Dear Dr. Middelburg,

Below please find a point-by-point summary of how we edited the manuscript in response to each of the reviewers' comments. Our responses are in bolded blue text. We thank you and the reviewers for careful evaluation. Addressing the comments has, in our view, significantly improved the manuscript. We look forward to your further evaluation.

Sincerely,
James Stegen, on behalf of all co-authors

############

Reviewer 1:

General comments

The authors present a theoretical review of factors affecting the relationship between absolute quantity of a compound vs. analytical response as measured by high resolution mass spectrometry, as well as an experiment showing the differences in response factor for standard compounds in (negative ion) ESI-MS when measured in different matrices. They then discuss the ramifications of the fact that equal quantities of different compounds in a sample can produce different mass peak intensities or that the same quantity of compound may have a variable response in different samples depending on matrix, for the treatment of HRMS data from DOM analysis with statistical approaches designed for population ecology. As my expertise is mostly in the area of MS I will focus most of my comments on those sections.

The theoretical factors governing the response and identification of compounds in MS analysis, that the authors describe, are well known. Within the LC/MS community it is well known that one should not compare apples with pears, and that even comparing apples may be tricky as quantitation is difficult and often semi-quantitative. Not understanding the confounding factors can lead to over-interpretation of lc/ms data. Whether this message is something that needs to (still) be learned within the DOM community is unclear to me. I hope a reviewer from that community has more to say about that.

Thank you for these comments. While we agree that quantitation has been discussed within the context of LC-MS, our survey of direct infusion DOM literature indicates a need to renew this discussion with respect to the analysis and interpretation of direct infusion FTICR-MS data. We plan to maintain the vision of the manuscript, as our literature review indicates there is a need for increased awareness of challenges and pitfalls across the DOM community. We observed that ~10 years ago, DOM manuscripts often included discussion points related to the issues of using peak intensities. More recent DOM manuscripts rarely mention or evaluate these issues, even if they are comparing apples to pears. We believe the current state of the field, in which methods using peak intensities have been used in previous papers and authors are picking up those methods

and applying them without necessarily understanding the pitfalls, necessitates the comprehensive treatment presented in our manuscript. In revision we did, however, significantly reduce the length of text across the sections discussing theory and empirical evaluation of peak intensities from mass spectrometry.

The experiments with a set of standard compounds measured in different matrices is a nice illustration of the effect of ion suppression by matrix but is probably not necessary for the message of the manuscript as it is mostly a text book experiment with accordingly predictable outcome.

We prefer to keep these analyses because while they may be textbook for mass spectrometry experts, we believe that many folks using high resolution mass spectrometry (HRMS) data would still benefit from these examples, particularly users of HRMS data whose formal training is in other disciplines such as ecology or biogeochemistry. In revision we did, however, significantly reduce the length of text across the sections discussing theory and empirical evaluation of peak intensities from mass spectrometry.

The more novel part of this manuscript lies in the discussion on how these quantitative errors and uncertainties might influence the data treatment and outcomes. However, as I indicate in the comments below, I don't think that the in silico data set was generated with sound choices.

Please see our comment below related to modifications to the *in silico* simulations.

Below I list several specific comments and questions to be addressed. In general my recommendation is to shorten and simplify the descriptions of the factors governing quantitative response in (LC)MS, and focus the manuscript on the consequences for data treatment. A welcome additions would be a discussion on how to remedy the problems. As this probably entails more than major revisions, I recommend rejection at this time.

While we agree with many of the suggestions from the reviewer, we believe that the recommended major structural revisions are related to the interpretation of LC-MS data rather than the focus of our manuscript, which is direct infusion (FTICR)-MS data. We believe that many of the sections regarding quantitative response which the reviewer recommends removing or shortening would directly benefit the FTICR-MS community. Thus, we prefer to maintain the overall manuscript structure. In revision we did, however, significantly reduce the length of text across the sections discussing theory and empirical evaluation of peak intensities from mass spectrometry.

Many HRMS data users have limited formal training in mass spectrometry. Our manuscript provides an overview of the issues to be aware of as well as an immediate solution in the form of a simulation model. We also provide ideas for additional solutions (e.g., hierarchical modeling).

