# Bayesian Least-Squares Supertrees (BLeSS): a flexible method for inferring large time-calibrated phylogenies

**David Černý** and **Graham J. Slater**
Department of the Geophysical Sciences, University of Chicago
david.cerny1@gmail.com

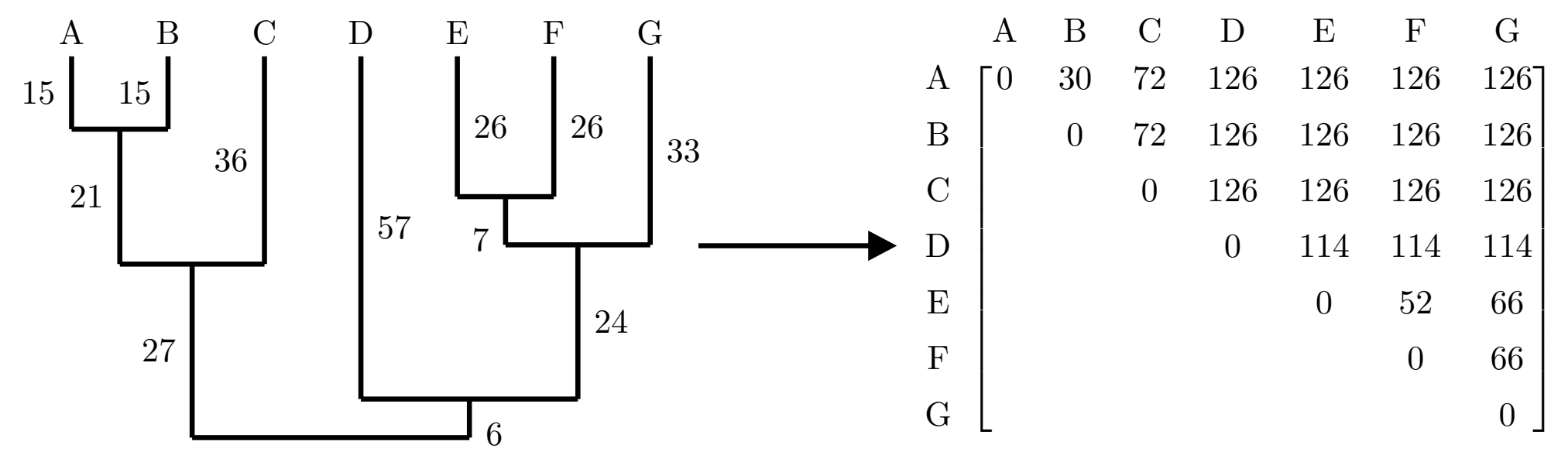
## Background

Time-calibrated phylogenies are essential for a variety of downstream inferences, but their estimation is difficult when the number of tips is large. Existing approaches either do not scale well (joint Bayesian inference of topology and divergence times), do little to accommodate uncertainty (*post hoc* time-scaling of maximum-likelihood phylograms), or rely on statistically invalid approximations and potentially problematic monophyly assumptions (the "backbone-and-patch" method). In the past, supertrees were often regarded as an attractive alternative to these frameworks. However, most supertree techniques are poorly suited to providing input for comparative methods, since they usually yield topological point estimates without branch lengths or any measure of uncertainty. Here, we present a new approach to supertree inference, the Bayesian Least-Squares Supertrees (BLeSS), which overcomes these obstacles.

## Method description

BLeSS combines elements of two previous supertree methods: the average distance matrix (Lapointe & Levasseur 2004) and the exponential error (Steel & Rodrigo 2008) frameworks.

