PhyloJunction: a computational framework for simulating, developing, and teaching evolutionary models

Supplementary Material

Fábio K. Mendes[1,*] and Michael J. Landis[1]

[1] *Department of Biology, Washington University in St. Louis, St. Louis, MO*

[*] *Corresponding author: E-mail: f.mendes@wustl.edu*

1 Comparisons with other software

PhyloJunction's (PJ) simulation code was compared to independently implemented counterparts whenever possible, which in some cases included multiple software packages. The latter included packages written in R, such as geiger [6], diversitree [2], phytools [7], TreeSim [9], and FossilSim [1], as well as in Java, such as MASTER [11]. As mentioned in the main text, each of those tools is unique in its conditioning of diversification models and filtering of simulated output.

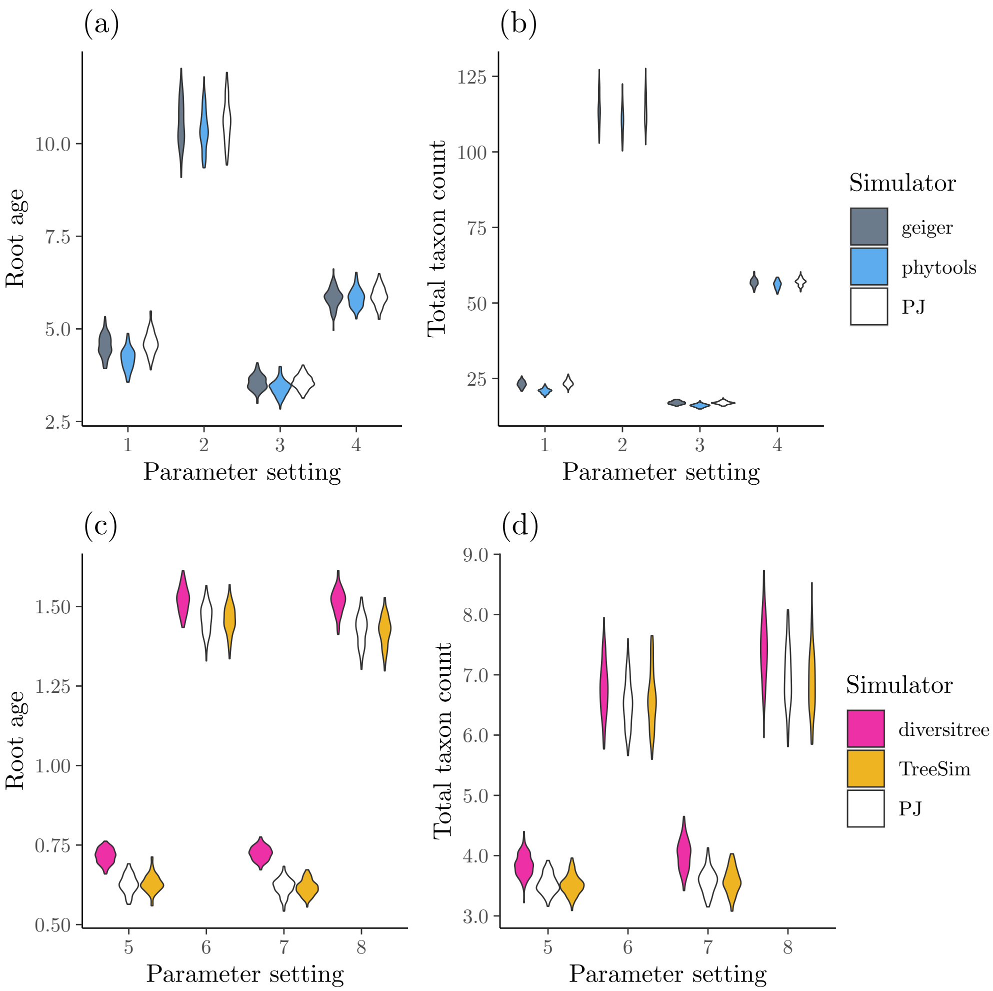
Comparisons under different models (Supplementary Figs. 1, 2, 3, 4, 5, and 6) were carried out in multiple arbitrary regions of parameter space (Supplementary Tables 1 to 4). Deciding which programs to compare in each scenario was largely determined by our perceived ability to match the model assumptions and output parsing of different simulators.