We will also note that the other reviewer (Reviewer 2) does not necessarily agree with all our interpretations of the experimental data. This demonstrates a clear need for further discussion to reconcile how data are interpreted. In other words, what may seem "textbook" to some (e.g., Reviewer 1) is not so simple to others. The only way to advance the field is direct sharing, discussing, and improving research outcomes via primary literature.

We also emphasize that the manuscript is not focused on LC-MS data. We are focused on direct infusion (FTICR)-MS data. We edited the manuscript to make this point clearer, especially in the early parts of the paper.

Specific comments

The second paragraph (lines 49 to 55) can be shortened and incorporated in the third paragraph, specifically in between line 67 and 68.

We feel there are important details in text and prefer to retain this material. Other parts of the manuscript have been significantly shortened, however, such that the overall manuscript is far more concise.

Line 75 to 76: peak intensity actually is proportional to differences in concentration for a certain compound, if the conditions are kept the same, otherwise no-one would be able to produce a standard curve. However the response factor (response per amount) may differ from compound to compound and for a compound depending on matrix and other factors. This should be rephrased.

We revised this text to clarify that we are talking about the relationship between peak intensity and concentration in complex, natural organic matter samples.

Line 96 to 99: move this section to the end of this section 2.3 as this is a concluding remark.

This text was removed during the streamlining of the associated section.

Line 101 to 112: Formation of (mixed) dimers and trimers should also be mentioned, as well as in-source fragmentation. There are several nice reviews on ion suppression , it would be good to reference a few here. The influence of the choice of ionization mode (+ vs - ionization; APCI vs ESI) should also be discussed. Also, a discussion on the use or (mis) use of internal standards would be a nice addition to this discussion

We have shortened this section of text and removed some specific details on different ion type formations with ESI, and so discussion on formation of clusters is no longer within the scope of this text. We agree that different ionization modes will also have impacts, however our main focus in this text is a general discussion of the most common technique. We have added a sentence explicitly stating that experiment-specific

considerations (sample preparation, ionization mode, instrument specific parameters) are outside the scope of our focus. We refer the reader to a recent review discussion on that topic. (10.1021/acs.est.1c01135)
We also have added a reference to a detailed review in matrix effects in LCMS (10.1002/mas.20298) highlighting that these effects are only exacerbated in direct infusion workflows.

Line 114 to 122: the issue described here is in fact not an ionization bias, but the inability to separate isobaric species. It simply describes the fact that when analyzing a complex (DOM) sample, any peak in any MS1 spectrum can in fact consist of the signal of multiple compounds with identical m/z (within the mass accuracy specifications) and is therefore a cumulative response of those compounds. This would still be the case even if the ionization of each compound was perfect at 100% efficiency.

We changed the title of this subsection to 'Ionization Efficiency and Isomers" and streamlined the text.

Section 2.3 The effect of dilution of any compound by abundant matrix should be discussed: in trapping instruments the fill rate of the trap (or ion target) is often a programmable parameter. If a compound, present at a given quantity is the most abundant compound there, than the trap is mostly filled with that compound. If the compound is present in the same quantity but with together with a lot of matrix ions, the matrix ions will take up part or most of the available space in the trap, effectively 'diluting" the compound and thereby to underestimation of its actual quantity. Maybe a better title for this section would be "ion transmission and collection"

We changed the subsection title to that suggested by the reviewer and added the following sentence: "Increases in the true abundance of other ions can decrease the measured peak intensity of a given ion due to a dilution effect resulting from a finite number of ions that can fit within the ion trap."

Lines 172 to 251: Section 3, in my opinion, can be removed in its entirety. These are predictable text book experiments and whatever extra information is discussed can be incorporated in section 2.

As discussed in response to earlier reviewer comments, we strongly prefer to retain this section. While mass spectrometrists may not be surprised by results shown in this section, the intended audience for this paper is much broader. We posit that the average ecology- or biogeochemistry-minded data user will not necessarily be aware of the issues or patterns addressed in this section. A key goal of our paper is to speak to researchers across the continuum, from ecologists to mass spectrometrists. At both ends of that continuum, researchers will find some of our examples very simple and 'textbook' in direct relation to their domain of training. However, few researchers are trained in all the concepts covered in our paper, and thus the paper provides common

ground for anyone working with direct infusion mass spectrometry, regardless of formal training domains.

