

Detecting Outlier Subtrees of Gene Trees Using SPR Moves and Machine Learning

Alan K. Mayer¹, Shayesteh Arasti², and Siavash Mirarab^{3*}

¹ Department of Mathematics, UC San Diego, CA 92093, USA

² Department of Computer Science and Engineering, UC San Diego, CA 92093, USA

³ Department of Electrical and Computer Engineering, UC San Diego, CA 92093, USA

Abstract. Abundant discordance among gene trees is widely documented, but the causes of this heterogeneity are varied. Discordance among estimated gene trees can stem from real sources such as coalescent processes, hybridization, and horizontal gene transfer (HGT). It can also stem from errors in data, such as hidden paralogy, mistaken homology, bad alignment, and contamination. While some of these processes create stochastic and subtle changes in gene tree topologies (e.g., human closer to gorilla than to chimp), others can produce unexpected patterns (e.g., guinea pig sister to gorilla). Given a large number of gene trees and a median species tree, one could attempt to automatically find these outliers among gene trees. In this paper, we develop a method that uses quartet-based subtree-prune-and-regraft (SPR) moves, paired with gradient-boosted decision trees, to predict whether parts of a gene tree disagree with species trees in unusual ways. We show that our method, GBOD, is quite accurate in finding HGT events, but less so in other scenarios. Nevertheless, this combination of machine learning and phylogenetic features provides a promising framework for outlier detection.

Keywords: Phylogenomics, machine learning, quartet-based methods, SPR moves, outlier detection, horizontal gene transfer (HGT)

1 Introduction

Phylogenomic analyses that infer species trees and gene trees from large collections of loci sampled across the genome have become standard [38]. While much attention has been paid to methods for inferring species trees and gene trees, a key challenge for practitioners is that input data often do not fully conform to the assumptions of these methods. On the one hand, large-scale datasets are often riddled with errors [24], stemming from incorrect homology and orthology [29, 35], sequence contamination [9, 14], and alignment error [15, 44]. Moreover, gene trees have their own added set of errors [34] due to the impact of fragmentary data [12, 30], sequence model misspecification [13, 27], and long branch attraction. Besides these errors, individual parts of the genomes can

* Corresponding author: smirarab@ucsd.edu

undergo unusual and unaccounted modes of evolution, such as recombination suppression [21], horizontal gene transfer in unexpected places [41], further creating modeling deficiencies. Regardless of cause, abnormalities in data impact gene trees and the species trees by extension [18, 23].

In response to these difficulties, many phylogenomic analyses adopt a wide range of filtering strategies, many operating on alignments [31, 37]. These methods tend to over-filter [25, 39] and lack the full evolutionary context to detect which patterns of discordance are expected or unusual [44]. Others have noted that errors and unmodelled processes often lead to unusual placements of taxa on gene trees, beyond what is expected based on typical causes of gene tree discordance such as incomplete lineage sorting (ILS). For example, a large number of unexpected topologies among mammalian gene trees (e.g., a guinea pig going with gorilla) was used to cast doubt on the validity of some earlier datasets, or even the whole notion of a gene tree [34].

In response to these observations, several methods have been developed to look for *outliers* in the gene trees [e.g., 4, 7, 16]. The guiding principle of these methods is that while some forms of discordance are expected for each dataset, some gene trees are so widely different that they can only be explained by errors or unmodelled processes (e.g., HGT in eukaryotes). We refer to these as outliers, remaining agnostic as to the exact cause. In this terminology, calling part of a gene tree an outlier simply means that the discordance it creates is above and beyond what is implied by the species tree or the distribution of gene trees. While automatic outlier detection methods exist and have been adopted increasingly, benchmarking shows that they all have room for improvement [4, 7]. Note that these methods do not look for outlier gene trees (an easier task), but rather, individual species or clades in individual gene trees that appear misplaced.

One way to formalize outliers in this context is to examine subtree prune and regraft (SPR) moves that remove a clade and regraft it onto another branch. Since SPRs are powerful, applying a few of them can create many forms of outliers. For outlier detection, we may expect that a highly erroneous clade would have to move very far on the gene tree to agree with a reference tree, such as an estimated species tree. Luckily, using an algorithm we recently developed, for each clade in a gene tree, one can determine the optimal SPR move that maximizes the quartet agreement with a reference tree in $O(n \log^2(n))$ time, amortized over all clades [3]. Thus, we can compute a quartet-based SPR (QSPR) for each clade of each gene tree in addition to several properties of the QSPR move (e.g., how far, what improvement in quartet score, etc.) in a scalable fashion.

As we will see, while QSPR does have some signal regarding outliers, interpreting that signal is far from trivial. How far is far enough? What improvement in quartet score is significant? The answers depend on the context in non-trivial ways. To address this challenge, we resort to machine learning, aiming to train a model that can determine whether a particular QSPR is outside the normal range. However, machine learning poses its own challenges, including a lack of access to labeled training data or knowledge of the underlying error mechanisms. Moreover, QSPR statistics and their interpretation are likely dataset dependent,

necessitating adjustments to the criteria per dataset. To address these challenges, we design a method called Gradient Boosted Outlier Detection (GBOD) based on cross-validation and data augmentation with positively labeled samples, trained for a given dataset. We show that QSPR statistics, when interpreted through this machine learning approach, have enough signal to detect outliers with levels of accuracy that rival existing methods. Our machine learning approach can, in the future, be combined with features beyond QSPR (e.g., those extracted using existing methods) to further improve accuracy.

2 Material and Methods

2.1 Notations

Let $T = (V_T, E_T)$ be a rooted binary tree, where V_T and E_T denote the sets of vertices and edges of T , respectively. For an edge $e \in E_T$, we use $l(e)$ to denote the length of e . Let $L_T \subset V_T$ be the leaf set of T . A *quartet* of T on four taxa $\{a, b, c, d\} \subseteq L_T$ is defined as the unrooted topology of T restricted to these four taxa. For a tree T with n leaves, there are $\binom{n}{4}$ such quartets. We say that two trees T_1 and T_2 *share a quartet* on $\{a, b, c, d\}$ if the quartets of T_1 and T_2 on these taxa have the same topology (among the three possible topologies). The *quartet score* of T_1 and T_2 is defined as the number of quartets shared between them. For each vertex $v \in V_T$, we define $C_T(v)$ as the subtree below v in T , and use the simpler notation $C(v)$ when T is clear from context. With a slight abuse of notation, we use $L_{C_T(v)}$ to denote the set of leaves below v in T . A subtree prune and regraft (SPR) move of a subtree $C_T(v)$ to an edge $(u, u') \in E_T$ is defined as follows: we prune $C_T(v)$ from T by removing the edge above v , and then regraft $C_T(v)$ onto the edge (u, u') by introducing a new vertex w , removing (u, u') , and adding the edges (u, w) , (w, u') , and (w, v) , resulting in a new tree T' . We call an SPR move of $C_T(v)$ to an edge $e \in E_T$ a *quartet-based SPR* (QSPR) with respect to a tree T_2 if, among all possible SPR moves of $C_T(v)$, this move results in the highest quartet score between T' and T_2 . When trees are unrooted, we define two SPRs per edge, placing each side of its bipartition on the other side.

2.2 GBOD Overview

Motivation We observed that when outliers are introduced to a subtree in a simulated dataset, the distance of the optimal QSPR moves on these subtrees is generally higher than that of typical subtrees (Fig. 1a). However, these distributions also have much overlap. Thus, this signal alone cannot definitively discriminate outliers from non-outliers. Moreover, the interpretation of the signal depends on the context. For example, even absent outliers, QSPR distances correlate with average tree branch length (Fig. 1b). Thus, to make sense of QSPR results, context needs to be considered. However, what information needs to be added and how it should affect our classification is not obvious. Instead of mathematical modeling of QSPR, we turn to the tools that machine learning offers to automatically learn meaningful relationships between the QSPR move and other features of the tree, in order to determine which parts are anomalous.

