Atmospheric
Measurement
Techniques

Orbitool: a software tool for analyzing online Orbitrap mass spectrometry data

Runlong Cai[1,★], **Yihao Li**[2,3,★], **Yohann Clément**[4], **Dandan Li**[5], **Clément Dubois**[5], **Marlène Fabre**[5], **Laurence Besson**[5], **Sebastien Perrier**[5], **Christian George**[5], **Mikael Ehn**[1], **Cheng Huang**[2], **Ping Yi**[3], **Yingge Ma**[2], and **Matthieu Riva**[5]

[1]Institute for Atmospheric and Earth System Research/Physics, Faculty of Science,
University of Helsinki, Helsinki, 00014, Finland
[2]State Environmental Protection Key Laboratory of Formation and Prevention of Urban Air Pollution Complex, Shanghai
Academy of Environmental Sciences, Shanghai, 200233, China
[3]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
[4]Univ. Lyon, Université Claude Bernard Lyon 1, CNRS, Institut des Sciences Analytiques, UMR 5280,
5 rue de la Doua, 69100 Villeurbanne, France
[5]Univ. Lyon, Université Claude Bernard Lyon 1, CNRS, IRCELYON, 69626, Villeurbanne, France
★These authors contributed equally to this work.

Correspondence: Matthieu Riva (matthieu.riva@ircelyon.univ-lyon1.fr) and Yingge Ma (mayg@saes.sh.cn)

Abstract. The Orbitrap mass spectrometer has recently been proved to be a powerful instrument to accurately measure gas-phase and particle-phase organic compounds with a greater mass resolving power than other widely used online mass spectrometers in atmospheric sciences. We develop an open-source software tool (Orbitool, https://orbitrap.catalyse.cnrs.fr, last access: 4 February 2021) to facilitate the analysis of long-term online Orbitrap data. Orbitool can average long-term data while improving the mass accuracy by re-calibrating each mass spectrum, assign molecular formulae of compounds and their isotopes to measured signals, and export time series and mass defect plots. The noise reduction procedure in Orbitool can separate signal peaks from noise and reduce the computational and storage expenses. Chemical ionization Orbitrap data from laboratory experiments on ozonolysis of monoterpenes and ambient measurements in urban Shanghai were used to test Orbitool. For the test dataset, the average mass accuracy was improved from < 2 to < 0.5 ppm by mass calibrating each spectrum. The denoising procedure removed 97 % of the noise peaks from a spectrum averaged for 30 min while maintaining the signal peaks, substantially helping the automatic assignment of unknown species. To illustrate the capabilities of Orbitool, we used the most challenging and complex dataset we have collected so far, which consists of ambient gas-phase measurements in urban Shanghai. These tests showed that Orbitool was able to automatically assign hundreds of molecular formulae as well as their isotopes with high accuracy.

1 Introduction

Biogenic and anthropogenic sources emit a wide variety of volatile organic compounds (VOCs) into the atmosphere (Hallquist et al., 2009; Shrivastava et al., 2017). Once emitted, VOCs can quickly react with different atmospheric oxidants (OH radicals, O_3, NO_3 radicals, or Cl atoms) yielding a wide variety of oxidized VOCs (OVOCs) spanning a broad range of chemical formulae and, thus, volatilities (Hallquist et al., 2009; Li et al., 2020; Wennberg et al., 2018). OVOCs play a central role in the formation of atmospheric aerosols by either condensing onto pre-existing aerosol particles or forming new particles (Hallquist et al., 2009; Jimenez et al., 2009; Kirkby et al., 2016; Shrivastava et al., 2017). Generally, the more oxidized OVOCs are, the lower their volatility and the greater their probability of partitioning to the particle phase. However, the quantitative evaluation of the impact of aerosols on climate is as yet inadequately constrained

due to many factors, including an incomplete understanding of how VOC oxidation processes contribute to new particle and secondary organic aerosol formation (Glasius and Goldstein, 2016). Indeed, the gas-phase oxidation of one single VOC can yield thousands of oxidized products (Glasius and Goldstein, 2016; Goldstein and Galbally, 2007; Li et al., 2020; Riva et al., 2019b). As a result, the chemical variety of OVOCs poses a major challenge in detecting, quantifying, and characterizing such a large number and wide variety of organic compounds.

Over the last decade, mass spectrometric techniques have made extensive improvements and are now well suited to detecting a large range of species simultaneously. This is highlighted by the key role of chemical ionization mass spectrometry (CIMS) in improving our understanding of atmospheric chemical composition (Breitenlechner et al., 2017; Ehn et al., 2014; Jokinen et al., 2012; Krechmer et al., 2018; Lindinger et al., 1998; Yuan et al., 2017). Chemical ionization is a soft ionization technique where the analytes are ionized through a clustering process with the reagent ions and undergo only minimal fragmentation. While CIMS provides very good sensitivities (i.e., as low as 10^4 molecules cm^{-3}) (Jokinen et al., 2012) and is suitable for measuring a wide variety of gaseous organic and inorganic compounds, it is mainly associated with time-of-flight (TOF) mass analyzers. 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 mass spectrometry (MS) used in atmospheric measurements is only up to 15 000 (Riva et al., 2019b). Although often able to distinguish some isobaric species, these mass resolving powers still limit accurate assignment and quantification of OVOCs in a complex air sample. The existence of multiple overlapping ions yields significant uncertainties (Cubison and Jimenez, 2015; Riva et al., 2019b; Stark et al., 2015). Computational approaches, including ion deconvolution procedures, are required to partly resolve this limitation in order to extract the maximum possible information content (Cubison and Jimenez, 2015; Meija and Caruso, 2004; Zhang et al., 2019, 2020). To overcome these limitations, we have coupled a high-resolution mass spectrometer (Orbitrap; Eliuk and Makarov, 2015) with a chemical ionization source (Riva et al., 2019a, 2020) and an extractive electrospray ionization inlet (Lee et al., 2020) for online analysis. Similarly, Zuth et al. (2018) have combined an atmospheric pressure chemical ionization (APCI) interface with an Orbitrap and have shown that such a technology can achieve much greater mass resolving power at a similar detection limit for the real-time measurement of organic aerosols. As a result, chemical ionization, extractive electrospray ionization, and/or APCI–Orbitrap represent very promising new techniques for online characterization of the chemical compositions of gaseous and particle phases at a very high mass resolving power.

