**Supplementary Material**

## 1. Methods

In the Bayesian framework, the marginal likelihood or evidence of data $D$ conditioned on model $\tau$ with associated parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$ is

$$p(D \mid \tau) = \int p(D \mid \boldsymbol{\theta}, \tau) p(\boldsymbol{\theta} \mid \tau) d\boldsymbol{\theta},$$

where $p(D \mid \boldsymbol{\theta}, \tau)$ is the probability of the data given parameters $\boldsymbol{\theta}$, $p(\boldsymbol{\theta} \mid \tau)$ is the prior on $\boldsymbol{\theta}$, and the integral is of dimension N.

Dependence on model $\tau$ is suppressed in the rest of the document to simplify notation.

### 1.1. Laplace method.

1.1.1. *Classical Laplace.* The Laplace method [Tierney and Kadane, 1986] approximates the marginal likelihood by approximating the posterior distribution using a multivariate normal distribution with mean equal to the maximum a posteriori estimates $\tilde{\boldsymbol{\theta}}$, and covariance $\tilde{\Sigma} = (-H)^{-1}$ where $H$ is the Hessian matrix of second derivatives of $\log(p(D \mid \boldsymbol{\theta})p(\boldsymbol{\theta}))$. Specifically, let us define $l(\boldsymbol{\theta}) = \log(p(D \mid \boldsymbol{\theta})p(\boldsymbol{\theta}))$ and Taylor-expand $l(\boldsymbol{\theta})$ around $\tilde{\boldsymbol{\theta}}$. Exponentiating this quadratic approximation leads to a normal distribution with $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\theta}}$ and $\tilde{\Sigma} = -H^{-1}$. Integrating the normal distribution yields the Laplace marginal likelihood estimator

$$\hat{p}_L(D) \approx (2\pi)^{d/2} \det(\tilde{\Sigma})^{1/2} p(D \mid \tilde{\boldsymbol{\theta}}) p(\tilde{\boldsymbol{\theta}}),$$

where $\det(\tilde{\Sigma})$ is the determinant of the covariance matrix.

Unfortunately, the above normal approximation is not always accurate in practice. In our specific phylogenetic setting, the positivity of branch lengths creates problems for the normal approximation. It is however possible to improve the normal approximation of the posterior and the Laplace method if we transform each variable $\theta_i$ using a one-to-one twice differentiable function $g$ such as $\theta_i = g(z_i)$ and $z_i = g^{-1}(\theta_i)$. Applying the chain rule, the Hessian of the posterior for the transformed parameters is

$$H^z_{i,j} = \frac{\partial^2 l}{\partial z_i \partial z_j} = \begin{cases} \frac{\partial l}{\partial \theta_i} \frac{\partial^2 \theta_i}{\partial z_i^2} + H_{ii} \left( \frac{\partial \theta_i}{\partial z_i} \right)^2 & \text{for } i = j, \\ H_{ij} \frac{\partial \theta_i}{\partial z_i} \frac{\partial \theta_j}{\partial z_j} & \text{otherwise.} \end{cases}$$

The transformation requires an adjustment to account for the distortion of the distribution hence insuring that the distribution integrates to 1. Therefore, given $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the density of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}) = \mathcal{N}(g^{-1}(\boldsymbol{\theta}) \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, |\det J_{g^{-1}}(\boldsymbol{\theta})|,$$

where $|\det J_{g^{-1}}(\boldsymbol{\theta})|$ is the absolute value of the determinant of the Jacobian matrix evaluated at $\boldsymbol{\theta}$. However, we find in practice that some branch length posteriors are monotonically decreasing functions with modes at 0, and thus the transformation approach is not sufficient to make the normal approximation accurate.

1.1.2. *The Laplus approximations.* However, while some transformations may work well for a branch or subset of branches, we find in practice that there is no one transformation that works well for all branches on a tree. As an alternative we use a family of approximations inspired by the Laplace that we call the Laplus approximations (in recognition of the fact that they are like the Laplace but designed for parameters on $\mathbb{R}^+$). We share with the Laplace approximation the assumption that the posterior is concentrated around the mode, $\tilde{\boldsymbol{\theta}}$. Unlike the Laplace approximation, we assume that branch lengths are mutually independent, such that we can make the approximation

$$p(\boldsymbol{\theta} \mid \tau, D) \approx \prod_i q(\theta_i; \boldsymbol{\phi}_i)$$

Here $q$ is a parametric distribution with known normalizing constant (such as the gamma distribution) that we will use to approximate the posterior distributions for each branch. For a given branch, $\boldsymbol{\phi}_i$ are the parameters of $q$ that approximate the marginal posterior of that branch. Let $C$ be a constant such that

$$p(\boldsymbol{\theta} \mid \tau, D) = C \times p(D \mid \tau, \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

That is, $C$ is the inverse of the marginal likelihood that we seek to estimate, and using our approximation above,

$$C = \frac{p(\boldsymbol{\theta} \mid \tau, D)}{p(D \mid \tau, \boldsymbol{\theta})p(\boldsymbol{\theta})} \approx \frac{\prod_i q(\theta_i; \boldsymbol{\phi}_i)}{p(D \mid \tau, \boldsymbol{\theta})p(\boldsymbol{\theta})}$$

Finally, by applying this equation at the posterior mode, our resulting estimate of the marginal likelihood is

$$\hat{p}_{\text{Laplus}}(D) = \hat{C}^{-1} = \frac{p(D|\tilde{\boldsymbol{\theta}})p(\tilde{\boldsymbol{\theta}})}{\prod_i q(\tilde{\theta}_i; \boldsymbol{\phi}_i)}$$

The general procedure for the Laplus approximations is similar regardless of parametric distributional family assumption $q$. Our goal is to take the joint MAP estimates of the branch lengths $\tilde{\boldsymbol{\theta}}$ and the vector of second derivatives of the log-posterior $(\frac{\partial^2 l}{\partial \theta_1^2}, \frac{\partial^2 l}{\partial \theta_2^2}, \ldots, \frac{\partial^2 l}{\partial \theta_n^2})$ and find the parameters of our approximating distributions for each branch, $\boldsymbol{\phi}_i$, by matching modes and second derivatives of the approximating and posterior distributions of branch lengths. The complete procedure is written here algorithmically.

(1) Find the (joint) MAP branch lengths, $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_n)$
(2) for $i$ in $1 : n$
    (i) Compute $\frac{\partial^2 l}{\partial \theta_i^2}$, the second derivative of the log unnormalized posterior with respect to the $i^{\text{th}}$ branch
    (ii) Find parameters of $\boldsymbol{\phi}_i$ by solving

$$\frac{d^2}{dx^2} \log(q(x; \boldsymbol{\phi}_i)) = \frac{\partial^2 l}{\partial \theta_i^2}\Big|_{\theta_i = \tilde{\theta}_i}$$
$$\text{mode}(q(x; \boldsymbol{\phi}_i)) = \tilde{\theta}_i$$

    (iii) Catch exceptions
(3) Compute the marginal likelihood as $\hat{p}_{\text{Laplus}}(D) = \frac{p(D|\tilde{\boldsymbol{\theta}})p(\tilde{\boldsymbol{\theta}})}{\prod_i q(\tilde{\theta}_i; \boldsymbol{\phi}_i)}$

Exceptions occur when elements of $\boldsymbol{\phi}_i$ are outside of the domain of support, when $H_{ii}$ is nonnegative (so the posterior has a mode at 0), or when elements of $\boldsymbol{\phi}_i$ are otherwise suspect (such as producing particularly high-variance distributions with very short branches). Exceptions and their handling depend on the distributional assumption, and so we describe exception handling in the section for each distribution individually. We consider three choices for $q$, the gamma distribution, the Beta′ distribution, and the lognormal distribution. Since the Laplus method is not derived through a Taylor expansion of the unnormalised posterior, it is not subject to some of the assumptions required by Laplace's method. Although both methods require the function to be twice differentiable, Laplace's method assumes that the global maxima $\tilde{\boldsymbol{\theta}}$ is not at the boundary of the interval of integration so that the first derivatives vanishes at $\tilde{\boldsymbol{\theta}}$. Zero-length branches have typically non-zero (i.e. negative) first derivatives and positive second derivatives making the Laplus method attractive. And while it is obvious that there must be some dependence between branch lengths, we find in practice that the posterior correlations between branch lengths are often quite small.

1.1.3. *Gamma-Laplus.* Here we seek to approximate the marginal posteriors of all branch lengths with gamma distributions. The vector $\boldsymbol{\phi}_i = (\alpha_i, \beta_i)$ contains the shape and rate parameters of the gamma distribution; the log probability density function of the gamma is

$$\log(\text{Gamma}(x; \alpha, \beta)) = \alpha \log(\beta) - \log(\Gamma(\alpha)) + (\alpha - 1)\log(x) - \beta x.$$

The first and second derivatives of the log gamma distribution with respect to $x$ are given by

$$\frac{d}{dx} \log(\text{Gamma}(x; \alpha, \beta)) = \frac{\alpha - 1}{x} - \beta,$$
$$\frac{d^2}{dx^2} \log(\text{Gamma}(x; \alpha, \beta)) = -\frac{\alpha - 1}{x^2}.$$

We make use of the second derivative of the log-posterior at the mode, $H_{ii} = \frac{\partial^2 l}{\partial \theta_i^2}\Big|_{\theta_i = \tilde{\theta}_i}$ to estimate $\hat{\alpha}_i$ using the second derivative of the log of the gamma distribution. Then we solve for $\hat{\beta}_i$ using the analytic

formula for gamma mode: $\tilde{\theta}_i = \frac{\alpha_i - 1}{\beta_i}$.

$$
\begin{aligned}
H_{ii} &= -\frac{\hat{\alpha}_i - 1}{\tilde{\theta}_i^2} \\
\hat{\alpha}_i &= 1 - \tilde{\theta}_i^2 H_{ii} \\
\hat{\beta}_i &= \frac{\hat{\alpha}_i - 1}{\tilde{\theta}_i} = \frac{-\tilde{\theta}_i^2 H_{ii}}{\tilde{\theta}_i} = -\tilde{\theta}_i H_{ii}
\end{aligned}
$$

We note two exceptions to handle with the GL approach. The first case are branches with a mode at 0, which have posteriors that are monotonically decreasing. The second case are branches that are short with oddly large variances. We detect branches of the first type by checking whether $\tilde{\theta}_i < \epsilon_1$ or $H_{ii} >= 0$. These branches are handled by fixing $\hat{\alpha_i} = 1$ (to ensure that the approximation is monotonically decreasing) and fitting $\hat{\beta}_i$ directly using the log-posterior calculated at $N$ points spaced evenly (on the log-scale) between $\tilde{\theta}_i$ and 0.5. We detect branches of the second type by checking whether $\tilde{\theta}_i < \epsilon_2$ and $\frac{\alpha_i}{\beta_i^2} > 0.1$. These branches are handled by fitting $\alpha_i, \beta_i$ to $N$ points spaced evenly (on the log-scale) between $\tilde{\theta}_i$ and 0.5, while constraining $\tilde{\theta}_i = \frac{\hat{\alpha}_i - 1}{\hat{\beta}_i}$ (such that the mode of the approximation to be the mode of the posterior). We use $N = 10$, $\epsilon_1 = 10^{-6}$, and $\epsilon_2 = 10^{-4}$.

