Quantifying the Impact of Fossil Age on Reconstructing Trait Evolution Using Phylogenetic Comparative Methods

William Gearty[1,2]*, Bethany J. Allen[3,4,5], Pedro L. Godoy[6], Alfio Alessandro Chiarenza[7]

1 *Division of Paleontology, American Museum of Natural History, New York, NY 10024, USA*
2 *Open Source Program Office, Syracuse University, Syracuse, NY 13210, USA*
3 *Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland*
4 *Computational Evolution Group, Swiss Institute of Bioinformatics, Lausanne, Switzerland*
5 *GFZ Helmholtz Centre for Geosciences, Potsdam, Germany*
6 *Department of Zoology, Institute of Bioscience, University of São Paulo, São Paulo, Brazil*
7 *Department of Earth Sciences, University College London, London, UK*

*Corresponding author:
Open Source Program Office, Syracuse University, 110 Smith Dr, Syracuse, NY 13210, USA,
willgearty@gmail.com

ABSTRACT

Collecting data for use in constructing phylogenies is a valuable but time- and resource-consuming pursuit. As a result, indicators of the potential value of including certain species in a phylogeny *a priori* could prove useful when planning this stage of research. Here, we used a simulation approach to investigate whether there are trends in the ability for phylogenetic comparative methods to recover the correct model of trait evolution based on certain characteristics of the phylogeny. First, we used multiple diversification rates to simulate phylogenies containing varying proportions of fossil and extant tips. We then simulated the evolution of a single trait across each phylogeny using multiple continuous trait evolution models, and compared the fit of the correct and incorrect models. This quantitative evaluation allowed us to discern whether there are certain tip characteristics associated with identifying the correct trait evolution models. Our results indicate that the inclusion of fossils can be highly beneficial to reconstructing certain trait histories (e.g., Ornstein-Uhlenbeck and ACDC) but not to others (e.g., Brownian motion). In fact, in many cases, increasing the proportion of fossils in a phylogenetic dataset is far more beneficial, and perhaps more time- and resource-efficient, than increasing the number of extant taxa in the dataset. Our results corroborate previous findings that the inclusion of fossil tips can vastly improve the reconstruction of trait histories, but also show that this effect is often stronger for older fossils.

Keywords: fossils continuous Ornstein-Uhlenbeck ACDC Brownian motion phylogeny simulation

INTRODUCTION

Phylogenies describe the hypothesized or inferred evolutionary relationships between biological entities. Phylogenetic comparative methods (PCMs) use phylogenies to investigate the intricate patterns and processes of evolution along the branches of the evolutionary tree (Felsenstein 1985; Cornwell and Nakagawa 2017; Cornwallis and Griffin 2024). Such analyses include estimating ancestral states of discrete or continuous characters (including historical biogeography), inferring the tempo and/or mode of evolution of one or more traits, and identifying diversification dynamics (Butler and King 2004; Stadler 2011; Slater 2013; Rabosky et al. 2018; Allen et al. 2019; Godoy et al. 2019; Harmon 2019; Gearty et al. 2021). However, a longstanding conundrum regarding the application of computational methods in evolutionary biology is how to ensure that the results of downstream analyses are valid, particularly as experimental data against which such methods can be verified often does not exist (Barido-Sottani et al. 2020a). This is particularly true of PCMs, with many of the most-commonly used methods and models being criticized for their potential inaccuracy (Boettiger et al. 2012; Pennell et al. 2015; Cooper et al. 2016; Louca and Pennell 2020; Cornuault 2022).

A major assumption underlying PCMs, which likely further contributes to the inaccuracy of the results they produce, is that the phylogeny being used is a fair representation of the underlying genealogical population, i.e. the clade of interest (Symonds 2002; Warnock et al. 2020; Hibbins et al. 2023). If the phylogeny is not representative of the clade's evolutionary history (due to, for example, bias in the branches sampled, incorrect relationships, or invalid branch lengths), it may result in incorrect inferences, which may then lead to false evolutionary conclusions. However, as a phylogeny approaches the true history, the inferences of PCMs should likewise approach the true evolutionary dynamics (provided that the underlying process of trait evolution can also be appropriately modeled). Choices concerning the data from which the phylogeny is inferred are fundamental to how likely the phylogeny is to truly represent evolutionary history. Collecting this data is a labor-intensive and resource-demanding (e.g., time, choices, money, computation) endeavor. Therefore, it is imperative to find ways to determine which of the decisions made when collecting data for constructing a phylogeny have the largest impact on PCM accuracy, so that researchers can make informed choices based on a cost-benefit approach (Mongiardino Koch and Parry 2020; Mongiardino Koch et al. 2021).

Recent computational developments have allowed for the construction of dated phylogenies, in which branch lengths represent time, to be conducted via the integration of diverse sources of information, such as genetic sequences, phenotypic characters, fossil ages, and ecological data (Ronquist et al. 2012; Zhang et al. 2016, 2023; Wright et al. 2022; Mulvey et al. 2025). The types of data to include is therefore one of the major choices when planning data collection for phylogenetic inference. One potential choice in the construction of phylogenies for PCMs is whether extinct taxa are included, and, if so, which ones are included (e.g., Puttick 2016). Paleontology has a long history of producing phylogenies based on extinct taxa using morphological character matrices (Vrba 1979; Gauthier 1986). However, the combination of morphological and molecular data is a much more recent phenomenon (Springer et al. 2001; Schrago et al. 2013; Zou and Zhang 2016). Since the inception of these combined data approaches (also known as “total-evidence”), paleontologists have advocated for the inclusion of fossils as tips in phylogenies (Wright et al. 2022; Zhang et al. 2023; Heckeberg et al. 2026).

Previous papers have investigated how well PCMs work when using calibrated phylogenies that do or do not include fossils. Several studies have shown that ancestral states estimated based on extant-only data can be extremely biased (Webster and Purvis 2002; Finarelli and Flynn 2006; Royer-Carenzi et al. 2013). Slater et al. (2012) demonstrated that including fossils as tips in phylogenies increases the power of PCMs to detect the true model of trait evolution. This point has been echoed in case studies examining the bird, caniform and monkey fossil records (Finarelli and Goswami 2013; Mitchell 2015; Silvestro et al. 2015), although see Puttick and Thomas (Puttick and Thomas 2015) for an example in Afrotheria in which separate inferences based on living taxa and fossils agree. Further, several studies have shown that diversity dynamics and paleobiogeography cannot be accurately estimated with extant-only phylogenies, even when node ages are calibrated (Quental and Marshall 2010; Rabosky 2010; Louca and Pennell 2020; Wisniewski et al. 2022; Faurby et al. 2024), although the benefits of including fossils remain unclear (Louca et al. 2021; Černý et al. 2022; Beaulieu and O’Meara 2023). However, despite these observations and cautions, extant-only time trees remain ubiquitous in studies conducting phylogenetic comparative methods (Jetz et al. 2012; Magallón et al. 2015; Jetz and Pyron 2018; Rabosky et al. 2018; Upham et al. 2019; Álvarez-Carretero et al. 2022; Pie et al. 2023; Smyčka et al. 2023).

Given the complexity of these issues and piecemeal nature of model adequacy investigation and reporting (Pennell et al. 2015), a comprehensive quantitative assessment of the impact of the inclusion of fossils in phylogenies that are used for PCMs is needed. A central issue, as highlighted by Oakley and Cunningham (2000), Slater et al. (2012), and Cooper et al. (2016), and further underscored by Grabowski et al. (2023), pertains to the challenge of model traceability, especially in the context of complex (multiparametrized) models (Hansen 1997). This challenge primarily emerges when dealing with incomplete and poorly sampled datasets, significantly affecting the ability of the model to accurately track the evolutionary trajectories of specific traits within a phylogenetic framework. This evaluation will not only clarify the suitability of complex but well-defined modes of evolution in comparative analyses, but also enhance our understanding of the broader challenges related to phylogenetic comparative methods in the face of incomplete datasets. Such constraints are imperative to ensure that researchers can make informed decisions when applying these methods and understand the implications of these decisions on the potential accuracy of their results.

Here, we determine how the inclusion of varying proportions of fossil tips, and their relative ages, affect the ability of PCMs to accurately recover specific and widely used models of continuous trait evolution. To achieve this, we implemented a multi-faceted strategy: we first simulated phylogenies of varying sizes, which contained varying proportions of root-biased, random, or recent-biased fossil (non-extant) tips. We then simulated the evolution of a single continuous trait across each phylogeny using multiple trait evolution models, before comparing the fit of correct and incorrect models to these simulated traits. Our results corroborate previous findings that the inclusion of fossil tips can vastly improve the reconstruction of trait histories, but also show that this effect is stronger for older fossils. The inclusion of fossil tips is also more beneficial for traits which evolved under multi-regime models, but less impactful for other, simpler models. Further, it is impossible to correctly infer a shift in the optimum trait value over time in phylogenies only containing extant tips. In some instances, increasing the proportion of extinct taxa within a phylogenetic dataset emerged as more efficient than increasing the overall number of tips by multiple orders of magnitude. These findings are a major step towards developing precise, quantitative guidelines for planning the construction of phylogenetic datasets for PCMs that incorporate both modern and fossil taxa. Further, they quantify how much caution should be taken when interpreting the results of PCMs.

MATERIALS & METHODS

Simulating Phylogenies

We developed a new workflow to simulate phylogenies that have both a fixed size and a set proportion of tips which are extinct (Fig. 1). First, to reflect the cladogenesis process (Fig. 1A), we generated a birth-death tree with specific birth (λ) and death (μ) rates using the "sim.bd.taxa" function from the *TreeSim* R package (Stadler 2019). This function simulates trees using rejection sampling based on a user-provided desired number of extant taxa. Simulated trees with lower extinction rates have fewer fossil tips relative to extant tips. Since we needed to ensure that we had enough fossil tips to sample for the final phylogeny, we needed to inflate the desired number of extant taxa when simulating under lower relative death rates. To accomplish this, we set the number of desired extant tips to $n * \frac{\lambda \div \mu}{1 \div 0.9}$, where *n* is the size of the desired final phylogeny. We used the complete = TRUE option to obtain the full birth-death tree including extinct tips. Then, to reflect the fossilization process (Fig. 1B), we simulated fossil occurrences under a Poisson process using the "sim.fossils.poisson" function from the *FossilSim* R package (Warnock et al. 2022). We used a Poisson sampling rate of 50 to ensure a large number of fossil occurrences were generated. We repeated this Poisson sampling until the number of extinct edges (i.e., non-extant and non-internal edges) with fossil occurrences met or exceeded the desired number. Then, to reflect the fossil sampling/recovery process (Fig. 1C), we defined a function that assigned an individual recovery potential value to each fossil occurrence based on its relative age (t: [0, 1]). By default, this workflow uses a uniform function ($f(t) = 1$) where all fossils are given the same recovery potential. Other functions can also be substituted in to impose increases or decreases in recovery potential through time. Once recovery potentials were assigned across the tree, we summed them up for all fossil occurrences along each fossil edge, yielding a single recovery potential per edge. We then randomly sampled the desired number of fossil tips from the full set of fossil tips (Fig. 1D), with the probability of sampling a given tip equal to its recovery potential. We used the *FossilSim* R package (Warnock et al. 2022) to drop the unsampled extinct tips from the phylogeny. Finally, to reflect the incomplete sampling of extant taxa (Fig. 1D), we randomly dropped extant tips until we had a final phylogeny with the desired proportion of extinct tips. This sampling process skews the birth and death rates away from those used to simulate the phylogeny but is necessary to create strong variation in the distribution of tips over time.

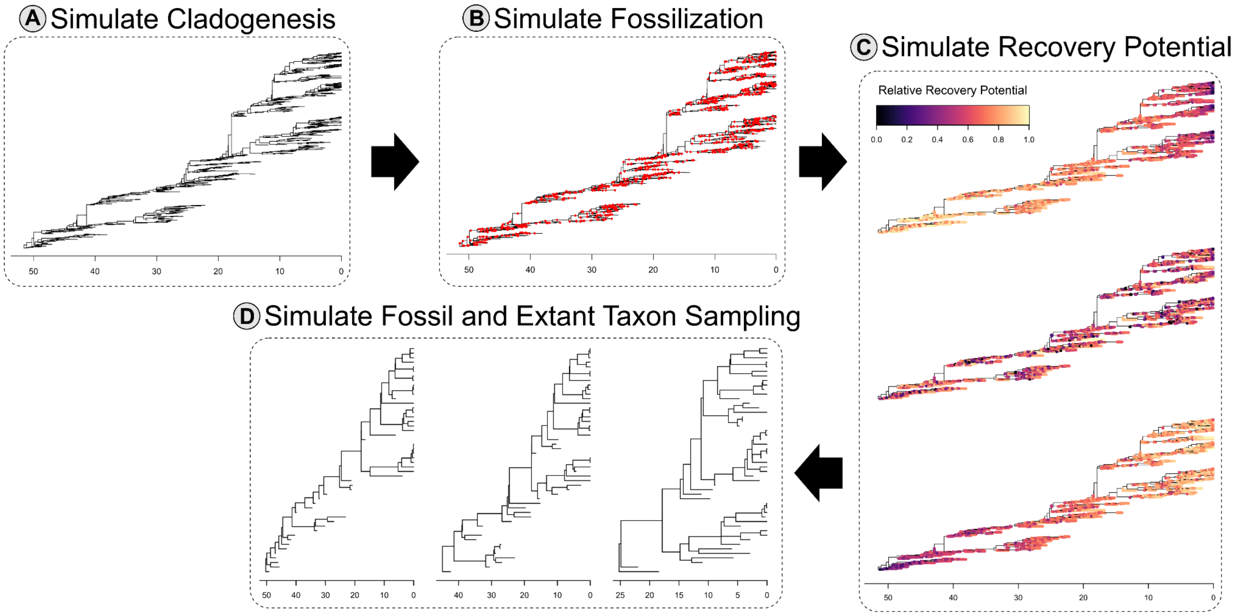


Figure 1: Diagram illustrating the simulation workflow that generates phylogenies that have both a fixed size and a set proportion of fossil tips. Step A: simulate the cladogenesis process by generating a phylogeny with a fixed number of extant tips under given birth and death rates. Step B: simulate the fossilization process by generating many fossils along the extinct tips under a Poisson process. Step C: simulate the fossil recovery process by assigning a recovery potential to each fossil based on a specified time-varying function (top: root-biased, middle: random, bottom: recent-biased). Step D: simulate the fossil collection process by randomly choosing fossil tips weighted by their summed recovery potentials.

We used this workflow to simulate a suite of phylogenies using a range of parameters relevant to the evolutionary dynamics of clades and the construction of morphological matrices and phylogenies (Barido-Sottani et al. 2020a). First, we used two different death rates (μ; 0.5 and 0.9) in combination with a static birth rate (λ; 1.0) to simulate evolutionary histories with low (0.5) and high (0.9) turnover (μ / λ), respectively. Sufficiently large phylogenies can be computationally difficult to simulate with extremely high turnover values because often the entire clade will go extinct before a sufficient number of lineages have evolved, so we did not attempt turnover values greater than 0.9. Second, we simulated these phylogenies with increasing total numbers of tips to reflect the different sizes of phylogenies that are commonly constructed in evolutionary biology (50, 100, 200, 500, and 1000). Based on our own testing, we decided to use exponential functions to reflect fossil recovery that increases (hereafter referred to as "recent-biased") or decreases (hereafter referred to as "root-biased") over time, in addition to the default uniform recovery function (hereafter referred to as “random”). For the recent-biased recovery function, where fossils are much more likely to be found towards the present (than in the uniform sampling), we used $f(t) = 100^x - 1$. For the root-biased recovery function, where fossils are much more likely to be found closer to the root of the tree, we used $f(t) = 10^{-(x-1)} - 1$. Note that the distributions of the sampled fossil tips do not look like these functions because of the exponential nature of increasing tips through time (Fig. 2 and S1). Finally, we generated these phylogenies with varying proportions of fossil tips to reflect the proportions of fossil taxa that are generally included in morphological matrices and paleobiological phylogenies (0, 0.1, 0.25, 0.5, and 0.95; n.b. the number of fossil tips is rounded down to the nearest integer). For each combination of these parameters, we simulated 100 different phylogenies. This resulted in a total of 15,000 simulated phylogenies.

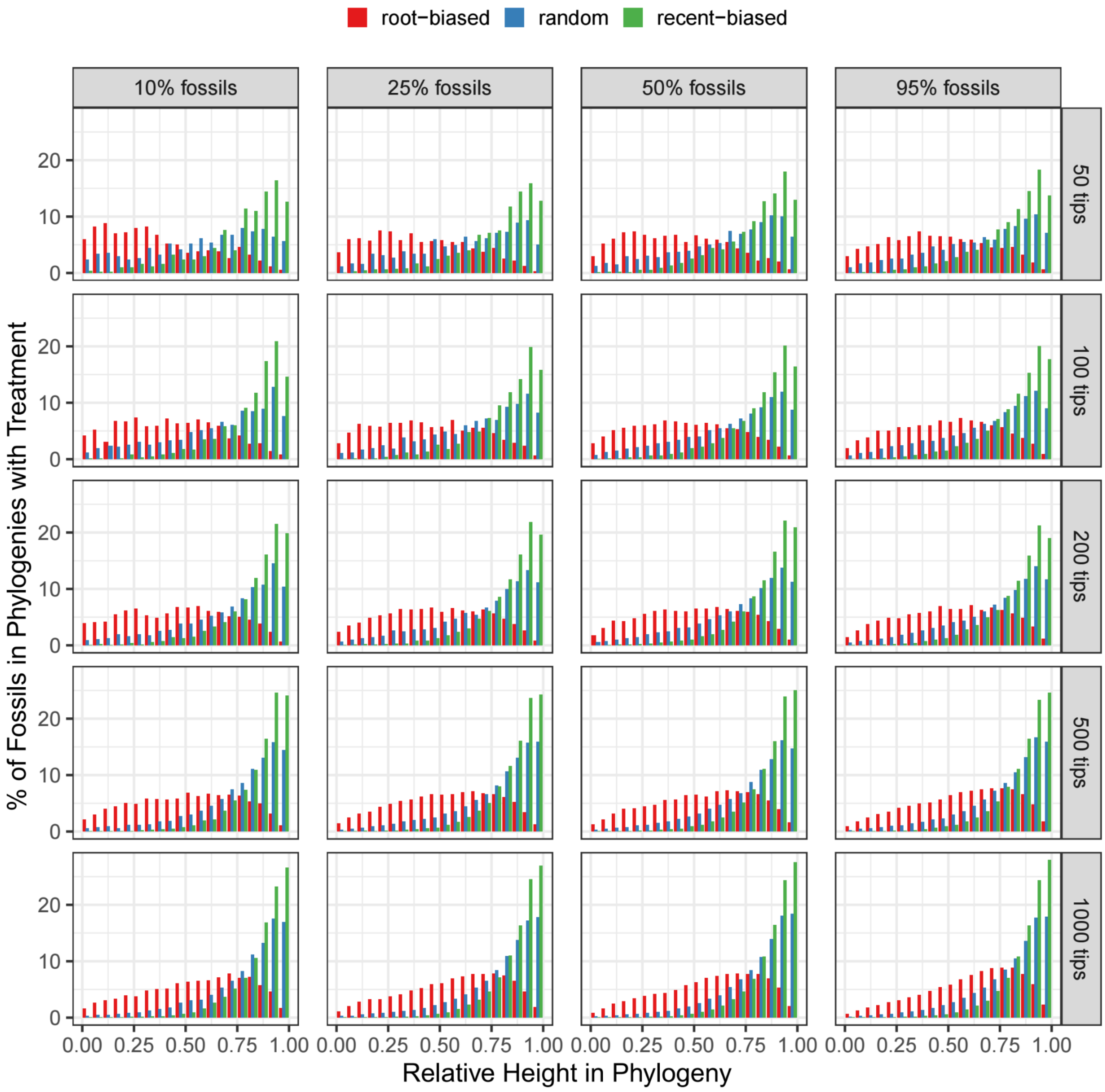


Figure 2: The realized relative height distributions of the fossils in simulated phylogenies. The results are split out by the number of tips in the simulated phylogeny (row headers), the proportion of fossils in the simulated phylogeny (column headers), and the temporal bias of the fossils (color). Each simulated phylogeny has a different raw height, so the ages of the fossils are presented here as relative heights. These results are only for analyses where $\mu = 0.9$, the results for analyses where $\mu = 0.5$ are in Fig. S1.

Simulating Traits

For each of our simulated phylogenies, we also simulated the evolution of continuous traits under 12 different evolutionary models (Table 1) using the *mvMORPH* R package (Clavel et al. 2015). The simplest model employed was Brownian motion (BM, Felsenstein 1985; O'Meara et al. 2006), whereby the trait evolves under a stochastic process akin to a random walk, governed by a strength of drift parameter σ. Two different σ values were chosen to represent weak (0.1) versus strong (0.5) drift. In addition to standard BM models, we also implemented BM models with trends, i.e. with a moving average trait value through time (trend, Pagel 2002; Hunt and Carrano 2010). These also used a weak (0.1) and strong (0.3) trend strength, with a fixed σ of 0.1.

The second family of models applied was Ornstein-Uhlenbeck (OU, Hansen 1997; Butler and King 2004; Beaulieu et al. 2012). In addition to the drift facilitated by Brownian motion, OU models have additional parameters to represent the strength of pull, α, towards an optimum trait value, θ. We implemented two types of OU models: the OUs model has a θ that is distinct from the ancestral state, representing adaptive evolution towards an optimum value which differs from the root state, while the OUc model has a θ that is identical to the ancestral state, which represents stabilizing evolution, where Brownian drift is counteracted by a spring-like pull towards the optimum (Beaulieu et al. 2012). To account for the varying heights of our simulated trees, instead of using a set α parameter, we calculated α individually for each simulated tree based on the desired phylogenetic half-life ($ln(2)/\alpha$) relative to the height of the trees (Hansen et al. 2008; Cornuault 2022). For the weak OU models, our desired half-life was 100% of the tree height, so we calculated α as $\frac{ln(2)}{tree\ height}$. For the strong OU models, our desired half-life was 20% of the tree height, so we calculated α as $\frac{ln(2)}{tree\ height\ /\ 5}$.

