Handling Character Dependency in Phylogenetic Inference: Extensive Performance Testing of Assumptions and Solutions Using Simulated Data

Tiago R. Simões[1,*], Oksana V. Vernygora[2], Bruno A.S. de Medeiros[3], and April M. Wright[4]

[1]*Department of Organismic and Evolutionary Biology & Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts, USA;*

[2]*Department of Entomology, University of Kentucky, Lexington, Kentucky, USA;*

[3]*Smithsonian Tropical Research Institute, Panama City, Panama;*

[4]*Department of Biological Sciences, Southeastern Louisiana University, Hammond, Louisiana, USA.*

[*]*Correspondence to be sent to: Department of Organismic and Evolutionary Biology & Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA.; Telephone: +1 617 955-1081; E-mail: tsimoes@fas.harvard.edu*

Abstract.— Character dependency is a major conceptual and methodological problem in phylogenetic inference of morphological datasets, as it violates the assumption of characters independency that is common to all phylogenetic methods. It is more frequently observed in higher-level phylogenies or in datasets characterizing major evolutionary transitions, as these represent parts of the tree of life where (primary) anatomical characters either originate or disappear entirely. As a result, secondary traits related to these primary characters become "inapplicable" across all sampled taxa in which that character is absent. Various solutions have been explored over the last three decades to handle character dependency, such as alternative character coding schemes and, more recently, new algorithmic implementations. However, the

accuracy of the proposed solutions, or the impact of character dependency across distinct optimality criteria, has never been directly tested using standard performance measures. Here, we utilize simple and complex simulated morphological datasets analyzed under different maximum parsimony optimization procedures and Bayesian inference to test the accuracy of various coding and algorithmic solutions to character dependency. We find that in small simulated datasets, absent coding performs better than other popular coding strategies available (contingent and multistate), whereas in more complex simulations (larger datasets controlled for different tree structure and character distribution models) contingent coding is favored more frequently. Under contingent coding, a recently proposed weighting algorithm produces the most accurate results for maximum parsimony. However, Bayesian inference outperforms all parsimony-based solutions to handle character dependency due to fundamental differences in their optimization procedures—a simple alternative that has been long overlooked. Yet, we show that the more primary characters bearing secondary (dependent) traits there are in a dataset, the harder it is to estimate the true phylogenetic tree, regardless of the optimality criterion. owing to a considerable expansion of the tree parameter space.

Keywords—character dependency, character coding, performance, phylogenetic accuracy, distance metrics, morphological phylogenetics, Bayesian inference, maximum parsimony.

One of the most important assumptions common to all phylogenetic methods, regardless of their optimality criteria, is that individual variables within any given dataset (e.g., morphological characters or molecular sites) are independent from each other (Farris et al. 1970, Felsenstein 2004). In practice, however, there may exist several variables within a given data matrix that share some level of dependency among each other. Such dependencies can be either logical—the state (or condition) of a variable depending directly on the state of another variable—or biological—e.g., evolutionary integration among two or more variables. Biological dependencies theoretically occur in molecular and morphological datasets (Brazeau et al. 2019), but both types of dependencies are conspicuous to morphological characters(Maddison 1993, Wilkinson 1995, Klingenberg 2008, Goswami and Polly 2010, Goswami et al. 2014). Despite existing guidelines to construct morphological characters in ways to minimize such dependencies (Sereno 2007, Simões et al. 2017a), it is almost impossible to completely avoid them for most empirical datasets. Consequently, character dependency has a direct and pervasive impact in datasets that can only be analyzed with morphological data (e.g., paleontological datasets), or which include morphological and molecular data to integrate fossils and extant taxa in total evidence phylogenetic inference—e.g., (Pyron 2011, Simões et al. 2018b, Mongiardino Koch and Thompson 2020, Ballesteros et al. 2022).

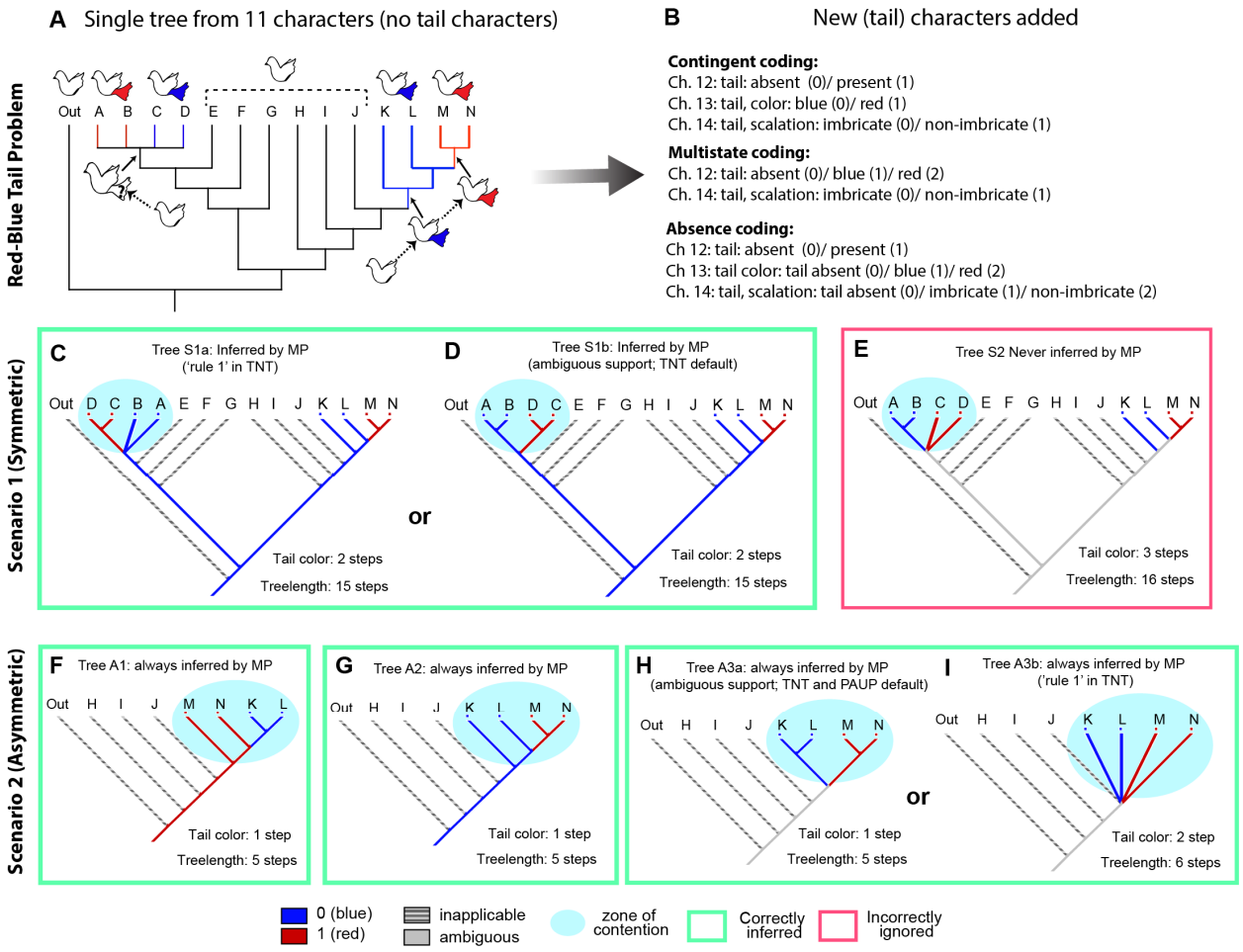
One of the most common forms of logical dependency in morphological phylogenetics are hierarchical characters—i.e., a set of two or more characters, including one primary character (governing the absence or presence of an anatomical structure) and one or more secondary characters (governing various properties of that same structure). A classic example of this logical dependency was introduced by (Maddison 1993) and is known as the Reb-Blue Tail (RBT) problem. In the latter, tails can be absent/present (primary character), but tail color (secondary

character) can only be determined for species in which the primary character is present (Fig. 1). Characters with such hierarchical structure are widespread in morphological datasets, especially those designed to assess higher-level phylogenetic relationships. The latter are more prone to include anatomical structures that originate (neomorphisms) or disappear across major branches of the tree of life, thus making all traits related to such structures secondary characters directly dependent on their presence. Examples of this include the origin of limb bones, which results in all limb related characters acting as secondary characters, during the fish-tetrapod transition (Simões and Pierce 2021); multiple independent limb losses within squamates (Wiens et al. 2006); the origin of wings in insects (Wipfler et al. 2019); the origin of all floral structures at the origin of angiosperms (Frohlich and Chase 2007). Therefore, hierarchical characters can be especially prone to impact phylogenetic datasets aimed towards understanding evolutionary transitions, as those are typically characterized by the origin or loss of major anatomical structures (and their dependent secondary characters). Furthermore, even datasets aimed at smaller taxonomic scales may include hierarchical characters, such as datasets focusing on early-deriving snakes, in which various cranial, limb, and pectoral girdle characters may be either absent or present (Garberoglio et al. 2019), directly impacting all secondary characters contingent upon those traits.

Historically, whenever a primary character is absent for any given taxon, secondary characters that are contingent on the presence of such primary character are traditionally treated as inapplicable (Maddison 1993). This is represented by the introduction of a gap (‘-’) or missing data symbol (‘?’) in the phylogenetic data matrix—in either case, that state is treated as the same by available algorithms in most phylogenetic programs (Brazeau 2011). This strategy, known as contingent (or traditional) character coding, incorporates the hierarchical relationships between

characters during tree inference, although it keeps these anatomical traits as separate characters despite their logical dependency (Maddison 1993, Strong and Lipscomb 1999). Additionally, the introduction of inapplicable or missing character state scores ('-' or '?') have an undesirable effect during tree search—the placement of taxa in a poorly resolved node in one sector of the tree being influenced by the placement of other taxa in another distant sector of the tree (Fig. 1). This is driven by the inability of most phylogenetic programs, especially by maximum parsimony (MP) algorithms, to find all possible resolutions for the tree node in question (Maddison 1993, Strong and Lipscomb 1999) —see also Supplementary Material.

Figure 1



An alternative to contingent coding—multistate coding—would merge all characters into a single character with multiple states, and it was the first proposed solution to the problem introduced by dependent characters by (Maddison 1993). Multistate coding removes the problem of dependency between anatomical traits but it does not recover the hierarchical relationship among them, thus removing the phylogenetic signal inherent to this important property and creating polytomic nodes that should have been resolved (Hawkins et al. 1997, Strong and Lipscomb 1999). In medium to large-sized datasets, it is also common for primary characters to have not just one, but multiple secondary characters dependent on it. In these cases, it is unfeasible to create a single multistate character including all possible combinations of secondarily dependent traits.

