Responses to Comments on Manuscript amt-2020-267

(Orbitool: A software tool for analyzing online Orbitrap mass spectrometry data)

We thank the reviewer for the comments that help to improve this manuscript. The manuscript was revised according to these comments. We added more details in the revised manuscript. The reviewer's comments are addressed in the following paragraphs. The comments are shown as sans-serif dark red texts and our responses are shown as serif black texts. Changes are highlighted in the revised manuscript and shown as "quoted underlined texts" in the responses. Reference are given at the end of the responses.

Reviewer #4

General comment:

The paper presents Orbitool, a novel online software tool for analyzing Orbitrap mass spectrometry data. The paper is within the scope of AMT but is often too superficial in the description of the working principles of the software and there is no real validation of the output. Some specific comments follow.

Response: We added more details on the working principles in the revised manuscript.

To validate the software, we have shown that the Orbitool is able to read and average raw spectra, fit a certain distribution to the profile of single peaks, conduct mass calibration, distinguish peaks and their corresponding abundances via peak fitting, assign molecular formulae to fitted peaks, etc. We agree with the reviewer's argument that they are not real validation; however, a real validation needs known true values as the benchmark, which is impossible for all the validations. Besides, the data processing methods used in Orbitool are mostly based on basic algebra and/or existing methods for the analysis with CI-ToF data. Therefore, we do not think the results presented in this manuscript lacks validation.

Specific comments:

p.4, l.1: the mass resolving power of the best TOF analyzers is now more than 50,000.

https://www.sciencedirect.com/science/article/pii/S0165993619302018

Response: We agree with the reviewer that TOF with greater mass resolving exists. However, to our knowledge the greatest mass resolving power for an online mass spectrometer is 15 000. TOF analyzers with a mass resolution of 50 000 are not suited to the needs of some applications in atmospheric researches (or have not been tested yet). For instance, to achieve such a mass resolving power, multi reflection technique is needed, which is often associated with poor ion transmission and sensibility. We revised this sentence as: "The mass resolving power of a TOF analyzer typically ranges from hundreds to less than 50 000 and the mass resolving power of online TOF-MS, used in atmospheric measurements, is only up to 15 000."

p.4, l.7-13: this paragraph on source comparison does not belong to this software development article

Response: This paragraph is not a comparison of different ionization sources. Instead, it introduces the potential applications of Orbitrap mass spectrometer in atmospheric researches. The contribution of Orbitool to the research community is affected by the broadness of its potential applications. As a result, it is important to first introduce these potential applications.

p.5, l.20: I do not understand what is meant by "long-term atmospheric measurements". Is it related to the averaging feature?

Response: This phrase was replaced by "long-term field or laboratory campaigns". Data averaging, noise reduction, mass calibration, and exporting time series are closed related to the demands of long-term field campaigns. "Long-term" indicates the amount of data to be analyzed, the importance of calculating and exporting time series over a period (e.g., a week or a month), and the importance of mass accuracy. "Atmospheric measurements" indicates low signal intensities and hence the importance of data averaging and noise reduction. These demands have been discussed in the Introduction section.

p.6, l.10: even if the orbitrap was recording all the data points, this would not generate an equally spaced grid since the size of the bins (i.e. resolution) is variable with the mass

Response: Thanks. This phrase was revised as "a dense grid-based spectrum".

p.6, l.12: the term "proper averaging" needs to be clarified: what are the different methods to make the average?

Response: We removed "proper" and added "The data recorded in the same file is averaged using an averaging function in the Thermo RawFileReader library". We compared this averaging algorithm with RawFileReader, an averaging method from a third-party python package, and averaging methods coded by us. RawFileReader is used for its accuracy and low computational expense. However, we prefer not to present these details in the manuscript to keep it concise.

p.6, l.15: "time bin" - this concept seems important afterwards but it is not defined

Response: We replaced it with "time interval for averaging".

p.6, l.17-18: why making weighted averages?

