Improving the Robustness of Phylogenetic Independent Contrasts: Addressing Abrupt Evolutionary Shifts with Outlier- and Distribution-Guided Correlation

Zheng-Lin Chen, Rui Huang, Hong-Ji Guo and Deng-Ke Niu*

MOE Key Laboratory for Biodiversity Science and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing, 100875, China

*Correspondence to be sent to: *Deng-Ke Niu, College of Life Sciences, Beijing Normal University, Beijing, 100875, China*; Emails: dkniu@bnu.edu.cn, dengkeniu@hotmail.com

ABSTRACT

Traditional implementations of phylogenetically independent contrasts (PIC), using Pearson correlation or ordinary least squares (OLS) regression—mathematically equivalent under linear assumptions in detecting trait associations—and regression-based methods like phylogenetic generalized least squares (PGLS), perform well under gradual evolutionary processes. However, abrupt evolutionary shifts—or macroevolutionary jumps characteristic of punctuated evolution—can produce extreme contrasts that inflate false positives or obscure genuine associations. We introduce O(D)GC (Outlier- and Distribution-Guided Correlation), a flexible workflow that first detects outliers by identifying contrasts exceeding *k* times the median absolute deviation (*k* scaled to sample size), then applies Spearman correlation if any outliers are found. If no outliers are detected, Pearson correlation is used—automatically for large datasets ($n \geq 30$), or guided by a normality test for small ones. We systematically compared seven methods—PIC-O(D)GC; PIC-Pearson; PIC-MM, a robust phylogenetic regression method; PGLS; PGLMM (phylogenetic generalized linear mixed model); MR-PMM (multi-response phylogenetic mixed model); and Corphylo, an Ornstein–Uhlenbeck-based method—on 220,000 simulated datasets spanning five evolutionary scenarios (two shift settings: single-trait shifts and dual-trait co-directional jumps; and three no-shift gradual evolution settings), each tested across 11 shift or noise gradients, three tree sizes (16, 128, 256 tips), and both balanced and random topologies. Across all shift gradients, PIC-O(D)GC consistently yielded the lowest or among the lowest false-positive rates, while maintaining accuracy comparable to PIC-MM. Under high-noise, no-shift conditions, it retained power similar to other methods but ran orders of magnitude faster than matrix-inversion models. Analyses of three empirical datasets likewise showed that PIC-O(D)GC and PIC-MM corrected shift-induced distortions that misled conventional methods. While matching PIC-MM in robustness, PIC-O(D)GC offers greater conceptual simplicity and markedly lower computational cost. Its correlation-only output provides less mechanistic detail than regression-based approaches like PGLS. However, when paired with PIC diagnostics, this outlier-guided strategy not only tests overall trait associations, but also highlights evolutionary jumps, distinguishes coupled from decoupled shifts, and—via clade partitioning or tip pruning—recovers background correlations, offering biologically informative insights into how punctuated events interact with gradual trends in trait evolution.

KEYWORDS

Phylogenetic comparative methods; Phylogenetically independent contrasts (PIC); Outlier- and distribution-guided correlation (PIC-O(D)GC); Abrupt evolutionary shifts; Robust regression; Trait correlation; Brownian motion (BM) evolution; Outlier detection.

INTRODUCTION

Phylogenetically closer species tend to exhibit more similar traits due to their shared ancestry, resulting in phylogenetic autocorrelation among samples (Felsenstein 1985, Garamszegi 2014, Cornwallis and Griffin 2024). Ignoring this autocorrelation and using conventional statistical methods can lead to erroneous conclusions. To address this challenge, evolutionary biologists have developed a range of phylogenetic comparative methods that explicitly account for these dependencies (Grafen 1989, Lynch 1991, Garamszegi 2014, O'Meara 2016, Cornwallis and Griffin 2024, Dewar et al. 2025).

The earliest method was phylogenetically independent contrasts (PIC) (Felsenstein 1985). The core principle behind PIC is that, while species at the tips of a phylogenetic tree are interdependent due to shared ancestry, the evolutionary changes along each branch can be treated as statistically independent events. In this context, PIC refers to the contrast values calculated between species pairs along the phylogenetic tree, inferring evolutionary changes along pairs of branches and standardizing these changes by branch length (or time) to generate transformed data points. These PIC values, typically subjected to ordinary least squares (OLS) regression, provide insight into the relationships between traits. This approach makes the evolutionary data more suitable for conventional statistical methods by removing the phylogenetic non-independence. Thus, PIC can refer both to the contrast values calculated from the phylogeny and to the subsequent OLS regression analysis applied to these values.

Applying OLS regression to PIC values assumes that traits evolve according to a Brownian motion (BM) model, under which trait variation accumulates gradually and randomly over time. While the computation of PIC values does not strictly require a BM model, the statistical justification for treating PIC values as independent and identically distributed (i.i.d.) variables in regression analyses typically relies on the BM assumption or its extensions. Under this framework, OLS regression through the origin on PIC values yields slope estimates, standard errors, and *p*-values that are numerically identical to those obtained from phylogenetic generalized least squares (PGLS), assuming BM holds (Rohlf 2001, Blomberg et al. 2012). PGLS generalizes this approach by directly modeling the phylogenetic covariance structure through a variance–covariance matrix of the residuals in the linear model, thereby accommodating phylogenetic autocorrelation among observations (Martins and Hansen 1997, Pagel 1997, Pagel 1999). This adjustment helps reduce false positives in hypothesis testing due to shared evolutionary history. Although some PGLS models, such as

Pagel's λ, Ornstein-Uhlenbeck (OU), and Early-burst (EB) models, can handle heterogeneous evolutionary rates (Hansen 1997, Freckleton et al. 2002, Harmon et al. 2010, Ho and Ane 2014), they fundamentally assume gradual trait changes (Landis et al. 2013). This assumption can be problematic when dealing with evolutionary jumps caused by singular events.

During biological evolution, certain branches may experience rare and dramatic trait changes (Maddison and FitzJohn 2015). Increasingly, studies and data have shown that abrupt evolutionary shifts, or evolutionary jumps, have a significant impact on biodiversity evolution (Landis and Schraiber 2017, Gao and Wu 2022, Smith et al. 2023, Sumner et al. 2023). A typical example of this is Felsenstein's worst-case scenario (Felsenstein 1985, Uyeda et al. 2018), where two traits undergo a significant jump near the root of the phylogenetic tree while evolving independently on other branches. Even if this dramatic change occurred only once, it could still lead to erroneous conclusions of significant correlation between the two traits using ordinary linear regression on PIC values or PGLS models (Uyeda et al. 2018).

Recently, more studies have focused on identifying jump events and providing more reasonable explanations for the processes underlying trait evolution. Uyeda et al. (2018) underscore the need for models like the singular events model, which can accommodate the complexities and unpredictabilities inherent in evolutionary biology. Some studies suggest that jump events cause trait changes to exhibit a "fat-tails" distribution, using Levy processes to handle this distribution (Landis et al. 2013, Duchen et al. 2017). Additionally, LASSO-based methods within models like OU have been applied to pinpoint shifts along specific branches of phylogenetic trees (Khabbazian et al. 2016, Bastide et al. 2018, Zhang et al. 2024). Others propose that abrupt shift events lead to outliers, using robust regression to reduce model sensitivity to violations (Slater and Pennell 2014, Adams et al. 2024). A variety of mixed-model approaches have also been explored to address model deviations and trait interdependence (Ives and Helmus 2011, Westoby et al. 2023, Halliwell et al. 2025).

Evolutionary shifts, especially those occurring in branches close to the root of the phylogenetic tree, typically do not manifest directly as outliers in the trait data but instead appear as outliers in the PIC values. As a result, robust statistical methods designed to handle outliers in the original trait data may not be as effective in addressing evolutionary shifts. Adams et al. (2024) demonstrated that when four different estimators (L1, S, M, and MM) were integrated into both PGLS and PIC analysis, methods like PIC-MM and PIC-S performed significantly better in detecting evolutionary relationships, especially when

accounting for outliers. Thus, computing PIC values is essential for effective analysis of the relationships between traits with evolutionary shifts.

Dramatic changes in some evolutionary branches can result in several outliers in the PIC dataset (Uyeda et al. 2018), distorting the statistical results of parametric tests. Nonparametric tests, which measure the strength and direction of association between two variables without assuming a specific distribution, can effectively handle data with outliers, reducing false positives. Therefore, we propose a hybrid framework, Outlier-Guided Correlation (OGC), to optimize the analysis of trait relationships under diverse evolutionary conditions. This framework integrates the existing PIC method with a dynamic selection mechanism that evaluates the presence of outliers in trait contrasts. For larger datasets, OGC selects Spearman or Pearson correlation depending on whether outliers are detected. For smaller datasets, where non-normality can significantly impact results, the framework incorporates both normality testing and outlier detection, referred to as Outlier- and Distribution-Guided Correlation (ODGC). Together, these approaches form a unified framework collectively referred to as O(D)GC, designed to address the challenges posed by varying data conditions. It should be noted that, unlike regression-based methods such as PGLS that aim to estimate model parameters, the O(D)GC framework is designed for null hypothesis testing—to assess whether significant pairwise trait correlations exist. Readers are encouraged to interpret its results accordingly. A more detailed discussion of its conceptual scope and limitations is provided in the Discussion.

Our analysis demonstrates that PIC-O(D)GC effectively reduces false positives in the presence of strong abrupt shifts, while maintaining high accuracy. When evolutionary shifts are weak or absent, PIC-O(D)GC performs comparably to established phylogenetic methods, including PGLS (optimized across five evolutionary models), PIC-Pearson, robust regression methods, Corphylo (Zheng et al. 2009), PGLMM, and MR-PMM. Its performance closely aligns with that of PIC-MM, with the added advantage of computational simplicity. This consistency across a wide range of evolutionary scenarios highlights O(D)GC as a reliable and scalable tool for analyzing trait correlations in phylogenetic datasets, complementing other robust methods such as PIC-MM.

MATERIALS AND METHODS

Phylogenetic Tree Construction and Overview of Simulation Strategy

We categorize the construction of phylogenetic trees into two types. The first is the fixed-balanced tree scenario, where all non-leaf nodes have left and right subtrees with identical topology and depth, forming a fully symmetric structure. In this setting, a single fixed-balanced phylogenetic tree with 128 species is used as the basis for all trait simulations. Specifically, traits are simulated under combinations of evolutionary scenarios and gradient settings (reflecting either trait noise or trait shifts), with 1,000 replicates per scenario-gradient combination, using different random seeds while keeping the phylogenetic tree fixed. As a result, for each trait evolutionary scenario-gradient combination, 1,000 datasets are generated based on the same phylogenetic structure. To assess the influence of tree size, additional analyses were conducted using 16- and 256-species fixed-balanced trees, following the same simulation protocol.

The second category is the random unfixed-balanced tree scenario (hereafter referred to as the random tree scenario), in which a new phylogenetic tree is generated prior to each trait simulation. These trees are simulated under a coalescent model with a fixed root split: the root ancestor always bifurcates into two subtrees of identical size and identical topology within a simulation, and this shared topology is randomly generated anew for each simulation. Specifically, for each replicate, a random phylogenetic tree is first generated, followed by trait simulation under a given evolutionary scenario and gradient condition. This pair of operations—tree construction and trait simulation—is repeated 1,000 times using different random seeds, ensuring variation in both tree topology and simulated trait values. As a result, for each trait evolutionary scenario-gradient combination, 1,000 distinct trees and 1,000 corresponding trait datasets are produced.

Subsequently, we evaluated the performance of each trait correlation inference method across all trait evolutionary scenario–gradient combinations using two complementary benchmarks for determining "true" trait correlations. The primary benchmark is a phylogenetically informed method that considers evolutionary independence and data characteristics, while the secondary benchmark is a model-based criterion derived directly from the simulation settings. For each of the 1,000 simulated datasets per scenario-gradient combination, either based on the same fixed phylogenetic tree or on 1,000 distinct trees generated per replicate under the coalescent model, the correlation results produced by each method were compared against the primary benchmark. A method was deemed correct for a given dataset if both the direction and statistical significance ($p < 0.05$ or not) of its inferred correlation matched the benchmark result; otherwise, it was marked as incorrect. By calculating the proportion of correct predictions among the 1,000 replicates, we obtained the

accuracy of each method for that trait evolutionary scenario–gradient combination. In addition, we computed the false positive rate—the proportion of datasets where no true correlation existed but a method incorrectly inferred significance—and the false negative rate—the proportion of datasets with a true correlation that a method failed to detect.

Results obtained using the secondary benchmark were largely consistent with those from the primary benchmark, confirming the robustness of the performance evaluations. Additional methodological details concerning the benchmarking strategy and simulation configurations are provided in the following sections.

Simulation of Independent Traits with Shared Directional Shifts on Deep Branches

With reference to previous simulation work of trait shift by Uyeda et al. (2018), two traits (X_1 and X_2) were independently simulated on a phylogenetic tree, with $X_2 = 0 \times X_1 + e$, where both X_1 and e underwent an abrupt shift of equal magnitude on one of the branches closest to the root, while on all other branches they evolved independently under the BM model with a variance of 1. The shift magnitude ranged from 4^{-3} to 4^{7}, where $4^{0} = 1$ represents no shift. Positive exponents indicate upsizing (increased trait values), while negative exponents indicate downsizing (decreased trait values). Each shift gradient was simulated 1000 times.

To simulate this deep-rooted abrupt shift, we first split the phylogenetic tree at the root into two subtrees—background_subtree and shift_subtree—using the *extract.clade()* function from the *ape* package (version 5.8) (Paradis and Schliep 2019). For both traits, the background_subtree used a baseline root value of 1, while the shift_subtree used a shifted root value (1 × shift magnitude). Trait evolution was then simulated independently under BM using the *rTraitCont()* from the *ape* package (evolutionary rate = 1). Independent random seeds were used to ensure X_1 and X_2 remained uncorrelated. This simulation was used mainly to evaluate the false positive rates of each phylogenetic method—i.e., whether a correlation is mistakenly detected when none truly exists.

Simulation of Correlated Traits with Shift in the Predictor Only

We simulated two positively correlated traits (X_1 and X_2) on a phylogenetic tree, with the relationship defined as $X_2 = 1 \times X_1 + e$, where X_1 evolves under a BM process with a variance of 1, and the error term e follows a normal distribution (mean = 0, variance = 0.01), following the phylogenetic regression framework of Revell (2010) and Mazel et al. (2016). To assess the impact of directional changes in only one of the traits, an abrupt shift of varying

magnitude was introduced only to X_1 on one of the branches closest to the root. The shift magnitude followed the same range and interpretation as described above (4^{-3} to 4^7, with 4^0 = 1 as no shift). Each shift magnitude was simulated 1,000 times.

