**Automated Optimization and Construction of Chemometric Models Based on Highly Variable Raw Chromatographic Data**

submitted to Analytica Chimica Acta, February 1, 2011

By:

Nikolai A. Sinkov,1 Brandon M. Johnston,1

P. Mark L. Sandercock,2 and James J. Harynuk1

1Department of Chemistry, University of Alberta, Edmonton, Alberta, T6G 2G2 Canada

2Trace Evidence Services, National Centre for Forensic Services – Alberta, Royal Canadian Mounted Police, Edmonton, AB, T5V 1B7 Canada

Corresponding Author: Dr. James Harynuk

Email: [james.harynuk@ualberta.ca](mailto:james.harynuk@ualberta.ca)

Phone: +1-780-492-8303

Fax: +1-780-492-8231

Key Words: Chemometrics; Data Alignment; Retention anchor; GC-MS; PLS-DA; Forensics; Arson

**Abstract**

Direct chemometric interpretation of raw chromatographic data (as opposed to integrated peak tables) has been shown to be advantageous in many circumstances. However, this approach presents two significant challenges: data alignment and feature selection. In order to interpret the data, the time axes must be precisely aligned so that the signal from each analyte is recorded at the same coordinates in the data matrix for each and every analyzed sample. Several alignment approaches exist in the literature and they work well when the samples being aligned are reasonably similar. In cases where the background matrix for a series of samples to be modeled is highly variable, the performance of these approaches suffers. Considering the challenge of feature selection, when the raw data are used each signal at each time is viewed as an individual, independent variable; with the data rates of modern chromatographic systems, this generates hundreds of thousands of candidate variables, or tens of millions of candidate variables if multivariate detectors such as mass spectrometers are utilized. Consequently, an automated approach to identify and select appropriate variables for inclusion in a model is desirable. In this research we present an alignment approach that relies on a series of deuterated alkanes which act as retention anchors for an alignment signal, and couple this with an automated feature selection routine based on our novel cluster resolution metric for the construction of a chemometric model. The model system that we use to demonstrate these approaches is a series of simulated arson debris samples analyzed by passive headspace extraction, GC–MS, and interpreted using partial least squares discriminant analysis (PLS-DA).**1. Introduction**

Modern chemical instrumentation provides incredibly rich data pertaining to highly complex samples analyzed in a quest to understand increasingly complex systems. Faced with these challenges, analysts are turning more and more often towards chemometric techniques to aid in data interpretation. Techniques that range from data visualization tools such as principal components analysis (PCA) [[[1]](#endnote-1),[[2]](#endnote-2),3] to classification techniques such as cluster analysis [[[3]](#endnote-3)] or partial least squares-discriminant analysis (PLS-DA) [[[4]](#endnote-4)] are frequently implemented for the interpretation of chemical data.

When preparing chromatographic data for chemometric analysis, many researchers utilize integrated peak tables where the data matrix represents the abundance of each identified analyte in each sample [1,[[5]](#endnote-5),[[6]](#endnote-6),[[7]](#endnote-7)]. With this approach, the resulting matrix is relatively small, making feature selection straightforward and obviating the need for data alignment. However, with even moderately complex mixtures, this approach faces challenges of coelutions and integration errors, which may influence the resultant model. An alternative approach is to use non-integrated, raw chromatographic profiles of each sample for modeling. In this approach, every data point is considered as an independent variable and a candidate for inclusion in the model [[[8]](#endnote-8),[[9]](#endnote-9),[[10]](#endnote-10)].This approach allows the analyst to utilize all of the chromatographic data, avoiding integration errors and the possibility of coelutions influencing results, especially when a multivariate detector is used. However, this approach presents the challenges of data alignment and feature selection, which must be performed prior to chemometric analysis.

With chromatographic data, small run-to-run variations in retention times of analytes are inevitable, which if not corrected can have a devastating effect on a chemometric model constructed from these data [[[11]](#endnote-11),[[12]](#endnote-12)]. A variety of chromatographic alignment techniques have been developed; examples being correlation optimized warping (COW) [11,[[13]](#endnote-13)], dynamic time warping (DTW) [13] and piecewise alignment algorithms based on peak matching [9,[[14]](#endnote-14),[[15]](#endnote-15)]. These alignment techniques look for similarities in chromatograms or seek to maximize the correlation between chromatograms. These approaches work very well when samples and their matrices are reasonably similar from one analysis to the next (i.e.: HPLC-DAD chromatograms of fungal culture extracts [11] or GC-FID chromatograms of diesel fuels [9]). However, when the chemical profiles of different sample classes are highly dissimilar, chromatographic alignment by these methods across sample types may suffer. If the background matrix of the samples is also highly variable, alignment is even more challenging.

