

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

3  
4           Tiago R. Simões<sup>1,\*</sup>, Oksana V. Vernygora<sup>2</sup>, Bruno A.S. de Medeiros<sup>3</sup>, and April M. Wright<sup>4</sup>

5  
6           <sup>1</sup>*Department of Organismic and Evolutionary Biology & Museum of Comparative Zoology, Harvard*  
7           *University, Cambridge, Massachusetts, USA;*

8           <sup>2</sup>*Department of Entomology, University of Kentucky, Lexington, Kentucky, USA;*

9           <sup>3</sup>*Smithsonian Tropical Research Institute, Panama City, Panama;*

10           <sup>4</sup>*Department of Biological Sciences, Southeastern Louisiana University, Hammond, Louisiana, USA.*

11  
12           \*Correspondence to be sent to: *Department of Organismic and Evolutionary Biology & Museum of*  
13           *Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA.; Telephone:*  
14           *+1 617 955-1081; E-mail: [tsimoes@fas.harvard.edu](mailto:tsimoes@fas.harvard.edu)*

15  
16           *Abstract.*— Character dependency is a major conceptual and methodological problem in  
17           phylogenetic inference of morphological datasets, as it violates the assumption of characters  
18           independency that is common to all phylogenetic methods. It is more frequently observed in  
19           higher-level phylogenies or in datasets characterizing major evolutionary transitions, as these  
20           represent parts of the tree of life where (primary) anatomical characters either originate or  
21           disappear entirely. As a result, secondary traits related to these primary characters become  
22           “inapplicable” across all sampled taxa in which that character is absent. Various solutions have  
23           been explored over the last three decades to handle character dependency, such as alternative  
24           character coding schemes and, more recently, new algorithmic implementations. However, the

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

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

42

43

44

## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

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

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

68 character) can only be determined for species in which the primary character is present (Fig. 1).  
69 Characters with such hierarchical structure are widespread in morphological datasets, especially  
70 those designed to assess higher-level phylogenetic relationships. The latter are more prone to  
71 include anatomical structures that originate (neomorphisms) or disappear across major branches  
72 of the tree of life, thus making all traits related to such structures secondary characters directly  
73 dependent on their presence. Examples of this include the origin of limb bones, which results in  
74 all limb related characters acting as secondary characters, during the fish-tetrapod transition  
75 (Simões and Pierce 2021); multiple independent limb losses within squamates (Wiens et al.  
76 2006); the origin of wings in insects (Wipfler et al. 2019); the origin of all floral structures at the  
77 origin of angiosperms (Frohlich and Chase 2007). Therefore, hierarchical characters can be  
78 especially prone to impact phylogenetic datasets aimed towards understanding evolutionary  
79 transitions, as those are typically characterized by the origin or loss of major anatomical  
80 structures (and their dependent secondary characters). Furthermore, even datasets aimed at  
81 smaller taxonomic scales may include hierarchical characters, such as datasets focusing on early-  
82 deriving snakes, in which various cranial, limb, and pectoral girdle characters may be either  
83 absent or present (Garberoglio et al. 2019), directly impacting all secondary characters  
84 contingent upon those traits.

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

## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

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

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

109         Besides multistate coding, numerous other solutions have been proposed over the past  
110 three decades to handle this simple but pervasive problem, from new character coding strategies  
111 (Maddison 1993, Hawkins et al. 1997, Strong and Lipscomb 1999, Hawkins 2000, Brazeau  
112 2011, Tarasov 2019) to new algorithmic solutions (Brazeau et al. 2019, Tarasov 2019, Hopkins  
113 and St John 2021). The vast array of character coding schemes, their benefits and limitations,

114 have been reviewed in many recent studies (Simões et al. 2017a, Brazeau et al. 2019, Hopkins  
115 and St John 2021), and so we refer the reader to these for further information (and also our  
116 Supplementary Material). In summary, despite the problems introduced by contingent coding,  
117 nearly all studies have agreed that contingent coding should be preferred over others as it is the  
118 least spurious solution to the problem of hierarchical characters (e.g., the RBT problem) (Strong  
119 and Lipscomb 1999, Sereno 2007, Brazeau 2011, Simões et al. 2017a).

120 As alternative coding schemes did not provide clear solutions to handle dependent  
121 characters, there was a recent shift in focus towards new algorithmic solutions rather than dataset  
122 construction ones. The first, the Morphy maximum parsimony algorithm introduced by (Brazeau  
123 et al. 2019), aims to escape the problem of inapplicable characters in contingent coding by  
124 providing a distinct treatment of inapplicable scores—referred to as the MP-M algorithm herein.  
125 Subsequently, (Hopkins and St John 2021) suggested down-weighting secondary characters  
126 relative to primary characters, also using maximum parsimony—referred as MP-HSJ herein.  
127 Subsequently, (Goloboff et al. 2021) advocated for the usage of Sankoff matrices to model  
128 character contingency in maximum parsimony.

129 The performance of these recent alternative algorithmic solutions, however, remains  
130 largely unknown. Simulated datasets, in which the “true” answer is known, have only been used  
131 once to test phylogenetic accuracy using a small synthetic dataset (with eight taxa) and restricted  
132 to maximum parsimony optimization approaches (Hopkins and St John 2021). Although  
133 analyzing small-sized simulated datasets can be useful as a proof of concept to better understand  
134 the behavior of alternative optimization methods, these do not explicitly test the model  
135 complexities that are inherent to much larger datasets that are closer in size to empirical ones.  
136 Important parameters that can be modeled in more complex simulated datasets include variable

137 levels of homoplasy, character evolutionary rates (contributing to branch lengths), tree  
 138 symmetry, the proportion of primary and secondary characters, among others. For instance,  
 139 previous studies have reported a significant performance disparity of various methods to  
 140 accurately infer datasets originated from symmetric and asymmetric trees (O'Reilly et al. 2018,  
 141 Puttick et al. 2019). Tree symmetry is also at the heart of the problem of hierarchical  
 142 characters—(Maddison 1993), see also Fig. 1 and Supplementary Material herein—and so we  
 143 should expect different performances from the proposed solutions to the problem of hierarchical  
 144 characters across different models.

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

158         Importantly, morphological datasets are now frequently analyzed by  
 159 probabilistic/statistical methods—maximum likelihood and Bayesian inference (BI)—across

160 various study systems—e.g.,(Lee et al. 2014, Giles et al. 2017, King et al. 2017, Simões et al.  
161 2017b, Paterson et al. 2019, Simões and Pierce 2021). Yet, the problem of hierarchical characters  
162 has rarely been discussed in the context of probabilistic inference methods. One major exception  
163 is a recent study suggesting the polymorphic re-coding of characters following the concept of  
164 structured and hidden states Markov models to incorporate the hierarchical structure of primary  
165 and secondary characters into Bayesian inference, as a solution to the problem of hierarchical  
166 characters (Tarasov 2019). However, no study to date has demonstrated if and how the problems  
167 introduced by hierarchical characters in MP impacts probabilistic phylogenetic algorithms to  
168 begin with, despite some previous suggestions that they would (Brazeau et al. 2019). At least in  
169 principle, theory suggests that likelihood-based methods should be less impacted by hierarchical  
170 characters. That is because all maximum likelihood and BI software implement variations of the  
171 Felsenstein likelihood optimization algorithm (Felsenstein 1973, 1981), which includes only a  
172 “down-pass” phase (from tips towards the root) for the calculation of likelihood scores at every  
173 node in the tree being reconstructed. The absence of an “up-pass” phase during the optimization  
174 of ancestral nodes—which is characteristic of maximum parsimony approaches (Brazeau 2011,  
175 Brazeau et al. 2019)—would suggest, for instance, that the dependency problem introduced by  
176 inapplicable state scores in contingent character coding should not impact tree inference using  
177 likelihood optimization procedures.

178         Here, we utilized a series of simulations of morphological datasets to address the  
179 following questions: how do different character coding schemes impact the relative performance  
180 of MP and BI in both simple and complex morphological datasets? Under a common coding  
181 scheme, how do classical and recently proposed optimization algorithms for MP perform relative  
182 to each other and to BI in morphological datasets? What is the impact of different tree and



## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

183 character models for the performance of each method? We find a striking contrast of results  
184 between simplistic and complex simulated datasets regarding best coding practices and a large  
185 disparity in performance among methods depending on tree or character distribution structures.  
186 As with other recent studies, our results are quite variable depending on the metric used for  
187 assessing accuracy and, using simulations, we demonstrate that quartet distance is less sensitive  
188 to tree resolution than bipartition-based metrics, being a better metric for differences in topology  
189 only. Finally, our results indicate that standard BI is significantly less impacted by contingent  
190 coding, displaying superior performance to all MP methods tested here, even those explicitly  
191 model to handle inapplicable characters.

192

### 193 MATERIALS AND METHODS

#### 194 *Simulation 1: Simplified Synthetic Datasets*

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

200 *Dataset 1 (Scenario 1, symmetric trees).*— Refers specifically to the RBT example of  
201 (Maddison 1993) with 14 taxa plus 1 outgroup with their internal relationships fully resolved and  
202 with each internal node supported by one synapomorphy, with the exception of the taxa within  
203 the so called zone of contention (Fig. 1a). A total of 11 characters are used to create this tree  
204 topology. The tree topology is symmetric and includes one fully resolved clade on the right side  
205 of the tree in which the primary character is present and the secondary character is applicable,

206 and one unresolved clade in which the primary character is convergently evolving on the left side  
207 of the tree. Subsequently, one or two extra characters are added to the dataset (depending on the  
208 coding scheme to be tested). For all coding schemes in which two characters are added,  
209 “character 12” is the primary character (denoting absence and presence of tail) and “character  
210 13” (denoting tail color) is the secondary character dependent on the primary character. Under  
211 multistate coding, a single “character 12” is present (Fig. 1b).

