# Fossils Do Not Substantially Improve, and May Even Harm, Estimates of Diversification Rate Heterogeneity

Jeremy M. Beaulieu[1] and Brian C. O'Meara[2]

[1]*Department of Biological Sciences, University of Arkansas, Fayetteville, Arkansas, 72701 USA*

[2]*Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, Tennessee, 37996-1610 USA*

[*]Email for correspondence: jmbeauli@uark.edu

*Abstract* — There is a prevailing view that the inclusion of fossil data could remedy identifiability issues related to models of diversification, by drastically reducing the number of congruent models. The fossilized birth-death (FBD) model is an appealing way of directly incorporating fossil information when estimating diversification rates. Here we explore the benefits of including fossils by implementing and then testing two-types of FBD models in more complex likelihood-based models that assume multiple rate classes across the tree. We also assess the impact of severely undersampling, and even not including fossils that represent samples of lineages that also had sampled descendants (i.e., *k*-type fossils), as well as converting a fossil set to represent stratigraphic ranges. Under various simulation scenarios, including a scenario that exists far outside the set of models we evaluated, including fossils rarely outperforms analyses that exclude them altogether. At best, the inclusion of fossils improves precision but does not influence bias. We also found that severely undercounting the number of *k*-type fossils produces highly inflated rates of turnover and extinction fraction. Similarly, we found that converting the fossil set to stratigraphic ranges results in turnover rates and extinction fraction estimates that are generally underestimated. While fossils remain essential for understanding diversification through time, in the specific case of understanding diversification given an existing, largely modern tree, they are not especially beneficial.

Keywords — fossilized birth-death, state speciation extinction, MiSSE, fossils, stratigraphic ranges, turnover

Diversification models, while fascinating to biologists, frequently lurch close to extinction themselves. For instance, Nee et al. (1994) demonstrated that estimating speciation and extinction rates from a molecular timetree is theoretically possible. This was quickly followed by work from Kubo and Iwasa (1995), which showed that if rates varied through time there is an infinite number of alternative sets of time-varying speciation and/or extinction rates that produce the same number of lineages at any given point in time (an observation largely ignored at the time). State speciation and extinction models (-SSE; Maddison et al. 2007) were derived as a seemingly robust framework for estimating the direct effects of discrete traits on diversification rates. However, Rabosky and Goldberg (2015) found that if a tree evolved under a heterogeneous branching process, completely independent from the evolution of the focal character, SSE models will almost always return erroneously strong support for a model of state-dependent diversification. While this issue was partially rescued by the hidden state models of Beaulieu and O'Meara (2016), there remains some confusion as to whether SSE models remain a viable means of assessing state-dependent diversification (Rabosky and Goldberg, 2017; but see Caetano et al. 2018).

With Louca and Pennell (2020) comes a new salvo of criticism regarding diversification methods, which echo and expand on the earlier points by Kubo and Iwasa (1995), as well as past concerns about the possibility of estimating extinction rates generally (i.e., Rabosky, 2010). One postulated source of salvation has been the inclusion of fossils (Mitchell et al. 2019; Louca and Pennell, 2020; but see Černý et al. 2021). At a minimum, the inclusion of fossils should drastically reduce the number of congruent models by excluding any congruent model that assumes no extinction. We also know that fossils, even without formal models, have contributed greatly to the understanding of diversification processes — take the discovery of mass

extinctions, for example. For these reasons, the fossilized birth-death (FBD) model (Stadler 2010) is seen as an appealing way of directly incorporating fossil information when estimating diversification rates, because it naturally assumes that the fossil information represents samples of extinct lineages in the past, in addition to the species sampled at the present. An important, but frequently overlooked, property of FBD is that it distinguishes between two types of sampled lineages (Fig. 1). The first are referred to as *m* fossils, which represent sampled branches that went completely extinct before the present and did not give rise to any additional sampled lineages. The second type are referred to as *k* fossils, which represent samples of lineages that had sampled descendants (other fossils, or extant tips). In other words, these are fossils that represent ancestors sampled on internal branches that eventually lead to sampled species. The model assumes a sampling rate of fossils that does not vary across taxa or time — a questionable assumption, but probably no worse than similar simplifications that speciation rate or extinction rate is similarly invariant. However, while the model makes this distinction among these fossil types, the rate by which each is sampled is governed by the same global sampling rate, $\psi$, such that $\psi_m = \psi_k$.

The linking of these parameters makes sense. When a dinosaur gets washed into a river, the probability of it being excavated later by a paleontologist does not depend on whether a descendant seventy million years later is sampled (making our unlucky dinosaur a *k* fossil) or not (making this an *m* fossil). However, in the model, an assumption is not only that the probability of sampling the fossil is the same regardless of whether it has descendants or not, but also that the probability of including it in the phylogeny is the same whether or not it has descendants. For example, Paleobiology Database (PBDB; accessed October 2021) has 67 collections for *Tyrannosaurus rex* (Osborn 1905) and only eight for *Stegosaurus stenops* (Marsh 1887). For the

model to fully apply, there should be 67/8 times as many *T. rex* fossils as *S. stenops* fossils included in the tree.

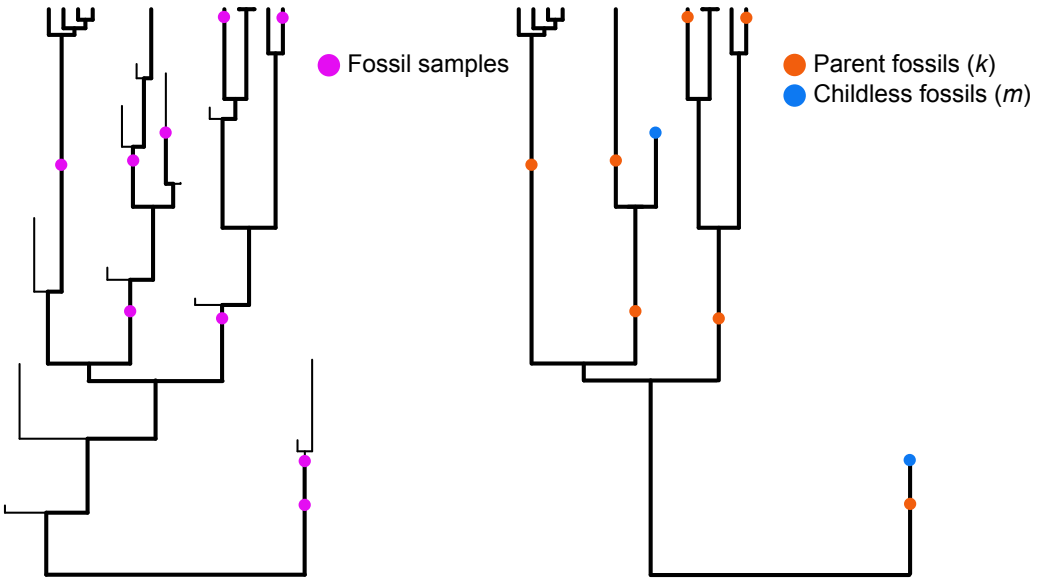


**Figure 1.** a) An example of a fossil set sampled from a complete tree generated by the birth-death process, which includes both surviving and extinct members. b) The sampled phylogeny pruned down from (a) to only includes sampled tips and fossils. The fossilized birth-death model distinguishes between two types of sampled lineages. An *m* fossil (denoted by a blue dot) represents sampled branches that went completely extinct before the present and did not give rise to any additional sampled lineages (i.e., "childless" fossils). A *k* fossil (denoted by an orange dot) represents samples of lineages that had sampled descendants (other fossils, or extant tips; "parent" fossils).