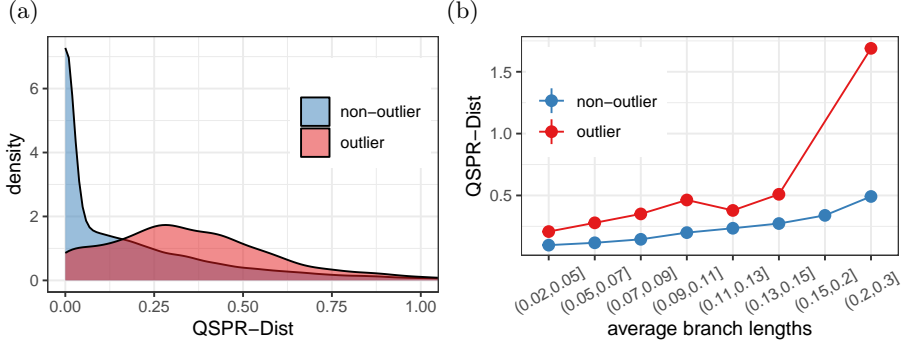


Fig. 1: (a) The distribution of how far a subtree moves with an optimal SPR move (QSPR-Dist in Table 1), for outlier and non-outlier subtrees in the simulated S200-perturbed dataset (Section 3.1). A subtree $C(v)$ is considered an outlier if all its leaves are outliers, and a non-outlier if all its leaves are non-outliers. (b) The correlation between the average branch lengths for a gene tree and the QSPR-Dist for outlier and non-outlier subtrees.

GBOD framework GBOD takes as input a set of gene trees and aims to train models that can learn to identify outlier subtrees in the gene trees based on QSPR statistics. The outliers are subtrees that have placements that are highly unusual given the other gene trees. Since the errors are unknown, we cannot train a supervised model on this data directly. To enable training, we use augmentation (Fig. 2): We inject the gene trees with known anomalies via SPR moves, train a model to identify these introduced outliers, and use the trained model to infer the anomalous parts of the original gene trees.

Leave-out mechanism: The unknown anomalies in the original gene trees we aim to detect persist in the augmented gene trees used as training data. Therefore, a procedure that naively trains a model on an augmented gene tree and uses the same model for inference on the original tree will be biased toward classifying the unknown anomalies as typical, since it was labeled as such in the training data. To remedy this, we adopt a procedure similar to k -fold cross-validation: We partition our original gene trees into k folds. Errors are added in $k - 1$ folds, and inference is done on the left-out fold (Fig. 2). Thus, inference is done on gene trees not seen during training. For computational efficiency, we reuse the augmented gene trees for training across different folds when possible.

The GBOD pipeline has four steps, which will be detailed in what follows: 1) creating positively labeled samples for training the model, 2) measuring a set of features for each tree or clade in a tree, 3) training a regression model for each clade to compute the probability that a randomly selected leaf under that clade is anomalous, and 4) using this model to infer outliers in the left out fold, with automatic adjustments to decide what threshold should indicate outliers.

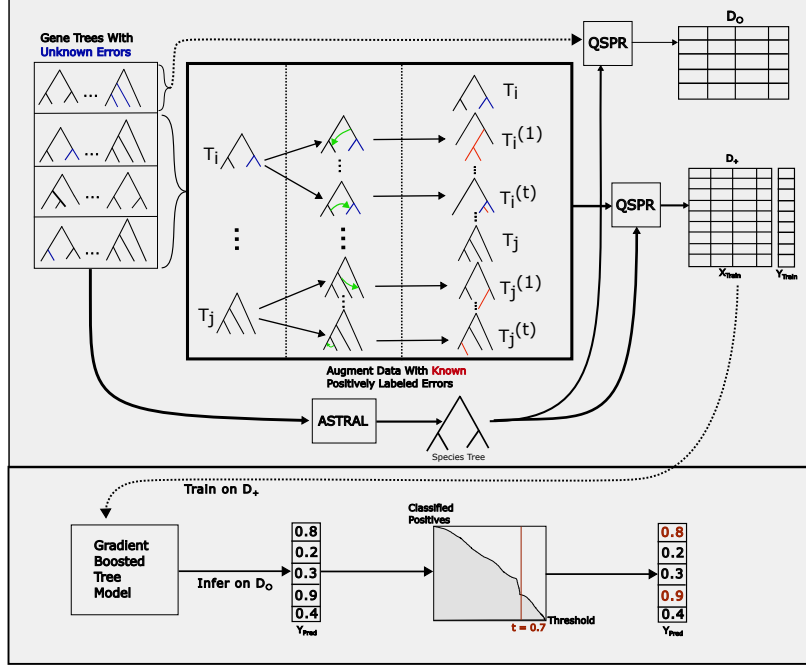


Fig. 2: An overview of the Gradient Boosted Outlier Detection pipeline. GBOD begins by partitioning gene trees into k folds. Then, $k - 1$ folds have augmented errors introduced via random SPR moves, creating the augmented dataset D_+ . We then use QSPR [3] with respect to the ASTRAL species tree to measure how far each clade is from its optimal position. Results are used to assign a feature vector to each clade of each gene tree (Fig. 3). We train our model on D_+ and infer outliers on the original, unperturbed gene trees in the left-out fold, D_o . Finally, we automatically threshold these predictions to classify outliers.

2.3 Creating labeled positive samples for training

On a set of N rooted gene trees, $G = \{T_1, T_2, \dots, T_N\}$, we begin by perturbing the data to inject known, labeled errors, which will be used to train a model applied downstream to detect errors in the original unperturbed data (from left-out fold). This augmentation transforms the problem into a supervised learning task well-suited for classical machine learning methods. In particular, for each gene tree, T_i , we augment it t times to get a set of gene trees with known errors $G'_i = \{T_i^{(1)}, T_i^{(2)}, \dots, T_i^{(t)}\}$. Each $T_i^{(j)}$ is constructed by choosing a source $e = (u, v)$ and a destination $e' = (u', v')$ on T_i at random, weighted by their branch lengths. Note that this selection is different from how HGT is often modeled where probability of a branch becoming destination is inversely proportional to

the path length between that branch and the source. We then perform an SPR move, by pruning $C(v)$ and regrafting on e' , resulting in the new tree $T_i^{(j)}$. This placement divides e' into two new edges $e'_1 = (u', w)$ and $e'_2 = (w, v')$, where $l(e'_1) + l(e'_2) = l(e')$. We choose $l(e'_1)$ uniformly at random from $[0, l(e')]$.

Let $L_+ = L_{C_{T_i}(v)}$ be the set of all leaves in the subtree $C_{T_i}(v)$ that are moved and thus labelled positive. For each node $n \in V_{T_i^{(j)}}$, we set the label $y_n = |L_{C(n)} \cap L_+| / |L_{C(n)}|$ as the proportion of the leaves in $C(n)$ that are outliers. The machine learning model is trained to predict y_n . After the augmentation process, we will have the training set of gene trees $\mathcal{G} = G \cup G'_1 \cup \dots \cup G'_t$, where $|\mathcal{G}| = N \times (t + 1)$. Note that we also include the original gene trees in \mathcal{G} , and we set $y_v = 0$ for all vertices of T_i .

2.4 Feature Generation

For each node, v in tree T , we define the feature vector x_v . This feature vector is composed of metrics describing both the subtree, $C(v)$, and the full tree, T . A comprehensive list of features with brief descriptions is presented in Table 1.

