Supplementary Information: Comprehensive detection of analytes in large chromatographic dataset using a coupled factor analysis/decision tree technique

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

[2]Aerodyne Research, Inc., Billerica, MA, 01821

S1. Example desciption of analysis process

The process of analyzing a 5-factor solution of a 10-second slice (335-345s) of an injection of known standards is presented here as an example. Four chromatograms of an alkane mixture sample are used, with one shown in Figure S1.

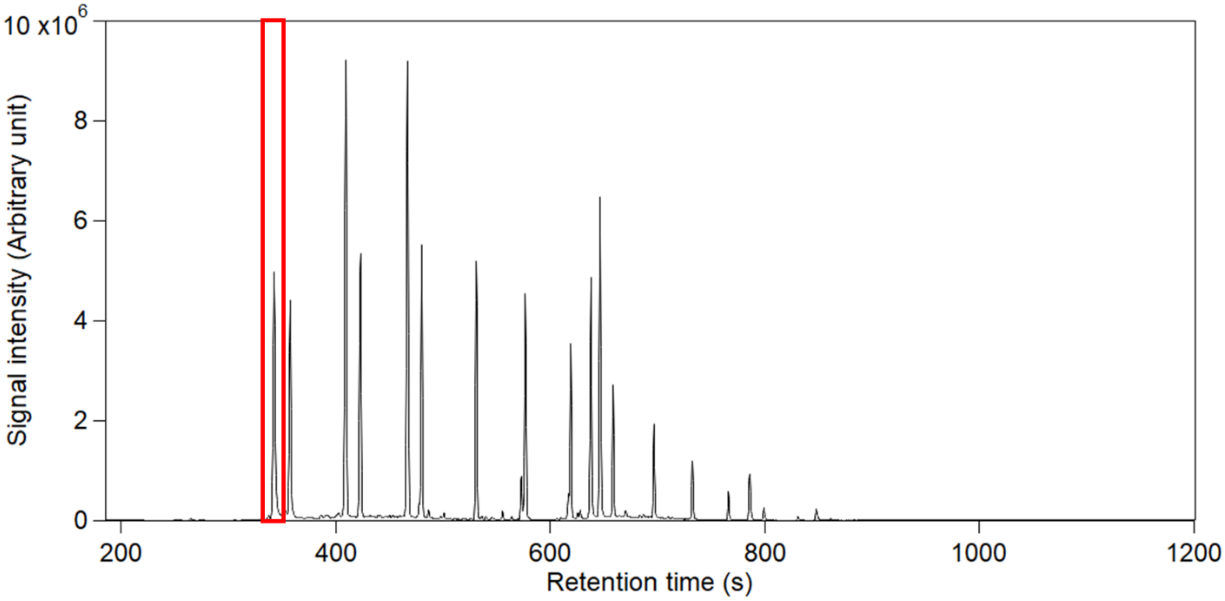

Figure S1. Chromatogram of an alkane and deuterated alkane mixture. A 10-second subsection (335-345s) selected for the purpose of demonstration is displayed in red.

In order to maximize our ability to determine whether peaks in the same slice of chromatograms are the same, a retention time refinement, modelled after correlation optimized warping (COW), is implemented. COW is a piecewise data preprocessing method that aligns the time profile of a sample towards a reference time profile by stretching or compressing the sample. The method used in this study, mode-ion COW, calculates the number of data points the single ion count (SIC) profile of the sample needs to be shifted by to maximize the correlation between the SIC profiles of the sample and reference. After calculating this parameter for all SICs within a slice, the most frequent number is used to shift the total ion count (TIC) of corresponding slice and chromatogram, and this step is repeated until all slices in all chromatograms are covered.