The demand for a new software tool for analyzing Orbitrap data comes along with applying an Orbitrap in long-term continuous online measurements. The commercial software, Xcalibur™, which was used in our initial studies (Lee et al., 2020; Riva et al., 2019a, 2020), provides an interface for basic data analysis, e.g., reading single or averaged scans and exporting the time evolution of selected signal peaks. However, more complicated data analysis is usually needed for investigating the complexity and chemical processes occurring within the atmosphere. Hence, some customized software tools, e.g., RawQuant (Kovalchik et al., 2018), RawTools (Kovalchik et al., 2019), and DIMSpy (Weber and Zhou, 2019) have been developed to meet different demands. In atmospheric sciences, the concentrations of some key chemical species (e.g., peroxy radicals) in a typical atmospheric environment are extremely low (< 1 ppt). The measured spectra must be averaged over a long period to decrease the noise level so that these low signals can be unambiguously identified among noise. This step implies averaging spectra across files while maintaining mass accuracy. Although facilitated by the high mass resolution of the Orbitrap, a minor proportion of the detected signal peaks still inevitably overlap with each other. Therefore, peak fitting is needed to separate the overlapped signal peaks. In addition, using a list (i.e., "peak list") of possible species obtained in similar atmospheric environments reduces the expense of data analysis. The aforementioned features have already been realized in existing software tools for analyzing data acquired with CIMS or proton-transfer-reaction (PTR) TOF MS, e.g., TofTools (Junninen et al., 2010), Tofware, and other software developed to analyze PTR TOF MS datasets (Holzinger, 2015; Müller et al., 2013). It would seem straightforward to convert the Orbitrap raw data into certain formats and then conduct data analysis using these existing software tools. However, due to the high mass resolution of the Orbitrap compared to TOF mass analyzers, the computational expense will be high if analyzing Orbitrap data using, for example, TofTools. This is because the TOF mass spectra are stored on equally spaced grids (with respect to TOF), with data points on the order of 10^4–10^5. For comparison, approximately 10^8 data points are needed to characterize a single Orbitrap spectrum using this grid-based data structure. Instead, saving only the signals with their adjacent zeros will reduce the computational and storing expenses by orders of magnitude, and the Orbitrap raw data are saved using this data structure. In addition, TofTools determines the noise level using the equally spaced data within a certain mass defect range, whereas such information is not recorded in the Orbitrap data. As a result, a new software tool for analyzing Orbitrap data is needed to facilitate its application in atmospheric research.

In this study, we develop a new software tool, named Orbitool, for analyzing online Orbitrap mass spectrometry data. The working principles and features of Orbitool are introduced. Several test examples using both laboratory and atmospheric data are presented to illustrate and validate its features.

2 Orbitool

Orbitool is a software tool with a graphical user interface (GUI) for analyzing online Orbitrap mass spectrometry data. It is optimized for long-term field or laboratory campaigns and laboratory studies. Flexible input and output interfaces facilitate the coupling between Orbitool and other software. Orbitool's basic working principle is shown in Fig. 1, which includes data reading and averaging, noise determination and reduction, peak shape determination, mass calibration, signal abundance determination via peak fitting, molecular formula assignment, and data output. The details of these features will be elaborated in the following sections. Orbitool is mainly coded in Python with support from several third-party libraries (Pedregosa et al., 2011; Goloborodko et al., 2013; Levitsky et al., 2018). To enhance computational speed, the mass spectra are processed using the Numba package (Lam et al., 2015) and chemical composition and signal abundance are determined using the Cython package (Behnel et al., 2011). RawFileReader from Thermo Fisher Scientific is used to read the raw data reported by Orbitrap. Orbitool was tested in Python 3.6 on the Microsoft Windows operating system. An executable version of Orbitool and its open-source codes are publicly available (see "Code availability").

2.1 Data averaging

The data collected by the Orbitrap are stored in a RAW file, containing segmented peak intensities and their corresponding masses of each scan, scan filters, event logs, instrumental information, etc. The RAW file records only the spectra of all the detected peaks rather than a dense grid-based spectrum. Most tuning parameters of Orbitrap, e.g., injection time of each microscan and the number of microscans for every single scan, do not affect the analysis with Orbitool. However, it should be specially clarified that Orbitool is compatible with the profile mode but not the centroid mode of Orbitrap, because attributing chemical formulae to measured signal peaks is based on fitting peaks to the profile data.

The first step to analyze raw mass spectrometry data with Orbitool is reading and averaging the spectra. Since the abundances of most detected species can be extremely low, especially under typical ambient concentrations, averaging is needed for identifying these signal peaks against noise. The raw data can be imported into Orbitool as single files or using a file folder. The user can choose to average all the imported files or a subset of them. The average spectra are calculated based on either scan number or time. For the time-based data averaging, time intervals are determined by the beginning time, the end time, and the duration of each time interval. Orbitool calculates the average spectrum for each time interval. The beginning and end times can be specified by the user, while they are by default obtained from the time range of the input data files. The data recorded in the same file are averaged using an averaging function in the Thermo RawFil-

eReader library, and the average is calculated with respect to scan number. Hence, it is assumed that the parameters determining the duration of every single scan (e.g., injection time and the number of microscans) are kept constant during each time interval. When multiple data files are found in the same time interval for averaging, the average spectra of each file are calculated first and then averaged with respect to time. Peaks close to each other in the averaged spectrum are merged or identified as separated peaks according to the mass resolution of the Orbitrap. To facilitate quick data analysis and minimize unnecessary expenses, averaging is not conducted until either the user or Orbitool calls the averaged spectrum. Further, data averaging can be skipped. In this case, single spectra are used instead of averaged spectra in the following steps. Orbitool also supports filtering data using the ion polarity. For instance, when the positive polarity is selected, only spectra from the positive mode are taken for the averaging while spectra from the negative mode are excluded. Hence, this feature can support the analysis of Orbitrap measurements that switch the ion polarity routinely (e.g., every 1 min) to realize near-real-time measurement of both positively and negatively charged compounds.

2.2 Noise determination and reduction

The noise is estimated and then removed from the average spectra during this step. When converting electronic signals into a spectrum via the Fourier transform, the instrument removes signals below the noise threshold defined by the Orbitrap. Signals exceeding this noise threshold are converted into peaks recorded in the data file. Since this procedure does not remove all the noise signals, some noise peaks are also recorded in the data file. In this study, "noise" refers to these noise peaks. It is difficult to separate the signal peaks for some compounds with low concentrations from noise peaks. However, after averaging for a sufficiently long period, the signal peaks usually exceed noise peaks because averaging reduces the height of noise peaks but not signal peaks. To estimate the noise level, Orbitool first takes all the detected peaks whose peak positions are located in the mass defect range of [0.5, 0.8] Da. The mass defect is the absolute difference between the accurate mass and the nominal mass of the given compound, which is determined by the type and quantity of atoms contained in a compound and its ion polarity. Such a mass defect range is chosen because most of the observed compounds in atmospheric measurements are located outside of this mass range (Zhang et al., 2019). Then, peaks lower than a certain percentile of the peak heights in this mass defect range (0.5–0.8) are taken as noise peaks. 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. The noise and discriminator levels are determined as μ and $\mu + 3\sigma$, respectively, where μ and σ are the mean and standard deviation of these noise peaks, respectively. The
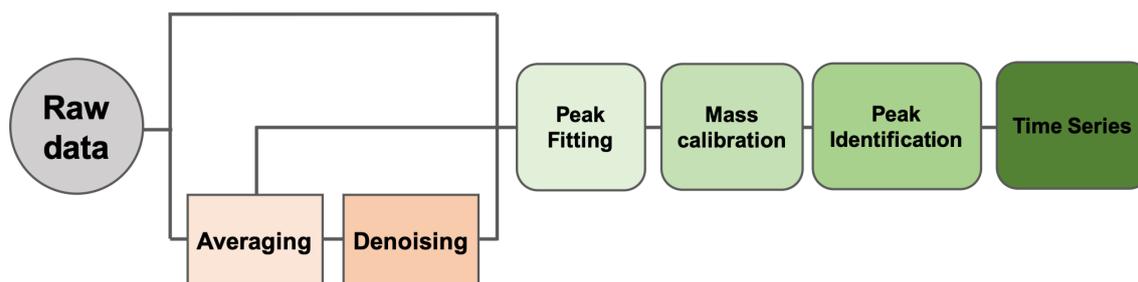
Figure 1. Simplified schematics of the working principle of Orbitool.

noise and discriminator levels as a function of mass can be exported into files for further analysis.

