Article

Effect of Polytomy on the Parameter Estimation and Goodness of Fit of Phylogenetic Linear Regression Models for Trait Evolution

Dwueng-Chwuan Jhwueng * and **Feng-Chi Liu**

Department of Statistics, Feng Chia University, Taichung 40724, Taiwan
* Correspondence: dcjhwueng@fcu.edu.tw

Abstract: Phylogenetic regression models for trait evolution incorporate phylogenetic trees for the analysis of comparative data, in order to study trait relationships among a group of related species. However, as such trees are estimated, it is unlikely that there will be no errors when describing the relationships among species. In particular, for polytomy trees, where the relationships within a particular clade of species are more/less certainly determined (hard/soft polytomy, respectively), results of comparative analyses obtained from models based on those phylogenetic trees may also be affected. In this study, through extensive simulations, the performances of several popular Gaussian process-based regression models (Brownian motion, BM; Ornstein–Uhlenbeck process, OU; early burst, EB), as well as branch-stretching models (Pagel's λ, δ, κ), were evaluated by assessing their fit and parameter estimation performance when soft polytomies are presented on either the root or a clade with insufficient phylogenetic information. Comparisons of the models are conducted by either assessing the accuracy of the estimator of regression and model parameters, or using a measure of fit (AIC, r^2, and mean square error). It is found that, although polytomy does not significantly impact the fit and parameter estimate within a specified model, distinguishable differences and effects may be observed among trees and models. In particular, Pagel λ model and the OU model yield more accurate estimates and provide better fitting effects than the other models (BM, EB, δ, κ). While correcting phylogeny is an essential step prior to analysis, users may also consider using more appropriate models when encountering the polytomy issue.

Keywords: phylogenetic comparative analysis; regression analysis; polytomy; Gaussian process; trait evolution

check for updates

Citation: Jhwueng, D.-C.; Liu, F.-C. Effect of Polytomy on the Parameter Estimation and Goodness of Fit of Phylogenetic Linear Regression Models for Trait Evolution. *Diversity* **2022**, *14*, 942. https://doi.org/10.3390/d14110942

Academic Editor: Márta Ladányi

Received: 20 September 2022
Accepted: 28 October 2022
Published: 2 November 2022

1. Introduction

Regression analysis has been broadly applied in the study of evolutionary relationships, through the use of comparative data and the phylogenetic tree [1–3]. A phylogenetic tree is a directed diagram, where the topology and branch lengths represent the magnitudes of evolutionary relationships among species [4]; for example, the magnitude of the relationship between a pair of taxa on the tree can be defined by their shared branch length, measured from the root of the tree to their most common ancestor. A phylogenetic tree can be constructed using various methods, and relevant software has been developed [5,6]. Comparative data are trait values collected from field studies, where scientists spend tremendous time, energy, and funding to search, measure, and record the characteristics of species. The phenotype, defined as the observable physical properties of an organism, for a group of species (e.g., the body mass, body length, or head shape) are usually recorded as quantitative values. These phenotypic trait values are used for studying research questions arising from ecology [7], evolutionary biology [8], paleontology [9], toxicology [10], and so on. For instance, comparative data can be used to answer questions, such as how to estimate ancestral status [11,12], how to calculate the speed of evolution [13,14], or how to compute the diversification rate [15,16] (see [1,2,17,18] for more comprehensive reviews).

To address these evolutionary questions, scientists have applied phylogenetic comparative methods (PCMs) to analyze comparative data. To study the relationships between traits, regression analysis is commonly applied to address questions, such as how the possession of a trait X influences the evolution of trait Y [19], the association between body mass and brain mass [20], and whether genome size increases with the water depth in marine fishes [21].

Thanks to current technology and the tremendous efforts of scientists in field studies, many online databases for accessing trees and traits are now available. For example, these important trait databases include AmphibiaWeb [22], The Reptile Database [23], Global Ants Database (GLAD) [24], and FishBase [25]; while important tree databases include the TreeFam database of animal gene trees [26], the TRY Plant Trait database [27], PHYLOtastic [28], and The Tree of Life Web (ToL) [29], among others [30–34] (see Appendix A.2 for a more detailed description). Such databases are also beneficial to scientists (ecologists, evolutionary biologists, paleontologist, statisticians, and so on), in terms of re-evaluating their research questions by applying/developing statistical methods and procedures to analyze those data sets. Note that comparative data are recorded as either categorical (nominal, ordinal) or quantitative (interval or ratio) values, and may be displayed by graphical visualization, in order to better understand the distribution of the data set [35]. The reader can also refer to the public free distributed software **R** [36], for which scientists have created a task view that describes a collection of **R** packages [37] implementing a variety of different comparative phylogenetic methods for analyzing historical patterns along phylogenetic trees.

As it has been widely accepted that all species share a certain evolutionary history, the relationships among species can be described by a rooted phylogenetic tree T_n comprised of n taxa. In fact, a tree T_n can be transformed into an $n \times n$ square matrix $\boldsymbol{G}_n = [g_{ij}]$, where each element of g_{ij} represents the magnitude of the affinity between a pair of species i and j [38]. An example of a five-taxa tree T_5 mapped to a 5×5 matrix $\boldsymbol{G}_5$ is shown in Figure 1.

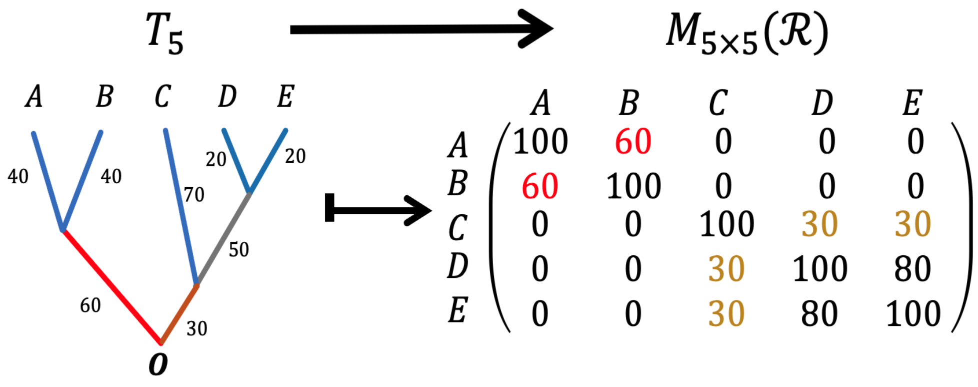

Figure 1. A rooted phylogenetic tree comprised of five taxa T_5 and its corresponding affinity matrix $\boldsymbol{G}_5$. Left panel: A rooted phylogenetic tree of five extant species A, B, C, D, E. Each branch of the tree has a length that represents the evolutionary time. Right panel: The affinity matrix $\boldsymbol{G}_5$, where the element g_{ij} represents the amount of shared branch length from the root to the most common ancestor. The diagonal elements of $\boldsymbol{G}_5$ have values of 100, as the distance from the root O to each tip A, B, C, D, E is 100. Each off-diagonal element g_{ij} is the sum of the shared branch length between a pair of tips i and j. For instance, species A and species B have a shared branch length of 60 (red branch); hence, $g_{AB} = g_{BA} = 60$; while species C and species E share a branch of length 30 (brown branch), such that $g_{CE} = g_{EC} = 30$.

1.1. Linear Regression Analysis

To conduct the phylogenetic regression analysis, the means of comparative data are typically used. In the five-taxon example shown in Figure 1, suppose $X = (x_A, x_B, x_C, x_D, x_E)$ and $Y = (y_A, y_B, y_C, y_D, y_E)$ are the two hypothetical trait vectors, where $x_i = \sum_{s=1}^{n_i} x_{i,s}/n_i$ and $y_i = \sum_{s=1}^{n_i} y_{i,s}/n_i, i = A, B, C, D, E$ are the mean trait values. The relationship between the two traits can be visualized using the scatter plot, as shown in Figure 2. In general, in the regression framework, the response trait for the n species is denoted as an $n \times 1$ vector $Y_{n\times 1} = (y_1, y_2, \cdots, y_n)^t$, while other traits of interest serve as the covariates, and are displayed as a design matrix X_n with n rows and $p+1$ columns $X_{n\times p} = (\mathbf{1}, X_1, X_2, \cdots, X_p)$, where **1** is a vector of '1s' and each covariate X_i is a column vector $X_i = (x_{i1}, x_{i2}, \cdots, x_{in})^t, i = 1, 2, \cdots, p$. The regression model

$$Y_{n\times 1} = \boldsymbol{X}_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

is applied to analyze the data, where $\beta_{p\times 1}$ is the regression parameter and $\epsilon_{n\times 1}$ is the error with zero mean.

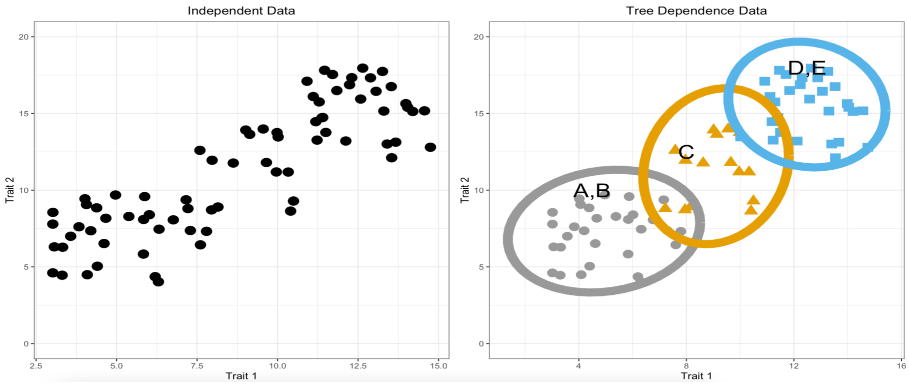

Figure 2. Scatter plots for a pair of hypothetical traits of the five species. **Left panel**: bivariate comparative data, assuming independence without phylogenetic relationships. **Right panel**: bivariate comparative data categorized by the species A, B, C, D, E assuming the relationships adapted from Figure 1.

Assuming there is no evolutionary relationship among species, $\epsilon_{n\times 1}$ has a diagonal covariance matrix $\sigma^2 \boldsymbol{I}_{n\times n}$, where σ^2 is the overall variance. Hence, the statistical distribution of the observed vector Y is $Y_{n\times 1} \sim \mathcal{N}(\boldsymbol{X}_{n\times p}\beta_{p\times 1}, \sigma^2 \boldsymbol{I}_{n\times n})$, where $\beta_p = (\beta_1, \cdots, \beta_p)$ is the regression parameter vector. The estimate for β under the ordinary least square (OLS) estimation approach is $\hat{\beta} = (\boldsymbol{X}^t\boldsymbol{X})^- \boldsymbol{X}^t Y$, with variance $var(\hat{\beta}) = (\boldsymbol{X}^t\boldsymbol{X})^{-1}$. On the other hand, when assuming that species share evolutionary history, the covariance matrix for the residual vector $\epsilon_{n\times 1}$ has non-zero off-diagonal elements (i.e., $var(\epsilon_{n\times 1}) = \sigma^2 \boldsymbol{V}_{n\times n}$), and $\boldsymbol{V}$ also has non-zero off-diagonal elements. When estimating the regression parameters, the estimate for β under the general least squares (GLS) procedure is $\hat{\beta} = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^- \boldsymbol{X}^t\boldsymbol{V}^{-1}Y$, with variance $var(\hat{\beta}) = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}$. A value $\beta_i = 0.32$, for instance, means that a unit increment in the i^{th} covariate x_i would result in a 0.32 unit increment in the expected value of the response Y.

