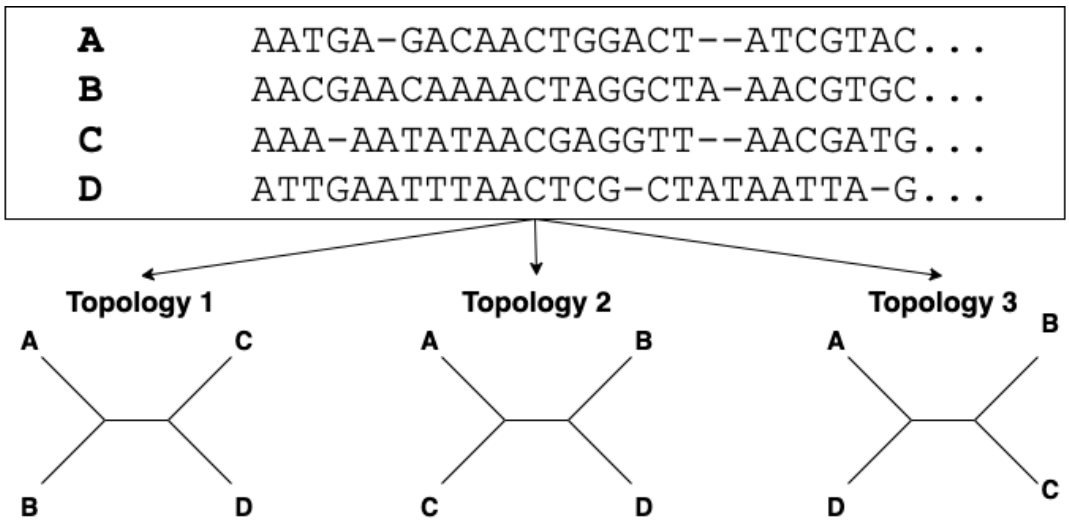

Figure 1: Three possible unrooted topologies for four taxa A, B, C, and D. Topology 1: $AB|CD$; Topology 2: $AC|BD$; and Topology 3: $AD|BC$.

2.2.1 Tree Topology Classifier

For any four-taxon set $\{A, B, C, D\}$, there are three possible unrooted tree topologies (Figure 1): Topology 1: $AB|CD$; Topology 2: $AC|BD$; and Topology 3: $AD|BC$.

Sequences within an alignment can appear in any order, creating $4! = 24$ possible permutations representing the same four-taxon alignment. These permutations may produce different site pattern frequencies, potentially affecting model predictions. To ensure that the classifier consistently returns the correct topology regardless of sequence order, we adopted the symmetry-preserving architecture proposed by Solís-Lemus et al. (2024).

The Adapted Symmetry-preserving Design

Permuting the input sequences may require tree topology changes to reflect the new ordering. However, not all permutations alter the underlying topology. For example, reordering the alignment from $[S_A, S_B, S_C, S_D]$ to $[S_B, S_A, S_C, S_D]$ preserves Topology 1: $AB|CD$.

To account for this, we identified permutation sets that preserve each topology. Let P_0 denote the set of permutations that preserve all three topologies; and P_1, P_2, and P_3 denote permutations preserving topologies 1, 2, and 3, respectively.

$P_0 = \{[S_A, S_B, S_C, S_D], [S_B, S_A, S_D, S_C], [S_C, S_D, S_A, S_B], [S_D, S_C, S_B, S_A]\}$

$P_1 = \{[S_A, S_B, S_D, S_C], [S_B, S_A, S_C, S_D], [S_C, S_D, S_B, S_A], [S_D, S_C, S_A, S_B]\}$

$P_2 = \{[S_A, S_D, S_C, S_B], [S_B, S_C, S_D, S_A], [S_C, S_B, S_A, S_D], [S_D, S_A, S_B, S_C]\}$

$P_3 = \{[S_A, S_C, S_B, S_D], [S_B, S_D, S_A, S_C], [S_C, S_A, S_D, S_B], [S_D, S_B, S_C, S_A]\}$

The feature representation F_i (where $0 \leq i \leq 3$) is computed by averaging the encoded features over all permutations in P_i:

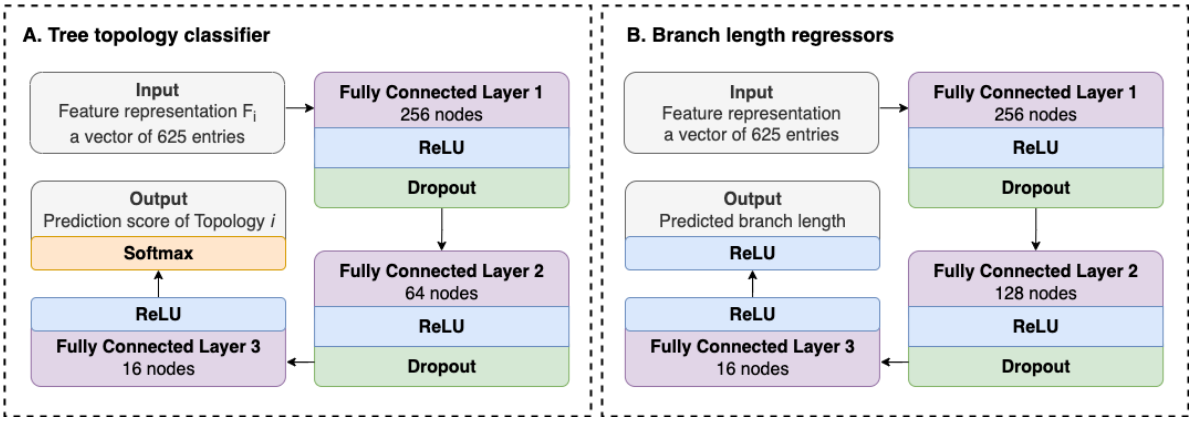

Figure 2: The network architecture of (A) the tree topology classifier; and (B) the branch length regressors.

$$F_i = \frac{1}{4} \sum_{p \in P_i} \psi(p)$$

Let f denote the classifier. The predicted score for Topology i (where $1 \leq i \leq 3$) is then obtained by applying f to the average of F_0 and F_i:

$$T_i = f\left(\frac{1}{2}(F_0 + F_i)\right)$$

This symmetry-preserving design ensures that the classifier produces consistent topology scores regardless of the input sequence order.

The Network Architecture

Figure 2A illustrates the architecture of the tree topology classifier, a five-layer fully-connected neural network. The network takes the feature representation F_i (see Section 2.2.1) - a 625-dimensional vector - as input. That input is processed through three hidden layers with decreasing dimensionality of 256, 64, 16, before outputting a prediction score for the Topology i. For each input alignment, the network is applied three times to predict the scores for all three possible topologies. Dropout regularization is applied at the first two hidden layers to reduce overfitting. The hidden layers use rectified linear unit (ReLU) activation functions (Glorot et al., 2011), while the output layer employs softmax activation to convert raw network outputs into class probabilities (Goodfellow et al., 2016).