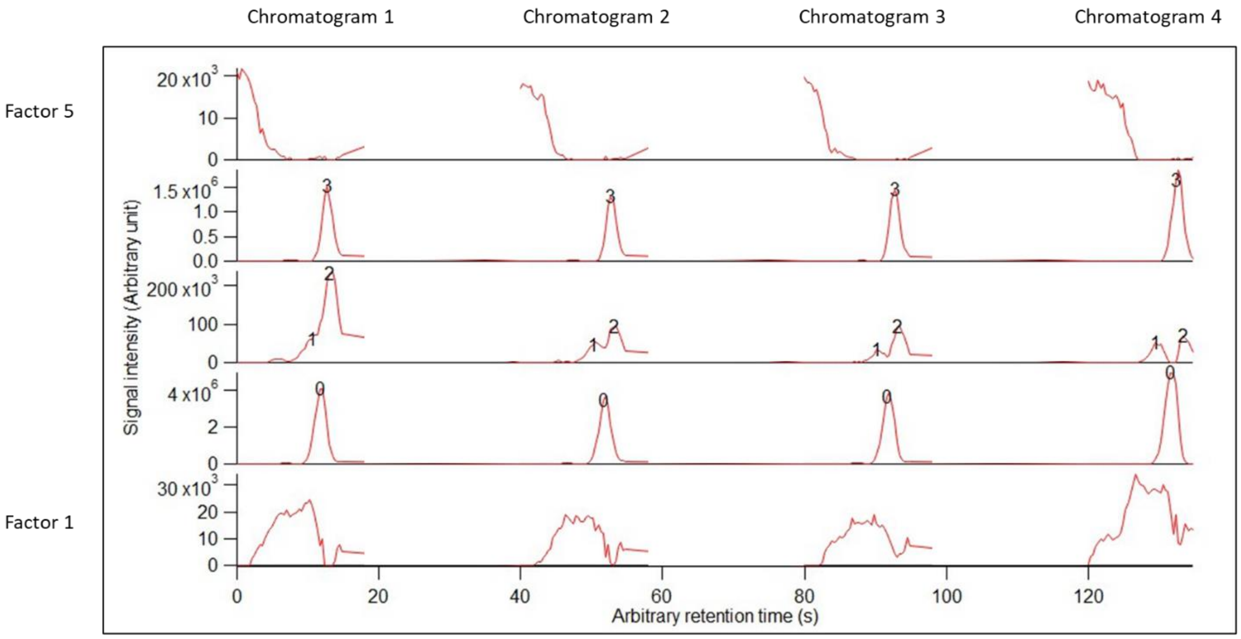

Figure S2. Time profiles of 5 factors used in PMF analysis. Peaks corresponding to each analyte are numbered from 0-3.

Four chromatograms are concatenated with an arbitrarily assigned time gap in between each chromatogram, with the chromatographic profiles of the resulting factors shown in Figure S2. Peak detection finds 16 peaks (numbered) across all 5 factors. Peak fitting results of Factor 3 are presented in Figure S3. Eight peaks are initially identified; these will eventually be sorted into two different analytes by the decision tree. From all four chromatograms, 16 Gaussian peaks out of 29 are sorted into four analytes (Figure S4) with known mass spectra (Figure S5).

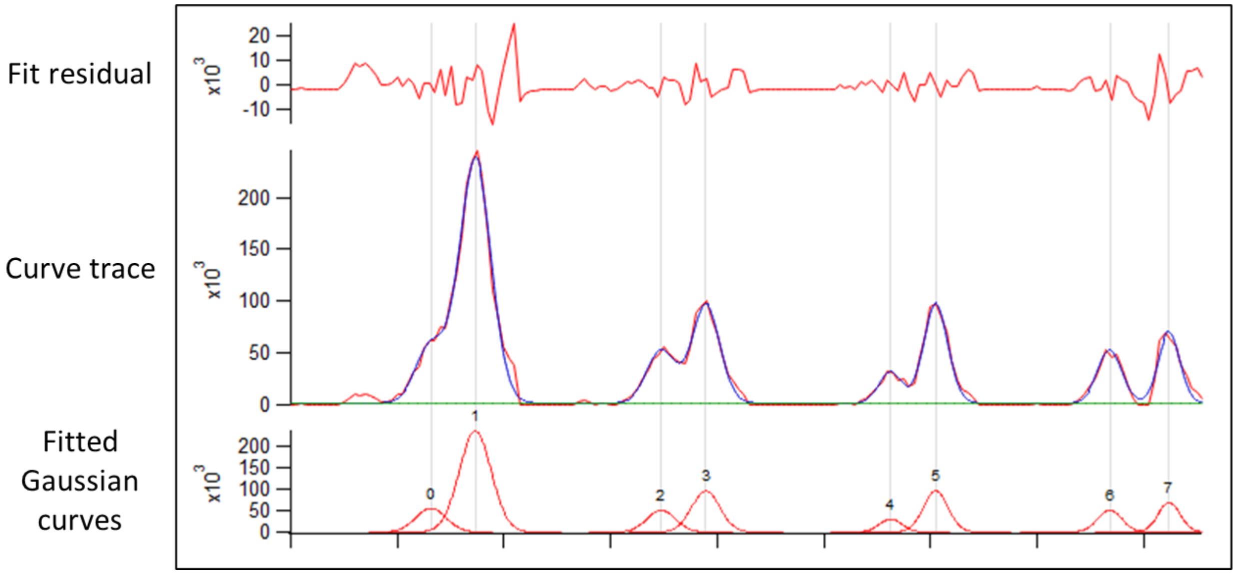

Figure S3. Gaussian curve fitting of factor 3. Fitted curves are numbered from 0-7 in increasing retention time order. The signal difference between the obtained time profile of factor 3 and the curve trace (the total added Gaussian signals), is labelled as fit residual and displayed on the top.

