Tractable time tree distributions

Tobia Ochsner[1, 2], Jonathan Klawitter[1], Alexei J. Drummond[1]*,

1 University of Auckland, Auckland, Aotearoa/New Zealand
2 ETH Zurich, Basel, Switzerland

* a.drummond@auckland.ac.nz

Abstract

In Bayesian phylogenetics, Markov Chain Monte Carlo (MCMC) methods generate samples of time trees to capture the posterior distribution of the phylogenetic tree. We explore novel approaches to fit distributions on these time trees samples. In particular, we extend the family of Conditional Clade Distributions (CCD) to model the tree topology and branch lengths simultaneously. Our distributions are designed to capture both central tendencies and dependencies within this high-dimensional and non-Euclidean space. They facilitate a straightforward reconstruction of a summary tree and manage to capture the tails of the time tree distribution. This property extends their utility beyond the robust reconstruction of a single summary tree, enabling more sophisticated downstream analyses like exploring credible regions or evaluating alternative evolutionary hypotheses.

Author summary

Introduction

Markov Chain Monte Carlo (MCMC) methods have become a cornerstone of Bayesian phylogenetics. Given a set of aligned molecular sequences and an evolutionary model (consisting of a tree model, a molecular clock model, a substitution model, and prior distributions on parameters), MCMC algorithms like Beast 2 [1], RevBayes [2] or MrBayes [3] generate a set of phylogenetic trees sampled from the posterior distribution. This posterior sample usually consists of thousands of trees and contains a wealth of information about the evolutionary relationships encoded in the provided sequences.

For simple continuous or indicator parameters of the phylogenetic model (like substitution or clock rates), analyzing the marginal distributions of the posterior sample is relatively straightforward. Tools like Tracer [4] provide readily available summaries and visualizations of these marginal distributions.

The space of phylogenetic trees presents a significantly greater challenge. The sheer number of possible rooted trees explodes super-exponentially with increasing taxa, making it highly unlikely that the MCMC samples contain the true tree topology [5]. The *curse of dimensionality* [6, 7] diminishes the effectiveness of common sample-based statistical methods, particularly distance-based approaches. We demonstrate in the Supplementary Information that MCMC samples can exhibit strikingly similar pairwise distances, contradicting the expectation that regions of higher probability density should yield samples with closer proximity. Furthermore, the space of possible node times is non-Euclidean, as the times are inherently constrained by the topology [8–12].

These challenges make it difficult to directly analyze the posterior sample of trees and branch lengths. One strategy to extract meaningful information from the sample involves dedicated post-processing techniques. The raw sample is distilled into representations that enable a more meaningful analysis and interpretation.

A common approach is to infer a point estimate, a single "best" tree that summarizes the posterior sample [13]. Several heuristics have been developed for this purpose, including DensiTree visualizations [14], Maximum Clade Credibility (MCC) trees, Hipster [15], and the Maximum A Posteriori tree based on Conditional Clade Distributions (CCD) [5, 16]. These methods often focus primarily on recovering a likely tree topology. Determining branch lengths is treated as an afterthought, potentially overlooking important information about evolutionary timescales.

Another class of algorithms simultaneously considers tree topology and branch lengths by embedding trees in a particular metric space. Examples include the Billera-Holmes-Vogtmann (BHV) space [12], the Wald space [9], the t and τ space [10], the palm tree space [11], the RNNI space [17], and the recently developed Wald space [9]. Within these spaces, a natural approach to obtain a point estimate is to calculate the Fréchet mean of the posterior tree sample [18]. However, these spaces often suffer from limitations such as the *sticky mean problem* (where the mean tree is almost independent of the actual tree samples) [19], high computational complexity associated with distance calculations, and the inherent difficulties of performing statistical inference in complex geometric spaces [20].

While obtaining a point estimate from the posterior sample is crucial, it is important to recognize that the MCMC sample contains information about the entire posterior distribution. Tractable tree distributions [5] try to capture the posterior distribution using a well-defined approximate distribution. This provides an elegant framework for testing alternative evolutionary hypotheses, determining whether a specific tree falls within a credible region, calculating statistics like information content [21], and analyzing dependencies between different parts of the tree (e.g., relationships between branch lengths and topological features). Existing methods that estimate distributions on the posterior sample for rooted trees focus on modeling tree topologies [5, 16], ignoring time information entirely. The family of Bayesian subsplit networks [22] has been extended to model branch lengths in multiple ways [23, 24]. However, Bayesian subsplit methods focus on direct variational inference, rather than posterior sample summarization.

Here, we address the challenge of inferring tractable distributions on time trees, encompassing both topology and branch lengths. We compare different methods and tree embeddings, and demonstrate that the described methods exhibit significant differences in terms of capturing central tendencies and the tails of the distribution. Finally, we introduce a software package that enables researchers to fit and analyze these distributions, facilitating more nuanced insight into phylogenetic uncertainty. This tool allows for a more complete utilization of the information within the posterior tree sample, moving beyond simple point estimates.

Materials and methods

In this section, we introduce the concept of tractable tree distributions on time trees and discuss desirable properties of such distributions. We then introduce a family of distributions satisfying these properties and show how to perform common operations such as parameter estimation and obtaining a point estimate. Finally, we outline the procedure used to assess and compare the distributions on both synthetic and real datasets.

Tractable tree distributions

Berling et al. [5] introduced *tractable tree distributions* as a set of properties of a distribution on tree topologies. A distribution is tractable if a range of operations can be performed efficiently in practice. We apply the same concept to distributions on *time trees*.