Hyperparameter Tunning

We employed Optuna (Akiba et al., 2019) to fine tune the hyperparameters of our classifier through 200 trials. Each trial samples one candidate configuration from predefined parameter ranges. The configuration yielding the lowest validation loss was selected for final model training. Supplementary Table S1 summarises the

search ranges and the best values.

2.2.2 Branch Length Regressor

The branch length regressor estimates five non-negative values: one internal and four external branch lengths. The internal branch length remains unchanged for any input sequence order, while the external lengths should permute accordingly. To enforce these constraints, we implemented separate regressors for the internal and external branches.

Let Q_i (where $i \in \{A, B, C, D\}$) denote the set of sequence permutations where S_i appears first in the alignment. For a four-taxon alignment, each set Q_i contains exactly six permutations. For instance, $Q_A = \{[S_A, S_B, S_C, S_D], [S_A, S_B, S_D, S_C], [S_A, S_C, S_B, S_D], [S_A, S_C, S_D, S_B], [S_A, S_D, S_B, S_C], [S_A, S_D, S_C, S_B]\}$.

The feature representation F_i (where $i \in \{A, B, C, D\}$) is computed by averaging the encoded features over all permutations in Q_i:

$$F_i = \frac{1}{6} \sum_{q \in Q_i} \psi(q)$$

Let g and h denote the regresors for the internal and external branches, respectively.

The internal branch length L_{int} is predicted as:

$$L_{int} = g\left(\frac{1}{4} \sum_{i \in \{A,B,C,D\}} F_i\right)$$

The external branch length L_i (where $i \in \{A, B, C, D\}$) is predicted as:

$$L_i = h(F_i)$$

This design ensures the predicted branch lengths remain consistent regardless of the input sequence order.

The Network Architecture

Figure 2B depicts the architecture of the branch length regressors, implemented as a five-layer fully connected neural network. This design resembles the topology classifier but with two key differences. The three hidden layers have sizes of 256, 128 (instead of 64), and 16, respectively. ReLU activation functions are applied to both hidden and output layers.

Hyperparameter Tunning

Similar to the tree topology classifier, we employed Optuna to fine-tune the hyperparameters of the branch length regressors. The search ranges and the best

hyperparameter settings are summarised in Supplementary Table S2.

2.3 Model Training

All networks were trained using the Adam optimiser (Kingma and Ba, 2014) with exponentially decaying learning rates. Cross-entropy and mean squared error (MSE) were used as the loss functions for the tree topology classifier and the branch length regressor, respectively. To mitigate overfitting, we implemented early stopping, terminating training if the validation accuracy failed to improve for 5 consecutive epochs.

Training was conducted on the Gadi supercomputing system at the National Computational Infrastructure (NCI), Australia (`https://nci.org.au/`), using an NVIDIA Tesla V100-SXM2-32GB GPU.

2.4 Evaluation

We conducted an extensive evaluation of IQ-NET through three key assessments. First, we benchmarked it against IQ-TREE and existing machine learning methods using the testing set from the EvoNAPS database. Second, we evaluated IQ-NET on an independent TreeBASE dataset to assess its generalization. Finally, we demonstrated a practical application of IQ-NET by providing quartet trees to ASTRAL for species tree reconstruction using the Turtle dataset (Chiari et al., 2012).

2.4.1 Benchmark IQ-NET against IQ-TREE and Existing Machine Learning Methods

We benchmarked IQ-NET against IQ-TREE and several state-of-the-art machine learning methods. For topology prediction, comparisons included Fusang (Wang et al., 2023), DeepNNPhylogeny (Kulikov et al., 2024), and the model by Suvorov et al. (2020) (hereafter referred to as 'Suvorov-topology'). Since DeepNNPhylogeny does not support gapped alignments, all gapped sites were removed prior to its evaluation. For branch length estimation, we compared IQ-NET with IQ-TREE. We did not benchmark the published machine learning model by Suvorov and Schrider (2024), because that study provides a collection of many pretrained models - each trained on data simulated under specific conditions - rather than a single general-purpose model. Consequently, no single pretrained model could be appropriately selected for real datasets.

Benchmarks were performed on the testing dataset extracted from the EvoNAPS database (see Section 2.1). EvoNAPS contains maximum likelihood

trees inferred from empirical alignments. We randomly subsampled quartet trees from these original maximum likelihood trees and used them as ground truth. This approach leverages additional phylogenetic information from the original many-taxon alignment to produce more reliable phylogenetic trees compared to direct inference from four-taxon alignments. In contrast, when IQ-TREE served as a benchmark method, it directly inferred four-taxon trees from four-taxon alignments; we refer to this as IQ-TREE-quartet to avoid confusion.

All benchmarks were conducted on Gadi using a single Intel i5-14600 CPU with 4 GB allocated RAM.

2.4.2 Evaluate IQ-NET's Generalization on the Independent TreeBASE Dataset

To further assess the generalization of IQ-NET, we evaluated it on an independently TreeBASE dataset (Piel et al., 2002). Noting that all TreeBASE-derived data were removed from EvoNAPS during the development of IQ-NET, ensuring complete independence between our training and testing data. We then downloaded all TreeBASE studies, but filtered out those containing multiple trees, non-DNA data, less than four sequences, or more than 40 sequences. The resulting set consists of 921 trees and their corresponding alignments. For computational feasibility, we generated three testing subsets by randomly sampling 0.1%, 1%, and 10% of all possible quartets from each original TreeBASE alignment, yielding 62,494, 234,151, and 2,336,286 quartets, respectively. Since many TreeBASE trees do not include branch lengths, we assessed IQ-NET's performance only on topology prediction.

2.4.3 Reconstruct Species Trees from Quartet Trees using the Turtle Dataset

We demonstrate the application of IQ-NET for species tree construction using a phylogenomic data set comprising 16 vertebrate taxa and 248 genes (Chiari et al., 2012). This data set contains approximately 187,000 base pairs and focuses on resolving the phylogenetic placement of turtles relative to birds and crocodiles. The original study demonstrated that the inferred relationships among crocodiles, birds, and turtles are sensitive to the choice of phylogenetic model: under single-site homogeneous DNA substitution models, crocodiles and turtles form a clade, whereas partitioned models group birds and crocodiles as a sister clade to turtles. Here, we assess IQ-NET's ability to recover the relationships among crocodiles, birds, and turtles.

For the analysis, we subsample 4-taxon alignments from each gene. Given a gene contains N species, we consider all possible combinations of 4-taxon alignments $\binom{N}{4}$

and randomly subsample a proportion of these combinations without replacement ranging from 10% to 100%. Sampling x% of all possible 4-taxon alignments from a gene results in $\frac{x}{100} \times \binom{N}{4}$ alignments. For each sub-sampled alignments, IQ-NET is used to infer a corresponding quartet tree. The resulting quartet trees are subsequently provided as input to ASTRAL version 5.7.8 (Mirarab et al., 2014) for species tree construction. To examine the effect of third codon positions, we repeated the analysis after removing the third codon position from each gene, reconstructed quartet trees, and inferred species trees using ASTRAL with these modified quartet trees.