The main signal is provided by features generated from a QSPR move using a dynamic programming algorithm [3], which identifies the optimal position for each subtree relative to a reference tree. A natural choice for this reference tree is an estimated species tree, S , which can be thought of as a summary of the gene trees. For a subtree $C(v)$ of a gene tree T , and the species tree S , QSPR identifies the optimal SPR move on $C(v)$ such that the number of quartets shared between the updated T and S is maximized. The primary subtree-specific signals that our model may learn from is the distance, both topological and in terms of path lengths (i.e., considering branch lengths), that $C(v)$ moved from its original position on the gene tree to its new position, as well as many derived features. An example of a few possible SPR moves, along with their associated features, can be found in Figure 3. For each vertex v , we perform QSPR both for $C(v)$ and its complement (i.e., the clade below v if the tree is rerooted somewhere inside $C(v)$). Therefore, we obtain two different feature vectors for each vertex v , corresponding to the bipartition defined by the branch above v . All other features, with the exception of clade size, are defined on the gene tree as a whole. These provide context for the model about the tree in which this clade was found, giving it more information to interpret the QSPR values.

2.5 Modeling and predicting

Let $x_i^{(j)}$ represent a feature vector for subtree j in the unperturbed gene trees $T_i \in G$. Recall that we have no associated ground truth $y_i^{(j)}$ on these gene trees, but hope to train a model that gives meaningful outputs when evaluated on these feature vectors. To train these models from our perturbed data, we adopt a procedure similar to k -fold cross-validation. Our unperturbed and randomly ordered gene trees, $G = \{T_1, T_2, \dots, T_N\}$ are partitioned into k folds, $G^{(1)}, \dots, G^{(k)}$. This induces a partition on the perturbed dataset, G' where if

Table 1: Features used by GBOD. The first set is outputs of QSPR [3], which runs in $O(kn^2 \log^2(n))$ time for all features across all k gene trees, each with n leaves. Others are calculated in $O(kn)$ time.

| name | W^\dagger | C^\ddagger | Description |
|---------------------|-------------|--------------|---|
| QSPR-Nodes | F | C | The number of nodes a clade moves by QSPR to S |
| QSPR-Nodes Norm. | F | C | QSPR-nodes / Avg. Clade Height |
| QSPR-Dist | T | C | The branch distance a clade moves by QSPR to S |
| QSPR-Dist Norm. | T | C | QSPR-Dist / Avg. Branch Height |
| Root Dist Before | T | C | Distance of clade to root before QSPR. |
| Root Dist After | T | C | Distance of clade to root after QSPR. |
| Root Dist Change | T | C | Root Dist After – Root Dist Before |
| Quartet Score Diff. | F | T | Change in gene tree quartet score vs. S after SPR. |
| Clade Size | F | C | The number of leaves in the clade. |
| Avg. Node Height | F | T | Average topological height of each vertex in the gene tree. |
| Std. Node Height | F | T | Standard Deviation of topological heights of each vertex. |
| Avg. Branch Height | T | T | Average of heights of each vertex. |
| Std. Branch Height | T | T | Standard Deviation of heights of each vertex. |
| Avg. Branch Length | T | T | Average length of all branches in tree. |
| Tree Diameter | T | T | Longest path between any two nodes |
| Branch Length Sum | T | T | Sum of all branch lengths |
| Treeness | T | T | Sum of internal branch lengths / Sum of all lengths |
| Cherry Proportion | F | T | Proportion of leaves sister to another leaf. |
| Colless Index | F | T | Sum of difference in child clade sizes. |
| Sackin Index | F | T | Sum of the depths (edges from the root) of all leaves |
| Tree Height | T | T | Maximum of the depths of all leaves |

\dagger : F: Branch lengths of gene trees are ignored; T: otherwise.

\ddagger : C: Defined per clade; T: defined per tree.

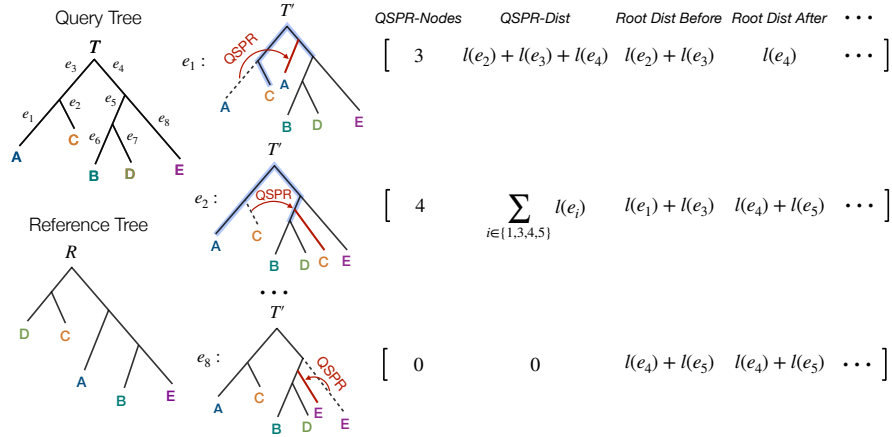


Fig. 3: Using QuartetSPR [3] to extract QSPR features. For each subtree C in the query (gene) tree, we find its optimal placement with respect to the reference (species) tree. We then compute QSPR features of C , e.g., QSPR-Nodes, QSPR-Dist, etc, based on the original and the optimal placement of C .

$G^{(i)} = \{T_{k \cdot i}, T_{k \cdot i + 1}, \dots, T_{k \cdot i + N/k}\}$, then $G'^{(i)} = \bigcup_{k \cdot i \leq j \leq k \cdot i + N/k} G'_j$. Then for a fold, i , we have labeled training data: $D_+^{(i)} = \bigcup_{j \neq i} G'^{(j)}$, and original unperturbed data: $D_o^{(i)} = G^{(i)}$. We train our model, $\phi^{(i)}$, on labeled pairs of data, $(x_j, y_j) \sim D_+^{(i)}$, where x_j is the feature vector for some subtree in an augmented gene tree of $D_+^{(i)}$, and y_j is its associated label. We then apply the trained model $\phi^{(i)}$ on the unlabeled data, $x_j \sim D_o^{(i)}$ to get $\hat{y}_j = \phi^{(i)}(x_j)$. Thus, test gene trees are absent from the training dataset, even as perturbed gene trees.

For $\phi^{(i)}$, we opt to use gradient boosted decision trees. These are ensemble machine learning models that train by iteratively building weak decision trees that correct the errors of the existing ensemble. Gradient boosted decision trees are computationally efficient to train compared to other methods. Additionally, they have been shown to often outperform deep models on tabular data [32]. While gene trees are not inherently tabular, the featurization and regression on each subtree independently of each other tabularize our graph-structured data, lending itself well to models optimized for working on this kind of data. The model is trained via logistic regression with the standard cross-entropy loss.

2.6 Thresholding

For each original gene tree, we now have a prediction for each subtree on the proportion of its leaves that are outliers. We need a threshold to use these predictions to determine which taxa are believed to be erroneous with sufficient confidence. We designed an automatic thresholding strategy that analyzes the proportion of leaves we remove at each threshold. Once a threshold is determined, a leaf is removed from the gene tree if it is contained in any subtree whose confidence is above that threshold. The thresholding strategy analyzes the *predicted* positive rate as a function of the decision threshold. Intuitively, our scheme aims to numerically find a point with high derivative for this function, corresponding to a value at which we get a relatively large increase in the number of filtered leaves. Formally, let the proportion of leaves removed at some threshold t be $F(t)$ where $F(0) = 1$ and $F(1) = 0$. For some small value, δ , and configurable parameter controlling how large the spike is, f_{\max} , we gradually decrease t from 1 to 0 and look for when:

$$F(t - \delta) - F(t) > f_{\max}.$$

We then opt to cut at the first threshold t that satisfies this condition. By default, we will decrease t in intervals of 0.01 at a time with $\delta = 0.01$ and $f_{\max} = 0.005$.

3 Experiments

We evaluate GBOD on two simulated datasets and one biological dataset. The simulated datasets, which will be discussed in detail below, each have known outliers defined on the leaves of some of the gene trees. We use ASTRAL-III [43]

for species tree inference from the original unperturbed gene trees to use as a reference for computing QSPR features. GBOD is then trained and tested separately for each dataset. When training the model in each fold, we use the left-out augmented data for the testing fold to determine early-stopping (without knowledge of true outliers). We evaluate our method on the simulated datasets both by analyzing the F_1 score on each of the datasets, which takes into account our automatic thresholding strategy, as well as the Receiver Operating Characteristic (ROC) curves, which are agnostic to our thresholding strategy. In our setting, a true positive is an outlier detected as such, and a false positive is a normal sequence marked as an outlier. On the biological dataset, we evaluate the effect that removal of the labeled outliers has on species tree inference.

