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

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

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

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

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

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

[*]*Correspondence to be sent to: Department of Organismic and Evolutionary Biology & Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA.;*

Telephone: +1 617 955-1081; E-mail: tsimoes@fas.harvard.edu

Abstract.— Logical character dependency is a major conceptual and methodological problem in phylogenetic inference of morphological datasets, as it violates the assumption of character independence that is common to all phylogenetic methods. It is more frequently observed in higher-level phylogenies or in datasets characterizing major evolutionary transitions, as these represent parts of the tree of life where (primary) anatomical characters either originate or disappear entirely. As a result, secondary traits related to these primary characters become "inapplicable" across all sampled taxa in which that character is absent. Various solutions have been explored over the last three decades to handle character dependency, such as alternative character coding schemes and, more recently, new algorithmic implementations. However, the accuracy of the proposed solutions, or the impact of character dependency across distinct optimality criteria, has never been directly tested using standard performance measures. Here, we utilize simple and complex simulated morphological datasets analyzed under different maximum parsimony optimization procedures and Bayesian inference to test the accuracy of various coding and algorithmic solutions to character dependency. This is complemented by empirical analyses using a recoded dataset on palaeognathid birds. We find that in small, simulated datasets, absent coding performs better than other popular coding strategies available (contingent and multistate), whereas in more complex simulations (larger datasets controlled for different tree structure and character distribution models) contingent coding is favored more frequently. Under contingent coding, a recently proposed weighting algorithm produces the most accurate results for maximum parsimony. However, Bayesian inference outperforms all parsimony-based solutions to handle character dependency due to

fundamental differences in their optimization procedures—a simple alternative that has been long overlooked. Yet, we show that the more primary characters bearing secondary (dependent) traits there are in a dataset, the harder it is to estimate the true phylogenetic tree, regardless of the optimality criterion, owing to a considerable expansion of the tree parameter space.

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

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

Logical dependency in morphological phylogenetics is usually in the form of hierarchical characters—i.e., a set of two or more characters, including one primary character (governing the absence or presence of an anatomical structure) and one or more secondary

characters (governing various properties of that same structure). For simplicity, hereafter we will use the acronym WDSC (with dependent secondary characters) to refer to such primary characters. A classic example of this logical dependency was introduced by Maddison (1993) and is known as the Red-Blue Tail (RBT) problem. In this problem, tails can be absent/present (primary character), but tail color (secondary character) can only be determined for species in which the primary character is present, creating a *zone of contention* (Fig. 1). Characters with such hierarchical structure are widespread in morphological datasets, especially those designed to assess higher-level phylogenetic relationships and/or major evolutionary transitions.

Some examples of such evolutionary transitions prone to include a large proportion of hierarchical characters include: the origin of limbs, which results in all limb related characters acting as secondary (dependent) characters for limbs during the fish-tetrapod transition (Simões and Pierce 2021); multiple independent limb losses within squamates (Wiens et al. 2006, Gauthier et al. 2012); the origin of wings in insects, making all wing structures dependent on the presence of wings (Wipfler et al. 2019); or the origin of all floral structures at the origin of angiosperms (Frohlich and Chase 2007). These secondary characters sometimes represent a substantial proportion of the characters related to specific anatomical units, especially in big morphological data sets—e.g., 15 secondary characters for the jugal and 15 for the squamosal bones (all dependent on the characters for the absence/presence of these respective elements) in the squamate dataset of Gauthier et al. (2012). The morphological characters for mammals by O'Leary et al. (2013) is another example—22 secondary characters for the nasal and 24 for the jugal and squamosal bones. The most extreme case known to us being represented by the 1,449 dental characters that are dependent on the two characters for absence/presence of lower and upper teeth (O'Leary et al. 2013). Furthermore, even small datasets (but also encompassing important evolutionary transitions)

may be subject to this problem. Such datasets do not usually have a large number of secondary characters dependent on a single primary character but may include many primary characters with (usually fewer) secondary dependencies. Some examples include datasets focusing on early-deriving snakes, in which various cranial, limb, and pectoral girdle characters may be either absent or present (Garberoglio et al. 2019), directly impacting all secondary characters contingent upon those traits.

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

There are three pieces of phylogenetic information universally present within primary and secondary characters as illustrated by the RBT problem (Fig. 1): i) the primary character WDSC (tail) groups all taxa with tails together and those without tails as a second clade; ii) the secondary character (tail color) groups red-tailed taxa together and blue-tailed taxa together as sub-clades within the clade where the primary character WDSC is present; iii) the logical dependency of the secondary character upon the primary character (tail) indicates that all aspects of the secondary character (defining the sub-clades with blue or red tail colors) should only be applicable to taxa in which the primary character is present (defining the larger clade for taxa with tail). Beyond these three aspects, there is no data provided by either the primary or secondary characters to inform which tail color evolved first. In fact, this is irrelevant for tree inference under either maximum parsimony (MP) or probabilistic methods, since reconstructing the direction of character state transformation (i.e., identifying synapomorphies) is only performed by MP upon the rooting of the tree once the most parsimonious solutions have been found (Nixon and Carpenter 1993, 2012). For probabilistic methods (maximum likelihood and Bayesian inference) outgroup comparison and the direction of character-state transformation is not taken into consideration during tree sampling (Felsenstein 1973, 2004). Therefore, in the absence of additional characters, there is

no single solution to the RBT problem as presented in Scenarios 1 and 2. Instead, any coding method or inference algorithm should allow the two possible solutions (i.e., red and blue-first hypotheses) to be equally likely *a priori*, and the secondary character (i.e., tail color) should only be considered within the clade composed by taxa where the primary character WDSC is present (i.e., tail). Therefore, any coding approach or inference method producing logically plausible and biologically realistic results must meet the following criteria:

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

Corollary 2.— All known states (e.g., red and blue tails) should be considered as equally parsimonious/likely to be the ancestral condition. Under BI, both hypotheses should also have similar posterior probabilities.

In the RBT problem, logically valid approaches based on these corollaries must estimate two or more distinct tree topologies, including both valid solutions within the zone of contention (e.g., blue-first vs red-first hypotheses). The consensus (strict or majority rule) tree estimated from the output trees meeting these criteria will necessarily include all taxa in the zone of contention as monophyletic (supported by the primary character WDSC), but with no preference for either blue or red evolving first. Hence, the consensus tree should necessarily be unresolved—i.e., depicting a polytomic relationship for the taxa within the zone of contention. These premises form the basis of our evaluation of our simulated approaches, as described below.

Available strategies to handle logically dependent (hierarchical) characters

Over the past three decades, numerous solutions have been proposed to handle this simple but pervasive problem, from new character coding strategies (Maddison 1993, Hawkins et al. 1997, Strong and Lipscomb 1999, Hawkins 2000, Brazeau 2011, Tarasov 2019) to new algorithmic solutions (Brazeau et al. 2019, Tarasov 2019, Hopkins and St John 2021). Alternative coding strategies include contingent, absence, and multi-state coding (Table 1). The vast array of character coding schemes, their benefits and limitations, have been reviewed in many recent studies (Simões et al. 2017a, Brazeau et al. 2019, Hopkins and St John 2021), and we refer the reader to these for further information (and also our Supplementary Material). In summary, despite the problems introduced by contingent coding, nearly all studies have agreed that it should be preferred over other strategies as it is the least spurious solution to the problem of hierarchical characters (e.g., the RBT problem) (Strong and Lipscomb 1999, Sereno 2007, Brazeau 2011, Simões et al. 2017a).

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

The performance of these recent alternative algorithmic solutions, however, remains largely unknown. Simulated datasets, in which the "true" answer is known, have only been used once to test phylogenetic accuracy using a small synthetic dataset (with eight taxa) and restricted to maximum parsimony optimization approaches (Hopkins and St John 2021). Large, simulated datasets are not only more similar in size to most empirical phylogenetic studies, but also allow evaluation of how different parameters affect inferences. These include variable levels of homoplasy, character evolutionary rates (contributing to branch lengths), tree symmetry (Maddison 1993, O'Reilly et al. 2018, Puttick et al. 2019), the proportion of primary and secondary characters (Hopkins and St John (2021), among others).

Importantly, morphological datasets are now frequently analyzed by probabilistic/statistical methods—maximum likelihood and Bayesian inference (BI)—across various living and fossil study systems—e.g., Lee et al. (2014), Giles et al. (2017), King et al. (2017), Simões et al. (2017b), Paterson et al. (2019), Simões and Pierce (2021). Yet, the problem of hierarchical characters has rarely been discussed in the context of probabilistic inference methods. One major exception is a recent study suggesting the polymorphic re-coding of characters following the concept of structured and hidden states Markov models to incorporate the hierarchical structure of primary and secondary characters into Bayesian inference as a solution to the problem of hierarchical characters (Tarasov 2019). However, no study to date has demonstrated if and how the problems introduced by hierarchical characters in MP impacts probabilistic phylogenetic algorithms to begin with, despite some previous suggestions that they would (Brazeau et al. 2019). At least in principle, theory suggests that likelihood-based methods should be less impacted by hierarchical characters. That is because all maximum likelihood and BI software implement variations of the Felsenstein likelihood optimization algorithm (Felsenstein 1973, 1981), which includes only a "down-pass" phase (from tips towards the root) for the calculation of likelihood scores at every node in the tree

being reconstructed. The absence of an "up-pass" phase during the optimization of ancestral nodes—which is characteristic of maximum parsimony approaches (Brazeau 2011, Brazeau et al. 2019)—would suggest, for instance, that the dependency problem introduced by inapplicable state scores in contingent character coding should not impact tree inference using likelihood optimization procedures.