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

221

### 222 *Simulation 2: Complex Synthetic Dataset*

223 It is well-established that number of taxa (Hillis 1996, 1998, Pollock et al. 2002, Zwickl  
224 and Hillis 2002, Hillis et al. 2003, Heath et al. 2008, Vernygora et al. 2020), number of  
225 characters (Wright and Hillis 2014, O'Reilly et al. 2016, Puttick et al. 2017, Puttick et al.  
226 2019)—but see (Keating et al. 2020)—and the relative number of taxa per character  
227 (taxon:character ratio) (Graybeal 1998) all impact the performance of phylogenetic analyses  
228 using both morphological and molecular data under different optimality criteria. Therefore, we

## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

229 kept the number of taxa, number of characters, and the taxon: character ratio all constant to avoid  
230 introducing the impact of those extra variables on tree inference accuracy. Specifically, we used  
231 the following fixed values: 31 taxa (30 ingroup taxa +1 outgroup) and 60 characters—and thus a  
232 fixed taxon:character ratio 1:2 for the ingroup, which approximates well the taxon:character ratio  
233 in empirical datasets (Scotland et al. 2003, Murphy et al. 2021).

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

243 As discussed above, another key factor is how secondary characters are distributed  
244 among primary characters. For instance, in approaches that down-weight secondary characters  
245 (e.g., HSJ), if 30 secondary characters are dependent upon a single primary character their total  
246 weight will add up to a maximum of 1 step for the total tree score, and their individual relative  
247 weights will be of only  $1/30$  ( $= 0.03$ ) (for a HSJ  $\alpha = 1$ ). However, if these 30 secondary characters  
248 come from 5 independent primary characters (e.g., 6 from each primary character), then their  
249 total contribution to the tree score will add up to a maximum of 5, and each secondary  
250 character's relative weight will be five times higher than in the previous example— $1/6$  ( $= 0.167$ ).  
251 Therefore, secondary characters may have quite different weights depending on the relative

252 distribution of secondary characters among primary characters. To account for the latter, we  
 253 introduced another variable to our simulations: the number of secondary characters per primary  
 254 characters ( $S_d$ ), with the relationship  $S_d = S_n/P_s$ , where  $P_s$  is the number of primary characters  
 255 with dependent secondary characters. For instance, if we have 30 secondary characters  
 256 dependent on just one primary character—as in all examples from (Hopkins and St John 2021),  
 257 where all secondaries are dependent on a single primary character— that would be a case where:

258

$$259 \quad 60(C) = 30(P_n) + 30(S_n)$$

260 and,

$$261 \quad S_n = 30 \text{ and } P_s = 1, \text{ then } S_d = S_n/P_s = \mathbf{30} \text{ secondary characters per primary character}$$

262

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

264

$$265 \quad S_n = 30 \text{ and } P_s = 5, \text{ then } S_d = S_n/P_s = \mathbf{6} \text{ secondary characters per primary character}$$

266

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

270

271 *Simulated tree construction.*—We generated two simulated master (“true”) trees, one  
 272 fully symmetrical and another with perfectly asymmetrical topology, to test for the impact of  
 273 different tree symmetries on phylogenetic performance. Each tree included 31 taxa (30 ingroup  
 274 and 1 outgroup) as defined in the previous section. To emulate the RBT problem, we designated

## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

275 10 ‘crown’ taxa in each sector of the symmetrical tree (total = 20 taxa) and 10 ‘crown’ taxa in  
276 the asymmetrical tree—therefore fixing to 10 the number of taxa with applicable secondary  
277 characters forming the zone of contention (Fig. S1). All ‘stem’ taxa lying rootward of the  
278 ‘crown’ were designated to have the primary character absent, thus being inapplicable in respect  
279 to secondary characters. Because our simulation design focused on generating morphological  
280 data with no common evolutionary mechanism (Puttick et al. 2019), the branch length parameter  
281 was omitted from the generated master trees.

282

283 *Simulated dataset construction.*—We used each simulated tree to generate 100 replicates  
284 of binary morphological data matrices for each set of parameters (Models 1-5; listed in Table 1),  
285 following the conceptual approach of (Puttick et al. 2019). This procedure does not use explicit  
286 molecular substitution models to simulate morphological datasets, as in most previous  
287 simulations of morphological datasets—e.g., (Wright and Hillis 2014, O'Reilly et al. 2016,  
288 Puttick et al. 2017, O'Reilly et al. 2018, Vernygora et al. 2020). Instead, each individual  
289 character is first defined as either homologous or homoplastic based on a probability function of  
290 character homoplasy derived from an extensive survey of empirical datasets (Goloboff et al.  
291 2017, Puttick et al. 2019). This approach is designed to generate morphological characters with a  
292 model that does not necessarily favor probabilistic inference approaches—in fact, possibly  
293 favoring MP (Puttick et al. 2019)—for directly comparing the performance of MP and  
294 probabilistic methods in phylogenetics (Puttick et al. 2019). For homologous characters, terminal  
295 taxa are assigned states that result in the minimum number of character state transformations and  
296 therefore have a consistency index (CI)=1. If a character was defined as homoplastic, character  
297 states were assigned to the terminal taxa to produce  $CI < 1$ . For our simulated datasets, we set a

298 target CI index for the entire matrix to be within an intermediary range between 0.4 – 0.5 [bin 5  
299 in (Puttick et al. 2019)].

300 We generated datasets using a two-step procedure. First, we generated all primary  
301 characters that were applicable to all taxa. Primary characters that were designated to have  
302 secondary characters were assigned a specific pattern of character state scores: present [char.state  
303 = 1] in the ‘crown’ ten taxa and absent [char. state = 0] in the outgroup and ‘stem’ taxa (Fig. S1).  
304 Next, we performed a second round of simulations to generate scores for the secondary traits  
305 only. The latter included pruned versions of the master trees only with taxa that were scored as  
306 having the primary characters as present. These simulated secondary data matrices were then  
307 merged with the primary data matrices. Taxa that were scored as ‘absent’ for the primary traits  
308 were scored as ‘inapplicable’ for the secondary traits in the final merged datasets. All simulated  
309 datasets contained variable characters only, which is typical of morphological datasets.

310

### 311 *Analyses of Simulated Datasets*

312 MP-F tree searches for the simplified datasets generated by Simulation 1 for distinct  
313 coding strategies were conducted using the “Implicit Enumeration” algorithm in the software  
314 TNT v.1.5 (Goloboff and Catalano 2016). For Simulation 2, tree searches were conducted using  
315 the *phangorn* R package (Schliep et al. 2017). For tree searches with MP-M optimization we  
316 used its implementation in the R package *TreeSearch* v1.0.1 (Smith 2018), which uses  
317 MorphyLib (Brazeau et al. 2017) to handle inapplicable data (Brazeau et al. 2019). Tree searches  
318 with MP-HSJ optimization, we used the “*dissimilarity*” and “*hsjScorer*” R functions from  
319 (Hopkins and St John 2021) in conjunction with the branch-swapping algorithms available in the  
320 package *TreeSearch* v1.0.1 (Smith 2018). Starting rooted trees were subject SPR and TBR

## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

321 branch swapping operations, the results of which were used as starting trees for further analyzes  
322 with a series of ratchet iterations (functions “*Ratchet*” and “*pratchet*”), switching to the next run  
323 if the best score was hit 10 times, and stopping all searches if best score from each run was the  
324 same for 20 runs. The best scoring tree was used as the starting point for multiple ratchet  
325 (function “*MultiRatchet*”) runs with the same criteria as above to obtain multiple most  
326 parsimonious trees.

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

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

339         All most parsimonious trees (MPTs) obtained from each optimization procedure were  
340 used to calculate a strict consensus tree. Posterior tree samples obtained by BI were used to  
341 calculate a majority rule consensus tree. Both consensus options were chosen as they are the  
342 standard output trees for each of those respective optimization procedures in most studies using

343 morphological data. Consensus trees were subsequently used for comparison with the master  
344 trees generated by simulations.

345

#### 346 *Performance Measures*

347         We measured accuracy based on the total similarity shared by the inferred trees to the  
348 generated master trees using both bipartition and quartet tree distance metrics. For bipartition  
349 comparisons, we used similarity scores based on the Mutual Clustering Information metric  
350 (MCI) (Smith 2020), an information theory-based metric that shows the amount of mutual  
351 clustering information shared by all bipartitions in two or more trees. The latter is part of a larger  
352 class of generalized Robinson-Foulds (RF) distance metrics that overcome the limitations from  
353 classical implementations of the RF distance, such as quick saturation of distance scores (Smith  
354 2020). Quartet similarity is based on the “tqDist” algorithm from (Sand et al. 2014)—  
355 implemented in the R package *Quartet* (Smith 2019)—to measure the number of shared four-  
356 taxon subtrees between two or more trees.

357         Quartet similarity is predicted to outperform bipartition metrics as it better reflects  
358 phylogenetic patterns at deeper internal nodes, thus better handling poorly resolved nodes  
359 (Mongiardino Koch et al. 2021)—a problem for previous tree distance metrics, including  
360 traditional RF and Matching split distances [e.g., (Vernygora et al. 2020)]. Further, quartet  
361 similarity is less prone to the influence of wildcard taxa and tree shape (Smith 2020,  
362 Mongiardino Koch et al. 2021). Accordingly, we found several instances in which MCI and  
363 quartet similarity differed when applied to the same trees, and so we simulated how each metric  
364 is impacted by decreased tree resolution or increased topological differences to test the precise  
365 conditions in which these metrics yield different results. For both the asymmetric and the



366 symmetric 30-taxa master trees, we randomly collapsed from 1 to 28 internal nodes and  
 367 calculated MCI and Quartet similarity to the starting tree. Similarly, we randomly applied from 1  
 368 to 45 nearest-neighbor interchange (NNI) moves and compared the resulting tree to the starting  
 369 tree under both metrics. For each number of collapsed nodes or NNI moves, we did 50 replicates.  
 370 Finally, we compared both metrics in terms of their sensitivity to the number of collapsed nodes  
 371 (tree resolution) or number of NNI moves (topological differences), and whether tree symmetry  
 372 affected either metric.

