# Handling Editor Comments to Author:

The author uses a simulated data set of morphological data to compare different phylogenetic approaches. The reviewers and the board member found the question to be important and liked the general approach, but had several major concerns. Importantly, the manuscript should be made more accessible for non-experts in the field, some parts are very technical.

I have rewritten the manuscript with a focus on making the prose easier to follow and more accessible to non-experts. Some technical information is now reported instead in supplementary information, to facilitate reproduction of results.

The manuscript is generally difficult to read due to a lack of structure and it would benefit from a typical structure including intro, methods, results and discussion.

These headings have now been added to help give the manuscript a suitable structure

In respect to the title, I can follow the reviewers’ problem with the word “equal”, but would agree with the author that the outcome of the work can be summarized in the title.

I have removed the word “equal”, providing a more modest summary of the principle conclusions.

Also the authors should refer to the tip-dating approaches which are an advantage of the Bayesian approach when dealing with extinct taxa.

A reference to tip dating has now been added, along with a statement emphasizing that the integration of non-morphological data is an additional strength of a Bayesian approach.

In addition, it would be important to know whether the general question only concerns morphological type analyses.

I have restructured the abstract and introduction to clarify that the issue is not unique to morphological datasets (though non-morphological data typically uses methods that can more readily be compared using standard statistical approaches).

Also the author fails to cite related studies, which already looked into the accuracy-precision issue.

Addressed; see detailed response to comments by referee 1

Finally, the study would be more accessible for the readers if traditional node support measures such as bootstrapping would be reported.

See detailed response to comments by referee 1

Overall, both reviewers gave many recommendations for improvement and I would like to ask the authors to change the manuscript accordingly.

I thank the editor and both referees for their detailed comments, which have helped me to make substantial improvements to the manuscript.

# Reviewers' Comments to Author: Referee 1

## Comments to the Author(s)

The author adds a new interesting take on the continuing debate comparing the efficacy of different phylogenetic inference methods for inferring topologies from discrete, morphological-type data. The author presents an argument that previously-used methods to judge efficacy of different phylogenetic inference methods have focussed too strongly on accuracy (i.e, whether the inferred topology is congruent with the true topology) over precision (i.e, the number of resolved nodes). As the author correctly points out, no tree topology is more accurate than a fully un-resolved, star phylogeny. Therefore, incorporating measures of precision is also important when assessing tree informativeness. However, I do have some major concerns at this time, and feel these need to be addressed prior to publication.

The key aim of the paper is to balance accuracy and precision in phylogenetic topological analyses. The paper appears to posit that some inaccurate precision (incorrect nodes) is preferable to a situation with no precision (unresolved nodes) for some nodes. Most phylogenetic analyses are used to elucidate relationships for a few key nodes; it would be better that these nodes are left unresolved (accurate but imprecise) rather than resolved incorrectly (inaccurate precision). Incorrect resolution of key nodes would lead to inaccurate biological conclusions. It is unclear if the key aim of the paper is to posit that some inaccurate precision is justified in any phylogenetic analyses, or just when comparing different methods for analysing morphological-type data. This needs to be clarified in the main aims of the paper

I have attempted to state the relevant recommendation for practice more clearly insofar as it pertains to the undertaking of phylogenetic analysis (i.e. collapsing nodes with least support). There are a variety of ways that researchers might apply this insight in this context, but this application is arguably secondary to what I would see as the main take-home message: that precision must be explicitly accounted for when comparing methods through simulation analyses.

The choice between accuracy-precision presented in the paper is not one necessarily faced when undertaking phylogenetic analyses. Whatever the method used to infer topology, the support for a key node(s) can be judge using various metrics, such as bootstrapping, Bremer supports, and posterior probabilities for Bayesian inference. Thus, there are ‘accepted’ levels of false imprecision for trees. By using node support measure such as bootstrapping a value of 50% support can be used: any nodes below this value will be collapsed to a polytomy. These measures already allow for a balance between accuracy and precision, but metrics, such as bootstrapping, are not considered for parsimony methods in the manuscript; posterior probabilities are used for Bayesian inference. Instead, the measure of support for parsimony is to include less-optimal trees in the consensus tree construction. However, it would be preferable for a future revision to employ bootstrapping on the optimal consensus trees for parsimony methods. This would have the advantage of having a definite level of acceptance (e.g., 50%) to compare trees across all methods

A lack of clarity in the description of my methods obscured the fact that I use the Bremer support metric to evaluate node support with parsimony methods. I apologise for the lack of clarity in the previous manuscript, which I hope the revision has addressed.

