Syst. Biol. XX(XX):XX–XX, 2024

https://doi.org/10.1093/sysbio/syae055
Advance Access Publication October 7, 2024

Assessing the Adequacy of Morphological Models Using Posterior Predictive Simulations

Laura P. A. Mulvey[1,*], Michael R. May[2], Jeremy M. Brown[3], Sebastian Höhna[4,5], April M. Wright[6], and Rachel C. M. Warnock[1]

[1]*GeoZentrum Nordbayern, Department of Geography and Geosciences, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Loewenichstraße 28, 91054 Erlangen, Germany*
[2]*Department of Evolution and Ecology, University of California Davis, Davis, 2320 Storer Hall, One Shields Avenue Davis, CA 95616, USA*
[3]*Department of Biological Sciences and Museum of Natural Science, Louisiana State University, 202 Life Science Bldg, Baton Rouge, LA 70803, USA*
[4]*GeoBio-Center, Ludwig-Maximilians-Universität München, Richard-Wagner-Str. 10, 80333 München, Germany*
[5]*Department of Earth and Environmental Sciences, Palaeontology & Geobiology, Ludwig-Maximilians-Universität München, Richard-Wagner-Str. 10, 80333 Munich, Germany*
[6]*Department of Biological Sciences, Biology Building, SLU 10736, Southeastern Louisiana University, Hammond, LA 70402, USA*
[*]*Correspondence to be sent to: Department of Geography and Geosciences, GeoZentrum Nordbayern, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Loewenichstraße 28, 91054 Erlangen, Erlangen, Germany; E-mail: lauramulvey479@gmail.com*
Associate Editor: Seraina Klopfstein

Received 22 January 2024; reviews returned 6 September 2024; accepted 4 October 2024

Abstract.—Reconstructing the evolutionary history of different groups of organisms provides insight into how life originated and diversified on Earth. Phylogenetic trees are commonly used to estimate this evolutionary history. Within Bayesian phylogenetics a major step in estimating a tree is in choosing an appropriate model of character evolution. While the most common character data used is molecular sequence data, morphological data remains a vital source of information. The use of morphological characters allows for the incorporation fossil taxa, and despite advances in molecular sequencing, continues to play a significant role in neontology. Moreover, it is the main data source that allows us to unite extinct and extant taxa directly under the same generating process. We therefore require suitable models of morphological character evolution, the most common being the Mk Lewis model. While it is frequently used in both palaeobiology and neontology, it is not known whether the simple Mk substitution model, or any extensions to it, provide a sufficiently good description of the process of morphological evolution. In this study we investigate the impact of different morphological models on empirical tetrapod datasets. Specifically, we compare unpartitioned Mk models with those where characters are partitioned by the number of observed states, both with and without allowing for rate variation across sites and accounting for ascertainment bias. We show that the choice of substitution model has an impact on both topology and branch lengths, highlighting the importance of model choice. Through simulations, we validate the use of the model adequacy approach, posterior predictive simulations, for choosing an appropriate model. Additionally, we compare the performance of model adequacy with Bayesian model selection. We demonstrate how model selection approaches based on marginal likelihoods are not appropriate for choosing between models with partition schemes that vary in character state space (i.e., that vary in Q-matrix state size). Using posterior predictive simulations, we found that current variations of the Mk model are often performing adequately in capturing the evolutionary dynamics that generated our data. We do not find any preference for a particular model extension across multiple datasets, indicating that there is no "one size fits all" when it comes to morphological data and that careful consideration should be given to choosing models of discrete character evolution. By using suitable models of character evolution, we can increase our confidence in our phylogenetic estimates, which should in turn allow us to gain more accurate insights into the evolutionary history of both extinct and extant taxa. [Bayesian phylogenetic analysis; model adequacy; model selection; morphological data; morphological models; palaeobiology.]

The origination and subsequent diversification of species is a fascinating, yet complex, process. Phylogenetic trees serve as a powerful tool to aid in our understanding of this process. They provide a hypothesis of the evolutionary history of a group, enabling us to make inferences about the relationships, timing of events, and patterns of evolution (Baum and Offner 2008). While molecular data may be more commonly used in phylogenetics (Lee and Palci 2015), morphological data was the original source of evidence (Farris et al. 1970) and remains extremely valuable to our interpretation of species diversification (López-Antõnanzas et al. 2022). As the majority of life on Earth is now extinct, the fossil record contains a wealth of knowledge about how species have adapted and diversified through time (Simpson 1952). Integrating this information into phylogenetic analysis, either in combination with morphological and molecular data of extant species, that is, in a total evidence approach (Pyron 2011; Ronquist et al. 2012; Gavryushkina et al. 2017; Mongiardino Koch et al. 2021) or independently, can therefore further our ability to resolve species relationships in deep time. Studies have also shown that incorporating fossil data into an analysis, even when the focus of the study is on extant taxa, can improve the topological resolution or even accuracy of a phylogenetic inference (Beck and Baillie 2018; Koch and Parry 2020; Mongiardino Koch et al. 2021). The use of morphological data in phylogenetics has been a topic of debate for many years, specifically, with regards to which approach should be

applied, that is, parsimony or model-based inference (Kolaczkowski and Thornton 2004; Wright and Hillis 2014; O'Reilly et al. 2016; Puttick et al. 2017; Goloboff et al. 2018, 2019; Sansom et al. 2018). Due to the complex nature of morphological data, there are doubts about our ability to correctly model its evolution, and that any assumptions made by the models will bias the resulting inference (Goloboff et al. 2019). Parsimony is often considered to be an assumption free approach; however, this is not entirely true, as there are still implicit assumptions about morphological evolution within a parsimony framework (Felsenstein 1983; Tuffley and Steel 1997; Steel and Penny 2000; Sober 2004). These 2 approaches have been compared many times throughout the literature, amassing in a large body of work which goes beyond the context of this study. Ultimately, model-based approaches have many more applications and statistical advantages, including the ability to select among competing models and assess model adequacy (Wright and Hillis 2014; O'Reilly et al. 2016; Puttick et al. 2017). Amidst this debate, however, an important question has yet to be addressed: are available models of morphological evolution in fact adequate for our data?

Morphological data collected from fossils or extant taxa, can be either discretized (e.g., presence/absence) or continuous (e.g., body size measurements). Discrete morphological data is the most widely used for phylogenetic inference (Lewis 2001; Wright and Hillis 2014; Harrison and Larsson 2015; Wright 2019) and will be the focus throughout this study. Discrete data, analogous to the format of a molecular alignment, where each site represents a morphological trait, must be manually collected. Traits are described using a character state which is indicative of the phenotype expressed by a given taxon. Traits can have any number of character states depending on the complexity. Presence/absence traits can be described by using only 0 and 1, that is, 2 character states. For complex traits, however, more character states may be required and are then referred to as multistate characters. An example of this could be describing the shape of part of a skull or a shell. In this scenario a state is assigned to a particular modification of the trait, where a number of different variants (or states) may be present in a group. Within a single morphological matrix some traits can have binary character states, while others require multiple states. It is important to emphasize that within a single matrix, a given character state label does not have a fixed biological meaning, that is, a 1 does not represent the same type of character across a matrix. This is markedly different when considering molecular data, where for example an A (adenine) represents the same entity across the matrix. For one given trait a 1 may represent the presence of that trait, for example, ornamentation, whereas for a different trait in the same matrix a 1 may be used to represent the type of ornamentation. See Wright (2019) for a more in-depth review of morphological data used in phylogenetics. The generation of this data is a challenging and time-intensive process, requiring an in-depth knowledge of the taxonomic group in question. Morphological data is, in turn, extremely valuable in helping us answer questions about the evolution of life that molecular data alone cannot answer (López-Antõnanzas et al. 2022).

Within a model-based phylogenetic analysis, the process that gives rise to discrete character data is described using a substitution model, (or morphological models in morphological phylogenetics). These models aim to capture the evolutionary dynamics resulting in the gain, loss or modification of discrete states. Substitution models are continuous-time Markov chain models. They allow states to change (evolve) stochastically at any point in time, and this change depends only on the current state that the evolving system is in. The assumptions of a substitution model are mathematically represented using a Q-matrix. A Q-matrix (also called a rate matrix) is a square matrix where each element represents the instantaneous rate of change between states. That is, Q[i, j] represents the rate of change from state i to state j. The probability of change over a given interval, or branch length v, is calculated using the Q-matrix. Developing models that can accurately describe the complex processes driving morphological evolution is extremely challenging and as a result, there is one main model that is commonly applied: the Mk model (Felsenstein 1992; Lewis 2001). This model is a generalization of the Jukes Cantor model (Jukes and Cantor 1969) used for molecular data, and as such, follows the same set of assumptions. It assumes equal transition rates between states, that is, the rate of transitioning from a state 0 to a 1 is the same as going from a state 0 to a 2. It also assumes equal base (character state) frequencies.

Morphological data are, needless to say, different to molecular. Thus, there are concerns about how well a model originally developed for molecular data can be applied to morphological data. Additionally, given that more complex models are often selected for molecular data, there is doubt about how well such a simple model can be applied to morphological data. As such, there have been a number of extensions implemented for the Mk model to relax these strict assumptions, and allow the model to better describe the reality of morphological evolution. Lewis immediately noted an important difference between morphological and molecular data collection (Lewis 2001). When taxonomists are creating a matrix, (character coding), they will typically exclusively choose traits which differ across species, resulting in a matrix where every site is variable. This is a markedly different behavior from molecular data collection, where there can be many sites where a nucleotide is conserved across all species. Not accounting for this phenomenon, known as ascertainment bias, (though referred to as acquisition bias in Lewis (2001)), can result in inferring trees with extremely long branch lengths. Lewis dealt with this by conditioning the likelihood calculation on there only being variable characters, developing the MkV model. Accounting for among-character rate variation has also been suggested

as important when modeling morphological evolution (Lewis 2001; Harrison and Larsson 2015). This allows different traits to transition at different rates, as some may be evolving faster than others. This is frequently achieved by drawing rates from a discretized gamma distribution and allowing a trait to transition according to a given rate category, the same as is done for molecular data (Yang 1994). It is worth noting that many extensions were suggested by Lewis (2001), however, their implementation was not achieved until later.

When carrying out an inference, it is common to partition data based on the maximum observed character state (Nylander et al. 2004; Khakurel et al. in press). This is the default in a number of phylogenetic software, for example MrBayes (Ronquist et al. 2012) and BEAST2 (Bouckaert et al. 2019), and works by constructing separate Q-matrices for traits with of a given number of observed maximum states. This ensures that traits are in a Q-matrix of the correct size. Here, we investigate the impact of unpartitioned and partitioned inference. In an unpartitioned analysis, the Q-matrix will take the size of the maximum character state in the morphological matrix, which could be for example 5. Transitions between binary characters will therefore also be calculated in this Q-matrix of size 5, meaning that there is some probability given to a binary character of transitioning to states 2, 3, or 4. As these states are not observed in this hypothetical binary trait we may be certain that this is incorrect. Therefore, using such an unpartitioned model would result in substantial model misspecification. Partitioning by character states such that all binary characters are in a Q-matrix of size 2 and so on, avoids this issue. Partitioning data can have an effect on branch lengths (Khakurel et al. in press) so it is important that it is done when necessary. Similarly, however, incorrect partitioning may lead to underestimation of rates as a result of observer bias.

The impact of these different variants of the Mk model is still not fully understood in terms of the effects on key parameter estimates, though some studies have looked at the impacts of using different partitioning schemes (Casali et al. 2023; Khakurel et al. in press). When deciding what model to use, there are 2 distinct questions that can be asked, (1) which is the best model for my data compared to other models? and/or (2) does this model fit my data? The first question, which is the more common of the 2, can be answered using model selection. Model selection approaches are common in molecular based studies although less frequently used for morphological data. For morphological studies there is a history of using substitution models that have been used in previous studies, choosing a model based on the structure of the dataset, or relying on software defaults, often without providing statistical justification for model choice. As previously stated, datasets are manually produced, meaning they can differ from each other depending on the taxonomist. If, for example, a substitution model had been applied to the taxonomic group of interest in the past, even if you are using similar taxa, if the morphological matrix is different, using the same substitution model as previous studies may not be logical. That being said, there are a number of studies where model selection has been applied to morphological datasets (e.g., Bapst et al. 2018; Caldwell et al. 2021; Rücklin et al. 2021; Wright et al. 2021). By using a model selection approach, any subjectivity in model choice can be reduced. One downside of model selection approaches, however, is that they give no indication of the absolute fit of the model to the data. Model selection tells you which model is the relative best, but that does not necessarily mean that the model provides a good description of the true data generating process, simply that it fits better than other models (Gatesy 2007). This is where question 2 becomes important. Asking if a single model is adequate allows you to understand how well a model can describe your data. These approaches, known as model adequacy, have gained popularity for molecular data (Duchêne et al. 2017, 2018; Brown and Thomson 2018) and have been sporadically applied to morphological datasets (Huelsenbeck et al. 2003; Slater and Pennell 2014) but have yet to be systematically assessed.