1.1.4. *Beta′-Laplus.* Here we seek to approximate the marginal posteriors of all branch lengths as beta′ distributions. In this case, the vector $\boldsymbol{\phi}_i = (\alpha_i, \beta_i)$ concatenates the shape parameters of the beta′ distribution with log probability density function is

$$\log(\text{Beta}'(x; \alpha, \beta)) = -\log(B(\alpha, \beta)) + (\alpha - 1)\log(x) - (\alpha + \beta)\log(x + 1),$$

where $B$ is the beta function.
The first and second derivatives of the log beta′ distribution with respect to $x$ are given by

$$
\begin{aligned}
\frac{d}{dx}\log(\text{Beta}'(x; \alpha, \beta)) &= \frac{\alpha - 1}{x} - \frac{\alpha + \beta}{x + 1}, \\
\frac{d^2}{dx^2}\log(\text{Beta}'(x; \alpha, \beta)) &= -\frac{\alpha - 1}{x^2} + \frac{\alpha + \beta}{(x + 1)^2}.
\end{aligned}
$$

When $\alpha \leq 1$, the beta′ distribution collapses to a monotonically decreasing distribution. When $\alpha = 1$,

$$
\begin{aligned}
\log(\text{Beta}'(x; 1, \beta_i)) &= -\log(B(1, \beta_i)) + (1 - 1)\log(x) - (1 + \beta_i)\log(x + 1), \\
\log(\text{Beta}'(x; 1, \beta_i)) &= -\log(B(1, \beta_i)) - (1 + \beta_i)\log(x + 1), \\
\frac{d}{dx}\log(\text{Beta}'(x; 1, \beta_i)) &= -\frac{1 + \beta_i}{x + 1}.
\end{aligned}
$$

We make use of the second derivative at the mode, $H_{ii} = \frac{\partial^2 l}{\partial \theta_i^2}\Big|_{\theta_i = \tilde{\theta}_i}$ to estimate $\hat{\beta}_i$. Then we solve for $\hat{\alpha}_i$ using the fact that $\tilde{\theta}_i = \frac{\hat{\alpha}_i - 1}{\hat{\beta}_i + 1}$.

$$
\begin{aligned}
H_{ii} &= -\frac{\hat{\alpha}_i - 1}{\tilde{\theta}_i^2} + \frac{\hat{\alpha}_i + \hat{\beta}_i}{(\tilde{\theta}_i + 1)^2} \\
&= -\frac{1}{\tilde{\theta}_i}\frac{\hat{\alpha}_i - 1}{\frac{\hat{\alpha}_i - 1}{\hat{\beta}_i + 1}} + \frac{1}{\tilde{\theta}_i + 1}\frac{\hat{\alpha}_i + \hat{\beta}_i}{\frac{\hat{\alpha}_i + \hat{\beta}_i}{\hat{\beta}_i + 1}} \\
&= -\frac{1}{\tilde{\theta}_i}(\hat{\beta}_i + 1) + \frac{1}{\tilde{\theta}_i + 1}(\hat{\beta}_i + 1) \\
&= (\hat{\beta}_i + 1)\Big(\frac{1}{\tilde{\theta}_i + 1} - \frac{1}{\tilde{\theta}_i}\Big) \\
&= \frac{\hat{\beta}_i + 1}{\tilde{\theta}_i(\tilde{\theta}_i + 1)} \\
\hat{\beta}_i &= -H_{ii}(\tilde{\theta}_i + 1)\tilde{\theta}_i - 1 \\
\hat{\alpha}_i &= \tilde{\theta}_i(\hat{\beta}_i + 1) + 1.
\end{aligned}
$$

We note two exceptions to handle with the BL approach. To start, we check if $\hat{\beta}_i < 0$, which implies $H_{ii} > 0$, meaning the posterior should be monotonically decreasing. In this case, we set $\hat{\alpha}_i = 1$ and use the equations outlined below to fit $\hat{\beta}_i$. We then check if $\hat{\beta}_i < 2$, in which case our approximate posterior has suspiciously high variance, in which case we fit $\alpha_i, \beta_i$ to $N$ points spaced evenly (on the log-scale) between $\tilde{\theta}_i$ and 0.5, while constraining $\tilde{\theta}_i = \frac{\hat{\alpha}_i - 1}{\hat{\beta}_i + 1}$ (such that the mode of the approximation to be the mode of the posterior).

When we set $\hat{\alpha}_i = 1$ we can use the first derivative of the log-posterior, $\nabla_i = \frac{\partial l}{\partial \theta_i}\Big|_{\theta_i = \tilde{\theta}_i}$, to fit $\hat{\beta}_i$:

$$
\begin{aligned}
\nabla_i &= -\frac{1 + \hat{\beta}_i}{\tilde{\theta}_i + 1}, \\
\hat{\beta}_i &= -\nabla_i(\tilde{\theta}_i + 1) - 1.
\end{aligned}
$$

1.1.5. *Lognormal-Laplus.* Here we seek to approximate the marginal posteriors of all branch lengths as lognormal distributions. The vector $\boldsymbol{\phi}_i = (\mu_i, \sigma_i)$ concatenates the mean and standard deviation parameters of the lognormal distribution with log probability density function

$$
\log(\text{Lognormal}(x; \mu_i, \sigma_i)) = -\frac{\log(2\pi)}{2} - \log(x) - \log(\sigma_i) - \frac{(\log(x) - \mu_i)^2}{2\sigma_i^2}.
$$

The first and second derivatives of the log lognormal distribution with respect to $x$ are given by

$$
\begin{aligned}
\frac{d}{dx}\log(\text{Lognormal}(x; \mu_i, \sigma_i)) &= -\frac{1}{x} - \frac{\log(x) - \mu_i}{x\sigma_i^2}, \\
\frac{d^2}{dx^2}\log(\text{Lognormal}(x; \mu_i, \sigma_i)) &= \frac{1}{x^2} - \frac{-\log(x) + \mu_i + 1}{x^2\sigma_i^2}.
\end{aligned}
$$

We make use of the second derivative at the mode, $H_{ii} = \frac{\partial^2 l}{\partial \theta_i^2}\Big|_{\theta_i=\tilde{\theta}_i}$, and the fact that $\tilde{\theta}_i = e^{\mu_i - \sigma_i^2}$ to estimate $\hat{\sigma}_i^2$. Then we solve for $\hat{\mu}_i$ using the fact that $\log(\tilde{\theta}_i) = \mu_i - \sigma_i^2$.

$$\begin{aligned}
H_{ii} &= \frac{1}{\tilde{\theta}_i^2} - \frac{-\log(\tilde{\theta}_i) + \hat{\mu}_i + 1}{\tilde{\theta}_i^2 \hat{\sigma}_i^2} \\
&= \frac{1}{\tilde{\theta}_i^2} - \frac{-(\hat{\mu}_i - \hat{\sigma}_i^2) + \hat{\mu}_i + 1}{\tilde{\theta}_i^2 \hat{\sigma}_i^2} \\
&= \frac{1}{\tilde{\theta}_i^2} - \frac{1}{\tilde{\theta}_i^2 \hat{\sigma}_i^2} - \frac{\hat{\sigma}_i^2}{\tilde{\theta}_i^2 \hat{\sigma}_i^2}, \\
\hat{\sigma}_i^2 &= -\frac{1}{\tilde{\theta}_i^2 H_{ii}}, \\
\hat{\mu}_i &= \log(\tilde{\theta}_i) + \hat{\sigma}_i^2.
\end{aligned}$$

We note two exceptions to handle with the LL approach. The first case are branches with a mode at 0, which have posteriors that are monotonically decreasing. The second case are branches that are short with oddly large variances. We nest the cases such that we first check for branches that fall in either category, checking $\tilde{\theta}_i < \epsilon_1$ or $H_{ii} >= 0$ or $\hat{\mu} > 5$ (which happens when $\hat{\sigma}$ is suspiciously large). As there is no parameter regime in which the lognormal is monotonically decreasing, and suspiciously high-variance branches are not fit any better by a lognormal distribution than a gamma distribution, at this point we switch to approximating branches as gamma distributions and proceed with exceptions as in the GL approach.

1.2. **Importance sampling.** Importance sampling uses a reference or importance distribution from which values are drawn, allowing summaries to be calculated for an unknown distribution by taking into account the importance weights (probabilities of drawing the sampled values). If $g$ is an importance distribution then

$$\begin{aligned}
p(D) &= \int p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&= \int \frac{p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} g(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&= \mathbb{E}_g \left( \frac{p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} \right).
\end{aligned}$$

For a normalized density $g$, the estimate is given by,

$$\hat{p}_{\mathrm{IS}}(D) = \frac{1}{N} \sum_{i=1}^{N} \frac{p(D \mid \tilde{\boldsymbol{\theta}}_i) p(\tilde{\boldsymbol{\theta}}_i)}{g(\tilde{\boldsymbol{\theta}}_i)}, \tilde{\boldsymbol{\theta}}_i \sim g(\boldsymbol{\theta}).$$

For an unnormalized density $q$, the self normalized importance sampling estimate [Owen, 2013] is given by

$$\hat{p}_{\mathrm{IS}}(D) = \frac{\sum_{i=1}^{N} p(D \mid \tilde{\boldsymbol{\theta}}_i) w(\tilde{\boldsymbol{\theta}}_i)}{\sum_{i=1}^{N} w(\tilde{\boldsymbol{\theta}}_i)}, \tilde{\boldsymbol{\theta}}_i \sim q(\boldsymbol{\theta}),$$

where $w(\tilde{\boldsymbol{\theta}}_i)$ is the importance weight given by $w(\tilde{\boldsymbol{\theta}}_i) = \frac{p(\tilde{\boldsymbol{\theta}}_i)}{q(\tilde{\boldsymbol{\theta}}_i)}$.

1.3. **Naive Monte Carlo.** The simplest Monte Carlo estimator of the marginal likelihood is defined as the expected value of the likelihood with respect to the prior distribution [Hammersley and Handscomb, 1964, Raftery and Banfield, 1991]. The so called naive Monte Carlo (NMC) estimator can be approximated by drawing $N$ samples $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N$ from the prior distribution and calculating the arithmetic mean of the likelihood.

$$\hat{p}_{\mathrm{NMC}}(D) = \frac{1}{N} \sum_{i=1}^{N} p(D \mid \tilde{\boldsymbol{\theta}}_i), \tilde{\boldsymbol{\theta}}_i \sim p(\boldsymbol{\theta}).$$

Although this approach is fast and unbiased, the high-likelihood region can be distant from the high-prior region. Most $\tilde{\boldsymbol{\theta}}_i$s will therefore be sampled from a region of the likelihood with low probability yielding high variance [Newton and Raftery, 1994].