To implement this asymmetrical shift while preserving the pre-defined correlation structure, we first simulated X_1 across the entire tree (root = 1, rate = 1), then constructed X_2 as $X_2 = X_1 + e$. Next, we re-simulated X_1 only in the shift_subtree using the same random seed but a different root value (1 × shift magnitude). The final X_1 was assembled by combining the original background_subtree values and the re-simulated shift_subtree values, while X_2 remained unchanged. This setup allowed us to assess, primarily, the false negative rates of each method by testing how often datasets with a true correlation were mistakenly identified as uncorrelated due to directional shifts applied only one trait.

Simulated Traits Without Shifts: Fixed Correlation with Varying Noise

Following the core simulation framework used by Revell (2010) and Mazel et al. (2016), in which one trait is modeled as a linear function of another plus noise, we generated two traits (X_1 and X_2) using the model $X_2 = 1 \times X_1 + e$, where e is a noise term with adjustable variance. This construction prescribes a causal link between the two traits; hence the simulations function mostly as power analyses—assessing how well each method recovers the intended association across noise gradients—while recognizing that sufficiently large noise can render some replicates only incidentally uncorrelated. Based on this framework, we simulated the following scenarios to reflect different evolutionary processes:

1. BM1 & BM1+BM2: X_1 evolves under a BM process with a variance of 1, while e follows another BM process with a variance of d, leading to X_2 representing the combined BM1 and BM2 processes (BM1+BM2).

2. BM & BM+Norm: X_1 evolves under a BM process with a variance of 1, and e follows a normal distribution (mean = 0, variance = d), resulting in X_2 combining phylogenetic and non-phylogenetic components. Because e is i.i.d., the residuals of X_2 are independent, satisfying the assumptions of OLS. This scenario thus serves as the only setting where OLS is theoretically appropriate and included for comparison (Revell 2010).

3. Norm & Norm+BM: X_1 follows a normal distribution (mean = 0, variance = 1), and e evolves under a BM process (variance = d), making X_2 represent the combination of the normal distribution and the BM process (Norm+BM).

The d takes values of 4^{-3}, 4^{-2}, 4^{-1}, 4^{0}, 4^{1}, 4^{2}, 4^{3}, 4^{4}, 4^{5}, 4^{6}, and 4^{7} in all the three scenarios. For each scenario and value of d, we conducted 1000 replicate simulations, totaling 33,000 simulations for each tree type.

This study did not include the fourth scenario from Revell (2010), Norm1 & Norm1+Norm2, in which Trait X_1 follows a normal distribution with a mean of 0 and a variance of 1, while X_2 is constructed as X_1 plus a noise term e that also follows a normal distribution (mean = 0, variance = d). Because both X_1 and e are phylogenetically unstructured, this scenario results in simulated data with little to no phylogenetic autocorrelation. While such a configuration is often used as a control for the absence of phylogenetic signal, we considered it redundant for our purposes. In our first three simulation scenarios, large values of the noise parameter d similarly weaken or eliminate phylogenetic structure—thereby incidentally producing datasets in which the traits are statistically uncorrelated despite a fixed causal relationship. As such, the role of a non-phylogenetic control is already fulfilled within our existing framework, and a separate scenario was deemed unnecessary.

Empirical Datasets

To evaluate the performance of different phylogenetic comparative methods when analyzing data that may contain abrupt evolutionary shifts, we used three empirical datasets. Two of them were previously analyzed by Adams et al. (2024) in a study of phylogenetically robust regression.

The first dataset was obtained from the supplementary material of Makino and Kawata (2019), and includes propagule size, genome size, and a phylogenetic tree for 44 invertebrate species. Following the original study, both traits were $\log_{10}$-transformed prior to analysis to reduce skewness and improve normality assumptions. Propagule size in this context refers to the physical size (cm) of a reproductive unit that is separated from the parent and capable of developing into a new individual. In ecological and evolutionary studies, propagule size is widely used as a key indicator of an organism's reproductive strategy, particularly along the r/K selection continuum (Romiguier et al. 2014). The phylogenetic tree provided by the authors was used without modification.

The second dataset was originally compiled by Scales et al. (2009) and includes the relative proportions of fast glycolytic (FG) and slow oxidative (SO) muscle fibers in the limb muscles of 22 lizard species. This dataset was later used by Uyeda et al. (2018) to illustrate

the limitations of phylogenetic comparative methods, particularly in cases involving abrupt shifts or convergent trait evolution. We used the phylogenetic tree provided by Uyeda et al. (2018).

The third dataset includes body mass and population density for 109 species of Artiodactyla. Trait data were obtained from the PanTHERIA database (accessed June 4, 2025) (Jones et al. 2009). Species with missing values for either trait were excluded. Both body mass (grams) and population density (individuals per km^2) were log_{10}-transformed prior to analysis to reduce skewness and better meet the assumptions of linear models and comparative methods. The phylogenetic tree used for the Artiodactyla clade was extracted from the time-scaled maximum clade credibility tree for mammals developed by Upham et al. (2019), available at VertLife.org (accessed on June 4, 2025).

Statistical Methods Compared

To evaluate the performance of PIC-OGC examining the evolutionary relationship between X_1 and X_2 across two major evolutionary scenarios—with and without abrupt evolutionary jumps—we compared it with several previously established phylogenetic comparative methods:

1. PIC-Pearson: Pearson correlation applied to PIC values; because it yields the same direction and significance inferences as an OLS regression on the same PICs, running a separate PIC-OLS analysis would be redundant, so we used only PIC-Pearson.

2. PGLS: Using five evolutionary models, BM (Felsenstein 1985), Pagel's λ (Pagel 1999), OU fixed, OU random (Hansen 1997), and EB (Harmon et al. 2010).

3. Robust phylogenetic regression with both PIC and PGLS frameworks: For each framework, we utilized four robust estimators: L1, S, M, and MM. These were implemented as PIC-L1, PIC-S, PIC-M, and PIC-MM for PIC-based methods, and PGLS-L1, PGLS-S, PGLS-M, and PGLS-MM for PGLS-based methods (Adams et al. 2024).

4. Corphylo: A method that calculates Pearson correlation coefficients for multiple continuous traits, allowing users to incorporate phylogenetic signal and measurement errors into the model (Zheng et al. 2009). This method assumes an OU process for trait evolution, making it particularly suitable for traits that deviate from the BM model. Corphylo serves a dual purpose: it enables both the detection of phylogenetic signals in individual traits and the examination of correlations between multiple traits. A key feature of Corphylo is its ability to

include measurement errors of trait data, potentially improving accuracy under real-world conditions.

5. PGLMM and **6. MR-PMM**: The phylogenetic generalized linear mixed models (PGLMM) incorporates both fixed and random effects to account for phylogenetic autocorrelation, offering flexibility for a range of evolutionary scenarios (Ives and Helmus 2011). The multi-response phylogenetic mixed models (MR-PMM) extends this framework to multiple traits, enabling simultaneous modeling of trait co-evolution (Westoby et al. 2023, Halliwell et al. 2025). These mixed-model approaches are particularly suitable when trait relationships are influenced by unobserved factors or involve hierarchical structure. They provide a theoretically robust alternative to conventional phylogenetic regression methods, especially when BM assumptions are violated.

7. Raw-Pearson: The standard Pearson correlation applied directly to the original trait values, without any phylogenetic correction. Because Pearson (or its OLS equivalent) assumes i.i.d. residuals, we report Raw-Pearson results only for the BM and BM + Norm scenarios—the only conditions in our simulations that satisfy this assumption. Although Pearson/OLS methods are seldom used in phylogenetic comparative analyses due to the non-independence created by shared ancestry, they remain a useful theoretical benchmark when their statistical requirements are met (Revell 2010).

Outlier- and Distribution-Guided Phylogenetic Correlation Analysis

To address the influence of evolutionary shifts and outliers in phylogenetic comparative analysis, we introduced a novel framework: PIC-OGC. In this method, the choice between Pearson and Spearman correlation analysis is guided by the identification of outliers within the PIC data. The process starts by calculating the PICs for each trait using the phylogenetic tree and determining their absolute deviations from the median PIC of the dataset. A PIC is defined as an outlier if its absolute deviation from the median exceeds $T_i = k \times MAD$, where T_i represents the threshold for detecting outliers, k is a sample-size-adjusted factor, and MAD (Median Absolute Deviation) is the median of the absolute deviations from the median PIC of the dataset (Wilcox 2003, Rindskopf and Shiyko 2010, Leys et al. 2013).

For a normal distribution, MAD is approximately equal to 0.67449 times the standard deviation (Iglewicz and Hoaglin 1993, Wilcox 2003). The expected coverage of the data under different multiples of MAD (Rindskopf and Shiyko 2010) is shown in Table 1, which also provides the corresponding expected number of false outliers detected under each

threshold. As the threshold increases, the data coverage percentage approaches 100%, leading to a sharp decline in the expected number of incorrectly flagged outliers. This reduction scales inversely with both the threshold and dataset size, reflecting how stricter thresholds provide greater specificity at the cost of identifying fewer potential outliers.

TABLE 1. Data coverage and expected false outliers across *MAD* thresholds.

Threshold ($k \times MAD$)	Coverage for normal distribution	Expected number of false outliers		
		16 Species	128 Species	256 Species
3	95.69753%	0.6	5	11
4	99.30234%	0.1	0.9	1.8
5	99.92550%	0.01	0.09	0.19
6	99.99481%	0.0008	0.007	0.013
7	99.99977%	0.00004	0.0003	0.0006

This table presents the expected number of false outliers and the corresponding data coverage percentages for various outlier detection thresholds, defined as multiples of the Median Absolute Deviation (*MAD*), under the assumption of a normal distribution. Since *MAD* is approximately equal to 0.67449 times the standard deviation (σ) under a normal distribution, the thresholds were converted to equivalent z-scores by multiplying k with the ratio MAD/σ=0.67449. The corresponding coverage ratios for a standard normal distribution were then calculated as the cumulative probability within the range $[-z,z]$. The expected number of false outliers is calculated for datasets with 15, 127, and 255 phylogenetic independent contrasts (PIC values), corresponding to phylogenetic trees with 16, 128, and 256 species, respectively. These outliers are considered “false outliers” because they are expected under a perfect normal distribution and result from probabilistic thresholding rather than true deviations.

For our datasets with 16 species (15 PIC values), 128 species (127 PIC values), and 256 species (255 PIC values), we set outlier detection thresholds at 5×*MAD* for the smallest sample size and 6×*MAD* for the larger two. These thresholds were chosen to ensure that only data points with substantial deviations are identified as outliers, thereby minimizing the risk of misclassifying data points that do not significantly impact the performance of statistical analysis methods. Using these thresholds, the expected number of falsely identified outliers for normal data is about one per hundred simulations, ensuring high specificity in identifying deviations caused by evolutionary shifts or other phenomena beyond normal variation.

To further evaluate the robustness of our approach, we repeated the study using a more stringent threshold to define outliers: 6×*MAD* for the datasets with 16 species and 7×*MAD* for

the datasets with 128 or 256 species. Although minor variations in specific metrics were observed under these stricter thresholds, the overall ranking of the performance of the compared methods was consistent (see the Results section for details).

If either of the two traits, X_1 or X_2 in a dataset, is identified as containing an outlier, the entire dataset is flagged as outlier-containing. The design of OGC, which independently evaluates outliers for each trait, enables it to handle scenarios where only one trait experiences a shift. This flexibility also extends to more complex cases where both traits experience shifts at different locations on the phylogeny, allowing OGC to remain robust when trait changes occur independently or asymmetrically across the tree. Specifically, when no outliers are detected in either trait (X_1 or X_2), Pearson correlation is used to capture the linear relationship. Conversely, when outliers are identified in one or both traits, Spearman's rank correlation is employed to mitigate the influence of these extreme values, ensuring that the method remains robust against deviations from normality.

Importantly, even when outliers are falsely detected—due to random variation rather than true evolutionary shifts— the cost of erroneously applying Spearman correlation to outlier-free data is typically limited to a minor reduction in statistical power. In contrast, failing to account for even a single influential outlier by applying Pearson correlation can lead to substantial misestimation or even inversion of the true correlation signal. Accordingly, OGC adopts a conservative yet robust approach by defaulting to Spearman in the presence of detected outliers.

For larger datasets with 128 species (127 PIC values), Pearson correlation, despite its reliance on normality assumptions, is generally robust to minor deviations from normality due to the central limit theorem. In such cases, the impact of skewness and kurtosis is reduced by the large sample size, ensuring reliable correlation estimates. Therefore, for these datasets, the choice of Pearson or Spearman correlation was determined solely by the presence or absence of outliers.

For smaller datasets with 16 species (15 PIC values), both normality testing and outlier detection were conducted. This approach, referred to as PIC-ODGC, dynamically selects the appropriate correlation method based on data characteristics. If no outliers were detected and normality assumptions were satisfied, the Pearson correlation was used. In cases where outliers were present, or normality was violated, Spearman correlation was applied to minimize the impact of extreme values on the results.

Benchmarking Statistically Detectable Trait Correlations

In evaluating the effectiveness of statistical methods for detecting trait correlations, it is important to distinguish among three types of correlation that may appear in comparative datasets: (1) Causal correlation, where two traits are functionally or mechanistically linked (e.g., one influences the evolution of the other or both respond to a shared environmental factor); (2) Incidental correlation, where no causal link exists, but the observed correlation arises due to stochastic similarity in evolutionary trajectories. This occurs when traits evolve independently, yet happen to change in the same direction or with similar patterns purely by chance. For example, under Brownian motion (BM), trait values drift randomly over time along the branches of a phylogenetic tree. Even when traits are simulated independently (i.e., with no explicit causal link), the random walks of two traits on the same tree may align by coincidence, producing detectable correlations. Such correlations are not artifacts of flawed methods or noise misinterpretation—they are real statistical associations present in the data, albeit without biological causation. From a biological interpretation perspective, they may be treated as false leads, but from a statistical evaluation standpoint, they must be recognized as true positives, since the methods correctly identify patterns actually present in the data. 3) Spurious correlation, which emerges solely due to methodological mismatch—such as applying standard OLS to raw species-level data when residuals exhibit strong phylogenetic structure—can lead to inflated type I error. Conversely, observed lack of correlation may result from a true absence of causality, incidentally divergent trait trajectories, or methodological limitations that obscure detectable associations.

This conceptual distinction has critical implications for simulation-based benchmarking. One might assume that the simulation configuration—i.e., whether a correlation was explicitly encoded into the model—could serve as a benchmark. However, this approach fails when incidental correlations are possible, particularly under stochastic processes such as BM or OU evolution. For example, traits simulated as “independent” may still exhibit detectable correlations in some replicates due to random alignment. If a method successfully identifies these, it should not be penalized as producing false positives. Doing so would underestimate the method's power and overvalue overly conservative approaches. To address this, we distinguish between the causal structure of the simulation model and the statistical structure of the simulated data. Specifically, benchmark performance was evaluated based on whether a method identified a significant correlation that was statistically present in each replicate, not merely whether one was causally embedded in the simulation. This distinction follows Ives (2022), who cautioned against conflating biological causality with statistical detectability when evaluating comparative methods under stochastic models.