Besides multistate coding, numerous other solutions have been proposed over the past three decades to handle this simple but pervasive problem, from new character coding strategies (Maddison 1993, Hawkins et al. 1997, Strong and Lipscomb 1999, Hawkins 2000, Brazeau 2011, Tarasov 2019) to new algorithmic solutions (Brazeau et al. 2019, Tarasov 2019, Hopkins and St John 2021). The vast array of character coding schemes, their benefits and limitations, have been reviewed in many recent studies (Simões et al. 2017a, Brazeau et al. 2019, Hopkins and St John 2021), and so we refer the reader to these for further information (and also our Supplementary Material). In summary, despite the problems introduced by contingent coding, nearly all studies have agreed that contingent coding should be preferred over others as it is the least spurious solution to the problem of hierarchical characters (e.g., the RBT problem) (Strong and Lipscomb 1999, Sereno 2007, Brazeau 2011, Simões et al. 2017a).

As alternative coding schemes did not provide clear solutions to handle dependent characters, there was a recent shift in focus towards new algorithmic solutions rather than dataset

construction ones. The first, the Morphy maximum parsimony algorithm introduced by (Brazeau et al. 2019), aims to escape the problem of inapplicable characters in contingent coding by providing a distinct treatment of inapplicable scores—referred to as the MP-M algorithm herein. Subsequently, (Hopkins and St John 2021) suggested down-weighting secondary characters relative to primary characters, also using maximum parsimony—referred as MP-HSJ herein. Subsequently, (Goloboff et al. 2021) advocated for the usage of Sankoff matrices to model character contingency in maximum parsimony.

The performance of these recent alternative algorithmic solutions, however, remains largely unknown. Simulated datasets, in which the "true" answer is known, have only been used once to test phylogenetic accuracy using a small synthetic dataset (with eight taxa) and restricted to maximum parsimony optimization approaches (Hopkins and St John 2021). Although analyzing small-sized simulated datasets can be useful as a proof of concept to better understand the behavior of alternative optimization methods, these do not explicitly test the model complexities that are inherent to much larger datasets that are closer in size to empirical ones. Important parameters that can be modeled in more complex simulated datasets include variable levels of homoplasy, character evolutionary rates (contributing to branch lengths), tree symmetry, the proportion of primary and secondary characters, among others. For instance, previous studies have reported a significant performance disparity of various methods to accurately infer datasets originated from symmetric and asymmetric trees (O'Reilly et al. 2018, Puttick et al. 2019). Tree symmetry is also at the heart of the problem of hierarchical characters—(Maddison 1993), see also Fig. 1 and Supplementary Material herein—and so we should expect different performances from the proposed solutions to the problem of hierarchical characters across different models.

Additionally, it has been shown that the number of secondary characters for each primary character will affect the performance of distinct optimization procedures, as demonstrated by (Hopkins and St John 2021). However, another key factor is the proportion of primary characters that are associated with secondary characters within a dataset. For instance, in approaches that down-weight secondary characters (e.g., MP-HSJ), if 30 secondary characters are dependent upon a single primary character, then their total weight will add up to a maximum of 1 step for the total tree score, and their individual relative weights will be of only 1/30 (= 0.03). However, if these 30 secondary characters come from 5 independent primary characters (e.g., 6 from each primary character), then their total contribution to the tree score can add up to a maximum of 5, and their individual relative weights will thus be five times higher than in the previous example—1/6 (= 0.167). To our knowledge, the proportion of primary characters with secondary characters has never been previously investigated for its impact on phylogenetic accuracy, and we predict it should considerably impact various phylogenetic inference approaches.

Importantly, morphological datasets are now frequently analyzed by probabilistic/statistical methods—maximum likelihood and Bayesian inference (BI)—across various study systems—e.g.,(Lee et al. 2014, Giles et al. 2017, King et al. 2017, Simões et al. 2017b, Paterson et al. 2019, Simões and Pierce 2021). Yet, the problem of hierarchical characters has rarely been discussed in the context of probabilistic inference methods. One major exception is a recent study suggesting the polymorphic re-coding of characters following the concept of structured and hidden states Markov models to incorporate the hierarchical structure of primary and secondary characters into Bayesian inference, as a solution to the problem of hierarchical characters (Tarasov 2019). However, no study to date has demonstrated if and how the problems introduced by hierarchical characters in MP impacts probabilistic phylogenetic algorithms to

begin with, despite some previous suggestions that they would (Brazeau et al. 2019). At least in principle, theory suggests that likelihood-based methods should be less impacted by hierarchical characters. That is because all maximum likelihood and BI software implement variations of the Felsenstein likelihood optimization algorithm (Felsenstein 1973, 1981), which includes only a "down-pass" phase (from tips towards the root) for the calculation of likelihood scores at every node in the tree being reconstructed. The absence of an "up-pass" phase during the optimization of ancestral nodes—which is characteristic of maximum parsimony approaches (Brazeau 2011, Brazeau et al. 2019)—would suggest, for instance, that the dependency problem introduced by inapplicable state scores in contingent character coding should not impact tree inference using likelihood optimization procedures.

Here, we utilized a series of simulations of morphological datasets to address the following questions: how do different character coding schemes impact the relative performance of MP and BI in both simple and complex morphological datasets? Under a common coding scheme, how do classical and recently proposed optimization algorithms for MP perform relative to each other and to BI in morphological datasets? What is the impact of different tree and character models for the performance of each method? We find a striking contrast of results between simplistic and complex simulated datasets regarding best coding practices and a large disparity in performance among methods depending on tree or character distribution structures. As with other recent studies, our results are quite variable depending on the metric used for assessing accuracy and, using simulations, we demonstrate that quartet distance is less sensitive to tree resolution than bipartition-based metrics, being a better metric for differences in topology only. Finally, our results indicate that standard BI is significantly less impacted by contingent

coding, displaying superior performance to all MP methods tested here, even those explicitly model to handle inapplicable characters.

MATERIALS AND METHODS

Simulation 1: Simplified Synthetic Datasets

To make our study directly comparable to previous ones addressing issues of character coding, we replicate the simplified synthetic datasets used to exemplify the RBT problem of (Maddison 1993), which was also used by others (Strong and Lipscomb 1999, Tarasov 2019). Specifically, this includes two datasets aimed towards replicating the two distinct problematic scenarios introduced by contingent coding and inapplicable character states.

Dataset 1 (Scenario 1, symmetric trees).— Refers specifically to the RBT example of (Maddison 1993) with 14 taxa plus 1 outgroup with their internal relationships fully resolved and with each internal node supported by one synapomorphy, with the exception of the taxa within the so called zone of contention (Fig. 1a). A total of 11 characters are used to create this tree topology. The tree topology is symmetric and includes one fully resolved clade on the right side of the tree in which the primary character is present and the secondary character is applicable, and one unresolved clade in which the primary character is convergently evolving on the left side of the tree. Subsequently, one or two extra characters are added to the dataset (depending on the coding scheme to be tested). For all coding schemes in which two characters are added, "character 12" is the primary character (denoting absence and presence of tail) and "character 13" (denoting tail color) is the secondary character dependent on the primary character. Under multistate coding, a single "character 12" is present (Fig. 1b).

Dataset 2 (Scenario 2, asymmetric trees). —Simulates the tree example used by Strong & Lipscomb (1999, Fig. 12 therein). The objective with this dataset is to explore potential biases introduced by primary absences and resulting secondary inapplicable characters at the base of the tree. This dataset includes 7 taxa plus 1 outgroup with their internal relationships fully resolved and with each internal node supported by one synapomorphy, except for the taxa within the zone of contention (Fig. 1f-i). A total of three characters are used to create this tree topology. The tree topology is strongly asymmetric and includes a single zone of contention. As for Dataset 1, one or two characters are added to represent primary and secondary characters for the various coding schemes.

Simulation 2: Complex Synthetic Dataset

It is well-established that number of taxa (Hillis 1996, 1998, Pollock et al. 2002, Zwickl and Hillis 2002, Hillis et al. 2003, Heath et al. 2008, Vernygora et al. 2020), number of characters (Wright and Hillis 2014, O'Reilly et al. 2016, Puttick et al. 2017, Puttick et al. 2019)—but see (Keating et al. 2020)—and the relative number of taxa per character (taxon:character ratio) (Graybeal 1998) all impact the performance of phylogenetic analyses using both morphological and molecular data under different optimality criteria. Therefore, we kept the number of taxa, number of characters, and the taxon: character ratio all constant to avoid introducing the impact of those extra variables on tree inference accuracy. Specifically, we used the following fixed values: 31 taxa (30 ingroup taxa +1 outgroup) and 60 characters—and thus a fixed taxon:character ratio 1:2 for the ingroup, which approximates well the taxon:character ratio in empirical datasets (Scotland et al. 2003, Murphy et al. 2021).

The approach above gives us the following fixed parameters: *T* (total number of taxa), C (total number of characters), *R* (taxon/character ratio). Additionally, the total number of characters (*C*) can be represented by: $C = P_n + S_n$, where P_n is the total number of primary characters and S_n is the total number of secondary characters. As previously acknowledged (Hopkins and St John 2021), the proportion of secondary characters in the dataset will impact the outcome of the results. Therefore, we simulated three groups of datasets with increasing amounts of secondary characters relative to the total number of characters: 10%, 25% and 50%. Given a constant total of 60 characters, the latter translates into S_n= 6, 15 and 30 secondary characters, respectively (Table 1).

TABLE 1. Combinations of characters distribution models.

Model	C	S_n(%C)	S_n (absolute)	P_n	P_s	S_d
M1	60	10	6	54	1	6
M2	60	25	15	45	1	15
M3	60	50	30	30	1	30
M4	60	50	30	30	2	15*
M5	60	50	30	30	5	6*

*Note that the number of secondary characters per primary character (S_d)on models M4 and M5 are the same as in models M2 and M1, respectively. However, the secondary characters in M4 and M5 are distributed across more primary characters (P_s), which will impact the final Fitch scores and tree lengths.

As discussed above, another key factor is how secondary characters are distributed among primary characters. For instance, in approaches that down-weight secondary characters (e.g., HSJ), if 30 secondary characters are dependent upon a single primary character their total

weight will add up to a maximum of 1 step for the total tree score, and their individual relative weights will be of only 1/30 (= 0.03) (for a HSJ α =1). However, if these 30 secondary characters come from 5 independent primary characters (e.g., 6 from each primary character), then their total contribution to the tree score will add up to a maximum of 5, and each secondary character's relative weight will be five times higher than in the previous example—1/6 (= 0.167). Therefore, secondary characters may have quite different weights depending on the relative distribution of secondary characters among primary characters. To account for the latter, we introduced another variable to our simulations: the number of secondary characters per primary characters (S_d), with the relationship $S_d = S_n/P_s$, where P_s is the number of primary characters with dependent secondary characters. For instance, if we have 30 secondary characters dependent on just one primary character—as in all examples from (Hopkins and St John 2021), where all secondaries are dependent on a single primary character— that would be a case where:

$$60(C) = 30(P_n) + 30(S_n)$$

and,

S_n = 30 and P_s = 1, then $S_d = S_n/P_s = 30$ secondary characters per primary character.