1.4. **Harmonic mean.** The harmonic mean (HM) estimator only requires samples from the posterior generated by a single MCMC or other samplers and is therefore appealing to the user [Newton and Raftery, 1994]. The harmonic mean estimator of marginal estimator is equivalent to an importance sampling estimator of $1/p(D)$ with importance distribution $p(\boldsymbol{\theta} \mid D)$:

$$\hat{p}_{\mathrm{HM}}(D) = \frac{1}{\frac{1}{N}\sum_{i=1}^{N} \frac{1}{p(D|\tilde{\boldsymbol{\theta}}_i)}}, \tilde{\boldsymbol{\theta}}_i \sim p(\boldsymbol{\theta} \mid D).$$

This estimator is unstable due to the possible occurrence of small likelihood values the estimator and hence this estimator has infinite variance. Although the Law of Large Numbers guarantees that this estimator is consistent, the number of samples required to get an accurate estimate can be prohibitively high.

1.5. **Stabilized harmonic mean.** Newton and Raftery [1994] also proposed the stabilized harmonic mean (SHM) estimator to address the instability of the HM estimator. The SHM estimator is based on importance sampling scheme where the importance sampling distribution is a mixture of the prior and the posterior: $p^\star(\boldsymbol{\theta}) = \delta p(\boldsymbol{\theta}) + (1-\delta)p(\boldsymbol{\theta} \mid D)$ where $\delta$ is small, such that

$$\hat{p}_{\mathrm{SHM}^*}(D) = \frac{\sum_{i=1}^{n} \frac{p(D|\tilde{\boldsymbol{\theta}}_i)}{\delta\hat{p}_{\mathrm{SHM}^*}(D)+(1-\delta)p(D|\tilde{\boldsymbol{\theta}}_i)}}{\sum_{i=1}^{n}\{\delta\hat{p}_{\mathrm{SHM}^*}(D) + (1-\delta)p(D \mid \tilde{\boldsymbol{\theta}}_i)\}^{-1}}, \tilde{\boldsymbol{\theta}}_i \sim p^\star(\boldsymbol{\theta}).$$

Unfortunately this estimator requires simulating from both the posterior and prior. Newton and Raftery proposed to simulate from the posterior and assume that a further $\frac{\delta n}{(1-\delta)}$ observations are drawn from the prior, all of them with their likelihoods equal to their expected value $p(D)$. The likelihood of the imaginary samples drawn from the prior is $p(D \mid \theta_j) = \hat{p}_{SHM}$ for $j = 1, \ldots, \frac{\delta n}{1-\delta}$. Then, the approximate marginal likelihood $\hat{p}_{\mathrm{SHM}}(D)$ satisfies the following equation:

$$\hat{p}_{\mathrm{SHM}}(D) = \frac{\frac{\delta n}{1-\delta} + \sum_{i=1}^{n} \frac{p(D|\tilde{\boldsymbol{\theta}}_i)}{\delta\hat{p}_{\mathrm{SHM}}(D)+(1-\delta)p(D|\tilde{\boldsymbol{\theta}}_i)}}{\frac{\delta n}{(1-\delta)\hat{p}_{\mathrm{SHM}}(D)} + \sum_{i=1}^{n}\{\delta\hat{p}_{\mathrm{SHM}}(D) + (1-\delta)p(D \mid \tilde{\boldsymbol{\theta}}_i)\}^{-1}}, \tilde{\boldsymbol{\theta}}_i \sim p(\boldsymbol{\theta} \mid D),$$

which is solved by an iterative scheme that updates an initial guess of the marginal likelihood (e.g. harmonic mean estimate) until a stopping criterion is satisfied. In our implementation the recursion stops when the absolute change in $\log \hat{p}_{\mathrm{SHM}}(D)$ is less than $10^{-7}$. Newton and Raftery [1994] advocate $\delta = 0.01$ while Lartillot and Philippe [2006] use $\delta = 0.1$. In this study we used the $\hat{p}_{\mathrm{SHM}}$ with $\delta = 0.01$.

1.6. **Bridge sampling.** Bridge sampling (BS) was initially developed to estimate Bayes factors [Kass and Raftery, 1995] and was more recently adapted to approximate the marginal likelihood of a single model [Overstall and Forster, 2010, Gronau et al., 2017]. Following a derivation by Gronau et al. [2017], the bridge sampling estimator is derived from the following identity:

$$1 = \frac{\int p(D \mid \boldsymbol{\theta})p(\boldsymbol{\theta})h(\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int p(D \mid \boldsymbol{\theta})p(\boldsymbol{\theta})h(\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}},$$

where $g(\boldsymbol{\theta})$ is the proposal distribution and $h(\boldsymbol{\theta})$ is the bridge function. The bridge function ensures that the denominator in the identity is not zero.

Multiplying both sides of the above identity by $p(D)$ the bridge sampling estimator of the marginal likelihood is

$$p_{\mathrm{BS}}(D) = \frac{\int p(D \mid \boldsymbol{\theta})p(\boldsymbol{\theta})h(\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int h(\boldsymbol{\theta})g(\boldsymbol{\theta})p(\boldsymbol{\theta} \mid D)d\boldsymbol{\theta}} = \frac{\mathbb{E}_{g(\boldsymbol{\theta})}(p(D \mid \boldsymbol{\theta})p(\boldsymbol{\theta})h(\boldsymbol{\theta}))}{\mathbb{E}_{p(\boldsymbol{\theta}|D)}(h(\boldsymbol{\theta})g(\boldsymbol{\theta}))}.$$

The marginal likelihood is approximated using $n_1$ samples from the posterior distribution and $n_2$ samples from the proposal distribution

$$\hat{p}_{\mathrm{BS}}(D) = \frac{1/n_2 \sum_{i=1}^{n_2}(p(D \mid \tilde{\boldsymbol{\theta}}_i)p(\tilde{\boldsymbol{\theta}}_i)h(\tilde{\boldsymbol{\theta}}_i))}{1/n_1 \sum_{j=1}^{n_1} h(\boldsymbol{\theta}_j^*)g(\boldsymbol{\theta}_j^*)}, \tilde{\boldsymbol{\theta}}_i \sim g(\boldsymbol{\theta}), \boldsymbol{\theta}_j^* \sim p(\boldsymbol{\theta} \mid D).$$

Several bridge functions can be used including the so called *optimal bridge function* [Meng and Wong, 1996]:

$$h(\boldsymbol{\theta}) = \frac{C}{s_1 p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) + s_2 p(D) g(\boldsymbol{\theta})},$$

where $s_1 = n_1/(n_1 + n_2)$ and $s_2 = n_2/(n_1 + n_2)$ and $C$ is a constant that cancels out.

The definition of the optimal bridge function depends on the marginal likelihood itself, suggesting an iterative scheme to approximate $p(D)$ starting from an initial guess, such as the HM estimate. Gronau et al. [2017] provide a detailed description of an algorithm.

1.7. **Thermodynamic integration (aka path sampling, power posterior).** The thermodynamic integration estimator was introduced by Lartillot and Philippe [2006] in the phylogenetic context, borrowing ideas from path sampling [Gelman and Meng, 1998] and the physics literature where a large body of research is dedicated to the estimation of normalisation constants. Lartillot and Philippe defined a path going from the prior to the unnormalised posterior $q$ using

$$q_\beta = p(D \mid \boldsymbol{\theta})^\beta p(\boldsymbol{\theta})$$

for $\beta \in [0, 1]$. The normalisation constant $Z_\beta$ of the tempered unnormalised posterior is therefore

$$Z_\beta = \int_{\boldsymbol{\theta}} p(D \mid \boldsymbol{\theta})^\beta p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

and the log marginal likelihood of the model follows from the path sampling identity:

$$\log p(D) = \log Z_1 - \log Z_0 = \int_0^1 \frac{\partial Z_\beta}{\partial \beta} d\beta = \int_0^1 E_{\boldsymbol{\theta}|D,\beta}(\log p(D \mid \boldsymbol{\theta})) d\beta.$$

Friel and Pettitt [2008] worked on similar ideas but differ in the choice of temperature schedule and how the integral over [0,1] is approximated. Lartillot and Philippe [2006] approximate the integral using the Simpson's rule while Friel and Pettitt [2008] applied the trapezoidal rule. The interval $\beta \in [0, 1]$ is discretized such that $0 = \beta_0 < \beta_1 < \cdots < \beta_K = 1$ and for each $\beta_i$ samples are drawn from $p(\boldsymbol{\theta} \mid D, \beta_i)$ to estimate $E_{\boldsymbol{\theta}|D,\beta_i}(\log p(D \mid \boldsymbol{\theta}))$. For example, using the trapezoidal rule the log marginal likelihood of a given model is

$$\log \hat{p}_{\mathrm{PS}}(D) \approx \sum_{i=1}^{K} (\beta_i - \beta_{i-1}) \left( \frac{E_{i-1} + E_i}{2} \right),$$

where $E_i = E_{\boldsymbol{\theta}|\beta_i} \log p(D \mid \boldsymbol{\theta})$ is the expectation of the log deviance at $\beta_i$.

Lartillot and Philippe [2006] used equally spaced inverse temperatures between 0 and 1, while Friel and Pettitt [2008] set $\beta_i = (i/K)^5$. It is clear that other temperature schedules can be exploited such as a schedule based on the quantiles of parametric distribution [Xie et al., 2010] (see stepping stone section) and the adaptive scheme proposed by Friel et al. [2014]. Friel et al. [2014] subsequently proposed a modified trapezoidal rule that uses the variance of the samples to improve the approximation:

$$\log \hat{p}_{\mathrm{MPS}}(D) \approx \sum_{i=1}^{K} (\beta_i - \beta_{i-1}) \left( \frac{E_{i-1} + E_i}{2} \right) - \sum_{i=1}^{K} \frac{(\beta_i - \beta_{i-1})^2}{12} (V_i - V_{i+1}),$$

where $V_i = V_{\boldsymbol{\theta}|\beta_i}(\log p(D \mid \boldsymbol{\theta}))$ is the variance of the log deviance at $\beta_i$.

1.8. **Stepping stone.** Xie et al. [2010] proposed the stepping stone (SS) algorithm that is related to the path sampling approach described in the previous section. It uses a series of distributions defining a path between the prior and posterior and therefore inherits the computational burden of path sampling. Thermodynamic integration and stepping stone differ in the choice of $\beta$ values: Xie et al. [2010] set $\beta_1, \ldots, \beta_n$ equal to the quantiles of a density with fixed parameters (e.g. beta distribution). This approach allows for a more intensive sampling of power posteriors with small $\beta$ values, for which the posterior is changing rapidly.