Accordingly, we constructed an empirical benchmark for each replicate by testing whether a statistically observable correlation existed, independent of causal assumptions. For each replicate, we computed Pearson or Spearman correlations between normalized trait changes ($\Delta X_1/L$ and $\Delta X_2/L$), where ΔX represents trait change across each branch and L the corresponding branch length. Because the simulation process records trait values for all nodes, including internal (ancestral) nodes, we were able to compute ΔX for every branch directly. Under BM, these normalized values are uncorrelated across branches, making conventional correlation tests valid and unbiased.

When a significant correlation was detected between $\Delta X_1/L$ and $\Delta X_2/L$ ($p < 0.05$), the replicate was labeled a benchmark hit; otherwise, a benchmark miss. This approach accommodates both causal and incidental correlations, avoiding misclassification of results as false positives or false negatives based solely on simulation setup. For instance, in cases where $X_2 = X_1 + e$ and the variance of e is extremely high, correlation may be obscured by noise—leading to a valid non-detection. Conversely, simulations lacking an explicit relationship (e.g., $X_2 = e$) may still produce observable correlations via stochastic alignment. In both cases, our benchmark reflects the actual statistical signal in the data.

Unlike our previous approach (Chen et al. 2023, Chen and Niu 2024), which relied on normality testing to select between Pearson and Spearman correlations, this study adopts an outlier-driven strategy. Outliers were identified as PIC values whose absolute deviation from the median exceeded a threshold $T_i = k \times MAD$, where k was adjusted based on the dataset size. This threshold was consistent with the one used in OGC, ensuring comparability in outlier detection across benchmarks. When no outliers were detected, Pearson correlation was applied; otherwise, Spearman rank correlation was used. This refined approach not only reduces the influence of extreme values but also ensures that true evolutionary relationships between traits with evolutionary shifts are accurately captured.

For smaller datasets with 16 species (15 PIC values), the benchmark also incorporated normality testing alongside outlier detection to better account for the data's characteristics. The same threshold applied in ODGC was used to define outliers, maintaining consistency in detection criteria across different analytical frameworks. If no outliers were identified and the normality assumption was satisfied, the Pearson correlation was used to assess the true evolutionary relationships. Conversely, in cases where outliers were present, or normality was violated, Spearman correlation was employed to mitigate the influence of extreme

values. This dual consideration of outlier presence and normality ensured that the benchmark remained robust and adaptable to the unique challenges posed by small datasets.

Secondary Benchmarking Framework for PIC-O(D)GC Evaluation

Despite the robustness of our benchmark, concerns remain about potential biases when assessing PIC-O(D)GC because the benchmark shares methodological similarities with the PIC-O(D)GC framework itself. For example, both rely on outlier detection using a shared threshold (e.g., $T_i = k \times MAD$) and employ the same decision-making process between Pearson and Spearman correlation based on the presence of outliers or normality testing. Such alignment raises the possibility that the benchmark could inadvertently favor PIC-O(D)GC by design. If PIC-O(D)GC exhibits superior performance relative to other methods is evident, it is prudent to question whether this advantage arises from genuine methodological robustness or inherent alignment with the benchmark criteria.

To address this concern, we implemented a second benchmark that is entirely independent of PIC-O(D)GC's design. In this evaluation framework, we directly used the simulated formula to assume the presence or absence of correlations between X_1 and X_2. Specifically, when the relationship was defined as $X_2 = X_1 + e$, we treated X_1 and X_2 as correlated. Conversely, when the simulation explicitly set X_1 and X_2 as uncorrelated (e.g., $X_2 = 0 \times X_1 + e$), we used the absence of a correlation as the gold standard.

Although the second benchmark introduces limitations—namely, incidental correlations caused by random noise could mislabel a method's correct detection as a false positive—its influence on all methods being compared remains consistent. This uniform impact ensures a logically unbiased assessment of relative performance, allowing us to validate PIC-O(D)GC's performance under a complementary and independent criterion.

By applying both benchmarks, we aimed to mitigate potential bias and provide a more comprehensive evaluation of the methods under consideration. In the Results section, unless otherwise specified, the comparisons of different methods are based on the first benchmark's determination of whether trait correlations exist within the data.

Programs and Packages

All simulations and analyses were performed in the R environment (version 4.1.3) (R Core Team 2020) on a Linux system. The phytools package (version 2.3.0) (Revell 2024) and ape package (version 5.8) (Paradis and Schliep 2019) were used for tree generation. The *ape*

package was also used for simulations and the calculation of PIC values. Normality tests were conducted using the Shapiro-Wilk test (Royston 1992) via the *shapiro.test()* function from the *stats* package (version 4.3.1). Pearson and Spearman correlation analyses were performed using the *cor.test()* function from the *stats* package. All PGLS regressions were conducted using the *phylolm()* function from the *phylolm* package (version 2.6.2) (Ho and Ane 2014). Corphylo was conducted using the *corphylo()* function from the *ape* package (version 5.8) (Zheng et al. 2009). PGLMM and MR-PMM were both conducted using the *MCMCglmm()* function from the *MCMCglmm* package (version 2.36) (Sorensen and Gianola 2002, Hadfield 2010, Hadfield and Nakagawa 2010). Robust phylogenetic regression were conducted using the *Conduct.Robust_PhylogeneticRegression()* function from ROBRT package (version 0.0.0.9000) (Adams et al. 2024).

RESULTS

Impact of Abrupt Evolutionary Shifts on Data Patterns and Correlation Analysis

To illustrate PIC-OGC's ability to handle abrupt evolutionary shifts, we simulated two traits on a fixed-balanced phylogenetic tree with 128 species (Fig. 1a), considering two shift configurations: (1) shifts applied to both traits simultaneously, and (2) shifts applied to X_1 only. For each configuration, we selected three shift levels—no shift, a weak shift, and a strong shift—as representative cases to illustrate how shift magnitude affects trait distributions (e.g., group separation) and contrasts (e.g., emergence of isolated outliers).

As shown in Figures 1b, 1d, and 1f, when both traits undergo an abrupt shift near the root, the raw trait values form two distinct clusters in the coordinate space, with the separation between these clusters increasing as the shift magnitude becomes larger. In the PIC coordinate system, the shared shift manifests as a single outlier in the PIC data, with the degree of outlier deviation becoming more pronounced as the shift magnitude increases (Figs. 1c, 1e, and 1g).

When phylogenetic autocorrelation is ignored, directly applying Pearson or Spearman correlation to the raw trait data frequently results in false positives, regardless of the shift magnitude (Figs. 1b, 1d, and 1f). In contrast, controlling for phylogenetic autocorrelation by using PIC values eliminates false positives when there is no shift or when the shift is small (Figs. 1c and 1e). However, under large upsizing shifts, Pearson correlation becomes highly sensitive to the outlier, leading to a false positive result and erroneously inferring a significant positive correlation between X_1 and X_2, akin to non-phylogenetic methods.

Conversely, the Spearman correlation applied to the PIC data under large shifts avoids false positives by correctly detecting no significant correlation between the two traits (Fig. 1g).

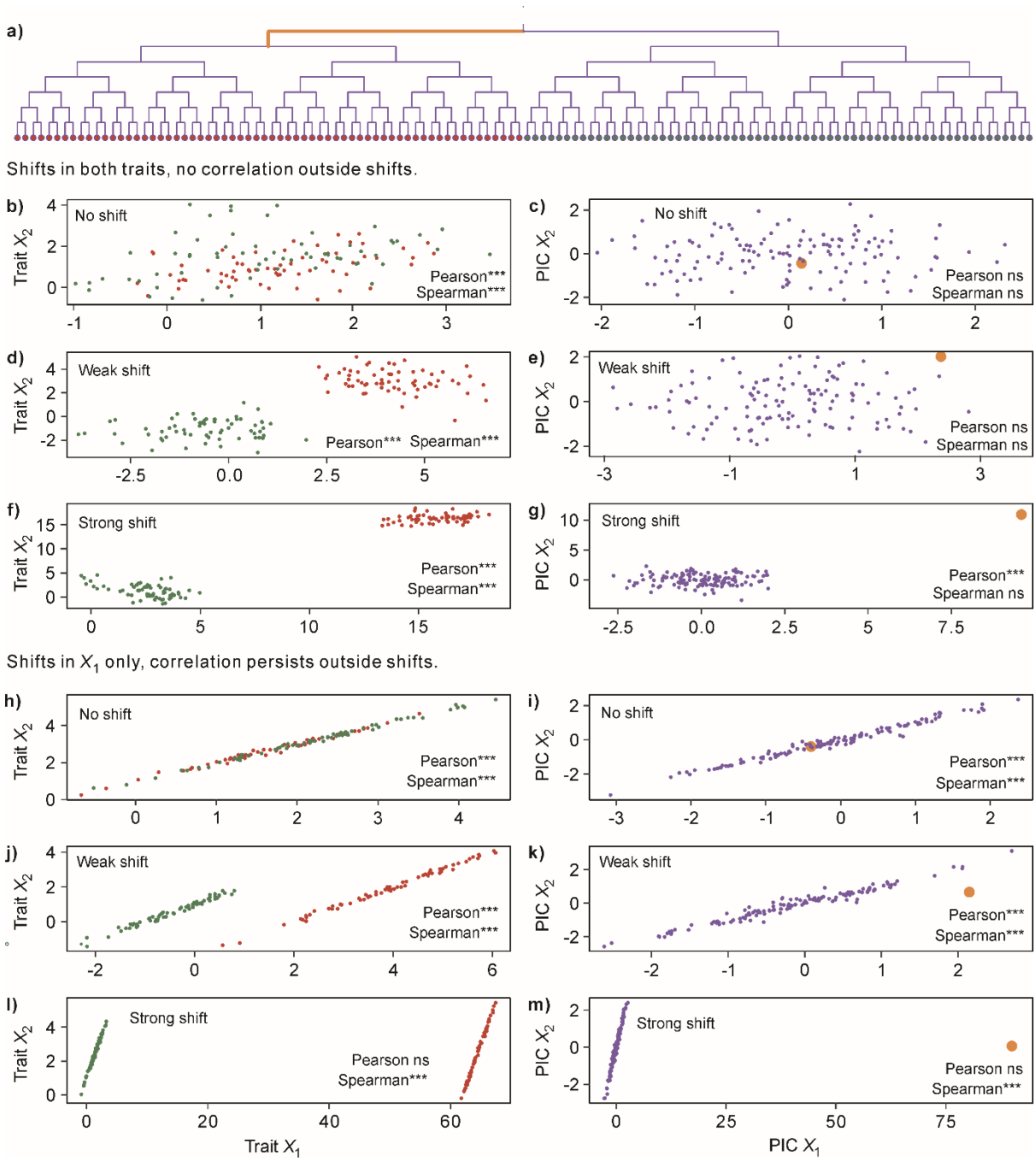

Figure 1. Effects of abrupt shifts on data pattern and correlation analysis. a) A fixed-balanced phylogenetic tree with 128 species, with the thick orange line indicating an abrupt evolutionary shift. b–c) Baseline scenario for the dual-shift simulations (d–g). Traits X_1 and X_2 evolve independently under Brownian motion (BM) with no abrupt shift (1-fold). Panels d–e and f–g represent increasing shift magnitudes (4-fold and 16-fold, respectively). h–i) Baseline scenario for the X_1-only shift simulations (j–m). Trait X_1 evolves under BM with no abrupt shift (1-fold), and X_2 is generated from X_1 via a consistent linear relationship with added noise. j–k) Moderate (4-fold) shift in X_1 only. l–m)

Strong (64-fold) shift in X_1 only. In all phylogenetically independent contrast (PIC) plots (c, e, g, i, k, m), the contrast corresponding to the branch marked by the orange line in panel a is highlighted as an enlarged orange dot. In panels e, g, k, and m, this contrast reflects an evolutionary shift; in panels c and i, it serves as a baseline reference for comparison. In the weak-shift dataset e, the orange point corresponds to a deviation of X_1 = 4×MAD and X_2 = 3×MAD, with MAD denoting the Median Absolute Deviation. In the strong-shift dataset g, the deviation increases to X_1 = 14×MAD and X_2 = 15×MAD. In the replicate dataset k, both traits again show moderate outlier values (X_1 = 4×MAD and X_2 = 5×MAD). In the extreme shift case m, X_1 exhibits a highly exaggerated deviation of 139×MAD, while X_2 remains at 4×MAD. Pearson and Spearman correlation tests: *** indicates $p < 0.001$; ns indicates not significant ($p \geq 0.05$). Data used to generate this figure are available in Supplementary Table S1.

In the second set of simulations, where only X_1 experiences an abrupt shift while X_2 follows unperturbed BM, the true correlation between the two traits remains intact outside the shifted lineage (Figs. 1h, 1j, and 1l). As in the dual-shift scenario, the raw data exhibit a visible cluster separation as the shift magnitude increases. Despite the presence of a shift in only one trait, the PIC data again display a prominent outlier under large upsizing (Fig. 1m). While Pearson correlation applied to the PIC values is affected by this outlier and fails to detect the existing correlation—resulting in a false negative (Fig. 1m)—Spearman correlation remains robust across shift levels, consistently recovering the true positive association (Figs. 1i, 1k, and 1m).

Together, these results show that large upsizing shifts can produce outliers in PIC data, which in turn may lead to false negatives or false positives in Pearson-based correlation analyses, depending on whether the shift affects one or both traits. While we simulated a deep-branch shift to illustrate maximal distortion, equally large shifts on intermediate or terminal branches can still mislead PIC-based correlation analyses. This reflects an inherent limitation of traditional PIC approaches, which weigh all contrasts equally without regard to their evolutionary context.

Refining Method Selection for Comparison with PIC-O(D)GC

To avoid visual and interpretive complexity, we limited comparisons to the best-performing method variants within each category, rather than exhaustively presenting all possible combinations. Specifically, we adopted objective criteria—such as model fit or simulation-based performance—to identify representative implementations for each statistical

framework. This streamlined design allowed for a focused and fair evaluation of PIC-O(D)GC against established alternatives.

For PGLS, the Akaike information criterion (AIC) was used to identify the optimal evolutionary model—BM, λ, OU fixed, OU random, or EB—for each dataset. Only the best-fitting PGLS model for each dataset was included in the final analysis.

For robust phylogenetic regression, which encompasses two methodological types (PGLS- and PIC-based regression) and four robust estimators (L1, S, M, and MM), we first evaluated their performance using simulations on the 128-species fixed-balanced tree.

Facing evolutionary shifts, PIC-MM stands out as the best-performing method. Under bivariate trait shift conditions (Fig. 2a), it maintained almost the lowest false positive rates across all shift magnitudes, and under univariate shift scenarios (Fig. 2c), it also exhibited the lowest false negative rates, demonstrating superior robustness in both dimensions. PIC-S also showed stability across different shift magnitudes; however, its false positive rates were consistently higher than those of PIC-MM in all bivariate shift conditions (Fig. 2a), indicating comparatively weaker control of type I error. In terms of false negatives (Fig. 2c), PIC-S performed similarly to PIC-MM, with no substantial differences observed. Other methods, whether based on PIC or PGLS, performed acceptably under weak shifts (shift ≤ 4), maintaining relatively low false positive and false negative rates (Figs. 2a and 2c). However, under stronger shifts (shift > 4), both error rates increased sharply—often approaching or reaching 1—highlighting their instability in the presence of abrupt evolutionary changes. Notably, PIC-L1 and PGLS-L1 exhibited lower false positive rates than PIC-MM under weak shifts (Fig. 2a and Supplementary Table S2), suggesting that the L1 estimator may have potential advantages in specific contexts.

