Dear Editor Niels de Winter,

Thank you for the final round of reviews. We implemented all the remaining suggestions from the reviewers according to the responses below (in red). We are looking forward to seeing our article published and thank you for your professional work.

Editor:

- Consider deleting or rephrasing the statement on line 85 about the use of data flagged by the original authors as altered if the procedure for excluding these problematic datapoints is not strictly followed in the dataset.

We rephrased this statement.

- Please clarify in the text whether median values and associated standard deviations or errors were computed by weighting data per species or in an unweighted manner (averaging all datapoints within a genus).

No weighting was used. As this was never implied anywhere in the previous version, we think that this should be clear already, but we nevertheless added a statement.

- Consider rephrasing the statement on line 614 in such a way that it is clear whether or not fixed thresholds for trace element chemistry would be appropriate as a way to screen for diagenesis. I feel that there is broad agreement between the reviewer and authors (and myself) that there is no single “one-size-fits-all” threshold for preservation and that the trace element concentration in a fossil (belemnite) does not solely depend on the “degree of alteration” but on circumstantial parameters such as the geological context and variations in the original in vivo carbonate composition for the taxon.

We rephrased this statement.

-Please clarify your recommendation for using stable isotope data in a phylogeochemical study. Perhaps adding a short statement highlighting that one should take into account long-term and spatial variability in the environment in such an approach would be helpful.

Yes, this is a good point, we clarified the statement.

- Consider clarifying in the discussion whether evolutionary trends can be confidently isolated from effects of diagenesis and environmental change on the trace element data. If this is not possible, I agree with the suggestion by the reviewer that "evolutionary" be dropped from the statements in the Abstract (line 10 and 14).

As outlined in the manuscript, it is challenging to separate these factors. Nevertheless, we believe that using only "rate" is somewhat unspecific, as it could also refer to a purely temporal perspective. The strength of our approach is specifically that we take evolutionary relationships into account. Therefore, in our opinion, it still makes sense to use "evolutionary", even if biological control is not the only factor. Furthermore, the term is established in evolutionary biology, where it is calculated using the same or similar methods. Thus, it would probably cause more confusion than would be solved by dropping "evolutionary". As a compromise, we replaced "evolutionary rate" with "phenotypic evolutionary rate" throughout the manuscript, as the phenotype may be influenced by both, genetic and environmental factors (in contrast to the genotype). For Mg/Ca and Sr/Ca, which are the only ones for which phenotypic evolutionary rates were calculated, diagenesis plays likely only a minimal role, as outlined before.

Reviewer:

Following the second round of reviews for the study "Phylogeochemistry: exploring evolutionary constraints on belemnite rostrum element composition", the authors have comprehensively responded to remaining comments and made further modifications, which I think have been very helpful to clarify the message of this contribution.

Thank you for your feedback!

Differences in opinion remain, but largely I feel that the last round of reviews and the extensive responses of the authors have been helpful in highlighting how certain points they were hoping to make were interpreted differently by me. At this stage it is clear that any differences in opinion will not be resolvable, which is not a problem in any case. I think the focus should now be on resolving residual potential for misunderstandings of what the authors stated they wanted to convey.

Thank you for your understanding. Even if different opinions remain, we believe that this is important to move science forward and thank the reviewer for the critical comments.

Use of data that have been flagged as problematic by previous authors

The authors have been clear and thorough in demonstrating that the inclusion – even of substantial numbers – of problematic data into their analysis has very limited impact on their interpretation due to the use of median values. I do not contest this to be true. I

nevertheless felt that the inclusion of data that have been shown to be unrepresentative of a taxon beyond reasonable doubt would be unfortunate. Readers may get the impression that these data are in fact valid. The authors note in L85 that they did not include data where evidence for alteration had been found, yet then state that they do so in certain cases (L 91). This approach seems unnecessarily arbitrary to me and – in my mind – does not really reduce circular reasoning, nor does it make data more comparable between studies (L90). For the benefit of the reader, I feel that either the statement in L85 ought to be deleted, or be followed stringently.

See comment above to editor.