The third family of models applied was accelerating and decelerating rate (ACDC, Blomberg et al. 2003; Harmon et al. 2010). Alongside the drift of Brownian motion, these models have a parameter β which describes the rate at which trait evolution changes over time. The decelerating rate model is particularly commonly applied, as it is thought to describe an "Early Burst" mode of evolution, under which a new clade experiences high rates of initial trait evolution which slowly decay over time (Simpson 1944; Harmon et al. 2010; Puttick 2018). These models were also implemented using a weak (0.1 and -0.1 respectively) and strong (0.3 and -0.3 respectively) strength, with reduced drift ($\sigma = 0.001$) compared to the other models in order to accentuate the effect of the overall rate trend.

Table 1. The parameter values used in individual models to simulate trait evolution on each phylogeny. In all models, the ancestral trait state was set to 0.

Model	**Strength**	**Strength of drift (σ)**	**Other parameters**	**Abbr.**
Brownian motion (BM)	Weak	0.1		wBM
	Strong	0.5		sBM
BM with trend	Weak	0.1	*trend* = 0.1	wtrend
	Strong	0.1	*trend* = 0.3	strend
Ornstein-Uhlenbeck with identical root and optimum states (centered, OUc)	Weak	0.1	$\theta = 0$; $\alpha = \frac{ln(2)}{tree\ height}$	wOUc
	Strong	0.1	$\theta = 0$; $\alpha = \frac{ln(2)}{tree\ height\ /\ 5}$	sOUc
OU with different root and optimum states (shifting, OUs)	Weak	0.1	$\theta = 2$; $\alpha = \frac{ln(2)}{tree\ height}$	wOUs
	Strong	0.1	$\theta = 2$; $\alpha = \frac{ln(2)}{tree\ height\ /\ 5}$	sOUs
Accelerating (AC)	Weak	0.001	$\beta = 0.1$	wAC
	Strong	0.001	$\beta = 0.3$	sAC
Decelerating (DC; “Early Burst”)	Weak	0.001	$\beta = -0.1$	wDC
	Strong	0.001	$\beta = -0.3$	sDC

Model Fitting

Having simulated the necessary data, we used the *mvMORPH* R package to fit five evolutionary models to the phylogenies and traits: BM (using the mvBM function), BM with a trend (mvBM with trend=TRUE), OU under stabilizing selection (mvOU with root=FALSE), OU with differing root and optimum states (mvOU with root=TRUE), and ACDC (mvEB with an upper bound of 1). Considering that we had simulated 12 traits across each of 15,000 trees, we ultimately ran 900,000 model fitting analyses in total. In 17 of these analyses, a model was unable to be fit to a particular trait simulation due to computational issues; we ignored these fits for all downstream comparisons and assessments.

Analyses without fossils (fossil proportion = 0) were incompatible with the three types of temporal bias employed for the analyses with fossils; these analyses were only performed once rather than three times to prevent biasing overall result calculations. Furthermore, phylogenies without fossils cannot provide information about possible trait value or optimum shifts through time, and therefore are always best fitted by models excluding these parameters (Slater et al. 2012; Clavel et al. 2015). While the *mvMORPH* package permits fitting such models to phylogenies without fossils, the package documentation states that models with trends are only identifiable with non-ultrametric trees, and that the `root=TRUE` option for the "mvOU" function (which we used across all OUs model fittings) can be problematic with ultrametric trees (Clavel et al. 2015). Therefore, we excluded all results for these models (BM with trend, OU with differing root and tip optima) that lacked fossil tips from downstream analyses, leaving 760,000 total analyses for further analysis.

For each simulated trait, we calculated the AICc weights of the five fitted models to assess their relative fit to the trait data (Sugiura 1978; Burnham and Anderson 2002). The model with the highest AICc weight was considered the best-fitting model. AICc is a modification of the Akaike information criterion (Akaike 1974) with an extra penalty term for the number of parameters in the fit model. Nonetheless, it is worth noting that previous studies have raised concerns about the use of AIC for comparing the fit of phylogenetic models, suggesting that this criterion may still be biased toward model overfitting (Ho and Ané 2014). Among the proposed alternatives, perhaps the most relevant is the phylogenetic Bayesian information criterion (pBIC, Khabbazian et al. 2016), which has been adopted in some subsequent studies (e.g., Benson et al. 2018; Godoy et al. 2019). Nevertheless, AIC remains the most commonly used criterion in model-fitting analyses (e.g., Troyer et al. 2022; Farina et al. 2023; Grossnickle et al. 2024; Sternes et al. 2024; Almeida et al. 2025), and we therefore adopted this approach here.

If the best-fitting model (i.e., with the highest AICc weight) was the same as the simulated model, we considered this a "correct" assessment; otherwise, if the best-fitting model was a different model, we considered this an "incorrect" assessment. We considered the proportion of these assessments that were "correct" as the "accuracy" under a particular set of parameters. We also calculated the ΔAICc between the best-fitting model (regardless of whether it was "correct") and the other four models to assess whether any other models also had "substantial" empirical support ($\Delta\text{AICc} \leq 2$, Burnham and Anderson 2002). If all of the other fitted models had $\Delta\text{AICc} > 2$, we considered this a "clear" result, otherwise we considered this an "unclear" result. We used these definitions to calculate the proportion of simulations in which 1) the best-fitting model was "correct" and "clear" (i.e., we got the correct model and it was substantially better than the other models), 2) the best-fitting model was "correct" and "unclear" (i.e., we got the correct model, but one or more other model(s) could not be ruled out), 3) the best-fitting model was "incorrect" and "unclear" (i.e., we got the wrong model, but one or more other models could not be ruled out), and 4) the best-fitting model was "incorrect" and "clear" (i.e., we got the wrong model, and it was substantially better than the other models, including the correct model). We used the first two categories to identify the "accuracy" (the proportion of simulations in categories 1 or 2) and "clear accuracy" (the proportion of simulations in category 1) of our model-fitting approach. We collated the results based on the identity of the simulated model, to assess whether any particular model had relatively high or low chances of being incorrectly chosen over the simulated model. We also collated the results based on the identity of the best-fitting model (BM, trend, OUc, OUs, ACDC), to illustrate interpretation under a more empirical-style scenario, to address the question "If this is my best-fitting model, what is the model that likely generated my data?" Finally, we also extracted the estimated parameters (σ, a, β, *trend*, and θ) from the results of the model fitting analyses when the generative and fitted models matched, to assess how well parameters are estimated across the range of simulation variables.

All analyses were conducted using version 4.5.2 of the R programming language (R Core Team 2024). We used the *dplyr*, *tidyr*, *tibble*, and *forcats* R packages to manipulate data (Müller and Wickham 2023; Wickham 2023; Wickham et al. 2023, 2024). We used the *ape*, *phytools*, *pcmtools*, *TreeSim*, and *FossilSim* R packages to simulate and manipulate phylogenies (Paradis and Schliep 2019; Stadler 2019; Warnock et al. 2022; Gearty 2023; Revell 2024). We used the *pbapply*, *future*, and *future.apply* R packages to conduct analyses in parallel (Bengtsson 2021; Solymos and Zawadzki 2023). We used the *ggplot2*, *ggh4x*, and *deeptime* R packages to generate visualizations of the results (Wickham 2016; Brand 2024; Gearty 2024). All code is available on GitHub at https://github.com/willgearty/PCM_sensitivity and archived on Zenodo (Gearty et al. 2026).

RESULTS

Here we focus on the analyses with relatively high turnover (where $\mu = 0.9$), which we consider to be more biologically realistic (Marshall 2017). For the analyses with lower relative turnover (where $\mu = 0.5$), results are presented in the Supplementary Materials. We note in the main text when outcomes differ dramatically between the two turnover scenarios.

Overall Accuracy

Across all high turnover model-fitting analyses, we found that 68.8% recovered the generative model as the best-fitting model (hereafter referred to as "accuracy"). When only those model-fits that are clearly better (ΔAICc > 2) than the second best-fitting model are counted, this accuracy (hereafter referred to as "clear accuracy") is only 46.4% (Fig. 3; pink portions only ["correct and clear"]). Across all analyses, the overall proportion of best-fitting models that are "incorrect and clear" (the worst-case scenario) is 9% (Fig. 3; green portions). In the absence of fossils, clear accuracy is always below 30%, regardless of phylogeny size (Fig. 3). However, including any fossils in simulations increases clear accuracy by 20.1% (from 24% with no fossils, to 44.1% when just 10% of tips are fossils). When overall accuracy is assessed, the improvement due to the inclusion of fossils is less dramatic, but still notable, with an increase from 58.3% with no fossils to 66.1% with just 10% of tips as fossils. For random and recent-biased fossils, as the proportion of fossils in the phylogeny increases, accuracy also increases, at least when the phylogeny has 200 or fewer tips (Figs. 3 and S2; middle and bottom rows).

However, a small proportion of root-biased fossils is often as beneficial for accuracy as a large proportion, regardless of the phylogeny size (Figs. 3 and S2; top row). There is also a general increase in accuracy as phylogeny size increases, with or without fossils. This size effect is largest when fossils are recent-biased and low in proportion, but is otherwise fairly minor compared to the accuracy increase due to the inclusion of any fossils.

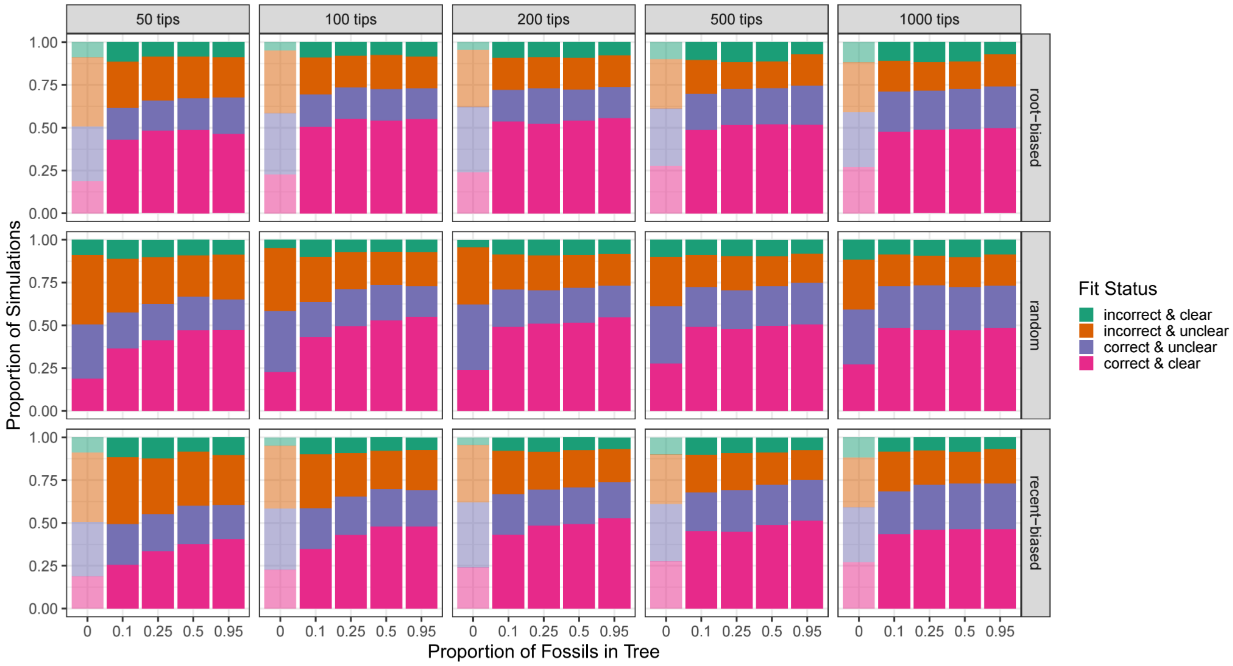


Figure 3: Summary of the best-fitting models across all simulations. The best-fitting model for a given simulation can be either correct (matching the simulated model) or incorrect (not matching the simulated model) and also either clear (all other models with ΔAICc > 2) or unclear (at least one other model with ΔAICc ≤ 2). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row headers), and the number of tips in the simulated phylogenies (column headers). Analyses without fossils are presented in full color in the row for randomly distributed fossils, and those same results are shown faded for the other rows. These results are only for analyses where $\mu = 0.9$, the results for analyses where $\mu = 0.5$ are in Fig. S2. Results split out by model instead of phylogeny size are shown in Fig. S3.

Brownian motion

When the simulating model was a Brownian motion model, this was also the best-fitting model in 67.2% of all simulations (Fig. 4). However, the clear accuracy for these models is much lower, at 5.6% (Fig. S3). These percentages are fairly consistent regardless of fossil distribution, phylogeny size, relative extinction rate, or the strength of the Brownian motion ("wBM" versus "sBM"). Notably, Brownian motion is the only model for which the clear accuracy drops when adding fossils, from ~20% without fossils to ~5% when any fossils are included (Fig S3). When a trend is added to the Brownian motion model ("wtrend" and "strend"), accuracy improves to

92.6% (with 85% clear accuracy). For strong trends (“strend”), the overall accuracy is 96.8% and all sets of parameters have >80% accuracy (Fig. 4). Accuracy generally increases or stays the same with increasing fossil tip proportion and/or phylogeny size. For weak trends (“wtrend”), accuracy is lower (88.4%), especially for simulations with recent-biased fossil distributions (82.6%). However, with random (89.2% average) or root-biased (93.2% average) fossil distributions, accuracy is >60% across the board. The BM simulation accuracy results are qualitatively the same for both death rates (μ), although simulations with a weak trend tend to have reduced accuracy (Fig. S4).

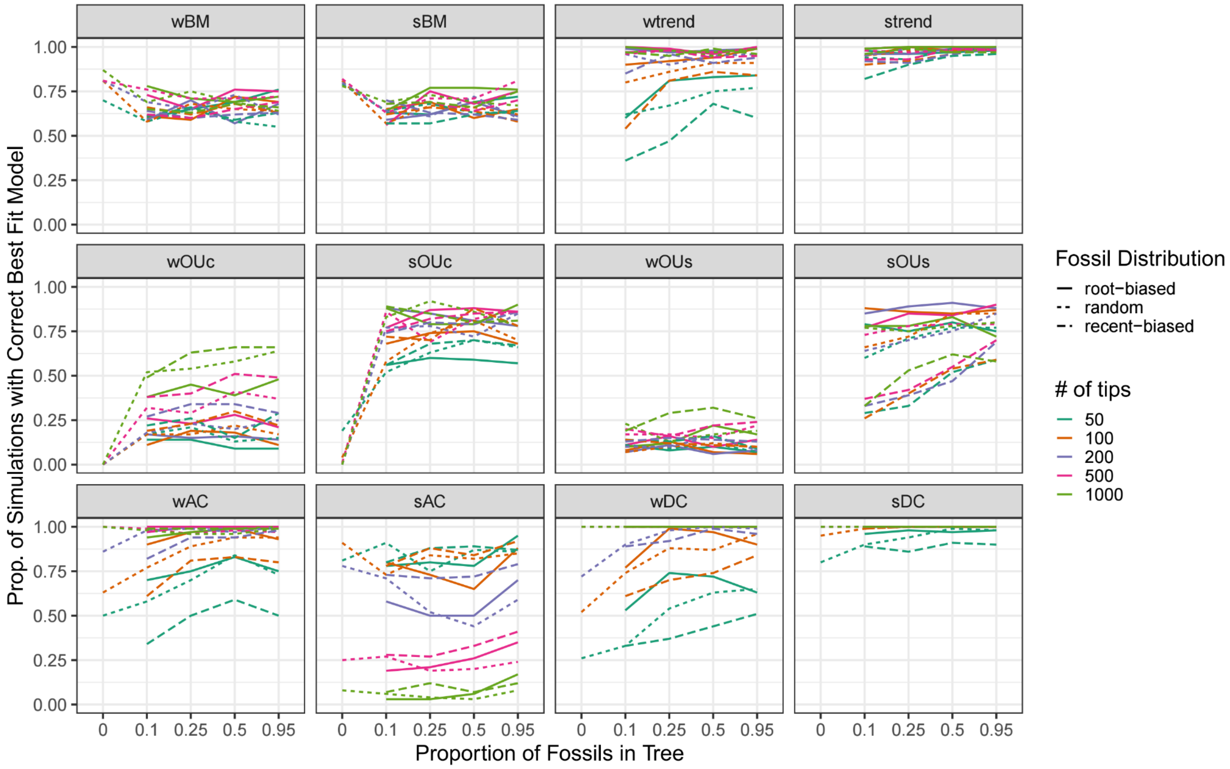


Figure 4: The proportion of simulations for which the best-fit model does match the simulated model (panel headers). These results are agnostic with regards to the ΔAICc score, and therefore combine the “correct & clear” and “correct & unclear” outcomes from Fig. 3. The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (line type), and the number of tips in the simulated phylogenies (line color). Analyses without fossils are plotted as part of the random distribution. These results are only for analyses where $\mu = 0.9$, the results for analyses where $\mu = 0.5$ are in Fig. S4.

For simulated Brownian motion models, when no fossils are included, ACDC models are nearly the only models that are chosen over a BM model, for 19.8% of simulations (Figs. 5 and S5). However, when any fossils are included, the probability of incorrectly choosing an ACDC model stays relatively the same, but the probability of choosing a trend (13%), OUc (4.2%), or OUs (1.4%) model increases slightly. For BM models with weak trends, 4.4% of simulations are incorrectly identified as BM, 1.8% of simulations are incorrectly identified as ACDC, and 5.5% of simulations are incorrectly identified as OUc or OUs. With a strong trend model, 3.1% of simulations are incorrectly identified as OUs, with a slight increase to 5.1% when the fossils are recent-biased (Figs. 5 and S5).

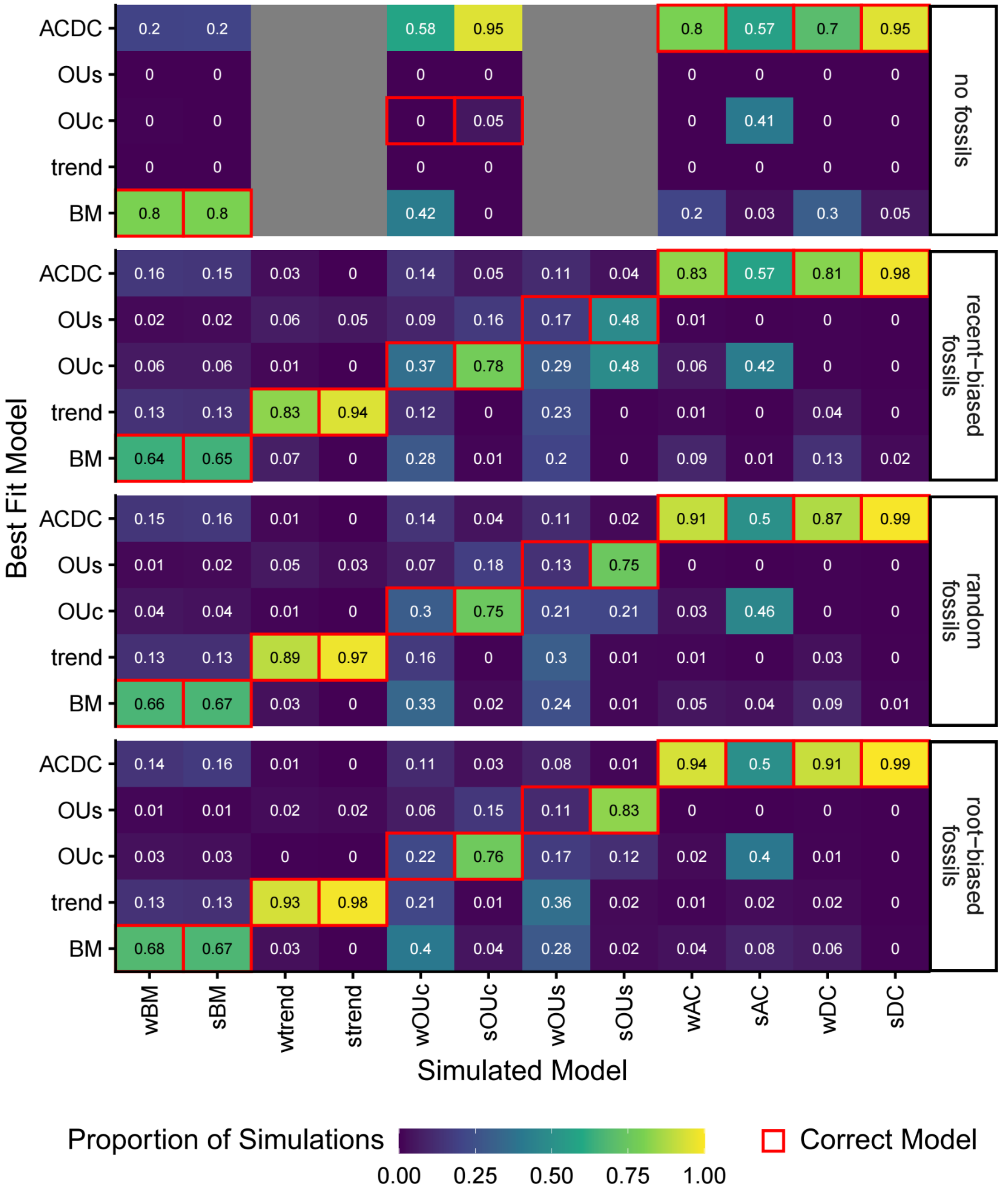


Figure 5: Confusion matrix showing the proportions of simulations that were best fit by each model (y-axis). The results are split out by the simulated model (x-axis) and the fossil

treatment (rows). Analyses without fossils are split out as a separate treatment. Red outlines represent the correct model given the simulated model (note that there is not a one-to-one relationship). Grey boxes indicate analyses that were skipped due to nonidentifiability. These results are only for analyses where $\mu = 0.9$, the results for analyses where $\mu = 0.5$ are in Fig. S5. Finer resolution breakdowns of these results are shown in Figs. S6-S11.