Of course, there are other biases and variation in fossilization rate (the whole field of taphonomy studies this rich and varied process), but we are concerned with a potential bias of experimentalists undersampling *k* fossils. This would have predictable effects on the estimates of $\psi$ as well as the extinction rate itself. To better understand the dynamics of the different fossil types, here is Eq. (5) from Stadler (2010), which is probability density of a tree, *T*, with *n* extant

taxa, $m \geq 0$ fossils samples, and $n \geq 0$ samples conditioned on the time since the most recent common ancestor of all taxa in the tree, $x_1$:

$$f[T|t_{mrca} = x_1] = \frac{\lambda^{n+m+2}\psi^{k+m}}{(1-\hat{p}_0(x_1))^2} p_1(x_1) \prod_{i=1}^{n+m+1} p_1(x_i) \prod_{i=1}^{m} \frac{p_0(y_i)}{p_1(y_i)}$$

For specifics on what the probabilities $p_0(t)$, $p_1(t)$, and $\widehat{p_0}(t)$ represent, we refer readers to the original derivation in Stadler (2010); for the purpose of this discussion, just note that they do not include *k*. We highlight two important observations here. The first is that the information provided by the *m*-type samples are used throughout the equation, including the time at which the extinct tip was sampled, $y_i$, as well as the time at which they split from their common ancestor, $x_i$,. The second observation is that the only information provided by the *k*-type fossils is simply how numerous they are in the fossil set. In fact, it is irrelevant where exactly they occur on a branch or which branches have them — that is, the likelihood is the same if they are dispersed evenly across the tree or sampled in one 10,000-year interval on one edge. This point is best illustrated by modifying the numerator in the first term in the equation above to separate out *k* and *m*,

$$f[T|t_{mrca} = x_1] = \psi^k \frac{\lambda^{n+m+2}\psi^{m}}{(1-\widehat{p_0}(x_1))^2} p_1(x_1) \prod_{i=1}^{n+m+1} p_1(x_i) \prod_{i=1}^{m} \frac{p_0(y_i)}{p_1(y_i)}.$$

Moving the $\psi^k$ to the front of the equation shows that the effect that *k*-type samples have on the probability is simply based on a factor of $\psi^k$. What this also says is that when both *m*- and *k*-type fossils are included, the overall log-likelihood can be calculated as, $logL = logL_{m\ only} + [k * log(\psi)]$. As an example, take the tree and set of sampled extinct lineages presented in Figure 2d. The overall log-likelihood for the maximum likelihood estimate (MLE) for this tree is -1224.937. The log-likelihood assuming only *m*-type fossils, but using the same parameter estimates, is

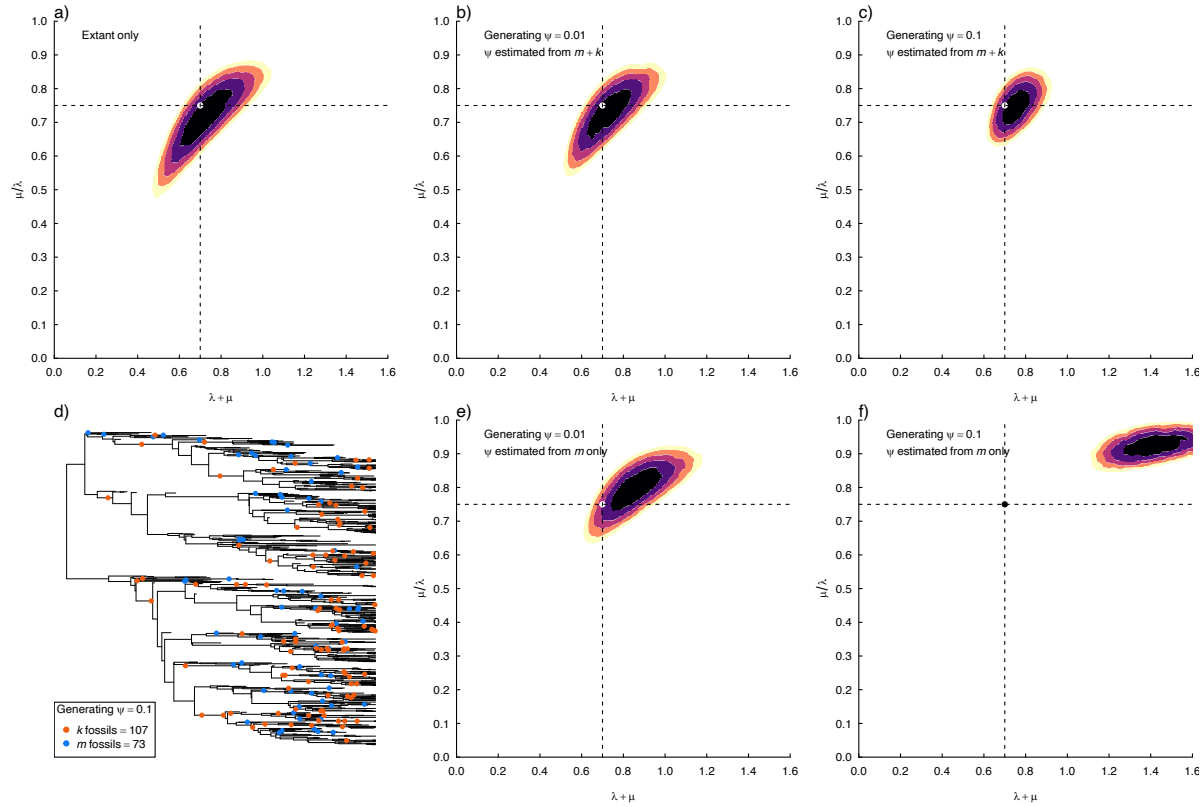


**Figure 2.** Contour plots of the likelihood surface under various scenarios for the same tree (topology shown in d) simulated under a birth-death process. Each surface is constrained such that turnover ($\tau = \lambda + \mu$) and extinction fraction ($\varepsilon = \mu/\lambda$) are fixed, but $\psi$, the rate of fossil preservation, is free to find their MLE given a pair of fixed parameter values. When provided with a sample of extinct lineages that is perfectly consistent with the generating model, the FDB performs and behaves well and generally reduces the variance in turnover and extinction fraction (b, c), relative to ignoring fossils completely (a). However, if $k$ fossils are removed completely (e,f), the likelihood surface begins to erroneously shift away from the generating parameters towards regions of parameter space of very high extinction rates as values of $\psi$ increase. The dashed vertical line represents the generating value for turnover ($\tau$=0.70) and the dashed horizontal line represents the generating value for extinction fraction ($\varepsilon$=0.75).

-983.6938, and we can add 107∗log($\hat{\psi}$=0.1049145) to obtain the overall log-likelihood of -1224.937 (for more details on these calculations see Supplemental Materials).

On the one hand, if a tree was generated by a Yule process (i.e., no extinction), the inclusion of only $k$-type samples has no effect on the parameter estimates. Their inclusion will simply slide the overall log-likelihood down, again, by a factor $k$*log($\psi$). On the other hand, in

the case of non-zero extinction rates, if $m$-type fossils are included and $k$-type not added, this will have a profound impact on the parameter estimates, particularly with regards to $\psi$ and $\mu$. Under the FBD formulation, the linking of $\psi_m = \psi_k$ forces the model to interpret the lack of $k$-type fossils as evidence of a low sampling rate. This will be in tension with the presence of only $m$-type samples, such that to explain a low sampling rate, but many $m$-type fossils in the set, the extinction rate must have been substantial. As shown in Figure 2e-f, this is exactly what happens when the $k$-type fossils are removed from the fossil sample set. The likelihood surface erroneously shifts away from the generating parameters towards regions of parameter space of very high extinction rates (based on estimates of turnover, $\lambda + \mu$, and extinction fraction, $\mu/\lambda$).

In any case, when provided with a sample of extinct lineages that is perfectly consistent with the generating model, the FDB performs and behaves exactly as it should (Fig. 2b-c), but so does a model that ignores fossils completely (Fig. 2a). Including fossils should help, especially with extinction rates (as shown by Mitchell et al. 2019). Our concern, however, is what happens when theory meets practice. It seems to us that aside from a few clades with remarkably rich fossil records (e.g., terrestrial arthropods, marine invertebrates, cetaceans), a vast majority, at best, may only have a small handful of fossils that can be placed with any confidence. This would mean that the decrease in the uncertainty (but likely not the bias) surrounding diversification rate estimates that include fossils would be small, though the inclusion of $m$-type fossils would help with our estimation of extinction rates. However, these benefits would quickly fade when estimating rates using a few fossils in a more parameter-rich model that contains many discrete changes in rates throughout the tree. We also suspect that, in practice, it is generally hard to know whether a fossil is best placed exactly on a branch or as an extinct sister to it given insufficient morphological evidence generally (see Parins-Fukuchi 2021). Many