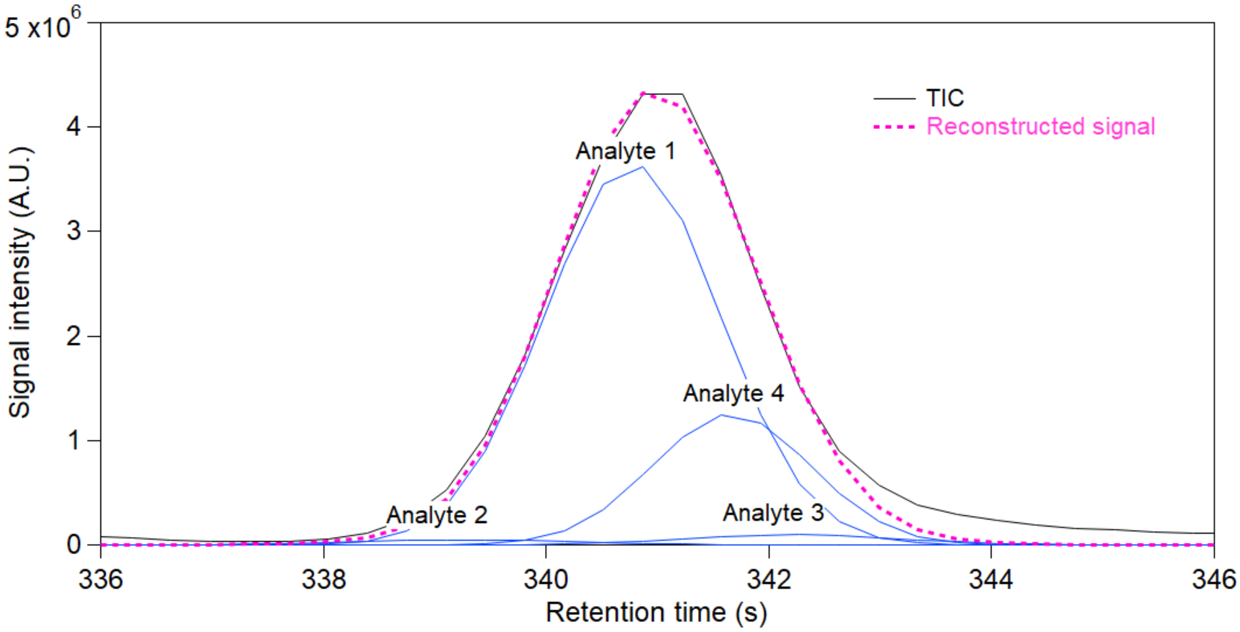

Figure S4. Gaussian curves of all analytes identified overlaid with the original chromatographic signal. The dashed line represents the total reconstructed signal of all analytes.