Accuracy reflects the combined influence of false positive and false negative rates by measuring the proportion of correct predictions (true positives and true negatives), offering a holistic view of the methods' overall performance. Figures 2b and 2d confirm that PIC-MM is the most reliable method, achieving both high accuracy and stability across all evolutionary shift gradients. PIC-S follows as the second-best performer. The trends in accuracy largely align with the false positive rates in Figures 2a and 2c, underscoring the strong correlation between these two metrics for robust methods.

Figures 2e-2g show the accuracy of various methods across three no-shift evolutionary scenarios: BM1 & BM1+BM2, BM & BM+Norm, and Norm & Norm+BM. PIC-MM is generally among the best-performing methods, though it is occasionally slightly outperformed in specific cases, such as under high variance in the BM & BM+Norm

scenario, where PIC-L1 and PGLS-L1 show a clear, though modest, advantage. PIC-S, on the other hand, is no longer the second-best performer; in fact, it ranks among the worst in BM1 & BM1+BM2 and Norm & Norm+BM, showing considerable performance degradation under these conditions.

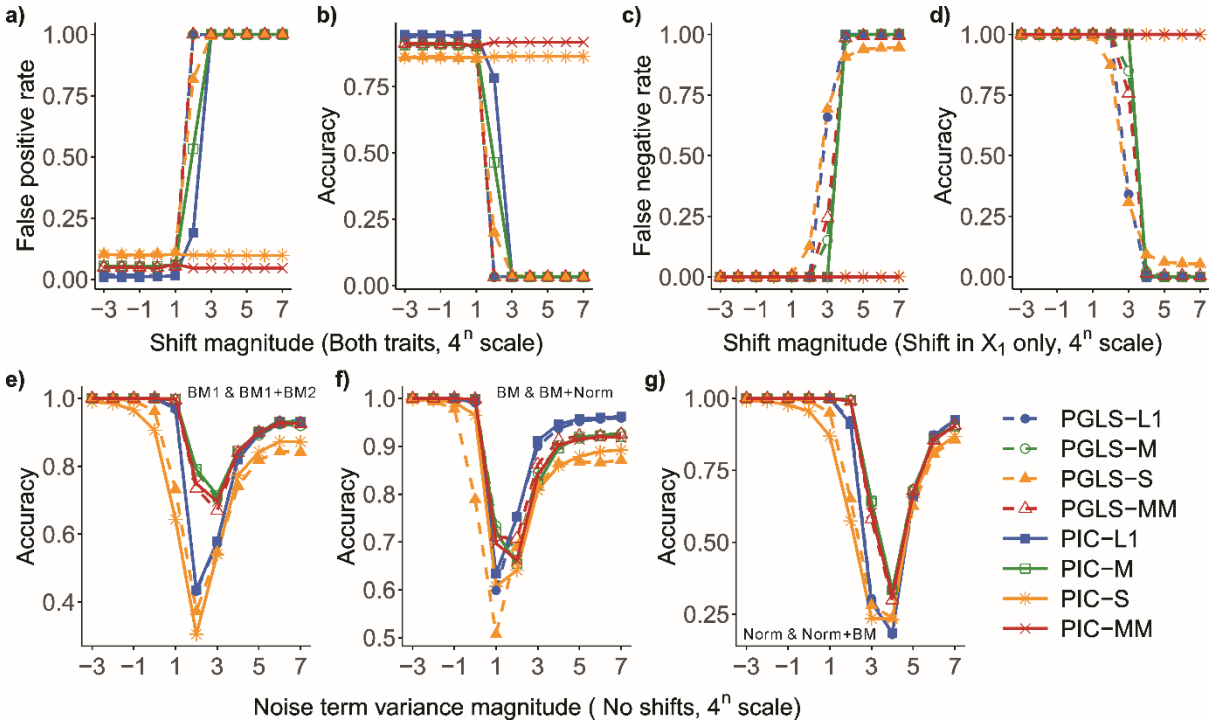

Figure 2. Evaluation of eight robust phylogenetic regression methods under simulations on a 128-species fixed-balanced tree. Panels a–b show false positive rates and accuracies under simultaneous shifts in both X_1 and X_2; panels c–d show false negative rates and accuracies when only X_1 is shifted. Panels e–g show accuracies under three background scenarios without abrupt shifts: (e) BM1 & BM1+BM2, (f) BM & BM+Norm, (g) Norm & Norm+BM. Shift/variance values are represented as 4^n, where n is the horizontal axis label. For example, –3 represents 4^{-3}, and 7 represents 4^7. Accuracy represents the proportion of simulations in which a method correctly detected both the direction and significance of the trait correlation, based on the benchmark. Additional details on the simulation design, variance gradient configurations, and performance metrics (including accuracy, false positive rate, and false negative rate) are described in the Materials and Methods. The data used to plot this figure are available in Supplementary Table S2.

When analyzing randomly generated 128-species trees, fully fixed-balanced 16-species trees, and fully fixed-balanced 256-species trees, PIC-MM consistently demonstrates the strongest ability to handle evolutionary shifts. Across the three simulation scenarios without evolutionary shifts, PIC-MM performs at least as well as, or not significantly worse than, the other methods (Supplementary Figs. S1-S3 and Supplementary Table S2). Even when the

outlier detection threshold in the benchmark was adjusted from 6×*MAD* to 7×*MAD* for 128-species datasets, PIC-MM maintained its top performance across both random and fixed-balanced trees (Supplementary Figs. S4-S7 and Supplementary Table S3). These results highlight PIC-MM's robustness and versatility, making it a reliable tool for analyzing phylogenetic data under both stable and dynamic evolutionary conditions. These findings align with the conclusions of Adams et al. (2024). Consequently, PIC-MM was selected as the representative robust regression method for comparison with PIC-O(D)GC.

Through this optimization and selection process, we reduced the total number of methods compared with PIC-O(D)GC to six, facilitating a more focused and interpretable evaluation.

PIC-OGC: Robust Performance Across Evolutionary Shifts

We compared the performance of PIC-OGC with six other methods using simulations on the 128-species fixed-balanced tree. As shown in Figure 3a, there are significant differences in false positive rates among the methods in bivariate shift conditions. Under weak or downsizing evolutionary shifts (shift ≤ 4), PIC-OGC consistently exhibits the same false positive rates as PIC-Pearson. These two methods consistently outperform the other five. The differences are subtle in Figure 3a and primarily discernible at the numerical level, as shown in Supplementary Table S4. When expressed as absolute percentage-point differences from PIC-OGC and PIC-Pearson, most alternative methods—including PIC-MM, PGLS, and PGLMM—show modest increases in false-positive rates, typically ranging between 1% and 2%. In contrast, MR-PMM exhibits a substantially higher rate, exceeding PIC-OGC and PIC-Pearson by more than 9%. Corphylo remains the closest in performance, with differences usually below 0.5%.

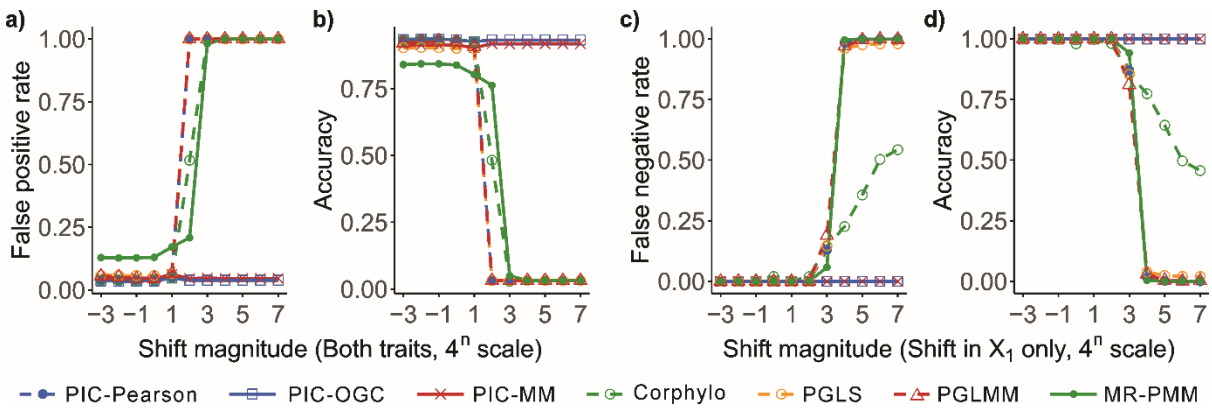

Figure 3. Performance of seven methods on simulations across evolutionary shift gradients simulated on a fixed-balanced tree of 128 species. Panels a–b show false positive rates and accuracies under simultaneous shifts in both X_1 and X_2; panels c–d show false negative rates and accuracies when only

X_1 is shifted. Shift values are represented as 4^n, where the horizontal axis shows only the exponent (n) instead of the full value. See the Materials and Methods section for additional details on the simulation design, gradient configurations, and performance metrics (including accuracy). The data used to plot this figure are available in Supplementary Table S4.

However, when the shift becomes stronger (> 4), the false positive rate of PIC-Pearson sharply increases to 1, while PIC-OGC maintains a stable false positive rate of 0.039. This result demonstrates that PIC-OGC is significantly more robust against evolutionary shifts than traditional PIC regression methods. PIC-MM likewise remains robust to evolutionary shifts, and its false-positive rate is virtually indistinguishable from that of PIC-OGC (Fig. 3a)— the small numeric differences are evident only upon inspection of the summary values used to generate the plot (Supplementary Table S4), not in the plotted curves. Other methods show a similar pattern to PIC-Pearson, with false positive rates rising sharply toward 1 as the shift magnitude increases, indicating limited robustness to stronger evolutionary shifts.

Figure 3b displays the accuracy of the same seven methods under the same 11 evolutionary shift gradients. Accuracy here serves as a comprehensive metric, reflecting both false positive and false negative rates. Notably, PIC-OGC demonstrates the highest accuracy or shares the top position across all shift gradients. PIC-MM also performs exceptionally well, with accuracy very similar to PIC-OGC (Supplementary Table S4). The other methods show acceptable performance under weak shifts but exhibit unacceptably low accuracy under stronger shifts. Despite PGLS results being derived from selecting the best-fitting model from five candidate models for each simulation, its performance still falls significantly short of PIC-OGC and PIC-MM.

Figures 3c–d illustrate results from simulations with evolutionary shifts in only a single trait (X_1). The performance patterns under this condition closely mirror those observed in the dual-trait shift scenarios (Figures 3a–b). PIC-OGC and PIC-MM again exhibit highly consistent false negative rates and accuracies, maintaining robustness across all shift gradients. Corphylo performs noticeably better under strong shifts in the single-trait scenario compared to its performance in dual-trait shifts, though it still underperforms relative to PIC-OGC and PIC-MM.

The fixed trees represent a fully fixed-balanced phylogenetic structure, where all simulations are conducted using the same fixed-balanced dichotomous tree. This consistent topology ensures uniformity in the branching structure, resulting in homogenized data. In contrast, random trees are generated based on a coalescent model. This variability results in

diverse tree topologies across simulations, reflecting a wider range of phylogenetic scenarios and introducing additional heterogeneity in the data. In our analysis with 128-species random trees (Supplementary Fig. S8 and Supplementary Table S4), most methods showed performance very similar to their results under fixed-balanced simulations, with PIC-OGC and PIC-MM consistently demonstrating the best performance overall. However, the slight yet consistent edge that PIC-OGC held over PIC-MM across all shift gradients under fixed-balanced tree simulations was no longer observed with random trees. Instead, each method exhibited slight advantages under specific gradients, emphasizing their complementary strengths in handling the complexities introduced by varying phylogenetic structures. Additionally, MR-PMM exhibited a significant improvement under low shift gradients (shift ≤ 4). While it had the worst performance with fixed-balanced trees, it became the method with the lowest false positive rate and highest accuracy in weak conditions, outperforming all other methods except PIC-OGC and PIC-MM.

All the above results were obtained using a $6\times MAD$ threshold to define outliers. Using a $7\times MAD$ threshold yielded highly similar results, with PIC-OGC and PIC-MM remaining the top-performing methods (Supplementary Figs. S9-S10 and Supplementary Table S5).

PIC-OGC: Strong Performance in Simulations Without Abrupt Shifts

The analysis here focuses on simulations conducted on a 128-species fixed-balanced tree, where two traits exhibit a relationship of $X_2 = X_1 + e$, with e representing random noise varying across 11 gradients of variance displayed on the x-axis of Figure 4. These simulations aim to evaluate the methods' performance in detecting relationships between evolutionary traits under conditions without abrupt shifts.

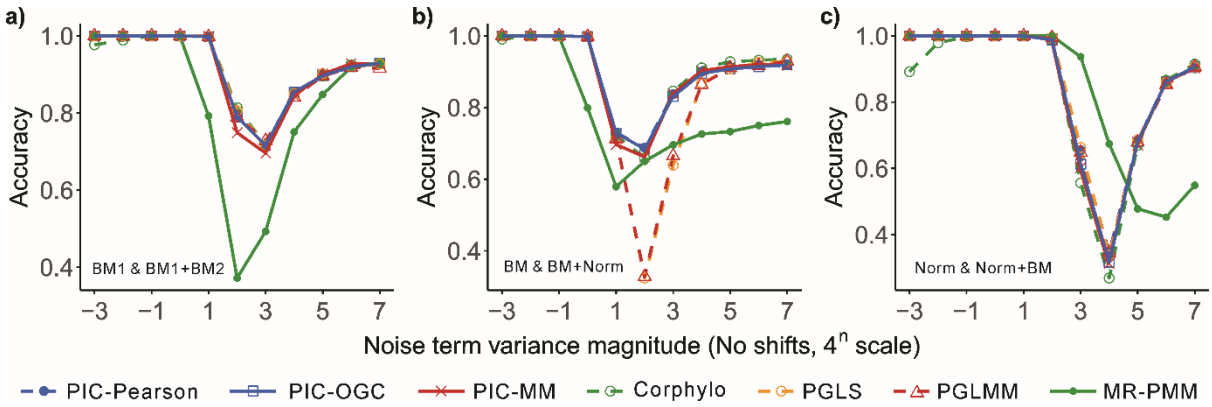

Figure 4. Accuracy of seven methods on simulations without abrupt shifts. Simulations were conducted on a fixed-balanced tree of 128 species, with relationships between traits defined as $X_2 = X_1 + e$, where e represents random noise, across three scenarios: a) BM1 & BM1+BM2, b) BM &

BM+Norm, and c) Norm & Norm+BM. Variance values are represented as 4^n, where the horizontal axis shows only the exponent (*n*) instead of the full value. For example, –3 corresponds to 4^{-3}, and 7 corresponds to 4^7. See the Materials and Methods section for additional details on the simulation design, gradient configurations, and performance metrics (including accuracy). The data used to plot this figure are available in Supplementary Table S4.