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

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

385

### 386 *Statistical Analyses*

387 To assess if there were significant differences between performance results among  
 388 different tree and character models by inference method type, we conducted nonparametric

389 pairwise Wilcoxon rank sum (Mann-Whitney) between all analyses (Supplementary Tables 1-3).  
390 Parametric tests were not possible considering the bimodal distribution of some of the results  
391 (e.g., Figs 3-5).

392

## 393 RESULTS

### 394 *“Solutions” to the RBT Problem—a Conceptual Paradox*

395         There are only three pieces of phylogenetic information universally present within  
396 primary and secondary characters as illustrated by the RBT problem (Fig. 1): i) the primary  
397 character groups all taxa with tails together and those without tails as a second clade; ii) the  
398 secondary character (tail color) groups red-tailed taxa together and blue-tailed taxa together; iii)  
399 the biological dependency of the secondary character upon the primary character indicates that  
400 all aspects of the secondary character should be only applicable to taxa in which the primary  
401 character is present (defining the clade with tail). Beyond these three aspects, there is no data  
402 provided by either the primary or secondary characters to inform which tail color evolved first.  
403 In fact, the latter is irrelevant for tree inference under either MP or probabilistic methods, since  
404 reconstructing the direction of character state transformation (i.e., identifying synapomorphies) is  
405 only performed by MP upon the rooting of the tree once the most parsimonious solutions have  
406 already been found (Nixon and Carpenter 1993, 2012). For probabilistic methods (maximum  
407 likelihood and Bayesian inference) outgroup comparison and the direction of character-state  
408 transformation is not taken into consideration during tree sampling (Felsenstein 1973, 2004).  
409 Therefore, in the absence of additional characters, there is no single solution to the RBT problem  
410 as presented in Scenarios 1 and 2—*contra* (Tarasov 2019). Instead, any coding method or  
411 inference algorithm should allow the two possible solutions (i.e., red and blue first hypotheses)

## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

412 to be equally likely, and the latter should only be considered within the clade composed by taxa  
413 where the primary character is present. Therefore, the information content of any set of primary  
414 and secondary characters do require that the following criteria should be met for any coding  
415 approach or inference method to produce logically plausible and biologically realistic results:

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

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

428 To meet expectations from both corollaries above, two or more distinct tree topologies  
429 should be estimated for a coding or inference approach to be valid, depicting tree topologies with  
430 both valid solutions within the zone of contention (e.g., blue-first vs red-first hypotheses).  
431 Additionally, all resulting trees should have the primary character grouping all taxa with the  
432 present condition within the zone of contention. Therefore, the consensus (strict or majority rule)  
433 tree estimated from the output trees meeting these criteria will necessarily include all taxa in the  
434 zone of contingency as monophyletic (supported by the primary character), but with no particular

435 preference for either blue or red evolving first. Hence, the consensus tree should necessarily be  
436 unresolved, depicting a polytomic relationship for the taxa within the zone of contention.

437

438 *Simplified synthetic datasets*

439 *Fitch MP (MP-F)* .—Under MP-F, we find that four combinations of coding schemes/tree  
440 topologies meet the two corollaries for logically sound resolutions of the RBT problem (Table  
441 2). One is provided by contingent coding under Scenario 2 (asymmetric trees), but which fails  
442 under Scenario 1 (symmetric trees), as illustrated in Fig. 1 (f-i) and discussed in the  
443 Supplementary Material. A second coding scheme to meet both corollaries is represented by  
444 multistate coding under Scenario 1 (symmetric trees), which had been highlighted by (Maddison  
445 1993) as a solution to the contingent coding problem (Table 2). However, multistate coding fails  
446 under Scenario 2 as it cannot recover the hierarchical relationship between primary and  
447 secondary characters— as previously observed by (Hawkins et al. 1997). The latter results in  
448 some taxa (in which the primary character is absent) to be estimated as nested within the zone of  
449 contention, and a strict consensus tree with reduced resolution relative to other coding schemes  
450 (Figs. S2-S7). Finally, all options including character ordering logically prevent the basic  
451 assumption set by corollary 2, as the ordering scheme will inevitably predetermine which  
452 secondary state (red or blue) will evolve first (Figs. S6 and S7, Table 2).

453 The only coding approach to successfully meet the conditions set by corollaries 1 and 2  
454 above under both symmetric and asymmetric trees (Scenarios 1 and 2) is “absent coding” (Fig.  
455 S8 and S9, Table 2). Despite being briefly discussed in the literature before, absent coding was  
456 tested only once (Strong and Lipscomb 1999), and its ability to meet both corollaries was never  
457 previously realized (Supplementary Material).

458

459 *Morphy MP (MP-M)* .—This approach correctly recovers the hierarchical relationship between  
 460 primary and secondary characters as well as correctly finding the blue-first and red-first  
 461 hypotheses as equally parsimonious among the MPTs (Figs. S10 and S11, Table 2). This  
 462 matches the expectations of both corollaries, as predicted (Brazeau et al. 2019).

463

464 *HSJ MP (MP-HSJ)* .—As with MP-M, this approach was designed to correctly recover blue-first  
 465 and red-first hypotheses as equally parsimonious (Hopkins and St John 2021). As expected, it  
 466 does recover those hypotheses among the MPTs (Figs. S12 and S13, Table 2). The hierarchical  
 467 relationship between primary and secondary characters are recovered, but we note that those  
 468 must be provided by the user in the form of text file indicating a priori what characters are the  
 469 primary characters and secondary characters.

470

471 *Bayesian Inference-Mkv model (BI)* .—Using traditional (non-clock) BI and traditional modelling  
 472 of morphological characters (Mkv model), we found a substantial contrast of performance  
 473 between scenarios 1 and 2 concerning hierarchy (corollary 1). Regardless of the character coding  
 474 scheme, BI analyses of symmetric trees always inferred the clade defined by the presence of the  
 475 primary character (i.e., tail) as monophyletic in more than 90% of the sampled posterior trees  
 476 (Figs. S14-S16, Table 2), and the posterior trees sampled successfully converged towards an  
 477 optimal tree topology solution (Fig. S14-16, c,d). Additionally, frequency among posterior trees  
 478 for the correct inference of the clade defined by the presence of the primary character (i.e., tail)  
 479 was slightly higher for absent coding (98.7%), compared to contingent coding (97%) or  
 480 multistate (92.9%) coding.

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

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

496

#### 497 *Complex synthetic datasets*

498 *Performance of tree distance metrics.*—We found that both metrics are insensitive to the  
499 symmetry of the starting tree (Fig. 2). For both MCI and Quartet similarity, similarity decreases  
500 approximately linearly with the number of NNI moves (Fig. 2a). MCI show signs of saturation  
501 earlier than Quartet similarity, with a decreasing slope as NNI moves increase, while for Quartet  
502 similarity the relationship continues approximately linear even when the number of NNI moves  
503 is greater than the number of internal nodes in the tree (Fig. 2a). The two metrics differ more

504 strongly in their response to decreased tree resolution, however. While MCI decreases  
 505 approximately linearly with the number of collapsed nodes, quartet similarity is less sensitive to  
 506 decreased tree resolution when the number of polytomies is small and decreases sharply when  
 507 trees approach a complete polytomy (Fig. 2b).

508

509 *Performance across coding and alpha schemes.*—Only two methods could be tested for different  
 510 coding schemes (MP-F and BI), since the two other MP methods (MP-M and MP-HSJ) were  
 511 designed to handle datasets constructed using contingent coding schemes specifically.  
 512 Additionally, we tested the performance across different weighting schemes for secondary  
 513 characters (alpha variable) for the MP-HSJ optimization (Hopkins and St John 2021), which was  
 514 previously untested.

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

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

525 For the MP-HSJ optimization, quartet distances indicate no substantial difference in  
 526 performance across distinct alpha values, whereas MCI indicates a likely worse performance for

527 alpha values of 0 relative to 0.5 and 1, which is induced by higher proportions of Type II error  
528 (Fig. 3c).

529  
530 *Performance across methods.*—When comparing all methods based on contingent coding—the  
531 best performing coding procedure (Fig. 3a and b) and the only one common to all inference  
532 methods—MP-F has the worst accuracy compared to all other methods (Fig. 4). This result is  
533 consistent with predictions in the literature and is consistent regardless of accuracy metric (Fig.  
534 S20). However, the best solution among the three remaining methods depends on the  
535 performance metric. Similarity scores based on MCI (Smith 2020) suggests MP-HSJ perform the  
536 best whereas quartet distances indicate BI performs more accurately than other inference  
537 methods (Fig. S20). However, quartet distances were found to be more robust to variations in  
538 tree resolution when compared to bipartition metrics here (Fig. 2)—an important factor when  
539 comparing consensus trees, as done herein. Considering this, we favor the results provided by  
540 quartet distances, which suggest BI outperforms all inference methods based on MP, even those  
541 specifically designed to handle inapplicable characters.