Response: This sentence was rewritten as "the average is calculated with respect to scan number". The original sentence wanted to emphasize that the average is calculated for scan number, yet the weight of each scan is equal.

p.6, l.19-20: why would the mathematical operation of average involve the resolution of the Orbitrap?

Response: Because two data points needed to be determined as the same one or two different ones before averaging. For instance, there are two data point, (x_1, y_1) and (x_2, y_2), measured at two different time, where x is the mass and y is the signal intensity. Before averaging, the software needs to decide whether $x_1 \approx x_2$. If the y values are stored with predetermined fixed array of x, the criterion for the approximate equality is determined by the round-off error of the used data type and it can be practically written as $x_1 = x_2$. However, the Orbitrap does store data in this way. Even for two adjacent scans, the values of x_1 and x_2 corresponding to the same data point are slightly different and this difference may increase with time. Hence, a criterion is needed to determine whether $|x_1-x_2|$ is no larger than this criterion. If so, the averaged x is the weight average of x_1 and x_2; otherwise, (x_1, y_1) and (x_2, y_2) are saved as two different data points. This criterion is determined by grid of the x array (although Orbitrap data is not save as a continuous array), and the density of the grid is determined by the mass resolution.

p.6, l.19-23: this should be at the beginning of the paragraph; first describe the procedure before going into the details of how it’s done and not the other way around

Response: We agree with the reviewer that the general procedure should be described before introducing the details. However, l.19-23 in the original manuscript are on how data points are identified as the same or different, when the averaging functions are triggered, and advanced features of data averaging. They are the details rather than general procedures. The general features, i.e., configuring data files for averaging, starting and end time, and temporal resolution for averaged data, are introduced at the beginning of this paragraph.

p.7, l.8: the mathematical effect of the average itself causes the noise level to go down (division of a constant level by a higher number of spectra), not the instrumental fact of acquiring more time

Response: We did not say “acquiring more time” or anything similar.

p.7, l.8-9: I am guessing that the mass defect range of [0.5, 0.8] Da was chosen because no signal peaks are expected with such a mass defect in the mass range studied?

Response: Following this sentence, we wrote “Such a mass defect range is chosen because most of the observed compounds in atmospheric measurements are located outside of this mass range.” Since some signals may locate in this range, we select only a proportion, not all, of peaks in this range to estimate the noise threshold.

p.7, l.13: what is "a certain percentile"? I had to wait until p.12, l.16-17 to find the explanation. It should be in the description section, not in the the discussion. See alsop.12, l.5-6

Response: We added “This certain percentile can be customized and it is by default 70%, i.e., the 70% low peaks within this mass defect range are taken as noise peaks and used to estimate the noise level.”

p.7, l.17: here, I think it is necessary to develop why the averaged spectrum would have different noise thresholds because according to me, even if the individual spectra have different noise thresholds, the resulting spectrum will in any case have a single noise level from a statistical point of view. Actually, I don’t understand the difference in the final result between option b and option a

Response: We revised “different amounts of noise” as “different amounts of noise peaks” to avoid confusion. The three strategies are on how to use the noise level instead of how to obtain it.

The difference between options a and b is whether the signals is subtracted by the noise level, as clarified in the manuscript. This subtraction mainly affects the intensity of peaks with intensities slightly above the discriminator level. For instance, when the noise level is 1 and the signal peak height is 4, there is a 25% difference for this peak between options a and b (4 and 3, respectively). When the noise peaks locates sparsely, the measured height of 4 is most likely to be contributed by the signals and option a is recommended. In contrast, when the noise peaks is dense after averaging for a long period, the measured peak height of 4 is most likely to be a combination of the signal peak and the noise. In this case, option b is recommended.

p.8, l.1-3: I understand that the Orbitrap doesn’t generate signals at 0 for the masses it doesn’t detect. However, I don’t see the link between removing noise, i.e. removing data, and the fact that the orbitrap doesn’t populate empty masses.

