Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., author comment AC2
https://doi.org/10.5194/amt-2022-16-AC2, 2022

Reply on RC2

Sungwoo Kim et al.

Author comment on "Comprehensive detection of analytes in large chromatographic datasets by coupling factor analysis with a decision tree" by Sungwoo Kim et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2022-16-AC2, 2022

We thank the reviewers for their careful consideration and recognition of the value of this work. Please find below our responses to all comments and associated changes to the manuscript. Original comments are included in bold, and changes to the manuscript are excerpted in italics.

Reviewer 2: Alexander Vogel

l. 17: Why is "peak width" not included in peak evaluation of the decision tree?

As originally written, this sentence was not clear, using the phrase "peak shape" to represent not only the distinctive pattern of time profile (e.g., Gaussian or Exponentially modified gaussian) but also the peak width, height, and their relative ratios. Line 17 has been revised as *"A decision tree based on peak parameters (e.g., location, width, and height), relative ratios of those parameters, noise, retention time, and mass spectrum is applied to…."*

l. 23 & l. 370: It is mentioned that 90% of the ~1100 "analytes" have no match with the NISTdatabase. Therefore, I suggest to use the term "features" instead of "analytes". This applies to the whole manuscript.

We acknowledge that the word 'analyte' is often used to refer to identified chemical substances, and we appreciate the suggestion to adjust the wording so as to avoid potential confusion for the readers. However, within this community, the term "analyte" has been frequently used to refer to a feature in a sample whether or not it has a known definitive identification (e.g., references from the manuscript: Amigo et al 2010, Isaacman-VanWertz et al 2017, and others in this journal: Li et al., 2022 and Grace et al., 2019). We are of the opinion that the term "analyte" will be less confusing to the broader atmospheric community and readership of this journal. To avoid any confusion, we have clarified our definition in the manuscript as excerpted below. We leave the final decision on this choice to the editorial staff of the journal and are open to making the switch from "analyte" to "feature" if the editor or journal staff feels it would be better suited.

"In this work, the term "analyte" is used to refer to a unique chromatographic peak (e.g., a chromatographic "feature") whether it has a known definitive identification or not, following the usage of this term other studies (Amigo et al., 2010; Grace et al., 2019; Isaacman-Vanwertz et al., 2017; Li et al., 2022)."

Li, H. et al.: Fragmentation inside proton-transfer-reaction-based mass spectrometers limits the detection of ROOR and ROOH peroxides, Atmos. Meas. Tech., 15, 1811–1827, https://doi.org/10.5194/amt-15-1811-2022, 2022.

Grace, D. N. et al.: Separation and detection of aqueous atmospheric aerosol mimics using

supercritical fluid chromatography–mass spectrometry, Atmos. Meas. Tech., 12, 3841–3851, https://doi.org/10.5194/amt-12-3841-2019, 2019.

l. 30-35 or somewhere else in the introduction: Might be worth mentioning 2D-GC approaches to resolve complex samples.
We agree that GCxGC holds promise for tackling some of these issues and warrants inclusion in the introduction. The following sentences have been added to the manuscript. *"The resolution of GC can be expanded by coupling multiple columns in series, and comprehensive two-dimensional gas chromatography (GC×GC) can provide greater sensitivity and resolution of complex mixtures (Bertsch, 1999; Phillips and Beens, 1999). This technique has yielded valuable insights into atmospheric composition (Hamilton, 2010), but the increased complexity of the instrumentation and more stringent requirements for the mass spectrometer (e.g., time resolution faster than ~50 Hz (Worton et al., 2012)) has limited adoption of GC×GC. Furthermore, despite the higher resolving power, co-elution of peaks still occurs (Potgieter et al., 2016) when highly complex samples are analyzed, and challenges remain in the data analysis. Therefore, it is consequently common for analyses of environmental data to focus on the resolution and quantification of only a subset of specific analytes of interest and leave a large fraction of data unprocessed and unused."*

