How to validate a Bayesian evolutionary model

Fábio K. Mendes[1†*], Remco Bouckaert[2†],
Luiz M. Carvalho[3†], Alexei J. Drummond[4]

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

[2]School of Computer Science, The University of Auckland

[3]Escola de Matemática Aplicada, Fundação Getulio Vargas

[4]School of Biological Sciences, The University of Auckland

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

[†]Authors contributed equally to this work

November 11, 2023

Abstract

Biology has become a highly mathematical discipline in which probabilistic models play a central role. As a result, research in the biological sciences is now dependent on computational tools capable of carrying out complex analyses. These tools must be validated before they can be used, but what is understood as validation varies widely among methodological contributions. This may be a consequence of the still embryonic stage of the literature on statistical software validation for computational biology. Our manuscript aims to advance this literature. Here, we describe and illustrate good practices for assessing the correctness of a model implementation, with an emphasis on Bayesian methods. We also introduce a suite of functionalities for automating validation protocols. It is our hope that the guidelines presented here help sharpen the focus of discussions on (as well as elevate) expected standards of statistical software for biology.

[Probabilistic model, Bayesian model, model validation, coverage]

Introduction

The last two decades have seen the biological sciences undergo a major revolution. Critical technological innovations such as the advent of massive parallel sequencing and the accompanying improvements in computational power and storage have flooded biology with unprecedented amounts of data ripe for analysis. Not only has intraspecific data from multiple individuals allowed progress in fields like medicine and epidemiology (e.g., The 1000 Genomes Project Consortium, 2015; Human Microbiome Project Consortium, 2012; Neafsey et al., 2015), population genetics (e.g., Lynch, 2007; Lack et al., 2016; de Manuel et al., 2016) and disease ecology (e.g., Rosenblum et al., 2013; Bates et al.,

2018), but now a large number of species across the tree of life have had their genomes sequenced, furthering our understanding of species relationships and diversification (e.g., Pease et al., 2016; Kawahara et al., 2019; Upham et al., 2019). Almost on par with with data accumulation is the rate at which new computational tools are being proposed, as evidenced by journals entirely dedicated to method advances, methodological sections in biological journals, and computational biology degrees being offered by institutions around the world.

One extreme case is the discipline of evolutionary biology, on which we focus our attention. While it could be said that many decade-old questions and hypotheses in evolutionary biology have aged well and stood up the test of time (e.g., the Red Queen hypothesis, Van Valen 1973; Lively 1987; Morran et al. 2011; Gibson and Fuentes-González 2015; the Bateson-Dobzhansky-Muller model, Dobzhansky 1936; Muller 1940; Hopkins and Rausher 2012; Roda et al. 2017), data analysis practices have changed drastically in recent years, to the point they would likely seem exotic and obscure to an evolutionary biologist active forty years ago. In particular, evolutionary biology has become highly statistical, with the development and utilization of probabilistic models now being commonplace.

Models are employed in the sciences for many reasons, and fall within a biological abstraction continuum (Servedio et al., 2014), going from fully verbal, highly abstract models (e.g., Van Valen 1973), through proof-of-concept models that formalize verbal models (e.g., Maynard Smith 1978; Reinhold et al. 1999), to models that interact directly with data through explicit mathematical functions (Yule, 1924; Felsenstein, 1973; Hasegawa et al., 1985; Hudson, 1990). Within the latter category, probabilistic models have seen a sharp surge in popularity within evolutionary biology, in conjunction with computational tools implementing them.

Despite the increasing pervasiveness of probabilistic models in the biological sciences, tools implementing such models show large variation not only with respect to code quality (from a software engineering perspective), but also to the provided evidence for correctness (Darriba et al., 2018). This is unsurprising given the challenges in funding software research (Siepel, 2019), and the multidisciplinary nature of method development. Much of the relevant information regarding good coding and statistical practices is out of reach of the average computational biologist, as it is spread across a variety of specialized sources, often obfuscated by its technical and theoretical presentation. The bioinformatics community is thus in dire need of synthetic and accessible resources that provide guidance for code improvement and validation.

Here, we summarize best practices in probabilistic model validation for method developers, with an emphasis on Bayesian methods. We execute two different validation protocols on variations of a simple phylogenetic model, discuss the results, and expand on how to interpret other potential outcomes. We further introduce a suite of methods for automating these protocols within the BEAST 2 platform (Bouckaert et al., 2019). Lastly, we propose method development guidelines for new model contributions, for researchers and reviewers who expect new software to meet not only a desirable standard, but also a reasonable one.

Probabilistic models

Probabilistic models mathematically formalize natural phenomena having an element of randomness. This is done through probability distributions describing both the observed empirical data – seen as the result of one or more random instantiations of the modeled process – as well the model parameters, which abstract relevant, but usually unknown aspects of the phenomenon at hand. In the domain of evolutionary biology specifically, the historical, stochastic, and highly dimensional nature of evolutionary processes makes the utility of probabilistic models self-evident.

The central component of a probabilistic model, $\Pr(D = d|\Theta = \theta)$, allows us to describe the probability distribution over the data (D) given the model parameters (Θ). This probability mass function (pmf; or its continuous counterpart, the probability density function, pdf, $f_D(d|\Theta = \theta)$) is sometimes referred to as the likelihood function. Just for this section, we will abuse and simplify notation, and drop variable subscripts, e.g., we will write $f_D(d|\Theta = \theta)$ as $f(d|\theta)$. As illustrated in the next sections, probabilistic models can be hierarchical, in which case there may be several likelihood functions. In a frequentist statistical framework, $f(d|\theta)$ is the sole component of an inferential procedure, and is maximized across parameter space during parameter estimation and model comparison.

In the present study we focus on Bayesian inference, where a probabilistic model $\mathcal{M}$ defines a posterior probability distribution for its parameters, $f(\theta|d) = \frac{f(d|\theta)f(\theta)}{f(d)}$. Here, our prior inferences or beliefs about the natural world – represented by the prior distribution $f(\theta)$ – are confronted with and updated by the data through the likelihood function. $f(d) = \int_\Theta f(d|\theta)f(\theta)d\theta$, the probability of the data, is also known as the marginal likelihood or the model evidence. Crucially, a Bayesian model includes a prior, $f(\theta)$: when models are compared, for example, $f(\theta)$ needs to be taken into account when computing the model evidence $f(d)$.

Models routinely used in evolutionary biology are often characterized by continuous parameters, and are normally complex enough to preclude analytical solutions for the posterior density $f(\theta|d)$, mainly due to the intractability of the integral appearing in the denominator – i.e., the marginal likelihood. In those cases, one can make use of the fact that $f(d)$ is a constant with respect to the parameters that can be ignored (i.e., $f(\theta|d) \propto f(\theta|d)f(\theta)$), and use techniques like Markov chain Monte Carlo (MCMC) to sample the posterior distribution. This is because MCMC is usually implemented in the form of the Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) algorithm, which only requires the posterior to be evaluated up to a constant.

In practice, the Metropolis-Hastings algorithm samples the posterior distribution (also referred to as the "target" distribution) by means of a transition mechanism. If the proposal distribution generated by this mechanism is irreducible, positive recurrent, and aperiodic, and the resulting chain is long enough, then the sampled posterior distribution will closely approximate the target distribution $f(\theta|d)$ (Smith and Roberts, 1993; Tierney, 1994; Gelman

et al., 2013).

We will spend time considering MCMC in particular, as it is the commonly chosen technique for obtaining samples from $f(\theta|d)$ under an implementation of model $\mathcal{M}$. A thorough validation effort thus entails verifying the correctness of (i) the model (i.e., $f(d|\theta)f(\theta)$), and (ii) the components involved in the MCMC transition mechanism. We note that the latter are not part of the model, however, and it is possible to sample $f(\theta|d)$ with other techniques such as importance sampling, Laplace approximations (Rue et al., 2009), or even by converting the sampling problem into an optimization one (e.g., Zhang and Matsen, 2019).

Finally, we stress that we are interested in practices for verifying model *correctness*. There are other tests and diagnostics employed to ensure that a particular MCMC analysis is converging as expected. Ascertaining whether one or more independent Markov chains have converged to a given posterior distribution is not a correctness test, as that distribution might be very different from the target distribution. We refer the reader interested in these and related topics to Warren et al. (2017), Fabreti and Höhna (2022), Magee et al. (2023) and references therein.

Validating a Bayesian model

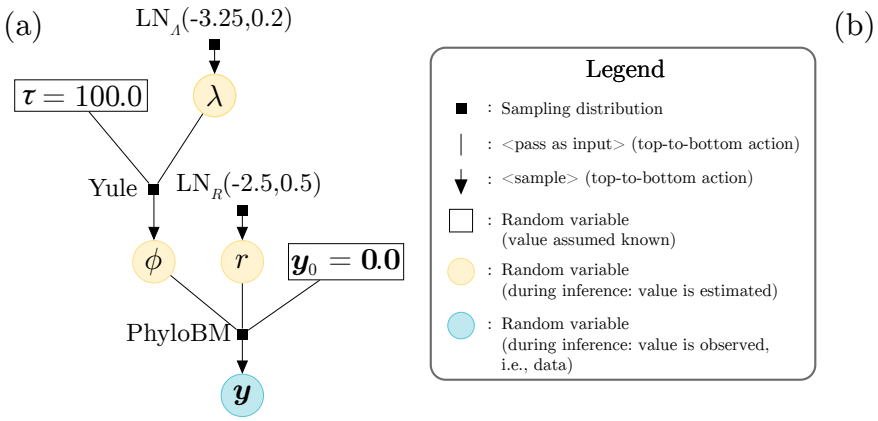
In this section we discuss procedures for validating an implementation of a Bayesian model $\mathcal{M}$. Whenever necessary, we will differentiate between a model implemented as a simulator ($\mathrm{S}[\mathcal{M}]$) and as a tool for inference ($\mathrm{I}[\mathcal{M}]$). Both $\mathrm{S}[\mathcal{M}]$ and $\mathrm{I}[\mathcal{M}]$ must be inspected in order to validate a model $\mathcal{M}$.

Validating the simulator, $\mathbf{S}[\mathcal{M}]$

When a new probabilistic model $\mathcal{M}$ is introduced an its inferential engine, $\mathrm{I}[\mathcal{M}]$ – what users employ in empirical analyses – is implemented for the first time, validating $\mathrm{I}[\mathcal{M}]$ requires that a simulator for $\mathcal{M}$, $\mathrm{S}[\mathcal{M}]$, be devised and itself validated. A simulator conventionally requires a parameter value as input (i.e., a value for Θ, θ, where θ might represent the values of more than one parameter), or a prior distribution on those values, $f_\Theta(\cdot)$. Note that we use "$\cdot$" when referring specifically to the generative function, rather than the value it takes given input. The simulator then outputs a sample of random variable(s), which for hierarchical models will include not only an instantiation d of data D, but also the parameters in Θ.