Let's define the unnormalized power posterior distribution $q_\beta = p(D \mid \boldsymbol{\theta})^\beta p(\boldsymbol{\theta})$ and normalized power posterior distribution $p_\beta = \frac{q_\beta}{c_\beta}$, where $c_\beta$ is the power marginal likelihood of the data. The aim of the method is to estimate the ratio $r_{\mathrm{SS}} = c_{1.0}/c_{0.0}$, which is equal to $c_{1.0}$ if the prior is proper. This ratio can be expanded into a series of telescopic product of ratios using intermediate power posteriors

$$r_{\mathrm{SS}} = \frac{c_{1.0}}{c_{0.0}} = \prod_{k=1}^{K} \frac{c_{\beta_k}}{c_{\beta_{k-1}}} = \prod_{k=1}^{K} r_{\mathrm{SS},k},$$

where $r_{\mathrm{SS},k} = c_{\beta_k}/c_{\beta_{k-1}}$ for $k = 1, \ldots, K$. Xie et al. [2010] estimate each ratio $c_{\beta_k}/c_{\beta_{k-1}}$ by importance sampling using $p_{\beta_{k-1}}$ as the importance distribution. Using the definition of importance sampling the $k^{th}$ ratio is

$$\hat{r}_{\mathrm{SS},k} = \frac{1}{n} \sum_{i=1}^{n} \frac{p(D \mid \boldsymbol{\theta}_{k-1,i})^{\beta_k}}{p(D \mid \boldsymbol{\theta}_{k-1,i})^{\beta_{k-1}}} = \frac{1}{n} \sum_{i=1}^{n} p(D \mid \boldsymbol{\theta}_{k-1,i})^{\beta_k - \beta_{k-1}},$$

where $p(D \mid \theta_{k-1,i})$ is the likelihood function evaluated at $\theta_{k-1,i}$, the $i^{th}$ MCMC sample sampled from $p_{\beta_{k-1}}$. The product of the $K$ ratios $\hat{r}_{\mathrm{SS},k}$ yields the estimate of the marginal likelihood

$$\hat{p}_{\mathrm{SS}} = \prod_{k=1}^{K} \hat{r}_{\mathrm{SS},k}.$$

1.9. **Generalized stepping stone.** Although stepping stone proved to be more accurate than other approaches, such as path sampling [Xie et al., 2010], sampling distributions close to the prior (i.e., small $\beta$ values) can be difficult, particularly if the prior is diffuse. Fan et al. [2010] proposed to generalize the stepping stone method using a reference distribution that approximates the posterior distribution of interest using samples from the posterior distribution to parametrize the reference distribution. The reference distribution can be independent probability densities from the same family as the prior distribution or the product of densities with the same support. In our study the priors are exponential distributions, but we used gamma distributions that are parametrized using the method of moments. The shape and rate parameters are estimated by matching the first two moments of the gamma distribution to the marginal posterior sample mean and variance.

In the same vein as the SS method, the unnormalized and normalized power posterior distributions in the generalized stepping stone (GSS) approach are

$$q_\beta = \big(p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})\big)^{\beta} \big(p_0(\boldsymbol{\theta}; \boldsymbol{\phi})\big)^{1-\beta},$$
$$p_\beta = \frac{q_\beta}{c_\beta},$$

where $p(D \mid \boldsymbol{\theta})$ is the likelihood function, $p(\boldsymbol{\theta})$ is the prior distribution, $p_0$ is the reference distribution parametrized by $\boldsymbol{\phi}$, and $c_\beta$ is the (power) marginal likelihood of the data. The key difference with the SS approach is that for $\beta = 0$ the power posterior is equivalent to the reference distribution.

As for the SS method, the aim of this method is to estimate the ratio $r_{\mathrm{GSS}} = c_{1.0}/c_{0.0}$ using importance sampling. The ratio $\hat{r}_{\mathrm{GSS},k}$ is estimated using $n$ samples from $p_{\beta_{k-1}}$:

$$\hat{r}_{\mathrm{GSS},k} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{p(D \mid \boldsymbol{\theta}_{k-1,i}) p(\boldsymbol{\theta}_{k-1,i})}{p_0(\boldsymbol{\theta}_{k-1,i}; \boldsymbol{\phi})} \right)^{\beta_k - \beta_{k-1}}.$$

Combining $\hat{r}_{\mathrm{GSS},k}$ for all $K$ ratios yields the marginal likelihood estimator:

$$\hat{p}_{\mathrm{GSS}} = \prod_{k=1}^{K} \hat{r}_{\mathrm{GSS},k}.$$

1.10. **Nested sampling.** Nested sampling is a Monte Carlo method that aims at calculating the marginal likelihood using a change of variable [Skilling, 2004, Skilling et al., 2006]. It transforms the multidimensional evidence integral over the parameter space into a more manageable one-dimensional integral over the likelihood space. Skilling defines the prior volume as $dX = p(\boldsymbol{\theta}) d\boldsymbol{\theta}$ so that

$$X(\lambda) = \int_{\mathcal{L}(\boldsymbol{\theta}) > \lambda} p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{1}$$

where $\mathcal{L}(\boldsymbol{\theta})$ is the likelihood function and the integral is taken over the region bounded by the iso-likelihood contour $\mathcal{L}(\boldsymbol{\theta}) = \lambda$. The marginal likelihood becomes a one-dimensional integral over unit range

$$p_{\mathrm{NS}}(D) = \int_0^1 L(X)dX,$$

where $L(X)$ is the inverse function of $X(\lambda)$.

Assuming that $L(X)$ can be computed for a sequence of decreasing values $0 < X_m < \cdots < X_0 = 1$, the unit integral can be approximated using quadrature techniques as the weighted sum:

$$\hat{p}_{\mathrm{NS}}(D) \approx \sum_{i=1}^{m} L(X_i)w_i,$$

where $w_i = X_i - X_{i-1}$.

The nested sampling algorithm uses a clever process of sampling from the prior (hence $dX$) and conditioning on the likelihood being above a given size (to achieve the likelihood condition of (1)) to approximate the input to such a quadrature technique [Skilling et al., 2006, Maturana Russel et al., 2018]. The algorithm is initialized with $N$ samples $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$ drawn from the prior and their corresponding likelihoods are calculated $\{\mathcal{L}(\boldsymbol{\theta}_1), \ldots, \mathcal{L}(\boldsymbol{\theta}_N)\}$. The sample with the lowest likelihood $L_{\min}$ is discarded from the set and replaced by a new sample $\boldsymbol{\theta}^*$ drawn from the prior subject to the constraint $L > L_{\min}$. When we use the discarded point as an $X_i$, the other points in the set of course satisfy the likelihood constraint. There are a variety of choices for terminating the algorithm [Maturana Russel et al., 2018]. We choose to terminate when the absolute change in $\log(\hat{p}_{\mathrm{NS}}(D))$ is less than $10^{-6}$.

1.11. **Posterior predictive model selection.** As an alternative to the marginal likelihood, the fit of a model can be assessed through the accuracy of its predictions [Gelman et al., 1996]. The probability distribution of a new data set $\tilde{D}$ having observed data set $D$ is defined as

$$p(\tilde{D} \mid D) = \int p(\tilde{D} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid D)d\boldsymbol{\theta}.$$

1.11.1. *Log pointwise predictive density.* A related quantity is the expected log pointwise predictive density [Vehtari et al., 2017] for a new data set, with $n$ data points, is defined as

$$\mathrm{elpd} = \sum_{i=1}^{n} \int p_t(\tilde{D}_i) \log p(\tilde{D}_i \mid D)d\tilde{D}_i,$$

where $p_t(\tilde{D}_i)$ is the distribution representing the true data-generating process for $\tilde{D}_i$. In the phylogenetic framework, the observation $D_i$ corresponds to a single site in the alignment. Since the $p_t$ is not known, one can use cross-validation to approximate elpd (see next section).

As in [Vehtari et al., 2017], we define the log pointwise predictive density

$$\mathrm{lpd} = \sum_{i=1}^{n} \log p(D_i \mid D) = \sum_{i=1}^{n} \log \int p(D_i \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid D)d\boldsymbol{\theta},$$

where $p(D_i \mid \boldsymbol{\theta})$ is the likelihood of the $i^{th}$ observation. The log pointwise predictive density can be estimated using $S$ draws $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_S$ from the posterior distribution $p(\boldsymbol{\theta} \mid D)$, by summing over the $n$ data points

$$\widehat{\mathrm{lpd}} = \sum_{i}^{n} \log\Big(\frac{1}{S}\sum_{s=1}^{S} p(D_i \mid \boldsymbol{\theta}_s)\Big), \boldsymbol{\theta}_s \sim p(\boldsymbol{\theta} \mid D).$$

We compared the fit of our topology models using the predictive accuracy approximation $\widehat{\mathrm{lpd}}$

$$\log \hat{p}_{\mathrm{PPD}}(D) = \widehat{\mathrm{lpd}}$$

as an estimate of the log marginal likelihood. Although we are not aware of others using it in this way, we have found that it provides a reasonable approximation. However, the lpd of observed data $D$ is an overestimate of the elpd for future data [Vehtari et al., 2017].

1.11.2. *Conditional predictive ordinates.* A related approach is the conditional predictive ordinates (CPO) method based on Bayesian leave-one-out (LOO).

The leave-one-out estimate of the predictive density for a datapoint is

$$\text{elpd}_{\text{loo}} = \sum_{i=1}^{n} \log p(D_i \mid D_{-i}) = \sum_{i=1}^{n} \log \int p(D_i \mid D_{-i}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D_{-i}) d\boldsymbol{\theta},$$

where $p(D_i \mid D_{-i})$ is the leave-one-out predictive density (aka conditional predictive ordinate) given the data without the $i^{th}$ data point.

The CPO estimate of this is given by

$$\hat{p}(D_i \mid D_{-i}) = \frac{1}{\frac{1}{S}\sum_{i=1}^{S} \frac{1}{p(D_i \mid \boldsymbol{\theta}_s)}}, \boldsymbol{\theta}_s \sim p(\boldsymbol{\theta} \mid D).$$

The resulting estimate of the log marginal likelihood (called the log pseudo-marginal likelihood by Lewis et al. [2013]) is given by

$$\log \hat{p}_{\text{CPO}}(D) = \widehat{\text{lpd}_{\text{loo}}} = \sum_{i=1}^{n} \log \hat{p}(D_i \mid D_{-i})$$

1.12. **Variational inference.** Variational Bayes methods provide an analytical approximation to the posterior probability and a lower bound for the marginal likelihood. The main idea is to choose a family of distributions $q$ parametrised with parameters $\boldsymbol{\phi}$ and to minimize the Kullback Leibler (KL) divergence from variational distribution $q$ to the posterior distribution $p$ of interest

$$\boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} \text{KL}(q(\boldsymbol{\theta}; \phi) \parallel p(\boldsymbol{\theta} \mid D)).$$

It is difficult to minimise the KL divergence directly but much easier to minimize a function that is equal to it up to a constant. Expanding the KL divergence we get

$$\begin{aligned} \text{KL}(q(\boldsymbol{\theta}; \boldsymbol{\phi}) \parallel p(\boldsymbol{\theta} \mid D)) &= \mathbb{E}[\log q(\boldsymbol{\theta}; \boldsymbol{\phi})] - \mathbb{E}[\log p(\boldsymbol{\theta} \mid D)] \\ &= \mathbb{E}[\log q(\boldsymbol{\theta}; \boldsymbol{\phi})] - \mathbb{E}[\log p(\boldsymbol{\theta}, D)] + \log p(D) \\ &= -\text{ELBO}(\boldsymbol{\phi}) + \log p(D), \end{aligned}$$

where $\text{ELBO}(\boldsymbol{\phi}) = \mathbb{E}[\log p(\boldsymbol{\theta}, D)] - \mathbb{E}[\log q(\boldsymbol{\theta}; \boldsymbol{\phi})]$. This equation suggests that the $\text{ELBO}(\boldsymbol{\phi})$ is the lower bound of the evidence: $\log p(D) \geq \text{ELBO}(\boldsymbol{\phi})$.