In all simulations for which BM is the best-fitting model, this is incorrect 39.4% of the time. As phylogeny size increases, the proportion of misidentified simulations decreases, from 58.3% for the phylogenies with 50 tips to 17.5% for phylogenies with 1000 tips (Figs. 6 and S12).

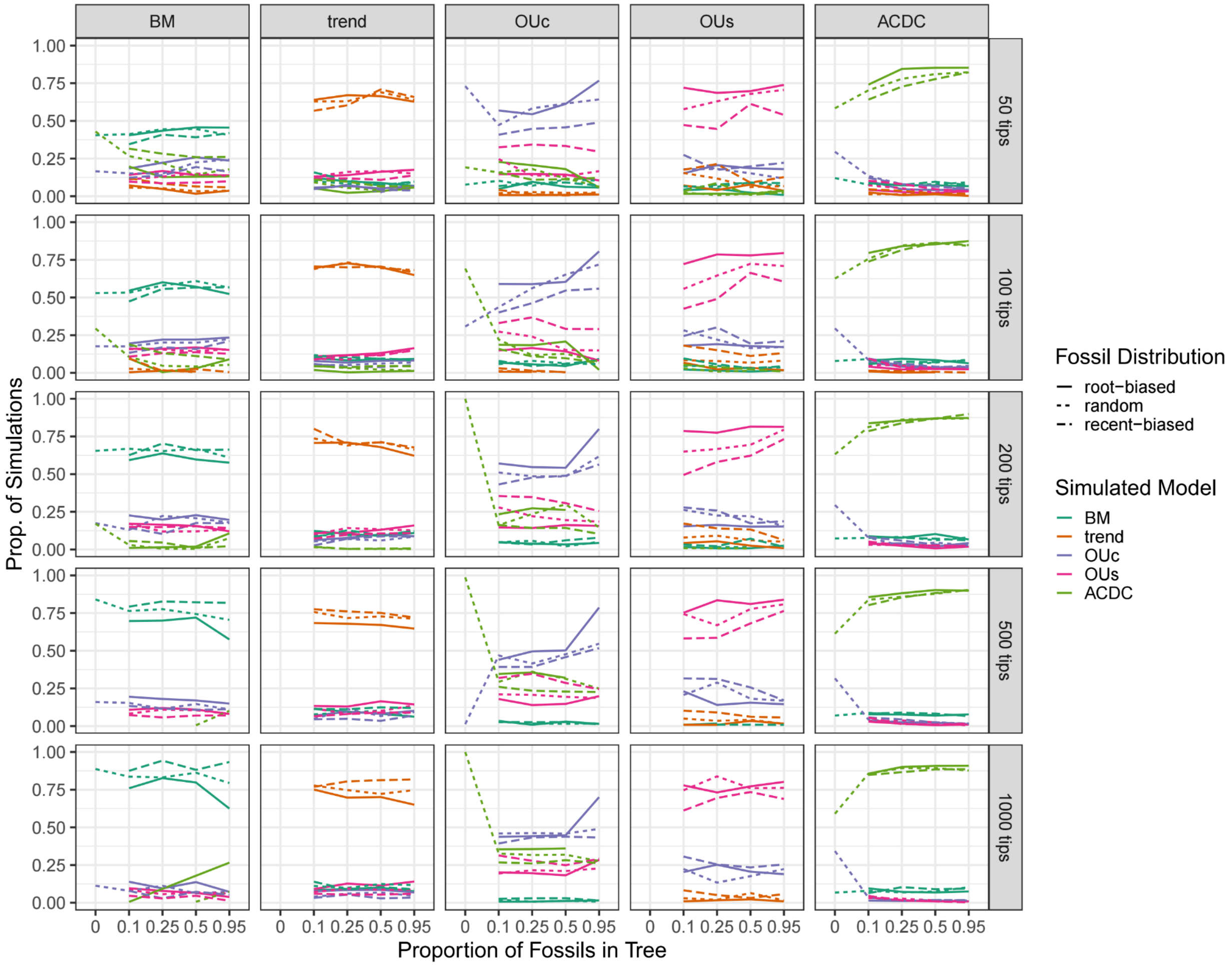


Figure 6: The proportion of simulations best fit by a given model (column headers) that were generated by each simulated model (line color). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the age distribution of the simulated fossils (line type), and the number of tips in the simulated phylogenies (row headers). Analyses without fossils were either not conducted due to the presence of a trend (*trend*, *OUs*) or are plotted as part of the random distribution. These results are only for analyses where $\mu = 0.9$, the results for analyses where $\mu = 0.5$ are in Fig. S12.

Ornstein-Uhlenbeck

The overall accuracy for OU simulations when no fossils are included is 2.4%, with a clear accuracy less than 1% (Figs. 4, S3, and S4). Note that this only includes OUc simulations since OUs were excluded due to nonidentifiability. When any fossils are included, this rises to an overall accuracy of 47% and a clear accuracy of 19.6%.

Weak OU simulations are particularly difficult to accurately infer. When no fossils are included, weak OU simulations with fixed optima ("wOUc") have an accuracy of 0%. When any fossils are included, this overall accuracy increases to 29.3% for wOUc and 13.7% for weak OU simulations with shifted optima ("wOUs"). Within both sets of simulations, those with root-biased fossil distributions have 21.8% overall accuracy for wOUc simulations and 10.7% overall accuracy for wOUs simulations (Figs. 4 and S4), while simulations with random fossil distributions show higher relative accuracy (up to 36.6% overall for wOUc simulations and up to 17.1% overall for wOUs simulations). The initial addition of 10% random or recent-biased fossils to the phylogeny provides the largest increase in accuracy, with additional fossils providing marginal improvements.

Simulations with strong OU signals ("sOUc" and "sOUs") generally have higher accuracy than their respective weak models. When no fossils are included, strong OU simulations with no shifted optima ("sOUc") have an overall accuracy of 4.8%. When any fossils are included, this overall accuracy increases to 76.4%. In general, accuracy appears to increase with increasing phylogeny size, although this increase (~5-25%) is much smaller than the increase related to adding fossils (~40-75%). Accuracies are generally lower for strong OU simulations with a shifting optimum ("sOUs") with overall accuracy of 68.5%. The age of the included fossils is important: models with recent-biased fossils have an average accuracy of 47.5%, models with random fossils have an average accuracy of 75.3%, and models with root-biased fossils have an average accuracy of 82.8%. Finally, for all of the OU models, when the simulations were produced with a lower death rate (μ), accuracy is generally higher for OUs models, with "wOUs" accuracies plateauing around 60% and "sOUs" accuracies reaching 100% for some simulations (Fig. S4).

When no fossils are included, 76.6% of OUc simulations are incorrectly identified as ACDC and 21% of simulations are incorrectly identified as BM (Fig. 5). When fossils are included, 15.4% of all OU simulations are incorrectly identified as BM, 12% of simulations incorrectly identified as trend, and 7.3% of simulations incorrectly identified as ACDC. Further, for wOUs, simulations with root-biased fossils tended to be incorrectly identified as trend (36.4% of simulations), whereas simulations with recent-biased fossils tended to be incorrectly identified as OUc (29% of simulations). Simulations with random fossils showed a gradual transition from tending to a trend model with smaller phylogenies, and tending to an OUc model with larger phylogenies. With a low relative extinction rate, when the sOUs model is incorrectly interpreted, it is almost always as the OUc model, and when the sOUc model is incorrectly interpreted, it is almost always as the OUs model (Fig. S5).

When centered OU (OUc) is the best-fitting model, 23.1% of simulations were generated under an OUs model, 23.2% of simulations were generated under ACDC, and 4% of simulations under BM (Figs. 6 and S12). With only 50 tips, 14% of simulations identified as OUc were actually generated under an ACDC model. This increases to 30.5% with 1000 tips. The inclusion of root-biased fossils results in more OUc simulations being correctly identified as OUc. Of all of the fitted models, the shifting Ornstein-Uhlenbeck (OUs) model was the best-fitting model for the fewest simulations (Figs. 6 and S12). When fossils are included, for simulations best fit by an OUs model, 20.4% of the simulations were generated by OUc, 6.6% of simulations were generated by trend, and 1% of simulations were simulated by ACDC. Misidentifications for OUs simulations under different relative death rates (μ) are qualitatively similar (Figs. 6 and S12).

Accelerating and Decelerating Rate

The accelerating and decelerating rate simulations show accuracy patterns that differ from the other simulation models (Fig. 4). The overall accuracy for ACDC simulations is 81.2%, with a clear accuracy of 75.4%. Fossil inclusion has a minimal impact on accuracy. The simulations with weak ACDC signals (“wAC” and “wDC”) show wide ranges of accuracies, varying from 15% to 100%. The strong AC (“sAC”) and strong DC (“sDC”) simulations show a much narrower range of accuracies (45% to 100%). Adding root-biased fossil tips in ACDC simulations generally increases accuracy, with some scenarios seeing accuracy increase by more than 25%. For most simulations, increasing phylogeny size also appears to increase accuracy. This is particularly impressive with weak AC and DC simulations which show increases from ~25% and ~15% accuracy, respectively, for 50 tip phylogenies, to 100% accuracy for 1000 tip phylogenies. Generally speaking, when the phylogenies were simulated with a lower death rate (μ), accuracy for ACDC simulations decreased (Fig. S3). Most notably, the strong DC simulations showed accuracy drops of up to 75%, except for those using the two largest phylogeny sizes (500 and 1000 tips).

For the ACDC simulations with a low relative extinction rate, 18.6% of simulations are incorrectly identified as BM, 4.1% of simulations are incorrectly identified as trend, and 2.2% of simulations are incorrectly identified as OUc (Figs. 5 and S5). When a relatively high extinction rate is simulated, 5.8% of simulations are incorrectly identified as BM, 11.5% of simulations are incorrectly identified as OUc, and 1.2% of simulations are incorrectly identified as trend. Notably, when the strong AC model was simulated with a high relative extinction rate, 42.3% of these simulations were misinterpreted as the OUc model (Fig. S8), with even higher rates when phylogenies with 500 or more tips were used (66.8% for 500 tips, 85.2% for 1000 tips).

When the ACDC model is the best-fitting model, 7.9% of simulations were actually generated under a BM model, 7% of simulations were generated under an OUc model, and 2.8% of simulations were generated under an OUs model (Figs. 6 and S12). When no fossils are included, the majority of misidentifications were generated under an OUc model (31% of simulations). However, when fossils are included, the majority of misidentifications were generated under a BM model (7.9% of simulations).

Parameter Estimation

Along with determining the best-fitting model type, the accuracy of parameter estimation is also an important factor in fitting models. To quantify this, we collated the parameter estimates produced during model fitting for the model matching the generative model (regardless of how well this model fitted). We found that σ estimates for untrended BM simulations tended to be fairly accurate (i.e., average close to the simulated value), regardless of the inclusion of fossils or the size of the phylogeny (Fig. 7). Simulations with weak σ (less Brownian noise) generally had higher precision (i.e., narrower range of estimates) than those with strong Brownian noise. Precision also increased with increasing phylogeny size. However, we found no noticeable relationship between accuracy or precision and the proportion of fossil tips or the age distribution of the fossils. These results are consistent across both relative death rates (Fig. S13).

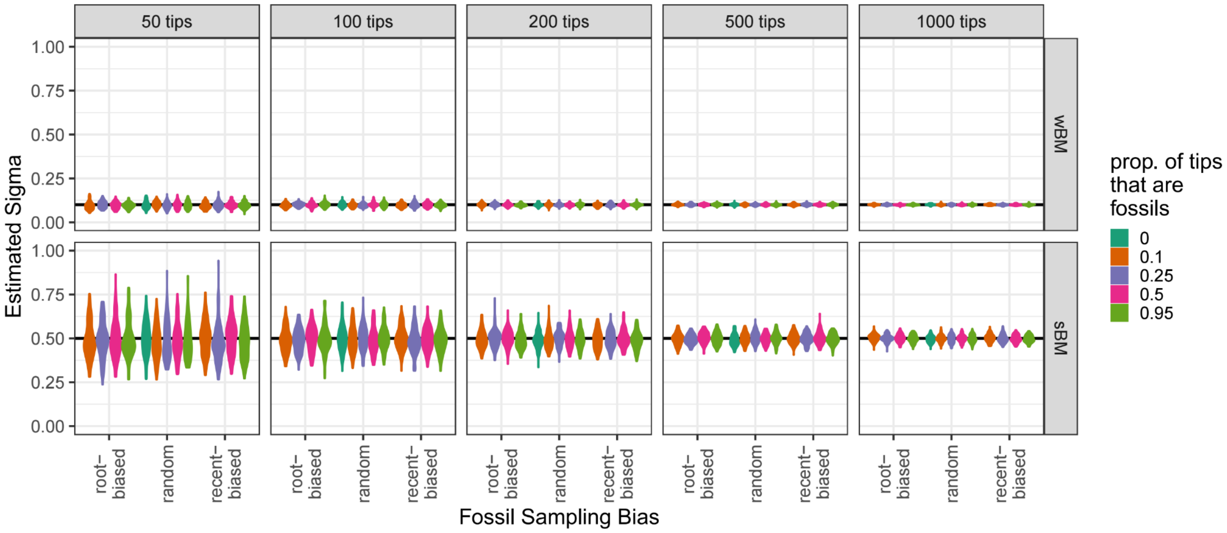


Figure 7: The distributions of the estimates for the σ parameter when a BM model was simulated on a phylogeny and then a BM model was fit to that simulated data. The top row represents simulated trends with a weak σ parameter (0.1) and the bottom row represents a trend model simulated with a strong σ parameter (0.5). The true (simulated) σ parameter is represented with a solid horizontal line, and the results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). Analyses without fossils are plotted as part of the random distribution. These results are only for analyses where $\mu = 0.9$, the results for analyses where $\mu = 0.5$ are in Fig. S13.

When Brownian motion was simulated with a temporal trend, accuracy of the estimation of the strength of the trend appears to be high, although including recent-biased fossils resulted in poor precision compared to root-biased and randomly distributed fossils (Fig. 8). As with untrended BM simulations, increasing phylogeny size is associated with increased precision. Furthermore, increasing the proportion of fossil tips appears to have a marginal increase in precision. These results are consistent across both relative death rates, with the higher death rate having marginally higher precision (Figs. 8 and S14).

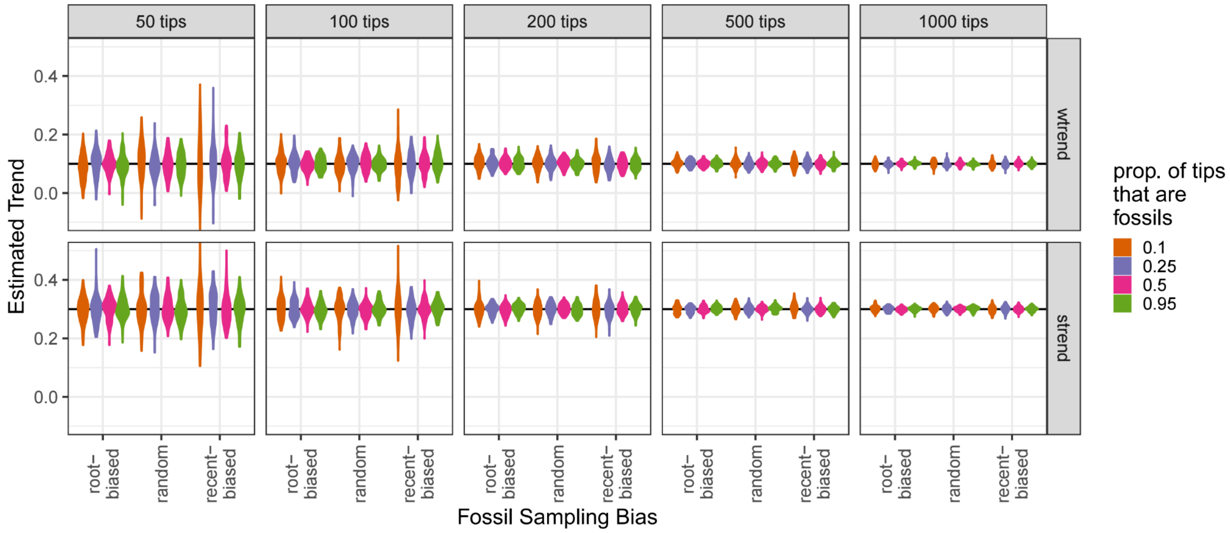


Figure 8: The distributions of the estimates for the *trend* parameter when a trend model was simulated on a phylogeny and then a trend model was fit to that simulated data. The top row represents simulated trends with a weak *trend* parameter (0.1) and the bottom row represents a trend model simulated with a strong *trend* parameter (0.3). The true (simulated) *trend* parameter is represented with a solid horizontal line, and the results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). These results are only for analyses where $\mu = 0.9$, the results for analyses where $\mu = 0.5$ are in Fig. S14.

For the centered Ornstein-Uhlenbeck models (OUc), estimation of the optimum trait value, the theta parameter (θ), has high accuracy, regardless of the number of tips, proportion of fossil tips, and fossil distribution (Fig. 9). Precision appears to be consistent, again regardless of any of these variables. However, for the weak shifting Ornstein-Uhlenbeck models (wOUs), recent-biased fossils generally result in less accuracy for these parameters. For the strong shifting Ornstein-Uhlenbeck models (sOUs), the optimum θ has high estimation accuracy regardless of fossil proportion, fossil age, or phylogeny size. The root θ has a similarly high estimation accuracy when root-biased or randomly-distributed fossils are included. However, when the fossils have a recent-biased age distribution, the root θ has notably very low accuracy. In fact, 22.4% of these model fits resulted in extreme root estimates that were classified as outliers. These θ results are consistent across both relative death rates, with the lower death rate having marginally higher precision (Fig. S15). The half-life parameter generally has high accuracy for strong OU simulations; however, the half-life parameter is often estimated to be relatively closer to zero for weak OU simulations (Fig. 10). Both of these trends appear to be consistent across simulations, regardless of fossil proportion, fossil age, and phylogeny size. These results are consistent for both relative death rates, with slightly higher precision for the lower death rate (Figs. 10 and S16).

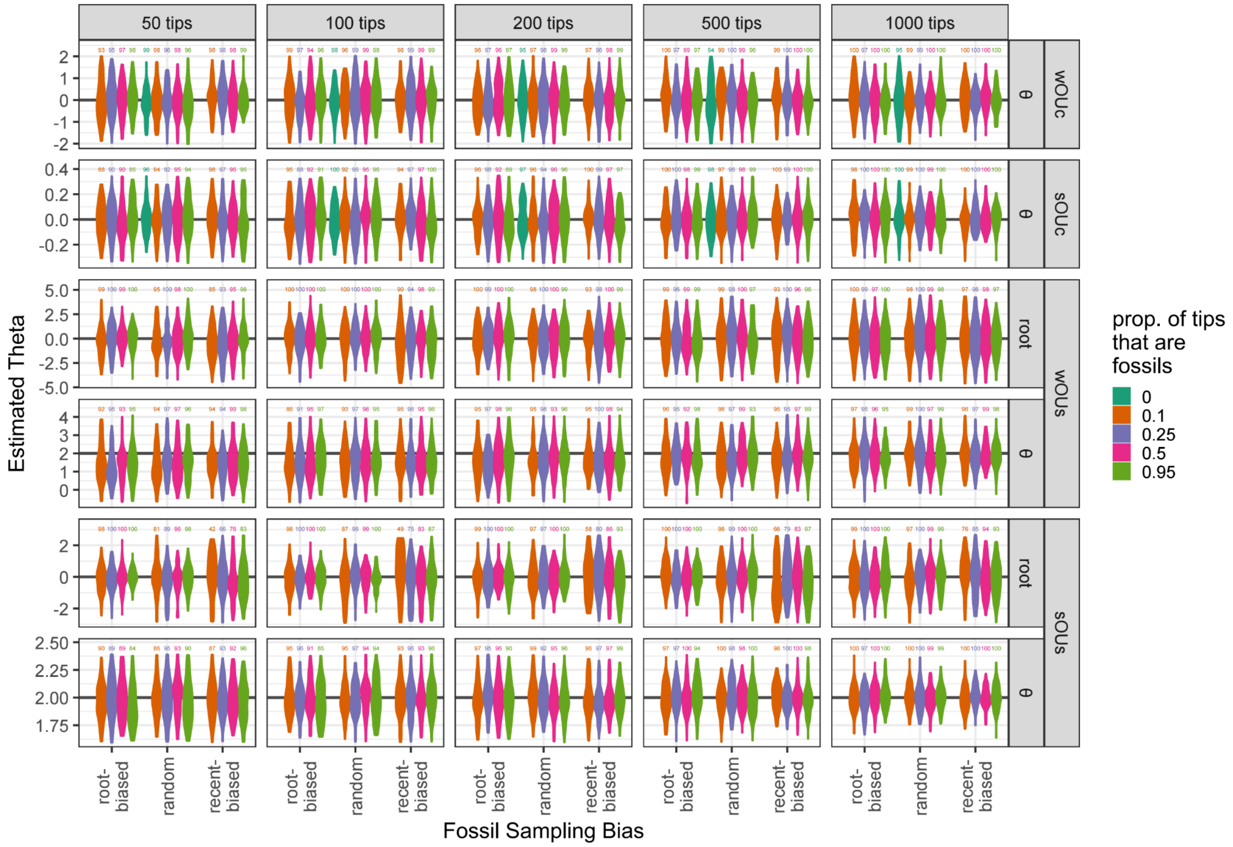


Figure 9: The distributions of the estimates for the θ parameter when an OU model was simulated on a phylogeny and then an OU model was fit to that simulated data. The true (simulated) θ parameter is represented with a solid horizontal line. Outliers have been removed separately for each row, and the numbers above violin plots represent the number of non-outliers that are represented by each violin plot. The results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). Note that y-axis limits vary from row to row, and analyses without fossils are plotted as part of the random distribution. These results are only for analyses where $\mu = 0.9$, the results for analyses where $\mu = 0.5$ are in Fig. S15.