171 studies that combine molecular and morphological information to place fossils in the context of
172 modern taxa, particularly within plant groups, often place fossils as extinct sisters, suggesting
173 that most data sets probably mistakenly contain far more m-type fossils than k-type ones. While
174 some studies have begun placing fossils on edges or tips (e.g., Heath et al. 2014; Gavryushkina et
175 al. 2017), much of the recent work is done in a Bayesian framework where it is hoped this
176 uncertainty is integrated out with no ill effects. Similarly, a recent extension of the FBD
177 explicitly treats the same fossil sets as multiple samples of the same species that specify distinct
178 stratigraphic ranges (Stadler et al. 2018). All that is required, then, is the oldest and youngest
179 fossil within a range, as opposed to the total number of fossils, which is then marginalized out,
180 potentially rendering issues related to counting *k* fossils moot in most cases.
181 These are the topics we seek to investigate here. Specifically, we explore the benefits of
182 including fossils in more complex models that assume multiple rate classes across the tree in
183 comparison to excluding them altogether. We also assess the impact of severely undersampling
184 and even not including *k*-type fossils into these analyses. In short, will fossils help diversification
185 analyses largely based on modern phylogenies? Our hypotheses at the outset were that well-
186 sampled fossils would have a very beneficial effect on estimation of diversification parameters
187 versus having no fossils and that models that exclude *k*-type fossils would perform poorly, likely
188 being outcompeted by models with no fossils. We expected approaches with stratigraphic
189 samples would have less benefit (reduced data, especially about fossilization rate) but also would
190 be relatively insensitive to sampling issues. We expected that fossils would be especially
191 valuable in a situation where the true model was more complex than any model used in analysis,
192 which approximates the situation in reality.
193

**Fossilized -SSE**

Incorporating the sampling rate, $\psi$, into a birth-death model that allows multiple discrete shifts has been implemented before (e.g., fossil BAMM, Mitchell et al. 2019). However, here we incorporate fossils within our hidden state speciation and extinction framework (HiSSE; Beaulieu and O'Meara, 2016), that includes any number of observed and/or hidden states. We focus our tests here exclusively using MiSSE (see Vasconcelos et al. 2021), a hidden state only model, like BAMM, but without priors (which could help, hurt, or have little effect on inference), but also, importantly, without the discontinuous inheritance of extinction probability that makes the likelihoods used in BAMM mathematically incorrect (see Moore et al. 2016). It is relatively straightforward to extend the equations of the canonical FBD model of Stadler (2010) to the state speciation and extinction models of Maddison et al. (2007) so that they also include both state transitions and the tree-wide rate of fossil sample rate parameter, $\psi$:

$$\frac{dE_i}{dt} = \mu_i - \left(\lambda_i - \mu_i - \psi + \sum_{i \neq j} q_{ij}\right)E_i(t) + \mu_i E_i(t)^2 + \sum_{i \neq j} q_{ij} E_j(t),$$

$$\frac{dD_i}{dt} = -\left(\lambda_i - \mu_i - \psi + \sum_{i \neq j} q_{ij}\right)D_i(t) + 2\lambda_i D_i(t) E_i(t) + \sum_{i \neq j} q_{ij} D_j(t). \quad (1)$$

The probability $E_i(t)$ is the probability that a lineage starting at time *t* in state *i* leaves no descendants at the present day (*t*=0), and $D_i(t)$ is the probability of a lineage in state *i* at time *t* before the present (*t*>0) evolved the exact branching structure as observed. These ordinary differential equations are generalized so that any number of observed or hidden states can be included in the model. For character-based models, like HiSSE and MuHiSSE, *i* and *j* represent the different observed and hidden states combinations specified in the model, whereas with MiSSE *i* and *j* represent hidden states only. For an extant tip, the initial condition for $D_i(0)$ is $\rho_i$,

which defines the probability that an extant individual observed in state $i$ is sampled in the tree and $1-\rho_i$ for $E_i(0)$. For an $m$-type fossil, $D_i(t)$=$\psi E_i(t_e)$, and $E_i(t)$=$E_i(t_e)$, where $t_e$ represents the time at which the sampled extinct lineage was sampled; for $k$-type fossil, $D_i(t)$=$\psi D(t)$, and $E_i(t)$=$E_i(t)$. At nodes, $N$, the initial condition is the combined probability of its two descendant branches, $L$ and $R$, such that, $D_{N,i} = D_{L,i}D_{R,i}\lambda_i$. The overall likelihood is the sum of $D_{N,i}$ calculated at the root.

When the fossil set represents stratigraphic ranges, the likelihood calculation is much more involved and requires designating three specific types of edge segments. First, we note that our implementation reflects the "symmetric speciation only" portion of Stadler et al. (2018; see Corollary 13), where only bifurcating speciation events are allowed — that is, a speciation event reflects the extinction of an ancestral species, and two new descendant species arise. This greatly reduces the complexity of the original model formulation, which at its most complex, also includes both anagenetic (ancestral species is replaced by a new descendant) and asymmetric (a single descendant "buds" off a persistent ancestral lineage) speciation processes with their own rates. Second, we note that a stratigraphic range is defined only by the oldest ($o$) and youngest ($y$) fossils that are unequivocally assigned to a species across a distinct time interval. Thus, a stratigraphic interval records the known duration of species on branches. These ranges are represented by a single fossil and a tip (i.e., fossil is $y$=0), by a branch fossil and an extinct tip ($y$=$t_e$), by two fossils representing a sampled ancestral stratigraphic range along an edge ($o$>$y$), or even by a single fossil (i.e., $o$=$y$). Fossils that occur within a given range are ignored.

When an edge segment is not associated with a stratigraphic range, the branch calculation and starting conditions proceed exactly as described in Eq. 1 above. However, when an edge segment represents a stratigraphic range, we must modify $D_i(t)$ to remove the possibility of an

unobserved speciation and subsequent extinction event within the interval given that we know that the entire segment [$o$,$y$] belongs to the same species:

$$\frac{d\widehat{D}_i}{dt} = -\left(\lambda_i - \mu_i - \psi + \sum_{i \neq j} q_{ij}\right)\widehat{D}_i(t) + \sum_{i \neq j} q_{ij}\widehat{D}_i(t). \quad (2)$$

When $y$=0, such that it is an extant species, the initial condition for $\widehat{D}_i(0)$ is $\rho_i$; if $y$ is an $m$-type fossil, $\widehat{D}_i(t)$=$\psi E_i(t_e)$; if $y$ denotes the beginning of a sampled ancestor stratigraphic range then $\widehat{D}_i(t)$=$D(t)$. Finally, for edge segments that represent intervening time intervals between two stratigraphic ranges without an observed speciation event, we modify the branch calculation to account for zero or more unobserved speciation events. Again, given the assumption that a stratigraphic range represents the sampling of a single species across a distinct interval of time, the presence of two stratigraphic ranges along the same edge would imply that at least one unobserved speciation event had occurred somewhere within this interval. To account for this, at the rootward end, $y_a$, which represents the youngest fossil of the older of the older of the two stratigraphic ranges, we correct $D_i(t)$ following the Stadler et al. (2018):

$$D_i(t) = D_i(y_a) * \left(1 - \frac{\widehat{D}_i(y_a)}{D_i(y_a)}\right) . \quad (3)$$

This probability then becomes the starting condition for the next stratigraphic range. Note that none of three specific branch types requires altering $E_i(t)$ from what is presented in Eq. 1, because this probability is based solely on time and is therefore unaffected by the tree topology. As before, at nodes, $N$, the initial condition is the combined probability of its two descendant branches, $L$ and $R$, such that, $D_{N,i} = D_{L,i}D_{R,i}\lambda_i$. The overall likelihood is again the sum of $D_{N,i}$ obtained at the root, but we must also marginalize over the number of fossils within a

stratigraphic range. This is done by multiplying the sum of $D_{N,i}$ at the root by $\psi^{K'}$ and $e^{\psi L_s}$, where $K'$ represents the total number of sampled fossils that represent the start and end times of a stratigraphic range (if $o=y$ for a given range then this counts as only one fossil), and where $L_s$ represents the sum of all stratigraphic range lengths ($\sum_{i=1}^{n} o_i - y_i$).

For both the canonical FBD, and FBD with stratigraphic ranges, we condition the overall likelihood by $\lambda_i[1 - E_i(t \mid \psi = 0)]^2$, where and $E_i(t \mid \psi = 0)$ is the probability that a descendant lineage of the root survived to the present and was sampled assuming no sampling in the past. We note that fossil BAMM does not condition on survival in this way, or even at all, when the tree includes extant taxa. For character-based models, such as HiSSE and MuHiSSE, we weight the overall likelihood by the probability that each possible state gave rise to the observed data (see FitzJohn et al. 2009). However, it was pointed out by Herrera-Alsina et al. (2018) that at the root, the individual likelihoods for each possible state should be conditioned prior to averaging the individual likelihoods across states. It is unclear to us which procedure is correct, but it does seem that both weighting schemes behave exactly as they should in the case of character-independent diversification — that is, the overall likelihood reduces to the likelihood of the tree multiplied by the likelihood of a trait model. We have also tested the behavior of both (not shown), and the likelihood differences are very small, and the parameter estimates in simulation are nearly indistinguishable from one another.