Instead of minimizing KL divergence, we maximize the evidence lower bound:

$$\text{ELBO}(\boldsymbol{\phi}) = \mathop{\mathbb{E}}_{q(\boldsymbol{\theta}; \boldsymbol{\phi})} [\log p(D, \boldsymbol{\theta}) - \log q(\boldsymbol{\theta}; \boldsymbol{\phi})].$$

Several variational distributions can be used including the mean-field and fullrank Gaussian distributions. The fullrank model uses a multivariate Gaussian distribution to model the correlation between variables while the meanfield distribution assumes a diagonal covariance matrix. In this study we used the meanfield model hence taking the assumption that there is no correlation between the branch lengths of the phylogeny:

$$q(\boldsymbol{\theta}; \boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2)) = \prod_{i=1}^{n} \mathcal{N}(\theta_i; \mu_i, \sigma_i^2).$$

It is common to use stochastic gradient ascent algorithm to maximise the ELBO as long as the model is differentiable [Ranganath et al., 2014, Kucukelbir et al., 2015]. In the phylogenetic context the derivative of posterior with respect to the branch lengths can be derived analytically without resorting to approximations such as finite differences. We used a log transform on the branch lengths to ensure that the variational distribution stays within the support of the posterior.

Given an optimized variational model we used the ELBO as an approximation of the marginal likelihood

$$\hat{p}_{\text{ELBO}}(D) = \max_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} \text{ELBO}(\boldsymbol{\phi}).$$

The ELBO estimates can have high variance and might be of little use to discriminate between closely related models (in the KL sense). We used importance sampling to calculate the marginal likelihood of a model using the variational distribution $q$ as the importance distribution. This yields the $\hat{p}_{\mathrm{VBIS}}(D)$ estimator:

$$\hat{p}_{\mathrm{VBIS}}(D) = \frac{1}{N}\sum_{i=1}^{N} \frac{p(D \mid \tilde{\boldsymbol{\theta}}_i)p(\tilde{\boldsymbol{\theta}}_i)}{q_{\mathrm{ELBO}}(\tilde{\boldsymbol{\theta}}_i)}, \tilde{\boldsymbol{\theta}}_i \sim q_{\mathrm{ELBO}}(\boldsymbol{\theta}).$$

## 2. Supplementary Figures

For completion, we include here equivalents of Figure 3 and Figure 2 for datasets DS1-4. We also include versions of Figure 4 and Figure 1 that use KL divergence instead of RMSD as the measure of accuracy. The KL and RMSD results are qualitatively similar.

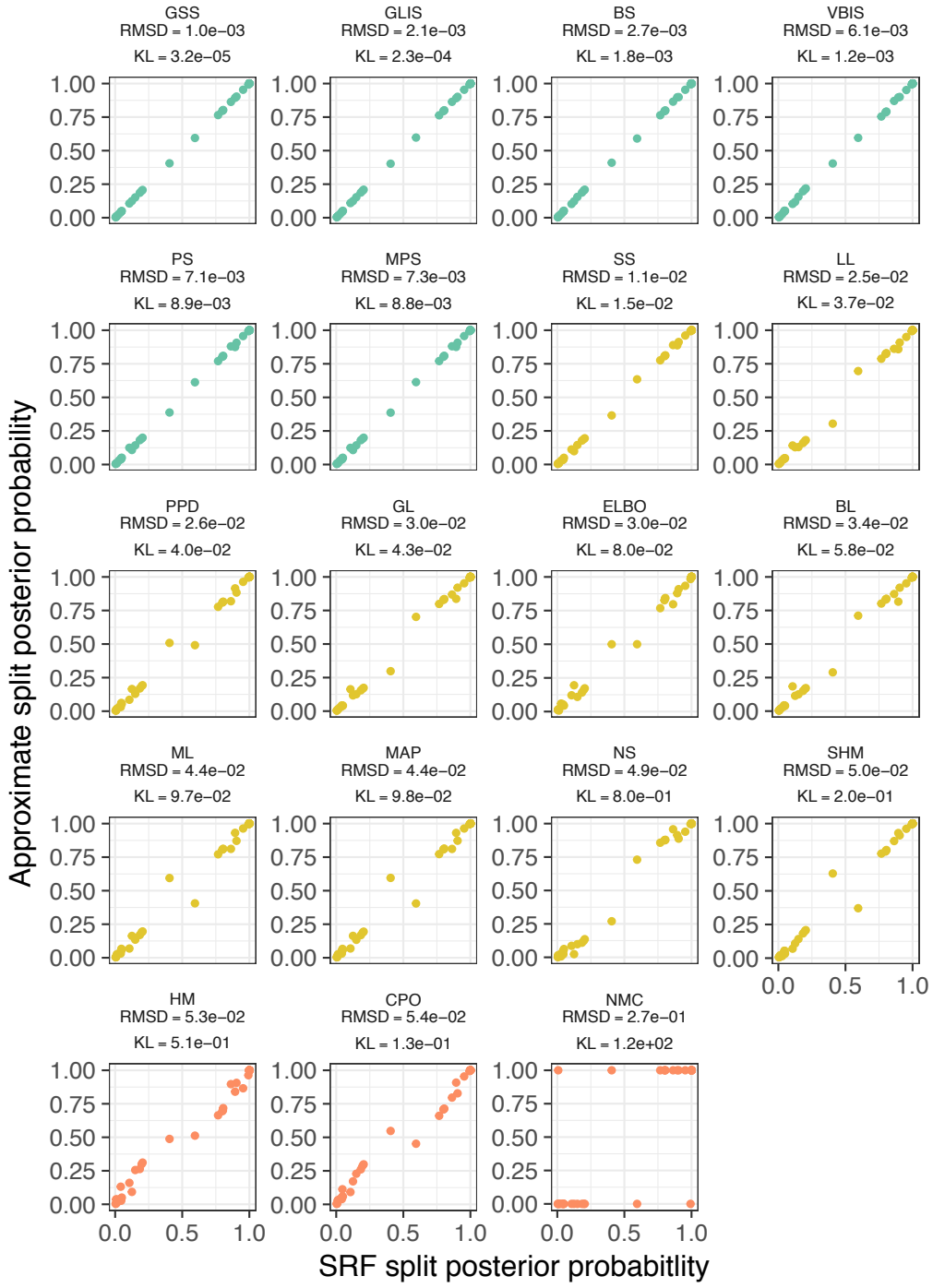


**Figure S1.** The posterior probabilities of all the splits observed in DS1 for a single replicate. MrBayes posteriors are plotted on the x-axis versus the denoted approximation on the y-axis. The line $y = x$ is provided for ease of interpretation, and points are colored by the thresholds we discuss: RMSD $< 0.01$ is a good approximation (green), $0.01 \leq$ RMSD $< 0.05$ is a potentially acceptable approximation (yellow), and RMSD $\geq 0.05$ is poor (red). Panels are ordered by RMSD in increasing order.

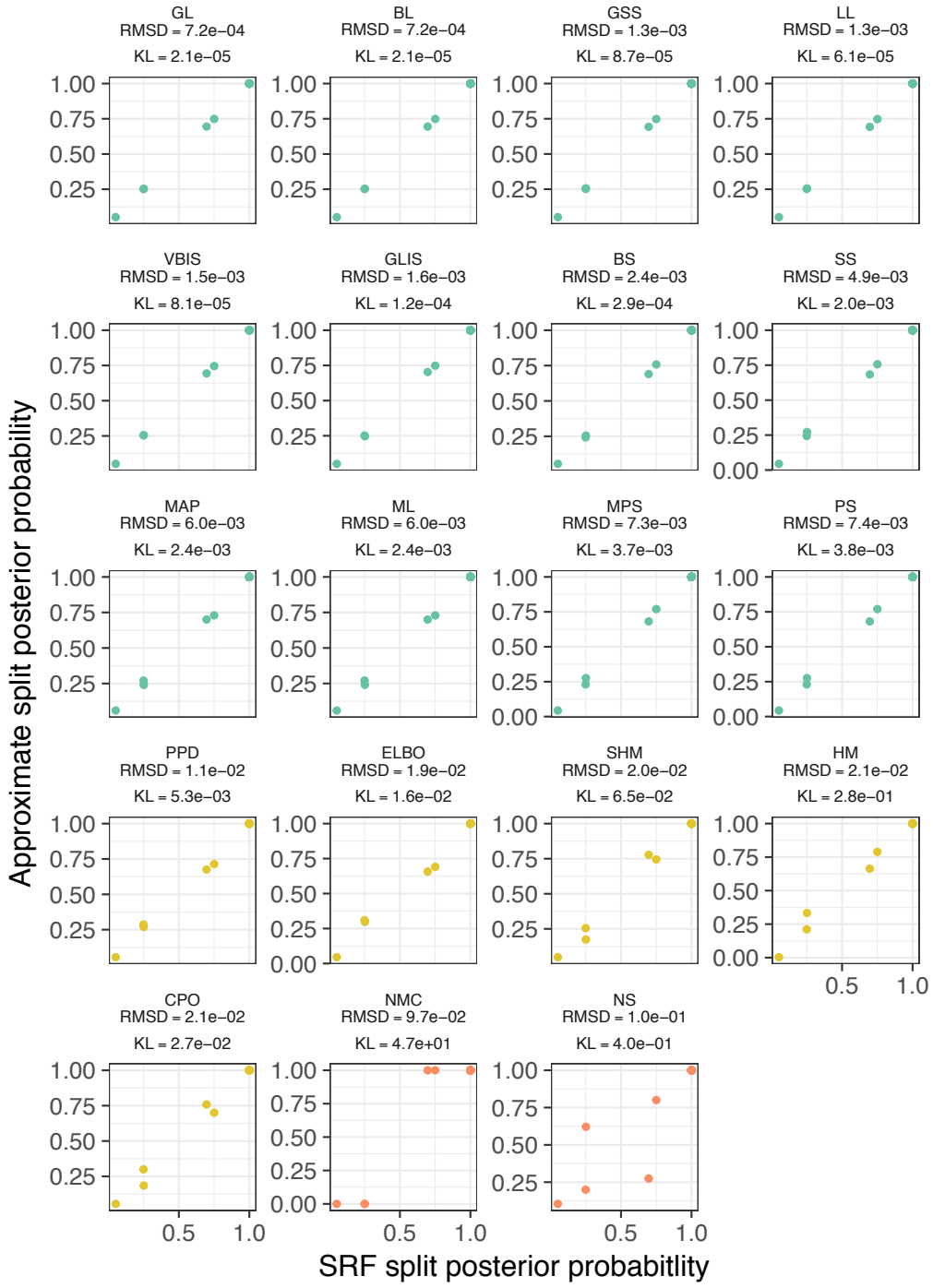


**Figure S2.** The posterior probabilities of all the splits observed in DS2 for a single replicate. MrBayes posteriors are plotted on the x-axis versus the denoted approximation on the y-axis. The line $y = x$ is provided for ease of interpretation, and points are colored by the thresholds we discuss: RMSD $< 0.01$ is a good approximation (green), $0.01 \leq$ RMSD $< 0.05$ is a potentially acceptable approximation (yellow), and RMSD $\geq 0.05$ is poor (red). Panels are ordered by RMSD in increasing order.