In order to confidently integrate morphological data from fossils and extant specimens into phylogenetic approaches, it is crucial we ensure that we have appropriate substitution models. Knowing that the models are behaving as expected can increase our confidence in the results and allow us to ask increasingly complex questions. Here we explored the impacts of different substitution models on key parameter estimates across a number of morphological datasets, as well as investigating the best approaches for choosing a model. We found that the models have a notable impact on both tree length and topology, highlighting the importance of validating a model before using it. In our simulation study, model adequacy performed well in predicting which model the data was simulated under. Ultimately, using model adequacy we found that of the 8 empirical datasets we investigate, 5 had at least one substitution model shown to be adequate, supporting the use of the Mk model for morphological data.

METHODS

Data

We used a collection of previously published morphological matrices from Sansom et al. (2018) (taken from http://graemetlloyd.com/matrdino.html). This dataset contained 166 morphological matrices of tetrapod taxa. The datasets vary in size in terms of taxa, from 12 to 219, traits, from 23 to 622, and number of different character states, from 2 to 10. They have also been used previously to examine the use of phylogenetic methods and as such were an ideal dataset for this study (Sansom et al. 2018). We removed matrices based on 2 criteria: (i) those that contained characters with more than 9 states or 80 taxa, as they became too computationally expensive, and (ii) those that

contained traits where only character state "0" and missing characters "?" were present for any trait. This resulted in a final dataset of 114 matrices. The datasets varied in size, with the number of taxa ranging from 12 to 80, and the number of characters being between 23 and 477.

Empirical Comparison of Morphological Models

Initially, our focus was on investigating how substitution models impact the estimation of key parameters. We chose 7 variants of the Mk model (Mk, MkV, MkV + G, Mk + G, MkVP, MkVP + G, MkP + G, see Table 1 for model assumptions) and compared differences in the resulting tree lengths and topologies. All morphological characters were treated as unordered throughout this study. Phylogenetic inference was performed in a Bayesian framework using the software RevBayes version 1.2.1 (Höhna et al. 2016). We ran an MCMC inference under each of the 7 models for all 114 datasets. This allowed us to determine whether there are any systematic differences in parameter estimates that could be attributed to the substitution model. For all models we assumed a uniform tree prior on the topology. Tree length was drawn from an exponential prior distribution with a rate parameter of 1. Relative branch lengths were drawn from a Dirichlet prior distribution (Zhang et al. 2012). The branch lengths were calculated as the product of the tree length and the relative branch lengths. Preliminary analyses were run using an exponential prior for branch length estimation, however, we found the Dirichlet tree prior to perform better in simulations. We used an Mk model, with the size of the Q-matrix being determined by the maximum character state of each dataset. When allowing for among character rate variation, ACRV, (+ G) the shape parameter of the gamma distribution α was estimated as the inverse of a random variable *alpha_inv* drawn from the exponential distribution with a rate parameter of 1. We discretized the gamma distribution into 4 discrete categories (Yang 1994). To account for ascertainment bias (+ V), we selected the variable coding option in RevBayes. Partitioned models (+P) split the dataset based on the maximum observed character states. Each grouping had its own Q matrix. That is, all binary traits were assigned to a Q-matrix of size 2, all tertiary traits were assigned to a Q-matrix of size 3 and so on. For this set up, we linked the gamma distribution for ACRV across partitions.

We ran the MCMC for 20,000 iterations with 2 simultaneous runs, sampling every 10 generations. The output of both chains was automatically combined in RevBayes, resulting in a posterior sample of 4000. Convergence was assessed using a custom R script with the R package coda (Plummer et al. 2006) to ensure ESS values >200 of all parameters estimated.

Posterior summaries.—Tree length was calculated as the sum of the branch lengths averaged across the entire posterior distribution. We calculated the percentage change in tree length relative to the Mk model for each dataset to make it easier to observe any consistent patterns across models. We then explored the differences in estimated tree topologies from the different substitution models for each dataset. Using a sample of 1000 trees from the posterior distribution for each substitution model, we calculated the normalized Robinson–Foulds distance between all trees. With this resulting matrix we performed a multivariate homogeneity of group dispersions analysis using the R package vegan (Oksanen et al. 2022). This calculated the distance between points and their group centroid. Plotting this as a PCoA allowed us to visualize where models were in tree space, relative to one another. In order to quantify these differences we carried out a permutation test to assess their significance using the permutest function in the vegan package (Oksanen et al. 2022). We could then determine if different variants of the Mk model inferred significantly different tree topologies.

Assessing the Performance of Model Adequacy and Model Selection Methods for Morphological Data

Choosing an appropriate model of evolution is an important step in any Bayesian phylogenetic analysis. The results from an inference will be conditioned on the assumptions of the evolutionary model. As such, if the model's assumptions are markedly different than that of the underlying process that generated the data, the results may be inaccurate. Methods for choosing an appropriate model often take a model selection approach, relying on estimation of the marginal likelihood (Brown 2014b). These methods provide the relative fit of competing models. Although a model may be selected as the best choice, it does not necessarily mean that the model is in any way adequate for the dataset being analyzed. That is, it may not provide a sufficiently realistic description of the data generating process (Gatesy 2007; Shepherd and Klaere 2019). Therefore, model selection provides no indication about how well the model actually fits your data, only its relative fit compared to other models. In contrast, model adequacy approaches provide information on the absolute fit of a model to a dataset. They can provide information about a model's ability to capture key characteristics of

TABLE 1. The assumptions of the Mk model and its variants tested in this study

Models and extensions	Assumptions
Mk	All transition rates are equal (Lewis 2001)
V	Accounts for ascertainment bias (Lewis 2001)
G	Allows for variation in substitution rates among characters (Yang 1994)
P	Partitions the data based on the number of character states

a given dataset, as well as highlight where the model may be inadequate. Importantly, model adequacy provides the ability to reject models, even if they are identified as the "best" using a model selection approach (Brown and Thomson 2018; Shepherd and Klaere 2019).

Posterior-predictive simulations (PPS) is a model adequacy approach that has been applied to a variety of data types, albeit with limited frequency in phylogenetics (Gelman et al. 1996; Bollback 2002; Brown 2014a; Brown and Thomson 2018; Höhna et al. 2018; Schwery et al. 2023). Briefly, it works by simulating data under a given model and comparing the similarity of the empirical data to the newly simulated data using a test statistic. The rationale here being that if the model adequately captures the underlying dynamics of the processes generating the data, the simulated data would be similar to the empirical (Gelman et al. 1996; Bollback 2002). To date, the use of PPS has been demonstrated more often for molecular data, for example Brown (2014a) and Duchêne et al. (2018), however, it has also been suggested for models of continuous trait evolution (Slater and Pennell 2014) and discrete character evolution (Huelsenbeck et al. 2003). Using simulations, we investigate the use of model adequacy and model selection for determining whether a morphological model fits our data.

Model adequacy using posterior predictive simulations.—To test the adequacy of morphological models we used posterior prediction simulations (PPS) following the workflow as described in Höhna et al. (2018) implemented in RevBayes. This can be broadly broken down into 4 mains steps. We provide a brief description of these steps here, but for a more thorough description see Höhna et al. (2018). (i) The first step is to analyze the empirical data under a given model. This involves a regular MCMC inference sampling parameter values from the posterior distribution. (ii) New datasets are then simulated in R using the phangorn package (Schliep 2011). Datasets are simulated under the same model as used in step 1 with trees and parameter estimates inferred in step 1. (iii) Inference under the same model is then carried out on all the newly simulated datasets from step 2. (iv) Test statistics are calculated and compared between the original empirical data and inference results, and the newly simulated data and inference results, see Figure 1. The overarching idea here being, the more similar the simulated data is to

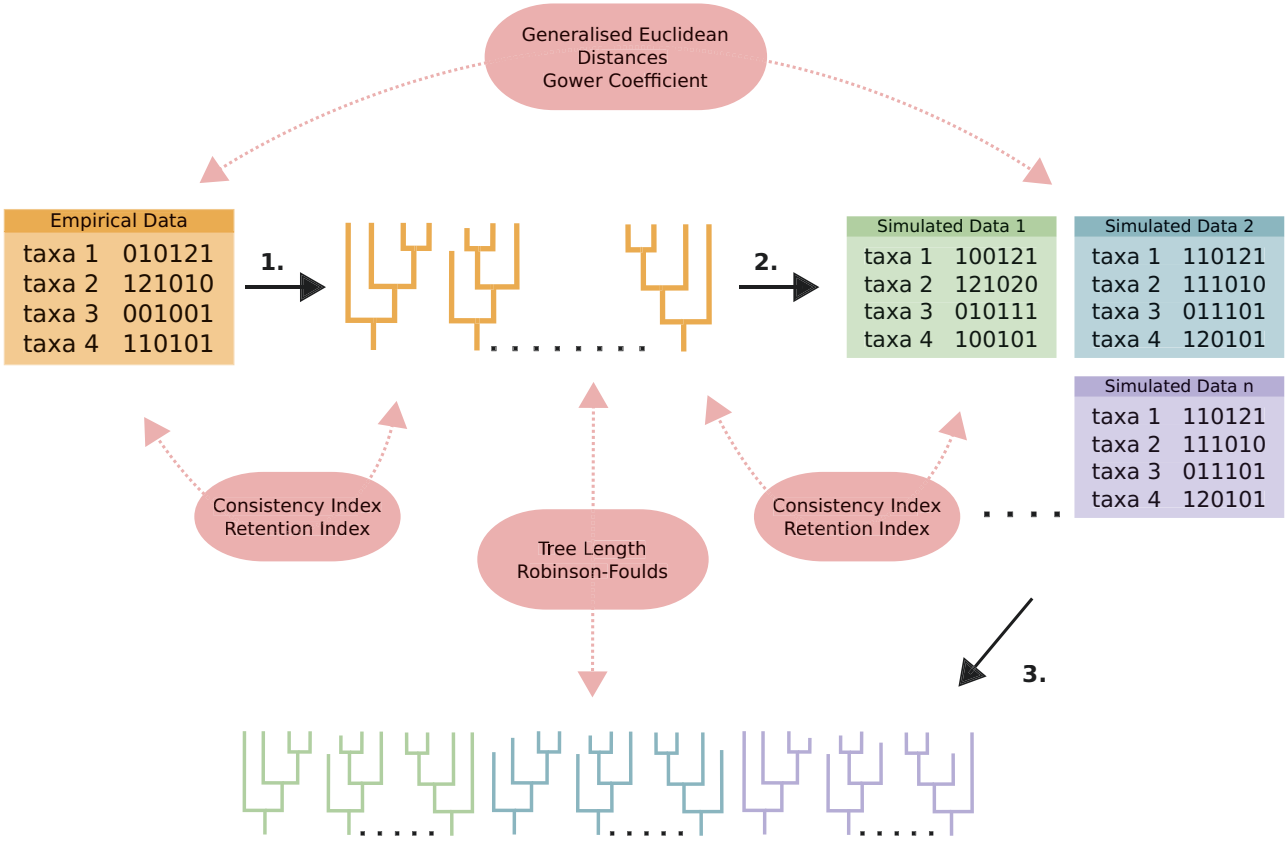

FIGURE 1. Posterior predictive simulation workflow. Step 1. an MCMC inference is carried out under a given model. Step 2. datasets are simulated under the same model based on parameter estimates from 1. Step 3. an MCMC inference is then carried out on the simulated datasets. The boxes show the test statistics that are applied to determine whether or not the model is adequate. Generalized Euclidean distances and Gower's coefficient are used to compare the datasets. Tree length and Robinson–Foulds are used to compare the inferred trees. Consistency index and retention index use the empirical trees and the empirical and simulated datasets to test for adequacy.

the empirical data, the better the model is at describing the underlying processes that produced your data. This in turn indicates whether we can have confidence in the results inferred under a given model. Note it is practical to simulate datasets in RevBayes, and we provide instructions for doing so in the associated tutorial (https://revbayes.github.io/tutorials/pps_morpho/). We chose to simulate data using phangorn as, at the time, it was slightly more computationally efficient given that our study featured an exceptionally large number of simulations (700,000 simulations for 160 individual datasets), but this should not be a concern for an empirical study, which would typically only contain one or a few individual datasets.