2 Machine learning example with PhyloJunction

We used phyddle [4] to train neural networks for phylogenetic parameter estimation using PhyloJunction as a simulator. phyddle is a Python package to design, manage, and deploy deep learning pipelines for phylogenetic modeling tasks.

Phylogenetic trees were simulated under a piecewise constant time-heterogeneous BiSSE model (Supplementary Table 5). The BiSSE process was set to start at $t = 0.0$ and terminate at $t = 10.0$, with the lineage

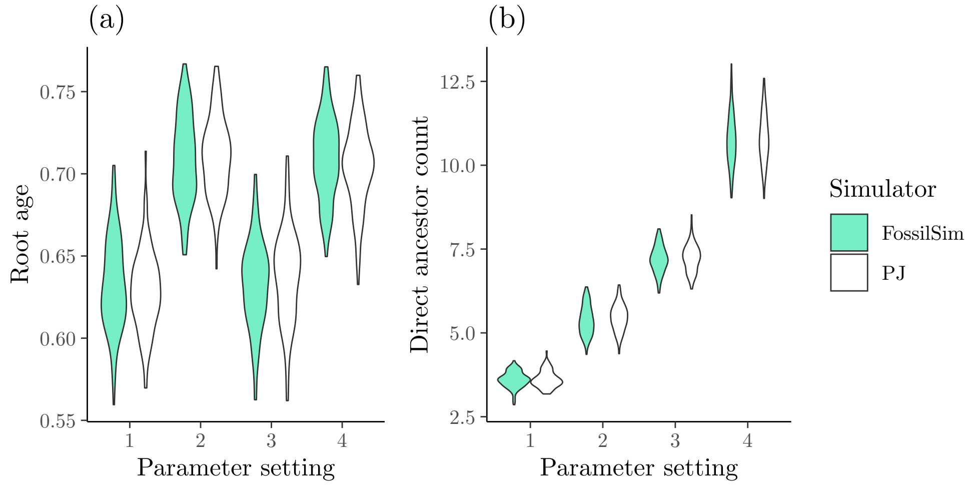
[1] PJ implements the "simple sampling approach" (SSA; see Stadler, 2011)



Supplementary Figure 1: Summaries of data sets simulated under the birth-death model with `PhyloJunction` (PJ), and the `geiger`, `phytools` and `diversitree` R packages. Quantities are summarized from complete (i.e., including extinct taxa) trees, assuming perfect sampling. Summaries include (a, c) root ages, and (b, d) total taxon count. Each violin plot comprises 100 values, each corresponding to the focal statistic (e.g, root age) averaged over 100 trees. Parameter settings are detailed in Supplementary Table 1.

birth rate in state 0 differing before and after time $t = 8.0$. Only extant taxa were retained for training. Datasets with fewer than 10 or more than 500 taxa were rejected and re-simulated.

We used `phyddle` to run a standard phylogenetic deep learning pipeline. Our pipeline used a small Python script and `PhyloJunction` to simulate 100,000 datasets and then converted them into tensor-format using a compact phylogenetic-state vector representation. This representation uses a compact diversity-based ladderization for extant-only phylogenies and an expansion of character state rows [12, 3, 10]. A subset of examples were next used to train a neural network to estimate λ_0 in epochs 1 ($0 \leq t < 8$)) and 2 ($t \geq 8$) using a mean-squared error loss function, as well their corresponding conformalized prediction intervals using a pinball loss function [8]. `phyddle` used its default settings with PyTorch [5] to design the network layers,