Bremer support (and its modification for implied weights) was selected for its simplicity and ease of computational calculation.

Bootstrapping is problematic in that it often assigns high credibility to spurious groups (Kopuchian, C., & Ramírez, M. J. (2010). Behaviour of resampling methods under different weighting schemes, measures and variable resampling strengths. Cladistics, 26(1), 86–97; Simmons, M. P., & Freudenstein, J. V. (2011). Spurious 99% bootstrap and jackknife support for unsupported clades. Molecular Phylogenetics and Evolution, 61(1), 177–191).

There is another potential pitfall with bootstrapping support values. Because both bootstrap support and posterior probabilities are expressed as percentages, some studies (e.g. Puttick et al 2018) have treated posterior probabilities (which represent the probability that clade is correct) as equivalent to bootstrap support values (which measure the proportion of different datasets under which a clade still appears in an optimal tree). A clade with a 50% posterior probability will not necessarily occur in 50% of bootstrap (or jackknife) replicates.

Because Bremer support values do not resemble percentages, there is less danger that a naïve reader will assume that the values of the support metrics are equivalent in this way.

This said, it would be very interesting for a future study to use the approaches advocated herein to compare the efficacy of different node support measures in identifying nodes that are most likely to be incorrect.

When the smaller datasets are used (for example 100 characters, 75 tips) the differences between methods are more pronounced with Bayesian methods appearing to perform better than the alternatives (Figure 2). These datasets are potentially more representative of most morphological matrices, especially from palaeontology, so more discussion is required on these results.

I’ve added a paragraph discussing the performance of methods at different sizes of datasets.

I feel the paper would be easier to read if it were separated into the usual sections: Introduction, Methods, Results, Discussion/Conclusions

These section headings have now been added

## Specific points:

Line 14. The term ‘informative’ needs to be clarified throughout as it is a major part of the paper

I’ve removed the word from the abstract, and provided an example when information is discussed in the main text.

Lines 14-15. Stating that previous studies have not addressed precision is not strictly true. For example, O’Reilly et al. (2016) investigated and plotted the relationship between accuracy and precision; these analyses showed that Bayesian methods are more accurate, but at the ‘cost’ of precision (resolution). Also, the study from Brown et al. (2017) investigated the relationship of accuracy and precision when using node support measures (bootstrapping).

I have added a further citation to Brown et al (2017) to emphasize their contribution. I’ve added an explicit remark discussing O’Reilly et al’s (2016) plots of RF distance vs precision. The latter do mention the role of resolution, but do not explore this in detail, even though it is potentially of great relevance to some of their conclusions – their discussion of implied weighting as a less accurate method, for example, does not mention its higher resolution as a likely explanation. I now describe the studies as ‘downplaying’ rather than ‘overlooking’ the role of precision, which I hope is a fair summary.

Lines 31. “An emphasis on accuracy therefore disadvantages methods that produce highly-resolved trees” Would this mean the opposite is also true – emphasising on precision disadvantages methods that don’t tend to be precise?

I’ve added ‘and vice versa’ to acknowledge that this issue does indeed both ways.

The study cited here (Brown et al. 2017) shows that the problem is too much imprecision for trees inferred under the Mk model using Maximum likelihood, and this issue can be improved by bootstrapping. By bootstrapping and incorporating this as a measure of support, the accuracy of the Maximum likelihood Mk model greatly increase when poorly-supported nodes are collapsed. Furthermore, it is not always the case that Bayesian implementation of the Mk model produces more poorly resolved trees compared to parsimony methods (Puttick et al. 2018).