Tractable time tree distributions are used as a tool to analyze a posterior tree sample. Therefore, we should be able to efficiently infer its parameters based on that tree sample. In order to compare different tree hypotheses, we need to evaluate the probability of any time tree under the model. Another important operation is sampling, as it enables to test if a tree is in the 95% credible region and to inspect the dependencies of different node heights. Lastly, a tractable time tree distribution should provide us with a *summary tree*, for instance by retrieving the highest-probability or expected tree.

Mathematical framework

A *time tree* $(T, \{H_v\}_{v \in V})$ for a set of taxa X and root R consists of an undirected tree $T = (V, E)$ and a set of random variables $\{H_v\}_{v \in V} \in \mathbb{R}^+$ denoting the vertex heights (the divergence times). The following properties have to be satisfied:

1. (V, E) describe a fully connected tree.

2. The root $R \in V$ has degree 2. The taxa $X \subset V$ have degree 1. All other vertices are *internal* vertices and have degree 3.

3. We only consider contemporary sampling of the taxa: $H_x = 0$ for all taxa x.

4. Each edge e in E consists of a parent p and child c such that the parent is closer to the root. The heights must describe a valid tree and evolution runs backward in time: $H_c < H_p$ for all edges (p, c) in E.

The probability of a time tree can be factorized into the probability of its topology T and the probability of the vertex heights $\{H_v\}$ given the topology:

$$P\left(T, \{H_v\}_{v \in V}\right) = P\left(T\right) P\left(\{H_v\}_{v \in V} | T\right).$$

To model $P\left(T\right)$, we use the conditional clade distributions *CCD0* and *CCD1* [5, 16]. They already exhibit the desired properties and have shown to work well in practice [5, 15, 25]. Thus, we focus on modelling the conditional probability $P\left(\{H_v\}_{v \in V} | T\right)$. It is crucial to note that we do not restrict ourselves to a single topology T—we simply model $P\left(\{H_v\}_{v \in V} | T\right)$ as a function of the topology T.

Credible regions

A key tool in our analysis is the *smallest γ-credible region* of the tree posterior. This region, defined as the minimum set of time trees with a combined probability mass of γ, provides a straightforward way to assess the compatibility of a tree with the posterior. It allows, for instance, to determine if a time tree obtained with a completely different methodology is in the smallest 95%-credible region (a random posterior tree has a 95% probability of being in the 95%-credible region). There is a surprisingly simple way to assess if a given time tree is within the smallest γ-credible region of a distribution: We generate a large number of samples from the distribution and rank the model probabilities of the generated trees. If the model probability of the given tree ranks within the top γ% of the sampled tree probabilities, it is in the smallest γ-credible region. This approach works because every point inside the smallest credible region has a probability density at least as large as any point outside the region.

Clade parameterisation

There are several main challenges when designing a distribution on the heights.

First, $P\left(\{H_v\}_{v\in V}|T\right)$ is a function of the topology T. However, we cannot assume that every possible topology is found in the MCMC sample. Hence, $P\left(\{H_v\}_{v\in V}|T\right)$ needs to be defined even for unknown topologies. Moreover, as the number of possible topologies can be extremly large, we have to share parameters amongst different topologies. In order to overcome this challenge, we employ the same strategy as the CCD0 model [5]: instead of random variables corresponding to vertices of a topology, they correspond to clades.

$$P\left(\{H_v\}_{v\in V}|T\right) = P\left(\{H_c\}_{c\in C(T)}|T\right),$$

where $C(T)$ is the set of clades in topology T. A clade is a subset of taxa that share a common ancestor. While every vertex corresponds to exactly one clade, modeling random variables on a clade-level allows to share parameters amongst different topologies that have common clades. We call a clade *non-trivial* if it is neither a singleton nor the root clade. In the following, we exclude the singleton clades corresponding to the taxa from $C(T)$, as H_c equals 0 for all singleton clades and does not need to be modelled.

Another difficulty is the fact that every topology enforces a different set of constraints on the heights. We expect $P\left(\{H_c\}_{c\in C(T)}|T\right)$ to vanish for heights invalid under the topology. We solve this by introducing the concept of a *clade embedding.* Given a topology T, a clade embedding f_T induces a bijection from the heights into another set of random variables in the *embedding space*:

$$f_T : \{H_c\}_{c\in C(T)} \mapsto \{Z_c\}_{c\in C(T)}.$$

The embedding can be designed such that the Z_c have nice constraints independent of the topology (for instance, $Z_c \in \mathbb{R}$ or $Z_c \in [0,1]$). Equipped with an embedding, we can define a probability distribution on $\{Z_c\}_{c\in C(T)}$ as well as the corresponding pulled-back probability distribution on $\{H_c\}_{c\in C(T)}$:

$$P\left(\{H_c\}_{c\in C(T)}|T\right) = P\left(\{Z_c\}_{c\in C(T)}|T\right)\left|\det\left(J_{f_t^{-1}}(\{Z_c\}_{c\in C(T)})\right)\right|,$$

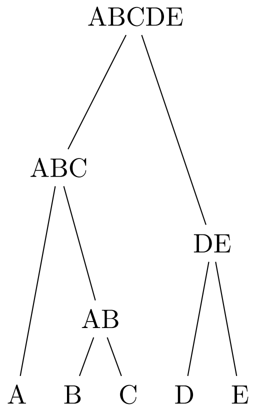
where $J_{f_t^{-1}}(\{Z_c\}_{c\in C(T)})$ is the Jacobian of the inverse of the embedding.

