Comprehensive detection of analytes in large chromatographic datasets by coupling factor analysis with a decision tree

Sungwoo Kim[1], Brian M. Lerner[2], Donna T. Sueper[2], Gabriel Isaacman-VanWertz[1]

[1]Charles E. Via Jr. Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA, 24061, USA
5   [2]Aerodyne Research, Inc., Billerica, MA, 01821, USA

Correspondence to: Gabriel Isaacman-VanWertz (ivw@vt.edu)

Abstract. Environmental samples typically contain hundreds or thousands of unique organic compounds, and even minor components may provide valuable insight into their sources and transformations. To understand atmospheric processes, individual components are frequently identified and quantified using gas chromatography/mass spectrometry. However, due
10   to the complexity and frequently variable nature of such data, data reduction is a significant bottleneck in analysis. Consequently, only a subset of known analytes is often reported for a dataset, and a large amount of potentially useful data are discarded. We present here an automated approach of cataloging and potentially identifying all analytes in a large chromatographic dataset and demonstrate the utility of our approach in an analysis of ambient aerosols. We use a coupled factor analysis/decision tree approach to de-convolute peaks and comprehensively catalog nearly all analytes in a dataset.
15   Positive Matrix Factorization (PMF) of small sub-sections of multiple chromatograms is applied to extract factors that represent chromatographic profiles and mass spectra of potential analytes, in which peaks are detected. A decision tree based on peak shape, noise, retention time, and mass spectrum is applied to discard erroneous peaks and combine peaks determined to represent the same analyte. With our approach, all analytes within the small section of the chromatogram are cataloged, and the process is repeated for overlapping sections across the chromatogram, generating a complete list of the retention
20   times and estimated mass spectra of all peaks in a dataset. We validate this approach using samples of known compounds and demonstrate the separation of poorly resolved peaks with similar mass spectra and the resolution of peaks that appear in only a fraction of chromatograms. As a case study, this method is applied to a complex real-world dataset of the composition of atmospheric particles, in which more than 1100 analytes are resolved.

1 Introduction

25   Atmospheric samples are highly complex and often contain multiple thousands of compounds (Goldstein and Galbally, 2007) with a potentially wide range of physicochemical properties and multiple isomers. Valuable information relating to the sources and chemistry of atmospheric components can be extracted from these compounds, however the complexity of the samples requires analytical techniques to effectively separate those compounds. Gas chromatography (GC), when combined with mass spectrometry (MS) as a detection method, is one of the most widely used analytical methods in chemical analysis

30   due to high sensitivity, low limits of detection, and high chemical resolution (HüBschmann, 2015). Though used frequently
for analysis of atmospheric samples, the complex nature of atmospheric data yields substantial challenges, in particular co-
elution of many chromatographic peaks. In some cases, co-elution can be so complex that resolution and integration of
individual components cannot be readily achieved, and data is treated as an "unresolved complex mixture" (Zhang et al.,
2014). It is consequently common for analyses of environmental data to focus on the resolution and quantification of only a
35   subset of specific analytes of interest and leave a large fraction of data unprocessed and unused.

"Traditional" processing of chromatographic data has relied on manual inspection of data to locate analytes of interest,
followed by integration of peaks using an algorithmically determined baseline. While software to perform these analyses are
readily available, this approach may require substantial user interaction, which can be time and resource intensive
(Isaacman-VanWertz et al., 2017). Furthermore, the algorithms implemented in these software show limited capability in
40   handling separation of co-eluted peaks, which leads to suboptimal utilization of data (Johnsen et al., 2013) and makes it
difficult to extract clean mass spectra of analytes for accurate identification. These challenges often result in discarding
potentially valuable information, particularly in large datasets that main contain hundreds or thousands of chromatograms
that need to be processed. As fast chromatography has improved and field-deployable gas chromatography has advanced in
fields like atmospheric chemistry (Lerner et al., 2017; Zhao et al., 2013; Apel et al., 2003; Goldan et al., 2004; Hornbrook et
45   al., 2011), the size and complexity of chromatographic datasets make manual processing approaches unfeasible. Field
instruments are also more impacted by shifts in operating conditions that may impact data reproducibility and peak co-
elution due to non-ideal laboratory conditions (e.g., temperature fluctuations, etc.). Efforts to tackle the analytical challenge
of integrating complex environmental datasets have focused on improved peak integration methods that use idealized
mathematical peak shapes and defined mass spectra to resolve and integrate even poorly resolved chromatographic peaks
50   (Blaško et al., 2009; Di Marco and Bombi, 2001; Isaacman-VanWertz et al., 2017; Jeansonne and Foley, 1991; Mydlová-
Memersheimerová et al., 2009; Naish and Hartwell, 1988). However, these methods still require manual inspection of the
data to identify and catalog peaks of interest.

To facilitate peak identification in complex samples, matrix decomposition methods have been proposed resolve complex
co-eluting peaks. In a relatively simple form, the co-variance of ions with chromatographic time can resolve representative
55   mass spectra that rise and fall together as chromatographic peak. This approach has, for instance, been implemented as the
Automated Mass Spectral Deconvolution & Identification System (AMDIS) (Zhang et al., 2006) and is a useful tool for the
identification of analytes within a single chromatogram (Meyer et al., 2010). However, large chromatographic datasets
present an opportunity to include an additional dimension of resolution, as not all chromatograms necessarily contain all the
same analytes. As an example, a peak that is unresolved from a neighboring peak in one chromatogram may not exist in
60   another sample, which would allow identification of the neighbor, and the neighbor could then be integrated more accurately
in the first chromatogram now that its spectrum and retention time (and potentially peak shape) is defined.

Several multi-dimensional co-variance or factorization approaches have been developed to identify peaks across multiple
chromatograms. One such approach is PARAFAC, a generalization of bilinear principal component analysis (PCA) to high

order arrays (Hubert et al., 2012) in which a data array consisting of multiple chromatograms is decomposed into loadings
and scores representing chromatographic profiles and mass spectra that can be more efficiently integrated. With relatively
high signal-to-noise ratio and the proper number of components, a unique solution that consists of true mass spectra of
analytes can be found (Skov and Bro, 2008). PARAFAC2 was further developed to perform similar decomposition but with
more robust handling of potential retention time shifts by not requiring all samples to have nearly identical time profiles, as
PARAFAC requires (Zhang et al., 2014). Similarly, positive matrix factorization (PMF), a matrix decomposition method
based on a weighted least squares fit (Paatero and Tapper, 1994), has been applied to deconvolve chromatographic data
(Zhang et al., 2014). Unlike PCA, the result matrices, scores, and loadings of PMF are constrained to be non-negative, which
reflects the characteristics of environmental data more accurately (Paatero, 1997). Another major difference is that, in
contrast to PCA, the factors obtained by PMF are not constrained to be orthogonal, are determined independently, and do not
form a hierarchy in which each successive factor captures less variance. These advantages make PMF well-suited to describe
environmental data, and PMF has become a preferred matrix size reduction technique, particularly in the field of
atmospheric chemistry (Ulbrich et al., 2009). Prior applications to chromatographic data have instead focused on either the
resolution and integration of major components (Amigo et al., 2010; Hoggard and Synovec, 2007), or the extraction of
average mass spectra and chromatographic profiles that provide binned information on broad classes of compounds (Zhang
et al., 2014). However, this approach has some limitations, as minor components may provide unique information regarding
the sources or chemical transformations of a sample.

This study presents an automated approach of cataloging and potentially identifying all analytes in a complex
chromatographic dataset. By coupling PMF, an established factor analysis technique with a decision tree, the approach
described here deconvolutes complex chromatograms into a list of all or nearly all unique analytes and their associated mass
spectra. This technique complements the new generation of tools described above developed to improve the efficiency and
accuracy of integrating a list of known chromatographic peaks.

2 Methods

Our approach to cataloging all analytes in a set of chromatograms consists of two major processes. The first process is PMF
analysis, and the second is a decision tree used to filtering, sorting, and cataloging PMF outputs into a list of analytes. We
first provide an overview of the PMF algorithm before describing the overall cataloging approach.

