

# Model Adequacy and the Macroevolution of Angiosperm Functional Traits

Matthew W. Pennell,<sup>1,\*</sup> Richard G. FitzJohn,<sup>2</sup> William K. Cornwell,<sup>3</sup> and Luke J. Harmon<sup>1</sup>

1. Department of Biological Sciences and Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, Idaho 83844; 2. Department of Biological Sciences, Macquarie University, Sydney, New South Wales 2109, Australia; 3. School of Biological, Earth, and Environmental Sciences, University of New South Wales, Sydney, New South Wales 2052, Australia

Submitted April 7, 2014; Accepted March 31, 2015; Electronically published June 12, 2015

Online enhancement: supplemental PDF.

**ABSTRACT:** Making meaningful inferences from phylogenetic comparative data requires a meaningful model of trait evolution. It is thus important to determine whether the model is appropriate for the data and the question being addressed. One way to assess this is to ask whether the model provides a good statistical explanation for the variation in the data. To date, researchers have focused primarily on the explanatory power of a model relative to alternative models. Methods have been developed to assess the adequacy, or absolute explanatory power, of phylogenetic trait models, but these have been restricted to specific models or questions. Here we present a general statistical framework for assessing the adequacy of phylogenetic trait models. We use our approach to evaluate the statistical performance of commonly used trait models on 337 comparative data sets covering three key angiosperm functional traits. In general, the models we tested often provided poor statistical explanations for the evolution of these traits. This was true for many different groups and at many different scales. Whether such statistical inadequacy will qualitatively alter inferences drawn from comparative data sets will depend on the context. Regardless, assessing model adequacy can provide interesting biological insights—how and why a model fails to describe variation in a data set give us clues about what evolutionary processes may have driven trait evolution across time.

**Keywords:** phylogenetic comparative methods, model adequacy, independent contrasts, angiosperm functional traits.

## Introduction

A statistical model may provide the best explanation for a data set compared to a few other models but still be a very poor explanation in terms of capturing the patterns of variation present in the data. For simple linear regression models, absolute model fit, or adequacy, is commonly assessed by simply plotting the data alongside the best re-

gression line. While not quantitative, visualizing the bivariate distribution can provide important insights regarding the fit of the model that are not captured by summaries such as the  $R^2$  or  $P$  value, such as whether the relationship is indeed linear (for a classic case study, see Anscombe 1973). For these types of models, there are also a wide variety of statistical tests of model adequacy (e.g., the relationship between the residuals and the independent variable,  $\chi^2$  goodness-of-fit test, etc.) that compliment our visual intuition about model adequacy. Such formal tests used alongside informal visualizations can help researchers assess whether the inferences drawn from the fitted model are meaningful and, more interestingly, suggest how a model can be improved (Gelman and Shalizi 2013).

Modern phylogenetic comparative methods for investigating trait evolution are almost exclusively model based (recently reviewed in O'Meara 2012; Pennell and Harmon 2013), meaning that inferences are contingent on both the phylogenetic tree and the model for the traits. Selecting a good model is therefore essential for making robust inferences. Researchers typically use likelihood ratio tests or information-theoretic measures (i.e., Akaike information criterion [AIC], Bayesian information criterion) to select among models (Mooers et al. 1999; Harmon et al. 2010; Hunt 2012), but these provide a measure of only relative fit. Unlike in linear regression models, for most phylogenetic models of trait evolution, it is usually very challenging to visually assess the adequacy of a model. This problem is compounded for relatively complex models such as multi-rate Brownian motion models (BM; O'Meara et al. 2006; Eastman et al. 2011) or multioptima Ornstein-Uhlenbeck models (OU; Hansen 1997; Butler and King 2004; Beaulieu et al. 2012; Uyeda and Harmon 2014). One can plot the trait values at the tips of the phylogeny, but determining “by eye” whether this distribution is consistent with the traits having evolved under the proposed model is difficult at small scales and impossible for large phylogenies.

\* Corresponding author; e-mail: mwpennell@gmail.com.

A number of statistical procedures have been proposed to quantitatively assess the absolute fit of a model of trait evolution (e.g., Garland et al. 1992, 1993; Purvis and Rambaut 1995; Díaz-Uriarte and Garland 1996; Freckleton and Harvey 2006; Boettiger et al. 2012; Beaulieu et al. 2013; Blackmon and Demuth 2014; Slater and Pennell 2014). These can be generally classified into two types of approaches. The first are tests for specific deviations from a particular model. In the early days of phylogenetic comparative biology, the focus was primarily on inferring character correlations in order to test hypotheses regarding adaptation (e.g., Felsenstein 1985; Grafen 1989; Harvey and Pagel 1991; Lynch 1991). Accordingly, a number of tests were developed to assess the reliability of assuming a BM model, which formed the basis for all phylogenetic tests of continuous character evolution at the time. Garland et al. (1992) proposed plotting the standardized independent contrasts (*sensu* Felsenstein 1985) against the standard deviation of each contrast. If the contrasts and their standard deviations are correlated, this would suggest that the model (or the phylogeny) is not adequate. Purvis and Rambaut (1995) suggested using the relationship between the contrasts and the height above the root at which they were generated (for a slight modification of this test, see also Freckleton and Harvey 2006). Similarly, Beaulieu et al. (2013) and Blackmon and Demuth (2014) used summary statistics to evaluate whether a set of discrete character data was consistent with some variant of a Mk model (Pagel 1994). These are all very useful ideas, and we have adopted many of these in the method we present below, but each approach is informative with respect to only a single type of misspecification for a single type of model.

The second class of approaches is to use Monte Carlo simulations to compare an observed data set to those expected under a model. Garland et al. (1993) and Díaz-Uriarte and Garland (1996) developed such an approach 2 decades ago. However, as this work preceded the development of analytical tools for fitting alternative (i.e., non-BM) models, the simulation parameters were not estimated directly from the data, and, therefore, “reasonable” parameter estimates had to be chosen a priori. More recently, two approaches have been suggested for assessing model adequacy using parameters estimated directly from the data. Boettiger et al. (2012) proposed simulating data under two candidate models using the maximum likelihood parameter estimates from each model and then fitting both models to both simulated data sets. They then computed the likelihood ratio between the two candidate models for each simulating condition. After many simulations, a distribution of likelihood ratios could be obtained for each case and these distributions compared to assess whether there was sufficient information in the data to favor one model over the other. Slater and Pennell (2014) used posterior pre-

dictive simulation (explained below) to assess the absolute fit of an “early-burst” (EB) model of trait evolution, in which rates of trait evolution declined through time, compared to that of a BM model. Both Boettiger et al. (2012) and Slater and Pennell (2014) focused on the ability to distinguish between two models using absolute fit. Our aim here is more general: we want to compare the fit of the model to the universe of possible models.

In this article, we propose a statistical framework for assessing the adequacy of phylogenetic models of quantitative trait evolution that generalizes previous approaches to a wide variety of alternative models. Our central thesis is that assessing model adequacy in a general way can provide valuable insights into evolutionary processes and patterns that are not evident from comparing a limited set of models. For example, one common application of phylogenetic trait models is to make inferences regarding the rate (tempo) of evolution using model selection (e.g., Mooers et al. 1999; Harmon et al. 2010; Hunt 2012; Slater 2013). Statements about rates are informative only in the context of a specific model (Hunt 2012). It is therefore imperative to know whether a model is really capturing the variation of the data in absolute terms.

In an oft-cited example of the model comparison approach, Harmon et al. (2010) compared three simple models of trait evolution across 49 clades and tallied the frequency with which the models were preferred in order to draw inferences about general patterns. We perform the same analysis but on a much larger scale. We analyze 337 data sets on three important angiosperm (flowering plants) functional traits using a recently published time-calibrated phylogeny (Zanne et al. 2014). We then assess the adequacy of the best-fitting model across all the data sets to determine how often one of these simple models would be adequate to make reliable inferences about rate of trait evolution.

### A General Framework for Assessing the Adequacy of Phylogenetic Models

We focus here on models that describe the evolution of a single, continuously valued trait. More specifically, our approach works for models that predict that trait values at the tips come from a multivariate normal distribution. This applies to most models of quantitative trait evolution that have been developed to date (see below for details on the scope of the method).

If we have a phylogenetic tree consisting of  $n$  lineages and data on the trait values observed at each tip  $X$  ( $X = x_1, x_2, \dots, x_n$ ), we can fit a model  $\mathcal{M}$  with parameters  $\theta$  to describe the pattern of trait evolution along the phylogeny. There are two primary ways of fitting models to comparative data. The first is used to obtain a point esti-

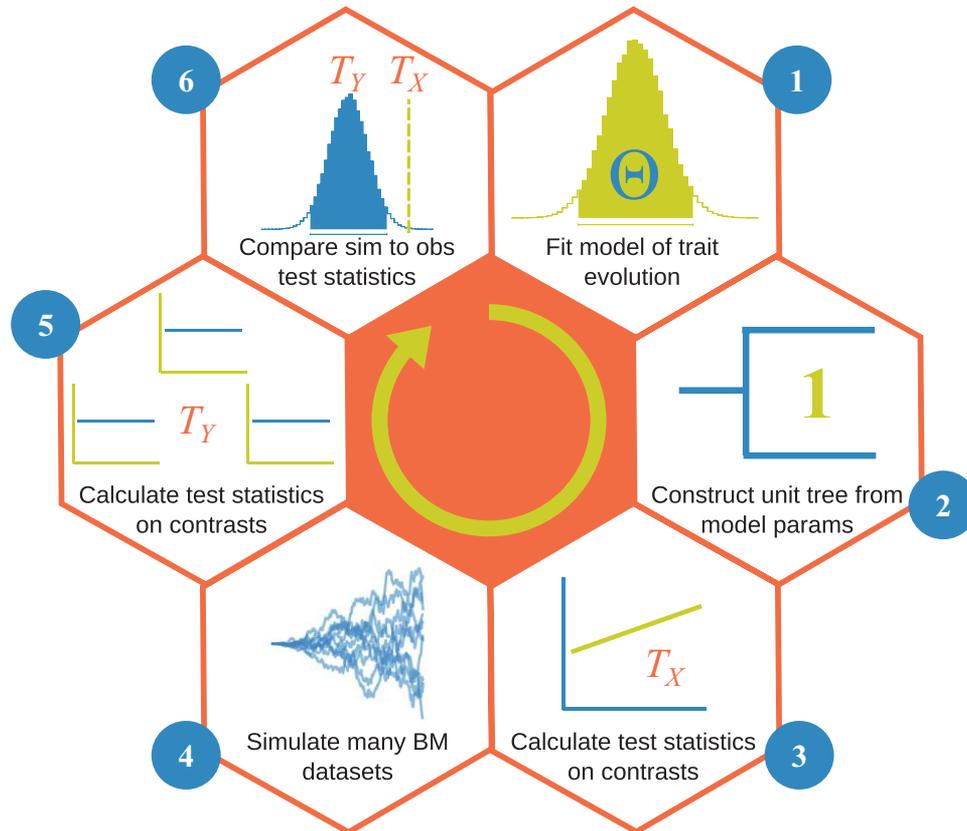
mate of  $\theta$  ( $\hat{\theta}$ ), via maximum likelihood (ML), restricted maximum likelihood (REML), least squares, and so on. The second is to estimate the posterior probability distribution  $\Pr(\theta|X, \mathcal{M})$  using Bayesian approaches. For the models used in comparative biology, estimating  $\Pr(\theta|X, \mathcal{M})$  requires using Markov chain Monte Carlo (MCMC) machinery to sample values of  $\theta$ .