1.2. Tree Polytomy

It is well-known that trees are estimated with some degree of error. Results of the comparative analysis through a phylogenetic regression analysis with *severe* tree errors may impact the model selection, parameter estimation, and the fit of the model(s). For instance, a polytomy tree is not a bifurcated tree, as it has at least one internal node that has more than two immediate descendants (i.e., sister taxa). Polytomies can represent two different cases: Hard polytomy, where the same ancestor is believed to have more than two daughter taxa; and *soft polytomy*, where a cladogram is uncertain without fewer expectations that the same ancestor gives rise to all daughter taxa. A common ancestral population is split through cladogenesis (i.e., speciation) into multiple lineages. As trees with soft polytomy usually have insufficient phylogenetic information, it would be interesting to investigate the fit of the phylogenetic regression model, as well as the parameter estimation, when the tree is not fully resolved at a certain level (see Figure 3 for illustration).

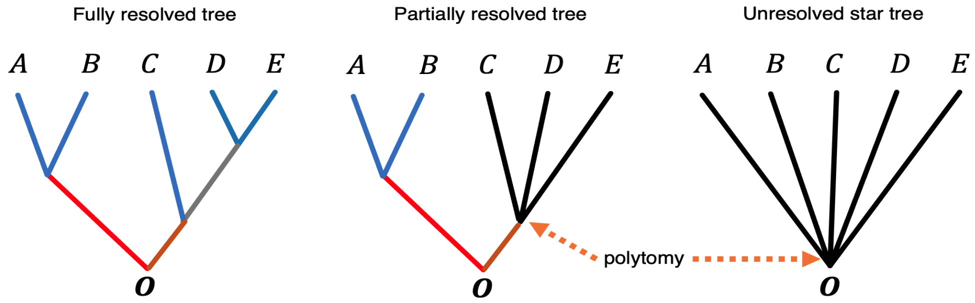

Figure 3. Tree with polytomy cases. **Left panel**: a fully resolved five-taxon tree (same tree in Figure 1), where any node only has two immediate descendants. **Middle panel**: a tree with a mixture of resolved nodes and polytomy by considering that three taxa C, D, E (60% of taxa) on the tips are independent. **Right panel**: a completely unresolved tree, termed a star-shaped tree (100% polytomy), where all taxa are independent.

The goal of this study is to evaluate whether there exists a tree effect on estimating parameters and choosing appropriate models for comparative data analysis, in terms of when polytomies occur on the tree. It is worth noting that an assessment of a phylogenetic regression model (assuming tree dependence with the branch effect), through evaluating the type I error and statistical power of the model, has been described in the literature [39]. Furthermore, it has been previously elucidated how tree misspecification is propagated, through a comparative analysis [40] in which phylogenetic regression under a Brownian motion model of evolution was investigated, considering the effect of local phylogenetic perturbations on the regression fit. This study provides another assessment, focusing on the effect of the tree for comparison among the phylogenetic regression models through qualitative measurements (AIC, r^2, and mean square error, MSE) [41], and the parameter estimation accuracy. This work extends the work in [40], through the evaluation of several other popular models (i.e., Brownian motion [42]; Ornstein–Uhlenbeck, OU [43], early burst, EB [44]; and Pagel's λ, δ, and κ [45] models).

The remainder of this paper is arranged as follows: The models and methods, as well as an explanation of the experiments, are provided in Section 2. The results of the comparison between tree- with non-tree-based regression models are provided in Section 3. Finally, the discussion is provided in Section 4.

2. Methods

In this section, the models for trait evolution considered in Section 2.1 are first described. Then, the experimental process considered in the assessment is detailed in Section 2.2.

2.1. Models Using Continuous Random Process on Trees for Trait Evolution

A scenario depicting the general trait evolution of two species along a fork tree is shown in Figure 4.

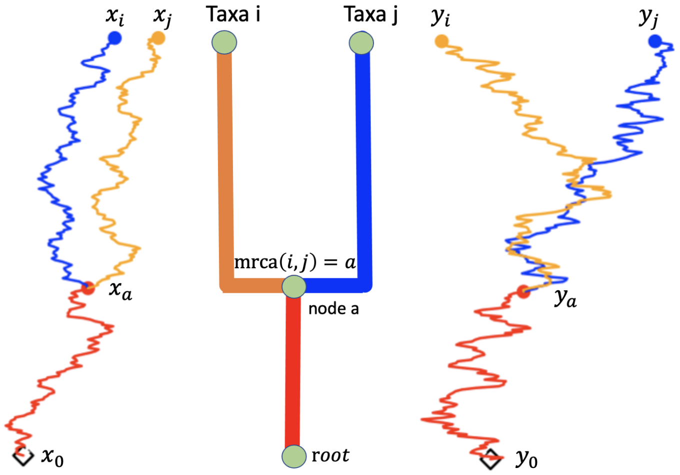

Figure 4. Trait evolution along a phylogenetic fork tree of two species i and j. The two species i and j diverged at time $t = t_a$ and evolved independently to their respective tips. The two unknown ancestral values at the root are denoted as x_0 and y_0, respectively. Two traits X, Y for the i^{th} species (x_i, y_i) and the j^{th} species (x_j, y_j) are observed at the tips. The two unknown ancestral states at time $t = t_a$ for trait X and trait Y are denoted as x_a and y_a, respectively. The dynamics of trait evolution for species along the tree vary with time, which may allow for a correlated relationship to exist between the species [46].

When describing trait evolution using a continuous random process, traits are treated as a pair of stochastic variables (x_t, y_t), adopted a certain pertinent process. Seven popular models were tested: Brownian motion (BM)[42] and the Ornstein–Uhlenbeck (OU) processes are conditioned on whether the root is a random variable or a fixed value [43]; while the early burst model [44], as well as Pagel's λ, δ, and κ models, stretch the branch lengths of the trees under various scenarios.

Brownian Motion: Normal distributions are common in biology. For quantitative trait evolution, it is often assumed that the traits are multivariate normal distributed. By the central limit theorem of probability and statistics, the sum of a set of independent identically distributed random variables (each with finite mean and variance, but no stronger requirements) is again a normal distribution. As a result, the logarithm of a product of many positive independent identical random variables is also normally distributed [47]. For example, if the body mass in a species continues to be multiplied by various factors (e.g., cooler climate leads to a 10% increase in size, lower food availability leads to a 5% decrease in size, or competition with other species leads to a 7% increase in size across many replicates of this evolutionary process), the value of log (body mass) should be normally distributed. This process is often called Brownian motion (BM) [17].

When applying Brownian motion for trait evolution along a tree, the character displacement presents stochastic randomness, as shown in the stochastic differential equation $dy_t = \sigma dB_t$, where $0 \leq t \leq T$. The distribution (see Appendix A.3) of the trait vector $Y_t = (y_{1,t}, y_{2,t}, \cdots, y_{n,t})^t, 0 < t \leq T$, under Brownian motion with tree dependency has the multivariate normal statistical distribution $Y_t \sim \mathcal{N}(\mu_t \mathbf{1}, \sigma^2 \mathbf{G}_t)$, where $\mathbf{G}_t[i,j] = Cov[y_{i,t}, y_{j,t}] = \begin{cases} t_a, & t \leq t_a \\ 0, & t > t_a \end{cases}$, where $t = t_a$ is the time that the two species i and j started to diverge. The covariance matrix element v_{ij} for a pair of species is represented as $v_{ij} = \sigma^2 g_{ij}$.

Ornstein–Uhlenbeck process model: The OU process can appropriately model the trait evolution under stabilizing selection phenomena [48]. Let y_t be an OU process random variable solving the stochastic differential equation $dy_t = \alpha(\theta - y_t)dt + \sigma dB_t$, where $0 \leq t \leq T$, where B_t is a Brownian motion, $\theta \in \mathbb{R}$ is the optimum state, $\alpha > 0$ is the force that pulls the trait back to the optimum, and $\sigma > 0$ is the rate of evolution. Given $Y_t = (y_{1t}, y_{1t}, \cdots, y_{nt})^t$, the joint distribution of the OU random vector Y_t is also a multivariate normal distribution (see Appendix A.3) [49] $Y_t \sim \boldsymbol{N}(\theta\mathbf{1}, \sigma^2\mathbf{V}_t)$, where $\mathbf{V}_t[i,j] = Cov[y_{i,t}, y_{j,t}] = \begin{cases} \frac{\sigma^2}{2\alpha}(\exp(-2\alpha(\mathrm{t}-s) - \exp(-2\alpha t)), & t \leq t_a \\ 0, & 0 \leq t \leq s. \end{cases}$ The covariance matrix element v_{ij} for a pair of species is represented as $v_{ij} = (\frac{\sigma^2}{2\alpha}(\exp(-2\alpha(g_{ii} - g_{ij})) - \exp(-2\alpha(g_{ii}))$.

Early burst model: The early burst model is a rate time-varying Brownian motion model, where the rate slows over time (branch length). In this model, the rate parameter $\sigma^2 = \sigma_0^2 \exp(rt)$, where $r < 0$ is a contraction parameter that slows down the rate of evolution and σ_0, is the rate at the root. Hence, $g_{ij} \mapsto \exp(rg_{ij})$ and the covariance matrix for a pair of species is represented as $v_{ij} = \sigma_0^2(\exp(rg_{ij} - 1))/r$.

Four realizations of the trajectories with the BM, OU, and EB models are shown in Figure 5.

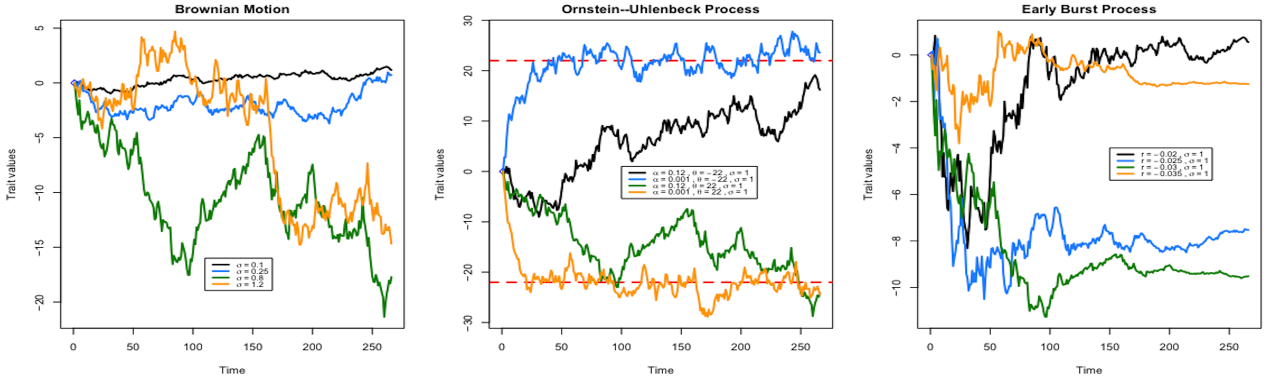

Figure 5. **Left panel**: Trajectories for the BM using four rates of evolution ($\sigma = 0.1, 0.25, 0.8$, and 1.2). **Middle panel**: Trajectories for the OU using different two optima ($\theta = -22, 22$), $\alpha = 0.001, 0.12$, and $\sigma = 1$. **Right panel**: Trajectories for the EB model using four different rates ($r = -0.02, -0.025, -0.03, -0.035$) with $\sigma = 1$.