However, if we have 30 secondary characters dependent upon 5 primary characters:

S_n = 30 and P_s = 5, then $S_d = S_n/P_s = 6$ secondary characters per primary character.

Therefore, here we simulated three categories for the distribution of secondary characters for datasets with 30 secondary characters: S_d = 6, 15, and 30 secondary characters per primary character (Table 1).

279 *Simulated tree construction.*—We generated two simulated master ("true") trees, one
280 fully symmetrical and another with perfectly asymmetrical topology, to test for the impact of
281 different tree symmetries on phylogenetic performance. Each tree included 31 taxa (30 ingroup
282 and 1 outgroup) as defined in the previous section. To emulate the RBT problem, we designated
283 10 'crown' taxa in each sector of the symmetrical tree (total = 20 taxa) and 10 'crown' taxa in
284 the asymmetrical tree—therefore fixing to 10 the number of taxa with applicable secondary
285 characters forming the zone of contention (Fig. S1). All 'stem' taxa lying rootward of the
286 'crown' were designated to have the primary character absent, thus being inapplicable in respect
287 to secondary characters. Because our simulation design focused on generating morphological
288 data with no common evolutionary mechanism (Puttick et al. 2019), the branch length parameter
289 was omitted from the generated master trees.
290
291 *Simulated dataset construction.*—We used each simulated tree to generate 100 replicates
292 of binary morphological data matrices for each set of parameters (Models 1-5; listed in Table 1),
293 following the conceptual approach of (Puttick et al. 2019). This procedure does not use explicit
294 molecular substitution models to simulate morphological datasets, as in most previous
295 simulations of morphological datasets—e.g., (Wright and Hillis 2014, O'Reilly et al. 2016,
296 Puttick et al. 2017, O'Reilly et al. 2018, Vernygora et al. 2020). Instead, each individual
297 character is first defined as either homologous or homoplastic based on a probability function of
298 character homoplasy derived from an extensive survey of empirical datasets (Goloboff et al.
299 2017, Puttick et al. 2019). This approach is designed to generate morphological characters with a
300 model that does not necessarily favor probabilistic inference approaches—in fact, possibly
301 favoring MP (Puttick et al. 2019)—for directly comparing the performance of MP and

probabilistic methods in phylogenetics (Puttick et al. 2019). For homologous characters, terminal taxa are assigned states that result in the minimum number of character state transformations and therefore have a consistency index (CI)=1. If a character was defined as homoplastic, character states were assigned to the terminal taxa to produce CI<1. For our simulated datasets, we set a target CI index for the entire matrix to be within an intermediary range between 0.4 – 0.5 [bin 5 in (Puttick et al. 2019)].

We generated datasets using a two-step procedure. First, we generated all primary characters that were applicable to all taxa. Primary characters that were designated to have secondary characters were assigned a specific pattern of character state scores: present [char.state = 1] in the 'crown' ten taxa and absent [char. state = 0] in the outgroup and 'stem' taxa (Fig. S1). Next, we performed a second round of simulations to generate scores for the secondary traits only. The latter included pruned versions of the master trees only with taxa that were scored as having the primary characters as present. These simulated secondary data matrices were then merged with the primary data matrices. Taxa that were scored as 'absent' for the primary traits were scored as 'inapplicable' for the secondary traits in the final merged datasets. All simulated datasets contained variable characters only, which is typical of morphological datasets.

Analyses of Simulated Datasets

MP-F tree searches for the simplified datasets generated by Simulation 1 for distinct coding strategies were conducted using the "Implicit Enumeration" algorithm in the software TNT v.1.5 (Goloboff and Catalano 2016). For Simulation 2, tree searches were conducted using the *phangorn* R package (Schliep et al. 2017). For tree searches with MP-M optimization we used its implementation in the R package *TreeSearch* v1.0.1 (Smith 2018), which uses

MorphyLib (Brazeau et al. 2017) to handle inapplicable data (Brazeau et al. 2019). Tree searches with MP-HSJ optimization, we used the "*dissimilarity*" and " *hsjScorer*" R functions from (Hopkins and St John 2021) in conjunction with the branch-swapping algorithms available in the package *TreeSearch* v1.0.1 (Smith 2018). Starting rooted trees were subject SPR and TBR branch swapping operations, the results of which were used as starting trees for further analyzes with a series of ratchet iterations (functions "*Ratchet*" and "*pratchet*"), switching to the next run if the best score was hit 10 times, and stopping all searches if best score from each run was the same for 20 runs. The best scoring tree was used as the starting point for multiple ratchet (function "*MultiRatchet*") runs with the same criteria as above to obtain multiple most parsimonious trees.

For the MP-HSJ optimization, we further tested the performance of distinct α rescaling parameter values—for details on its implementation, see (Hopkins and St John 2021). In summary, when $\alpha = 0$, secondary characters are disregarded entirely from the analysis (weight = 0), and when $\alpha = 1$, secondary characters will not be further penalized, although all characters that are secondary to a primary character will still have a combined maximum score value of 1. To see the impact of different α values on the performance of MP-HSJ optimization, we tested for a range of three possible α values: 0, 0.5 and 1.

Bayesian analyses used the Mk model for morphological characters assuming the presence of variable characters only (Mkv model), with rate variation among characters sampled from a gamma distribution. Each analysis consisted of two independent runs using four chains each, sampling at every 1000 generation, for a total of 10 million generations using the software Mr. Bayes v 3.2.6 (Ronquist et al. 2012).

All most parsimonious trees (MPTs) obtained from each optimization procedure were used to calculate a strict consensus tree. Posterior tree samples obtained by BI were used to calculate a majority rule consensus tree. Both consensus options were chosen as they are the standard output trees for each of those respective optimization procedures in most studies using morphological data. Consensus trees were subsequently used for comparison with the master trees generated by simulations.

Performance Measures

We measured accuracy based on the total similarity shared by the inferred trees to the generated master trees using both bipartition and quartet tree distance metrics. For bipartition comparisons, we used similarity scores based on the Mutual Clustering Information metric (MCI) (Smith 2020), an information theory-based metric that shows the amount of mutual clustering information shared by all bipartitions in two or more trees. The latter is part of a larger class of generalized Robinson-Foulds (RF) distance metrics that overcome the limitations from classical implementations of the RF distance, such as quick saturation of distance scores (Smith 2020). Quartet similarity is based on the "tqDist" algorithm from (Sand et al. 2014)—implemented in the R package *Quartet* (Smith 2019)—to measure the number of shared four-taxon subtrees between two or more trees.

Quartet similarity is predicted to outperform bipartition metrics as it better reflects phylogenetic patterns at deeper internal nodes, thus better handling poorly resolved nodes (Mongiardino Koch et al. 2021)—a problem for previous tree distance metrics, including traditional RF and Matching split distances [e.g., (Vernygora et al. 2020)]. Further, quartet similarity is less prone to the influence of wildcard taxa and tree shape (Smith 2020,

Mongiardino Koch et al. 2021). Accordingly, we found several instances in which MCI and quartet similarity differed when applied to the same trees, and so we simulated how each metric is impacted by decreased tree resolution or increased topological differences to test the precise conditions in which these metrics yield different results. For both the asymmetric and the symmetric 30-taxa master trees, we randomly collapsed from 1 to 28 internal nodes and calculated MCI and Quartet similarity to the starting tree. Similarly, we randomly applied from 1 to 45 nearest-neighbor interchange (NNI) moves and compared the resulting tree to the starting tree under both metrics. For each number of collapsed nodes or NNI moves, we did 50 replicates. Finally, we compared both metrics in terms of their sensitivity to the number of collapsed nodes (tree resolution) or number of NNI moves (topological differences), and whether tree symmetry affected either metric.

As discussed in detail in our Results, we found a superior performance of quartet distances over bipartition metrics (e.g., MCI) in instances of poor node resolution (Fig. 2). This limits our ability to infer resolution error, since this metric is calculated based on bipartition tree distances (Smith 2020). Hence, we only evaluated resolution error when results from MCI matched the results obtained by quartet distances.

Finally, considering the BI is not intended to provide a point tree estimate, we also examined the size of the parameter space using different coding schemes for BI results. We did that by calculating the mean and variance of RF distances among the post-burnin trees of the posterior sample *sensu* Wright and Lloyd (2020). Since the trees in the posterior sample do not contain polytomies, the RF distance metric is not impacted by differences in tree resolution (see Results). This metrics provide a perspective on tree disparity in the posterior sample (i.e., how loosely or tightly scattered trees are in the posterior distribution).

393

394 *Statistical Analyses*

395 To assess if there were significant differences between performance results among
396 different tree and character models by inference method type, we conducted nonparametric
397 pairwise Wilcoxon rank sum (Mann-Whitney) between all analyses (Supplementary Tables 1-3).
398 Parametric tests were not possible considering the bimodal distribution of some of the results
399 (e.g., Figs 3-5).

400

401 RESULTS

402 *"Solutions" to the RBT Problem—a Conceptual Paradox*

403 There are only three pieces of phylogenetic information universally present within
404 primary and secondary characters as illustrated by the RBT problem (Fig. 1): i) the primary
405 character groups all taxa with tails together and those without tails as a second clade; ii) the
406 secondary character (tail color) groups red-tailed taxa together and blue-tailed taxa together; iii)
407 the biological dependency of the secondary character upon the primary character indicates that
408 all aspects of the secondary character should be only applicable to taxa in which the primary
409 character is present (defining the clade with tail). Beyond these three aspects, there is no data
410 provided by either the primary or secondary characters to inform which tail color evolved first.
411 In fact, the latter is irrelevant for tree inference under either MP or probabilistic methods, since
412 reconstructing the direction of character state transformation (i.e., identifying synapomorphies) is
413 only performed by MP upon the rooting of the tree once the most parsimonious solutions have
414 already been found (Nixon and Carpenter 1993, 2012). For probabilistic methods (maximum
415 likelihood and Bayesian inference) outgroup comparison and the direction of character-state

transformation is not taken into consideration during tree sampling (Felsenstein 1973, 2004). Therefore, in the absence of additional characters, there is no single solution to the RBT problem as presented in Scenarios 1 and 2—*contra* (Tarasov 2019). Instead, any coding method or inference algorithm should allow the two possible solutions (i.e., red and blue first hypotheses) to be equally likely, and the latter should only be considered within the clade composed by taxa where the primary character is present. Therefore, the information content of any set of primary and secondary characters do require that the following criteria should be met for any coding approach or inference method to produce logically plausible and biologically realistic results:

Corollary 1.—Secondary characters (e.g., tail color) can only evolve within a clade where the primary character is present (e.g., tail is present). This hierarchical relationship is important both biologically and methodologically, as the inability to recover this hierarchical relationships will inevitably lead to the loss of tree resolution (Hawkins et al. 1997).