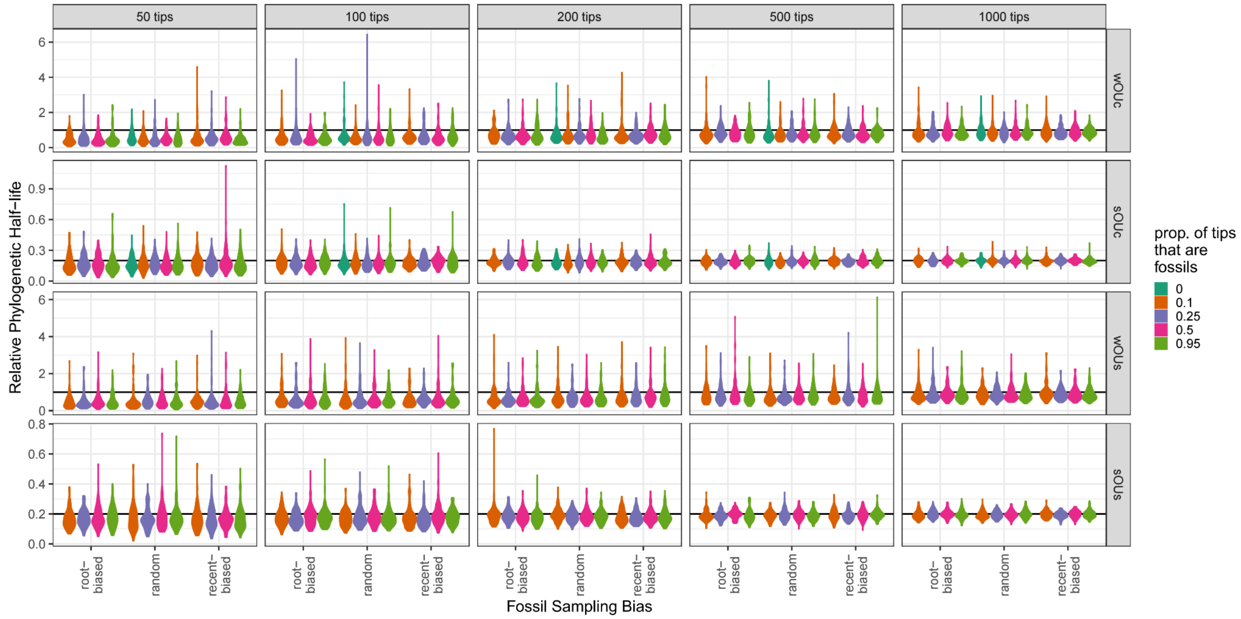


Figure 10: The distributions of the estimates for the relative half-life when an OU model was simulated on a phylogeny and then an OU model was fit to that simulated data. Relative half-life is calculated as $\frac{log(2)\ /\ \alpha}{tree\ height}$. The first and second rows represent centered OU models simulated with weak (1) and strong (0.1) relative half-lives, respectively. The third and fourth rows represent shifting OU models simulated with weak (1) and strong (0.2) relative half-lives, respectively. The true (simulated) relative half-life is represented with a solid horizontal line, and the results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). Note that y-axis limits vary from row to row, and analyses without fossils are plotted as part of the random distribution. These results are only for analyses where μ = 0.9, the results for analyses where μ = 0.5 are in Fig. S16.

Generally speaking, estimation accuracy of the beta parameter (β) is good for accelerating or decelerating simulations (Figs. 11 and S17). Accuracy and precision of β increases with phylogeny size in most cases. The inclusion of fossils also appears to result in a large increase in accuracy, especially under the higher relative death rate; however, the proportion of fossil tips and the ages of the fossils do not seem to matter. The one exception is for the strong AC model under the higher death rate, where increasing phylogeny size often results in estimating values of β at or near zero. For simulations with phylogenies with at least 100 tips, we see a bimodal distribution of parameter estimates, with one mode centered on the simulated value and one mode centered on 0. As phylogeny size increases, estimates of this parameter trend away from the simulated value of 0.5 and towards the alternative mode centered around 0.

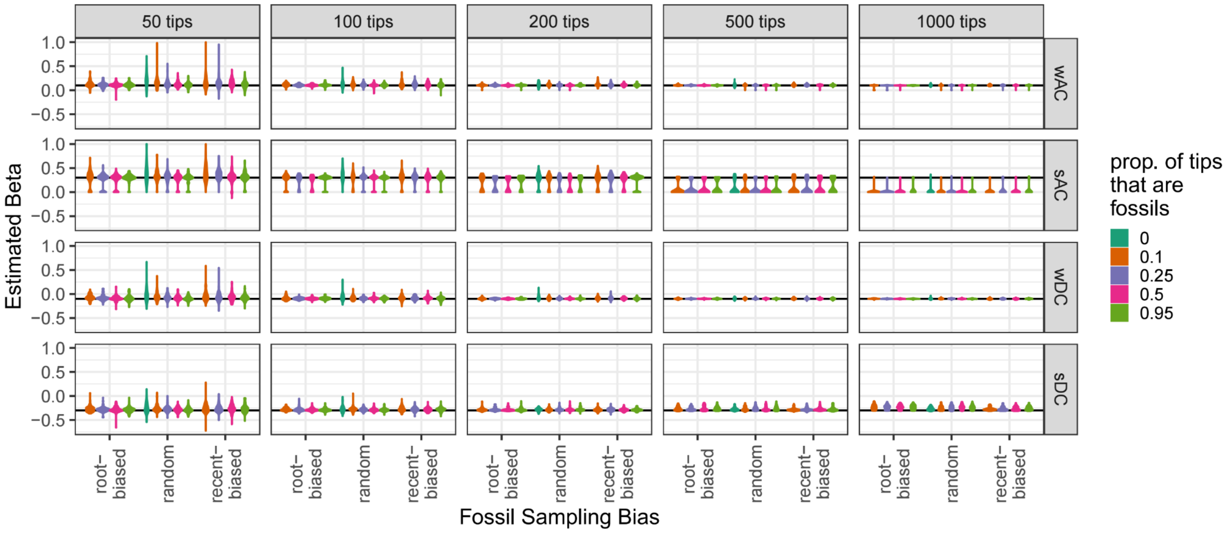


Figure 11: The distributions of the estimates for the *β* parameter when an ACDC model was simulated on a phylogeny and then an ACDC model was fit to that simulated data. The true (simulated) *β* parameter is represented with a solid horizontal line. The results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). Analyses without fossils are plotted as part of the random distribution. These results are only for analyses where $\mu = 0.9$, the results for analyses where $\mu = 0.5$ are in Fig. S17.

Discussion

The ability to accurately and precisely infer how continuous characters have evolved over time is crucial for testing hypotheses in evolutionary biology (Slater 2013; Benson et al. 2014; Gearty et al. 2018; Cornwallis and Griffin 2024). Our results echo past research (Slater et al. 2012; Mitchell 2015; Tejada et al. 2024; Mulvey et al. 2025), showing that the inclusion of fossil tips in phylogenies can be critical for inferring both the correct tempo and mode of continuous trait evolution. On average, including any fossil taxa in our simulations dramatically increased statistical power (Fig. 3). Generally speaking, the first 10% of fossil tips lead to the largest increase in accuracy, with diminishing returns beyond this proportion. This result suggests that adding a handful of fossils to an extant-only phylogeny can dramatically improve the reliability of modelling continuous character evolution.

Beyond the mere inclusion of fossil taxa in such analyses, we demonstrate that the relative age of these fossil taxa is fundamental to this increase in accuracy. When relatively young (recent-biased) fossil taxa are included, especially in smaller trees, the degree to which accuracy improves can be marginal (Figs. 3 and 4), indicating that young fossil taxa provide about as much statistical information as extant taxa. However, when relatively old (root-biased) or evenly distributed (random) fossil taxa are included, the accuracy improves dramatically (Figs. 3 and 4). Root-biased fossils provide insight into older parts of evolutionary history which are harder to infer from extant tips alone; it is therefore reasonable to expect that including this information would drastically improve our ability to recover the true (simulated) tempo and mode of evolution (Woolley et al. 2022). Notably, the improvement observed when fossil taxa are evenly distributed (random) is particularly informative given that the precise phylogenetic placement of many fossils is inherently uncertain, suggesting that this scenario may more closely reflect the structure of most empirical datasets.

Modes of evolution with an expected value (i.e., mean) that is heterogeneous through time, such as trend and OUs models (Fig. 4), are two of the most common models used in PCMs to test for adaptive trait evolution. They are hypothesized to represent scenarios where either (a) a clade evolves from some suboptimal ancestral state towards a different, more optimal, state, or (b) a clade consistently residing at its optimality peak exhibits trait value shift in line with shifts in the peak (Pagel 2002; Butler and King 2004; Cooper et al. 2016; Gill et al. 2017; Gearty et al. 2018; Godoy et al. 2019; Chiarenza et al. 2021, 2024; Farina et al. 2023; Cornwallis and Griffin 2024). These models cannot be tested when fossils are not included in the phylogeny (e.g., Slater et al. 2012; Clavel et al. 2015). Here, we show that adding root-biased fossils provides the most notable increase in our ability to successfully fit these models. Furthermore, when a split OU or trend model is the best-fitting model, and root-biased fossils have been included in the phylogeny, this is rarely an error (Fig. 6). Therefore, it is extremely encouraging that positive results for the split OU model and trend BM model generally can be assumed to be trustworthy.

The inclusion of fossils has mixed results for some other simulated trajectories of evolution. For example, weak OUc and OUs modes are generally very difficult to correctly infer regardless of the parameters we tested (Figs. 4-6). Including fossils does improve inference accuracy, but these modes are unique in that randomly distributed fossils provide the greatest increase in accuracy, while root-biased fossils provide the smallest increase in accuracy (Figs. 4 and S4). This does not appear to be related to either particular mode, since strong versions of these modes do not behave this way. Instead, this appears to be related to the weak strength of the α parameter. In these cases, the phylogenetic half-life is simulated at 1 tree height, which is fairly weak selection (in the case of OUs) or stabilization (in the case of OUc). Most analyses usually misinterpret these evolutionary histories as BM or trend, and, as the proportion of root-biased fossils and the total size of the phylogeny are increased, these simulations are increasingly misinterpreted as trends (Fig. 6). The misinterpretation as BM is expected because OU models with longer phylogenetic half-lives approach a BM process (Uyeda and Harmon 2014; Cornuault 2022; Bartoszek et al. 2023). It is also understandable that the wOUs model is misinterpreted as a trend model, especially when it was simulated with such a long phylogenetic half-life; OUs models with low selection behave as trend-like with gradual changes in the mean over time (Benson et al. 2018). However, this is extremely concerning for wOUc, as this mode represents evolutionary stasis, which is usually interpreted in a fundamentally different manner to a trend model. Further, increasing phylogeny size, and therefore taxonomic coverage, is often a major goal in phylogenetics (Zwickl and Hillis 2002). The fact that this potential for misinterpretation increases with increasing phylogeny size is counterintuitive, and there is great need for further investigation into this pattern in the future, particularly under a variety of OU simulation parameter values.

Additionally, we found mixed behavior in the ACDC family of models (Martin et al. 2023). The addition of fossils, especially root-biased fossils, dramatically improves the accuracy of inferring models of decelerating trait evolution, regardless of tempo (DC; Figs. 4-6). However, the inclusion of fossils almost always has a negative impact on the inference accuracy of rapidly accelerating models (sAC; Figs. 4-6). For weak accelerating models (wAC), random or recent-biased fossils have a similar effect. Further, while increases in phylogeny size usually have a positive influence on inference accuracy for other models, larger phylogenies lead to increased misinterpretation of the sAC model as OUc, even when fossils are included (Fig. 6). It has been shown that an AC model is indistinguishable from an OU model when analyzing ultrametric trees (Uyeda et al. 2015). Slater et al. (2012) found that distinguishing between an AC and an OUc-like model ("SSP") was easier when fossils were included, but we do not find that to be the case here when relative death rate is high. This may be due to slight differences between our OUc model and the OUc-like model of Slater et al. (2012) or the fact that Slater et al. (2012) did not consider trees with more than 377 tips. Regardless, under these parameters, the OUc model is apparently easily misspecified for a trait that has been simulated under the AC model regardless of fossil inclusion, especially with increasing phylogeny size (Fig. 6). It should be noted that these strange behaviors do not occur for AC models simulated with low relative death rates (Fig. S12), but this remains an unlikely scenario across the majority of Earth's history (Marshall 2017). Overall, we echo Slater et al. (2012) by strongly recommending caution when trying to fit AC models without fossils, particularly when using large, ultrametric phylogenies.

The inference of some particular evolutionary histories is not improved by the inclusion of fossils, regardless of their number or age distribution. Brownian motion is unique in that the model accuracy is relatively consistent regardless of fossil age, fossil abundance, or even phylogeny size (Figs. 4-6). This suggests that the information that can be gleaned from individual tips to reconstruct a character history evolving under Brownian motion is quickly saturated. Additional tips and non-extant tips provide little to no benefit once this saturation point (<50 tips in our simulations) is reached. In fact, the inclusion of fossil tips appears to cause a marginal decrease across the board in the ability to infer Brownian motion as the correct model. It appears that including these fossils results in favor of the trend model (Fig. 6). We intuit that this occurs when the distribution of the simulated values of the included fossils is, purely by chance, significantly different from the distribution of simulated values for the extant tips, providing support for a changing mean through time. Regardless, inference accuracy of histories simulated under Brownian motion remains relatively high. This would be encouraging; however, the error rate when Brownian motion is recovered as the best-fitting model also is relatively high (Fig. 6). The inclusion of fossils in a phylogeny reduces this error rate, especially for larger trees. Therefore, we argue that the marginal decrease in overall inference accuracy due to the inclusion of fossils is well outweighed by this reduction in error rate. However, caution should be taken when Brownian motion is inferred to be the best-fitting model in an empirical analysis of a small phylogeny, even when fossils are included. Finally, our results indicate lower overall accuracy (~75%) for inferring BM histories than those of previous studies (Silvestro et al. 2015). This suggests that the method of model comparison (AIC vs. likelihood ratio test) and the total number and type of tested models are highly influential when conducting both simulation and empirical studies and reaffirms that AIC, despite its wide use, could be sensitive to model overfitting (Bartoszek et al. 2023; Cornwallis and Griffin 2024).

For parameter estimation, the size of the phylogeny appears to be the most important factor for determining the average amount of error across simulations, with increasing phylogeny size leading to decreased degrees of overall error (Figs. 7-11, S13-17). This is especially apparent for the σ parameter of the BM model (Figs. 7 and S13). For this parameter, the inclusion of fossils has no discernible effect on accuracy. On the other hand, increasing phylogeny size appears to have an overall negative effect on the accuracy of β for the strong AC model with a high relative death rate (Figs. 11). In the case of the trend model, the *trend* parameter is estimated fairly accurately, with a decrease in average estimate error with increasing proportion of fossils and/or phylogeny size. However, the average estimate error increases when recent-biased fossils are used. For weak OU simulations, many of the estimates of the half-life parameter are estimated to be closer to zero, regardless of the inclusion of fossils or the size of the phylogeny. This likely explains why these weak OU models are often not selected as the best-fitting models when they are the simulating models (Fig. 4). Regardless, given that these parameter estimates are often used to estimate the strength of selection on different traits or for hypothesis testing (Grabowski et al. 2023), this directional misestimation of the strength of selection is concerning, and caution should be taken when directly interpreting these parameter estimates.

Finally, we accounted for a wide variety of sources of variability in evolutionary histories and the use of phylogenetic comparative methods in our simulations, including the size of the phylogeny, the relative death rate, the proportion of fossils included, the relative ages of the fossils, and the tempo and mode of character evolution. However, evolution and the implementation of PCMs are both highly complex and we acknowledge that there are various other sources of variability and error that we did not account for. First, we used a standard birth-death procedure with high and moderate relative death rates to simulate phylogenies with roughly similar tree shape. However, empirical phylogenies indicate that there is a large degree of diversification rate heterogeneity among lineages which results in a wide variety of tree shapes, both in terms of the distributions of branch lengths and tree balance (Mooers and Heard 1997; Martins and Housworth 2002; Boettiger et al. 2012). Second, we assumed that the topology and branch lengths of the simulated phylogeny were completely known. However, in empirical studies this is rarely the case, and inaccuracies related to these two sources are known to cause issues with other comparative methods (Purvis et al. 1994; Symonds 2002). In empirical studies, particularly those using the results of Bayesian phylogenetic analyses or gene tree analyses, it is becoming more common to perform analyses over a sample of phylogenies to account for uncertainty in branch length and topology (Gearty et al. 2018; Godoy et al. 2019;

Gearty and Payne 2020; Grossnickle 2020; Soul and Wright 2021; Farina et al. 2023; Hibbins et al. 2023). Third, and relatedly, fossils are notoriously incomplete and are often associated with more topological uncertainty than extant taxa (Sansom and Wills 2013; Pattinson et al. 2015; Barido-Sottani et al. 2020b; Mongiardino Koch et al. 2023; Tejada et al. 2024). For example, fossils may actually be samples from other lineages in the tree, fragmentary fossils may be missing key traits, and fossil taxa may represent a derived condition that differs from other extinct and surviving relatives. All of these factors could result in an inaccurate and/or uncertain placement of a fossil in the phylogeny. Therefore, an increase in the proportion of fossils in a phylogenetic inference would therefore also likely increase the degree of uncertainty in the topology. Future work should investigate whether this source of error is outweighed by the benefits of adding fossils which are described above (O'Reilly and Donoghue 2020; Wright and Hopkins 2025). Finally, we assumed that character states for all taxa were known without error (Smith et al. 2023). However, empirical studies have measurement error due to the intraspecific variability of natural populations and from instrumental/human imprecision, and these sources of error are known to introduce biases in the reconstruction of trait evolution (Ives et al. 2007; Felsenstein et al. 2008; Silvestro et al. 2015). Accounting for all of these sources of variability and error was outside the scope of this project, but future simulation studies should address the interplay of the results presented herein and these other factors.

We acknowledge that researchers working exclusively with sequence data may be concerned about the time and effort required to incorporate fossils into their phylogenies as we have recommended. Historical conventions required that all of the taxa were coded for a suite of morphological characters, which can take much more time than is required to sequence the extant taxa. However, recent work has shown that taxonomic constraints and/or backbones can be just as reliable as morphological matrices when integrating fossils into broader phylogenetic contexts (Barido-Sottani et al. 2023; Mulvey et al. 2025; Heckeberg et al. 2026) and conducting phylogenetic comparative methods (Soul and Friedman 2015). Furthermore, our results indicate that only a small number of fossils (e.g., 10% of tips) is required to reap their benefits. Ultimately, including this few fossils may have the same statistical effect as would sequencing and adding thousands of additional extant taxa but for far less work.

All-in-all, we find the results of this simulation study to be encouraging. Although we focus on the instances when simulated and fitted trait evolution models do not match, our intention is to equip readers to make informed decisions about how to plan and carry out their entire research pipelines. This includes, but is not limited to, specimen identification, phylogenetic matrix assembly, phylogenetic inference, phylogenetic comparative analysis, and result interpretation. While there is still much more work to be done in understanding trait data and evolutionary models, we hope that our simulations can be used to highlight clear avenues for future research.

CONCLUSIONS

The development and use of phylogenetic comparative methods in evolutionary biology and paleobiology has blossomed since the work of Felsenstein (Felsenstein 1985). Although the amount of paleontological information available to biologists has increased over this time, the use of PCMs without fossils has remained abundant in the literature. Here we reiterate that including fossils as tips in phylogenetic analyses results in superior downstream reconstructions of the evolution of continuous characters within a phylogenetic context. More importantly, fossils farther in time from the extant tips potentially hold much more information than their extant-adjacent counterparts and can play a profound role in our ability to accurately infer evolutionary tempo and mode. The models that are used to reconstruct these evolutionary histories have several shortcomings, but the inclusion of fossils helps alleviate many of these. Given all of this, we strongly assert that the vast majority of phylogenetic comparative analyses should include fossil tips moving forward, except for the minority of clades where fossils are not (yet) known. This will lead to a more robust field of study and more reliable biological interpretations. While the inclusion of fossils may present a hurdle to some evolutionary biologists, we argue that the extra work is well worth it: a fossil taxon that is included within a phylogeny is, at worst, just as useful as an extant taxon for model inference, and is, at best, far more useful. Ultimately, collaboration between neontologists and paleontologists provides the surest foundation for maximizing the potential of both genetic and fossil data (Hunt and Slater 2016; Smith et al. 2025).

ACKNOWLEDGMENTS

W.G. was supported by the Lerner-Gray Postdoctoral Research Fellowship from the Richard Gilder Graduate School at the American Museum of Natural History, a Norman Newell Early Career Grant from the Paleontological Society, and as part of a grant from the Alfred P. Sloan Foundation (#G2023-20946). B.J.A was supported by ETH Zurich, and by Deutsche Forschungsgemeinschaft Collaborative Research Centre 1211 (grant no. 268236062). P.L.G. was supported by the São Paulo Research Foundation (FAPESP, grant numbers 2022/05697-9 and 2024/09825-7). A.A.C. was supported by a European Research Council (ERC) starting grant under the European Union's Horizon 2020 Research and Innovation Programme (grant no. 947921), a Juan de la Cierva-formación 2020 fellowship (FJC2020-044836-I) funded by the Ministry of Science and Innovation from the European Union Next Generation EU/PRTR, and a Royal Society Newton International Fellowship (NIF\R1\231802).

DATA ACCESSIBILITY STATEMENT

All code is archived on GitHub at https://github.com/willgearty/PCM_sensitivity and on Zenodo (https://doi.org/10.5281/zenodo.13361631, Gearty et al. 2026). There are no data associated with this manuscript.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control. 19:716–723.