Pagel's λ model: The λ model (equivalent to the phylogenetic mixed model [50]) accounts for a mixture of independent evolution and BM. The λ model makes a branch transformation, where all internal branches are multiplied by a factor λ in $[0,1]$, while the tips are kept at the original distance from the root. Hence, the λ model transforms the off-diagonal elements $g_{ij}, i \neq j$ in $\mathbf{G}$ into a linear combination $g_{ij} \mapsto \lambda g_{ij}$, where $0 \leq \lambda \leq 1$ is the shrink parameter, which strengthens the tree into a more independent type. While a value of λ close to 1 would maintain the original tree, a smaller value of λ (i.e., closer to 0) would result in lower covariance between a pair of species. The covariance matrix element v_{ij} for a pair of species is represented as $v_{ij} = \sigma^2\lambda g_{ij}$ for $i \neq j$ and $v_{ii} = \sigma^2\lambda g_{ii}$.

Pagel's δ model: The δ model directly transforms the element in the covariance matrix $g_{ij} \mapsto g_{ij}^\delta$, where $\delta > 0$. Usually, the tree is scaled to a unit tree height, such that all elements in $\mathbf{G}$ are smaller than 1 ($0 \leq g_{ij} \leq 1$). Clearly, when $\delta = 1$, the δ model is equivalent to the Brownian motion model. Given $\delta < 1$, the elements $g_{ij}^\delta > g_{ij}$ (e.g., $0.25^{0.5} = 0.5 > 0.25$), indicating that the node heights are reduced with longer branch lengths, while $\delta > 1$ yields $g_{ij}^\delta < g_{ij}$ (e.g., $0.25^2 = 0.0625 < 0.25$) and the transformation stretches the branches with shorter branch lengths. The covariance matrix element v_{ij} for a pair of species is represented as $v_{ij} = \sigma^2 g_{ij})^\delta$. In short, the delta transformation raises the distance from the root to all nodes to the power $\delta > 0$.

Pagel's κ model: The κ model concerns the transformation of the piece-wise branch lengths through the power element. First, the element g_{ij} is decomposed into a sum of the branch lengths from the root to the most recent ancestor of tips i and j (i.e., $g_{ij} = \sum_{s=1}^{d} b_{s,ij}$, where $b_{s,ij}$ is the branch and d is the total number of branches the i and j shared, respectively). Then, the transformation is carried out as $g_{ij} \mapsto \sum_{b=0}^{s} g_{b,ij}^{\kappa}, \kappa > 0$; that is, the branch lengths themselves are transformed. For $\kappa < 1$, (e.g., $g_{ij} = 0.25 = 0.16 + 0.09 \mapsto 0.16^{0.5} + 0.09^{0.5} = 0.4 + 0.3 = 0.7 > 0.25$), the branch is stretched with a deeper branch length; while, for $\kappa > 1$ (e.g., $g_{ij} = 0.25 = 0.16 + 0.09 \mapsto 0.16^2 + 0.09^2 = 0.256 + 0.0081 = 0.037 < 0.25$), when compared to the original tree, the branches are compressed in the same scale. Note that, as the diagonal elements are transformed, the tree is no longer an ultrametric tree (all tips have the same tree height). The covariance matrix element v_{ij} for a pair of species is represented as $v_{ij} = \sigma^2 \sum_{b=0}^{s} g_{b,ij}^{\kappa}$. In short, the κ transformation raises all branch lengths to a power κ in $[0,1]$. An example illustrating the transformed covariance matrices is provided in Appendix A.3. Readers may also refer to [51,52], for more comprehensive and instructional descriptions of these models.

The likelihood for the regression model: The general statistical distribution for the regression model assuming a Gaussian process for trait evolution is a normal distribution. Hence, the trait data vector $Y = (y_1, y_2, \cdots, y_n)^t$ observed at the tip for n species follows a multivariate normal distribution $Y \sim \mathcal{N}(\mu \mathbf{1}_n, \sigma^2 \boldsymbol{V})$, where $\boldsymbol{V}$ is the covariance structure and $\mathbf{1}_n = (1, 1, \cdots, 1)^t$ is a vector of 1s. The negative log-likelihood function is

$$-\log L(\Theta, \beta | Y, \boldsymbol{X}, \boldsymbol{V}) = \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\sigma^2 \boldsymbol{V}| + \frac{1}{2\sigma^2} (Y - \boldsymbol{X}\beta)^t \boldsymbol{V}^{-1} (Y - \boldsymbol{X}\beta), \quad (1)$$

where $\Theta = \sigma^2$ for the BM model [42], $\Theta = (r, \sigma^2)$ for the EB model [44], $\Theta = (\alpha, \theta, \sigma^2)$ for the OU model [53], $\Theta = (\lambda, \sigma^2)$ for the λ model, $\Theta = (\delta, \sigma^2)$ for the δ model, and $\Theta = (\kappa, \sigma^2)$ for the κ model [45].

The generalized least squares estimate for the regression parameter is

$$\hat{\beta} = (\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X})^{-} \boldsymbol{X}^t \boldsymbol{V}^{-1} Y, \quad (2)$$

where $\boldsymbol{X}^{-1}$ is the generalized inverse of $\boldsymbol{X}$ and $\boldsymbol{X}^t$ is the transpose of $\boldsymbol{X}$. As additional model parameters (i.e., α for the OU model, r for the EB model, λ for the λ model, *delta* for the delta model, and κ for the κ model) are embedded in the covariance matrix $\boldsymbol{V}$, a heuristic procedure for estimating the model and regression parameters can be performed; in particular, to estimate the model parameter embedded in $\boldsymbol{V}$ and the regression parameter β. The above parameter estimation was executed using the R package `phylolm` [54].

2.2. Assessment Through Extensive Simulation

The main objective of this study was to evaluate the effect of using trees with errors on the model selection. The models for trait evolution (BM, OU fixed root, OU random root, EB, λ, δ, κ models) were used as the models for evaluation. Consider the following scenario for evaluating the model performance under varying levels and types of polytomy. For the simulation, four types of trees were set: (1) balanced tree with taxa size; (2) random split tree; (3) pure birth tree; and (4) birth–death tree. The balanced tree consisted of $8, 16, 32, 64, 128, 256$, or 512 taxa, while the random split, pure birth, and the birth–death trees used taxa sizes of $20, 50, 100, 150, 200, 250, 300, 350, 400, 450$, or 500. The simulated traits and trees were fitted to the seven models (BM, EB, OU fixed root, OU random root, δ, κ, λ). The regression parameters (intercept a, slope b, model parameter rate of evolution σ), as well as the model fitting measures AIC ($AIC = 2k - 2\log L_i(\hat{\Theta}|Y, \boldsymbol{X})$, where k is the number of parameters, L is the likelihood, and $\hat{\Theta}$ is the maximum likelihood estimator), r^2 (which measures the variation of the response by the variation of the predictor/independent variables, where higher values represent smaller differences between the observed data and the fitted values), and the mean squared error ($MSE = \sum_{i=1}^{m} (\theta - \hat{\theta})^2 / m$, where m is the number of replicates), were assessed.

As for the polytomy error of the tree, two polytomy cases, considering the type of polytomy (clade vs. root) and the level of polytomy (level 1 to level 8, determined by the node level from the root), were used (as shown in Figure 6). A selected node was chosen and the tree was transformed to introduce the polytomy on the clade, as executed by the R package `RRphylos` function `fix.poly` [55]. The eight levels were then grouped into two categories (high and low) for comparison in the next step, which was accomplished using the R package `TreeTools` function `CollapseNode, NDescendants` [56], where the number of tips was chosen based on the ratio of polytomy taxa and the total taxa.

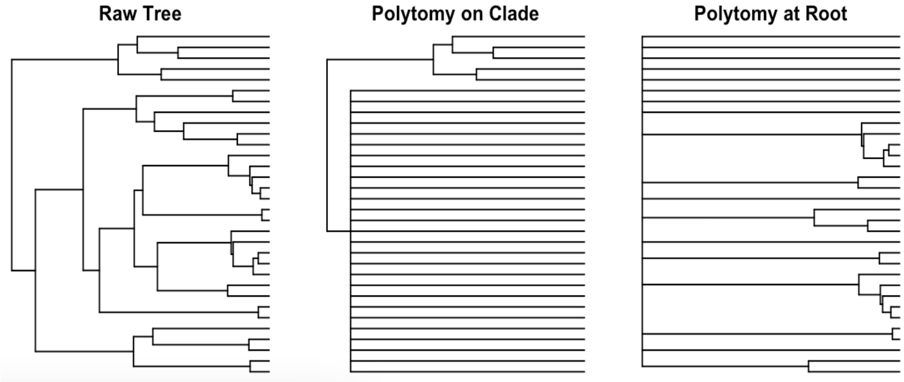

Figure 6. Trees with severe polytomy. **Left panel**: a pure birth tree. **Middle panel**: polytomy on a clade. **Right panel**: polytomies on the root.

The simulation starts by setting up the true parameters (intercept $a = 5$, slope $b = 3$, and $\sigma^2 = 1.2$). For the model-specific parameter, the rate $r = -0.10$ was assigned to the EB model, $\delta = 0.25$ to the delta model, $\kappa = 1.25$ to the kappa model, $\lambda = 0.625$ to the lambda model, and $\alpha = 0.10$ to the OU fixed root model and the OU random root model. Trait data were then simulated using the tree models (BM, OU, EB, λ, δ, κ model). Trees were then transformed into trees with polytomy at a clade or root, following which the polytomy tree was incorporated to fit the models, thus obtaining the parameter estimates and fitting measures. A total of 100 replicates of the trait data were generated for each combination of tree type (4 types), taxa size (11 sizes), model type (7 models), polytomy type (2 types), and polytomy level (8 levels). This yielded 100 replicates $\times$ 4 tree types $\times$ 7 models $\times$ 11 taxa size $\times$ 8 levels of polytomy = 246,400 replicates, estimates, and measures used for assessing model performance.

3. Results

The simulations differed from each other. By considering the variable trees, regarding their polytomy type and polytomy level, the results for parameter estimation and fitting measures were dissected into several categories, as follows. The results for the overall estimates are provided in Section 3.1; the results grouped by polytomy type and level are provided in Section 3.2; the results grouped by the tree type are provided in Section 3.3; finally, the results grouped by the model are provided in Section 3.4. All results are reproducible (see the link in the Appendix A.1).

3.1. Overall Estimate

The comparison between the raw parameters and the simulated values is shown in Table 1, where the result in the *raw* row considers the data simulated from the true parameters, then fitting the models to the data to obtain the estimates and measures; while the result in the *sim* row uses the estimator from the raw data to simulate 100 replicates under each scenario (polytomy type and level), after which the models were fitted to the data to obtain the estimates and measures.

From Table 1, it can be seen that the difference between the raw and simulation data was insignificant. For the intercept parameter, the test statistic was 0.147 with p-value = 0.884; for the slope parameter, the test statistic was -0.680 with p-value = 0.497; and, for the σ^2 parameter, the test statistic was -0.114 with p-value = 0.909. The boxplots for all replicates of the intercept, slope and rate of square estimates are shown in Figure 7.

Table 1. The overall mean estimates across tree types, models, and taxa sizes for parameters (intercept, slope, and σ^2) as well as fitting measures (AIC, r^2, and MSE). In the raw row: Trait and tree data were simulated using the true parameters. Sim: Estimates from 100 replicates simulated from the parameter estimates using the raw data across trees, models, taxa size, polytomy type, and polytomy level.