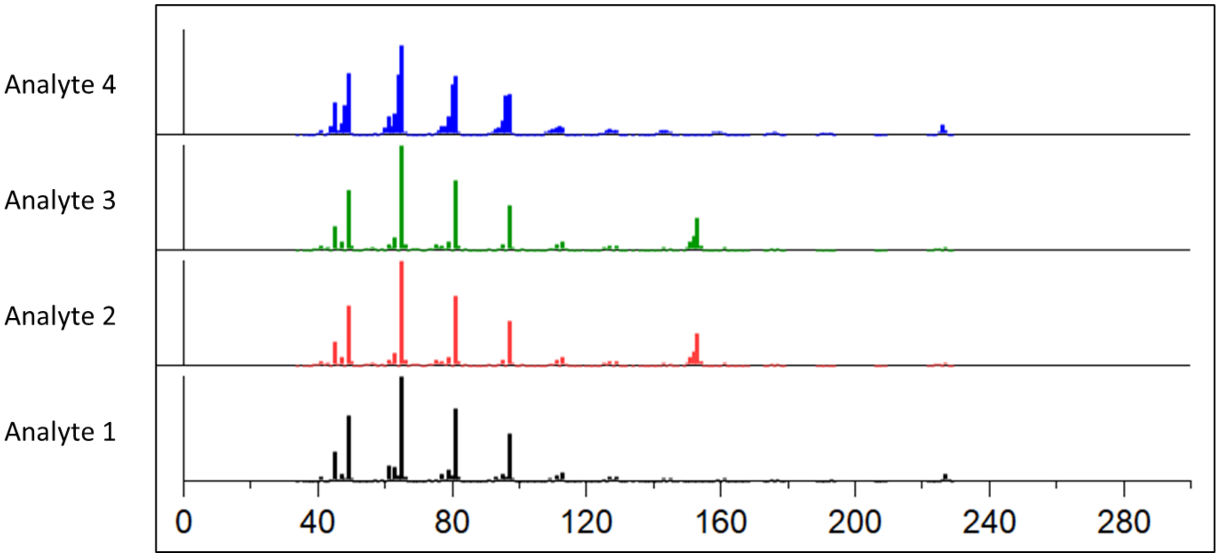

Figure S5. Mass spectra of 4 analytes identified.

S2. Analysis results of a subsection containing tetradecane (alkane mixture sample)