We can choose commonly used probability distributions in the embedding space. In combination with the embedding and inverse embedding, this leads to a probability distribution that is both tractable and respects the inherent constraints of vertex heights. In order to get the probability of a set of heights under a topology, we first transform the set of heights into the embedding space. Then, we can use the pulled-back probability distribution to calculate the probability. If we want to sample from the distribution, we can sample from the embedding space and transform the samples back into the original height space.

There is another challenge that arises when designing a distribution on the heights. Dependencies between node heights are fundamentally tied to the height constraints and the restriction to contemporary sampling. However, other factors can influence branch length dependencies as well. For instance, a root height calibration prior can cause an anti-correlation between the pendant branch lengths and external branch lengths. The chosen embedding method can also affect these dependencies, possibly decreasing the correlation of the embedded random variables.

In the following sections, we introduce three clade embeddings equipped with different strategies for assigning probabilities in the embedding space (see figures 1 and 2 for examples). Our general approach is to find the best approximation of the marginal time tree posterior distribution of the MCMC model. Thus, biological interpretation of the designed distributions is not a primary goal. Instead, we focus on a wide range of embeddings and distributions that fulfill the properties of tractable time tree distributions and then compare their fit with the MCMC posterior distribution on various real and simulated datasets.

Fig 1. This example shows the same time tree embedded into the different embedding spaces. We assign a value to every non-singleton clade. The second column (Absolute heights) contains the absolute vertex height. The third column (Last divergence) denotes the time since the last divergence for the root vertex and the length of the branch leading to a vertex for the internal vertices. The third embedding (Shorter branch) denotes the length of the branch leading to the older child. The last column (Shorter branch) represents the total tree height for the root clade and the ratio of the internal vertices to their parents.



Clade	Absolute heights	Last divergence	Shorter branch	Height ratio
ABCDE	50	10	20	50
ABC	30	20	20	0.6
AB	10	20	10	0.33
DE	20	30	20	0.4

Last divergence embedding

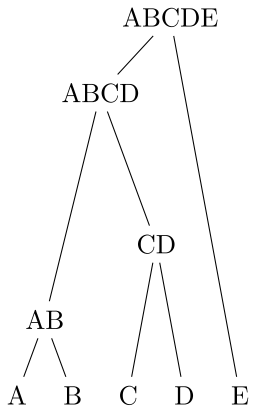
Under the birth death model, internal branch lengths are assumed to be independent realizations of a stochastic evolutionary process. In contrast, pendant branch lengths are primarily influenced by the time of sampling. We propose to model internal branch lengths using independent distributions, while separatly accounting for the time between sampling and the last divergence event. This approach aligns with theoretical findings suggesting that internal and pendant branch lengths follow notably different distributions [26, 27].

Specifically, we consider the branch length leading to a vertex v corresponding to a non-trivial clade c:

$$Z_c = H_{p_v} - H_v.$$

Furthermore, we consider the time between the present and the last divergence event and assign it to the random variable corresponding to the root clade X:

Fig 2. A second example.



Clade	Absolute heights	Last divergence	Shorter branch	Height ratio
ABCDE	50	10	10	50
ABCD	40	10	20	0.8
AB	10	30	10	0.25
CD	20	20	20	0.5

$$Z_X = \min_{x \in X} H_x.$$

This embedding allows to choose arbitrary positive Z_c while always producing a valid tree. Thus, we propose to use either independent lognormal, independent gamma, or independent Weibull distributions:

$$P(\{Z_c\}_{c \in C(T)}) = \begin{cases} \prod_{c \in C(T)} LogNormal(Z_c; \mu_c, \sigma_c), \\ \prod_{c \in C(T)} Gamma(Z_c; \alpha_c, \beta_c), \\ \prod_{c \in C(T)} Weibull(Z_c; k_c, \lambda_c). \end{cases}$$

Different models for divergence suggest different distributions on the branch lengths [28]. For instance, a constant rate of divergence results in an exponential distribution, which is a special case of both the gamma and the weibull distribution. If the rate of divergence changes with the amount of time since the last divergence, the resulting distribution will be closer to a Weibull distribution.

Shorter branch embedding

We introduce an alternative embedding that directly utilizes absolute branch lengths. For a clade c corresponding to vertex v, we define Z_c as the branch length leading to the older child of v.

$$Z_c = H_v - \max{(H_{a_v}, H_{b_v})},$$

where a_v and b_v represent the two child vertices of v. This parameterisation ensures that any set of positive Z_c, given a topology, results in a valid contemporary sampled time tree. To assign probabilities to these branch lengths, we consider same the three probabilistic models:

$$P(\{Z_c\}_{c\in C(T)}) = \begin{cases} \prod_{c\in C(T)} LogNormal(Z_c;\mu_c,\sigma_c), \\ \prod_{c\in C(T)} Gamma(Z_c;\alpha_c,\beta_c), \\ \prod_{c\in C(T)} Weibull(Z_c;k_c,\lambda_c). \end{cases}$$

Height ratio embedding

Inspired by TreeFlow [29], we introduce the height ratio embedding to capture the relative temporal structure of the vertices in time trees. For any non-trivial clade c correspoding to vertex v, we define Z_c as the ratio of the height of v to the height of its parent vertex p_v.

$$Z_c = \frac{H_v}{H_{p_v}}.$$

In addition to the relative positions of the vertices, we model the overall height of the time tree using the random variable corresponding to the root clade X:

$$Z_X = H_R.$$

Consequently, Z_X takes values in $\mathbb{R}^+$. For any non-trivial clade c, Z_c is constrained to the interval $[0,1]$. The primary advantage of the height ratio embedding lies in its ability to mitigate the effects of dependencies introduced by contemporary sampling and potential tree height constraints.