Concerning different quantities of noise peaks in an averaged spectrum, Orbitool provides three strategies to deal with the obtained noise: (a) remove noise peaks below the discriminator level; (b) remove noise peaks below the discriminator level and subtract the noise level (μ) from all signals above the discriminator level; (c) preserve all the noise peaks. Option (a) is the default setting. Option (b) is for averaging over a long period so that the signal peaks are likely to be overlapped by noise peaks, and hence the abundances of signal peaks have contributions by noise. Option (c) is for averaging over a very short period so that it is difficult to distinguish signals from noise. In addition to correcting the impact of noise on the signal abundance, removing noise peaks reduces the expenses for both peak fitting and data storage. This is because each averaged spectrum in Orbitool is stored as a series of peaks rather than the signal intensities at each given mass grid.

2.3 Single-peak-shape determination

Although the mass resolution of the Orbitrap (e.g., 140 000 at mass 200 Da) is considerably higher than those of conventional online mass spectrometers in atmospheric research (Riva et al., 2019a), some peaks (e.g., ^{13}C or ^{18}O isotopes) in an Orbitrap spectrum might still need to be distinguished via fitting multiple peaks to the measured signals. A known single peak shape is necessary for the peak fitting, where a single peak herein refers to the signals contributed by compounds with a single molecular formula. A signal peak hereafter refers to a measured peak, which may be composed of a single ion or multiple different ions. Orbitool uses a normal distribution to characterize the single peak shape (see Fig. 2a). To obtain the standard deviation of the normal distribution at different masses, Orbitool first takes the x-highest signal peaks in the spectrum, where x can be set from 1 to 999 and is 100 by default. The mass resolution of Orbitrap is inversely proportional to $\sqrt{m}$ (Zubarev and Makarov, 2013), where m is mass; hence, the peak width is theoretically proportional to $\sqrt{m}$. According to this relationship, the measured widths of these selected peaks are normalized to widths at 200 Da by dividing them by $\sqrt{m/200\,\mathrm{Da}}$. A reference value of 200 Da is chosen because the instrumental resolution of the Orbitrap is usually reported at this mass. After this normalization, the mass resolution at 200 Da is determined via least-squares fitting. The shapes of some selected signal peaks composed of multiple ions may deviate from others. The user can remove these peaks manually on the GUI before the fitting. As shown in Fig. 2a, the normalized peaks can be well characterized using a normal distribution and the fitted resolution at 200 Da is close to the instrumental mass resolution.

2.4 Mass calibration

Due to the high mass resolving power of the Orbitrap, determining the mass accurately is critical to assigning the molecular formula corresponding to a measured peak. The Orbitrap is able to maintain its mass accuracy (i.e., < 2 ppm) for long time series with the help of constantly existing peaks with relatively high abundances, named lock masses (Olsen et al., 2005). Despite this feature, additional mass calibration might be needed, especially for long-term measurements, to ensure an optimal assignment of molecular formulae across the entire mass range. The user can also skip the mass calibration.

During the mass calibration step, Orbitool can automatically find the peaks of user-specified calibration species within a given uncertainty range. The mass positions of these calibration species are shown on the top of the averaged spectrum on the GUI to guide the eyes. Then, a calibration curve is fitted using a polynomial function and shown on the GUI. The users can set the degree of the polynomial and adjust the calibration curve by using different calibration species. The spectra saved in the same raw data file share the same calibration parameters. The difference between the measured mass and the calibration mass inferred from the chemical formula is shown on the GUI. If the analyzed dataset contains more than one raw file, the trends in these differences for all the calibration species are displayed (Fig. 2b). The calibration information can also be exported for further analysis. Using lock mass(es) during Orbitrap measurements can improve mass accuracy, but it requires a constantly existing compound in the measured spectra. When charging compounds by chemical ionization, the reagent ion and its dimer are usually in high abundance. Therefore, it is recommended to use
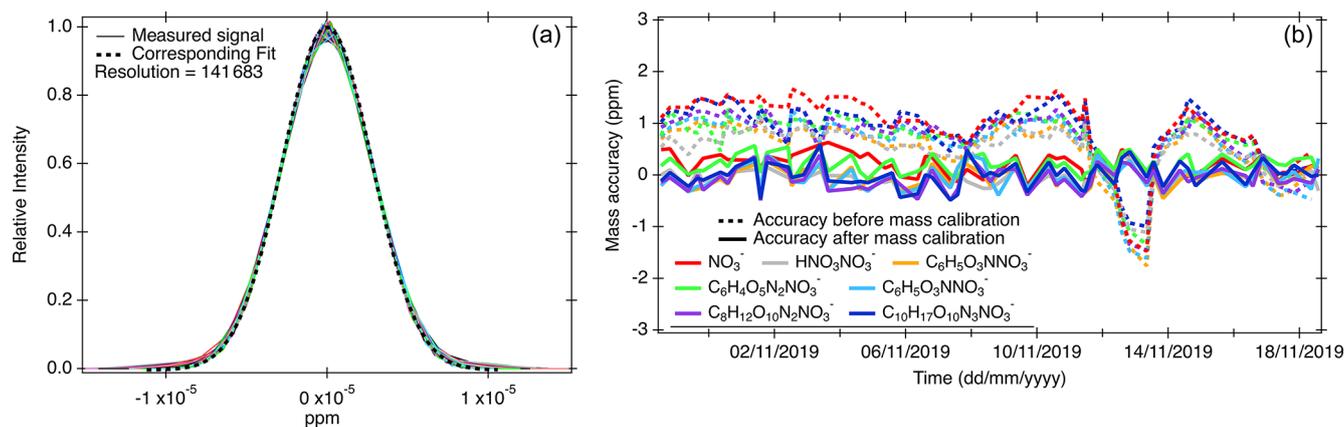
Figure 2. (a) Peak shape determined using the highest 100 peaks observed by the mass spectrometer. The position and width of signal peaks were normalized to 200 Da. The intensity of each signal peak was normalized by dividing by its peak intensity. **(b)** Mass accuracy of product ions at mass (in Da) 62 (red), 125 (grey), 183 (orange), 201 (green), 246 (blue), 358 (purple), and 401 (dark blue) before and after mass calibration. Each data point (as indicated by the fluctuation of the lines) corresponds to a raw data file.

the reagent ions as lock masses when using a CI-Orbitrap. Instead of excluding the lock masses from the spectra (i.e., the default setting in Xcalibur™), Orbitool treats the lock masses as normal signals since the abundance of the reagent ion is usually a critical parameter for characterizing the performance of a CI mass spectrometer.