Here, we utilized a series of simulations of morphological datasets and one empirical dataset to address the following questions: how do different character coding schemes impact the relative performance of MP and BI in both simple and complex morphological datasets? Under a common coding scheme, how do classical and recently proposed optimization algorithms for MP perform relative to each other and to BI in morphological datasets? What is the impact of different tree shapes and character models for the performance of each method? We find a striking contrast of results between simplistic and complex simulated datasets regarding best coding practices and a large disparity in performance among methods depending on tree or character distribution structures. Our results indicate that standard BI is significantly less impacted by contingent coding, displaying superior performance to all MP methods tested here, even those explicitly modelled to handle inapplicable characters.

Materials and Methods

Simulation 1: Simplified Synthetic Datasets

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

Dataset 1 (Scenario 1, symmetric trees).— Refers specifically to the RBT example of Maddison (1993) with 14 taxa plus 1 outgroup with their internal relationships fully resolved and with each internal node supported by one synapomorphy, with the exception of the taxa within the so called zone of contention (Fig. 1a). A total of 11 characters were used, which is the minimum number of characters to create this symmetric tree topology with one fully resolved clade one side of the tree and another clade of unresolved relationships on the other side of the tree. Subsequently, one or two extra characters were added to the dataset (depending on the coding scheme to be tested). For all coding schemes in which two characters are added, "character 12" is the primary character WSDC—(denoting absence and presence of tail) and "character 13" (denoting tail color) is the secondary character that is dependent on the primary "character 12" (Fig. 1b). Under multistate coding, both characters are merged into a single "character 12" (Fig. 1b). The primary character WSDC is convergently evolving the "present state" on the both sides of the tree, where the secondary character becomes applicable, creating the zone of contention *sensu* Maddison (1993). The basic tree topology (outside the zone of contention) is not impacted by either the primary character WDSC or its secondary characters (i.e., characters 12 and 13).

Dataset 2 (Scenario 2, asymmetric trees).—Simulates the tree example used by Strong & Lipscomb (1999, Fig. 12 therein). The objective with this dataset is to explore potential biases introduced by primary absences and resulting secondary inapplicable characters at the base of the tree. This dataset includes 7 taxa plus 1 outgroup with their internal relationships fully resolved and with each internal node supported by one synapomorphy, except for the taxa within the zone of contention (Fig. 1f-i). It is the equivalent of only one side of the tree in Scenario 1 in order to create tree asymmetry while keeping the same number of taxa between the outgroup and the zone of contention. A total of three characters were used, which is the minimum number of characters to create this tree topology. The tree topology is

strongly asymmetric and includes a single zone of contention. As for Dataset 1, one or two characters are added to represent primary and secondary characters for the various coding schemes.

Simulation 2: Complex Synthetic Datasets

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

The approach above gives us the following fixed parameters: T (total number of taxa), C (total number of characters), R (taxon/character ratio). Additionally, the total number of characters (C) can be represented by: $C = P_n + S_n$, where P_n is the total number of primary characters and S_n is the total number of secondary characters. We simulated three groups of datasets to cover the large range of proportions of secondary characters found in empirical studies. These include low (10%), intermediate (25%), to a high (50%) proportion of secondary characters relative to the total number of characters. Given a constant total of 60 characters, these proportions translate into S_n= 6, 15 and 30 secondary characters, respectively (Table 2).

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

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

and,

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

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

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

Therefore, here we simulated three categories for the distribution of secondary characters for datasets with 30 secondary characters: S_d = 6, 15, and 30 secondary characters per primary character WDSC (Table 2).

Simulated tree construction.—We generated two simulated master ("true") trees, one fully symmetrical and another with perfectly asymmetrical topology, to test for the impact of different tree symmetries on phylogenetic performance. Each tree included 31 taxa (30 ingroup and 1 outgroup) as defined in the previous section. To emulate the RBT problem, we designated 10 'crown' taxa in each sector of the symmetrical tree (total = 20 taxa) and 10 'crown' taxa in the asymmetrical tree—therefore fixing to 10 the number of taxa with applicable secondary characters forming the zone of contention (Fig. S1). All 'stem' taxa lying rootward of the 'crown' were designated to have the primary character WDSC absent, thus being inapplicable in respect to secondary characters.

Simulated dataset construction.—We used each simulated tree to generate 100 replicates of binary morphological data matrices containing 60 characters each distributed among the 31 taxa, with the proportion of primary and secondary characters as detailed in Table 2 for models M1-M5. We followed the conceptual approach of Puttick et al. (2019) known as "no common evolutionary mechanism", which does not use explicit molecular substitution models to simulate morphological datasets, as in most previous simulations of morphological datasets—e.g., (Wright and Hillis 2014, O'Reilly et al. 2016, Puttick et al. 2017, O'Reilly et al. 2018, Vernygora et al. 2020). Instead, each individual character is given a consistency index (CI governing its degree of homoplasy) based on a probability function of character homoplasy derived from an extensive survey of empirical datasets (Goloboff et al. 2017, Puttick et al. 2019). This approach is designed to generate morphological characters with a model that does not necessarily favor probabilistic inference approaches—in fact, possibly favoring MP (Puttick et al. 2019)—for directly comparing the performance of MP and probabilistic methods in phylogenetics (Puttick et al. 2019). For our simulated datasets,

we set a target CI index for the entire matrix to be within an intermediary CI range, between 0.4 – 0.5 [bin 5 in (Puttick et al. 2019)]. Therefore, all primary characters (regardless of having secondary dependencies) may have character state transitions across all taxa—i.e., in both 'crown' and 'stem' taxa—to replicate circumstances found in empirical datasets.

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

Empirical Dataset

To assess the impact on tree topology and character evolution from each algorithmic alternative in the empirical dataset, we utilized the morphological dataset on Palaeognathae (ratites and tinamids) published by Mitchell et al. (2014). This dataset includes several desirable properties for our research question, as it has: i) almost exactly the same number of taxa as in our complex simulation procedures (30 taxa)—and small enough to not be impacted by tree search efficacy of each method being tested; ii) primary characters with secondary dependences scored with the gap symbol ("-") to distinguish it from missing data ("?"); iii) a very small proportion of missing data. Over the last decade, phylogenomic data

has revealed the independent loss of flight in palaeognath birds, with tinamids (flight capable) inferred as deeply nested within flightless "ratites" (Baker et al. 2014, Mitchell et al. 2014), instead of forming their sister clade as traditionally inferred by morphological datasets (Worthy and Scofield 2012). This implies either convergence of morphological characters towards flightlessness or re-gain of flight in tinamids, resulting in independent losses or gains of several characters related to flight and cursoriality. Hence, key characters for the evolution of this group follow the scenarios described above (Fig. 1a), in which character hierarchy is expected to have the greatest impact on phylogenetic inference.

We recoded this dataset to match the expectations of contingent coding (see details in Supplementary Material) and reanalyzed it using the same analytical procedures tested for simulated datasets (MP-F, MP-M, MP-HSJ, and BI). To map the impact of convergent evolution on primary and secondary characters associated with flight and cursoriality, we conducted a second series of analyses in which we constrained the tree topology to a partial molecular backbone (i.e.,tinamids were constrained to be the sister group to moas).

Analyses of Simulated and Empirical Datasets

MP-F tree searches for the simplified datasets generated by Simulation 1 for distinct coding strategies were conducted using the "Implicit Enumeration" algorithm in the software TNT v.1.5 (Goloboff and Catalano 2016). For Simulation 2, tree searches were conducted using the *phangorn* R package (Schliep et al. 2017). For tree searches with MP-M optimization we used its implementation in the R package *TreeSearch* v1.0.1 (Smith 2018), which uses MorphyLib (Brazeau et al. 2017) to handle inapplicable data (Brazeau et al. 2019). For tree searches with MP-HSJ optimization, we used the "*dissimilarity*" and "*hsjScorer*" R functions from Hopkins and St John (2021), which work in conjunction with the branch-swapping functions available in the package *TreeSearch* v1.0.1 (Smith 2018)—i.e.,

"*TreeSearch()*", "*Ratchet()*", and "*MultiRatchet()*". For both MP-M and MP-HSJ approaches, starting rooted trees were subject SPR and TBR branch swapping operations, the results of which were used as starting trees for further analyzes with a series of ratchet iterations (functions "*Ratchet*" and "*pratchet*"), switching to the next run if the best score was hit 10 times, and stopping all searches if best score from each run was the same for 20 runs. The best scoring tree was used as the starting point for multiple ratchet (function "*MultiRatchet*") runs with the same criteria as above to obtain multiple most parsimonious trees.

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

Bayesian analyses used the Mk model for morphological characters (Lewis 2001) assuming the presence of variable characters only (Mkv model), with rate variation among characters sampled from a gamma distribution. Each analysis consisted of two independent runs using four chains each, sampling at every 1,000 generation, for a total of 10 million generations using the software Mr. Bayes v 3.2.6 (Ronquist et al. 2012).