Across all three simulation scenarios (Figs. 4a-c), MR-PMM exhibits a distinct and inconsistent performance pattern, clearly separating it from the other methods. Unlike the other approaches, its accuracy fluctuates across noise levels and simulation settings, lacking a consistent advantage or disadvantage.

In both the BM1 & BM1+BM2 (Fig. 4a) and Norm & Norm+BM (Fig. 4c) scenarios, the remaining six methods—PIC-O(D)GC, PIC-MM, PIC-Pearson, Corphylo, PGLS, and PGLMM—perform similarly, with only slight variations across the noise gradients. These methods exhibit a non-linear accuracy pattern, with a significant reduction in performance at moderate-to-high noise levels compared to both low and high extremes.

The BM & BM+Norm scenario (Fig. 4b) offers a particularly informative contrast. At moderate-to-high noise levels, the six methods separate into two distinct performance groups: a higher-performing group comprising PIC-O(D)GC, PIC-MM, PIC-Pearson, and Corphylo, and a lower-performing group including PGLS and PGLMM.

Interestingly, as noise levels become extreme (e.g., variance = 4^7), the correlation between traits is effectively eliminated by random variation, and benchmark labels shift to "non-correlation." In this regime, except MR-PMM, all methods appropriately return non-significant results, which paradoxically leads to higher accuracy.

Results from simulations on the 128-species random trees show overall trends consistent with those observed on the fixed-balanced tree, reaffirming the strong performance of the three PIC-based methods across diverse phylogenetic topologies (Supplementary Fig. S11 and Supplementary Table S4). While their overall performance remains excellent, PIC-OGC shows a slight drop at two specific variance gradients (4^4 and 4^5) of the last simulation scenario (Norm & Norm+BM), where it no longer ranks among the top three methods. These results indicate a minor, context-dependent limitation of PIC-OGC in random tree simulations.

In the BM & BM+Norm scenario, where the data-generating process conforms to the assumptions of standard linear regression, correlation analysis on the original trait values (Raw–Pearson, equivalent to OLS) was also included as a theoretical benchmark

(Supplementary Fig. S12a and Supplementary Table S6). The accuracy trajectory of Raw–Pearson closely mirrored those of PGLS and PGLMM, forming a consistently low-performing group under moderate-to-high levels of added noise. This consistency in performance patterns among the three methods was also observed across alternative tree structures, including the 128-species random tree and the 16- and 256-species fixed-balanced trees (Supplementary Fig. S12b–d). These results reinforces theoretical expectations: when residuals are phylogenetically independent, PGLS effectively reduces to OLS (Revell 2010, Symonds and Blomberg 2014).

The above results on methods' accuracies detecting relationships between evolutionary traits under conditions without abrupt shifts were obtained using a 6×*MAD* threshold to define outliers. Using a 7×*MAD* threshold yielded highly similar results, with the three PIC-based methods remaining the top-performing methods (Supplementary Figs. S13-S14 and Supplementary Table S5).

Performance of the Methods Under Non-BM Evolutionary Conditions

While the application of OLS regression or Pearson correlation on PIC values requires the assumption that traits evolve under a BM model—so that PICs are statistically i.i.d.—the computation of PIC values themselves does not strictly rely on BM. Instead, PIC values are defined as algebraic contrasts on a phylogeny, and remain meaningful under a wide range of evolutionary models (Garland et al. 1992). It is the statistical interpretation (e.g., *p*-values) of PIC-OLS or PIC-Pearson that breaks down when the BM assumption is violated. In contrast, methods such as PIC-MM and PIC-OGC—though still operating on PIC values—employ robust estimators or nonparametric correlations that do not require the strict assumptions of BM. Thus, they are retained in our comparison alongside PGLS, PGLMM, MR-PMM, and Corphylo when evaluating method performance under simulated non-BM conditions. Importantly, this also means that their performance can be fairly compared to methods that are explicitly designed to accommodate a range of evolutionary models.

To classify simulated datasets as exhibiting BM or non-BM condition, we calculated the AIC values from PGLS regression fitted under five evolutionary models: BM, lambda, OU random, OU fixed, and EB. The datasets were classified as BM condition if the BM model yielded the lowest AIC; otherwise, they were considered non-BM condition.

Generally, the two PIC-based methods, PIC-OGC and PIC-MM, maintained their superiority under non-BM conditions involving bivariate shifts (Figs. 5a–5b). PIC-OGC performed exceptionally well, achieving almost the lowest false positive rates and highest

accuracy across all shift gradients. PIC-MM also demonstrated strong overall performance, though it exhibited a high false positive rate and low accuracy at a specific shift gradient (4). Despite being first optimized for performance by selecting the best-fitting evolutionary model, PGLS was among the weakest-performing methods under non-BM conditions, showing high false positive rates and low accuracy, particularly at higher shift gradients.

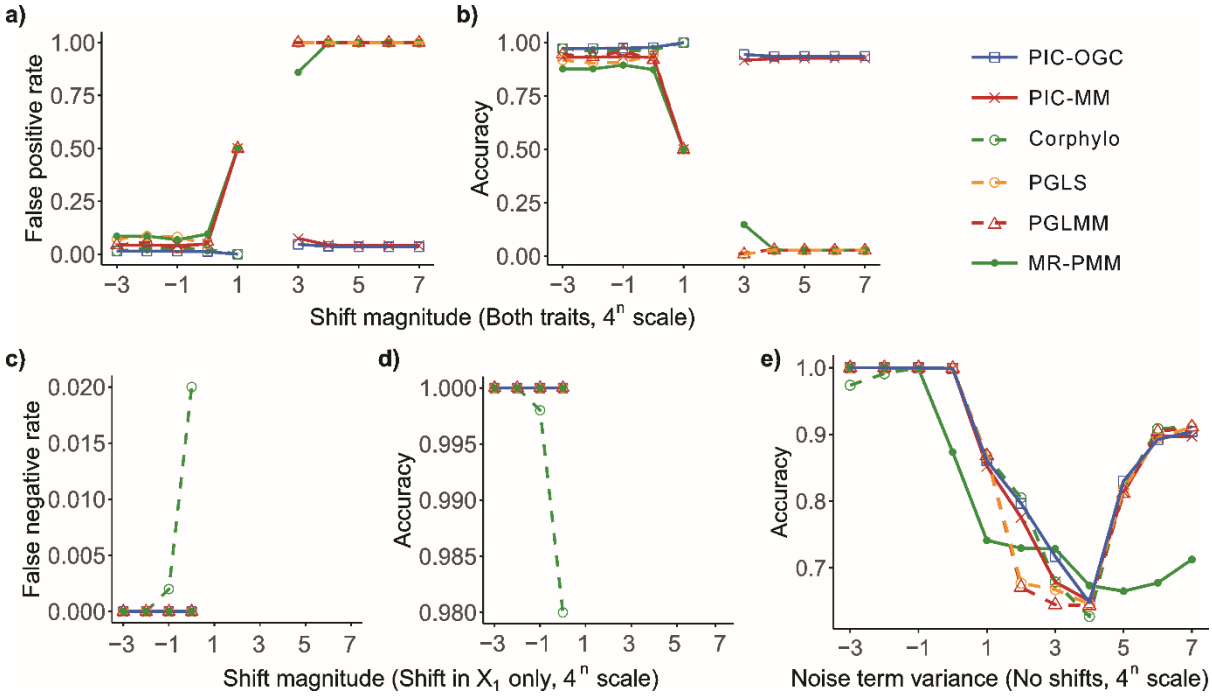

Figure 5. Performance of six methods across datasets that deviate from BM assumptions. a) False positive rates of six methods under non-BM datasets with evolutionary shifts in both X_1 and X_2. b) Accuracy of the methods under the same conditions. c) False negative rates of six methods under non-BM datasets with evolutionary shifts only in X_1. d) Accuracy of the methods under the same conditions. e) Accuracy of the methods under non-BM datasets without evolutionary shifts. Shift/variance values are represented as 4^n, where the horizontal axis shows only the exponent (n) instead of the full value. For example, –3 corresponds to 4^{-3}, and 7 corresponds to 4^7. All the datasets were simulated using a fixed-balanced tree of 128 species. They were classified as BM or non-BM conditions based on AIC values from PGLS regression models fitted under five evolutionary models (BM, lambda, OU random, OU fixed, and EB). Data were classified as BM condition if the BM model had the lowest AIC; otherwise, they were considered non-BM condition. Blank regions in the figure indicate positions where all simulated datasets were classified as BM conditions, and thus no corresponding non-BM datasets were available for evaluation. See the Materials and Methods section for additional details on the simulation design, gradient configurations, and performance metrics (including accuracy). The data used to plot this figure are available in Supplementary Table S7.

Under single-trait shift conditions, non-BM datasets only occur at lower shift magnitudes, and the differences among methods are minimal (Fig. 5c–d). PIC-OGC, PIC-MM, PGLS, PGLMM, and MR-PMM all maintain nearly identical performance. The only notable deviation is observed in Corphylo, which exhibits elevated false negative rates and reduced accuracy at the two highest available shift magnitudes.

In scenarios without evolutionary shifts (Fig. 5e), none of the six retained methods showed a consistent advantage or disadvantage. Performance varied across different variance levels, with no single method consistently emerging as the clear best or worst.

Similar trends were observed when using random trees (Supplementary Fig. S15 and Supplementary Table S7) or adjusting the threshold for defining outliers (Supplementary Fig. S16-S17 and Supplementary Table S8), reinforcing the robustness of these conclusions across varying tree structures and outlier detection criteria.

Comparison of the accuracies under conditions with and without shifts (e.g., Fig. 3b, d vs Fig. 4a-c; Fig. 5b vs Fig. 5e) reveals a clear pattern: PIC-OGC and PIC-MM maintain considerably stronger performance under medium-to-high shifts than under non-shifted conditions with variance of comparable magnitude, highlighting their robustness to directional evolutionary change. By contrast, the other methods perform substantially better under medium and high noise variance when no shifts are present than when exposed to directional shifts of similar magnitude.

PIC-O(D)GC: Robust Performance with Small and Large Samples

To evaluate robustness across sample sizes, we examined performance using a small fixed-balanced tree (16 species) and a large one (256 species), in addition to the default 128-species setting. For small samples consisting of 16 species (15 PIC values), the choice between Pearson and Spearman correlation is influenced not only by the presence or absence of outliers but also by whether the data follow a normal distribution. To address this, we enhanced PIC-OGC to PIC-ODGC, incorporating both outlier detection and normality testing. For outlier detection, we applied size-adjusted thresholds: 5×MAD and 6×MAD for the 16-species dataset, and the stricter 6×MAD and 7×MAD thresholds for the 256-species dataset, accounting for increased variability in larger samples.

Results from the 16-species datasets closely mirrored those of the 128-species analysis in scenarios with evolutionary shifts (Supplementary Figs. S12, S18-S23 and Supplementary Tables S4–S8). PIC-ODGC and PIC-MM consistently ranked among the top-performing methods.

Analyses on the 256-species balanced tree also yielded performance patterns similar to those observed in the 128-species dataset, further reinforcing the scalability and robustness of PIC-ODGC and PIC-MM across diverse sample sizes. For complete accuracy scores and comparative visualizations, see Supplementary Figures S12, S24–S29 and Supplementary Tables S4–S8.

Secondary Benchmark Confirms Robustness of PIC-Based Methods Addressing Concerns about PIC-Based Performance

PIC-O(D)GC and PIC-MM, being rooted in PIC calculations, might initially raise concerns about whether their observed strong performance stems from a bias due to the gold standard's reliance on calculating $\Delta X_1/L$ and $\Delta X_2/L$, a process that closely mirrors the PIC estimation procedure. This concern is particularly pronounced for PIC-O(D)GC, as it also employs the same outlier detection thresholds and criteria for selecting Pearson or Spearman correlation.

To address potential biases in the original benchmark, we validated PIC-O(D)GC and PIC-MM using a second, entirely independent benchmark. Unlike the first benchmark, this alternative framework does not rely on any assumptions or calculations related to PIC or any specific method. Instead, it directly determines the presence or absence of correlations based on the simulated formula: X_1 and X_2 are treated as correlated when defined as $X_2 = X_1 + e$, and uncorrelated when defined as $X_2 = 0 \times X_1 + e$. Although this benchmark may classify incidental correlations caused by random noise as false positives, its design is method-agnostic, ensuring no bias toward or against any specific approach.

Using the second, entirely independent benchmark, we found that the relative performance rankings of all methods remained largely consistent across simulation settings. This result indicates that the strong performance previously observed for PIC-O(D)GC and PIC-MM is not an artifact of benchmark design, but reflects their inherent robustness.

Among the eight phylogenetically robust methods—including both PGLS- and PIC-based regression, as well as four robust estimators (L1, S, M, and MM)—PIC-MM consistently achieved the best performance across a range of shift scenarios (Supplementary Figs. S30–S33 and Supplementary Table S9). It maintained the lowest false negative rates under single-trait shifts and the lowest false positive rates under dual-trait shifts, particularly under moderate and large shift magnitudes. However, in no-shift scenarios with variance heterogeneity—especially under larger trees (e.g., 128-taxon balanced or random trees and 256-taxon balanced trees)—performance differences among methods were generally smaller,

and PIC-MM was among the less accurate methods under these specific conditions (Supplementary Figs. S30–S33 and Supplementary Table S9). Nevertheless, because the central goal of the analysis is to detect directional evolutionary shifts, these results reinforce the value of robust regression approaches—particularly PIC-MM—as a preferred choice for shift-sensitive inference.

When compared to non-robust methods such as PIC-Pearson and standard PGLS, PIC-MM and PIC-O(D)GC again demonstrated superior accuracy under moderate and large shift magnitudes (Supplementary Figs. S34–S37 and Supplementary Table S10), consistent with findings under the first benchmark. Under low-shift or no-shift scenarios, differences among methods diminished, and no method exhibited universally superior performance. Overall, PIC-MM and PIC-O(D)GC continued to show strong and stable behavior, reaffirming their robustness regardless of the benchmark framework.

Application of PIC-O(D)GC to Selected Empirical Datasets with Trait Shifts

Simulation results, validated against internal benchmarks from the generative model or recorded ancestral states, showed PIC-O(D)GC and PIC-MM to be most robust under abrupt shifts. To evaluate PIC-O(D)GC's performance on real-world data, we analyzed three empirical trait datasets, derived either in full or in part from published phylogenetic resources or comparative studies (Jones et al. 2009, Scales et al. 2009, Makino and Kawata 2019).