2.5 Peak assignment

Identifying the abundances and assigning molecular formulae of compounds contained in each signal peak are the main advantages of Orbitool. Before peak assignment, the user can merge the peaks in a given mass range using a customized mass tolerance. A merge sort algorithm (Knuth, 1998) is used to improve computational efficiency. The first step of peak assignment is to fit single peaks to the calibrated averaged spectrum. Since the width of a certain single peak is determined by its position (mass), the fitted parameters are only peak position and peak area (which can be converted to peak height). Orbitool usually conducts a default peak fitting to the whole spectra. The number of single peaks used to fit a signal peak is determined automatically according to the number of local maxima of intensities. After a default fitting, the user can look into each signal peak. The signal peak, fitted single peaks, and fitting residues are shown on the GUI to indicate the quality of the default fitting. Orbitool supports repeating the peak fitting for each signal peak with a user-determined single-ion number, and the updated fitting results for the signal peak will replace the default fitting results in the whole spectra.

The possible molecular formulae corresponding to each fitted single peak are retrieved during fitting. Prior to assigning the molecular formulae from a given mass, the user can customize the possible elements contained in the detected compounds and the ion polarity. The minimum and maxi-

mum numbers of atoms contained in each measured ion, the range of equivalent double-bond numbers, and the maximum tolerance of mass precision are used as boundary conditions. Orbitool searches for all possible chemical formulae satisfying these boundary conditions for a given mass. If needed, the nitrogen rule can be used to help in constraining the peak identification (Pellegrin, 1983). An empty string is returned if no possible chemical formula is found under the given conditions. The user can edit the retrieved chemical formulae after this automatic screening. A separate window on the GUI also supports the conversion of chemical formulae to mass at any analyzing step and vice versa.

Orbitool also supports the identification of isotopes. The users can adjust the possible isotopes in an isotope list, and all the isotopes in this list will be considered. Using the default algorithm for peak assignment, the molecular formula containing a less abundant isotope can be determined only when its corresponding formula containing the most abundant isotope is found in the spectrum. In addition, this default algorithm also checks whether the abundance ratio of these two species is consistent with their natural abundances within the uncertainty range. To facilitate studies such as isotope-labeling experiments, Orbitool also provides an algorithm that does not restrict isotope abundances during peak assignment. However, it is important to mention that the Orbitrap provides non-linear responses when the concentrations of the analytes are very low (i.e., $< 1 \times 10^6$ molecules cm^{-3}, at a 10 min integration time). As a result, the calculated isotope abundance may be substantially smaller than its true value (Riva et al., 2020).

After assigning molecular formulae to all the single peaks, a mass list is generated from the fitted peak positions and their corresponding molecular formulae. The user may add all the molecular formulae into the mass list or select only a proportion of them. The unidentified chemical species are

shown with empty strings. The mass list can be exported and saved into a CSV file. Alternatively, an existing mass list with a certain format can be imported into Orbitool. The new mass list imported from a file or retrieved from a new mass spectrum can either substitute the existing list or merge with it. Orbitool supports fitting the signal peaks using a mass list instead of fitting all the signal peaks, which reduces the user's workload when analyzing the chemical species in a known system. For instance, when measuring ambient air, the user can average data over a whole day to obtain a mass list. After this, the user may restart data analysis from the beginning and average the raw data with a higher temporal resolution (e.g., 30 min). Then, the user may fit the signals using the former mass list derived from the whole day instead of repeating the default fitting and a manual correction.

2.6 Time series and mass defect

The time series (temporal evolution of abundance) of given chemical species can be calculated based on the measured time-evolving spectra. The chemical species for the time series are usually chosen from the mass list. Due to the unavoidable random error in the measured mass even after mass calibration, Orbitool searches for these chosen peaks within a customized uncertainty range (e.g., 1 or 2 ppm). The time series of all the chosen species are displayed on the GUI, and they can be exported into other file formats. The temporal resolution of these time series is configured in the data-averaging step. To facilitate a quick analysis, Orbitool can also show the time series of a given compound instead of using a mass list. The input for specifying this compound is either its accurate mass (or its molecular formula which will be automatically converted into accurate mass) with a mass tolerance or a mass range. Using this feature, the user can also obtain the abundance of all the signals within a wide mass range. Orbitool is also able to show the mass defect plot on the GUI for intuitive and rapid characterization of an averaged spectrum. The size of data points is determined by their intensity, which can be tuned on the GUI. These data points are colored by the number of a chosen element (e.g., C, O, N) or the double-bond equivalent. The peaks without assigned molecular formulae can be either omitted or displayed in grey.

3 Results and discussion

As illustrated in Sect. 2.2, Orbitool estimates the noise level and discriminator level based on the statistics of the measured spectra. Except for the mass defect range of [0.5, 0.8] Da, no prior information is used for determining the noise level. This indicates that one cannot unambiguously distinguish every signal peak from the noise, which may theoretically introduce uncertainties in retrieving molecular formulae and abundances of the measured compounds.

However, such uncertainties are practically negligible after averaging the spectrum. To validate the algorithm used in Orbitool for estimating the noise level, Fig. 3 shows the total number of peaks in each single or averaged spectrum measured in Shanghai. The noise and discriminator levels were first determined using a certain percentile of peaks in the mass defect range of [0.5, 0.8] Da. Peaks below the determined discriminator level in the entire mass range were then taken as noise and removed from the spectrum. Considering that some species may not be detected in a single scan (i.e., present at a very low concentration), the total number of peaks in the averaged spectrum should grow with the increasing averaging time when the period is considerably short, and then it should converge to a certain value when the averaging period is sufficiently long. This is consistent with the observed trend in the total number of peaks after noise reduction in Fig. 3, in which the total number of peaks after noise reduction with the 50th percentile increases slightly with an increasing averaging time. In contrast, the total number of peaks without noise reduction increases dramatically with the increasing averaging time, indicating that most of the peaks observed in the averaged spectrum are noise. In addition, using the 25th or 70th percentile does not lead to a significant difference in the total number of peaks after noise reduction (within 5 %–10 %), which indicates that the noise level is governed by the instrument rather than by the customized percentile value. Accordingly, Orbitool uses the 70th percentile as the default value to estimate the noise and discriminator levels.