Procedures for the empirical analyses were the same as for simulated datasets. However, for the second set of analyses on the empirical dataset (with the molecular backbone constrain), it was not possible to use MP-HSJ, as the tree search algorithms for this method—functions "*TreeSearch()*", "*Ratchet()*", and "*MultiRatchet()*" in the R package *TreeSearch* v1.0.1 (Smith 2018)—do not allow for the constraining of the tree topology. For

MP-M, the wrapper function "*MaximizeParsimony()"*in *TreeSearch* v1.0.1 (Smith 2018) was used to run parsimony analyses using the MP-M approach with constrained tree topologies to match the molecular backbone.

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

Performance Measures

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

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

similarity has been suggested to be less prone to the influence of wildcard taxa and tree shape (Smith 2020, Mongiardino Koch et al. 2021). However, to our knowledge there are no quantitative direct comparisons between the two metrics to assess their individual assessment of tree topologies in the face of variations in tree resolution. Tree resolution is a frequently overlooked component of tree metrics but has a direct impact on the assessment of tree performance (Vernygora et al. 2020).

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

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

Finally, considering the BI is not intended to provide a point tree estimate, we also examined the size of the parameter space using different coding schemes for BI results. We did that by calculating the mean and variance of RF distances among the post-burnin trees of the posterior sample *sensu* Wright and Lloyd (2020). Since the trees in the posterior sample

do not contain polytomies, the RF distance metric is not impacted by differences in tree resolution (see Results). This metric provides a perspective on tree disparity in the posterior sample (i.e., how loosely or tightly scattered trees are in the posterior distribution).

Statistical Analyses

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

RESULTS

Simplified synthetic datasets

Fitch MP (MP-F) .—Under MP-F, we find that four combinations of coding schemes/tree topologies meet the two corollaries for logically sound resolutions of the RBT problem (Table 3). Contingent coding is successful under Scenario 2 (asymmetric trees), but fails under Scenario 1 (symmetric trees), as illustrated in Fig. 1 (f-i) and discussed in the Supplementary Material. A second coding scheme to meet both corollaries is represented by unordered multistate coding under Scenario 1 (symmetric trees), which had been highlighted by Maddison (1993) as a solution to the contingent coding problem (Table 3). However, multistate coding fails under Scenario 2 as it cannot recover the hierarchical relationship between primary and secondary characters— as previously observed by Hawkins et al. (1997). This failure results in some taxa (in which the primary character WDSC is absent) to be estimated as nested within the zone of contention, and a strict consensus tree with reduced resolution relative to other coding schemes (Figs. S2-S7). Finally, all options including

character ordering logically prevent the basic assumption set by corollary 2, as the ordering scheme will inevitably and arbitrarily predetermine which secondary state (red or blue) will evolve first (Figs. S6 and S7, Table 3).

The only coding approach to successfully meet the conditions set by corollaries 1 and 2 above under both symmetric and asymmetric trees (Scenarios 1 and 2) is "absent coding" (Fig. S8 and S9, Table 3). Despite being briefly discussed in the literature before, absent coding was tested only once (Strong and Lipscomb 1999), and its ability to meet both corollaries was never previously realized (see further discussion in Supplementary Material).

Morphy MP (MP-M) .—This approach correctly recovers the hierarchical relationship between primary and secondary characters as well as correctly finding the blue-first and red-first hypotheses as equally parsimonious among the MPTs (Figs. S10 and S11, Table 3). This matches the expectations of both corollaries, as predicted (Brazeau et al. 2019).

HSJ MP (MP-HSJ) .—As with MP-M, this approach was designed to correctly recover blue-first and red-first hypotheses as equally parsimonious (Hopkins and St John 2021). As expected, it does recover those hypotheses among the MPTs (Figs. S12 and S13, Table 3) and the hierarchical relationship between primary and secondary characters is recovered, since those must be provided a priori by the user. .

Bayesian Inference-Mkv model (BI) .—We found a substantial contrast of performance between scenarios 1 and 2 concerning hierarchy (corollary 1). Regardless of the character coding scheme, BI analyses of symmetric trees always inferred the clade defined by the presence of the primary character WDSC (i.e., tail) as monophyletic in more than 90% of the sampled posterior trees (Figs. S14-S16, Table 3), and the posterior trees sampled successfully

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

In contrast, asymmetric trees were much harder to estimate using BI across all coding schemes, with the posterior sample of trees not converging towards similar topologies (Fig. S17-19) and with the focal clade defined by the primary character WDSC being inferred at drastically lower frequencies compared to symmetric trees (Table 3). However, the absent coding scheme still was the best performing one relative to competing coding schemes in this aspect (ca. 50% compared to 21 and 23% from other schemes).

Additionally, we expected the frequency of posterior trees inferring red and blue-first hypotheses to be similar to each other under corollary 2. We found exactly this pattern with almost identical sampling frequencies (<1% of difference) in the frequency of trees with blue or red first hypotheses under absent and multistate coding for symmetric trees (Scenario 1) (Table 3). We found similar results using absent and contingent coding for asymmetric trees (Scenario 2). However, contingent coding in Scenario 1 strongly favored a blue-first hypothesis (similarly to MP-F), whereas multistate coding in Scenario 2 favored a red-first hypotheses more strongly.

As with MP-F, absent coding was the overall best performing coding scheme for BI. However, whereas for MP-F absent coding met both corollaries for both simulated scenarios, in BI absent coding met both corollaries for Scenario 1, but only corollary 1 for Scenario 2.

Complex synthetic datasets

Performance of tree distance metrics.—We found that both metrics are insensitive to the symmetry of the starting tree (Fig. S27). For both MCI and Quartet similarity, similarity decreases approximately linearly with the number of NNI moves (Fig. S27a). MCI show

signs of saturation earlier than Quartet similarity, with a decreasing slope as NNI moves increase, while for Quartet similarity the relationship continues approximately linear even when the number of NNI moves is greater than the number of internal nodes in the tree (Fig. S27a). However, the two metrics differ more strongly in their response to decreased tree resolution. While MCI decreases approximately linearly with the number of collapsed nodes, quartet similarity is less sensitive to decreased tree resolution when the number of polytomies is small and decreases sharply when trees approach a complete polytomy (Fig. S27b).

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

Under MP-F, all coding methods had extremely similar performances regardless of the tree distance metric used (Fig 2a). Given the extremely similar results presented by both metrics, we evaluated the resolution error incurred by different coding schemes—see Methods. Resolution error was also identical across all three coding methods for both Type I (incorrectly resolved notes) and Type II (incorrectly unresolved nodes) for all coding schemes.

Under BI, however, mean, median, and modal accuracy values were significantly higher for contingent coding relative to absent and multistate coding under both MCI and quartets tree distance metrics (Fig. 2b). Furthermore, resolution error results indicate contingent coding induces a slightly lower amount of Type I and II errors compared to absent and multistate coding.

For the MP-HSJ optimization, quartet distances indicate no substantial difference in performance across distinct alpha values, whereas MCI indicates a likely worse performance for alpha values of 0 relative to 0.5 and 1, which is induced by higher proportions of Type II error (Fig. 2c).

Performance across methods.—When comparing all methods based on contingent coding—the best performing coding procedure (Fig. 2a and b) and the only one common to all inference methods—MP-F has a significantly worse accuracy compared to all other methods (Fig. 3). This result is consistent with predictions in the literature and is consistent regardless of accuracy metric (Fig. S20). However, the best solution among the three remaining methods depends on the performance metric. Similarity scores based on MCI (Smith 2020) suggest MP-HSJ perform the best whereas quartet distances indicate BI performs more accurately than other inference methods (Fig. S20). However, quartet distances were found to be more robust to variations in tree resolution when compared to bipartition metrics here (Fig. S27)—an important factor when comparing consensus trees. Considering this, we favor the results provided by quartet distances, which suggest BI outperforms all inference methods based on MP, even those specifically designed to handle inapplicable characters.

Performance across tree and character models.—The larger data dispersal and bimodality in the results for each inference method (Fig. 3) suggest that other factors influence their respective performance, two of which were explicitly modeled here: tree symmetry and distribution of secondary characters among primary characters WDSC.

Using quartets distances, MP-F performs significantly better for asymmetric trees compared to symmetric trees (Fig. 4a, Figs. S23 and S24, and Table S2), as predicted by the RBT problem (Maddison 1993) and in our simplified synthetic datasets (Fig. 1 and Table 3). MP-M performs significantly better than MP-F for both tree models, and with asymmetric

trees also significantly more accurately inferred compared to symmetric trees. MP-HSJ and BI have significantly greater accuracy relative to MP-M and MP-F (Fig. 4a, Figs. S25 and S26, Table S1). The latter two methods perform relatively similarly for datasets used to reconstruct symmetric and asymmetric trees, with a slight advantage for symmetric trees (although nonsignificant for MP-HSJ, Table S2). The greatest improvement in performance for MP-HSJ and BI relative to MP-F and MP-M is observed on the inference of symmetric trees (Fig. 4a,), suggesting they are more capable than MP-M of removing the problems introduced by inapplicable characters.

In contrast, the MCI metric suggests that accuracy in MP-F tree inference is similar for symmetric and asymmetric trees (Figs. S23 and S24). This surprising result contrasts with previous evidence from the literature and herein indicating symmetric trees (as in Figs. 1a, c-e) are considerably harder to estimate using MP-F compared to asymmetric trees (as in Fig. 1f-i) in the presence of inapplicable scores for hierarchical characters. This further suggests this metric is not capable of detecting meaningful differences in performances across methods.