542       When examining the tree-to-tree distances within each posterior sample (Fig. S21), we  
543 observed that simulation conditions in which secondary characters are spread more evenly  
544 among primary characters showed higher mean RF distances (i.e., models 3, 4, and 5). It should  
545 be noted that unlike in accuracy comparisons between methods, a higher RF score does not mean  
546 more differences from a “true” or simulation tree. This is a metric of within-posterior sample  
547 differences. In this case, a higher RF means that more different trees are being proposed and  
548 evaluated in these simulation conditions. We confirmed this by calculating a per-posterior



549 variance in the RF distance. This measure, too, indicated that greater dispersal of secondary  
 550 characters is associated with exploring more disparate phylogenetic trees (Fig. S22).

551  
 552 *Performance across tree and character models.*—The larger data dispersal and bimodality in the  
 553 results for each inference method (Fig. 4) suggest that other factors influence their respective  
 554 performance, two of which were explicitly modeled here: tree symmetry and distribution of  
 555 secondary characters among primary characters.

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

567         In contrast, the MCI metric suggests that accuracy in MP-F tree inference is similar for  
 568 symmetric and asymmetric trees (Figs. S23 and S24), thus going against all predictions above  
 569 and previous evidence from the literature indicating symmetric trees (as in Figs. 1a, c-e) are  
 570 considerably harder to estimate using MP-F compared to asymmetric trees (as in Fig. 1f-i) in the

571 presence of inapplicable scores for hierarchical characters. This further suggests this metric is not  
572 capable of detecting meaningful differences in performances across methods.

573         The performance of distinct inference methods when considering different primary and  
574 secondary character distribution models (Table 1) indicates a significant decrease in accuracy of  
575 MP-F when increasing the number of secondary characters per primary character (M1-M3), or  
576 when increasing the number of primary characters bearing secondary characters (M3-M5) (Figs.  
577 5b, S25 and S26, and Table S3). Such decrease in accuracy is also observed among other  
578 methods under the same circumstances, but to a much lower extent, except for model M5. In the  
579 latter, the increase in the number of primary characters bearing secondary characters dependent  
580 upon them substantially decreases performances across all methods (Fig. 5b).

581

## 582 DISCUSSION

### 583 *Differences between quartet and bipartition metrics to measure method accuracy*

584         Here we found that quartet and bipartition metrics favor different inference methods. Our  
585 simulations show that this is likely due to a difference in the sensitivity of each metric to tree  
586 resolution in summary trees and topological differences, but not to tree symmetry. MCI  
587 decreases approximately linearly with tree resolution and small topological differences (Fig. 2).  
588 As a result, when trees being compared include polytomies (e.g., most summary or consensus  
589 trees from MP and non-clock BI studies), the underlying cause of distances estimated may be  
590 ambiguous. Quartet similarity, on the other hand, appears to be less sensitive to polytomies  
591 except for extreme cases, better reflecting differences in topology. When applied only to fully  
592 resolved trees, MCI possesses several desirable properties in relation to other metrics, including  
593 Quartet Similarity (Smith, 2020). When trees vary both in topology and resolution, however,

594 interpretation from MCI can be problematic. By using of both metrics, we are able to find that BI  
 595 results in more accurate but less resolved trees, while MP-HSJ results in trees with higher  
 596 information content shared with true trees because they are better resolved, although less  
 597 accurate (i.e., include more false positives).

598

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

600 It has long been suggested that contingent coding is the less spurious solution to the  
 601 problem of dependent characters despite the introduction of inapplicable character states in  
 602 secondary characters (Strong and Lipscomb 1999, Sereno 2007, Brazeau 2011, Simões et al.  
 603 2017a). However, this assumption had never been tested using complex simulated morphological  
 604 datasets, and nearly all conclusions regarding distinct coding strategies come from small,  
 605 simulated datasets (Strong and Lipscomb 1999, Brazeau et al. 2019, Hopkins and St John 2021),  
 606 equivalent in size and scope to our Simulations 1 (simplified synthetic datasets). By examining  
 607 both symmetric and asymmetric tree structures for Simulations 1 and ancestral state  
 608 reconstructions for each of the three optimization procedures tested here (contingent, absent, and  
 609 multistate), we find new results and interpretations concerning the utilization of these coding  
 610 schemes. We find that the problems introduced by character dependency are most easily avoided  
 611 by using absent coding instead of contingent or multistate coding (Table 2, Figs. S8 and 9), thus  
 612 going against previous suggestions concerning this particular coding strategy using similarly  
 613 small synthetic datasets (Strong and Lipscomb 1999, Brazeau et al. 2019, Hopkins and St John  
 614 2021).

615 We attribute some of this difference to the fact that ancestral state reconstructions were  
 616 not conducted for all outputs of distinct coding strategies by (Strong and Lipscomb 1999),

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

627         By comparing the results of our Simulations 1 with more complex simulation scenarios  
628 (Simulations 2) we find important contrasts in our results and to previous conclusions using  
629 simplified datasets. When simulating larger datasets with explicit tree and character model  
630 variations, there is no significant difference in accuracy or resolution error among distinct coding  
631 strategies for traditional MP (MP-F), regardless of the performance metric (Fig. 3). We attribute  
632 this difference to the fact that the detected advantages of absent coding in simplified simulations  
633 (the only coding method meeting the assumptions of corollaries 1 and 2 discussed above) is  
634 counterbalanced by the negative bias introduced by the repeated occurrence of the absent state.  
635 As the number of secondary characters increases for larger datasets, it also increases the number  
636 of secondary characters with the absent condition, disproportionately overweighting the absent  
637 state. Although we did not explicitly test for a variable number of characters, we predict that  
638 datasets with a larger number of characters analyzed by traditional MP (MP-F) might see an even  
639 greater negative impact from the overweighting of the absent condition with absent coding,

640 potentially leading contingent coding to become the most accurate coding, as previously  
 641 suggested (Strong and Lipscomb 1999, Sereno 2007, Brazeau 2011, Simões et al. 2017a).

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

650

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

652 Perhaps the first attempt toward solving the problem of character dependency, outside the  
 653 scope of character coding schemes, was the utilization of step-matrices of costs—or Sankoff  
 654 matrices—as they could embed hierarchical relationships among characters (Forey and Kitching  
 655 2000). These have long been criticized for the amount of time required to build individual  
 656 matrices for every collection of primary character and their dependent secondary characters,  
 657 among other issues—e.g., (Brazeau et al. 2019). Recently, such problems were ameliorated by  
 658 faster methods to construct Sankoff matrices in the program TNT (Goloboff et al. 2021).  
 659 However, as the number of secondary characters increases in a dataset, this solution becomes  
 660 less practical as it surpasses the total possible number of states allowed by TNT (32 states). The  
 661 latter creates a maximum limit of four binary dependent characters (Goloboff et al. 2021). Even  
 662 if a larger number of character states are enabled the future implementations of TNT, the costs of

663 character state transformations would still have to be subjectively customized and without  
664 accounting for uncertainty around such transformational costs. Therefore, Sankoff matrices may  
665 never be a feasible universal solution to the problem of character dependency.

666         The Morphy (MP-M) approach (Brazeau et al. 2019) is, to our knowledge, the first  
667 algorithmic attempt to revise traditional parsimony optimization schemes for discrete characters  
668 (Fitch 1971), allowing for a more appropriate treatment of dependent characters. It was analyzed  
669 conceptually and empirically by subsequent studies, which criticized MP-M for not controlling  
670 for primary characters and their relationship to secondary characters (the same major limitation  
671 of the Fitch algorithm), leading to overweighting of absences for controlling primary characters  
672 (Hopkins and St John 2021). It was also detected that, by increasing the number of secondary  
673 characters, the MP-M approach (just as with MP-F) would result in a larger set of MPTs,  
674 including solutions where secondary characters were treated as applicable, thus contrary to its  
675 primary goal—a behavior not detected for the MP-HSJ method.

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

686 when increasing the number of secondary characters for a single controlling primary character  
 687 (models M1-M3 herein) as previously suspected (Hopkins and St John 2021), but it also  
 688 becomes less accurate when increasing the number of primary characters with dependent  
 689 characters (models M3-M5 herein).

690         Among all parsimony-based methods, MP-HSJ is consistently recovered as the best  
 691 performing method to handle the problem of inapplicable scores for dependent characters,  
 692 regardless of accuracy metric, tree structure, and character models simulated herein (Figs. 4, 5,  
 693 S21-S24). We attribute this performance to the fact that this is the only approach that specifically  
 694 identifies primary characters and each of their secondary character dependencies (Hopkins and St  
 695 John 2021). However, MP-HSJ downweights secondary characters to only a small fraction of the  
 696 relative weight attributed to primary characters, and this penalization increases proportionally to  
 697 the number of secondary characters in a dataset. The downweighting of secondary characters may  
 698 even be further boosted through the  $\alpha$  parameter introduced by this optimization procedure  
 699 (Hopkins and St John 2021). Our tests revealed that the downweighting of secondary characters  
 700 by MP-HSJ is so extensive that performance results under this approach are nearly identical  
 701 regardless of the chosen value of  $\alpha$  (even for  $\alpha = 0$ , which completely eliminates secondary  
 702 characters from the analysis) (Fig. 3c). Such heavy downweighting of secondary characters may  
 703 pose a limitation for datasets in which those characters are the only ones available to resolve  
 704 relationships within the zone of contention (e.g., Fig. 1). The latter might be one of the key  
 705 reasons for the superior performance of BI relative to MP-HSJ under the most accurate metric  
 706 (quartets), even though BI does not distinguish primary and secondary characters.