Line 331 to 354: generation of in silico data set. I find the use of errors well above 1 (like 1.5) debatable. Although ion enhancement does happen it is very rare and seldom that pronounced. Even in fig 4, the increase shown in panel C from 0 to 2 ppm matrix added, is attributed to the addition of endogenous compounds from the matrix. So is this realistic? What would happen if you make the error non-gaussian?

We understand the questions raised by the reviewer and have made one targeted change to the simulation model; we modified the error range to more closely match the empirical data. This had no influence on the simulation outcomes; we have included a set of supplemental figures to show this and the code and figures are all in the publicly available GitHub repository.

More generally, we acknowledge that 1) there are multiple ways to set up the simulations, 2) we expect future developments in this vein, and 3) we are quite sure that different researchers will suggest and prefer different implementations of multiple aspects of the model. Part of the reason that many valid setups exist is that we are simulating phenomena that are not deeply characterized and may never be truly known. However, our simulation model's primary value lies not in the particular conditions simulated in our paper, but rather in its ability to be adjusted to more closely match a dataset of interest, thus providing a flexible framework by which mass spectrometry users can evaluate their own data. To this end, we provided the code used to produce the simulation model in a publicly available format.

We also edited the text to more clearly highlight the general utility of the model in the revised manuscript. A great follow-up manuscript would be a deep exploration of how the ecological metrics respond to different assumptions in the model, as well as extending to additional ecological metrics as pointed to in the manuscript.

Line 375: Again a comment about the choices underlying the in silico data set: Ionization efficiencies do not randomly vary for a compound across samples. In general, more complex samples with more matrix will have less ionization efficiency for all compounds in that sample. The deviations are not truly random.

In summary, we edited the manuscript to include additional ideas for simulation model development, such as the idea suggested here. Below, we provide more rationale for this plan, as opposed to myriad sensitivity analyses.

As noted in the manuscript, the presented simulation model is intended as a first look at what biases and/or issues may arise when peak intensities are used to calculate common ecological metrics. We feel it is important to keep the scope of the modeling as constrained as possible while still providing informative outcomes. While the point made

here by the reviewer is interesting, the practical question is whether we should expand the scope of the simulation model to account for it, or if we should provide this in the manuscript text as an example of future development needs.

More specifically, we ran the model with either ~100 or ~1000 peaks within each sample. The more peaks in a sample is, we believe, what the reviewer refers to as 'more matrix.' In this case, we could make ionization efficiency of all peaks inversely proportional to the number of peaks in a sample. We believe that this change will have a limited impact on the ecological metrics because all samples in a given simulation run have about the same number of peaks. Where this change could have more influence is in comparisons across samples that vary widely in the number of peaks (e.g., comparing one sample with 100 peaks vs. one sample with 10000 peaks). In theory, we could pursue this direction and implement a function that decreases ionization efficiency for all peaks in a sample based on how many peaks are in that sample. However, the extent to which this function accurately models physical phenomena is unclear, and this change will also add a significant amount of scope to the simulation component of the manuscript. For instance, we would need to add another large set of figures across both of our current error scenarios and include sensitivity analyses for how the effect of this mechanism changes depending on details of the function.

We are reluctant to expand the simulation scope in this way, which will add numerous figures and substantial length to the (already long) manuscript. In addition, we are not fully convinced that more peaks in a sample will, by itself, decrease ionization efficiency. Peak intensities may decrease, but that is because the concentration of each peak/molecule must be lower if there are more kinds of peaks/molecules and total dissolved organic carbon concentration is kept constant. Decreasing peak intensity due to lower concentration is not the same as decreasing ionization efficiency.

Given the above considerations, we added some new text in the manuscript laying out additional ideas for simulation model development, such as that suggested here. The added text is in the Conclusion section and reads: "The model should be expanded by including additional ecological metrics/analyses, more than two-sample situations, sample-to-sample variation in peak richness, links between peak richness and peak intensity, other ways of modeling error, and measured levels of error between concentrations and peak intensities."