Combining data averaging and noise reduction, Orbitool can remove most of the noise while maintaining most of the signal peaks. Figure 4 shows the averaged spectra before and after noise reduction measured from the ozonolysis of monoterpenes in ambient indoor air conditions and ambient air measured in Shanghai. The noise levels were calculated using the 70th percentile of peaks in the mass defect range of [0.5, 0.8] Da. They can also be identified by eye in the mass range from 600 to 750 Da, where only a few signal peaks were detected and noise peaks are at approximately the same height. The noise level in Fig. 4b is lower than that in Fig. 4a because it was averaged for a longer period. Of the measured peaks in the Shanghai dataset after the 30 min averaging, 97 % were identified as noise and hence removed. Since no significant mass dependencies on noise and discriminator levels were observed, Orbitool uses size-independent noise and discriminator levels. It should be clarified that some signal peaks with low abundances may be mistaken as noise peaks during noise reduction. Even after averaging for a longer period, some detected compounds with extremely low abundances may still be below the discriminator level.

To illustrate the capabilities of Orbitool, measured ions depicted in Fig. 4b were characterized. A total of 1850 molecular formulae were identified. Figure 5 shows the corresponding mass defect plot of gaseous compounds measured in ur-
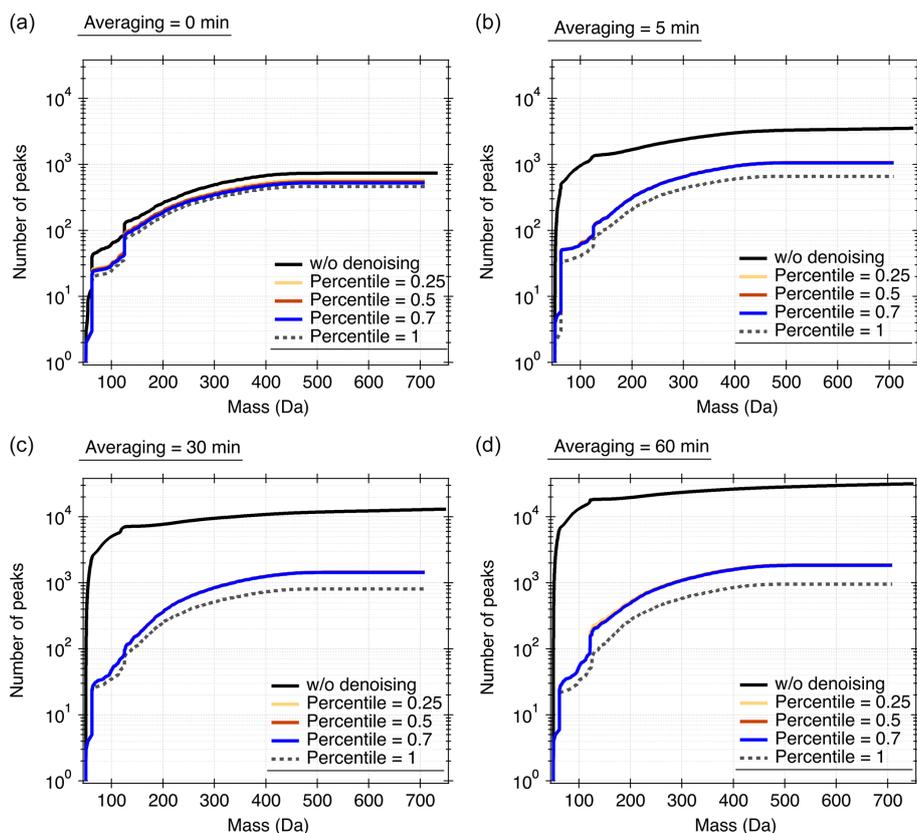
Figure 3. Number of peaks automatically detected in ambient air by Orbitool with and without noise reduction and for different averaging times. When the 100th percentile is used (i.e., the percentile is 1 in the figure), all the peaks in the mass defect range of [0.5, 0.8] were removed.
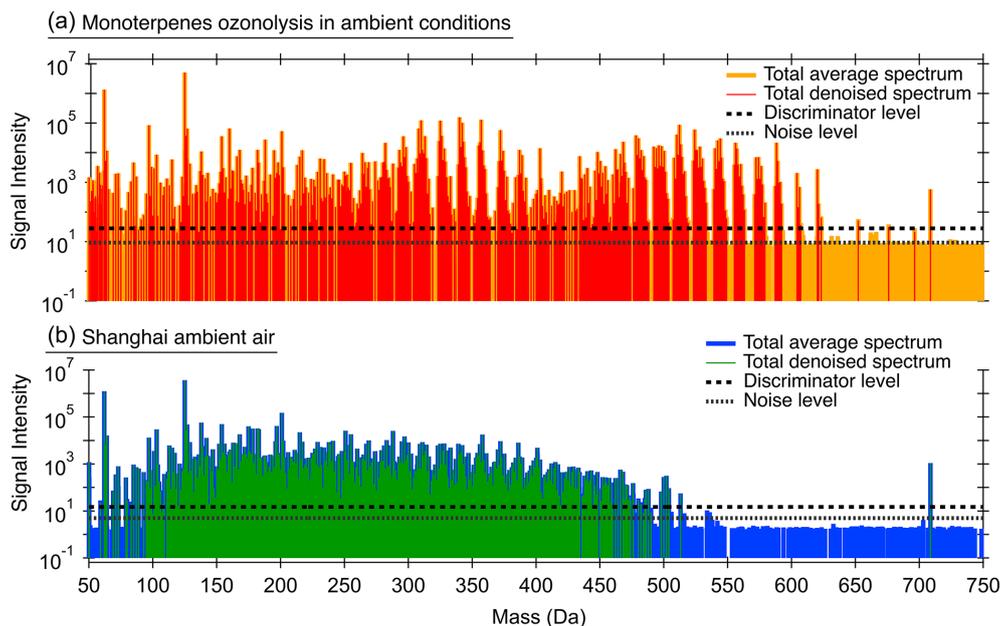
Figure 4. The noise and discriminator levels of mass spectra obtained from the ozonolysis of monoterpenes in **(a)** ambient indoor air conditions (via peeling an orange in front of the Orbitrap) and **(b)** ambient air measured in Shanghai (31 October 2019 between 14:00–14:30 UTC+8). The averaging span for raw data in **(a, b)** was 5 and 30 min, respectively. The 70th percentile was used to estimate the noise and discriminator level.
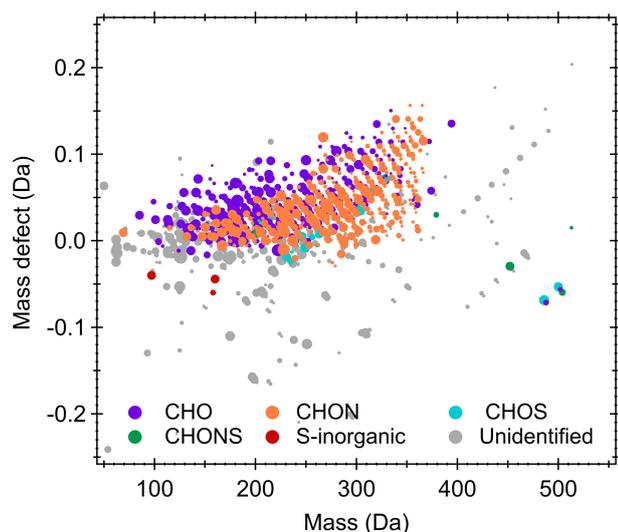
Figure 5. Mass defect of gaseous compounds observed using the CI-Orbitrap in Shanghai (31 October 2019 between 14:00–14:30 UTC+8).

surements. The noise reduction was demonstrated to be able to separate signal peaks from noise and reduce the number of saved data. For instance, the total number of peaks of a test spectrum averaged over 30 min was reduced by 97 %. The mass defect plot retrieved using Orbitool indicates that most of the chemical species in urban Shanghai measured by a CI-Orbitrap were organic compounds and reveals the presence of gaseous compounds with unique compositions (i.e., CHONS, CHOS).