707

708 *The inapplicable states problem is mostly restricted to MP*

709           The primary cause for the problem of contingent coding and its impact on tree inference  
710 relates to the two-steps approach towards the optimization of ancestral state in MP—the “down-  
711 pass” and “up-pass” phases of the Fitch algorithm (Fitch 1971, Brazeau 2011). Since BI  
712 programs use the Felsenstein optimization (Felsenstein 1973, 1981) when calculating likelihoods  
713 for internal nodes, which has only a “down-pass” phase, it would be expected that the impact of  
714 inapplicable characters from contingent coding would be strongly reduced, or at least  
715 substantially minimized, relative to MP. Our results in Simulations 1 support our predictions in  
716 finding that contingent coding in MP-F will favor a blue-first hypothesis 100% of the time and  
717 never return any trees with a red-first hypotheses in Scenario 1 (Fig. 1, Table 2). On the other  
718 hand, BI will favor a similar hypothesis (blue-first = 46.1%) but it retrieves the competing  
719 hypotheses at frequencies much higher than 0% (i.e., red-first = 21%) (Table 2). As expected by  
720 their design, both MP-M and MP-HSJ accurately find most parsimonious trees with both blue  
721 and red-first hypotheses.

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

729           Using more complex simulations combining several parameters and larger numbers of  
730 taxa and characters (Simulation 2), BI again consistently recovers more accurate trees than MP  
731 using the traditional Fitch algorithm (MP-F). How BI compares in performance to other



## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

732 approaches designed to correct for the impact of inapplicable characters (MP-M and MP-HSJ)  
733 depends on the measure of accuracy. BI performs equally well under various scenarios to MP-M,  
734 but less accurately than MP-HSJ using the MCI metric. When analyzed under the quartet  
735 similarity metric, which is less influenced by tree resolution (Figs. 4 and 5), BI is significantly  
736 more accurate than the two parsimony approaches that correct for inapplicable characters.

737         Interestingly, solutions to character dependency have also been proposed in the context of  
738 Bayesian inference in recent years, such as for the utilization of structured (SMM) and hidden-  
739 state Markov models (HMM) (Tarasov 2019). While the latter study demonstrates that these  
740 newer methods can adequately deal with inapplicable states in dependent characters, no study  
741 has ever shown that traditional BI using the Mk model has a poor performance. Tarasov’s  
742 comparison between traditional BI and SMM/HMM models is limited to a 4-taxon case example,  
743 which may not generalize well to larger trees. The proposed solution to the RBT problem from  
744 Tarasov’s SMM model (2019, Fig. 5 therein)—equivalent to our simplistic Simulations 1 herein  
745 using a symmetric tree topology—is the result in which red and blue tailed clades evolve  
746 “simultaneously” and receive similar posterior support in the majority rule consensus tree . This  
747 is the same result obtained here by using standard Fitch parsimony with the default collapsing  
748 rule in TNT (Fig. 1d), or when using the Mk model for BI under absence or unordered multistate  
749 coding (Figs, S14-16, Table 2)—the best performing coding strategy detected here for such small  
750 data sets. As demonstrated above, these results are expected for BI analyses due the way that  
751 maximum likelihood optimization operates, and not something unique to the SMM or HMM  
752 models.

753

754 *Limitations of BI and how to move forward.*

755           It should be noted that BI performing more accurately than alternative MP approaches  
756 does not mean it is completely exempt of biases introduced by inapplicable character states in  
757 contingent coding. The sampling of the posterior distribution via the MCMC algorithm is  
758 strongly impacted by the number of primary characters with dependencies. In simulation models  
759 with an increasingly larger number of primary characters with dependent secondary characters  
760 (M4 and M5), there is only a small difference in performance of BI relative to MP-M and MP-  
761 HSJ— although all the latter three still outperform traditional Fitch maximum parsimony (Fig.  
762 5b).

763           Additionally, by quantifying the distribution of posterior trees from BI across the tree  
764 parameter space (Figs. S20 and S21), we find that the mean RF distance between the posterior  
765 trees within each simulation for models M1 and M2 is considerably lower than for models with a  
766 larger proportion of secondary characters (M3) or with more primary characters bearing  
767 secondary characters for each dataset (M4 and M5), irrespective of coding strategy. The total  
768 variance (or disparity) of RF values is also considerably higher for models M3 to M5, except for  
769 contingent coding, which is only higher for symmetric trees under models M3 to M5. Overall,  
770 this indicates a substantial increase in the size of the tree space when there is a large amount of  
771 secondary characters in the dataset (30% herein), and especially when there is an increase in the  
772 number of primary characters bearing secondary traits within the same dataset. This increase in  
773 the tree space (most notably in absent and multistate coding) makes it harder for the MCMC to  
774 sample across all local optima and reach the global optimum, which is the most likely cause  
775 further significant reduction in accuracy for models M4 and M5. The similarity of this result with  
776 that observed for the results from MP analyses suggests the same phenomenon might be  
777 impacting MP tree inference.

## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

778           These results demonstrate the pervasive and detrimental role of increasing the number of  
779 primary characters with dependent characters in phylogenetic datasets, even when there is a  
780 decrease in the proportion of secondary characters for each primary character decreases (models  
781 M3 to M5). The unfortunate practical consequence of our findings is that, considering there is a  
782 finite number of anatomical structures from which morphological characters can be created in  
783 the context of any given organismal study system, increasing the number of morphological  
784 characters in a dataset will strongly rely on increasing the number of secondary characters that  
785 are dependent on the presence of these anatomical structures (primary characters). Therefore,  
786 increasing the number of characters in morphological datasets—a clearly recognizable pattern  
787 for the past two decades and which is likely to continue into the future (Simões et al. 2017a,  
788 2018a)—will almost invariably expand the tree parameter space in a way that both MP or BI  
789 algorithms will struggle to find answers closer to the global optimum, decreasing the accuracy of  
790 such inferences.

791

## 792 CONCLUSIONS

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

798           Using different simulation scenarios, we demonstrate that alternative maximum  
799 parsimony algorithms designed to handle character dependency can generally produce more  
800 accurate results than traditional (Fitch) maximum parsimony, especially in cases with symmetric

801 tree topologies and with low numbers of secondary characters. The MP-HSJ algorithm is  
802 generally more accurate than the competing approach MP-M, but traditional (non-clock)  
803 Bayesian inference is significantly more accurate than all MP approaches. This simple  
804 alternative to analyze datasets with dependent secondary characters has long been overlooked,  
805 and its superior performance derives from the fact that the likelihood optimization approach  
806 utilized by all probabilistic methods of phylogenetic inference does not include an “up-pass”  
807 phase, which is the cause of the issues introduced by secondary characters in MP. Importantly,  
808 increasing the number of secondary characters, and most importantly, increasing the number of  
809 primary characters with secondary characters that become inapplicable, substantially reduces  
810 phylogenetic accuracy regardless of optimality criterion or character coding strategy.

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

## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

824 inapplicable) to a portion of the sampled taxa may introduce the detrimental effects phylogenetic  
825 accuracy detected herein.

826

### 827 SUPPLEMENTARY MATERIAL

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

830

### 831 FUNDING

832       This work was supported by the Natural Science and Engineering Research Council of  
833 Canada postdoctoral fellowship to T.R.S. Work on this manuscript was supported by NSF  
834 DEB - 2113425 and NSF - DEB - 2045842 and an Institutional Development Award (IDeA)  
835 from the National Institute of General Medical Sciences of the National Institutes of Health  
836 under grant number P20 GM103424 - 20 to A.M.W. This work was supported by a Smithsonian  
837 Institution postdoctoral researcher fellowship to B.A.S.M.

838

### 839 AUTHOR CONTRIBUTIONS

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

843

844

845

## 846 REFERENCES

- 847 Ballesteros JA, Santibáñez-López CE, Baker CM, Benavides LR, Cunha TJ, Gainett G, Ontano  
848 AZ, Setton EVW, Arango CP, Gavish-Regev E, *et al.* 2022. Comprehensive Species  
849 Sampling and Sophisticated Algorithmic Approaches Refute the Monophyly of Arachnida.  
850 Mol. Biol. Evol., 39.
- 851 Barido-Sottani J, van Tiel NMA, Hopkins MJ, Wright DF, Stadler T, Warnock RCM. 2020.  
852 Ignoring Fossil Age Uncertainty Leads to Inaccurate Topology and Divergence Time  
853 Estimates in Time Calibrated Tree Inference. *Frontiers in Ecology and Evolution*, 8:1-13.
- 854 Brazeau MD. 2011. Problematic character coding methods in morphology and their effects. *Biol.*  
855 *J. Linn. Soc.*, 104:489-498.
- 856 Brazeau MD, Guillerme T, Smith MR. 2019. An algorithm for Morphological Phylogenetic  
857 Analysis with Inapplicable Data. *Syst. Biol.*, 68:619-631.
- 858 Brazeau MD, Smith MR, Guillerme T. 2017. MorphyLib: a library for phylogenetic analysis of  
859 categorical trait data with inapplicability (<http://www.morphyproject.org/>). Zenodo doi.
- 860 Farris JS, Kluge AG, Eckardt MJ. 1970. A Numerical Approach to Phylogenetic Systematics.  
861 *Syst. Zool.*, 19:172-189.
- 862 Felsenstein J. 1973. Maximum Likelihood and Minimum-Steps Methods for Estimating  
863 Evolutionary Trees from Data on Discrete Characters. *Syst. Zool.*, 22:240-249.
- 864 Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach.  
865 *J. Mol. Evol.*, 17:368-376.
- 866 Felsenstein J. 2004. *Inferring phylogenies*. Sunderland, MA, Sinauer Associates Sunderland.
- 867 Fitch WM. 1971. Toward defining the course of evolution: minimum change for a specific tree  
868 topology. *Syst. Biol.*, 20:406-416.

## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

- 869 Forey PL, Kitching I. 2000. Experiments in coding multistate characters. In: Scotland RW,  
870 Pennington RT editors. Homology and systematics : coding characters for phylogenetic  
871 analysis. London & New York, Taylor & Francis, p. 54-80.
- 872 Frohlich MW, Chase MW. 2007. After a dozen years of progress the origin of angiosperms is  
873 still a great mystery. *Nature*, 450:1184-1189.
- 874 Garberoglio FF, Apesteguía S, Simões TR, Palci A, Gómez RO, Nydam RL, Larsson HCE, Lee  
875 MSY, Caldwell MW. 2019. New skulls and skeletons of the Cretaceous legged snake  
876 *Najash*, and the evolution of the modern snake body plan. *Sci. Adv.*, 5:eaax5833.
- 877 Giles S, Xu G-H, Near TJ, Friedman M. 2017. Early members of ‘living fossil’ lineage imply  
878 later origin of modern ray-finned fishes. *Nature*, 549:265.
- 879 Goloboff PA, Catalano SA. 2016. TNT version 1.5, including a full implementation of  
880 phylogenetic morphometrics. *Cladistics*, 32:221-238.
- 881 Goloboff PA, De Laet J, Ríos-Tamayo D, Szumik CA. 2021. A reconsideration of inapplicable  
882 characters, and an approximation with step-matrix recoding. *Cladistics*.
- 883 Goloboff PA, Torres A, Arias JS. 2017. Weighted parsimony outperforms other methods of  
884 phylogenetic inference under models appropriate for morphology. *Cladistics*, 34:407-437.
- 885 Goswami A, Polly PD. 2010. The influence of character correlations on phylogenetic analyses: a  
886 case study of the carnivoran cranium. In: Goswami A, Friscia A editors. *Carnivoran*  
887 *Evolution: New Views on Phylogeny, Form and Function*. Cambridge, Cambridge University  
888 Press, p. 141-164.
- 889 Goswami A, Smaers JB, Soligo C, Polly PD. 2014. The macroevolutionary consequences of  
890 phenotypic integration: from development to deep time. *Philosophical Transactions of the*  
891 *Royal Society B: Biological Sciences*, 369.

- 892 Graybeal A. 1998. Is It Better to Add Taxa or Characters to a Difficult Phylogenetic Problem?  
893 *Syst. Biol.*, 47:9-17.
- 894 Hawkins JA. 2000. A survey of primary homology assessment: different botanists perceive and  
895 define characters in different ways. In: Scotland RW, Pennington RT editors. *Homology and*  
896 *systematics: coding characters for phylogenetic analysis*. London and New York, The  
897 Systematics Association, p. 22-53.
- 898 Hawkins JA, Hughes CE, Scotland RW. 1997. Primary Homology Assessment, Characters and  
899 Character States. *Cladistics*, 13:275-283.
- 900 Heath TA, Hedtke SM, Hillis DM. 2008. Taxon sampling and the accuracy of phylogenetic  
901 analyses. *J Syst Evol*, 46:239-257.
- 902 Hillis DM. 1996. Inferring complex phylogenies. *Nature*, 383:130.
- 903 Hillis DM. 1998. Taxonomic Sampling, Phylogenetic Accuracy, and Investigator Bias. *Syst.*  
904 *Biol.*, 47:3-8.
- 905 Hillis DM, Pollock DD, McGuire JA, Zwickl DJ. 2003. Is sparse taxon sampling a problem for  
906 phylogenetic inference? *Syst. Biol.*, 52:124.
- 907 Hopkins MJ, St John K. 2021. Incorporating Hierarchical Characters into Phylogenetic Analysis.  
908 *Syst. Biol.*, Advance article.
- 909 Keating JN, Sansom RS, Sutton MD, Knight CG, Garwood RJ. 2020. Morphological  
910 Phylogenetics Evaluated Using Novel Evolutionary Simulations. *Syst. Biol.*, 69:897-912.
- 911 King B, Qiao T, Lee MSY, Zhu M, Long JA. 2017. Bayesian Morphological Clock Methods  
912 Resurrect Placoderm Monophyly and Reveal Rapid Early Evolution in Jawed Vertebrates.  
913 *Syst. Biol.*, 66:499-516.



## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

- 914 Klingenberg CP. 2008. Morphological Integration and Developmental Modularity. *Annu. Rev.*  
915 *Ecol. Evol. Syst.*, 39:115-132.
- 916 Lee MSY, Cau A, Naish D, Dyke GJ. 2014. Morphological Clocks in Paleontology, and a Mid-  
917 Cretaceous Origin of Crown Aves. *Syst. Biol.*, 63:442-449.
- 918 Maddison WP. 1993. Missing Data Versus Missing Characters in Phylogenetic Analysis. *Syst.*  
919 *Biol.*, 42:576-581.
- 920 Mongiardino Koch N, Garwood RJ, Parry LA. 2021. Fossils improve phylogenetic analyses of  
921 morphological characters. *Proc. R. Soc. Lond., Ser. B: Biol. Sci.*, 288:20210044.
- 922 Mongiardino Koch N, Thompson JR. 2020. A Total-Evidence Dated Phylogeny of Echinoidea  
923 Combining Phylogenomic and Paleontological Data. *Syst. Biol.*, 70:421-439.
- 924 Murphy JL, Puttick MN, O'Reilly JE, Pisani D, Donoghue PC. 2021. Empirical distributions of  
925 homoplasy in morphological data. *Palaeontology*, Online First. DOI: 10.1111/pala.12535.
- 926 Nixon KC, Carpenter JM. 1993. On outgroups. *Cladistics*, 9:413-426.
- 927 Nixon KC, Carpenter JM. 2012. On homology. *Cladistics*, 28:160-169.
- 928 O'Reilly JE, Puttick MN, Parry L, Tanner AR, Tarver JE, Fleming J, Pisani D, Donoghue PCJ.  
929 2016. Bayesian methods outperform parsimony but at the expense of precision in the  
930 estimation of phylogeny from discrete morphological data. *Biol. Lett.*, 12.
- 931 O'Reilly JE, Puttick MN, Pisani D, Donoghue PC. 2018. Probabilistic methods surpass  
932 parsimony when assessing clade support in phylogenetic analyses of discrete morphological  
933 data. *Palaeontology*, 61:105-118.
- 934 Paterson JR, Edgecombe GD, Lee MSY. 2019. Trilobite evolutionary rates constrain the duration  
935 of the Cambrian explosion. *Proc. Natl. Acad. Sci. USA*, 116:4394-4399.

- 936 Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. 2002. Increased taxon sampling is  
937 advantageous for phylogenetic inference. *Syst. Biol.*, 51:664.
- 938 Puttick MN, O'Reilly JE, Pisani D, Donoghue PC. 2019. Probabilistic methods outperform  
939 parsimony in the phylogenetic analysis of data simulated without a probabilistic model.  
940 *Palaeontology*, 62:1-17.
- 941 Puttick MN, O'Reilly JE, Tanner AR, Fleming JF, Clark J, Holloway L, Lozano-Fernandez J,  
942 Parry LA, Tarver JE, Pisani D, *et al.* 2017. Uncertain-tree: discriminating among competing  
943 approaches to the phylogenetic analysis of phenotype data. *Proc. R. Soc. Lond., Ser. B: Biol.*  
944 *Sci.*, 284.
- 945 Pyron RA. 2011. Divergence time estimation using fossils as terminal taxa and the origins of  
946 Lissamphibia. *Syst. Biol.*:syr047.
- 947 Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L,  
948 Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference  
949 and model choice across a large model space. *Syst. Biol.*, 61:539-542.
- 950 Sand A, Holt MK, Johansen J, Brodal GS, Mailund T, Pedersen CNS. 2014. tqDist: a library for  
951 computing the quartet and triplet distances between binary or general trees. *Bioinformatics*,  
952 30:2079-2080.
- 953 Schliep K, Potts AJ, Morrison DA, Grimm GW. 2017. Intertwining phylogenetic trees and  
954 networks. *Methods Ecol. Evol.*, 8:1212-1220.
- 955 Scotland RW, Olmstead RG, Bennett JR. 2003. Phylogeny reconstruction: the role of  
956 morphology. *Syst. Biol.*, 52:539-548.
- 957 Sereno PC. 2007. Logical basis for morphological characters in phylogenetics. *Cladistics*,  
958 23:565-587.

## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

- 959 Simões TR, Caldwell MW, Palci A, Nydam RL. 2017a. Giant taxon-character matrices: quality  
960 of character constructions remains critical regardless of size. *Cladistics*, 33:198-219.
- 961 Simões TR, Caldwell MW, Palci A, Nydam RL. 2018a. Giant taxon-character matrices II: a  
962 response to Laing et al. (2017). *Cladistics*, 34:702-707.
- 963 Simões TR, Caldwell MW, Tałanda M, Bernardi M, Palci A, Vernygora O, Bernardini F,  
964 Mancini L, Nydam RL. 2018b. The origin of squamates revealed by a Middle Triassic lizard  
965 from the Italian Alps. *Nature*, 557:706-709.
- 966 Simões TR, Pierce SE. 2021. Sustained High Rates of Morphological Evolution During the Rise  
967 of Tetrapods. *Nat. Ecol. Evol.*, 5:1403–1414.
- 968 Simões TR, Vernygora O, Paparella I, Jimenez-Huidobro P, Caldwell MW. 2017b. Mosasauroid  
969 phylogeny under multiple phylogenetic methods provides new insights on the evolution of  
970 aquatic adaptations in the group. *PloS one*, 12:e0176773.
- 971 Smith M. 2018. TreeSearch: phylogenetic tree search using custom optimality criteria. *Compr. R*  
972 *Archive Network*.
- 973 Smith MR. 2019. Bayesian and parsimony approaches reconstruct informative trees from  
974 simulated morphological datasets. *Biol. Lett.*, 15:20180632.
- 975 Smith MR. 2020. Information theoretic generalized Robinson–Foulds metrics for comparing  
976 phylogenetic trees. *Bioinformatics*, 36:5007-5013.
- 977 Strong EE, Lipscomb D. 1999. Character Coding and Inapplicable Data. *Cladistics*, 15:363-371.
- 978 Tarasov S. 2019. Integration of Anatomy Ontologies and Evo-Devo Using Structured Markov  
979 Models Suggests a New Framework for Modeling Discrete Phenotypic Traits. *Syst. Biol.*,  
980 68:698-716.