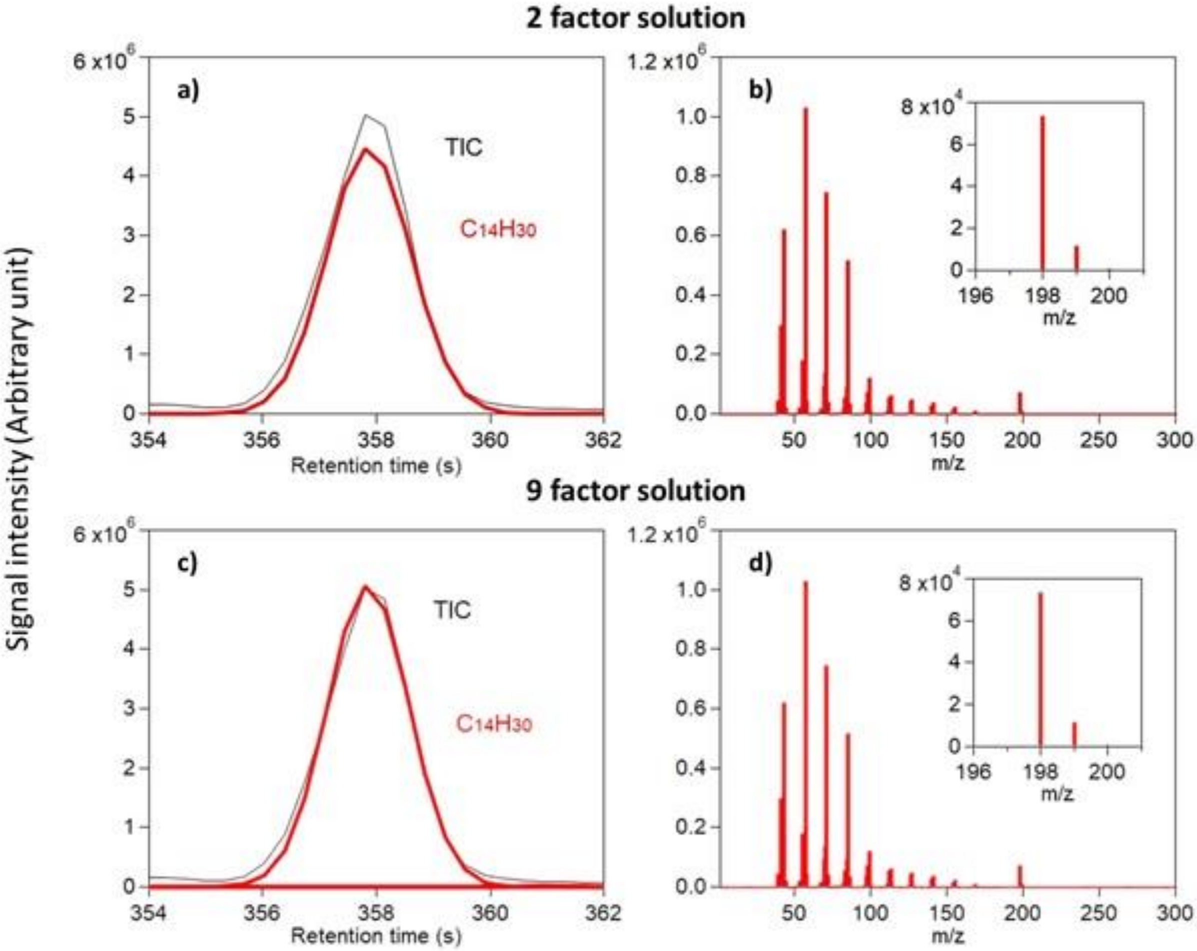

Figure S6. Analysis results of tetradecane, analogous to Figure 4). TIC and the reconstructed time profile of each analyte found in a (a) 2-factor solution and (c) 9-factor solution are displayed. Measured mass spectra at the location of each analyte found in a (b) 2-factor solution and (d) 9-factor solution are stacked and displayed with a magnified view of their molecular weight peaks.

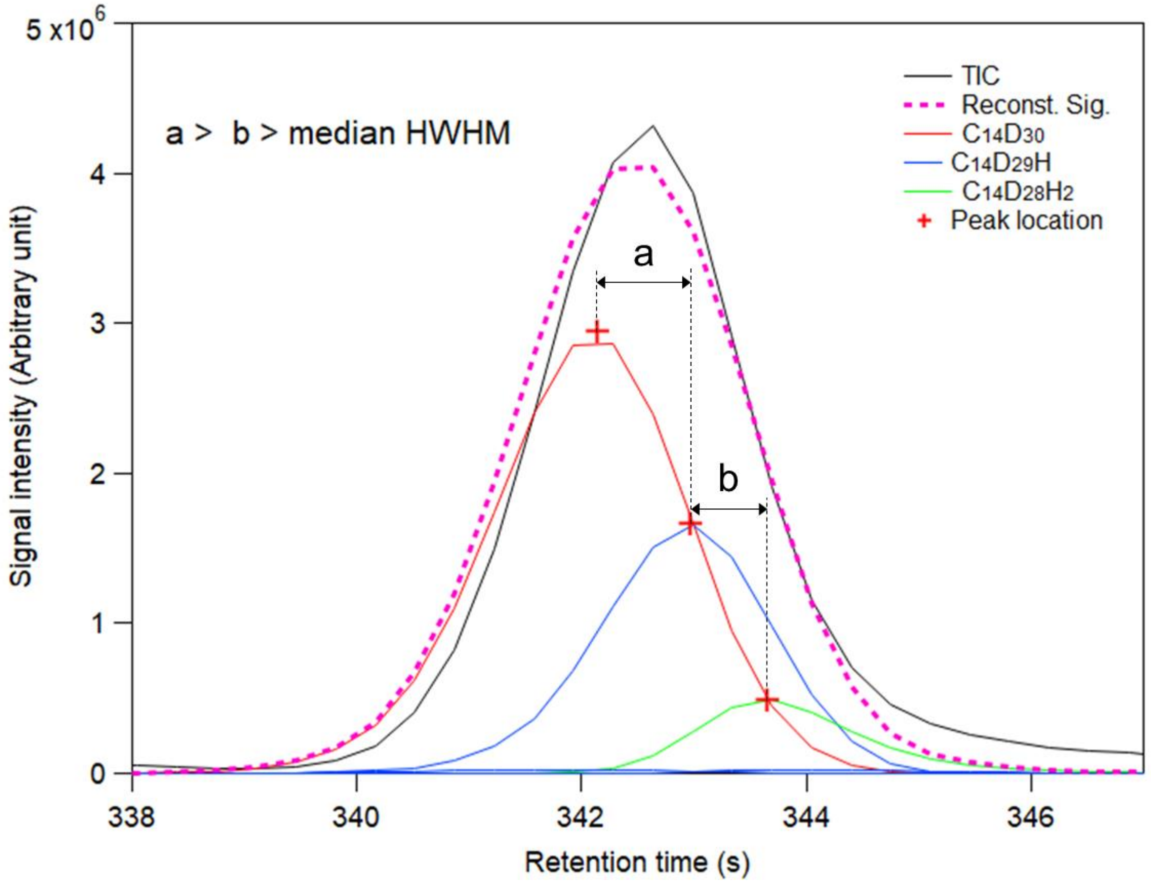

Figure S7. Analysis results of deuterated tetradecane with an exponentially modified gaussian (EMG) as a peak fitting model. EMG curves of all analytes identified overlaid with the original chromatographic signal. The purple dashed line represents the total reconstructed signal of all analytes. Red cross markers represent the location and height of each corresponding analyte. The retention time difference between $C_{14}D_{30}$ and $C_{14}D_{29}H$ is labelled as 'a' and the latter as 'b'.