Allen B.J., Stubbs T.L., Benton M.J., Puttick M.N. 2019. Archosauromorph extinction selectivity during the Triassic–Jurassic mass extinction. Palaeontology. 62:211–224.

Almeida F.C., Helgen K.M., Simmons N.B., Giannini N.P. 2025. Evolution and ecology of body size in the world's largest bats. Proc. R. Soc. B Biol. Sci. 292:20250743.

Álvarez-Carretero S., Tamuri A.U., Battini M., Nascimento F.F., Carlisle E., Asher R.J., Yang Z., Donoghue P.C.J., dos Reis M. 2022. A species-level timeline of mammal evolution integrating phylogenomic data. Nature. 602:263–267.

Barido-Sottani J., Pohle A., De Baets K., Murdock D., Warnock R.C.M. 2023. Putting the F into FBD analysis: tree constraints or morphological data? Palaeontology. 66:e12679.

Barido-Sottani J., Saupe E.E., Smiley T.M., Soul L.C., Wright A.M., Warnock R.C.M. 2020a. Seven rules for simulations in paleobiology. Paleobiology. 46:435–444.

Barido-Sottani J., van Tiel N.M.A., Hopkins M.J., Wright D.F., Stadler T., Warnock R.C.M. 2020b. Ignoring Fossil Age Uncertainty Leads to Inaccurate Topology and Divergence Time Estimates in Time Calibrated Tree Inference. Front. Ecol. Evol. 8.

Bartoszek K., Fuentes-González J., Mitov V., Pienaar J., Piwczyński M., Puchałka R., Spalik K., Voje K.L. 2023. Model Selection Performance in Phylogenetic Comparative Methods Under Multivariate Ornstein–Uhlenbeck Models of Trait Evolution. Syst. Biol. 72:275–293.

Beaulieu J.M., Jhwueng D.-C., Boettiger C., O'Meara B.C. 2012. Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. Evol. Int. J. Org. Evol. 66:2369–83.

Beaulieu J.M., O'Meara B.C. 2023. Fossils Do Not Substantially Improve, and May Even Harm, Estimates of Diversification Rate Heterogeneity. Syst. Biol. 72:50–61.

Bengtsson H. 2021. A Unifying Framework for Parallel and Distributed Processing in R using Futures. R J. 13:208–227.

Benson R.B.J., Frigot R.A., Goswami A., Andres B., Butler R.J. 2014. Competition and constraint drove Cope's rule in the evolution of giant flying reptiles. Nat. Commun. 5:3567.

Benson R.B.J., Hunt G., Carrano M.T., Campione N. 2018. Cope's rule and the adaptive landscape of dinosaur body size evolution. Palaeontology. 61:13–48.

Blomberg S.P., Garland JR. T., Ives A.R. 2003. Testing for Phylogenetic Signal in Comparative Data: Behavioral Traits Are More Labile. Evolution. 57:717–745.

Boettiger C., Coop G., Ralph P. 2012. IS YOUR PHYLOGENY INFORMATIVE? MEASURING THE POWER OF COMPARATIVE METHODS. Evolution. 66:2240–2251.

Brand T. van den. 2024. ggh4x: Hacks for "ggplot2." .

Burnham K.P., Anderson D.R. 2002. Model Selection and Multimodel Inference. New York, NY: Springer New York.

Butler M., King A. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am. Nat. 164:683–695.

Černý D., Madzia D., Slater G.J. 2022. Empirical and Methodological Challenges to the Model-Based Inference of Diversification Rates in Extinct Clades. Syst. Biol. 71:153–171.

Chiarenza A.A., Cantalapiedra J.L., Jones L.A., Gamboa S., Galván S., Farnsworth A.J., Valdes P.J., Sotelo G., Varela S. 2024. Early Jurassic origin of avian endothermy and thermophysiological diversity in dinosaurs. Curr. Biol. 34:2517-2527.e4.

Chiarenza A.A., Fabbri M., Consorti L., Muscioni M., Evans D.C., Cantalapiedra J.L., Fanti F. 2021. An Italian dinosaur Lagerstätte reveals the tempo and mode of hadrosauriform body size evolution. Sci. Rep. 11:23295.

Clavel J., Escarguel G., Merceron G. 2015. mvMORPH: An R package for fitting multivariate evolutionary models to morphometric data. Methods Ecol. Evol. 6:1311–1319.

Cooper N., Thomas G.H., Venditti C., Meade A., Freckleton R.P. 2016. A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. Biol. J. Linn. Soc. 118:64–77.

Cornuault J. 2022. Bayesian Analyses of Comparative Data with the Ornstein–Uhlenbeck Model: Potential Pitfalls. Syst. Biol. 71:1524–1540.

Cornwallis C.K., Griffin A.S. 2024. A Guided Tour of Phylogenetic Comparative Methods for Studying Trait Evolution. Annu. Rev. Ecol. Evol. Syst. 55:181–204.

Cornwell W., Nakagawa S. 2017. Phylogenetic comparative methods. Curr. Biol. 27:R333–R336.

Farina B.M., Godoy P.L., Benson R.B.J., Langer M.C., Ferreira G.S. 2023. Turtle body size evolution is determined by lineage-specific specializations rather than global trends. Ecol. Evol. 13:e10201.

Faurby S., Silvestro D., Werdelin L., Antonelli A. 2024. Reliable biogeography requires fossils: insights from a new species-level phylogeny of extinct and living carnivores. Proc. R. Soc. B Biol. Sci. 291:20240473.

Felsenstein J. 1985. Phylogenies and the Comparative Method. Am. Nat. 125:1–15.

Felsenstein J., Otto A.E.S.P., Whitlock E.M.C. 2008. Comparative Methods with Sampling Error and Within-Species Variation: Contrasts Revisited and Revised. Am. Nat. 171:713–725.

Finarelli J.A., Flynn J.J. 2006. Ancestral State Reconstruction of Body Size in the Caniformia (Carnivora, Mammalia): The Effects of Incorporating Data from the Fossil Record. Syst. Biol. 55:301–313.

Finarelli J.A., Goswami A. 2013. Potential pitfalls of reconstructing deep time evolutionary history with only extant data, a case study using the Canidae (Mammalia, Carnivora). Evolution. 67:3678–3685.

Gauthier J. 1986. Saurischian monophyly and the origin of birds. Mem. Calif. Acad. Sci. 8:1--55.

Gearty W. 2023. pcmtools: Tools for Phylogenetic Comparative Methods. .

Gearty W. 2024. deeptime: Plotting Tools for Anyone Working in Deep Time. .

Gearty W., Allen B., Godoy P.L., Chiarenza A. 2026. willgearty/PCM_sensitivity: Manuscript resubmission. .

Gearty W., Carrillo E., Payne J.L. 2021. Ecological Filtering and Exaptation in the Evolution of Marine Snakes. https://doi.org/10.1086/716015. 198:506–521.

Gearty W., McClain C.R., Payne J.L. 2018. Energetic tradeoffs control the size distribution of aquatic mammals. Proc. Natl. Acad. Sci. 115:4194–4199.

Gearty W., Payne J.L. 2020. Physiological constraints on body size distributions in Crocodyliformes. Evolution. 74:245–255.

Gill M.S., Tung Ho L.S., Baele G., Lemey P., Suchard M.A. 2017. A Relaxed Directional Random Walk Model for Phylogenetic Trait Evolution. Syst. Biol. 66:299–319.

Godoy P.L., Benson R.B.J., Bronzati M., Butler R.J. 2019. The multi-peak adaptive landscape of crocodylomorph body size evolution. BMC Evol. Biol. 19:167.

Grabowski M., Pienaar J., Voje K.L., Andersson S., Fuentes-González J., Kopperud B.T., Moen D.S., Tsuboi M., Uyeda J., Hansen T.F. 2023. A cautionary note on “A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies.” Syst. Biol. 72:955–963.

Grossnickle D.M. 2020. Feeding ecology has a stronger evolutionary influence on functional morphology than on body mass in mammals. Evolution. 74:610–628.

Grossnickle D.M., Sadier A., Patterson E., Cortés-Viruet N.N., Jiménez-Rivera S.M., Sears K.E., Santana S.E. 2024. The hierarchical radiation of phyllostomid bats as revealed by adaptive molar morphology. Curr. Biol. 34:1284-1294.e3.

Hansen T.F. 1997. Stabilizing Selection and the Comparative Analysis of Adaptation. Evolution. 51:1341.

Hansen T.F., Pienaar J., Orzack S.H. 2008. A comparative method for studying adaptation to a randomly evolving environment. Evolution. 62:1965–1977.

Harmon L.J. 2019. Phylogenetic Comparative Methods: Learning From Trees. EcoEvoRxiv.

Harmon L.J., Losos J.B., Jonathan Davies T., Gillespie R.G., Gittleman J.L., Bryan Jennings W., Kozak K.H., McPeek M.A., Moreno-Roark F., Near T.J., Purvis A., Ricklefs R.E., Schluter D., Schulte II J.A., Seehausen O., Sidlauskas B.L., Torres-Carvajal O., Weir J.T., Mooers A.Ø. 2010. Early Bursts of Body Size and Shape Evolution Are Rare in Comparative Data. Evolution. 64:2385–2396.

Heckeberg N.S., Capobianco A., Khakurel B., Darlim G., Höhna S. 2026. Practical Guide and Review of Fossil Tip-Dating in Phylogenetics. Syst. Biol. 75:156–192.

Hibbins M.S., Breithaupt L.C., Hahn M.W. 2023. Phylogenomic comparative methods: Accurate evolutionary inferences in the presence of gene tree discordance. Proc. Natl. Acad. Sci. 120:e2220389120.

Ho L.S.T., Ané C. 2014. Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. Methods Ecol. Evol. 5:1133–1146.

Hunt G., Carrano M.T. 2010. Models and Methods for Analyzing Phenotypic Evolution in Lineages and Clades. Paleontol. Soc. Pap. 16:245–269.

Hunt G., Slater G. 2016. Integrating Paleontological and Phylogenetic Approaches to Macroevolution. Annu. Rev. Ecol. Evol. Syst. 47:189–213.

Ives A.R., Midford P.E., Garland T. 2007. Within-Species Variation and Measurement Error in Phylogenetic Comparative Methods. Syst. Biol. 56:252–270.

Jetz W., Pyron R.A. 2018. The interplay of past diversification and evolutionary isolation with present imperilment across the amphibian tree of life. Nat. Ecol. Evol. 2:850–858.

Jetz W., Thomas G.H., Joy J.B., Hartmann K., Mooers A.O. 2012. The global diversity of birds in space and time. Nature. 491:444–448.

Khabbazian M., Kriebel R., Rohe K., Ané C. 2016. Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models. Methods Ecol. Evol. 7:811–824.

Louca S., McLaughlin A., MacPherson A., Joy J.B., Pennell M.W. 2021. Fundamental Identifiability Limits in Molecular Epidemiology. Mol. Biol. Evol. 38:4010–4024.

Louca S., Pennell M.W. 2020. Extant timetrees are consistent with a myriad of diversification histories. Nature. 580:502–505.

Magallón S., Gómez-Acevedo S., Sánchez-Reyes L.L., Hernández-Hernández T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytol. 207:437–453.

Marshall C.R. 2017. Five palaeobiological laws needed to understand the evolution of the living biota. Nat. Ecol. Evol. 1:1–6.

Martin B.S., Bradburd G.S., Harmon L.J., Weber M.G. 2023. Modeling the Evolution of Rates of Continuous Trait Evolution. Syst. Biol. 72:590–605.

Martins E.P., Housworth E.A. 2002. Phylogeny Shape and the Phylogenetic Comparative Method. Syst. Biol. 51:873–880.

Mitchell J.S. 2015. Extant-only comparative methods fail to recover the disparity preserved in the bird fossil record. Evolution. 69:2414–2424.

Mongiardino Koch N., Garwood R.J., Parry L.A. 2021. Fossils improve phylogenetic analyses of morphological characters. Proc. R. Soc. B Biol. Sci. 288:20210044.

Mongiardino Koch N., Garwood R.J., Parry L.A. 2023. Inaccurate fossil placement does not compromise tip-dated divergence times. Palaeontology. 66:e12680.

Mongiardino Koch N., Parry L.A. 2020. Death is on Our Side: Paleontological Data Drastically Modify Phylogenetic Hypotheses. Syst. Biol. 69:1052–1067.

Mooers A.O., Heard S.B. 1997. Inferring Evolutionary Process from Phylogenetic Tree Shape. Q. Rev. Biol. 72:31–54.

Müller K., Wickham H. 2023. tibble: Simple Data Frames. .

Mulvey L.P.A., Nikolic M.C., Allen B.J., Heath T.A., Warnock R.C.M. 2025. From fossils to phylogenies: exploring the integration of paleontological data into Bayesian phylogenetic inference. Paleobiology. 51:214–236.

Oakley T.H., Cunningham C.W. 2000. INDEPENDENT CONTRASTS SUCCEED WHERE ANCESTOR RECONSTRUCTION FAILS IN A KNOWN BACTERIOPHAGE PHYLOGENY. Evolution. 54:397–405.

O’Meara B.C., Ané C., Sanderson M.J., Wainwright P.C. 2006. TESTING FOR DIFFERENT RATES OF CONTINUOUS TRAIT EVOLUTION USING LIKELIHOOD. Evolution. 60:922.

O'Reilly J.E., Donoghue P.C.J. 2020. The Effect of Fossil Sampling on the Estimation of Divergence Times with the Fossilized Birth–Death Process. Syst. Biol. 69:124–138.

Pagel M. 2002. Modelling the evolution of continuously varying characters on phylogenetic trees: the case of Hominid cranial capacity. Morphology, Shape and Phylogeny. CRC Press.

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.

Pattinson D.J., Thompson R.S., Piotrowski A.K., Asher R.J. 2015. Phylogeny, Paleontology, and Primates: Do Incomplete Fossils Bias the Tree of Life? Syst. Biol. 64:169–186.

Pennell M.W., FitzJohn R.G., Cornwell W.K., Harmon L.J. 2015. Model Adequacy and the Macroevolution of Angiosperm Functional Traits. Am. Nat. 186:E33–E50.

Pie M.R., Divieso R., Caron F.S. 2023. Clade density and the evolution of diversity-dependent diversification. Nat. Commun. 14:4576.

Purvis A., Gittleman J.L., Luh H.-K. 1994. Truth or Consequences: Effects of Phylogenetic Accuracy on Two Comparative Methods. J. Theor. Biol. 167:293–300.

Puttick M.N. 2016. Partially incorrect fossil data augment analyses of discrete trait evolution in living species. Biol. Lett. 12:20160392.

Puttick M.N. 2018. Mixed evidence for early bursts of morphological evolution in extant clades. J. Evol. Biol. 31:502–515.

Puttick M.N., Thomas G.H. 2015. Fossils and living taxa agree on patterns of body mass evolution: a case study with Afrotheria. Proc. R. Soc. B Biol. Sci. 282:20152023.

Quental T.B., Marshall C.R. 2010. Diversity dynamics: molecular phylogenies need the fossil record. Trends Ecol. Evol. 25:434–441.

R Core Team. 2024. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rabosky D.L. 2010. Extinction rates should not be estimated from molecular phylogenies. Evolution. 64:1816–1824.

Rabosky D.L., Chang J., Title P.O., Cowman P.F., Sallan L., Friedman M., Kaschner K., Garilao C., Near T.J., Coll M., Alfaro M.E. 2018. An inverse latitudinal gradient in speciation rate for marine fishes. Nature. 559:392–395.

Revell L.J. 2024. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). PeerJ. 12:e16505.

Ronquist F., Klopfstein S., Vilhelmsen L., Schulmeister S., Murray D.L., Rasnitsyn A.P. 2012. A Total-Evidence Approach to Dating with Fossils, Applied to the Early Radiation of the Hymenoptera. Syst. Biol. 61:973–999.

Royer-Carenzi M., Pontarotti P., Didier G. 2013. Choosing the best ancestral character state reconstruction method. Math. Biosci. 242:95–109.

Sansom R.S., Wills M.A. 2013. Fossilization causes organisms to appear erroneously primitive by distorting evolutionary trees. Sci. Rep. 3:2545.

Schrago C.G., Mello B., Soares A.E.R. 2013. Combining fossil and molecular data to date the diversification of New World Primates. J. Evol. Biol. 26:2438–2446.

Silvestro D., Kostikova A., Litsios G., Pearman P.B., Salamin N. 2015. Measurement errors should always be incorporated in phylogenetic comparative analysis. Methods Ecol. Evol. 6:340–346.

Simpson G.G. 1944. Tempo and mode in evolution. New York: Columbia University Press.

Slater G.J. 2013. Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the Cretaceous-Palaeogene boundary. Methods Ecol. Evol. 4:734–744.

Slater G.J., Harmon L.J., Alfaro M.E. 2012. Integrating fossils with molecular phylogenies improves inference of trait evolution. Evolution. 66:3931–3944.

Smith J.A., Dowding E.M., Abdelhady A.A., Abondio P., Araújo R., Aze T., Balisi M.A., Buatois L.A., Carvajal-Chitty H., Chattopadhyay D., Coiro M., Dietl G.P., Arango C.G., Kevrekidis C., Kimmig J., Mychajliw A.M., Pimiento C., Fernández O.R.R., Schroeder K.M., Warnock R.C.M., Yang T.-R., Yasuhara M., Akita L.G., Allen B.J., Anderson B.M., Andréoletti J., Archuby F.M., Ballen G.A., Bari M.I., Benton M.J., Bergh E.W., Brambilla L., Brombacher A., Chan Y.K.S., Chiarenza A.A., Chinzorig T., Coates K.M., Cordie D.R., Cortés-Sánchez M., Cruz-Vega E.J., Cybulski J.D., Baets K.D., Entrambasaguas J.D., Dillon E.M., Du A., Dunhill A.M., Erlandson J.M., Forel M.-B., Foster W.J., Gates T.A., Gavryushkina A., Grace M.K., Grossart H.-P., Hänsel P., Harnik P.G., Hopkins M.J., Hopkins S.S.B., Hu K., Huang H.-H.M., Irmis R.B., Jaques V.A.J., Jenkins X.A., Jukar A.M., Kelley P.H., Kihn R.G., Klompmaker A.A., Kocsis Á.T., Kriwet J., Lazarus D., Liao C.-C., Lin C.-H., Louys J., Lozano-Fernandez J., Lozano-Francisco M.C., Lueders-Dumont J.A., Malvé M.E., Martindale R.C., Mazzini I., Modenini G., Mondal S., Mondini M., Monferran M.D., Mulvey L.P.A., Nanglu K., Nguyen J.M.T., Norris R., O'Dea A., Ollendorf A.L., Orihuela J., Pandolfi J.M., Pereira

T., Piro A., Plotnick R.E., Plaza-Torres S.M., Porto A., Prieto-Márquez A., Punyasena S.W., Quental T.B., Raja N.B., Ranaivosoa V., Ribas-Deulofeu L., Rivals F., Roden V.J., Rosso A., Saleh F., Salvador R.B., Saupe E.E., Schneider S., Sclafani J.A., Smith M.R., Souron A., Steinbauer M.J., Stewart M., Tambussi C.P., Thomas E., Tschopp E., Tütken T., Varela S., Vezzosi R.I., Villaseñor A., Weinkauf M.F.G., Zanno L.E., Zhang C., Zhao Q., Kiessling W. 2025. Identifying the Big Questions in paleontology: a community-driven project. Paleobiology. 51:408–431.

Smith T.J., Sansom R.S., Pisani D., Donoghue P.C.J. 2023. Fossilization can mislead analyses of phenotypic disparity. Proc. R. Soc. B Biol. Sci. 290:20230522.

Smyčka J., Toszogyova A., Storch D. 2023. The relationship between geographic range size and rates of species diversification. Nat. Commun. 14:5559.

Solymos P., Zawadzki Z. 2023. pbapply: Adding Progress Bar to "*apply" Functions. .

Soul L.C., Friedman M. 2015. Taxonomy and Phylogeny Can Yield Comparable Results in Comparative Paleontological Analyses. Syst. Biol. 64:608–620.

Soul L.C., Wright D.F. 2021. Phylogenetic Comparative Methods: A User's Guide for Paleontologists. Elem. Paleontol.

Springer M.S., Teeling E.C., Madsen O., Stanhope M.J., de Jong W.W. 2001. Integrated fossil and molecular data reconstruct bat echolocation. Proc. Natl. Acad. Sci. 98:6241–6246.

Stadler T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. Proc. Natl. Acad. Sci. 108:6187–6192.

Stadler T. 2019. TreeSim: Simulating Phylogenetic Trees. .

Sternes P.C., Schmitz L., Higham T.E. 2024. The rise of pelagic sharks and adaptive evolution of pectoral fin morphology during the Cretaceous. Curr. Biol. 34:2764-2772.e3.

Sugiura N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections: Further analysis of the data by akaike' s. Commun. Stat. - Theory Methods. 7:13–26.

Symonds M.R.E. 2002. The Effects of Topological Inaccuracy in Evolutionary Trees on the Phylogenetic Comparative Method of Independent Contrasts. Syst. Biol. 51:541–553.

Tejada J.V., Antoine P.-O., Münch P., Billet G., Hautier L., Delsuc F., Condamine F.L. 2024. Bayesian Total-Evidence Dating Revisits Sloth Phylogeny and Biogeography: A Cautionary Tale on Morphological Clock Analyses. Syst. Biol. 73:125–139.

Troyer E.M., Betancur-R R., Hughes L.C., Westneat M., Carnevale G., White W.T., Pogonoski J.J., Tyler J.C., Baldwin C.C., Ortí G., Brinkworth A., Clavel J., Arcila D. 2022. The impact of paleoclimatic changes on body size evolution in marine fishes. Proc. Natl. Acad. Sci. 119:e2122486119.

Upham N.S., Esselstyn J.A., Jetz W. 2019. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. PLOS Biol. 17:e3000494.

Uyeda J.C., Caetano D.S., Pennell M.W. 2015. Comparative Analysis of Principal Components Can be Misleading. Syst. Biol. 64:677–689.