We model the ratios and the tree height independently. For the height ratios, we consider the following probability distributions:

$$P(\{Z_c\}_{c\in C(T)\setminus X}) = \begin{cases} \prod_{c\in C(T)\setminus X} Beta(Z_c;\alpha_c,\beta_c), \\ \prod_{c\in C(T)\setminus X} LogitNormal(Z_c;\mu_c,\sigma_c). \end{cases}$$

We further experimented with a generalized Dirichlet distribution, parameterized by a single parameter for each non-trivial clade. It is a special case of the independent beta distributions, defined such that the marginal distribution of any path from the root vertex to a leaf follows a Dirichlet distribution. However, it did not perform nearly as well as the other distributions and parameter estimation was not as straightforward, leading to an exclusion from the comparison.

We model the overall tree height with a lognormal distribution:

$$P(Z_X) = \text{LogNormal}(Z_X;\mu_X,\sigma_X).$$

Operations on tractable time tree distributions

We use the maximum likelihood estimator to estimate the parameters for each embedding and the corresponding distributions. The stated independence assumptions significantly simplify this process, reducing it to the maximum likelihood estimation of basic univariate continuous distributions. For the lognormal and logitnormal distribution, the maximum likelihood estimators exhibit well-known closed form solutions. We apply Newton's method for the gamma, beta and weibull distributions. A critical assumption of the maximum likelihood estimators is that the samples are independent. To meet this assumption for the MCMC time trees, we estimate the effective sample size (tree ESS) and then use a corresponding subsample to fit the distribution. The ESS is estimated using a procedure described in the supporting information.

In order to calculate the probability for a given time tree, we first transform the set of heights into the embedding space. Then, we can use the pulled-back probability distribution to calculate the probability. Sampling from the distribution is achieved by drawing samples from the embedding space and mapping them back to the original height space using the inverse transformation. Another important operation is constructing a point estimate based on a fitted tractable time tree distribution. One natural choice is the highest probability tree, as it is done for the CCD distributions [5]. For models based on gamma or weibull distributions, this can lead to a tree with branch lengths of 0. Hence, we use the expected vertex heights as a point estimate.

To determine if a tree is in the smallest γ-credible region of the approximated posterior distribution, we first sample from the distribution and rank the model probabilities of the generated time trees. The given tree is in the smallest γ-credible region if the model probability of the given tree ranks within the top γ% of the sampled tree probabilities.

A final operation that becomes tractable for the given embeddings is the calculation of the (differential) entropy of the distribution:

$$\begin{aligned} h(\{H_v\}_{v\in V}) =& h(\{Z_c\}_{c\in C(T)}) \\ &+ \int P(\{Z_c\}_{c\in C(T)}) \log \left|\det\left(J_{f_t^{-1}}(\{Z_c\}_{c\in C(T)})\right)\right| d\{Z_c\}_{c\in C(T)}. \end{aligned}$$

Validation

We use two complementary approaches to evaluate the quality of our proposed tractable time tree distributions: *model comparison* and *goodness-of-fit (GoF) testing*. We perform a train-validation split of the decorrelated subsample of the posterior time trees. The first 75% of the posterior samples is used as the training set for parameter estimation, while the remaining 25% is used as the validation set.

Given a posterior set of time trees, we calculate the data likelihood to compare the relative performance of our distributions on that set. Since all our models have the same number of parameters, we do not have to use a metric like the Akaike Information Criterion (AIC) to balance the likelihood with the parameter count. While the likelihood allows us to rank different models, it does not determine if any model accurately reflects the posterior distribution.

Hence, we also perform a visual goodness-of-fit test to assess how well our tractable tree distributions capture the posterior distribution. The test is based on the credible regions of the fitted distributions and compares them to the MCMC posterior samples. Specifically, we count the number of MCMC trees in the smallest γ-credible regions of the distributions for different values of γ. If the approximation is successful, we expect γ% of the MCMC trees to be in the smallest γ-credible region. Plotting the fraction for every credible region should thus result in a straight line. We use the *Cramér-von Mises* criterion [30] to quantify the deviation from the $x = y$ line by calculating the sum of the squared differences between the observed and expected fraction of trees in the smallest γ-credible region. The smaller the value, the better the fit.

There is a close connection to the *rank-uniformity plots* used in Bayesian model validation (first introduced by [31], later extended by [32] and applied to phylogenetics by [33]). Given any probabilistic model, rank-uniformity plots are used to test the interplay of a simulator generating samples under the model and an inference algorithm approximating the posterior distribution of the model parameters using the samples. For example, we can model time trees using a Yule model with a lognormal prior on the birth rate. A simulator would generate k samples of birth rates and tree topologies. We can run MCMC for each sample to obtain a posterior sample of birth rates. As the birth rate is a continuous parameter, we can apply the rank-uniformity plots as follows: For each of the k replications, we calculate the rank of the true birth rate among the MCMC-sampled birth rates. If everything is working as expected, the distribution of ranks is uniformly distributed. Plotting the empirical cumulative distribution function (ECDF) of the ranks for each replication should thus result in a straight line.

There are two differences between the rank-uniformity plots and the credible-set plots. First, while the rank-uniformity plots compare the true distribution with the posterior MCMC distribution, the credible-set plots compare the MCMC distribution with its approximation. Second, the rank-uniformity plots directly operate on univariate continuous parameters, whereas the credible-set plots transform each time tree into a univariate continuous space by computing its probability under the fitted model.