Corollary 2.—As we have no prior information on which state of the secondary character (e.g., tail color) evolved first, all known states (e.g., red and blue tails) should be considered as equally parsimonious/likely to be the ancestral condition. Under MP, solutions S1-S2 (Fig. 1c-e) and A1-A3 (Fig. 1f-i) should all be inferred as equally most parsimonious. For BI, tree topologies with blue evolving first and red evolving first should all be equally likely to be inferred and subsequently sampled from the posterior distribution with similar frequencies, considering all other parameters remaining constant. Both hypotheses should also have similar posterior probabilities.

To meet expectations from both corollaries above, two or more distinct tree topologies should be estimated for a coding or inference approach to be valid, depicting tree topologies with both valid solutions within the zone of contention (e.g., blue-first vs red-first hypotheses).

Additionally, all resulting trees should have the primary character grouping all taxa with the present condition within the zone of contention. Therefore, the consensus (strict or majority rule) tree estimated from the output trees meeting these criteria will necessarily include all taxa in the zone of contingency as monophyletic (supported by the primary character), but with no particular preference for either blue or red evolving first. Hence, the consensus tree should necessarily be unresolved, depicting a polytomic relationship for the taxa within the zone of contention.

Simplified synthetic datasets

Fitch MP (MP-F) .—Under MP-F, we find that four combinations of coding schemes/tree topologies meet the two corollaries for logically sound resolutions of the RBT problem (Table 2). One is provided by contingent coding under Scenario 2 (asymmetric trees), but which fails under Scenario 1 (symmetric trees), as illustrated in Fig. 1 (f-i) and discussed in the Supplementary Material. A second coding scheme to meet both corollaries is represented by multistate coding under Scenario 1 (symmetric trees), which had been highlighted by (Maddison 1993) as a solution to the contingent coding problem (Table 2). However, multistate coding fails under Scenario 2 as it cannot recover the hierarchical relationship between primary and secondary characters— as previously observed by (Hawkins et al. 1997). The latter results in some taxa (in which the primary character is absent) to be estimated as nested within the zone of contention, and a strict consensus tree with reduced resolution relative to other coding schemes (Figs. S2-S7). Finally, all options including character ordering logically prevent the basic assumption set by corollary 2, as the ordering scheme will inevitably predetermine which secondary state (red or blue) will evolve first (Figs. S6 and S7, Table 2).

461 The only coding approach to successfully meet the conditions set by corollaries 1 and 2
462 above under both symmetric and asymmetric trees (Scenarios 1 and 2) is “absent coding” (Fig.
463 S8 and S9, Table 2). Despite being briefly discussed in the literature before, absent coding was
464 tested only once (Strong and Lipscomb 1999), and its ability to meet both corollaries was never
465 previously realized (Supplementary Material).
466
467 *Morphy MP (MP-M)* .—This approach correctly recovers the hierarchical relationship between
468 primary and secondary characters as well as correctly finding the blue-first and red-first
469 hypotheses as equally parsimonious among the MPTs (Figs. S10 and S11, Table 2). This
470 matches the expectations of both corollaries, as predicted (Brazeau et al. 2019).
471
472 *HSJ MP (MP-HSJ)* .—As with MP-M, this approach was designed to correctly recover blue-first
473 and red-first hypotheses as equally parsimonious (Hopkins and St John 2021). As expected, it
474 does recover those hypotheses among the MPTs (Figs. S12 and S13, Table 2). The hierarchical
475 relationship between primary and secondary characters are recovered, but we note that those
476 must be provided by the user in the form of text file indicating a priori what characters are the
477 primary characters and secondary characters.
478
479 *Bayesian Inference-Mkv model (BI)* .—Using traditional (non-clock) BI and traditional modelling
480 of morphological characters (Mkv model), we found a substantial contrast of performance
481 between scenarios 1 and 2 concerning hierarchy (corollary 1). Regardless of the character coding
482 scheme, BI analyses of symmetric trees always inferred the clade defined by the presence of the
483 primary character (i.e., tail) as monophyletic in more than 90% of the sampled posterior trees

(Figs. S14-S16, Table 2), and the posterior trees sampled successfully converged towards an optimal tree topology solution (Fig. S14-16, c,d). Additionally, frequency among posterior trees for the correct inference of the clade defined by the presence of the primary character (i.e., tail) was slightly higher for absent coding (98.7%), compared to contingent coding (97%) or multistate (92.9%) coding.

In contrast, asymmetric trees were much harder to estimate using BI across all coding schemes, with the posterior sample of trees not converging towards similar topologies (Fig. S17-19) and with the focal clade defined by the primary character being inferred at drastically lower frequencies compared to symmetric trees (Table 2). However, the absent coding scheme still was the best performing one relative to competing coding schemes in this aspect (ca. 50% compared to 21 and 23% from other schemes).

Additionally, we expected the frequency of posterior trees inferring red and blue-first hypotheses to be similar to each other under corollary 2. We found exactly this pattern with almost identical sampling frequencies (<1% of difference) in the frequency of trees with blue or red first hypotheses under absent and multistate coding for symmetric trees (Scenario 1) (Table 2). We found similar results using absent and contingent coding for asymmetric trees (Scenario 2). However, contingent coding in Scenario 1 strongly favored a blue-first hypothesis (similarly to MP-F), whereas multistate coding in Scenario 2 favored a red-first hypotheses more strongly. As with MP-F, absent coding was the only coding scheme meeting both corollaries for both simulated scenarios.

TABLE 2. Results for the simplified synthetic datasets using various coding schemes. Coding schemes meeting expectations from corollaries 1 and 2 are highlighted with blue background. Coding schemes with results pre-established by users (ordered characters) highlighted in gray**.** Results for coding schemes that are not applicable to particular methods are marked with "NA". Abbreviations: Abs, absence coding; B, blue tail-first hypothesis; Cont, contingent coding; Cor, corollaries; M, method; Multi, multistate coding; P-S, primary and secondary character hierarchy; ord, ordered; R, red tail-first hypothesis; unord, unordered.

		Scenario 1 (Symmetric/two zones)					**Scenario 2 (Asymmetric/one zone)**				
M	**Cor**	**Abs**		**Cont**	**Multi**		**Abs**		**Cont**	**Multi**	
		Ord	**Unord**		**Ord**	**Unord**	**Ord**	**Unord**		**Ord**	**Unord**
MP-F	1	yes	yes	yes	yes	yes	yes	yes	yes	yes	no
	2	no	yes	no	no	yes	no	yes	yes	no	yes
MP-M	1	NA	NA	yes	NA	NA	NA	NA	yes	NA	NA
	2	NA	NA	yes	NA	NA	NA	NA	yes	NA	NA
MP-HSJ	1	NA	NA	yes	NA	NA	NA	NA	yes	NA	NA
	2	NA	NA	yes	NA	NA	NA	NA	yes	NA	NA
BI	1	yes	yes (98.7%)	yes (97%)	yes	yes (92.9%)	yes	no (50.2%)	no (21.13%)	yes	no (23%)
	2	no	yes (B-R <1%)	no (B-R=26%)	no	yes (B-R <1%)	no	yes (B-R <1%)	yes (B-R <1%)	no	no (B-R=15.7%)

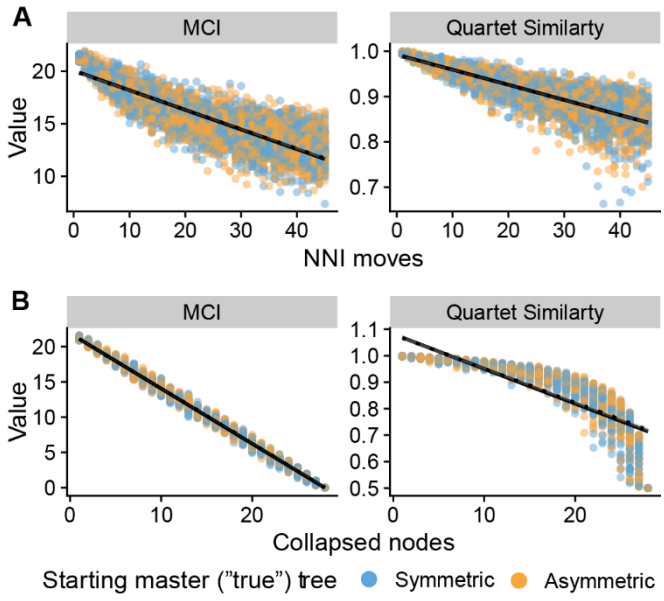
* Yes if >90% of posterior trees infer the focal clade (defined by primary character being present) as monophyletic.

**Yes if difference in frequency between blue (B) and red (R)-first hypotheses <1%.

Complex synthetic datasets

Performance of tree distance metrics.—We found that both metrics are insensitive to the symmetry of the starting tree (Fig. 2). For both MCI and Quartet similarity, similarity decreases approximately linearly with the number of NNI moves (Fig. 2a). MCI show signs of saturation earlier than Quartet similarity, with a decreasing slope as NNI moves increase, while for Quartet similarity the relationship continues approximately linear even when the number of NNI moves is greater than the number of internal nodes in the tree (Fig. 2a). The two metrics differ more strongly in their response to decreased tree resolution, however. While MCI decreases approximately linearly with the number of collapsed nodes, quartet similarity is less sensitive to decreased tree resolution when the number of polytomies is small and decreases sharply when trees approach a complete polytomy (Fig. 2b).