Lines 385-408: as wordy as some of the other section are, as quickly the authors go through the consequences for the application of the statistical models. Some parts were truly unreadable to me as a person not involved in statistics and models. The concept of mean trait values and figure 11 are hardly introduced and poorly explained.

Our inference is that the primary challenge here is with the mean trait values. In turn, we add additional explanation that reads: "We also assigned an arbitrary trait value to each peak and calculated true and observed sample-level mean trait values; the mean values

for each sample were weighted by true abundance (true mean) or observed peak intensity (observed mean). This is analogous to the commonly used approach in ecological studies of computing community-level abundance-weighted trait values, such as plant leaf area index or animal body size (Muscarella and Uriarte, 2016). This approach is also common with HRMS data, such as sample-level peak-intensity-weighted values of hydrogen-to-carbon ratios and molecular weight (Roth et al., 2019; Wen et al., 2021).”

Line 483 – 489; After reading the entire manuscript the conclusion that the ecological metrics actually perform quite well, came as a bit of surprise to me. The tone of the manuscript as a whole is quite negative towards these concepts.

Based on this suggestion, as well as similar comments from the other reviewer, we edited the text throughout the manuscript to have a more positive and future-looking tone.

Fig 4: What does relative intensity mean here? Relative to what? As no bar is ever reaching 100%, is the response of the compound with no matrix or SRFA added 100%? Please clarify. Which of the differences between bars is statistically significant? However, I recommend to remove this figure along with section 3.

As discussed above, we greatly prefer to retain this section and this figure. By ‘relative intensity,’ we mean that the data were scaled to the largest signal in any replicate from that series of spectra. We are combining replicates to show the mean values with a 95% confidence interval, which is why the bars do not always reach 100%. These details have been added to the manuscript by editing the Figure 4 caption, the front end of which reads: “A) Barplot visualization of the relationship between signal intensity (relative intensity) and concentration of analyte for three chemically distinct molecules analyzed contemporaneously but independently in pure methanol solvent. Relative intensity indicates data were scaled to the largest signal in any replicate from the associated series of spectra. Replicates are combined to show their mean and 95% confidence interval.”

The primary value of the figure is to show how observed peak intensities deviate from expectations and assumptions made by ecological metrics commonly applied to direct infusion mass spectrometry data. As such, we prefer to not add details of statistical significance, beyond providing the 95% confidence intervals, as adding this information would introduce significant complexity to the existing figure, likely contributing to more confusion than clarity.

Fig. 8. Panel A, B and D are data clouds, which seems a logical outcome of the way errors are assigned in the in silico data set. But why the convergence to 0 for the true to observed difference in panel C? I do not see the strong resemblance between panel A and C described in line 383 of the text.

We added some explanation for why the data converged to zero in the middle of the data cloud. The following text was added to section 5: "In Figure 8B the differences collapse when near zero because when two samples have essentially the same peak intensity for a given peak, introducing the same error to that peak in both samples has little influence on the between-sample difference in peak intensity."

The point about lack of resemblance is due to an error whereby figure labels were misplaced (i.e., panel B should be C and vice versa); we corrected this error and thank the reviewer for catching it.

Citation: https://doi.org/10.5194/egusphere-2022-1105-RC1

Reviewer 2:

Given the growing literature on the application of ecological analyses to high-resolution mass spectrometry, a careful assessment of the bias, flaws, and assumptions is timely and a welcome contribution to the expanding subdiscipline. In particular, the suggestion to expand modeling approaches, including machine learning or hierarchical modeling, to quantify the magnitude of errors is an important one for future research.

Thank you for the encouraging remarks.

Meanwhile, the empirical results in this study show that many ecological metrics derived from the peak intensities provide valid patterns. This is an important result and could be highlighted alongside papers demonstrating the acquisition of quantitative data from HMRS (e.g., Kruve et al., 2020; Groff et al., 2022).

We added a reference to Kruve 2020 which highlights strategies for quantitative analysis of non-targeted LC-MS workflows. This was added in section 2.1, and the revised text reads: "Ionization suppression can be mitigated by online separation whereby non-targeted LC-MS approaches may yield more quantitative data (Kruve, 2020), but matrix effects remain a significant issue even for LC-MS (Trufelli et al., 2011)." Kruve 2020 and Groff et al. 2022 are both focused on LC-MS workflows, however. Our manuscript is contextualized around direct infusion analysis which lacks online separation. As such, we also edited the manuscript throughout to clarify that the focus of our manuscript is on direct infusion analysis.