The other challenge for the chemometric interpretation of raw chromatographic data is that of feature selection. When a mass spectrometer is used for detection, the number of individual ion intensities measured during the entire analysis numbers easily in the millions. Consequently, identifying and selecting the relatively small number of meaningful variables while excluding the irrelevant variables is crucial. Most variables in, for example, a GC-MS separation will include only background noise and irrelevant variations; their inclusion in the model will do much more harm than good [[[16]](#endnote-16),[[17]](#endnote-17)]. To reduce the chromatogram to a relatively small number of meaningful variables, several approaches are available. The use of total ion chromatograms (TICs) [[[18]](#endnote-18),[[19]](#endnote-19)] or extracted ion chromatograms (EICs) [18,[[20]](#endnote-20)] is relatively straightforward, but these approaches may eliminate useful information (using TICs results in the loss of detail in the mass spectral domain, and EICs impose preconceived notions of what ions are important on the model). Other methods such as analysis of variance (ANOVA) [9,[[21]](#endnote-21),[[22]](#endnote-22)] or the discriminating variable (DIVA) test [[[23]](#endnote-23)] are more complex. These approaches calculate a metric by which variables can be ranked based on their perceived ability to distinguish between sample classes. When using these approaches, the metric is calculated for all variables, and only the top-ranked variables are used in model construction. Choosing the appropriate number of variables to include is the subsequent challenge and a variety of methods have been applied to guide users in this decision, including the mean correct classification rate [23], Euclidian distance-based metrics [10], and cluster resolution, a metric that we have introduced previously [17]. It is worth noting that each route to feature selection and optimization has its own advantages and disadvantages, and there is always a risk of including uninformative variables or excluding important variables. In cases where the number of samples is small while number of variables is very high, the problem can become more severe [24]. This is especially true for techniques such as PLS-DA, and validation of any model is always recommended [25].

In this research, we present a combined approach for automatically constructing classification models based on raw GC–MS data where there is significant variability in the background sample matrix. The example used herein is the chemometric interpretation of simulated fire debris analyzed as would be done over the course of an arson investigation. Reliable automated chemometric analysis of arson debris is an inherently difficult task. The goal of the arson investigation is to determine whether the fire was accidental or if it was started intentionally. This is one of the more difficult crimes to investigate because much of the evidence at the scene is destroyed by fire [26]. An important aspect of the investigation is the determination of the presence of an ignitable liquid (IL) in the fire debris. In most cases involving arson, a petroleum-based IL (most often gasoline) [27,28] is used because these are inexpensive and readily available [29]. The fluids are mixtures that may easily contain hundreds of individual compounds. Additionally, the matrix is highly variable and complex, containing numerous pyrolysis products that may interfere with the analysis [30]. In order to identify traces of an IL in fire debris, techniques based upon the concentration of headspace vapours [31], typically passive headspace sampling using activated carbon strips [32], or techniques such as solid phase microextraction (SPME) [33] are used. Extracted analytes are then separated and analyzed using GC–MS. In a time-consuming process, the data are usually studied by two or more analysts to confirm the identity of any IL that may be present in the sample.

Considering chemometric interpretation of fire debris, the signatures of ILs are reasonably consistent within each IL class, but differ significantly from one class to the next (e.g. gasoline is very different from camp fuel or turpentine). Thus aligning the raw chromatographic profiles across classes is somewhat challenging. This is compounded by the fact that the relative abundance of an IL pro- file in the background matrix signal is essentially uncontrolled, as is the exact make-up of the background matrix signal, which depends entirely upon what was present at the scene of the fire. To overcome this aspect of the challenge, we present the use of a series of deuterated alkanes, which act as a series of retention anchors to which the entire chromatogram can be warped, regardless of the matrix and sample profile. To construct a chemometric model for identifying the presence or absence of an IL, automated feature selection for a PLS-DA model based on ANOVA rankings and guided by cluster resolution is used.