Consider a profile of ultrametric time trees $\mathcal{P} = \{\mathcal{T}_1, \ldots, \mathcal{T}_K\}$ defined over a set of leaves $\mathcal{L}$, where $\mathcal{L}(\mathcal{T}_k)$ is the leaf set of tree $\mathcal{T}_k$ such that $\mathcal{L}(\mathcal{T}_k) \subseteq \mathcal{L}$. First, each source tree in $\mathcal{P}$ is represented as a path-length distance matrix:



$$
\begin{array}{c|ccccccc}
 & A & B & C & D & E & F & G \\
\hline
A & 0 & 30 & 72 & 126 & 126 & 126 & 126 \\
B & & 0 & 72 & 126 & 126 & 126 & 126 \\
C & & & 0 & 126 & 126 & 126 & 126 \\
D & & & & 0 & 114 & 114 & 114 \\
E & & & & & 0 & 52 & 66 \\
F & & & & & & 0 & 66 \\
G & & & & & & & 0
\end{array}
$$

If we denote by $\pi_{i,j}(\mathcal{P})$ the (possibly empty) set of indices of all trees in $\mathcal{P}$ that contain both $i$ and $j$, that is, $\pi_{i,j}(\mathcal{P}) = \{k \in \{1, \ldots, K\} \ : \ i \in \mathcal{L}(\mathcal{T}_k) \wedge j \in \mathcal{L}(\mathcal{T}_k)\}$, the *average distance matrix* $\bar{\mathbf{D}}$ can be calculated element-wise as follows:

$$\bar{d}(i,j) = \begin{cases} \dfrac{1}{|\pi_{i,j}(\mathcal{P})|} \displaystyle\sum_{k \in \pi_{i,j}(\mathcal{P})} d_k(i,j) & \text{if } \pi_{i,j}(\mathcal{P}) \neq \varnothing \\ \text{undefined} & \text{if } \pi_{i,j}(\mathcal{P}) = \varnothing \end{cases} \tag{1}$$

The undefined entries in $\bar{\mathbf{D}}$, corresponding to missing distances, can either be treated as such and disregarded in subsequent steps, or filled in using a variety of imputation schemes.

Once $\bar{\mathbf{D}}$ has been calculated, the goal is to find a supertree $\hat{\mathcal{T}}$ with the full leaf set $\mathcal{L}$ that minimizes the least-squares difference between its own path-length matrix $\hat{\mathbf{D}}$ and $\bar{\mathbf{D}}$:

$$\delta(\hat{\mathbf{D}}, \bar{\mathbf{D}}) = \sum_{i=1}^{|\mathcal{L}|} \sum_{j=1}^{|\mathcal{L}|} \left[\hat{d}(i,j) - \bar{d}(i,j)\right]^2 \tag{2}$$

To this end, we use a modified version of the exponential error model introduced by Steel & Rodrigo (2008) and parameterized by a single penalty term $\lambda_e$:

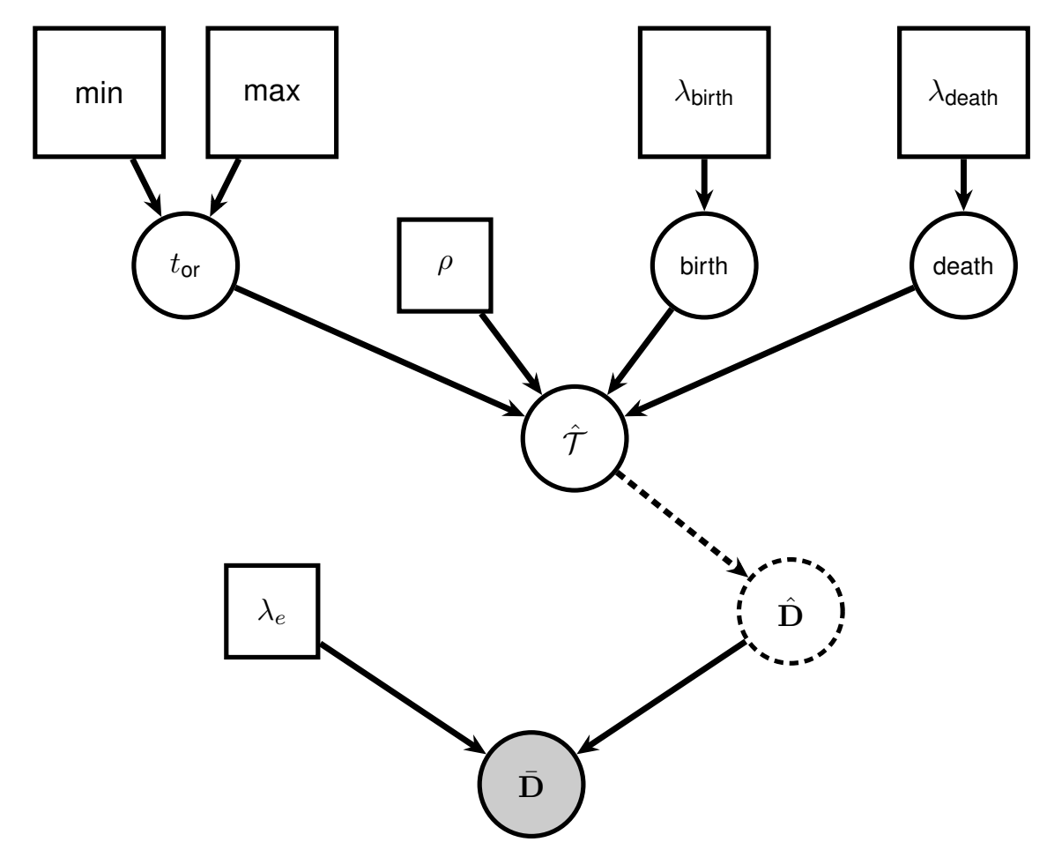
$$\begin{aligned} f(\bar{\mathbf{D}} \mid \hat{\mathcal{T}}, \lambda_e) &= \lambda_e e^{-\lambda_e \delta(\hat{\mathbf{D}}, \bar{\mathbf{D}})} \\ &= \lambda_e \exp\left\{-\lambda_e \sum_{i=1}^{|\mathcal{L}|} \sum_{j=1}^{|\mathcal{L}|} \left[\hat{d}(i,j) - \bar{d}(i,j)\right]^2\right\} \end{aligned} \tag{3}$$

This likelihood function can be combined with an arbitrary tree prior, monophyly constraints, or node calibrations. If the relevant hyperparameters are jointly denoted as $\Phi$ and estimated along with the supertree, the joint posterior distribution of $(\mathcal{T}, \Phi)$ is given by:

$$P(\mathcal{T}, \Phi \mid \bar{\mathbf{D}}, \lambda_e) = \frac{f(\bar{\mathbf{D}} \mid \lambda_e, \mathcal{T})\, P(\mathcal{T} \mid \Phi)\, P(\Phi)}{\displaystyle\int_{\mathcal{T}} f(\bar{\mathbf{D}} \mid \lambda_e, \mathcal{T}) \left[\int_{\Phi} P(\mathcal{T} \mid \Phi)\, P(\Phi)\, \mathrm{d}\Phi\right] \mathrm{d}\mathcal{T}} \tag{4}$$