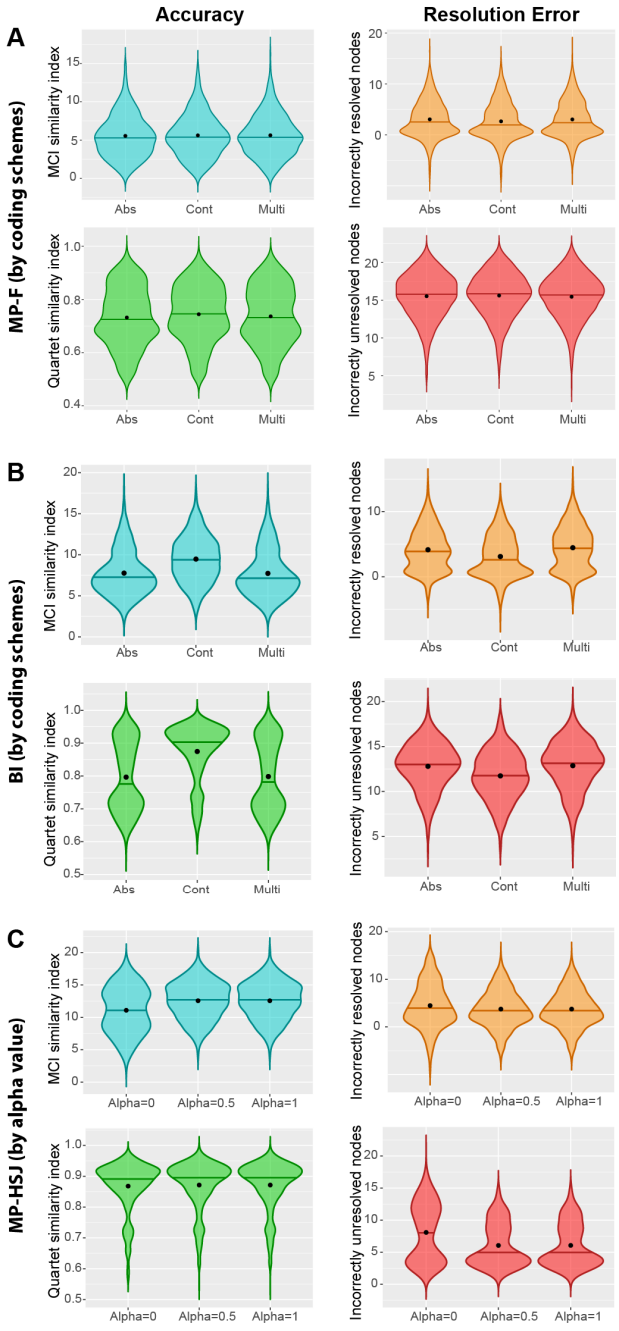
Figure 2



Performance across coding and alpha schemes.—Only two methods could be tested for different coding schemes (MP-F and BI), since the two other MP methods (MP-M and MP-HSJ) were

538 designed to handle datasets constructed using contingent coding schemes specifically.
539 Additionally, we tested the performance across different weighting schemes for secondary
540 characters (alpha variable) for the MP-HSJ optimization (Hopkins and St John 2021), which was
541 previously untested.

542 **Figure 3**
543



544

Under MP-F, all coding methods had extremely similar performances regardless of the tree distance metric used (Fig 3a). Given the extremely similar results presented by both metrics, we evaluated the resolution error incurred by different coding schemes—see Methods. Resolution error was also identical across all three coding methods for both Type I (incorrectly resolved notes) and Type II (incorrectly unresolved nodes) for all coding schemes.

Under BI, however, mean, median, and modal accuracy values were significantly higher for contingent coding relative to absent and multistate coding under both MCI and quartets tree distance metrics (Fig. 3b). Furthermore, resolution error results indicate contingent coding induces a slightly lower amount of Type I and II errors compared to absent and multistate coding.

For the MP-HSJ optimization, quartet distances indicate no substantial difference in performance across distinct alpha values, whereas MCI indicates a likely worse performance for alpha values of 0 relative to 0.5 and 1, which is induced by higher proportions of Type II error (Fig. 3c).

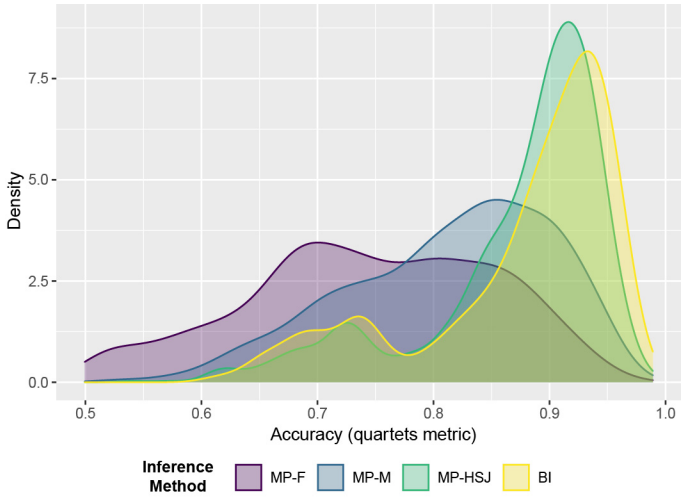
Performance across methods.—When comparing all methods based on contingent coding—the best performing coding procedure (Fig. 3a and b) and the only one common to all inference methods—MP-F has the worst accuracy compared to all other methods (Fig. 4). This result is consistent with predictions in the literature and is consistent regardless of accuracy metric (Fig. S20). However, the best solution among the three remaining methods depends on the performance metric. Similarity scores based on MCI (Smith 2020) suggests MP-HSJ perform the best whereas quartet distances indicate BI performs more accurately than other inference methods (Fig. S20). However, quartet distances were found to be more robust to variations in

568 tree resolution when compared to bipartition metrics here (Fig. 2)—an important factor when
569 comparing consensus trees, as done herein. Considering this, we favor the results provided by
570 quartet distances, which suggest BI outperforms all inference methods based on MP, even those
571 specifically designed to handle inapplicable characters.

572 When examining the tree-to-tree distances within each posterior sample (Fig. S21), we
573 observed that simulation conditions in which secondary characters are spread more evenly
574 among primary characters showed higher mean RF distances (i.e., models 3, 4, and 5). It should
575 be noted that unlike in accuracy comparisons between methods, a higher RF score does not mean
576 more differences from a "true" or simulation tree. This is a metric of within-posterior sample
577 differences. In this case, a higher RF means that more different trees are being proposed and
578 evaluated in these simulation conditions. We confirmed this by calculating a per-posterior
579 variance in the RF distance. This measure, too, indicated that greater dispersal of secondary
580 characters is associated with exploring more disparate phylogenetic trees (Fig. S22).

581

582 **Figure 4**



583

584

585 *Performance across tree and character models.*—The larger data dispersal and bimodality in the
586 results for each inference method (Fig. 4) suggest that other factors influence their respective
587 performance, two of which were explicitly modeled here: tree symmetry and distribution of
588 secondary characters among primary characters.

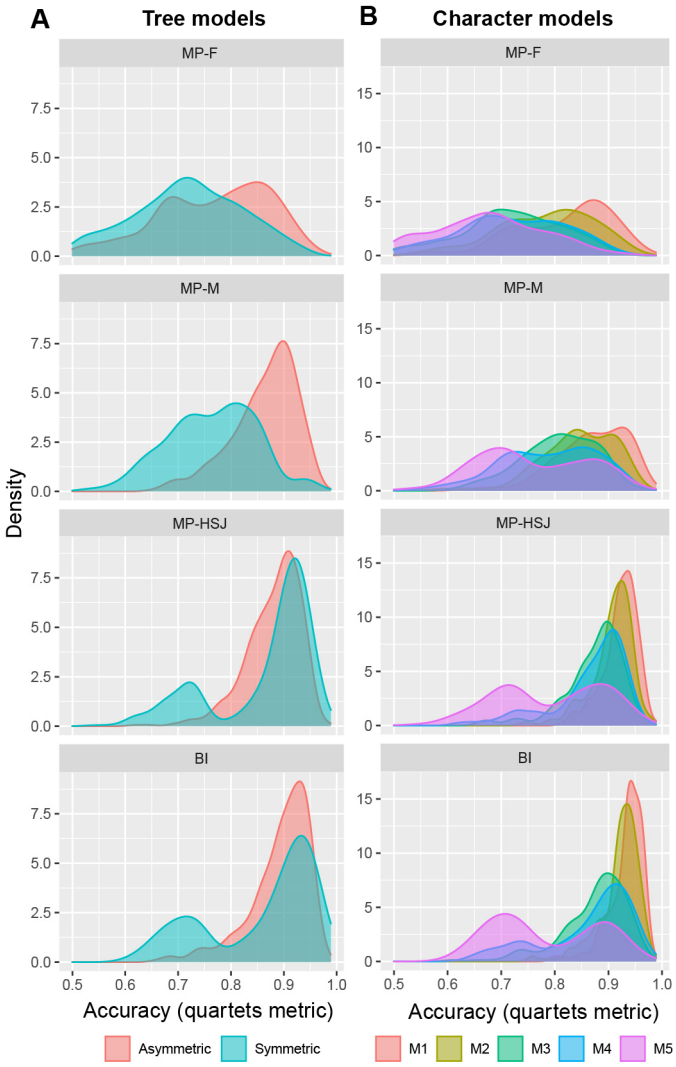
589 Using quartets distances, MP-F performs significantly better for asymmetric trees
590 compared to symmetric trees (Fig. 5a, Figs. S23 and S24, and Table S2), as predicted by the
591 RBT problem (Maddison 1993) and in our simplified synthetic datasets (Fig. 1 and Table 2).
592 MP-M performs significantly better than MP-F for both tree models, and with asymmetric trees
593 also significantly more accurately inferred compared to symmetric trees. MP-HSJ and BI have
594 greater accuracy relative to MP-M and MP-F (Fig. 5a, Figs. S23 and S24). The latter two
595 methods perform relatively similarly for datasets used to reconstruct symmetric and asymmetric
596 trees, with a slight advantage for symmetric trees (although nonsignificant for MP-HSJ). The
597 greatest improvement in performance for MP-HSJ and BI relative to MP-F and MP-M is
598 observed on the inference of symmetric trees (Fig. 5a,), suggesting they are more capable than
599 MP-M of removing the problems introduced by inapplicable characters.

600 In contrast, the MCI metric suggests that accuracy in MP-F tree inference is similar for
601 symmetric and asymmetric trees (Figs. S23 and S24), thus going against all predictions above
602 and previous evidence from the literature indicating symmetric trees (as in Figs. 1a, c-e) are
603 considerably harder to estimate using MP-F compared to asymmetric trees (as in Fig. 1f-i) in the
604 presence of inapplicable scores for hierarchical characters. This further suggests this metric is not
605 capable of detecting meaningful differences in performances across methods.

606 The performance of distinct inference methods when considering different primary and
607 secondary character distribution models (Table 1) indicates a significant decrease in accuracy of

608 MP-F when increasing the number of secondary characters per primary character (M1-M3), or
609 when increasing the number of primary characters bearing secondary characters (M3-M5) (Figs.
610 5b, S25 and S26, and Table S3). Such decrease in accuracy is also observed among other
611 methods under the same circumstances, but to a much lower extent, except for model M5. In the
612 latter, the increase in the number of primary characters bearing secondary characters dependent
613 upon them substantially decreases performances across all methods (Fig. 5b).

614 **Figure 5**



615

616

DISCUSSION

Differences between quartet and bipartition metrics to measure method accuracy

Here we found that quartet and bipartition metrics favor different inference methods. Our simulations show that this is likely due to a difference in the sensitivity of each metric to tree resolution in summary trees and topological differences, but not to tree symmetry. MCI decreases approximately linearly with tree resolution and small topological differences (Fig. 2). As a result, when trees being compared include polytomies (e.g., most summary or consensus trees from MP and non-clock BI studies), the underlying cause of distances estimated may be ambiguous. Quartet similarity, on the other hand, appears to be less sensitive to polytomies except for extreme cases, better reflecting differences in topology. When applied only to fully resolved trees, MCI possesses several desirable properties in relation to other metrics, including Quartet Similarity (Smith, 2020). When trees vary both in topology and resolution, however, interpretation from MCI can be problematic. By using of both metrics, we are able to find that BI results in more accurate but less resolved trees, while MP-HSJ results in trees with higher information content shared with true trees because they are better resolved, although less accurate (i.e., include more false positives).