In the case of hierarchical models, it is sometimes useful to consider $\mathrm{S}[\mathcal{M}]$ as a collection of component simulators, each characterized by a different sampling distribution. For example, $\mathrm{S}[\mathcal{M}]$ for model we will work with below (Fig. 1; Table 1) can be seen as an ensemble comprised by:

1. $\mathrm{S}[f_\Theta(\cdot)]$ (where $\Theta = \{T, \Lambda, R, \boldsymbol{Y_0}\}$), which jointly simulates $\theta = \{\tau, \lambda, r, \boldsymbol{y_0}\}$,



Node name	Meaning
λ	Birth rate
τ	Origin age
ϕ	Phylogeny
r	Morphological rate
$\boldsymbol{y}_0$	Character state at origin
$\boldsymbol{y}$	Character states at tips

Figure 1: A simple probabilistic graphical (Bayesian) model we will validate in this work. (a) When read from top to bottom, the graphical model describes a generative process (see the legend for the meaning of vertical lines and downward-pointing arrows). If read from bottom to top, the graphical model describes the process of inference (assuming arrows having opposite orientation denoting the flow of information); in this case, the blue and yellow circles represent the data and the parameters being estimated, respectively. A random variable within a rectangular box signifies a parameter whose value is assumed known by the user; these are normally nuisance hyperparameters, or parameters that are not of immediate interest perhaps because they have been estimated elsewhere. (b) Each random variable node in the model, and how they should be interpreted. Table 1 presents more detail on each of the sampling distributions. Briefly, "LN" stands for log-normal, "Yule" for a Yule process also known as a pure-birth model, and "PhyloBM" stands for a phylogenetic Brownian motion model.

2. S$[f_{\Phi|T,\Lambda}(\cdot|T=\tau,\Lambda=\lambda)]$, which simulates a Yule tree ϕ given an origin age value τ and a λ (the birth-rate) simulated in (1),

3. S$[f_{Y|\Phi,R,\boldsymbol{Y_0}}(\cdot|\Phi=\phi,R=r,\boldsymbol{Y_0}=\boldsymbol{y_0})]$, which simulates an array with n continuous-trait values (one value per species), $\boldsymbol{y}$, given a phylogeny ϕ with n species, an evolutionary rate r and ancestral character values $\boldsymbol{y_0}$ (simulated in (1) and (2), respectively).

Being able to isolate the building blocks of a hierarchical model simulator helps divide and conquer the validation task, especially when some but not all of the sampling distributions are well-known parametric distributions, or when they result from well characterized stochastic processes (see below).

Table 1: Sampling distributions used in the probabilistic model validated in this work (Fig. 1). Columns "During simulation" and "During inference" specify how the sampling distributions should be read and interpreted, following the notation in the main text.

Label (Fig. 1)	Full name or alias	During simulation	During inference
$\text{LN}_\Lambda(-3.25, 0.2)$	Log-normal	$f_{\Lambda\|M_\Lambda,\Sigma_\Lambda}(\cdot\|M_\Lambda{=}-3.25,\Sigma_\Lambda{=}0.2)$	$f_{\Lambda\|M_\Lambda,\Sigma_\Lambda}(\lambda\|M_\Lambda{=}-3.25,\Sigma_\Lambda{=}0.2)$
$\text{LN}_R(-3.25, 0.2)$	Log-normal	$f_{R\|M_R,\Sigma_R}(\cdot\|M_R{=}-2.5,\Sigma_R{=}0.5)$	$f_{R\|M_R,\Sigma_R}(\lambda\|M_R{=}-2.5,\Sigma_R{=}0.5)$
Yule	Pure-birth	$f_{\Theta\|\mathcal{O},\Lambda}(\cdot\|T{=}\tau,\Lambda{=}\lambda)$	$f_{\Theta\|\Lambda}(\lambda\|T{=}\tau,\Lambda{=}\lambda)$
PhyloBM	Phylogenetic Brownian motion	$f_{\boldsymbol{Y}\|\Theta,R,Y_0}(\cdot\|\Theta{=}\theta,R{=}r,Y_0{=}y_0)$	$f_{\boldsymbol{Y}\|\Theta,R,Y_0}(\boldsymbol{y}\|\Theta{=}\theta,R{=}r,Y_0{=}y_0)$

One way to validate a probabilistic model simulator is by using it to produce (sample) a large number of data sets given a set of parameters. These parameters can be seen as characterizing a "population" of the entities being modeled.

For each data set, one can then construct $\alpha \times 100\%$-confidence intervals (where $\alpha \in (0,1)$ gives the confidence level) for certain summary statistics (e.g., mean, variance, covariance). If the simulator is behaving as expected, one should be able to verify that the (population or "true") summary statistic is contained approximately $\alpha\%$ of the time within their $\alpha \times 100\%$-confidence intervals. An example is the Yule model (also known as the pure-birth model; Yule 1924), a continuous-time Markov process that has been classically employed in phylogenetics to model the number of species in a clade (Yule, 1924; Aldous, 2001). Under a Yule process with a species birth rate of λ, the expected tree height, $\mathrm{E}[t_{\text{root}}]$, for a tree with n tips is:

Box 1: Models characterized by well-known parametric distributions

One commonly used model in macroevolution for the study of continuous traits is the phylogenetic Brownian motion model ("PhyloBM" in Fig. 1; Felsenstein 1973). The pdf characterizing this model's sampling distribution is in fact the pdf of the multivariate normal (MVN) probability distribution:

$$\begin{aligned} \log f(\boldsymbol{y} \mid \boldsymbol{y_0}, r, \boldsymbol{T}) = &-\frac{1}{2}\Big[n\log(2\pi) + \log|r\boldsymbol{T}|\Big] \\ &-\frac{1}{2}\Big[(\mathbf{y}-\boldsymbol{y_0})^T(r\boldsymbol{T})^{-1}(\mathbf{y}-\boldsymbol{y_0})\Big], \end{aligned} \tag{1}$$

where $\boldsymbol{y}$ corresponds to the observed values of a trait scored for n species, $\boldsymbol{y_0}$ is the trait value at the root of the tree, r is the instantaneous rate of change (i.e., the evolutionary rate, and sometimes represented by σ^2), and $r\boldsymbol{T}$ is the variance-covariance matrix. $\boldsymbol{T}$ is a matrix whose elements are deterministically defined by tree ϕ's topology and branch lengths; see Fig. 1 below).

The probability density function in equation (1) describes the distribution that would result from an infinite number of BM "experiments" (each experiment being non-mean-reverting, and representing an independent evolutionary trajectory). Under this model, $\theta = \{\boldsymbol{y_0}, r, \boldsymbol{T}\}$ and $d = \{\boldsymbol{y}\}$ (but note that sometimes researchers treat ϕ and consequently $\boldsymbol{T}$ as data).

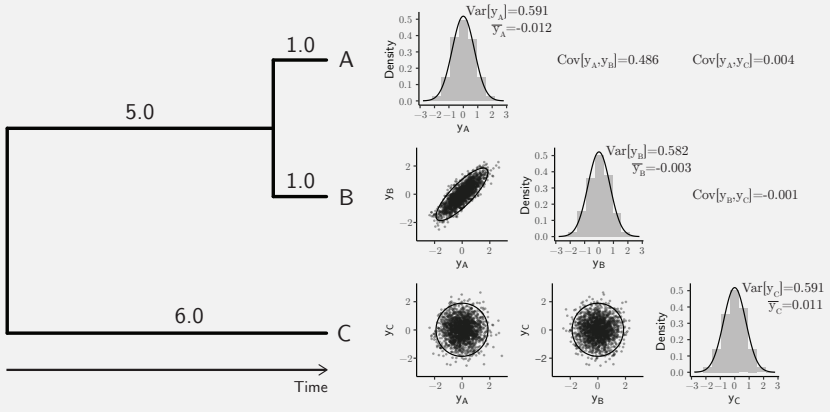


Figure 2: A sample of 1000 draws from a MVN distribution, each representing the evolutionary trajectory of one continuous trait along the species tree on the left. The root trait value, $\boldsymbol{y_0}$, and the evolutionary rate of the process, r, were set to 0.0 and 0.1, respectively. The panel on the right shows histograms of 1000 trait values sampled from the MVN for each species, as well as their covariation.

Validating a phylogenetic BM simulator

The MVN is a well-characterized parametric distribution. When used as the sampling distribution of the phylogenetic BM process, it explicitly defines the expected trait value for each species ($\boldsymbol{y_0}$), as well as their trait value variances and covariances. The latter comes from the variance-covariance matrix; for the tree shown in Figure 2 and with $r = 0.1$, this matrix is:

$$r\boldsymbol{T} = 0.1 \begin{bmatrix} 6 & 5 & 0 \\ 5 & 6 & 0 \\ 0 & 0 & 6 \end{bmatrix} \tag{2}$$

Together, variance-covariance matrix $r\boldsymbol{T}$ and $\boldsymbol{y_0} = [0.0, 0.0, 0.0]$ characterize a population of phylogenetically related species trait values whose means are 0.0, variances are 6.0, and co-variances are 5.0 (between species "A" and "B") and 0.0 (between species "C"and either "A" or "B").

Figure 1 shows the distributions of trait values and their variances and covariances for one sample of one thousand independent realizations of phylogenetic BM processes. One can see that the sample's average trait value and the variances and covariances approach their expectations. More rigorously, one can follow the method described in the main text and verify that those expectations fall within their 95% confidence intervals 95% of the time, as calculated from a large number of samples (Supplementary Fig. 1 and Supplementary Table 1).

$$\mathrm{E}[t_{\mathrm{root}}] = \sum_{i=2}^{n} \frac{1}{i\lambda}. \tag{3}$$

One can then verify, for example, if $\mathrm{E}[t_{\mathrm{root}}]$ is 95% of the time within ± 1.96 standard errors of the average Yule-tree height (from each sampled data set). Confirming that this is the case indicates $\mathrm{S}[f_{\Phi|\Lambda}(\cdot|T = \tau, \Lambda = \lambda)]$ is correctly implemented (Fig. 3). In Box 1, we illustrate this procedure for the (parametric) sampling distribution underlying the phylogenetic Brownian motion model ("PhyloBM"; Felsenstein, 1973). Protocols for validating $\mathrm{I}[\mathcal{M}]$ (see below) will also normally validate $\mathrm{S}[\mathcal{M}]$ at the same time.