Candidate test statistics for morphological data.—PPS are only as good as the test statistics used, meaning if the test statics are not able to capture differences that result from the underlying dynamics of the data generating processes, it will not be possible to use PPS to understand the adequacy of a given model. Using test statistics allows us to convert the empirical data and output into numerical values that we can use to summarize the differences between empirical and simulated data. The test statistics can then be compared using effect sizes, which provide a way of quantifying variation in model fit and allow us to distinguish between the fit of a given model. Previous studies have used posterior-predictive *P*-values to accept or reject a model. In this study we chose to focus on effect sizes over *P*-values for 2 reasons. First, given that the fit of morphological models to empirical data had not been tested previously, we wanted to determine how different models performed and compare their fit to empirical data. Second, effect sizes provide a more intuitive way of comparing the fit of different models. By applying *P*-values only we can assess whether a model is adequate or not, but not how the models perform relative to each other (Brown 2014a; Duchêne et al. 2017). Effect sizes therefore allow us to gain a better understanding of the impact of different morphological models, and ultimately address the main questions of this study. In an empirical study, researchers can choose either approach, and we do include the use of *P*-values for our empirical analysis.

Here, the effect sizes were calculated by:

$$ES = \frac{empTS - simTS}{stdSimTS} \quad (1)$$

where *empTS* is the empirical value for a given test statistic, *simTS* is the value of the test statistic from a single simulated replicate, and *stdSimTS* is the standard deviation across all simulated replicates. The closer this number is to zero, the better the model is at explaining your data. Test statistics can be divided into 3 categories: (i) data based, (ii) inference based, and (iii) data inference hybrid or mixed. Data based test statistics compare the actual morphological datasets themselves, inference based compare the inferred trees and mixed statistics uses both the data and the trees to compare your empirical and simulated values.

Data based test statistics. As the name suggests, these test statistics focus on characterizing the matrices themselves, here meaning the morphological data. As PPS studies in phylogenetics have previously focused on molecular data, many of the data based statistics are only suited to DNA. For example, quantifying the GC content or number of invariant sites (Höhna et al. 2018). Summarizing morphological datasets in a similar way requires different metrics. To do this we explore the use of disparity metrics. Disparity is a measure of the morphological variation observed among species (Hopkins et al. 2017). It is important to note, we are not interested in the actual measure of disparity, we are interested in how the value differs between the original empirical data and the simulated data. We tested 2 metrics of disparity.

(i) *Generalized Euclidean Distances* (GED) (Wills 1998) is a popular disparity metric commonly used in vertebrate research (Brusatte et al. 2011; Lehmann et al. 2019). This measure is similar to the basic Euclidean distances but incorporates adjustments to accommodate missing characters. Lloyd (2016) (modified from Wills (2001)) defines GED as:

$$S_{ij} = \sqrt{\sum_{k=1}^{v} S_{ijk}^2 W_{ijk}} \quad (2)$$

where S_{ij} is the total distance between taxa i and j, v is the total number of characters in the matrix, W_{ijk} is the weight of the kth character, and S_{ijk} is the distance between taxa i and j at the kth character. S_{ijk} equals 0 when the ith and jth sequence match in the kth position and 1 when there is a mismatch. To account for missing data, a mean estimate of disparity is first calculated across all comparisons for which we have observations:

$$\bar{S}_{ijk} = \frac{\sum_{k=1}^{v} S_{ijk} W_{ijk}}{\sum_{k=1}^{v} S_{(ijk)_{max}} W_{ijk}}$$

where $S_{(ijk)max}$ is the maximum possible distance between taxa i and j for the kth character, which equals 1 for discrete characters. The term $\bar{S}_{ijk}.S_{(ijk)max}$ is then substituted into Equation 2 for missing S_{ijk} values. In all cases, we treat characters as equally weighted, that is, $W_{ijk} = 1$.

(ii) *Gower's coefficient* (GC) (Gower 1971) is commonly used in invertebrate studies (Hopkins and Smith 2015). This metric calculates disparity differently to the GED, notably in regards to how it deals with missing characters. Here this is achieved by normalizing by the available data. GC can be written as (Lloyd 2016)

$$S_{ij} = \frac{\sum_{k=1}^{v} S_{ijk} W_{ijk}}{\sum_{k=1}^{v} \delta_{ijk} W_{ijk}} \quad (3)$$

where δ_{ijk} is coded as 1 if both taxa i and j can be coded for k (i.e., character states are observed for both taxa), and zero if not. As above, we use assume equal weights, that is, $W_{ijk} = 1$.

For both the above metrics, we used the R package Claddis (Lloyd 2016). In the calculations we set characters as unordered. The output from this gives a matrix of the pairwise distance between taxa. We took the average disparity across the matrix for the calculation of the effect size, that is, for *empTS* and *simTS*.

Inference based statistics. Inference based test statistics aim to characterize the inferred trees in the posterior distribution.

(i) *Mean Tree Length* (TL) was calculated using all trees sampled in the posterior distribution, defined as (Höhna et al. 2018):

$$\frac{1}{K}\sum_{i=1}^{K} TL_i \tag{4}$$

where K is the total number of posterior samples and TL is defined as the sum of branch lengths $TL = \sum_{2N-3}^{i=1} bl_i$. This calculation was done in RevBayes. We took the mean tree lengths across the posterior distribution of trees as the input for the effect sizes.

(ii) *Mean Robinson-Foulds Distance* (RF) was used to measure the topological uncertainty within the posterior distribution (Robinson and Foulds 1981). The RF distance was calculated for each posterior distribution of trees, such that we had a measure of topological uncertainty for the empirical inference and compared this to the simulated inferences. This can be defined as (Höhna et al. 2018):

$$RF = \frac{2}{K(K-1)}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} RF\left(\Psi_i, \Psi_j\right) \tag{5}$$

where Ψ_i and Ψ_j are any pair of trees from the posterior. This value was calculated in RevBayes.

Mixed test statistics. These test statistics take both the data and the tree into consideration. Again, we investigate the use of 2 test statistics here. (i) *Consistency Index* (CI) (Kluge and Farris 1969) is a measure of homoplasy within the dataset. It can be calculated as (Murphy et al. 2021):

$$CI = \frac{m}{s} \tag{6}$$

where m is the minimum possible number of steps or changes along a tree and s is the reconstructed number, that is, the number observed along estimated trees (Kluge and Farris 1969). This metric has been used to characterise datasets in paleontology (Murphy et al. 2021) and has been applied to model adequacy studies focusing on molecular data (Duchêne et al. 2018). A CI of 1 indicates no homoplasy and gets closer to zero as the amount of homoplasy increases.

(ii) The *retention index* (RI) (Farris 1989), builds on the consistency index to calculate the level of potential synapomorphy observed along the tree and is calculated as (Murphy et al. 2021):

$$RI = \frac{g-s}{g-m} \tag{7}$$

where g is the maximum number of possible steps on a given tree. For both consistency and retention index, we used the maximum clade credibility (MCC) tree generated from inference of the empirical data for all calculations. We carried out preliminary analysis where we used the entire posterior distribution of trees for this calculation. This increased computation time from a number of minutes to 24 hours and produced extremely similar results, see Supplementary Figure S2. For this reason, we continued to use the MCC tree only for the rest of the analyses.

Model selection using stepping stone sampling.—For model selection, Bayes factors are computed to compare between models. In order to do this, we first have to calculate the marginal likelihood of the data. The marginal likelihood is an important quantity in Bayesian model selection as it provides a measure of the goodness of fit of the model to the data, while accounting for model complexity. The marginal probability is the probability of the data integrated over all possible parameter values weighted by their prior probabilities for a given model. This is tricky to calculate and can be extremely computationally expensive. As such we avoid calculating it in regular MCMC inference using the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970). We therefore need to use a different approach in order to approximate this value. One such approach is stepping stone sampling. Stepping stone sampling is a Monte Carlo method that uses a sequence of intermediate distributions, or steps, between the prior and posterior distributions to compute the marginal likelihood. Stepping stone sampling has been demonstrated to be a reliable method for calculating marginal likelihoods and therefore performing model selection with molecular data (Xie et al. 2011; Höhna et al. 2021). Marginal likelihoods has also been used for model selection with morphological data (Rosa et al. 2019; Wright et al. 2021; Casali et al. 2022), though the appropriateness of this approach has never been explored. Further, we wanted to determine if model adequacy and model selection agreed on what model fit a given dataset.

Simulated data.—We based our simulation study on 2 empirical datasets, one on Proboscideans (the group containing elephants and their nearest extinct relatives) (Shoshani et al. 2006) and the other on Hyaenodontidae (Egi et al. 2005). For simplicity we will refer to each dataset as simulated elephants and simulated hyaenodonts, respectively. The simulated elephant dataset is larger, having 40 taxa, 125 characters with 6 states compared to the simulated hyaenodonts which has 15 taxa, 65 characters and 5 states. For each dataset, we used 20 trees from the posterior distribution inferred under a given model and simulated character data under the same model in R using phagnorn (Schliep 2011). We did not simulate any traits with missing data. We did this for the MkV, MkVP, MkV + G and MkVP + G models for each dataset (160 simulated replicates in total).

Analysis of simulated data.—We carried out PPS following Section Model adequacy using posterior predictive simulations on all simulated elephant and simulated hyaenodont datasets. This allowed us to jointly validate the candidate test statistics and determine how well PPS can detect the correct model, as well as how it handles incorrect models. We analyzed each of the simulated datasets under the same 7 models as in Section Empirical Comparison of Morphological Models (Mk, MkV, MkV + G, Mk + G, MkVP, MkVP + G, MkP + G) and kept all model parameters the same. This resulted in 560 inferences per simulation set up (i.e., we simulated 20 datasets under 4 models and then each was analyzed under the 7 models stated above) totaling 1200 inferences across both simulated elephant and simulated hyaenodont. The MCMC was run for 10,000 iterations, with 2 individual chains. Convergence was assessed by calculating the ESS values for the likelihood, prior, posterior, tree length, and when present in the model, the estimated alpha values using the R package coda (Plummer et al. 2006). MCMC chains that produced ESS values <200 were run again with an increase in the chain length. For the simulated hyaenodont datasets, 533 converged after 10,000 iterations, 24 after 50,000 iterations and 3 after 100,000. For the simulated elephant data, 548 reached convergence after 10,000 iterations and 12 required 50,000 iterations.

The number of simulations required for PPS is not strictly defined. Given that the number of simulation replicates will increase both the computation time and memory requirements, having more than required should be avoided. To explore this, we used both the simulated elephant and simulated hyaenodont datasets generated under the MkV + G model. We ran an MCMC inference as described above with 1000 simulation replicates. We calculated the cumulative means for each test statistic inferred under each model. Following Robinson et al. (2004), we plotted the cumulative means, thereby taking a graphical approach that shows the point at which the line becomes flat, indicating the required number of replicates Supplementary Figure S3 (Robinson et al. 2004). We found that after 500 replicates the lines were flat and we determined this to be sufficient. To ensure that this number of simulation replicates was not affecting the calculation of the actual effect sizes, we compared the effect sizes for each test statistic with 500 and 1000 replicates. For ~92% of the effect sizes calculated, we found that the difference was less than 0.1 with a median of ~0.03. The largest change in effect sizes we saw was between 500 and 1000 replicates, ~0.5. This was calculated for the 2 data based test statistics both inferred under the model MkVP + G model and for the same replicate. This result was thus considered an outlier. All other differences were less than 0.25, and did not change whether a model was considered to be adequate or not. As a result of these tests, we determined that within a PPS analysis, simulating 500 datasets is sufficient to determine the fit of a given model. Following this, for all further analyses, at step 2 in the PPS workflow we simulated 500 datasets. We then used stepping stone sampling to estimate the marginal likelihoods under each of the models. We kept all model parameters the same as above, and used 48 stones.

Analysis of Empirical Data

We chose to analyze 8 empirical datasets here; Agnolin (2007); Egi et al. (2005); Bourdon et al. (2009); Shoshani et al. (2006); Archibald et al. (2001); Schoch and Sues (2013); Bloch et al. (2001); Tomiya (2011). This was limited by the computational costs of running the analysis multiple times. Datasets were chosen to cover a range of sizes, in terms of taxa, characters, and states. We tested the same 7 models we used throughout (Mk, MkV, MkV + G, Mk + G, MkVP, MkVP + G, MkP + G) and kept all model parameters the same as in Section *Model adequacy using posterior predictive simulations*. We also used stepping stone sampling on each of the datasets in order to see how the models chosen by model selection compared to those identified as most appropriate by model adequacy. Posterior *P*-values were calculated in R for each of the test statistics to compare with the results obtained using effect sizes.