2.1 Positive matrix factorization

Positive matrix factorization is a bilinear model that approximates an observed data matrix, X, by finding the weighted least
squares solutions for a set of factors that can describe the dataset (Paatero and Tapper, 1994). In the case of chromatographic
data, the data is a matrix in which rows represent the average mass spectra of each averaging time period (typically 1-5 Hz)
and columns are the time series of each mass spectral mass-to-charge ratio (*m/z*). The model is represented as:

95    $X = GF + E$ (1)

where X is the data matrix, E is the residual matrix, and G and F are the score and loading matrices, respectively. The chromatographic data matrix is thus described by a set of factors, each of which has an average mass spectrum (the loading) and a time-dependent score that represents the chromatographic profile. The elements of both G and F matrices are constrained to be non-negative and are therefore expected to more accurately represent real-world data than PCA or other

100   PCA-based matrix decomposition methods.

The number of factors is prescribed *a priori* by the user, which represents a major source of uncertainty and subjective interpretation in typical PMF applications. In contrast to other PMF applications, the primary goal in this work is not to optimally describe the complete data set, but rather to increase the number of factors to a point where even minor components are extracted as separate factors, even at the risk of over-fitting the data (which will be rectified by a subsequent

105   decision tree). Assigning too many factors (e.g., greater than the number of compounds in the sample) in the model can cause "factor splitting," a phenomenon in which a factor that might carry real-world meaning or interpretation is divide into multiple factors that cannot be readily interpreted (Hoggard and Synovec, 2007). In the context of this work, splitting would result in the separation of an analyte into multiple different chromatographic peaks that all represent the same analyte. We address this case using the decision tree presented below and therefore do not rely on existing metrics for evaluating the

110   optimality of the factor solution (e.g., the ratio of error in the solution to expected error, "Q/Q_{exp}").

In this study the PMF Evaluation Tool, PET version 3.04 (Ulbrich et al., 2009) software package in Igor Pro 8 (WaveMetrics, Inc.) was used to run the PMF2 (Paatero and Hopke, 2009) algorithm on the dataset and obtain PMF outputs. Chromatograms were analyzed using the freely available TERN software package within the same programming environment (Isaacman-VanWertz et al., 2017).

115   **2.2 Description of analyte cataloging method**

2.2.1 Batch process of positive matrix factorization

An overview of the complete process is shown in Fig. 1. Multiple chromatograms representing time-varying mass spectra are stacked to yield a three-dimensional data array (I x J x K), where I and J constitute a chromatogram (elution profile x mass spectra) and K is the number of samples (Amigo et al., 2010). Each chromatogram is first aligned to the same retention time

120   basis by using a small number of known compounds or introduced standards in each sample to define known retention times. To achieve high chemical resolution and identify minor constituents, PMF is not performed on the full matrix, but rather on sub-sectioned "slices" that represent short periods of the chromatogram. As samples of environmental data can be highly complex and heterogenous, applying PMF to small portions of the elution time successively is significantly more effective than extracting factors from the full range of the elution time data (Zhang et al., 2014). Each slice is comprised of the same

125   period of elution time from each sample chromatogram. Though retention times of chromatograms have been globally aligned, some small shifts in retention time may still exist at the timescale of a slice; consequently, a secondary, "fine-scale"

retention time correction is applied to each sample within the slice to yield a unified time basis (typically the retention time of one of the chromatograms). A form of correlation optimized warping (COW) (Nielsen et al., 1998) is used, in which the retention time offset is determined that maximizes the correlation of the maximum number of single ion chromatograms (SIC) within the slice. This approach is expected to work so long as a dominant fraction of peaks are present in all chromatograms but does not require all peaks to be present. This fine-scale retention time adjustment is not strictly necessary for the application of PMF; without it, the same analytes will generally be found in the chromatograms. However, minimizing retention time differences between chromatograms is very useful for the subsequent decision tree to determine whether peaks in different chromatograms represent the same analyte, as opposed different analytes with highly similar retention times and mass spectra (e.g., isomers).

For each slice of the three-dimensional data array, the dimensionality is reduced by concatenating samples, such that the first row of the two-dimensional matrix of one chromatogram is positioned after the last row of the matrix of another chromatogram. The resulting two-dimensional slice matrix represents repeating periods of elution time. PMF is then applied to this concatenated slice matrix, yielding a set of factors that represent the elution profiles of mass spectra that co-vary (i.e., chromatographic peaks of analytes). For the reasons discussed above, tens of factors are used in the PMF solution, which is higher than used in most other more common PMF applications (Ulbrich et al., 2009). This number of factors is on the order of the length of the slice divided by the typical peak width (i.e., one or two factors per resolvable peak), but a more detailed discussion of optimizing the number of factors is discussed in the results section. Each resulting PMF factor from the slice consists of the chromatographic profiles of a given mass spectrum in each chromatogram (an example is shown in Fig. S2). Factor data are stored for subsequent processing, and PMF is performed iteratively through slices until the entire range of chromatographic time is covered. Each slice overlaps in part with the slices before and after to capture potential peaks that may be cut off at the edges of each slice; overlap must equal or exceed the typical width of a chromatographic peak to ensure this outcome. The full set of PMF results for all slices are compiled and addressed through a decision tree, shown broadly in Fig. 1 and addressed in more detail below.

The optimal number/length of slices is expected to be data-dependent and is expected to control the number of factors needed to fully deconvolute the data. Shorter slices can likely be analyzed with fewer factors, but require more slices to analyze the full chromatogram, making estimation of computational time somewhat complex. In this work we use slices of 5-15 seconds and PMF solutions with up to 35 factors, then discuss tradeoffs in the results. We demonstrate in this work that the decision tree described below addresses potential factor splitting caused by high factor solutions, so there is little disadvantage to increasing the number of factors other than the additional required computational time. In contrast, decreasing the number of factors can result in the detection of fewer analytes. Consequently, it is necessarily a decision of the user of this method to balance potential decrease in data extraction against computational resources.

2.2.2 Decision tree

A detailed description of steps involved in the decision tree is presented in Fig. 2. Steps are described in detail below, but the
160 overall approach proceeds as follows:

(1) Peaks are detected in each PMF factor and cataloged by quantitative parameters describing their idealized mathematical form.

(2) Spurious peaks are removed by several filters to eliminate noise.

(3) Peaks in each factor are sorted into potential analytes based on retention times.

165 (4) Potential analytes are sorted and combined into a catalog of unique analytes by comparing retention times and mass spectra.

Peak detection and fitting. Within a factor, peaks are detected by using first and second derivatives to find local minima and maxima. All found peaks are then simultaneously fit to mathematically idealized forms; we assume a Gaussian curve as the ideal chromatographic peak shape (Anderson et al., 1970), though experimental peaks are often perturbed from the ideal
170 shape by instrumental factors and a certain level of mixing of signals is introduced (Isaacman-VanWertz et al., 2017). Implementation of this approach was performed using built-in packages within the Igor Pro 8 programming environment (specifically, "Multipeak fitting 2"). A more complex approach could include modified peak shapes (e.g., convolution with an exponential (Isaacman-VanWertz et al., 2017)), but because the goal is to catalog all peaks as opposed to perfectly integrate them, a refined peak shape is likely unnecessary. The outcome of this peak fitting is a set of peaks present within
175 each factor chromatographic profile (example shown in Fig. S3), including their known retention times within each chromatogram and their retention times relative to peaks in all the other chromatograms (i.e., retention times both uncorrected and corrected to a unified time basis). Not all factor chromatographic profiles necessarily contain any peaks at all, with noise or background factors frequently returned (as seen in the example shown in Fig. S2). The parameters that quantitatively describe the peak (location, width, and height) are stored alongside corresponding time profiles and mass
180 spectra from the PMF. The location of a peak (i.e., retention time) is the mean of the Gaussian curve, and the width of a peak is described by the standard deviation of the curve. The uncertainty in these three parameters determined by the fit are also stored. By performing peak detection and initial fitting on all factors within the slice, an index of all potential peaks is generated within the region of the chromatogram. These peaks are then validated and cataloged into analytes in subsequent steps by a decision tree.