**2. Experimental**

*2.1 Materials and reagents:*

Nine different gasoline samples were obtained from three local gas stations as IL samples. These samples represented a range of octane ratings (87, 89, and 91) from each vendor. Samples of three perfumes were also collected locally to be used as negative controls. Samples of new lumber, painted scrap lumber, plywood, carpet, fabric (50:50 cotton polyester blend), cotton cloth, glossy magazine pages, linoleum, vinyl siding, asphalt shingles, tar paper, Tyvek® and polyethylene vapour barrier were obtained locally and used in the generation of simulated debris. Carbon disulfide (Omnisolv; VWR, Mississauga, ON) was used as the solvent for extraction of analytes from activated carbon strips 8 mm × 20 mm (Albrayco Technologies, Cromwell, CT). The solvent was spiked with a deuterated alkane ladder consisting of n-heptane (d16), n-nonane (d20), n-undecane (d24), n-tridecane (d28), and n-pentadecane (d32) (CDN Isotopes, Pointe-Claire, QC) at concentrations of 1.3 μL·L­-1 each.

*2.2 Preparation of gasoline samples:*

Samples of weathered gasoline were prepared from aliquots of each gasoline sample that were evaporated to levels of 50, 75 and 90% by weight at room temperature. Unweathered samples were also retained for use. A jet of compressed air was passed over each sample of fresh gasoline to speed its evaporation, which was stopped when the flask reached its target weight for the evaporation. A 3-cm bed of granular activated carbon (Fisher Scientific, Nepean, ON) in the pipet used as a nozzle for the air jet prevented contamination of samples with any compounds that could have been present in the air stream. The carbon bed was held in place by a plug of glass wool. The weathering procedure was repeated for each of the nine gasoline samples at the three levels of evaporation to generate a total of 36 different IL samples (including unweathered gasolines).

*2.3 Preparation of debris:*

Seven different combinations of materials for debris generation, shown in Table 1, were placed into aluminum roasting pans and set alight in a fire-resistant fume hood using a propane torch to avoid leaving an IL signature. Samples were allowed to burn until the contents were well-charred, at which point the fire was extinguished by suffocation and the debris was allowed to cool. Samples were placed in PTFE-lined 1-gallon paint cans (General Paint, Edmonton, AB), and stored for later use.

*2.4 Headspace extraction of volatile analytes:*

For passive sampling, a debris sample was placed in a 1 L mason jar and 1 μL of gasoline or perfume was spiked directly onto the debris, depending on whether the sample was to contain gasoline or be a negative control. For debris blanks the debris was placed into the jar without the addition of gasoline or perfume. Jars were then capped with an activated carbon strip suspended from a safety pin on the inside of the jar lid. Safety pins were held in place by magnets placed on the outside of the lid. The jars were placed in an oven at 60 °C for 16 h to equilibrate. For quality control, one empty jar containing nothing but a suspended activated carbon strip was included in each batch of jars placed into the oven. After the equilibration time, the carbon strips were removed and each coiled into a 1.8 mL GC vial (Chromatographic Specialties, Brockville, ON). 1.0 mL of CS2 containing the deuterated alkane ladder was then added to each vial to extract the analytes.

*2.5 GC-MS analysis:*

Samples were analyzed using an Agilent Technologies 7890A gas chromatograph (GC) with a 5975 quadrupole mass spectrometer (MS) and 7683 auto sampler (Agilent Technologies, Mississauga, ON). Data acquisition and automation were accomplished using MS ChemStation (Agilent). A 30 m × 250 μm × 0.25 μm HP-5 column (Agilent) was used for the separation. The oven program that was used was 50 °C (held for 3.5 min) followed by a ramp to 280 °C at a rate of 20 °C⋅min-1. Samples were injected in split mode to an injector held at 250 °C. The injection volume was 1 μL, with a split ratio of 20:1. The transfer line and source temperatures were 185 and 230 °C, respectively. Mass spectral searching was performed against the 2005 edition of the NIST MS Database (NIST, Gaithersburg, MD).