Uyeda J.C., Harmon L.J. 2014. A Novel Bayesian Method for Inferring and Interpreting the Dynamics of Adaptive Landscapes from Phylogenetic Comparative Data. Syst. Biol. 63:902–918.

Vrba E.S. 1979. Phylogenetic analysis and classification of fossil and recent Alcelaphini Mammalia: Bovidae. Biol. J. Linn. Soc. 11:207–228.

Warnock R., Barido-Sottani J., Pett W., O'Reilly J. 2022. FossilSim: Simulation of Fossil and Taxonomy Data. .

Warnock R.C.M., Heath T.A., Stadler T. 2020. Assessing the impact of incomplete species sampling on estimates of speciation and extinction rates. Paleobiology. 46:137–157.

Webster A.J., Purvis A. 2002. Testing the accuracy of methods for reconstructing ancestral states of continuous characters. Proc. R. Soc. Lond. B Biol. Sci. 269:143–149.

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. New York: Springer International Publishing.

Wickham H. 2023. forcats: Tools for Working with Categorical Variables (Factors). .

Wickham H., François R., Henry L., Müller K., Vaughan D. 2023. dplyr: A Grammar of Data Manipulation. .

Wickham H., Vaughan D., Girlich M. 2024. tidyr: Tidy Messy Data. .

Wisniewski A.L., Lloyd G.T., Slater G.J. 2022. Extant species fail to estimate ancestral geographical ranges at older nodes in primate phylogeny. Proc. R. Soc. B Biol. Sci. 289:20212535.

Woolley C.H., Thompson J.R., Wu Y.-H., Bottjer D.J., Smith N.D. 2022. A biased fossil record can preserve reliable phylogenetic signal. Paleobiology. 48:480–495.

Wright A.M., Bapst D.W., Barido-Sottani J., Warnock R.C.M. 2022. Integrating Fossil Observations Into Phylogenetics Using the Fossilized Birth–Death Model. Annu. Rev. Ecol. Evol. Syst. 53:251–273.

Wright D.F., Hopkins M.J. 2025. Assessing the impact of character evolution models on phylogenetic and macroevolutionary inferences from fossil data. Palaeontology. 68:e70031.

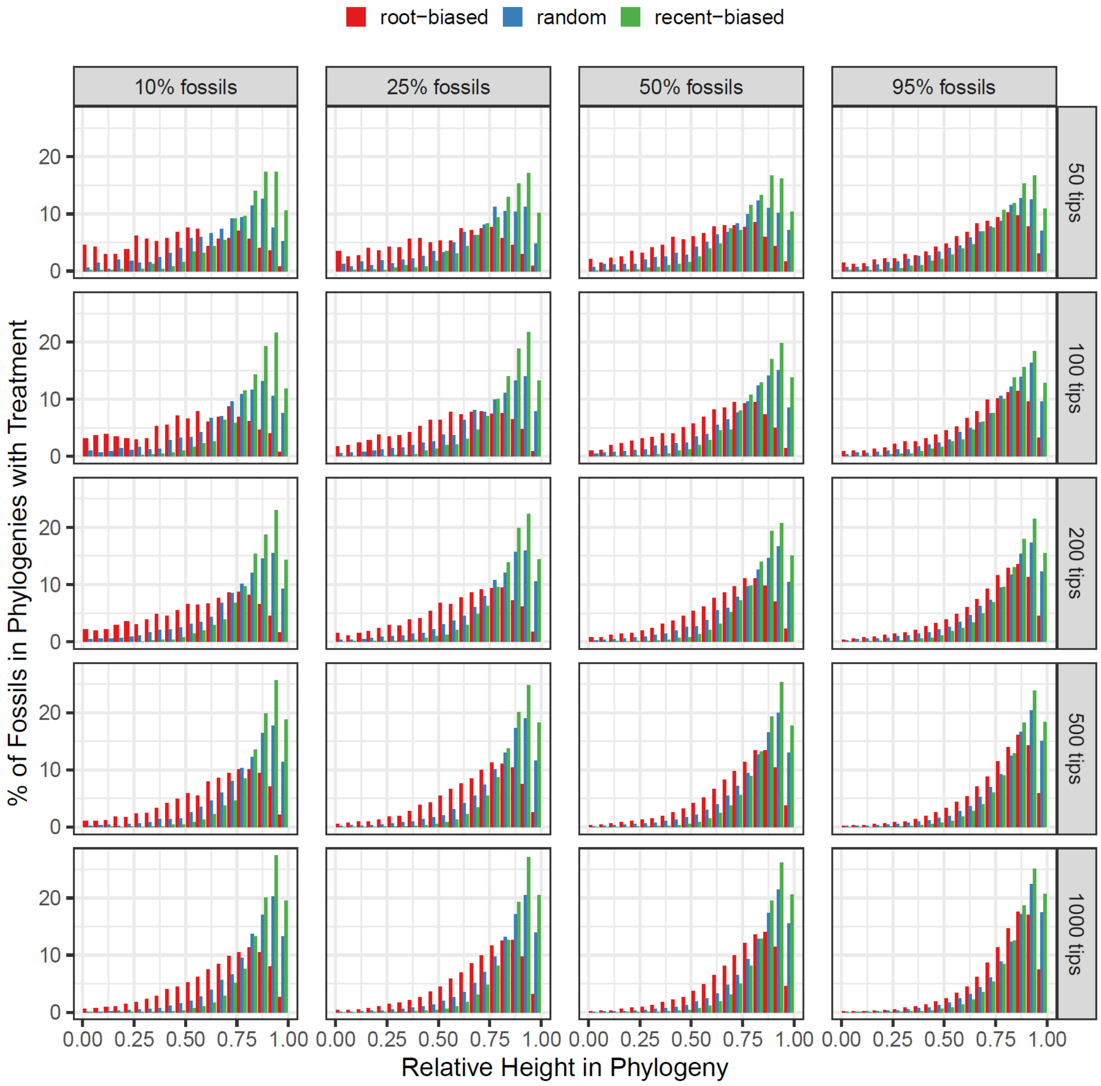
Zhang C., Ronquist F., Stadler T. 2023. Skyline Fossilized Birth–Death Model is Robust to Violations of Sampling Assumptions in Total-Evidence Dating. Syst. Biol. 72:1316–1336.

Zhang C., Stadler T., Klopfstein S., Heath T.A., Ronquist F. 2016. Total-Evidence Dating under the Fossilized Birth–Death Process. Syst. Biol. 65:228–249.

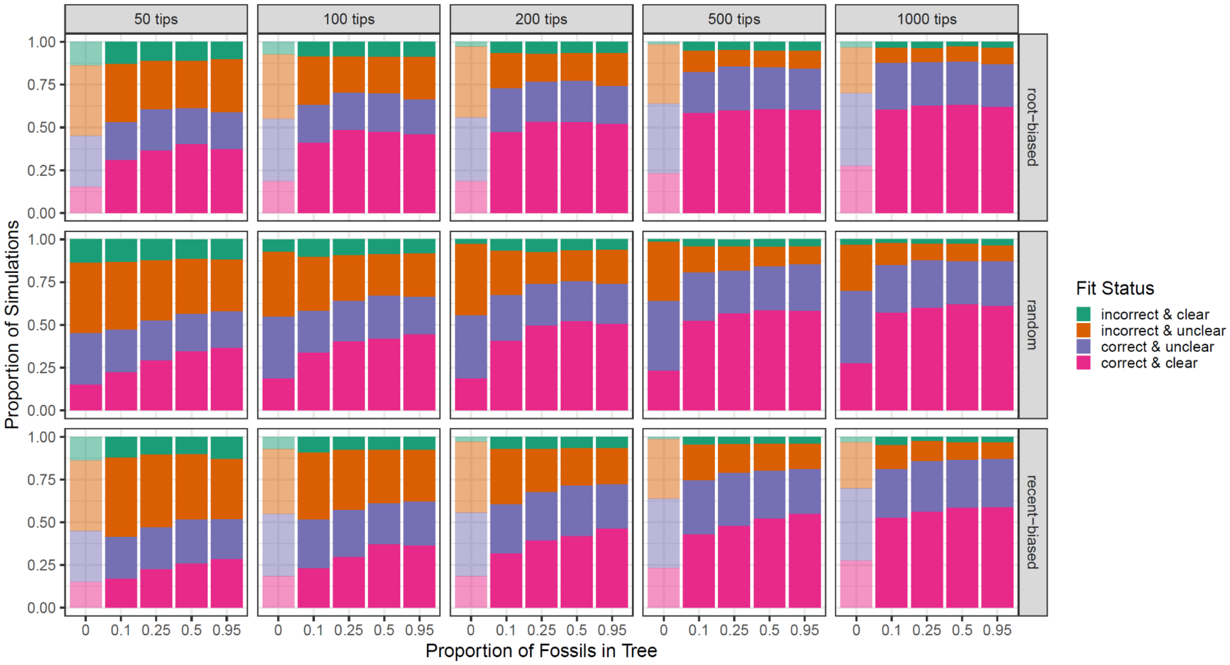
Zou Z., Zhang J. 2016. Morphological and molecular convergences in mammalian phylogenetics. Nat. Commun. 7:12758.

Zwickl D.J., Hillis D.M. 2002. Increased Taxon Sampling Greatly Reduces Phylogenetic Error. Syst. Biol. 51:588–598.

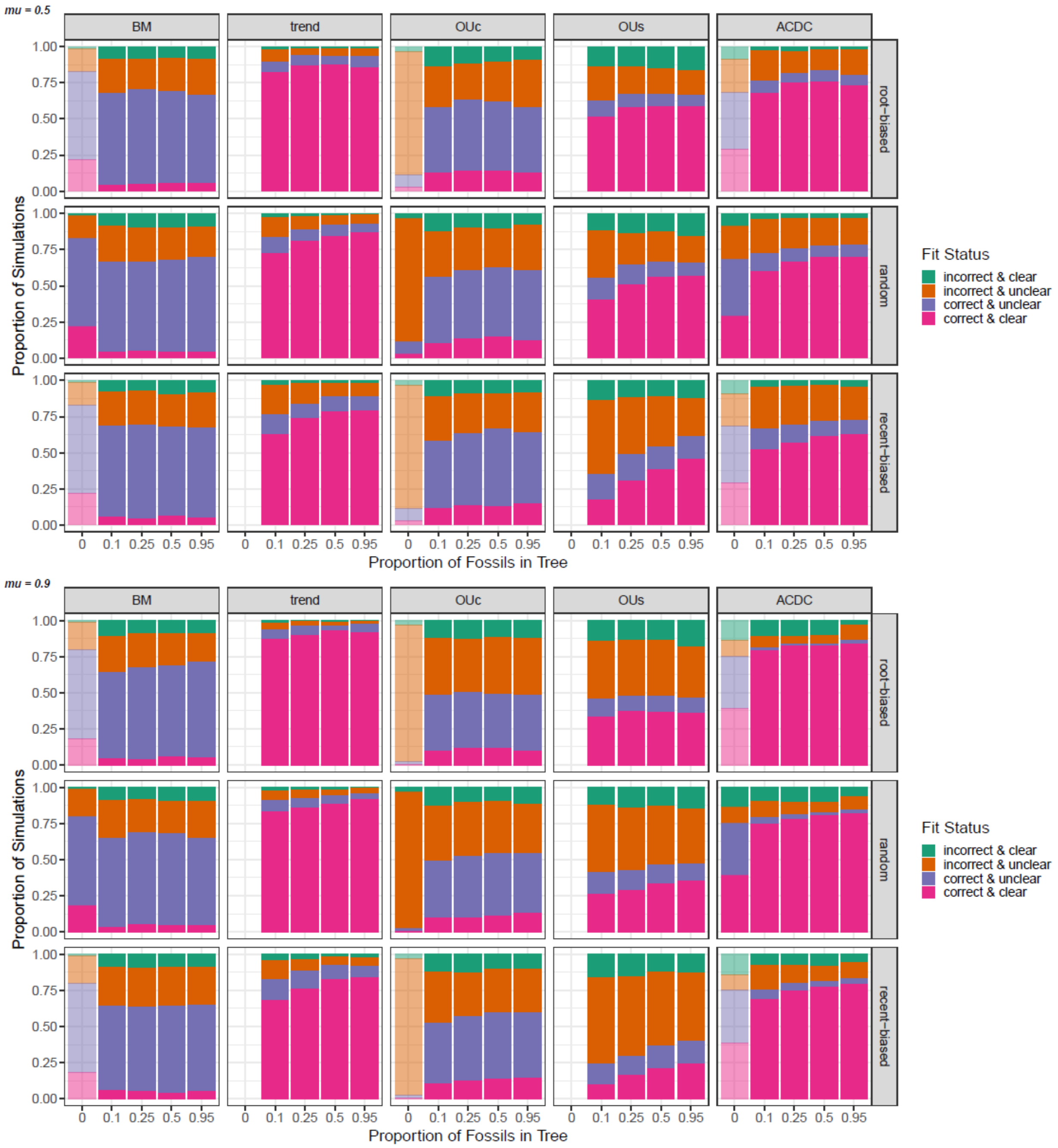
SUPPLEMENTAL FIGURES



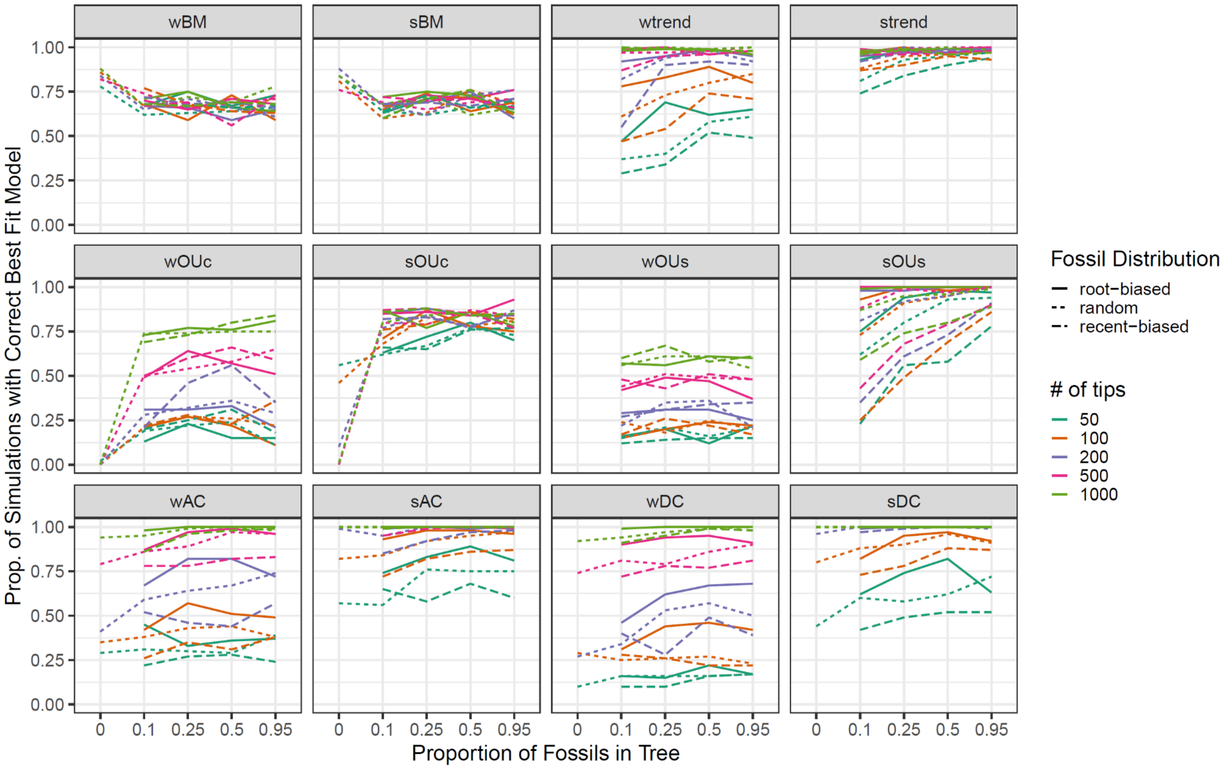
Supplemental Figure 1: The realized age distributions of the fossils in simulated phylogenies. The results are split out by the number of tips in the simulated phylogeny (row headers), the proportion of fossils in the simulated phylogeny (column headers), and the temporal bias of the fossils (color). Each simulated phylogeny has a different raw height, so the ages of the fossils are presented here as relative heights. These results are only for analyses where $\mu = 0.5$, the results for analyses where $\mu = 0.9$ are in Fig. 2.



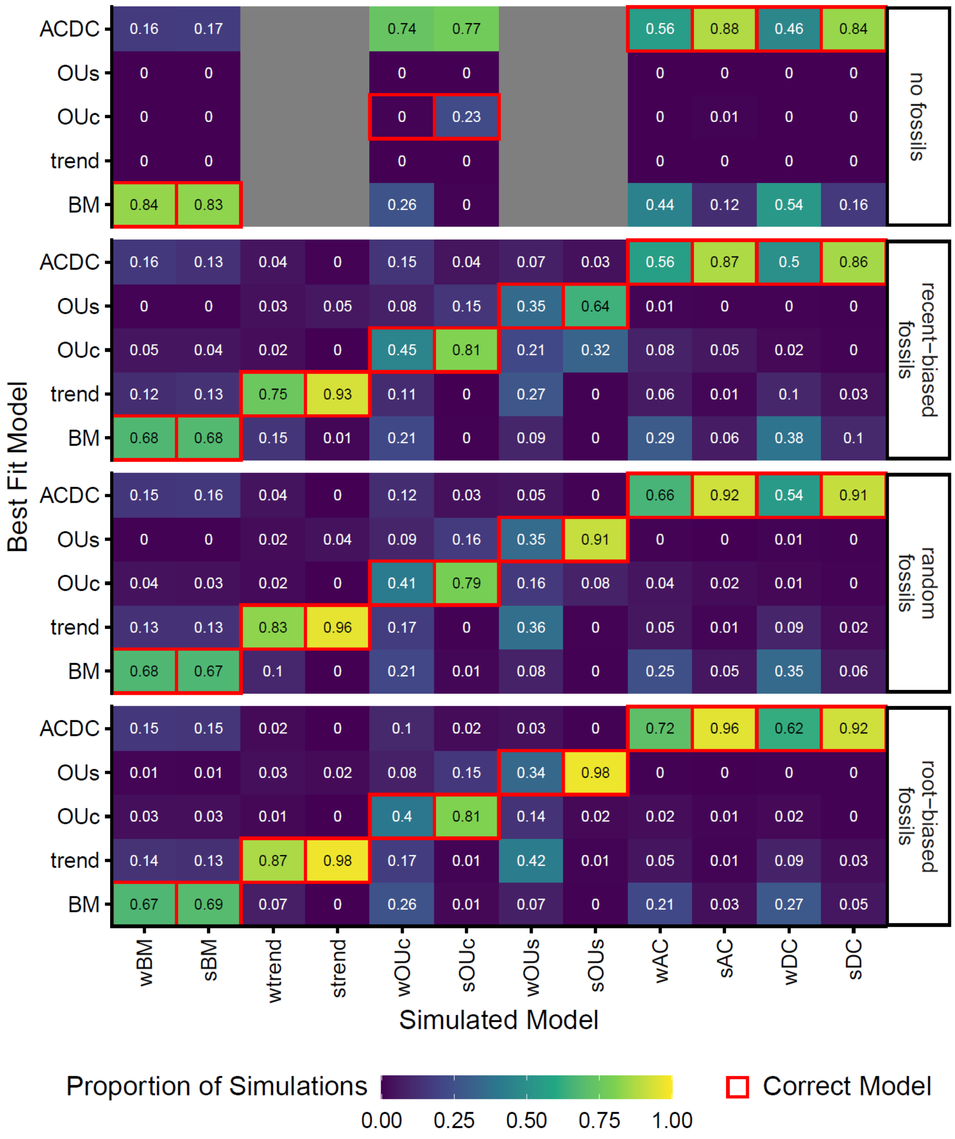
Supplemental Figure 2: Summary of the best-fitting models across all simulations ($\mu = 0.5$). The best-fitting model for a given simulation can be either correct (matching the simulated model) or incorrect (not matching the simulated model) and also either clear (all other models with ΔAICc > 2) or unclear (at least one other model with ΔAICc ≤ 2). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row headers), and the number of tips in the simulated phylogenies (column headers). Analyses without fossils are presented in full color in the row for randomly distributed fossils, and those same results are shown faded for the other rows. These results are only for analyses where $\mu = 0.5$, the results for analyses where $\mu = 0.9$ are in Fig. 3.