Most analyses using comparative data aim to answer one of the following questions: What values of  $\theta$  best explain  $X$  given  $\mathcal{M}$ ? Or, does  $\mathcal{M}_1$  explain the data better than  $\mathcal{M}_0$ ? Our approach is conceptually distinct in that we want to ask how likely it is that model  $\mathcal{M}$  with parameters  $\theta$  would produce a data set similar to  $X$  if we reran evolution.

While optimizing and Bayesian approaches to model fitting are philosophically different from one another, our approach to assessing model adequacy is the same for both: (1) fit the model of trait evolution; (2) rescale the branch lengths of the phylogeny to place the data on a standard

scale; (3) calculate a set of test statistics,  $T_X$ , that provide statistical summaries of the observed data; (4) simulate many new data sets  $Y_1, Y_2, \dots, Y_m$  under the model using the estimated parameters; (5) calculate test statistics on the simulated data  $T_{Y,1}, T_{Y,2}, \dots, T_{Y,m}$ ; (6) compare  $T_X$  to the distribution of  $T_Y$ . If  $T_X$  deviates significantly from the distribution of  $T_Y$ , we can consider the model an inadequate descriptor (see fig. 1).

If we have a point estimate of the model parameters, we simulate  $Y_1, Y_2, \dots, Y_m$  on the phylogeny according to  $\hat{\theta}$  and  $\mathcal{M}$ . We then compare a single set of test statistics  $T_X$  calculated from our observed data to the distribution of values for  $T_Y$  computed across all  $m$  simulated data sets. In statistical terminology, this procedure is known as parametric bootstrapping. Parametric bootstrapping is likely familiar to phylogenetic biologists in the form of the Goldman-Cox test (Goldman 1993) for assessing the adequacy of sequence evolution models and more recently



**Figure 1:** Schematic diagram representing our approach for assessing model adequacy. 1, Fit a model of trait evolution to the data. 2, Use the estimated model parameters to build a unit tree. 3, Compute the contrasts from the data on the unit tree and calculate a set of test statistics  $T_X$ . 4, Simulate a large number of data sets on the unit tree, using a Brownian motion (BM) model with  $\sigma^2 = 1$ . 5, Calculate the test statistics on the contrasts of each simulated data set  $T_Y$ . 6, Compare the observed and simulated test statistics. If the observed test statistic lies in the tails of the distribution of simulated test statistics, the model can be rejected as inadequate. The rotational circle in the center of the diagram indicates that assessing model adequacy is an iterative process. If a model is rejected as inadequate, the next step is to propose a new model and repeat the procedure.

the phylogenetic Monte Carlo approach of Boettiger et al. (2012).

If we have a posterior probability distribution  $\Pr(\theta|X, \mathcal{M})$ , we can assess model adequacy using posterior predictive simulation (Rubin 1984; Gelman et al. 1996). We obtain new data sets by sampling from a second distribution, the posterior predictive distribution

$$\Pr(Y|X, \mathcal{M}) = \int \Pr(Y|\theta, \mathcal{M})\Pr(\theta|X, \mathcal{M})d\theta, \quad (1)$$

where  $\Pr(Y|X, \mathcal{M})$  is the probability of a new data set  $Y$  given  $X$  and  $\mathcal{M}$ , averaged over the posterior distribution of the parameters.  $\Pr(Y|X, \mathcal{M})$  can be approximated by simulating data sets using parameters drawn from the posterior distribution. Therefore, the data sets  $Y_1, Y_2, \dots, Y_m$  are each generated from different values of  $\theta$ . Posterior predictive simulation approaches have been previously developed for models in molecular phylogenetics (Bollback 2002; Brown 2014; Lewis et al. 2014; Reid et al. 2014) and recently for phylogenetic comparative methods (PCMs; Slater and Pennell 2014) but have not been widely adopted in either field.

#### *Test Statistics*

No simulated data set will ever be exactly the same as our observed data set. We therefore need to choose informative test statistics in order to evaluate whether the model predicts data sets that are similar to our observed data set in meaningful ways. As the states at the tips of the phylogeny are not independent—this is why we are using PCMs in the first place—calculating test statistics on the data directly is not generally informative for models in comparative biology. We account for the nonindependence of the observed data by calculating test statistics on the set of contrasts (i.e., “phylogenetically independent contrasts”; Felsenstein 1985) computed at each node. (For details on how contrasts are calculated, we refer readers to Felsenstein 1985; Rohlf 2001; Blomberg et al. 2012.) Under BM the contrasts will be independent and identically distributed (i.i.d.) according to a normal distribution with mean 0 and standard deviation  $\sigma$ ; that is, contrasts are  $\sim \mathcal{N}(0, \sigma)$ , where  $\sigma^2$  is the BM rate parameter (Felsenstein 1985). This i.i.d. condition allows us to perform standard statistical tests on the contrasts.

The choice of what test statistics to use for assessing model adequacy is ultimately one of balancing statistical intuition and computational effort. We have chosen a set of six test statistics, described in table 1, to compute on the contrasts because they capture a range of possible model violations and have well-understood statistical properties. All of these essentially evaluate whether the contrasts come from the distribution expected under BM.

Alternative test statistics are certainly possible. One could, for instance, calculate the median of the squared contrasts, the skew of the distribution of contrasts, and so on. If the generating model was known, we could use established procedures for selecting a set of sufficient (or approximately sufficient; Joyce and Majoram 2008) test statistics for that model, as is typically done when computing likelihood ratio tests. However, the aim of our approach is to assess the fit of a proposed model without reference to a true model. Our test statistics will detect many types of model misspecification, but this does not mean that they will necessarily detect every type of model misspecification. We encourage researchers interested in specific questions to explore alternative test statistics that capture deviations relevant to the problem at hand.

An additional challenge is determining how to deal with the statistical problems (i.e., inflated type I error rates) that may be introduced when using many test statistics. In our analyses, we chose not to correct our  $P$  values for multiple comparisons (using Bonferroni, false discovery rates, etc.). We did this for a number of reasons. First, our tests are not truly independent, and the degree of correlation between test statistics will necessarily depend on the “true” model of trait evolution. Second, as argued by Gelman (2006, p. 175), we might be interested in the specific aspects of the data that differ from the expectations under the model; rather than focus on whether a model should be accepted or rejected, we “want to understand the limits of its applicability in realistic replications.”

#### *Beyond Brownian Motion*

All of our test statistics are designed to evaluate the adequacy of a BM model of trait evolution. However, if we propose a different model for the evolution of the trait, such as an OU (Hansen 1997) process, then the expected distribution of the contrasts is different. For example, under an OU model, contrasts will not be i.i.d. (Hansen 1997). The expected distribution of contrasts under most models of trait evolution, aside from BM, is not formally characterized, and even if it were, this would necessitate a specific set of test statistics for every model proposed.

Our solution to this problem is to create what we term a “unit tree,” which is a phylogenetic tree transformation that captures the dynamics of trait change under a particular evolutionary model. For a particular evolutionary model  $\mathcal{M}$  (with parameter values  $\theta$ ), we define a unit tree as a phylogenetic tree that has the following property: the length of each branch is equal to the amount of variance expected to accumulate over it under  $\mathcal{M}, \theta$ . The variance is standardized, such that the expected distribution of the trait data on the unit tree is equal to that of a BM model with a rate  $\sigma^2$  equal to 1.

**Table 1:** Description of test statistics used to assess model adequacy

Test statistic	Description
$M_{\text{SIG}}$	The mean of the squared contrasts. This is equivalent to the restricted maximum likelihood estimator of the Brownian motion rate parameter $\sigma^2$ (Garland et al. 1992; Rohlf 2001). $M_{\text{SIG}}$ is a metric of overall rate. Violations detected by $M_{\text{SIG}}$ indicate whether the overall rate of trait evolution is over- or underestimated.
$C_{\text{VAR}}$	The coefficient of variation (standard deviation/mean) of the absolute value of the contrasts. If $C_{\text{VAR}}$ calculated from the observed contrasts is greater than that calculated from the simulated contrasts, it suggests that we are not properly accounting for rate heterogeneity across the phylogeny. If $C_{\text{VAR}}$ from the observed is smaller, it suggests that contrasts are even more than the model assumes. We use the coefficient of variation rather than the variance because the mean and variance of contrasts can be highly correlated.
$S_{\text{VAR}}$	The slope of a linear model fitted to the absolute value of the contrasts against their expected variances (following Garland et al. 1992). Each (standardized) contrast has an expected variance proportional to the sum of the branch lengths connecting the node at which it is computed to its daughter lineages (Felsenstein 1985). Under a model of Brownian motion, we expect no relationship between the contrasts and their variances. We use it to test whether contrasts are larger or smaller than we expect based on their branch lengths. If, for example, more evolution occurred per unit time on short branches than long branches, we would observe a negative slope. If $S_{\text{VAR}}$ calculated from the observed data deviates substantially from the expectations, a likely explanation is branch length error in the phylogenetic tree.
$S_{\text{ASR}}$	The slope of a linear model fitted to the absolute value of the contrasts against the ancestral state inferred at the corresponding node. We estimated the ancestral state using the least squares method suggested by Felsenstein (1985) for the calculation of contrasts. (We note that this is not technically an ancestral state reconstruction [see Felsenstein 1985]; it is more properly thought of as a weighted average value for each node.) We used this statistic to evaluate whether there is variation in rates relative to the trait value. For example, do larger organisms evolve proportionally faster than smaller ones?
$S_{\text{HGT}}$	The slope of a linear model fitted to the absolute value of the contrasts against node depth (after Purvis and Rambaut 1995). This is used to capture variation relative to time. It is alternatively known as the “node-height test” and has been used to detect early bursts of trait evolution during adaptive radiations (for uses and modifications of this test see Freckleton and Harvey 2006; Slater and Pennell 2014).
$D_{\text{CDF}}$	The $D$ statistic obtained from a Kolmogorov-Smirnov test from comparing the distribution of contrasts to that of a normal distribution with mean 0 and standard deviation equal to the root of the mean of squared contrasts (the expected distribution of the contrasts under Brownian motion; see Felsenstein 1985; Rohlf 2001). We chose this to capture deviations from normality. For example, if traits evolved via a “jump-diffusion”-type process (Landis et al. 2013) in which there were occasional bursts of rapid phenotypic evolution (Pennell et al. 2013), the tip data would no longer be multivariate normal owing to a few contrasts throughout the tree being much larger than the rest (i.e., the distribution of contrasts would have heavy tails).