In the absence of character information, assuming a model with a single birth rate and death rate, and a $\psi>0$ the likelihoods are identical to that of Stadler (2010) and Stadler et al. (2018). To ensure that our math and the implementations are correct with two or more birth and death rates, we relied on the property of SSE models described by Beaulieu and O'Meara (2016) and Caetano et al. (2018): when a trait has no differential effect on the diversification process

(i.e., the character-independent model), the overall likelihood is the product of the tree likelihood and likelihood of a Markov model applied to the character data (it is the sum of the two likelihoods in log space). Thus, for a binary character that is independent of the diversification process on a tree with two rate classes and a non-zero fossil preservation rate, the BiSSE likelihood should reduce to the product of the likelihood of a MiSSE model with two rate classes and the likelihood of the transition rate model applied in the character data. As we show in the Supplemental Materials, we can confirm that our implementation of the FBD for MiSSE, HiSSE, and MuHiSSE (two binary characters; Nakov et al. 2019) are indeed correct. We also show that the different branch calculations for FBD with stratigraphic ranges conforms to the analytical calculations provided by Stadler et al. (2018). All models are available in the R package *hisse* (Beaulieu and O'Meara 2016).

**Impact of Fossils Under Various MiSSE Models**

Under a single rate model of birth and death, the inclusion of fossil data reduces uncertainty, but not necessarily the bias, in the estimates of functions of speciation and extinction, namely, the turnover rate ($\tau = \lambda + \mu$; Beaulieu and O'Meara 2016) and the extinction fraction ($\varepsilon = \mu/\lambda$). We explored these results further by conducting additional single rate regime simulations where trees were generated using a speciation rate of $\lambda$=0.4 and setting the extinction rate to take on three values, $\mu$= 0.1, 0.2, 0.3, to ensure three different values of extinction fraction, $\varepsilon$ = 0.25, 0.50, and 0.75, respectively. In each scenario, trees were grown until they reached 200 taxa. We then used a range of fossil sampling rates ($\psi$=0.01, 0.05, and 0.10 fossils/Myr), as well as varying the number of *k*-type fossils, where either only half were randomly chosen to be included in the analysis (i.e., half *k* fossils missing), none were included

at all (i.e., only *m*-type fossils included), or converted to stratigraphic ranges. When converting a fossil set into a set of stratigraphic ranges, we simply take the oldest and youngest fossil on each edge, removing all others, prior to pruning unsampled extinct lineages from the full tree.

In all cases, and across three metrics of diversification (turnover, extinction fraction, and net diversification), when comparing extant-only versus a full fossil sample under a given sampling scenario, there was minimal improvement in the rate estimates. Perhaps the most dramatic differences occurred when the extinction fraction was the smallest (i.e., $\varepsilon = 0.25$) and when the fossil sampling rates were highest ( $\psi$=0.10 fossils/Myr), which resulted in the full fossil set having both the bias and the variance being an order of magnitude smaller than estimates from extant-only. However, when extinction fraction was the highest, at $\varepsilon = 0.75$, the inclusion of fossils had very minor improvements in terms of bias and variance. When a portion of the *k*-type fossils were removed, there was a notable upward bias, but not the variance, in the rate estimates (see Table S1), which is consistent with the contour plot shown in Figure 2. We also note that when the fossil set is converted to stratigraphic ranges, there is rather severe downward bias in estimates of both turnover rate and extinction fraction as both extinction fraction and $\psi$ increases. Interestingly, bias in the estimates of net diversification remains nominal generally across every combination of generating parameters (see Table S1).

By contrast, Mitchell et al (2019) showed that with their fossil implementation within BAMM, which allows for shifts in discrete regimes of speciation and extinction, there were substantial improvements to the rate estimates, most notably with the extinction fraction. These results seem at odds with our results presented thus far. While their simulations were focused on understanding the behavior of a much more complex model, it should be noted that the most striking improvement in extinction fraction occurred when $\psi$=1 fossils/Myr, which is an

extremely high rate of preservation. In our case, the maximum value $\psi$=0.1 fossils/Myr resulted in a fossil set that was at least as numerous as the number of extant taxa ($N$>200), which seemed reasonable to us. Still, even when the addition of fossils resulted in doubling the amount of data available to the birth-death model, there was very little impact on the rate estimates, especially when the generating extinction fraction was high ($\varepsilon$ = 0.75). We also wondered how other functions of speciation and extinction, such as turnover rates and net diversification rates, are impacted not only by the inclusion of fossils, but also by the inclusion of biased samples. The simulations of Mitchell et al. (2019) demonstrated that estimates of speciation are insensitive to the inclusion of fossils, with extinction being strongly impacted, which suggests that this would also impact additional functions of speciation and extinction in different ways.

We explored this by performing simulations to test the behavior of our MiSSE model when fossils are included. As mentioned above, the underlying model is like BAMM in that it ignores character information altogether. Our "hidden state only" model assumes that a "shift" in diversification denotes a lineage tracking some unobserved, hidden state. We evaluated the performance of our MiSSE model by simulating trees under three scenarios and then estimating the fit and bias of the inferred rates from these trees. The first simulation scenario (scenario 1), we introduced two turnover rate regimes, where regime $B$ ($\tau_B$=0.50 events/Myr) had a turnover rate that was 2x the rate of regime $A$ ($\tau_A$=0.25 events/Myr). The extinction fraction was set at 0.75 for both regimes. In the second scenario (scenario 2), the rates were set such that there were two unique turnover rate and extinction fraction regimes. Specifically, we assumed regime $B$ ($\tau_B$=0.50 events/Myr) had a turnover rate that was 2x the rate of regime $A$ ($\tau_A$=0.25 events/Myr), but where regime $A$ had a lower extinction fraction ($\varepsilon_A$ =0.55) than regime $B$ ($\varepsilon_B$=0.75). With the turnover rates and extinction fractions structured in this way the net diversification rates between