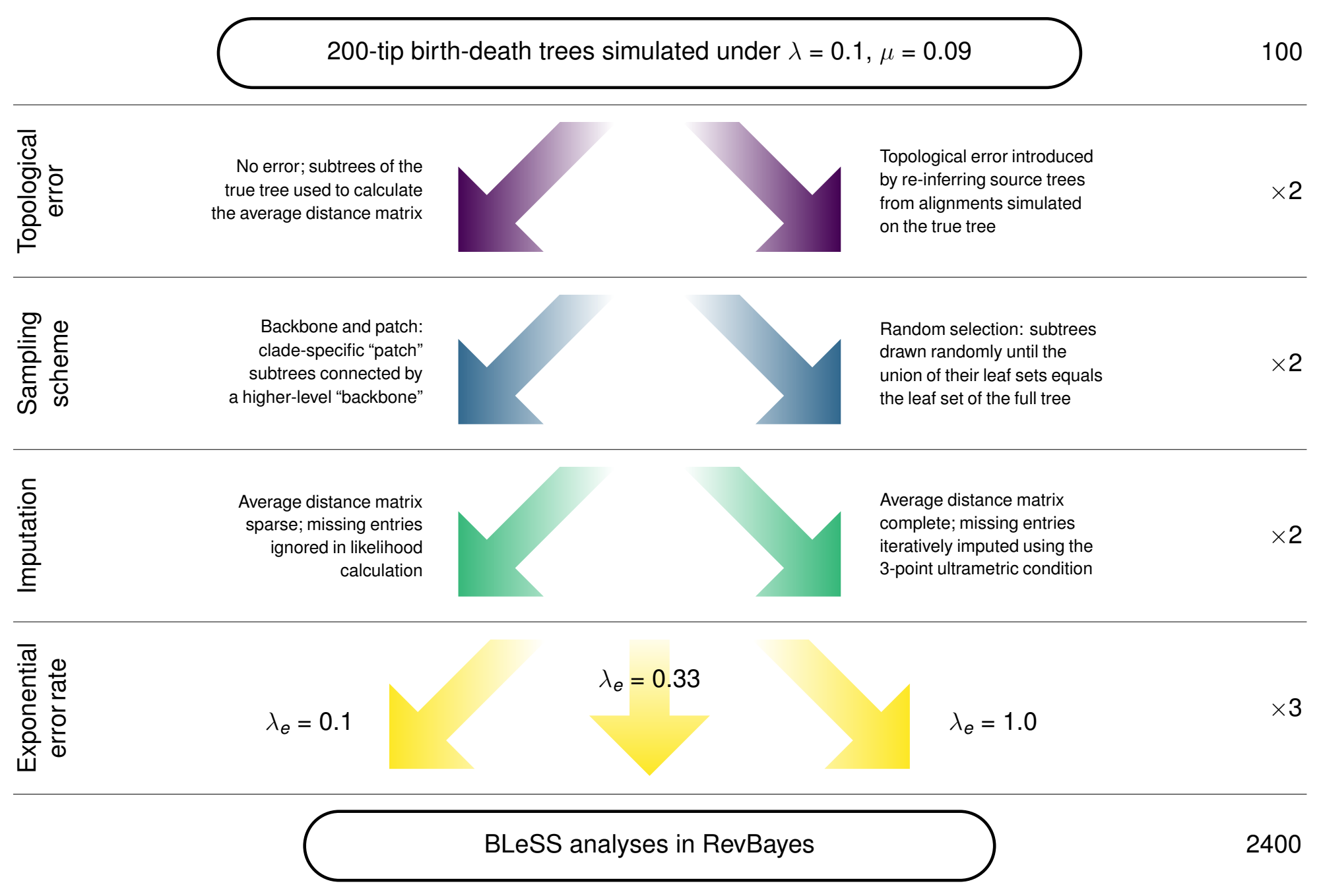
## Implementation

BLeSS is implemented in `RevBayes`, an open source C++ program built around probabilistic graphical models (Höhna et al. 2016). A realistic case of BLeSS inference can be represented as follows:



## Evaluation: simulations

To evaluate BLeSS in a setting approaching a plausible supertree analysis in terms of tree size, we used a hierarchical simulation scheme involving trees of 200 tips:



200-tip birth-death trees simulated under $\lambda = 0.1$, $\mu = 0.09$ — 100

Topological error: No error; subtrees of the true tree used to calculate the average distance matrix / Topological error introduced by re-inferring source trees from alignments simulated on the true tree — ×2