If the fitted model is adequate, the trait data at the tips of the unit tree will have the same distribution as data generated under a BM process with a rate of 1, and the contrasts will be distributed according to a standard normal distribution (hence the name unit tree). Creating the unit tree from the estimated model parameters before computing the contrasts generalizes the test statistics to most models of quantitative trait evolution (but for exceptions see Landis et al. 2013; Schraiber and Landis 2015).

We also emphasize that because the contrasts are calculated on the unit tree, the test statistics all must depend on both the data and the model; for this reason, the Bayesian version of our approach produces a distribution of observed test statistics. Once we have created the unit tree from the estimated parameters, new data sets can be simulated under the model simply using a BM process with  $\sigma^2 = 1$ , which has the added benefit of being computationally efficient.

The distribution of test statistics calculated on these simulated data sets can then be compared to the test statistics from the observed data.

#### *Details of Unit Tree Construction and the Scope of This Approach*

Here we formalize our definition of the unit tree and delimit the scope of our approach. Readers can skip this section without missing the main point. A unit tree can be constructed from any evolutionary model where the trait has expected variance-covariance matrix  $\mathbf{V}$  that satisfies the (generalized) three-point condition proposed by Ho and Ané (2014), and the data follow a multivariate normal distribution. A matrix  $\mathbf{V}$  has a strict three-point structure if the following condition holds: for any lineages  $i, j, k$ , the two smallest of  $V_{ij}, V_{ik}, V_{jk}$  are equal. Under a simple BM model,

it is straightforward to show that this condition holds. If  $\mathbf{C}$  is the matrix representation of the phylogeny (such that  $C_{ij}$  is the shared path length between lineages  $i$  and  $j$ ), then by the nature of the tree structure, the three-point condition will hold for  $\mathbf{C}$ . Since under BM  $\mathbf{V} = \sigma^2\mathbf{C}$ , then  $\mathbf{V}$  will also be three-point structured. The same holds true for any evolutionary model that is a branch length transformation of a BM model including the  $\lambda$ ,  $\delta$ ,  $\kappa$  models (Pagel 1997, 1999) and models where rates change through time (the EB model, also referred to as the accelerating/decelerating change model; Blomberg et al. 2003; Harmon et al. 2010) or across the tree (O'Meara et al. 2006; Thomas et al. 2006; Eastman et al. 2011; Revell et al. 2012; Thomas and Freckleton 2012). Standard error can be incorporated into any of these models by simply adding a species-specific scalar to each element of the diagonal. For all of the models where the three-point condition applies, we can construct a unit tree by setting the length  $\nu$  of the edge  $\{(i, j), k\}$  connecting the most recent common ancestor (MRCA) of lineages  $i$  and  $j$  to the MRCA of lineages  $i$  and  $k$  to be

$$v_{\{(i,j),k\}} = V_{ij} - V_{ik}, \quad (2)$$

where  $V_{ij}$  and  $V_{jk}$  are, by the requirements of the three-point-structured condition, equal to one another. Once all branches have been transformed, the contrasts computed on the unit tree will be i.i.d.  $\sim \mathcal{N}(0, 1)$  under the model in question.

The OU model of trait evolution also generates three-point-structured matrices when the tree is ultrametric; this is true of both single-optimum and multioptima models (Ho and Ané 2014). However, while the variance structure can easily be transformed to a BM-like tree, the contrasts on this tree will not necessarily be distributed according to a standard normal. For example, while it is often assumed when fitting a single-regime OU model that the ancestor is at the optimum trait value (see, e.g., Harmon et al. 2010), this need not be the case. Furthermore, if there are multiple optima on the phylogeny (Hansen 1997; Butler and King 2004; Ingram and Mahler 2013; Uyeda and Harmon 2014), lineages will necessarily be tracking optima that are different from the root state. Therefore, a transformation must also be made to the data in addition to the branch lengths of the phylogeny to produce contrasts that are i.i.d. according to a standard normal.

To accomplish this, we again turn to the recent work of Ho and Ané (2014). In addition to three-point-structured matrices, Ho and Ané defined a broader condition; a matrix of the form

$$\mathbf{V} = \mathbf{D}_1 \tilde{\mathbf{V}} \mathbf{D}_2$$

is considered to have a generalized three-point structure if  $\tilde{\mathbf{V}}$  is three-point structured and  $\mathbf{D}_1$  and  $\mathbf{D}_2$  are diagonal

matrices. Ho and Ané (2014) prove that many phylogenetic models are indeed of this class, including multioptimum OU models (Butler and King 2004; Ingram and Mahler 2013; Uyeda and Harmon 2014), those with varying rates and models across the tree (e.g., Beaulieu et al. 2012), and OU models fitted to non-ultrametric trees. For any model that satisfies the generalized three-point condition and where the data are assumed to come from a multivariate normal distribution, there exists some transformation to the tree (applying eq. [2] to  $\tilde{\mathbf{V}}$ ) and data (using  $\mathbf{D}_1$  and  $\mathbf{D}_2$ ) that will produce a unit tree with standard normal contrasts. We note that Slater (2014) recently pointed out that for OU models fitted to non-ultrametric trees, there is no valid transformation that can make  $\mathbf{V}$  BM-like. While this is indeed correct, it is possible to get a BM-like tree by adding a species-specific scalar to the data matrix (Ho and Ané 2014). Therefore, once the proper tree and data transformations have been made, all the test statistics described above can apply.

The above also applies to phylogenetic regression models (Grafen 1989; Lynch 1991; Martins and Hansen 1997) of the form

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

In these models, the error variance is structured by phylogeny assuming some model of trait evolution such that  $\epsilon \sim \mathcal{N}(0, \mathbf{V})$ . In these regression models,  $\mathbf{V}$  represents the variance-covariance matrix of the residuals rather than the traits (Rohlf 2001). Therefore, if  $\mathbf{V}$  is either three-point or generalized three-point structured, the tree (and possibly the data) can be transformed such that the contrasts on the residuals will be i.i.d. standard normal. This fact allows researchers to use our approach to assess the adequacy of a trait model for understanding correlations between traits. We note, however, that as  $\mathbf{V}$  affects only the error structure for these models, alternative approaches (see, e.g., Gelman et al. 2003, chap. 6) will be required to assess the adequacy of the mean structure  $Y = \beta_0 + \beta_1 X$  of the model.

### Simulations

As a verification of our method, we conducted a brief simulation study. We focused here on assessing type I error rates. As above, we emphasize that these are not necessarily the most important quantities when thinking about model adequacy, but they do provide a useful metric for demonstrating that our code is functioning correctly. The philosophy behind approaches such as ours is that the "true" model is outside of the candidate set. We want to ask whether a given model can adequately describe the variation in the data. If it does, we can consider it statisti-

cally adequate even if it is not the true model or even the best model in our set (see “Discussion” for comments on the relationship between model adequacy and model selection). Furthermore, while it is certainly interesting to examine what types of deviations in model space produce what types of deviations in the various test statistics, the number of possible simulation conditions is infinitely broad.

We simulated data under BM, single-optimum OU, and EB (the same models we used in the analysis; see below). For each set of conditions, we simulated trees of 50, 100, and 200 taxa under a pure-birth process and then rescaled the tree to be unit height. For BM, we set  $\sigma^2 = 1$ . For OU, we used  $\sigma^2 = 1$  but varied the “selection” parameter  $\alpha$  ( $\alpha = \{1, 2, 4\}$ ). For EB, we again set  $\sigma^2 = 1$  and varied  $a$ , the exponential rate of decline (for details see Harmon et al. 2010; Slater and Pennell 2014), which was set to  $a = \{\log(0.01), \log(0.02), \log(0.04)\}$ . For each parameter combination, we ran 500 simulations under two sets of conditions: (1) assuming no measurement error and (2) assuming known error rates of 5% of the expected variance in trait values across the phylogeny. We then fitted the corresponding model using ML and evaluated the type I error under each set of conditions. All simulations were conducted using diversitree (FitzJohn 2012).

### The Adequacy of Models for the Evolution of Plant Functional Traits

#### Data

We used a phylogeny of angiosperms, containing 30,535 species, from a recent study by Zanne et al. (2014). We conducted all analyses on the maximum likelihood estimate (MLE) of the phylogeny (available on Dryad, <http://datadryad.org/resource/doi:10.5061/dryad.63q27> [Zanne et al. 2013]). We used existing large data sets on three functionally important plant traits: specific leaf area (SLA; defined as fresh area/dry mass), seed mass, and leaf nitrogen content (% mass). Seed mass is a crucial part of species’ life-history strategy (Leishman et al. 2000; Westoby et al. 2002), and SLA and leaf nitrogen content are important and widely measured components of species’ carbon capture strategies (Wright et al. 2004). Understanding the macroevolutionary patterns of these three traits can provide key insights into the evolutionary processes that have shaped much of plant diversity (Cornwell et al. 2014). The SLA and leaf nitrogen data come from Wright et al. (2004), with additional SLA data from the LEDA project (Kleyer et al. 2008). Seed mass data come from the Kew database (Royal Botanical Gardens, Kew 2014). We used an approximate grepping approach to find and correct spelling mistakes and synonymy tools from the Plant List (2014) to match the trait databases to the Zanne et al. phylogeny. The full data set includes 3,293 species for

SLA, of which 2,200 match species in the Zanne et al. (2014) tree. For seed mass, the data set included 22,817 species, with 11,107 matching the phylogeny. For leaf nitrogen content, we have data for 1,574 species, with 936 included in the tree. See <https://github.com/richfitz/modeladequacy> for specific locations and scripts to access and process the original data.

We log transformed all data before analysis. We did this for biological reasons rather than to conform the data to the assumptions of the model (Houle et al. 2011). It is more meaningful to model trait evolution as a multiplicative process rather than an arithmetic one. An increase of 2 g is much more significant for the seed of an orchid than the seed of a palm tree. However, we should recognize that both of these rationales are essentially statements about model adequacy, and thus, the validity of the log transformation can be quantitatively assessed.

Because the vast majority of the species are represented by only a single record, it was not possible to use a species-specific estimate of trait standard error (SE) to account for either measurement error or intraspecific variation. As an alternative, we estimated a single SE for each trait by calculating the mean standard deviation for all species for which we had multiple measurements. The assumption of a constant SE across all species is unlikely to be correct, but even an inaccurate estimate of error is better than assuming none at all (Hansen and Bartoszek 2012).

#### Analysis

We first matched our trait data to the whole phylogeny and then extracted subclades from this data set in a three ways: (1) by family, (2) by order, and (3) by cutting the tree at 50-million-year (myr) intervals and extracting the most inclusive clades (named or unnamed) for which the most recent common ancestor of a group was younger than the time slice. (The crown age of angiosperms is estimated to be ~243 myr in the MLE tree, and the tree was cut at 50, 100, 150, and 200 myr.) We kept only subclades for which there were at least 20 species present in both the phylogeny and the trait data so that we had a reasonable ability to estimate parameters and distinguish between models (Boettiger et al. 2012; Slater and Pennell 2014). For SLA, this left us with 72 clades, seed mass, 226 clades, and leaf nitrogen content, 39 clades (337 in total). We note that these data sets are not independent, as many of the same taxa were included in family, order, and multiple time slice subtrees.