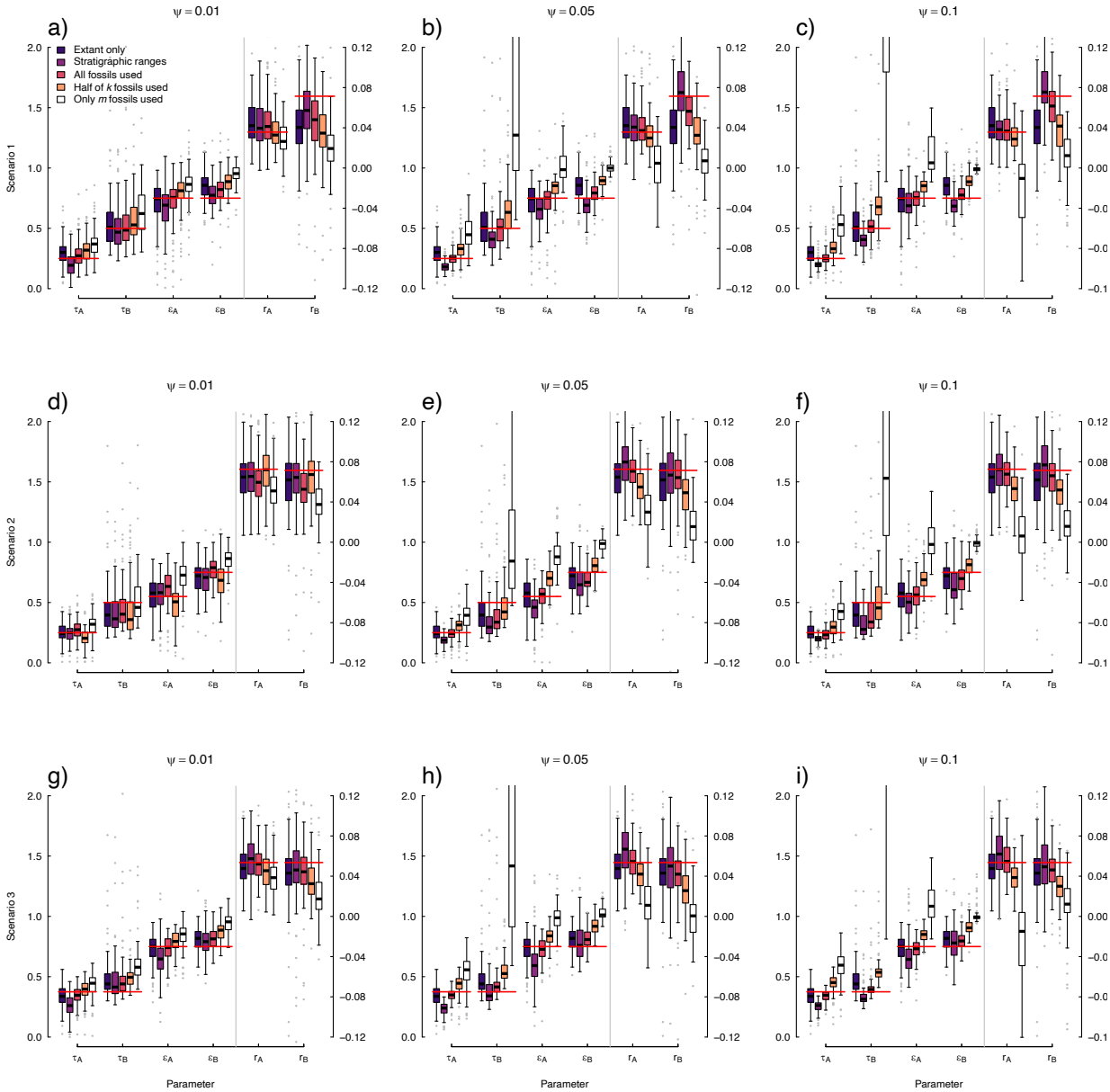


**Figure 3.** The uncertainty surrounding estimates of turnover ( $\tau_A$ and $\tau_B$), extinction fraction ( $\varepsilon_A$ and $\varepsilon_B$), and net diversification ($r_A$ and $r_B$) when the generating model assumes: (a-c) regime *B* had a turnover rate that was 2x the rate of regime *A* ($\tau_A$=0.25, $\tau_B$=0.50) and extinction fraction was set at 0.75 for both regimes; (d-f) there were two unique turnover rate and extinction fraction regimes, specifically, regime *B* had a turnover rate that was 2x the rate of regime *A* ($\tau_A$=0.25, $\tau_B$=0.50), but where regime *A* had a lower extinction fraction than regime *B* ($\varepsilon_A$ =0.55, $\varepsilon_B$=0.75); or (g-h) both regime *A* and *B* had identical turnover rates ($\tau_A$=$\tau_B$=0.375) and extinction fractions ($\varepsilon_A = \varepsilon_B$=0.75). In all cases, different assumptions about the fossils were fit to the same data set, and parameters were model averaged across a set of four models that variously allowed turnover and extinction fraction to vary (see text). The solid red lines represent the generating value for each parameter in each regime; the right y-axis represent the net diversification scale.

these regimes are nearly identical. Finally, we simulated a scenario (scenario 3) that assumed both regime *A* and *B* had identical turnover rates (*A*=*B*=0.375 events/Myr) and extinction fractions ($\varepsilon_A = \varepsilon_B$=0.75). This scenario was meant to examine the impact that including fossils has on the support for more complex models when the true model contains homogeneous diversification rates. In all three scenarios we assumed that the transition between regimes *A* and *B* was *q*=0.005 transitions/Myr.

For each simulation replicate, within each scenario, we fit four MiSSE models that broadly capture the complexity of the scenarios. Specifically, we fit a single regime model, a model that only allows turnover to vary ($\tau_A \neq \tau_B$), a model that only allowed extinction fraction to vary, and a model that allowed both to vary ($\tau_A \neq \tau_B$ and $\varepsilon_A \neq \varepsilon_B$). Each model was evaluated using a two-step optimization routine. The first step consists of a bounded stochastic simulated annealing run for 5000 iterations, followed by a bounded subplex routine that searches parameter space until the maximum likelihood is found. Since the state space allowed each model to be nested within the most complex model where turnover and extinction fraction were allowed to vary, we model-averaged the rate estimates using the AIC weights. However, before doing so, we culled the model set to remove any resulting redundant model fits. If the maximum likelihood estimates for, say, $\tau_A$ and $\tau_B$, in the turnover varying only model, take on the same value, it is essentially the same as including a single turnover rate model twice, and this would lower the weight of other models as a consequence. It is recommended in these situations to remove the more complex of the two from the set (see Burnham and Anderson 2003).

As with the single regime simulations above, the main effect of including fossils is to reduce the variance of the estimates, not the bias. In fact, scenario 2, which assumed differences in both turnover and extinction fraction, but not net diversification, proved difficult regardless of

whether fossils were included into the analyses. There was a consistent downward bias in the rates associated with shifts to regime *B* even as $\psi$ increased, meaning even when many fossils were included. With regards to converting the fossil set to stratigraphic ranges, there was, again, a general tendency for a downward bias in both turnover and extinction fraction. However, when compared to extant-only and a full fossil set assuming the canonical FBD, the estimates of net diversification for stratigraphic ranges generally performed better, albeit with higher variance associated with the estimates compared to the other two analysis types. We do note that with scenario 3, which assumed no rate differences among regimes *A* and *B*, the use of stratigraphic ranges consistently and erroneously inferred higher net diversification rates for regime *A*. Finally, removing *k* fossils completely from the set results in very high biases related to turnover rates specifically, as $\psi$ increases, and rather severe downward biases generally in estimates of net diversification rates.

**How Do Fossils Help in a World Where All Models are Simply Wrong?**

An easy and valid criticism of the simulations described above is that they are too simplistic. In other words, trees were generated under an SSE model that shifted between, at most, two different rate classes, with the rate of these shifts set by a single transition rate, *q*. While we were genuinely surprised that extant-only trees performed as well as trees that also included fossil sample data under the same conditions, the simulation scenarios are hardly realistic. The processes that generate most empirical trees are likely very complex, likely carrying signatures of non-random extinction, even mass extinctions, with diversification rates varying substantially among lineages and across time. The simplicity of our MiSSE models is out of mathematical convenience and tractability. Even still, we might expect that even when the

true model is not included in the set of models evaluated, the inclusion of fossil information attached to branches distributed across the tree would provide more weight and better parameter estimation for the more complex models that are included in the set. Most models will work fine with data that meets their assumptions, but with complex and messy data, maybe fossils are needed to rescue their performance.

To explore this further we devised a fourth simulation scenario that was meant to closely imitate processes and conditions that are more realistic in empirical settings, which would also prove challenging to our MiSSE model, but also be biased in a way that would favor including fossils. Specifically, we assumed that there were four discrete "regimes" of turnover rates ($\tau_A$= 0.3, $\tau_B$= 0.6, $\tau_C$= 1.5, $\tau_D$= 1.0 events/Myr) and extinction fractions ($\varepsilon_A$= 0.7, $\varepsilon_B$= 0.9, $\varepsilon_C$= 0.95, $\varepsilon_D$= 0.8) that controls the diversification dynamics. We also assumed a rather extreme heterogeneous transition matrix, **Q**, that governed the dynamics of transitions, $q_{ij}$, among these four regimes,

$$\mathbf{Q} = \begin{array}{c} A \\ B \\ C \\ D \end{array} \begin{bmatrix} - & 0.01 & 0.01 & 0.01 \\ 0.05 & - & 0.04 & 0.02 \\ 0.02 & 0.03 & - & 0.03 \\ 0.06 & 0.02 & 0.08 & - \end{bmatrix}$$

The magnitude of these rates was chosen somewhat at random, but cumulatively ensured that approximately 10% of the branches across the entire history of the tree (including extinct lineages) contained at least one change in rate regime. We also encoded two mass extinction events. The first occurred after the first 40-time units and removed, at random, 70% of lineages alive at that time point. The second occurred 30-time units later, 70-time units in total from the start of the simulation and removed 90% of the lineages alive at the time point (see Fig. 4a).

Finally, simulations were terminated once the tree reached 267 taxa. However, we non-randomly chose 67 taxa and removed them from the final tree. The purpose of this procedure was to assume that our final tree had a biased sampling fraction of 75%. The bias was generated by simulating a trait under Brownian motion, then normalizing the values so that they were between 0 and 1 and had a phylogenetic signal. These values were then used as probabilities for removing taxa.