We note that we have so far used $\mathrm{S}[\mathcal{M}]$ to represent a *direct* simulator under model $\mathcal{M}$ (Table 2), meaning each and every sample generated by $\mathrm{S}[\mathcal{M}]$ is independent. This is contrast with other simulation strategies, such as conducting MCMC under model $\mathcal{M}$ with no data (i.e., "sampling from the prior"), given specific parameter values (θ). This latter approach may be the only option if $\mathrm{S}[\mathcal{M}]$ has not been yet implemented, and it is predicated upon the existence of correct implementations of both an inferential engine $\mathrm{I}[\mathcal{M}]'$ and of proposal functions. We distinguish $\mathrm{I}[\mathcal{M}]'$ from $\mathrm{I}[\mathcal{M}]$ because simulations are being carried out precisely to validate $\mathrm{I}[\mathcal{M}]$. Unless MCMC simulations are done with $\mathrm{I}[\mathcal{M}]'$ – an independent implementation of $\mathrm{I}[\mathcal{M}]$ – they can introduce circularity to the validation task.

(a)

Birth rate (λ)	$\mathrm{E}[t_{\mathrm{root}}] \in$ 95% CI (%)
0.5	94
0.6	91
0.7	98
0.8	95
0.9	96
1.0	92

(b)

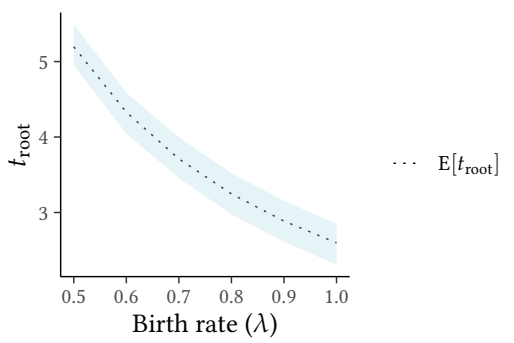


Figure 3: Validation of Yule tree simulator. (a) Number of simulated data sets (out of 100) for which the expected tree height (t_{root}) was inside the 95% CI about its sample average – if the simulator is correct, we expect this number to be between 90 and 99 about 95% of the time. Each data set consisted of 50 twenty-taxon simulated Yule trees. (b) The area shaded in light blue represents the 95% confidence interval about the average tree height, obtained from the 5,000 Yule trees simulated in (a). Simulations were carried out with the `TreeSim` R package (Stadler, 2011).

Table 2: A non-exhaustive list of direct simulation software used for various models in evolutionary biology.

Software package	Model type	Platform	Reference
Seq-Gen	Molecular sequence evolution models	Standalone	Rambaut and Grass, 1997
ms	Coalescent model	Standalone	Hudson, 2002
msprime	Coalescent model	Python	Kelleher et al., 2016
SLiM	Population genetic models	Standalone	Haller and Messer, 2019
TreeSim	Birth-death models	R	Stadler, 2011
mvMORPH	Continuous trait evolution models	R	Clavel et al., 2015
diversitree	Several birth-death models	R	FitzJohn, 2012
MASTER	Several birth-death models	BEAST 2	Vaughan and Drummond, 2013
LPhy	Several evolutionary models	Standalone	Drummond et al., 2023

Validating the inferential engine, $\mathbf{I}[\mathcal{M}]$

The more complex the natural phenomenon under study, the more difficult it will be to strike a good balance between model practicality and realism (Levins, 1966). The popular aphorism rings true: "all models are wrong but some are useful" (Box, 1979). Very simple models are easier to implement in efficient inference tools, but will commonly make assumptions that are likely to be broken by the data. Conversely, complex models will fit the data better, but may become unwieldy with increasing levels of realism.

A large number of parameters can cause overfitting and weak identifiability, and inference under highly complex models might be prohibitively slow (Shapiro et al., 2000). Deciding on the utility of a model for real-world problems is a daunting task (Brown and Thomson, 2018; Shepherd and Klaere, 2018), and is a challenge we do not address

in the present contribution. Such model appraisals (what we call "model characterization" below) are normally carried out after a model is published, often in multiple contribution bouts, and are critical for a model's longevity. Analyses of model fit against data are normally accompanied by discussions on assumption validity, and more rarely by benchmarking and scrutinization of model behavior and implementation (e.g., Maddison et al., 2007; Stadler, 2010; Rabosky et al., 2013; Rabosky and Goldberg, 2015; Moore et al., 2016).

When a new model $\mathcal{M}$ is initially proposed, however, authors must ensure that their methods can at the very least robustly recover generating parameters. In this section, we discuss a few techniques that can be employed to assess the correctness of a parameter-estimation routine. These techniques assume that one can accurately simulate from a probabilistic data-generating process (see previous section).

Coverage validation

Our discussion on how to ensure a Bayesian model is well-calibrated and thus correct will mostly follow the ideas in Cook et al. (2006) and Talts et al. (2018). The basic idea is presented in the flowchart in figure 4 (acquamarine dotted box and what is above it), and consists of three stages: simulation, inference, and coverage calculation. Once we have a validated simulator for model $\mathcal{M}$, we start by sampling n parameter sets $\boldsymbol{\theta} = \{\theta_i : 1 \leq i \leq n\}$ from its prior, $f_\Theta(\cdot)$, i.e.:

$$\theta_i \sim f_\Theta(\cdot).$$

For each parameter set θ_i, we then sample a data set d_i from $f_{D|\Theta}(\cdot|\Theta = \theta_i)$:

$$d_i \sim f_{D|\Theta}(\cdot|\Theta = \theta_i),$$

These two steps conclude the "simulation" stage of this validation protocol. With $\boldsymbol{d} = \{d_i : 1 \leq i \leq n\}$, we use the inferential machinery $\mathrm{I}[\mathcal{M}]$ under evaluation to compute $f_{\Theta|D}(\theta_i|D = d_i)$ for each d_i. Recall that we assume the posterior distribution defined by $f_{\Theta|D}(\theta|D = d)$ over Θ will be approximated with MCMC, an algorithm that generates a large sample of size L of parameter values from that posterior distribution, $\boldsymbol{\theta'} = \{\theta_i^j : 1 \leq i \leq n, 1 \leq j \leq L\}$. At this point, we have concluded the inference stage of this validation pipeline.

The third stage and final stage consists of investigating coverage properties of uncertainty intervals. The critical expectation here is that if the inferential engine is correct, we will be able to obtain interval estimates with precise coverage properties. More concretely, let us first define the highest posterior density (HPD) interval. For a credibility level $\alpha \in (0, 1)$, we define $I_\alpha(d) := (a(d, \alpha), b(d, \alpha))$ such that:

$$\frac{1}{f_D(d)} \int_{a(d,\alpha)}^{b(d,\alpha)} f_{D|\Theta}(d|\Theta = \theta) f_\Theta(\theta)\, d\theta = \alpha,$$

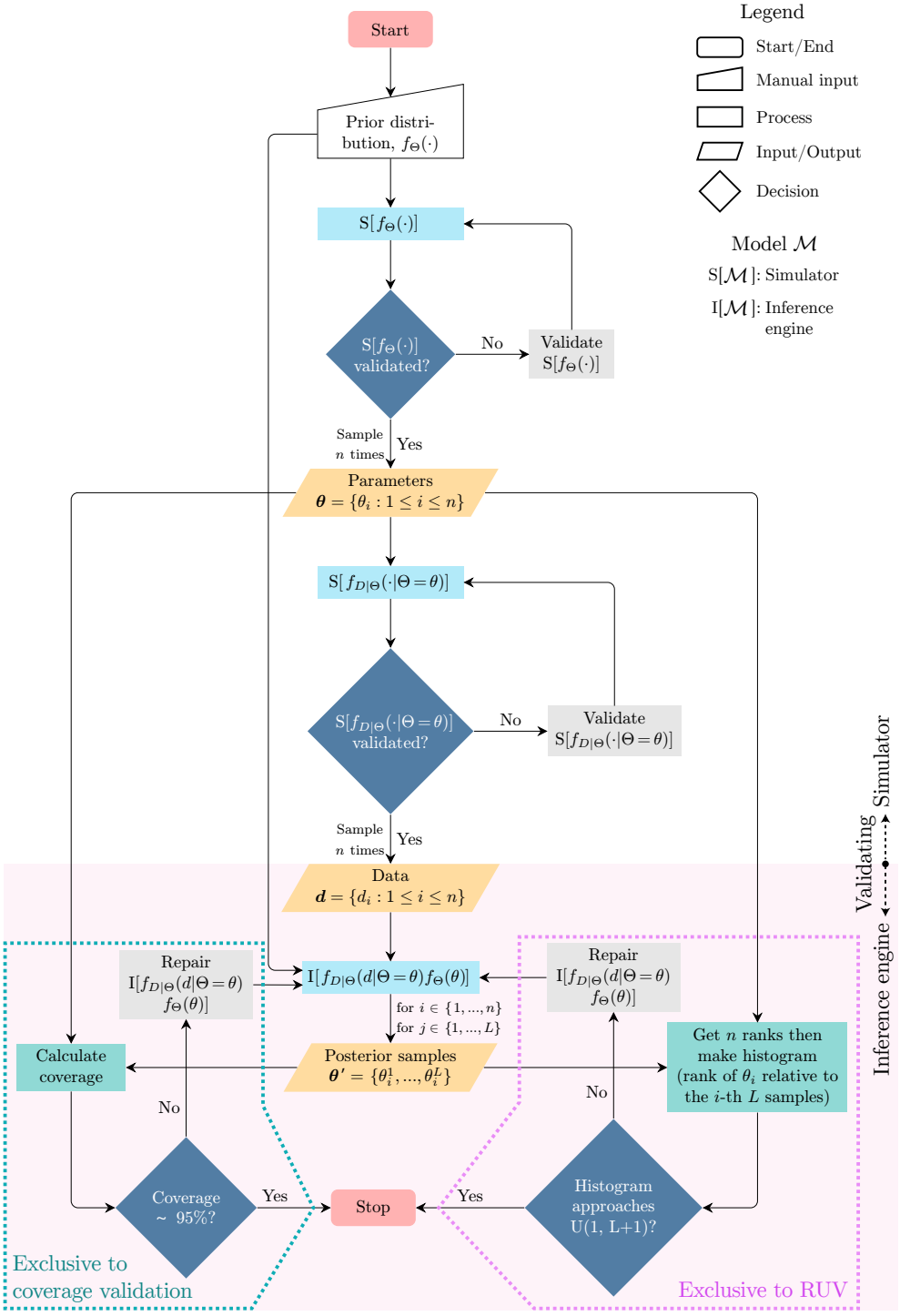


Figure 4: Flowchart of the validation of a Bayesian model. Standard flowchart symbols are explained in the legend. The flowchart area with a clear background is where (true) parameters and data are generated, and where the model simulator(s) is validated. The flowchart area shaded in pink mark the steps involved in validating the inference engine once the data has been generated. $\boldsymbol{\theta}$ denotes a vector with n elements, where each element is an i.i.d. parameter(s) sample from its (their) corresponding prior(s) $f_\Theta(\cdot)$. Analogously, $\boldsymbol{d}$ denotes a vector with n elements, where each element is an i.i.d. data sample from the corresponding likelihood(s) $f_{D|\Theta}(\cdot|\Theta=\theta)$. $\boldsymbol{\theta'}$ holds $n \times L$ elements, with each being one of the L posterior samples for each of the n parameter samples in $\boldsymbol{\theta}$. All L posterior samples obtained from the i-th data set d_i comprise together what one would call the posterior distribution over θ_i. Posterior samples are commonly obtained through MCMC. $\mathrm{U}(l, u)$ denotes a uniform distribution with and including lower and upper bounds l and u, respectively. The acquamarine dotted box encloses the stages of the pipeline that are exclusive to the coverage validation procedure. The pink dotted box encloses the stages of the pipeline that are exclusive to the rank-uniformity validation (RUV) procedure.