The performance of distinct inference methods when considering different primary and secondary character distribution models (Table 2) indicates a significant decrease in accuracy of MP-F when increasing the number of secondary characters per primary character WDSC (i.e from M1 to M2 and M3), or when increasing the number of primary characters WDSC (i. e. from M3 to M4 and M5) (Figs. 4b, S25 and S26, and Table S3). Other methods show a similar pattern but with more attenuated differences. Model M5, with the largest number of primary characters WDSC performs poorly across all methods (Fig. 4b, Table S3).

When examining the tree-to-tree distances within each posterior sample from the BI analyses (Fig. S21), we observed that simulation conditions in which secondary characters are spread more evenly among primary characters WDSC showed higher mean RF distances

(i.e., models M3, M4, and M5). It should be noted that, unlike in accuracy comparisons between methods, a higher RF score does not mean more differences from a "true" simulated tree. This is a metric of within-posterior sample differences. In this case, a higher RF means that more different trees are being proposed and evaluated under these simulation conditions—i.e., they result in a larger tree parameter space. We confirmed this by calculating a per-posterior variance in the RF distance. This measure, too, indicated that greater dispersal of secondary characters is associated with exploring more disparate phylogenetic trees (Fig. S22).

Empirical dataset results.— Analyses of the recoded dataset for palaeognath birds resulted into various differences among their optimal solutions depending on the algorithms used. Unconstrained analyses produced the same single MPT under MP-F and MP-HSJ (α=0.5) and three MPTs using MP-M (yielding a well resolved strict consensus), which were all similar to the results from MP-F and MP-HSJ, except for variations in the placement of a *Struthio* and *Rhea+Pterocnemia* within the clade they form with casuarids (Figs. S28-S30). The MRCT from BI is well resolved and finds a similar topology to MP-F and MP-HSJ, except for the more basal placement of kiwis (*Apteryx*), which is found as the sister group to nearly all other ratites, instead of forming a clade with moas and elephant birds as in MP-F and MP-HSJ (Fig. S31). These results are all relatively consistent with previous morphological analyses of this dataset, especially in finding all ratites as a clade with tinamids as their sister group (and a single loss of flight in palaeognaths).

When constraining the analyses to a molecular backbone to enforce the convergent evolution of flightlessness (i.e., tinamous as the sister group to moas instead of a sister group to all ratites—Fig. 5, clade Y), MP-F resulted into four most parsimonious trees (MPTs—e.g., Fig.5a and b) that produce a poorly resolved strict consensus tree (Figs. S32-34). MP-M

resulted inferred a single MPT (Fig. S33) and BI produced a well resolved MRCT (Fig. 5c and d; Fig. S36). The greatest difference between these results is in the placement of kiwis—within the clade also formed by casuarids, ostriches, and elephant birds (Fig. 5, clade X) by most MPTs from MP-F and also by the MRCT from BI, whereas inferred as the sister clade to all other palaeognaths by MP-M (Figs. S34-36).

When mapping the evolution of anatomical traits evolving convergently among flightless lineages (i. e. in the constrained analyses), we observe some patterns which reflect our simulated results. The primary character WDSC (absence/presence of a scar for *lig. Collaterale medialis*) evolves the present condition convergently in flightless groups, whereas one of its dependent secondary characters (size of the scar—treated as a discrete character by this dataset), has variations in one of these groups (clade X, Fig. 5) but not on the other flightless group (moas, within clade Y, Fig. 5b and c). The ancestral state reconstruction for moas is state 1, seemingly enforcing state 1 as the ancestral state for most of the MPTs for clade X (Fig. 5b). This is the case for three out of four MPTs indicating state 1 evolving first, and a single MPT with ambiguous ancestral state for clade X (Figs. S32-34). On the other hand, the consensus tree from BI indicates states 0 and 1 as being equally likely to have evolved first (Fig. 5d). The result from MP-M indicates state 1 as the ancestral for Clade X (similarly to MP-F), but MP-M does not recover kiwis within clade X, making interpretation of this result harder to compare with MP-F and BI. These patterns are in general agreement with our predictions from the simulated results, in which MP-F favors state 1 evolving first among most of its MPTs, whereas BI results in either state as equally likely to have evolved first.

DISCUSSION

Differences between quartet and bipartition metrics to measure method accuracy

Here we found that quartet and bipartition metrics favor different inference methods. Our simulations show that this is likely due to a difference in the sensitivity of each metric to tree resolution in summary trees and topological differences, but not to tree symmetry. MCI decreases approximately linearly with tree resolution and small topological differences (Fig. S27). As a result, when trees being compared include polytomies (e.g., most summary or consensus trees from MP and non-clock BI studies), the underlying cause of distances estimated may be ambiguous. Quartet similarity, on the other hand, appears to be less sensitive to polytomies except for extreme cases, better reflecting differences in topology. When applied only to fully resolved trees, MCI possesses several desirable properties in relation to other metrics, including Quartet Similarity (Smith, 2020). When trees vary both in topology and resolution, however, interpretation from MCI can be problematic. By using both metrics, we are able to find that BI results in more accurate but less resolved trees, while MP-HSJ results in trees with higher information content shared with true trees because they are better resolved, although less accurate—i.e., include more false positives. Overall, we suggest quartet distances should be preferred over bipartition metrics for similar performance studies in the future.

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

Although contingent coding has traditionally been considered the less spurious solution to the problem of dependent characters, all conclusions regarding distinct coding strategies come from small, simulated datasets (Strong and Lipscomb 1999, Brazeau et al. 2019, Hopkins and St John 2021), equivalent in size and scope to our Simulations 1 (simplified synthetic datasets). By examining both symmetric and asymmetric tree structures

for Simulations 1 and ancestral state reconstructions for each of the three optimization procedures tested here (contingent, absent, and multistate), we find new results and interpretations concerning the utilization of these coding schemes. We find that the problems introduced by logical character dependency are most easily avoided by using absent coding, the only coding method meeting the assumptions of corollaries 1 and 2 discussed above (Table 3, Figs. S8 and 9). The better performance in comparison to contingent or multistate coding contradicts previous suggestions concerning this particular coding strategy using similarly small synthetic datasets (Strong and Lipscomb 1999, Brazeau et al. 2019, Hopkins and St John 2021).

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

By comparing the results of our Simulations 1 with more complex simulation scenarios (Simulations 2) we find important contrasts in our results and to previous conclusions using simplified datasets. When simulating larger datasets with explicit tree and

character model variations, there is no significant difference in accuracy or resolution error among distinct coding strategies for traditional MP (MP-F), regardless of the performance metric (Fig. 2). We attribute this difference to the fact that the detected advantages of absent coding in simplified simulations is counterbalanced by the negative bias introduced by the repeated occurrence of the absent state. As the number of secondary characters increases for larger datasets, it also increases the number of secondary characters with the absent condition, disproportionally overweighting the absent state. Although we did not explicitly test for a variable number of characters, we predict that datasets with a larger number of characters analyzed by traditional MP (MP-F) might see an even greater negative impact from the overweighting of the absent condition with absent coding, potentially leading contingent coding to become the most accurate coding, as previously suggested (Strong and Lipscomb 1999, Sereno 2007, Brazeau 2011, Simões et al. 2017a).

Under BI, contingent coding has a slightly superior performance compared to other coding schemes for the complex simulated datasets (Simulations 2) (Fig. 2b). This is expected from theory since BI is not as strongly impacted by inapplicable scores introduced by contingent coding as the Fitch algorithm for MP (MP-F) due to the absence of an "up-pass" phase in the former. Therefore, the advantages of absent relative to contingent coding detected for very small datasets under MP-F are not observed under BI. However, as BI also suffers from the biases introduced by the overweighting of the absent condition, there is an overall negative balance for the performance of absent coding relative to other coding schemes. Overall, absent coding demonstrates to be advantageous only for extremely small datasets (e.g., 7-14 tips) characterized by a very low number of secondary characters. However, many morphological datasets are around the size of the simulation dataset (Barido-Sottani et al. 2020)—in which contingent coding performs equally well or better than absent coding (Fig. 2). Additionally, considering the ongoing trend towards much larger

morphological data sets, contingent coding would provide a better fit for the vast majority of morphological datasets. In rare instances where very small datasets are still used (e.g., ca. 15-20 tips), we suggest users should test between potential topological differences between absent vs contingent coding schemes.

Limitations of approaches designed to deal with logical character dependency

Perhaps the first attempt towards solving the problem of logical character dependency, outside the scope of character coding schemes, was the utilization of step-matrices of costs—or Sankoff matrices—as they could embed hierarchical relationships among characters (Forey and Kitching 2000). These have long been criticized for the amount of time required to build individual matrices for every collection of primary character WDSC and their dependent secondary characters, among other issues—see further discussions in Brazeau et al. (2019). Recently, such problems were ameliorated by faster methods to construct Sankoff matrices in the program TNT (Goloboff et al. 2021). However, as the number of secondary characters increases in a dataset, this solution becomes less practical as it surpasses the total possible number of states allowed by TNT (32 states). This creates a maximum limit of four binary dependent characters (Goloboff et al. 2021). In addition to this limitation, costs of character state transformations are arbitrary and do not include any uncertainty measure. Therefore, Sankoff matrices may never be a feasible universal solution to the problem of logical character dependency.