We know fossilization rates vary, sometimes dramatically, by time and taxa (Wagner and Marcott 2013). Incorporating this would make the simulation far more realistic as well, but it would provide additional complications given the difficulties in separating variable extinction rates from variable preservation rates (e.g., Foote et al. 2019). The fact that the fossilization process matches that of the model gives the fossils a chance to perform better than not including them. That is, can well-modeled fossils help an analysis where the true diversification model is more complex and different than any analyzed? If fossils cannot help in this case, they probably will not help. Had we used a more complex fossilization process, and fossils failed to help in that case, it could just be that we tried too extreme a difference and set up a bias against the utility of fossils.

Our model set contained the same models as before (though now with $f_i$= 0.75 to account for incomplete sampling, though assuming, incorrectly, that all taxa have the same sampling rate; see Beaulieu 2020), thus ensuring that the generating model was not included in the model set. We also fit the same set of models for increasing values of $\psi$ (i.e., 0.01, 0.05, 0.10 events/Myr), and for different sampling of *k*-type fossils (i.e., full, half *k*, only *m*-type fossils used), including converting the set to stratigraphic intervals. We chose to summarize results based on the diversification rates model-averaged across only the tips that survived to the

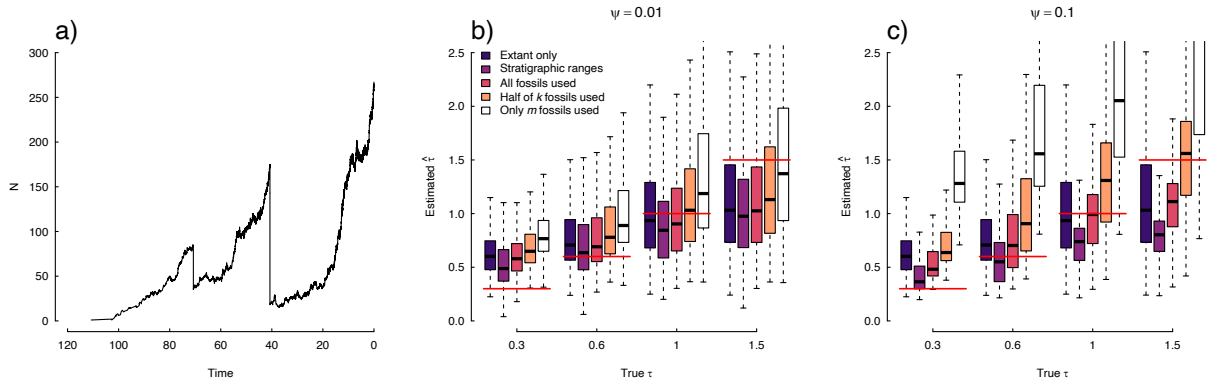


**Figure 4.** The uncertainty surrounding estimates of turnover when the models being evaluated does not include the true model. Specifically, we assumed that there were four discrete "regimes" of turnover rates ($\tau_A$= 0.3, $\tau_B$= 0.6, $\tau_C$= 1.5, $\tau_D$= 1.0) and extinction fractions ($\varepsilon_A$= 0.7, $\varepsilon_B$= 0.9, $\varepsilon_C$= 0.95, $\varepsilon_D$= 0.8) that controls the diversification dynamics; we also assumed a rather extreme transition dynamics among these four regimes (see text). We also encoded two mass extinction events (a), and we also assumed biased sampling among the tips that survived to the present. The boxplots represent the distribution of model-averaged rates based on the true rate classes at the tips. (b-c) Despite the true model being far outside the models being evaluated, the turnover rates at the tips at hints at there being more complexity within the data than even the most complex two-rate models in the set might suggest. However, the inclusion of fossils does not seem to perform remarkably better than not including them.

present, because the rate regimes in the generating model do not map to the rate regimes in the model set, namely, regime *D* in a four-rate model does not easily map to regime *A* or *B* in a two-rate model. For a given model, the marginal probability of each rate regime is obtained for every tip, and the rates for each regime are averaged together using the marginal probability as a weight: a weighted average of these rates is then obtained across all models using Akaike weights. The use of model-averaging is particularly advantageous here because the tip rates are not conditional on a single best model that would be far less complex than the model that generated the data. Different models reflect different aspects of the data. By fitting multiple models, each good model might tell us something about how the data evolved. Model-averaging is one way to summarize these different elements in a comprehensible way.

Despite the true generating model being far outside the set of models evaluated, the use of model-averaging at least hints at there being more complexity within the data than even the most complex two-rate models allow. When the model-averaged tip rates are aligned with the four true tip rate regimes for each simulation iteration there are consistent significant positive slopes across all treatments. For fossil-based analyses, particularly for those where the number of fossils outpaced the number of extant tips (i.e., $\psi$ = 0.10 fossils/Myr), there are clearer differences between the four rate regimes compared to the extant-only estimates. However, as with all other analyses presented here, when the fossil set is perfectly consistent with the FBD process by which the samples were taken, the main effect of the fossils is to reduce the overall uncertainty in the tip rate estimates compared to extant-only inferences.

There were some improvements with respect to the bias with the inclusion of the full fossil sets overall, most notably at two ends of the rate distribution under the generating model (i.e., $\tau_A$= 0.3 and $\tau_C$= 1.5 events/Myr). Incidentally, both rate classes were by far the most frequent states across the entire history of almost every simulated tree. The highest turnover rate regime also had the highest extinction fraction, and this coupled with several mass extinctions, resulted in very few survivors at the tips. Even with the addition of a substantial number of fossils placed throughout the tree, the improvement in inferring this particular rate category from model-average rates was minimal. From a qualitative standpoint, by ignoring biases in rate and focusing on comparisons among tips, it seems that all treatments, regardless of the fossil makeup, can consistently assign the correct sign differences among pairs of tip rates. That is, even when the trees were analyzed with only half of the *k*-type fossils used, if they were removed entirely, or converted to stratigraphic ranges, the expected proportion of sign differences are consistent with the extant-only and full fossil set estimates (Table S2). However, as the number

of *k*-type fossils are removed, the model-averaged estimates of turnover become greatly inflated, from erroneously high estimates of the extinction, and are associated with much greater uncertainty. Also, when the fossil sets are converted to stratigraphic ranges, the turnover rates are, again, generally underestimated, with the improvement of the estimates with the lower turnover regimes being simply incidental. Taken together, this is consistent with the MiSSE scenarios above (Fig. 3), as well as the constant birth-death models (Fig. 2e-f), which suggests that gross omissions of *k*-type samples in a FBD erroneously shifts estimates towards areas of very high turnover rates and extinction fraction. Moreover, converting to stratigraphic intervals results in erroneous shifts towards lower turnover rates and extinction fraction.

**A Caution About Stratigraphic Ranges**

Our simulation results highlight a curious pattern in that when a fossil set is converted to stratigraphic ranges, estimates of turnover and extinction fraction are increasingly underestimated as the generating $\psi$ increases. A clear illustration of the pattern is shown in the likelihood surfaces depicted in Figure 5a-b. When the generating $\psi$ is low ($\psi$=0.01 fossils/Myr; Fig. 5a), the likelihood surface includes the generating parameters. However, when $\psi$ is increased by an order of magnitude ($\psi$=0.1 fossils/Myr; Fig. 5a-b), the likelihood surface begins to shift away from the generating parameters towards regions of parameter space represented by lower turnover rates and extinction fraction. While this downward bias in this particular example is somewhat subtle, this behavior is consistent across all models and all simulation scenarios examined here. It should be noted that it is possible that we have incorrectly interpreted the math described in Stadler et al. (2018). Admittedly, the implementation of this model was quite challenging, and is further complicated by the fact that FBD with stratigraphic ranges is not