3 Results

The tree topology classifier converged at epoch 13 and completed training after 18 epochs (Suppl. Figure S1A), while the branch length regressor converged at epoch 2 and finished training after 7 epochs (Suppl. Figure S1B).

3.1 Benchmark Results

We first compared the accuracy and runtime of IQ-NET with IQ-TREE-quartet (IQ-TREE inferring trees directly from quartet alignments) and existing machine learning methods. This benchmark used the testing set extracted from the EvoNAPS database, where the ground-truth quartet trees were subsampled from larger maximum likelihood trees inferred by IQ-TREE from the original multi-taxon alignments (see Section 2.1).

3.1.1 IQ-NET Outperforms Existing Methods in Tree Topology Prediction

IQ-NET achieved the highest accuracy in predicting tree topologies, with an average accuracy of 82.3%, followed by IQ-TREE-quartet (79.9%), DeepNNPhylogeny (79.4%), Fusang (76.6%), and Suvorov-topology (57.7%). Moreover, IQ-NET demonstrated balanced performance across all three topologies (82-83% accuracy) without bias toward any particular topology (Suppl. Figure S2).

We also evaluated the impact of internal branch length and sequence length on prediction accuracy. Generally, increasing either internal branch length or sequence length tends to improve the accuracy of all methods (Figure 3), as longer branches and sequences provide more evolutionary signals (i.e., more mutations). However, accuracy declined when the internal branch length exceeded 0.24 or when the sequence length reached 3,950 sites. A possible explanation for this

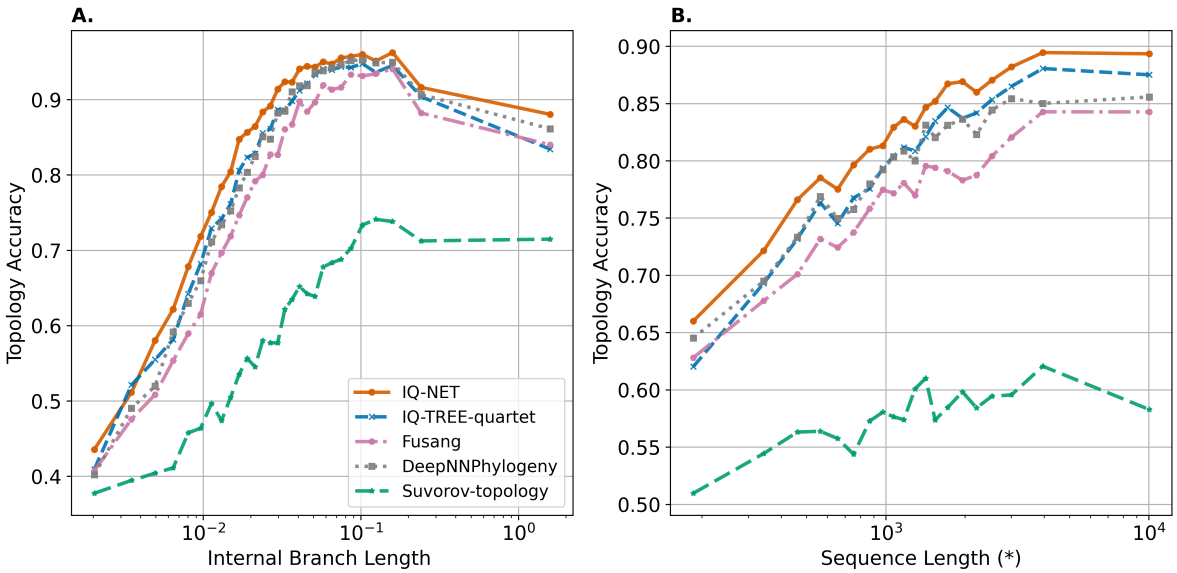

Figure 3: Topology prediction accuracy vs. (A) internal branch length; and (B) sequence length. (*) Sequence length excludes fully gapped sites, which contain no evolutionary signal.

phenomenon lies in the characteristics of our testing dataset: samples with long internal branches often derive from short sequences (Suppl. Figure S3), making topology reconstruction more difficult. Conversely, long sequences in our dataset tend to be highly similar, resulting in short internal branch lengths (Suppl. Figure S3) that also challenge the topology inference. Another explanation is that long sequences may contain multiple genes that have evolved independently under different evolutionary processes, potentially supporting multiple tree topologies rather than a single one.

3.1.2 IQ-NET Excels in Branch Length Prediction

Figure 4 shows scatter plots of true versus predicted branch lengths estimated by IQ-NET and IQ-TREE-quartet. Branch lengths predicted by IQ-NET closely match the true values, with a Pearson correlation of 0.9007 and a slope of 0.8112, compared with 0.6754 and 0.6905 for IQ-TREE-quartet.

Supplementary Table S3 presents the estimation errors of IQ-NET and IQ-TREE-quartet. IQ-NET outperformed IQ-TREE-quartet on most metrics, achieving lower Mean Squared Error (MSE), Mean Absolute Error (MAE), and Branch Score Distance (BSD) Kuhner and Felsenstein (1994). However, IQ-TREE-quartet obtained a lower Mean Relative Error (MRE). Since MRE is computed by dividing the absolute error by the true value, shorter branches have a greater influence on this metric. Thus, the lower MRE suggests that