Morphy (MP-M) (Brazeau et al. 2019) is, to our knowledge, the first algorithmic attempt to revise traditional parsimony optimization schemes to handle logical dependency between phylogenetic characters. Morphy (MP-M) was analyzed conceptually and empirically by subsequent studies, which criticized it for not controlling for primary characters and their relationship to secondary characters (the same major limitation of the

MP-F algorithm), leading to overweighting of absences for primary characters WDSC (Hopkins and St John 2021). Moreover, with a large number of secondary characters, both MP-M and MP-F approaches result in a larger set of MPTs, including solutions where secondary characters are treated as applicable, thus contrary to its primary goal. This behavior was not detected for the MP-HSJ method (Hopkins and St John 2021).

Our results support and expand upon those findings by establishing that MP-M optimization can improve on the performance of MP-F in datasets with inapplicable scores when reconstructing asymmetric trees (Figs. 4b), but not in symmetric ones (Figs. 4b, S17 and S23). Accordingly, the negative effects of inapplicable scores for contingent coding are expected to be the greatest in symmetric trees (Maddison 1993, Brazeau et al. 2019, Hopkins and St John 2021). Additionally, MP-M has greater accuracy across different models of primary and secondary character distribution in the dataset compared to MP-F, but MP-HSJ and BI are significantly more accurate under these same conditions (Figs. 4b, S23 and S24). This disparity in performance to MP-HSJ and BI increases both with the number of secondary characters for a single primary character WDSC (models M1-M3), as previously suspected (Hopkins and St John 2021), and with increasing the number of primary characters WDSC (models M3-M5).

Among all parsimony-based methods, MP-HSJ is consistently recovered as the best performing method, regardless of accuracy metric, tree structure, and character models simulated (Figs. 3, 4, S21-S24). We attribute this performance to the fact that this is the only approach that specifically identifies primary characters WDSC and each of their secondary character dependencies (Hopkins and St John 2021). However, MP-HSJ downweighs secondary characters to only a small fraction of the relative weight attributed to primary characters. This penalization increases proportionally to the number of secondary characters in a dataset and can be further boosted through its α parameter (Hopkins and St John 2021).

Our tests revealed that performance results under this approach are almost entirely insensitive to value of α, including a complete elimination of secondary characters with α = 0 (Fig. 2c). Such heavy downweighing of secondary characters may pose a limitation for datasets in which those characters are the only ones available to resolve relationships within the zone of contention (e.g., Fig. 1). This might be one of the key reasons for the superior performance of BI relative to MP-HSJ under the most accurate metric (quartets), even though BI does not distinguish primary and secondary characters.

The inapplicable states problem is mostly restricted to MP

The primary cause for the problem of contingent coding and its impact on tree inference relates to the two-steps approach towards the optimization of ancestral state in MP—the "down-pass" and "up-pass" phases of the Fitch algorithm (Fitch 1971, Brazeau 2011). Since BI programs use the Felsenstein optimization (Felsenstein 1973, 1981) when calculating likelihoods for internal nodes, which only has a "down-pass" phase, it would be expected that the impact of inapplicable characters from contingent coding would be strongly reduced, or at least substantially minimized, relative to MP. Our results in Simulations 1 support our predictions in finding that contingent coding in MP-F will favor a blue-first hypothesis 100% of the time and never return any trees with a red-first hypotheses in Scenario 1 (Fig. 1, Table 3). On the other hand, BI will favor a similar hypothesis (blue-first = 46.1%) but it retrieves the competing hypotheses at frequencies much higher than 0% (i.e., red-first = 21%) (Table 3). As expected by their design, both MP-M and MP-HSJ accurately find most parsimonious trees with both blue and red-first hypotheses.

The advantage of BI under Simulations 1 is limited to the better-studied Scenario 1 (symmetric trees). The difficulty of retrieving hierarchical relationships and reaching topological convergence in small asymmetric trees causes BI to fail corollaries 1 and 2 more

frequently than MP-F when estimating asymmetric trees (Table 3). Our findings thus corroborate previous studies suggesting symmetric trees can be more accurately reconstructed than asymmetric trees using phenotypic data under BI (Puttick et al. 2017, Puttick et al. 2019), although we do not recover such performance disparity for distinct tree models under MP-F. The exact cause for this difference in performance between tree symmetries is currently unknown. However, in this study the reduced accuracy of BI under Simulations 1 (Scenario 2) might be simply a result of the smaller size of this dataset compared to Scenario 1, providing less characters to infer this tree topology, as suggested by the high disparity in tree topologies in the posterior sample (Figs. S17-S19). Therefore, we do not see this as major limitation of BI, but simply a result of the boundary conditions for this particularly simulation scenario (with extremely small-sized datasets for Simulations 1), designed for comparison with both historical and recent studies using small datasets to address on the problem of logical character dependency,

Using more complex simulations combining several parameters and larger numbers of taxa and characters (Simulation 2), BI consistently recovers more accurate trees than MP using the traditional Fitch algorithm (MP-F). When analyzed under the quartet similarity metric, which is less influenced by tree resolution (Figs. 3 and 4), BI is also significantly more accurate than the two parsimony approaches that correct for inapplicable characters (MP-M and MP-HSJ).

Algorithmic solutions to logical character dependency have also been proposed in the context of Bayesian inference in recent years, such as for the utilization of structured (SMM) and hidden-state Markov models (HMM) (Tarasov 2019). While these newer methods can adequately deal with inapplicable states in dependent characters, no study had ever shown whether traditional BI using the Mk model would have a poor performance. Tarasov's comparison between traditional BI and SMM/HMM models is limited to a 4-taxon case

example, which may not generalize well to larger trees. The proposed solution to the RBT problem from Tarasov's SMM model (2019, Fig. 5 therein)—equivalent to our simplistic Simulations 1 herein using a symmetric tree topology—is the result in which red and blue tailed clades evolve "simultaneously" and receive similar posterior support in the majority rule consensus tree. This is the same result obtained here by using standard MP-F with the default collapsing rule in TNT (Fig. 1d), or when using the Mk model for BI under absence or unordered multistate coding (Figs, S14-16, Table 3)—the best performing coding strategy detected here for such small data sets. As demonstrated above, these results are expected for BI analyses due to the way that maximum likelihood optimization operates, and not something unique to the SMM or HMM models.

Limitations of BI and suggestions on how to move forward.

It should be noted that BI performing more accurately than alternative MP approaches does not mean it is completely exempt of biases introduced by inapplicable character states in contingent coding. The sampling of the posterior distribution via the MCMC algorithm is strongly impacted by the number of primary characters WDSC. In simulation models with an increasingly larger number of primary characters WDSC (M4 and M5), there is only a small difference in performance of BI relative to MP-M and MP-HSJ, with all three methods outperforming MP-F (Fig. 4b).

Additionally, by quantifying the distribution of posterior trees from BI across the tree parameter space (Figs. S20 and S21), we find that the mean RF distance between the posterior trees within each simulation for models M1 and M2 is considerably lower than for models with a larger proportion of secondary characters (M3) or with more primary characters WDSC for each dataset (M4 and M5), irrespective of coding strategy. The total variance (or disparity) of RF values is also considerably higher for models M3 to M5, except

for contingent coding, which is only higher for symmetric trees under models M3 to M5. Overall, this indicates a substantial increase in the size of the tree space when there is a large amount of secondary characters in the dataset (50% for models M3-M5), and especially when there is an increase in the number of primary characters WDSC within the same dataset (models M4-M5). This increase in the tree space (most notably in absent and multistate coding) makes it harder for the MCMC to sample across all local optima and reach the global optimum, which is the most likely cause further significant reduction in accuracy for models M4 and M5.

Our results demonstrate the pervasive and detrimental role of increasing the number of primary characters with logically dependent characters in phylogenetic datasets even whenthe proportion of secondary characters for each primary character WDSC decreases (models M3 to M5). The unfortunate consequence of our findings is that, considering a finite number of anatomical structures from which morphological characters can be created, increasing the number of morphological characters in a dataset will strongly rely on increasing the proportion of secondary characters that are dependent on the presence of these anatomical structures (primary characters WDSC). For instance, squamate (lizards and snakes) morphological datasets are typically high in the proportion of taxa with inapplicable secondary characters, given the high number of clades independently losing their limbs. As a result, limb-related characters become inapplicable for a substantial portion of the dataset—e.g., 71 limb and girdle characters (11.6% of all characters) for 35% of all taxa in Gauthier et al. (2012) and 67 limb and girdle characters (19.3% of all characters) for 25% of all squamates sampled in Simões et al. (2020). However, these secondary characters are important towards resolving relationships in other areas of squamate phylogeny, especially among early-branches of squamate evolution (Gauthier et al. 2012, Simões and Pyron 2021).

For those reasons, we do not recommend ignoring secondary characters altogether. Additionally, the strong negative impact introduced by logically dependent characters detected here are a direct result of our simulation parameters. These were explicitly designed to replicate the conditions under which logical dependency is expected to cause the greatest impact on phylogenetic performance (Fig. 1). How frequently these exact conditions are found in empirical datasets is hard to determine and are most likely dataset-dependent.