On line 483, it is stated that HRMS has many weaknesses, just like any analytical platform. Most practitioners of HRMS would agree that the biases presented in this paper are present (as reviewed in Urban et al., 2016; Kujawinski et al., 2010). Many of these biases (Viera-Silva et al., 2019) exist with other compositional data as well (unlike the statement on lines 73-77), such as microbiome data and have produced solutions such as those reviewed in Gloor et al 2017, such as the creation of internal standards (Hardwick et al., 2018).

We added references to the suggested papers and pointed out there are methods developed in other fields that could be helpful for FTMS data. That text is in the final paragraph and reads: "In summary, FTMS has many strengths and weaknesses just like any analytical platform. Other types of compositional data also contain biases and uncertainties, such as the lack of true quantitation in sequence-based microbiome data (Gloor et al., 2017). Careful use of FTMS peak intensity data informed by objective, model-based guidance can overcome some of its weaknesses. We encourage further development of the model presented here and inclusion of additional methods developed to address issues that arise in similar data types (e.g., Gloor et al., 2017; Hardwick et al., 2018; Vieira-Silva et al., 2019)."

We also edited the text around lines 73-77 to clarify that we are referring specifically to FTMS data. That text now reads: "These studies may be discarding useful information, though it is unclear what biases and uncertainties are introduced into ecological metrics when using FTMS peak intensities."

Together, these two foci highlight the most problematic aspect of this paper: The tone is much too negative to accurately reflect the reality of the (very common) use of peak intensities. A more balanced and contrasted view should be taken so as not to alienate specialist readers or mislead those less well-versed. The tone leaves the feeling that the bulk of the literature in the past decade is highly flawed and not to be trusted. This is not impossible, but if this is what the authors are trying to convey, the analysis must be made much more robust. I suggest the authors change the tone to ensure that the key messages (the utility of ecological metrics and some of their drawbacks) are most effectively conveyed.

The other reviewer had a similar comment. We did a thorough edit of the whole manuscript to present a more positive and forward looking tone.

I am also concerned that some of the assumptions of the empirical work have significant technical flaws. This may be improved by providing greater transparency for the selection choices.

Please see comments below regarding our choices of compounds to empirically study.

1) The errors of their simulation model. How can a random selection between 0 and 100 for the simulated errors be justified (lines 352 and 369)? Why is 0 included in the random selection? The decision for this range should be motivated by actual evidence, such as from the experiments measuring variation in peak intensities of analytes of known concentration. Examining Fig, 4b. it looks like a better error selection would be between 1-8. Without further justification, the results of the in silica simulation model appear quite arbitrary.

The reason for including zero is to allow molecules to drop below the limit of detection of the instrument, such as in the cases of very poor ionization given either their inherent chemistry and/or interactions with other molecules in the sample/matrix. To address the

question about sensitivity of the simulation outcomes to the range error range, we used the empirical data (as suggested) and re-ran the simulations using an error range from 0 to 8. This had no influence on the results. The outcomes are in the GitHub repository and are included as supplemental figures.

The revised text is within section 5 and reads: "The inclusion of 0 indicates situations in which a given peak (i.e., ion) does not ionize well enough to be observed. The results should not be sensitive to the selected range, but as a sensitivity analysis we also used a distribution of errors ranging from 0 to 8. This narrower range is suggested by our empirical data (Fig. 4B), but simulation results were not affected (Supplementary Figs. S3-S8)."

2) Peak intensities are normally distributed in HRMS data (e.g. He et al., 2020). The way the authors generate random intensities does not reproduce the normal distribution in peak intensities.

Typical FTMS spectra of NOM, especially highly processed standards like SRFA, show an envelope of peaks across the m/z domain which looks similar to a normal distribution. However, peak *intensities* are not normally distributed. That is, if you take all the peak intensities and sort them by intensity - rather than by m/z - the distribution has an apex at low intensities, and is closer to a poisson distribution. Further, NOM samples are typically more heterogeneous than SRFA due to a higher mix of 'fresh' organic matter components. Given this heterogeneity, we believe our approach to random sampling is appropriate. We have also made recommendations in the conclusions section that the modeling approach can be expanded for more experiment-specific analysis.