Following Harmon et al. (2010), we considered three simple models of trait evolution: (1) BM, which can be associated with genetic drift (Lande 1976; Lynch and Hill 1986; Felsenstein 1988; Lynch 1990; Hansen and Martins 1996), randomly varying selection (Felsenstein 1973), or the summation of many independent processes over macroevolutionary time (Hansen and Martins 1996; Uyeda et al.

2011; Pennell et al. 2013); (2) single-optimum OU, which is often assumed to represent stabilizing selection (following Lande 1976), though we think a more meaningful interpretation is that it represents an “adaptive zone” (Hansen 2012; Pennell and Harmon 2013); and (3) EB, which was developed as a phenomenological representation of a niche-filling process during an adaptive radiation (Blomberg et al. 2003; Harmon et al. 2010). We fitted each of these models to all 337 subclades in our data set. We then used the approach we developed to assess the adequacy of each fitted model.

All of the analyses conducted in this article were conducted using both likelihood and Bayesian inference. We did so to demonstrate the scope of our approach and because both ML and Bayesian inference are commonly used in comparative biology. We emphasize that our approach is not tied to any single statistical paradigm.

For the likelihood analyses, we fitted the three models (BM, OU, and EB) using ML with the *diversitree* package (FitzJohn 2012). We calculated the AIC score for each model. We then constructed a unit tree for each subtree, trait, and model combination using the ML estimates of the parameters. We calculated the six test statistics described above ( $M_{\text{SIG}}$ ,  $C_{\text{VAR}}$ ,  $S_{\text{VAR}}$ ,  $S_{\text{ASR}}$ ,  $S_{\text{HGT}}$ ,  $D_{\text{CDF}}$ ) on the contrasts of the data. We simulated 1,000 data sets on each unit tree using a BM model with  $\sigma^2 = 1$  and calculated the test statistics on the contrasts of each simulated data set.

For the Bayesian analysis, we fitted the same models as above using a MCMC approach, sampling parameter values using slice sampling (Neal 2003), as implemented in *diversitree* (FitzJohn 2012). For the BM model, we set a broad uniform prior on  $\sigma^2 \sim \mathcal{U}[0, 2]$ , the upper bound being substantially larger than the ML estimate of  $\sigma^2$  for any clade. For the OU model, we used the same prior for  $\sigma^2$  and drew  $\alpha$  values, the strength of attraction to the optimum, from a lognormal ( $\mu = \log(0.5)$ ,  $\sigma = \log(1.5)$ ) distribution. A complication involved in fitting OU models is deciding what assumptions to make about the state at the root  $z_0$ . Here, we follow other authors (Butler and King 2004; Harmon et al. 2010) and assume that  $z_0$  is at the optimum. For the EB model, we again used the same prior for  $\sigma^2$  and a uniform prior on  $a$ , the exponential rate of decrease in  $\sigma^2$ , such that  $a \sim \mathcal{U}[-1, 0]$  (the minimum value is much more negative than we would typically expect; Slater and Pennell 2014).

Again, for each model/trait/subtree combination, we ran a Markov chain for 10,000 generations. Preliminary investigations demonstrated that this was more than sufficient to obtain convergence and proper mixing for these simple models. After removing a burn-in of 1,000 generations, we calculated the deviance information criterion, a Bayesian analog of AIC (Spiegelhalter et al. 2002), for each model. We drew 1,000 samples from the joint posterior distribution. For

each of the sampled parameter sets, we used the parameter values to construct a unit tree and calculated our six test statistics on the contrasts. We then simulated a data set on the same unit tree and calculated the test statistics on the contrasts of the simulated data.

In the likelihood analyses, for each data set, we had one set  $\mathcal{T}_X$  of observed test statistics and 1,000 sets  $\mathcal{T}_{Y,1}, \mathcal{T}_{Y,2}, \dots, \mathcal{T}_{Y,1,000}$  of test statistics calculated on data simulated on the same unit tree. In the Bayesian version, we had 1,000 sets of observed test statistics  $\mathcal{T}_{X,1}, \mathcal{T}_{X,2}, \dots, \mathcal{T}_{X,1,000}$  using a different unit tree for each set and 1,000 sets of simulated test statistics  $\mathcal{T}_{Y,1}, \mathcal{T}_{Y,2}, \dots, \mathcal{T}_{Y,1,000}$ , with each  $\mathcal{T}_{Y,i}$  corresponding to the unit tree used to compute  $\mathcal{T}_{X,i}$ .

For both types of analyses, we report two-tailed  $P$  values (i.e., the probability that a simulated test statistic was more extreme than the observed). As a multivariate measure of model adequacy, we calculated the Mahalanobis distance, a scale-invariant metric, between the observed test statistics and the mean of our simulated test statistics, taking into account the covariance structure between the simulated test statistics. We took the log of the Kolmogorov-Smirnov  $D$  statistic,  $D_{\text{CDF}}$ , as the Mahalanobis measure assumes data are multivariate normal and the  $D$  statistic is bounded between 0 and 1. For the Bayesian analyses, we report the mean of the distribution of Mahalanobis distances. All analyses were conducted in R v3.0.2 (R Development Core Team 2013). Scripts to fully reproduce all analyses are available at <https://github.com/richfitz/modeladequacy>.

#### *A Case Study: Seed Mass Evolution in the Meliaceae and Fagaceae*

As an illustration of our approach, we present a case study examining seed mass evolution in two tree families, the Meliaceae, the “mahogany family,” and Fagaceae, which contains oaks, chestnuts, and beech trees. The trait data and phylogeny for both groups are subsets of the larger data set used in the analysis. Superficially, these data sets are quite similar. Both are of similar size (Meliaceae: 44 species in the data set, 550 in the clade; Fagaceae: 70 species in the data set, 600 in the clade), age (crown age of Meliaceae: ~53 myr; Fagaceae: ~40 myr) and are ecologically comparable in terms of dispersal strategy and climatic niche.

As described above, we fitted three simple models of trait evolution (BM, OU, EB) to both data sets using ML and computed AIC weights ( $AIC_w$ ; Akaike 1974; Burnham and Anderson 2004) for the three models. For both data sets, an OU model was overwhelmingly supported ( $AIC_w > 0.97$  for both groups). Therefore, looking only at relative model support, we might conclude that similar evolutionary processes are important in these two clades of trees.

Examining model adequacy provides a different perspective. We took the MLE of the parameters from the OU models for each data set and constructed a unit tree based on those parameters. We calculated our six test statistics on the contrasts of the data and then simulated 1,000 data sets on the unit tree and calculated the test statistics on the contrasts of each simulated data set (fig. 2). For seed mass evolution in Meliaceae, the OU model was an adequate model; all six observed test statistics were in the middle of the distribution of simulated test statistics ( $M_{\text{SIG}}$ ,  $P = .921$ ;  $C_{\text{VAR}}$ ,  $P = .605$ ;  $S_{\text{VAR}}$ ,  $P = .979$ ;  $S_{\text{ASR}}$ ,  $P = .485$ ;  $S_{\text{HGT}}$ ,  $P = .170$ ;  $D_{\text{CDF}}$ ,  $P = .657$ ). In contrast, for Fagaceae we found that the test statistics calculated with an OU model lay outside the expected values for  $S_{\text{VAR}}$  ( $P \approx 0$ ) and  $S_{\text{HGT}}$  ( $P = .014$ ), suggesting that the process of evolution that gave rise to these data was more complex than that captured by a simple OU process.

More specifically, the slope of the contrasts against their variances  $S_{\text{VAR}}$  is negative, meaning that contrasts computed on short branches are larger than expected (or, conversely, contrasts computed on long branches are too small). Such a pattern could be generated by phylogenetic error: the terminal branches in the Fagaceae tree are very short and are likely underestimated relative to the longer internal branches. This explanation is further supported by the fact that  $S_{\text{HGT}}$  is also negative—the standardized contrasts close to the tips are much larger than expected. The rest of the observed test statistics did not differ significantly from the simulated test statistics ( $M_{\text{SIG}}$ ,  $P = .298$ ;  $C_{\text{VAR}}$ ,  $P = .837$ ;  $S_{\text{ASR}}$ ,  $P = .074$ ;  $D_{\text{CDF}}$ ,  $P = .551$ ). This example illustrates the distinction between the conventional approach to model selection in PCMs and model adequacy. Selecting among a limited pool of models does not give a complete picture of the amount of variation that a chosen model is actually capturing.

## Results

### Simulations

In our simulations, we found that when we assessed the adequacy of the generating model, all of the test statistics showed type I errors that were consistently around or less than 0.05. This was true across models, parameters, and tree sizes and did not depend on whether we included a known SE (figs. S1–S3 in the supplemental material; figs. S1–S7 and supplemental material available online). These results demonstrate that our unit tree construction is working properly; if the MLE is equal to the generating value, then the contrasts will be i.i.d.  $\mathcal{N}(0, 1)$  and standard normal statistical properties will apply. Some of the test statistics are very conservative (have very low type I error rates) under some models. We are not aware of any general statistical