	Intercept	Slope	σ^2	AIC	R^2	MSE
Raw	4.992	2.981	25.742	1114.414	0.324	2.317
Sim	4.997	2.999	26.638	1117.378	0.364	2.328

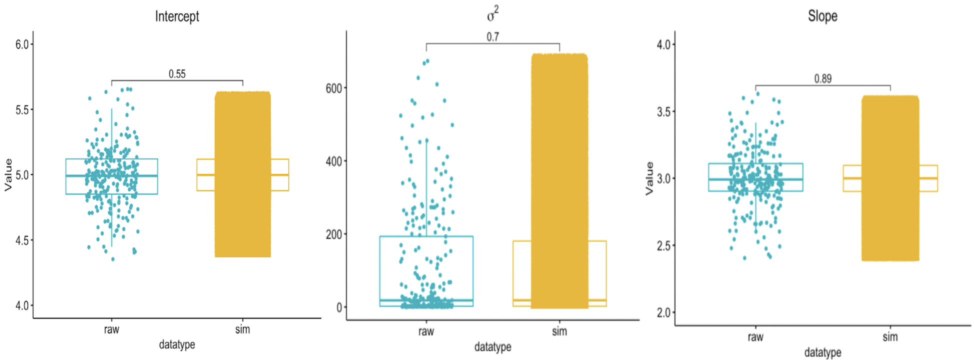

Figure 7. Boxplots for all replicates. Raw: Parameter estimates using data simulated by true parameter values. Sim: Parameter estimates using data simulated by the raw estimates. Overall the mean estimates between the raw and sim is insignificant. The *p*-values reported in plot are 0.55, 0.7, and 0.89 for intercept, σ^2, and slope, respectively.

From Table 1, as there was no significant difference between using the raw or simulated data, the results provide preliminary support for the validity of the set-up for the simulation, as the models performed well to return reasonable estimates.

We next conducted a *t*-test, regarding the overall difference between the polytomy conditions; the results are shown in Table 2.

Table 2. Differences between polytomy types and levels. All results are insignificant (high *p* values are reported) except for comparing the r^2 between the root and clade (*p*-value of 0.01)

	Polytomy Type (Root vs. Clade)	Polytomy Level (High vs. Low)
Intercept *a*	$(-0.014, 0.023)$, $p = 0.63$	$(-0.023, 0.013)$, $p = 0.61$
Slope *b*	$(-0.006, 0.021)$, $p = 0.28$	$(-0.019, 0.008)$, $p = 0.44$
σ^2	$(-531.447, 15.805)$, $p = 0.06$	$(-179.965, 367.877)$, $p = 0.5$
AIC	$(-77.071, 101.029)$, $p = 0.79$	$(-91.487, 86.618)$, $p = 0.96$
r^2	$(0.011, 0.086)$, ***p* = 0.01** *	$(-0.065, 0.01)$, $p = 0.15$
MSE	$(-2.752, 3.143)$, $p = 0.9$	$(-3.276, 2.619)$, $p = 0.83$

The test considering the type and level of polytomy (high vs. low) on the simulation data showed that there were mainly insignificant differences in the parameters and measures. The confidence intervals and their *p*-value are reported. For instance, for the intercept parameter, the values $(-0.014, 0.023)$,0.63 in the first column and first row indicate that the 95% confidence interval for the difference between the intercept estimated under the two groups (clade and root) was $(-0.014, 0.023)$, and the test H_0 : $\text{intercept}_{\text{clade}} - \text{intercept}_{\text{root}} = 0$ had a *p*-value of 0.63, indicating that there was no significant difference for the intercept estimated by the clade and root type, respectively.

Note that, from Table 2, the r^2 measure presented a significant difference with respect to the polytomy type. Further investigation of this difference was facilitated through more fine dissection by graphical visualization; see the following sections.

3.2. Polytomy Type vs. Polytomy Level

Two polytomy cases were analyzed, and comparisons were made with respect to polytomy type (clade vs. root), as well as polytomy level (high vs. low).

The results for the parameter estimates are reported in Figure 8. Overall, although there were outliers for the estimates, the polytomy type and level did not significantly affect the parameter estimation, as reasonable estimates were obtained while varying both the type and the level (the overall median was close to the true parameter value).

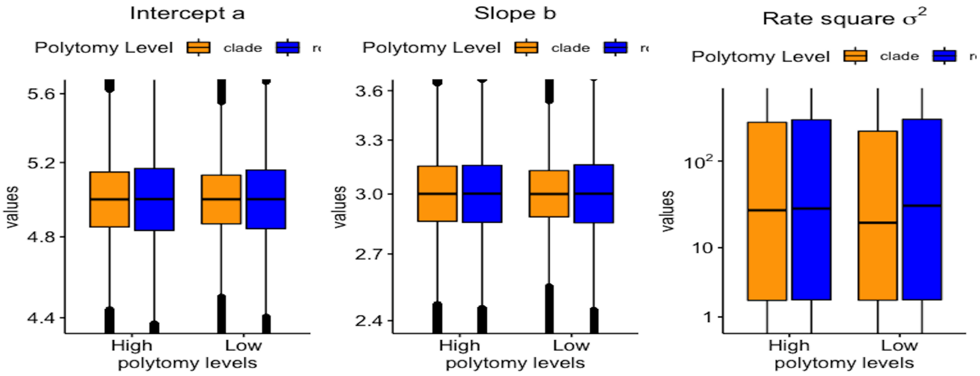

Figure 8. Boxplots for the intercept a, the slope b, and the square of rate of evolution σ^2 with varying polytomy type and level.

For the intercept parameter, the regression model using the intercept estimate as the response was Intercept $\sim$ poly type + poly level. For the polytomy type, the regression estimate for the slope of clade vs. root was -0.004 with the t-value of -0.769 and a p-value of 0.442. In the paired sample t-test (clade vs. root), the 95% confidence interval between the difference of clade vs. root was $(-0.006, 0.013)$, with the p-value of 0.442. For the polytomy level, the regression estimate for the slope of low vs. high was 0.005 with a t-value of 0.869 and a p-value of 0.385. In the paired t-test (low vs. high), the 95% confidence interval was $(-0.016, 0.006)$, with a p-value of 0.396.

For the slope parameter, the regression model using the slope estimate as the response was slope $\sim$ poly type + poly level. For the polytomy type (clade vs. root), the regression estimate for the slope of clade vs. root was -0.004 with a t-value of -0.969 and a p-value of 0.332. In the paired sample t-test (clade vs. root), the 95% confidence interval was $(-0.004, 0.123)$, with a p-value of 0.332; the regression estimate for the slope of low vs. high was 0.005 with a t-value of 1.105 and a p-value of 0.269. In the paired sample t-test (low vs. high), the 95% confidence interval was $(-0.015, 0.004)$, with a p-value of 0.263.

For the σ^2 parameter, the regression model using the σ^2 estimate as the response was $\sigma^2 \sim$ poly type + poly level. For the polytomy type (clade vs. root), the regression estimate for the slope of clade vs. root was 209.10 with a t-value of 3.354 and p-value of 8×10^{-4} (i.e., significant). In the paired sample t-test (clade vs. root), the 95% confidence interval was $(-331.283, -86.918)$, with a p-value of 0.0008. For polytomy level (low vs. high), the regression estimate for the slope of low vs. high was -93.96 with a t-value of 71.98 and a p-value of 0.19. In the paired sample t-test (low vs. high), the 95% confidence interval was $(-40.636, 228.548)$, with a p-value of 0.171.

For the regression parameters, the above results were consistent with those shown in [40], where only a modest effect was reported when the tree effect was encountered. For the model parameter σ^2, when polytomy was present, σ^2 was over-estimated. Therefore, one should be careful about the estimated rate of evolution σ when the tree has more uncertainty due to polytomy.

For AIC, the results are shown in Figure 9, where polytomy overall did not affect the fit AIC tremendously. The linear regression for AIC on the polytomy type level also reported insignificant results. The regression analysis considered aic $\sim$ poly type + poly level. For the polytomy type, the regression estimate for the slope of clade vs. root was -0.004 with a t-value of -0.969 and a p-value of 0.332. In the paired t-test (clade vs. root), the 95% confidence interval was $(4.721, 13.739)$, with a p-value $\approx 6 \times 10^{-5}$. For the polytomy level, the regression estimate for the slope of low vs. high showed an estimate of 0.005 with a t-value of 1.105 and a p-value of 0.269. In the paired sample t-test (low vs. high), the 95% confidence interval was $(-7.687, 2.817)$, with a p-value of 0.364. Note that the violin plots have multi-modal shapes with multiple peaks, due to the larger taxa sizes (from $20, 50, 100, 150, 200, \cdots, 500$ taxa) contributing to larger likelihoods, thus yielding a larger value of the AIC.

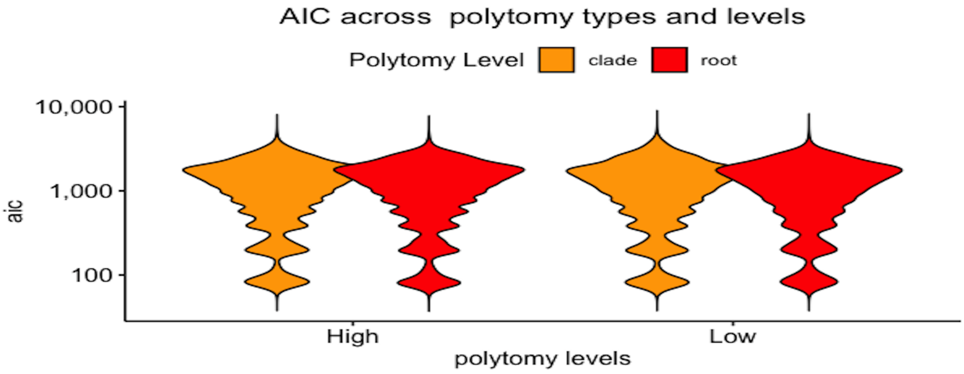

Figure 9. Violin plots of AIC by polytomy type and level, conditioned through taxa size, tree type, and model.

For r^2, the results are shown in Figure 10, where polytomy overall had major effects on r^2. The linear regression for r^2 on the polytomy type and level reported significant results (i.e., the median between the clade and root in the low category is distinguishable). In particular, when a low level of polytomy is presented (Figure 10, right panel), the r^2 presented different values between the clade and root. The regression analysis considered $r^2 \sim$ poly type + poly level. For the polytomy type, the regression estimate for the slope clade vs. root was -0.030, with a t-value of -29.68 and a p-value $< 2 \times 10^{-16}$. In the paired sample t-test (clade vs. root), the 95% confidence interval was $(0.028, 0.032)$, with p-value $< 2 \times 10^{-16}$. For the polytomy level, the regression estimate for the slope low vs. high was 0.028 with a t-value of 23.87 and a p-value $< 2 \times 10^{-16}$. In the paired sample t-test (low vs. high), the 95% confidence interval was $(-0.030, -0.025)$, with a p-value $< 2 \times 10^{-16}$.