Supplementary Figure 2: Summaries of data sets simulated under the fossilized birth-death (FBD) model with `PhyloJunction` (PJ) and the `FossilSim` R package. Quantities are summarized from complete (i.e., including extinct taxa) trees, assuming perfect sampling. Summaries include (a) root ages, and (b) direct ancestor taxon count. Each violin plot comprises 100 values, each corresponding to the focal statistic (e.g, root age) averaged over 100 trees. Parameter settings are detailed in Supplementary Table 1.

apply activation functions, and train the network. Training was terminated when the summed loss score of a separate validation dataset increased across three consecutive training epochs, i.e., to prevent overfitting. We then used the trained network to estimate model parameters and prediction intervals for a batch of new 250 simulated data points that were withheld from the training procedure.

Neural networks trained using `phyddle` with `PhyloJunction` as a simulator accurately estimated the two targeted birth rate parameter values (Supplementary Fig 7).

Supplementary Table 1: Model configurations used in PhyloJunction validation. Dots denote parameters that do not apply to a given model. "BD" stands for birth-death, "FBD" for fossilized birth-death and "BiSSE" for binary state-dependent speciation and extinction models.

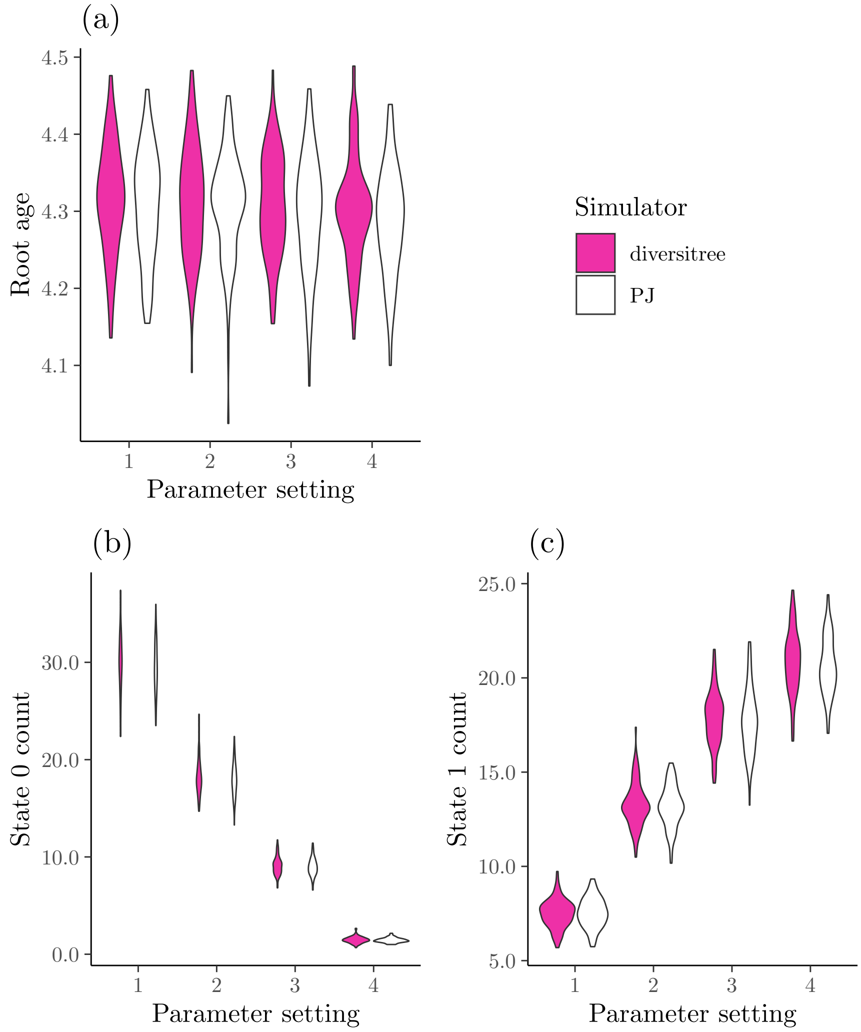
Tree model	Parameter setting	λ or λ_0, λ_1	μ	ψ	q_{01}, q_{10}	Max. age	Max. extant taxon count[1]	Start from
BD	1	1.0	0.8	·	·	·	10	Root
	2	1.0	0.8	·	·	·	30	Root
	3	1.0	0.5	·	·	·	10	Root
	4	1.0	0.5	·	·	·	30	Root
	5	1.0	0.8	·	·	1.0	·	Origin
	6	1.0	0.8	·	·	2.0	·	Origin
	7	1.0	0.5	·	·	1.0	·	Origin
	8	1.0	0.5	·	·	2.0	·	Origin
FBD	1	1.0	1.0	2.0	·	1.0	·	Origin
	2	2.0	1.0	2.0	·	1.0	·	Origin
	3	1.0	1.0	4.0	·	1.0	·	Origin
	4	2.0	1.0	4.0	·	1.0	·	Origin
BiSSE	1	1.0, 0.75	0.5	·	0.25, 0.75	5.0	·	Origin
	2	1.0, 0.75	0.5	·	0.5, 0.5	5.0	·	Origin
	3	1.0, 0.75	0.5	·	0.75, 0.25	5.0	·	Origin
	4	1.0, 0.75	0.5	·	1.0, 0.0	5.0	·	Origin