l. 170 ff.: Peak-shapes of chromatograms are important and mostly not a perfect Gaussian. Peak shapes depend on the properties of the compound, but also on the condition of the column. That means, when analysing a few hundred of samples, the peak shape for one compound can become worse over time. How does the algorithm deal with that?
As described in lines 101-110, positive matrix factorization (PMF) extracts a commonly observed pattern from the dataset as a factor which consists of a chromatographic signal and corresponding mass spectrum. The time profile of a factor is a true reflection of the peak shape (Gaussian or otherwise) in each chromatogram. A changing peak shape over time will be reflected in the profiles. Fitting to a Gaussian (or other shape) peak is used simply to catalog or describe the peak in each chromatogram. As peak shape changes from sample to sample, the parameters in the fit would represent that, and if a deteriorating shape begins to deviate from Gaussian, the width of the fit may become less representative of the true width. However, the location, mass spectrum, and approximate height of the peak should be reasonably well documented in the catalog, enabling a user to find and integrate the peak using whatever peak shape best describes the data.

The authors used a Gaussian for chromatographic peak fitting and mention that a more complex approach is not necessary, although possible to include modified peak shapes. I disagree with the statement that non-Gaussian peak models are not necessary for proper peak fitting. An exponentially modified Gaussian (EMG) actually allows evaluating the Gaussian shape of chromatographic peaks by fitting with four variables (area, elution time, peak width and exponential) instead of just area, elution time and peak width (e.g. see Goodman & Brenna 1994). The mentioned paper by IsaacmanVan Wertz et al. (2017) is missing in the references.
Incorporating an EMG model for peak fitting increases the ability of a fitting algorithm to describe real world chromatographic peaks, though with increased possibility of poor fitting of poorly resolved peaks due to the additional degrees of freedom. In this study, the primary goal is catalog peaks, with peak fitting used primarily to extract the parameters to describe the peak (location, height, width); indeed, peak detection is agnostic toward peak shape, relying on first and second derivatives to identify inflection points in the factor profiles. Consequently, we apply only a Gaussian model to lower the level of complexity in the calculation of the parameters describing each peak. However,

we agree that an EMG peak shape may improve the accuracy of these parameters (particularly peak width), and have clarified our reasoning and language in the lines 191-197 and 460-464 of the revised manuscript:
"A more complex approach could include modified peak shapes (e.g., convolution with an exponential (Isaacman-VanWertz et al., 2017)), *which would likely enable more accurate characterization of the parameters describing a peak. However, in this work the goal is to catalog all peaks by their approximate parameters as opposed to perfectly integrate them, so increasing the complexity of peak fitting by incorporating refined peak shapes has not been implemented. Implementation of exponentially modified gaussian (EMG) as a peak fitting model has been inspected on the samples containing deuterated tetradecane presented in Fig.4 and discussed in the supplementary information (Fig. S7). Optimal peak shapes could be used in subsequent processing for accurate integration of data."*
"Three major future developments would further enhance this approach: improved retention time correction without human interaction (e.g., by parametric optimized warping (Eilers, 2004)), incorporation of modified peak shapes (e.g., exponentially modified gaussian) as a peak fitting model, and algorithmic optimization of decisions around the length of each slice, the number of factors, and the number of chromatograms."
A far more expansive investigation of EMG peak shapes has also now been included as discussed in response to the comment below.

How robust is the finding of the three different compounds (C14D30, C14D29H, C14D28H2) of the nine-factor solution when a non-Gaussian peak shape model is employed that allows to fit peak tailing? Since the difference is in the mass spectrum at m/z 226-230, I assume it is robust, but I can be wrong.
We would like to thank the reviewer for raising this discussion, which prompted us to examine more deeply decisions around the retention time difference parameter in the decision tree. We have added a discussion of the reviewer's specific comment and significantly revised discussion surrounding peak shape and width. As described in lines 242-246, the retention time difference and cosine similarity value are the two main criteria used in the separation of peaks in the decision tree. Prompted by this comment, we have considered the retention time difference more deeply; while it is inherently a function of peak width, there is also some flexibility and implications in how the critical retention time difference should be set so we have switched from calling this parameter peak width to "critical retention time difference" and discuss it in more detail.
Though there exists some difference in the mass spectra of the compounds $C_{14}D_{30}$, $C_{14}D_{29}H$, and $C_{14}D_{28}H_2$, but due to heavy fragmentation of alkanes, the mass spectra are sufficiently similar ($\varepsilon > 0.8$) that it is difficult to algorithmically separate with cosine similarity as a sole metric. These peaks are instead being separated primarily by retention time difference. Considering that the differences in the high molecular weight ion distributions, and the predictable rightward shift of the retention times of each isotopologue, we are reasonably confident that these are unique analytes and thus present an opportunity to examine the impacts of how the retention time difference criterion is selected and the potential impact of peak shape. We find, overall, that there is a strong argument for allowing some control by the user over that argument and examine the impacts of this control, which balances a tendency for negative and positive errors.
In the discussion of peak sorting, in which peaks in the same factor are separated into analytes with same mass spectrum, we have added:
"Selection of a critical retention time difference is somewhat dependent on the goals of the user but is inherently related to peak widths. A conservative estimate of a critical width is several times the standard deviation (e.g., FWHM = 2.355σ), which would ensure that only peaks that are truly chromatographically resolved are regarded as unique. However, in many cases, isomers may not be well resolved but nevertheless represent unique analytes, which may be apparent in small changes in ion ratios or signal intensities across chromatograms. In these cases, a more aggressive (i.e., smaller) approach to critical retention time differences may be appropriate, which might include HWHM (~1.18 σ), or,