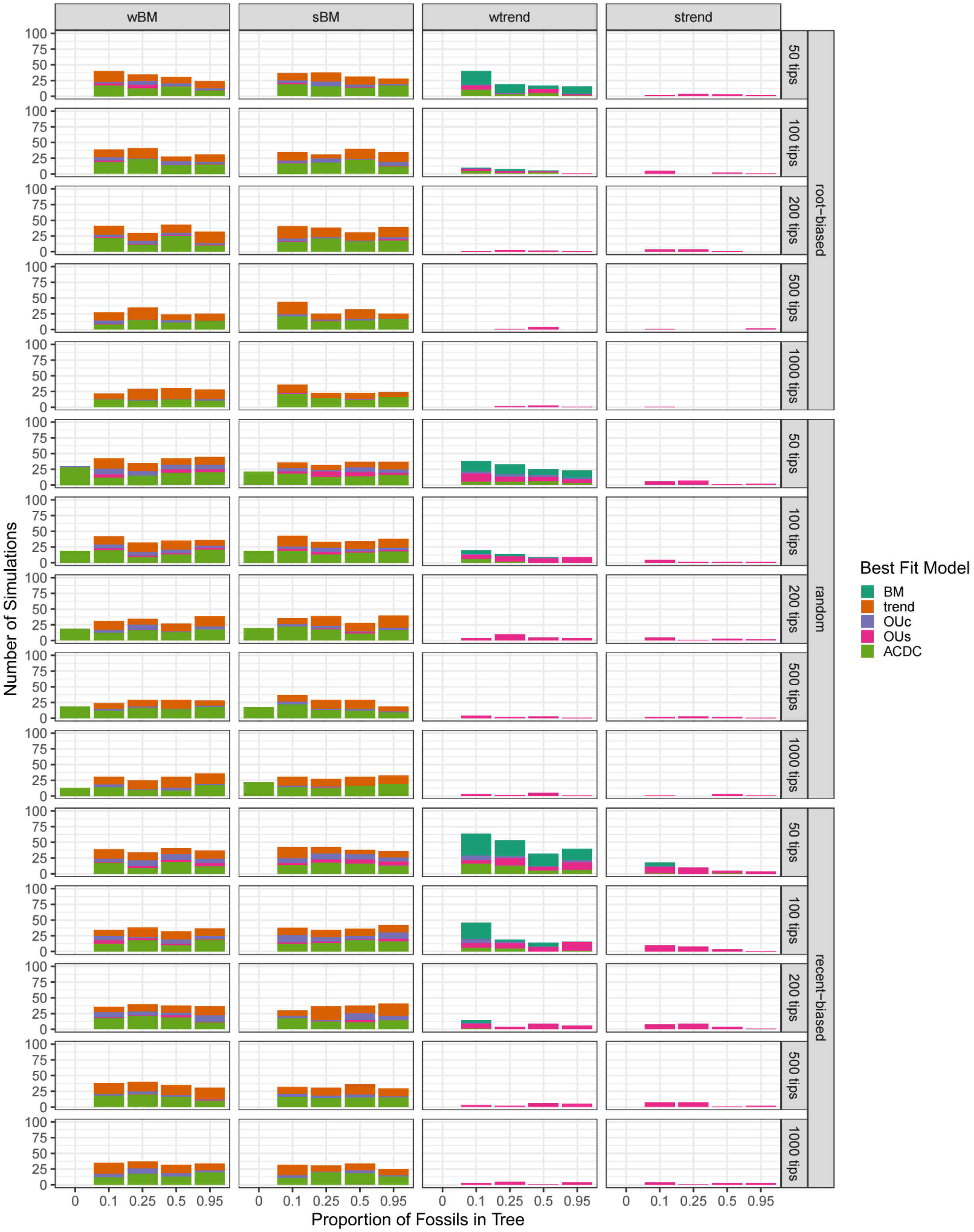
Supplemental Figure 3: Alternate summary of the best-fitting models across all simulations. The best-fitting model for a given simulation can be either correct (matching the simulated model) or incorrect (not matching the simulated model) and also either clear (all other models with ΔAICc > 2) or unclear (at least one other model with ΔAICc ≤ 2). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row headers), and the type of simulated model (column headers). Analyses without fossils are presented in full color in the row for randomly distributed fossils, and those same results are shown faded for the other rows. Missing bars indicate analyses that were skipped due to nonidentifiability. The upper panel shows the results for analyses where $\mu = 0.5$, and the lower panel shows the results for analyses where $\mu = 0.9$.



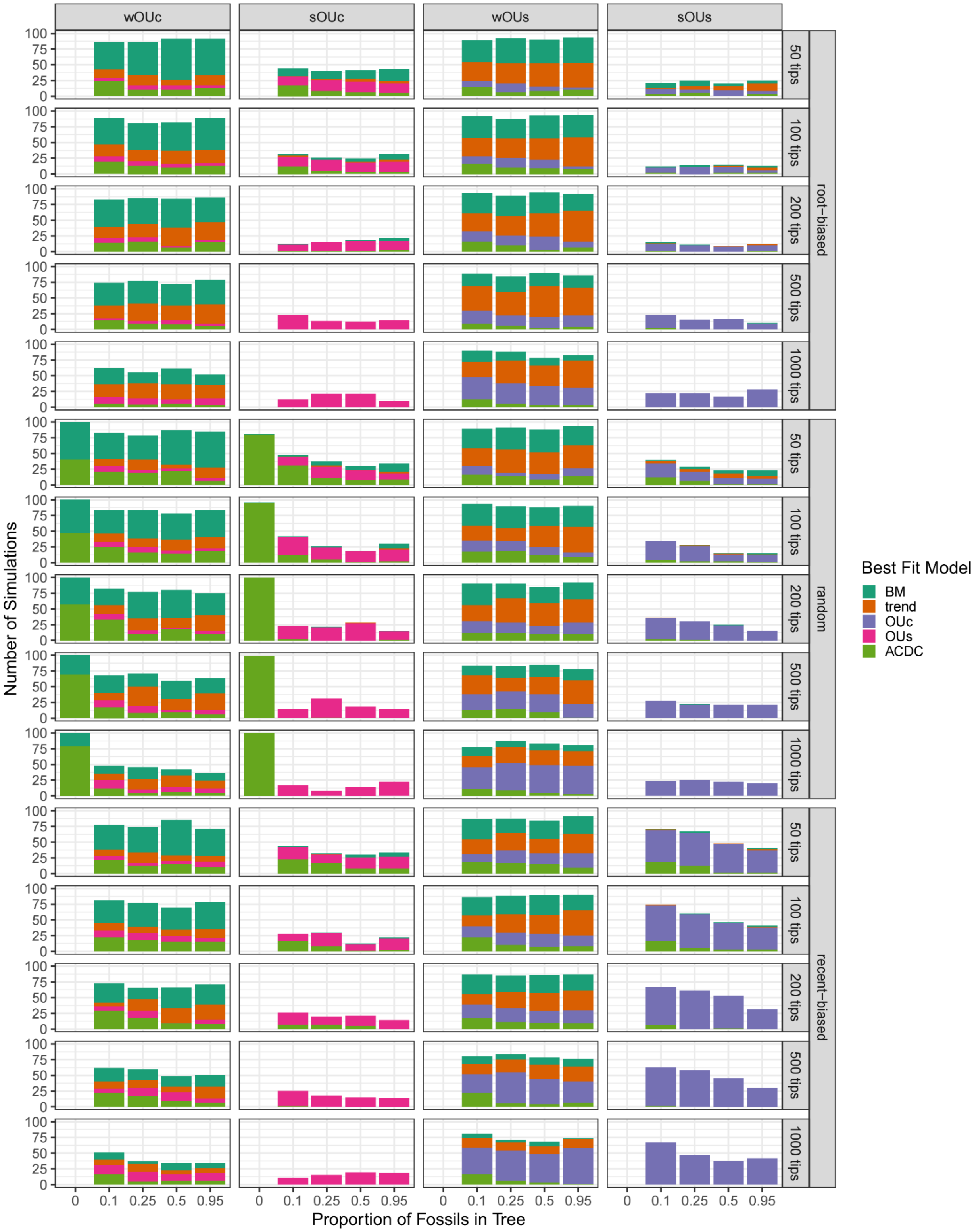
Supplemental Figure 4: The proportion of simulations for which the best-fit model <u>does</u> match the simulated model (panel headers) (μ = 0.5). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (line type), and the number of tips in the simulated phylogenies (line color). These results are only for analyses where μ = 0.5, the results for analyses where μ = 0.9 are in Fig. 4.



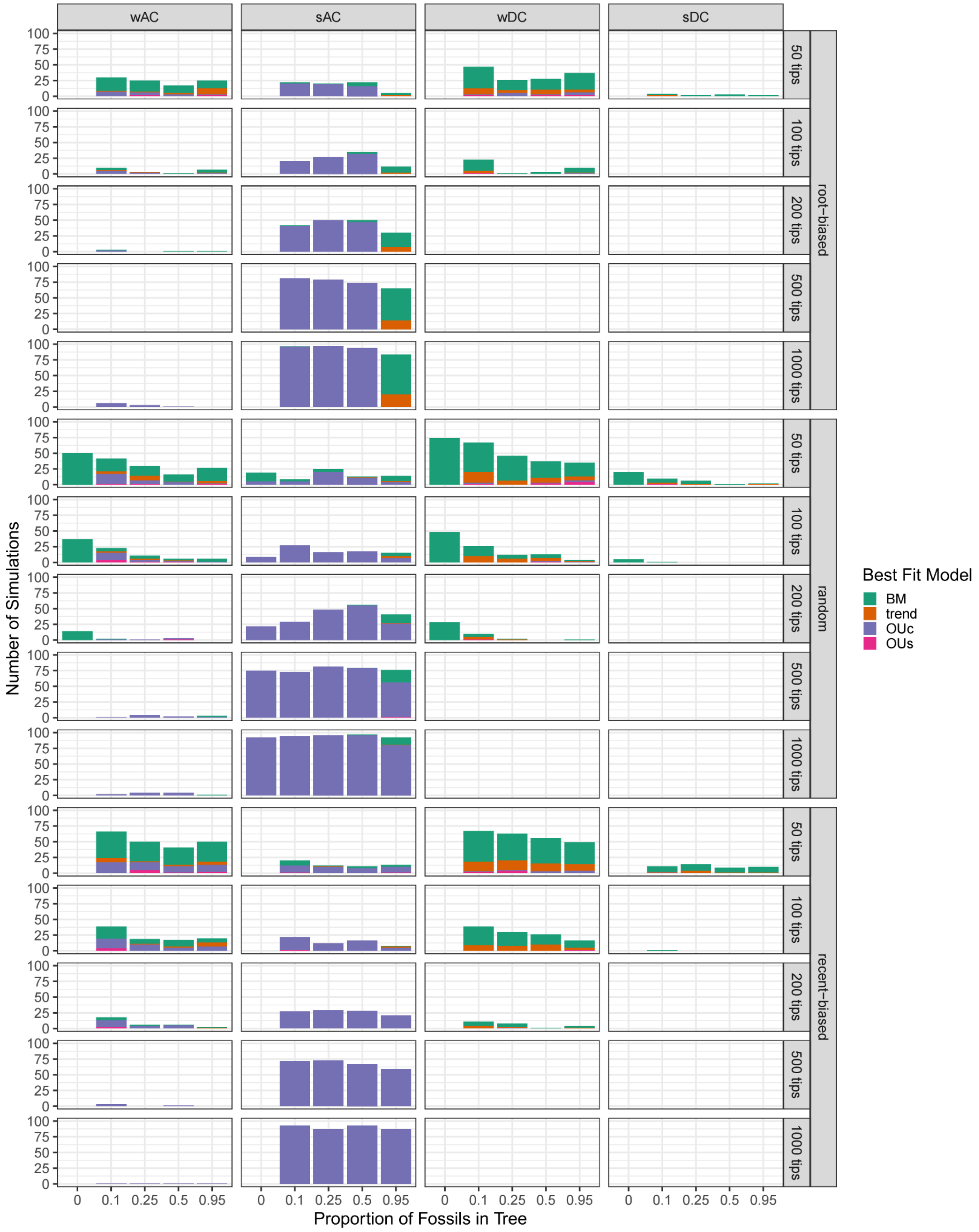
Supplementary Figure 5: Confusion matrix showing the proportions of simulations that were best fit by each model (y-axis) (μ = 0.5). The results are split out by the simulated model (x-axis) and the fossil treatment (rows). Analyses without fossils are split out as a separate treatment. Red outlines represent the correct model given the simulated model (note that there is not a one-to-one relationship). Grey boxes indicate analyses that were skipped due to nonidentifiability. These results are only for analyses where μ = 0.5, the results for analyses where μ = 0.9 are in Fig. 5.



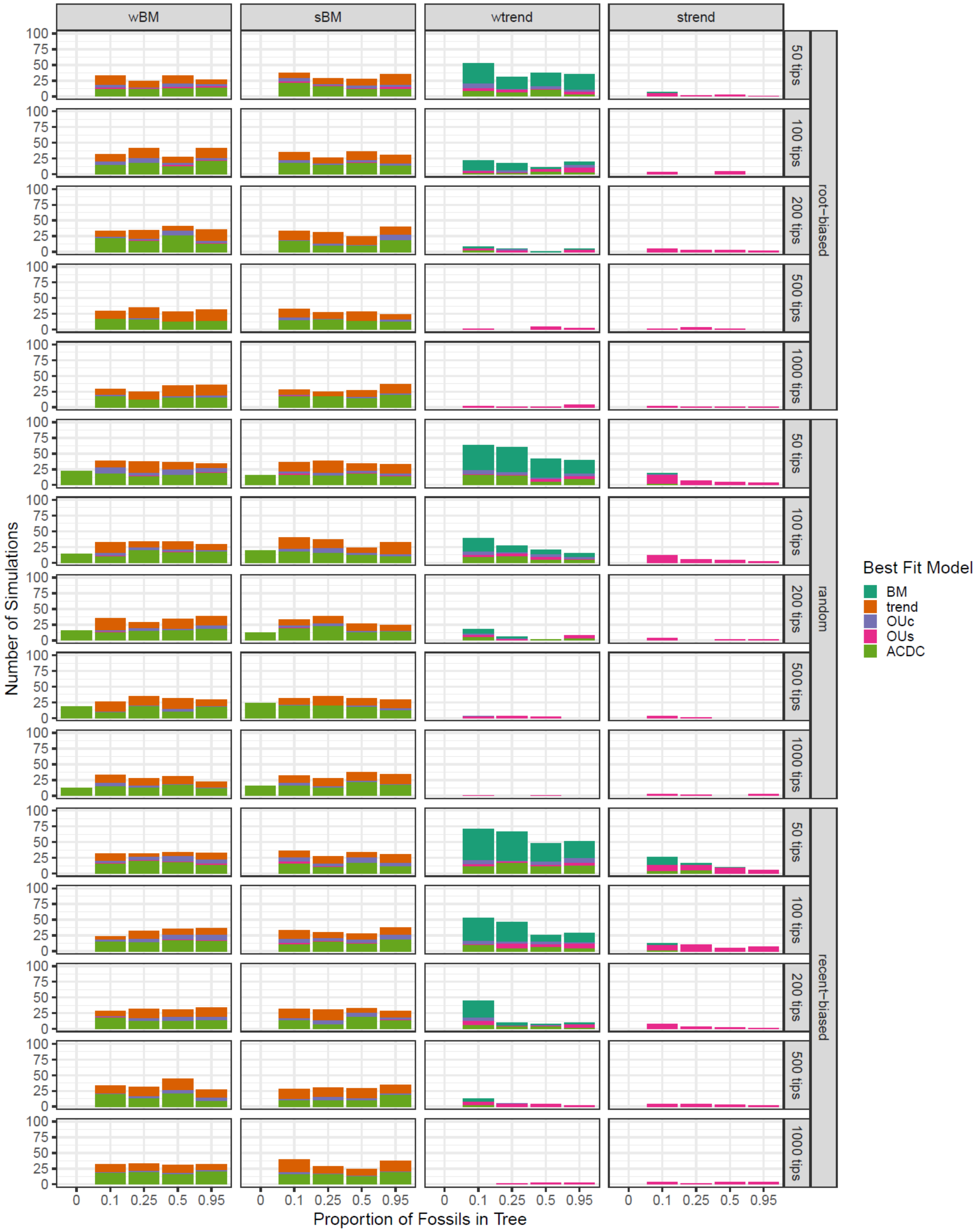
Supplemental Figure 6: The number of BM/trend simulations for which the best-fit model (color) <u>does not</u> match the simulated model (panel headers). Only BM and trend simulations are included (see Figs. S7 and S8 for other simulations). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row groups), and the number of tips in the simulated phylogenies (rows). These results are only for analyses where $\mu = 0.9$, the results for analyses where $\mu = 0.5$ are in Fig. S9.



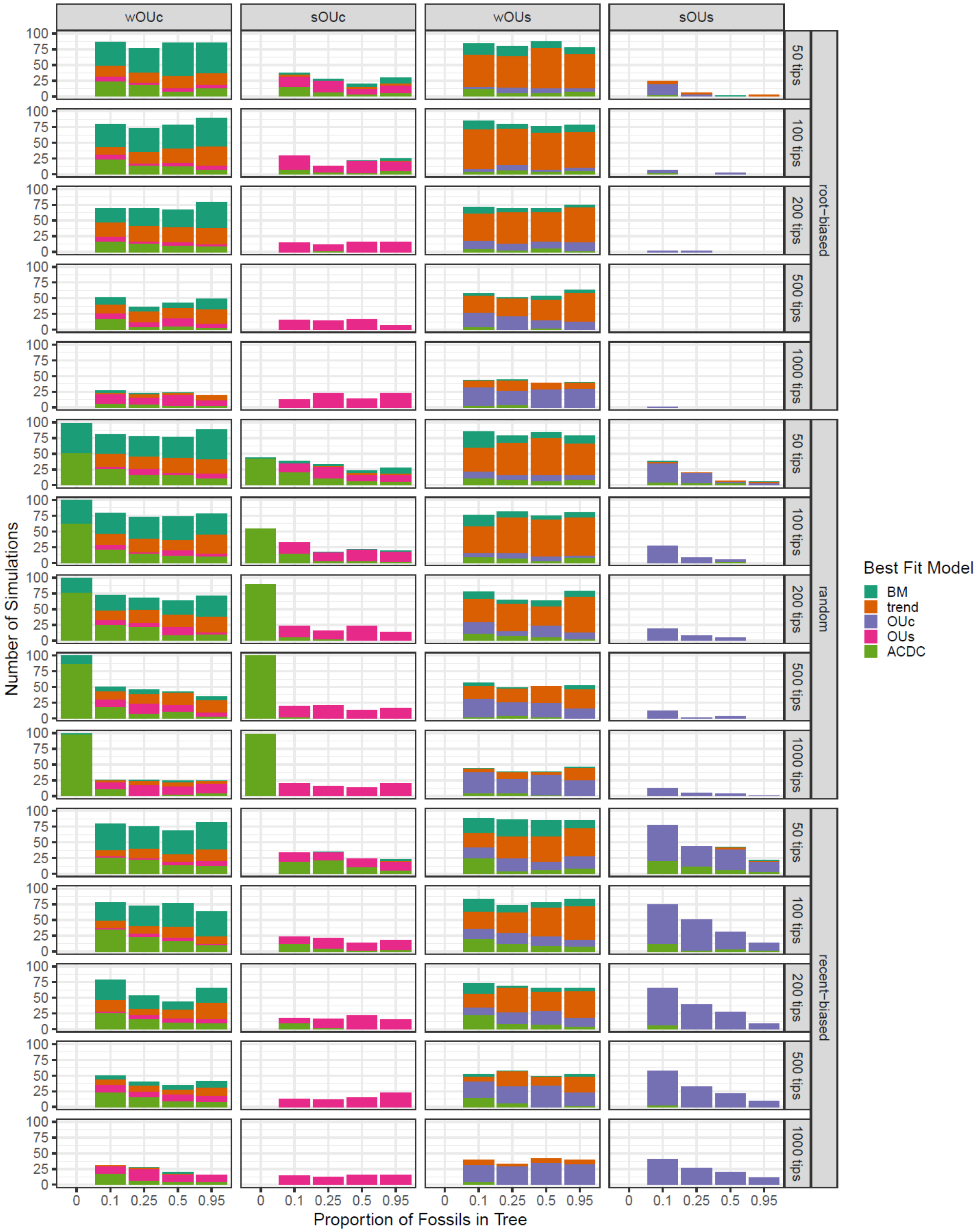
Supplemental Figure 7: The number of OU simulations for which the best-fit model (color) <u>does not</u> match the simulated model (panel headers). Only OU simulations are included (see Figs. S6 and S8 for other simulations). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row groups), and the number of tips in the simulated phylogenies (rows). These results are only for analyses where μ = 0.9, the results for analyses where μ = 0.5 are in Fig. S10.



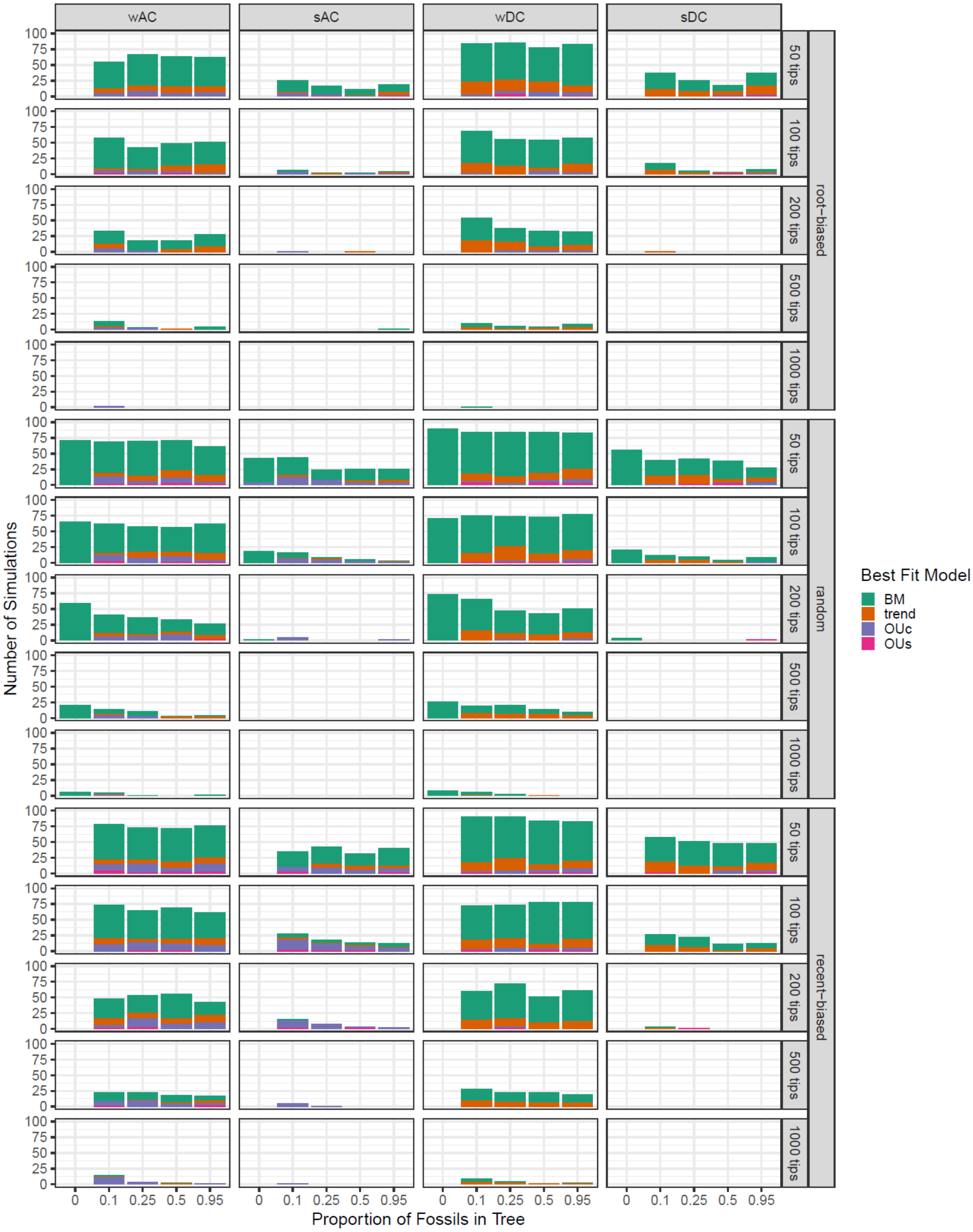
Supplemental Figure 8: The number of ACDC simulations for which the best-fit model (color) <u>does not</u> match the simulated model (panel headers). Only AC and DC simulations are included (see Figs. S6 and S7 for other simulations). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row groups), and the number of tips in the simulated phylogenies (rows). These results are only for analyses where $\mu = 0.9$, the results for analyses where $\mu = 0.5$ are in Fig. S11.