Puttick et al (2018) summarise that “The Bayesian implementation of the Mk model recovers the fewest correct nodes […] the majority-rule consensus topologies of the most parsimonious trees from the equal weights and implied weights analyses tend to recover a large number of nodes, albeit with high variance”. This is to say, even if the high variance picks up some individual exceptions, the general rule is that parsimony tends to produce (by default) more precision – particularly under implied weights, for the reasons discussed by Goloboff et al (2015, 2018; see reference list in main paper)

Line 40. The quartet dissimilarity metric used here arguably conflates accuracy and precision by summing them as the numerator in the equation on line 40. It would arguably be more useful to look at these two measures (accuracy, precision) separately – this needs more discussion

The quartet dissimilarity method has been chosen precisely because it incorporates both these aspects of information in a single metric. I’ve added a line to the methods to emphasize that the Ternary plots serve to distinguish the contributions of accuracy and precision (i.e. resolution) to overall dissimilarity.

Line 52-53. I think it is a good thing to assess tree accuracy using various methods but I do not agree that quartet dissimilarity necessarily generates ‘more useful results’ than the normalised symmetric differences. Firmer arguments in support of the quartet dissimilarity over the symmetric differences are required.

I have provided a clearer account of the problems with the normalized symmetric difference in the main text, and a more detailed exposition in the supplementary information.

Line 58. Was this the Pyron (2011) tree originally used by Wright and Hillis (2014) I believe this tree contained polytomies – were these randomly resolved?

I’ve provided more details on the bifurcating tree topology, O’Reilly et al (2016) reused after Wright & Hillis (2014), who selected it from the topologies generated by Pyron (2011).

Lines 74-77. The use of less optimal trees to collapse character support seems odd. It would be better to establish support on the nodes of these strict parsimony trees (e.g., using bootstrapping); then compare the outcome of collapsing poorly-supported nodes to polytomies (as has been done for the Bayesian trees). Similar studies have already been done by Brown et al. (2017), O’Reilly et al. (2017), and Puttick et al. (2018). These studies also showed that differences diminished when support values were used to collapse nodes, but Bayesian approaches tending to achieve both higher accuracy and precision when data have little phylogenetic signal (Puttick et al. 2018).

This paragraph was poorly written, for which I apologise. I’ve now replaced it to make it clearer that the referee’s suggestion is very close to what I have done: the analyses establish support values for the nodes of the strict parsimony trees (using the traditional method of Bremer supports, rather than bootstrapping), and collapse poorly supported nodes to polytomies (as done for the Bayesian trees).

Line 84. I do not understand the reference to auto-correlation here, please clarify

I’ve rewritten this to clarify.

Lines 108-111. Could an acceptable, universal cut-off value be used via this method, or would it not be easier to use bootstrapping and collapse poorly-supported nodes? For example, using the approach employed by Brown et al. (2017).

I have rewritten the text to make it clearer that I am proposing that resolution is reduced by collapsing poorly supported nodes. I consider it unlikely that a universal cut-off value exists to determine which nodes should be collapsed: the optimal point in the trade-off is likely to depend on the nature of data from which trees are generated (e.g. number of characters, amount and distribution of missing and inapplicable data, conflict between characters).

Lines 118-125. Although not practical in a simulation study, the results of key nodes of interest in a phylogeny can also be tested using all methods. Thus, it would be worth discussing that for any method testing between key topological hypotheses can be conducted, without having to consider informativeness. For example, if researchers are testing monophyly of a clade with parsimony inference, a Templeton Test could be used, and Bayes Factor comparisons can be used for hypothesis testing under Bayesian inference. If the tree is being used for further down-stream analyses that require fully resolved trees (e.g., some comparative methods) then it would be better to conduct these across full uncertainty by using, for example, the full posterior from a Bayesian analysis or all optimal parsimony trees.

I agree with the referee that Bayes Factor comparisons are an excellent way to compare competing hypotheses, and that these should feature more prominently in the literature. Given the concise format of Biology Letters articles, I am not sure that the present contribution can do justice to a full discussion of explicit hypothesis testing or other comparative methods that might be applied downstream.