Credibility level % ($100 \times \alpha$)	n (replicates)	Lower quantile	Upper quantile
90	100	84	95
	200	171	188
	500	436	463
95	100	90	99
	200	184	196
	500	465	484
99	100	97	100
	200	195	200
	500	490	499

Table 3: The 95% central interquantile intervals for the number of HPD intervals covering the true parameter value (obtained during coverage validation), under different credibility levels and numbers of replicates. Assuming model correctness, the number of true simulated values that fall within their corresponding $100 \times \alpha$%-HPDs (coverage) is binomially distributed with n trials and probability of success α.

where $f_D(d)$ is a constant that can be ignored. By defining $\text{Cred}(I_\alpha(d)) = \alpha$,

$$\inf_{b(d,\alpha)-a(d,\alpha)} \{I_\alpha(d) : \text{Cred}(I_\alpha(d)) = \alpha\}$$

yields the shortest interval with the required credibility. Note that we approximate a particular $I_\alpha(d_i)$ from the i-th L samples obtained with MCMC, in $\boldsymbol{\theta}'$.

Now taking a set of parameter values θ_i sampled from $f_\Theta(\cdot)$ it can be shown that $\Pr(\theta_i \in I_\alpha(d)) = \alpha$, i.e., that $100 \times \alpha$% HPDs have nominal coverage under the true generative model (a proof is provided in the supplementary material). More formally, the coverage of n intervals obtained as above will be distributed as binomial random variable with n trials and success probability α. When $n = 100$ and $\alpha = 0.95$, the 95% central interquantile interval for the number of simulations containing the correct data-generating parameter is between 90 and 99 (Table 3). If we ascertain that $\text{I}[\mathcal{M}]$ of a Bayesian model produces coverage lying within the expected bounds, we say the model has passed the coverage validation, and is well-calibrated and correct.

At this point, we will take a moment to remark that the usefulness of model coverage analysis in Bayesian inference is only manifest when θ_i is sampled from $f_\Theta(\cdot)$. Method developers may be tempted, for example, to calculate coverage for specific parameter values – perhaps chosen across a grid over parameter space – using a different prior during inference. In such cases, we emphasize that obtaining a coverage lower than 95% (for 95% HPDs) does not necessarily mean that a model is incorrectly implemented; conversely, obtaining exactly 95% coverage does not imply model correctness. Coverage values only have bearing on model correctness if, and only if, random variables are sampled from the same prior distribution used in inference.

We provide examples of coverage validation attempts in Figure 5, which shows coverage graphical summaries for data simulated under the model represented in Figure 1. This model is deliberately simple for the sake of brevity and clarity in the discussion below. The parameters in this model are the phylogenetic tree Φ, the species birth rate Λ, and the continuous-trait evolutionary rate R (we assume the continuous trait value at the root, $\boldsymbol{Y_0}$, is known and set it

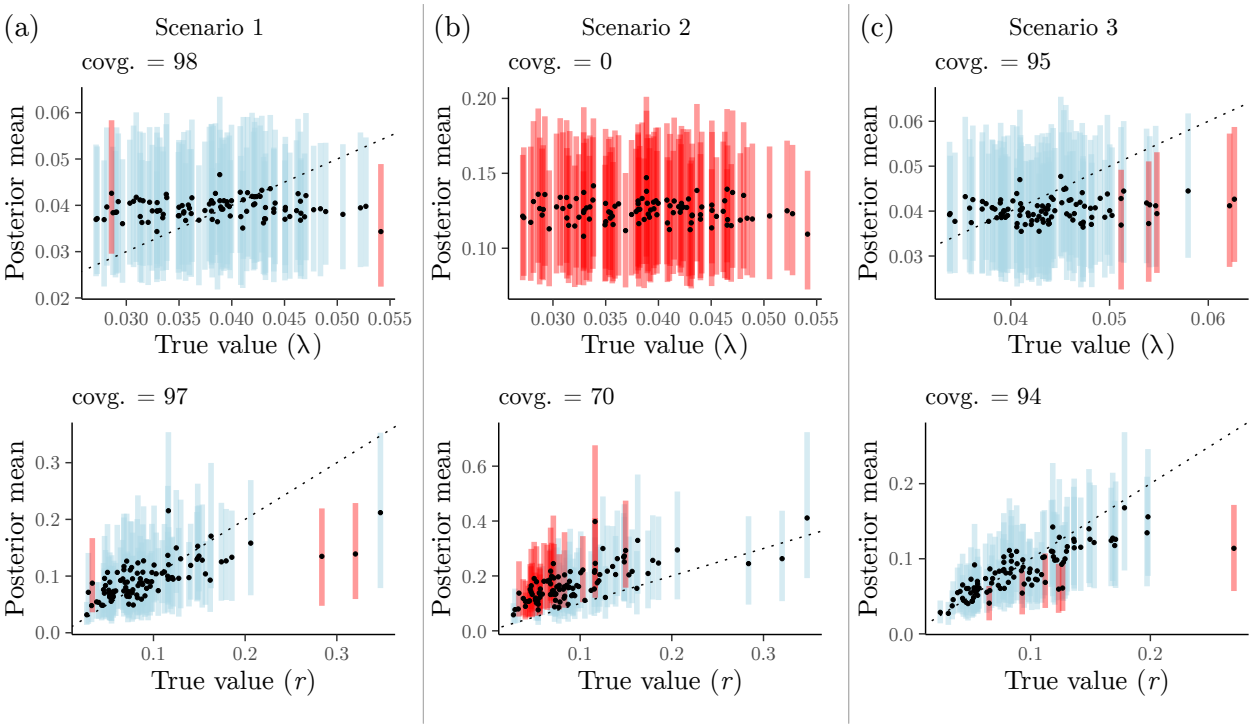


Figure 5: Coverage validation analyses of the Bayesian hierarchical model in Fig. 1. Panels show the true (i.e., simulated) parameter values plotted against their mean posteriors (the dashed line shows $x = y$). Dots and lines (100 per panel) represent true values and their 95%-HPDs, respectively. Simulations for which 95%-HPDs contained the true value are highlighted in blue, otherwise are presented in red. (a) In "Scenario 1", the model was correctly specified, and we simulated trees with 3 to 300 taxa using rejection sampling (approximately one in ten trees were rejected). (b) In "Scenario 2", the model was incorrectly specified in inference (see main text), and we used the same data sets simulated in "Scenario 1". (c) In "Scenario 3" the model was specified just as in "Scenario 1", with the difference that rejection sampling was substantially greater (we rejected a large number of trees, approximately 90%, keeping those having between 100 to 200 tips).

to **0.0** for all simulated data sets). When the model is correctly specified between simulation and inference ("Scenario 1", Fig. 5a), coverage is close to 95% and adequate for both Λ and R, which indicates that $\mathrm{I}[\mathcal{M}]$ – as implemented in BEAST 2, the software we used – is well-calibrated and correct.

In "Scenario 2" of Figure 5 (Fig. 5b), however, we misspecify the model during inference, setting the prior distribution on Λ to be a log-normal with a mean of -2.0 (rather than -3.25, as specified in the simulation procedure; Fig. 1). In contrast with scenario 1, coverage is 0.0 for Λ and 70% for R, both much lower and outside the expected coverage bounds (Table 3). These numbers indicate that one or more of the parts comprising model $\mathcal{M}$ used in $\mathrm{I}[\mathcal{M}]$ differs from their counterparts in $\mathrm{S}[\mathcal{M}]$. This result was expected because we purposefully made the models in simulation and inference differ; we know $\mathrm{I}[\mathcal{M}]$ is correct because of the results from scenario 1. Of course, in a real-world validation experiment the model should be correctly specified, and such a result would suggest a problem with the inferential machinery (provided the the simulator had been previously validated).

Lastly, in "Scenario 3" of Figure 5 (Fig. 5c), we specified the model just like in scenario 1, but carried out substantial rejection sampling during simulation. Approximately 90% of all simulated trees were rejected based on their taxon count; trees were rejected if they had fewer than 100 or more than 200 taxa. As with scenario 1, coverage fell within the expected ranges for a correct model implementation. This result may strike the reader as odd: if $\mathrm{I}[\mathcal{M}]$ expects trees with a wide range of tip numbers, and we feed it simulated trees within a narrow tip number interval, should this not lower coverage? For example, one may have expected the estimated λ to be consistently higher or lower than the true λ. Λ is nonetheless challenging to infer under the current model, as suggested by estimates falling around the corresponding prior mean value; unlike in scenario 2, however, here the prior mean parameter was correctly specified. As a result, coverage validation was not capable of detecting any symptoms arising from the rejection of tree samples.

Scenario 3 brings home the point that an incorrect model implementation may pass coverage validation, unless model misspecification is sufficiently severe (e.g., scenario 2), or parameter estimate location is highly responsive to the evidence in the data – unlike Λ in the examined model. (In the supplement we expand on this point using a different and simpler model, and show that, if extreme, rejection schemes will be detected by coverage validation as a model misspecification issue; Supplementary Table 1.) Put simply, obtaining appropriate coverage may not be enough to ascertain that a model is correct. Potential biases in parameter estimates may remain undetected unless more investigation is done (see rank-uniformity validation below).

The three scenarios we explored above illustrate how coverage validation results can be interpreted in terms of model implementation correctness. One can additionally capitalize on this validation setup and gauge how accurate our inferential tool can be for different parameters. The easier it is to estimate a parameter, the higher should be the correlation between its posterior mean and its generating "true" value. In our scenarios 1 and 3, the species birth rate Λ was hard to estimate given the sizes of the phylogenetic trees. Conversely, the continuous-trait evolutionary

rate, R, was more easily identifiable, as revealed by the higher correlation between its true values and their posterior means. We conclude this section by noting that the absence of correlation between parameter estimates and their true values (sometimes referred to as "weak unidentifiability") should not be taken as a sign that a model is incorrect – inappropriate coverage values should.