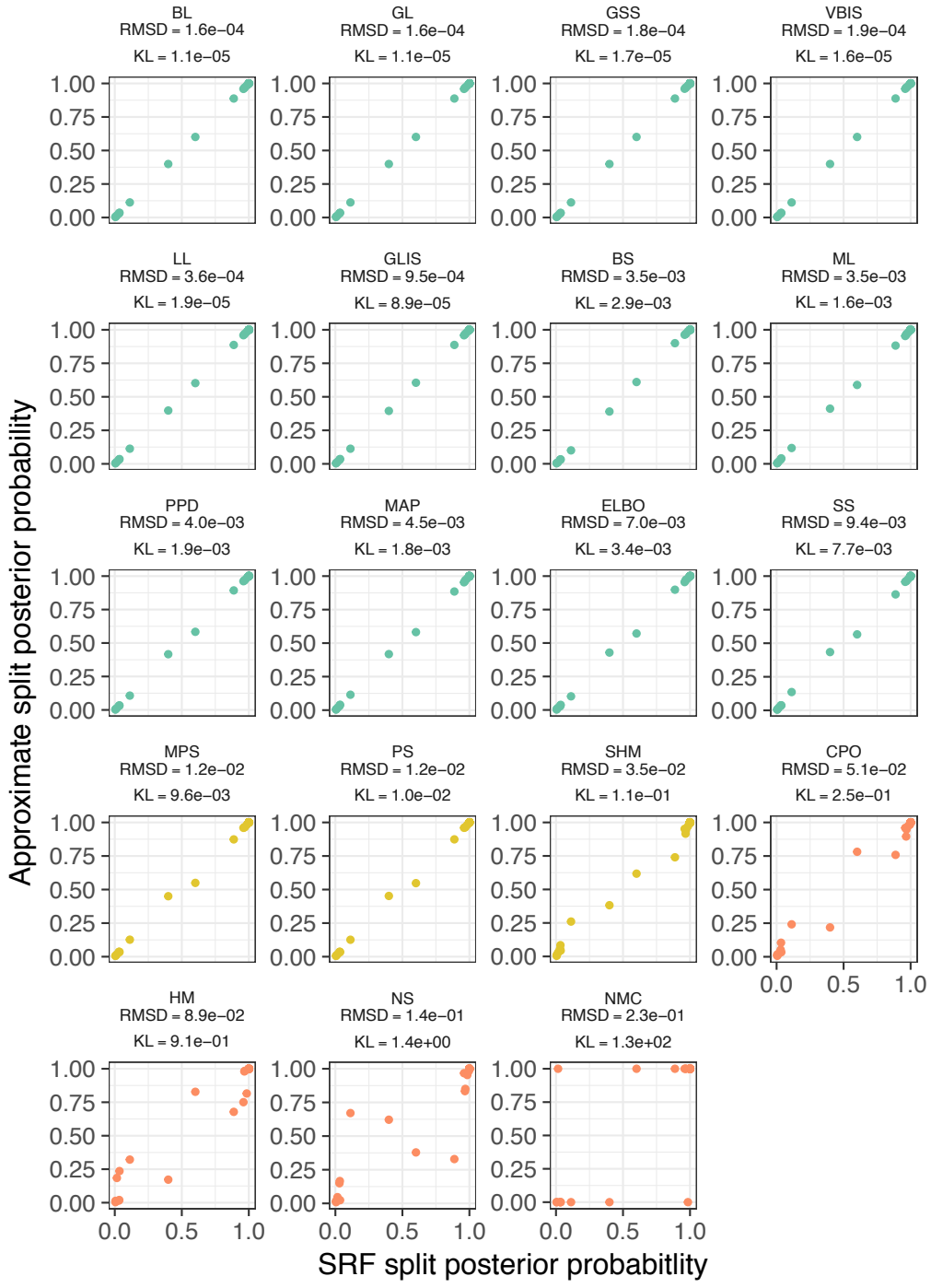


**Figure S3.** The posterior probabilities of all the splits observed in DS3 for a single replicate. MrBayes posteriors are plotted on the x-axis versus the denoted approximation on the y-axis. The line $y = x$ is provided for ease of interpretation, and points are colored by the thresholds we discuss: RMSD $< 0.01$ is a good approximation (green), $0.01 \leq$ RMSD $< 0.05$ is a potentially acceptable approximation (yellow), and RMSD $\geq 0.05$ is poor (red). Panels are ordered by RMSD in increasing order.

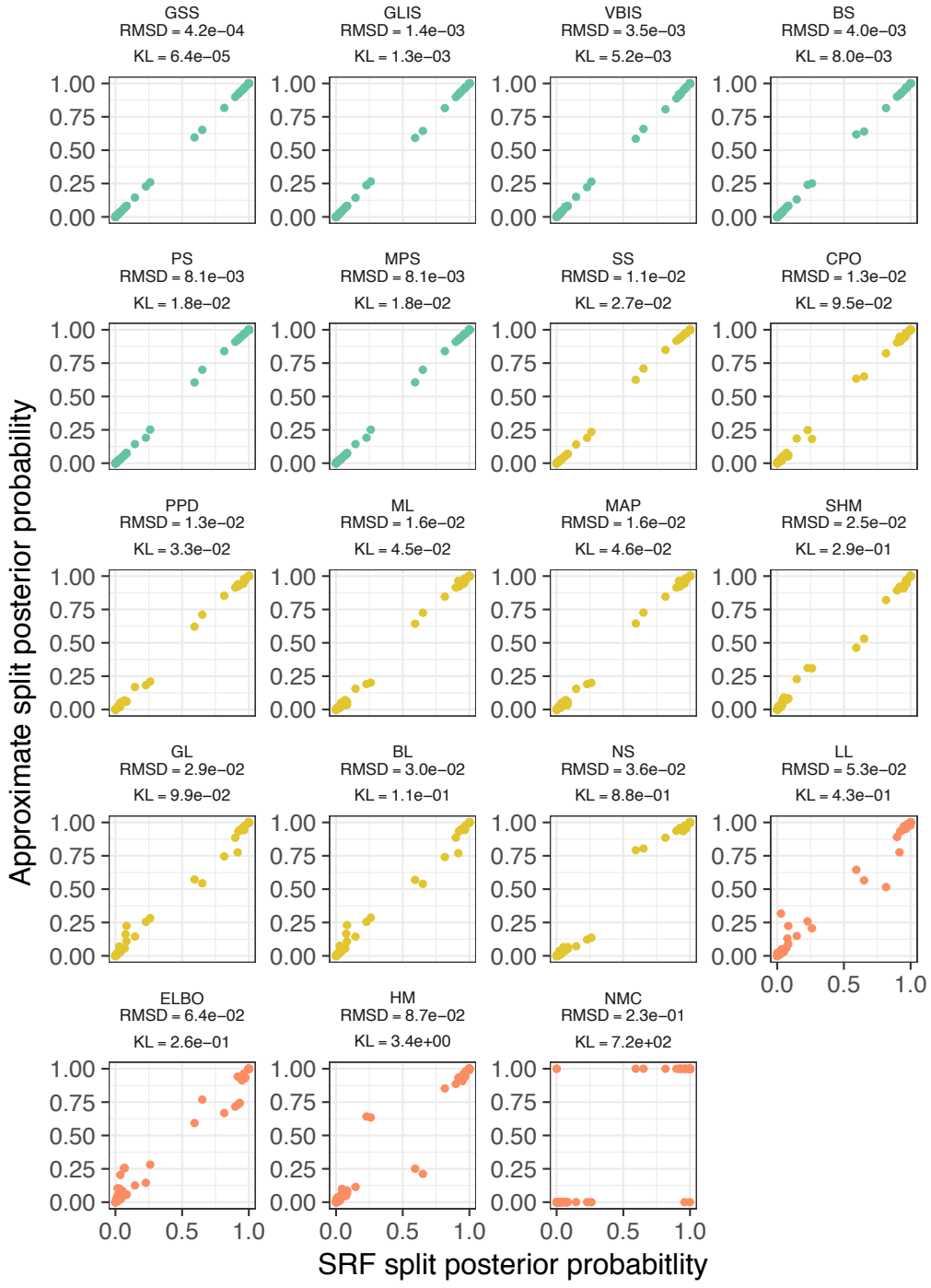


**Figure S4.** The posterior probabilities of all the splits observed in DS4 for a single replicate. MrBayes posteriors are plotted on the x-axis versus the denoted approximation on the y-axis. The line $y = x$ is provided for ease of interpretation, and points are colored by the thresholds we discuss: RMSD $< 0.01$ is a good approximation (green), $0.01 \leq$ RMSD $< 0.05$ is a potentially acceptable approximation (yellow), and RMSD $\geq 0.05$ is poor (red). Panels are ordered by RMSD in increasing order.

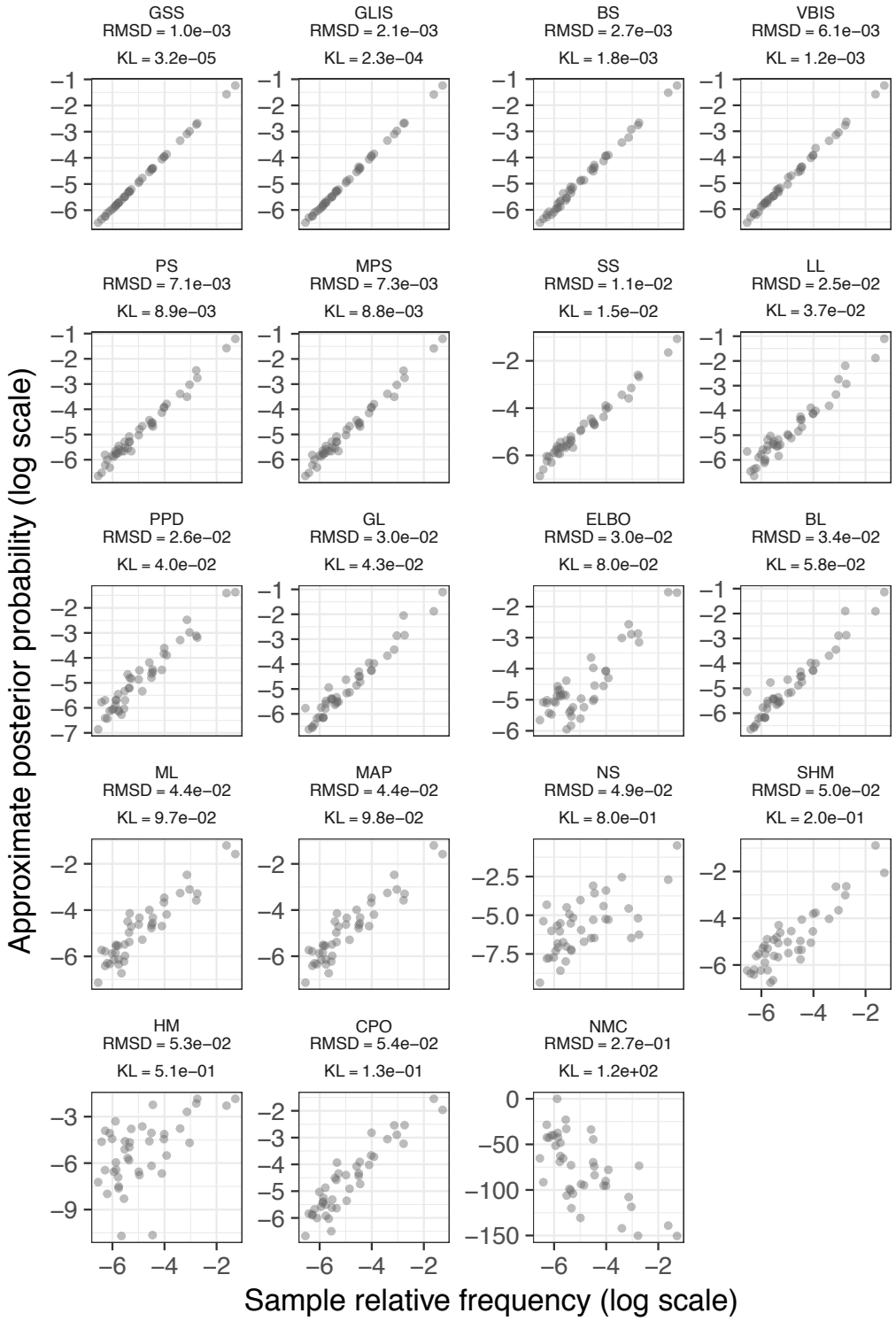


**Figure S5.** The approximate posterior probabilities of the topologies in DS1 versus the ground truth posterior probabilities from MrBayes, plotted on the log scale for clarity. Results are for a single run of each method. Panels are ordered by RMSD in increasing order.