Supplementary Table 2: Geographic state-dependent speciation and extinction (GeoSSE) model configurations used in PhyloJunction validation. Processes started at the origin, stopped at a maximum age of 4.0, and were conditioned on the survival of at least one living taxon. (parameter names within parentheses follow 'diversitree"s notation).

Tree model	Parameter setting	λ_1 (sA)	λ_2 (sB)	$\lambda_{0,1,2}$ (sAB)	μ_1 (xA)	μ_2 (xB)	$q_{1,0}$ (dA)	$q_{2,0}$ (dB)
GeoSSE	1	1.25	1.25	0.75	1.0	1.0	1.0	1.0
	2	1.25	1.25	1.0	1.0	1.0	1.0	1.0
	3	1.25	1.25	1.25	1.0	1.0	1.0	1.0
	4	1.25	1.25	1.5	1.0	1.0	1.0	1.0

Supplementary Table 3: Time-heterogeneous Yule model configurations used in PhyloJunction validation. All model configurations specified processes starting at the origin, and stopping at a maximum age of 6.0. All epoch starting times t are defined in forward time units, with t_0, t_1 and t_2 corresponding to the starts of the first, second and third epochs, respectively. Each epoch, from oldest to youngest, was specified its own birth-rate, λ^0, λ^1, and λ^2.