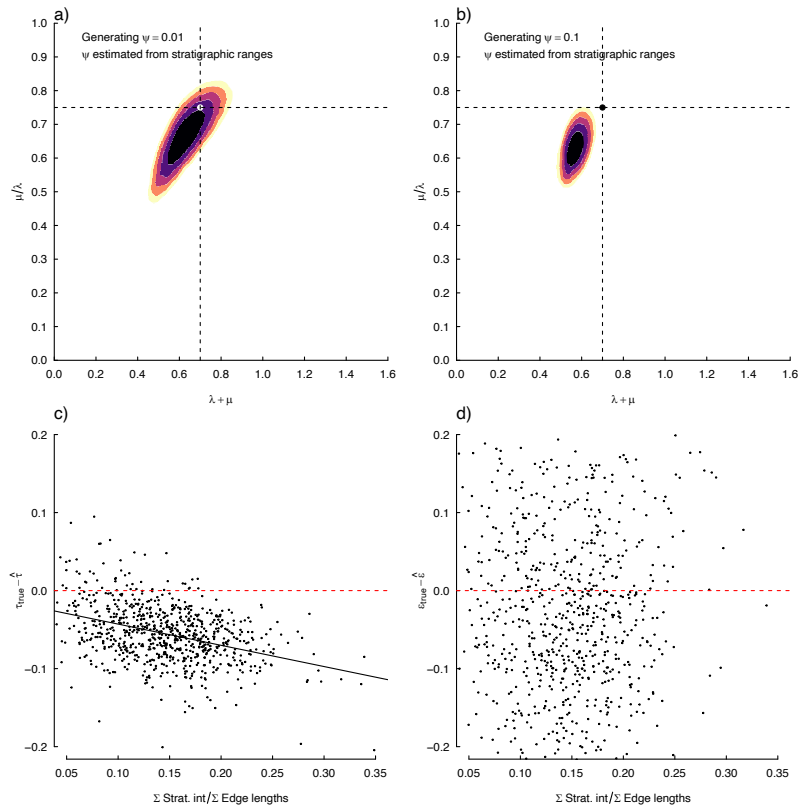


**Figure 5.** (a-b) Contour plots of the likelihood surface for the fossilized birth-death with stratigraphic ranges using the same tree and data set as shown in Figure 2. When the generating $\psi$ is low, the likelihood surface includes the generating parameters, but when $\psi$ is increased by an order of magnitude ($\psi$=0.1), the likelihood surface also begins to drift away from the generating parameters towards regions of parameter space represented by lower turnover rates and extinction fraction. (c) A stratigraphic range records the known duration of species on branches, and so we assume that no unobserved speciation or extinction takes place along these intervals. Since the same rates apply to all branches, and since turnover is a measure of the total number of speciation and extinction events, as the number of stratigraphic ranges and duration of each increase, turnover rates will increasingly be underestimated as the proportion of the stratigraphic ranges relative to the total tree length increases. Interestingly, this does not seem to impact estimates of extinction fraction (d).

naturally nested within the canonical FBD model. That is, if all fossils are treated as single representatives of a stratigraphic range, the likelihood does not simply revert to the likelihood of the canonical FBD model. This is because of the assumption about how to correct for unobserved speciation occurring along an intervening edge segment of two distinct stratigraphic ranges on the same edge. This makes unit testing of this model challenging. That being said, in

the constant birth-death case, we implemented the model using the analytic equations presented in Stadler et al. (2018) as well as using the ordinary differential equations above and the likelihoods were identical across a range of parameter values.

In the absence of an implementation error, we suspect that the underestimation of turnover is likely due to the assumptions made when computing the probability of the portion of an edge representing a stratigraphic range. Under this model a stratigraphic range records the known duration of species on branches, and so we assume that no unobserved speciation or extinction takes place along these intervals. Since the same rates apply to all branches, and since turnover is a measure of the total number of speciation and extinction events, we might expect that as the number of stratigraphic ranges and duration of each increase, the rate estimates would likely reflect a balance between a large proportion of time in a tree where no unobserved speciation and extinction events have occurred. Thus, we would predict that the downward bias in turnover rates would be a function of the proportion of the tree consisting of stratigraphic ranges. We conducted a simulation where we evolved a set of 100 trees for 50 Myr, assuming the same generating rates as the ones that produced the trees in Figure 2. We also used a range of fossil sampling rates starting from $\psi$=0.05 fossils/Myr and increasing at 0.01 increments until reaching $\psi$=0.15 fossils/Myr. Consistent with our hypothesis, we found that as the proportion of the stratigraphic ranges relative to the total tree length increased (i.e., stratigraphic ranges accounted for increasingly more of the time represented in the tree), we found that this was significantly and negatively associated with an increased downward bias in turnover rates ($R^2$=0.15, slope=-0.272, $P$<0.001). Interestingly, we did not find a similar trend with extinction fraction ($R^2$<0.001, slope=0.09, $P$=0.246).

## Concluding Remarks

Our results show that, if the trees fit the generating model well, then adding fossil taxa correctly might help with precision in rate estimations, but not with bias. However, for most cases if there is undersampling of fossils along branches (*k*-type fossils) the results are worse than ignoring fossils altogether, even when the placement of the sampled fossils was correct temporally and topologically. These results occurred in a simulation framework where the diversification models used were far simpler than the simulated data and where we expected fossils to be able to somewhat rescue the results given that they were nearly 3-times more numerous than extant species in certain scenarios (i.e., scenario 4).

We hasten to emphasize that fossils remain key for understanding macroevolution — that is, important extinct groups like trilobites, sauropods, and extinct lycophytes are essential to our understanding of evolution in deep time. However, the idea that neontological studies of diversification is dramatically enhanced by sprinkling carefully chosen fossils on a tree and applying sophisticated fossilized birth-death models to understand rate variation does not seem supported by our simulations. At best, fossils have minor effects. At worst, they lead to less accurate inferences than removing them altogether. We even carefully biased our study to purposely give fossils the best possible chance, namely, by perfectly identifying each fossil, placing them with full certainty in a phylogeny, and assuming that their true sampling rate was constant through time. All this is impossible with empirical data, which is likely the product of variable fossilization rates (Wagner and Marcott 2013) that further complicates teasing apart variable extinction from variable preservation rates (see Foote et al. 2019). Taken together, our simulations provide us with little hope for the utility of fossils in real-world applications, at least with regards to estimating diversification rate heterogeneity across a tree.

One area where we suspect fossils will continue to be of great importance in neontological applications is in molecular divergence time analyses. Much of the information used in fitting most diversification models is the timing between events. So, if the relaxed clock does not properly account for clades with a faster rate of molecular substitution, most methods to smooth branches will yield ages that are much older than they should be (since the branches are still too long relative to variation implicit in the clock; see Beaulieu et al. 2015), which will tend to also decrease the net diversification rate across the tree. However, having more fossils throughout the tree, we suspect, might alleviate this issue. It thus remains an open question whether methods that jointly estimate the topology, branching times, and diversification parameters using an FBD-type model give a better estimate of diversification rate parameters than alternate approaches (e.g., *r8s*; Sanderson 2002) that infer a chronogram agnostic to the diversification rate process, which is then used as input for subsequent diversification analyses. Our study simply demonstrates that in the case of a perfectly accurate (albeit perhaps undersampled) tree, fossils do not help substantially and, in some circumstances, can hurt.

One motivation for this research, in the light of recent and well-earned concern about diversification methods using phylogenies (Kubo and Iwasa, 1995; Maddison and FitzJohn, 2015; Rabosky and Goldberg, 2015; Louca and Pennell, 2020) is to help guide empiricists seeking to create improvements in diversification rate estimation. There is understandable interest in fossils to improve diversification models. In our own reading of the foundational work of Stadler (2010), which was motivated by our own curiosity about a path for improvement, we were struck by the *m* versus *k* fossil distinction and what this might mean for its usability in practice, as well as how stratigraphic intervals might mitigate this issue. We were surprised that, at least for our particular use case, fossils provided at best little benefit to, and at worst harmed,

inferences of diversification from trees of mostly modern taxa. This does not imply they may not help substantially in other areas of phylogenetics of extant taxa, but as a solution to the field's present diversification difficulties, they are less helpful than we hoped, and researchers might be more fruitfully rewarded attempting other solutions first.

**Funding**

This work was funded by the National Science Foundation grants DEB–1916558 and DEB-1916539.

**Supplementary Materials**

Data available from the following github repository:

https://github.com/thej022214/Fossils_impact_BO

**Acknowledgements**

We thank members of the Beaulieu and O'Meara labs for their comments and discussions of the ideas presented here. We would also like to thank Caroline Parins-Fukuchi and James Pease for edits at various stages of this work.

## References


Beaulieu J.M. 2020. The problem with clade-specific sampling fractions. Available at:https://rdrr.io/cran/hisse/f/inst/doc/Clade-specific-sampling.pdf.

Beaulieu J.M., O'Meara B.C. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Systematic Biology 65:583-601.

Beaulieu J.M., O'Meara B.C., Crane P., Donoghue M.J. 2015. Heterogeneous rates of molecular evolution and diversification could explain the Triassic age estimate for angiosperms. Systematic Biology 64:869-878.

Burnham K.P., and Anderson D.R. 2003. Model selection and multimodel inference: a practical information-theoretic approach. New York:Springer.

Caetano D.S., O'Meara B.C., Beaulieu J.M. 2018. Hidden state models improve state-dependent diversification approaches, including biogeographic models. Evolution 72:2308-2324.