We use $k = 10$ folds in the experiments. When computationally feasible, we augment each gene tree, T_i , 99 times, to get an augmented dataset size 100 times larger than the original. When this is too computationally demanding, we augment each gene tree 9 times. When training our gradient boosted tree models, we opt to use the XGBoost Python package [6]. In each fold during model training, we use the augmented gene trees associated with the left-out fold to perform early stopping to prevent overfitting.

3.1 Datasets

HGT Carnivora Simulated Dataset We simulated a dataset, following the same exact procedure as the simulated Carnivora dataset used to test PhylteR [7]. This dataset is simulated using Simphy [17] and uses a 53-taxa Carnivora tree [2] as a reference species tree. In each replicate, 500 gene trees were simulated with the rate of HGT being $1e-8$ and varying the levels of ILS. Each gene tree in this simulation contains at most one HGT event, and all taxa affected by such HGT events are considered outliers, and it is of interest if our procedure is able to detect these HGT events. From the 500 simulated trees, the first 100 gene trees were retained due to runtime restrictions. We only analyzed the most general case in which we do not limit the number of outliers per gene, so up to 53 can undergo an HGT event (on average: 1.25 outliers per gene tree).

S200-perturbed dataset We use a second dataset described in earlier work [4]. This dataset starts with a 201-taxon Simphy-simulated dataset [17] with gene trees that undergo ILS. Errors are then algorithmically introduced into the sequence alignment by performing between one and four rounds of perturbation to the sequence data (not gene trees). In a single round of perturbations, $k \sim \text{Poisson}(40)$ genes are modified uniformly at random, and in each of the genes, a small proportion ($t \sim \text{Beta}(10, 90)$) of the taxa are chosen to be modified. The taxa chosen to be modified are sampled according to a strategy that leads them to be closer to each other on the trees than sampling taxa uniformly at random would, but they are not monophyletic. Once the taxa for the round are selected to be modified, the species tree is randomly rerooted, and the root of the tree becomes the “source” taxa. Some portion of the alignment of the source

taxa replaces the portions at the same sites in each of the taxa chosen to be modified, leading to chimeric sequences between the chosen taxa and the source taxa. This procedure introduced between 0 and 81 perturbations to each gene (3.2 perturbations in expectation). All perturbed sequences in each gene tree are marked as (unknown) outliers for that gene tree.

Once alignments are perturbed, gene trees are re-inferred with FastTree II [26] and these gene trees become the input to our method. We omit two of the replicates due to extremely large polytomies in some of the original gene trees. This dataset was too large to run with 99 augmentations on each gene tree, so we opted to run it with 9 augmentations per gene tree instead.

Biological Data We study six real datasets. We use a mammalian dataset [33] with 37 species and 424 genes that is known to have errors identified in previous studies [11, 20]. We use the gene trees inferred by RAxML [36] from a reanalysis of this dataset [19]. There are two datasets studying Xenacoelomorpha, one studied by Cannon *et al.*, [5], which we call XenCannon, and one studied by Rouse *et al.*, [28], which we call XenRouse. XenCannon has 78 species and 213 genes, while XenRouse has 26 species and 393 genes. Our next dataset is on frogs [8] and contains 164 species and 95 genes. We study a dataset focused on plants and algae [42] containing 104 species and 852 genes. Lastly, we use the transcriptomic insect dataset [22] containing 144 species and 1478 genes. Only the insect and plant datasets were too large to run GBOD with 99 augmentations of each gene tree, so we opted to run it with 9 instead.

3.2 Method Comparison

We compare GBOD to PhylteR [7] and TCMM [4]. PhylteR detects local outliers in gene trees by representing each gene tree as a distance matrix on the leaves. Then, using DISTATIS [1], a generalization of multidimensional scaling built to compare distance matrices, PhylteR represents each taxa by a point in a low-dimensional Euclidean space, called the compromise space. By doing this, PhylteR can identify which leaves in which gene trees exhibit abnormally large deviations from expectations in this compromise space and classify them as outliers. TCMM provides a way to transform the branch lengths of the gene trees according to some reference tree. It is parameterized by a parameter λ , which determines how much it changes branch lengths. TCMM’s output can be used as input to PhylteR to control its ability to detect either topological (with $\lambda = 0$) or branch length outliers (with very large λ), or some compromise between the two. We run TCMM with its default $\lambda = 1$ value, looking for both topological and branch length outliers.

PhylteR and TCMM identify outliers using a modified Tukey method that defines outliers as values above the third quartile, adjusted for distributional skew. The aggressiveness of this method is controlled by the parameter k : lower values of k result in more taxa being classified as outliers. We evaluate how true and false positive rates change as we vary this k parameter, yielding ROC curves.

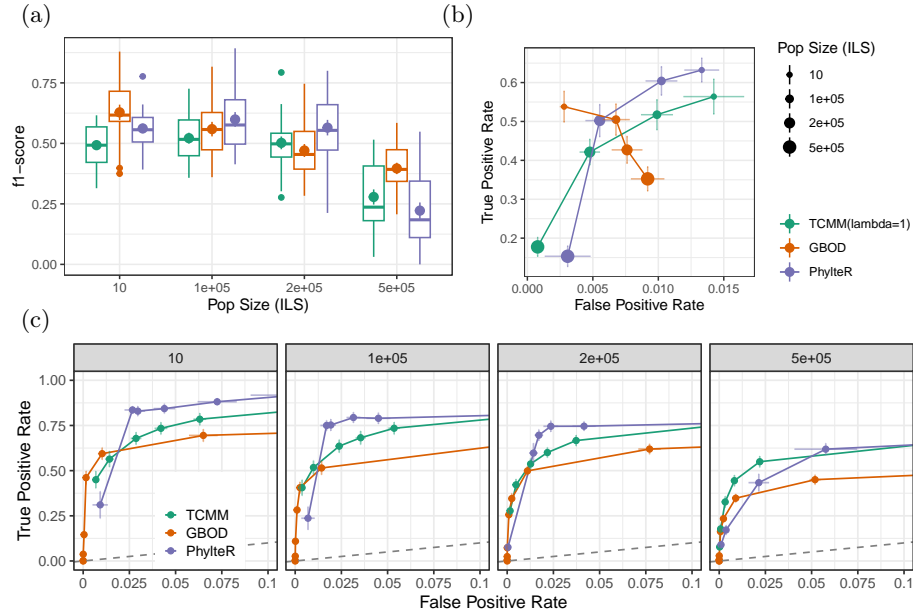


Fig. 4: (Carnivora Simulated Dataset) (a) Comparing GBOD to PhylteR and TCMM on various levels of ILS using F_1 -score. Mean and standard error are shown on top of the box plots. (b) False Positive Rate (FPR) versus True Positive Rate (TPR) for each method across various levels of ILS. (c) ROC curve for each method across all levels of ILS, showing $FPR < 0.1$. See Fig. S1 for full figure without restrictions and Table S1 for area under curve (AUC).

GBOD differs in that it directly predicts a confidence score for each subtree. We then make a determination of outliers by considering some threshold, t . We classify a taxon as being an outlier if it is contained in any subtree whose confidence is above this threshold. By default, this threshold is determined automatically.