185 *Peak filtering*. Once all potential peaks are indexed, spurious potential peaks are eliminated. The automated peak detection and fitting algorithm may find peaks in factors that do not qualitatively appear to contain any chromatographic peaks or may imply that the inclusion of a negative peak improves the fit (Anderson et al., 1970). Filtering thresholds are introduced to remove these peaks. Specifically, peaks with negative parameters and/or an estimated error greater than the corresponding parameter (i.e., width) are considered to be a result of a bad fit and thereby rejected. In addition, peaks with weak signal
190 (peak height/baseline signal < 10) are removed. Furthermore, because true chromatographic peaks are expected to have a

peak widths that can be reasonably well-defined (i.e., the range of possible widths in a dataset may be variable but is generally not very broad), outliers of peak widths are identified using Tukey's fences method (Tukey, 1977) with a conservative range for reasonable peak width. The range is defined as:

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)] \tag{2}$$

195   where Q1 and Q3 are lower and upper quartiles of the sorted peak width, respectively, and k = 3. Peaks with widths outside this range are rejected, i.e., those with widths outside the interquartile range by more than three times the magnitude of the interquartile range. Finally, a small number of peaks whose parameters are near both the upper boundary of the peak width and the lower boundary of the peak height to base ratio (i.e., low abundance, broad peaks, those with a height/width ratio of, empirically, ~10000) are rejected as they indicate either poor fitting of the peaks or a fitting of noise.

200   *Peak sorting*. Ideally, using the optimal number of factors will result in each factor representing one chemical compound, but in practice more than one analyte may be detected within a given factor. This may be due to the true presence of multiple analytes within the slice that have mass spectra too similar to deconvolve (e.g., branched alkane isomers), or may be due to an error introduced during peak detection or fitting. Peaks in a given factor by definition share a mass spectrum, so those that are chromatographically separated by less than the median peak width (i.e., have nearly the same retention time) are assumed

205   to represent the same analyte. When such peaks are found in the same chromatogram, they are assumed to be the product of factor and/or peak splitting and are combined. Peaks that meet these criteria across multiple chromatograms (i.e., found within one factor at the same relative retention time) are assumed to represent the same unique analyte across each chromatogram. Peaks within a factor that are separated by more than a peak width are considered unique analytes. This process sorts the peak catalog to yield a list of all potentially unique analytes across all factors, with some factors containing

210   multiple analytes (all with the same mass spectrum) and some factors containing no analytes. Potential analytes each have an associated retention time and mass spectrum, and a known peak height and width in each chromatogram in which it was found.

It is theoretically also possible that multiple analytes are present in a factor not because they have similar spectra, but because the number of factors is substantially lower than the number of analytes present so PMF yields some approximate

215   convolution of analytes. However, in practice, this issue is largely avoided by using a large number of factors. Furthermore, the approach to combining peaks with similar retention times and spectra within a chromatogram may capture a small number of isomers that share a mass spectrum, are rarely resolved, and covary between samples, (e.g., *m*- and *p*-xylene). This limitation is likely inherent to the approach, as their resolution by an automated peak detection algorithm would require an assumption of peak shape, which would limit its application to complex data.

220   *Analyte sorting*. PMF followed by peak fitting, peak detection, and peak sorting is performed for all slices, generating a list of potential analytes across the full chromatographic range. Because these potential analytes were generated by examining peaks within each factor, this process does not account for the possibility that PMF factor splitting generates multiple factors containing the same analyte with slight variations in their mass spectra due to instrument drifts, etc. To address the issue of

factor splitting, all the potential analytes are intercompared to remove and combine possible repeats. Mass spectra of all

225 analytes (i.e., the mass spectrum of the factor in which they were found) are compared by cosine similarity:

$$\epsilon = \frac{M_1 \cdot M_2}{||M_1||\,||M_2||} \tag{3}$$

where M_1 and M_2 are normalized mass spectra of two analytes. This is the preferred approach of commonly used mass spectral libraries and search programs (Stein, 2014). Two identical mass spectra will have $\epsilon = 1$. Values of 0.8 and higher are generally considered to indicate two mass spectra that may represent the same analyte (Stein, 1994; Worton et al., 2017).

230 Analytes with mass spectral cosine similarity values, $\epsilon \geq 0.8$, are compared by their retention times. In cases where the difference is greater than the median width – in other words, if the peaks are considered sufficiently distant to each other – they are cataloged as two unique analytes. Analytes with matching mass spectra and retention times that are not sufficiently resolved are considered to be the same compound detected by two different factors or are two analytes that cannot be resolved by the instrument either chromatographically or by their mass spectra. When found within one slice, analytes are

235 combined by summing peak heights and weighted averaging of their defining parameters (width, mass spectra, etc.). In overlapping sections between slices, any repeat analytes (i.e., found in both slices with matching retention times and spectra) are simply filtered out.

The outcome of this analyte sorting process is a catalog of unique analytes with associated retention times and mass spectra, including information about their widths and heights in each chromatogram used in the analysis. Examples of analytes found

240 are provided in Figs. S4 and S5, which are discussed in more detail in the results section below. This catalog of analytes is the end goal of the present work, but could be used as a template for subsequent analyses, or as a dataset to be matched against existing libraries or authentic standards for identification (Worton et al., 2017).

2.3 Sample datasets

The method developed is tested using two GC-MS datasets: a laboratory-generated dataset of known standards and a dataset

245 of ambient aerosol samples. Both datasets were collected using a Semi-Volatile Thermal desorption Aerosol Gas chromatograph (SV-TAG). This instrument has been described elsewhere in detail (Isaacman et al., 2014; Williams et al., 2006; Zhao et al., 2013). In brief, sample is collected on a passivated metal fiber filter housed in a temperature-controlled cell, either by introducing a liquid standard or pulling through sampling ambient air. The cell is then thermally desorbed with a programmed temperature ramp (25 to 315 ºC over 8 minutes) and analytes are transferred to a GC column ramped from 50

250 °C to 310 °C. GC eluent is then analyzed by electron ionization mass spectrometry (Agilent Technologies). The two datasets differ in their column ramp rate and dimensions. Laboratory data was collected with a MTX-5 column (15 m × 0.25 mm × 0.25 μm, Restek) at a ramp rate of 12.5 °C/min. Ambient data was collected with a Rtx-5Sil MS column (20 m × 0.18 mm × 0.18 μm, Restek) at a ramp rate of 23.6 °C/min.

For analysis of known standards, liquid standards were injected into the sample collection cell through the automated liquid

255 injection system of the TAG (Isaacman et al., 2011). Standards included 10 ng of n-alkanes (C$_8$-C$_{40}$ diluted from 500 μg/ml,

supplied by AccuStandard) and 15 ng of select perdeuterated n-alkanes: C_{14}, C_{15}, C_{16}, C_{20}, C_{24}, and C_{26} (diluted from stock mixtures made of pure compounds, supplied by C/D/N Isotopes approximately two years prior to use).

Collection of ambient air data took place near Manaus, Brazil as part of the GoAmazon2014/5 campaign (Martin et al., 2016). Details of sampling and the SV-TAG instrument used to collect this dataset has been previously published (Isaacman-VanWertz et al., 2016). The data presented in this work were collected during the wet season in February and March 2014. Samples of atmospheric particles and semi-volatile gases were collected during the first 22 minutes of every hour. During desorption of the collection cell, analytes were derivatized by introducing N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA) into the desorption flow; this method silylates all hydroxyl groups, improving transfer through the GC column (Isaacman et al., 2014). Approximately 100 compounds have been previously resolved and cataloged in this dataset (Isaacman-VanWertz et al., 2016), of which only a small fraction were identified as compounds with known molecular structures and identities.

2.4 Method validation

The analyte cataloging method is investigated using real-world GC-MS data collected on known calibrants and under field conditions, as described below. Two major failure modes are examined: (1) negative errors in the form of uncatalogued analytes due to underfitting, and (2) positive errors in the form of false analytes identified due to factor splitting or overfitting. The former can theoretically be addressed in large part by increasing the number of factors, but this approach increases the potential for the latter. To examine this interplay, and the ability of the decision tree to compensate for potential positive errors, we examine sections of chromatograms containing known *n*-alkanes and perdeuterated isotopologues. These samples are analyzed with a varying number of factors to understand the ability of the method to identify major and minor components, and address factor splitting caused by high numbers of factors. Slices of 4 chromatograms of a 5-15 second window containing known analytes are investigated under a range of method parameters. Application to complex field data provides an addition test for negative errors by challenging the method with data that has previously been catalogued by an expert operator, as well as providing insight into the power of the proposed method.