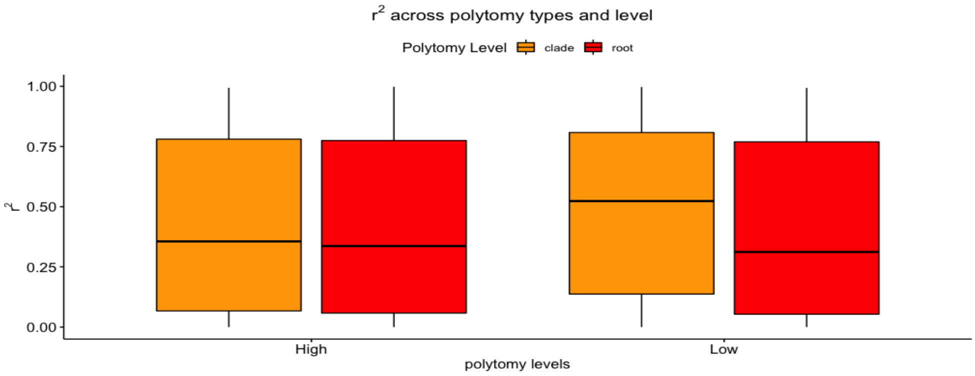

Figure 10. Box plots of r^2 by polytomy type and level conditioned through polytomy type and polytomy level.

For MSE, the results are shown in Figure 11, from which it can be seen that polytomy overall had a minor effect on the MSE. The linear regression for MSE on the polytomy type level also reported insignificant results. The regression analysis considered MSE $\sim$ poly type + poly level. For the polytomy type, the regression estimate for the slope clade vs. root was −0.123 with a t-value of −0.162 and a p-value of 0.871. In the paired sample t-test (clade vs. root), the 95% confidence interval was $(-1.370, 1.616)$, with a p-value of 0.872. For the polytomy level, the regression estimate for the slope (low vs. high) was estimated as 0.329 with a t-value of 0.374 and a p-value of 0.709. By paired sample t-test (low vs. high), the 95% confidence interval was $(-2.512, 1.855)$, with a p-value of 0.768.

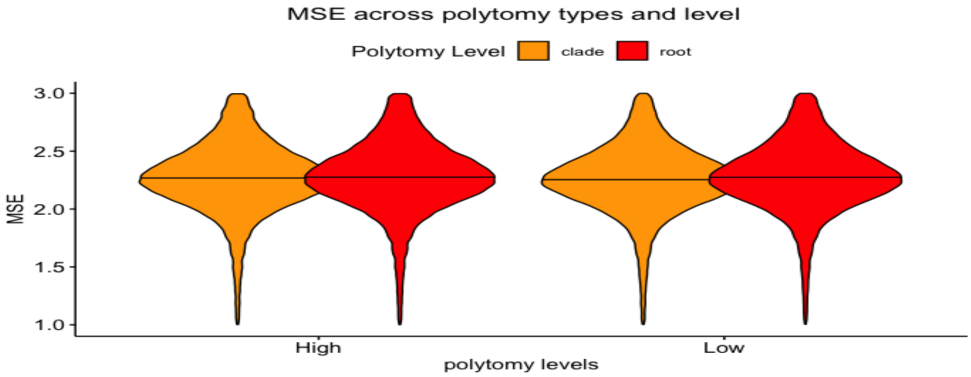

Figure 11. Violin plots of MSE by the polytomy type and level conditioned by taxa size, tree type, and model.

3.3. Tree vs. Measures

In this section, the comparison of the measures, with respect to the tree type and polytomy type, is reported.

For AIC, the result shown in Figure 12 indicates the trend that polytomy type and level overall did not overly affect the fit, as measured by the AIC. One can conduct a median test (e.g., Mood's median test or Kruskal–Wallis test) for the four tree groups.

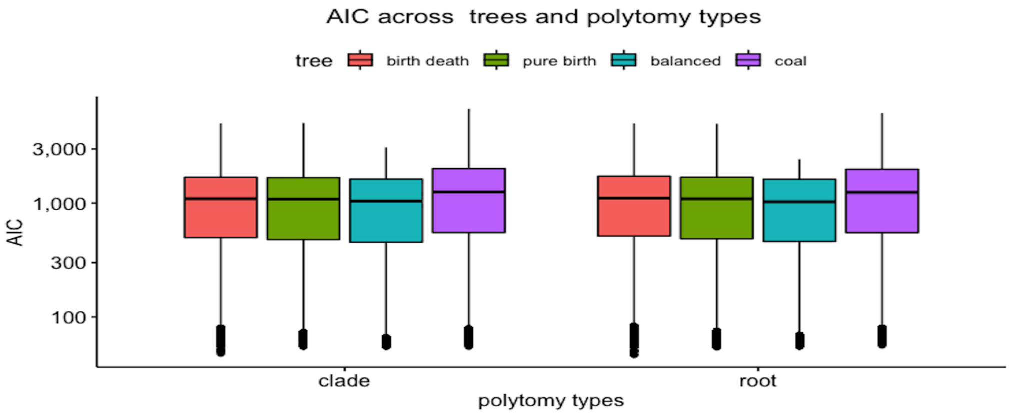

Figure 12. Comparison of AIC across tree type and polytomy type.

For r^2, the results are shown in Figure 13, which indicates that there was a tremendous effect on the r^2 when polytomy was present. Specifically, while the birth death tree was the most robust tree type, mostly retaining high r^2, followed by the pure birth tree, the balanced tree and the random split tree gave lower r^2 values.

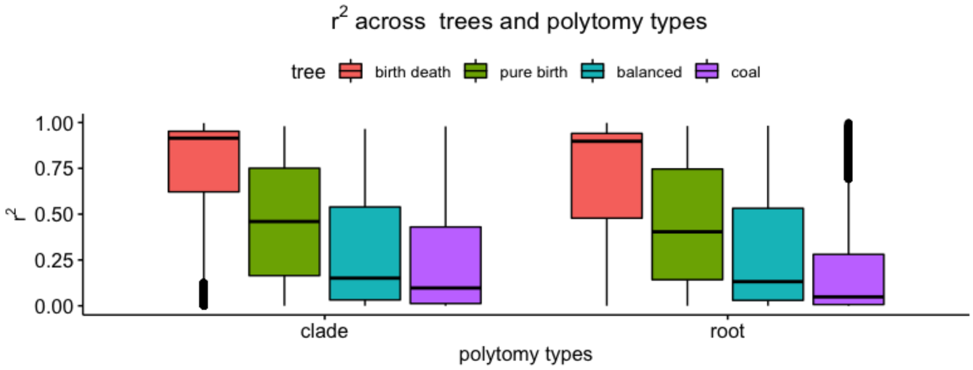

Figure 13. Comparison of r^2 across tree type and polytomy type. Higher values represent smaller differences between the observed data and the fitted values.

For MSE, the result is shown in Figure 14. Overall, there was no tremendous difference in the median of the mean square error across tree types. However, the random split tree (rcoal), in general, presented a larger variation (i.e., wider interquartile range) than the other three tree types.

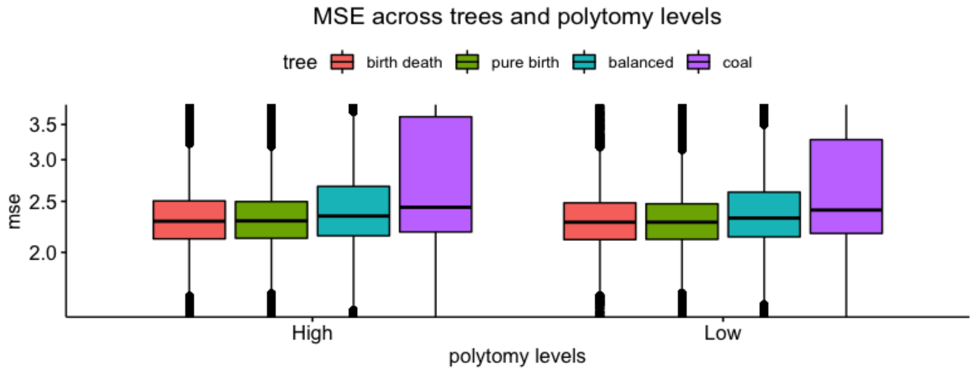

Figure 14. Comparison of MSE across tree type and polytomy level.

3.4. Model vs. Measure

In this section, comparisons of the measures with respect to the polytomy and model types are reported.

For the intercept parameter, the result is shown in Figure 15. When polytomy was present (with various types and levels), the λ model, OU random root model, and OU fixed root model (the three boxplot in the rightmost in each panel) returned better estimates with narrower interquartile ranges than the others (δ, EB, BM, κ model).

For the slope parameter, the result is shown in Figure 16. Similar to the result for the intercept, when encountering polytomy (at various types and levels), the λ model, OU random root model, and OU fixed root model (the three boxplots in the rightmost in each panel) performed better than the other four models (the δ model, the EB model, the BM model, the κ model).

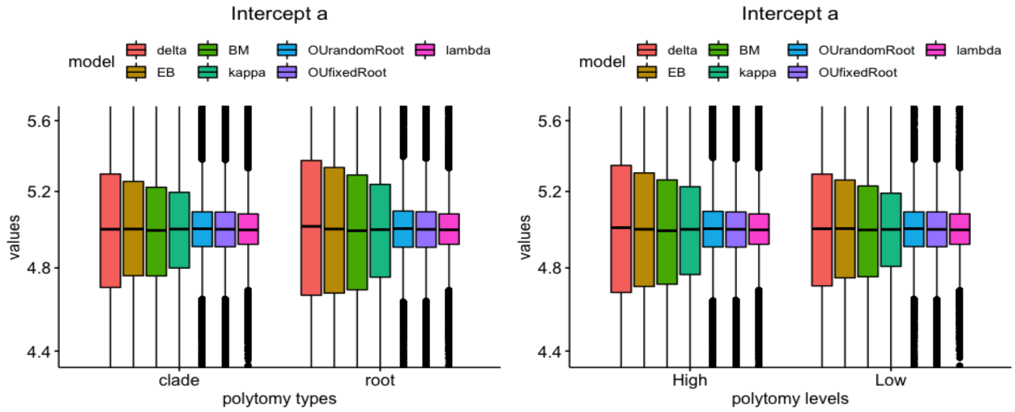

Figure 15. Regression intercept estimation concerning different models and polytomy types. **Left panel**: polynomial type vs. models. **Right panel**: polynomial level vs. models.

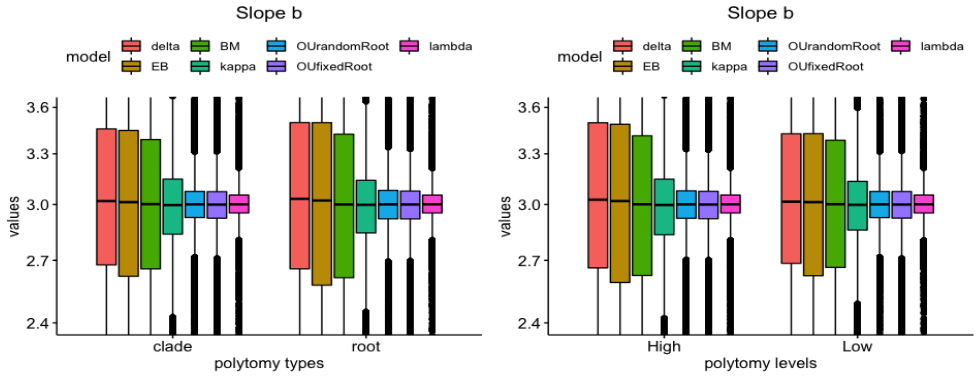

Figure 16. Regression slope estimation concerning different models and polytomy types. **Left panel**: polynomial type vs. models. **Right panel**: polynomial level vs. models.

For the σ^2 parameter, the result is shown in Figure 17. When encountering polytomy (at various types and levels), the λ model and the κ model remained robust, returning reasonable estimates, while the other models (δ, EB, BM, OU random root, OU fixed root) returned over-estimated values.