Use of data that are (likely) erroneous

It is helpful that the authors included now in L114 that there is definite potential for some of the used Mg/Ca ratios to be incorrectly reported in the original paper and outline in L277 onwards what the consequences for this would be. From my point of view, a number of points speak for the listed ratios in the supplement to be computed mistakenly: The data were published by the same lead author in close temporal proximity, the same type of fossils and region were studied, there is the occurrence of a simple conversion error relating to Sr, the supplementary data files for both studies look nearly identical in design, and if one was to follow this interpretation, the Mg offsets that are otherwise observed would largely disappear.

However, a remote possibility remains that the data are in fact valid, and that has implications for other interpretations: If the published Mg/Ca data are correct and indeed low, what is the reason for belemnites for the studied region to be chemically distinct? Following on from this, if intra-generic differences can be so big, how much weight can be given to perceived evolutionary changes in Mg/Ca at higher level, especially where based on few measurements of only a small number of rostra?

The authors added some text towards this matter in L266 onwards, but I am not sure how methodologically this was computed and how robust this statement is. Were intrageneric s.d. values derived from s.d. of all individual measurements for this genus, i.e. unweighted, or for the averages (median) of individual species studied within the genus, i.e. weighted. My guess is the former, having tested this using data on Passaloteuthis: I have replicated the dataset as plotted in figure 3c and found that the median Mg/Ca for these species is 4.0 mmol/mol with 2 s.d. of 5.9 mmol/mol when using only the median values for the individual species and for Passaloteuthis sp. (n = 10). When using the entire dataset for Passalotheuthis, which is dominated by P. sp. and P. bisulcata, the result instead is a median of 9.1 mmol/mol with 2 s.d. of 4.4 mmol/mol (n = 569). If opting to show standard deviations, I think it would be insufficient to only use 1 s.d. values for comparison – 2 s.d. at least give 95 % confidence in the range. However, I am not sure whether it would be more appropriate to use the standard error

of the mean here, if the authors want to justify the robustness of the derived median values? However, as evident from the above, given the heterogeneity of the dataset it would drastically change the median value at least for this genus depending on how it was computed – much more so than what the standard error of the mean would suggest. Based on the entire dataset, a robust average of any sample taken from any Passaloteuthis would be expected to be 9.1 +/- 0.2 mmol/mol (2 s.e., n = 569), while the expected median for any species of Passaloteuthis would be 4 +/- 2 mmol/mol (2 s.e. n = 10).

Regarding the issues which are most prominent with Passaloteuthis, but impact also on other taxa, I have contacted the lead author on the affected studies to ask for clarification. Unfortunately, I have been unable to clarify the matter with her entirely before the review was due, but I do hope the authors can do so and will follow up on this point with her.

See comment above to editor. Note that species identification had no influence on the calculation of our median values and we never calculated any summary statistics at the species level. This would probably make not much sense, as (i) many species are represented only by a single sample; (ii) there is no way to verify the identifications; (iii) there are many taxa in open nomenclature (e.g., *Passaloteuthis* sp.), meaning that they may contain multiple species or are, in part, identical with other species in the dataset.

Regarding intra-generic differences between species: The taxonomy of belemnites can in no way be considered “final”. From a modern biological standpoint, only monophyletic groups should be accepted, but this has never been tested, and phylogenetic approaches have just been started by our group. Genera and to a certain extent even species are categories defined by humans and not “natural units”. There are many species today that can hardly be distinguished by morphology alone, sometimes this even extends to the genus level. Thus, it is very well possible that some of the species in the dataset should actually belong to a different genus. If geochemical signatures can be corroborated to be characteristic for certain taxa and correlate with other traits (e.g., morphometric data), this could someday maybe be used in taxonomic diagnoses. As of now, this is again a simplification but is currently the best we have. It highlights again that dedicated studies focusing on individual species with very restricted temporal and spatial range is the way forward. Such a study would be independent of the genus, as in the end, this is just a label.

Statement on diagenetic screening:

I am more comfortable with the way in which the text in section 4.3 has been updated and from the responses to the earlier review I sense that the authors are thinking largely along the same lines as I do. Just for the benefit of avoiding any confusion amongst the readers then I think a few points could benefit from final tweaks:

L608 notes that earlier studies often applied too high limiting element concentrations, which they tentatively and partially (L608-9) link to the lack of knowledge about pristine composition of fossil. They further caution in L612-13 that taxonomy and local diagenetic context need to be taken into consideration for confident assessment.