Invertebrate propagule size vs. genome size.— We analyzed the relationship between propagule size and genome size with their phylogenetic context across 44 invertebrates. Instead of presenting trait values directly, the analysis is framed through PICs, which isolate evolutionary changes along branches while correcting for shared ancestry. Three univariate outliers were identified—two in genome size and one in propagule size (Fig. 6a). Their phylogenetic locations, shown in Figure 6b, span a diverse set of clades and illustrate that extreme trait shifts can occur independently across lineages. Unfortunately, examining the phylogenetic positions of these outlier contrasts provides little clarity as to which lineage experienced an upsizing or downsizing shift. Nevertheless, these isolated contrasts illustrate how lineage-specific shifts can disproportionately affect correlation estimates (Fig. 6c). PIC-OGC, which identifies and downweights extreme contrasts, and PIC-MM, which reweights contrasts based on residual magnitude, both yielded non-significant results in this case—unlike most other methods. This divergence emphasizes the importance of outlier treatment in comparative analyses.

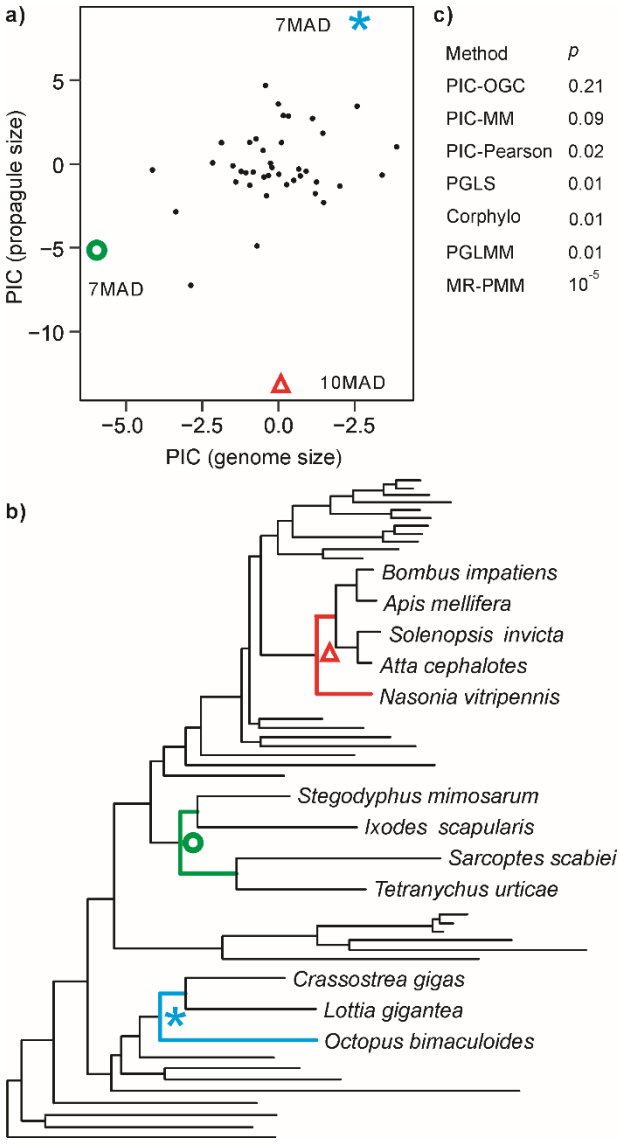

Figure 6. Relationship between propagule size and genome size across 44 invertebrates. a) Phylogenetically independent contrasts (PICs) between $\log_{10}$-transformed propagule size and genome size. A blue star and a red triangle represent contrasts that are outliers in propagule size only, while a green circle indicates an outlier in genome size only. None of these contrasts are joint outliers. Although visual inspection may suggest extremeness in both traits, univariate outlier analysis confirmed that each deviates by less than 4 Median Absolute Deviations (MAD) from the median in the other trait. b) Phylogenetic tree of the 44 invertebrates. Highlighted branches indicate the taxa involved in each outlier contrast shown in panel (a), with colored symbols placed between them to represent the corresponding contrast. The two branches involved in each outlier contrast are colored and slightly thickened to match the symbol color, aiding visual identification. Each outlier contrast likely reflects an abrupt evolutionary shift in either propagule size or genome size. c) Statistical significance of the association between genome size and propagule size as assessed by seven comparative methods. Detailed correlation coefficients and *p*-values (reported to additional decimal

places), along with the original trait values and the PIC values from which the scatterplots were derived, are provided in Supplementary Table S11.

Lizard fast glycolytic vs. slow oxidative fiber proportion.—We further examined the relationship between SO and FG muscle fiber proportions in a dataset of 22 lizard species. A single univariate outlier was identified in FG fiber proportion (Fig. 7a), corresponding to a contrast between *Phrynosoma* and its sister clade (Fig. 7b). Inspection of the raw data revealed that this contrast is driven almost entirely by a marked reduction in FG fiber proportion across the four sampled *Phrynosoma* species, indicating a directional shift within this lineage. Across the full dataset, phylogenetic comparative methods produced mixed results (Fig. 7c). While PIC-ODGC, PIC-MM, and PGLMM detected a statistically significant association between SO and FG fiber proportions, other methods such as PIC-Pearson, PGLS, and MR-PMM failed to do so.

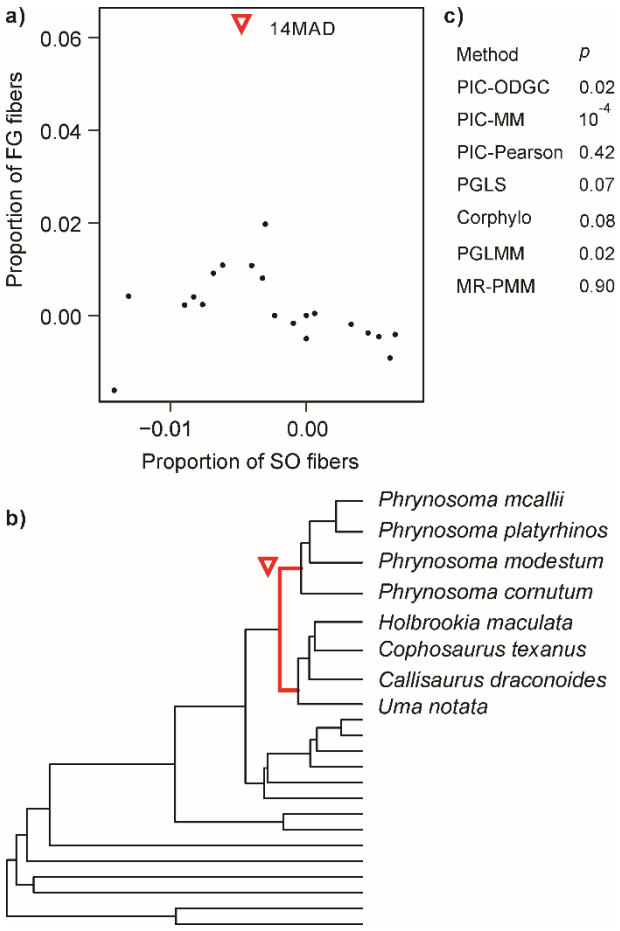

Figure 7. Relationship between slow oxidative (SO) and fast glycolytic (FG) muscle fiber proportions in 22 lizard species. a) Phylogenetically independent contrasts (PICs) between the proportions of SO and FG muscle fibers. A single univariate outlier (red triangle) is detected in FG fiber proportion, with a deviation of 14 Median Absolute Deviations (MAD) from the median; its corresponding

deviation in SO fiber proportion falls within the expected range. b) Phylogenetic tree showing the location of the outlier contrast. The two taxa involved are connected by a red internal branch, with the red triangle placed on the *Phrynosoma* lineage, which exhibits a pronounced reduction in FG fiber proportion relative to its sister group. c) The *p*-values from multiple phylogenetic comparative methods testing for association between SO and FG fiber proportions. Detailed correlation coefficients and *p*-values (reported to additional decimal places), along with the original trait values and PIC values from which the scatterplots were derived, are provided in Supplementary Table S12.

To test the influence of the directional shift in *Phrynosoma*, we re-ran all analyses after removing the four species in this genus. Following this exclusion, all methods except MR-PMM returned *p*-values below 0.05, consistently indicating a negative association between SO and FG fiber proportions (Supplementary Table S12). This result suggests that the localized evolutionary shift within *Phrynosoma* both introduces statistical heterogeneity and suppresses correlation signals under certain comparative frameworks.

Body mass and population density in even-toed ungulates.—Figure 8 integrates statistical associations between body mass and population density in 109 species of even-toed ungulates (*Artiodactyla*) using PICs. The distribution of contrasts revealed two clear univariate outliers (Fig. 8a). One involved a 15-MAD deviation in body mass between *Syncerus caffer* (African buffalo) and *Ammodorcas clarkei* (dibatag), while the other showed a 6-MAD deviation in population density between *Kobus leche* and *Kobus megaceros*. In both cases, deviation in the non-outlier trait was less than 5 MADs, and no contrast exhibited extreme values in both dimensions simultaneously.

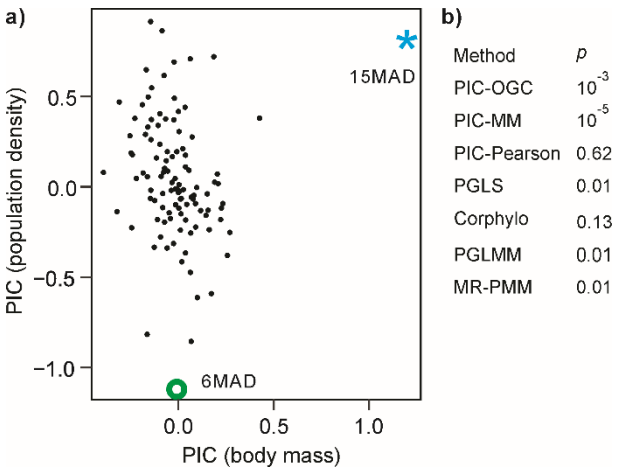

Figure 8. Relationship between body mass and population density in even-toed ungulates (Artiodactyla). a) Phylogenetically independent contrasts (PICs) between $\log_{10}$-transformed body mass and population density for 109 artiodactyl species. Two contrasts are identified as univariate

outliers: a blue star marks a 15 Median Absolute Deviations (MAD) in body mass (*Syncerus caffer* vs *Ammodorcas clarkei*), and a green circle marks a 6 MAD in population density (*Kobus leche* vs *Kobus megaceros*). In both cases, deviation in the other trait was less than 5 MADs. b) *p*-values from multiple comparative methods testing for an association between body mass and population density. Detailed correlation coefficients and *p*-values (reported to additional decimal places), along with the original trait values and PIC values from which the scatterplots were derived, are provided in Supplementary Table S13.

All methods returned negative associations between the two traits, but their statistical significance varied (Fig. 8b). Notably, PIC-Pearson and Corphylo yielded non-significant results ($p = 0.621$ and 0.127, respectively), in contrast to other methods such as PIC-OGC ($p = 0.001$) and PGLS ($p = 0.010$).

Because the direction of evolutionary change could not be confidently inferred from the phylogenetic positions of the outlier contrasts, we adopted a conservative approach by removing both taxa involved in each contrast. Given the large sample size (109 species), excluding these four terminal taxa had minimal impact on overall statistical power. To examine the influence of the two extreme contrasts, we re-ran all methods after removing the four terminal taxa involved in these contrasts from the dataset. This exclusion led to consistent results across all methods, with *p*-values below 0.01 and negative effect estimates in every case (Supplementary Table S13). These findings indicate that a small number of extreme observations can substantially suppress signal strength in otherwise robust trait associations.

Interestingly, body mass and population density in even-toed ungulates showed a pattern that differed from the other two datasets. In both the invertebrate and lizard datasets, PIC-Pearson and PGLS yielded consistent results in terms of statistical significance, even though the exact *p*-values differed. In contrast, for the artiodactyl data, PIC-Pearson failed to detect a significant association ($p = 0.621$), while PGLS identified a significant negative relationship ($p = 0.010$). This discrepancy stands in contrast to the common expectation that PIC-Pearson and PGLS tend to produce similar results under standard comparative conditions.

Further inspection revealed that the phylogenetic location of outliers may help explain the observed discrepancy between PIC-Pearson and PGLS in the artiodactyl dataset. The two methods yielded consistent results in the first two datasets, where extreme contrasts occurred between internal nodes positioned slightly deeper in the tree—each involving clades of more than two species. In the artiodactyl dataset, however, outliers were confined to contrasts

between terminal species pairs. This difference in phylogenetic depth may underlie the divergent outcomes of these two methods. PIC-Pearson assigns equal weight to all contrasts regardless of phylogenetic position, allowing outliers on terminal branches to disproportionately influence the overall correlation. In contrast, PGLS distributes influence across all taxa via a covariance matrix, rendering it less sensitive to abrupt shifts involving only a few species. These findings highlight how the statistical and phylogenetic positioning of outliers can influence comparative inferences, even between methods that are otherwise closely aligned.

Runtime Comparison of Phylogenetic Correlation Methods

To assess computational efficiency, we compared the runtime of PIC-O(D)GC with the other six phylogenetic comparative methods across datasets of increasing size (Fig. 9). Among all methods, PIC-O(D)GC and PIC-MM showed the smallest increase in runtime as sample size grew, highlighting their strong scalability. The overall time cost followed a clear pattern: PIC-Pearson < PIC-O(D)GC < PIC-MM < PGLS < PGLMM < MR-PMM < Corphylo.

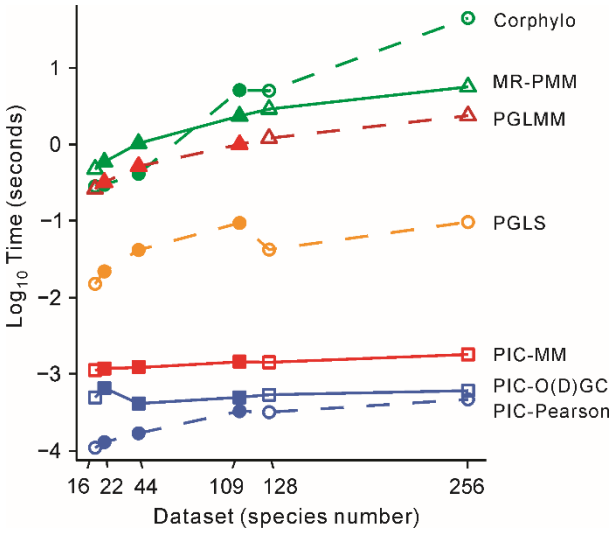

Figure 9. Computational time comparison across phylogenetic correlation methods. Average runtime (in seconds, log10-transformed) was measured for each method across six datasets, including three empirical datasets used in the above analyses (solid symbols) and three simulated datasets based on fixed-balanced phylogenetic trees (hollow symbols). All simulated datasets were generated under the two-trait shift scenario with an effect size of 4^{-3}. Each method was repeated 100 times per dataset to obtain mean values. All runtime measurements were performed on a Dell server running CentOS Linux 7.9, equipped with an AMD EPYC 7702 64-core processor and 1 TB of RAM. Computational time did not always increase with sample size. For instance, both PGLS and Corphylo ran slower on the 109-species dataset than on the 128-species one. This was likely due to relatively long tip labels (e.g., Latin species names) in the 109-species phylogeny, which increased the time required for tree–

trait matching. For PIC-O(D)GC, slightly longer runtime was observed in small datasets ($n < 30$), where both outlier detection and normality testing steps were applied. Due to noticeable fluctuations in system load on the shared computational environment, absolute runtimes varied considerably across replicates. However, the relative performance ranking of methods was stable and therefore provides a more reliable basis for comparison. The raw timing data used to generate this figure are provided in Supplementary Table S14.