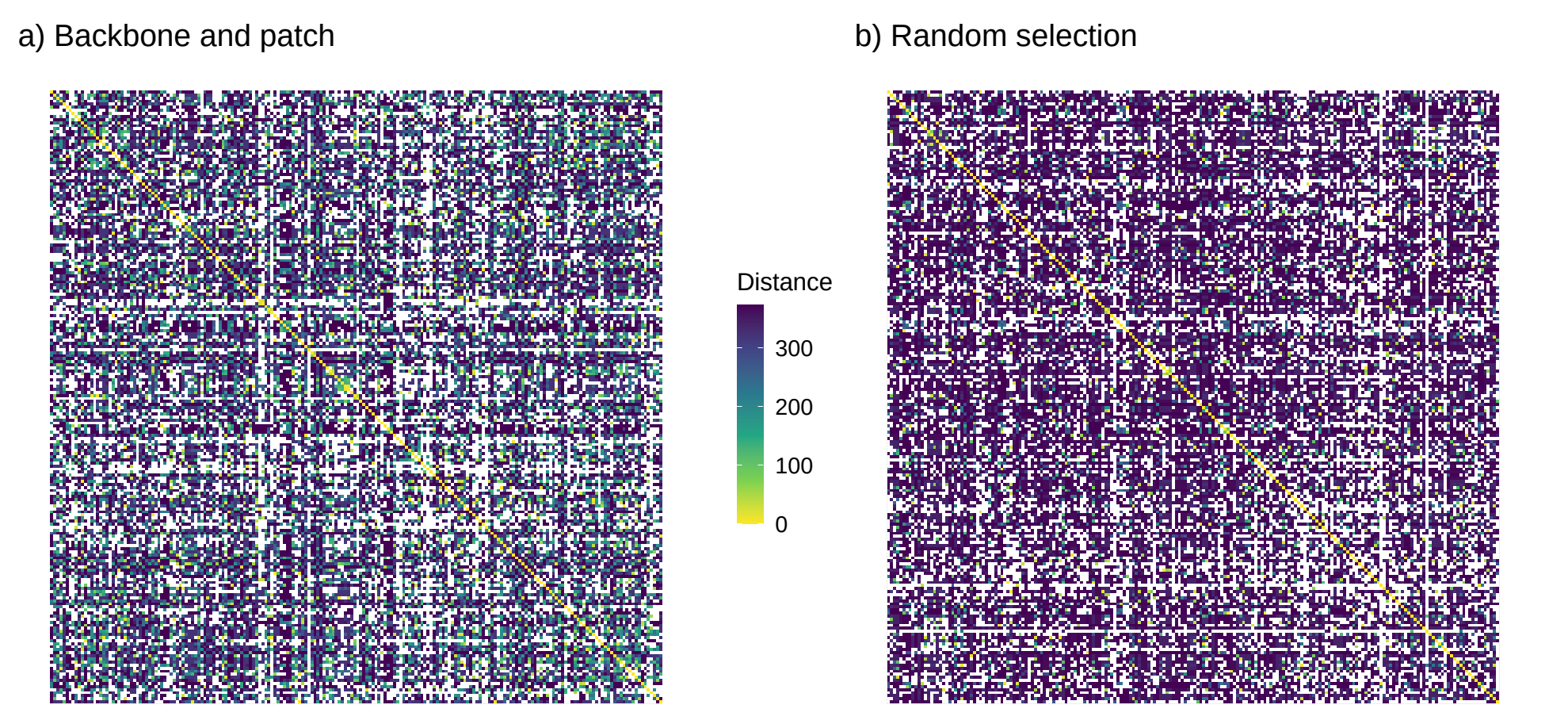
Sampling scheme: Backbone and patch: clade-specific "patch" subtrees connected by a higher-level "backbone" / Random selection: subtrees drawn randomly until the union of their leaf sets equals the leaf set of the full tree — ×2

Imputation: Average distance matrix sparse; missing entries ignored in likelihood calculation / Average distance matrix complete; missing entries iteratively imputed using the 3-point ultrametric condition — ×2