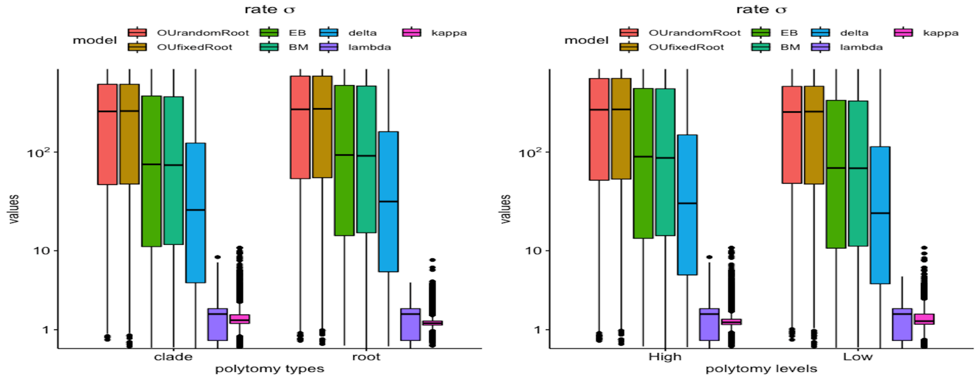

Figure 17. σ^2 estimation concerning different models and polytomy types. **Left panel**: polynomial type vs. models. **Right panel**: polynomial level vs. models.

For AIC, the result is shown in Figure 18. From the median value in the boxplots, under various polytomy types and polytomy levels, there existed a transparent difference between two groups of models, where the δ, EB, BM, and κ models had higher AIC (worse fit) than the OU random root, OU fixed root, and λ models.

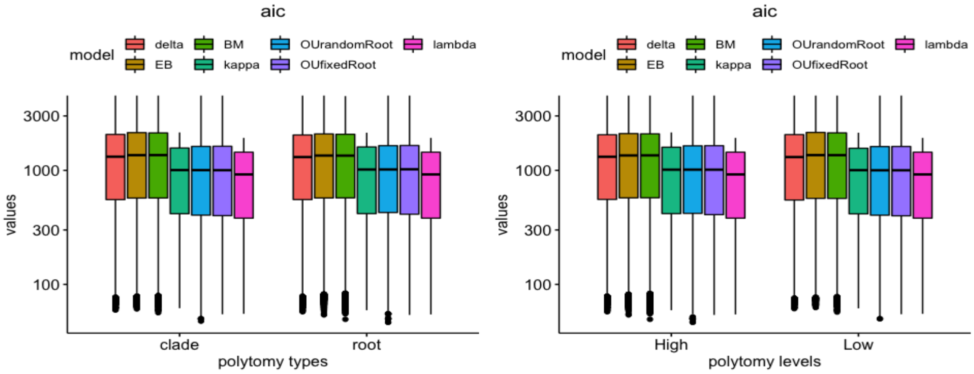

Figure 18. AIC concerning different models and polytomy types. **Left panel**: polynomial type vs. models. **Right panel**: polynomial level vs. models.

For r^2, the result is shown in Figure 19. When encountering polytomy (various types and levels), the λ model overall outperformed the other models (followed by the OU-type models) and provided a better fit than the others (δ, EB, BM, and κ models).

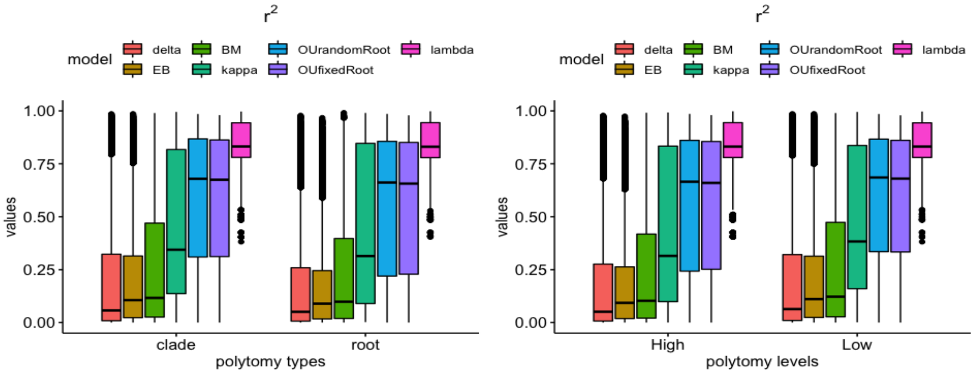

Figure 19. r^2 concerning different models and polytomy types. **Left panel**: polynomial type vs. models. **Right panel**: polynomial level vs. models.

For MSE, the result is shown in Figure 20. When encountering polytomy at various types and levels, the λ model (followed by the OU type models) provided a better fit, with lower MSE, than the others (δ, EB, BM, κ models).

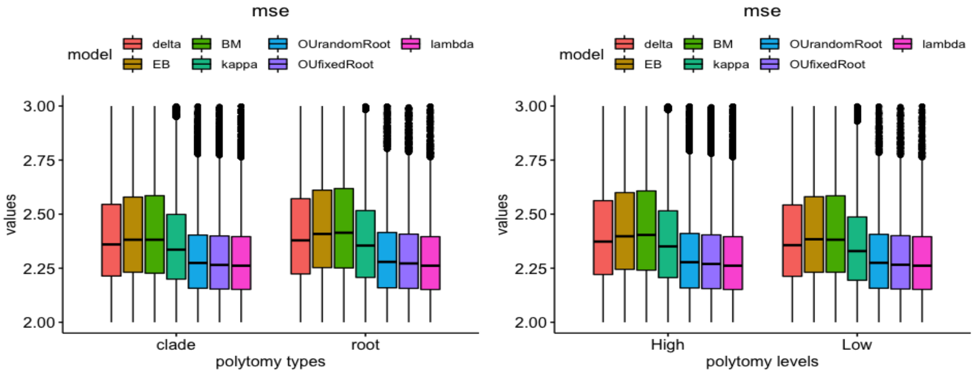

Figure 20. MSE concerning different models and polytomy types. **Left panel**: polynomial type vs. models. **Right panel**: polynomial level vs. models.

From the results reported in Figures 18–20, encountering polytomy issues, such as the λ model (lengthen/shrink the tip) and the OU models (the larger force parameter α leads to more independent structures in the covariance matrix) can transform the tree into a more independent style (i.e., polytomy at the root); the λ and OU-type models provided a better fit than other models.

4. Discussion

This work aimed to compare the influence of error in phylogenetic trees on the results of phylogenetic regressions. Several phylogenetic regression models were reviewed, and their goodness of fit to large data sets were evaluated through the AIC, r^2, and mean square error measures. While the parameter estimates were not significantly impacted by the tree effects, the fitting comparison of the models suggested that, overall, Pagel's λ model and the OU models possess the best functionality, transforming the tree into a more independent style (i.e., polytomy at the root); one may infer that these two model types provide a better fit when polytomy is present.

There are many other regression models feasible for other types of response variables, such as positive [57], binary [58], counts [59], and proportional [60] data. Popular univariate models of trait evolution concerning quantitative response variables were investigated in this study; in the future, it would be interesting to investigate how polytomy could impact the comparative methods when extending the univariate models, as well as other models [13,49,61–64]. It would be also interesting to determine whether polytomy would have a stronger effect on the regression analysis that includes hybrid species, where the phylogenetic network is implemented to describe the dependence structure [65,66]. Note that, while more parameter-rich models provide more evolutionary information, they often come with a sophisticated covariance structure, thus increasing the difficulty of parameter estimation in higher dimensional parameter space. Under this circumstance, using the Bayesian parameter estimation technique would allow more flexibility in this framework [67,68].

One should be careful when using phylogenetic regression models with more sophisticated covariance structure, which might violate the assumptions required for linear regression (i.e., linear relationship, multivariate normality, no or little multi-collinearity, no auto-correlation, and homoscedasticity). Violation of these assumptions could yield the same estimates, but misleading results. See Anscombe's quartet in ordinary linear regression [69] and phylogenetic regression [70].

As the comparison of species helps us to understand systematic differences in biological traits, incorporating trees for the testing of evolutionary hypotheses has become an essential procedure for verifying relevant assumptions. Finally, users should correct for phylogeny when studying the relationships between traits, and choose appropriate models for analyzing data prior to conducting the phylogenetic comparative data analysis.

Author Contributions: Conceptualization, D.-C.J.; data curation, F.-C.L. and D.-C.J.; methodology, F.-C.L. and D.-C.J.; formal analysis, F.-C.L. and D.-C.J.; investigation, F.-C.L. and D.-C.J.; resources, D.-C.J.; visualization, F.-C.L. and D.-C.J.; writing—original draft preparation, D.-C.J.; writing—review and editing, D.-C.J. and F.-C.L.; funding acquisition, D.-C.J. and F.-C.L. All authors have read and agreed to the published version of the manuscript.

Funding: Jhwueng's research was funded by National Science and Technology Council Taiwan: MOST 111-2118-M-035-004-; Liu's research was funded by National Science and Technology Council Taiwan: MOST 111-2118-M-035-005-.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Acknowledgments: The authors thank Olaf R.P. Bininda-Emonds, Vasileios Maroulas, Folashade B. Agusto and the two anonymous reviewers for their constructive suggestions, which helped tremendously in improving the manuscript. Thanks to En-Ping Lin, Chia-Wang Kuo, Chia-Yi Lin, and Hung-Lin Liao for their technical assistance.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Script and Link for Reproducing Result

The script and files can be accessed via the following links: https://tonyjhwueng.info/pcmreg (accessed date: 19 September 2022).

1. Figure 1: https://tonyjhwueng.info/pcmreg/ttev2.pptx (accessed date: 19 September 2022).
2. Figure 2: https://tonyjhwueng.info/pcmreg/scatterclusterv2.html (accessed date: 19 September 2022).
3. Figure 3: https://tonyjhwueng.info/pcmreg/ttevprog.pptx (accessed date: 19 September 2022).
4. Figure 4: https://tonyjhwueng.info/pcmreg/bm1bm2reg.html (accessed date: 19 September 2022).
5. Figure 5: https://tonyjhwueng.info/pcmreg/bmpathv3_outpathv3_ebpathv2.html (accessed date: 19 September 2022).
6. Figure 6: https://tonyjhwueng.info/pcmreg/makepolytree.html (accessed date: 19 September 2022).
7. Figure 7: https://tonyjhwueng.info/pcmreg/mainsimSummarWrapabsig2.html (accessed date: 19 September 2022).
8. Tables 1 and 2: https://tonyjhwueng.info/pcmreg/mainsimSummarWrapTable.html (accessed date: 19 September 2022).
9. Figures 8–20: https://tonyjhwueng.info/pcmreg/mainsimSummarWrapBoxplotv4.html (accessed date: 19 September 2022).

Appendix A.2. Database for Accessing Comparative Data and Tree

Several popular databases are listed in Table A1.

Table A1. Online databases for phylogenetic trees and trait data sets.

Logo	Database	Reference	Link
	AmphibiaWeb	[22]	https://amphibiaweb.org/ (accessed date: 19 September 2022).
	The Reptile Database	[23]	http://www.reptile-database.org/ (accessed date: 19 September 2022).
	GLAD	[24]	http://globalants.org/ (accessed date: 19 September 2022).
DATElife	DateLife	[71]	http://datelife.opentreeoflife.org (accessed date: 19 September 2022).
EZBioCloud	EzBioCloud	[72]	https://www.ezbiocloud.net (accessed date: 19 September 2022).
FishBase	FishBase	[25]	https://www.fishbase.se/ (accessed date: 19 September 2022).
OPEN Tree of Life	Open Tree of Life	[32]	https://tree.opentreeoflife.org/ (accessed date: 19 September 2022).