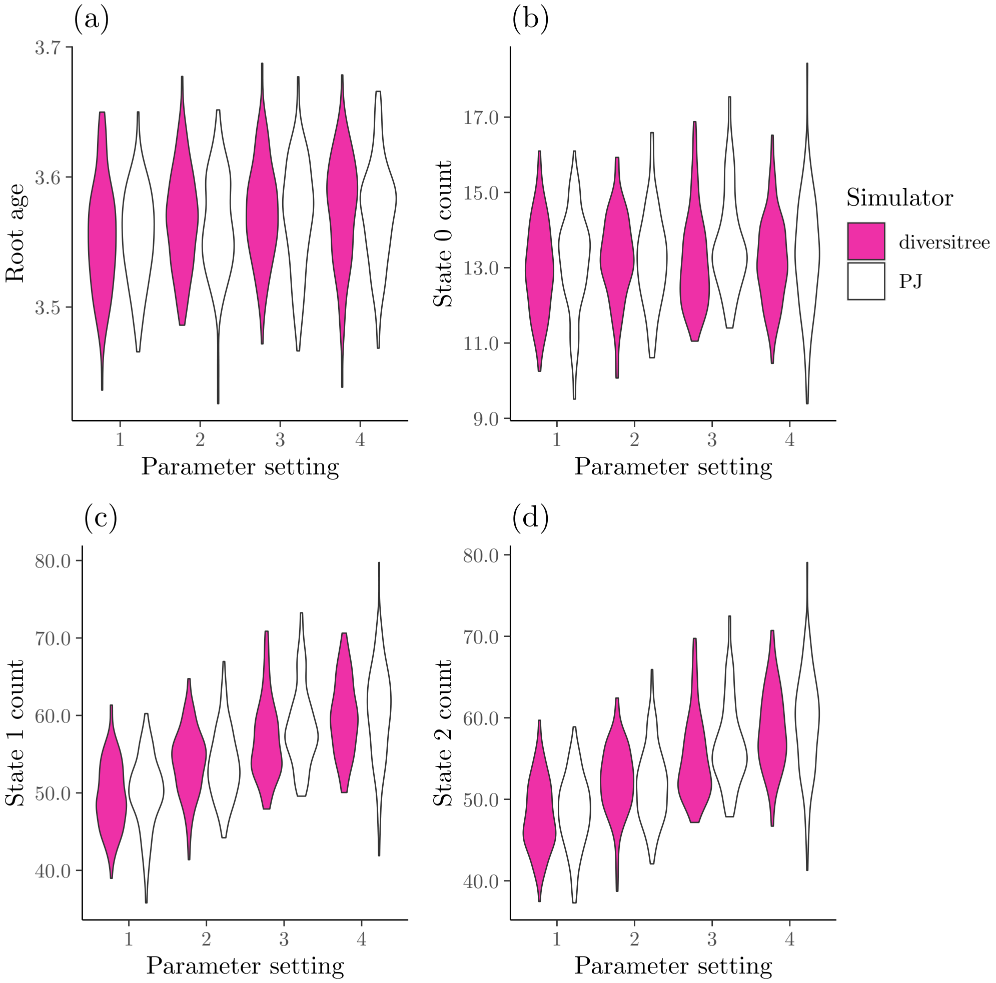
Tree model	Parameter setting	t_0	t_1	t_2	λ^0	λ^1	λ^2
Yule	1	0.0	1.0	2.0	2.0	0.5	0.1
	2	0.0	1.0	3.0	2.0	0.5	0.1
	3	0.0	1.0	4.0	2.0	0.5	0.1
	4	0.0	1.0	5.0	2.0	0.5	0.1



Supplementary Figure 3: Summaries of data sets simulated under the binary state-dependent speciation and extinction (BiSSE) model with `PhyloJunction` (PJ) and the `diversitree` R package. Quantities are summarized from complete (i.e., including extinct taxa) trees, assuming perfect sampling. Summaries include (a) number of taxa at state 0, (b) number of taxa at state 1, and (c) root ages. Each violin plot comprises 100 values, each corresponding to the focal statistic (e.g, root age) averaged over 100 trees. Parameter settings are detailed in Supplementary Table 1.

Supplementary Table 4: Time-heterogeneous binary state-dependent speciation and extinction (BiSSE) model configurations used in PhyloJunction validation. All model configurations specified processes starting at the origin, and stopping at a maximum age of 6.0. All epoch starting times t are defined in forward time units, with t_0 and t_1 corresponding to the starts of the first and second epochs, respectively. Birth-rates and transition rates were kept contant across epochs. Each epoch, from oldest to youngest, was specified its own death-rate for state 1, μ_1^1.