To validate the method, we discuss below the results of three specific tests. In the first test (Section 3.1), we investigate the potential for positive errors by using high-factor PMF solutions to generate the catalog of peaks used by the decision tree. In the second test (Section 3.2), we investigate the potential for negative errors by examining the deconvolution of poorly resolved analytes with similar mass spectra. In the third test (Section 3.3), we investigate the utility of the method in real-world data by applying the method to a complex environmental sample and examine the potential for negative errors by comparing the analyte catalog to a previously published analysis (Isaacman-VanWertz et al., 2016).

3 Results and discussion

3.1 Effects of increasing factors

285 The proposed cataloging method was applied to a 15-second chromatographic window that included the peak known to represent injected perdeuterated tetradecane ($C_{14}D_{30}$), with the number of PMF factors ranging from 1 to 20 (Fig. 3). In a one-factor solution, one analyte was found, representing the known compound (Fig. 3a). The number of analytes found

290 increased as more factors were used, with the injected compound always found and minor analytes found in higher factor solutions as discussed below. The relationship between the number of analytes found and the number of factors is non-linear, approaching an apparent plateau. This plateauing behavior for analytes is in contrast to growth in the number of peaks found, which continues to increase linearly with number of factors. An ever-growing number of analytes is physically improbable given the relative simplicity of the data, and these peaks likely represent factor splitting. The decision tree addresses this

295 issue by rejecting and combining these found peaks, eventually yielding six unique and distinct analytes that remain relatively stable across solutions.

This result agrees with the trend in the percent of the total signal that is not described by the found analytes, i.e., percent residual calculated as the sum of absolute difference between the total ion chromatogram and the reconstructed signal curve at each point in time relative to the sum of total ion signal. With increasing factors, percent residual first drops, from

300 approximately 17% with one factor down to less than 10% with a few factors. Though identifying the main injected compound and describing 83% of the measured data is independently quite compelling, the 9-factor solution (Fig. 3b) suggests that the measured data can be better described by increasing the number of factors. Though these analytes appear to represent splitting of the chromatographic peak, we demonstrate in the following section that these data represent real analytes that might be overlooked by a manual operator. The 3 to 4 major analytes are therefore cataloged with only a few

305 factors, and the same major analytes are cataloged in a 9-factor solution as a 19-factor solution (Fig. 3c). Subsequent increases in the number of factors, from 4 to 20, yield little additional information to describe the measured data, detecting only a small number of low-abundance peaks. Overall, these results demonstrate that the approach can identify minor components, while increasing the number of factors beyond the minimum necessary neither provides additional information nor impedes the method (other than the additional computational resources used).

3.2 Deconvolution of poorly resolved analytes

310 Analysis of perdeuterated tetradecane discussed above indicates that most of the signal can be described by 3 analytes and while a few additional analytes may be present, their inclusion does little to describe the overall signal. These 3 analytes are those shown in Figs. 3b and 3c as poorly resolved peaks, and they are commonly identified in all factor solutions that found three or more analytes (i.e., 4-factor solutions and higher). The presence of 3 analytes under this peak is curious, as the

315 known sample was comprised of only perdeuterated tetradecane ($C_{14}D_{30}$). However, we demonstrate here that these data can

be described as two isotopologues, $C_{14}D_{29}H$ and $C_{14}D_{28}H_2$. (Fig. 4), which are not unexpected in isotopically labeled standards mixed into methanol, in particular those that were purchased several years prior as is the case here.

The total ion chromatographic peak appears to be normally distributed with minimal skew, but the analyte cataloging method we develop here finds three analytes with slight shifts in retention time and differences in mass spectra. One analyte is

320 deconvolved using only a 2-factor solution (Fig. 4a) and the third is found in higher-factor solutions (Fig. 4c). Retention times are shifted later, as expected for replacement of a deuterium with a hydrogen, and projection of this trend forward (i.e., replacement of all deuterium with hydrogen) predicts a retention time roughly that of non-labelled tetradecane as expected. Similarly, the fragmentation patterns are highly similar, but there are some critical differences in their intensities. This is clearest in the fragmentation patterns at their molecular weight, with $C_{14}D_{30}$, $C_{14}D_{29}H$, and $C_{14}D_{28}H_2$ having substantial

325 signal at *m/z* 228, 227, and 226, respectively. At lower *m/z*, all compounds have large signal on mass *m/z* 66 and differences of 16 (CD_2), but isotopologues also have higher signal at masses shifted by one or two amu (e.g., higher m/z 65 for $C_{14}D_{29}H$). The separation of isotopologues presents one of the most difficult challenges for separation methods like chromatography (Valleix et al., 2006; Filer, 1999) due to the tendency of isotopologues with relatively smaller signal to be completely embedded in the peak of their counterpart and their similar spectral signals (Amigo et al., 2010). Figure 4

330 demonstrates these issues, and the ability of the method to overcome them.

It is a clear possibility that the deconvolution of these three analytes is a case of positive error – that these found analytes are an error within the method as opposed to real analytes. To test for that possibility, we performed the same analysis on non-labelled tetradecane and found all signal reasonably described by a single analyte with no co-eluting analytes as expected (Fig. S6). This result supports the conclusion that the additional peaks found for deuterated alkanes are not artifacts due to

335 the high factor solutions but rather represent true co-eluting peaks that demonstrate the ability to find difficult-to-resolve analytes.

3.3 Cataloging analytes in real-world data

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. Doing so requires user decisions on the

340 optimal parameters (e.g., number of factors, slice length). Figure 3 demonstrates the tendency of the method to find increasing numbers of analytes with increasing factors until reaching a certain threshold. It is reasonable to expect that the maximum number of analytes found in each slice is also a function of slice size (i.e., length of the chromatographic window). As slice length increases, the number of slices decreases (roughly linearly scaling computational time) and the number of necessary factors increases (roughly exponentially scaling computational time, Fig. S8). Because the decision tree

345 is effective at mitigating positive errors, the results of the methods (i.e., the catalog of analytes) are not strongly impacted by optimization decisions, which instead primarily impact efficiency (i.e., minimizing computational time). For the real-world data tested here, the maximum number of analytes observed in each slice roughly approaches a plateau when the number of factors used is 2-3 times the length of each slice (in seconds) (Fig. S7). Due to the balance between factor number and

length, it is generally somewhat more efficient to use lower-factor solutions for a larger number of shorter slices, but
350 computational time is not substantially different across different sets of parameters that meet the necessary number of factors
per slice length (Fig. S9). For this data, we use a 25-factor PMF on 10-second slices with 2-second overlap between slices
(based on a typical peak width of 1-2 seconds).

A sample of four chromatograms within the retention time window 200 – 650 second were analyzed by this approach (Fig.
5), constituting 56 slices. These chromatograms were selected from the complete dataset based on their similar sampling
355 times, minimizing differences in instrument operating conditions (retention time, mass spectrometer tuning) over time. We
recognize that the variance obtained by a larger sample size may increase the amount of information extracted, however this
would significantly increase the computational expense. In essence, the optimization of sample size is dependent on sample-
to-sample variability and processing capability. In this work, we use a sample of four chromatograms to demonstrate the
effectiveness of this approach, and optimization of sample size is dataset-dependent and will be explored in future work.
360 From this sample of chromatograms, a total of 1169 analytes were identified, with a computational time of 290 minutes. In
contrast, a previously published analysis of this dataset focused on only ~100 compounds cataloged by manual inspection.
We probe the present analysis for negative errors by comparing the analyte catalog against analytes identified in the
previously published analysis. The analyte catalog in this work was found to detect every peak in the dataset with a known
identification, including introduced isotopically labeled internal standards, analytes identified by authentic standards, and
365 tracer compounds of interest for known atmospheric processes such as oxidation products of naturally emitted gases and
emissions from biomass burning. For example, the identified peaks observed in the inset of Fig. 5 at 350 and 356 seconds are
the known, highly studied oxidation products of isoprene, 2-methylthreitol and 2-methylerythritol (Claeys et al., 2004;
Surratt et al., 2010; Wang et al., 2005), while the peak co-eluting earlier at 350 seconds is the α-pinene oxidation product
pinic acid. The analytes were also compared to the NIST library and only 96 (~8%) of the cataloged analytes had mass
370 spectral matches in the library that were in the "good" or "excellent" range (Stein, 2008) (Fig. S10). Previous work has
shown that matches below these thresholds indicate the found spectra does not represent the unknown analyte (Worton et al.,
2017), suggesting that 90% of analytes in these samples do not exist in the mass spectral library. These results demonstrate
both the utility of the proposed approach to detect and identify known analytes of interest, and to catalog hundreds of
unknown analytes by their retention time and mass spectra.