For the simulated datasets, the true tree is known. Thus, we can use them to assess the accuracy of the point estimators. Note that we use the same point estimate for the topology (CCD0 MAP) for all the models, hence the only difference is the heights of the vertices. Furthermore, we can compare the point estimators to the current default method implemented in TreeAnnotator [34]: for any given internal node corresponding to a clade C, the average height of the most common recent ancestor of C in the posterior samples is used. We employ the following metric to calculate the distance between the true and estimated tree:

$$SRBS\left(\{H_c^{(1)}\}_{c\in C(T_1)}, \{H_c^{(2)}\}_{c\in C(T_2)}\right) = \sum_{c\in C(T_1)\cap C(T_2)} \left(H_c^{(1)} - H_c^{(2)}\right)^2 .$$

Note that this is a modified version of the *Squared Branch Score* [13], where only the clades in common are considered.

Datasets

We conduct a comprehensive comparison of different distributions using both simulated and real datasets. For the simulated datasets, we used LPhyBEAST [35] to generate 100 trees corresponding sequence alignments under the Yule process with 100, 200 and 400 taxa respectively. Afterwards, we performed MCMC inference with BEAST 2 [1] on each of the obtained 300 sequence alignments. We refer to the Supplementary Information for more details on the simulations.

To assess the performance on biological datasets, we obtained 44 posterior time tree distributions from Dryad [36]. Specifically, we filtered the database for files with a *.trees* extension and manually selected those containing a posterior sample of ultrametric time trees obtained using BEAST 2 or a similar software. An exhaustive list of the studies considered and potential reasons for exclusion are provided in the supporting information (see also table 1). The datasets are part of a growing collection of phylogenetic posterior trees [37] curated to enable better method validation.

Results

We now apply the presented methods on simulated and real datasets. We use CCD1 to model the topology. All results are on the held-out validation set except when explicitly stated otherwise.

Data likelihood

The first metric we consider is the likelihood of the validation samples for each of the fitted Distributions. Because we downsample the MCMC trees to its effective sample

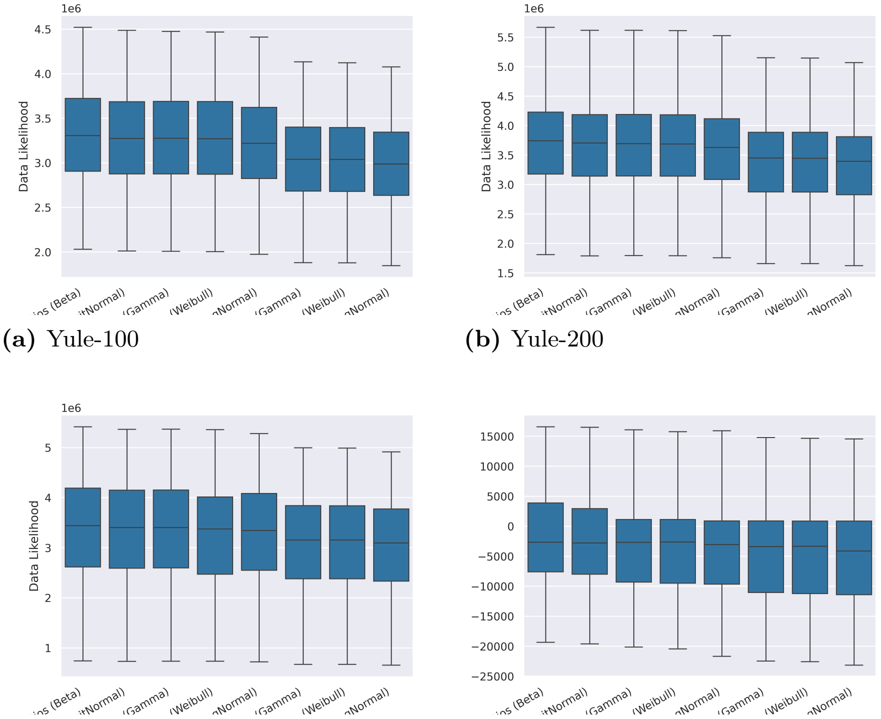
Table 1. Some statistics on the biological datasets used in this study. The datasets were obtained from Dryad [36] by filtering for files with a *.trees* extension and manually selecting those containing a posterior sample of ultrametric time trees obtained by BEAST 2 or a similar software.

Summary of the biological datasets	
Number of datasets	56
Number of datasets on plants	24
Number of datasets on animals	32
Number of studies	34
Number of taxa (min, max, mean)	5-441 (93)
Number of trees (min, max, mean)	100-100'001 (12'468)

size, we can assume that the samples are independent. The joint data likelihood is thus given by the product of the probability for each tree in the validation set.

Figure 3 shows the distribution of the log data likelihoods for the different distributions and datasets. There is a clear pattern where the height ratio distributions have the highest likelihood, followed by the shortest branch and last divergence embeddings. This pattern is also visible when looking at the ranking of each method for every single MCMC sample in figure 4.

Fig 3. The log data likelihoods for the different distributions and datasets. (The higher the better.)