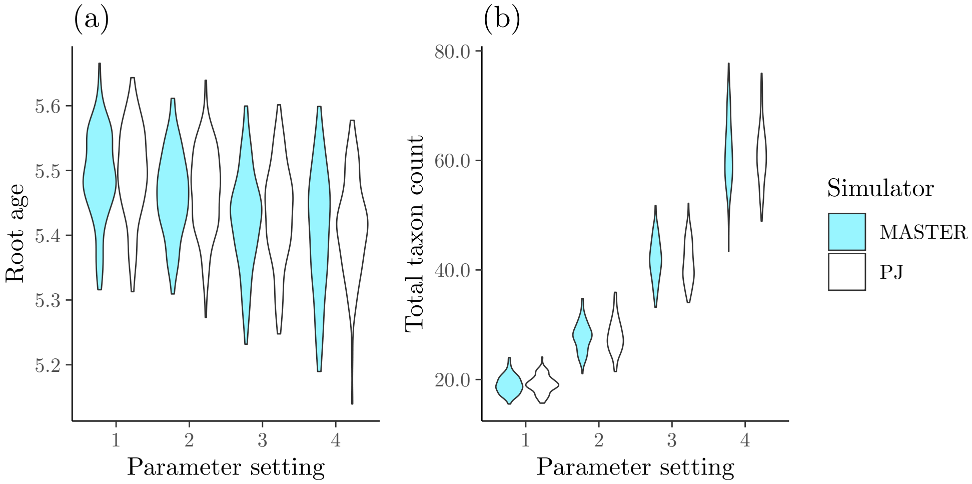
Tree model	Parameter setting	t_0	t_1	λ_0^0	λ_0^0	μ_0^0	μ_1^0	μ_0^1	μ_1^1	$q_{0,1}^0$	$q_{0,1}^0$	$q_{0,1}^1$	$q_{0,1}^1$
BiSSE	1	0.0	2.0	1.1	1.2	1.0	0.0	1.0	0.85	0.0	0.0	1.0	0.5
	2	0.0	3.0	1.1	1.2	1.0	0.0	1.0	0.9	0.0	0.0	1.0	0.5
	3	0.0	4.0	1.1	1.2	1.0	0.0	1.0	0.95	0.0	0.0	1.0	0.5
	4	0.0	5.0	1.1	1.2	1.0	0.0	1.0	1.0	0.0	0.0	1.0	0.5



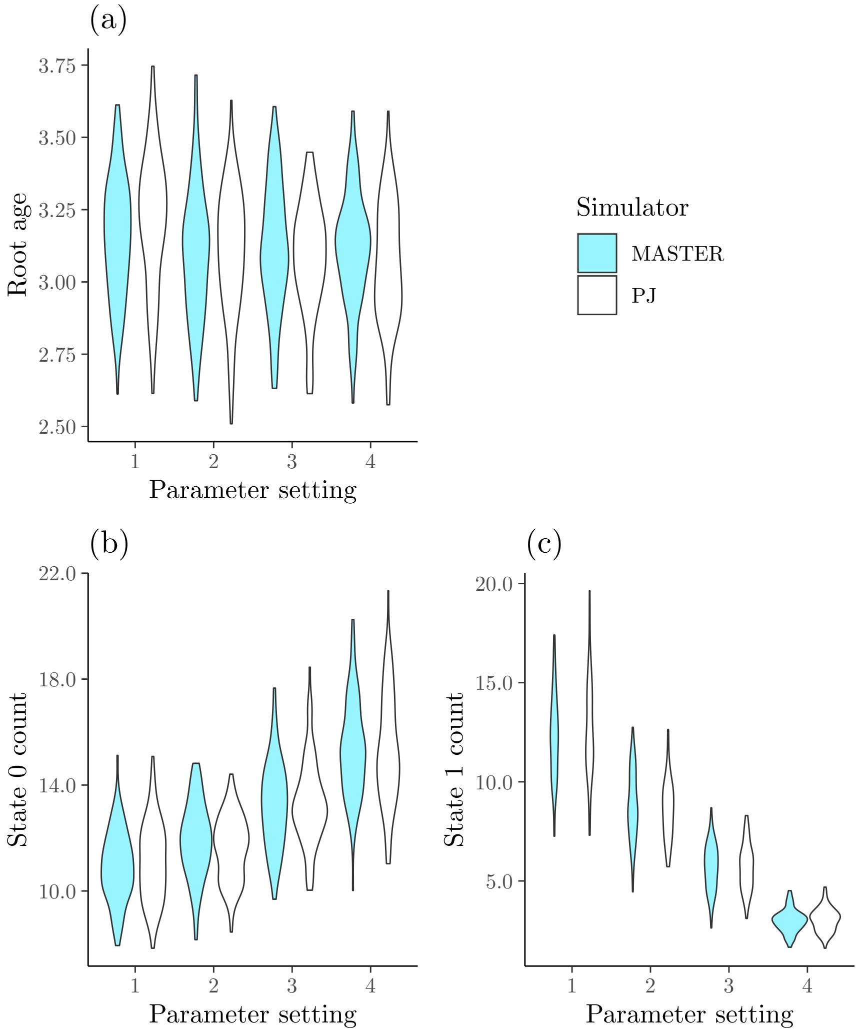
Supplementary Figure 4: Summaries of data sets simulated under the geographic state-dependent speciation and extinction (GeoSSE) model with `PhyloJunction` (PJ) and the `diversitree` R package. Quantities are summarized from complete (i.e., including extinct taxa) trees, assuming perfect sampling. Summaries include (a) root ages, (b) number of taxa at state 0, (c) number of taxa at state 1, and (d) number of taxa at state 2. Each violin plot comprises 100 values, each corresponding to the focal statistic (e.g, root age) averaged over 100 trees. Parameter settings are detailed in Supplementary Table 2.