375 **Conclusions**

In this work, we describe and evaluate a method to catalog analytes in a set of chromatograms representing complex
environmental data. Analysis of known standards demonstrates high skill at finding minor analytes even when poorly
resolved, with no strong tendency to find spurious analytes that are not actually present. This approach will consequently be
valuable for the automated processing of complex chromatographic data and will enable new information to be extracted that
380 might otherwise be ignored or discarded by conventional approaches due to technical difficulties or limited resources.

Analysis of real-world data found roughly an order of magnitude more analytes than previous manual inspection, with little or no human interaction. Two major future developments would further enhance this approach: improved retention time correction without human interaction (e.g., by parametric optimized warping (Eilers, 2004)), and algorithmic optimization of decisions around the length of each slice, the number of factors, and the number of chromatograms. However, even without

385   these advancements, the required levels of operator interaction are limited, and the proposed method has the potential to substantially improve and expand data analyses of both new and previously collected data.

Author contributions

GIVW conceptualized and supervised the project; SK developed the software code, performed the analysis, and wrote the manuscript draft; BML, DTS, and GIVW contributed to the analysis, and reviewed and edited the manuscript.

390   **Competing interests**

BML and DTS are employed by Aerodyne Research, Inc., which commercializes GC and MS instrumentation.

Acknowledgement

This work was supported by the National Oceanic and Atmospheric Administration Small Business Innovative Research Program (WC133R18CN0064). We thank Chenyang Bi for assistance with generating laboratory data and Allen Goldstein

395   for sharing ambient data.

References

Amigo, J. M., Popielarz, M. J., Callejon, R. M., Morales, M. L., Troncoso, A. M., Petersen, M. A., and Toldam-Andersen, T. B.: Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis, J Chromatogr A, 1217, 4422-4429, 10.1016/j.chroma.2010.04.042, 2010.

400   Anderson, A. H., Gibb, T. C., and Littlewood, A. B.: Computer Resolution of Unresolved Convoluted Gas-Chromatographic Peaks, Journal of Chromatographic Science, 8, 640-646, 10.1093/chromsci/8.11.640, 1970.

Apel, E. C., Hills, A. J., Lueb, R., Zindel, S., Eisele, S., and Riemer, D. D.: A fast-GC/MS system to measure C2 to C4 carbonyls and methanol aboard aircraft, Journal of Geophysical Research: Atmospheres, 108, https://doi.org/10.1029/2002JD003199, 2003.

405   Blaško, J., Kubinec, R., Ostrovský, I., Pavlíková, E., Krupčík, J., and Soják, L.: Chemometric deconvolution of gas chromatographic unresolved conjugated linoleic acid isomers triplet in milk samples, Journal of Chromatography A, 1216, 2757-2761, https://doi.org/10.1016/j.chroma.2008.11.019, 2009.

Claeys, M., Wang, W., Ion, A. C., Kourtchev, I., Gelencsér, A., and Maenhaut, W.: Formation of secondary organic aerosols from isoprene and its gas-phase oxidation products through reaction with hydrogen peroxide, Atmospheric Environment, 38,

410   4093-4098, https://doi.org/10.1016/j.atmosenv.2004.06.001, 2004.

Di Marco, V. B. and Bombi, G. G.: Mathematical functions for the representation of chromatographic peaks, Journal of Chromatography A, 931, 1-30, https://doi.org/10.1016/S0021-9673(01)01136-0, 2001.

Eilers, P. H. C.: Parametric Time Warping, Analytical Chemistry, 76, 404-411, 10.1021/ac034800e, 2004.

Filer, C. N.: Isotopic fractionation of organic compounds in chromatography, Journal of Labelled Compounds and

415   Radiopharmaceuticals, 42, 169-197, 10.1002/(SICI)1099-1344(199902)42:2<169::AID-JLCR178>3.0.CO;2-0, 1999.

Goldan, P. D., Kuster, W. C., Williams, E., Murphy, P. C., Fehsenfeld, F. C., and Meagher, J.: Nonmethane hydrocarbon and oxy hydrocarbon measurements during the 2002 New England Air Quality Study, Journal of Geophysical Research: Atmospheres, 109, https://doi.org/10.1029/2003JD004455, 2004.

Goldstein, A. H. and Galbally, I. E.: Known and Unexplored Organic Constituents in the Earth's Atmosphere, Environmental

420   Science & Technology, 41, 1514-1521, 10.1021/es072476p, 2007.

Hoggard, J. C. and Synovec, R. E.: Parallel Factor Analysis (PARAFAC) of Target Analytes in GC × GC−TOFMS Data: Automated Selection of a Model with an Appropriate Number of Factors, Analytical Chemistry, 79, 1611-1619, 10.1021/ac061710b, 2007.

Hornbrook, R. S., Blake, D. R., Diskin, G. S., Fried, A., Fuelberg, H. E., Meinardi, S., Mikoviny, T., Richter, D., Sachse, G.

425   W., Vay, S. A., Walega, J., Weibring, P., Weinheimer, A. J., Wiedinmyer, C., Wisthaler, A., Hills, A., Riemer, D. D., and Apel, E. C.: Observations of nonmethane organic compounds during ARCTAS − Part 1: Biomass burning emissions and plume enhancements, Atmos. Chem. Phys., 11, 11103-11130, 10.5194/acp-11-11103-2011, 2011.

Hubert, M., Van Kerckhoven, J., and Verdonck, T.: Robust PARAFAC for incomplete data, Journal of Chemometrics, 26, 290-298, 10.1002/cem.2452, 2012.

430   Hübschmann, H.-J.: Fundamentals, in: Handbook of GC‑MS, 7-354, https://doi.org/10.1002/9783527674305.ch2, 2015.

Isaacman-VanWertz, G., Sueper, D. T., Aikin, K. C., Lerner, B. M., Gilman, J. B., de Gouw, J. A., Worsnop, D. R., and Goldstein, A. H.: Automated single-ion peak fitting as an efficient approach for analyzing complex chromatographic data, J Chromatogr A, 1529, 81-92, 10.1016/j.chroma.2017.11.005, 2017.

Isaacman-VanWertz, G., Yee, L. D., Kreisberg, N. M., Wernis, R., Moss, J. A., Hering, S. V., de Sa, S. S., Martin, S. T.,

435   Alexander, M. L., Palm, B. B., Hu, W., Campuzano-Jost, P., Day, D. A., Jimenez, J. L., Riva, M., Surratt, J. D., Viegas, J., Manzi, A., Edgerton, E., Baumann, K., Souza, R., Artaxo, P., and Goldstein, A. H.: Ambient Gas-Particle Partitioning of Tracers for Biogenic Oxidation, Environ Sci Technol, 50, 9952-9962, 10.1021/acs.est.6b01674, 2016.

Isaacman, G., Kreisberg, N. M., Worton, D. R., Hering, S. V., and Goldstein, A. H.: A versatile and reproducible automatic injection system for liquid standard introduction: application to in-situ calibration, Atmos. Meas. Tech., 4, 1937-1942,

440   10.5194/amt-4-1937-2011, 2011.

Isaacman, G., Kreisberg, N. M., Yee, L. D., Worton, D. R., Chan, A. W. H., Moss, J. A., Hering, S. V., and Goldstein, A. H.: Online derivatization for hourly measurements of gas- and particle-phase semi-volatile oxygenated organic compounds by thermal desorption aerosol gas chromatography (SV-TAG), Atmos. Meas. Tech., 7, 4417-4429, 10.5194/amt-7-4417-2014, 2014.

445 Jeansonne, M. and Foley, J.: Review of the Exponentially Modified Gaussian (EMG) Function Since 1983, Journal of Chromatographic Science, 29, 258-266, 10.1093/chromsci/29.6.258, 1991.

Johnsen, L. G., Amigo, J. M., Skov, T., and Bro, R.: Automated resolution of overlapping peaks in chromatographic data, Journal of Chemometrics, 28, 71-82, 10.1002/cem.2575, 2013.

Lerner, B. M., Gilman, J. B., Aikin, K. C., Atlas, E. L., Goldan, P. D., Graus, M., Hendershot, R., Isaacman-VanWertz, G.