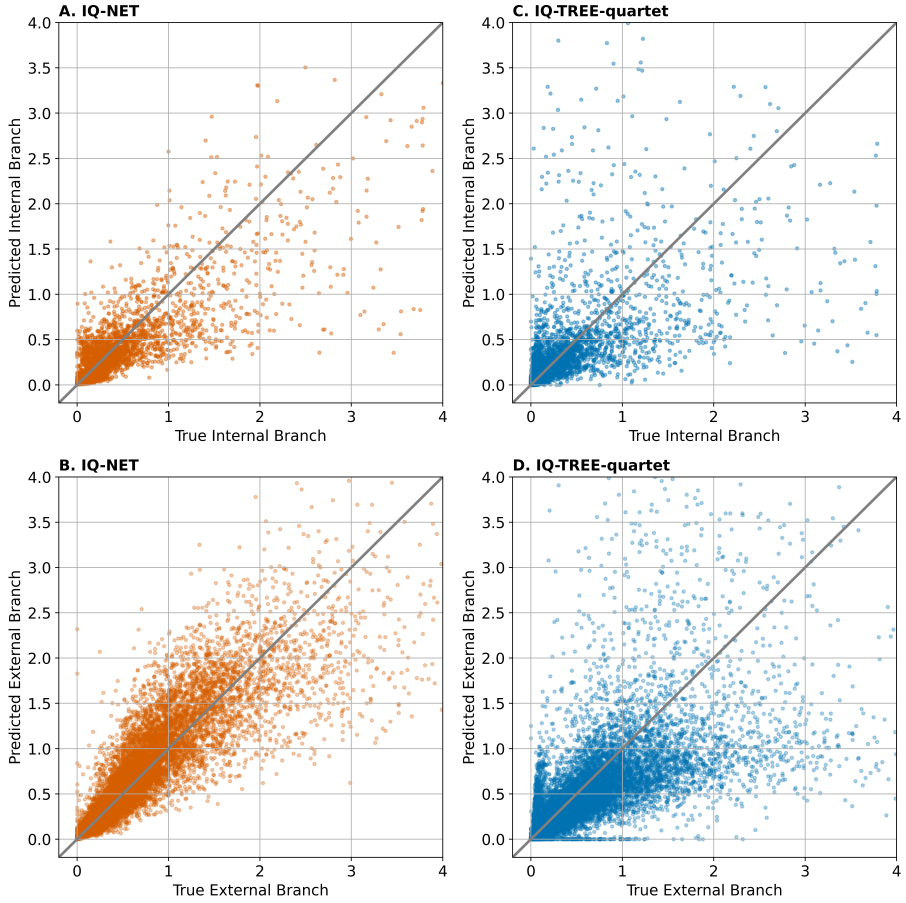

Figure 4: Scatter plots of True vs Predicted branch lengths estimated by IQ-NET (Panels A and B) and IQ-TREE-quartet (Panels C and D). Panels A and C correspond to internal branches, whereas B and D correspond to external branches.

IQ-TREE-quartet is more accurate at predicting shorter branches.

We also examined the influence of branch length and sequence length on branch length estimation. Increasing branch length tends to increase prediction error for both methods (Suppl. Figure S4A). IQ-NET performs comparably to IQ-TREE-quartet on short branches but clearly surpasses IQ-TREE-quartet on longer ones (> 0.02).

< 0.3.

In contrast, increasing sequence length tends to enhance the accuracy of both methods (Suppl. Figure S4B). For short alignments, IQ-NET is more accurate than IQ-TREE-quartet, while for long alignments ($> 1,000$ sites), the performance gap becomes negligible.

3.1.3 IQ-NET Achieves 24-Fold Speedup in Quartet Tree Inference

We compared the runtime of IQ-NET and IQ-TREE-quartet since they are the only two methods among our benchmarks that can perform complete quartet tree

inference, including branch length estimation. IQ-NET required 6.7 minutes to reconstruct 51,241 quartet trees compared with 2.67 hours for IQ-TREE, a 24-fold speedup.

For fairness, it should be noted that IQ-NET, as a machine learning method, required initial training, which took 13.75 and 15.37 GPU minutes for the tree topology classifier and branch length regressors, respectively.

3.2 IQ-NET Demonstrates Strong Generalization on the Independent TreeBASE Dataset

The prediction accuracy of tree topology of IQ-NET on the TreeBASE dataset are compareable to that of IQ-TREE-quartet. For example, on 1% TreeBASE dataset, IQ-NET has 83.7% accuracy, IQ-TREE-quartet has 83.9% accuracy. Fusang and DeepNNPhylogeny have lower accuracies of 78.2% and 81.7%, respectively. Please check Table 1 for detail results on 0.1%, 1%, and 10% of TreeBASE dataset. These results not only demonstrate the superior performance of IQ-NET in quartet topology prediction but, more importantly, highlight its ability to generalize to an empirical dataset independent of its training data.

TreeBASE percentage	0.1%	1%	10%
IQ-NET	**86.9%**	83.7%	82.8%
IQ-TREE-quartet	85.9%	**83.9%**	**83.7%**
Fusang	82.2%	78.2%	78.2%
DeepNNPhylogeny	84.5%	81.7%	81.7%
Suvorov-topology	75.7%	73.2%	73.2%

Table 1: The topology prediction accuracies of IQ-NET, IQ-TREE-quartet, Fusang, DeepNNPhylogeny, and Suvorov-topology on the independent TreeBASE dataset. Best methods are highlighted in bold.

3.3 IQ-NET + ASTRAL Recovers Turtle Phylogeny Consistent with Existing Studies

We compared species trees inferred using several approaches: (i) **IQ-NET + ASTRAL**, (ii) **concatenation-based analysis with IQ-TREE**, (iii) **ASTRAL** using gene trees inferred with IQ-TREE and ModelFinder, and (iv) **ASTRAL** using gene trees inferred with IQ-TREE and MixtureFinder. When all three codon positions were included, the IQ-NET + ASTRAL approach recovered turtles as the sister group to crocodiles, with this clade forming a sister group to birds (Supplementary Figure S5B). This topology was also recovered by the

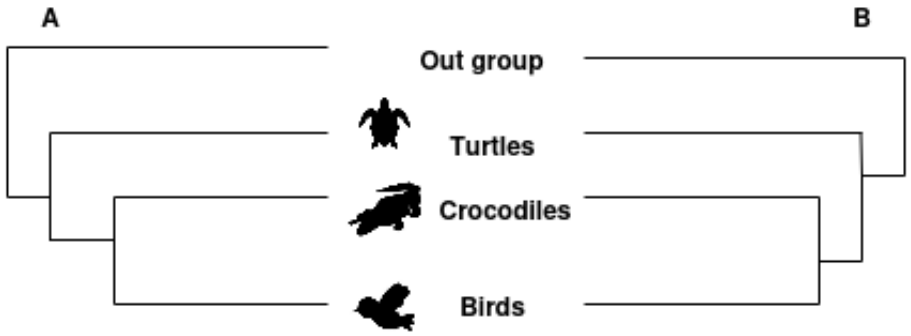

Figure 5: (A) Accepted tree for placement of turtles as sister clade to birds and crocodiles(Chiari et al., 2012). (B) Tree constructed under methods of IQ-NET+ASTRAL after removing third codon positions, ASTRAL using gene trees estimated with IQ-TREE and MixtureFinder.

concatenation-based method and by ASTRAL when using gene trees inferred under site-homogeneous models with ModelFinder.

To further examine the effect of codon positions, we repeated the analysis after removing the third codon position from every gene as recommended by Chiari et al. (2012). Quartet trees were estimated with IQ-NET from the reduced alignments, and a species tree was subsequently inferred with ASTRAL. Under this setting, turtles were placed as the sister clade to the combined group of birds and crocodiles (Figure 5B). This topology is congruent with the results obtained from ASTRAL using gene trees estimated with MixtureFinder, as well as with the conclusions of the original study (Chiari et al., 2012). Notably, subsampling only 10% of the possible quartet trees was sufficient to recover this relationship, highlighting the efficiency of the IQ-NET + ASTRAL approach.