I fully concur that taxonomy and local diagenetic context are crucial to this, and it may also be true that previous authors have been a little optimistic when it comes to their screening thresholds. Where I think the disagreement lies is whether or not the original composition of the belemnite rostra is unknown, or at least what the extent of "unknown" is. In my opinion, to a practical limit it actually is known relatively well for a number of geochemical proxies and taxa, but I sense that the authors feel otherwise.

The example that the authors chose is found in L95: Given current knowledge we can assume that a representative Mg/Ca ratio of B. mammillatus is c. 12 mmol/mol, and for Sr/Ca the ratio is around 2.1 mmol/mol regardless of whether one thinks any of the published data are overprinted or not. To me this means that for these two element/Ca ratios, we have good knowledge of original composition. We can even constrain the likely data spread to allow future studies to compare against: Mg/Ca of 9.6-13.3 mmol/mol for 90% quantile and Sr/Ca of 1.9-2.4 mmol/mol for 90% quantile of screened data.

The authors rightly caution that we do have no proof that original composition for Mn/Ca and Fe/Ca is zero, an assumption that is implicitly made when assigning Mn and Fe thresholds. The authors (if I understand correctly) feel that this invalidates the use of these ratios. While ultimately I fully agree that we have no exact knowledge of the primary range of Mn and Fe values in any taxon, we tend to have quite reasonable constraints on this regardless. The question is whether this meaningfully impacts on their use for screening purposes, i.e. if it makes any practical difference if original Mn/Ca and Fe/Ca were zero, or some small finite quantity.

Following on with the same example that the authors note in the paper, we can only discuss Mn/Ca, but Fe/Ca would follow analogous reasoning. The observation is that the median Mn/Ca ratio in the tested materials was found to be 9 µmol/mol, so following above arguments for Mg and Sr this would be a reasonable maximum estimate for primary median Mn. For the studied specimens – distinct geochemical co-variation with Sr, C isotopes and Mg/Ca arises at Mn/Ca levels > 20 µmol/mol. C isotope values heavier than any other observed values and Sr/Ca ratios lower than any other values are seen in the samples from 20 µmol/mol upward in two out of three specimens. From this it follows that 20 µmol/mol signify (for this basin, time and taxon), a level where alteration is prominently developed, even though it cannot be excluded that some samples may inadvertently have been excluded that originally had Mn/Ca ratios > 20 µmol/mol. It can also not be excluded that samples with Mn/Ca ratios lower than 20 µmol/mol were retained despite changes to their geochemistry, but the geochemical proxy data for these samples do not markedly differ from seemingly well-preserved

material. It is important also to add, that samples that are excluded on the grounds of Mn/Ca ratios exceeding Mn/Ca ratios of 20 µmol/mol are nearly invariably found either in coherent sequences including the apical line or the rim of the rostra, which are areas that are well known to be most affected by diagenesis. It would thus be an odd circumstance for primary incorporation of Mn being most pronounced in these areas which are otherwise known to suffer from preservation issues and co-occur with values in other geochemical proxies that are most removed from the typical values for the taxon. The apical zone of the studied taxon has been shown to still be partially porous using petrographic techniques, increasing the likelihood of encountering diagenetic cements here, and geochemical trends prescribed to alteration point towards values akin to Late Cretaceous chalk (as known from other studies).

Given the above, does it practically matter that primary Mn/Ca ratios are not exactly known? Certainly not for any of the studied proxies whose distribution is not meaningfully affected by any choice of Mn limit lower than 20 µmol/mol. Mn itself is not interpreted and this will probably always be the case given the ubiquitous minor enrichments of this element even in samples that have seen very little diagenetic overprint. I would also add that, while it ultimately boils down to just one number (20 µmol/mol), this number was chosen based on varied evidence from multiple screening techniques, respecting the taxonomic and regional geological context. This is I think in keeping with what the authors argue should be done as well and includes (but goes beyond) “getting a statistical approximation of the “original” chemistry of the rostra”. The same approach would have been taken in numerous other studies that opted for other, locally appropriate limiting values based on varied considerations and a multi-proxy approach.

The authors note in their response to the last review that they “never suggested that there is a single “one-size-fits-all” threshold for each element”. As a reader I would take the statement in L614 differently. This line indicates that a single value may “give hints for more refined thresholds”, even though the cited table does not consider taxonomy, nor local geological context. I would just like there to be no opportunity to misunderstand the approach that I think both authors and myself feel is more appropriate.