In the revised manuscript, I have emphasized that the primary purpose of my study is to improve the criteria by which different phylogenetic approaches are evaluated. To take an example, if it had turned out that parsimony methods systematically failed to identify the true tree (as other studies have insinuated, but which this study shows not to be the general case), then a Templeton Test will simply measure how confidently parsimony methods ‘support’ a false hypothesis. Only once a method has been shown to deliver useful results are further analyses of its output warranted.

## Figures

The figures are difficult to interpret at present. Using more informative labels for each axis would aid interpretation.

I now label the axes rather than the vertices, and provide additional detail.

Including all Implied Weights k values makes it difficult to differentiate which line is representing each method, so it would be better to reduce the number of lines and points on the plots.

Referee 2 suggested using a single colour for all implied weights analyses. Implementing this suggestion makes the graph much less busy. I’ve also reduced the opacity of all but one implied weight plots. This means that graphs still emphasize the importance of selecting an appropriate concavity constant, whilst making it easy to distinguish the different methodological approaches.

What does the top right plot inset represent in Figure 1 and the bottom insets on Figure 2? Are these zoomed-in version of each full plot? Please clarify

The insets have now been identified in the caption, and the corresponding region of the adjacent panel marked for easy visual recognition.

## References

Brown JW, Parins-Fukuchi C, Stull GW, Vargas OM, and Smith SA. 2017. Bayesian and likelihood phylogenetic reconstructions of morphological traits are not discordant when taking uncertainty into consideration: a comment on Puttick et al. Proc. R. Soc. B, 284: 20170986.

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

O'Reilly JE., Puttick MN, Parry L, Tanner AR, Tarver JE, Fleming J, Pisani D, and Donoghue PCJ. 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. Biology Letters 12: 20160081.

Puttick MN, O'Reilly JE, Pisani D, and Donoghue PCJ. 2018. Probabilistic methods outperform parsimony in the phylogenetic analysis of data simulated without a probabilistic model. Palaeontology.

Pyron RA. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Systematic Biology, 60:466-481.

Wright AM and Hillis DM. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. PLoS One, 9: e109210.

# Referee: 2

## Comments to the Author(s)

comments on M.Smith manuscript for Biology Letters

1 Oct 2018

Overall I very much like the concise & informative analyses in this manuscript, which makes a key contribution to ongoing debates on the nuances of various phylogenetic methods to reconstruct the Tree of Life. The data & interpretations herein are not just relevant to paleontological datasets, but for anyone hoping to extract phylogenetic signal from biological data. I recommend publication with minor revisions.

That said, the text could be made more accessible & thereby would have a greater impact on a broad audience.

I thank the referee for their positive review, and for their detailed suggestions, which have helped me (I hope) to better communicate my methods and the significance of my conclusions

One of these minor revisions is the title. I'd recommend deleting "equally" as it implies "equal" performance, whereas what (I think) is meant is that both methods are informative, broadly speaking (but not equal). I'm also not a fan of declarative titles, but in the age of Twitter I don't blame the author for putting conclusions in his title. How about something like "Probabilistic and parsimony approaches to reconstruct phylogeny in simulated morphological datasets"?

I have followed the handling editor’s suggestion of keeping the conclusion in the title (with apologies to the referee). I have dropped the word ‘equally’, and have revised the conclusions to make it clearer that I am finding that the methods do exhibit ‘equal’ (or at least similar) performance (with the caveats mentioned by Referee 1 regarding tip dating, Bayes Factor comparisons etc possible under Bayesian Inference)

Also the brief content of the supp-data files would be better as a brief table or appendix connected to the main text, rather than a PDF which is itself embedded as a supp-data file.

I prefer to have the supplementary files as separate plain-text files in the appropriate format, so that individual scripts can be installed and run on a reader’s computer immediately on unzipping. To integrate the code into a single PDF or HTML file would require a reader to copy and paste the text into an appropriate text editor and create files in the appropriate formats, which may represent an obstacle to less technically inclined users, and which (from experience) risks introducing errors or frustrations in the copying-and-pasting process.

If permissible, section headers would help the flow of the text ("Intro", "Methods", "Results", "Discussion & Conclusions", etc.).

Section headings have now been introduced.

line 15: for the abstract I'd replace "precision" with a slightly longer (but concise) definition of what this means, noting how precision differs from accuracy.