Table A1. *Cont.*

Logo	Database	Reference	Link
	PhylomeDB	[73]	http://phylomedb.org/ (accessed date: 19 September 2022).
	PHYLOtastic	[28]	https://phylotastic.org/ (accessed date: 19 September 2022).
	Traitbase	[74]	https://traitbase.info/ (accessed date: 19 September 2022).
	TreeBASE	[75]	https://www.treebase.org/ (accessed date: 19 September 2022).
	Treefam	[26]	http://www.treefam.org (accessed date: 19 September 2022).
	Tree of Life Web Project	[29]	http://tolweb.org/tree/ (accessed date: 19 September 2022).
	The Open Traits Network	[30]	https://opentraits.org/ (accessed date: 19 September 2022).
	TRY Plant Trait Database	[27]	https://www.try-db.org/ (accessed date: 19 September 2022).

Appendix A.3. Covariance Matrix

The Brownian motion model: For an infinitesimal time change Δt, one can discretize $dy_t = \sigma dB_t$ into $y_t - y_{t-\Delta t} = \sigma(B_t - B_{t-\Delta t}), \Delta t = \frac{T}{m}$, m is an integer , which yields $y_t = y_{t-\Delta t} + \sigma(B_t - B_{t-\Delta t}) = \cdots = y_0 + \sigma(B_t - B_0)$. For a pair of species i and j, the trait variables $y_{i,t}$ and $y_{j,t}$ have covariance $cov(y_{i,t}, y_{j,t}) = cov[y_{i,0} + \sigma(B_{i,t} - B_{i,0}), y_{j,0} + \sigma(B_{j,t} - B_{j,0})] = \sigma^2 cov[B_{i,t}, B_{j,t}] = \sigma^2 t_a$ where $B_{i,t} = B_{j,t}, 0 \leq t \leq t_a$ and $B_{i,t}, B_{j,t}$ are independent for $s < t \leq T$.

Referring to Figure 1, the matrix transformed from the tree is represented in Equation (A1).

$$\mathbf{G} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 40+60 & 60 & 0 & 0 & 0 \\ 60 & 40+60 & 0 & 0 & 0 \\ 0 & 0 & 70+30 & 30 & 30 \\ 0 & 0 & 30 & 20+50+30 & 50+30 \\ 0 & 0 & 30 & 50+30 & 20+50+30 \end{pmatrix} \end{array}. \tag{A1}$$

The Ornstein–Uhlenbeck model:

For an infinitesimal time change Δt, one can discretize Equation (A2) to $y_{t+\Delta t} = y_t + \alpha(\theta - y_t)\Delta t + \sigma(B_{t+\Delta t} - B_t)$. The mapping is $t_a \mapsto \frac{e^{-2\alpha(t-t_a)}}{2\alpha} - \frac{e^{-2\alpha t}}{2\alpha}$. Referring to Figure 1, the matrix transformed from the tree is represented in Equation (A2).

$$
\begin{aligned}
\boldsymbol{G}_\alpha = \frac{1}{2\alpha}
\begin{array}{cc}
 & \begin{array}{ccccc} A & B & C & D & E \end{array} \\
\begin{array}{c} A \\ B \\ C \\ D \\ E \end{array} &
\begin{pmatrix}
1 & e^{-2\alpha(40)} & 0 & 0 & 0 \\
e^{-2\alpha(40)} & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & e^{-2\alpha(70)} & e^{-2\alpha(70)} \\
0 & 0 & e^{-2\alpha(70)} & 1 & e^{-2\alpha(20)} \\
0 & 0 & e^{-2\alpha(70)} & e^{-2\alpha(20)} & 1
\end{pmatrix}
\end{array} \\
- \frac{1}{2\alpha} e^{-2\alpha(100)}
\begin{array}{cc}
 & \begin{array}{ccccc} A & B & C & D & E \end{array} \\
\begin{array}{c} A \\ B \\ C \\ D \\ E \end{array} &
\begin{pmatrix}
1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 & 1
\end{pmatrix}
\end{array}.
\end{aligned}
\tag{A2}
$$

The early burst model:

This model includes a rate function, represented as $\sigma_t^2 = \sigma_0^2 \exp(rt), r < 0$. Referring to Figure 1, the matrix transformed from the tree is represented in Equation (A3).

$$
\boldsymbol{G}_r =
\begin{array}{cc}
 & \begin{array}{ccccc} A & B & C & D & E \end{array} \\
\begin{array}{c} A \\ B \\ C \\ D \\ E \end{array} &
\begin{pmatrix}
e^{-(40+60)r} & e^{-60r} & 0 & 0 & 0 \\
e^{-60r} & e^{-(40+60)r} & 0 & 0 & 0 \\
0 & 0 & e^{-(70+30)r} & e^{-30r} & e^{-30r} \\
0 & 0 & e^{-30r} & e^{-(20+50+30)r} & e^{-(50+30)r} \\
0 & 0 & e^{-30r} & e^{-(50+30)r} & e^{-(20+50+30)r}
\end{pmatrix}
\end{array}.
\tag{A3}
$$

The Pagel's λ, κ, and δ models:

Considering branch length stretching, referring to Figure 1, the matrices transformed from the tree are represented in Equations (A4)–(A6).

The Pagel's λ model

$$
\boldsymbol{G}_\lambda =
\begin{array}{cc}
 & \begin{array}{ccccc} A & B & C & D & E \end{array} \\
\begin{array}{c} A \\ B \\ C \\ D \\ E \end{array} &
\begin{pmatrix}
40+60 & \lambda \times 60 & 0 & 0 & 0 \\
\lambda \times 60 & 40+60 & 0 & 0 & 0 \\
0 & 0 & 70+30 & \lambda \times 30 & \lambda \times 30 \\
0 & 0 & \lambda \times 30 & 20+50+30 & \lambda \times (50+30) \\
0 & 0 & \lambda \times 30 & \lambda \times (50+30) & 20+50+30
\end{pmatrix}
\end{array}.
\tag{A4}
$$

The Pagel's δ model

$$
\boldsymbol{G}_\delta =
\begin{array}{cc}
 & \begin{array}{ccccc} A & B & C & D & E \end{array} \\
\begin{array}{c} A \\ B \\ C \\ D \\ E \end{array} &
\begin{pmatrix}
(40+60)^\delta & 60^\delta & 0 & 0 & 0 \\
60^\delta & (40+60)^\delta & 0 & 0 & 0 \\
0 & 0 & (70+30)^\delta & 30^\delta & 30^\delta \\
0 & 0 & 30^\delta & (20+50+30)^\delta & (50+30)^\delta \\
0 & 0 & 30^\delta & (50+30)^\delta & (20+50+30)^\delta
\end{pmatrix}
\end{array}.
\tag{A5}
$$

The Pagel's κ model

$$\boldsymbol{G}_{\kappa} = \left(\begin{array}{cccccc} & A & B & C & D & E \\ A & 40^{\kappa}+60^{\kappa} & 60^{\kappa} & 0 & 0 & 0 \\ B & 60^{\kappa} & 40^{\kappa}+60^{\kappa} & 0 & 0 & 0 \\ C & 0 & 0 & 70^{\kappa}+30^{\kappa} & 30^{\kappa} & 30^{\kappa} \\ D & 0 & 0 & 30^{\kappa} & 20^{\kappa}+50^{\kappa}+30^{\kappa} & 50^{\kappa}+30^{\kappa} \\ E & 0 & 0 & 30^{\kappa} & 50^{\kappa}+30^{\kappa} & 20^{\kappa}+50^{\kappa}+30^{\kappa} \end{array}\right). \tag{A6}$$

References

1. Garamszegi, L.Z. *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice*; Springer: Berlin/Heidelberg, Germany, 2014.
2. Paradis, E. An introduction to the phylogenetic comparative method. In *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 3–18.
3. Martins, E.P.; Martins, E.P. *Phylogenies and the Comparative Method in Animal Behavior*; Oxford University Press on Demand: Oxford, UK, 1996.
4. Choudhuri, S. *Bioinformatics for Beginners: Genes, Genomes, Molecular Evolution, Databases and Analytical Tools*; Elsevier: Amsterdam, The Netherlands, 2014.
5. Hall, B.G. Building phylogenetic trees from molecular data with MEGA. *Mol. Biol. Evol.* **2013**, *30*, 1229–1235.
6. Bouckaert, R.; Heled, J.; Kühnert, D.; Vaughan, T.; Wu, C.H.; Xie, D.; Suchard, M.A.; Rambaut, A.; Drummond, A.J. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **2014**, *10*, e1003537.
7. Furness, A.I.; Capellini, I. The evolution of parental care diversity in amphibians. *Nat. Commun.* **2019**, *10*, 1–12.
8. Woolley, C.H.; Thompson, J.R.; Wu, Y.H.; Bottjer, D.J.; Smith, N.D. A biased fossil record can preserve reliable phylogenetic signal. *Paleobiology* **2022**, *48*, 1–16.
9. Polly, P.D.; Stayton, C.T.; Dumont, E.R.; Pierce, S.E.; Rayfield, E.J.; Angielczyk, K.D. Combining geometric morphometrics and finite element analysis with evolutionary modeling: Towards a synthesis. *J. Vertebr. Paleontol.* **2016**, *36*, e1111225.
10. Davies, E.L.; Arbuckle, K. Coevolution of snake venom toxic activities and diet: Evidence that ecological generalism favours toxicological diversity. *Toxins* **2019**, *11*, 711.
11. Krüger, O. The evolution of reversed sexual size dimorphism in hawks, falcons and owls: A comparative study. *Evol. Ecol.* **2005**, *19*, 467–486.
12. Pagel, M.; Meade, A.; Barker, D. Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.* **2004**, *53*, 673–684.
13. Beaulieu, J.; Jhwueng, D.C.; Boettiger, C.; O'Meara, B. Modeling stabilizing selection: Expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution* **2012**, *66*, 2369–2383.
14. Besnard, F.; Picao-Osorio, J.; Dubois, C.; Félix, M.A. A broad mutational target explains a fast rate of phenotypic evolution. *Elife* **2020**, *9*, e54928.
15. Pannetier, T.; Martinez, C.; Bunnefeld, L.; Etienne, R.S. Branching patterns in phylogenies cannot distinguish diversity-dependent diversification from time-dependent diversification. *Evolution* **2021**, *75*, 25–38.
16. Xue, B.; Guo, X.; Landis, J.; Sun, M.; Tang, C.; Soltis, P.; Soltis, D.; Saunders, R. Accelerated diversification correlated with functional traits shapes extant diversity of the early divergent angiosperm family Annonaceae. *Mol. Phylogenet. Evol.* **2020**, *142*, 106659.
17. O'Meara, B.C. Evolutionary inferences from phylogenies: A review of methods. *Annu. Rev. Ecol. Evol. Syst.* **2012**, *43*, 267–285.
18. Vasconcelos, T.; O'Meara, B.C.; Beaulieu, J.M. A flexible method for estimating tip diversification rates across a range of speciation and extinction scenarios. *Evolution* **2022**, *76*, 1420–1433.
19. Duchen, P.; Alfaro, M.L.; Rolland, J.; Salamin, N.; Silvestro, D. On the effect of asymmetrical trait inheritance on models of trait evolution. *Syst. Biol.* **2021**, *70*, 376–388.
20. Smaers, J.B.; Dechmann, D.K.; Goswami, A.; Soligo, C.; Safi, K. Comparative analyses of evolutionary rates reveal different pathways to encephalization in bats, carnivorans, and primates. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 18006–18011.
21. Medeiros, A.P.; Santos, B.A.; Betancur-R, R. Does Genome Size Increase with Water Depth in Marine Fishes? *Evolution* **2022**, *76*, 1578–1589.
22. Bickford, D.; Lohman, D.; Navjot, S.; Ng, P.; Meier, R.; Winker, K.; Ingram, K.; Feinberg, J.; Newman, C.; Watkins-Colwell, G.; et al. AmphibiaWeb. 2017. Available online: http://amphibiaweb.org (accessed on 19 September 2022).
23. Uetz, P.; Koo, M.S.; Aguilar, R.; Brings, E.; Catenazzi, A.; Chang, A.; Wake, D. A quarter century of reptile and amphibian databases. *Herpetol. Rev.* **2021**, *52*, 246–255.
24. Parr, C.L.; Dunn, R.R.; Sanders, N.J.; Weiser, M.D.; Photakis, M.; Bishop, T.R.; Fitzpatrick, M.C.; Arnan, X.; Baccaro, F.; Brandão, C.R.; et al. GlobalAnts: A new database on the geography of ant traits (Hymenoptera: Formicidae). *Insect Conserv. Divers.* **2017**, *10*, 5–20.
25. Pauly, D.; Froese, R. FishBase. 2010. Available online: https://www.fishbase.se/search.php (accessed on 19 September 2022)
26. Schreiber, F.; Patricio, M.; Muffato, M.; Pignatelli, M.; Bateman, A. TreeFam v9: A new website, more species and orthology-on-the-fly. *Nucleic Acids Res.* **2014**, *42*, D922–D925.