most aggressively, peaks that are separated by only one or two datapoints (i.e., a peak in a different time period of instrument acquisition). Setting this parameter more aggressively increases the possibility of positive errors, discussed in Section 2.4."
And we note that xylenes are an example of this case:
"Isomers such as these represent an example of the potential impact of a user-specified critical retention time difference, as an aggressive value (e.g., one or two datapoints) may separate these analytes if there is at least some separation by retention time and some variability in ratios between samples that may be detected by the PMF, while a more conservative approach (e.g., FWHM) is unlikely to separate poorly resolved isomers."
We further discuss the selection of the critical retention time difference in the analyte sorting discussion when peaks in separate factors are combined. We note that in this case, there is some measurable difference in spectra and/or sample-to-sample variability, since peaks have been separated by PMF even if cosine similarities are not strictly above the threshold:
"Again, the selection of the critical retention time difference exerts some control on the opposing tendencies of this approach to either consider peaks unique (potentially leaving multiple peaks representing the same analyte) or combine peaks (potentially binning multiple analytes). In this step, any potential analytes being compared must exhibit at least some difference in mass spectrum and sample variability, since they were separated by the PMF, so a more aggressive critical retention time difference is likely warranted here."
With this in mind, we have specified our critical retention time differences throughout the manuscript. Specifically, we note that in our analysis of analytes used for Fig. 3 we use a relatively aggressive value:
"The critical retention time difference used in this analysis was relatively aggressive (median HWHM, which equals 0.7 s in this data) in order to examine the capability of the method to find unique peaks; the effects of this selection are discussion below."
An extensive discussion of the effect of retention time difference on isotopologue separation has also been added in direct response to this comment:
"Separation of these isotopologues presents an opportunity to examine the impact of the critical retention time difference, and the impact of assumed Gaussian peak shapes on this separation. Though exhibiting interpretable differences in their higher molecular weight ions, the heavy fragmentation of alkanes yields mass spectra that are not sufficiently different to be separated by the cosine similarity threshold (i.e., comparisons between all three isotopologues have □≥0.8), despite sufficient differences to be separated into different factors in the PMF. Consequently, resolution of these peaks relies on separation by retention time in the analyte sorting step. Separation between each peak is roughly 0.75 s second in retention time, while median peak width in the dataset (σ) is 0.6 s, peak widths of these analytes are roughly on the order of 0.7 s, and a mass spectrum is collected every 0.3 s. Peaks are consequently separated by more than two datapoints, and more than the median HWHM of the dataset (0.71 s), but not by more than the HWHM of these specific peaks (0.82 s) or by more than the median FWHM of the dataset (1.4 s). In other words, only more aggressive screening methods (i.e., using σ or median HWHM as the critical retention time difference) would separate these isotopologues. This approach also increases the chance of chromatographic artifacts being cataloged as real analytes (positive error), but a more conservative approach increases the possibility of overlooking poorly resolved and similar analytes such as these (negative error). Ultimately, it is up to the user to decide the optimal critical retention time difference."
The effect of a modified peak shape is found to be relatively minor. Because PMF is agnostic to peak shape and so is peak detection (which uses first and second derivatives), the impact of EMG fitting is only on the widths of the peak and the potential for combining them in peak sorting and analyte sorting. So the issue is again one of setting the critical retention time threshold. We find the results to be essentially the same as those achieved using Gaussian fitting, that more aggressive values separate the isotopologues while less aggressive values do not. That has been included in the revised manuscript as well as an associated figure and discussion in the SI:

"The effect of a non-Gaussian peak shape was also examined. Because peak detection relies on derivatives to identify potential peaks based on inflection points in the data, the number of peaks found is agnostic toward peak shape; instead, peak shape primarily impacts peak widths. Using an exponentially modified Gaussian peak shape to the analysis of isotopologues does not substantially change the result (Fig. S7). With this peak shape, isotopologues remain separated using more aggressive critical retention time differences (median FWHM or more than two datapoints) but are combined by more conservative thresholds. This result is of course limited to the shown case, in which a Gaussian curve reasonably describes the observed data. Datasets containing highly non-Gaussian peak shapes may be more impacted and should be examined closely for the potential impact of peak tailing on positive errors."

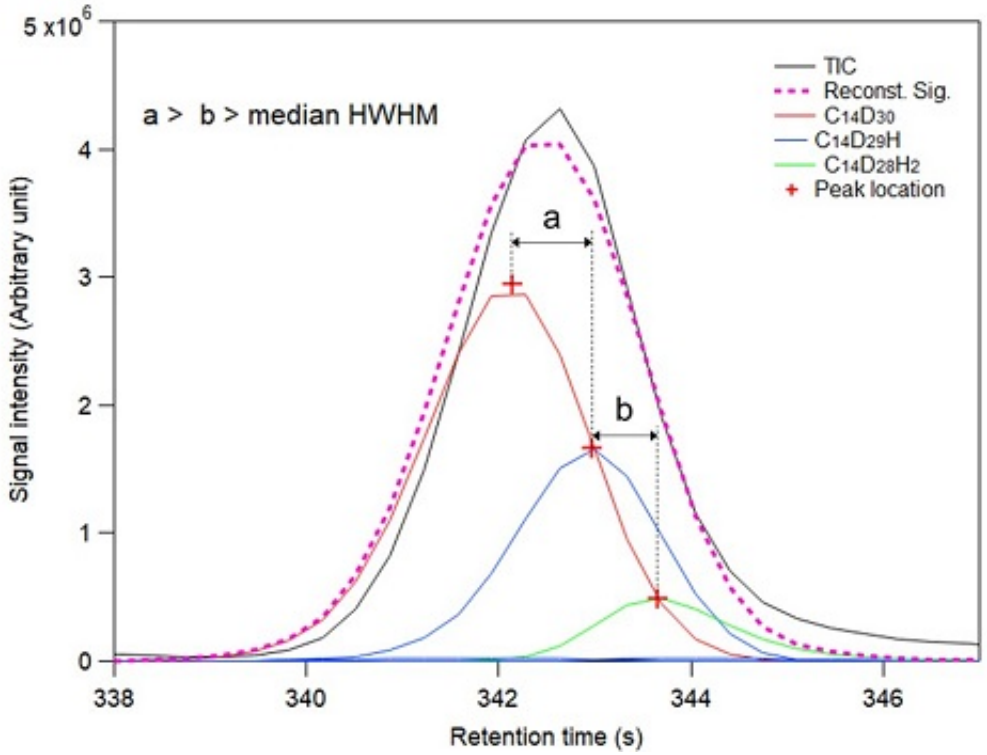

Figure S7. Analysis results of deuterated tetradecane with an exponentially modified gaussian (EMG) as a peak fitting model. EMG curves of all analytes identified overlaid with the original chromatographic signal. The purple dashed line represents the total reconstructed signal of all analytes. Red cross markers represent the location and height of each corresponding analyte. The retention time difference between C14D30 and C14D29H is labelled as 'a' and the latter as 'b'.