Rewritten

In line 32 "precision" is again hinted at but then referred to by multiple names, "Resolution (precision)...". I'd stick with just one, with the one you favor not in parentheses at first use. Having gone through the manuscript it seems that precision is used synonymously with resolution to indicate the extent to which a given topology is resolved, i.e., shows bifurcating nodes. "Resolution" seems to me to be more widely used & familiar in a phylogenetic context. So I'd delete "precision" throughout and just use "resolution".

This is a very helpful suggestion; I now use ‘resolution’ throughout.

line 26: In stating "...Such studies generally find probabilistic approaches to be most accurate..." the text here does not accurately represent what has actually transpired in the literature. This was partly true of papers like O'Reilly et al. & Wright & Hillis because they (inaccurately) generated hypothetical "known" trees (as noted by Goloboff et al. 2018, cited here) "using branch lengths common for all characters" and because "The results defended by O’Reilly et al. (2017) strongly depend on generating data with this model...". Even more damning, I think, is that some authors have misunderstood the relevance of ultrametricity to post-hoc evaluations of the performance of MP vs probabilistic methods:

"Unsurprisingly, the model trees of Puttick et al. (2017a) which produce the datasets where parsimony performs most poorly relative to maximum likelihood or bayesian analyses are the asymmetric trees, which (by virtue of being ultrametric; see below) have the greatest disparity in branch lengths" (Goloboff et al. 2018: 626). Also Goloboff et al. 2017: 14, in reference to Puttick et al. 2017 state "The successive splits branching off terminals in their symmetric trees have decreasing lengths ... The symmetric model trees that they used, instead, have all their branches with exactly the same length (0.2). Incredibly, Puttick et al. (2017) never discuss this difference between their two model trees, or the crucial implication that ultrametricity on an asymmetric tree forces a mixture of very long and very short branches."

Anyway none of this is news to the author but I'd be careful about making a straw-man along the lines of "the literature" thinks MP sucks, etc. Some authors have tried to make this case, but "the literature" as I read it remains rightfully unconvinced. (And this manuscript is an important contribution to this ongoing debate.)

I’ve tried to balance concision with detail in summarising the extensive literature surrounding this topic. I’ve now rephrased the summary of Goloboff *et al.*’s position such that the introductory paragraph does not leave the impression that previous studies have conclusively answered the debate.

line 29: Use a colon after a complete clause and before an incomplete clause. The phrase "no tree..." is complete and should be preceded by a semicolon or period ([**https://www.bartleby.com/141/**](https://www.bartleby.com/141/)).

Rephrased

line 33: "Barter" (= trade) sounds odd here & implies that precision & accuracy are mutually exclusive, but I don't think they are. There can be lots of one and less of the other in some cases, both maximized or minimized in others. Your intent is to recognize the extent of both across a range of examples, right? If so I'd just write that without implying that they're mutually exclusive.

This sentence has been removed, allowing the paragraph to convey this point more clearly.

line 46: "...properties of Shannon-Weiner..." is too jargony. Could you replace this with a plain-English explanation of what you mean (and/or provide citations)?

I’ve removed the explicit link to information theory to avoid the need for a detailed tangent here.

line 55: Perhaps it would be possible to make the paper (in "Biology Letters") more biological by noting if any of the matrices (like Congreve & Lamsdell) represent real or particular groups (gastropods, vertebrates, etc.), and refer to such matrices in the text by their biological name(s). Or are the "known trees" purely hypothetical? (I guess given the "simulated" in the title they are hypothetical.)

I’ve given a little more details about the trees (which are drawn from biology), which helps to emphasize that the matrices are simulated.

line 65: It would make this manuscript more accessible if you noted briefly why k=1 is unreasonable as a concavity function (and/or cite the paper(s) where this is explained).

A brief explanation is now included in the text, with a detailed exposition of the mathematics in the supplementary material.

lines 66-70: It would be helpful to ensure your readers know that each of your runs consisted of 4 chains EACH, so 16 chains total, right?

I’ve clarified that each run is independent, and contained four Metropolis-coupled Markov chains.