*2.6 Data handling:*

Chromatograms for 204 samples containing gasoline and 84 containing no gasoline or containing perfume (negative controls) were exported from Chemstation as .csv text files and then imported into MATLAB 7.10.0 (The Mathworks, Natick, MA). Retention time alignment algorithms and chemometric analysis algorithms were performed using lab-built routines in MATLAB with some chemometric analysis functions from the PLS Toolbox 5.2 (Eigenvector Research Inc, Wenatchee, WA).

**3. Results and Discussion**

In this research, we present an approach for aligning chromatographic data where the sample matrix is highly variable and poorly reproducible, resulting in chromatographic signals that are poorly-aligned by other alignment techniques. Additionally, we present a method for automated feature selection for performing PLS-DA analysis of raw chromatographic-mass spectrometric data. The example data set which we have chosen to demonstrate these approaches is a set of simulated fire debris samples, though the approaches could equally be applied to data collected for other studies or by other instrumentation.

Considering fire debris, numerous methods have been proposed for sampling and analysis. In this work, we have imposed constraints that the extraction and instrumental analysis steps must remain essentially unchanged from those presently used in a typical arson laboratory. This will demonstrate the relative ease with which these techniques could be adopted by any laboratory with no requirement for new instrumentation, minimal training, and protecting the utility of libraries of historical chromatographic data for reference purposes. In this initial work, we have also constrained the type of IL to gasoline as it is the most frequently encountered in cases of arson [27,28]. Future studies will address the plethora of other ILs that could be encountered.

*3.1 Data Alignment*

Two alignment routines were compared. The first alignment routine was based on a piecewise alignment algorithm where the features to be aligned were automatically identified based on peak apexes [9], with an additional mass spectral comparison performed before peak matches were assigned. This method requires a target chromatogram to which all chromatograms are aligned. Due to the highly variable nature of the chromatograms, a composite target was required. Choosing a single IL-containing chromatogram as the target did not yield a data set from which a usable model could be generated (data not shown). The composite was constructed by selecting one chromatogram at random. A second chromatogram was then aligned to the initial target using the piecewise alignment algorithm. These two aligned chromatograms were summed, forming a new target chromatogram. A third chromatogram was aligned to this composite using the same alignment algorithm and then added into the composite target. This process was repeated until all 288 chromatograms were included in the target, at which point the intensities of all points were divided by 288 to obtain an average value. After generation of the composite, the aligned chromatograms generated during its construction were discarded. This avoided issues with over-weighting chromatograms added early in the process and was computationally faster than trying to weight the chromatograms while they were being added. As the alignment routine does not depend on the absolute intensity of a peak for matching, skewing the abundances of peaks during this process did not affect the final result. Once the composite target had been constructed, each of the 288 unaligned chromatograms were aligned to the composite target, with the exception that the target was not modified during this process.

For the second alignment routine, an initial filter was applied to each chromatogram by taking the product of the abundances of ions characteristic of the deuterated alkanes (m/z = 46, 50, 66, 82, and 98) in each mass spectrum of the chromatogram. This new vector provided a series of unambiguous peaks due to the presence of the deuterated alkane anchors which were then used to determine the warping to be applied to the entire chromatogram (Fig. 1). When this method was used, one chromatogram was chosen at random as the target and all other chromatograms were aligned to it.

*3.2 Variable selection and PLS-DA analysis:*

For model construction and testing, the data were separated into a training set containing 240 chromatograms and a test set containing 48 chromatograms. The training set was used for feature selection and model construction while the test set was used only for model validation. Assignment of chromatograms to either set was random, and this process was repeated ten times.

Each chromatogram contained 7 300 scans with m/z values from 30 to 300, providing approximately 2×106 variables per chromatogram (each ion abundance in each scan of the mass spectrometer was treated as an individual variable). Each GC-MS chromatogram was unfolded along the retention time axis. ANOVA-based feature ranking [10] was subsequently performed. Briefly, the perceived utility of a variable for distinguishing between classes of samples is directly correlated with its F-statistic. Other variable ranking techniques could have been used; ANOVA was chosen as it is computationally inexpensive and easy to implement. A comparison of feature ranking techniques is beyond the scope of this work. Once the variables were ranked, variable selection was performed using an automated forward-selection approach, with an endpoint based on maximizing the cluster resolution [17] observed between the IL-containing samples and those that did not contain any IL on the PLS-DA scores plot. Briefly, this metric is based on detecting collisions between confidence ellipses drawn around the clusters, thus cluster resolutions of 0.75 or 0.99 mean that 75 or 99% confidence ellipses can be drawn around the clusters without collisions, respectively.