4 Conclusion and Future Work

We present IQ-NET, an end-to-end machine learning framework for quartet tree reconstruction. Unlike traditional likelihood-based approaches that rely on explicit substitution models, IQ-NET learns evolutionary relationships directly from empirical data, enabling joint inference of tree topology and branch lengths from gapped alignments without model assumptions. By adapting the symmetry-preserving architecture proposed by Solís-Lemus et al. (2024), IQ-NET ensures consistent predictions regardless of sequence order. Trained and tested exclusively on empirical datasets with the generalization validated on the independent TreeBase benchmark, IQ-NET demonstrates superior accuracy over existing machine learning methods and outperforms IQ-TREE in both accuracy and runtime. Notably, IQ-NET reconstructed over 50,000 trees in only 6.7 minutes,

compared to 2.67 hours with IQ-TREE - a 24-fold speedup.

Despite these advances, there remains room for further development. First, complex evolutionary processes such as incomplete lineage sorting and introgression cannot be fully captured by a single phylogenetic tree. Since IQ-NET outputs scores for all three quartet topologies, these values could potentially serve as support measures, offering a foundation for detecting signals of incomplete lineage sorting or introgression. However, systematic validation is required to confirm this utility.

Second, the accuracy and generalization of machine learning methods remain dependent on training data quality. Expanding training to larger and more diverse empirical datasets may further improve IQ-NET's performance.

Third, while we have demonstrated integration with ASTRAL, IQ-NET could also be combined with other quartet puzzling approaches (Strimmer and von Haeseler, 1996; Schmidt et al., 2002) to scale reconstruction to larger trees.

Finally, alternative machine learning architectures such as transformer, which are not constrained by fixed input size, may enable direct extension of IQ-NET to larger tree reconstruction, while ensemble learning could enhance robustness and stability across diverse evolutionary scenarios.

Data Availability

All source code and scripts used in the development of IQ-NET, including those for data generation, training, testing, and the pre-trained models, are publicly available at `https://github.com/Clipper1331757/IQ_Net/`.

The data underlying this study are available in the Supplementary Material and the Zenodo Repository, at `https://link.com`.

Acknowledgements

This research was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government. We thank Professor Robert Lanfear, Dr. Thomas Wong, and Hashara Kumarasinghe for their valuable suggestions and discussions. We used AI tools to assist in polishing the writing of this manuscript.

Funding

References

Akiba T, Sano S, Yanase T, et al (2019) Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, pp 2623–2631, https://doi.org/10.1145/3292500.3330701

Attwood SW, Hill SC, Aanensen DM, et al (2022) Phylogenetic and phylodynamic approaches to understanding and combating the early sars-cov-2 pandemic. Nature Reviews Genetics 23(9):547–562. https://doi.org/10.1038/s41576-022-00483-8

Chiari Y, Cahais V, Galtier N, et al (2012) Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (archosauria). BMC Biology 10:65. https://doi.org/10.1186/1741-7007-10-65

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics 6(5):361–375. https://doi.org/10.1038/nrg1603

Drummond AJ, Rambaut A (2007) Beast: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7:214. https://doi.org/10.1186/1471-2148-7-214

Felsenstein J (2004) Inferring Phylogenies. Sinauer Associates, Inc., Sunderland, Massachusetts, USA

Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. In: Gordon G, Dunson D, Dudk M (eds) Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, vol 15. PMLR, Fort Lauderdale, Florida, USA, pp 315–323

Gómez-Carballa A, Bello X, Pardo-Seco J, et al (2020) Mapping genome variation of sars-cov-2 worldwide highlights the impact of covid-19 super-spreaders. Genome Research 30(10):1434–1448. https://doi.org/10.1101/gr.266221.120

Goodfellow I, Bengio Y, Courville A (2016) Softmax units for multinoulli output distributions. In: Deep Learning. MIT Press, Cambridge, Massachusetts, USA, chap 6.2.2.3, p 180–183

Grenfell BT, Pybus OG, Gog JR, et al (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303(5656):327–332. https://doi.org/10.1126/science.1090727

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52(5):696–704. https://doi.org/10.1080/10635150390235520

Hodcroft EB, Zuber M, Nadeau S, et al (2021) Spread of a sars-cov-2 variant through europe in the summer of 2020. Nature 595:707–712. https://doi.org/10.1038/s41586-021-03677-y

Izquierdo-Carrasco F, Stamatakis A (2011) Computing the phylogenetic likelihood function out-of-core. In: 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and PhD Forum. IEEE, Anchorage, Alaska, USA, pp 444–451, https://doi.org/10.1109/IPDPS.2011.185

Jermiin LS, Jayaswal V, Ababneh FM, et al (2016) Identifying optimal models of evolution. In: Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution. Springer, New York, USA, p 379–420

Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. https://doi.org/10.48550/arXiv.1412.6980, preprint at https://arxiv.org/abs/1412.6980

Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular Biology and Evolution 11(3):459–468. https://doi.org/10.1093/oxfordjournals.molbev.a040126

Kulikov N, Derakhshandeh F, Mayer C (2024) Machine learning can be as good as maximum likelihood when reconstructing phylogenetic trees and determining the best evolutionary model on four taxon alignments. Molecular Phylogenetics and Evolution 200:108181. https://doi.org/10.1016/j.ympev.2024.108181

Leuchtenberger AF, Crotty SM, Drucks T, et al (2020) Distinguishing felsenstein zone from farris zone using neural networks. Molecular Biology and Evolution 37(12):3632–3641. https://doi.org/10.1093/molbev/msaa164

Li T, Liu D, Yang Y, et al (2020) Phylogenetic supertree reveals detailed evolution of sars-cov-2. Scientific Reports 10:22366. https://doi.org/10.1038/s41598-020-79484-8

Mai U, Mirarab S (2018) Treeshrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics 19:272. https://doi.org/10.1186/s12864-018-4620-2

Mirarab S, Reaz R, Bayzid MS, et al (2014) Astral: genome-scale coalescent-based species tree estimation. Bioinformatics 30(17):i541–i548. https://doi.org/10.1093/bioinformatics/btu462

Nesterenko L, Blassel L, Veber P, et al (2025) Phyloformer: fast, accurate, and versatile phylogenetic reconstruction with deep neural networks. Molecular Biology and Evolution 42(4):msaf051. https://doi.org/10.1093/molbev/msaf051

Nguyen LT, Schmidt HA, von Haeseler A, et al (2015) Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution 32(1):268–274. https://doi.org/10.1093/molbev/msu300

Piel WH, Donoghue MJ, Sanderson MJ (2002) Treebase: a database of phylogenetic information. In: Proceedings of the 2nd International Workshop of Species 2000. National Institute for Environmental Studies, Tsukuba, Japan, pp 41–47

Ronquist F, Huelsenbeck JP (2003) Mrbayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19(12):1572–1574. https://doi.org/10.1093/bioinformatics/btg180

Schmidt HA, Strimmer K, Vingron M, et al (2002) Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18(3):502–504. https://doi.org/10.1093/bioinformatics/18.3.502