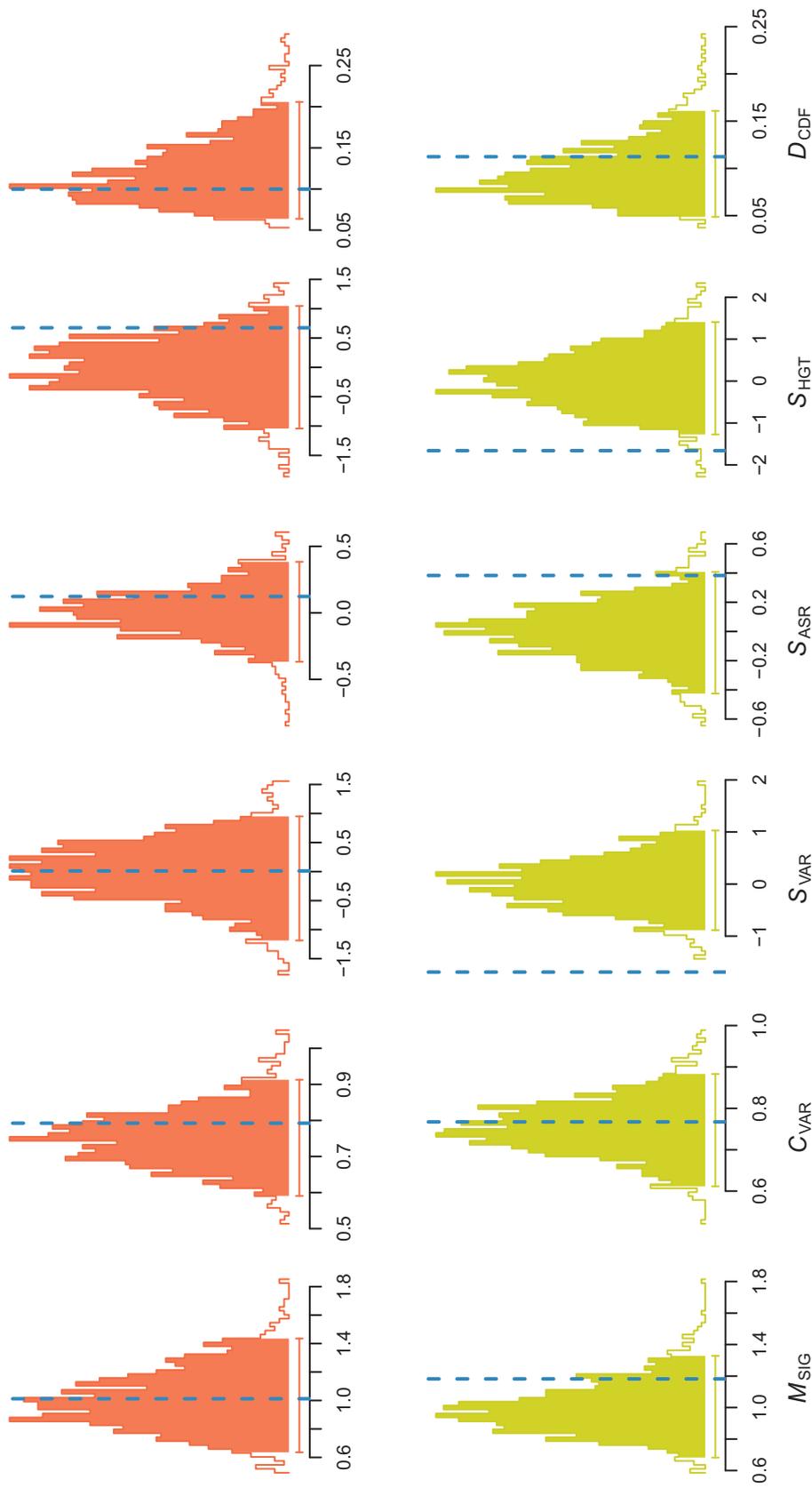
theory that will allow us to predict the conditions under which a test statistic will have low power to detect deviations from the expected distributions. However, there is an intuitive explanation for this pattern. Consider, for example, our test statistic. As mentioned above, this is equivalent to the REML estimate of  $\sigma^2$ . When we fit BM (or a more general model, of which BM is a special case) and then rescale the tree with  $\hat{\sigma}^2$ , the observed contrasts on the unit tree will effectively be minimized with respect to this quantity, and all of the contrasts on the simulated data set will tend to be larger than our observed contrasts. So if the quantity captured by the test statistic is tightly correlated with one of the parameters being optimized in the model, this test statistic will tend to have low power to detect deviations from this model.

We also found that by using multiple test statistics and reporting a type I error if any of the test statistics deviated significantly from expectations, the error rate increased substantially (up to around 20% under some conditions). However, as we discuss above, we do not think that this is necessarily a defect of the analysis and are not overly concerned with this error rate. Looking at what test statistics were violated and how they were violated is much more meaningful than simply rejecting or accepting a model based on the overall  $P$  value. Furthermore, the degree to which the type I error rate will rise with multiple comparisons will be a complex function of the generating model and the size of the data set, and there is no suitable general correction that we know of.

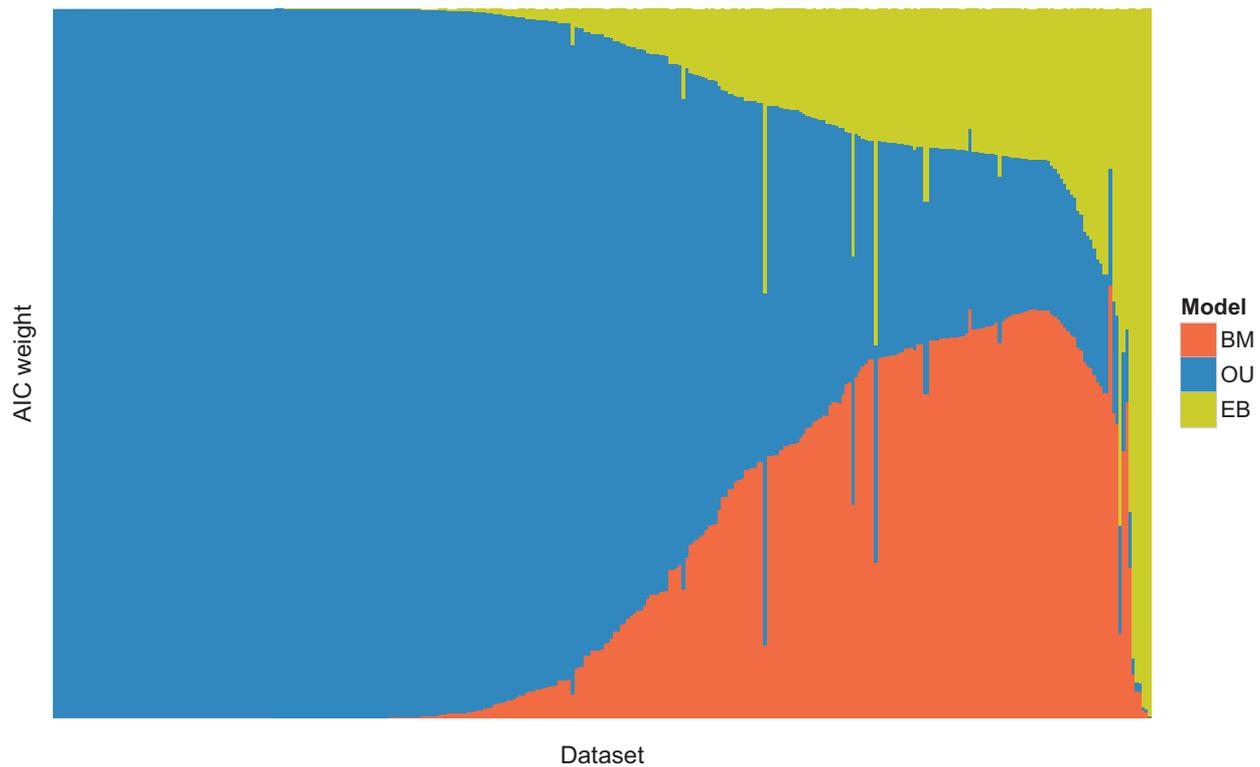
### Models for the Evolution of Angiosperm Functional Traits

Our results for likelihood and Bayesian inference were broadly similar; for conciseness, we present only the results from the likelihood analyses here. Results from the Bayesian analysis are presented in the supplemental material. Full results from all analyses can be reproduced using code and workflows available at <https://github.com/richfitz/modeladequacy>.

Across the 337 subclades, we found widespread support for OU models. For 235 clades, OU had the highest  $AIC_w$ . OU had ~100% of the  $AIC_w$  in 27 clades and >75% of the weight in 189 clades (fig. 3). Consistent with Harmon et al. (2010) we found that EB models rarely had high support (only six clades supported EB with >75%  $AIC_w$ ), though we acknowledge that even if EB dynamics were important to long-term patterns of trait evolution, these may be difficult to detect with extant species alone (Slater et al. 2012; Slater and Pennell 2014). Larger clades commonly had very high support for a single model (of the 101 clades consisting of more than 100 taxa, 44 had >90% of the  $AIC$  weight on a single model), and that was overwhelmingly likely to be an OU model (42/44 clades).



**Figure 2:** Illustration of our approach to model adequacy. We fitted three models (Brownian motion, Ornstein-Uhlenbeck [OU], and early burst) to seed mass data from two different tree families, the Meliaceae (top, red) and the Fagaceae (bottom, yellow). In both cases, an OU model (analyzed here) was strongly supported when fitted with maximum likelihood. The plotted distributions are the test statistics ( $M_{SIG}$ ,  $C_{VAR}$ ,  $S_{VAR}$ ,  $S_{ASR}$ ,  $S_{HGT}$ ,  $D_{CDF}$ ) calculated from the contrasts of the simulated data; the bars underneath the plots represent 95% of the density. The dashed vertical lines are the values of the test statistics calculated on the contrasts of the observed data. Using our test statistics, an OU model appears to be an adequate model for the evolution of seed mass in the Meliaceae; for all of the test statistics, the observed test statistic lies in the middle of the distribution of simulated test statistics. For the Fagaceae, the slopes of the contrasts against their expected variances  $S_{VAR}$  and node height  $S_{HGT}$  are much lower than the expectations under the model.



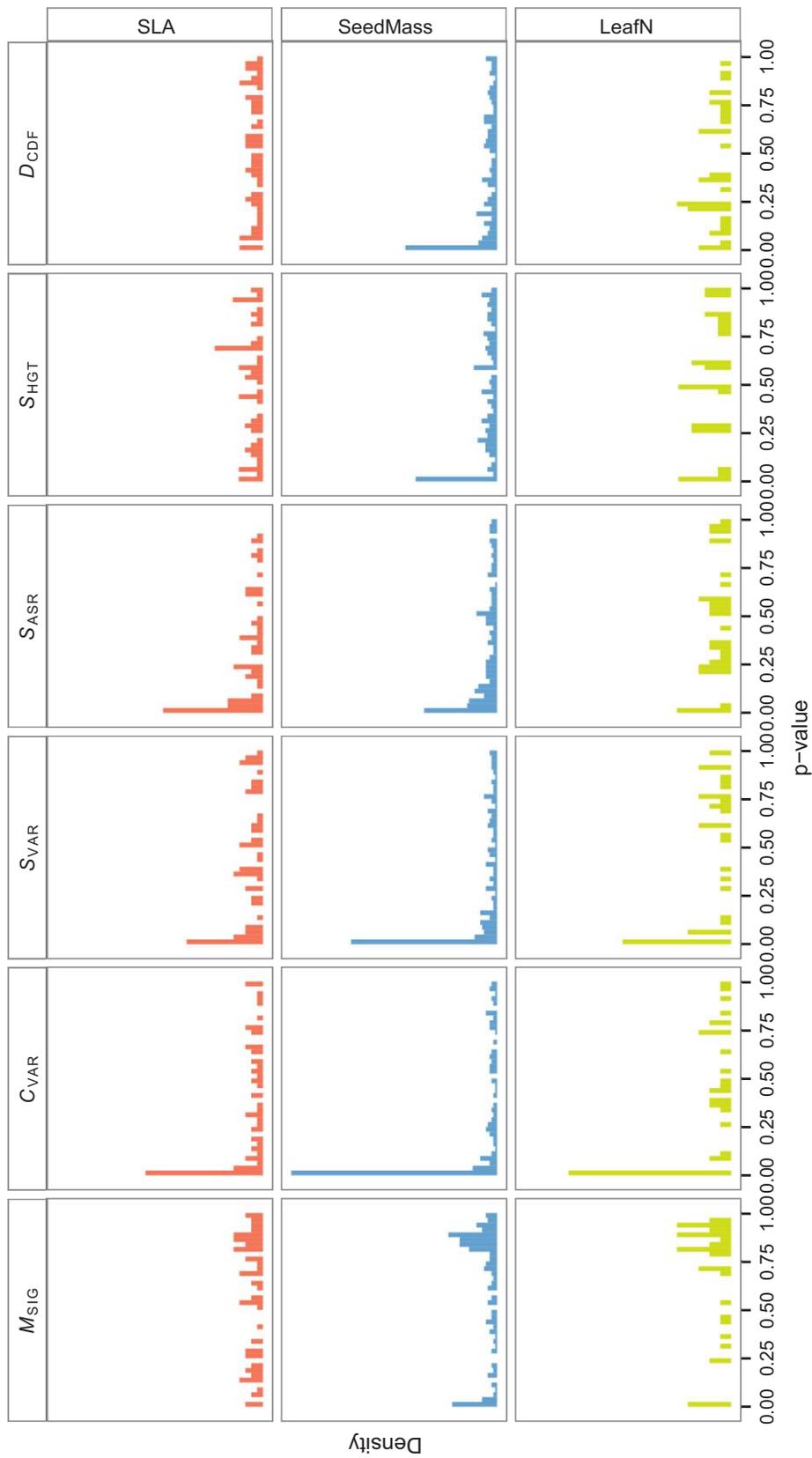
**Figure 3:** The relative support, as measured by Akaike information criterion (AIC) weight, for the three models used in our study (Brownian motion [BM], Ornstein-Uhlenbeck [OU], and early burst [EB]) across all 337 data sets. An OU model is highly supported for a majority of the data sets.