PLS-DA analysis with two latent variables (LVs) was performed on the ten sets of cross-validation data using both alignment methods. Two LVs were chosen as the cluster resolution metric requires a two-dimensional score-plot space, and two LVs should, in principle, be sufficient for classifying two groups of samples. The endpoint chosen for the feature selection was a cluster resolution of 0.9999, or in cases where this degree of resolution could not be reached, the maximum achievable resolution. The results are summarized in Table 2.

*3.3 Comparison of alignment routines:*

In the case of arson debris, there is a high degree of dissimilarity between samples within both classes (those containing an IL and those without an IL). Debris samples without traces of IL show significant variation in their chromatographic profiles due to the combusted matrix: variability that is unavoidable in real samples. Furthermore, these samples contain few peaks due to ILs. With debris containing traces of gasoline, there is significant variation between gasoline traces: gasoline samples from different vendors and different octane ratings have slightly different profiles, and most gasoline samples underwent varying degrees of weathering prior to analysis. This is in addition to the previously mentioned background variability. Consequently, the conventional alignment approach resulted in models of a generally poorer quality, as evidenced by the results presented in Table 2. These data show that the deuterated alkane ladder-based alignment outperforms peak matching-based alignment. With the ladder-based alignment there are no false positives or false negatives observed, and the minimum probablility for a true positive (i.e. the probability for most ambiguously-classified IL-containing sample) is at worst 99.66%. Likewise the maximum probability for a true negative (i.e. the probability that the most ambiguous blank sample contains an IL) is 0.01%. While more extensive testing with samples covering a broader range of IL types, and ideally samples from actual crime scenes is required, the performance of this model appears to be sufficient for eventual application in an arson laboratory, where false negatives and false positives must be avoided at all costs. The performance of this alignment algorithm is contrasted by the models generated with the conventional alignment algorithm. These are plagued by false positives, false negatives, and generally more ambiguous results.

A careful inspection of the raw chromatograms after being aligned by the two techniques demonstrated that the alignment of individual peaks was often superior using the peak matching algorithm, there were also numerous mismatched peaks, and when these occurred the effects were devastating. The reason for these mismatches was likely due to the incredible variability in the matrix background. Conversely, the deuterated alkane ladder provided a series of unambiguous anchors for alignment. With the ladder, the alignment of individual peaks was not as precise; however, there were no mismatched peaks. Consequently the overall alignment of the chromatograms was superior. It is worth noting that in cases where the samples are not as variable as those examined in this work, the ladder-based alignment could be used to effectively adjust for gross misalignment of data, followed by an additional “fine-tuning” alignment (e.g. by feature-based alignment or COW) either globally or across windows of data if needed.

A further comparison of the results for the two alignment approaches is demonstrated in Figures 2 through 4 which depict the scores, y-predicted, and probability plots for sets 5 and 10 which represent the worst and one of the best results for the peak matching-based alignment, respectively. For both of these sets, the ladder-based alignment delivers unambiguous results.

Considering the original size of the data (about 2 × 106 features) a relatively small number of variables (on average less than 300 or 0.015%) allow efficient discrimination between the classes. To verify that the features selected by the algorithm were reasonable in a chemical sense, the selected features were traced backwards to identify the corresponding components in the IL-containing debris samples. Mass spectra of compounds corresponding to the identi- fied features were compared to the NIST MS Database and retention indices were estimated based on the deuterated alkane ladder signal and compared to literature values (Table 3). It is worth noting that in GC, deuterated alkane retention times will be slightly earlier than the corresponding non-deuterated alkanes. The end result of this being a slight increase in our estimated retention indices. Additionally, it is well documented that retention indices are temperature programming rate dependent, with the retention index increasing with temperature programming rate [34]. The literature values presented were collected at a temperature programming rate of 6 ◦C min−1 and the temperature programming rate used in our research was 20 ◦C min−1. Thus even though our estimates do not match exactly with the literature values, the positive variation of about 25–30 units is easily rationalized and the relative values for each compound are consistent with each other. Furthermore, these compounds are consistent with the presence of gasoline in a sample.