An experiment has been conducted to investigate the impact of employment of exponentially modified gaussian (EMG) as a peak fitting model, using this same subsection as an example. The analysis method was applied with nine factors on a 15-second chromatographic window of known liquid standards samples described in section 2.3. Since the notable difference in the mass spectra is only detected at the molecular mass ions, their cosine similarity values are above the predetermined threshold ($\varepsilon > 0.8$). However, the median half width at half maximum (HWHM), is smaller than the separation between peaks, but more conversative thresholds for the critical retention time difference would combine some or all of these peaks. The calculated values of median HWHM, a, and b are 0.612, 0.825, and 0.677, respectively.

S3. Results of a single slice of GoAmazon data.

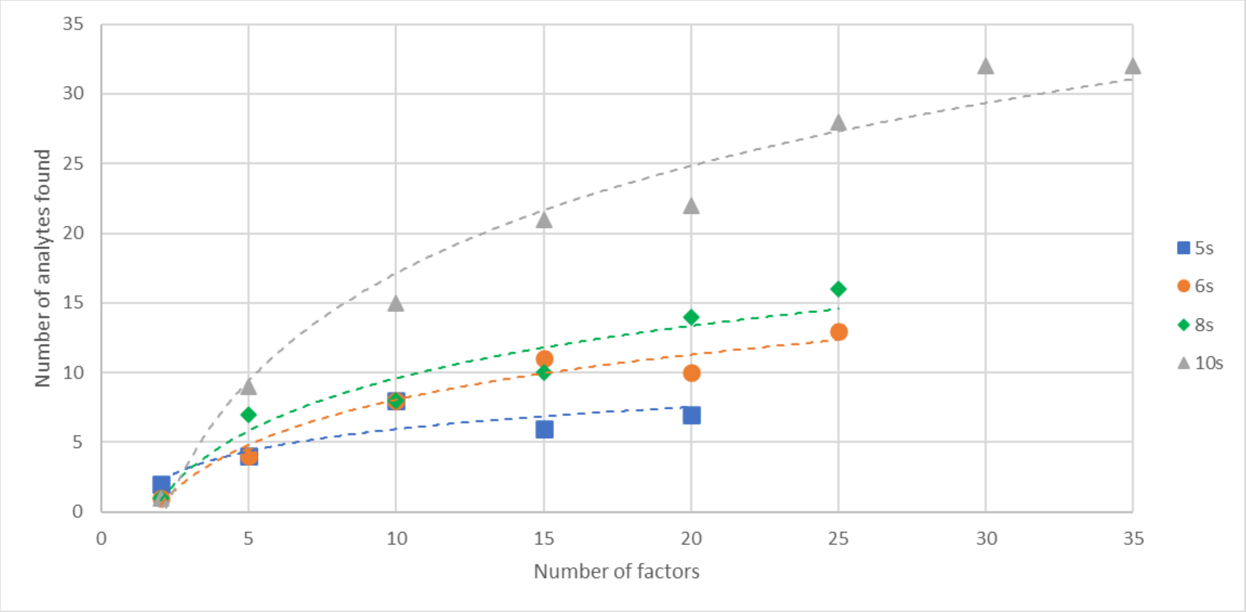

Figure S8. Number of analytes identified using various subsection sizes and a range of factors (2-35).

The analysis results of a known alkane mixture clearly show the rate of new information acquired decreases as the number of factors increases. This indicates that a balance between the level of information and computational time can be found. With this knowledge, multiple analyses were performed on a randomly selected slice of the real-world test data to determine a preferred subsection size and number of factors to be used for the analysis of the entire dataset. Number of analytes found in various subsection sizes (5,6,8, and 10s) were recorded as the number of factors were ranged from 2 to 35. An approximate ratio of roughly 2.5 factors needed per second of time in the slice was determined as the approximate level beyond which increasing factors does not substantially increase the amount of information extracted. Additional factors may provide some additional information but would substantially increase computational time (Figures S9 and S10); the balance between comprehensive analysis and computational time inherently needs to be weighed based on the preferences and available resources of the user.