Supplementary Table 5: Time-heterogeneous BiSSE model used in training data simulations. Birth rates for state 1 (λ_1), death rates (μ's) and state transition rates (q's) are the same across epochs.

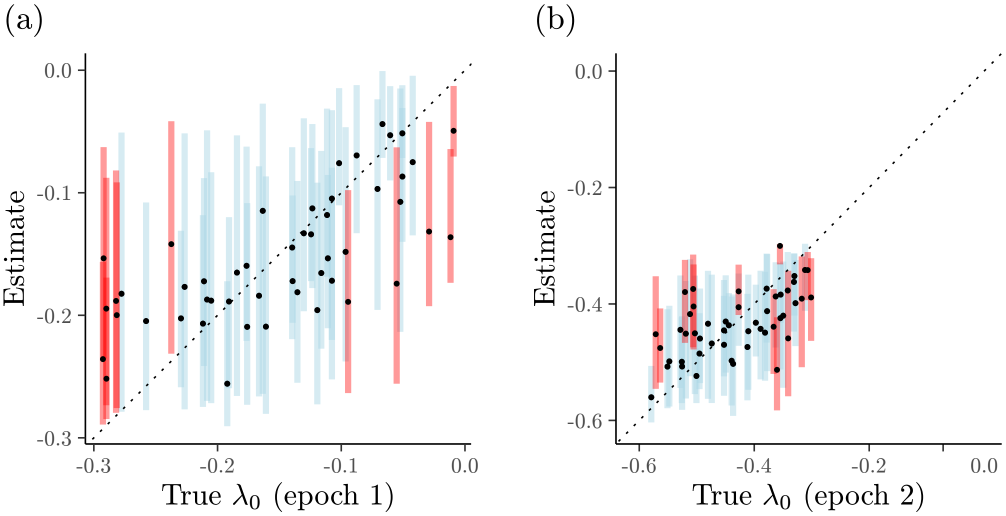
Parameter	Prior or value
λ_0 (epoch 1)	Uniform(0.5, 1.0)
λ_0 (epoch 2)	Uniform(1.0, 2.0)
λ_1	0.5
$\mu_0 = \mu_1$	0.1
$q_{01} = q_{10}$	0.2



Supplementary Figure 5: Summaries of data sets simulated under a time-heterogeneous Yule model with `PhyloJunction` (PJ) and the `MASTER` BEAST 2 package. Quantities are summarized from perfectly sampled trees. Summaries include (a) root ages, and (b) total taxon count. Each violin plot comprises 100 values, each corresponding to the focal statistic (e.g, root age) averaged over 100 trees. Parameter settings are detailed in Supplementary Table 3.