450 A., Koss, A., Kuster, W. C., Lueb, R. A., McLaughlin, R. J., Peischl, J., Sueper, D., Ryerson, T. B., Tokarek, T. W., Warneke, C., Yuan, B., and de Gouw, J. A.: An improved, automated whole air sampler and gas chromatography mass spectrometry analysis system for volatile organic compounds in the atmosphere, Atmos. Meas. Tech., 10, 291-313, 10.5194/amt-10-291-2017, 2017.

Martin, S. T., Artaxo, P., Machado, L. A. T., Manzi, A. O., Souza, R. A. F., Schumacher, C., Wang, J., Andreae, M. O.,

455 Barbosa, H. M. J., Fan, J., Fisch, G., Goldstein, A. H., Guenther, A., Jimenez, J. L., Pöschl, U., Silva Dias, M. A., Smith, J. N., and Wendisch, M.: Introduction: Observations and Modeling of the Green Ocean Amazon (GoAmazon2014/5), Atmos. Chem. Phys., 16, 4785-4797, 10.5194/acp-16-4785-2016, 2016.

Meyer, M. R., Peters, F. T., and Maurer, H. H.: Automated Mass Spectral Deconvolution and Identification System for GC-MS Screening for Drugs, Poisons, and Metabolites in Urine, Clinical Chemistry, 56, 575-584,

460 10.1373/clinchem.2009.135517, 2010.

Mydlová-Memersheimerová, J., Tienpont, B., David, F., Krupcik, J., and Sandra, P.: Gas chromatography of 209 polychlorinated biphenyl congeners on an extremely efficient nonselective capillary column, Journal of chromatography. A, 1216, 6043-6062, 10.1016/j.chroma.2009.06.049, 2009.

Naish, P. J. and Hartwell, S.: Exponentially Modified Gaussian functions—A good model for chromatographic peaks in

465 isocratic HPLC?, Chromatographia, 26, 285-296, 10.1007/BF02268168, 1988.

Nielsen, N.-P. V., Carstensen, J. M., and Smedsgaard, J.: Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, Journal of Chromatography A, 805, 17-35, https://doi.org/10.1016/S0021-9673(98)00021-1, 1998.

Paatero, P.: Least squares formulation of robust non-negative factor analysis, Chemometrics and Intelligent Laboratory

470 Systems, 37, 23-35, https://doi.org/10.1016/S0169-7439(96)00044-5, 1997.

Paatero, P. and Hopke, P. K.: Rotational tools for factor analytic models, Journal of Chemometrics, 23, 91-100, 10.1002/cem.1197, 2009.

Paatero, P. and Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, Environmetrics, 5, 111-126, https://doi.org/10.1002/env.3170050203, 1994.

475   Skov, T. and Bro, R.: Solving fundamental problems in chromatographic analysis, Anal Bioanal Chem, 390, 281-285, 10.1007/s00216-007-1618-z, 2008.

Stein, S. E.: Estimating probabilities of correct identification from results of mass spectral library searches, Journal of the American Society for Mass Spectrometry, 5, 316-323, https://doi.org/10.1016/1044-0305(94)85022-4, 1994.

Stein, S. E.: National Institute and Standards and Technology

480   (NIST) Mass Spectral Search Program, 2008.

Stein, S. E.: NIST/EPA/NIH Mass Spectral Library with Search Program

Data Version: NIST v20

Software Version: 2.4, 10.18434/T4H594, 2014.

Surratt, J. D., Chan, A. W. H., Eddingsaas, N. C., Chan, M., Loza, C. L., Kwan, A. J., Hersey, S. P., Flagan, R. C.,

485   Wennberg, P. O., and Seinfeld, J. H.: Reactive intermediates revealed in secondary organic aerosol formation from isoprene, Proceedings of the National Academy of Sciences, 107, 6640, 10.1073/pnas.0911114107, 2010.

Tukey, J. W.: Exploratory data analysis, Addison-Wesley series in behavioral science, Addison-Wesley Pub. Co., Reading, Mass.1977.

Ulbrich, I. M., Canagaratna, M. R., Zhang, Q., Worsnop, D. R., and Jimenez, J. L.: Interpretation of organic components

490   from Positive Matrix Factorization of aerosol mass spectrometric data, Atmos. Chem. Phys., 9, 2891-2918, 10.5194/acp-9-2891-2009, 2009.

Valleix, A., Carrat, S., Caussignac, C., Leonce, E., and Tchapla, A.: Secondary isotope effects in liquid chromatography behaviour of 2H and 3H labelled solutes and solvents, J Chromatogr A, 1116, 109-126, 10.1016/j.chroma.2006.03.078, 2006.

Wang, W., Kourtchev, I., Graham, B., Cafmeyer, J., Maenhaut, W., and Claeys, M.: Characterization of oxygenated

495   derivatives of isoprene related to 2-methyltetrols in Amazonian aerosols using trimethylsilylation and gas chromatography/ion trap mass spectrometry, Rapid Communications in Mass Spectrometry, 19, 1343-1351, https://doi.org/10.1002/rcm.1940, 2005.

Williams, B., Goldstein, A., Kreisberg, N., and Hering, S.: An In-Situ Instrument for Speciated Organic Composition of Atmospheric Aerosols: T hermal Desorption A erosol G C/MS-FID (TAG), Aerosol Science and Technology, 40, 627-638,

500   10.1080/02786820600754631, 2006.

Worton, D. R., Decker, M., Isaacman-VanWertz, G., Chan, A. W. H., Wilson, K. R., and Goldstein, A. H.: Improved molecular level identification of organic compounds using comprehensive two-dimensional chromatography, dual ionization energies and high resolution mass spectrometry, Analyst, 142, 2395-2403, 10.1039/c7an00625j, 2017.

Zhang, W., Wu, P., and Li, C.: Study of automated mass spectral deconvolution and identification system (AMDIS) in

505   pesticide residue analysis, Rapid Commun Mass Spectrom, 20, 1563-1568, 10.1002/rcm.2473, 2006.

Zhang, Y., Williams, B. J., Goldstein, A. H., Docherty, K., Ulbrich, I. M., and Jimenez, J. L.: A Technique for Rapid Gas Chromatography Analysis Applied to Ambient Organic Aerosol Measurements from the Thermal Desorption Aerosol Gas Chromatograph (TAG), Aerosol Science and Technology, 48, 1166-1182, 10.1080/02786826.2014.967832, 2014.

Zhao, Y., Kreisberg, N. M., Worton, D. R., Teng, A. P., Hering, S. V., and Goldstein, A. H.: Development of an In Situ
510   Thermal Desorption Gas Chromatography Instrument for Quantifying Atmospheric Semi-Volatile Organic Compounds, Aerosol Science and Technology, 47, 258-266, 10.1080/02786826.2012.747673, 2013.