Orbitool is open-source and freely available for download (see "Code availability" below). We hope that it will evolve in the future with added features, optimally as a community effort as Orbitrap mass spectrometers become increasingly utilized as tools for online studies within, and possibly outside, atmospheric sciences.

ban Shanghai, whose data were retrieved and exported by Orbitool. Organic compounds contributed to a major proportion of the measured compounds. The reagent ion (i.e., NO_3^-) is not considered when sorting out the different ions into family groups. Hence the chemical analysis reveals that due to its higher mass resolving power compared to other online mass spectrometers, the CI-Orbitrap identifies a large number of gas-phase signals with elemental composition categories including CHO, CHON, CHONS, and CHOS. In addition, inorganic sulfuric acid was observed, which has been reported to be one of the major gaseous precursors for nucleation in Shanghai (Yao et al., 2018). Interestingly, the presence of highly oxygenated CHONS and CHOS species might indicate the existence of uncharted chemical processes.

4 Summary

We developed a software tool named Orbitool for analyzing long-term online Orbitrap mass spectrometry data in atmospheric research. Orbitool is capable of averaging raw data across files while improving the mass accuracy, distinguishing signal peaks against noise peaks, assigning chemical compounds and their isotopes via peak fitting, and exporting time series and mass defect plots. Compared to similar software tools for atmospheric science studies such as TofTools, Orbitool redesigned the storage method for the measured signals so that it reduces the computational and storage expenses. The data measured in a laboratory study on ozonolysis of monoterpenes and in the ambient air of Shanghai were used to test Orbitool. The mass axis uncertainty after calibration using Orbitool was < 0.5 ppm in this test, which is lower than the average uncertainty of 2 ppm calibrated during mea-

and the French national program LEFE (Les Enveloppes Fluides et l'Environnement).

Review statement. This paper was edited by Pierre Herckes and reviewed by four anonymous referees.

References

Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., and Smith, K.: Cython: The Best of Both Worlds, Comput. Sci. Eng. 13, 31–39, https://doi.org/10.1109/MCSE.2010.118, 2011.

Breitenlechner, M., Fischer, L., Hainer, M., Heinritzi, M., Curtius, J., and Hansel, A.: PTR3: An Instrument for Studying the Lifecycle of Reactive Organic Carbon in the Atmosphere, Anal. Chem., 89, 5824–5831, https://doi.org/10.1021/acs.analchem.6b05110, 2017.

Cubison, M. J. and Jimenez, J. L.: Statistical precision of the intensities retrieved from constrained fitting of overlapping peaks in high-resolution mass spectra, Atmos. Meas. Tech., 8, 2333–2345, https://doi.org/10.5194/amt-8-2333-2015, 2015.

Ehn, M., Thornton, J. A., Kleist, E., Sipilä, M., Junninen, H., Pullinen, I., Springer, M., Rubach, F., Tillmann, R., Lee, B., Lopez-Hilfiker, F., Andres, S., Acir, I.-H., Rissanen, M., Jokinen, T., Schobesberger, S., Kangasluoma, J., Kontkanen, J., Nieminen, T., Kurtén, T., Nielsen, L. B., Jørgensen, S., Kjaergaard, H. G., Canagaratna, M., Maso, M. D., Berndt, T., Petäjä, T., Wahner, A., Kerminen, V.-M., Kulmala, M., Worsnop, D. R., Wildt, J., and Mentel, T. F.: A large source of low-volatility secondary organic aerosol, Nature, 506, 476–479, https://doi.org/10.1038/nature13032, 2014.

Eliuk, S. and Makarov, A.: Evolution of Orbitrap Mass Spectrometry Instrumentation, Annu. Rev. Anal. Chem., 8, 61–80, https://doi.org/10.1146/annurev-anchem-071114-040325, 2015.

Glasius, M. and Goldstein, A. H.: Recent Discoveries and Future Challenges in Atmospheric Organic Chemistry, Environ. Sci. Technol., 50, 2754–2764, https://doi.org/10.1021/acs.est.5b05105, 2016.

Goldstein, A. H. and Galbally, I. E.: Known and Unexplored Organic Constituents in the Earth's Atmosphere, Environ. Sci. Technol., 41, 1514–1521, https://doi.org/10.1021/es072476p, 2007.

Goloborodko, A. A., Levitsky, L. I., Ivanov, M. V., and Gorshkov, M. V.: Pyteomics – a Python framework for exploratory data analysis and rapid software prototyping in proteomics, J. Am. Soc. Mass Spectr., 24, 301–304, https://doi.org/10.1007/s13361-012-0516-6, 2013.

Hallquist, M., Wenger, J. C., Baltensperger, U., Rudich, Y., Simpson, D., Claeys, M., Dommen, J., Donahue, N. M., George, C., Goldstein, A. H., Hamilton, J. F., Herrmann, H., Hoffmann, T., Iinuma, Y., Jang, M., Jenkin, M. E., Jimenez, J. L., Kiendler-Scharr, A., Maenhaut, W., McFiggans, G., Mentel, Th. F., Monod, A., Prévôt, A. S. H., Seinfeld, J. H., Surratt, J. D., Szmigielski, R., and Wildt, J.: The formation, properties and impact of secondary organic aerosol: current and emerging issues, Atmos. Chem. Phys., 9, 5155–5236, https://doi.org/10.5194/acp-9-5155-2009, 2009.

Holzinger, R.: PTRwid: A new widget tool for processing PTR-TOF-MS data, Atmos. Meas. Tech., 8, 3903–3922, https://doi.org/10.5194/amt-8-3903-2015, 2015.