Further investigation of the relevant features, some of which are shown in Fig. 5, shows that they are often highly correlated. This is understandable given that numerous masses are measured for each individual compound, and when a relevant compound elutes, its ions will all appear together. The present algorithm does not include a feature reduction step to reduce highly correlated data, but this tool is under development.

**4. Conclusions**

We have demonstrated the use of a deuterated alkane ladder to provide a series of retention time anchors that improves the overall quality of data alignment over conventional approaches with highly variable samples. The ladder-based approach is robust to misalignment issues that can arise in these situations.

The data aligned by both methods were then used to demonstrate the use of PLS-DA for the identification of IL-containing arson debris. These PLS-DA models were constructed using a fully-automated feature selection and model construction routine which required no user intervention to find the relatively small number of features relevant for discriminating between debris containing an IL and debris that did not contain an IL. The ~300 features selected (0.015% of the original data) were found to be derived from seven compounds consistent with the presence of gasoline.

**5. Acknowledgements**

We wish to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and Alberta Innovates Technology Futures for funding this research through a New Faculty Award to J.J. Harynuk.

**6. References**

**Table 1.** Compositions of arson debris ; HDPE = high density polyethylene; MDF = medium density fibreboard

|  |  |
| --- | --- |
| Debris | Contents |
| 1 | Wood, carpet, cotton fabric, 50:50 cotton:polyester fabric, glossy magazine |
| 2 | Wood, polyethylene film, old linoleum, carpet |
| 3 | Wood, siding, shingles, tar paper, Tyvek™ |
| 4 | Wood |
| 5 | Wood, polyethylene film, newpaper |
| 6 | Wood, HDPE food container, new linoleum, MDF |
| 7 | Wood, carpet, cotton fabric, 50:50 cotton:polyester fabric, glossy magazine, linoleum, siding, shingle, tarpaper, Tyvek™ |

**Table 2.** Results of cross-validation of PLS-DA models.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Deuterated alkane ladder-based alignment | | | | | | | | | | |
| set number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Number of  variables selected1 | 320 | 320 | 300 | 40 | 30 | 260 | 220 | 270 | 220 | 240 |
| False positive | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| False negative | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Minimum probability  for true positive | 1 | 0.9998 | 1 | 0.9988 | 1 | 0.9999 | 0.9976 | 0.9991 | 0.9998 | 0.9966 |
| Maximum probability  for true negative | 0.0001 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|  | | | | | | | | | | |
| Peak matching-based alignment | | | | | | | | | | |
| set number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Number of  variables selected1 | 460 | 570 | 400 | 390 | 200 | 400 | 430 | 520 | 520 | 500 |
| False positive | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| False negative | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| Minimum probability  for true positive | 0.6592 | 0.5782 | 0.9199 | 0.7636 | 0 | 0.8852 | 0.7468 | 0.9968 | 0.8404 | 0.7509 |
| Maximum probability for true negative | 1 | 1 | 1 | 0.0355 | 1 | 0.1737 | 0.1866 | 0.7709 | 0.0759 | 0.0674 |

1Minimum number of variables required to reach a cluster resolution of 99.99%, or the maximum achievable resolution in cases where 99.99% was unachievable using the training data.

**Table 3.** Tentative identification of compounds from which the features responsible for identifying gasoline are derived. \*Literature values are for 6 °C⋅min-1 taken from [34].

|  |  |  |  |
| --- | --- | --- | --- |
| **Compound** | **Estimated Retention Index** | **Literature Retention Index\*** | **Relevant Masses** |
| Ethylbenzene | 870 | 857 | 51,77,91,106 |
| Para-xylene | 880 | 866 | 52,76,98,106 |
| Ortho-xylene | 908 | 891 | 105,106 |
| Propylbenzene | 980 | 952 | 62,89,91,120 |
| 1-ethyl-3-methyl benzene | 989 | 961 | 41,65,91,120 |
| 1,2,4-trimethyl benzene | 996 | 967 | 91,103,105,120 |
| 1,3,5-trimethyl benzene | 1022 | 992 | 41,89,116,119 |

**Figure Captions:**

Figure 1: Deuterated alkane ladder signal for retention alignment superimposed over a TIC of a debris sample. Red trace is the ladder signal constructed from the product of the abundances of ions m/z = 50, 66, 82, and 98.