Rank-uniformity validation (RUV)

Talts et al. (2018) showed that one can devise other tests that might be more powerful to detect problems than just looking at the coverage of Bayesian HPD intervals. In particular, given $\boldsymbol{\theta} = \{\theta_i : 1 \leq i \leq n\}$ (produced according to the protocol in Fig. 4), those authors demonstrated (Theorem 1 therein) that if the inference machinery $\mathrm{I}[\mathcal{M}]$ works as intended, the distribution of the rank r_i of θ_i relative to $\boldsymbol{\theta}'_i$ – i.e., the rank of the i-th parameter value relative to its corresponding L MCMC chain samples – will follow a uniform distribution on $[1, L+1]$ (Fig. 4, pink dotted box; Fig. 6a). In other words, if one were to sort all true parameter values θ_i against $\boldsymbol{\theta}'_i$ – their corresponding L MCMC posterior samples – the first (smallest ranking) 10% out of n θ_i values should account for approximately 10% of the total rank mass; the next 10% of (higher ranking) θ_i values should account, again, for approximately 10% of the total rank mass, and so on.

Adherence to this distribution can be investigated by constructing histograms (Talts et al., 2018) as well as by looking at the empirical cumulative distribution function (ECDF) and their confidence bands (Säilynoja et al., 2021). When a model implementation fails RUV, it can do so in different ways. For instance, when the inference machinery leads to consistent overdispersed estimates, it produces a pattern of ranks concentrating around the middle rank (Fig. 6b). When underdispersion is present, on the other hand, ranks tend to bunch up towards the ends (Fig. 6c), creating a pattern of "horns", which can also be caused by high autocorrelation in the MCMC draws. This is also why we recommend thinning MCMC draws in order to reduce autocorrelation. Figure 6d shows the rank patterns when the inference machinery produces biased estimates: ranks will bunch up against one of the ends, depending on whether estimates are biased downwards of upwards. In the particular case shown in figure 6c, the parameter at hand is being overestimated.

We conducted RUV on the three scenarios described in the previous section, which make use of the model depicted in Figure 1. In the interest of brevity, we only show the histograms and ECDFs for R, and leave the remaining plots for Λ to the supplement (but see Box 1 below). As expected, under scenario 1 our model implementation passes the RUV – as indicated by histogram bars and ECDF values falling within their 95% confidence intervals (Fig. 7a).

Under scenario 2, again as expected, our method failed RUV (Fig. 7b). In particular, we observes great overestimation of the Brownian motion rate (R). In a real world-analysis, these results would point to one or more faulty implementations (e.g., one or more model components, MCMC machinery, the simulator, etc). We remind the reader that in our experiment, scenario 2 was purposefully set up so that the (prior) models used in simulation and inference

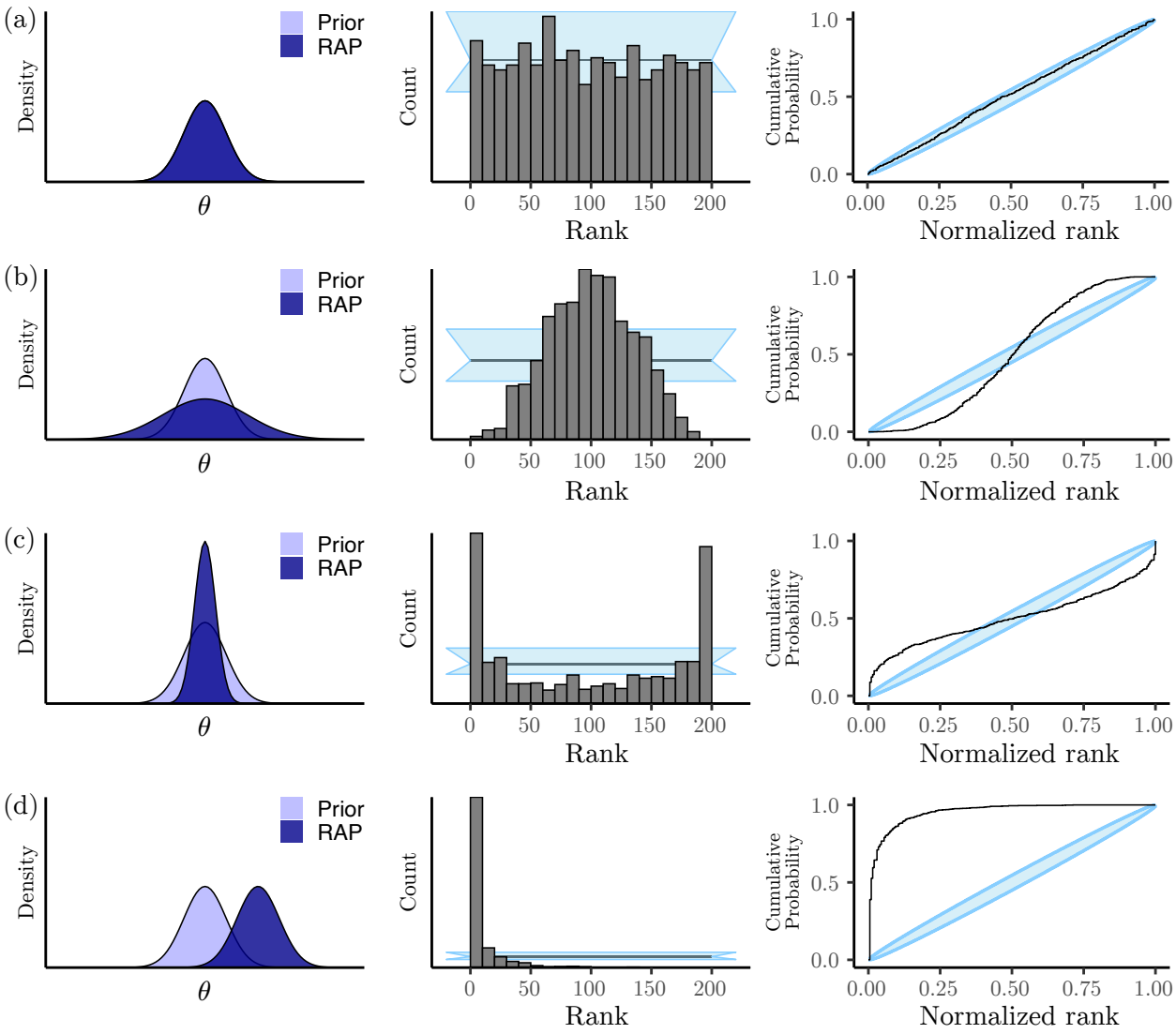


Figure 6: Patterns observable after inference in rank-uniformity validation (RUV). We explain how to interpret the histogram of ranks (middle column) and ECDF plots (right-hand side column) in the main text. (a) Model implementation is correct. (b) Parameter estimates are overdispersed relative to their true values. (c) Parameter estimates are underdispersed relative to their true values. (d) Parameter estimates are consistently overestimated relative to their true values. In the left-hand side column, the prior and replicate-averaged posterior (also known as the data-averaged posterior) distributions over some parameter θ are shown in light blue and dark blue, respectively. In the middle graphs, light-blue bands represent the 95% confidence interval about the expected rank count, and horizontal black lines mark the rank count mean. Light-blue ellipses in the rightmost graphs represent confidence intervals about the empirical cumulative distribution function (ECDF).

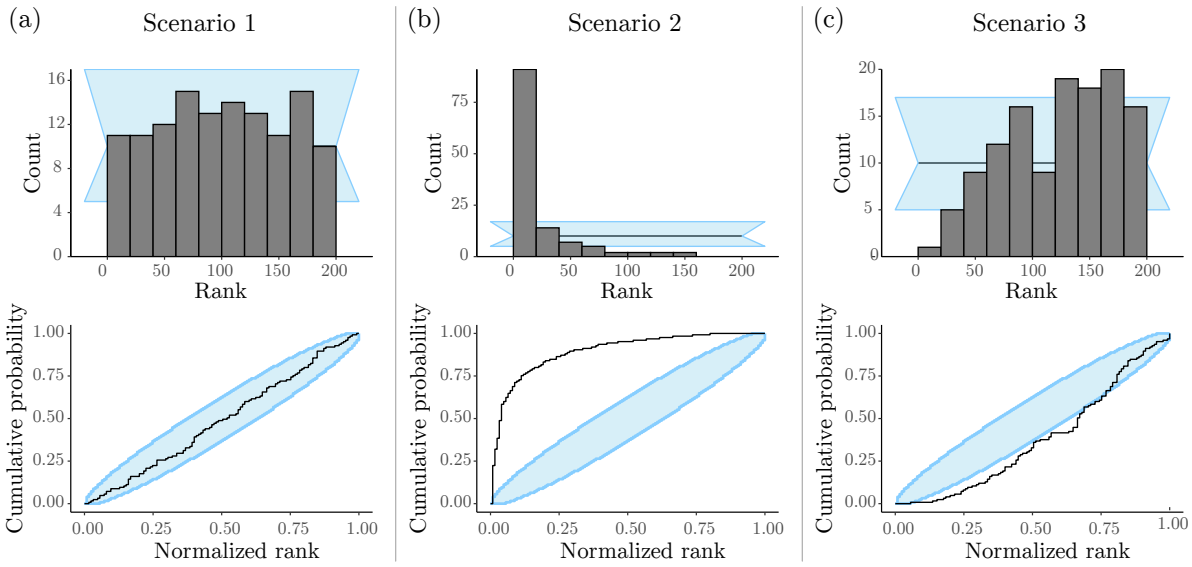


Figure 7: Rank-uniformity validation (RUV) of the Bayesian hierarchical model in Fig. 1. Panels in the top row show the histograms of $n = 100$ ranks, for parameter R in each scenario, obtained after 10% burnin and thinning of posterior samples down to 200 out of 10,000. Panels in the bottom row show the corresponding ECDF plots, for parameter R in each scenario. (a) In "Scenario 1", the model was correctly specified and we can see that the ranks are compatible with a uniform distribution (within the blue band). (b) In "Scenario 2", the inference machinery was misspecified and a clear pattern of overestimation shows up in the ranks, meaning the ranks for the data-generating values are usually smaller than expected under correctness. (c) In "Scenario 3" we can see a pattern of underestimation, evidenced by ranks bunching up to the right.

differed; our implementations are actually correct, but were induced to fail the RUV procedure.