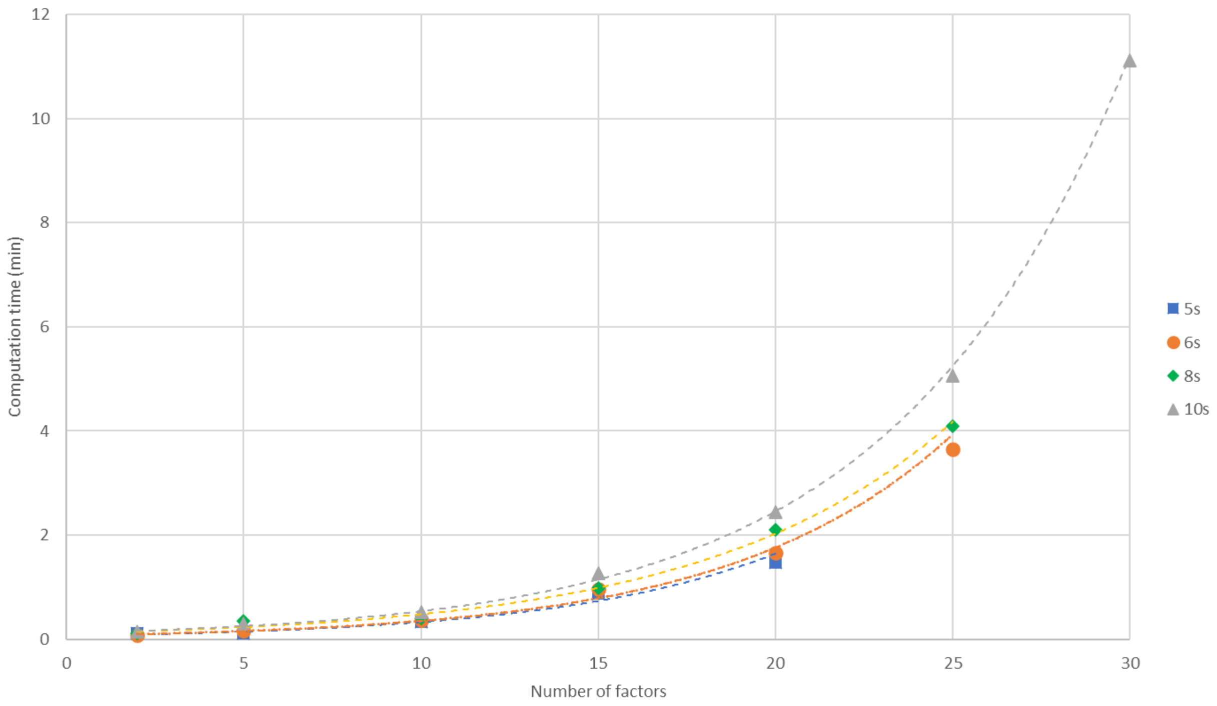

Figure S9. Computation time of one slice using various subsection sizes and a range of factors (2-35).