Jimenez, J. L., Canagaratna, M. R., Donahue, N. M., Prevot, A. S. H., Zhang, Q., Kroll, J. H., DeCarlo, P. F., Allan, J. D., Coe, H., Ng, N. L., Aiken, A. C., Docherty, K. S., Ulbrich, I. M., Grieshop, A. P., Robinson, A. L., Duplissy, J., Smith, J. D., Wilson, K. R., Lanz, V. A., Hueglin, C., Sun, Y. L., Tian, J., Laaksonen, A., Raatikainen, T., Rautiainen, J., Vaattovaara, P., Ehn, M., Kulmala, M., Tomlinson, J. M., Collins, D. R., Cubison, M. J., E., Dunlea, J., Huffman, J. A., Onasch, T. B., Alfarra, M. R., Williams, P. I., Bower, K., Kondo, Y., Schneider, J., Drewnick, F., Borrmann, S., Weimer, S., Demerjian, K., Salcedo, D., Cottrell, L., Griffin, R., Takami, A., Miyoshi, T., Hatakeyama, S., Shimono, A., Sun, J. Y., Zhang, Y. M., Dzepina, K., Kimmel, J. R., Sueper, D., Jayne, J. T., Herndon, S. C., Trimborn, A. M., Williams, L. R., Wood, E. C., Middlebrook, A. M., Kolb, C. E., Baltensperger, U., and Worsnop, D. R.: Evolution of Organic Aerosols in the Atmosphere, Science, 326, 1525–1529, https://doi.org/10.1126/science.1180353, 2009.

Jokinen, T., Sipilä, M., Junninen, H., Ehn, M., Lönn, G., Hakala, J., Petäjä, T., Mauldin III, R. L., Kulmala, M., and Worsnop, D. R.: Atmospheric sulphuric acid and neutral cluster measurements using CI-APi-TOF, Atmos. Chem. Phys., 12, 4117–4125, https://doi.org/10.5194/acp-12-4117-2012, 2012.

Junninen, H., Ehn, M., Petäjä, T., Luosujärvi, L., Kotiaho, T., Kostiainen, R., Rohner, U., Gonin, M., Fuhrer, K., Kulmala, M., and Worsnop, D. R.: A high-resolution mass spectrometer to measure atmospheric ion composition, Atmos. Meas. Tech., 3, 1039–1053, https://doi.org/10.5194/amt-3-1039-2010, 2010.

Kirkby, J., Duplissy, J., Sengupta, K., Frege, C., Gordon, H., Williamson, C., Heinritzi, M., Simon, M., Yan, C., Almeida, J., Tröstl, J., Nieminen, T., Ortega, I. K., Wagner, R., Adamov, A., Amorim, A., Bernhammer, A.-K., Bianchi, F., Breitenlechner, M., Brilke, S., Chen, X., Craven, J., Dias, A., Ehrhart, S., Flagan, R. C., Franchin, A., Fuchs, C., Guida, R., Hakala, J., Hoyle, C. R., Jokinen, T., Junninen, H., Kangasluoma, J., Kim, J., Krapf, M., Kürten, A., Laaksonen, A., Lehtipalo, K., Makhmutov, V., Mathot, S., Molteni, U., Onnela, A., Peräkylä, O., Piel, F., Petäjä, T., Praplan, A. P., Pringle, K., Rap, A., Richards, N. A. D., Riipinen, I., Rissanen, M. P., Rondo, L., Sarnela, N., Schobesberger, S., Scott, C. E., Seinfeld, J. H., Sipilä, M., Steiner, G., Stozhkov, Y., Stratmann, F., Tomé, A., Virtanen, A., Vogel, A. L., Wagner, A. C., Wagner, P. E., Weingartner, E., Wimmer, D., Winkler, P. M., Ye, P., Zhang, X., Hansel, A., Dommen, J., Donahue, N. M., Worsnop, D. R., Baltensperger, U., Kulmala, M., Carslaw, K. S., and Curtius, J.: Ion-induced nucleation of pure biogenic particles, Nature, 533, 521–526, https://doi.org/10.1038/nature17953, 2016.

Kovalchik, K. A., Moggridge, S., Chen, D. D. Y., Morin, G. B., and Hughes, C. S.: Parsing and Quantification of Raw Orbitrap Mass Spectrometer Data Using RawQuant, J. Proteome Res., 17, 2237–2247, https://doi.org/10.1021/acs.jproteome.8b00072, 2018.

Kovalchik, K. A., Colborne, S., Spencer, S. E., Sorensen, P. H., Chen, D. D. Y., Morin, G. B., and Hughes, C. S.: RawTools: Rapid and Dynamic Interrogation of Orbitrap Data Files for Mass Spectrometer System Management, J. Proteome Res., 18, 700–708, https://doi.org/10.1021/acs.jproteome.8b00721, 2019.

Krechmer, J., Lopez-Hilfiker, F., Koss, A., Hutterli, M., Stoermer, C., Deming, B., Kimmel, J., Warneke, C., Holzinger, R., Jayne, J. T., Worsnop, D. R., Fuhrer, K., Gonin, M., and de Gouw, J. A.: Evaluation of a New Reagent-Ion Source and Focusing Ion-Molecule Reactor for use in Proton-Transfer-Reaction Mass Spectrometry, Anal. Chem., 90, 12011–12018, https://doi.org/10.1021/acs.analchem.8b02641, 2018.

Knuth, D.: Sorting by Merging, Sect. 5.2.4, in: The Art of Computer Programming 3: Sorting and Searching, 2nd edn., Addison-Wesley, Massachusetts, 1998.

Lam, S. K., Pitrou, A., and Seibert, S.: Numba: A LLVM-Based Python JIT Compiler, in: Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC (LLVM '15), Association for Computing Machinery, New York, NY, USA, Article 7, 1–6, https://doi.org/10.1145/2833157.2833162, 2015.

Lee, C. P., Riva, M., Wang, D., Tomaz, S., Li, D., Perrier, S., Slowik, J. G., Bourgain, F., Schmale, J., Prevot, A. S. H., Baltensperger, U., George, C., and El Haddad, I.: Online Aerosol Chemical Characterization by Extractive Electrospray Ionization–Ultrahigh-Resolution Mass Spectrometry (EESI-Orbitrap), Environ. Sci. Technol., 54, 3871–3880, https://doi.org/10.1021/acs.est.9b07090, 2020.

Levitsky, L. I., Klein, J., Ivanov, M. V., and Gorshkov, M. V.: Pyteomics 4.0: five years of development of a Python proteomics framework, J. Proteome Res., 18, 709–714, https://doi.org/10.1021/acs.jproteome.8b00717, 2018.

Li, H., Riva, M., Rantala, P., Heikkinen, L., Daellenbach, K., Krechmer, J. E., Flaud, P.-M., Worsnop, D., Kulmala, M., Villenave, E., Perraudin, E., Ehn, M., and Bianchi, F.: Terpenes and their oxidation products in the French Landes forest: insights from Vocus PTR-TOF measurements, Atmos. Chem. Phys., 20, 1941–1959, https://doi.org/10.5194/acp-20-1941-2020, 2020.

Lindinger, W., Hansel, A., and Jordan, A.: On-line monitoring of volatile organic compounds at pptv levels by means of proton-transfer-reaction mass spectrometry (PTR-MS) medical applications, food control and environmental research, Int. J. Mass Spectrom., 173, 191–241, https://doi.org/10.1016/S0168-1176(97)00281-4, 1998.