Advantages of contingent coding over other coding schemes under MP and BI

It has long been suggested that contingent coding is the less spurious solution to the problem of dependent characters despite the introduction of inapplicable character states in secondary characters (Strong and Lipscomb 1999, Sereno 2007, Brazeau 2011, Simões et al. 2017a). However, this assumption had never been tested using complex simulated morphological datasets, and nearly all conclusions regarding distinct coding strategies come from small,

simulated datasets (Strong and Lipscomb 1999, Brazeau et al. 2019, Hopkins and St John 2021), equivalent in size and scope to our Simulations 1 (simplified synthetic datasets). By examining both symmetric and asymmetric tree structures for Simulations 1 and ancestral state reconstructions for each of the three optimization procedures tested here (contingent, absent, and multistate), we find new results and interpretations concerning the utilization of these coding schemes. We find that the problems introduced by character dependency are most easily avoided by using absent coding instead of contingent or multistate coding (Table 2, Figs. S8and 9), thus going against previous suggestions concerning this particular coding strategy using similarly small synthetic datasets (Strong and Lipscomb 1999, Brazeau et al. 2019, Hopkins and St John 2021).

We attribute some of this difference to the fact that ancestral state reconstructions were not conducted for all outputs of distinct coding strategies by (Strong and Lipscomb 1999), among other issues in the interpretation their results—see Supplementary Material. Additionally, the other two studies (Brazeau et al. 2019, Hopkins and St John 2021) used a distinct, although analogous, approach to absent coding as defined here, in which inapplicable scores were interpreted as a new character state—i.e., gaps (‘-‘) interpreted as a third character state for otherwise binary characters. Therefore, some of the difference in results may derive from the fact that interpreting inapplicable scores as a distinct third state is not, strictly speaking, the same as scoring it with the absent state, as the latter is homologous to the absent state on the primary character. Additionally, the simplistic simulations of (Hopkins and St John 2021) introduced more secondary characters, which might have increased the negative impact of overweighting the new character state—a problem also pervasive to absent coding, as described below.

By comparing the results of our Simulations 1 with more complex simulation scenarios (Simulations 2) we find important contrasts in our results and to previous conclusions using simplified datasets. When simulating larger datasets with explicit tree and character model variations, there is no significant difference in accuracy or resolution error among distinct coding strategies for traditional MP (MP-F), regardless of the performance metric (Fig. 3). We attribute this difference to the fact that the detected advantages of absent coding in simplified simulations (the only coding method meeting the assumptions of corollaries 1 and 2 discussed above) is counterbalanced by the negative bias introduced by the repeated occurrence of the absent state. As the number of secondary characters increases for larger datasets, it also increases the number of secondary characters with the absent condition, disproportionally overweighting the absent state. Although we did not explicitly test for a variable number of characters, we predict that datasets with a larger number of characters analyzed by traditional MP (MP-F) might see an even greater negative impact from the overweighting of the absent condition with absent coding, potentially leading contingent coding to become the most accurate coding, as previously suggested (Strong and Lipscomb 1999, Sereno 2007, Brazeau 2011, Simões et al. 2017a).

Under BI, contingent coding has a slightly superior performance compared to other coding schemes for the complex simulated datasets (Simulations 2) (Fig. 3). This is expected from theory since BI is not as strongly impacted by inapplicable scores introduced by contingent coding as the Fitch algorithm for MP (MP-F) due to the absence of an “up-pass” phase in the former. Therefore, the advantages of absent relative to contingent coding detected for small datasets under MP-F are not observed under BI. However, as BI also suffers from the biases introduced by the overweighting of the absent condition, there is an overall negative balance for the performance of absent coding relative to other coding schemes.

685

686 *Limitations of approaches explicitly designed to deal with character dependency*

687 Perhaps the first attempt toward solving the problem of character dependency, outside the
688 scope of character coding schemes, was the utilization of step-matrices of costs—or Sankoff
689 matrices—as they could embed hierarchical relationships among characters (Forey and Kitching
690 2000). These have long been criticized for the amount of time required to build individual
691 matrices for every collection of primary character and their dependent secondary characters,
692 among other issues—e.g., (Brazeau et al. 2019). Recently, such problems were ameliorated by
693 faster methods to construct Sankoff matrices in the program TNT (Goloboff et al. 2021).
694 However, as the number of secondary characters increases in a dataset, this solution becomes
695 less practical as it surpasses the total possible number of states allowed by TNT (32 states). The
696 latter creates a maximum limit of four binary dependent characters (Goloboff et al. 2021). Even
697 if a larger number of character states are enabled the future implementations of TNT, the costs of
698 character state transformations would still have to be subjectively customized and without
699 accounting for uncertainty around such transformational costs. Therefore, Sankoff matrices may
700 never be a feasible universal solution to the problem of character dependency.

701 The Morphy (MP-M) approach (Brazeau et al. 2019) is, to our knowledge, the first
702 algorithmic attempt to revise traditional parsimony optimization schemes for discrete characters
703 (Fitch 1971), allowing for a more appropriate treatment of dependent characters. It was analyzed
704 conceptually and empirically by subsequent studies, which criticized MP-M for not controlling
705 for primary characters and their relationship to secondary characters (the same major limitation
706 of the Fitch algorithm), leading to overweighting of absences for controlling primary characters
707 (Hopkins and St John 2021). It was also detected that, by increasing the number of secondary

characters, the MP-M approach (just as with MP-F) would result in a larger set of MPTs, including solutions where secondary characters were treated as applicable, thus contrary to its primary goal—a behavior not detected for the MP-HSJ method.

Our results support and expand upon those findings, by establishing that MP-M optimization can improve on the performance of datasets with inapplicable scores when reconstructing asymmetric trees (Figs. 5b). However, MP-M faces similar difficulties as traditional parsimony (MP-F) in the case of symmetric trees (Figs. 5b, S17 and S23), precisely where the negative effects of inapplicable scores for contingent coding are expected to be the greatest (Maddison 1993, Brazeau et al. 2019, Hopkins and St John 2021). Additionally, MP-M has greater accuracy across different models of primary and secondary character distribution in the dataset compared to MP-F, but we note that significantly higher levels of accuracy are obtainable by MP-HSJ and BI under these same conditions (Figs. 5b, S23 and S24). The latter suggests that not only MP-M becomes less accurate than alternative methods (MP-HSJ and BI) when increasing the number of secondary characters for a single controlling primary character (models M1-M3 herein) as previously suspected (Hopkins and St John 2021), but it also becomes less accurate when increasing the number of primary characters with dependent characters (models M3-M5 herein).

Among all parsimony-based methods, MP-HSJ is consistently recovered as the best performing method to handle the problem of inapplicable scores for dependent characters, regardless of accuracy metric, tree structure, and character models simulated herein (Figs. 4, 5, S21-S24). We attribute this performance to the fact that this is the only approach that specifically identifies primary characters and each of their secondary character dependencies (Hopkins and St John 2021). However, MP-HSJ downweighs secondary characters to only a small fraction of the

relative weight attributed to primary characters, and this penalization increases proportionally to the number of secondary characters in a dataset. The downweighing of secondary characters may even be further boosted through the α parameter introduced by this optimization procedure (Hopkins and St John 2021). Our tests revealed that the downweighing of secondary characters by MP-HSJ is so extensive that performance results under this approach are nearly identical regardless of the chosen value of α (even for $\alpha = 0$, which completely eliminates secondary characters from the analysis) (Fig. 3c). Such heavy downweighing of secondary characters may pose a limitation for datasets in which those characters are the only ones available to resolve relationships within the zone of contention (e.g., Fig. 1). The latter might be one of the key reasons for the superior performance of BI relative to MP-HSJ under the most accurate metric (quartets), even though BI does not distinguish primary and secondary characters.

The inapplicable states problem is mostly restricted to MP

The primary cause for the problem of contingent coding and its impact on tree inference relates to the two-steps approach towards the optimization of ancestral state in MP—the "down-pass" and "up-pass" phases of the Fitch algorithm (Fitch 1971, Brazeau 2011). Since BI programs use the Felsenstein optimization (Felsenstein 1973, 1981) when calculating likelihoods for internal nodes, which has only a "down-pass" phase, it would be expected that the impact of inapplicable characters from contingent coding would be strongly reduced, or at least substantially minimized, relative to MP. Our results in Simulations 1 support our predictions in finding that contingent coding in MP-F will favor a blue-first hypothesis 100% of the time and never return any trees with a red-first hypotheses in Scenario 1 (Fig. 1, Table 2). On the other hand, BI will favor a similar hypothesis (blue-first = 46.1%) but it retrieves the competing

754 hypotheses at frequencies much higher than 0% (i.e., red-first = 21%) (Table 2). As expected by
755 their design, both MP-M and MP-HSJ accurately find most parsimonious trees with both blue
756 and red-first hypotheses.

757 The advantage of BI under Simulations 1 is limited to the better-studied Scenario 1
758 (symmetric trees). The difficulty of retrieving hierarchical relationships and reaching topological
759 convergence in small asymmetric trees causes BI to fail corollaries 1 and 2 more frequently than
760 MP-F when estimating asymmetric trees (Table 2). Our findings thus corroborate previous
761 studies suggesting symmetric trees can be more accurately reconstructed than asymmetric trees
762 using phenotypic data under BI (Puttick et al. 2017, Puttick et al. 2019), although we do not
763 recover such performance disparity for distinct tree models under MP-F.

764 Using more complex simulations combining several parameters and larger numbers of
765 taxa and characters (Simulation 2), BI again consistently recovers more accurate trees than MP
766 using the traditional Fitch algorithm (MP-F). How BI compares in performance to other
767 approaches designed to correct for the impact of inapplicable characters (MP-M and MP-HSJ)
768 depends on the measure of accuracy. BI performs equally well under various scenarios to MP-M,
769 but less accurately than MP-HSJ using the MCI metric. When analyzed under the quartet
770 similarity metric, which is less influenced by tree resolution (Figs. 4 and 5), BI is significantly
771 more accurate than the two parsimony approaches that correct for inapplicable characters.