It seems obvious that removing data (noise) reduces the size of the file and thus the number of points to be processed afterwards

Response: We do not understand this comment because we did not mention Orbitrap here. Removing noise refers to the procedure to remove the noise peaks after reading from file and averaging. Practically, it reduces the length of the array to store the averaged spectrum but not the size of the raw data file. The raw data file is readable only for Orbitool.

p.8, l.9-10: my personal experience is that peaks are not really Gaussian, and I don't seem to see any discussion on why the Gaussian model works in your case. It seems necessary to develop what you say in the caption in the body of the text

Response: Figure 2A shows that the normal distribution is consistent with the measured peak profiles. In the main text, we wrote "As shown in Fig. 2A, the normalized peaks can be well characterized using a normal distribution."

We have not encountered a failure of the normal distribution for our datasets. Other fitting functions can be readily added to the Orbitool, in case there is such a demand in the future.

p.10, l.9-10: this statement needs a reference or at least more explanations

Response: We have revised the sentence as follow: "If needed, the nitrogen rule can be used to help constraining the peak identification (Pellegrin, 1983)."

p.12, l.22: Th here and Da in figures. The same unit should be used throughout the paper

Response: We revised "Th" as "Da" throughout the manuscript.

p.13, l.14-15: I'm having a hard time with this sentence (and this paragraph, actually). It's a software article. A software needs to be validated with a systematic validation protocol with other software/publications and the quantification of the performance. It cannot just be like "oh, that's cool, we are seeing things we've already seen and things we don't know anything about"

Response: We disagree with the reviewer that a software article should exclude application examples. To its contrary, an application example provides a straight forwarding understand of the capability of this software. Besides, the beginning phrase "To illustrate the capabilities of Orbitool…" indicates that this paragraph is not on the validation of the software.

As aforementioned, the key scientific features of Orbitool is tested and reported in this manuscript. The validation of algorithms is either reported if they are not only based on simple algebra. True values are naturally absent for a real validation. The usual validation routines mainly focus on the technical part of software development, e.g., bug and error management, and hence we prefer not to include them in this manuscript. Finally, it is worth mentioning that bugs and error managements are mentioned in the readme file that is available when downloading the software.

p.13, l.21: is it Orbitool or rather the fundamental difference between ToF and Orbitrap data? Because nowhere is there any mention of a data storage strategy that is not based on the fact that Orbitrap data is generated in the right way

Response: The data structure of a spectrum can be easily changed upon reading from the raw data file. For instance, if the number of data points is relatively low after filling the blanks of an Orbitrap spectrum, ToFTools is probably able to analyze Orbitrap data after some modifications. However, due to the much greater mass resolving power of the Orbitrap, such a grid-based data structure will cause an ultra-high computational expense. Note that a different data structure corresponds to different algorithms, classes, functions, and variable, i.e., a different software.

In short, the data structures of Orbitrap raw data and the spectrum class in Orbitool both result from the high resolving power of Orbitrap and there is no causal relationship between the data structures of Orbitrap raw data and Orbitool.

p.14, l.1: I am not convinced. See my previous comment above

Response: We removed "successfully". We agree with the reviewer that this part is an application example instead of a validation.

p.14, l5-7: This sentence does not belong to this article

Response: We removed this sentence.

Figure 5: The mass defect as presented in this figure does not match with the definition of the mass defect in p.7, l.9-10: depending on your definition of the nominal mass should be either in the range of [0, 1] or of [-0.5, 0.5]

Response: We added "The mass defects larger than 0.5 are subtracted by 1.0 so that the domain of mass defect in this figure is [-0.5, 0.5]" in the caption of Fig. 5. The range of vertical axis shown in Fig.5 is [-0.25, 0.25] because no molecular formulae are identified outside this range. This is consistent with the algorithm to estimate the noise level using the mass defect range of [-0.5, -0.2] (i.e., [0.5, 0.8]).