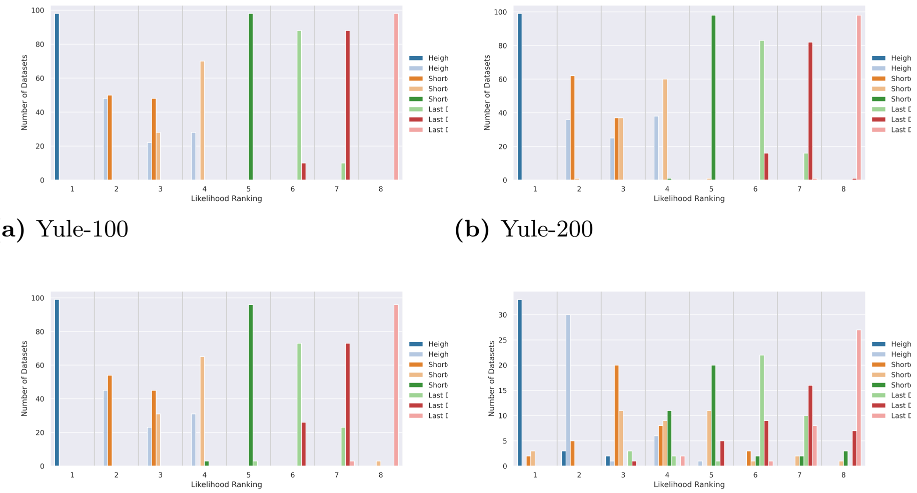
(a) Yule-100

(b) Yule-200

(c) Yule-400

(d) Real-world datasets

Fig 4. This figure shows how often a specific method ranks at a certain position when compared to the other methods. For example, the height ratio embedding with the beta distribution ranked first for almost all datasets. (The lower the rank the better.)



(a) Yule-100 **(b)** Yule-200

(c) Yule-400 **(d)** Real-world datasets

Goodness-of-fit

Accuracy of the point estimators

For the simulated datasets, the true tree is known. Thus, we can use them to assess the accuracy of the point estimators. Note that we use the same point estimate for the topology (CCD0 MAP) for all the models, hence the only difference is the heights of the vertices. Figure 5 shows the distance to the true tree for the different distributions (and MRCA) and simulated datasets. Figure 6 shows how often a specific method ranks at a certain position when compared to the other methods. It is apparant that none of the presented methods performs better than using MRCA for a majority of datasets. However, ignoring the last divergence embeddings, all methods produce point estimates of similar quality.

Discussion

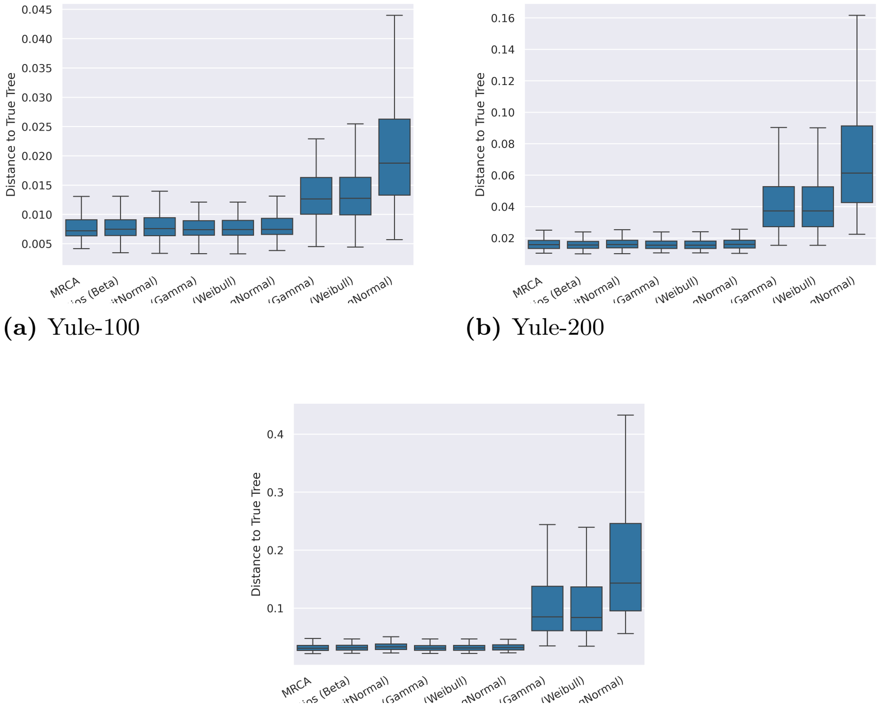
Conclusion

Supporting information

Julia package

A Julia package [38] was developed implementing all the methods presented in this paper. It can be found at `https://github.com/tochsner/TractableTimeTreeDistributions` and can be used to fit and analyze the distributions for any time tree sample. A web version accessible from a browser is in development.

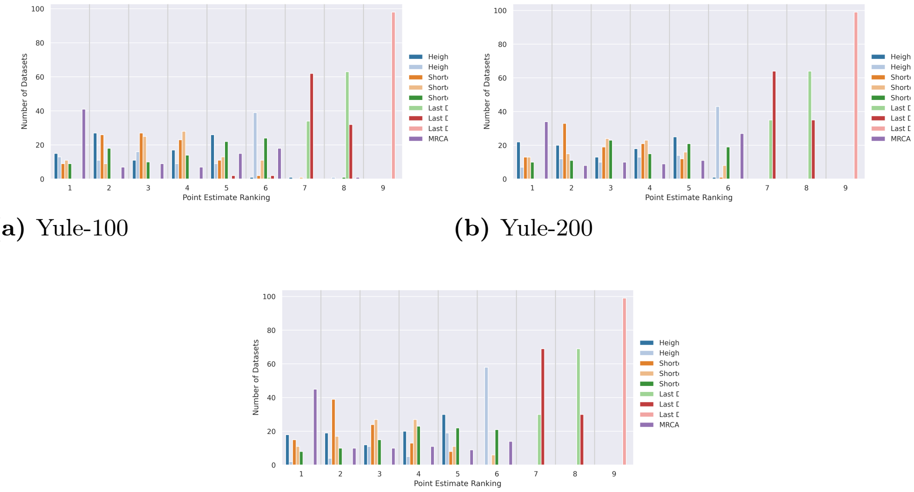
Fig 5. The log squared branch distance to the true tree for point estimates using the different distributions and datasets. (The lower the better.)