772 Interestingly, solutions to character dependency have also been proposed in the context of
773 Bayesian inference in recent years, such as for the utilization of structured (SMM) and hidden-
774 state Markov models (HMM) (Tarasov 2019). While the latter study demonstrates that these
775 newer methods can adequately deal with inapplicable states in dependent characters, no study
776 has ever shown that traditional BI using the Mk model has a poor performance. Tarasov's

777 comparison between traditional BI and SMM/HMM models is limited to a 4-taxon case example,
778 which may not generalize well to larger trees. The proposed solution to the RBT problem from
779 Tarasov's SMM model (2019, Fig. 5 therein)—equivalent to our simplistic Simulations 1 herein
780 using a symmetric tree topology—is the result in which red and blue tailed clades evolve
781 "simultaneously" and receive similar posterior support in the majority rule consensus tree . This
782 is the same result obtained here by using standard Fitch parsimony with the default collapsing
783 rule in TNT (Fig. 1d), or when using the Mk model for BI under absence or unordered multistate
784 coding (Figs, S14-16, Table 2)—the best performing coding strategy detected here for such small
785 data sets. As demonstrated above, these results are expected for BI analyses due the way that
786 maximum likelihood optimization operates, and not something unique to the SMM or HMM
787 models.
788
789 *Limitations of BI and how to move forward.*
790 It should be noted that BI performing more accurately than alternative MP approaches
791 does not mean it is completely exempt of biases introduced by inapplicable character states in
792 contingent coding. The sampling of the posterior distribution via the MCMC algorithm is
793 strongly impacted by the number of primary characters with dependencies. In simulation models
794 with an increasingly larger number of primary characters with dependent secondary characters
795 (M4 and M5), there is only a small difference in performance of BI relative to MP-M and MP-
796 HSJ— although all the latter three still outperform traditional Fitch maximum parsimony (Fig.
797 5b).
798 Additionally, by quantifying the distribution of posterior trees from BI across the tree
799 parameter space (Figs. S20 and S21), we find that the mean RF distance between the posterior

800 trees within each simulation for models M1 and M2 is considerably lower than for models with a
801 larger proportion of secondary characters (M3) or with more primary characters bearing
802 secondary characters for each dataset (M4 and M5), irrespective of coding strategy. The total
803 variance (or disparity) of RF values is also considerably higher for models M3 to M5, except for
804 contingent coding, which is only higher for symmetric trees under models M3 to M5. Overall,
805 this indicates a substantial increase in the size of the tree space when there is a large amount of
806 secondary characters in the dataset (30% herein), and especially when there is an increase in the
807 number of primary characters bearing secondary traits within the same dataset. This increase in
808 the tree space (most notably in absent and multistate coding) makes it harder for the MCMC to
809 sample across all local optima and reach the global optimum, which is the most likely cause
810 further significant reduction in accuracy for models M4 and M5. The similarity of this result with
811 that observed for the results from MP analyses suggests the same phenomenon might be
812 impacting MP tree inference.

813 These results demonstrate the pervasive and detrimental role of increasing the number of
814 primary characters with dependent characters in phylogenetic datasets, even when there is a
815 decrease in the proportion of secondary characters for each primary character decreases (models
816 M3 to M5). The unfortunate practical consequence of our findings is that, considering there is a
817 finite number of anatomical structures from which morphological characters can be created in
818 the context of any given organismal study system, increasing the number of morphological
819 characters in a dataset will strongly rely on increasing the number of secondary characters that
820 are dependent on the presence of these anatomical structures (primary characters). Therefore,
821 increasing the number of characters in morphological datasets—a clearly recognizable pattern
822 for the past two decades and which is likely to continue into the future (Simões et al. 2017a,

2018a)—will almost invariably expand the tree parameter space in a way that both MP or BI algorithms will struggle to find answers closer to the global optimum, decreasing the accuracy of such inferences.

CONCLUSIONS

Dependency among morphological characters has been a long-recognized issue in phylogenetic inference (Maddison 1993), and which has been considered a problem without a clear solution for nearly three decades. New algorithmic solutions to this problem have been proposed in recent years, but without benchmarks studies assessing the accuracy of those solutions to character dependency.

Using different simulation scenarios, we demonstrate that alternative maximum parsimony algorithms designed to handle character dependency can generally produce more accurate results than traditional (Fitch) maximum parsimony, especially in cases with symmetric tree topologies and with low numbers of secondary characters. The MP-HSJ algorithm is generally more accurate than the competing approach MP-M, but traditional (non-clock) Bayesian inference is significantly more accurate than all MP approaches. This simple alternative to analyze datasets with dependent secondary characters has long been overlooked, and its superior performance derives from the fact that the likelihood optimization approach utilized by all probabilistic methods of phylogenetic inference does not include an “up-pass” phase, which is the cause of the issues introduced by secondary characters in MP. Importantly, increasing the number of secondary characters, and most importantly, increasing the number of primary characters with secondary characters that become inapplicable, substantially reduces phylogenetic accuracy regardless of optimality criterion or character coding strategy.

Most studies have historically found that increasing the number of morphological characters generally produces more accurate phylogenetic reconstructions—e.g., (Wiens 2004, Wright and Hillis 2014, O'Reilly et al. 2018, Puttick et al. 2019, Barido-Sottani et al. 2020). However, more recent simulations that assume the non-randomness of homoplastic distributions across the tree have found that an absolute increase in the number of characters does not produce more accurate phylogenetic trees due to convergent evolution (Keating et al. 2020). Our findings suggest that, if increasing the number of characters is majorly performed by increasing the number of secondary characters, then performance may in fact the reduced. We expect that the future development of more efficient algorithms to explore the larger tree parameter space created by secondary characters more thoroughly (especially for BI) might alleviate some of the existing limitations demonstrated here. Additionally, we urge caution when increasing the number of characters in morphological datasets, as the indiscriminate expansion of secondary characters that are dependent on primary characters that become absent (and therefore inapplicable) to a portion of the sampled taxa may introduce the detrimental effects phylogenetic accuracy detected herein.

SUPPLEMENTARY MATERIAL

Supplementary Material (text and figures) is available online. Supplementary files (all data and codes) are available at Dryad [NNNNNN].

FUNDING

This work was supported by the Natural Science and Engineering Research Council of Canada postdoctoral fellowship to T.R.S. Work on this manuscript was supported by NSF

DEB‐2113425 and NSF‐DEB‐2045842 and an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P2O GM103424‐20 to A.M.W. This work was supported by a Smithsonian Institution postdoctoral researcher fellowship to B.A.S.M.

AUTHOR CONTRIBUTIONS

Project conceptualization: TRS; experimental design: TRS and OV; analyses: all authors; discussions and interpretation of results: all authors; manuscript writing: TRS (with input by all authors).

REFERENCES

Ballesteros JA, Santibáñez-López CE, Baker CM, Benavides LR, Cunha TJ, Gainett G, Ontano AZ, Setton EVW, Arango CP, Gavish-Regev E*, et al.* 2022. Comprehensive Species Sampling and Sophisticated Algorithmic Approaches Refute the Monophyly of Arachnida. Mol. Biol. Evol., 39.

Barido-Sottani J, van Tiel NMA, Hopkins MJ, Wright DF, Stadler T, Warnock RCM. 2020. Ignoring Fossil Age Uncertainty Leads to Inaccurate Topology and Divergence Time Estimates in Time Calibrated Tree Inference. Frontiers in Ecology and Evolution, 8:1-13.

Brazeau MD. 2011. Problematic character coding methods in morphology and their effects. Biol. J. Linn. Soc., 104:489-498.

Brazeau MD, Guillerme T, Smith MR. 2019. An algorithm for Morphological Phylogenetic Analysis with Inapplicable Data. Syst. Biol., 68:619-631.

Brazeau MD, Smith MR, Guillerme T. 2017. MorphyLib: a library for phylogenetic analysis of categorical trait data with inapplicability (http://www.morphyproject.org/). Zenodo doi.

Farris JS, Kluge AG, Eckardt MJ. 1970. A Numerical Approach to Phylogenetic Systematics. Syst. Zool., 19:172-189.

Felsenstein J. 1973. Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. Syst. Zool., 22:240-249.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol., 17:368-376.

Felsenstein J. 2004. Inferring phylogenies. Sunderland, MA, Sinauer Associates Sunderland.

Fitch WM. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Biol., 20:406-416.

Forey PL, Kitching I. 2000. Experiments in coding multistate characters. In: Scotland RW, Pennington RT editors. Homology and systematics : coding characters for phylogenetic analysis. London & New York, Taylor & Francis, p. 54-80.

Frohlich MW, Chase MW. 2007. After a dozen years of progress the origin of angiosperms is still a great mystery. Nature, 450:1184-1189.

Garberoglio FF, Apesteguía S, Simões TR, Palci A, Gómez RO, Nydam RL, Larsson HCE, Lee MSY, Caldwell MW. 2019. New skulls and skeletons of the Cretaceous legged snake *Najash*, and the evolution of the modern snake body plan. Sci. Adv., 5:eaax5833.

Giles S, Xu G-H, Near TJ, Friedman M. 2017. Early members of 'living fossil'lineage imply later origin of modern ray-finned fishes. Nature, 549:265.

Goloboff PA, Catalano SA. 2016. TNT version 1.5, including a full implementation of phylogenetic morphometrics. Cladistics, 32:221-238.

Goloboff PA, De Laet J, Ríos-Tamayo D, Szumik CA. 2021. A reconsideration of inapplicable characters, and an approximation with step-matrix recoding. Cladistics.

Goloboff PA, Torres A, Arias JS. 2017. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. Cladistics, 34:407-437.

Goswami A, Polly PD. 2010. The influence of character correlations on phylogenetic analyses: a case study of the carnivoran cranium. In: Goswami A, Friscia A editors. Carnivoran Evolution: New Views on Phylogeny, Form and Function. Cambridge, Cambridge University Press, p. 141-164.

Goswami A, Smaers JB, Soligo C, Polly PD. 2014. The macroevolutionary consequences of phenotypic integration: from development to deep time. Philosophical Transactions of the Royal Society B: Biological Sciences, 369.

927 Graybeal A. 1998. Is It Better to Add Taxa or Characters to a Difficult Phylogenetic Problem?
928 Syst. Biol., 47:9-17.
929 Hawkins JA. 2000. A survey of primary homology assessment: different botanists perceive and
930 define characters in different ways. In: Scotland RW, Pennington RT editors. Homology and
931 systematics: coding characters for phylogenetic analysis. London and New York, The
932 Systematics Association, p. 22-53.
933 Hawkins JA, Hughes CE, Scotland RW. 1997. Primary Homology Assessment, Characters and
934 Character States. Cladistics, 13:275-283.
935 Heath TA, Hedtke SM, Hillis DM. 2008. Taxon sampling and the accuracy of phylogenetic
936 analyses. J Syst Evol, 46:239-257.
937 Hillis DM. 1996. Inferring complex phytogenies. Nature, 383:130.
938 Hillis DM. 1998. Taxonomic Sampling, Phylogenetic Accuracy, and Investigator Bias. Syst.
939 Biol., 47:3-8.
940 Hillis DM, Pollock DD, McGuire JA, Zwickl DJ. 2003. Is sparse taxon sampling a problem for
941 phylogenetic inference? Syst. Biol., 52:124.
942 Hopkins MJ, St John K. 2021. Incorporating Hierarchical Characters into Phylogenetic Analysis.
943 Syst. Biol., Advance articled.
944 Keating JN, Sansom RS, Sutton MD, Knight CG, Garwood RJ. 2020. Morphological
945 Phylogenetics Evaluated Using Novel Evolutionary Simulations. Syst. Biol., 69:897-912.
946 King B, Qiao T, Lee MSY, Zhu M, Long JA. 2017. Bayesian Morphological Clock Methods
947 Resurrect Placoderm Monophyly and Reveal Rapid Early Evolution in Jawed Vertebrates.
948 Syst. Biol., 66:499-516.