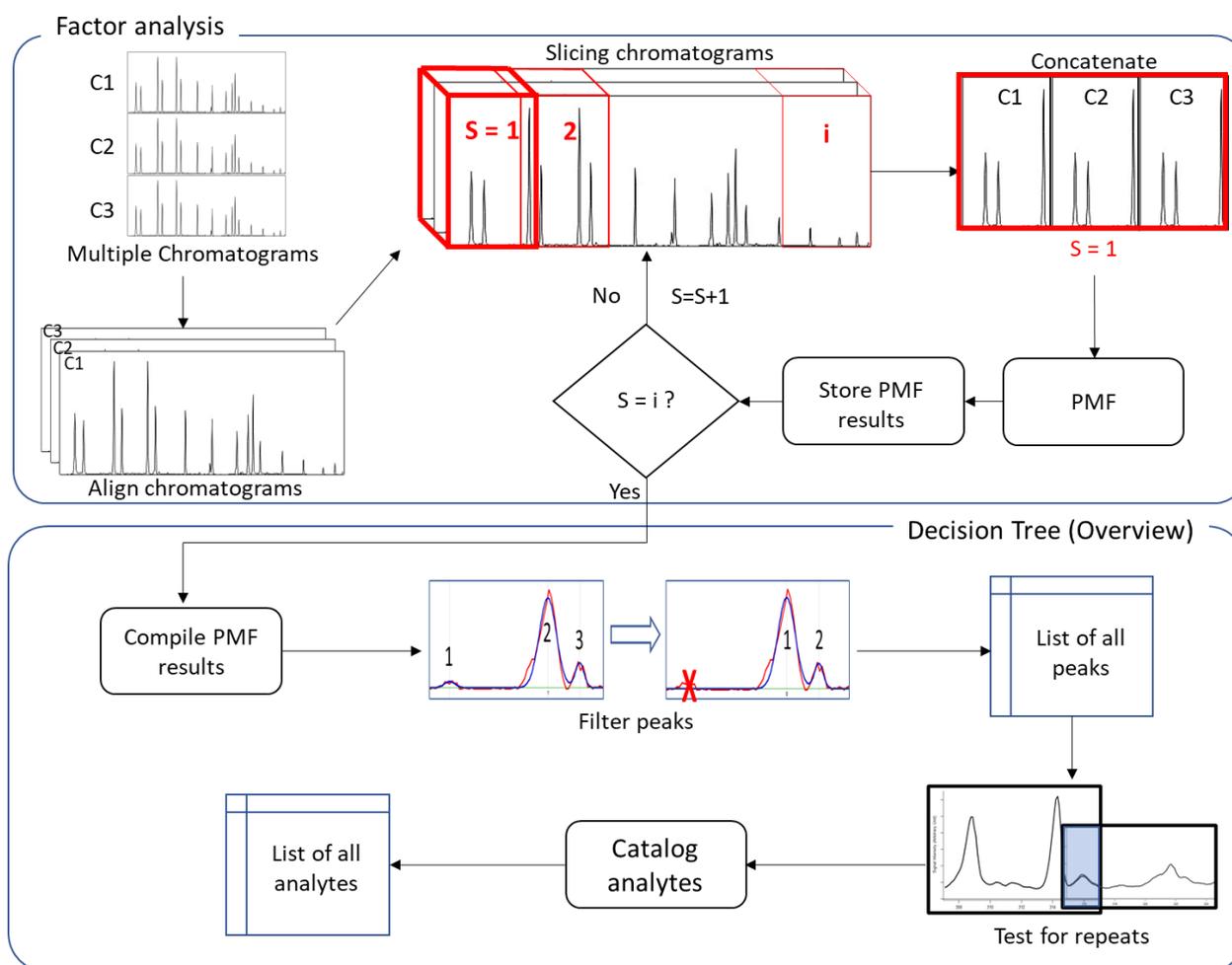



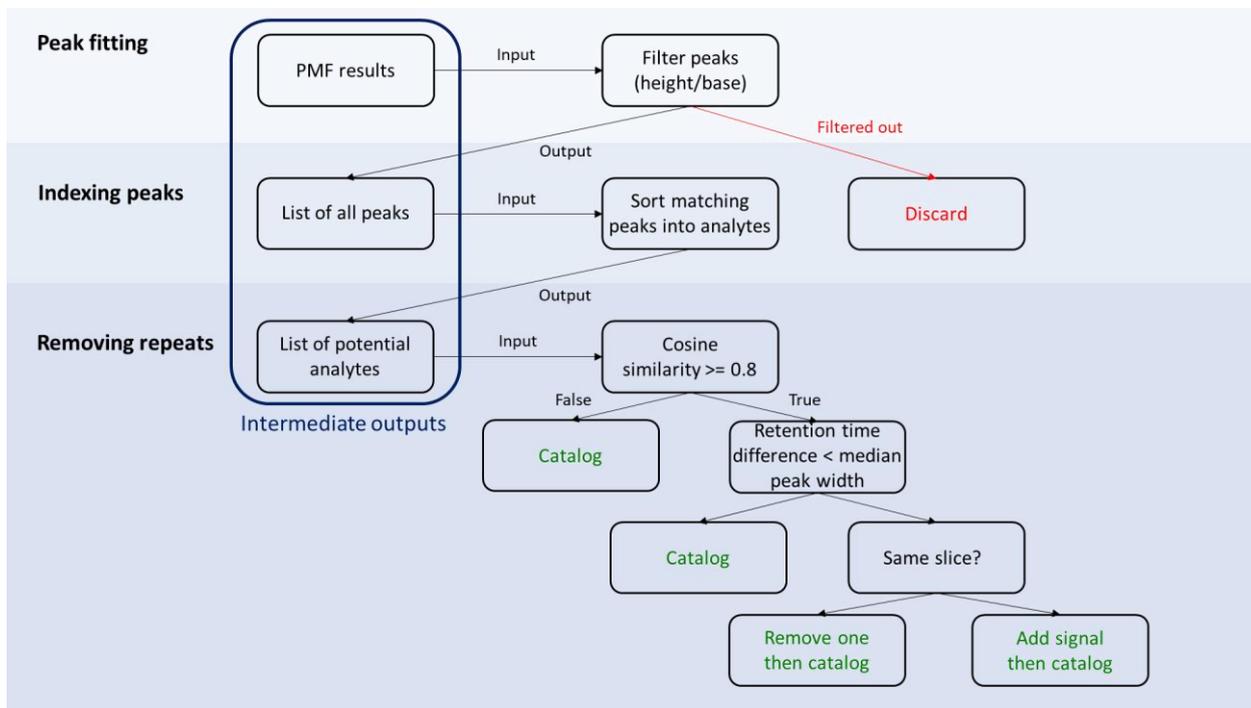
Figure 1: Flowchart of analysis process including an overview of the developed decision tree implemented in this study. Chromatograms and subsections are labeled as C and S, respectively.

515

Figure 2: Detailed work flow of the decision tree. Steps involved in cataloging individual analytes are displayed in green.