Černý D., Madzia D., Slater G.J. 2021. Empirical and methodological challenges to the model-based inference of diversification rates in extinct clades. Systematic Biology https://doi.org/10.1093/sysbio/syab045.

FitzJohn R.G., Maddison W.P., Otto S.P. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Systematic Biology 58:595–611.

Foote M., Sadler P.M., Cooper R.A., Crampton J.S. 2019. Completeness of the known graptoloid palaeontological record. Journal of the Geological Societ 176:1038-1055.

Gavryushkina A., Heath T.A., Ksepka D.T., Stadler T., Welch D., Drummond A.J. 2017. Bayesian total-evidence dating reveals the recent crown radiation of penguins. Systematic Biology 66:57-73.

Heath T.A., Huelsenbeck J.P., Stadler T. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences, USA 111:E2957-E2966.

Herrera-Alsina L., van Els P., Etienne R.S. 2018. Detecting the dependence of diversification on multiple traits from phylogenetic trees and trait data. Systematic Biology, 68:317-328.

Kubo T., Iwasa Y. 1995. Inferring the rates of branching and extinction from molecular phylogenies. Evolution 49:694-704.

Louca S., Pennell M.W. 2020. Extant timetrees are consistent with a myriad of diversification histories. Nature 580:502-506.

Maddison W.P., R.G. FitzJohn. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. Systematic Biology 64: 127-136.

Maddison W.P., Midford P.E., Otto S.P. 2007. Estimating a binary character's effect on speciation and extinction. Systematic Biology 56:701–710.

Marsh O.C. Principal characters of American Jurassic dinosaurs. Part IX: The skull and dermal armor of Stegosaurus. American Journal of Science 34: 413–417.

Mitchell J.S., Etienne R.S., Rabosky D.L. 2019. Inferring diversification rate variation from phylogenies with fossils. Systematic Biology 68:1-18.

Moore B.R., Hohna S., May M.R., Rannala B., Huelsenbeck J.P. 2016. Critically evaluating the theory and performance of Bayesian analysis of macroevolutionary mixtures. Proceedings of the National Academy of Sciences, USA 113:9569-9574.

Nakov T., Beaulieu J.M., Alverson A.J. 2019. Diatoms diversify and turn over faster in freshwater than marine environments. Evolution 73:2497-2511.

Nee S., Holmes E.C., May R.M., Harvey P.H. 1994. Extinction rates can be estimated from molecular phylogenies. Philosophical Transactions of the Royal Society B 344:77–82.

Osborn H.F. 1905. Tyrannosaurus and other Cretaceous carnivorous dinosaurs. Bulletin American Museum of Natural History 21:259-265.

Parins-Fukuchi C. 2021. Morphological and phylogeographic evidence for budding speciation: an example in hominins. Biology Letters 17, doi: https://doi.org/10.1098/rsbl.2020.0754

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

Rabosky D.L., Goldberg E.E. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. Systematic Biology 64:340–355.

Rabosky D.L., Goldberg E.E. 2017. FiSSE: A simple nonparametric test for the effects of a binary character on lineage diversification rates. Evolution, 71:1432–1442.

Sanderson M.J. 2002. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. Molecular Biology Evolution 14:1218–1231.

Stadler T. 2010. Sampling-through-time in birth-death trees. Journal of Theoretical Biology 267:396-404.

Stadler T., Gavryushkina A., Warnock R.C.M., Drummond A.J., Heath T.A. 2018. The fossilized birth-death model for the analysis of stratigraphic range data under different speciation modes. Journal of Theoretical Biology 447:41-55.

Vasconcelos T., O'Meara B.C., Beaulieu J.M. 2021. Hidden state-only speciation and extinction models provide accurate tip-estimates of diversification rates. Submitted.

Wagner P.J., Marcot J.D. 2013. Modelling distributions of fossil sampling rates over time, space and taxa: assessment and implications for macroevolutionary students. Methods in Ecology and Evolution 4:703-713.

## Figure Captions

**Figure 1.** a) An example of a fossil set sampled from a complete tree generated by the birth-death process, which includes both surviving and extinct members. b) The sampled phylogeny pruned down from (a) to only includes sampled tips and fossils. The fossilized birth-death model distinguishes between two types of sampled lineages. An *m* fossil (denoted by a blue dot) represents sampled branches that went completely extinct before the present and did not give rise to any additional sampled lineages (i.e., "childless" fossils). A *k* fossil (denoted by an orange dot) represents samples of lineages that had sampled descendants (other fossils, or extant tips; "parent" fossils).

**Figure 2.** Contour plots of the likelihood surface under various scenarios for the same tree (topology shown in d) simulated under a birth-death process. Each surface is constrained such that turnover ($\tau = \lambda + \mu$) and extinction fraction ($\varepsilon = \mu/\lambda$) are fixed, but $\psi$, the rate of fossil preservation, is free to find their MLE given a pair of fixed parameter values. When provided with a sample of extinct lineages that is perfectly consistent with the generating model, the FDB performs and behaves well and generally reduces the variance in turnover and extinction fraction (b, c), relative to ignoring fossils completely (a). However, if *k* fossils are removed completely (e,f), the likelihood surface begins to erroneously shift away from the generating parameters towards regions of parameter space of very high extinction rates as values of $\psi$ increase. The dashed vertical line represents the generating value for turnover ($\tau$=0.70) and the dashed horizontal line represents the generating value for extinction fraction ($\varepsilon$=0.75).

**Figure 3.** The uncertainty surrounding estimates of turnover ( $\tau_A$ and $\tau_B$), extinction fraction ( $\varepsilon_A$ and $\varepsilon_B$), and net diversification ($r_A$ and $r_B$) when the generating model assumes: (a-c) regime *B* had a turnover rate that was 2x the rate of regime *A* ($\tau_A$=0.25, $\tau_B$=0.50) and extinction fraction was set at 0.75 for both regimes; (d-f) there were two unique turnover rate and extinction fraction regimes, specifically, regime *B* had a turnover rate that was 2x the rate of regime *A* ($\tau_A$=0.25, $\tau_B$=0.50), but where regime *A* had a lower extinction fraction than regime *B* ($\varepsilon_A$ =0.55, $\varepsilon_B$=0.75); or (g-h) both regime *A* and *B* had identical turnover rates ($\tau_A$=$\tau_B$=0.375) and extinction fractions ($\varepsilon_A = \varepsilon_B$=0.75). In all cases, different assumptions about the fossils were fit to the same data set, and parameters were model averaged across a set of four models that variously allowed turnover and extinction fraction to vary (see text). The solid red lines represent the generating value for each parameter in each regime; the right y-axis represent the net diversification scale.

**Figure 4.** The uncertainty surrounding estimates of turnover when the models being evaluated does not include the true model. Specifically, we assumed that there were four discrete "regimes" of turnover rates ($\tau_A$= 0.3, $\tau_B$= 0.6, $\tau_C$= 1.5, $\tau_D$= 1.0) and extinction fractions ($\varepsilon_A$= 0.7, $\varepsilon_B$= 0.9, $\varepsilon_C$= 0.95, $\varepsilon_D$= 0.8) that controls the diversification dynamics; we also assumed a rather extreme transition dynamics among these four regimes (see text). We also encoded two mass extinction events (a), and we also assumed biased sampling among the tips that survived to the present. The boxplots represent the distribution of model-averaged rates based on the true rate classes at the tips. (b-c) Despite the true model being far outside the models being evaluated, the turnover rates at the tips at hints at there being more complexity within the data than even the most complex two-rate models in the set might suggest. However, the inclusion of fossils does not seem to perform remarkably better than not including them.

**Figure 5.** (a-b) Contour plots of the likelihood surface for the fossilized birth-death with stratigraphic ranges using the same tree and data set as shown in Figure 2. When the generating $\psi$ is low, the likelihood surface includes the generating parameters, but when $\psi$ is increased by an order of magnitude ($\psi$=0.1), the likelihood surface also begins to drift away from the generating parameters towards regions of parameter space represented by lower turnover rates and extinction fraction. (c) A stratigraphic range records the known duration of species on branches, and so we assume that no unobserved speciation or extinction takes place along these intervals. Since the same rates apply to all branches, and since turnover is a measure of the total number of speciation and extinction events, as the number of stratigraphic ranges and duration of each increase, turnover rates will increasingly be underestimated as the proportion of the stratigraphic ranges relative to the total tree length increases. Interestingly, this does not seem to impact estimates of extinction fraction (d).