949 Klingenberg CP. 2008. Morphological Integration and Developmental Modularity. Annu. Rev.
950 Ecol. Evol. Syst., 39:115-132.
951 Lee MSY, Cau A, Naish D, Dyke GJ. 2014. Morphological Clocks in Paleontology, and a Mid-
952 Cretaceous Origin of Crown Aves. Syst. Biol., 63:442-449.
953 Maddison WP. 1993. Missing Data Versus Missing Characters in Phylogenetic Analysis. Syst.
954 Biol., 42:576-581.
955 Mongiardino Koch N, Garwood RJ, Parry LA. 2021. Fossils improve phylogenetic analyses of
956 morphological characters. Proc. R. Soc. Lond., Ser. B: Biol. Sci., 288:20210044.
957 Mongiardino Koch N, Thompson JR. 2020. A Total-Evidence Dated Phylogeny of Echinoidea
958 Combining Phylogenomic and Paleontological Data. Syst. Biol., 70:421-439.
959 Murphy JL, Puttick MN, O'Reilly JE, Pisani D, Donoghue PC. 2021. Empirical distributions of
960 homoplasy in morphological data. Palaeontology, Online First. DOI: 10.1111/pala.12535.
961 Nixon KC, Carpenter JM. 1993. On outgroups. Cladistics, 9:413-426.
962 Nixon KC, Carpenter JM. 2012. On homology. Cladistics, 28:160-169.
963 O'Reilly JE, Puttick MN, Parry L, Tanner AR, Tarver JE, Fleming J, Pisani D, Donoghue PCJ.
964 2016. Bayesian methods outperform parsimony but at the expense of precision in the
965 estimation of phylogeny from discrete morphological data. Biol. Lett., 12.
966 O'Reilly JE, Puttick MN, Pisani D, Donoghue PC. 2018. Probabilistic methods surpass
967 parsimony when assessing clade support in phylogenetic analyses of discrete morphological
968 data. Palaeontology, 61:105-118.
969 Paterson JR, Edgecombe GD, Lee MSY. 2019. Trilobite evolutionary rates constrain the duration
970 of the Cambrian explosion. Proc. Natl. Acad. Sci. USA, 116:4394-4399.

971 Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. 2002. Increased taxon sampling is
972 advantageous for phylogenetic inference. Syst. Biol., 51:664.
973 Puttick MN, O'Reilly JE, Pisani D, Donoghue PC. 2019. Probabilistic methods outperform
974 parsimony in the phylogenetic analysis of data simulated without a probabilistic model.
975 Palaeontology, 62:1-17.
976 Puttick MN, O'Reilly JE, Tanner AR, Fleming JF, Clark J, Holloway L, Lozano-Fernandez J,
977 Parry LA, Tarver JE, Pisani D*, et al.* 2017. Uncertain-tree: discriminating among competing
978 approaches to the phylogenetic analysis of phenotype data. Proc. R. Soc. Lond., Ser. B: Biol.
979 Sci., 284.
980 Pyron RA. 2011. Divergence time estimation using fossils as terminal taxa and the origins of
981 Lissamphibia. Syst. Biol.:syr047.
982 Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L,
983 Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference
984 and model choice across a large model space. Syst. Biol., 61:539-542.
985 Sand A, Holt MK, Johansen J, Brodal GS, Mailund T, Pedersen CNS. 2014. tqDist: a library for
986 computing the quartet and triplet distances between binary or general trees. Bioinformatics,
987 30:2079-2080.
988 Schliep K, Potts AJ, Morrison DA, Grimm GW. 2017. Intertwining phylogenetic trees and
989 networks. Methods Ecol. Evol., 8:1212-1220.
990 Scotland RW, Olmstead RG, Bennett JR. 2003. Phylogeny reconstruction: the role of
991 morphology. Syst. Biol., 52:539-548.
992 Sereno PC. 2007. Logical basis for morphological characters in phylogenetics. Cladistics,
993 23:565-587.

994 Simões TR, Caldwell MW, Palci A, Nydam RL. 2017a. Giant taxon-character matrices: quality
995 of character constructions remains critical regardless of size. Cladistics, 33:198-219.
996 Simões TR, Caldwell MW, Palci A, Nydam RL. 2018a. Giant taxon-character matrices II: a
997 response to Laing et al. (2017). Cladistics, 34:702-707.
998 Simões TR, Caldwell MW, Tałanda M, Bernardi M, Palci A, Vernygora O, Bernardini F,
999 Mancini L, Nydam RL. 2018b. The origin of squamates revealed by a Middle Triassic lizard
1000 from the Italian Alps. Nature, 557:706-709.
1001 Simões TR, Pierce SE. 2021. Sustained High Rates of Morphological Evolution During the Rise
1002 of Tetrapods. Nat. Ecol. Evol., 5:1403–1414.
1003 Simões TR, Vernygora O, Paparella I, Jimenez-Huidobro P, Caldwell MW. 2017b. Mosasauroid
1004 phylogeny under multiple phylogenetic methods provides new insights on the evolution of
1005 aquatic adaptations in the group. PloS one, 12:e0176773.
1006 Smith M. 2018. TreeSearch: phylogenetic tree search using custom optimality criteria. Compr. R
1007 Archive Network.
1008 Smith MR. 2019. Bayesian and parsimony approaches reconstruct informative trees from
1009 simulated morphological datasets. Biol. Lett., 15:20180632.
1010 Smith MR. 2020. Information theoretic generalized Robinson–Foulds metrics for comparing
1011 phylogenetic trees. Bioinformatics, 36:5007-5013.
1012 Strong EE, Lipscomb D. 1999. Character Coding and Inapplicable Data. Cladistics, 15:363-371.
1013 Tarasov S. 2019. Integration of Anatomy Ontologies and Evo-Devo Using Structured Markov
1014 Models Suggests a New Framework for Modeling Discrete Phenotypic Traits. Syst. Biol.,
1015 68:698-716.

1016 Vernygora OV, Simões TR, Campbell EO. 2020. Evaluating the Performance of Probabilistic
1017 Algorithms for Phylogenetic Analysis of Big Morphological Datasets: A Simulation Study.
1018 Syst. Biol., 69:1088-1105.
1019 Wiens JJ. 2004. The role of morphological data in phylogeny reconstruction. Syst. Biol., 53:653-
1020 661.
1021 Wiens JJ, Brandley MC, Reeder TW. 2006. Why does a trait evolve multiple times within a
1022 clade? Repeated evolution of snakeline body form in squamate reptiles. Evolution, 60:123-
1023 141.
1024 Wilkinson M. 1995. A Comparison of Two Methods of Character Construction. Cladistics,
1025 11:297-308.
1026 Wipfler B, Letsch H, Frandsen PB, Kapli P, Mayer C, Bartel D, Buckley TR, Donath A,
1027 Edgerly-Rooks JS, Fujita M*, et al.* 2019. Evolutionary history of Polyneoptera and its
1028 implications for our understanding of early winged insects. Proc. Natl. Acad. Sci. USA,
1029 116:3024-3029.
1030 Wright AM, Hillis DM. 2014. Bayesian Analysis Using a Simple Likelihood Model Outperforms
1031 Parsimony for Estimation of Phylogeny from Discrete Morphological Data. PLoS ONE,
1032 9:e109210.
1033 Wright AM, Lloyd GT. 2020. Bayesian analyses in phylogenetic palaeontology: interpreting the
1034 posterior sample. Palaeontology, 63:997-1006.
1035 Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error.
1036 Systematic Biology, 51:588-598.

1037

1038

FIGURES CAPTIONS

FIGURE 1. Problems stemming from contingent coding and introduced by inapplicable character states. a) Single tree from the analysis of 11 characters with homoplastic evolution of a primary character in distantly related clades that are separated by intervening taxa in which the primary character is inapplicable. b) Distinct coding schemes for new (tail) characters. c-e) Alternative resolutions for the ambiguous node in this case (Scenario 1, symmetric trees): the optimization of ancestral nodes on the right side of the tree will determine the ancestral state optimization on an unresolved clade (zone of contention) on the opposite side of the tree. Although there are three possible resolutions for the taxa in the zone of contention, most programs will only infer one of the S1 trees (depending on collapsing rules). One tree (Tree S2) will never be inferred by MP. f-i) Alternative resolutions for the ambiguous node in a distinct case (Scenario 2, asymmetric trees): when the primary character is inapplicable on the outgroup/earliest evolving taxa. In this case, all three solutions are inferred by MP programs, but the third solution (trees A3) can be presented in either one of two ways: supporting ambiguous nodes, as set by default in TNT and PAUP (tree A3a) or collapsing all nodes with zero branch lengths (‘rule 1’in TNT) (tree A3b).

FIGURE 2. Comparison of tree distance metrics. Lines show linear relationships between variables. Symmetric and asymmetric starting trees are the same used in the simulation of complex datasets. a) Decrease in similarity with number of random NNI moves from starting tree. b) Decrease in similarity with number of randomly collapsed nodes from starting tree.

FIGURE 3. Accuracy and resolution error for different coding and weighting schemes across distinct phylogenetic inference procedures. Results for absent (Abs), contingent (Cont), and multistate (Multi) coding schemes for MP using the traditional Fitch optimization—MP-F (a), for Bayesian inference—BI (b), and distinct weighting schemes for secondary characters as implemented by MP using HSJ optimization—MP-HSJ (c). For each quadrant, accuracy measured by MCI similarity (top left, in cyan) and quartets similarity (bottom left, in green), followed by resolution error measured by the proportion of incorrectly resolved nodes—Type I error (top right, in orange), and incorrectly unresolved nodes—Type II error (bottom right, in red).

FIGURE 4. Overall accuracy of each phylogenetic inference method using the best performing accuracy metric (quartets distance) regardless of simulated tree or character models. All methods are significantly different in performance based on pairwise Mann-Whitney tests (Supplementary Table 1). For method abbreviations, see Methods.

FIGURE 5. Accuracy of each phylogenetic inference method using the best performing accuracy metric (quartets distance) for distinct simulated tree and character models. Difference in performance between symmetric (Scenario 1) and asymmetric (Scenario 2) tree models (a), and between different character models (see Table 1) (b), for distinct phylogenetic inference methods. There is a steady increase in accuracy from MP-F (top row) to BI (bottom row) for both model classes (a and b). Most results are significantly different in performance based on pairwise Mann-Whitney tests (Supplementary Tables 2 and 3), with notable exceptions:

nonsignificant between tree models for MP-HSJ, and between character models M3-M4 for all inference methods. For method abbreviations, see Methods.