Results

Empirical Comparison of Morphological Models

Assuming different models of morphological evolution produced different estimates of key parameters of interest. Figure 2A shows the percentage difference in mean tree lengths relative to that of the Mk model for all 114 datasets. There are some general trends that emerged here. As expected, the MkV model produced smaller estimates of tree length relative to the Mk model for all but one dataset. The Mk + G model produced longer trees for 96% of the datasets compared to the Mk model. However, when used in combination, these 2 extensions produced the smallest trees compared to all models in 96% of datasets. Partitioned models estimated larger trees, with the MkP + G model estimating larger trees in 100% of the datasets, consistent with the findings of Khakurel et al. (in press). Interestingly, the MkVP + G model generated both larger and smaller trees compared to the Mk model, with only 35% of the trees being larger. Figure 2B shows the tree length plotted for 2 datasets, of Hyaenodontidae (Egi et al. 2005) and Proboscideans (Shoshani et al. 2006), respectively. This is to highlight, that while there are some general trends, models still behave differently depending on the dataset. It is worth noting that the Shoshani et al. (2006) dataset (Fig. 2B (i)) is the larger of the 2, both in terms of number of taxa and characters. The influence of different models on tree length tended to increase with larger datasets, both in terms of taxa and character number, see Supplementary Figure S1.

Figure 2C shows the tree space for the same 2 datasets. It is clear that the different models are plotting in

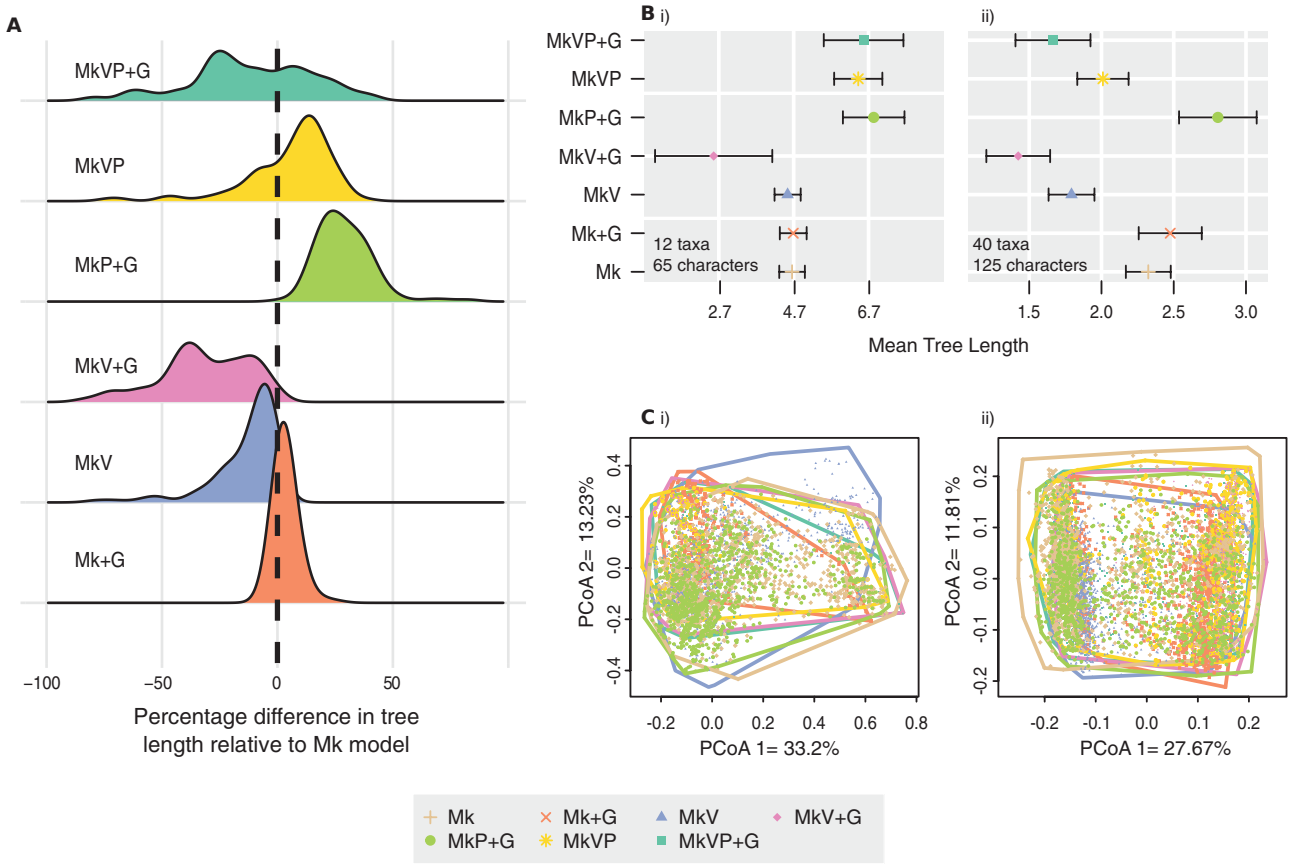

FIGURE 2. Analysis from 114 datasets under the 7 different models Mk, MkV, MkV + G, Mk + G, MkVP, MkVP + G, MkP + G. a) The changes in mean tree length of the posterior inferred using each model relative to the Mk model. b) The tree length calculated for each model for 2 different datasets from Egi et al. (2005) (Hyaenodontidae) and Shoshani et al. (2006) (Proboscideans), respectively. c) The tree space of the same 2 datasets as for B.

different parts of the tree space, therefore producing different posterior distribution of trees. Using the permuted *P*-values estimated from the pairwise distances using Robinson–Foulds, we found that for both datasets the majority of models occupied a different tree space, that is, differences in topology were significant. For the dataset from Egi et al. (2005), trees inferred using MkV, MkV + G and Mk + G models grouped in a similar tree space, whereas all other models occupied different spaces. For the dataset from Shoshani et al. (2006) we found 2 clusters, one consisting of trees inferred using Mk + G and MkV models, and the other of trees from MkV and MkV + G models, though there was no overlap between Mk + G and MkV + G posteriors. These results highlight that, not only do the substitution models have an impact on key parameter estimates but this impact is not uniform across datasets.

Assessing the Performance of Model Adequacy and Model Selection Methods for Morphological Data

Candidate test statistics for morphological data.—We explored the use of 6 test statistics for morphological models. The desired characteristic of test statics considered here is their ability to indicate the adequacy of a particular model while also pointing out the inadequacy of another, that is, we want the effect size of the correct model to be consistently around zero, while being larger for the incorrect models. We will focus on the results from both hyaenodont and elephant datasets simulated under the MkV + G and MkVP + G models. We carried out the same investigation on datasets simulated under the MkV and MkVP models and reached the same conclusions, see Supplementary Figure S6–8. The data based test statistics, Gower's coefficient and generalized Euclidean distance, both show a similar pattern, shown in Figure 3. For the unpartitioned models there is no discernible preference for a given model. That is, they all fall within a similar range of effect sizes. For data simulated under a partitioned model, there was a stronger separation of effect sizes, where all the partitioned models are closer to zero and fall within a similar range. Neither of the inference based test statistics, shown in Figure 4, show any strong or meaningful separation of effect sizes, that is, there is no preference for any of the models and it is unclear what explains this pattern. As for the mixed test statistics, consistency index and retention index, shown in Figure 5, there is a

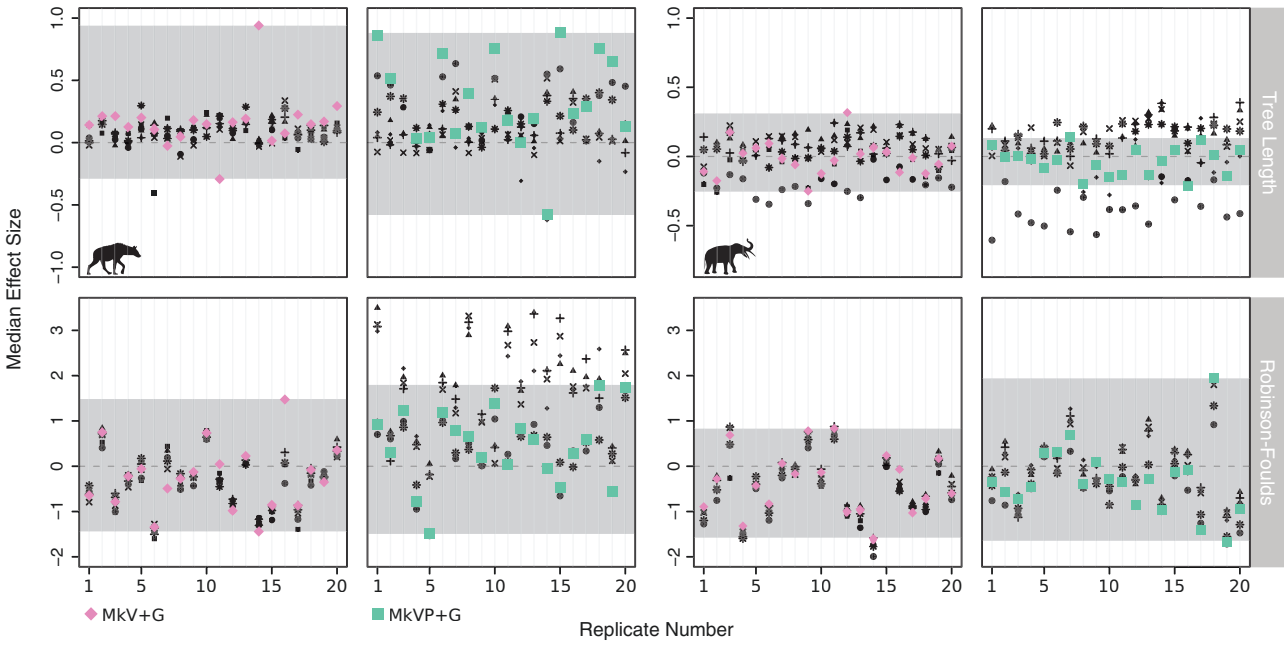

FIGURE 3. Validation of the data based test statistics (Gowers's coefficient and generalised Euclidean distance). Plots show the output from each simulated dataset with 20 replicates for each test statistic. Plots on the left show results of the simulated hyaenodont datasets, with 15 taxa and 65 characters, and on the right from the simulated elephant datasets with 40 taxa and 125 characters. The colored points indicate the correct model, with the gray horizontal bar marking the range of effect sizes calculated for the correct model. + = Mk, × = Mk + G, ▲ = MkV, ◆ = MkV + G, *= MkVP, ●= MkP + G, and ■ = MkVP + G.

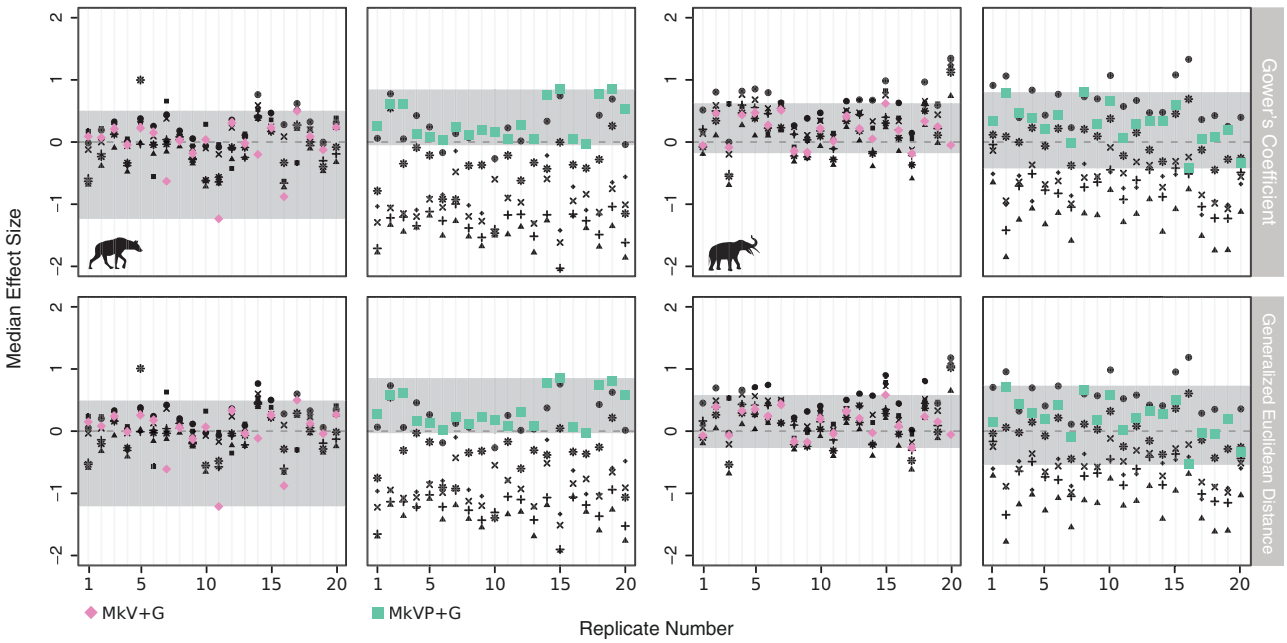

FIGURE 4. Validation of the inference based test statistics (tree length and Robinson Foulds). Plots show the output from each simulated dataset with 20 replicates for each test statistic. Plots on the left show results of the simulated hyaenodont datasets, with 15 taxa and 65 characters, and on the right from the simulated elephant datasets with 40 taxa and 125 characters. The colored points indicate the correct model, with the grey horizontal bar marking the range of effect sizes calculated for the correct model. + = Mk, × = Mk + G, ▲ = MkV, ◆ = MkV + G, *= MkVP, ●= MkP + G, and ■ = MkVP + G.

similar pattern to that of the data based test statistics, however, with the differences in effect sizes between models being more pronounced.