27. Kattge, J.; Bönisch, G.; Díaz, S.; Lavorel, S.; Prentice, I.C.; Leadley, P.; Tautenhahn, S.; Werner, G.D.; Aakala, T.; Abedi, M.; et al. TRY plant trait database–enhanced coverage and open access. *Glob. Chang. Biol.* **2020**, *26*, 119–188.
28. Stoltzfus, A.; Lapp, H.; Matasci, N.; Deus, H.; Sidlauskas, B.; Zmasek, C.M.; Vaidya, G.; Pontelli, E.; Cranston, K.; Vos, R.; et al. Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient. *BMC Bioinform.* **2013**, *14*, 1–17.
29. Maddison, D.R.; Schulz, K.S.; Maddison, W.P. The tree of life web project. *Zootaxa* **2007**, *1668*, 19–40.
30. Gallagher, R.; Falster, D.; Maitner, B.; Salguero-Gomez, R.; Vandvik, V.; Pearse, W.; Enquist, B. Open Science principles for accelerating trait-based science across the Tree of Life. *Nat. Ecol. Evol.* **2020** *4*, 294–303.
31. Kleyer, M.; Bekker, R.; Knevel, I.; Bakker, J.; Thompson, K.; Sonnenschein, M.; Poschlod, P.; Van Groenendael, J.; Klimeš, L.; Klimešová, J.; et al. The LEDA Traitbase: A database of life-history traits of the Northwest European flora. *J. Ecol.* **2008**, *96*, 1266–1274.
32. Michonneau, F.; Brown, J.W.; Winter, D.J. rotl: An R package to interact with the Open Tree of Life data. *Methods Ecol. Evol.* **2016**, *7*, 1476–1481.
33. McTavish, E.J.; Sánchez-Reyes, L.L.; Holder, M.T. OpenTree: A Python package for Accessing and Analyzing data from the Open Tree of Life. *Syst. Biol.* **2021**, *70*, 1295–1301.
34. Huerta-Cepas, J.; Bueno, A.; Dopazo, J.; Gabaldón, T. PhylomeDB: A database for genome-wide collections of gene phylogenies. *Nucleic Acids Res.* **2007**, *36*, D491–D496.
35. Smith, S.D.; Ané, C.; Baum, D.A. The role of pollinator shifts in the floral diversification of Iochroma (Solanaceae). *Evol. Int. J. Org. Evol.* **2008**, *62*, 793–806.
36. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.
37. O'Meara, B. CRAN Task View: Phylogenetics, Especially Comparative Methods. 2020. Available online: https://cran.r-project.org/web/views/Phylogenetics.html (accessed on 2 November 2022).
38. Ané, C. Analysis of comparative data with hierarchical autocorrelation. *Ann. Appl. Stat.* **2008**, *2*, 1078–1102.
39. Revell, L.J. Phylogenetic signal and linear regression on species data. *Methods Ecol. Evol.* **2010**, *1*, 319–329.
40. Stone, E.A. Why the phylogenetic regression appears robust to tree misspecification. *Syst. Biol.* **2011**, *60*, 245–260.
41. Anderson, D.; Burnham, K. *Model Selection and Multi-Model Inference*, 2nd ed.; Springer: New York, NY, USA, 2004; Volume 63, p. 10.
42. Felsenstein, J. Phylogeny and the comparative method. *Am. Nat.* **1985**, *125*, 1–15.
43. Hansen, T.F. Stabilizing selection and the comparative analysis of adaptation. *Evolution* **1997**, *51*, 1341–1351.
44. Harmon, L.J.; Losos, J.B.; Jonathan Davies, T.; Gillespie, R.G.; Gittleman, J.L.; Bryan Jennings, W.; Kozak, K.H.; McPeek, M.A.; Moreno-Roark, F.; Near, T.J.; et al. Early bursts of body size and shape evolution are rare in comparative data. *Evol. Int. J. Org. Evol.* **2010**, *64*, 2385–2396.
45. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **1999**, *401*, 877.
46. Adams, D.C.; Collyer, M.L. Multivariate phylogenetic comparative methods: Evaluations, comparisons, and recommendations. *Syst. Biol.* **2018**, *67*, 14–31.
47. Billingsley, P. *Probability and Measure*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
48. Hansen, T.; Martins, E. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution* **1996**, *50*, 1404–1417.
49. Butler, M.; King, A. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *Am. Nat.* **2004**, *164*, 683–695.
50. Housworth, E.A.; Martins, E.P.; Lynch, M. The phylogenetic mixed model. *Am. Nat.* **2004**, *163*, 84–96.
51. Harmon, L.J. Phylogenetic Comparative Methods. 2019. Available online: https://lukejharmon.github.io/pcm/ (accessed on 2 November 2022).
52. Revell, L.J.; Harmon, L.J. *Phylogenetic Comparative Methods in R*; Princeton University Press: Princeton, NJ, USA, 2022.
53. Hansen, T.; Pienaar, J.; Orzack, S. A Comparative Method for Studying Adaptation to a Randomly Evolving Environment. *Evolution* **2008**, *62*, 1965–77.
54. Ho, L.S.T.; Ane, C.; Lachlan, R.; Tarpinian, K.; Feldman, R.; Yu, Q.; van der Bijl, W.; Maspons, J.; Vos, R.; Ho, M.L.S.T. Package 'Phylolm'. 2016. Available online: http://cran.r-project.org/web/packages/phylolm/index.html (accessed 2 February 2021)
55. Castiglione, S.; Serio, C.; Mondanaro, A.; Melchionna, M.; Carotenuto, F.; Di Febbraro, M.; Profico, A.; Tamagnini, D.; Raia, P. Ancestral state estimation with phylogenetic ridge regression. *Evol. Biol.* **2020**, *47*, 220–232.
56. Smith, M.R. TreeTools: Create, Modify and Analyse Phylogenetic Trees. Comprehensive R Archive Network. 2019. R Package Version 1.7.3. Available online: https://doi.org/10.5281/zenodo.3522725 (accessed on 19 September 2022).
57. Bossio, M.C.; Cuervo, E.C. Gamma regression models with the Gammareg R package. *Comun. En Estadística* **2015**, *8*, 211–223.
58. Ives, A.R.; Garland, T., Jr. Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.* **2010**, *59*, 9–26.
59. Paradis, E.; Claude, J.; Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **2004**, *20*, 289–290. https://doi.org/10.1093/bioinformatics/btg412.
60. Cribari-Neto, F.; Zeileis, A. Beta regression in R. *J. Stat. Softw.* **2010**, *34*, 1–24.
61. O'Meara, B.; Ané, C.; Sanderson, M.; Wainwright, P. Testing different rates of continuous trait evolution using likelihood. *Evolution* **2006**, *60*, 922–933.

62. Manceau, M.; Lambert, A.; Morlon, H. A unifying comparative phylogenetic framework including traits coevolving across interacting lineages. *Syst. Biol.* **2017**, *66*, 551–568.
63. Jhwueng, D.C.; Maroulas, V. Phylogenetic ornstein–uhlenbeck regression curves. *Stat. Probab. Lett.* **2014**, *89*, 110–117.
64. Jhwueng, D.C.; Maroulas, V. Adaptive trait evolution in random environment. *J. Appl. Stat.* **2016**, *43*, 2310–2324.
65. Bastide, P.; Solís-Lemus, C.; Kriebel, R.; William Sparks, K.; Ané, C. Phylogenetic comparative methods on phylogenetic networks with reticulations. *Syst. Biol.* **2018**, *67*, 800–820.
66. Jhwueng, D.C. Reduced drought tolerance during domestication and the evolution of weediness results from tolerance–growth trade-offs. *Evol. Int. J. Org. Evol.* **2012**, *66*, 3803–3814.
67. Uyeda, J.C.; Harmon, L.J. A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Syst. Biol.* **2014**, *63*, 902–918.
68. Villemereuil, P.d.; Wells, J.A.; Edwards, R.D.; Blomberg, S.P. Bayesian models for comparative analysis integrating phylogenetic uncertainty. *BMC Evol. Biol.* **2012**, *12*, 1–16.
69. Anscombe, F.J. Graphs in statistical analysis. *Am. Stat.* **1973**, *27*, 17–21.
70. Revell, L.J.; Schliep, K.; Valderrama, E.; Richardson, J.E. Graphs in phylogenetic comparative analysis: Anscombe's quartet revisited. *Methods Ecol. Evol.* **2018**, *9*, 2145–2154.
71. Liang, X.; Wang, Z.; Zhou, Z.; Huang, Z.; Zhou, J.; Cen, K. Up-to-date life cycle assessment and comparison study of clean coal power generation technologies in China. *J. Clean. Prod.* **2013**, *39*, 24–31.
72. Yoon, S.H.; Ha, S.M.; Kwon, S.; Lim, J.; Kim, Y.; Seo, H.; Chun, J. Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int. J. Syst. Evol. Microbiol.* **2017**, *67*, 1613.
73. Huerta-Cepas, J.; Capella-Gutierrez, S.; Pryszcz, L.P.; Marcet-Houben, M.; Gabaldon, T. PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.* **2014**, *42*, D897–D902.
74. Adler, P.B.; Fajardo, A.; Kleinhesselink, A.R.; Kraft, N.J. Trait-based tests of coexistence mechanisms. *Ecol. Lett.* **2013**, *16*, 1294–1306.
75. Piel, W.H.; Donoghue, M.; Sanderson, M.; Netherlands, L. TreeBASE: A database of phylogenetic information. In Proceedings of the 2nd International Workshop of Species, Copenhagen, Denmark, 22–26 May 2000; Volume 2000.