4 Results

4.1 Simulated datasets

HGT detection on the Carnivora Dataset With default settings (automatic thresholding for GBOD, $k = 3$ for PhylteR and TCMM, and $\lambda = 1$ for TCMM), the relative F_1 accuracy heavily depends on the ILS level (Fig. 4a). In a setting with almost no ILS (population size set to 10), our method outperforms both TCMM and PhylteR. However, the F_1 score of GBOD consistently degrades with the increase in ILS, exhibiting both a lower TPR and higher FPR as ILS increases (Fig. 4b). These together lead to a consistent decrease in F_1 across ILS levels for GBOD (Fig. 4a). This contrasts the behavior of both TCMM and PhylteR, which

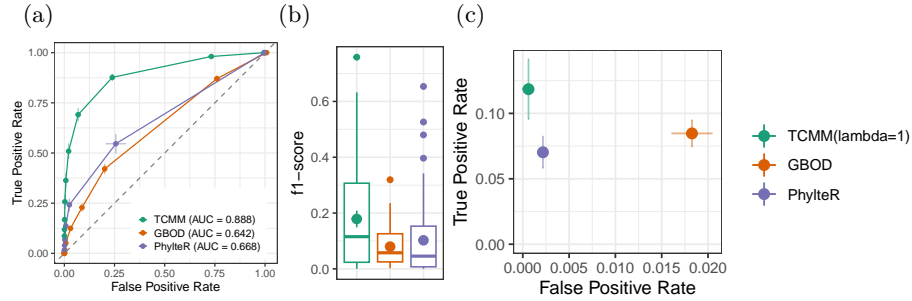


Fig. 5: (S200-perturbed Dataset) (a) ROC curve comparing all methods. Area under curve (AUC) values are shown for each method. (b) Comparing GBOD to PhylteR and TCM using F_1 -score. Mean and standard error are shown on top of the box plots. (c) False Positive Rate (FPR) versus True Positive Rate (TPR) for each method. Mean and standard error are shown.

suffer from lower TPR but enjoy lower FPR as well with higher ILS (Fig. 4b). Thus, in their default mode, these methods simply find fewer outliers with high ILS, and as a result, they can potentially maintain, or even slightly improve, F_1 score as ILS increases to the next two levels (Fig. 4a). Nevertheless, at the highest ILS level, these methods find very few outliers, leading to GBOD having a substantially higher F_1 than the alternatives, achieved through much better recall. We note that our choice of threshold is not necessarily optimal in these cases, as other thresholds with similar recall but better precision exist (Fig. S2).

When exploring various thresholds, several patterns emerge. Across ILS levels, GBOD displays a lower Area Under the Curve (AUC) than either PhylteR or TCM (Table S1). Importantly, this drop in AUC comes from GBOD’s poorer performance only at significantly high levels of FPR (Fig. S1). In the context of outlier detection, outliers are assumed to be quite rare, so in practice, each of these methods will choose thresholds corresponding to low false positive rates. Thus, this AUC cannot be taken as the best comparison method. Focusing on cases where FPR is below 10%, we see that GBOD is able to achieve lower FPR rates, and for very low FPR (e.g., below 2%), is better than or competitive with other methods (Fig. 4c). In all conditions except for the low ILS case, our automatic choice of threshold flags too many outliers compared to the threshold that maximizes F_1 (Fig. S2), leaving room for better threshold tuning.

Error detection in the S200 dataset We next turn to outliers caused by data errors (chimera), as simulated in the S200-perturbed dataset. On this dataset, none of the methods were particularly effective at finding outliers in the default settings, detecting 7–10% of errors on average and resulting in low F_1 scores (Fig. 5). The FPR was also low, showing that the errors introduced were simply missed in most cases in the background of real heterogeneity introduced by ILS.

Among the methods, GBOD has the lowest F_1 score, followed closely by PhylteR and then TCMM, with the slightly lower F_1 compared to PhylteR being due to higher FPR, which offsets its slightly higher recall. When examining the ROC curve, we observe that TCMM with better choices of threshold could find close to 70% of errors with an acceptable FPR (7%). In contrast, GBOD and PhylteR achieve only about 25% error detection if FPR is controlled at 10% or lower. And between the two, GBOD performs slightly worse across all thresholds. Therefore, the GBOD model does not appear to have been able to learn and generalize from our augmented outliers to the signal in the underlying errors introduced to the perturbed sequences on this dataset. The lower accuracy is only mildly affected by the reduced number of perturbations we were forced to perform; using 99 augmented trees on a subset of 10 replicate simulated datasets improved accuracy only slightly (Fig. S3).

4.2 Biological Data

Across the six biological datasets we studied, removal of taxa classified by GBOD as outliers improves the gene trees' similarity to the species tree in all cases (Fig. 6b and Fig. S4). In the Mammal and XenRouse datasets, while TCMM and PhylteR do not identify many outlier taxa in each gene tree (Fig. 6a), and subsequently do not see a large change in RF distance, GBOD's clade-based approach identifies an order of magnitude more taxa as outliers. We also observe that it is in these two datasets that we see the largest normalized RF improvement across all methods and datasets, followed by the plants dataset, where TCMM and PhylteR lead to higher RF improvements than GBOD (Fig. 6b). For frogs, all methods achieve similar levels of improvement, though PhylteR removes far more. On insects, PhylteR and GBOD achieve the best improvement, both removing far more than TCMM.

In general, GBOD tends to find different outliers compared to both TCMM and PhylteR (Fig. 6a). In each dataset except Frogs, only 0.5%-10% of the outliers GBOD identified were shared by either TCMM or PhylteR. For example, in the insect dataset, PhylteR and GBOD remove almost disjoint sets of outliers, yet they achieve an almost identical increase in normalized quartet score. On the other hand, in the Frogs dataset, the majority (54%) of the outliers GBOD identified were also identified by PhylteR. While PhylteR removed many more taxa than GBOD or TCMM, the increase in normalized quartet score was extremely similar across all three methods.

5 Discussion and Future Work

We introduced a general machine learning framework with a novel scheme for augmenting gene trees with positively labeled errors, which can be used to train models to recognize subtrees with unusual placements. This task is more computationally demanding than detecting outlier gene trees as a whole. The goal of our GBOD model is to detect which exact subtrees of a gene tree are positioned

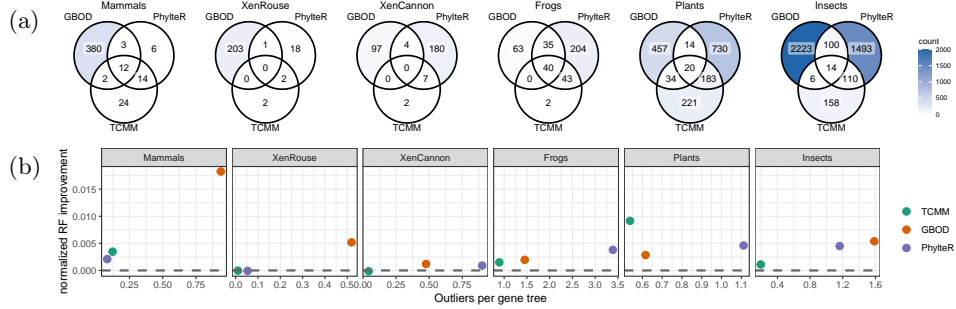


Fig. 6: (Biological Dataset) (a) Venn diagrams for the outliers marked by the three methods for each dataset, and their intersection. The number of outliers is shown on each partition. (b) The average number of outliers per gene tree versus the improvement in normalized RF distance (nrf) between the gene trees and the estimated species tree, defined as the difference between nrf after removing detected outliers from the gene trees and the nrf of original unfiltered gene trees with respect to the species tree. Above the dashed line shows improvement over the original gene trees.

in anomalous positions with respect to a species tree. These questions would be easy to answer using QSPR moves if gene trees were always extremely similar to the species tree; any QSPR to a sufficiently distant place would be anomalous. However, true gene tree discordance due to normal processes such as ILS makes it harder to distinguish outliers from normal heterogeneity (Fig. 1). The goal of the machine learning component is to distinguish the two.