Lastly, RUV results for scenario 3 contrasted with what we observed for this scenario's coverage validation (Fig. 5c). While the model specified in scenario 3 passed its coverage validation (coverage was acceptable for both Λ and R), it did not pass the RUV procedure. The corresponding rank histogram and ECDF plots indicate that R is underestimated (Fig. 7c). This result suggests that rank-uniformity validation can be more sensitive than coverage validation, at least for certain types of model misspecification, such as those affecting parameter estimate location.

Unlike scenario 2, in which we caused an explicit mismatch between the distributions used in simulation and inference, model misspecification under scenario 3 was subtler: simulation and inference models were identical (as in scenario 1), but tree samples from $f_\Phi(\cdot)$ (the Yule prior) were often rejected as $\boldsymbol{\theta}$ was generated. The model used in scenario 3 failed RUV because rejecting tree samples induced an implicit Yule model in simulation that differed from the Yule model used in inference. Indeed, using a much simpler model and an analogous rejection scheme (Supplementary Fig. 3), we were able to recapitulate the results in Figure 7. While here we focus on so-called continuous parameters like R and Λ, it is also possible to conduct RUV to assess correctness on the space of phylogenies, a topic we leave for future research.

Tree models

Tree models are stochastic processes that can organically capture the most fundamental tenet in evolutionary biology, namely common descent, at multiple time scales. Over the last few decades, pivotal theoretical work has not only characterized many properties of the more elementary tree models (for examples and an overview, see Nee, 2006; Wakeley, 2009; Stadler, 2013; Harmon, 2018), but also generalized them to be more realistic. Popular among empiricists, for example, are tree models that allow for lineage-affecting event rates that vary over time and across taxa, and that are state-dependent ("state" here meaning the attributes of a lineage's genotypic, phenotypic, ecological or geographic character(s)). Such models lend themselves to the study evolutionary phenomena such as species diversification and infectious disease spread.

Although convenient evolutionary abstractions, tree models can nonetheless be challenging to formalize depending on their level of realism. The parameter space of a tree model is difficult to handle: it includes both a combinatorily complicated discrete component (the tree topology) and a continuous component (the branch lengths) (Semple et al., 2003). The theoretical properties, summarization, and exploration of tree space are all active topics of research in mathematical and computational biology (Gavryushkina et al., 2013; Gavryushkin and Drummond, 2016; Brown et al., 2020).

Given the interest in tree models shown by empirical, computational and theoretical biologists, in this section we will cover how tree models have been commonly validated in the past, with an emphasis on tree space. Our treatment is not meant to be an exhaustive review, but a short synthesis, and in keeping with the subject of the present work, we will not discuss protocols for the development and validation of Bayesian proposals in tree space. This topical subject is multifaceted (e.g., Douglas et al., 2021; Bouckaert, 2022; Douglas et al., 2022) and deserves a dedicated contribution we leave for the future.

Of the many tree models implemented in computational tools, only a select few have had analytical solutions derived for their mathematical properties. Among the most theoretically understood models are the birth-death process ([add citations]; of which the Yule model, above, is a subcase) and the Kingman's coalescent (of which the popular multispecies coalescent, MSC, is a generalization; [add citations]). Typical theoretical properties known for these models include, for example, ...

- PG1/2: Simplest models (Yule, birth-death, coalescent): theoretical expectations; TODO: plot root age sampling from the prior (put in supplement);
- PG1/2: Looking at theoretical expectations for serially sampled genealogies under Kingman coalescent, and gene tree composition (Rosenberg, 2002). This has been done for simulator (Mendes et al., 2018), and packages exist for exact calculations (Kim et al., 2019);
- PG1/2: Looking at coverage of tree-space parameters, e.g., root height (continuous space) and number of

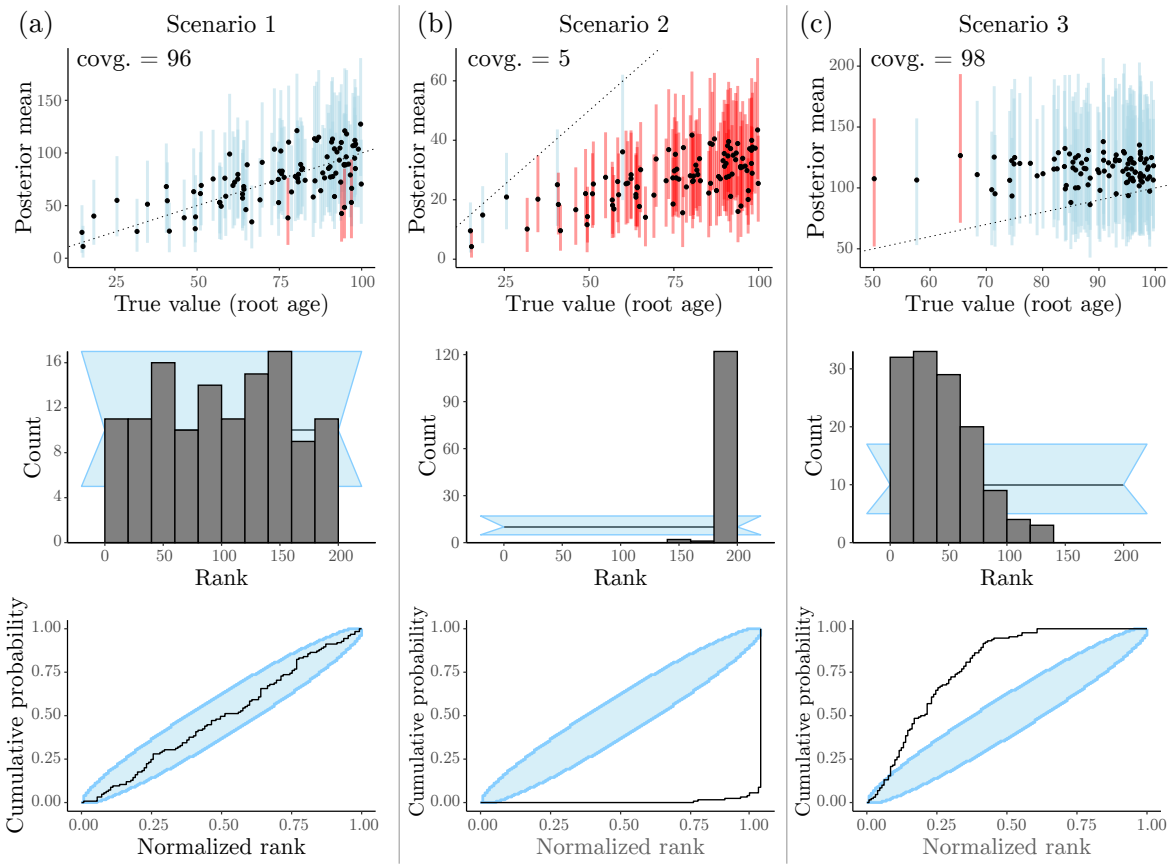


Figure 8: Coverage validation and rank-uniformity validation (RUV) of the Bayesian hierarchical model in Fig. 1, for each scenario described in the main text (also see Figs. 5 and 7), with respect to the height of ϕ (i.e., the phylogeny's root age). Panels in the top row show the true (i.e., simulated) root age values plotted against their mean posteriors (the dashed line shows $x = y$). Dots and lines (100 per panel) represent true values and their 95%-HPDs, respectively. Simulations for which 95%-HPDs contained the true value are highlighted in blue, otherwise are presented in red. Panels in the middle row show the RUV histograms of $n = 100$ ranks in each scenario, obtained after 10% burnin and thinning of posterior samples down to 200 out of 10,000. Panels in the bottom row show the corresponding RUV ECDF plots in each scenario.

sampled ancestors (discrete space) (see Fig. 3 in Gavryushkina et al., 2014). Also coverage of tree models like the MSC (Ogilvie et al., 2022).

- PG3: Balance statistics building on Albert Soewongsono's work; TODO: plot "set" β statistic (y-axis, violin plots) for Yule model with different tree sizes; plot theoretical "set" β statistic on the x-axis and β-MLE estimate on y-axis for 100 sets of 100 BiSSE trees (fixed parameters) – plot 1:1 line and get $\sim$ 95% CI bars overlapping that line;
- PG4: More complicated models are troublesome. Exact likelihood values are computed for sub-cases of a new general model that have theoretical expectations (BDSS) or against an independent implementation (Andréoletti et al., 2022). Comparing sampling distributions between direct simulator (inverse CDF) and sampling from the prior (BDSS implementation in Zhang et al., 2023);
- PG5: Models that cannot be written as generative models, like node-calibration models; discuss Heled's calibrated-Yule prior is a solution.

Software

We implemented a suite of methods for automating many of the steps involved in coverage validation and RUV. These methods were developed in Java and integrated into the BEAST 2 platform (Bouckaert et al., 2019). Code and a tutorial are available on https://github.com/rbouckaert/DeveloperManual.

Bayesian model validation guidelines for developers and reviewers

In the previous sections, we described and executed two procedures for validating Bayesian models, namely, coverage validation and rank-uniformity validation (RUV). Once both simulation and inference can be conducted under a model $\mathcal{M}$, executing these procedures amounts to following relatively straightforward protocols (Fig. 4). Importantly, following these protocols should validate any Bayesian model, regardless of the nature of its parameter space and its component sampling distributions. This is because such protocols provide clear, objective rules for assessing model correctness, based on the coverage and rank distribution of parameters values; both can be computed for any and all Bayesian models. For the reasons above, carrying out an analysis of coverage and/or of the distribution of parameter-value ranks (with respect to their posterior samples) should on one hand be a requirement, and on another should suffice for introducing a new Bayesian model implementation.

My model implementation failed correctness tests, what now?

Method developers should expect their software to often fail validation, especially at early development stages. New model implementations visit the loops in Fig. 4 many times – it is then that the most glaring, and later the more

insidious, backend coding mistakes and model misspecification errors are uncovered and fixed. Our words are meant as encouragement: in our experience, the validation procedure is almost always arduous and repetitive, but very effective in revealing issues and later in giving modelers peace of mind when releasing their software.

As we showed above, however, a model inference machinery can still fail a validation test despite being correctly implemented in code. This can happen when there is model misspecification as a result of users filtering out simulations that do not generate the types of data sets observed in the real world. If upon such failure researchers are confident that all coding mistakes have been ironed out, a delicate stage of method develop begins, when further analysis can take place, decisions must be made, and different resolutions are possible.

When the observed coverage is distant from its expectation or if RUV reveals over- or underestimated parameters, for example, a researcher may want to hypothesize about why validation is failing, and test those hypotheses with further analysis of simulated data. If validation success is contingent upon simulation conditions that are empirically unrealistic or computationally intractable (e.g., trees with zero or an immense number of extant tips must be included), regularities emerging from further simulation experiments might illuminate the nature of the model misspecification. For example, method developers may want to progressively tweak an aspect of simulation, and then check if their implementation seems to approach correctness (e.g., section 3 in the supplement).