(a) Yule-100 **(b)** Yule-200

(c) Yule-400

Fig 6. How often each method ranked at a certain position when compared to the other methods. For example, MRCA ranked first for the majority of datasets. (The lower the rank the better.)



(a) Yule-100 **(b)** Yule-200

(c) Yule-400

The curse of dimensionality

This section examines the susceptibility of time trees to the *curse of dimensionality* [6, 7], a phenomenon where the complexity of a problem increases exponentially with the number of dimensions. Specifically, we look at the pairwise Robinson–Foulds (RF) distances between the MCMC trees of the synthetic Yule datasets. We find that the distances increase with the number of taxa while the relative variation of the distances decreases (see figures 7 and 8). This defies the intuition that regions of higher probability density should yield samples with closer proximity and might complicate the application of conventional distance-based methods to analyze the posterior sample. Previous work suggests that the curse of dimensionality might not be limited to simulated trees. The datasets analyzed by [39] show a median intrinsic dimensionality of around 7 when using the RF distance—potentially high enough to degrade the performance of distance-based methods [40].

Fig 7. The average pairwise Robinson–Foulds distance between the MCMC trees of the synthetic Yule datasets. The distances increase with the number of taxa.

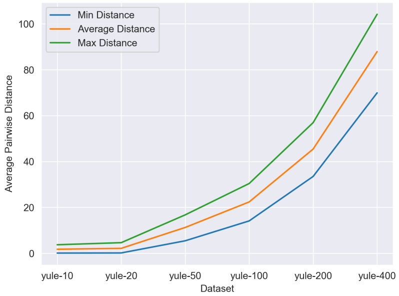
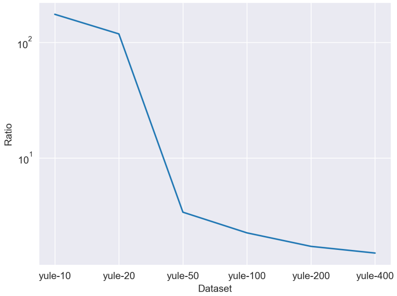


Fig 8. The average ratio of the minimum to maximum pairwise Robinson–Foulds distance between the MCMC trees of the synthetic Yule datasets. The ratios decrease with the number of taxa.



Correlations

Tree ESS

MCMC generates posterior time tree samples by performing a walk through the tree space, where new tree candidates are proposed by modifying the current tree using specific operations. Each tree depends on the previous one, resulting in autocorrelation in the samples. The maximum likelihood estimation and the data likelihood calculation assume that the samples are independently sampled. Thus, we estimate the effective sample size (ESS) of the MCMC samples and use it to subsample the posterior samples down to a set of (approximately) independent samples.

We use two methods to estimate the tree ESS. Both methods work by converting each tree into a continuous parameter and assessing the effective sample size of the resulting set of continuous parameters (trace) [41, 42]. The first trace consists of the sum of the branch lengths for each tree. The second trace corresponds to the Robinson-Foulds distance to the CCD1-MAP tree [5]. We then use the smaller ESS of these two traces to estimate the ESS of the posterior tree samples.

Simulated datasets

The following LinguaPhylo script was used to generate the simulated datasets. Only the number of taxa (n) was changed for the different runs.

```
data {
        L = 300;
        clockRate = 1.0;
        nCatGamma = 4;
        birthRate = 25.0;
        n = 100;
}
model {
        frequencies ~ Dirichlet(conc=[5.0, 5.0, 5.0, 5.0]);
        kappa ~ LogNormal(meanlog=1.0, sdlog=1.25);
        Q = hky(kappa=kappa, freq=frequencies);
        shape ~ LogNormal(meanlog=-1.0, sdlog=0.5);
        siteRates ~ DiscretizeGamma(
                shape=shape, ncat=nCatGamma, replicates=L
        );
        phi ~ Yule(lambda=birthRate, n=n);
        D ~ PhyloCTMC(
                L=L, Q=Q, mu=clockRate,
                siteRates=siteRates, tree=phi
        );
}
```

Real-world datasets

See the Excel file in the supporting information for a list of studies considered and possible reasons for exclusion of a dataset. All datasets can be downloaded from the Dryad repository [36] and the curated *Phylogenetic Posterior Tree Datasets* repository [37] found at `https://github.com/tochsner/phylogenetic-tree-posterior-datasets`.

References

1. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS computational biology. 2014;10(4):e1003537.

2. Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, et al. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic biology. 2016;65(4):726–736.

3. Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic biology. 2012;61(3):539–542.

4. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Systematic biology. 2018;67(5):901–904.

5. Berling L, Klawitter J, Bouckaert R, Xie D, Gavryushkin A, Drummond AJ. Accurate Bayesian phylogenetic point estimation using a tree distribution parameterized by clade probabilities. PLOS Computational Biology. 2025;21(2):e1012789.

6. Altman N, Krzywinski M. The curse (s) of dimensionality. Nat Methods. 2018;15(6):399–400.

7. Bellman R. Dynamic programming. science. 1966;153(3731):34–37.

8. Semple C, Steel M, et al. Phylogenetics. vol. 24. Oxford University Press on Demand; 2003.

9. Lueg J, Garba MK, Nye TM, Huckemann SF. Wald space for phylogenetic trees. In: Geometric Science of Information: 5th International Conference, GSI 2021, Paris, France, July 21–23, 2021, Proceedings 5. Springer; 2021. p. 710–717.