We limit our analyses of model adequacy to only the most highly supported model in the candidate set, as supported by AIC. We did this to present a best-case scenario; if a model had very little relative support, it would be unremarkable if it also had poor adequacy (but see Ripplinger and Sullivan 2010). Even considering only the best of the set, in general, the data sets often deviated from the expectations of the model in at least some ways (fig. 4). Of the 72 comparative data sets of SLA, we detected deviations from the expectations in 32 data sets (using a cutoff of  $P = .05$ ), 22 by at least two, and 15 by three or more. Results were similar in the seed mass data (of the 226 seed mass data sets, we detected deviations in 153 data sets with at least one test statistic, 95 by at least two, and 65 by three or more) and leaf nitrogen content (of the 39 data sets, we detected deviations in 19 by at least one, 12 by at least two, and eight by three or more test statistics).

Some test statistics were much more likely to detect model violations than others. In 163 cases  $C_{VAR}$  revealed that the data deviated significantly from the expectations of the best model. In 118 cases,  $S_{VAR}$  did. The rate of deviation was somewhat lower for the other test statistics ( $M_{SIG}$ , 39;  $S_{ASR}$ , 84;  $S_{HGT}$ , 54;  $D_{CDF}$ , 67).

For 133 of the 337 data sets, we did not detect any deviations from the expectations of the best-supported model. As stated above, the numbers of models that showed deviations with at least one test statistic may be somewhat overinflated. However, the proportion of clades in which  $P$  values were less than 0.05 is much, much greater than the error rates we found in our simulations. And the proportions for each individual test statistic are much higher than would be expected by chance.

As the subclades are not independent (overlapping sets of taxa are present in family, order, and time slice phylogenies), conventional statistics, such as linear regression, are not straightforward to apply across data sets. Nonetheless, the trend is clear: the larger the phylogeny is, the more likely OU is to be highly supported and the stronger the evidence is that the model is inadequate. There is a strong relationship between the size of a subclade and the overall distance between observed and simulated test statistics, as measured by the Mahalanobis distance (fig. S4). While it is likely that evolutionary rates and processes are more heterogeneous when one considers larger clades, it is also true that violations from model expectations are more easily detected when considering more data: the more contrasts that are



**Figure 4:** The distribution of  $P$  values for our six test statistics over all 337 data sets in our study after fitting the models using maximum likelihood. The  $P$  values are from applying our model adequacy approach to the best supported of the three models (as evaluated with Akaike information criterion). Many of the data sets deviate from the expectations under the best model along a variety of axes of variation. Deviations are particularly common for the coefficient of variation  $C_{VAR}$  and the slope of the contrasts against their expected variances  $S_{VAR}$ .  $SLA$  = specific leaf area.

examined, the lower the variance in the distribution of simulated test statistics, and, therefore, we are more likely to detect model misspecification (see “Discussion”). We also note here that if the fitted model was identical to the “true” model of trait evolution, we would not expect a relationship between the Mahalanobis distance and the size of the phylogeny.

## Discussion

### *Why Does Model Adequacy Matter?*

Whatever inferences we want to make from comparative data—for example, characterizing broadscale patterns of evolution through time, investigating correlations between characters, or testing hypotheses about the processes that have driven trait evolution over macroevolutionary time—it is important that our chosen statistical model captures variation in the data relevant to the question being addressed. If, for example, the goal is to assess variation in macroevolutionary rates over time, it is essential that the model does a good job of explaining temporal heterogeneity. If we want to know about the slope of an evolutionary allometric relationship, we need a model that provides a meaningful estimate of this parameter (Hansen and Bartoszek 2012). Comparing the fit of a model to a set of alternatives (using likelihood ratio tests, information-theoretic metrics, Bayes factors, etc.) can allow for only a relative assessment of the suitability of the model for the task. Such a model comparison approach does not provide any information about whether a model will allow us to actually get at the question we are interested in.

The flip side of this is that tests of model adequacy, such as ours, are designed to measure the absolute fit but not the absolute appropriateness of the model. We know that all of the models used in comparative biology are wrong. Whether they are useful will depend on the question being addressed. We are far from the first to suggest that model adequacy is important to consider when using comparative methods (see, e.g., Felsenstein 1985, 1988, 2012; Harvey and Pagel 1991; Garland et al. 1992, 1999; Díaz-Uriarte and Garland 1996; Hansen and Martins 1996; Price 1997; Garland and Ives 2000; Hansen and Orzack 2005; Boettiger et al. 2012; Hansen and Bartoszek 2012; Slater and Pennell 2014; Beaulieu et al. 2013; Blackmon and Demuth 2014). The contribution of our article is to generalize many of these previous approaches into a single, flexible statistical framework.

Again, we emphasize that simply because a data set deviates from the expectations of the model does not imply that the model should necessarily be rejected. In our analyses of model adequacy across the 337 angiosperm clades, we were focused on whether the model was suitable for measuring rates of evolution, which is dependent on the

model being a good one (Hunt 2012). For other questions, the fact that a model fails to capture some aspects of the variation in the data may not be that important. For example, if our question was that of Harmon et al. (2010)—are early bursts of evolution common in macroevolution?—we could conclude with good certainty that they are not. Our data sets may not be well described by an OU model, but they are certainly nothing like what we would expect under an EB scenario. Likewise, if we are interested primarily in whether there is a pattern of correlation between two traits, the fact that the model we used is not adequately describing much of the variation will in many cases not greatly impact the qualitative conclusions. A nuanced view of model adequacy is particularly important when analyzing large phylogenies: the more data we consider, the greater our ability to detect subtle deviations from model expectations (figs. S4, S7). Focusing only on the test statistic  $P$  values may lead us to reject models that are actually reasonably suitable for addressing our question of interest.

However, we view the most interesting cases to be where the best model does not adequately describe the variation of interest. The way in which a model fails can provide a richer understanding of our data and the processes that have driven the patterns we observe (Gelman and Shalizi 2013). First, model inadequacy can point to problems in the data. We suspect that this is likely a common cause of poor model fit. For the empirical analyses, we used a very large phylogeny of angiosperms that was constructed to test specific global-scale biodiversity questions (Zanne et al. 2014). We recognize that the tree is poorly resolved in many places (particularly near the tips) and is likely ill suited for addressing more detailed, clade-specific questions (see the recent critique by Donoghue and Edwards 2014). Specifically, the inaccurate placement of species will, on average, cause evolutionary rates to be inflated, which is precisely what we find (see below). However, we emphasize that phylogenetic error is likely ubiquitous, and this problem is certainly not limited to the tree we used. Likewise, the data set we assembled is rather heterogeneous in terms of quality; the data were originally collected for a diverse set of reasons, and some groups have been measured much more carefully than others. And while we have done our best to clean the data, errors undoubtedly remain.

Second, and most excitingly, the failure of a model to adequately describe relevant aspects of the data can provide insight into the processes we have failed to consider in our model (Gelman and Shalizi 2013). For example, if a model fails to capture variation relative to time (evaluated by the test statistic  $S_{\text{HGT}}$ ), this suggests that temporal heterogeneity has been greater than we allowed for. The causes of such heterogeneity have long been a topic of interest in macroevolutionary studies (e.g., Simpson 1944; Foote 1997), and there has been a great deal of recent development to-

ward more complex rate-varying models (e.g., O'Meara et al. 2006; Thomas et al. 2006; Eastman et al. 2011; Weir and Mursleen 2013; Rabosky et al. 2014). Likewise, failure to adequately describe variation across the clade may indicate that the existence of multiple macroevolutionary optima (sensu Hansen 2012) are driving the dynamics of traits over time (for models that have been used to capture these dynamics, see Hansen 1997; Butler and King 2004; Beaulieu et al. 2012; Ingram and Mahler 2013; Uyeda and Harmon 2014).

Model inadequacy may also suggest types of models that have not previously been considered. For example, if recently diverged species tend to be more dissimilar than can be accounted for under a simple diffusion model such as BM or OU, this may be the result of character displacement. However, almost no phylogenetic models have been put forth that explicitly model interactions between lineages (but see Nusimer and Harmon 2015). Or if traits have lower variance than expected under an OU process, this may be the result of hard bounds. Boucher et al. (2014) recently argued that this is the case for climatic niches and that alternative models need to be developed for this case. Of course, a researcher may discover that his or her data set is poorly described by all of the currently available models. Aside from deriving new models specific to his or her question and data set, the researcher should at least carefully examine the extent to which model misspecification is likely to affect the major conclusions and proceed forward with due caution.

#### *Implications for Empirical Studies*

In our analysis of angiosperm functional traits, we found common macroevolutionary models to often be poor descriptors for the patterns of variation and likely inadequate for estimating evolutionary rates. While there are certainly a number of important caveats to our analysis (discussed above), the overall trends are clear. This should certainly give researchers some pause about the models routinely used in our field—especially as they are often used in a model comparison framework to evaluate the “tempo and mode” of macroevolution. We argue that our results strongly suggest that we may often be missing a large part of the story.

The 337 comparative data sets we analyzed varied in terms of traits, size, age, and placement in the angiosperm phylogeny. Nonetheless, several general patterns emerge. An OU model was, by and large, the most supported of the three we examined. In an analysis of 67 comparative data sets consisting of size and shape data from a variety of animal taxa, Harmon et al. (2010) also found substantial support for OU models, though for their data sets, BM was more commonly chosen by AIC. (We note, however, that many of their data sets were quite small; see Slater and Pennell

2014.) Since their article, a substantial number of studies conducted in a diverse array of groups have also found OU models to be preferred over BM models (e.g., Burbrink et al. 2012; López-Fernández et al. 2013; Quintero and Wiens 2013; Thomas et al. 2014).