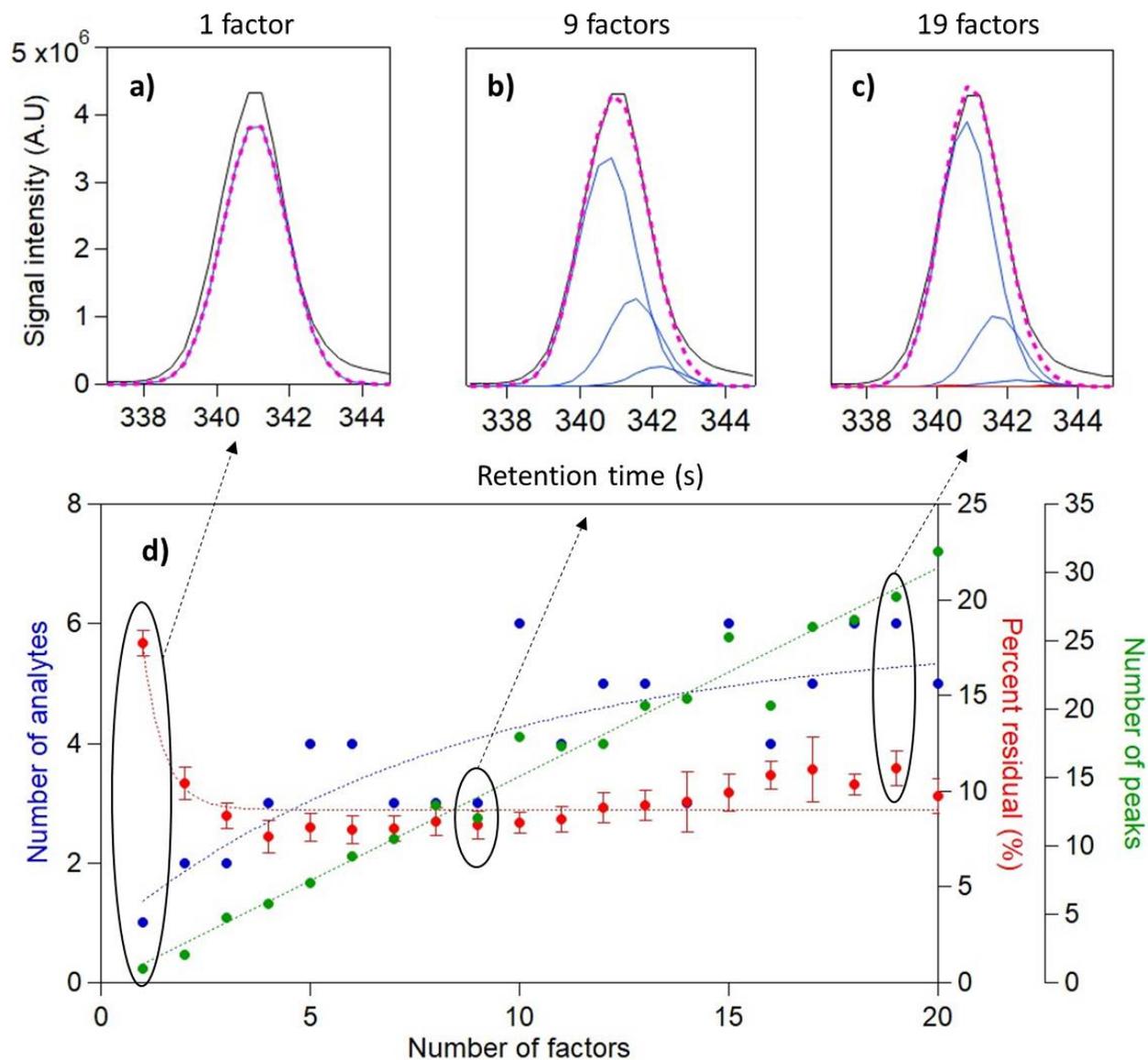



Figure 3: Analysis results of n-alkane mixture with four chromatograms and a 15-second time window containing perdeuterated tetradecane over a range of factors (1-20). Total ion signal and the reconstructed total signal of (a) 1-factor, (b) 9-factor, and (c) 19-factor solutions are displayed above. The number of analytes identified and calculated percent residual over varying number of factors are displayed at the bottom (d) in blue and red, respectively. Error bars represent the standard error of percent residual values from all chromatograms.

520

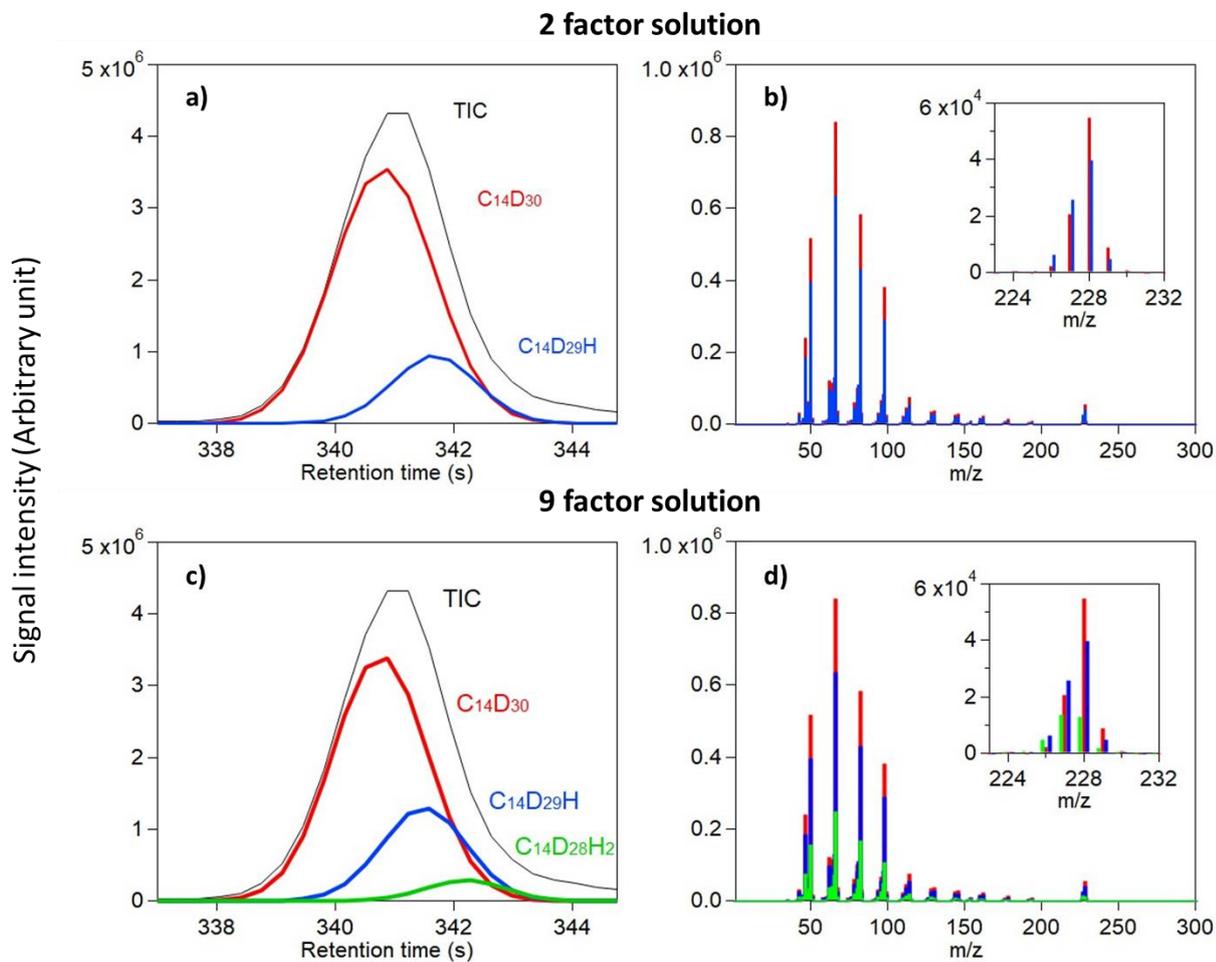



Figure 4: Analysis of deuterated tetradecane. Total ion chromatogram and the reconstructed chromatographic profile of each
525 **analyte found in a (a) 2-factor solution and (c) 9-factor solution are displayed. Measured mass spectra at the corresponding**
location of each analyte found in a (b) 2-factor solution and (d) 9-factor solution are overlaid and displayed with a magnified view
of their molecular weight peaks.

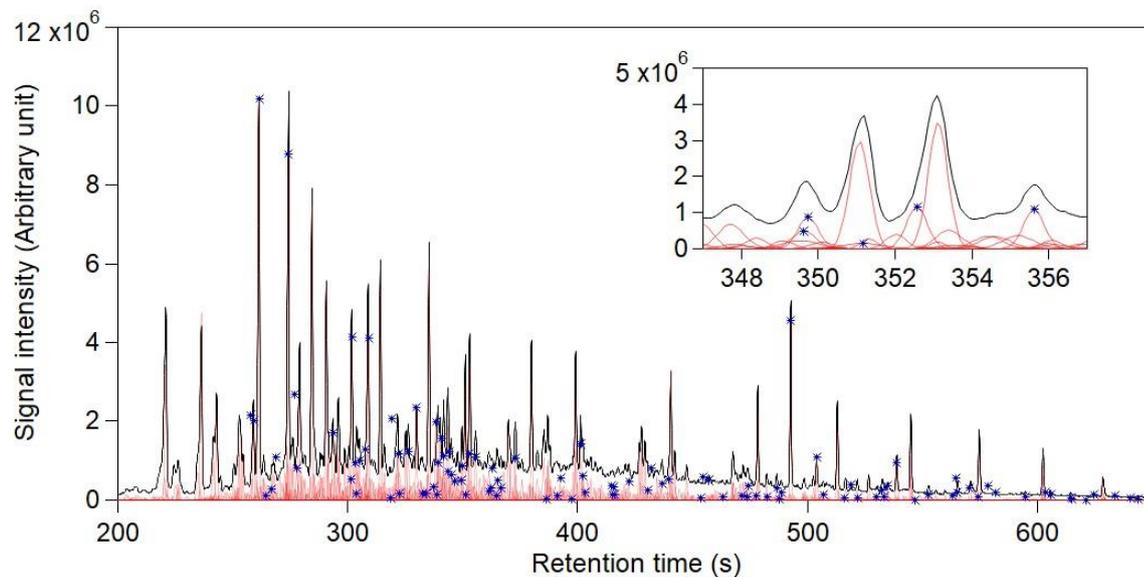



**Figure 5: Analysis of GoAmazon data using 4 chromatograms across the full 450-second chromatogram (200 – 650s). 10-second
slices were used with 2-second overlap. Total ion chromatogram and reconstructed signal of each analyte are displayed in black
and light red, respectively. The magnified view of a 10-second window (347-457s) is presented at the top right corner to display
successful separation of co-eluted peaks. Blue stars indicate the analytes that were identified in a previously published manual
analysis of this data.**