Scornavacca C, Belkhir K, Lopez J, et al (2019) Orthomam v10: scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. Molecular Biology and Evolution 36(4):861–862. https://doi.org/10.1093/molbev/msz015

Smith ML, Hahn MW (2023) Phylogenetic inference using generative adversarial networks. Bioinformatics 39(9). https://doi.org/10.1093/bioinformatics/btad543

Solís-Lemus C, Tang X, Zepeda-Nuñez L, et al (2024) Novel symmetry-preserving neural network model for phylogenetic inference. Bioinformatics Advances 4(1):vbae022. https://doi.org/10.1093/bioadv/vbae022

Stamatakis A (2006) Phylogenetic models of rate heterogeneity: a high performance computing perspective. In: Proceedings of the 20th IEEE International Parallel and Distributed Processing Symposium. IEEE, Rhodes, Greece, p 8, https://doi.org/10.1109/IPDPS.2006.1639535

Stamatakis A (2014) Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Strimmer K, von Haeseler A (1996) Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Molecular Biology and Evolution 13(7):964. https://doi.org/10.1093/oxfordjournals.molbev.a025664

Suvorov A, Schrider DR (2024) Reliable estimation of tree branch lengths using deep neural networks. PLoS Computational Biology 20(8):e1012337. https://doi.org/10.1371/journal.pcbi.1012337

Suvorov A, Hochuli J, Schrider DR (2020) Accurate inference of tree topologies from multiple sequence alignments using deep learning. Systematic Biology 69(2):221–233. https://doi.org/10.1093/sysbio/syz060

Turakhia Y, Thornlow B, Hinrichs AS, et al (2021) Ultrafast sample placement on existing trees (usher) enables real-time phylogenetics for the sars-cov-2 pandemic. Nature Genetics 53(6):809–816. https://doi.org/10.1038/s41588-021-00862-7

Wang Z, Sun J, Gao Y, et al (2023) Fusang: a framework for phylogenetic tree inference via deep learning. Nucleic Acids Research 51(20):10909–10923. https://doi.org/10.1093/nar/gkad805

Whelan S, de Bakker PI, Quevillon E, et al (2006) Pandit: an evolution-centric database of protein and associated nucleotide domains with inferred trees. Nucleic Acids Research 34:D327–D331. https://doi.org/10.1093/nar/gkj087

Yang Z, Rannala B (2012) Molecular phylogenetics: principles and practice. Nature Reviews Genetics 13:303–314. https://doi.org/10.1038/nrg3186

Zhu Y, Li Y, Li C, et al (2025) A critical evaluation of deep-learning based phylogenetic inference programs using simulated datasets. Journal of Genetics and Genomics 52(5):714–717. https://doi.org/10.1016/j.jgg.2025.01.006

Zou Z, Zhang H, Guan Y, et al (2020) Deep residual neural networks resolve quartet molecular phylogenies. Molecular Biology and Evolution 37(5):1495–1507. https://doi.org/10.1093/molbev/msz307

Supplementary

Hyperparameter	Description	Search Range	Sampling Strategy	Best Setting
Learning Rate	Step size for weight updates during training	[e−4, 2e−3]	Log-uniform	7.844e−4
Learning Rate Decay	Factor to reduce the learning rate over epochs	[0.85, 1.0]	Uniform	0.9044
Dropout	Fraction of neurons randomly deactivated to prevent overfitting	[0.0, 0.3]	Uniform	0.08314
β_1(Adam)	Exponential decay rate for first moment estimates	[0.85, 0.95]	Uniform	0.8687
β_2(Adam)	Exponential decay rate for second moment estimates	[0.9, 0.999]	Uniform	0.9976
Batch Size	Number of samples processed per training iteration	{8, 16, 32, 64, 128, 256, 512}	Categorical	128

Table S1: Hyperparameter search ranges and the best setting for the tree topology classifier.

Hyperparameter	Description	Search Range	Sampling Strategy	Best Setting
Learning Rate	Step size for weight updates during training	[e−5, e−2]	Log-uniform	6.297e−4
Learning Rate Decay	Factor to reduce the learning rate over epochs	[0.85, 1.0]	Uniform	0.9667
Dropout	Fraction of neurons randomly deactivated to prevent overfitting	[0.0, 0.3]	Uniform	0.05618
Weight Decay	L_2 regularization to prevent large weights	[e−6, e−3]	Log-uniform	6.297e−4
β_1(Adam)	Exponential decay rate for first moment estimates	[0.85, 0.95]	Uniform	0.9281
β_2(Adam)	Exponential decay rate for second moment estimates	[0.9, 0.999]	Uniform	0.9717
Batch Size	Number of samples processed per training iteration	{8, 16, 32, 64, 128, 256, 512}	Categorical	32

Table S2: Hyperparameter search ranges and the best setting for the branch length regressor.

Metric	Branch Type	IQ-NET	IQ-TREE-quartet
MAE	Internal branch	**0.0322**	0.0388
	External branch	**0.0273**	0.042
MSE	Internal branch	**0.0202**	0.0538
	External branch	**0.0122**	0.0442
MRE	Internal branch	0.7824	**0.6564**
	External branch	0.2519	**0.2286**
BSD	All branches	**0.0946**	0.1485

Table S3: Branch length estimation errors of IQ-NET and IQ-TREE-quartet. For IQ-NET and IQ-TREE, only trees with correct topology prediction were counted. Bold numbers denote the best results.