The tendency of OU to explain data better than BM has inspired diverse process-based explanations, including stabilizing selection, evolutionary constraints, and the presence of “adaptive zones” (Hansen and Martins 1996; Butler and King 2004; Hansen 2012; Pennell and Harmon 2013). If the widespread support for OU models was indeed caused by the biological processes that have been proposed, we would expect that an OU model would also be widely adequate. However, this is not what we found. The data sets deviated significantly from the distributions expected under OU models, most often detected with  $C_{VAR}$  and  $S_{VAR}$  but frequently with others as well. OU models often failed to capture other important types of heterogeneity—variation with respect to rate variation ( $M_{SIG}$ ), trait values ( $S_{ASR}$ ), and time ( $S_{HGT}$ ). Additionally, a substantial number of data sets were not well modeled by a multivariate normal distribution ( $D_{CDF}$ ). These results suggest a statistical explanation for the high support for OU models. OU predicts higher variance near the tips of the phylogeny than do BM or EB models (see fig. 1 in Harmon et al. 2010).

Heterogeneous evolutionary processes, phylogenetic misestimation, and measurement error could also produce such a pattern. In light of our results from model adequacy, it seems likely that OU is often supported because it is able to accommodate more “slop” (phylogenetic and trait error in addition to model misspecification) than the other models. This is not to say that the processes captured by OU models are unimportant in macroevolution but rather that OU models may be favored for reasons that are more statistical than biological. Future, and hopefully more widely adequate, models of trait evolution could be developed that include aspects of the OU model, especially the bounds on trait values, while incorporating additional biological realism (for a recent example of such a model, see Nusimer and Harmon 2015).

The way in which the observed test statistics deviate from the simulated values also supports the claim that the widespread support for OU is largely a statistical artifact. Model violations were most frequently detected by the variance estimate,  $C_{VAR}$ . If the evolutionary process (or, alternatively, phylogenetic/measurement error) is heterogeneous across the tree, the lineages in some parts of the clade will be much more divergent than in others. The only way for the model to account for the highly divergent groups is to estimate a large  $\sigma^2$  (and/or a small  $\alpha$  parameter for the OU model). The unit tree formed by these parameter estimates will have long branches across the entire tree. In

the less divergent parts of the tree, the contrasts calculated on this unit tree will be small, relative to what we expect under BM. So perhaps counterintuitively, when heterogeneity in processes across taxa causes the estimated global rates of divergence to be inflated, this results in a higher value for  $C_{VAR}$ .

The second major take-home from the empirical analyses is that error, in both trait values and phylogenies, can have serious consequences for model adequacy. We frequently detected deviations from model expectations with  $S_{VAR}$ , the slope between the contrasts and their expected variances. This indicates that the rate of evolution appears to be varying with regard to branch length over which it is measured. This seems unlikely to be attributable to any biological process; it is far more probable that this reflects phylogenetic error (particularly branch length error). Above, we outlined some of the deficiencies of the data sets we used in this article but argue that these are likely to be widespread in comparative data. The test statistics outlined above can serve as useful diagnostics to aid researchers in identifying outliers that may be driving the pattern. We recommend that researchers faced with an inadequate model plot the magnitude of the contrasts onto the unit tree; this will usually be much more informative with regard to the model fit than plotting the magnitude of the contrasts on the original phylogeny. Exceptionally large or small contrasts on the unit tree can provide clues as to where the data may be erroneous. If phylogenetic error were causing poor model fits, we would predict that many of the anomalous contrasts would occur in parts of the tree that are poorly supported.

#### *Extensions of Our Approach*

There are a number of additional ways our approach could be extended. First, we have considered only a limited set of test statistics. We chose them because each of these has a clear statistical expectation and observed deviations from them have intuitive biological explanations. However, they are certainly a subset of all possible test statistics that could be applied. For example, because contrasts are i.i.d., there should be no autocorrelation between neighboring contrasts; the test statistics could be expanded to detect nonzero autocorrelation. Second, as stated above, our approach can be applied equally well to phylogenetic regression models, such as phylogenetic generalized least squares models (Grafen 1989; Martins and Hansen 1997) or phylogenetic mixed models (Lynch 1991; Housworth et al. 2004; Hadfield and Nakagawa 2010), where concerns regarding model adequacy are just as pertinent (Hansen and Bartoszek 2012). While our approach can be used to assess the adequacy of the phylogenetic component of regression models “out of the box,” additional steps are required to assess the adequacy of the

linear component. Third, our method was designed for quantitative trait models that assume that data can be modeled with a multivariate normal distribution. We need general model adequacy approaches for other types of traits, such as discrete traits (i.e., binary, multistate, ordinal; for recent discussions of this see Beaulieu et al. 2013; Blackmon and Demuth 2014; Maddison and FitzJohn 2015), traits that influence speciation rates (Maddison et al. 2007; FitzJohn 2010; Rabosky and Goldberg 2015), and quantitative trait models that do not predict a multivariate normal distribution of traits (Landis et al. 2013; Schraiber and Landis 2015).

It may also be possible to extend our approach with an eye toward model selection. Slater and Pennell (2014) developed their posterior predictive simulation approach (which is related to our method) to distinguish between a BM model and one where rates of evolution decreased through time. They chose test statistics specifically to address this question. Slater and Pennell found using posterior predictive fit as a model selection criterion to be much more powerful than comparing models using AIC or likelihood ratio tests, particularly when “outlier taxa” (lineages where the pattern of evolution deviates from the overall model) were included in the analysis. The logic of Slater and Pennell could be extended to other scenarios; to test some evolutionary hypotheses, we may care a lot about whether a model explains variation along some axes but be less concerned about others. This is a question-specific approach to model selection and has been developed in the context of molecular phylogenetics (Bollback 2002; Lewis et al. 2014). This is also the essence of the decision-theoretic approach to model selection (Robert 2007), which has also been well used in phylogenetics (Minin et al. 2003) but has not previously been considered in PCMs.

#### **Arbutus**

We have implemented our approach in a new R package *arbutus*. It is available on github (<https://github.com/mwpennell/arbutus>). For this project, we have also adopted code from the *ape* (Paradis et al. 2004), *geiger* (Pennell et al. 2014), and *diversitree* (FitzJohn 2012) libraries. We have written functions to parse the output of a number of different programs for fitting trait evolution models (see the *arbutus* website for an up-to-date list of supported models and packages). As this approach was developed to be general, we have written the code in such a way that users can include their own test statistics and trait models in the analyses.

#### **Concluding Remarks**

Attempts to assess the adequacy of phylogenetic models are almost as old as modern comparative phylogenetic biology. In the 1980s and 1990s much discussion surrounded

the appropriateness of various methods and models (Felsenstein 1985, 1988; Harvey and Pagel 1991; Garland et al. 1992, 1999; Díaz-Uriarte and Garland 1996; Price 1997; Garland and Ives 2000). We argue that this discussion is key to progressing in our field. This is not simply because we are concerned that many inferences may not be robust to model violations. Rather, we believe that considering model adequacy can help suggest new ways of thinking about how to combine data and models to test macroevolutionary hypotheses.

### Acknowledgments

We would like to thank the members of the Tempo and Mode of Trait Evolution Working Group at the National Evolutionary Synthesis Center (NESCent) for their suggestions and encouragement. We thank D. Caetano, A. Mooers, S. Otto, W. Pearse, and J. Uyeda for their thoughtful comments on the manuscript. C. Ané, J. Brown, P. Joyce, and G. Slater also provided valuable insights into this project. This manuscript greatly benefitted from the comments of T. Day, S. Stepan, and two anonymous reviewers. Last, we are grateful to the researchers who made their data available; this project would not have been possible without them. M.W.P. was supported by a NESCent graduate fellowship and a Natural Sciences and Engineering Research Council of Canada postgraduate fellowship. This work was also supported by National Science Foundation grants awarded to L.J.H. (DEB 0919499 and 1208912).

### Literature Cited

- Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19:716–723.
- Ancombe, F. J. 1973. Graphs in statistical analysis. *American Statistician* 27:17–21.
- Beaulieu, J. M., D.-C. Jhwueng, C. Boettiger, and B. C. O'Meara. 2012. Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution* 66:2369–2383.
- Beaulieu, J. M., B. C. O'Meara, and M. J. Donoghue. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. *Systematic Biology* 62:725–737.
- Blackmon, H., and J. P. Demuth. 2014. Estimating tempo and mode of Y chromosome turnover: explaining Y chromosome loss with the fragile Y hypothesis. *Genetics* 197:561–572.
- Blomberg, S. P., T. Garland Jr., and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57:717–745.
- Blomberg, S. P., J. G. Lefevre, J. A. Wells, and M. Waterhouse. 2012. Independent contrasts and PGLS regression estimators are equivalent. *Systematic Biology* 61:382–391.
- Boettiger, C., G. Coop, and P. Ralph. 2012. Is your phylogeny informative? measuring the power of comparative methods. *Evolution* 66:2240–2251.
- Bollback, J. P. 2002. Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution* 19:1171–1180.
- Boucher, F. C., W. Thuiller, T. J. Davies, and S. Lavergne. 2014. Neutral biogeography and the evolution of climatic niches. *American Naturalist* 183:573–584.
- Brown, J. M. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. *Systematic Biology* 63:334–348.
- Burbrink, F. T., X. Chen, E. A. Myers, M. C. Brandley, and R. A. Pyron. 2012. Evidence for determinism in species diversification and contingency in phenotypic evolution during adaptive radiation. *Proceedings of the Royal Society B: Biological Sciences* 279:4817–4826.
- Burnham, K., and D. Anderson. 2004. *Model selection and multi-model inference: a practical information-theoretic approach*. Springer, New York.
- Butler, M. A., and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *American Naturalist* 164:683–695.
- Cornwell, W. K., M. Westoby, D. S. Falster, R. G. FitzJohn, B. C. O'Meara, M. W. Pennell, D. J. McGlenn, et al. 2014. Functional distinctiveness of major plant lineages. *Journal of Ecology* 102:345–356.
- Díaz-Uriarte, R., and T. Garland Jr. 1996. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian motion. *Systematic Biology* 45:27–47.
- Donoghue, M. J., and E. J. Edwards. 2014. Biome shifts and niche evolution in plants. *Annual Reviews of Ecology, Evolution, and Systematics* 45:547–572.
- Eastman, J. M., M. E. Alfaro, P. Joyce, A. L. Hipp, and L. J. Harmon. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution* 65:3578–3589.
- Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics* 25:471–492.
- . 1985. Phylogenies and the comparative method. *American Naturalist* 125:1–15.
- . 1988. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics* 19:445–471.
- . 2012. A comparative method for both discrete and continuous characters using the threshold model. *American Naturalist* 179:145–156.
- FitzJohn, R. G. 2010. Quantitative traits and diversification. *Systematic Biology* 59:619–633.
- . 2012. Diversitree: comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution* 3:1084–1092.
- Foote, M. 1997. The evolution of morphological diversity. *Annual Review of Ecology and Systematics* 28:129–152.
- Freckleton, R. P., and P. H. Harvey. 2006. Detecting non-Brownian trait evolution in adaptive radiations. *PLoS Biology* 4:e373.
- Garland, T., Jr., A. W. Dickerman, C. M. Janis, and J. A. Jones. 1993. Phylogenetic analysis of covariance by computer simulation. *Systematic Biology* 42:265–292.
- Garland, T., Jr., P. H. Harvey, and A. R. Ives. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Systematic Biology* 41:18–32.
- Garland, T., Jr., and A. R. Ives. 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *American Naturalist* 155:346–364.