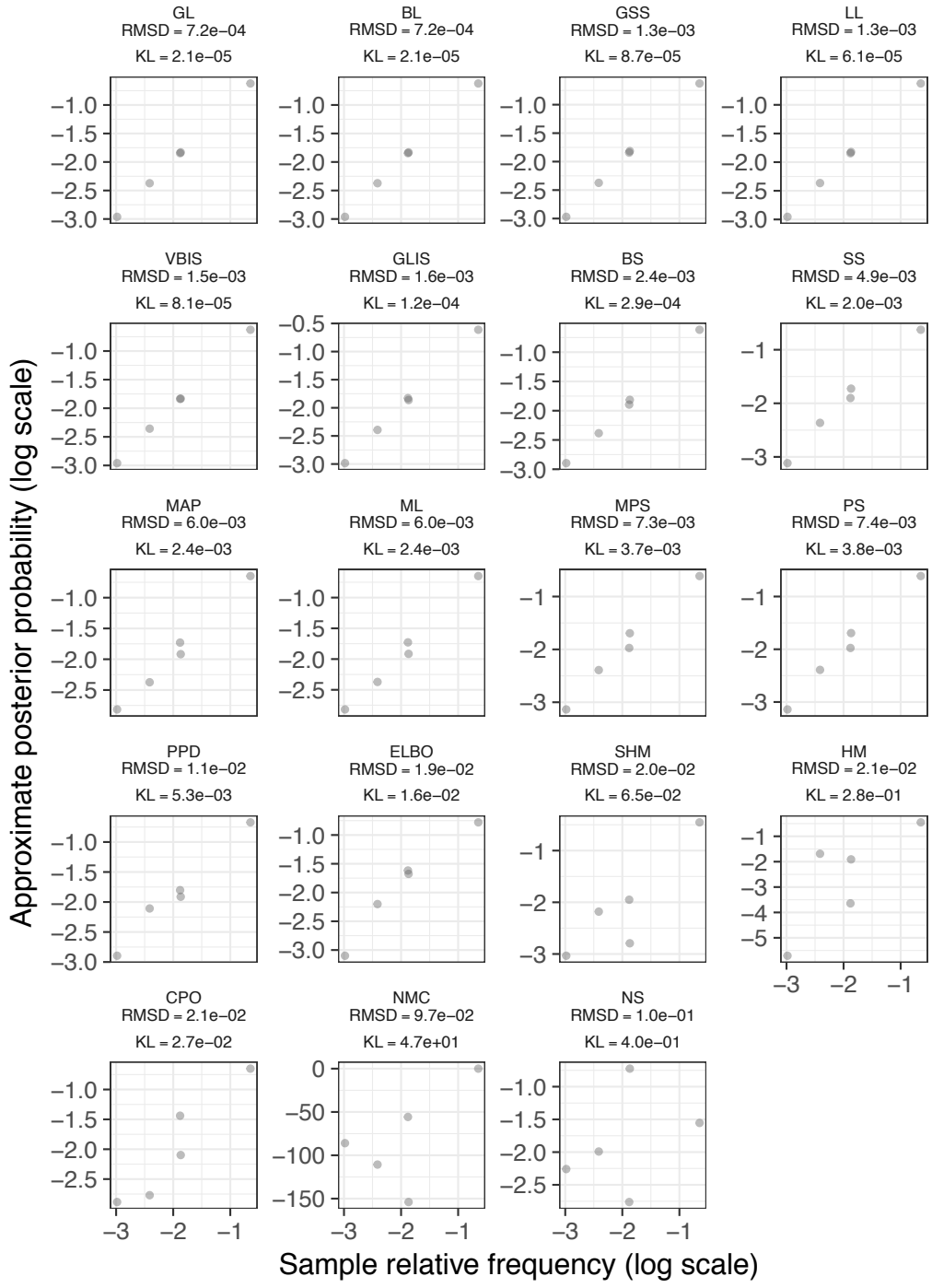


**Figure S6.** The approximate posterior probabilities of the topologies in DS2 versus the ground truth posterior probabilities from MrBayes, plotted on the log scale for clarity. Results are for a single run of each method. Panels are ordered by RMSD in increasing order.

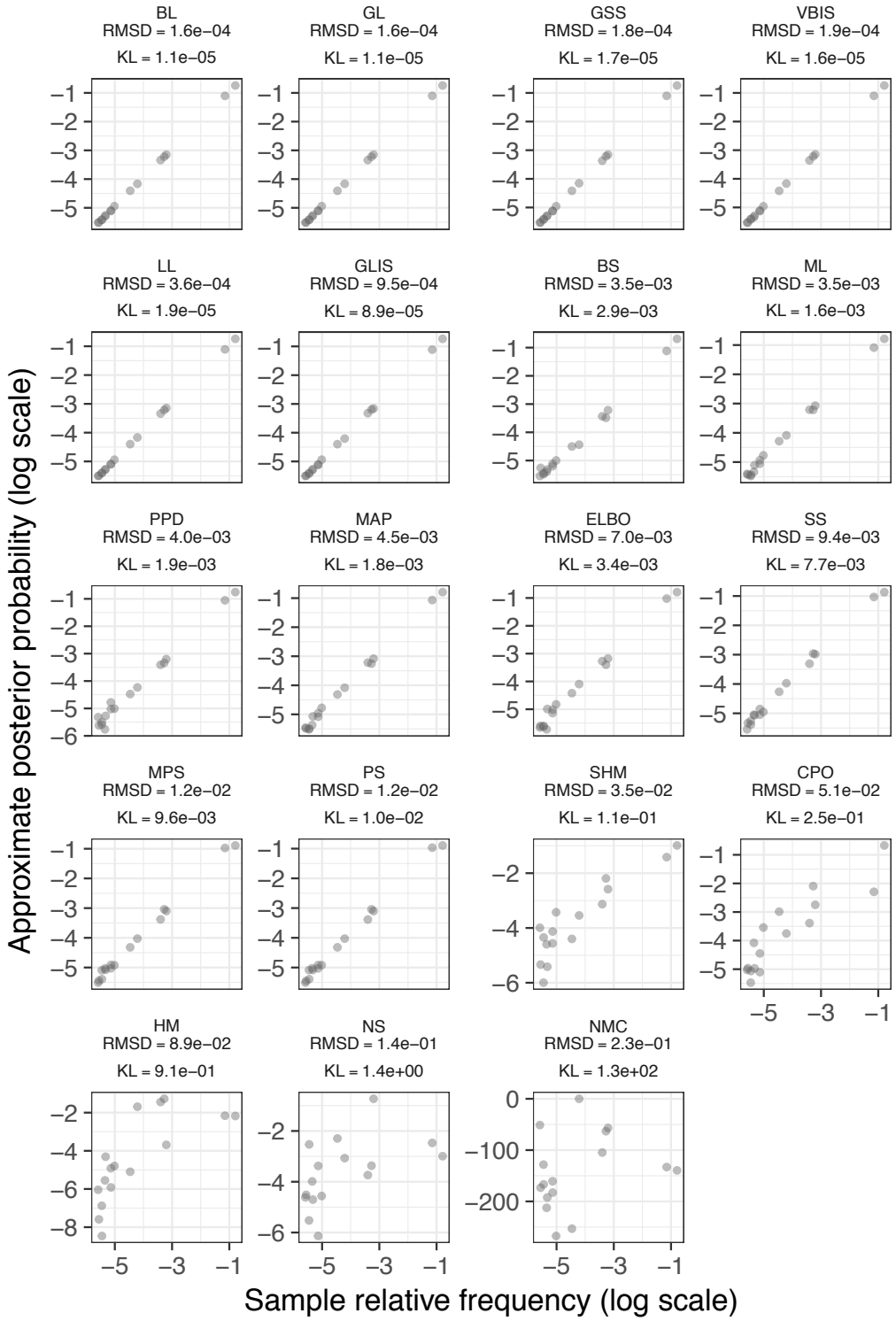


**Figure S7.** The approximate posterior probabilities of the topologies in DS3 versus the ground truth posterior probabilities from MrBayes, plotted on the log scale for clarity. Results are for a single run of each method. Panels are ordered by RMSD in increasing order.