The type of experimentation described above should be further useful to those interested in characterizing model behavior (see below), even after it is published. We expect that the different extents to which an RUV procedure, in particular, can be made to fail (and how) will help method developers identify biases in parameter estimates [cite Sean's pre-print/paper?]. Failing to obtain appropriate coverage should also be informative, but to a lesser degree. First, because coverage validation is a less sensitive validation procedure, and second, because a coverage number does not point to the direction of bias. Here, graphical summaries of post-processed MCMC results (e.g., Fig. 5) can help uncover parameter estimate biases.

Lessons about models can thus be learned by attempting to validate a model implementation even if ultimately it cannot be shown to be correct. At this point, method developers will have to choose whether to move forward and release their tool, or to have it undergo further testing until evidence for correctness is produced. In the former case, researchers should in the very least be expected to report all attempts made to validate an implementation, why they seemed to fail, and what biases were uncovered, if any. Ideally, authors should also provide guidelines for interpreting results obtained with the proposed tool.

Should passing validation procedures appear hopeless, the method developer may understandably find themselves downplaying the importance of the validation effort. If "all models are wrong" and if the necessary conditions for successful validation are unrealistic or hard to meet, then maybe validating a model is not all that important. Those with real data set to analyze will likely find pragmatic comfort in this line of reasoning. But while a tempting thought, we urge researchers to consider how this perspective runs roughshod over statistical hypothesis testing, a ubiquitous

analysis in empirical studies. Although understanding inference biases may alleviate it, once an implementation cannot be shown to be correct, it becomes a non-trivial exercise, for example, to determine if a parameter estimate is different from another value. The task of model comparison, in turn, is even more hampered – here, more than one (potentially incorrect) implementation must be compared, and biases in model evidence estimation will be particularly hard to characterize. Successful model validation is evidently something to strive for and in the absence of which interpreting the results will be challenging at best, and impossible in many cases.

When confronted with utter validation failure, method developers may have to ask the hard question of whether their models are reasonable in the first place. If simulations must be conducted within a substantially constrained parameter space so as to yield realistic data – or data whose probabilities can be calculated – this could be a sign that the model needs to be modified. Independent implementations that also do not pass validation tests provide further evidence that the issue is potentially in the model assumptions themselves. Historically, model design has often gone in the direction of incrementally conditioning the statistical process so as to match empirical observations (e.g., Gelman and Meng, 1995; Gelman et al., 2020), but re-imagining the model entirely might be the best solution [ideas for citations?].

Model characterization

In addition to the model validation we detailed above, there is an infinite number of ways in which a new or published model can have its behavior inspected. Researchers may want to know, given a model, how sensitive parameter estimates are to data set size, prior choice, model complexity, violation of model assumptions, to name a few. Studies have examined how these factors affect estimation accuracy and precision (e.g., Zhang et al., 2023; Luo et al., 2023), as well as the mixing and convergence of MCMC chains (e.g., Nylander et al., 2004; Zhang et al., 2023). We collectively refer to these examinations as "model characterization": any analysis of model behavior beyond assessing its correctness. Model characterization is rarely carried out to satisfy the curiosity of the theoretician (but see, e.g., Tuffley and Steel, 1997; Steel and Penny, 2000); it is instead normally motivated by a model's empirical applications. These investigations are thus critical for the longevity and popularity of a model, as domain experts will only adopt a model widely if they know when to trust the results and how to interpret them.

It is possible to characterize certain aspects of model behavior while simultaneously verifying its correctness, as discussed in the coverage validation section above. For example, one can observe how accurate parameter estimates are (e.g., if the points in Fig. 5 fall on the identity line) under both correctly and incorrectly specified models. However, the requirement of simulating parameter values from a prior distribution $f_\Theta(\cdot)$ during the validation of a model can complicate its characterization. Depending on the characterization experiment's goals and design, researchers may find themselves rejecting a large fraction of simulated data sets – perhaps because data sets do not resemble those in real life, or because they are too large to analyze. But as we showed, rejecting draws in simulation may then

be picked up by the validation protocol as an incorrectly implemented model. This problem can only worsen the more dimensions of parameter space are allowed to vary. In most cases, it may thus make more sense to first verify model correctness by following the procedures we described above, and then characterize model behavior further in a subsequent batch of analyses.

We conclude this section by proposing that scientists contributing or reviewing a new model ask the following question: Is the contribution at hand carrying out an empirical analysis that will specifically profit from scrutinizing model behavior? If not, then model characterization efforts will likely serve their purpose better elsewhere, and profit from being shouldered by the scientific community at large.

Concluding remarks

In order to keep up with the large amounts of data of different kinds accumulating in public databases, researchers in the life sciences must constantly update their computational tool boxes. New models are implemented in computational methods every day, but if they are not properly validated, downstream conclusions from using those methods may be void of any significance.

In the present study, we described and executed two distinct validation protocols that verify a Bayesian model has been correctly implemented. Although we looked at examples from evolutionary biology, specifically statistical phylogenetics, these two simulation-based protocols work for any and all Bayesian models.

We further elaborate on the difference between experiments in model validation versus model characterization. Newly implemented models can only profit from validation experiments, which are strictly concerned with theoretical expectations (e.g., about coverage) a model must meet if correctly implemented. Model characterization, on the other hand, is about inspecting model behavior as a variety of data set and model attributes interact; here, exact quantitative predictions may not be theoretically guaranteed. Such experiments are best designed and justified when empirically motivated.

We hope the guidelines described here can enhance both the release rate and standards of statistical software for biology, by assisting its users, developers and referees in quickly finding common ground when evaluating new modeling work.

Funding

F.K.M. was supported by Marsden grant 16-UOA-277 and by the National Science Foundation (DEB-2040347). R.B. was supported by Marsden grant 18-UOA-096. L.M.C was partially supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and by the School of Applied Mathematics, Getulio

Vargas Foundation. A.J.D. was supported by Marsden grant 16-UOA-277.

References

Aldous, D. J. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Stat. Sci.*, 16:23–24.

Andréoletti, J., Zwaans, A., Warnock, R. C. M., Aguirre-Fernández, G., Barido–Sottani, J., Gupta, A., Stadler, T., and Manceau, M. (2022). The occurrence birth-death process for combined-evidence analysis in macroevolution and epidemiology. *Syst. Biol.*, 71:1440–1452.

Bates, K. A., Clare, F. C., O'Hanlon, S., Bosch, J., Brookes, L., Hopkins, K., McLaughlin, E. J., Daniel, O., Garner, T. W. J., Fisher, M. C., and Harrison, X. A. (2018). Amphibian chytridiomycosis outbreak dynamics are linked with host skin bacterial community structure. *Nature Comm.*, 9:1–11.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N. D., Matschiner, N., Müller, N. F., Ogilvie, H., du Plessis, L., Popinga, A., Mendes, F. K., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T., and Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comp. Biol.*, 15:e1006650.

Bouckaert, R. R. (2022). An efficient coalescent epoch model for Bayesian phylogenetic inference. *Syst. Biol.*, 71:1549–1560.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In Launer, R. L. and Wilkinson, G. N., editors, *Robustness in statistics*, pages 201–236. Academic Press.

Brown, J., Mount, G. G., Gallivan, K. A., and Wilgenbusch, J. C. (2020). The diverse applications of tree set visualization and exploration. *EcoEvoRxiv*.

Brown, J. M. and Thomson, R. C. (2018). Evaluating model performance in evolutionary biology. *Annu. Rev. Ecol. Evol. Syst.*, 49:95–114.

Clavel, J., Escarguel, G., and Merceron, G. (2015). mvMORPH: an R package for fitting multivariate evolutionary models to morphometric data. *Methods Ecol. Evol.*, 6:1311–19.

Cook, S. R., Gelman, A., and Rubin, D. B. (2006). Validation of software for Bayesian models using posterior quantiles. *J. Comput. Graph. Stat.*, 15(3):675–692.

Darriba, D., Flouri, T., and Stamatakis, A. (2018). The state of software for evolutionary biology. *Mol. Biol. Evol.*, 35:1037–46.

de Manuel, M. et al. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, 354:477–81.

Dobzhansky, T. (1936). Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. *Genetics*, 21:113–35.

Douglas, J., Jiménez-Silva, C. L., and Bouckaert, R. (2022). StarBeast3: adaptive parallelized Bayesian inference under the multispecies coalescent. *Syst. Biol.*, 71:901–915.

Douglas, J., Zhang, R., and Bouckaert, R. (2021). Adaptive dating and fast proposals: revisiting the phylogenetic relaxed clock model. *PLoS Comp. Biol.*, 17:e1008322.

Drummond, A. J., Chen, K., Mendes, F. K., and Xie, D. (2023). LinguaPhylo: a probabilistic model specification language for reproducible phylogenetic analyses. *PLoS Comp. Biol.*, 19:e1011226.

Fabreti, L. G. and Höhna, S. (2022). Convergence assessment for bayesian phylogenetic analysis using mcmc simulation. *Methods in Ecology and Evolution*, 13(1):77–90.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.*, 25:471–92.

FitzJohn, R. G. (2012). Diversitree: comparative phylogenetic analyses of diversification in R. *Methods Ecol. Evol.*, 3:1084–1092.

Gavryushkin, A. and Drummond, A. J. (2016). The space of ultrametric phylogenetic trees. *J. Theor. Biol.*, 403:197–208.

Gavryushkina, A., Welch, D., and Drummond, A. J. (2013). Recursive algorithms for phylogenetic tree counting. *Algorithms Mol. Biol.*, 8(26):1–13.

Gavryushkina, A., Welch, D., Stadler, T., and Drummond, A. J. (2014). Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Comp. Biol.*, 10:e1003919.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC Press, Boca Raton, Florida.

Gelman, A. and Meng, X.-L. (1995). Model checking and model. *Markov chain Monte Carlo in practice*, page 189.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808*.

Gibson, A. K. and Fuentes-González, J. A. (2015). A phylogenetic test of the Red Queen Hypothesis: outcrossing and parasitism in the Nematode phylum. *Evolution*, 69:530–40.

Haller, B. C. and Messer, P. W. (2019). SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. *Mol. Biol. Evol.*, 36.

Harmon, L. K. (2018). *Phylogenetic comparative methods: learning from trees*.

Hasegawa, M., Kishino, H., and Yano, T. A. (1985). Dating of the human age splitting by a molecular clock of mitochondrial-DNA. *J. Mol. Evol.*, 22:160–74.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57.

Hopkins, R. and Rausher, M. D. (2012). Pollinator-mediated selection on flower color allele drives reinforcement. *Science*, 335:1090–92.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.*, 11:1–44.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model. *Bioinformatics*, 18:337–8.