We also edited the text to highlight the possibility of modifying the simulation model to reflect different distributions of peak intensities. The text in section 6 now reads: "It should be possible to include the number of samples, the number of peaks in each sample, the peak intensity distributions, number of replicates, and the specific ecological analyses that will be applied."

3) The number of peaks. The simulation models use either 100 or 1000 peaks. These are not environmentally relevant. Most environmental studies have several thousand peaks, where the authors nicely and unequivocally show in Fig. S2 that there is absolutely no bias in the calculation of ecological metrics, that is, the observed vs true R2 values approximate 1. For this reason, any simulations with small numbers of peaks are misleading and not relevant to most studies.

With respect to the 100 and 1000 peak simulations, we note that real datasets vary tremendously in the number of peaks being used for analyses. There are many reasons for large variation in the number of peaks actually used in the calculation of ecological metrics. For instance, some researchers may only examine peaks with assigned formulas and/or peaks observed consistently across technical replicates. It is important for

researchers to be aware that biases and uncertainty will increase as the number of peaks used decreases. Thus we feel that the continued inclusion of the analyses, figures, and discussion informed by simulations with 100 or 1000 peaks is warranted.

We edited text at the end of section 5 to read: "The variation in observed values explained by true values (via a linear model) increases rapidly with the number of peaks, and sharply asymptotes beyond ~500-1000 peaks per sample (Fig. S2). Sample-to-sample changes in the value of ecological metrics can, therefore, be interpreted with increasing confidence as the number of peaks increases. Qualitative gradients are, therefore, more robust with more peaks. The absolute magnitude of some ecological metrics, however, are shifted away from their true magnitude even when there are large numbers of peaks (e.g., Fig. 10D). Quantitative comparisons from one dataset to another may, therefore, require further simulation-based evaluation. We also caution that the number of peaks needed to reach the asymptote, thereby minimizing error, is likely dataset dependent and 500-1000 peaks should not be taken as a general rule for real-world datasets. We encourage researchers to complete such simulations using the numbers of peaks present across their real-world datasets to better understand their ability to make statistical and conceptual inferences."

4) Why these specific standards? What are their features beyond just an absence in natural DOM. I think there needs to be a description of what makes the molecular structures, ionization properties, etc… of these analytes appropriate spike-ins?

We added a sentence in section 3.1 to explain the rationale behind these standards. The sentence reads: "We selected chemical standards which are natural products with molecular formula and chemistries typical of compounds commonly observed in organic matter, and were amenable to negative mode ESI analysis."

5) No evidence is presented why these higher analyte concentrations (>200 ppb) or with relative intensities of individual peaks >1% will ever be realistic. I similarly don't understand why the summed relative intensities exceed 100% in Fig 4a in the absence of SRFA.

We agree that the higher concentrations we used may be greater than typical concentrations for individual endogenous molecules in NOM, but again note that in a typical NOM mass spectrum, any given peak is the sum of many different isomeric compounds. By using individual compounds, we have to increase their concentration to compensate for this reduced complexity. We have text to this effect in section 3.1, which reads: "These standards were analyzed at higher concentrations than typically observed for NOM because they were single compounds rather than formula-summed features (with multiple isomers) within a NOM spectrum; higher concentrations were required to compensate for lower isomeric diversity."

We clarified the relative intensity scale in the figure caption. Specifically, we explain that 'relative intensity' means 'scaled to the tallest feature in the spectrum'. *i.e.,* the base peak

is at 100%, and other peaks are relative to that. This is why the sum of relative intensities exceeds 100%. Further, the panels C,D,E are all on the same scale as each other - which highlights the signal suppression from using a BondElut matrix rather than pure methanol alone.

There are also several strong statements that I do not believe are sufficiently supported by the scientific evidence presented in the paper. These are on line 245 “Strategies to use calibration curves will fail” and line 324 “The previous sections show that between-peak changes in peak intensity do not accurately reflect between-peak changes in abundance”.