Our simulations indicated that GBOD is successful in making such distinctions when outliers are due to HGT, but less so for chimeric sequences. On biological data, GBOD detected outliers that were quite different from the existing methods. In particular, it appeared that many of the issues with the known problematic mammals dataset were detected by GBOD. These data are extremely unlikely to have HGT. High effectiveness of GBOD on these data likely indicates that *some* forms of data errors can be detected by GBOD. Thus, the low recall of GBOD on the S200-perturbed dataset may be specific to its particular form of introducing errors by creating chimeric sequences. Since that mode of error introduction is not necessarily the most realistic scenario, future work should further explore other methods of simulating errors in the data.

The underlying augmentation strategy we use here is implemented using random SPR moves to change gene trees. SPR were chosen because they allow general modifications to the tree topology and are easy to implement. Our results showed both the promise and limitations of this approach. On the dataset with HGT outliers, which are similar though not identical to SPR moves [40], the results were strong. On the S200 perturbed dataset, where errors were not similar to SPR, the results were far less strong. On this dataset, TCMM, which accounts for branch lengths more explicitly than GBOD performed far better.

Thus, our results demonstrate that training a model to recognize SPRs alone has inductive biases that carry over to inference. Future work can expand on the data augmentation mechanisms used. Our choice of SPR-based data augmentation methods was motivated by their ease of application. A simple extension is to apply multiple rounds of SPR (one on top of another) to change gene trees more dramatically. Other approaches to explore in the future could include p -Edge-Contract-and-Refine moves [10] and tree-bisection and reconnection. Finally, we performed between 10 and 100 rounds of augmentation per gene tree; preliminary results show that the number of rounds does impact accuracy (Fig. S5). However, augmenting further will increase the running time of feature extraction in our current implementation. With some preprocessing, we may be able to perform QSPR moves on SPR-adjacent gene trees without recomputing all the counters, saving much of the running time.

Beyond data augmentation, both the choice of model and features are easily expanded upon. While our results showed that QSPR moves do have the signal to detect at least some types of outliers, there is no reason that our machine learning framework could not incorporate other features, such as results of TCMM and PhylteR, as extra features. Such combinations seem promising and should be explored in the future. In particular, our current method is mostly focused on topological outliers, but branch length outliers can be equally informative, motivating adding such features.

Gradient boosted trees, while computationally efficient to train, work best on tabular data. As such, in this framework, we tabularize the inherently tree-based data by assigning each internal node a feature vector and performing regression on each vector independently. A natural next step, then, is to directly incorporate trees into the model. To this point, the augmented, positively labeled clades are created in a dependent manner (e.g., all smaller subtrees of a moved clade are labeled as outliers); however, due to our tabularization of the data, our current framework ignores these dependencies and treats each subtree as an independent training point. A modification to this framework that more naturally lends itself to modeling the tree structure may significantly improve some results.

Acknowledgments. This study was funded by NIH MIRA grant 1R35GM142725.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

- [1] Abdi, H., O'Toole, A., Valentin, D., Edelman, B.: DISTATIS: the analysis of multiple distance matrices. In: Proceedings of the IEEE Computer Society: International Conference on Computer Vision and Pattern Recognition. pp. 42–42 (Jul 2005). <https://doi.org/10.1109/CVPR.2005.445>
- [2] Allio, R., Tilak, M.K., Scornavacca, C., Avenant, N.L., Kitchener, A.C., Corre, E., Nabholz, B., Delsuc, F.: High-quality carnivoran genomes from roadkill samples enable comparative species delineation in aardwolf and bat-eared fox. *eLife* **10**, e63167 (Feb 2021). <https://doi.org/10.7554/eLife.63167>, <https://doi.org/10.7554/eLife.63167>
- [3] Arasti, S., Mirarab, S.: Median quartet tree search algorithms using optimal subtree prune and regraft. *Algorithms for Molecular Biology* **19**(1), 12 (Mar 2024). <https://doi.org/10.1186/s13015-024-00257-3>, <https://doi.org/10.1186/s13015-024-00257-3>
- [4] Arasti, S., Tabaghi, P., Tabatabaee, Y., Mayer, Alan K., Mirarab, S.: Optimal Tree Metric Matching Enables Phylogenomic Branch Length Reconciliation. *Systematic Biology* **in press**, bioRxiv: 2023.11.13.566962 (2026). <https://doi.org/10.1101/2023.11.13.566962>
- [5] Cannon, J.T., Vellutini, B.C., Smith, J.r., Ronquist, F., Jondelius, U., Hejnol, A.: Xenacoelomorpha is the sister group to nephrozoa. *Nature* **530**(7588), 89–93 (Feb 2016)
- [6] Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). <https://doi.org/10.1145/2939672.2939785>, <https://doi.org/10.1145/2939672.2939785>
- [7] Comte, A., Tricou, T., Tannier, E., Joseph, J., Siberchicot, A., Penel, S., Allio, R., Delsuc, F., Dray, S., De Vienne, D.M.: PhylteR: Efficient Identification of Outlier Sequences in Phylogenomic Datasets. *Molecular Biology and Evolution* **40**(11), msad234 (Nov 2023). <https://doi.org/10.1093/molbev/msad234>, <https://academic.oup.com/mbe/article/doi/10.1093/molbev/msad234/7330000>
- [8] Feng, Y.J., Blackburn, D.C., Liang, D., Hillis, D.M., Wake, D.B., Cannatella, D.C., Zhang, P.: Phylogenomics reveals rapid, simultaneous diversification of three major clades of gondwanan frogs at the cretaceous–paleogene boundary. *Proceedings of the National Academy of Sciences* **114**(29), E5864–E5870 (2017)
- [9] Francois, C.M., Durand, F., Figuet, E., Galtier, N.: Prevalence and Implications of Contamination in Public Genomic Resources: A Case Study of 43 Reference Arthropod Assemblies. *G3* **10**(2), 721–730 (Feb 2020). <https://doi.org/10.1534/g3.119.400758>, <http://g3journal.org/lookup/doi/10.1534/g3.119.400758>

- [10] Ganapathy, G., Ramachandran, V., Warnow, T.: Better Hill-Climbing Searches for Parsimony. In: Algorithms in Bioinformatics, pp. 245–258 (2003). https://doi.org/10.1007/978-3-540-39763-2_19
- [11] Gatesy, J., Springer, M.S.: Phylogenetic Analysis at Deep Timescales: Unreliable Gene Trees, Bypassed Hidden Support, and the Coalescence/Concatalaescence Conundrum. *Molecular phylogenetics and evolution* **80**, 231–266 (2014). <https://doi.org/10.1016/j.ympev.2014.08.013>, <http://www.ncbi.nlm.nih.gov/pubmed/25152276>
- [12] Hosner, P.A., Faircloth, B.C., Glenn, T.C., Braun, E.L., Kimball, R.T.: Avoiding Missing Data Biases in Phylogenomic Inference: An Empirical Study in the Landfowl (Aves: Galliformes). *Molecular Biology and Evolution* **33**(4), 1110–1125 (Apr 2016). <https://doi.org/10.1093/molbev/msv347>, <https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msv347>
- [13] Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H.: Phylogenomics: the beginning of incongruence? *Trends in Genetics* **22**(4), 225–231 (2006). <https://doi.org/10.1016/j.tig.2006.02.003>, ISBN: 0168-9525
- [14] Laurin-Lemay, S., Brinkmann, H., Philippe, H.: Origin of land plants revisited in the light of sequence contamination and missing data. *Current Biology* (2012). <https://doi.org/10.1016/j.cub.2012.06.013>, ISBN: 1879-0445 (Electronic)\r0960-9822 (Linking)
- [15] Li-San Wang, Leebens-Mack, J., Wall, P.K., Beckmann, K., de Pamphilis, C.W., Warnow, T.: The Impact of Multiple Protein Sequence Alignment on Phylogenetic Estimation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **8**(4), 1108–1119 (Jul 2011). <https://doi.org/10.1109/TCBB.2009.68>, <http://www.ncbi.nlm.nih.gov/pubmed/21566256>
- [16] Mai, U., Mirarab, S.: TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* **19**(S5), 272 (May 2018). <https://doi.org/10.1186/s12864-018-4620-2>, <https://bmcbgenomics.biomedcentral.com/articles/10.1186/s12864-018-4620-2>, ISBN: 9783319679785
- [17] Mallo, D., De Oliveira Martins, L., Posada, D.: SimPhy : Phylogenomic Simulation of Gene, Locus, and Species Trees. *Systematic Biology* **65**(2), 334–344 (Mar 2016). <https://doi.org/10.1093/sysbio/syv082>, <http://sysbio.oxfordjournals.org/content/early/2015/12/04/sysbio.syv082.short?rss=1>
- [18] Mirarab, S., Bayzid, M.S., Boussau, B., Warnow, T.: Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* **346**(6215), 1250463–1250463 (Dec 2014). <https://doi.org/10.1126/science.1250463>, <http://www.sciencemag.org/cgi/doi/10.1126/science.1250463>
- [19] Mirarab, S., Bayzid, M.S., Boussau, B., Warnow, T.: Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* **346**(6215), 1250463 (2014)
- [20] Mirarab, S., Bayzid, M.S., Warnow, T.: Evaluating Summary Methods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage

- Sorting. *Systematic Biology* **65**(3), 366–380 (May 2016). <https://doi.org/10.1093/sysbio/syu063>, iSBN: 1063-5157
- [21] Mirarab, S., Rivas-González, I., Feng, S., Stiller, J., Fang, Q., Mai, U., Hickey, G., Chen, G., Brajuka, N., Fedrigo, O., Formenti, G., Wolf, J.B.W., Howe, K., Antunes, A., Schierup, M.H., Paten, B., Jarvis, E.D., Zhang, G., Braun, E.L.: A region of suppressed recombination misleads neoavian phylogenomics. *Proceedings of the National Academy of Sciences* **121**(15), e2319506121 (Apr 2024). <https://doi.org/10.1073/pnas.2319506121>, <https://doi.org/10.1073/pnas.2319506121>
- [22] Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A.J., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T.R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermini, L.S., Kawahara, A.Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D.D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von Reumont, B.M., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N.U., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walz, M.G., Wiegmann, B.M., Wilbrandt, J., Wipfler, B., Wong, T.K.F., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D.K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K.M., Zhou, X.: Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**(6210), 763–767 (Nov 2014)
- [23] Patel, S.: Error in Phylogenetic Estimation for Bushes in the Tree of Life. *Journal of Phylogenetics & Evolutionary Biology* **01**(02), 110 (2013). <https://doi.org/10.4172/2329-9002.1000110>, <http://www.esciencecentral.org/journals/error-in-phylogenetic-estimation-for-bushes-in-the-tree-of-life-2329-9002.1000110.php?aid=154>
- [24] Philippe, H., Vienne, D.M.d., Ranwez, V., Roure, B., Baurain, D., Delsuc, F.: Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy* (2017). <https://doi.org/10.5852/ejt.2017.283>
- [25] Portik, D.M., Wiens, J.J.: Do Alignment and Trimming Methods Matter for Phylogenomic (UCE) Analyses? *Systematic Biology* (Aug 2020). <https://doi.org/10.1093/sysbio/syaa064>, <https://doi.org/10.1093/sysbio/syaa064>
- [26] Price, M.N., Dehal, P.S., Arkin, A.P.: FastTree-2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5**(3), e9490 (Mar 2010). <https://doi.org/10.1371/journal.pone.0009490>, <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2835736&tool=pmcentrez&rendertype=abstract>

- [27] Reddy, S., Kimball, R.T., Pandey, A., Hosner, P.A., Braun, M.J., Hackett, S.J., Han, K.L., Harshman, J., Huddleston, C.J., Kingston, S.: Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. *Systematic biology* **66**(5), 857–879 (2017). <https://doi.org/10.1093/sysbio/syx041>
- [28] Rouse, G.W., Wilson, N.G., Carvajal, J.I., Vrijenhoek, R.C.: New deep-sea species of xenoturbella and the position of xenacoelomorpha. *Nature* **530**(7588), 94–97 (Feb 2016)
- [29] Salichos, L., Rokas, A.: Evaluating ortholog prediction algorithms in a Yeast Model Clade. *PLoS ONE* (2011). <https://doi.org/10.1371/journal.pone.0018755>, iSBN: 1932-6203 (Electronic)\r1932-6203 (Linking)
- [30] Sayyari, E., Whitfield, J.B., Mirarab, S.: Fragmentary Gene Sequences Negatively Impact Gene Tree and Species Tree Reconstruction. *Molecular Biology and Evolution* **34**(12), 3279–3291 (Dec 2017). <https://doi.org/10.1093/molbev/msx261>, <http://dx.doi.org/10.1093/molbev/msx261>
- [31] Sela, I., Ashkenazy, H., Katoh, K., Pupko, T.: GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Research* **43**(W1), W7–W14 (Jul 2015). <https://doi.org/10.1093/nar/gkv318>, <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv318>
- [32] Shwartz-Ziv, R., Armon, A.: Tabular data: Deep learning is not all you need. *Information Fusion* **81**, 84–90 (2022). <https://doi.org/https://doi.org/10.1016/j.inffus.2021.11.011>, <https://www.sciencedirect.com/science/article/pii/S1566253521002360>
- [33] Song, S., Liu, L., Edwards, S.V., Wu, S.: Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences of the United States of America* **109**(37), 14942–14947 (2012). <https://doi.org/10.1073/pnas.1211733109>
- [34] Springer, M.S., Gatesy, J.: The gene tree delusion. *Molecular Phylogenetics and Evolution* **94**(Part A), 1–33 (Jan 2016). <https://doi.org/10.1016/j.ympev.2015.07.018>, <http://www.sciencedirect.com/science/article/pii/S1055790315002225>, iSBN: 1095-9513 (Electronic)\r1055-7903 (Linking)
- [35] Springer, M.S., Gatesy, J.: On the importance of homology in the age of phylogenomics. *Systematics and Biodiversity* **16**(3), 210–228 (Apr 2018). <https://doi.org/10.1080/14772000.2017.1401016>, <https://www.tandfonline.com/doi/full/10.1080/14772000.2017.1401016>
- [36] Stamatakis, A.: RAXML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**(9), 1312–1313 (2014). <https://doi.org/10.1093/bioinformatics/btu033>, arXiv: 10.1093/bioinformatics/btu033 iSBN: 1367-4811
- [37] Steenwyk, J.L., Buida, T.J., Li, Y., Shen, X.X., Rokas, A.: ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLOS Biology* **18**(12), e3001007 (Dec 2020). <https://doi.org/10.1371/journal.pbio.1001007>