10. Gavryushkin A, Drummond AJ. The space of ultrametric phylogenetic trees. Journal of theoretical biology. 2016;403:197–208.

11. Monod A, Lin B, Yoshida R, Kang Q. Tropical geometry of phylogenetic tree space: a statistical perspective. arXiv preprint arXiv:180512400. 2018;.

12. Billera LJ, Holmes SP, Vogtmann K. Geometry of the space of phylogenetic trees. Advances in Applied Mathematics. 2001;27(4):733–767.

13. Heled J, Bouckaert RR. Looking for trees in the forest: summary tree from posterior samples. BMC evolutionary biology. 2013;13:1–11.

14. Bouckaert RR, Heled J. DensiTree 2: Seeing trees through the forest. BioRxiv. 2014; p. 012401.

15. Baele G, Carvalho LM, Brusselmans M, Dudas G, Ji X, McCrone JT, et al. HIPSTR: highest independent posterior subtree reconstruction in TreeAnnotator X. bioRxiv. 2024; p. 2024–12.

16. Larget B. The estimation of tree posterior probabilities using conditional clade probability distributions. Systematic biology. 2013;62(4):501–511.

17. Gavryushkin A, Whidden C, Matsen IV FA. The combinatorics of discrete time-trees: theory and open problems. Journal of mathematical biology. 2018;76(5):1101–1121.

18. Brown DG, Owen M. Mean and variance of phylogenetic trees. Systematic biology. 2020;69(1):139–154.

19. Lammers L, Van DT, Nye TM, Huckemann SF. Types of Stickiness in BHV Phylogenetic Tree Spaces and Their Degree. In: International Conference on Geometric Science of Information. Springer; 2023. p. 357–365.

20. Said S, Mostajeran C, Heuveline S. Gaussian distributions on Riemannian symmetric spaces of nonpositive curvature. In: Handbook of Statistics. vol. 46. Elsevier; 2022. p. 357–400.

21. Lewis PO, Chen MH, Kuo L, Lewis LA, Fučíková K, Neupane S, et al. Estimating Bayesian phylogenetic information content. Systematic biology. 2016;65(6):1009–1023.

22. Zhang C, Matsen IV FA. Generalizing Tree Probability Estimation via Bayesian Networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc.; 2018.Available from: `https://proceedings.neurips.cc/paper_files/paper/2018/file/b137fdd1f79d56c7edf3365fea7520f2-Paper.pdf`.

23. Zhang C. Improved variational Bayesian phylogenetic inference with normalizing flows. Advances in neural information processing systems. 2020;33:18760–18771.

24. Zhang C, IV FAM. A Variational Approach to Bayesian Phylogenetic Inference. Journal of Machine Learning Research. 2024;25(145):1–56.

25. Berling L, Ochsner T, Bouckaert R, Drummond AJ. CCD0 revisited; 2025. Available from: https://www.beast2.org/2025/02/25/CCD0-revisited.html.

26. Paradis E. The distribution of branch lengths in phylogenetic trees. Molecular phylogenetics and evolution. 2016;94:136–145.

27. Mooers A, Gascuel O, Stadler T, Li H, Steel M. Branch lengths on birth–death trees and the expected loss of phylogenetic diversity. Systematic biology. 2012;61(2):195–203.

28. Venditti C, Meade A, Pagel M. Phylogenies reveal new interpretation of speciation and the Red Queen. Nature. 2010;463(7279):349–352.

29. Swanepoel C, Fourment M, Ji X, Nasif H, Suchard MA, Matsen IV FA, et al. TreeFlow: probabilistic programming and automatic differentiation for phylogenetics. arXiv preprint arXiv:221105220. 2022;.

30. Cramér H. On the composition of elementary errors. Skand Aktuarietids. 1928;11:13–74.

31. Cook SR, Gelman A, Rubin DB. Validation of software for Bayesian models using posterior quantiles. Journal of Computational and Graphical Statistics. 2006;15(3):675–692.

32. Talts S, Betancourt M, Simpson D, Vehtari A, Gelman A. Validating Bayesian inference algorithms with simulation-based calibration. arXiv preprint arXiv:180406788. 2018;.

33. Mendes FK, Bouckaert R, Carvalho LM, Drummond AJ. How to validate a Bayesian evolutionary model. Systematic Biology. 2025;74(1):158–175.

34. Rambaut A, Drummond A. TreeAnnotator v1. 7.0. Available as Part of the BEAST package; 2013.

35. Drummond AJ, Chen K, Mendes FK, Xie D. LinguaPhylo: a probabilistic model specification language for reproducible phylogenetic analyses. PLOS Computational Biology. 2023;19(7):e1011226.

36. Dryad;. http://datadryad.org/.

37. Ochsner T. Phylogenetic Posterior Tree Datasets; 2025. Available from: https://github.com/tochsner/phylogenetic-tree-posterior-datasets/tree/main.

38. Ochsner T. Tractable Time Tree Distributions; 2025. Available from: https://github.com/tochsner/TractableTimeTreeDistributions.

39. Smith MR. Robust analysis of phylogenetic tree space. Systematic Biology. 2022;71(5):1255–1270.

40. Venkat N. The curse of dimensionality: Inside out. Pilani (IN): Birla Institute of Technology and Science, Pilani, Department of Computer Science and Information Systems. 2018;10(10.13140).

41. Lanfear R, Hua X, Warren DL. Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses. Genome biology and evolution. 2016;8(8):2319–2332.

42. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC. Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC (with discussion). Bayesian analysis. 2021;16(2):667–718.