Meija, J. and Caruso, J. A.: Deconvolution of isobaric interferences in mass spectra, J. Am. Soc. Mass Spectrom., 15, 654–658, https://doi.org/10.1016/j.jasms.2003.12.016, 2004.

Müller, M., Mikoviny, T., Jud, W., D'Anna, B., and Wisthaler, A.: A new software tool for the analysis of high resolution PTR-TOF mass spectra, Chemometr. Intell. Lab., 127, 158–165, https://doi.org/10.1016/j.chemolab.2013.06.011, 2013.

Olsen, J. V., de Godoy, L. M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., and Mann, M.: Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap, Mol. Cell. Proteomics, 4, 2010–2021, https://doi.org/10.1074/mcp.T500030-MCP200, 2005.

Orbitool: Analysis of online Orbitrap mass spectrometric data, available at: https://orbitrap.catalyse.cnrs.fr, last access: 4 February 2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res., 12, 2825–2830, 2011.

Pellegrin, V.: Molecular formulas of organic compounds: the nitrogen rule and degree of unsaturation, J. Chem. Educ., 60, 626, https://doi.org/10.1021/ed060p626, 1983.

Riva, M., Ehn, M., Li, D., Tomaz, S., Bourgain, F., Perrier, S., and George, C.: CI-Orbitrap: An Analytical Instrument To Study Atmospheric Reactive Organic Species, Anal. Chem., 91, 9419–9423, https://doi.org/10.1021/acs.analchem.9b02093, 2019a.

Riva, M., Rantala, P., Krechmer, J. E., Peräkylä, O., Zhang, Y., Heikkinen, L., Garmash, O., Yan, C., Kulmala, M., Worsnop, D., and Ehn, M.: Evaluating the performance of five different chemical ionization techniques for detecting gaseous oxygenated organic species, Atmos. Meas. Tech., 12, 2403–2421, https://doi.org/10.5194/amt-12-2403-2019, 2019b.

Riva, M., Brüggemann, M., Li, D., Perrier, S., George, C., Herrmann, H. and Berndt, T.: The capability of CI-Orbitrap for gas-phase analysis in atmospheric chemistry: A comparison with the CI-APi-TOF technique, Anal. Chem., 92, 8142–8150, https://doi.org/10.1021/acs.analchem.0c00111, 2020.

Shrivastava, M., Cappa, C. D., Fan, J., Goldstein, A. H., Guenther, A. B., Jimenez, J. L., Kuang, C., Laskin, A., Martin, S. T., Ng, N. L., Petaja, T., Pierce, J. R., Rasch, P. J., Roldin, P., Seinfeld, J. H., Shilling, J., Smith, J. N., Thornton, J. A., Volkamer, R., Wang, J., Worsnop, D. R., Zaveri, R. A., Zelenyuk, A., and Zhang, Q.: Recent advances in understanding secondary organic aerosol: Implications for global climate forcing: Advances in Secondary Organic Aerosol, Rev. Geophys., 55, 509–559, https://doi.org/10.1002/2016RG000540, 2017.

Stark, H., Yatavelli, R. L. N., Thompson, S. L., Kimmel, J. R., Cubison, M. J., Chhabra, P. S., Canagaratna, M. R., Jayne, J. T., Worsnop, D. R., and Jimenez, J. L.: Methods to extract molecular and bulk chemical information from series of complex mass spectra with limited mass resolution, Int. J. Mass Spectrom., 389, 26–38, https://doi.org/10.1016/j.ijms.2015.08.011, 2015.

Weber, R. J. M. and Zhou, J.: DIMSpy: Python package for processing direct-infusion mass spectrometry-based metabolomics and lipidomics data, Version v1.4.0, Zenodo, https://doi.org/10.5281/zenodo.3764110, 2019.

Wennberg, P. O., Bates, K. H., Crounse, J. D., Dodson, L. G., McVay, R. C., Mertens, L. A., Nguyen, T. B., Praske, E., Schwantes, R. H., Smarte, M. D., St Clair, J. M., Teng, A. P., Zhang, X., and Seinfeld, J. H.: Gas-Phase Reactions of Isoprene and Its Major Oxidation Products, Chem. Rev., 118, 3337–3390, https://doi.org/10.1021/acs.chemrev.7b00439, 2018.

Yao, L., Garmash, O., Bianchi, F., Zheng, J., Yan, C., Kontkanen, J., Junninen, H., Mazon, S. B., Ehn, M., Paasonen, P., Sipilä, M., Wang, M., Wang, X., Xiao, S., Chen, H., Lu, Y., Zhang, B., Wang, D., Fu, Q., Geng, F., Li, L., Wang, H., Qiao, L., Yang, X., Chen, J., Kerminen, V.-M., Petäjä, T., Worsnop, D. R., Kulmala, M., and Wang, L.: Atmospheric new particle formation from sulfuric acid and amines in a Chinese megacity, Science, 361, 278–281, https://doi.org/10.1126/science.aao4839, 2018.

Yuan, B., Koss, A. R., Warneke, C., Coggon, M., Sekimoto, K., and de Gouw, J. A.: Proton-Transfer-Reaction Mass Spectrometry: Applications in Atmospheric Sciences, Chem. Rev., 117, 13187–13229, https://doi.org/10.1021/acs.chemrev.7b00325, 2017.

Zhang, Y., Peräkylä, O., Yan, C., Heikkinen, L., Äijälä, M., Daellenbach, K. R., Zha, Q., Riva, M., Garmash, O., Junninen, H., Paatero, P., Worsnop, D., and Ehn, M.: A novel approach for simple statistical analysis of high-resolution mass spectra, At-

mos. Meas. Tech., 12, 3761–3776, https://doi.org/10.5194/amt-12-3761-2019, 2019.

Zhang, Y., Peräkylä, O., Yan, C., Heikkinen, L., Äijälä, M., Daellenbach, K. R., Zha, Q., Riva, M., Garmash, O., Junninen, H., Paatero, P., Worsnop, D., and Ehn, M.: Insights into atmospheric oxidation processes by performing factor analyses on subranges of mass spectra, Atmos. Chem. Phys., 20, 5945–5961, https://doi.org/10.5194/acp-20-5945-2020, 2020.

Zubarev, R. A. and Makarov, A.: Orbitrap Mass Spectrometry, Anal. Chem., 85, 5288–5296, https://doi.org/10.1021/ac4001223, 2013.

Zuth, C., Vogel, A. L., Ockenfeld, S., Huesmann, R., and Hoffmann, T.: Ultrahigh-Resolution Mass Spectrometry in Real Time: Atmospheric Pressure Chemical Ionization Orbitrap Mass Spectrometry of Atmospheric Organic Aerosol, Anal. Chem., 90, 8816–8823, https://doi.org/10.1021/acs.analchem.8b00671, 2018.