Supplementary Figure 6: Summaries of data sets simulated under a time-heterogeneous binary state-dependent speciation and extinction (BiSSE) model with `PhyloJunction` (PJ) and the `MASTER` BEAST 2 package. Quantities are summarized from complete, perfectly sampled trees. Summaries include (a) root ages, (b) number of taxa at state 0, (c) number of taxa at state 1. Each violin plot comprises 100 values, each corresponding to the focal statistic (e.g, root age) averaged over 100 trees. Parameter settings are detailed in Supplementary Table 4.



Supplementary Figure 7: Neural network-based parameter estimates for a time-heterogeneous BiSSE model. Results were produced using `phyddle` with `PhyloJunction` as the training dataset simulator. For the sake of visual clarity, with show only the first 50 (out of 250) parameter estimates. Training examples were simulated under a BiSSE process (starting at time $t = 0.0$ and ending at time $t = 10.0$) where lineage birth rates in state 0 (λ_0) differ between time interval $0.0 \leq t < 8.0$ (epoch 1) and $t \geq 8.0$ (epoch 2). All remaining model rates are constant within each time interval (details in text). Estimated birth rates (y-axis) predict true simulated birth rates (x-axis) for a test dataset that was not used to train the network. The diagonal dashed line is the identity line and represents perfect parameter estimation. Conformalized predictive intervals (bars) were trained to contain the true simulated value with frequency 0.8. Individual intervals are colored in blue if they contain the true value (i.e., if they cross the diagonal dashed line), and in red otherwise.

References

[1] Joëlle Barido-Sottani, Walker Pett, Joseph E. O'Reilly, and Rachel C. M. Warnock. FossilSim: an R package for simulating fossil occurrence data under mechanistic models of preservation and recovery. *Methods Ecol. Evol.*, 10:835–840, 2019.

[2] Richard G. Fitzjohn. Diversitree: comparative phylogenetic analyses of diversification in R. *Methods Ecol. Evol.*, 3:1084–1092, 2012.

[3] Sophia Lambert, Jakub Voznica, and Hélène Morlon. Deep learning from phylogenies for diversification analyses. *Syst. Biol.*, 72:1262–1279, 2023.

[4] Michael J. Landis and Ammon Thompson. phyddle: software to fit phylogenetic models with deep learning. `https://github.com/mlandis/phyddle`, 2024.

[5] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *arXiv*, abs/1912.01703, 2019.

[6] Matthew W. Pennell, Jonathan M. Eastman, Graham J. Slater, Jeremy W. Brown, J. C. Uyeda, R. G. Fitzjohn, Michael E. Alfaro, and Luke J. Harmon. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, 30:2216–2218, 2014.

[7] Liam J. Revell. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.*, 3:217–223, 2012.

[8] Y Romano, E Patterson, and E J Candès. Conformalized quantile regression. *arXiv*, 2019.

[9] Tanja Stadler. TreeSim. Available from `http://cran.r-project.org/web/packages/TreeSim/index.html`. [Internet]: 2010.

[10] Ammon Thompson, Benjamin Liebeskind, Erik J. Scully, and Michael Landis. Deep learning and likelihood approaches for viral phylogeography converge on the same answers whether the inference model is right or wrong. *Syst. Biol.*, 73:183–206, 2024.

[11] Tim G. Vaughan and Alexei J. Drummond. A stochastic simulator of birth-death master equations with application to phylodynamics. *Mol. Biol. Evol.*, 30:1480, 2013.

[12] J Voznica, A. Zhukova, V. Boskova, E. Saulnier, F. Lemoine, M. Moslonka-Lefebvre, and O. Gascuel. Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks. *Nature Comm.*, 23(3896), 2022.