- 981 Vernygora OV, Simões TR, Campbell EO. 2020. Evaluating the Performance of Probabilistic  
982 Algorithms for Phylogenetic Analysis of Big Morphological Datasets: A Simulation Study.  
983 *Syst. Biol.*, 69:1088-1105.
- 984 Wiens JJ. 2004. The role of morphological data in phylogeny reconstruction. *Syst. Biol.*, 53:653-  
985 661.
- 986 Wiens JJ, Brandley MC, Reeder TW. 2006. Why does a trait evolve multiple times within a  
987 clade? Repeated evolution of snakeline body form in squamate reptiles. *Evolution*, 60:123-  
988 141.
- 989 Wilkinson M. 1995. A Comparison of Two Methods of Character Construction. *Cladistics*,  
990 11:297-308.
- 991 Wipfler B, Letsch H, Frandsen PB, Kapli P, Mayer C, Bartel D, Buckley TR, Donath A,  
992 Edgerly-Rooks JS, Fujita M, *et al.* 2019. Evolutionary history of Polyneoptera and its  
993 implications for our understanding of early winged insects. *Proc. Natl. Acad. Sci. USA*,  
994 116:3024-3029.
- 995 Wright AM, Hillis DM. 2014. Bayesian Analysis Using a Simple Likelihood Model Outperforms  
996 Parsimony for Estimation of Phylogeny from Discrete Morphological Data. *PLoS ONE*,  
997 9:e109210.
- 998 Wright AM, Lloyd GT. 2020. Bayesian analyses in phylogenetic palaeontology: interpreting the  
999 posterior sample. *Palaeontology*, 63:997-1006.
- 1000 Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error.  
1001 *Systematic Biology*, 51:588-598.

1002

1003

## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

1004 FIGURES CAPTIONS

1005

1006 FIGURE 1. Problems stemming from contingent coding and introduced by inapplicable character  
1007 states. a) Single tree from the analysis of 11 characters with homoplastic evolution of a primary  
1008 character in distantly related clades that are separated by intervening taxa in which the primary  
1009 character is inapplicable. b) Distinct coding schemes for new (tail) characters. c-e) Alternative  
1010 resolutions for the ambiguous node in this case (Scenario 1, symmetric trees): the optimization of  
1011 ancestral nodes on the right side of the tree will determine the ancestral state optimization on an  
1012 unresolved clade (zone of contention) on the opposite side of the tree. Although there are three  
1013 possible resolutions for the taxa in the zone of contention, most programs will only infer one of  
1014 the S1 trees (depending on collapsing rules). One tree (Tree S2) will never be inferred by MP. f-  
1015 i) Alternative resolutions for the ambiguous node in a distinct case (Scenario 2, asymmetric  
1016 trees): when the primary character is inapplicable on the outgroup/earliest evolving taxa. In this  
1017 case, all three solutions are inferred by MP programs, but the third solution (trees A3) can be  
1018 presented in either one of two ways: supporting ambiguous nodes, as set by default in TNT and  
1019 PAUP (tree A3a) or collapsing all nodes with zero branch lengths ('rule 1' in TNT) (tree A3b).

1020

1021 FIGURE 2. Comparison of tree distance metrics. Lines show linear relationships between  
1022 variables. Symmetric and asymmetric starting trees are the same used in the simulation of  
1023 complex datasets. a) Decrease in similarity with number of random NNI moves from starting  
1024 tree. b) Decrease in similarity with number of randomly collapsed nodes from starting tree.

1025

1026

1027 FIGURE 3. Accuracy and resolution error for different coding and weighting schemes across  
1028 distinct phylogenetic inference procedures. Results for absent (Abs), contingent (Cont), and  
1029 multistate (Multi) coding schemes for MP using the traditional Fitch optimization—MP-F (a),  
1030 for Bayesian inference—BI (b), and distinct weighting schemes for secondary characters as  
1031 implemented by MP using HSJ optimization—MP-HSJ (c). For each quadrant, accuracy  
1032 measured by MCI similarity (top left, in cyan) and quartets similarity (bottom left, in green),  
1033 followed by resolution error measured by the proportion of incorrectly resolved nodes—Type I  
1034 error (top right, in orange), and incorrectly unresolved nodes—Type II error (bottom right, in  
1035 red).

1036

1037 FIGURE 4. Overall accuracy of each phylogenetic inference method using the best performing  
1038 accuracy metric (quartets distance) regardless of simulated tree or character models. All methods  
1039 are significantly different in performance based on pairwise Mann-Whitney tests (Supplementary  
1040 Table 1). For method abbreviations, see Methods.

1041

1042 FIGURE 5. Accuracy of each phylogenetic inference method using the best performing accuracy  
1043 metric (quartets distance) for distinct simulated tree and character models. Difference in  
1044 performance between symmetric (Scenario 1) and asymmetric (Scenario 2) tree models (a), and  
1045 between different character models (see Table 1) (b), for distinct phylogenetic inference  
1046 methods. There is a steady increase in accuracy from MP-F (top row) to BI (bottom row) for  
1047 both model classes (a and b). Most results are significantly different in performance based on  
1048 pairwise Mann-Whitney tests (Supplementary Tables 2 and 3), with notable exceptions:

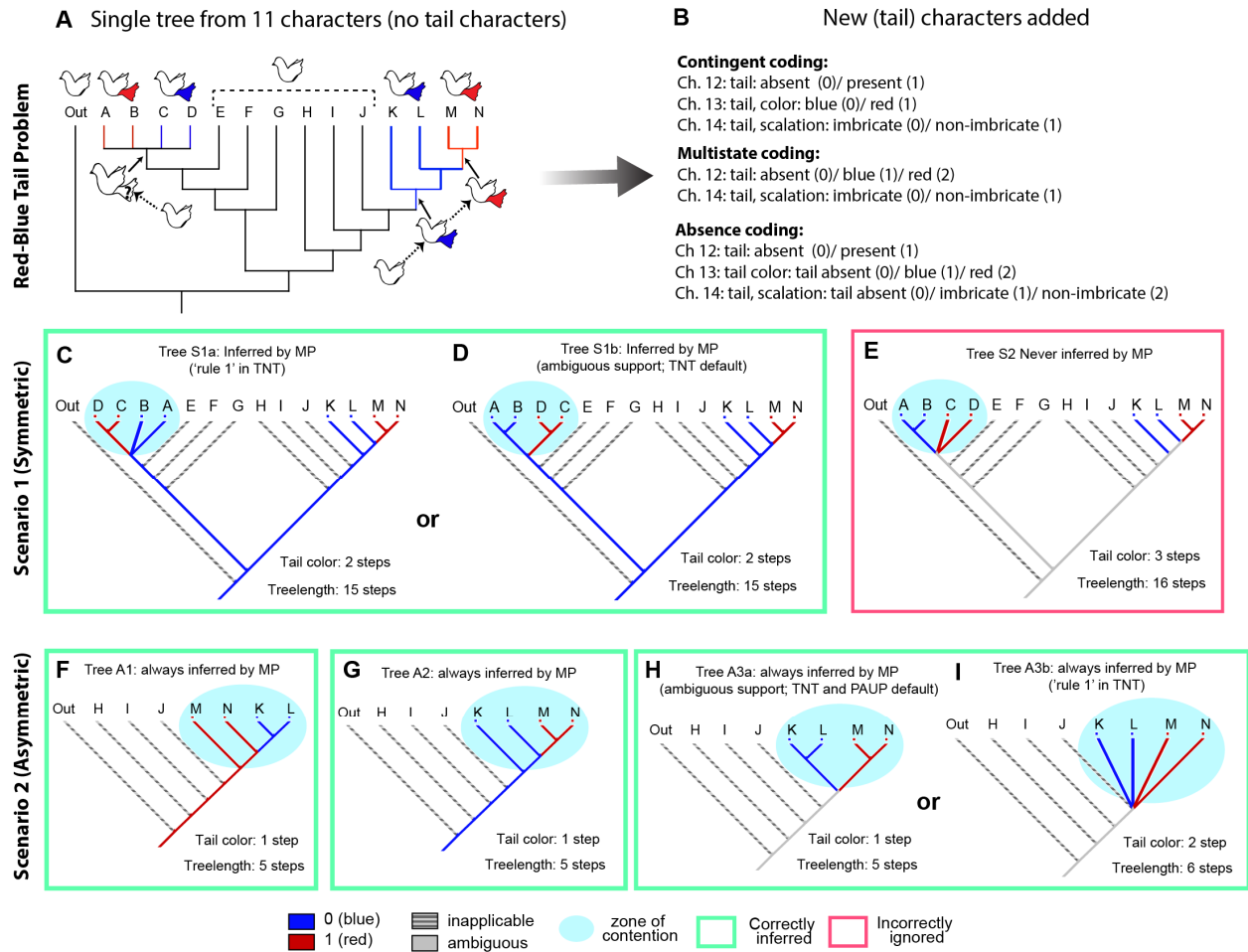
## METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

1049 nonsignificant between tree models for MP-HSJ, and between character models M3-M4 for all

1050 inference methods. For method abbreviations, see Methods.