Considering the importance of secondary characters, long-term solutions to this issue should stem from more efficient exploration of the tree parameter space, either by more efficient algorithms [e.g., (Zhang et al. 2020)] and continued expansion of available computational power for phylogenetic research. It also may be advisable to conduct analyses with BI using Fossilized Birth-Death models, in which tip ages and other evolutionary parameters can be used to discriminate among topologies when complex dependencies are present. Yet, this remains untested, and the relative performances of competing BI approaches in the face of character dependency, including Tarasov (2019) new coding method, shall be the focus of future investigations.

CONCLUSIONS

Here we provide the first comparable benchmarks for recent algorithmic solutions for the problem of logical dependency among morphological characters, a relevant issue that remains unsolved for nearly three decades (Maddison 1993). We demonstrate that alternative maximum parsimony algorithms designed to handle logical character dependency can generally produce more accurate results than traditional (Fitch) maximum parsimony, especially in cases with symmetric tree topologies and with low numbers of secondary characters. The MP-HSJ algorithm is generally more accurate than the competing approach MP-M, but undated Bayesian inference is significantly more accurate than all MP

approaches. This simple alternative to analyze datasets with dependent secondary characters has long been overlooked. Importantly, increasing the proportion of secondary characters and of primary characters with dependent secondary characters that become inapplicable substantially reduces phylogenetic accuracy regardless of optimality criterion or character coding strategy. We expect that the development of more efficient algorithms to explore the larger tree parameter space created by secondary characters (especially for BI) might alleviate some of the existing limitations demonstrated here.

FUNDING

This work was supported by the Natural Science and Engineering Research Council of Canada postdoctoral fellowship to T.R.S. Work on this manuscript was supported by NSF DEB-2113425 and NSF-DEB-2045842 and an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P2O GM103424-20 to A.M.W. This work was supported by a Smithsonian Institution postdoctoral researcher fellowship to B.A.S.M.

AUTHOR CONTRIBUTIONS

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

REFERENCES

Baker AJ, Haddrath O, McPherson JD, Cloutier A. 2014. Genomic support for a moa–tinamou clade and adaptive morphological convergence in flightless ratites. Mol. Biol. Evol., 31:1686-1696.

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

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

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

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

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

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

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

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

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

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

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

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

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

Gauthier JA, Kearney M, Maisano JA, Rieppel O, Behlke ADB. 2012. Assembling the squamate tree of life: perspectives from the phenotype and the fossil record. Bull. Peabody Mus. Nat. Hist., 53:3-308.

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

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

Goloboff PA, De Laet J, Ríos-Tamayo D, Szumik CA. 2021. A reconsideration of

inapplicable characters, and an approximation with step-matrix recoding. Cladistics.

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

Goswami A, Polly PD. 2010. The influence of character correlations on phylogenetic analyses: a case study of the carnivoran cranium. In: Goswami A, Friscia A editors.

Carnivoran Evolution: New Views on Phylogeny, Form and Function. Cambridge, Cambridge University Press, p. 141-164.

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

Graybeal A. 1998. Is It Better to Add Taxa or Characters to a Difficult Phylogenetic Problem? Syst. Biol., 47:9-17.

Hawkins JA. 2000. A survey of primary homology assessment: different botanists perceive and define characters in different ways. In: Scotland RW, Pennington RT editors. Homology and systematics: coding characters for phylogenetic analysis. London and New York, The Systematics Association, p. 22-53.

Hawkins JA, Hughes CE, Scotland RW. 1997. Primary Homology Assessment, Characters and Character States. Cladistics, 13:275-283.

Heath TA, Hedtke SM, Hillis DM. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol, 46:239-257.

Hillis DM. 1996. Inferring complex phytogenies. Nature, 383:130.

Hillis DM. 1998. Taxonomic Sampling, Phylogenetic Accuracy, and Investigator Bias. Syst. Biol., 47:3-8.

Hillis DM, Pollock DD, McGuire JA, Zwickl DJ. 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol., 52:124.

Hopkins MJ, St John K. 2021. Incorporating Hierarchical Characters into Phylogenetic Analysis. Syst. Biol., Advance articled.

Keating JN, Sansom RS, Sutton MD, Knight CG, Garwood RJ. 2020. Morphological Phylogenetics Evaluated Using Novel Evolutionary Simulations. Syst. Biol., 69:897-912.

King B, Qiao T, Lee MSY, Zhu M, Long JA. 2017. Bayesian Morphological Clock Methods Resurrect Placoderm Monophyly and Reveal Rapid Early Evolution in Jawed Vertebrates. Syst. Biol., 66:499-516.

Klingenberg CP. 2008. Morphological Integration and Developmental Modularity. Annu. Rev. Ecol. Evol. Syst., 39:115-132.

Lee MSY, Cau A, Naish D, Dyke GJ. 2014. Morphological Clocks in Paleontology, and a Mid-Cretaceous Origin of Crown Aves. Syst. Biol., 63:442-449.

Lewis PO. 2001. A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. Syst. Biol., 50:913-925.

Maddison WP. 1993. Missing Data Versus Missing Characters in Phylogenetic Analysis. Syst. Biol., 42:576-581.

Mitchell KJ, Llamas B, Soubrier J, Rawlence NJ, Worthy TH, Wood J, Lee MS, Cooper A. 2014. Ancient DNA reveals elephant birds and kiwi are sister taxa and clarifies ratite bird evolution. Science, 344:898-900.

Mongiardino Koch N, Garwood RJ, Parry LA. 2021. Fossils improve phylogenetic analyses of morphological characters. Proc. R. Soc. Lond., Ser. B: Biol. Sci., 288:20210044.

Mongiardino Koch N, Thompson JR. 2020. A Total-Evidence Dated Phylogeny of Echinoidea Combining Phylogenomic and Paleontological Data. Syst. Biol., 70:421-439.

Murphy JL, Puttick MN, O'Reilly JE, Pisani D, Donoghue PCJ. 2021. Empirical distributions of homoplasy in morphological data. Palaeontology, 64:505-518.

Nixon KC, Carpenter JM. 1993. On outgroups. Cladistics, 9:413-426.

Nixon KC, Carpenter JM. 2012. On homology. Cladistics, 28:160-169.

O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, Goldberg SL, Kraatz BP, Luo Z-X, Meng J, *et al.* 2013. The Placental Mammal Ancestor and the Post–K-Pg Radiation of Placentals. Science, 339:662-667.

O'Reilly JE, Puttick MN, Parry L, Tanner AR, Tarver JE, Fleming J, Pisani D, Donoghue PCJ. 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. Biol. Lett., 12.

O'Reilly JE, Puttick MN, Pisani D, Donoghue PC. 2018. Probabilistic methods surpass parsimony when assessing clade support in phylogenetic analyses of discrete morphological data. Palaeontology, 61:105-118.

Paterson JR, Edgecombe GD, Lee MSY. 2019. Trilobite evolutionary rates constrain the duration of the Cambrian explosion. Proc. Natl. Acad. Sci. USA, 116:4394-4399.

Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol., 51:664.

Puttick MN, O'Reilly JE, Pisani D, Donoghue PC. 2019. Probabilistic methods outperform parsimony in the phylogenetic analysis of data simulated without a probabilistic model. Palaeontology, 62:1-17.

Puttick MN, O'Reilly JE, Tanner AR, Fleming JF, Clark J, Holloway L, Lozano-Fernandez J, Parry LA, Tarver JE, Pisani D*, et al.* 2017. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. Proc. R. Soc. Lond., Ser. B: Biol. Sci., 284.

Pyron RA. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Syst. Biol.:syr047.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol., 61:539-542.

Sand A, Holt MK, Johansen J, Brodal GS, Mailund T, Pedersen CNS. 2014. tqDist: a library for computing the quartet and triplet distances between binary or general trees. Bioinformatics, 30:2079-2080.

Schliep K, Potts AJ, Morrison DA, Grimm GW. 2017. Intertwining phylogenetic trees and networks. Methods Ecol. Evol., 8:1212-1220.

Scotland RW, Olmstead RG, Bennett JR. 2003. Phylogeny reconstruction: the role of morphology. Syst. Biol., 52:539-548.

Sereno PC. 2007. Logical basis for morphological characters in phylogenetics. Cladistics, 23:565-587.

Simões TR, Caldwell MW, Palci A, Nydam RL. 2017a. Giant taxon-character matrices: quality of character constructions remains critical regardless of size. Cladistics, 33:198-219.

Simões TR, Caldwell MW, Palci A, Nydam RL. 2018a. Giant taxon-character matrices II: a response to Laing et al. (2017). Cladistics, 34:702-707.

Simões TR, Caldwell MW, Tałanda M, Bernardi M, Palci A, Vernygora O, Bernardini F, Mancini L, Nydam RL. 2018b. The origin of squamates revealed by a Middle Triassic lizard from the Italian Alps. Nature, 557:706-709.

Simões TR, Pierce SE. 2021. Sustained High Rates of Morphological Evolution During the Rise of Tetrapods. Nat. Ecol. Evol., 5:1403–1414.

Simões TR, Pyron RA. 2021. The Squamate Tree of Life. Bull. Mus. Comp. Zool., 163:47-95.

Simões TR, Vernygora O, Paparella I, Jimenez-Huidobro P, Caldwell MW. 2017b. Mosasauroid phylogeny under multiple phylogenetic methods provides new insights on the evolution of aquatic adaptations in the group. PloS one, 12:e0176773.