PIC-Pearson and PIC-O(D)GC were the fastest methods, both benefiting from linear or near-linear computational complexity. PIC-O(D)GC involves median-based outlier detection (O(n)) and, depending on outlier presence, applies either Pearson or Spearman correlation. The use of Spearman introduces an O(n log n) cost due to ranking, which remains efficient even for large trees. In small datasets, a normality test (e.g., Shapiro-Wilk) is also applied to guide correlation choice, which adds some additional computational time, particularly in small sample sizes.

PIC-MM required noticeably more time than PIC-O(D)GC, reflecting its robust two-stage estimation procedure: an initial S-estimation step (up to $O(n^3)$), followed by iterative M-estimation steps (each $O(n^2)$). PGLS was even slower than PIC-MM, driven by repeated matrix inversions. The remaining model-based approaches—PGLMM, MR-PMM, and Corphylo—incurred the highest computational time overall. For example, on the largest dataset ($n = 256$), PIC-Pearson, PIC-O(D)GC, and PIC-MM completed in 5×10^{-4}, 6×10^{-4}, and 2×10^{-3} seconds, respectively, while PGLS, PGLMM, and MR-PMM required 0.10, 2.3, and 5.6 seconds, respectively. Corphylo was particularly slow, taking 44 seconds on the same dataset.

These results underscore the scalability of PIC-O(D)GC, which achieves high computational efficiency without compromising robustness. This dual strength makes it particularly suitable for large-scale phylogenetic comparative analyses.

DISCUSSION

Equivalence of Pearson Correlation and OLS in Context

Both Pearson correlation and OLS regression can be used to assess trait associations using PIC (Felsenstein 1985, Felsenstein 1988, Murray et al. 2020, Horvath et al. 2022, Bai et al. 2023, Wang et al. 2023). When applied to PIC data under linear assumptions, these two approaches yield mathematically equivalent results in terms of statistical significance and correlation direction. In this study, we chose to use Pearson correlation for its computational

simplicity and consistency with other nonparametric methods, such as Spearman correlation, which is a key component of the PIC-O(D)GC framework. The equivalence between Pearson correlation and OLS regression ensures logical and statistical consistency with previous findings while reinforcing the relevance and robustness of our results within the context of phylogenetic comparative analyses.

Comparing Statistical Methods for Detecting Trait Correlations

Our study evaluated the performance of various statistical methods in detecting trait correlations under diverse evolutionary scenarios, including with and without abrupt shifts. PIC-O(D)GC and PIC-MM consistently demonstrated robust performance, particularly excelling under shift scenarios by minimizing false positives as well as false negatives, thereby maintaining high overall accuracy—consistent with prior evaluations of robust regression strategies in the presence of model violations and trait shifts (Adams et al. 2024). In contrast, traditional methods like PIC-Pearson and PGLS, as well as robust variants such as PGLS-MM, struggled under high shift gradients, highlighting their limitations in challenging conditions.

In no-shift scenarios, both PIC-MM and PIC-O(D)GC performed comparably to other reliable methods, such as PIC-Pearson, PGLS, PGLMM, and Corphylo, underscoring that neither method was uniformly superior across all conditions. These complementary strengths highlight the importance of aligning method choice with expected data characteristics and analytical goals.

Notably, the abrupt trait shifts introduced in our simulations can be considered an extreme form of evolutionary heterogeneity—a concept emphasized by Mazel et al. (2016) as a key source of model violation in phylogenetic regression. While Mazel et al. focused on rate heterogeneity, our results show that discrete directional shifts can also distort inference and compromise methods assuming homogeneous evolutionary processes. This underscores the importance of methods capable of accommodating both continuous and discontinuous heterogeneity.

Conceptual Simplicity and Practical Accessibility

In addition to the substantial runtime advantage demonstrated in Figure 6—where PIC-O(D)GC was considerably faster than other commonly used methods—PIC-O(D)GC also offers strengths in conceptual simplicity and user accessibility. By dynamically adapting to the presence or absence of outliers and data normality, it strikes an effective balance between

robustness and interpretability. Its reliance on straightforward statistical principles, such as Spearman's rank correlation, ensures that the method is logically intuitive and easy to understand, even for researchers with limited statistical training (Redmond and Keenan 2002). By avoiding rigid assumptions about data distribution, PIC-O(D)GC provides transparent and interpretable results, reducing the risk of misinterpretation and enhancing its accessibility to a broader range of biologists.

Limitations of Traditional Data Transformations in Phylogenetic Correlation Analysis

A potential concern readers may raise is the complexity of the PIC-O(D)GC framework. Given its focus on handling non-normal data distributions and outliers, why not simply apply data transformations (e.g., logarithmic or square root transformations) to standardize the data before using the traditional PIC-OLS regression? Such transformations are well-known for their ability to address skewness and improve normality, potentially achieving comparable results with much less complexity.

While this rationale is common in general statistical analyses, it faces critical conceptual and practical limitations when applied to phylogenetic contrasts. First, PICs are calculated as differences between trait values at bifurcating nodes and are thus centered around zero, frequently including both positive and negative values. Logarithmic and square root transformations, however, are only defined for positive values. Shifting the data by adding an arbitrary constant can enable transformation, but doing so introduces artificial scaling and biases that are especially problematic when contrasts span a wide numerical range. More importantly, PIC values represent an interval scale, where only additive operations are mathematically and conceptually justified. Logarithmic transformation, being the inverse of exponentiation, requires ratio scale data and thus violates the scale properties of contrasts (Houle et al. 2011).

Second, even when applied to the original trait data rather than contrasts, transformations may distort biologically meaningful signals. For instance, when a simultaneous evolutionary shift in two traits occurs near the root of a phylogenetic tree, the original trait distributions may exhibit a bimodal pattern. Log-transforming the traits before calculating PICs might compress the two modes and bring the distribution closer to normality—making the PICs appear more regular and the regression model seemingly more stable. However, this comes at a cost: the shift signal itself is attenuated or even obscured. From a biological perspective, this shift may reflect an important event in evolutionary history, such as a rapid radiation or

punctuated equilibrium. Discarding it in favor of statistical smoothness risks losing insight into the very processes under study.

Even setting aside concerns about distortion of trait distributions caused by data transformations, logarithmic transformation alone often fails to mitigate the influence of outliers. In our empirical datasets—such as propagule size versus genome size in invertebrates (Makino and Kawata 2019) and body mass versus population density in artiodactyls (Jones et al. 2009)—we observed that even after $\log_{10}$ transformation, the resulting PICs still exhibited pronounced outliers. These outliers remained capable of skewing correlation estimates in methods like PIC-Pearson and PGLS, highlighting the inadequacy of standard transformations in mitigating the statistical impact of lineage-specific shifts.

Advantages of Correlation-Based Methods over Regression-Based Approaches

Correlation-based methods such as PIC-O(D)GC, Corphylo, and MR-PMM offer notable advantages over regression-based approaches like PIC-OLS, PGLS, and PGLMM, particularly in the presence of measurement error or uncertain causal structure. Regression-based methods, including those adapted to phylogenetic contexts, are rooted in the logic of classical linear regression, where slope estimates are often interpreted causally—with changes in an independent variable assumed to drive changes in a dependent one. This framework generally presumes precise measurement of predictors and minimal confounding (McArdle 1988, Osborne and Waters 2002, Casson and Farmer 2014, Jarantow et al. 2023).

When the independent variable is subject to substantial measurement error, however, the resulting slope estimates may be biased and misleading—a phenomenon known as measurement error bias. In extreme cases, the bias may obscure real relationships or generate spurious ones, as demonstrated in empirical studies such as those examining the effects of lead exposure on human IQ (Marais and Wecker 1998, Lamadrid-Figueroa et al. 2011). While approaches like standardized major axis regression attempt to address these issues by accounting for error in both variables (Warton et al. 2006), they still rely on directional modeling frameworks and retain assumptions about the structure of the relationship. Under such conditions, correlation-based methods—which assess statistical associations without assuming causal directionality, primacy, or error-free predictors—can offer more robust and interpretable insights. Because they are agnostic to causal structure and are symmetrical with respect to variable roles, correlation approaches are particularly suitable for exploratory

analyses or contexts where both variables may contain measurement error (Ives et al. 2007, Grabowski et al. 2016).

Even without causal assumptions, slope estimates can offer meaningful descriptive summaries of trait covariation. In evolutionary studies where statistical significance testing may not be central, regression models can be used to characterize patterns rather than test hypotheses. For example, Grabowski et al. (2016) conducted phylogenetic regressions without reporting *p*-values, interpreting slope magnitudes as indicators of morphological integration and evolutionary constraint across hominoid clades. In such contexts, slopes reflect the strength and direction of covariation, even when variable roles are ambiguous or measurement error is present.

Limitations of Correlation-Based and Nonparametric Methods in Phylogenetic Analysis

While correlation-based and nonparametric methods like PIC-O(D)GC offer robustness to outliers and do not require distributional assumptions, they share several fundamental limitations. First, they can only detect monotonic associations, failing to capture more complex or non-monotonic evolutionary relationships, such as U-shaped or bell-shaped patterns. Second, these methods cannot provide parameter estimates, such as regression coefficients or curvature terms, which are crucial for mechanistic interpretations of trait evolution. Furthermore, they typically cannot accommodate multiple predictors or interaction effects, restricting their utility in multifactorial evolutionary scenarios. These limitations have prompted concerns that such methods, while statistically conservative, offer little insight beyond rejecting unrealistic null models (Caetano et al. 2018). In contrast, parametric approaches like PGLS allow for the estimation of meaningful parameters and hypothesis testing regarding the strength, direction, and functional form of evolutionary relationships (Garamszegi 2014).

Despite these limitations, correlation-based methods remain valuable tools for addressing specific challenges in phylogenetic data analysis. They should not be viewed as replacements for parametric models but rather as complementary approaches that provide robust and computationally efficient solutions, particularly in scenarios with non-normal data distributions, outliers, or small sample sizes.

From Statistical Artifacts to Evolutionary Signals: Interpreting PIC Outliers

Beyond its utility in assessing trait correlations, the outlier detection step in PIC-O(D)GC holds promise for identifying evolutionary shifts or jumps. Recent studies have highlighted

the importance of models that accommodate abrupt changes in trait evolution, such as the singular events model (Uyeda et al. 2018). These rare but significant transitions may leave identifiable signatures in PIC data (Uyeda et al. 2018), including extreme contrast values, or localized changes along specific branches.

The importance of examining PIC scatterplots cannot be overstated. Even when an overall correlation appears statistically significant, the result may be disproportionately driven by a few extreme contrast values. Visual inspection of the PIC plot is therefore essential, not only as a diagnostic check, but also as a guide to more structured interpretation. When such outliers are present, we recommend analyzing the relationship between two traits from three complementary perspectives: 1) Overall correlation can be assessed using robust methods such as PIC-MM or PIC-O(D)GC, which account for outliers and non-normality while maintaining statistical stability across heterogeneous evolutionary scenarios; 2) Shift-to-shift relationships should be interpreted separately, considering whether shifts are aligned or opposed in direction, which may reveal patterns of evolutionary coupling or divergence. 3) Background correlation refers to the relationship between traits during non-shift, gradual phases of evolution. When a shift is located deep in the phylogeny—particularly near the root, as seen in Figure 1 (panels d, f, j, l)—an effective strategy is to partition the data into two clades at the shift node. This allows separate estimation of trait relationships within each group, providing insight into background evolutionary dynamics before and after the divergence introduced by the shift. For shifts occurring on terminal or near-terminal branches (e.g., Figures 8 and 9), removing the taxa involved in the shift can help recover the underlying correlation structure. These strategies respect phylogenetic hierarchy and help isolate gradual evolutionary signals from singular, large-effect shifts. Once the influence of shifts is appropriately filtered—whether through clade partitioning or tip removal—regression-based methods such as PIC-OLS, PGLS, or PGLMM (Felsenstein 1985, Martins and Hansen 1997, Pagel 1997, Pagel 1999, Ives et al. 2007, Revell 2010, Ives and Helmus 2011) become more appropriate and informative. These approaches are preferred in background-phase analysis due to their interpretability and modeling efficiency. In contrast, robust contrast-based methods like PIC-O(D)GC or PIC-MM (Adams et al. 2024) may be overly conservative in this refined context.

These findings highlight that PIC-based outlier detection, beyond its statistical role, can reveal deeper evolutionary processes. By linking extreme contrast values to evolutionary shifts, researchers can move beyond global associations to explore how trait relationships vary across distinct evolutionary contexts. This stratified perspective—separating shift-

driven, background, and shift-to-shift dynamics—enhances the precision of macroevolutionary interpretation.

Empirical evidence increasingly supports the idea that phenotypic evolution involves rare but consequential jumps embedded within longer periods of gradual change or stasis. Such discontinuities have been observed across a range of systems, including body size evolution in crocodylomorphs (Godoy et al. 2019), bursts of morphological innovation in lizards (Brennan et al. 2024), shifts in venom gene evolution linked to ecological opportunity in snakes (Barua and Mikheyev 2020), and abrupt changes in genomic traits such as GC content and genome size in bacteria (Gao and Wu 2022, Hu et al. 2022, Mahajan and Agashe 2022). These cases underscore the importance of not only identifying where and when such evolutionary shifts occur, but also understanding how they alter patterns of trait covariation. In particular, they highlight the need to distinguish among overall correlations across the tree, background correlations within stable lineages, and shift-to-shift relationships that may reflect directional evolutionary coupling or divergence between regimes.

The combined use of contrast-based diagnostics such as PIC-O(D)GC and targeted analyses of trait relationships offers a tractable framework for detecting and interpreting evolutionary pulses. Unlike fully model-based methods, PIC analysis leverages branch-level resolution to identify candidate shifts without prior assumptions about evolutionary regimes. This approach complements theoretical and algorithmic models such as OU-with-jumps (Landis et al. 2013), Lévy process-based jump models (Duchen et al. 2017, Landis and Schraiber 2017), LASSO-regularized shift detection (Khabbazian et al. 2016, Bastide et al. 2018, Zhang et al. 2024), and singular events models (Uyeda et al. 2011, Uyeda et al. 2018), all of which recognize discrete shifts as critical components of macroevolutionary dynamics.

Further refinement of this framework may be achieved by incorporating phylogenetic features such as branch lengths and trait-specific evolutionary rates. These additions could improve the specificity of shift detection and help distinguish genuine evolutionary transitions from background variation. Integrating PIC outlier diagnostics into hypothesis-driven macroevolutionary studies may thus provide a more nuanced view of how trait relationships evolve across lineages.