1051

1052



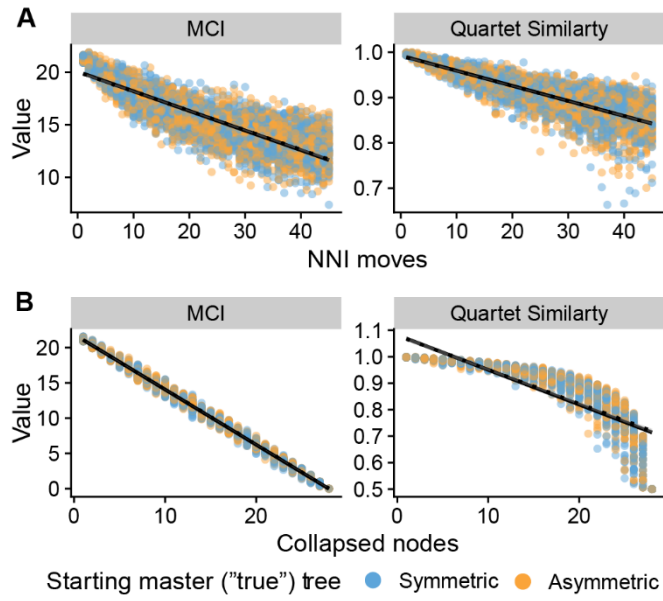
1053

1054

1055

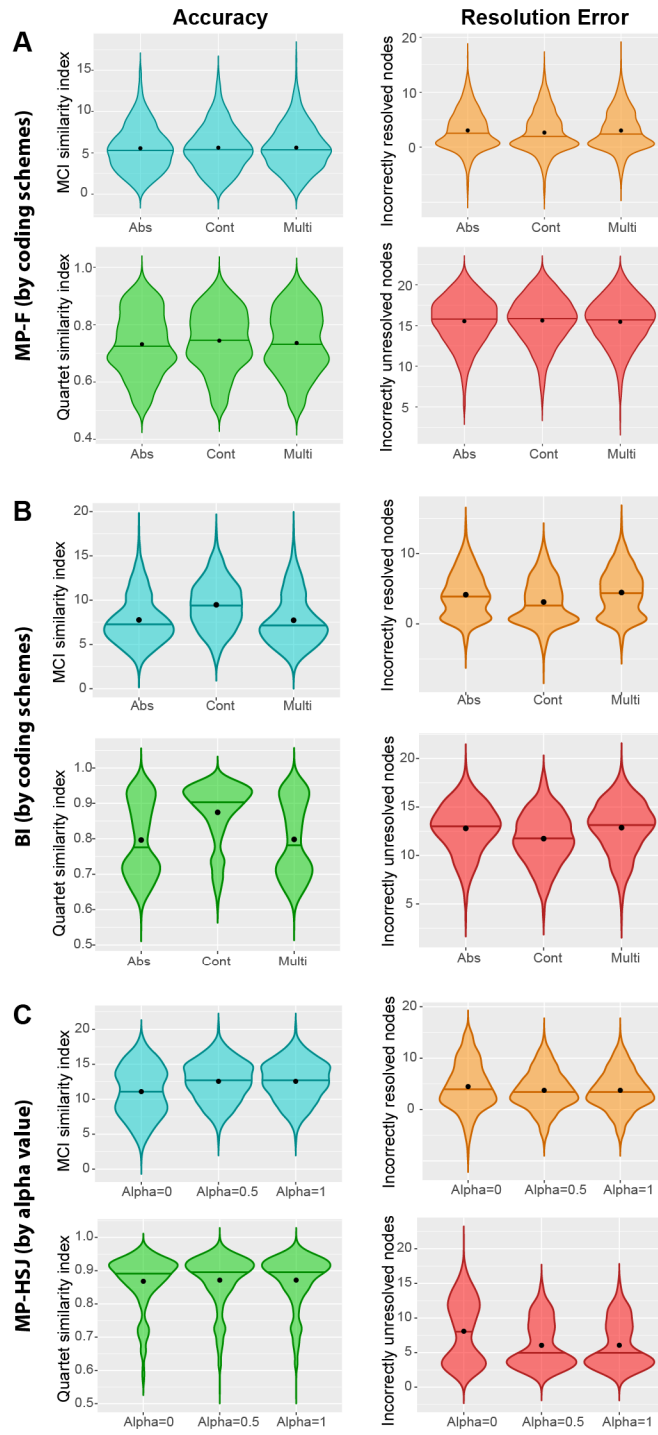


METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

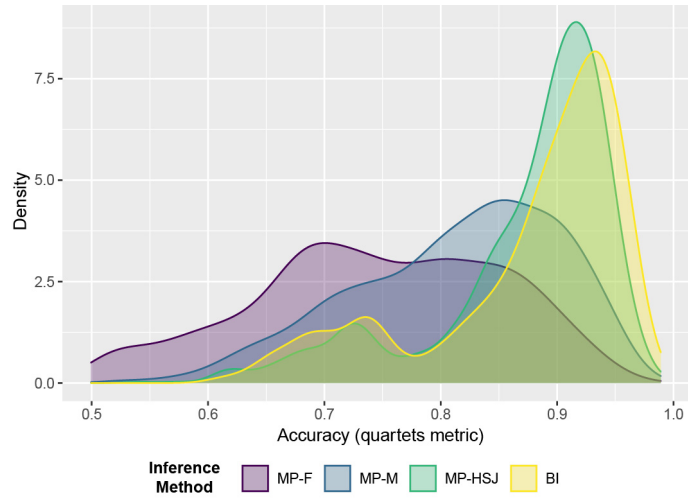


1056

1057



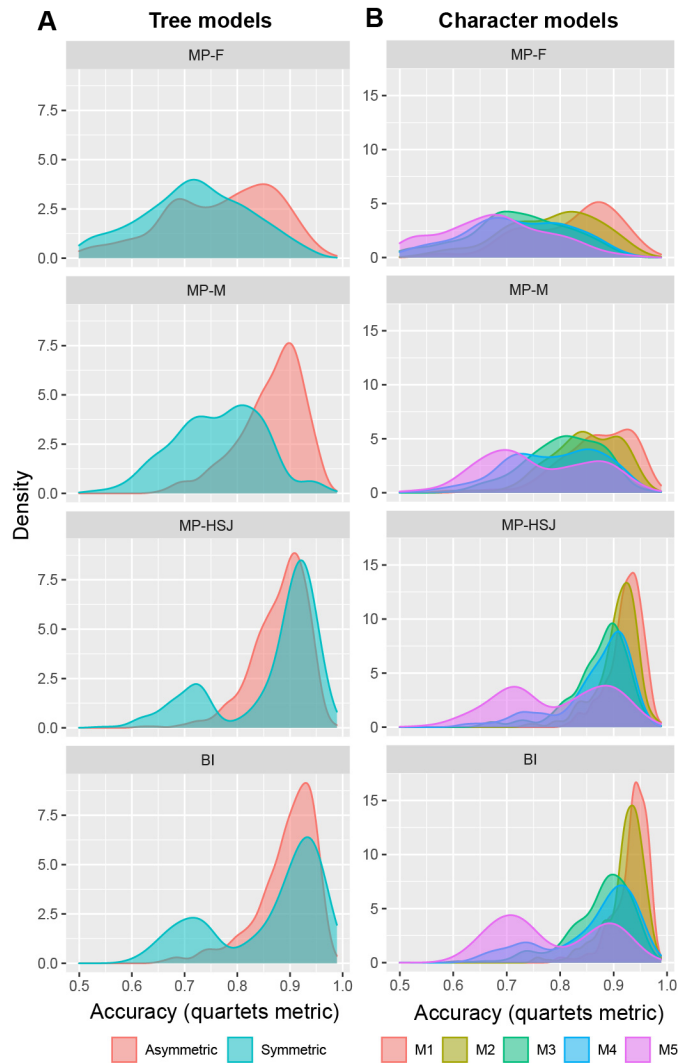
METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY



1059

1060

1061



1062

1063 TABLES

1064

1065 TABLE 1. Combinations of characters distribution models.

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

1066

1067 \*Note that the number of secondary characters per primary character ( $S_d$ ) on models M4 and M5  
1068 are the same as in models M2 and M1, respectively. However, the secondary characters in M4  
1069 and M5 are distributed across more primary characters ( $P_s$ ), which will impact the final Fitch  
1070 scores and tree lengths.  
1071

METHODOLOGICAL PERFORMANCE GIVEN CHARACTER DEPENDENCY

1072 TABLE 2. Results for the simplified synthetic datasets using various coding schemes. Coding  
 1073 schemes meeting expectations from corollaries 1 and 2 are highlighted with blue background.  
 1074 Coding schemes with results pre-established by users (ordered characters) highlighted in gray.  
 1075 Results for coding schemes that are not applicable to particular methods are marked with “NA”.  
 1076 Abbreviations: Abs, absence coding; B, blue tail-first hypothesis; Cont, contingent coding; Cor,  
 1077 corollaries; M, method; Multi, multistate coding; P-S, primary and secondary character  
 1078 hierarchy; ord, ordered; R, red tail-first hypothesis; unord, unordered.

M	Cor	Scenario 1 (Symmetric/two zones)					Scenario 2 (Asymmetric/one zone)				
		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%)

1079  
 1080 \* Yes if >90% of posterior trees infer the focal clade (defined by primary character being  
 1081 present) as monophyletic.

1082 \*\*Yes if difference in frequency between blue (B) and red (R)-first hypotheses <1%.

1083

1084