Human Microbiome Project Consortium (2012). A framework for human microbiome research. *Nature*, 486:215–221.

Kawahara, A. Y., Plotkin, D., Espeland, M., Meusemann, K., Toussaint, E. F. A., Donath, A., Gimnich, F., Frandsen, P. B., Zwick, A., dos Reis, M., Barber, J. R., Peters, R. S., Liu, S., Zhou, X., Mayer, C., Podsiadlowski, L., Storer, C., Yack, J. E., Misof, B., and Breinholdt, J. W. (2019). Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc. Natl. Acad. Sci. USA.*, 116:22657–63.

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comp. Biol.*, 12:e1004842.

Kim, A., Rosenberg, N. A., and Degnan, J. H. (2019). Probabilities of unranked and ranked anomaly zones under birth–death models. *Mol. Biol. Evol.*, 37:1480–1494.

Lack, J. B., Lange, J. D., Tang, A. D., Corbett-Detig, R. B., and Pool, J. E. (2016). A thousand fly genomes: an expanded *Drosophila* genome nexus. *Mol. Biol. Evol.*, 33:3308–13.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, 54(4):421–431.

Lively, C. M. (1987). Evidence from a New Zealand snail for the maintenance of sex by parasitism. *Nature*, 328:519–21.

Luo, A., Zhang, C., Zhou, Q.-S., Ho, S. Y. W., and Zhu, C.-D. (2023). Impacts of taxon-sampling schemes on Bayesian tip dating under the fossilized birth-death process. *Syst. Biol.*, 72:781–801.

Lynch, M. (2007). Population genomics of *Daphnia pulex*. *Genetics*, 206:315–32.

Maddison, W. P., Midford, P. E., and Otto, S. P. (2007). Estimating a binary character's effect on speciation and extinction. *Syst. Biol.*, 56:701–710.

Magee, A., Karcher, M., Matsen IV, F. A., and Minin, V. M. (2023). How trustworthy is your tree? Bayesian phylogenetic effective sample size through the lens of Monte Carlo error. *Bayesian Anal.*, 1(1):1–29.

Maynard Smith, J. (1978). *The evolution of sex*. Cambridge University Press.

Mendes, F. K., Fuentes-González, J. A., Schraiber, J., and Hahn, M. W. (2018). A multispecies coalescent model for quantitative traits. *eLife*, 7:e36482.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092.

Moore, B. R., Höhna, S., May, M. R., Rannala, B., and Huelsenbeck, J. P. (2016). Critically evaluating the theory and performance of bayesian analysis of macroevolutionary mixtures. *Proc. Natl. Acad. Sci. USA.*, 113:9569–9574.

Morran, L. T., Schmidt, O. G., Gelarden, I. A., II, R. C. P., and Lively, C. M. (2011). Running with the Red Queen: host-parasite coevolution selects for biparental sex. *Science*, 333:216–18.

Muller, H. J. (1940). Bearing of the *Drosophila* work on systematics. In Huxley, J. S., editor, *The new systematics*, pages 185–268. Clarendon Press, Oxford.

Neafsey, D. E., Waterhouse, R. M., Abai, M. R., Aganezov, S. S., Alekseyev, M. A., Allen, J. E., Amon, J., Arcà, B., Arensburger, P., Artemov, G., Assour, L. A., Basseri, H., Berlin, A., Birren, B. W., Blandin, S. A., Brockman, A. I., Burkot, T. R., Burt, A., Chan, C. S., Chauve, C., Chiu, J. C., Christensen, M., Costantini, C., Davidson, V. L., Deligianni, E., Dottorini, T., Dritsou, V., Gabriel, S. B., Guelbeogo, W. M., Hall, A. B., Han, M. V., Hlaing, T., Hughes, D. S., Jenkins, A. M., Jiang, X., Jungreis, I., Kakani, E. G., Kamali, M., Kemppainen, P., Kennedy, R. C., Kirmitzoglou, I. K., Koekemoer, L. L., Laban, N., Langridge, N., Lawniczak, M. K., Lirakis, M., Lobo, N. F., Lowy, E., MacCallum, R. M., Mao, C., Maslen, G., Mbogo, C., McCarthy, J., Michel, K., Mitchell, S. N., Moore, W., Murphy, K. A., Naumenko, A. N., Nolan, T., Novoa, E. M., O'Loughlin, S., Oringanje, C., Oshaghi, M. A., Pakpour, N., Papathanos, P. A., Peery, A. N., Povelones, M., Prakash, A., Price, D. P., Rajaraman, A., Reimer, L. J., Rinker, D. C., Rokas, A., Russell, T. L., Sagnon, N., Sharakhova, M. V., Shea, T., Simão, F. A., Simard, F., Slotman, M. A., Somboon, P., Stegniy, V., Struchiner, C. J., Thomas, G. W., Tojo, M., Topalis, P., Tubio, J. M., Unger, M. F., Vontas, J., Walton, C., Wilding, C. S., Willis, J. H., Wu, Y. C., Yan, G., Zdobnov, E. M., Zhou, X., Catteruccia, F., Christophides, G. K., Collins, F. H., Cornman, R. S., Crisanti, A., Donnelly, M. J., Emrich, S. J., Fontaine, M. C., Gelbart, W., Hahn, M. W., Hansen, I. A., Howell,

P. I., Kafatos, F. C., Kellis, M., Lawson, D., Louis, C., Luckhart, S., Muskavitch, M. A., Ribeiro, J. M., Riehle, M. A., Sharakhov, I. V., Tu, Z., Zwiebel, L. J., and Besansky, N. J. (2015). Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science*, 347(6217):1258522.

Nee, S. (2006). Birth-death models in macroevolution. *Annu. Rev. Ecol. Evol. Syst.*, 37:1–17.

Nylander, J. A. A., Ronquist, F., Huelsenbeck, J. P., and Nieves-Aldrey, J. (2004). Bayesian phylogenetic analysis of combined data. *Syst. Biol.*, 53(1):47–67.

Ogilvie, H. A., Mendes, F. K., Matzke, N. J., Stadler, T., Welch, D., and Drummond, A. J. (2022). Novel integrative modeling of molecules and morphology across evolutionary timescales. *Systematic Biology*, 71:208–220.

Pease, J. B., Haak, D. C., Hahn, M. W., and Moyle, L. C. (2016). Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biol.*, 14(2):e1002379.

Rabosky, D. L. and Goldberg, E. E. (2015). Model inadequacy and mistaken inferences of trait-dependent speciation. *Syst. Biol.*, 64:340–355.

Rabosky, D. L., Santini, F., Eastman, J., Smith, S. A., Sidlauskas, B., Chang, J., and Alfaro, M. E. (2013). Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nat. Commun.*, 4(1958):1–8.

Rambaut, A. and Grass, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13:235–38.

Reinhold, K., Engqvist, L., Misof, B., and Kurtz, J. (1999). Meiotic drive and evolution of female choice. *Proc. R. Soc. Lond. B*, 266:1341–45.

Roda, F., Mendes, F. K., Hahn, M. W., and Hopkins, R. (2017). Genomic evidence of gene flow during reinforcement in Texas *Phlox*. *Mol. Ecol.*, 26:2317–30.

Rosenberg, N. A. (2002). The probability of topological concordance of gene trees and species trees. *Theor. Popul. Biol.*, 61:225–247.

Rosenblum, E. B., James, T. Y., Zamudio, K. R., Poorten, T. J., Ilut, D., Rodriguez, D., Eastman, J. M., Richards-Hrdlicka, K., Joneson, S., Jenkinson, T. S., Longcore, J. E., Olea, G. P., Toledo, L. F., Arellano, M. L., Medina, E. M., Restrepo, S., Flechas, S. V., Berger, L., Briggs, C. J., and Stajich, J. E. (2013). Complex history of the amphibian-killing chytrid fungus revealed with genome resequencing data. *Proc. Natl. Acad. Sci. USA.*, 110:9385–90.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *J. R. Stat. Soc., B: Stat.*, 71(2):319–392.

Säilynoja, T., Bürkner, P.-C., and Vehtari, A. (2021). Graphical test for discrete uniformity and its applications in goodness of fit evaluation and multiple sample comparison. *arXiv preprint arXiv:2103.10522*.

Semple, C., Steel, M., et al. (2003). *Phylogenetics*, volume 24. Oxford University Press.

Servedio, M. R., Brandvain, Y., Dhole, S., Fitzpatrick, C. L., Goldberg, E. E., Stern, C. A., Cleve, J. V., and Yeh, D. J. (2014). Not just a theory – the utility of mathematical models in evolutionary biology. *PLoS Biology*, 12:e1002017.

Shapiro, B., Rambaut, A., and Drummond, A. J. (2000). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Genetics*, 23:7–9.

Shepherd, D. A. and Klaere, S. (2018). How well does your phylogenetic model fit your data? *Syst. Biol.*, 68:157–167.

Siepel, A. (2019). Challenges in funding and developing genomic software: roots and remedies. *Genome Biol.*, 20(147).

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Stat. Soc. B*, 55:3–23.

Stadler, T. (2010). Sampling-through-time in birth–death trees. *J. Theor. Biol.*, 267(3):396–404.

Stadler, T. (2011). Simulating trees with a fixed number of extant species. *Syst. Biol.*, 60:676–84.

Stadler, T. (2013). Recovering speciation and extinction dynamics based on phylogenies. *J. Evol. Biol.*, 26:1203–1219.

Steel, M. and Penny, D. (2000). Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol. Biol. Evol.*, 17:839–850.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2018). Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526:68–74.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Stat.*, 22:1701–62.

Tuffley, C. and Steel, M. (1997). Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull. Math. Biol.*, 59:581–607.

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

Van Valen, L. (1973). A new evolutionary law. *Evolutionary Theory*, 1:1–30.

Vaughan, T. G. and Drummond, A. J. (2013). A stochastic simulator of birth–death master equations with application to phylodynamics. *Mol. Biol. Evol.*, 30:1480.

Wakeley, J. (2009). *Coalescent theory: an introduction.* Greenwood Village: Roberts and Company Publishers.

Warren, D. L., Geneva, A. J., and Lanfear, R. (2017). Rwty (r we there yet): an r package for examining convergence of bayesian phylogenetic analyses. *Mol. Biol. Evol.*, 34(4):1016–1020.

Yule, G. U. (1924). A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, FRS. *Philos. Trans. R. Soc. London Ser. B*, 213:21–87.

Zhang, C. and Matsen, F. A. (2019). Variational bayesian phylogenetic inference. In *International Conference on Learning Representations*.

Zhang, R., Drummond, A. J., and Mendes, F. K. (2023). Fast Bayesian inference of phylogenies from multiple continuous characters. *bioRxiv*.