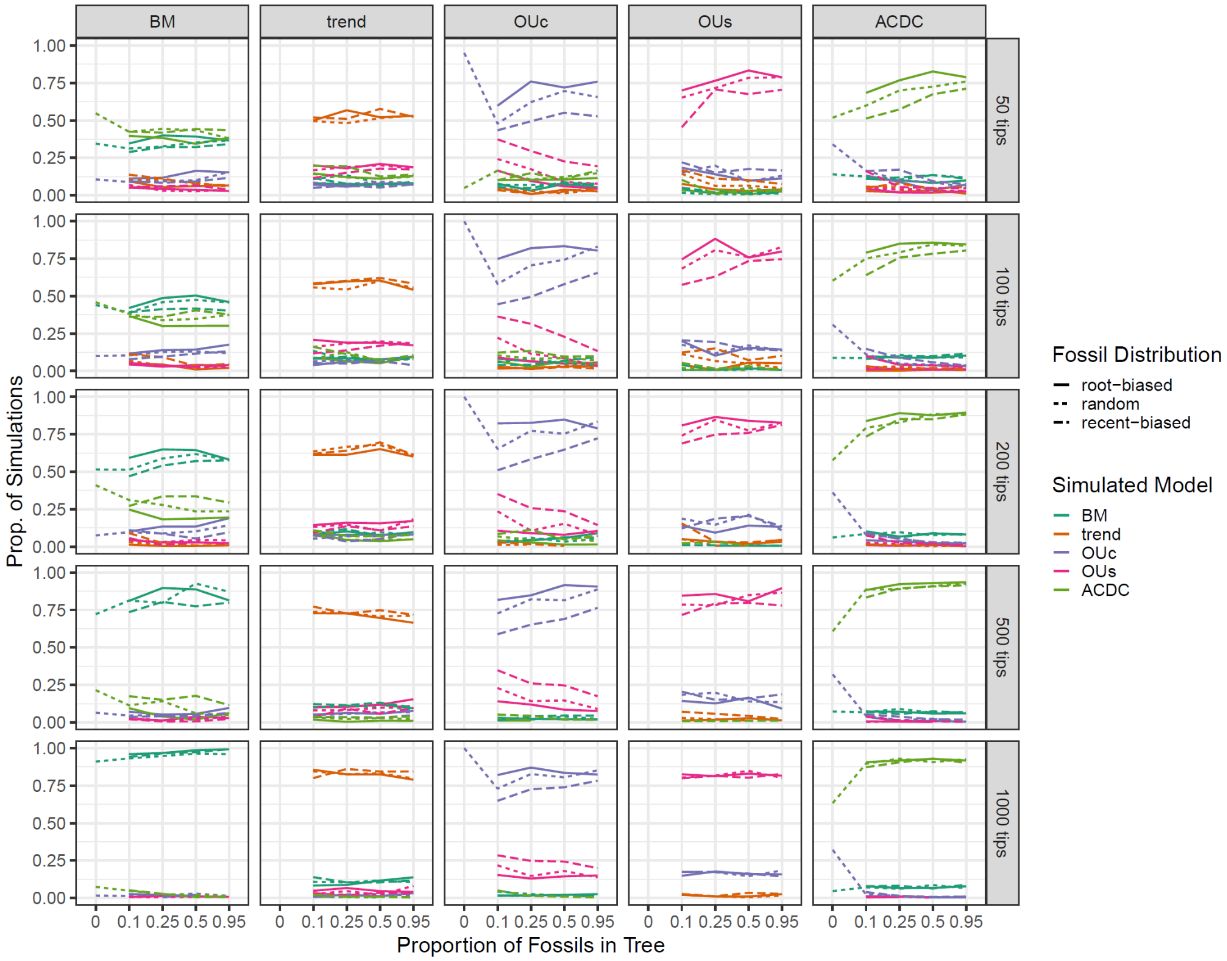
Supplementary Figure 9: The number of BM/trend simulations for which the best-fit model (color) does not match the simulated model (panel headers) (μ = 0.5). Only BM and trend simulations are included (see Figs. S10 and S11 for other simulations). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row groups), and the number of tips in the simulated phylogenies (rows). These results are only for analyses where $\mu = 0.5$, the results for analyses where $\mu = 0.9$ are in Fig. S6.



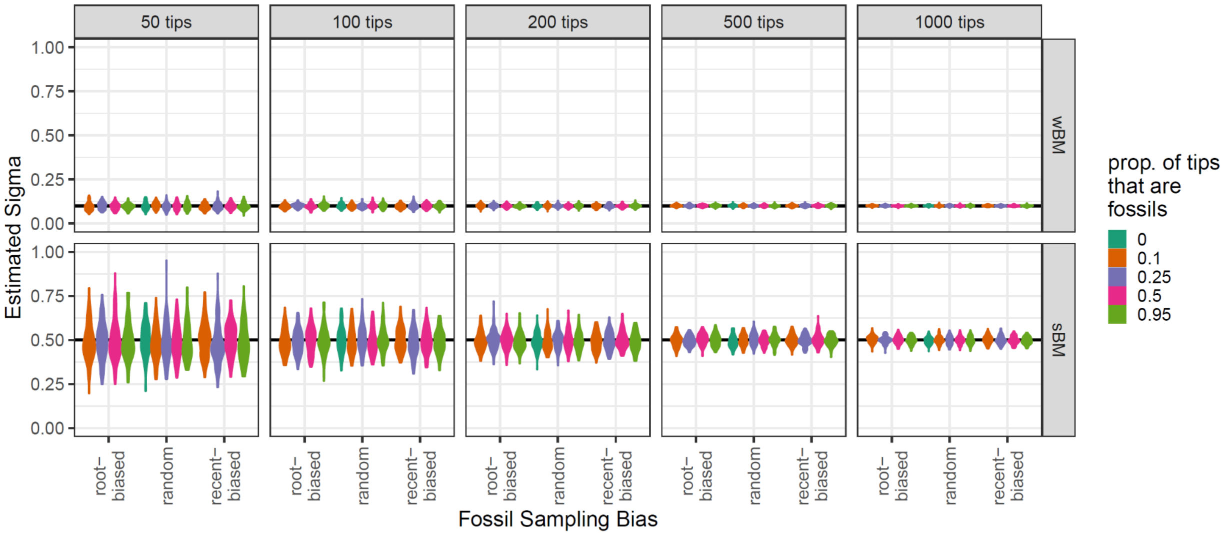
Supplementary Figure 10: The number of OU simulations for which the best-fit model (color) <u>does not</u> match the simulated model (panel headers) (μ = 0.5). Only OU simulations are included (see Figs. S9 and S11 for other simulations). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row groups), and the number of tips in the simulated phylogenies (rows). These results are only for analyses where $\mu = 0.5$, the results for analyses where $\mu = 0.9$ are in Fig. S7.



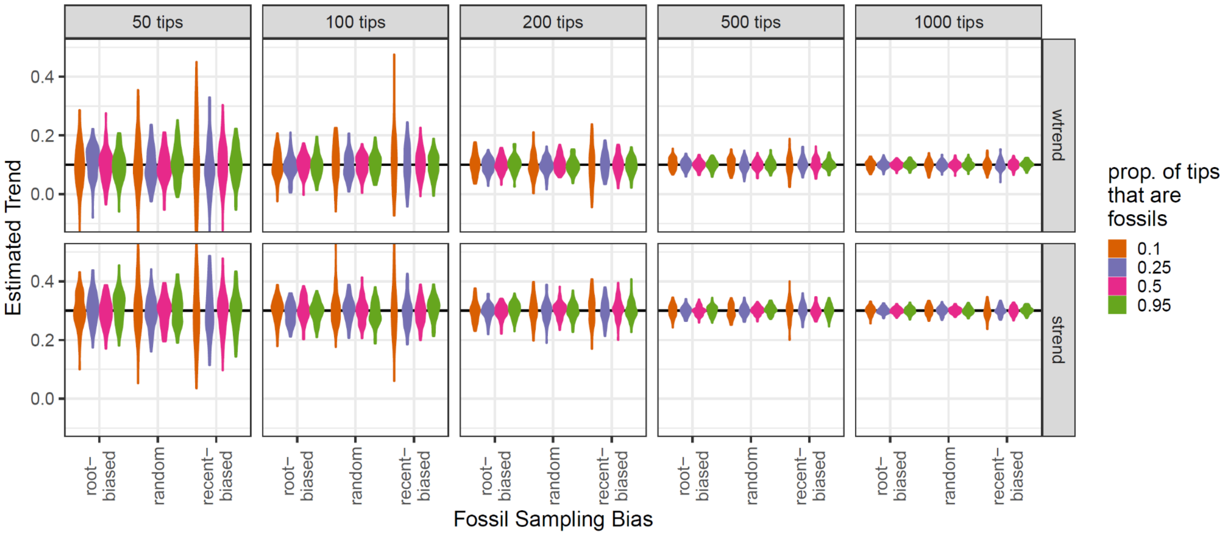
Supplementary Figure 11: The number of BM/trend simulations for which the best-fit model (color) does not match the simulated model (panel headers) (μ = 0.5). Only BM and trend simulations are included (see Figs. S9 and S10 for other simulations). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the distribution of the simulated fossils (row groups), and the number of tips in the simulated phylogenies (rows). These results are only for analyses where $\mu = 0.5$, the results for analyses where $\mu = 0.9$ are in Fig. S8.



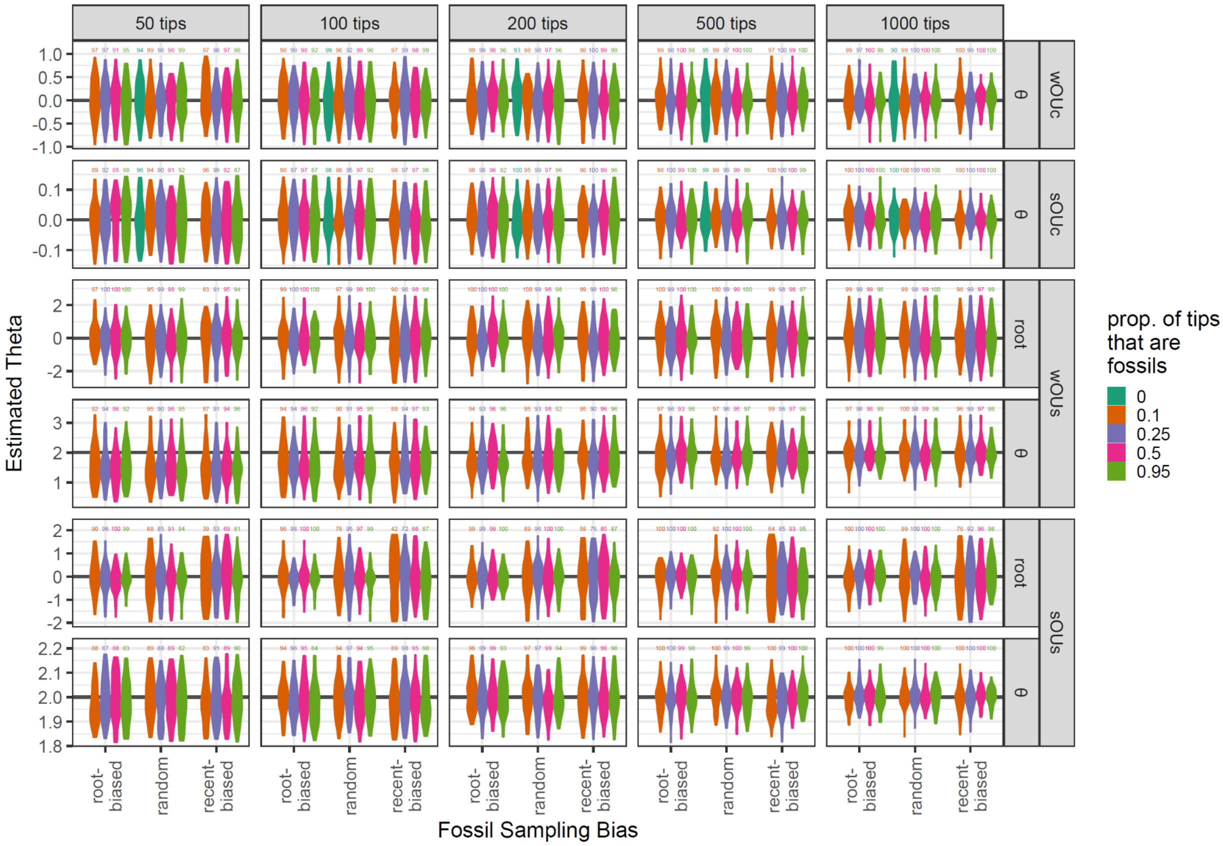
Supplemental Figure 12: The proportion of simulations best fit by a given model (column headers) that were generated by each simulated model (line color). The results are split out by the proportion of fossils in the simulated phylogeny (x-axis), the age distribution of the simulated fossils (line type), and the number of tips in the simulated phylogenies (row headers). Analyses without fossils were either not conducted due to the presence of a trend (*trend*, *OUs*) or are plotted as part of the random distribution. These results are only for analyses where $\mu = 0.5$, the results for analyses where $\mu = 0.9$ are in Fig. 6.



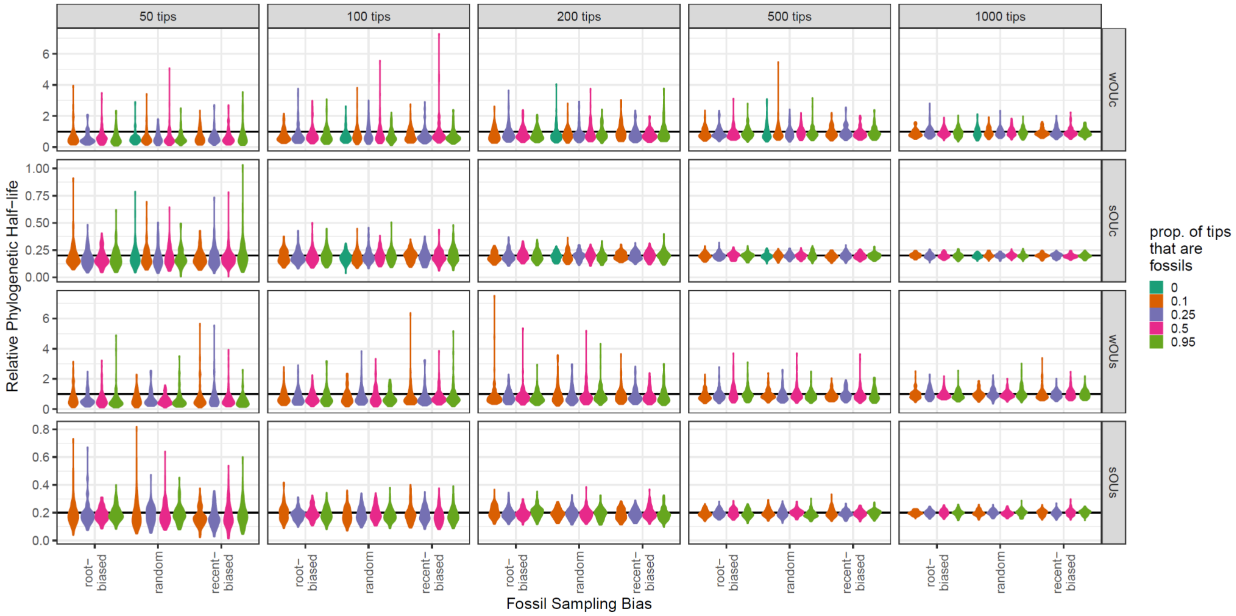
Supplemental Figure 13: The distributions of the estimates for the σ parameter when a BM model was simulated on a phylogeny and then a BM model was fit to that simulated data (μ = 0.5). The top row represents simulated trends with a weak σ parameter (0.1) and the bottom row represents a trend model simulated with a strong σ parameter (0.5). The true (simulated) σ parameter is represented with a solid horizontal line, and the results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). Analyses without fossils are plotted as part of the random distribution. These results are only for analyses where μ = 0.5, the results for analyses where μ = 0.9 are in Fig. 7.



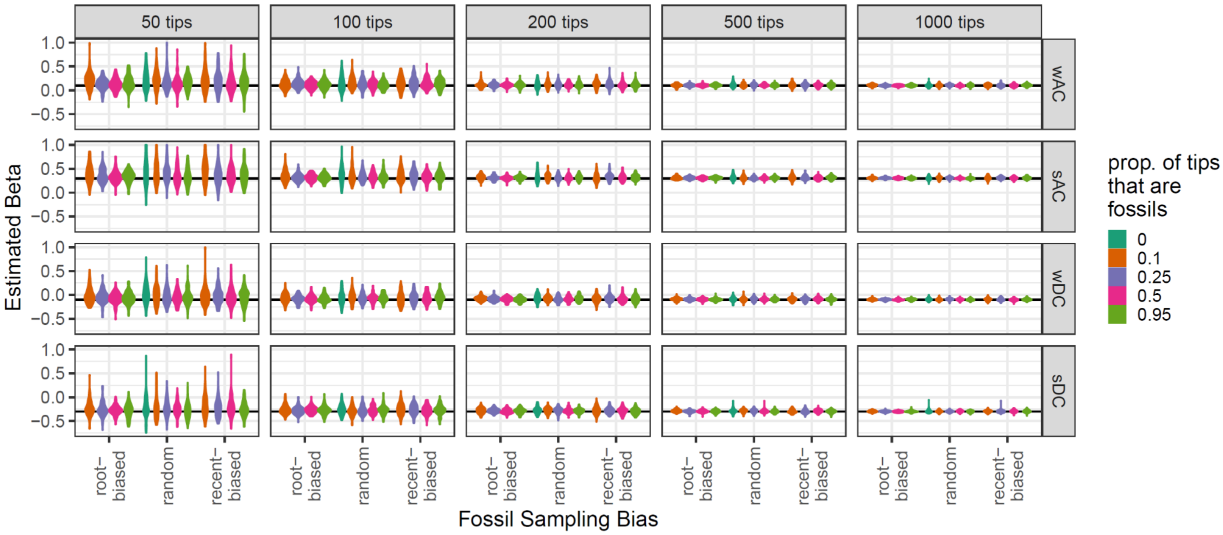
Supplemental Figure 14: The distributions of the estimates for the *trend* parameter when a trend model was simulated on a phylogeny and then a trend model was fit to that simulated data (μ = 0.5). The top row represents simulated trends with a weak *trend* parameter (0.1) and the bottom row represents a trend model simulated with a strong *trend* parameter (0.3). The true (simulated) *trend* parameter is represented with a solid horizontal line, and the results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). These results are only for analyses where μ = 0.5, the results for analyses where μ = 0.9 are in Fig. 8.



Supplemental Figure 15: The distributions of the estimates for the *θ* parameter when an OU model was simulated on a phylogeny and then an OU model was fit to that simulated data (*μ* = 0.5). The true (simulated) *θ* parameter is represented with a solid horizontal line. Outliers have been removed separately for each row, and the numbers above violin plots represent the number of non-outliers that are represented by each violin plot. The results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). Note that y-axis limits vary from row to row, and analyses without fossils are plotted as part of the random distribution. These results are only for analyses where *μ* = 0.5, the results for analyses where *μ* = 0.9 are in Fig. 9.



Supplemental Figure 16: The distributions of the estimates for the relative half-life when an OU model was simulated on a phylogeny and then an OU model was fit to that simulated data (*μ* = 0.5). Relative half-life is calculated as $\frac{log(2) / \alpha}{tree\ height}$. The first and second rows represent centered OU models simulated with weak (1) and strong (0.1) relative half-lives, respectively. The third and fourth rows represent shifting OU models simulated with weak (1) and strong (0.2) relative half-lives, respectively. The true (simulated) relative half-life is represented with a solid horizontal line, and the results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). Note that y-axis limits vary from row to row, and analyses without fossils are plotted as part of the random distribution. These results are only for analyses where *μ* = 0.5, the results for analyses where *μ* = 0.9 are in Fig. 10.



Supplemental Figure 17: The distributions of the estimates for the β parameter when an ACDC model was simulated on a phylogeny and then an ACDC model was fit to that simulated data (μ = 0.5). The true (simulated) β parameter is represented with a solid horizontal line. The results are split out by the proportion of fossils in the simulated phylogeny (color), the distribution of the simulated fossils (x-axis), and the number of tips in the simulated phylogenies (column headers). Analyses without fossils are plotted as part of the random distribution. These results are only for analyses where μ = 0.5, the results for analyses where μ = 0.9 are in Fig. 11.