Also presumably all of your simulated matrices reached convergence across these 16 chains, and you defined 40% burnin based on some ESS, SDSF, LnL asymptotes, etc identified with your sumt/sump commands, Tracer, or similar. I'd recommend making all these details RE: convergence clear rather than a blanket "40% samples ...burnin".

This paragraph was originally written to correspond to the parameter values expected by the MrBayes program. I’ve now provided the same information in a more human-readable format that corresponds more closely to how the analysis actually works in practice. At the risk of introducing more jargon, I’ve clarified that samples are only taken from the cold chain (not all 16 chains) in MCMCMC runs.

The PRSF and ESS metrics establish that, taken together, the parameter choices (sampling frequency, number of generations, and burnin percentage) were sufficient for the four runs to converge and the sampled trees to accurately reflect the posterior distribution.

line 71: Again I'm not sure what "barter" means here (see also line 93). Maybe something like "To explore the behavior of precision and accuracy..."

Reworded to “To explore the relationship between resolution and accuracy”

line 74: "...from 50-100" Do you mean 20 levels of relaxed posterior probabilities between 0.5-1.0? (not sure if this corresponds to the "20 further trees" mentioned previously).

I’ve rewritten this paragraph to make it more comprehensible.

line 77: As for the array of "20 further trees" from Bayesian analyses above, exactly what this log distribution means is a bit opaque. I think you're selecting levels of suboptimality given a certain concavity value that correspond to integers 1-20 in an equally-weighted analysis, right? I'd add a line or two of explanation around here.

I’ve rewritten this paragraph to make it more comprehensible.

line 81: I would recommend noting here the specific R commands you've used in the main text, unless they're lengthy scripts in which case reference to a specific supp-data file would be helpful. E.g., for symmetric distances I would guess RF.dist or treedist in phangorn. I note the current "guide to supplementary files.pdf" does not mention any particular R scripts, beyond a link to SlowQuartet. Note that some asymmetric indices (e.g., distinct.edges in Distory) count arbitrarily resolved root nodes that differ between two topologies. I've not yet figured out how to prevent this but it's predictable at least and can yield sums of asymmetric distances (length(distinct.edges(A,B)+length(distinct.edges(B,A)) that are slightly greater than RF.dist(A,B). So far this doesn't seem to be an issue for the methods outlined here but I'd still like to know at least broadly the R commands the author has used to compare topologies.

I’ve now linked to the specific functions in the SlowQuartet library, which are robust to arbitrarily resolved root nodes. The SlowQuartet source code can be inspected internally within R, or via the GitHub link provided in the manuscript text.

line 84: As you can probably tell my preference is to read self-contained papers without necessarily having to frequently consult the cited refs to understand what's going on. Lazy on my part, I know, but maybe a sentence or two here would be helpful to explain the issue with "autocorrelation when graphing results".

I’ve clarified the problem at more length

line 93: again accuracy & resolution (= "precision") are not necessarily mutually exclusive.

Reworded

line 94: why the extra comma after "per", "[per, 7,22]"?

Deleted

The last two paragraphs represent a very nice summary of the overall conclusions of this paper (better than the current abstract!)

I’m glad to hear that these paragraphs work well! Hopefully my revisions have brought the rest of the paper closer to this clarity.

Fig. 1: Perhaps the lines representing various concavity functions could be shown in one color, distinguished by the various polygons, to better set them apart from Mk and equal weights.

Referee one also suggested simplifying the figures. Using a single colour for all implied weights analyses is an excellent suggestion and really helps to simplify the graphs. In the interests of clarity, I’ve not identified every possible value of *k*; rather, I’ve marked the best and worst values for each analysis, which I feel conveys the necessary information without overwhelming the reader with detail.

# Editorial office comments to authors:

Please ensure that you include;

\*Please also confirm in your cover letter whether all the figures are your own or whether permission has been obtained for their use

Yes, I generated all figures myself

\*Please upload your original figure files as eps, tif or jpeg files, rather than PDFs

EPS files are provided, alongside PDFs to verify correct rendering.

\*If you have any images that can be used to promote your article on social media (should this be accepted) please upload them as a supplementary file

No figures suitable for this purpose are available.