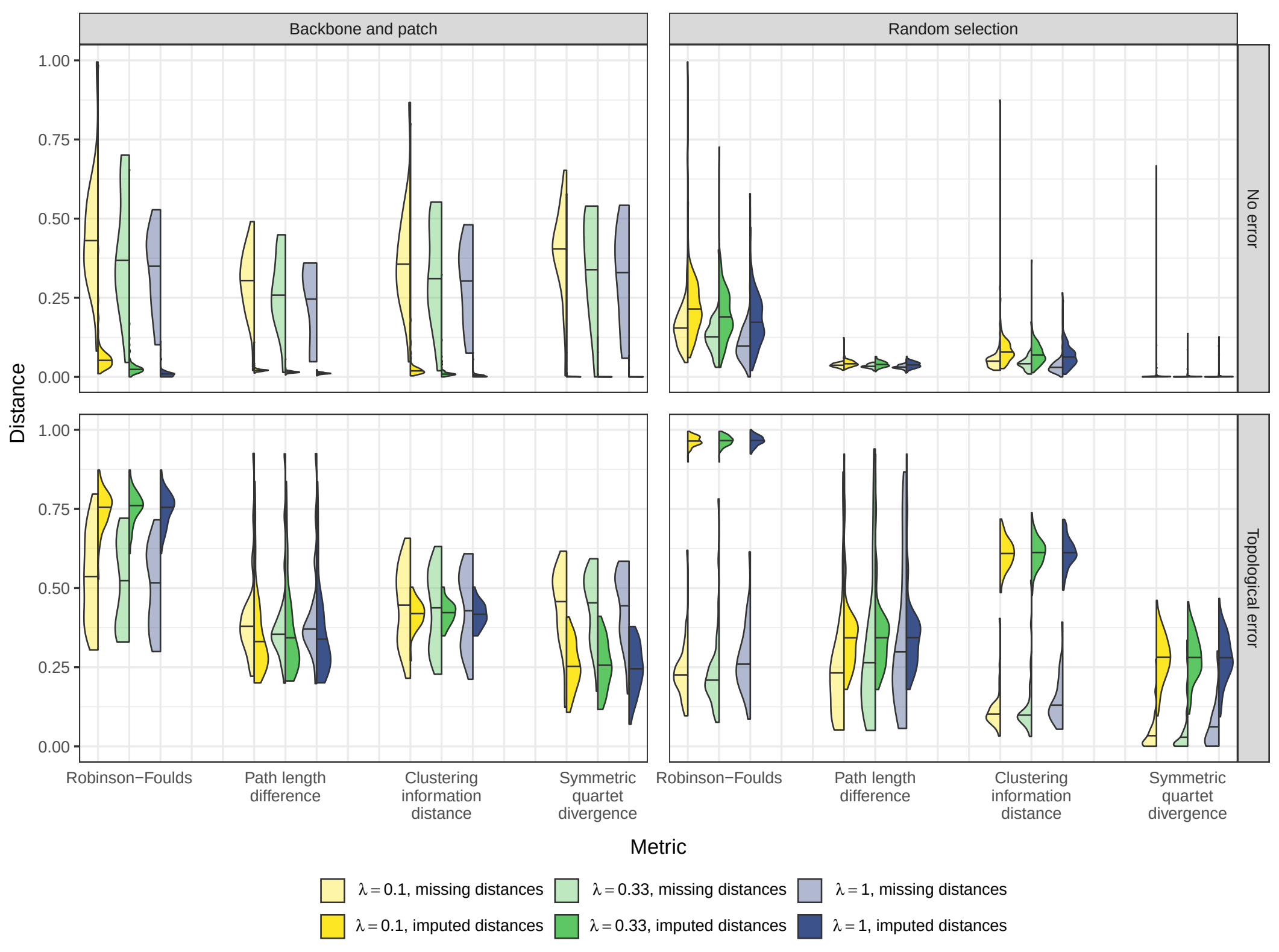
Exponential error rate: $\lambda_e = 0.1$ / $\lambda_e = 0.33$ / $\lambda_e = 1.0$ — ×3

BLeSS analyses in RevBayes — 2400

The backbone-and-patch vs. random selection schemes were used to evaluate the sensitivity of the method to the distribution of missing entries in the average distance matrix:



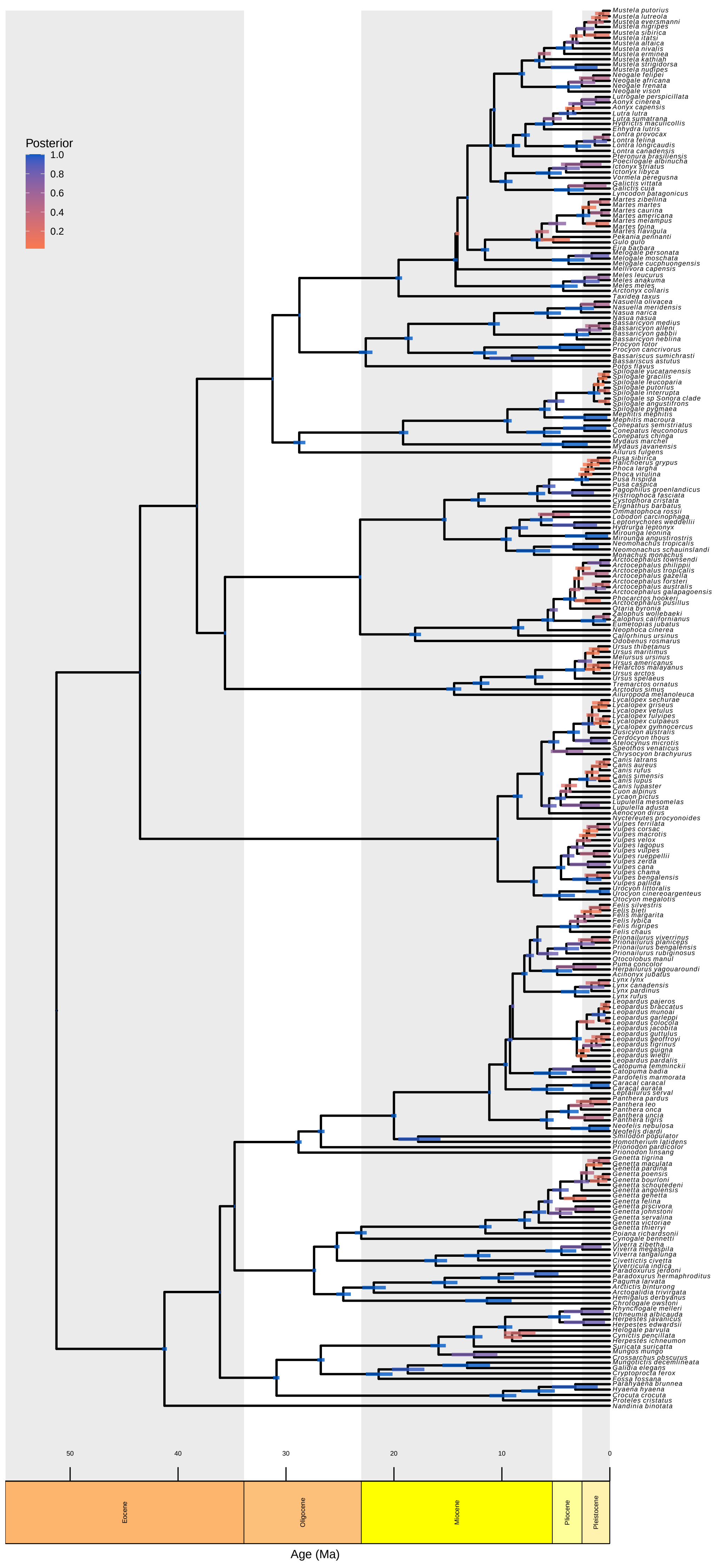
The similarity in topology and/or branch lengths between each resulting MCC tree and the corresponding generating tree was assessed using a variety of metrics:



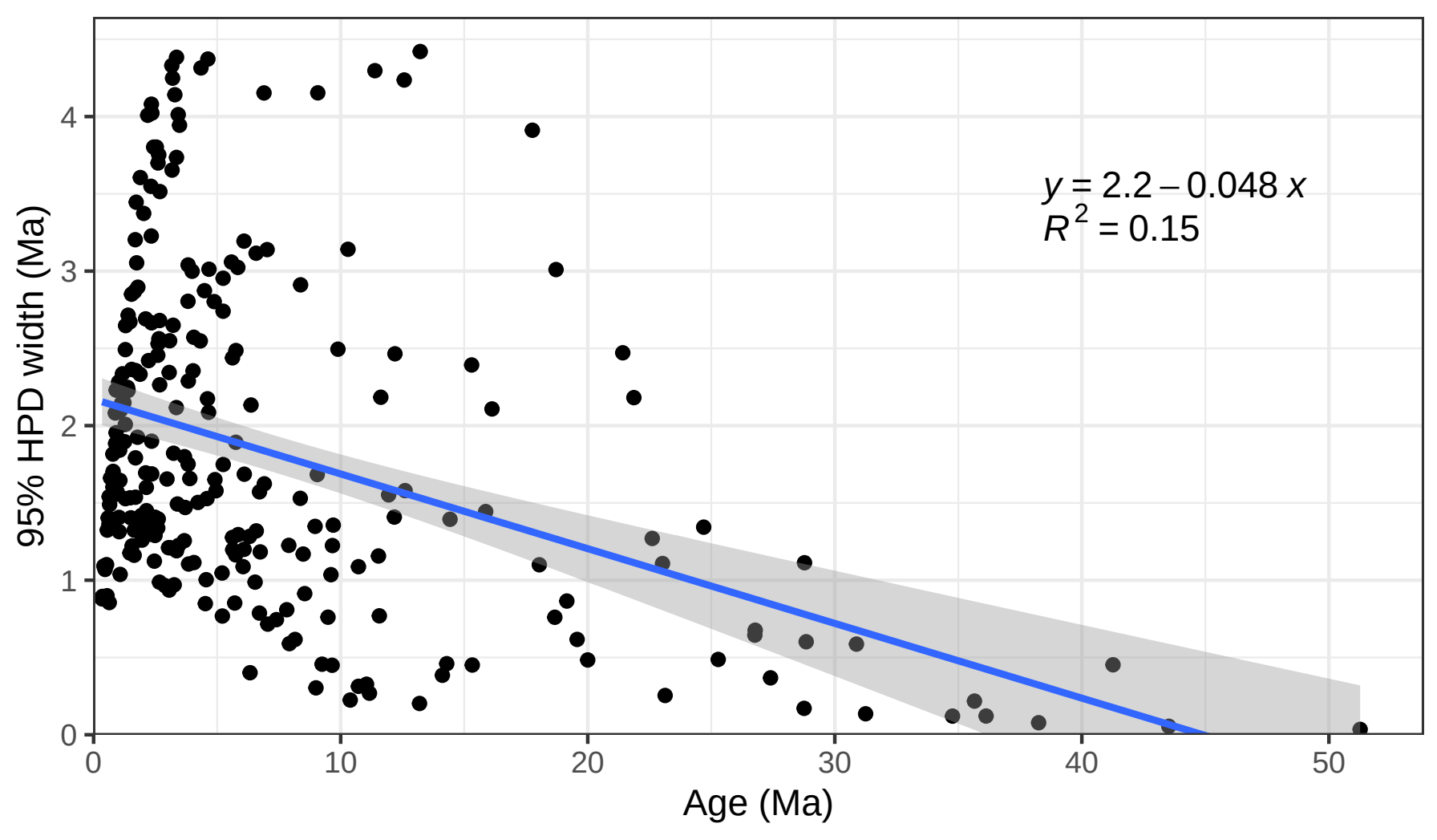
Imputation outperforms treating missing data as such when source trees are free of topological error, especially under the backbone-and-patch sampling scheme. The relationship is reversed when topological error is present; detailed examination confirms that overall topological error is driven by imputation rather than misestimated source trees.

## Evaluation: empirical data

We further used BLeSS to estimate a time-calibrated supertree for the order Carnivora based on previously published time trees inferred from nuclear ($n$ = 13) or mitochondrial ($n$ = 20) data. Mitogenome-based source trees were downweighted by a factor of 10 when constructing the average distance matrix.



The MCC tree above contained 45 out of 53 benchmark clades defined prior to the analysis, indicating good topological performance. Node age precision increased toward the root, as deeper divergences were informed by a greater number of distances.



## Acknowledgments & References

D.Č would like to thank Tracy Heath, Will Pett, and Joëlle Barido-Sottani for their invaluable guidance on implementing BLeSS in `RevBayes`, Michael Landis for assistance with simulation design, and Luke Kelly for his help with diagnosing MCMC convergence. All simulations were performed on the Midway3 Research Computing Cluster at the University of Chicago under proposals #39336 and #40114..

- Höhna et al. (2016): http://doi.org/10.1093/sysbio/syw021
- Lapointe & Levasseur (2004): http://doi.org/10.1007/978-1-4020-2330-9_5
- Steel & Rodrigo (2008): http://doi.org/10.1080/10635150802033014