Simões TR, Vernygora OV, Caldwell MW, Pierce SE. 2020. Megaevolutionary dynamics and the timing of evolutionary innovation in reptiles. Nat. Comm., 11:3322.

Smith M. 2018. TreeSearch: phylogenetic tree search using custom optimality criteria. Compr. R Archive Network.

Smith MR. 2019. Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. Biol. Lett., 15:20180632.

Smith MR. 2020. Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. Bioinformatics, 36:5007-5013.

Strong EE, Lipscomb D. 1999. Character Coding and Inapplicable Data. Cladistics, 15:363-371.

Tarasov S. 2019. Integration of Anatomy Ontologies and Evo-Devo Using Structured Markov Models Suggests a New Framework for Modeling Discrete Phenotypic Traits. Syst. Biol., 68:698-716.

Vernygora OV, Simões TR, Campbell EO. 2020. Evaluating the Performance of Probabilistic Algorithms for Phylogenetic Analysis of Big Morphological Datasets: A Simulation Study. Syst. Biol., 69:1088-1105.

Wiens JJ, Brandley MC, Reeder TW. 2006. Why does a trait evolve multiple times within a clade? Repeated evolution of snakeline body form in squamate reptiles. Evolution, 60:123-141.

Wilkinson M. 1995. A Comparison of Two Methods of Character Construction. Cladistics, 11:297-308.

Wipfler B, Letsch H, Frandsen PB, Kapli P, Mayer C, Bartel D, Buckley TR, Donath A, Edgerly-Rooks JS, Fujita M, *et al.* 2019. Evolutionary history of Polyneoptera and its implications for our understanding of early winged insects. Proc. Natl. Acad. Sci. USA, 116:3024-3029.

Worthy TH, Scofield RP. 2012. Twenty-first century advances in knowledge of the biology of moa (Aves: Dinornithiformes): a new morphological analysis and moa diagnoses revised. N. Z. J. Zool., 39:87-153.

Wright AM, Hillis DM. 2014. Bayesian Analysis Using a Simple Likelihood Model Outperforms Parsimony for Estimation of Phylogeny from Discrete Morphological Data. PLoS ONE, 9:e109210.

Wright AM, Lloyd GT. 2020. Bayesian analyses in phylogenetic palaeontology: interpreting the posterior sample. Palaeontology, 63:997-1006.

Zhang C, Huelsenbeck JP, Ronquist F. 2020. Using Parsimony-Guided Tree Proposals to Accelerate Convergence in Bayesian Phylogenetic Inference. Syst. Biol., 69:1016-1032.

Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. Systematic Biology, 51:588-598.

FIGURES CAPTIONS

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

FIGURE 2. Accuracy and resolution error for different coding and weighting schemes across distinct phylogenetic inference procedures. Results for absent (Abs), contingent (Cont), and multistate (Multi) coding schemes for MP using the traditional Fitch optimization—MP-F (a), for Bayesian inference—BI (b), and distinct weighting schemes for secondary characters as implemented by MP using HSJ optimization—MP-HSJ (c). For each quadrant, accuracy measured by MCI similarity (top left, in cyan) and quartets similarity (bottom left, in green), followed by resolution error measured by the proportion of incorrectly resolved nodes—Type

I error (top right, in orange), and incorrectly unresolved nodes—Type II error (bottom right, in red).

FIGURE 3. Overall accuracy of each phylogenetic inference method using the best performing accuracy metric (quartets distance) regardless of simulated tree or character models. All methods are significantly different in performance based on pairwise Mann-Whitney tests (Supplementary Table 1). For method abbreviations, see Methods.

FIGURE 4. Accuracy of each phylogenetic inference method using the best performing accuracy metric (quartets distance) for distinct simulated tree and character models. Difference in performance between symmetric (Scenario 1) and asymmetric (Scenario 2) tree models (a), and between different character models (see Table 2) (b), for distinct phylogenetic inference methods. There is a steady increase in accuracy from MP-F (top row) to BI (bottom row) for both model classes (a and b). Most results are significantly different in performance based on pairwise Mann-Whitney tests (Supplementary Tables 2 and 3), with notable exceptions: nonsignificant between tree models for MP-HSJ, and between character models M3-M4 for all inference methods. For method abbreviations, see Methods.

FIGURE 5. Results of analyses on empirical dataset (palaeognathid birds) from Mitchell et al. (2014), after character recording for the contingent coding scheme. a) Parsimony-based ancestral state reconstruction of the primary character 192 WDSC using the first MPT produced by MP-F (clade Y constrained—see Methods). b) Same as in (a), but for the secondary character 153, which is applicable only when character 192 is present (state 1). c) Parsimony-based ancestral state reconstruction of the primary character 192 WDSC using the

MRCT produced by BI (clade Y constrained—see Methods). d) Same as in (c), but for the secondary character 153, which is applicable only when character 192 is present (state 1). Note that all taxa with character 192 as present also represent flightless taxa and all taxa with character 192 as absent represent flight capable taxa. This represents just one instance out of 17 detected in this dataset of logically dependent morphological characters (see additional ones in Supplementary Material), some of which are directly related with the convergent evolution of flightlessness in this group, and which can be impacted by the treatment of logically dependent characters.

TABLE1. Character coding strategies to deal with logically dependent characters. See Supplementary Material for a detailed discussion on character coding strategies.

Coding Strategy	Description	Advantages	Limitations
Contingent	Introducing inapplicable states in secondary characters	• Maintains hierarchical relationships	• Logically dependent characters are treated as independent • Placement of taxa affected by relationships in distant parts of the tree
Absence	Using "absent" state in all primary and secondary characters	• Maintains hierarchical relationships • Avoids inapplicable states	• Overweighting of absent state • Logically dependent characters are independent
Multistate	Merging primary and associated secondary characters	• Avoids inapplicable states • Merges logically dependent characters	• Hierarchical relationships lost • Potentially unfeasible number of states

TABLE 2. Combinations of characters distribution models. Abbreviations: C, total number of characters; P_n, number of primary characters; P_s, number of primary characters WDSC; S_d, number of secondary characters per primary characters; S_n, number of secondary characters.

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

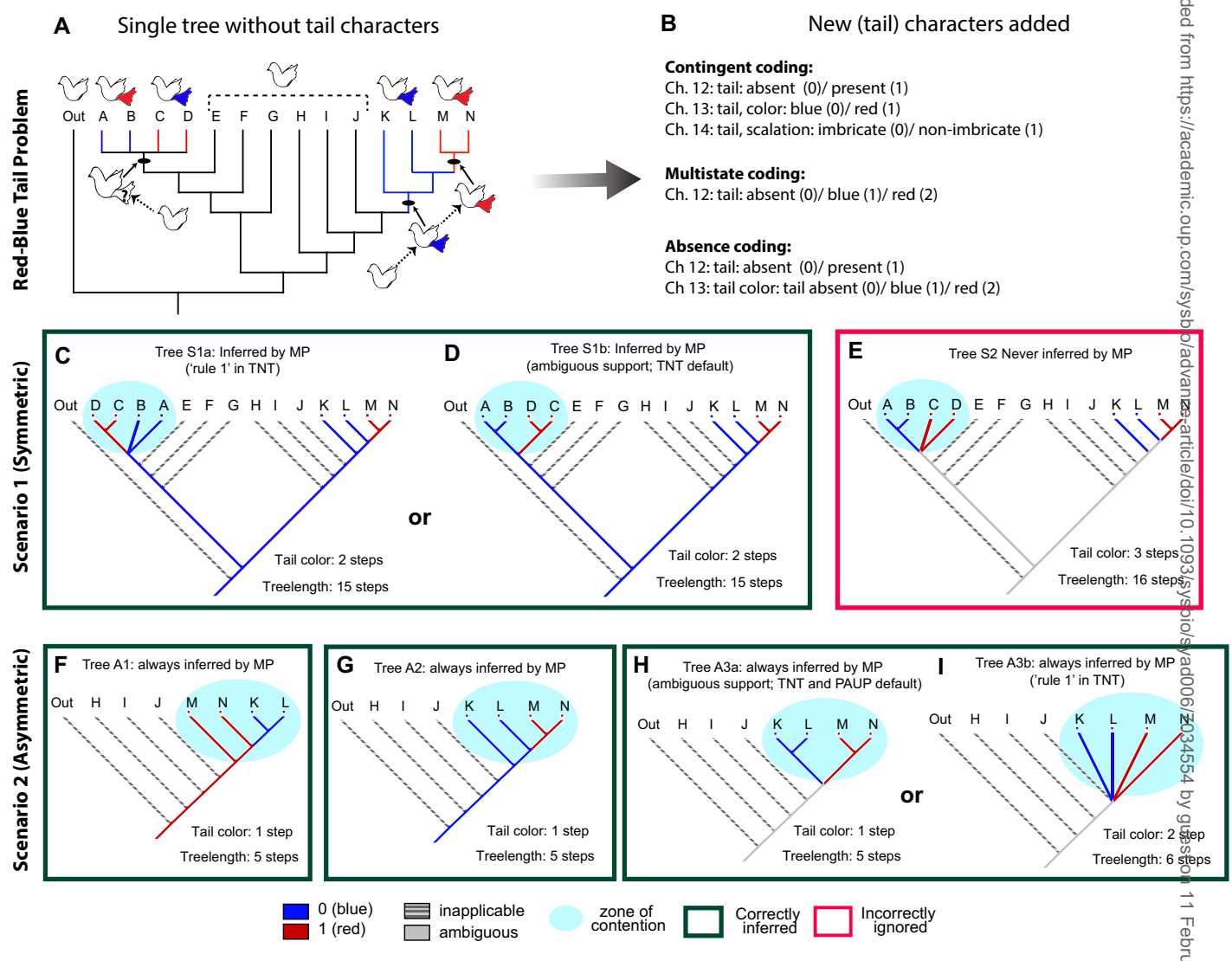
*Note that the number of secondary characters per primary character (S_d)on models M4 and M5 are the same as in models M2 and M1, respectively. However, the secondary characters in M4 and M5 are distributed across more primary characters (P_s), which will impact the final Fitch scores and tree lengths.