- 1371/journal.pbio.3001007, <https://dx.plos.org/10.1371/journal.pbio.3001007>
- [38] Steenwyk, J.L., Li, Y., Zhou, X., Shen, X.X., Rokas, A.: Incongruence in the phylogenomics era. *Nature Reviews Genetics* **24**(12), 834–850 (Dec 2023). <https://doi.org/10.1038/s41576-023-00620-x>, <https://www.nature.com/articles/s41576-023-00620-x>
 - [39] Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., Dessimoz, C.: Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Systematic Biology* **64**(5), 778–791 (Sep 2015). <https://doi.org/10.1093/sysbio/syv033>, <https://academic.oup.com/sysbio/article-lookup/doi/10.1093/sysbio/syv033>
 - [40] Tannier, E., Tricou, T., Benali, S., De Vienne, D.M.: HGTs are not SPRs: In the Presence of Ghost Lineages, Series of Horizontal Gene Transfers do not Result in Series of Subtree Pruning and Regrafting. *Molecular Biology and Evolution* **42**(6), msaf128 (Jun 2025). <https://doi.org/10.1093/molbev/msaf128>, <https://academic.oup.com/mbe/article/doi/10.1093/molbev/msaf128/8154856>
 - [41] Van Etten, J., Bhattacharya, D.: Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? *Trends in Genetics* **36**(12), 915–925 (Dec 2020). <https://doi.org/10.1016/j.tig.2020.08.006>, <https://linkinghub.elsevier.com/retrieve/pii/S0168952520302067>
 - [42] Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., Ruhfel, B.R., Wafula, E., Der, J.P., Graham, S.W., Mathews, S., Melkonian, M., Soltis, D.E., Soltis, P.S., Miles, N.W., Rothfels, C.J., Pokorny, L., Shaw, A.J., DeGironimo, L., Stevenson, D.W., Surek, B., Villarreal, J.C., Roure, B., Philippe, H., dePamphilis, C.W., Chen, T., Deyholos, M.K., Baucom, R.S., Kutchan, T.M., Augustin, M.M., Wang, J., Zhang, Y., Tian, Z., Yan, Z., Wu, X., Sun, X., Wong, G.K.S., Leebens-Mack, J.: Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences* **111**(45), E4859–E4868 (2014)
 - [43] Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S.: ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**(S6), 153 (May 2018). <https://doi.org/10.1186/s12859-018-2129-y>, <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2129-y>
 - [44] Zhang, C., Zhao, Y., Braun, E.L., Mirarab, S.: TAPER: Pinpointing errors in multiple sequence alignments despite varying rates of evolution. *Methods in Ecology and Evolution* **12**(11), 2145–2158 (Nov 2021). <https://doi.org/10.1111/2041-210X.13696>, <https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.13696>

Supplementary Material

Table S1: (Carnivora Simulated Dataset) Area under the curve (AUC) for all methods. Best Performance for each ILS level is highlighted.

| Pop Size (ILS) | GBOD | PhylteR | TCMM($\lambda = 1$) |
|-----------------|-------|--------------|-----------------------|
| 10 | 0.820 | 0.936 | 0.931 |
| 1×10^5 | 0.763 | 0.886 | 0.912 |
| 2×10^5 | 0.782 | 0.855 | 0.896 |
| 5×10^5 | 0.671 | 0.795 | 0.835 |

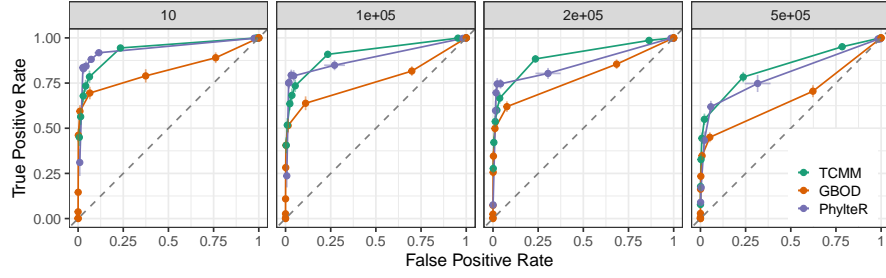


Fig. S1: (Carnivora Simulated Dataset) ROC curve for each method across all levels of ILS, showing the entire range.

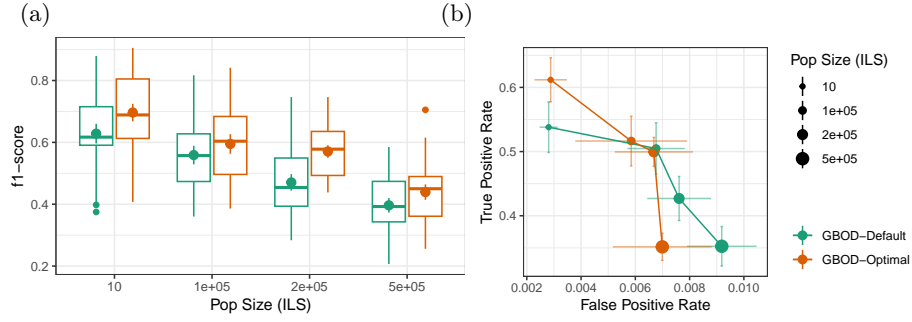


Fig. S2: (Carnivora Simulated Dataset) (a) Comparing GBOD's default threshold selection versus the optimal possible on various levels of ILS using F_1 -score. (b) False Positive Rate (FPR) versus True Positive Rate (TPR) for each method across various levels of ILS.

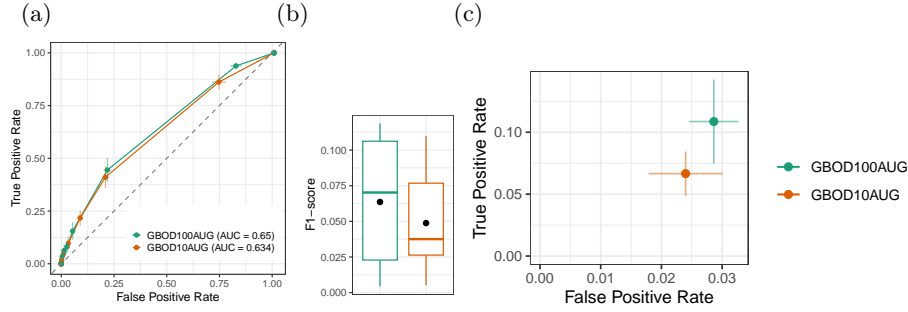


Fig. S3: (S200-perturbed Dataset - 10 Sampled Replicates) (a) ROC curve comparing our method with 99 augmentations per gene tree versus 9. Area under curve (AUC) values are shown for each method. (b) Comparing differences in the number of augmented gene trees in GBOD using F_1 -score. Mean and standard error are shown on top of the box plots. (c) False Positive Rate (FPR) versus True Positive Rate (TPR) for each method. Mean and standard error are shown.

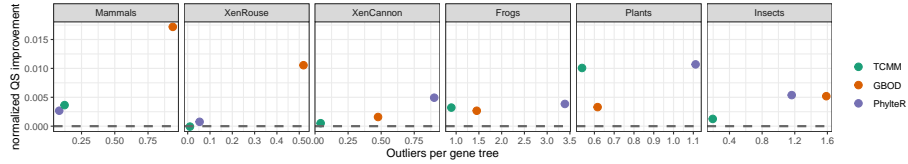


Fig. S4: (Biological Dataset) The average number of outliers per gene tree versus the improvement in normalized quartet score (nqs) between the gene trees and the estimated species tree, defined as the difference between nqs after removing detected outliers from the gene trees and the nqs of original unfiltered gene trees with respect to the species tree. Above the dashed line shows improvement over the original gene trees.

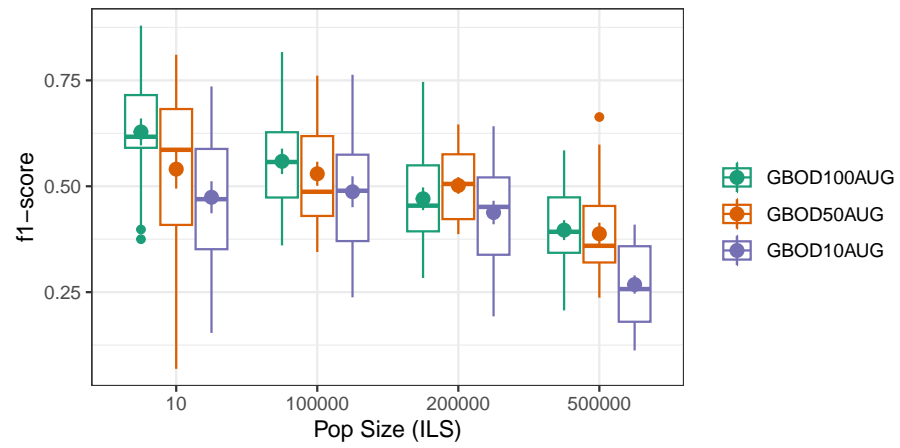


Fig. S5: (Carnivora Simulated Dataset) F_1 scores varying the number of times each gene tree is augmented to produce the training dataset.