Thank you for the opportunity to review this manuscript, which is a response to a recent paper published in Scientific Reports (DePalma et al 2022). The manuscript calls attention to a number of issues regarding the Scientific Reports publication, especially regarding the lack of data availability, methods details, and possible inconsistencies in the presentation of results. The manuscripts argues that these issues represent evidence that the results may be fabricated, a very serious allegation indeed. I approach this as someone with 10+ years of experience conducting stable isotope analysis on biological and paleontological samples, including the collection, handling, and (micro)sampling of samples, running (and customizing/repairing) isotope ratio mass spectrometers and associated peripheral devices (including those mentioned in this manuscript), and handling datasets ranging in size from tens to thousands of stable isotope measurements.

My overall impression of this manuscript (perhaps shared by the authors), is that all of these issues ought to have been raised during the peer review and editorial process at Scientific reports prior to the publication of DePalma et al 2022, and thus are reasonable to raise in some form. However, a number of these issues are minor (not naming the analytical facility, not providing sample weights, not naming specific standards used) and do not either individually, or in combination, provide evidence one way or the other regarding the possibility of data fabrication. Some issues raised in this manuscript regarding the graphs in DePalma 2022 are potentially more serious, and are indeed worth raising, but I don't see a smoking gun. As such, I would ask the authors of this manuscript consider revising their manuscript such that it clearly acknowledges alternative interpretations of the issues raised, such as unintentional mistakes, database (copy/paste) errors, or graphing software misuse cannot be discounted.

Specific comments:

Lines 67-69: I agree the lack of data availability is unfortunate, and that the authors of the Scientific Reports publication should have included results with their paper. Some fault here also lies on the editor and reviewers of that paper, and as such this issue does not itself constitute evidence of fault solely on the part of the authors of the Scientific Reports publication.

Lines 71-74: I agree that it is good practice to include this information, but many papers do not and this does not constitute a major anomaly as long as there is some clear indication where the analyses were conducted and by whom, which the original Scientific Reports publication does clearly provide.

Lines 77-78: I agree that it is good practice to include such information, but this does not constitute a major anomaly but rather a minor omission that is often caught in the course of the peer review/editorial process. The authors of the PCI preprint might specify what other information they would wish to know regarding the techniques. For instance, one might wish to see a statement explaining that phosphoric acid was used to analyze carbonate component of fossil samples, and the reaction temperature.

Lines 83-89: This is a reasonable question to raise, and I agree here that additional

information should have been provided by the authors of the Scientific Reports publication regarding their sampling strategy, especially regarding the typical area over which powder was collected for each analysis.

Lines 110-113: How do the authors define failure of either measurement? Do they mean the software does or does not provide a value? Regardless, it is not correct to say that situations where either carbon or oxygen analyses fail (however defined), the other cannot still be used. Rather, it depends on why how failure is defined. For instance, if high inter-peak variation is observed in d18O for an individual sample, the d13C value could still be used if the its inter-peak variation is 'normal'.

Lines 113-117: I wonder if this could also be explained by repeated micro sampling of the same areas, measured multiple times, or potentially by errors in spreadsheet management and/or data use in graphing software.

Lines 117-118: Could this also be the result of 'sloppy' use of graphing software?

Lines 119-127: I agree the difference in error bar length is an issue worth raising, but as with the other issues raised here more innocent explanations such as simple sloppy graphing software use cannot be discounted. The parenthetical statement is not relevant here and should be removed.

Lines 128-133: The conclusions here are one possibility, but their case is very far from conclusive. I do not mean to suggest that such errors are unimportant, but sloppy handling of data and graphic software (perhaps by a student) could very easily result in such issues, which indeed should be corrected but are nonetheless not equivalent to intentional forgery. Thus, the authors should soften their language a bit, especially by changing "demonstrate" to "suggests the possibility" and also by acknowledging other possible explanations.

Lines 137-138: This sentence is too vague, please provide more information.

Lines 141-154: These are interesting points raised here, which are perhaps the most (really, only) compelling evidence to even raise the possibility of data fabrication.

Figures: Could the authors please define "misaligned" data points?