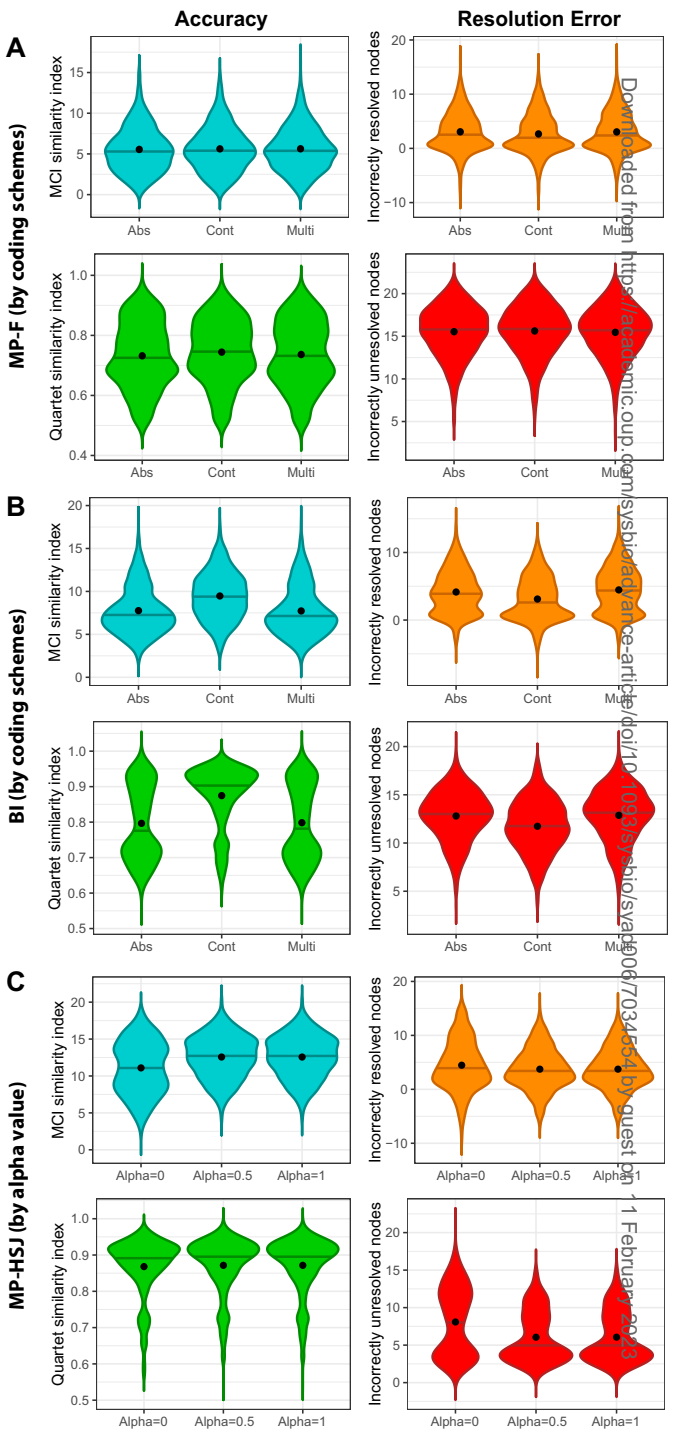
TABLE 3. Results for the simplified synthetic datasets using various coding schemes. Coding schemes meeting expectations from corollaries 1 and 2 are highlighted with blue background. Coding schemes with results pre-established by users (ordered characters) highlighted in gray**.** Results for coding schemes that are not applicable to particular methods are marked with "NA". Abbreviations: Abs, absence coding; B, blue tail-first hypothesis; Cont, contingent coding; Cor, corollaries; M, method; Multi, multistate coding; P-S, primary and secondary character hierarchy; ord, ordered; R, red tail-first hypothesis; unord, unordered.

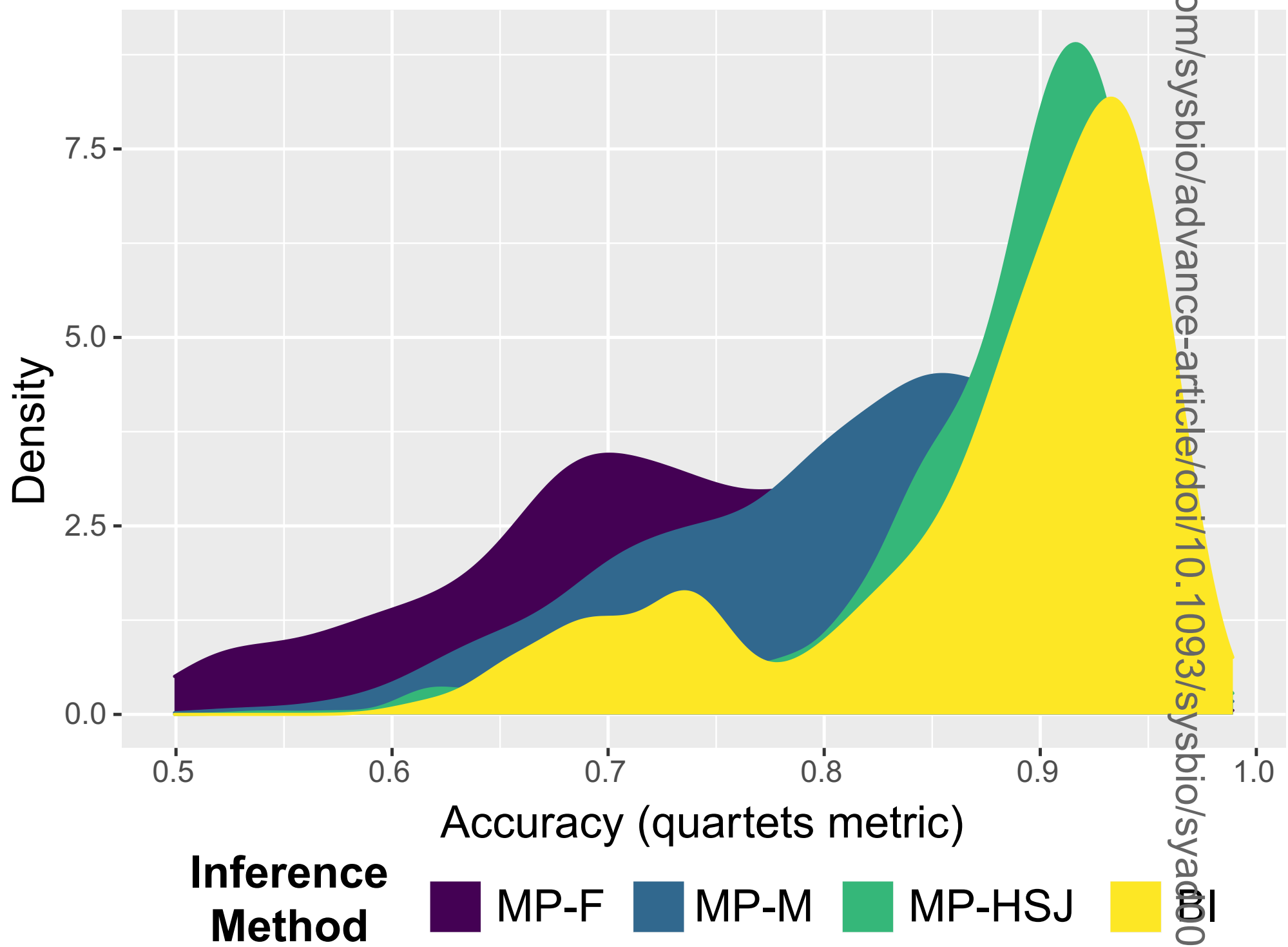
		Scenario 1 (Symmetric/two zones)					**Scenario 2 (Asymmetric/one zone)**				
M	**Cor**	**Abs**		**Cont**	**Multi**		**Abs**		**Cont**	**Multi**	
		Ord	**Unord**		**Ord**	**Unord**	**Ord**	**Unord**		**Ord**	**Unord**
MP-F	1	yes	yes	yes	yes	yes	yes	yes	yes	yes	no
	2	no	yes	no	no	yes	no	yes	yes	no	yes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

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

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






Density
7.5
5.0
2.5
0.0
0.5
0.6
0.7
0.8
0.9
1.0
Accuracy (quartets metric)
Inference Method
MP-F
MP-M
MP-HSJ
ML
om/sysbio/advance-article/doi/10.1093/sysbio/syaa00