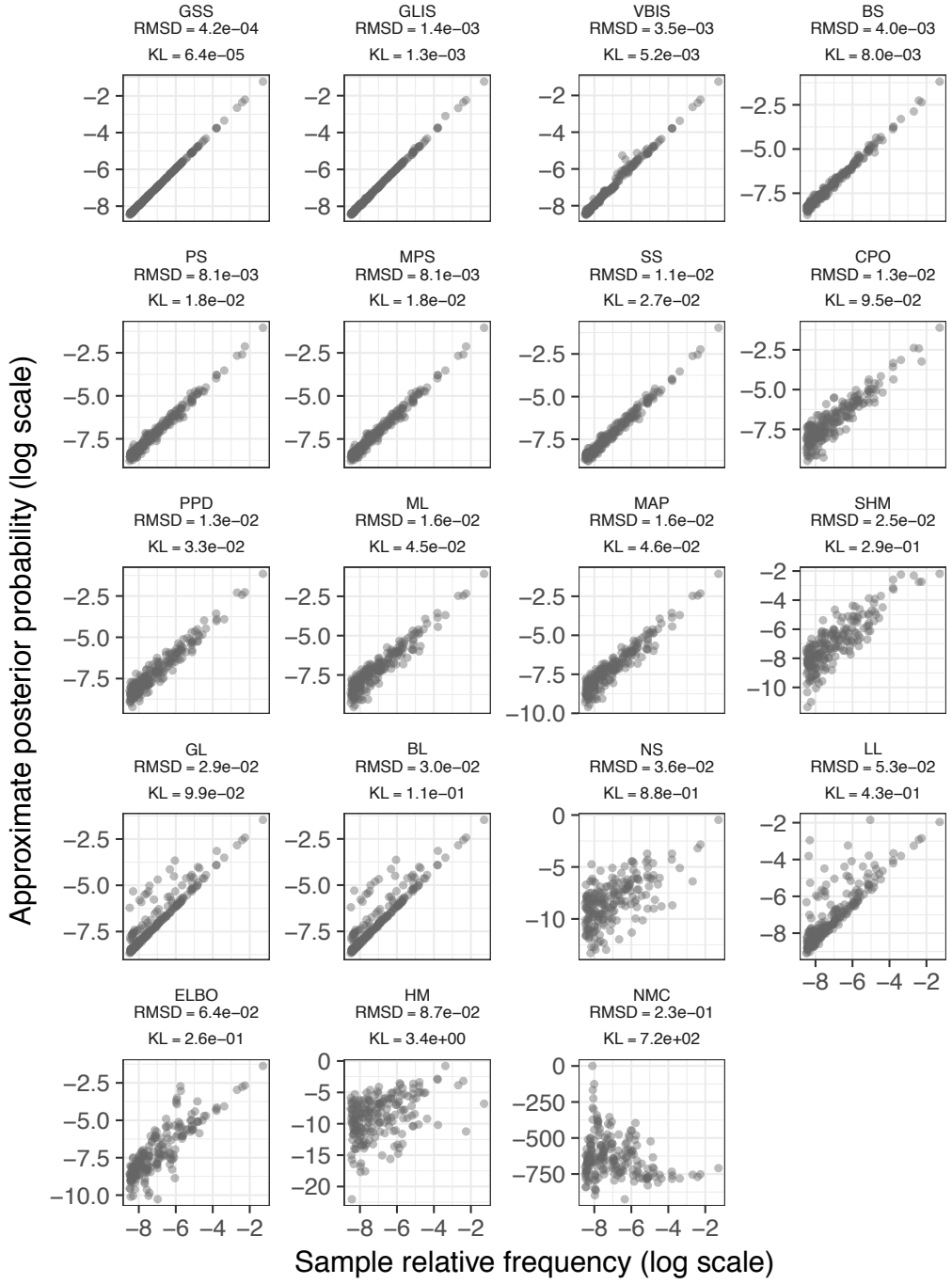


**Figure S8.** The approximate posterior probabilities of the topologies in DS4 versus the ground truth posterior probabilities from MrBayes, plotted on the log scale for clarity. Results are for a single run of each method. Panels are ordered by RMSD in increasing order.

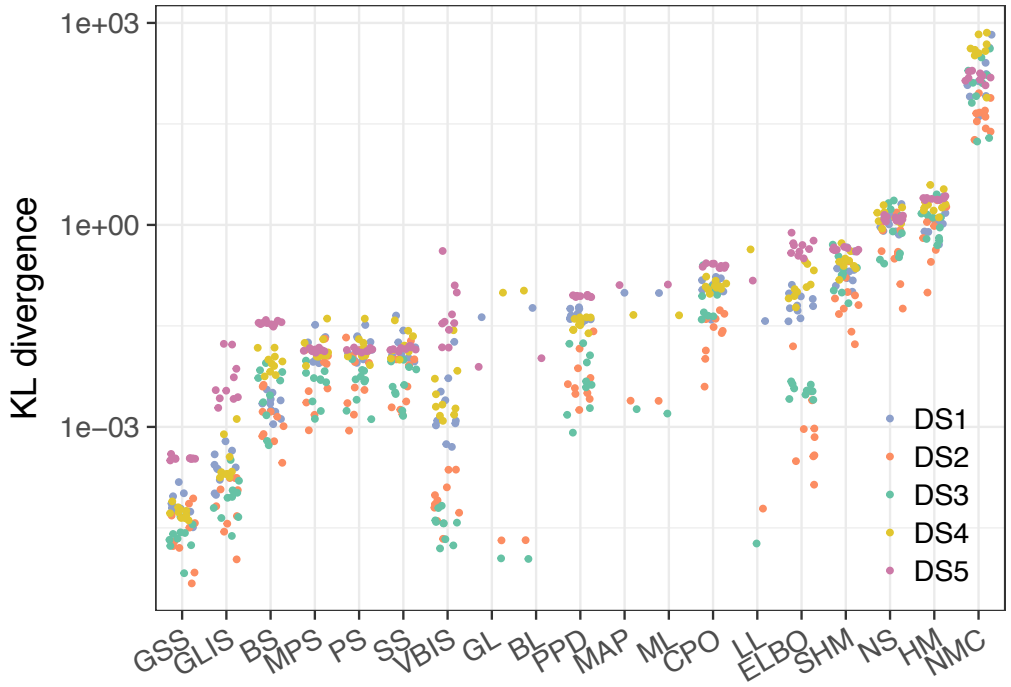


**Figure S9.** Average Kullback-Leibler (KL) divergence from MrBayes posteriors to approximate posteriors for each method on each dataset for 10 replicates. LL, GL, BL, MAP, and ML are deterministic and therefore only one replicate is shown.

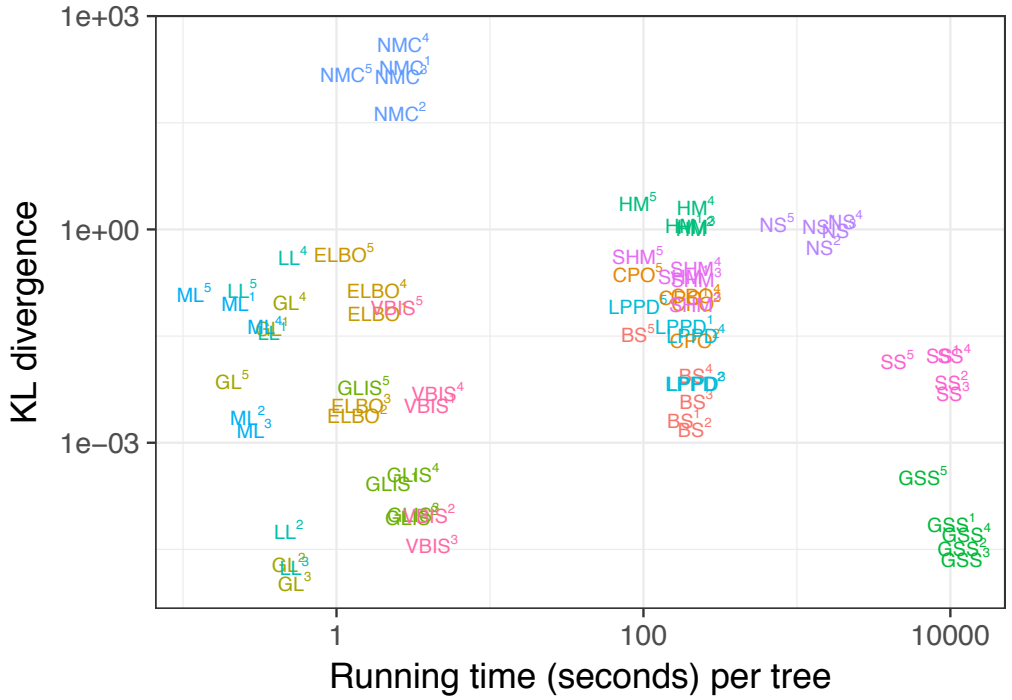


**Figure S10.** Average Kullback-Leibler (KL) divergence from MrBayes posteriors to approximate posteriors of splits in the approximate posterior against running time. Text denotes method used, while superscripts label applications to individual datasets. Four methods are omitted for visual clarity: MAP is essentially identical to ML, BL is nearly identical to GL, and PS and MPS are both similar to SS. The horizontal dashed and solid lines depict RMSDs of 0.01 and 0.05 respectively. The KL divergence is calculated using the average marginal likelihood of each tree from each of 10 replicate analyses. The running time is calculated using the average running time of each tree from each of 10 replicate analyses.

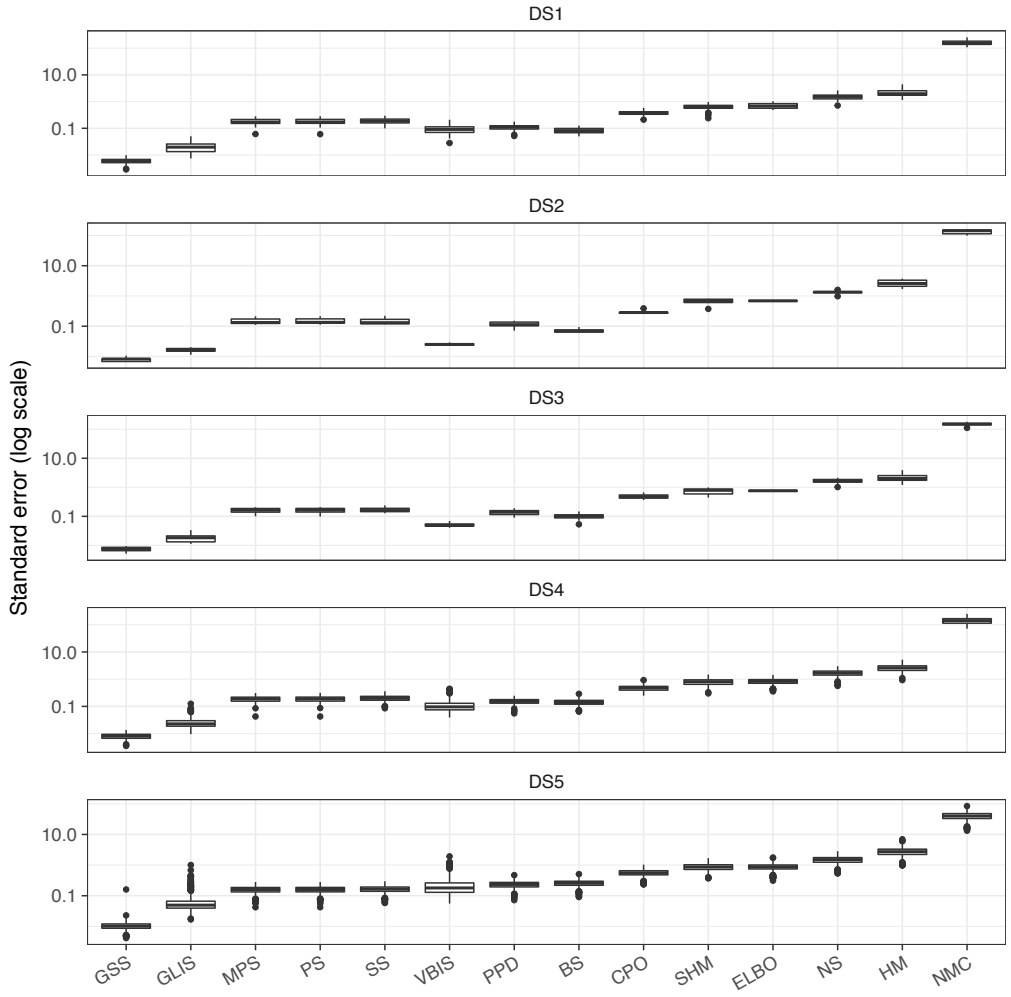


**Figure S11.** Standard error of the Monte-Carlo-based estimators. Each point represents the standard error of an individual tree across the 10 replicate analyses for each estimator.