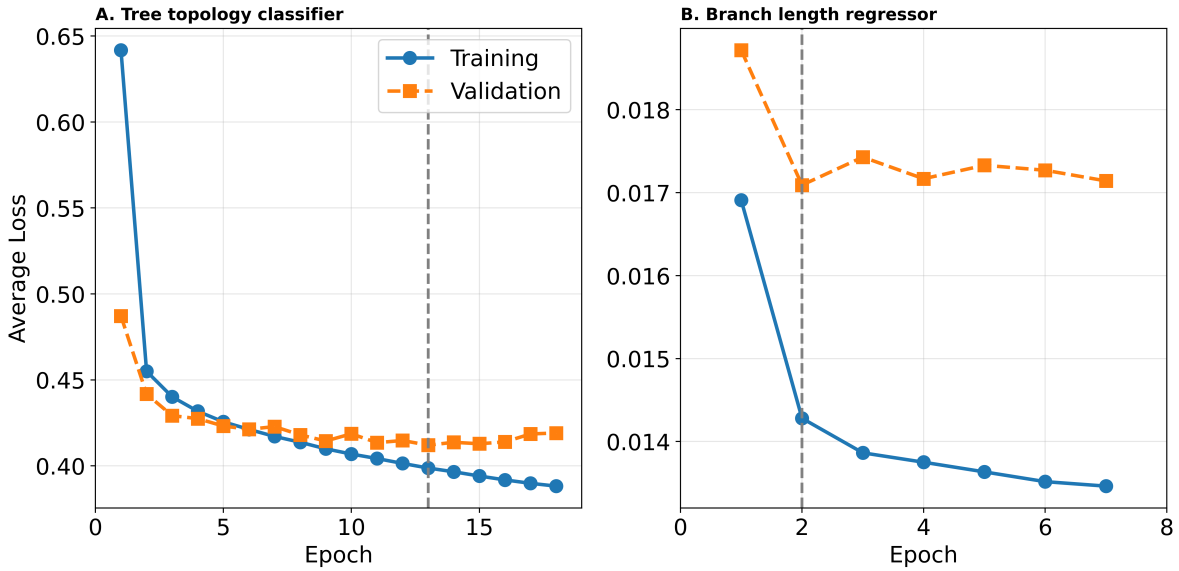

Figure S1: The training and validation loss of (A) the tree topology classifier; and (B) the branch length regressor. The regressors for internal and external branches were jointly trained, resulting in a single loss curve. The black dashed vertical line indicates the epoch at which the best parameters were obtained (epoch 13 for the topology classifier and epoch 2 for the branch length regressor) before training was terminated by early stopping.

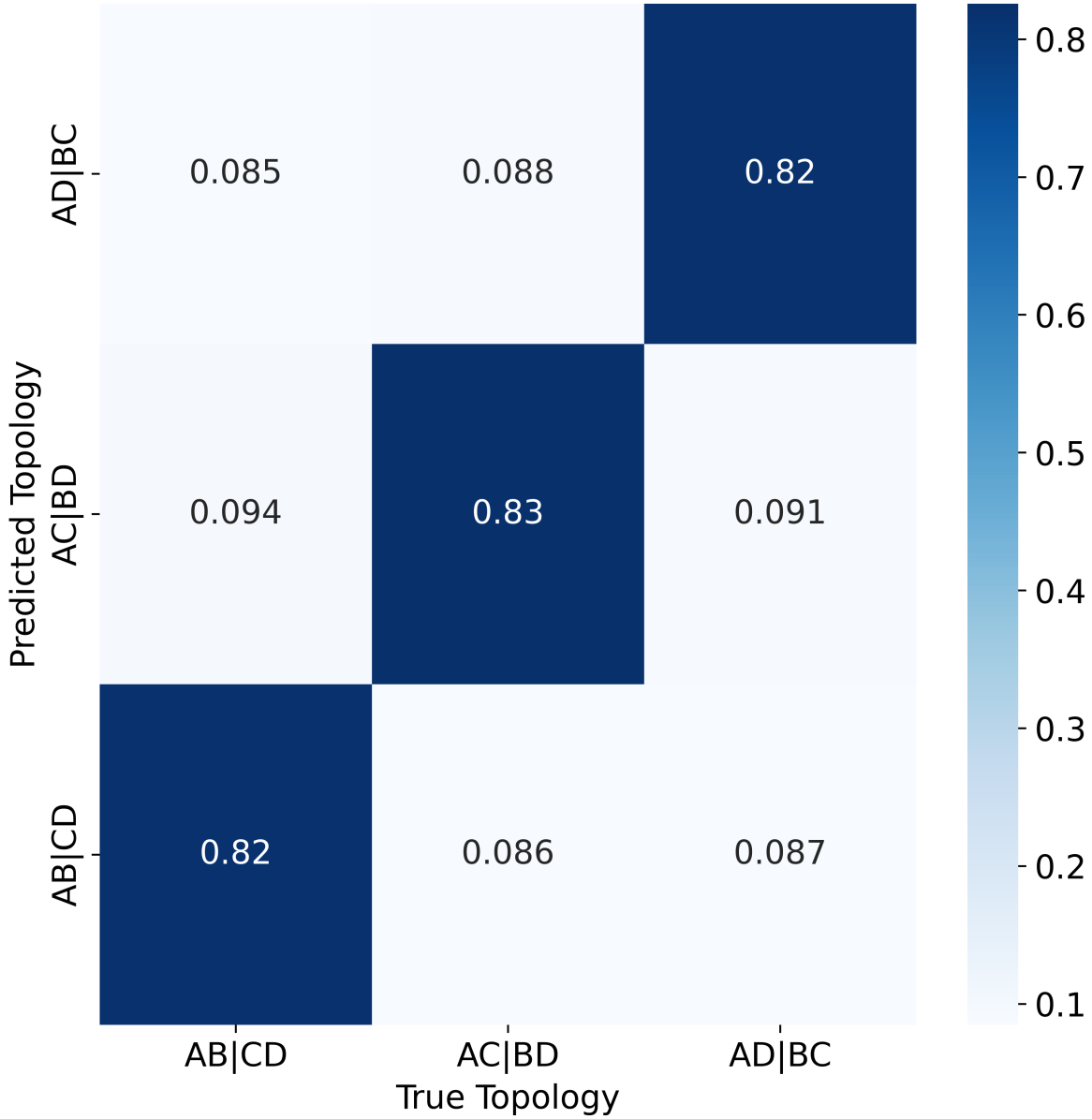

Figure S2: Confusion Matrix of True vs Predicted tree topologies.