With an EMG model, the compounds are separated since the median HWHM is smaller than both a and b. However, it is notable that the HWHM of C14D30 is greater than a. Ideally, each pair of co-eluting peaks is investigated, and a proper width needs to be determined for each pair and used in the decision tree. Programmatically, this would be challenging and thus a representative value, median HWHM, is compared to the retention time differences. Ultimately, it is up to the user to decide what definition of width should be used as an efficient screening tool.

l. 174: The authors should provide evidence that "a refined peak shape is likely unnecessary", or otherwise should argue more carefully.
As described in response to the reviewer's comments above, lines 191-195 have been revised:
"A more complex approach could include modified peak shapes (e.g., convolution with an exponential (Isaacman-VanWertz et al., 2017)), *which would likely enable more accurate*

characterization of the parameters describing a peak. However, in this work the goal is to catalog all peaks by their approximate parameters as opposed to perfectly integrate them, so increasing the complexity of peak fitting by incorporating refined peak shapes has not been implemented. Optimal peak shapes could be used in subsequent processing for accurate integration of data."

l.227: "M1 and M2 are normalized mass spectra of two analytes". This is confusing, because if the result is that epsilon>=0.8, then it is two normalized mass spectra of one analyte. I suggest rephrasing "M1 and M2 are normalized mass spectra selected for comparison".

We agree this can be confusing. The phrase 'potential analytes' have been used throughout the manuscript to refer to these undetermined substances. We have clarified using a combination of this and the suggested language: *"M1 and M2 are normalized mass spectra of two potential analytes being compared to determine whether they represent the same analyte."*

Figure 5 shows several low-abundant compounds that were detected manually (blue asterisks). Were these compounds also detected by the presented PMF method? Why are the large prominent peaks in the TIC not detected by the manual method? It looks if there is a homologue series of alkanes in the chromatogram (visible as an evenly-spaced series of decreasing peaks from 400-650s). Why has this not been identified in the manual analysis? As a consequence, the manual inspection could easily identify much more compounds, with implication on the statement of the "one order of magnitude" (line 381).

The reviewer makes a good point, that as described in the original manuscript, the previously published analysis of these data "focused on only ~100 compounds," not that those were the only analytes that could be possibly identified in this dataset. Indeed, the alkanes could be identified, though there were not a focus of that analysis or present in the published work (which focused on oxidation products of biogenic emissions). We examine this dataset in this work with two goals in mind: to test for negative artifacts (overlooking analytes known to exist in the dataset) and to determine the number of analytes cataloged as a representative application. We have clarified these goals at the start of section 3.3:

"To evaluate the proposed method in a real-world application, we apply it across the full chromatographic range for data representing the gas- and particle-phase composition of atmospheric samples. *The goal of this analysis is to both provide an estimate of the number of analytes found in representative atmospheric samples and evaluate the ability of the cataloging approach to identify analytes known to exist in a complex, real-world dataset."*

We have further clarified that the 100 compounds focused on may not represent all manually identifiable compounds, on line 430, and note the additional benefits beyond simply finding a higher number of analytes:

"In contrast, a previously published analysis of this dataset focused on only ~100 compounds cataloged by manual inspection, *though additional compounds are observed to exist in the dataset that were not a focus on this previous analysis. We note that a major advantage of the proposed approach is not only the larger number of analytes cataloged (with significantly less manual interaction), but also that each of these analytes has a well-defined mass spectrum that can be used for identification or comparison to existing mass spectral libraries."*

We have also removed the language from the conclusion regarding "an order of magnitude" as we agree that is dependent on how many peaks manual inspection can find, which is a function of time and effort costs, and instead focus on the more quantifiable conclusions that *"more than 1000 analytes were cataloged with little or no*

human interaction."
In addition, we have examined how our decision to set a critical retention time difference may impact these numbers. We find that aggressive thresholds increase the value to ~1200, and very conservative thresholds (FWHM) decrease it to 950, so roughly 1000 analytes appear to be a fairly robust conclusion. This has been included in the manuscript as well as a table of the results in the SI.
"This analysis uses a moderately aggressive critical retention time difference (1.4 σ), but the number of analytes found is slightly reduced by more conservative approaches (e.g., only 20% lower at using the much more conservative FWHM, Table S1)."

Table S1. Number of analytes cataloged by using various critical retention time difference values.

Name	Definition	Retention time (s)	Number of analytes
HWHM	1.177σ	0.328	1216
MPF*	√2σ	0.394	1169
4 datapoints	1.5σ	0.420	1145
2σ	2σ	0.558	1018
FWHM	2.355σ	0.656	943

Please also note the supplement to this comment:
https://amt.copernicus.org/preprints/amt-2022-16/amt-2022-16-AC2-supplement.pdf