Thank you for your detailed explanation. We implemented the change suggested by the editor (see above), so this potential for misunderstanding should now be avoided. At a very broad level, Table 4 would provide information on the order level (Belemnitida), but of course, it is preferable to consider at least the genus-level.

The use of C and O isotopes for comparable studies

The authors emphatically note in their response to the last review that C and O isotope ratios would not be used in the same way as median Sr/Ca and Mg/Ca ratios. My remark

in the previous review was to ask specifically, why such data were not utilised if they are available and useful for this kind of analysis, driven by the authors' proposal in the conclusions to do so, and by my attempts to understand how these data would be usefully applied. I agree that knowing temperature values for individual species would be great, but this is not something that can be readily obtained for belemnite species using a global median of O isotope data for each species. Things become even trickier for C isotopes. For instance, Passaloteuthis is known from the Pliensbachian and lower Toarcian from European sections as well as Russia; this period saw large-scale palaeoclimate and carbon cycle shifts. Which – if any – C and O isotope value out of the range recorded by the genus throughout its existence as a consequence of changes and heterogeneity in carbon cycle, palaeotemperature, and other environmental parameters, would meaningfully describe this genus?

I cannot help but read L637 which calls for "comparable studies for isotope data" in a way that C and O isotope data should be used in analogy to Sr/Ca and Mg/Ca. As a reader I would assume that this means one should use median C and O isotope data to add them into the mix of proxies that were studied in the present contribution. If this is not the intention, then perhaps just rephrasing this statement a bit would be useful?

Thank you for this point. Of course, there will always be necessary adjustments in the study design. The reviewer rightly points out that global averages of genera would not be appropriate. Nevertheless, this can be solved by only sampling specimens (ideally only a single species) from the same locality and a very restricted stratigraphic range). We clarified this in the revised discussion.

Clarity of message

I am now not fully sure what ultimately the authors' assessment of the main underlying cause of the observed patterns in their data is given that they contest my interpretation (in relation to L599 of previous review) that they concluded the data to largely reflect biological controls.

It is correct that the authors repeatedly state that interpretation of Mg/Ca and Sr/Ca data is complex, for example prominently so in the abstract (L10) and point this out also in their last set of responses: "This study highlights the complex interplay between evolutionary, ontogenetic, environmental and diagenetic effects".

However, they equally state that signatures are dominantly taxon-specific (L10), and display high evolutionary rates (L11), with "five evolutionary transitions" (L14). Within the text it feels to me that the authors ultimately build towards section 4.2 which promotes the idea that taxon-specific element uptake is the principle feature targeted by their dataset, and the statements in lines L545-547 introduce this by relativising any of the ontogenetic, environmental and diagenetic effects as things that can be

circumvented by specific sampling routines. Diagenesis had also already been discounted earlier in the study as a relevant factor (L95).

As a reader I would assume that “evolutionary rates” and “evolutionary transitions” are controlled by the taxa, and not by environmental change. Otherwise they would just be “rates” and “transitions”? I feel this is more than a semantic issue. If the element/Ca data are changing through time because of external factors, they do not perhaps mean very much in terms of evolutionary lineages of belemnites. If instead they are an expression of genetic changes in belemnites, and display large, repeated swings, this would lead to much different conclusions. If the authors feel unsure about the underlying causes of these changes, then perhaps “evolutionary” should be dropped or replaced for clarity?

See comment above to editor. Note that “taxon-specific” does not imply a genetic mechanism – different taxa live at different times in different habitats/environments, which could be another reason for these patterns. We replaced “evolutionary transitions” from the abstract.

Figure 7: I appreciate the extra work adding a chronostratigraphic chart as a new x-axis here. I think that this really improves readability of the graphs, even though perhaps the Cenozoic part of the plots is not required. It seems, however, that accidentally the wrong TJ-boundary age was added here.

Thank you for pointing this out, the age of the Lower/Upper Jurassic boundary was accidentally used in the previous version, which is now corrected. We agree that the Cenozoic part is perhaps not required, but since the names of some of the belemnite genera are relatively long, it avoids having excessive amounts of white space.