In order to quantify these results, we focused on 3 key features: (i) the variance in effect sizes for the correct model, meaning the total range of effect sizes for

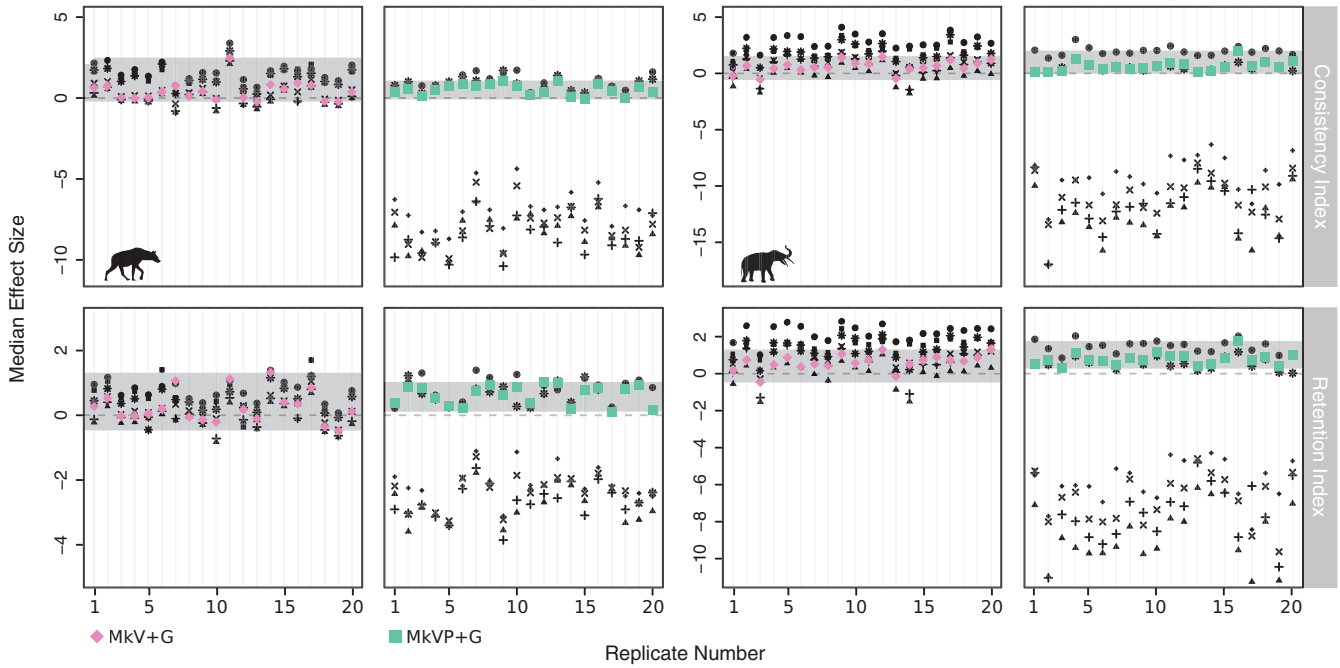

FIGURE 5. Validation of the mixed test statistics (consistency index and retention index). Plots show the output from each simulated dataset with 20 replicates for each test statistic. Plots on the left show results of the simulated hyaenodont datasets, with 15 taxa and 65 characters, and on the right from the simulated elephant datasets with 40 taxa and 125 characters. The colored points indicate the correct model, with the grey horizontal bar marking the range of effect sizes calculated for the correct model. + = Mk, × = Mk + G, ▲ = MkV, ◆ = MkV + G, *= MkVP, ●= MkP + G, and ■ = MkVP + G.

TABLE 2. Validation of test statistics from the simulated hyaenodont datasets

Model	Test statistic	Correct ES	Overall ES	Num in Correct
MkV + G	GC	1.7	2.2	5.7
	GED	1.7	2.2	5.6
	TL	1.2	1.4	6
	RF	2.9	3.1	5.9
	CI	2.7	4.3	5
	RI	1.8	2.5	5.6
MkVP + G	GC	0.9	2.9	1
	GED	0.9	2.8	1
	TL	2.8	2.8	6
	RF	3.3	5.0	4.1
	CI	1.2	11.7	1.40
	RI	1.1	5.3	1.6

Correct ES gives the total range of effect sizes for a given test statistic with the correct model. Overall ES gives the total range of effect sizes for a given test statistic across all models. Num in Correct gives the number of models which fall into the Correct ES range. Num in Correct only looks at incorrect models, which means the maximum value here can be 6. GC = Gower's coefficient, GED = generlized Euclidean distance, TL = tree length, RF = Robinson Foulds, CI = consistency index, and RI = retention index. Consistency index and retention index have the largest overall ES range with, on average, the fewest models falling in the same range as that of the correct model.

a given test statistic with the correct model, (ii) how incorrect models performed, meaning the total range of effect sizes for a given test statistic across all models, and (iii) how easily we could differentiate between adequate and inadequate models by calculating the number of models that fall into the correct model effect size (ES) range. A numerical summary of these results can be found in Tables 2 and 3. Consistency index and retention index demonstrated the best performance of these 3 aspects, with the correct models being consistently close to zero, incorrect models having larger ES values, and the fewest number of models on average falling within the correct model effect size range. While the data based test statistics seem promising, the difference in effect sizes were less than that of the mixed test statistics. As such, in the empirical analyses we relied solely on the mixed test statistics, the consistency and retention indices. An added advantage of using only the mixed test statistics is that we do not need to carry out an inference on the simulated data.

Table 3. Validation of test statistics from the simulated elephant datasets

Model	Test statistic	Correct ES	Overall ES	Num in Correct
MkV + G	GC	0.8	2.0	4.1
	GED	0.9	1.9	4.7
	TL	0.7	0.7	5.7
	RF	2.4	2.9	5.5
	CI	2.0	5.9	2.8
	RI	1.8	4.3	3
MkVP + G	GC	1.2	3.2	2.1
	GED	1.2	4.0	2.95
	TL	0.3	1.0	2.95
	RF	3.7	3.7	5.95
	CI	1.9	20.0	1.45
	RI	1.5	13	1.6

Correct ES gives the total range of effect sizes for a given test statistic with the correct model. Overall ES gives the total range of effect sizes for a given test statistic across all models. Num in Correct gives the number of models which fall into the Correct ES range. Num in Correct only looks at incorrect models, which means the maximum value here can be 6. GC = Gower's coefficient, GED = generalized Euclidean distance, TL = tree length, RF = Robinson Foulds, CI = consistency index, and RI = retention index. Consistency index and retention index have the largest overall ES range with, on average, the fewest models falling in the same range as that of the correct model.

We used the MCC tree from the empirical inference, therefore saving on computational time and memory requirements.

Model adequacy versus model selection.—Here we compared the use of model adequacy and model selection using simulated datasets. To reiterate, unlike model selection, model adequacy approaches do not rank potential models in the same way, indicating that one model is the best. Therefore, for any given dataset, if multiple models are investigated, as was the case here, several models may be adequate according to a particular test statistic. We will focus on the same 4 datasets as in Section *Candidate test statistics for morphological data*.

In the above section, to identify appropriate test statistics, we focused on the pattern of median ES values. When considering individual replicates we required more information than just the median ES value to determine the adequacy of a model for a given dataset. Using this value alone makes it difficult to determine a model's adequacy unless the median value is zero. We explored the use of upper and lower quartiles, and minimum and maximum limits and found the latter to be the more informative approach for identifying a model's adequacy. We propose that if the minimum and maximum limits pass through zero, this indicates that the model is adequate using our chosen test statistics, as shown in Figure 6. Following this criteria, we could quantify the percentage of simulation replicates where the model was deemed adequate/inadequate. Table 4 shows the percentage of times a model met the above criteria using the consistency index and the retention index.

Model selection produced surprising results. We consistently found support for partitioned models, regardless of the model used to simulate the data. Table 5 shows the percentage of times a model was chosen as the best model according to Bayes factors. For this reason, using Bayes factors is not a reliable approach for deciding between partitions with morphological data, at least not using the standard approach we applied to partition characters, i.e., by the maximum observed state number (see the Discussion for a full explanation).

Analysis of Empirical data

We then applied PPS with the newly validated test statistics to 8 empirical datasets. This allowed us to answer our main question: are current morphological models adequate for empirical data? Of the 8 datasets, 5 had at least one model that was adequate. Figure 6 shows the effect sizes from 4 datasets (see also supplementary Fig. S9). The MkVP + G model was found to be adequate for all 5 datasets. Of those 5 datasets, 4 also fit an MkVP model. We found the MkP + G model to be adequate for 3 datasets. For one of the datasets, Figure 6C, we found all models apart from the MkP + G model to be adequate. We do not see any clear pattern in terms of adequate models, with respect to the size of the datasets, that is, number of taxa, characters or state number. This suggests that these variables are not informative when choosing a model. For the 2 largest datasets, in terms of taxa, we did not find any models to be adequate. These datasets had 40 taxa (Shoshani et al. 2006) and 50 taxa (Tomiya 2011). However, no models were adequate for a third dataset with only 25 taxa (Schoch and Sues 2013). Table 6 shows the *P*-values calculated for consistency index and retention index for the same datasets as in Figure 6. See Supplementary Table S1 for *P*-values calculated for an additional 4 datasets. Values below 0.025 and above 0.975 are considered to be significant, although these thresholds can be considered as conservative (see Fabreti et al. 2024). This would indicate that the simulated data is significantly different from the empirical data, and that the model does not capture the underlying data generating processes and therefore is not adequate for that dataset. Results using effect sizes and *P*-values agree on the same models for all datasets. There is

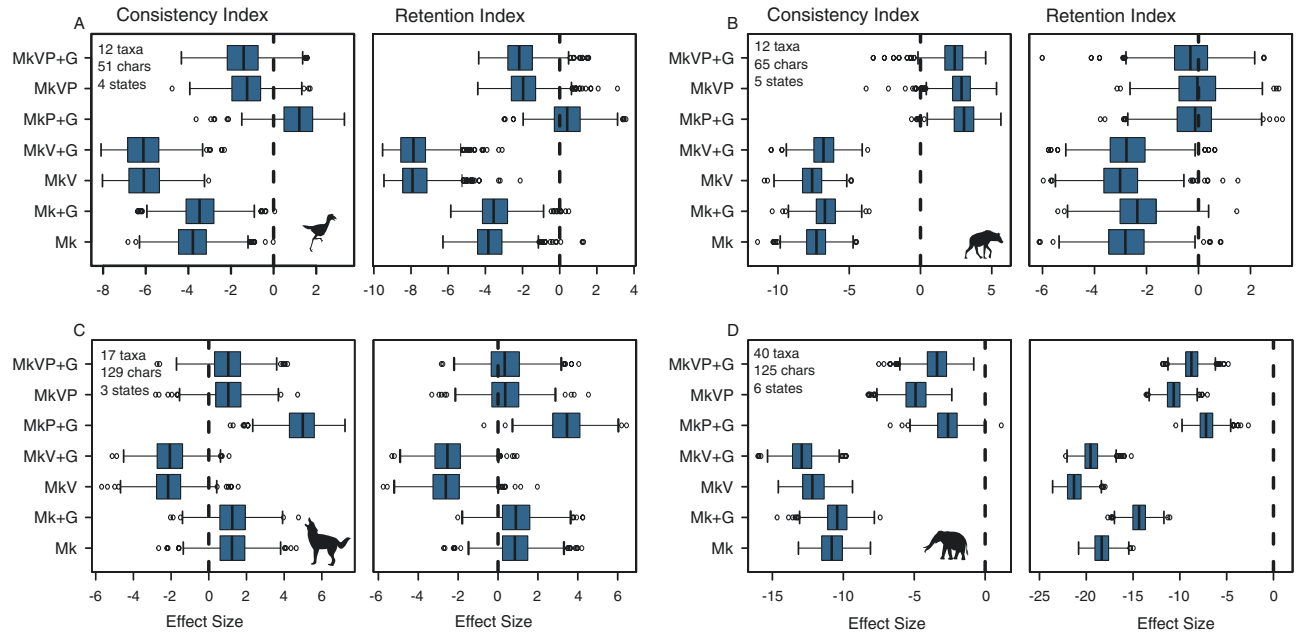

FIGURE 6. Effect sizes for 4 empirical datasets for the consistency index and retention index. The dashed black line is at zero is there to help identify adequate models. The datasets are taken from a) Agnolin (2007), b) Egi et al. (2005), c) Bourdon et al. (2009), and d) Shoshani et al. (2006). Silhouettes from PhyloPic (http://phylopic.org). A by Zimices (CC BY-NC 3.0), B by Margot Michaudand (CC0 1.0), and C by Gabriela Palomo-Munoz (CC BY-NC 3.0).