FUNDING

This study was supported in part by the National Natural Science Foundation of China (grant number 31671321), with leftover resources used for occasional research expenses.

ACKNOWLEDGMENTS

We sincerely thank Professor Josef Uyeda (Associate Editor) and the anonymous reviewer for their constructive and insightful comments. Many sentences in this article were drafted with the assistance of ChatGPT-4 and its later versions, large language models developed by OpenAI. The authors thoroughly reviewed and revised each sentence and took full responsibility for the language and content of the entire article.

DATA AVAILABILITY

Supplementary Tables S1–S14 and Supplementary Figures S1–S37 are available in the files Supplementary_Tables.xlsx and Supplementary_Figures.pdf, respectively. The raw data for the simulated phylogenetic analyses, including trait correlation data for 16-, 128-, and 256-species fixed-balanced trees, as well as 128-species random trees across shift and non-shift scenarios, are available in RawData.zip. The scripts used to generate the results presented in the main text— covering phylogenetic tree construction and trait data simulation (01Simulation_BalancedTree.R, 01Simulation_RandomTree.R, 02Simulation_BM1&BM1+BM2.R, 02Simulation_BM&BM+Norm.R, 02Simulation_Norm&Norm+BM.R, 02Simulation_ShiftsinBothTraits.R and 02Simulation_ShiftsinOneTrait.R), outlier detection (03Analysis_05_MAD.R), regression model fitting (03Analysis_01_PIC.R, 03Analysis_02_PGLS_RPR.R, 03Analysis_03_Corphylo.R, 03Analysis_04_PGLMM_MRPMM.R), are provided in Code.zip. All three files, Supplemental_Tables.xlsx, RawData.zip, and Code.zip are available from the Dryad Digital Repository: http://datadryad.org/stash/share/fLN-7CFO0F_Mw7aL0hSfl9drAh-r0RmN8BNE1q25xFE. We also have implemented a suite of functions for conducting PIC-O(D)GC in the R package PICODGC (Phylogenetic Independent Contrasts with Outlier- and Distribution-Guided Correlation). PICODGC is freely available for open use by the community (https://github.com/zhenglinchen/PICODGC).

REFERENCES

Adams R., Cain Z., Assis R., DeGiorgio M. 2024. Robust phylogenetic regression. Syst. Biol., 73:140-157.

Bai X.-L., Yang D., Sher J., Zhang Y.-B., Zhang K.-Y., Liu Q., Wen H.-D., Zhang J.-L., Slot M. 2023. Divergences in stem and leaf traits between lianas and coexisting trees in a subtropical montane forest. Journal of Plant Ecology, 17:rtad037.

Barua A., Mikheyev A.S. 2020. Toxin expression in snake venom evolves rapidly with constant shifts in evolutionary rates. Proc. R. Soc. B, 287:20200613.

Bastide P., Ané C., Robin S., Mariadassou M. 2018. Inference of adaptive shifts for multivariate correlated traits. Syst. Biol., 67:662-680.

Blomberg S.P., Lefevre J.G., Wells J.A., Waterhouse M. 2012. Independent contrasts and PGLS regression estimators are equivalent. Syst. Biol., 61:382-391.

Brennan I.G., Chapple D.G., Keogh J.S., Donnellan S. 2024. Evolutionary bursts drive morphological novelty in the world's largest skinks. Curr. Biol., 34:3905-3916.e3905.

Caetano D.S., O'Meara B.C., Beaulieu J.M. 2018. Hidden state models improve state-dependent diversification approaches, including biogeographical models. Evolution, 72:2308-2324.

Casson R.J., Farmer L.D. 2014. Understanding and checking the assumptions of linear regression: a primer for medical researchers. Clin. Exp. Ophthalmol., 42:590-596.

Chen Z.-L., Guo H.-J., Niu D.-K. 2023. Dependent variable selection in phylogenetic generalized least squares regression analysis under Pagel's lambda model. bioRxiv:2023.2005.2021.541623.

Chen Z.-L., Niu D.-K. 2024. Optimizing variable selection in phylogenetic eigenvector regression for trait correlation analysis. bioRxiv:2024.2004.2014.589420.

Cornwallis C.K., Griffin A.S. 2024. A guided tour of phylogenetic comparative methods for studying trait evolution. Annu. Rev. Ecol. Evol. Syst., 55:181-204.

Dewar A.E., Belcher L.J., West S.A. 2025. A phylogenetic approach to comparative genomics. Nat. Rev. Genet., 26:395-405.

Duchen P., Leuenberger C., Szilagyi S.M., Harmon L., Eastman J., Schweizer M., Wegmann D. 2017. Inference of evolutionary jumps in large phylogenies using Levy processes. Syst. Biol., 66:950-963.

Felsenstein J. 1985. Phylogenies and the comparative method. Amer. Natur., 125:1-15.

Felsenstein J. 1988. Phylogenies and quantitative characters. Annu. Rev. Ecol. Syst., 19:445-471.

Freckleton R.P., Harvey P.H., Pagel M. 2002. Phylogenetic analysis and comparative data: A test and review of evidence. Amer. Natur., 160:712-726.

Gao Y., Wu M. 2022. Microbial genomic trait evolution is dominated by frequent and rare pulsed evolution. Sci. Adv., 8:eabn1916.

Garamszegi L.Z. 2014. Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice. Berlin, Springer.

Garland T., Jr, Harvey P.H., Ives A.R. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol., 41:18-32.

Godoy P.L., Benson R.B.J., Bronzati M., Butler R.J. 2019. The multi-peak adaptive landscape of crocodylomorph body size evolution. BMC Evol. Biol., 19:167.

Grabowski M., Voje K.L., Hansen T.F. 2016. Evolutionary modeling and correcting for observation error support a 3/5 brain-body allometry for primates. Journal of Human Evolution, 94:106-116.

Grafen A. 1989. The phylogenetic regression. Philos. Trans. R. Soc. Lond. B Biol. Sci., 326:119-157.

Hadfield J.D. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. J. Stat. Softw., 33:1-22.

Hadfield J.D., Nakagawa S. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J. Evol. Biol., 23:494-508.

Halliwell B., Holland B.R., Yates L.A. 2025. Multi-response phylogenetic mixed models: concepts and application. Biol. Rev., 100:1294-1316.

Hansen T.F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution, 51:1341-1351.

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., Purvis A., Ricklefs R.E., Schluter D., Schulte Ii J.A., Seehausen O., Sidlauskas B.L., Torres-Carvajal O., Weir J.T., Mooers A.Ø. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution, 64:2385-2396.

Ho L.S.T., Ane C. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Syst. Biol., 63:397-408.

Horvath S., Lu A.T., Haghani A., Zoller J.A., Li C.Z., Lim A.R., Brooke R.T., Raj K., Serres-Armero A., Dreger D.L., Hogan A.N., Plassais J., Ostrander E.A. 2022. DNA methylation clocks for dogs and humans. Proc. Natl. Acad. Sci. USA, 119:e2120887119.

Houle D., xe, labon C., Wagner G., xfc, nter, xa, P, Hansen T., xa, F. 2011. Measurement and meaning in biology. Q. Rev. Biol., 86:3-34.

Hu E.-Z., Lan X.-R., Liu Z.-L., Gao J., Niu D.-K. 2022. A positive correlation between GC content and growth temperature in prokaryotes. BMC Genomics, 23:110.

Iglewicz B., Hoaglin D. 1993. Volume 16: How to Detect and Handle Outliers. ASQ Quality Press.

Ives A.R. 2022. Random errors are neither: On the interpretation of correlated data. Methods Ecol. Evol., 13:2092-2105.

Ives A.R., Helmus M.R. 2011. Generalized linear mixed models for phylogenetic analyses of community structure. Ecol. Monographs, 81:511-525.

Ives A.R., Midford P.E., Garland T., Jr. 2007. Within-species variation and measurement error in phylogenetic comparative methods. Syst. Biol., 56:252-270.

Jarantow S.W., Pisors E.D., Chiu M.L. 2023. Introduction to the use of linear and nonlinear regression analysis in quantitative biological assays. Curr. Protoc., 3:e801.

Jones K.E., Bielby J., Cardillo M., Fritz S.A., O'Dell J., Orme C.D.L., Safi K., Sechrest W., Boakes E.H., Carbone C., Connolly C., Cutts M.J., Foster J.K., Grenyer R., Habib M., Plaster C.A., Price S.A., Rigby E.A., Rist J., Teacher A., Bininda-Emonds O.R.P., Gittleman J.L.,

Mace G.M., Purvis A. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology, 90:2648-2648.

Khabbazian M., Kriebel R., Rohe K., Ané C. 2016. Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models. Methods Ecol. Evol., 7:811-824.

Lamadrid-Figueroa H., Téllez-Rojo M.M., Ángeles G., Hernández-Ávila M., Hu H. 2011. Bias correction by use of errors-in-variables regression models in studies with K-X-ray fluorescence bone lead measurements. Environmental Research, 111:17-20.

Landis M.J., Schraiber J.G. 2017. Pulsed evolution shaped modern vertebrate body sizes. Proc. Natl. Acad. Sci. USA, 114:13224-13229.

Landis M.J., Schraiber J.G., Liang M. 2013. Phylogenetic analysis using Levy processes: finding jumps in the evolution of continuous traits. Syst. Biol., 62:193-204.

Leys C., Ley C., Klein O., Bernard P., Licata L. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J. Exp. Soc. Psychol., 49:764-766.

Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology. Evolution, 45:1065-1080.

Maddison W.P., FitzJohn R.G. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. Syst. Biol., 64:127-136.

Mahajan S., Agashe D. 2022. Evolutionary jumps in bacterial GC content. G3 Genes|Genomes|Genetics, 12:jkac108.

Makino T., Kawata M. 2019. Invasive invertebrates associated with highly duplicated gene content. Mol. Ecol., 28:1652-1663.

Marais M.L., Wecker W.E. 1998. Correcting for omitted-variables and measurement-error bias in regression with an application to the effect of lead on IQ. J Am Stat Assoc, 93:494-505.

Martins E.P., Hansen T.F. 1997. Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. Amer. Natur., 149:646-667.

Mazel F., Davies T.J., Georges D., Lavergne S., Thuiller W., Peres-Neto P.R. 2016. Improving phylogenetic regression under complex evolutionary models. Ecology, 97:286-293.

McArdle B.H. 1988. The structural relationship: regression in biology. Can. J. Zool., 66:2329-2339.

Murray G.G.R., Charlesworth J., Miller E.L., Casey M.J., Lloyd C.T., Gottschalk M., Tucker A.W., Welch J.J., Weinert L.A. 2020. Genome reduction is associated with bacterial pathogenicity across different scales of temporal and ecological divergence. Mol. Biol. Evol., 38:1570-1579.

O'Meara B. 2016. Phylogenetic Comparative Method. In: Kliman RM editor. Encyclopedia of Evolutionary Biology. Oxford, Academic Press, p. 254-256.

Osborne J.W., Waters E. 2002. Four assumptions of multiple regression that researchers should always test. Pract. Assess. Res. Eval., 8:2.

Pagel M. 1997. Inferring evolutionary processes from phylogenies. Zool. Scr., 26:331-348.

Pagel M. 1999. Inferring the historical patterns of biological evolution. Nature, 401:877-884.

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics, 35:526-528.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Redmond A.C., Keenan A.M. 2002. Understanding statistics. Putting p-values into perspective. J Am Podiatr Med Assoc, 92:297-305.

Revell L.J. 2010. Phylogenetic signal and linear regression on species data. Methods Ecol. Evol., 1:319-329.

Revell L.J. 2024. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). PeerJ, 12:e16505.

Rindskopf D., Shiyko M. 2010. Measures of Dispersion, Skewness and Kurtosis. In: Peterson P, Baker E, McGaw B editors. International Encyclopedia of Education (Third Edition). Oxford, Elsevier, p. 267-273.

Rohlf F.J. 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution, 55:2143-2160.

Romiguier J., Gayral P., Ballenghien M., Bernard A., Cahais V., Chenuil A., Chiari Y., Dernat R., Duret L., Faivre N., Loire E., Lourenco J.M., Nabholz B., Roux C., Tsagkogeorga G., Weber A.A.T., Weinert L.A., Belkhir K., Bierne N., Glémin S., Galtier N. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature, 515:261-263.

Royston P. 1992. Approximating the Shapiro-Wilk W-test for non-normality. Statistics and Computing, 2:117-119.

Scales J.A., King A.A., Butler M.A. 2009. Running for your life or running for your dinner: what drives fiber-type evolution in lizard locomotor muscles? Amer. Natur., 173:543-553.

Slater G.J., Pennell M.W. 2014. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. Syst. Biol., 63:293-308.

Smith E.G., Surm J.M., Macrander J., Simhi A., Amir G., Sachkova M.Y., Lewandowska M., Reitzel A.M., Moran Y. 2023. Micro and macroevolution of sea anemone venom phenotype. Nat. Commun., 14:249.

Sorensen D., Gianola D. 2002. Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics. New York, Springer.

Sumner S., Favreau E., Geist K., Toth A.L., Rehan S.M. 2023. Molecular patterns and processes in evolving sociality: lessons from insects. Philos Trans R Soc Lond B Biol Sci, 378:20220076.

Symonds M.R.E., Blomberg S.P. 2014. A primer on phylogenetic generalised least squares. In: Garamszegi LZ editor. Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice. Berlin, Heidelberg, Springer Berlin Heidelberg, p. 105-130.

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

Uyeda J.C., Hansen T.F., Arnold S.J., Pienaar J. 2011. The million-year wait for macroevolutionary bursts. Proc. Natl. Acad. Sci. USA, 108:15908-15913.

Uyeda J.C., Zenil-Ferguson R., Pennell M.W. 2018. Rethinking phylogenetic comparative methods. Syst. Biol., 67:1091-1109.

Wang B., Tong Z.-Y., Xiong Y.-Z., Wang X.-F., Scott Armbruster W., Huang S.-Q. 2023. The evolution of flower–pollinator trait matching, and why do some alpine gingers appear to be mismatched? Ann. Bot., 132:1073-1088.

Warton D.I., Wright I.J., Falster D.S., Westoby M. 2006. Bivariate line-fitting methods for allometry. Biol. Rev., 81:259-291.

Westoby M., Yates L., Holland B., Halliwell B. 2023. Phylogenetically conservative trait correlation: Quantification and interpretation. J Ecol, 111:2105-2117.

Wilcox R.R. 2003. 3 - Summarizing data. In: Wilcox RR editor. Applying Contemporary Statistical Techniques. Burlington, Academic Press, p. 55-91.

Zhang W., Kenney T., Ho L.S.T. 2024. Evolutionary shift detection with ensemble variable selection. BMC Ecology and Evolution, 24:11.

Zheng L., Ives A.R., Garland T., Larget B.R., Yu Y., Cao K. 2009. New multivariate tests for phylogenetic signal and trait correlations applied to ecophysiological phenotypes of nine *Manglietia* species. Funct. Ecol., 23:1059-1069.