As part of our thorough revision to the tone of the paper, we revised these statements to be more positive and forward looking. With respect to line 245 in the original manuscript, we edited the end of section 3.2 to read: “Combining the empirical results from this subsection and the previous subsection with instrument theory discussed above suggests significant uncertainty in relationships between true concentrations and peak intensities from direct infusion FTICR-MS. Calibration curves can be used in the simplest of situations, but may be challenging when there are structural isomers and sample-to-sample variation in matrix composition. Modeling of constrained systems may, however, allow for data-driven and mechanistic data normalization strategies for enhanced use of peak intensity data.”

With respect to line 324 in the original manuscript we edited the start of section 5 to read: “The previous sections highlight challenges in connecting between-peak changes in peak intensity to between-peak changes in abundance (Fig. 4).”

We also note that Reviewer 1 appears to agree with our interpretations and feels that the data shown are too basic to warrant publication. They suggest removing an entire section of the manuscript because it is ‘textbook.’ Reviewer 2 interprets the data differently, however, which indicates there is variation in how researchers view and use peak intensity data from direct infusion FTMS. This emphasizes the importance of retaining all sections of the manuscript and more generally to continue publishing the type of empirical studies we provide to further the literature-based discussion.

Fig. 4 seems to show that the relative intensity scales with concentration. The authors can predict this by a nonlinear model or GAM for each analyte, or with a single model where the slope varies with the m/z (for example). Without having attempted such an analysis it is difficult to understand how this statement is supported.

Figure 4 does show that for an individual, known compound, matrix-matched, dilution ladder, concentration is proportional to signal intensity. This is indeed how quantitative mass spectrometry approaches work (i.e., through calibration curves). However, those calibration curves are molecule-specific and dependent on matrix effects. This is evidenced by the changed behaviors in Fig. 4D and 4E, relative to 4C. Thus, it is not possible to generalize such calibration curves when considering NOM samples

containing thousands of different unknown molecules with unknown isomeric complexity. Fortunately, our simulations show that it may not be necessary to model or otherwise explicitly derive quantitative relationships between intensity and concentration because the ecological metrics perform well without that information.

Figure 4 also nicely shows that you can reach quantitative assumptions between-peaks. As shown in Figure 4a, at low concentrations of the three different molecules (<100 ppb), the signal intensities seem statistically indistinguishable. A similar result is seen, especially in the MeOH Matrix, at higher concentrations of SRFA that effectively dilute the analytes to representative concentrations. These results suggest that the analytes are performing quantitatively.

While this is an interesting observation by the reviewer, we do not feel it is generally applicable in complex mixture data. Our data show that there is inconsistency in how well differences in peak intensity map to differences in concentration. It may be that there is a clearer link at low concentration, but in natural samples that vary widely in complexity, freshness, etc., the true concentrations of each peak are unknown. Put another way, one could not tell if a low-signal ion is due to a high-concentration analyte with poor ionization efficiency, or due to a low concentration analyte with high ionization efficiency, nor if a high-signal ion consisted of a single high concentration analyte or many low-concentration isomeric analytes. Thus, we have no way of knowing which peaks have intensities that carry valid information and those that do not. As such, we cannot generally recommend using between-peak differences in intensity to infer between-peak shifts in concentration.

We added text to section 3.1 to help address this point, which reads: "We note that absolute differences in signal intensity may be smaller between molecules at lower concentrations, but this does not necessarily mean that low intensity signals consistently indicate low concentrations and this does not aid in quantitatively interpreting higher intensity signals."

More generally, we again highlight the contrast between the two reviewers on these topics. Reviewer 1 feels our empirical data show outcomes that are 'textbook' and thus are not needed because it is well-known that peak intensities cannot be used quantitatively. Reviewer 2 feels our data support the quantitative use of peak intensities. Clearly, there are diverging opinions across the research community, highlighting the importance of this section of the paper. We need to come to a resolution on these topics as a research community and that can only happen by presenting, discussing, and improving evidence.

Additional comments:

Line 60: The authors should cite the reference of the first use of 21T FT-ICR-MS (Smith et al., 2018).

The citation has been added.

L195-205 – This point is already made on L122

The text has been streamlined to minimize length and duplication.

Figure 4- relative intensity is not explained. Relative to what?

Text has been edited to clarify this (see comments above).