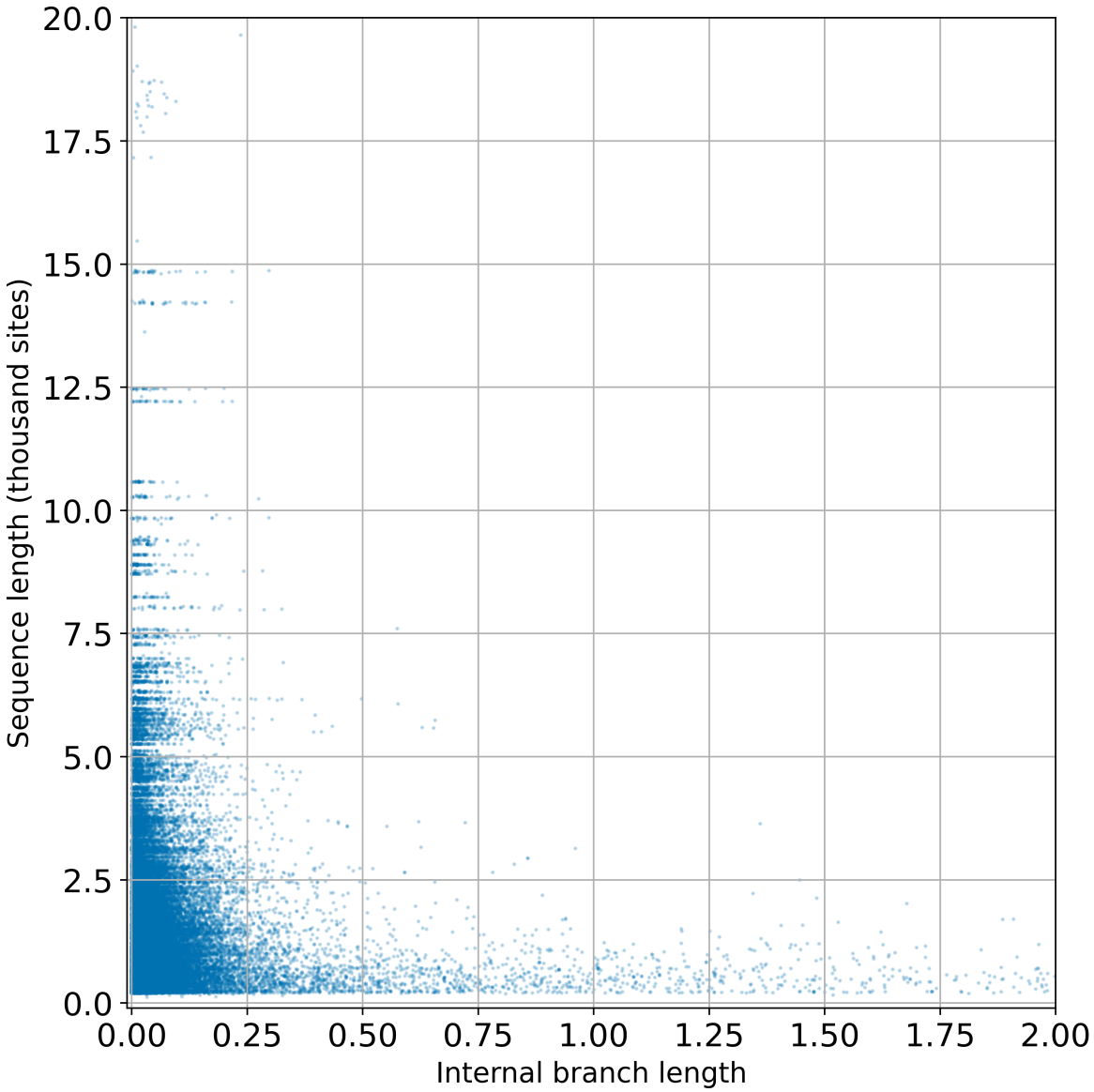

Figure S3: Distribution of Internal branch lengths vs Sequence length across testing samples.

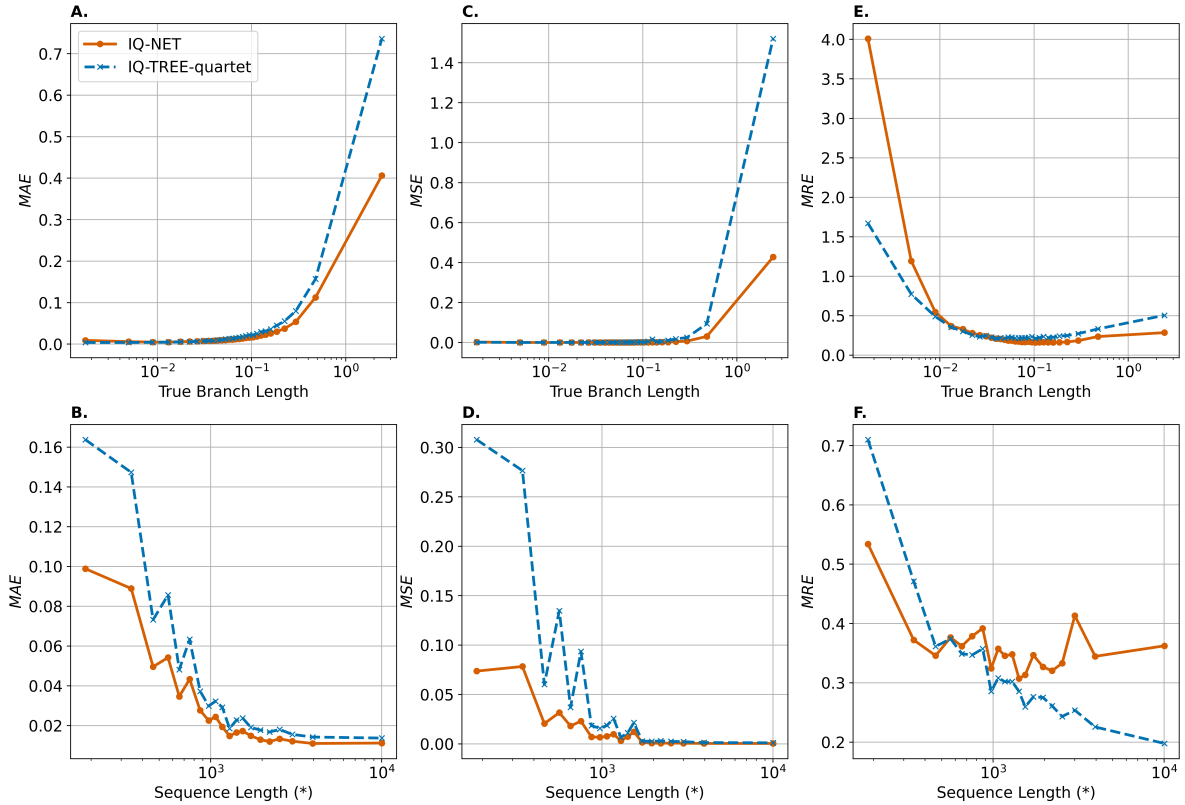

Figure S4: Mean absolute error (MAE), Mean square error (MSE), and Mean relative error (MRE) of branch length estimated by IQ-NET and IQ-TREE-quartet. (A) MAE vs branch length; (B) MAE vs sequence length; (C) MSE vs branch length; (D) MSE vs sequence length; (E) MRE vs branch length; and (F) MRE vs sequence length. (*) Sequence length excludes fully gapped sites, which contain no evolutionary signal.

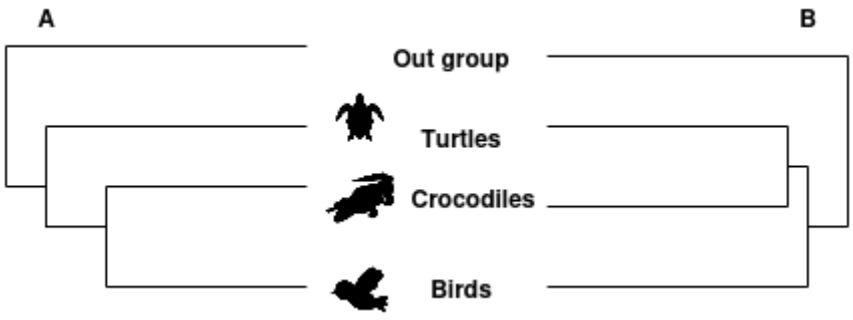

Figure S5: (A) Accepted tree for placement of turtles as sister clade to birds and crocodiles (Chiari et al., 2012). (B) Tree constructed under methods of IQ-NET+ASTRAL, IQ-TREE with concatenated MSA, ASTRAL using gene trees estimated with IQ-TREE and ModelFinder.