TABLE 4. The percentage of times a model was found to be adequate across all simulated replicates using consistency index (CI) and retention index (RI) as tests statistics

Sim model	Dataset	Test Statistic	Mk	Mk + G	MkV	MkV + G	MkP + G	MVP	MkVP + G
MkV + G	Hyaenodont	CI	100%	95%	100%	100%	95%	95%	95%
MkV + G	Hyaenodont	RI	100%	100%	100%	100%	100%	100%	100%
MkVP + G	Hyaenodont	CI	–	–	–	–	100%	100%	100%
MkVP + G	Hyaenodont	RI	50%	65%	45%	75%	100%	100%	100%
MkV + G	Elephant	CI	100%	100%	100%	100%	40%	85%	80%
MkV + G	Elephant	RI	100%	100%	100%	100%	70%	100%	100%
MkVP + G	Elephant	CI	–	–	–	–	100%	100%	100%
MkVP + G	Elephant	RI	–	–	–	–	100%	100%	100%

In order for a model to be considered adequate the effect sizes need to meet the criteria put forward here, where the range of minimum and maximum values contain zero. The dashed lines indicate 0%.

TABLE 5:. Models chosen using Bayes factors

Model	Dataset	Mk	Mk + G	MkV	MkV + G	MkP + G	MVP	MkVP + G
MkV + G	Hyaenodont	–	–	-	–	5%	15%	80%
MkVP + G	Hyaenodont	–	–	-	–	5%	30%	65%
MkV + G	Elephant	–	–	-	–	–	–	100%
MkVP + G	Elephant	–	–	-	–	–	–	100%

Cells show the percentage of times a model was selected across the 20 replicates from each simulation set up. The dashed line indicates the model was never selected.

one instance when there is a disagreement using retention index. For the dataset from Egi et al. (2005), the Mk + G model was accepted using the threshold that we defined for effect sizes and rejected using *P*-values. Both metrics rejected the model according to consistency index, however, so the Mk + G was ultimately rejected using both approaches.

DISCUSSION

Understanding morphological evolution is an extremely difficult task. Within morphological phylogenetics we rely on a small number of relatively simple models to describe this complex process (Wright 2019). Until now, the impact of these different substitution models on parameter estimates was not well understood. Our analysis on the

Table 6. Posterior *P*-values from the empirical analyses

Model	Agnolin		Egi		Bourdon		Shoshani	
	CI	RI	CI	RI	CI	RI	CI	RI
Mk	0	0.003	0	0.005	0.8895	0.812	0	0
Mk + G	0.001	0.004	0	0.006	0.898	0.8235	0	0
MkV	0	0	0	0.005	0.019	0.011	0	0
MkV + G	0	0	0	0.006	0.033	0.01	0	0
MkP + G	0.835	0.659	0.994	0.446	1	0.999	0.001	0
MkVP	0.105	0.041	0.992	0.483	0.859	0.655	0	0
MkVP + G	0.095	0.034	0.974	0.376	0.848	0.6245	0	0

CI refers to consistency index and RI to retention index. Values below 0.025 and above 0.975 are considered to be significant. This would indicate that the simulated data is significantly different than the empirical data and that the model is not adequate for that dataset. The results here agree with those produced using effect sizes. Agnolin (2007): 12 taxa with 51 characters, Egi et al. (2005): 12 taxa with 65 characters, Bourdon et al. (2009): 17 taxa with 129 characters, Shoshani et al. (2006): 40 taxa with 125 characters.

influence of these models using empirical datasets, focusing on tree length and topology, demonstrates that different models can produce contrasting reconstructions of the evolutionary history of a group, emphasizing the importance of model choice (Fig. 2). Although the impact of models on parameter estimates is not uniform across datasets, the most consistent pattern we observe is whether or not the data is partitioned by the number of states. Further, we found using model adequacy, that partitioned models are often a good fit to empirical datasets (for 5 out of 8 tested here), and that there can be more than one model adequate for a given dataset.

Partitioned Models

In all the partitioned models explored here, traits were partitioned based on the number of observed character states. This is a reasonable approach, both in terms of the biology of the traits being described and the way in which the characters tend to be coded. We found that for all but 2 datasets, the unpartitioned models produced smaller trees. To further investigate the cause of this, we ran an analysis using a binary dataset and increased the Q-matrix size from 2 to 5. The objective here was to mirror what happens when we have characters with a lower number of observed states than the maximum number of states in the matrix. For example, placing binary characters in a partition with a maximum of 5 character states. We show that, as the size of the Q-matrix increases, tree length gets smaller (Supplementary Fig. S10). The effect of partitioning that we observe on empirical estimates of tree length, is therefore a direct result of how morphological data is typically partitioned (see also Equations 8 and 9). Characters are partitioned by the maximum number of observed states, for example, binary characters are all together in one partition and assigned to a rate matrix of size 2, characters with 3 states are assigned to a rate matrix of size 3 and so on. For unpartitioned models, however, all of the characters will be in a single Q-matrix that is the size of the maximum number of observed states across the whole dataset. This means that for a given branch length v, under a model that assumes there are n states, for characters where we observe $< n$ states (e.g., a binary character in a rate matrix of size 5), the probability of observing no change will be underestimated. Similarly, the probability of observing a given change will also be lower if there are more (unobserved) possible states. Both cases will result in shorter branch lengths. Partitioning morphological data by character state number is a practical approach. However, this requires making an assumption that we know the number of states for each character, when in reality we might not. For molecular data of course, this is not something we need to consider, as we know there are 4 nucleotides. By assuming we know the number of states, based on the number of observed states, we may be biasing our results. The effects of whether or not a dataset is partitioned are considerable in terms of parameter estimates. As such, it is important to consider how the data is being partitioned and whether or not it makes biological sense for your dataset to do so.

Here we focused exclusively on partitioning by the number of character states. This is the most common partitioning scheme and is even a default in some phylogenetic software programs, for example BEAST2 (Bouckaert et al. 2019) and MrBayes (Ronquist et al. 2012). Yet this is not the only way that data could be partitioned. A researcher could partition the data based on different anatomical regions, or based on subsets of anatomical, ecological or behavioral traits (Klopfstein et al. 2015; Casali et al. 2023). Thus, one may need to decide between various partitioning schemes or no partitioning at all. To date, model selection is regarded as the gold standard for choosing between substitution models and partition schemes (Xie et al. 2011). Within a Bayesian framework, comparing marginal likelihoods has been shown to be effective for choosing between partition schemes with molecular data. Our results, however, show that for morphological data, model selection consistently selects a partitioned model, regardless of the model used to simulate the data. This result can be explained by taking into account how partitioning morphological data affects the likelihood calculation, importantly how it affects the transition probabilities and the stationary frequencies.

For example, assume you have a tree consisting of 2 tips, one with discrete state 0 and the other with discrete state 1, as shown here.

$$0 \text{ ——————— } 1$$
$$\leftarrow 2v \rightarrow$$

The tips share a common ancestor v time units in the past. The transition probability for this scenario under the Mk model is calculated as:

$$p_{01}(2v) = \frac{1}{k} - \frac{1}{k}e^{-2v} \tag{8}$$

where k is the number of states. Further, the likelihood of this data is:

$$P(0,1|v) = \frac{1}{k}\, x\, \frac{1}{k}\, \left[1 - e^{-2v}\right] \tag{9}$$

Here k would be set to 2 as we observe 2 states. However, in cases where there are other traits, some of which have a higher maximum observed state, k would increase, as happens in unpartitioned inference. Higher values of k would result in a lower likelihood. This change in likelihood is a direct result of the partitioning scheme. When partitioning molecular data, we do not change the size of the Q-matrix (k), which is why we do not see the same effects on the likelihood. Figure 7 shows the impact on the log likelihood of changing the size of the Q-matrix (k) along different branch lengths (v) for these 2 tips.

To empirically demonstrate the impact of partitioning by state space on the likelihood we ran 2 experiments. First, using an empirical binary morphological matrix we calculated the marginal likelihood under an unpartitioned MkV + G model increasing the Q-matrix size from 2 to 5. Supplementary Figure S12 shows the decrease in marginal likelihood as we increase the number of transition possibilities (Q-matrix size). We then wanted to investigate the impact of adding the "correct" partitions. Here, we used an empirical morphological matrix with a maximum of 6 states. We first calculated the marginal likelihood under an unpartitioned MkV + G model. We then created 2 partitions, one partition for all binary traits and the second for all other traits. Then we increased the number of partitions to 3, with 1 for binary traits, 1 for tertiary traits, and kept all others in the third partition. This method of adding partitions was continued until there were 5 in total and all traits were in the appropriately sized Q-matrix. Supplementary Figure S11 shows that the marginal likelihood increases as partitions are added to the model. This is expected, given Equations 8 and 9. This suggests that the results from model selection will not be indicative of any meaningful biological signal in this context. For this reason, using model selection to differentiate between partitions for morphological data is not appropriate when the Q-matrix size varies.

Test Statistics

Overall, our results show that model adequacy, in particular PPS, currently offers the most effective way of identifying the most suitable model for morphological data. In addition, we demonstrate that PPS can reliably determine whether a given model is adequate or not. Understanding the absolute fit of available models can lend support to the use of model based phylogenetics for the analysis of morphological data. Here we carried out the first thorough investigation into the use of PPS with discrete morphological substitution models.

One of the most important aspects of PPS to consider is the choice of test statistics. As this was the first systematic application of PPS to discrete character data, we first validated available test statistics using simulations. We explored the use of 6 test statistics and ultimately found consistency index and retention index to be the most informative. Neither of the inference based test statistics we explored, Robinson–Foulds or tree length, were able to give a clear indication of model adequacy.

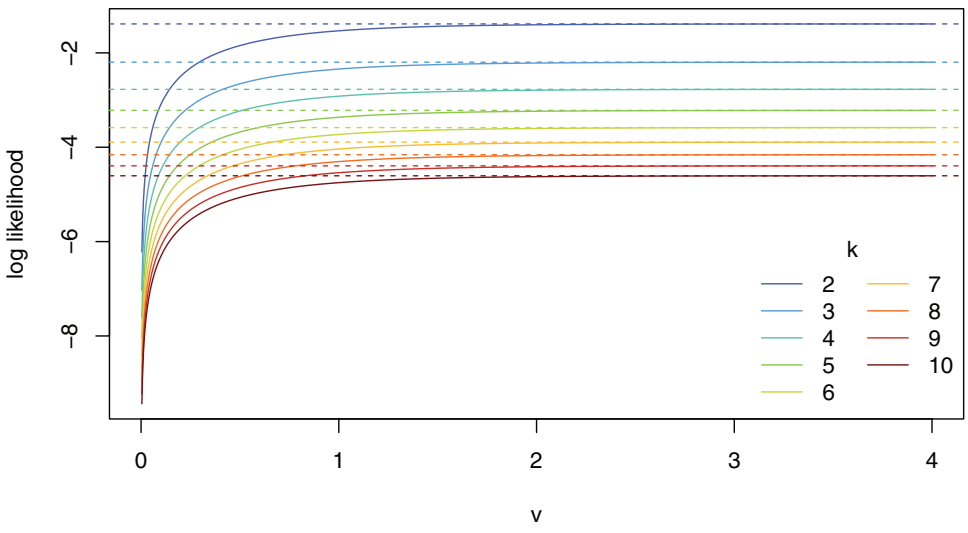

FIGURE 7. Log likelihoods calculated for different sizes Q-matrices (k) along as a function of branch lengths (v). The log likelihoods level off as v increases and the transition probability approaches the stationary frequencies.