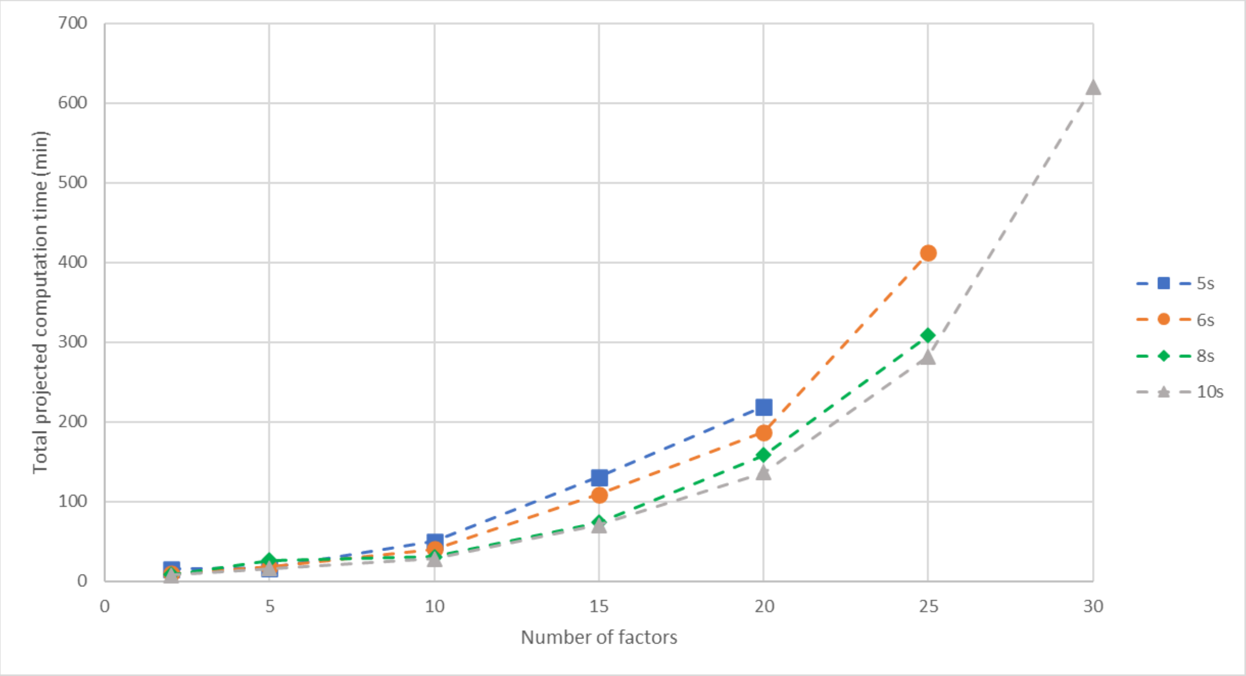

Figure S10. Calculated computation time for analysis of the entire retention time range (200-650 s) using various subsection sizes and a range of factors (2-35).

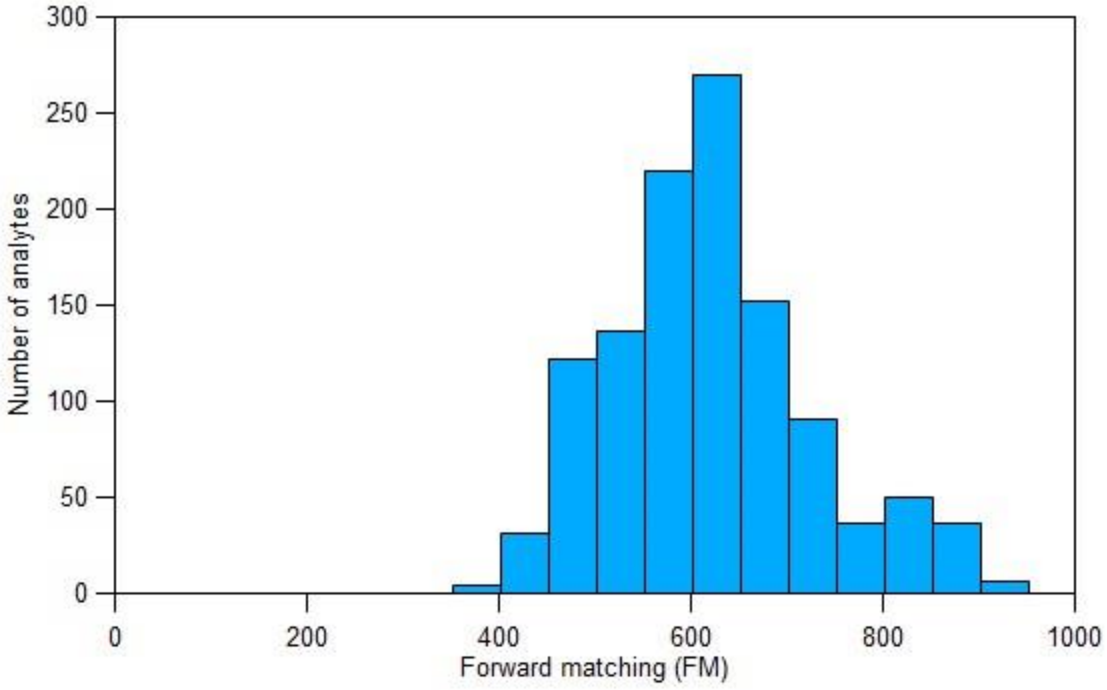

Figure S11. Distribution of mass spectra forward matching values compared to the NIST MS library.

The mass spectra of all 1169 analytes were compared to the NIST MS library using forward matching metric. NIST guideline describes a match value ranging from 800 to 900 as a "good match" and from 900 to 1000 as an "excellent match". A total of 96 analytes were confirmed with match values above 800.

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

**Used by the Multi-peak fitting 2 package used for analysis to simplify calculations*

The impact of screening potential analytes with various critical retention time difference values has been examined. The ambient aerosol samples described in Section 2.3 were analyzed using the method described in Section 3.3 with several critical retention time differences. The results show that the number of analytes cataloged decreases as a more conservative approach is used. As compared to the analysis results presented in Fig. 5, approximately 20 percent fewer analytes were cataloged by using the most conservative (approximately 66 percent larger in retention time) critical retention time difference. Since all other conditions for analysis are kept identical throughout the experiment except for the critical widths, the reduction in the number of cataloged analytes is independent of their mass spectra.