- Garland, T., Jr., P. E. Midford, and A. R. Ives. 1999. An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. *American Zoologist* 39:374–388.
- Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1:515–534.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2003. *Bayesian data analysis*. 2nd ed. Chapman & Hall/CRC, Boca Raton, FL.
- Gelman, A., X. Meng, and H. Stern. 1996. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* 6:733–807.
- Gelman, A., and C. R. Shalizi. 2013. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology* 66:8–38.
- Goldman, N. 1993. Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* 36:182–198.
- Grafen, A. 1989. The phylogenetic regression. *Philosophical Transactions of the Royal Society B: Biological Sciences* 326:119–157.
- Hadfield, J. D., and S. Nakagawa. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology* 23:494–508.
- Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution* 51:1341–1351.
- . 2012. Adaptive landscapes and macroevolutionary dynamics. Pages 205–221 in E. Svensson and R. Calsbeek, eds. *The adaptive landscape in evolutionary biology*. Oxford University Press, Oxford.
- Hansen, T. F., and K. Bartoszek. 2012. Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. *Systematic Biology* 61:413–425.
- Hansen, T. F., and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution* 50:1404–1417.
- Hansen, T. F., and S. H. Orzack. 2005. Assessing current adaptation and phylogenetic inertia as explanations of trait evolution: the need for controlled comparisons. *Evolution* 59:2063–2072.
- Harmon, L. J., J. B. Losos, T. Jonathan Davies, R. G. Gillespie, J. L. Gittleman, W. Bryan Jennings, K. H. Kozak, et al. 2010. Early bursts of body size and shape evolution are rare in comparative data. *Evolution* 64:2385–2396.
- Harvey, P. H., and M. D. Pagel. 1991. *The comparative method in evolutionary biology*. Oxford University Press, Oxford.
- Ho, L. S. T., and C. Ané. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology* 63:397–408.
- Houle, D., C. Pelabon, G. P. Wagner, and T. F. Hansen. 2011. Measurement and meaning in biology. *Quarterly Review of Biology* 86:3–34.
- Housworth, E. A., E. P. Martins, and M. Lynch. 2004. The phylogenetic mixed model. *American Naturalist* 163:84–96.
- Hunt, G. 2012. Measuring rates of phenotypic evolution and the inseparability of tempo and mode. *Paleobiology* 38:351–373.
- Ingram, T., and D. L. Mahler. 2013. Surface: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise AIC. *Methods in Ecology and Evolution* 4:416–425.
- Joyce, P., and P. Majoram. 2008. Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology* 7:1–16.
- Kleyer, M., R. Bekker, I. Knevel, J. Bakker, K. Thompson, M. Sonnenschein, P. Poschlod, et al. 2008. The LEDA traitbase: a database of life-history traits of the northwest European flora. *Journal of Ecology* 96:1266–1274.
- Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. *Evolution* 30:314–334.
- Landis, M. J., J. G. Schraiber, and M. Liang. 2013. Phylogenetic analysis using Lévy processes: finding jumps in the evolution of continuous traits. *Systematic Biology* 62:193–204.
- Leishman, M. R., I. J. Wright, A. T. Moles, and M. Westoby. 2000. The evolutionary ecology of seed size. Pages 31–57 in M. Fenner, ed. *Seeds: the ecology of regeneration in plant communities*. CABI, Wallingford.
- Lewis, P. O., W. Xie, M.-H. Chen, Y. Fan, and L. Kuo. 2014. Posterior predictive Bayesian phylogenetic model selection. *Systematic Biology* 63:309–321.
- López-Fernández, H., J. H. Arbour, K. O. Winemiller, and R. L. Honeycutt. 2013. Testing for ancient adaptive radiations in Neotropical cichlid fishes. *Evolution* 67:1321–1337.
- Lynch, M. 1990. The rate of morphological evolution in mammals from the standpoint of the neutral expectation. *American Naturalist* 136:727–741.
- . 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution* 45:1065–1080.
- Lynch, M., and W. G. Hill. 1986. Phenotypic evolution by neutral mutation. *Evolution* 40:915–935.
- Maddison, W. P., and R. G. FitzJohn. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Systematic Biology* 64:127–136.
- Maddison, W. P., P. E. Midford, and S. P. Otto. 2007. Estimating a binary character's effect on speciation and extinction. *Systematic Biology* 56:701–710.
- Martins, E. P., and T. F. Hansen. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *American Naturalist* 149:646–667.
- Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. *Systematic Biology* 52:674–683.
- Mooers, A. Ø., S. M. Vamosi, and D. Schluter. 1999. Using phylogenies to test macroevolutionary hypotheses of trait evolution in cranes (Gruinae). *American Naturalist* 154:249–259.
- Neal, R. M. 2003. Slice sampling. *Annals of Statistics* 31:705–741.
- Nusimer, S. L., and L. J. Harmon. 2015. Predicting rates of interspecific interaction from phylogenetic trees. *Ecology Letters* 18:17–27.
- O'Meara, B. C. 2012. Evolutionary inferences from phylogenies: a review of methods. *Annual Review of Ecology, Evolution, and Systematics* 43:267–285.
- O'Meara, B. C., C. Ané, M. J. Sanderson, and P. C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. *Evolution* 60:922–933.
- Pagel, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society B: Biological Sciences* 255:37–45.
- . 1997. Inferring evolutionary processes from phylogenies. *Zoologica Scripta* 26:331–348.
- . 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877–884.

- Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20: 289–290.
- Pennell, M. W., J. M. Eastman, G. J. Slater, J. W. Brown, J. C. Uyeda, R. G. FitzJohn, M. E. Alfaro, and L. J. Harmon. 2014. Geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* 15:2216–2218.
- Pennell, M. W., and L. J. Harmon. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Annals of the New York Academy of Sciences* 1289:90–105.
- Pennell, M. W., L. J. Harmon, and J. C. Uyeda. 2013. Is there room for punctuated equilibrium in macroevolution? *Trends in Ecology and Evolution* 29:23–32.
- Price, T. 1997. Correlated evolution and independent contrasts. *Philosophical Transactions of the Royal Society B: Biological Sciences* 352:519–529.
- Purvis, A., and A. Rambaut. 1995. Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. *Computer Applications in the Biosciences* 11:247–251.
- Quintero, I., and J. J. Wiens. 2013. Rates of projected climate change dramatically exceed past rates of climatic niche evolution among vertebrate species. *Ecology Letters* 16:1095–1103.
- R Development Core Team. 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.
- Rabosky, D. L., S. C. Donnellan, M. Grundler, and I. J. Lovette. 2014. Analysis and visualization of complex macroevolutionary dynamics: an example from Australian scincid lizards. *Systematic Biology* 63:610–627.
- Rabosky, D. L., and E. E. Goldberg. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. *Systematic Biology* 64:340–355.
- Reid, N. M., S. M. Hird, J. M. Brown, T. A. Pelletier, J. D. McVay, J. D. Satler, and B. C. Carstens. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical data. *Systematic Biology* 63:322–333.
- Revell, L. J., D. L. Mahler, P. R. Peres-Neto, and B. D. Redelings. 2012. A new phylogenetic method for identifying exceptional phenotypic diversification. *Evolution* 66:135–146.
- Ripplinger, J., and J. Sullivan. 2010. Assessment of substitution model adequacy using frequentist and Bayesian methods. *Molecular Biology and Evolution* 27:2790–2803.
- Robert, C. P. 2007. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. 2nd ed. Springer, New York.
- Rohlf, F. J. 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. *Evolution* 55:2143–2160.
- Royal Botanical Gardens, Kew. 2014. Seed information database (SID). Version 7.1. <http://data.kew.org/sid>.
- Rubin, D. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* 12:1151–1172.
- Schraiber, J. G., and M. J. Landis. 2015. Sensitivity of quantitative traits to mutational effects and number of loci. *Theoretical Population Biology* 102:85–93.
- Simpson, G. G. 1944. *Tempo and mode of evolution*. Columbia University Press, New York.
- Slater, G. J. 2013. Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the Cretaceous-Palaeogene boundary. *Methods in Ecology and Evolution* 4:734–744.
- . 2014. Correction to “Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the Cretaceous-Palaeogene boundary” and a note on fitting macroevolutionary models to comparative paleontological data sets. *Methods in Ecology and Evolution* 5:714–718.
- Slater, G. J., L. J. Harmon, and M. E. Alfaro. 2012. Integrating fossils with molecular phylogenies improves inference of trait evolution. *Evolution* 66:3931–3944.
- Slater, G. J., and M. W. Pennell. 2014. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. *Systematic Biology* 63:293–308.
- Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B: Statistical Methodology* 64:583–639.
- The Plant List. 2014. Version 1.1. <http://www.theplantlist.org>.
- Thomas, G. H., N. Cooper, C. Venditti, A. Meade, and R. P. Freckleton. 2014. Bias and measurement error in comparative analyses: a case study with the Ornstein-Uhlenbeck model. *BioRxiv*. doi:10.1101/004036.
- Thomas, G. H., and R. P. Freckleton. 2012. Motmot: models of trait macroevolution on trees. *Methods in Ecology and Evolution* 3: 145–151.
- Thomas, G. H., R. P. Freckleton, and T. Székely. 2006. Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. *Proceedings of the Royal Society B: Biological Sciences* 273:1619–1624.
- Uyeda, J. C., T. F. Hansen, S. J. Arnold, and J. Pienaar. 2011. The million-year wait for macroevolutionary bursts. *Proceedings of the National Academy of Sciences of the USA* 108:15908–15913.
- Uyeda, J. C., and L. J. Harmon. 2014. Bayesian reversible-jump modeling of adaptive shifts for studying macroevolutionary patterns of trait evolution. *Systematic Biology* 63:902–918. doi:10.1093/sysbio/syu057.
- Weir, J. T., and S. Mursleen. 2013. Diversity-dependent cladogenesis and trait evolution in the adaptive radiation of the auks (Aves: Alcidae). *Evolution* 67:403–416.
- Westoby, M., D. S. Falster, A. T. Moles, P. A. Vesk, and I. J. Wright. 2002. Plant ecological strategies: some leading dimensions of variation between species. *Annual Review of Ecology and Systematics* 33:125–159.
- Wright, I. J., P. B. Reich, M. Westoby, D. D. Ackerly, Z. Baruch, F. Bongers, J. Cavender-Bares, et al. 2004. The worldwide leaf economics spectrum. *Nature* 428:821–827.
- Zanne, A. E., D. C. Tank, W. K. Cornwell, J. M. Eastman, S. A. Smith, R. G. FitzJohn, D. J. McGlenn, et al. 2013. Data from: Three keys to the radiation of angiosperms into freezing environments. *Nature, Dryad Digital Repository*. <http://dx.doi.org/10.5061/dryad.63q27/3>.
- . 2014. Three keys to the radiation of angiosperms into freezing environments. *Nature* 506:89–92.

Associate Editor: Scott J. Stepan  
Editor: Troy Day