In this context, Robinson–Foulds distance is used to quantify variance across the posterior distribution of trees, therefore reflecting topological uncertainty. Given that morphological datasets tend to be small, the uncertainty in topology can be high, regardless of the model used for inference (Barido-Sottani et al. 2020). The uninformativeness of tree length is more puzzling, since competing models have a clear impact on the estimated tree length. Tree length has also previously been shown to be a poor test statistic for molecular data (Duchêne et al. 2018). Both Gower's coefficient and generalized Euclidean distance did show some potential value as test statistics (Fig. 3), although the mixed test statistics, the consistency index and retention indices, were substantially better (Fig. 5), Tables 2–3). Having test statistics exclusively for the data would be favorable. Future studies could focus on alternative ways of including disparity metrics as test statistics. For example, we used the mean pairwise distance of disparity across the matrix, perhaps looking at the sum of the variance or sum of the ranges could be more informative for model adequacy (Smith et al. 2023).

Practical Considerations and Outlook

Our simulation study allowed us to identify ways of reducing the overall computational costs. As with many Bayesian analyses, there can be high computational costs associated with running a PPS analysis. To mitigate any unnecessary computation, we assessed the maximum number of simulation replicates required to reach stability in the mean effect sizes. By doing so, we were able to ensure that we were not running unnecessary replicates. Further, the most expensive part of running a PPS analysis comes from the inference of the simulation replicates. Based on our simulation study, we did not find any benefit to including inference based test statistics (tree length and Robinson–Foulds, Fig. 4), meaning this expensive step can be skipped. Taking both of these findings into account, the time, and memory required to run a PPS analysis becomes a lot smaller. For example, when compared to a stepping stone analysis, we found PPS to take half the time per model.

From our simulation study, relying exclusively on the mixed test statistics, consistency index, and retention index, we found that for all replicates, more than one model was adequate (Table 4). When interpreting these results it is important to remember simulated data is often "neater" than empirical data. In our simulation set up, all characters in a given matrix were simulated under the same model and the model extensions we used are not proposing conflicting statements about the underlying process. As such, it is not surprising that we found multiple models to be adequate for our simulated data. The choice of substitution model may have less impact on our simulated data, as the topology is easier to infer. For example, taking all simulation replicates of the simulated hyaenodont data under an MkV + G model, the mean variance in tree length across the 7 different models was 0.74. In contrast, for the empirical data used as the basis for the simulations, the variance in tree length across models was 4.29 (Fig. 2B(i)). Our simulation study was valuable in determining which test statistics were sensitive to model choice under exemplar conditions, but it is not alarming that differentiating between similar models, that is, all partitioned models, was not possible. Future work could investigate model adequacy when data is simulated under more complex models, for example, generating matrices that contain conflicting characters associated with different models or topologies (Sansom et al. 2017; Weisbecker et al. 2023).

The results from our empirical datasets show a larger difference in the effect sizes for different models (Fig. 6). Based on our criteria of using the minimum and maximum effect sizes (after removing outliers) we determined that for 5 of the datasets, at least one of the models tested here was adequate. This leaves the other 3 without an adequate model. While initially this result may seem negative, in that no models were adequate, it is actually more reasonable than not. The expectation that all datasets would have a model available that fits would have been unrealistic, given the complexity of the data versus the simplicity of the models. Having a method that allows the researcher to detect the limits of available models is much more useful than picking the best out of a group of models without considering whether any of them fit. This result highlights the benefit of using such an approach. In the situation where no models are considered adequate for a dataset, it would be up to the researcher to determine how to proceed. For instance, if the effect sizes are not markedly far from zero one may still opt to use a model—however, appreciating its limitations would be important before drawing any conclusions based on the inference results. It is also encouraging to see that the most complex model, the MkVP + G model, was identified as adequate for all 5 of the datasets for which we found an adequate model, indicating that we are moving in the right direction in terms of our assumptions about the data generating processes. This strongly supports the above discussed rationale of partitioning the data based on character state, lending confidence to our biological interpretation of the evolution of the data.

Here we have demonstrated how PPS outperforms a model selection approach in several respects. Making this a standard approach in morphological phylogenetics would be beneficial to the field in allowing for a better appreciation of how well our models are performing. In this study we explored the use of 7 extensions of the Mk model, as they are the most commonly applied. This is not an exhaustive list of available models and there are a number of alternatives that further relax assumptions of the Mk model.

These models aim to better capture the underlying biological processes that generated the data. For example, Nylander et al. (2004) introduced an approach to relax the assumption of symmetric probabilities of change between characters through the use of priors. Subsequent exploration by Wright et al. (2016) showed

how this can improve model fit and phylogenetic estimation. Similarly, Klopfstein et al. (2015) explored the use of accounting for directional evolution by allowing character state frequency to vary. It is also possible to incorporate ordered characters into a model. In this scenario the model will only allow transitions in pre-defined orders, that is, traits can go from 0 to 1 but not 0 to 2 (Slowinski 1993; Brocklehurst and Haridy 2021). A lot of work has been carried out exploring appropriate partitioning schemes for morphological data. Partitioning based on biological properties, such as anatomical region, function, or using evolutionary rates has also been suggested (Clarke and Middleton 2008; Close et al. 2015; Simões et al. 2020; Casali et al. 2022, 2023). For feasibility we focused our investigation on partitioning based on maximum observed character state. Additionally, models employed in biogeographic probabilistic analyses may have potential applications for discrete traits (Sanmartín et al. 2008, 2010; Lemey et al. 2009). These models can allow for independent stationary frequencies and independent pairwise transition rates which further relax the assumptions of the Mk model. Alternative models and partition schemes mentioned above can all be assessed using the workflow presented here, the only requirements being that the model can be used for both simulation and inference.

There are also a number of models of continuous character evolution that are often used in phylogenetic comparative methods (Alvarez-Carretero et al. 2022; Hansen et al. 2022), which previously were explored using model adequacy (Slater and Pennell 2014). We focused exclusively on discrete data as it remains the most widely used for tree inference. Finally, our results have implications for studies focused on divergence time estimation and ancestral state reconstruction which rely on discrete traits for inference. The same model validation can be applied before either of these types of analyses are carried out. Ultimately, fossils are our only direct source of information about extinct taxa. Collection and character coding of extinct and extant taxa for phylogenetic analysis requires huge effort, both in terms of time and knowledge required. Ensuring that we are using the best available models can help provide confidence in our results and support us in asking more complex questions with the data.

Conclusions

As the use of morphological data in Bayesian phylogenetic analysis increases in popularity, it is important that we understand the adequacy of models available for describing morphological evolution. Here we show that substitution model choice impacts estimates of both branch lengths and topology. By providing a workflow for posterior predictive simulations to validate the adequacy of a model, researchers can gain insights into the absolute rather than the relative model fit, and can have more confidence in their choice of substitution model going forward. We show that, despite the arguably simplistic assumptions of available morphological models, they are often able to approximate the underlying generating processes of discrete morphological datasets. However, we also show that no single model fit all datasets examined here, so we recommend researchers use model adequacy to assess model fit as a first step in phylogenetic inference. Given the substantial taxonomic effort invested into collecting such datasets, the importance of utilizing accurate models cannot be overstated. Our work reinforces the significance of these considerations, particularly as fossil data remains the primary avenue for gaining a comprehensive understanding of evolutionary history in deep time.

Supplementary Material

Data available from the Dryad Digital Repository: https://dx.doi.org/10.5061/dryad.4f4qrfjkq.

Acknowledgements

AMW was supported on NSF DEB-2045842. SH was supported by the Deutsche Forschungsge meinschaft (DFG) Emmy Noether-Program (Award HO 6201/1-1 to S.H.) and by the European Union (ERC, MacDrive, GA 101043187). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. MRM was supported on NSF DEB-1754705. The majority of this research was conducted with high-performance computational resources provided by the Erlangen National High-Performance Computing Center (NHR@FAU). We thank the editors, and Mario Coiro, Mike Lee, and one anonymous reviewer for their helpful comments on the manuscript.

Data Availability

All data sets and code used for this study and are available from the Dryad Digital Repository (https://dx.doi.org/10.5061/dryad.4f4qrfjkq) and GitHub (https://github.com/laumul/PPS_Morphology). The associated RevBayes tutorial is available here (https://revbayes.github.io/tutorials/pps_morpho/).

References

Agnolin F. 2007. Brontornis burmeisteri moreno & mercerat, un anseriformes (aves) gigante del mioceno medio de patagonia, argentina. Rev. Mus. Argent. Cienc. Nat. Nueva Ser. 9:15–25.

Alvarez-Carretero S., Tamuri A.U., Battini M., Nascimento F.F., Carlisle E., Asher R.J., Yang Z., Donoghue P.C., Dos Reis M. 2022.

A species-level timeline of mammal evolution integrating phylogenomic data. Nature. 602:263–267.
Archibald J.D., Averianov A.O., Ekdale E.G. 2001. Late Cretaceous relatives of rabbits, rodents, and other extant eutherian mammals. Nature. 414:62–65.
Bapst D.W., Schreiber H.A., Carlson S.J. 2018. Combined analysis of extant Rhynchonellida (Brachiopoda) using morphological and molecular data. Syst. Biol. 67:32–48.
Barido-Sottani J., Van Tiel N.M., Hopkins M.J., Wright D.F., Stadler T., Warnock R.C. 2020. Ignoring fossil age uncertainty leads to inaccurate topology and divergence time estimates in time calibrated tree inference. Front. Ecol. Evol. 8:183.
Baum D.A., Offner S. 2008. Phylogenics & tree-thinking. Am. Biol. Teach. 70:222–229.
Beck R.M., Baillie C. 2018. Improvements in the fossil record may largely resolve current conflicts between morphological and molecular estimates of mammal phylogeny. Proc. R. Soc. B. 285:20181632.
Bloch J.I., Fisher D.C., Rose K.D., Gingerich P.D. 2001. Stratocladistic analysis of Paleocene Carpolestidae (Mammalia, Plesiadapiformes) with description of a new late Tiffanian genus. J. Vert. Paleontol. 21:119–131.
Bollback J.P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171–1180.
Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchˆene S., Fourment M., Gavryushkina A., Heled J., Jones G., Kühnert D., De Maio N., et al. 2019. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 15:e1006650.
Bourdon E., De Ricqlès A., Cubo J. 2009. A new Transantarctic relationship: morphological evidence for a Rheidae–Dromaiidae–Casuariidae clade (Aves, Palaeognathae, Ratitae). Zool. J. Linn. Soc. 156:641–663.
Brocklehurst N., Haridy Y. 2021. Do meristic characters used in phylogenetic analysis evolve in an ordered manner? Syst. Biol. 70:707–718.
Brown J.M. 2014a. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst. Biol. 63:334–348.
Brown J.M. 2014b. Predictive approaches to assessing the fit of evolutionary models. Syst. Biol. 63:289–292.
Brown J.M., Thomson R.C. 2018. Evaluating model performance in evolutionary biology. Annu. Rev. Ecol. Evol. Syst. 49:95–114.
Brusatte S.L., Montanari S., Yi H.-y, Norell M.A. 2011. Phylogenetic corrections for morphological disparity analysis: new methodology and case studies. Paleobiology. 37:1–22.
Caldwell M.W., Simões T.R., Palci A., Garberoglio F.F., Reisz R.R., Lee M.S., Nydam R.L. 2021. *Tetrapodophis amplectus* is not a snake: reassessment of the osteology, phylogeny and functional morphology of an Early Cretaceous dolichosaurid lizard. J. Syst. Paleontol. 19:893–952.
Casali D.M., Boscaini A., Gaudin T.J., Perini F.A. 2022. Reassessing the phylogeny and divergence times of sloths (mammalia: Pilosa: Folivora), exploring alternative morphological partitioning and dating models. Zool. J. Linn. Soc. 196:1505–1551.
Casali D.M., Freitas F.V., Perini F.A. 2023. Evaluating the impact of anatomical partitioning on summary topologies obtained with Bayesian phylogenetic analyses of morphological data. Syst. Biol. 72:62–77.
Clarke J.A., Middleton K.M. 2008. Mosaicism, modules, and the evolution of birds: results from a Bayesian approach to the study of morphological evolution using discrete character data. Syst. Biol. 57:185–201.
Close R.A., Friedman M., Lloyd G.T., Benson R.B. 2015. Evidence for a mid-Jurassic adaptive radiation in mammals. Curr. Biol. 25:2137–2142.
Duchêne D.A., Duchêne S., Ho S.Y. 2017. New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. Mol. Biol. Evol. 34:1529–1534.
Duchêne D.A., Duchêne S., Ho S.Y. 2018. Differences in performance among test statistics for assessing phylogenomic model adequacy. Genome Biol. Evol. 10:1375–1388.
Egi N., Holroyd P.A., Tsubamoto T., Soe A.N., Takai M., Ciochon R.L. 2005. Proviverrine hyaenodontids (Creodonta: Mammalia) from the Eocene of Myanmar and a phylogenetic analysis of the proviverrines from the Para-Tethys area. J. Syst. Paleontol. 3:337–358.
Fabreti L.G., Coghill L.M., Thomson R.C., Höhna S., Brown J.M. 2024. The expected behaviors of posterior predictive tests and their unexpected interpretation. Mol. Biol. Evol. 41:msae051.
Farris J.S. 1989. The retention index and the rescaled consistency index. Cladistics. 5:417–419.
Farris J.S., Kluge A.G., Eckardt M.J. 1970. A numerical approach to phylogenetic systematics. Syst. Zool. 19:172–189.
Felsenstein J. 1983. Parsimony in systematics: biological and statistical issues. Annu. Rev. Ecol. Syst. 14:313–333.
Felsenstein J. 1992. Phylogenies from restriction sites: a maximum-likelihood approach. Evolution. 46:159–173.
Gatesy J. 2007. A tenth crucial question regarding model use in phylogenetics. Trends Ecol Evol. 22:509–510.
Gavryushkina A., Heath T.A., Ksepka D.T., Stadler T., Welch D., Drummond A.J. 2017. Bayesian total-evidence dating reveals the recent crown radiation of penguins. Syst. Biol. 66:57–73.
Gelman A., Meng X.-L., Stern H. 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statisticasinica 6:733–760.
Goloboff P.A., Pittman M., Pol D., Xu X. 2019. Morphological data sets fit a common mechanism much more poorly than DNA sequences and call into question the Mkv model. Syst. Biol. 68:494–504.
Goloboff P.A., Torres A., Arias J.S. 2018. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. Cladistics. 34:407–437.
Gower J.C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27:857–871.
Hansen T.F., Bolstad G.H., Tsuboi M. 2022. Analyzing disparity and rates of morphological evolution with model-based phylogenetic comparative methods. Syst. Biol. 71:1054–1072.
Harrison L.B., Larsson H.C. 2015. Among-character rate variation distributions in phylogenetic analysis of discrete morphological characters. Syst. Biol. 64:307–324.
Hastings W.K. 1970. Monte carlo sampling methods using Markov chains and their applications. Biometrika. 57:97–109.
Höhna S., Coghill L.M., Mount G.G., Thomson R.C., Brown J.M. 2018. P3: phylogenetic posterior prediction in RevBayes. Mol. Biol. Evol. 35:1028–1034.
Höhna S., Landis M.J., Heath T.A., Boussau B., Lartillot N., Moore B.R., Huelsenbeck J.P., Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst. Biol. 65:726–736.
Höhna S., Landis M.J., Huelsenbeck J.P. 2021. Parallel power posterior analyses for fast computation of marginal likelihoods in phylogenetics. PeerJ. 9:e12438.
Hopkins M., Gerber S., de la Rosa L.N., Muller G. 2017. Morphological Disparity. In: Nuño de la Rosa L., Müller G.B., Lorenz K., editors. *Evolutionary developmental biology*. Springer Nature Switzerland. p. 965–976.
Hopkins M.J., Smith A.B. 2015. Dynamic evolutionary change in post-paleozoic echinoids and the importance of scale when interpreting changes in rates of evolution. Proc. Natl. Acad. Sci. U.S.A. 112:3758–3763.
Huelsenbeck J.P., Nielsen R., Bollback J.P. 2003. Stochastic mapping of morphological characters. Syst. Biol. 52:131–158.
Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. Mammalian Protein Metab 3:21–132.
Khakurel B., Grigsby C., Tran T.D., Zariwala J., Höhna S., Wright A.M. in press. The fundamental role of character coding in Bayesian morphological phylogenetics. Syst. Biol.: syae033.
Klopfstein S., Vilhelmsen L., Ronquist F. 2015. A nonstationary Markov model detects directional evolution in hymenopteran morphology. Syst. Biol. 64:1089–1103.
Kluge A.G., Farris J.S. 1969. Quantitative phyletics and the evolution of anurans. Syst. Biol. 18:1–32.
Koch N.M., Parry L.A. 2020. Death is on our side: paleontological data drastically modify phylogenetic hypotheses. Syst. Biol. 69:1052–1067.
Kolaczkowski B., Thornton J.W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature. 431:980–984.
Lee M.S., Palci A. 2015. Morphological phylogenetics in the genomic age. Curr. Biol. 25:R922–R929.
Lehmann O.E., Ezcurra M.D., Butler R.J., Lloyd G.T. 2019. Biases with the generalized Euclidean distance measure in disparity analyses with high levels of missing data. Palaeontology. 62:837–849.

Lemey P., Rambaut A., Drummond A.J., Suchard M.A. 2009. Bayesian phylogeography finds its roots. PLoS Comput. Biol. 5:e1000520.
Lewis P.O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.
Lloyd G.T. 2016. Estimating morphological diversity and tempo with discrete character-taxon matrices: implementation, challenges, progress, and future directions. Biol. J. Linn. Soc. 118:131–151.
López-Antõnanzas R., Mitchell J., Simões T.R., Condamine F.L., Aguilée R., Peláez Campomanes P., Renaud S., Rolland J., Donoghue P.C. 2022. Integrative phylogenetics: tools for palaeontologists to explore the tree of life. Biology. 11:1185.
Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E. 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092.
Mongiardino Koch N., Garwood R.J., Parry L.A. 2021. Fossils improve phylogenetic analyses of morphological characters. Proc. Biol. Sci. 288:20210044.
Murphy J.L., Puttick M.N., O'Reilly J.E., Pisani D., Donoghue P.C. 2021. Empirical distributions of homoplasy in morphological data. Palaeontology. 64:505–518.
Nylander J.A., Ronquist F., Huelsenbeck J.P., Nieves-Aldrey J. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53:47–67.
O'Reilly J.E., Puttick M.N., Parry L., Tanner A.R., Tarver J.E., Fleming J., Pisani D., Donoghue P.C. 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. Biol. Lett. 12:20160081.
Oksanen, J., G. L. Simpson, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. O'Hara, P. Solymos, M. H. H. Stevens, E. Szoecs, H. Wagner, M. Barbour, M. Bedward, B. Bolker, D. Bor card, G. Carvalho, M. Chirico, M. De Caceres, S. Durand, H. B. A. Evangelista, R. FitzJohn, M. Friendly, B. Furneaux, G. Hannigan, M. O. Hill, L. Lahti, D. McGlinn, M.-H. Ouellette, E. Ribeiro Cunha, T. Smith, A. Stier, C. J. Ter Braak, and J. Weedon. 2022. vegan. R package version 2.6-4.
Plummer M., Best N., Cowles K., Vines K., Sarkar D., Bates D., Almond R., Magnusson A. 2006. CODA: convergence diagnosis and output analysis for MCMC. R News. 6:7–11.
Puttick M.N., O'Reilly J.E., Tanner A.R., Fleming J.F., Clark J., Holloway L., Lozano Fernandez J., Parry L.A., Tarver J.E., Pisani D., et al. 2017. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. Proc R Soc B: Biol Sci. 284:20162290.
Pyron R.A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of lissamphibia. Syst. Biol. 60:466–481.
Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
Robinson T.J., Borror C.M., Myers R.H. 2004. Robust parameter design: a review. Qual. Reliab. Eng. Int. 20:81–101.
Ronquist F., Teslenko M., Mark P.V.D., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.
Rosa B.B., Melo G.A., Barbeitos M.S. 2019. Homoplasy-based partitioning outperforms alternatives in Bayesian analysis of discrete morphological data. Syst. Biol. 68:657–671.
Rücklin M., King B., Cunningham J.A., Johanson Z., Marone F., Donoghue P.C. 2021. Acanthodian dental development and the origin of gnathostome dentitions. Nat Ecol Evol. 5:919–926.
Sanmartín I., Anderson C.L., Alarcon M., Ronquist F., Aldasoro J.J. 2010. Bayesian island biogeography in a continental setting: the rand flora case. Biol. Lett. 6:703–707.
Sanmartín I., Mark P.V.D., Ronquist F. 2008. Inferring dispersal: a Bayesian approach to phylogeny-based island biogeography, with special reference to the canary islands. J. Biogeogr. 35:428–449.
Sansom R.S., Choate P.G., Keating J.N., Randle E. 2018. Parsimony, not Bayesian analysis, recovers more stratigraphically congruent phylogenetic trees. Biol. Lett. 14:20180263.
Sansom R.S., Wills M.A., Williams T. 2017. Dental data perform relatively poorly in recon structing mammal phylogenies: morphological partitions evaluated with molecular benchmarks. Syst. Biol. 66:813–822.
Schliep K. 2011. phangorn: phylogenetic analysis in R. Bioinformatics. 27:592–593.
Schoch R.R., Sues H.-D. 2013. A new dissorophid temnospondyl from the Lower Permian of north-central Texas. C.R. Palevol 12:437–445.
Schwery O., Freyman W.A., Goldberg E.E. 2023. adequaSSE: Model adequacy testing for trait-dependent diversification models. bioRxiv: 2023–2003.
Shepherd D.A., Klaere S. 2019. How well does your phylogenetic model fit your data? Syst. Biol. 68:157–167.
Shoshani J., Walter R.C., Abraha M., Berhe S., Tassy P., Sanders W.J., Marchant G.H., Lib sekal Y., Ghirmai T., Zinner D. 2006. A proboscidean from the late Oligocene of Eritrea, a "missing link" between early Elephantiformes and Elephantimorpha, and biogeographic implications. Proc. Natl. Acad. Sci. U.S.A. 103:17296–17301.
Simões T.R., Caldwell M.W., Pierce S.E. 2020. Sphenodontian phylogeny and the impact of model choice in Bayesian morphological clock estimates of divergence times and evolutionary rates. BMC Biol. 18:1–30.
Simpson G.G. 1952. How many species? Evolution. 6:342–342.
Slater G.J., Pennell M.W. 2014. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. Syst. Biol. 63:293–308.
Slowinski J.B. 1993. "unordered" versus "ordered" characters. Syst. Biol. 42:155–165.
Smith T.J., Sansom R.S., Pisani D., Donoghue P.C. 2023. Fossilization can mislead analyses of phenotypic disparity. Proc R Soc B. 290:20230522.
Sober E. 2004. The contest between parsimony and likelihood. Syst. Biol. 53:644–653.
Steel M., Penny D. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17:839–850.
Tomiya S. 2011. A new basal caniform (Mammalia: Carnivora) from the middle Eocene of North America and remarks on the phylogeny of early carnivorans. PLoS One. 6:e24146.
Tuffley C., Steel M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59:581–607.
Weisbecker V., Beck R.M., Guillerme T., Harrington A.R., Lange-Hodgson L., Lee M.S., Mar don K., Phillips M.J. 2023. Multiple modes of inference reveal less phylogenetic signal in marsupial basicranial shape compared with the rest of the cranium. Philos. Trans. R. Soc. B. 378:20220085.
Wills M.A. 1998. Crustacean disparity through the Phanerozoic: comparing morphological and stratigraphic data. Biol. J. Linn. Soc. 65:455–500.
Wills, M. A. 2001. Morphological disparity: a primer. In: Adrain J.M., Edgecombe G.D., Lieberman B.S., editors. Fossils, phylogeny, and form: an analytical approach. Springer Science+Business Media New York. p. 55–144.
Wright, A., P. J. Wagner, and D. F. Wright. 2021. Testing character evolution models in phylogenetic paleobiology: a case study with Cambrian echinoderms. University Printing House, Cambridge CB2 8BS, United Kingdom: Cambridge University Press.
Wright A.M. 2019. A systematist's guide to estimating Bayesian phylogenies from morphological data. Insect Syst. Diversity. 3:2.
Wright A.M., Hillis D.M. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. PLoS One. 9:e109210.
Wright A.M., Lloyd G.T., Hillis D.M. 2016. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. Syst. Biol. 65:602–611.
Xie W., Lewis P.O., Fan Y., Kuo L., Chen M.-H. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60:150–160.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.
Zhang C., Rannala B., Yang Z. 2012. Robustness of compound Dirichlet priors for Bayesian inference of branch lengths. Syst. Biol. 61:779–784.