Figure 2: Score plots for PLS-DA models for the worst classification (Set 5) and one of the best (Set 10) when considering feature-based alignment. Ladder-based results are shown for comparison. Samples containing an IL are indicated by blue circles; samples with no IL are indicated by red triangles. Samples in the training set are indicated by hollow markers while samples in the test set are indicated by filled markers. Confidence ellipses are constructed at 95 % confidence intervals using data in the training set. **A** Feature-based alignment, Set 5; **B** Ladder-based alignment, Set 5; **C** Feature-based alignment, Set 10; **D** Ladder-based alignment, Set 10.

Figure 3: Predicted Y-value plots for PLS-DA models for Set 5 and Set 10 using both feature-based and ladder-based alignment. Samples containing an IL are indicated by blue circles; samples with no IL are indicated by red triangles. Samples in the training set are indicated by hollow markers while samples in the test set are indicated by filled markers. **A** Feature-based alignment, Set 5; **B** Ladder-based alignment, Set 5; **C** Feature-based alignment, Set 10; **D** Ladder-based alignment, Set 10.

Figure 4: Predicted probability plots for identifying ILs in samples for PLS-DA models for Set 5 and Set 10 using both feature-based and ladder-based alignment. Samples containing an IL are indicated by blue circles; samples with no IL are indicated by red triangles. Samples in the training set are indicated by hollow markers while samples in the test set are indicated by filled markers. **A** Feature-based alignment, Set 5; **B** Ladder-based alignment, Set 5; **C** Feature-based alignment, Set 10; **D** Ladder-based alignment, Set 10.

Figure 5: Region of GC-MS data containing features of interest, plotted as a contour plot. Retention time is indicated on the x-axis and m/z on the y-axis. **A** Raw GC-MS chromatogram; **B** Features of interest in the raw data identified by algorithm.

Figure 1:

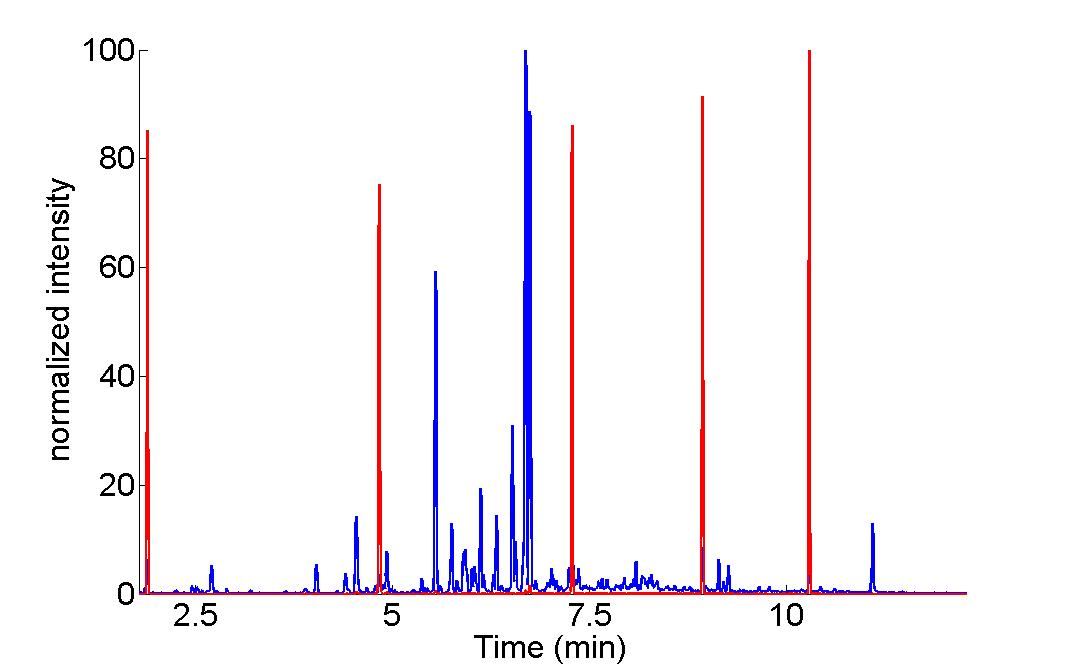


Figure 2:

A:



B:



C:



D:



Figure 3:

A:



B:



C:



D:



Figure 4:

A:



B:



C:

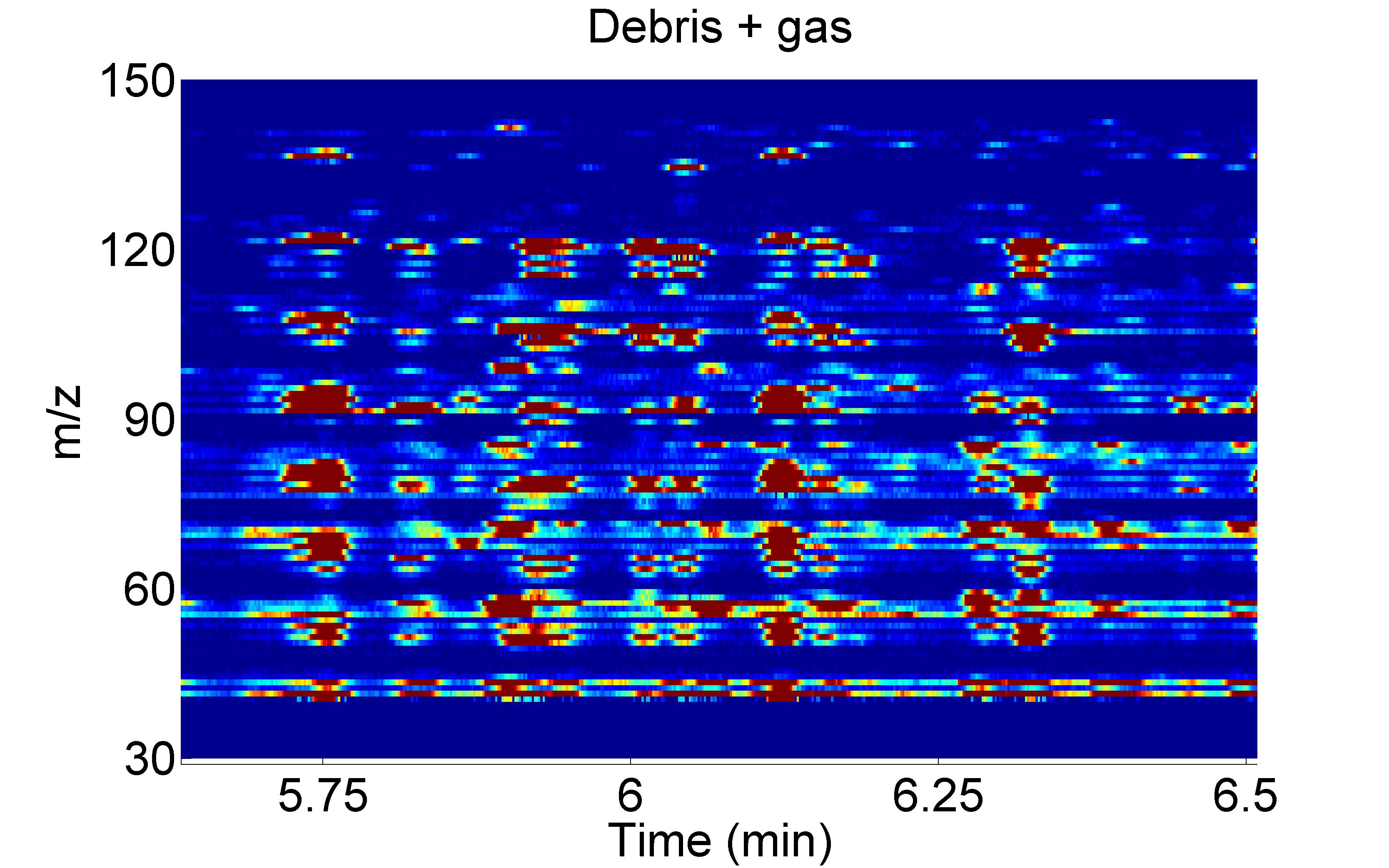


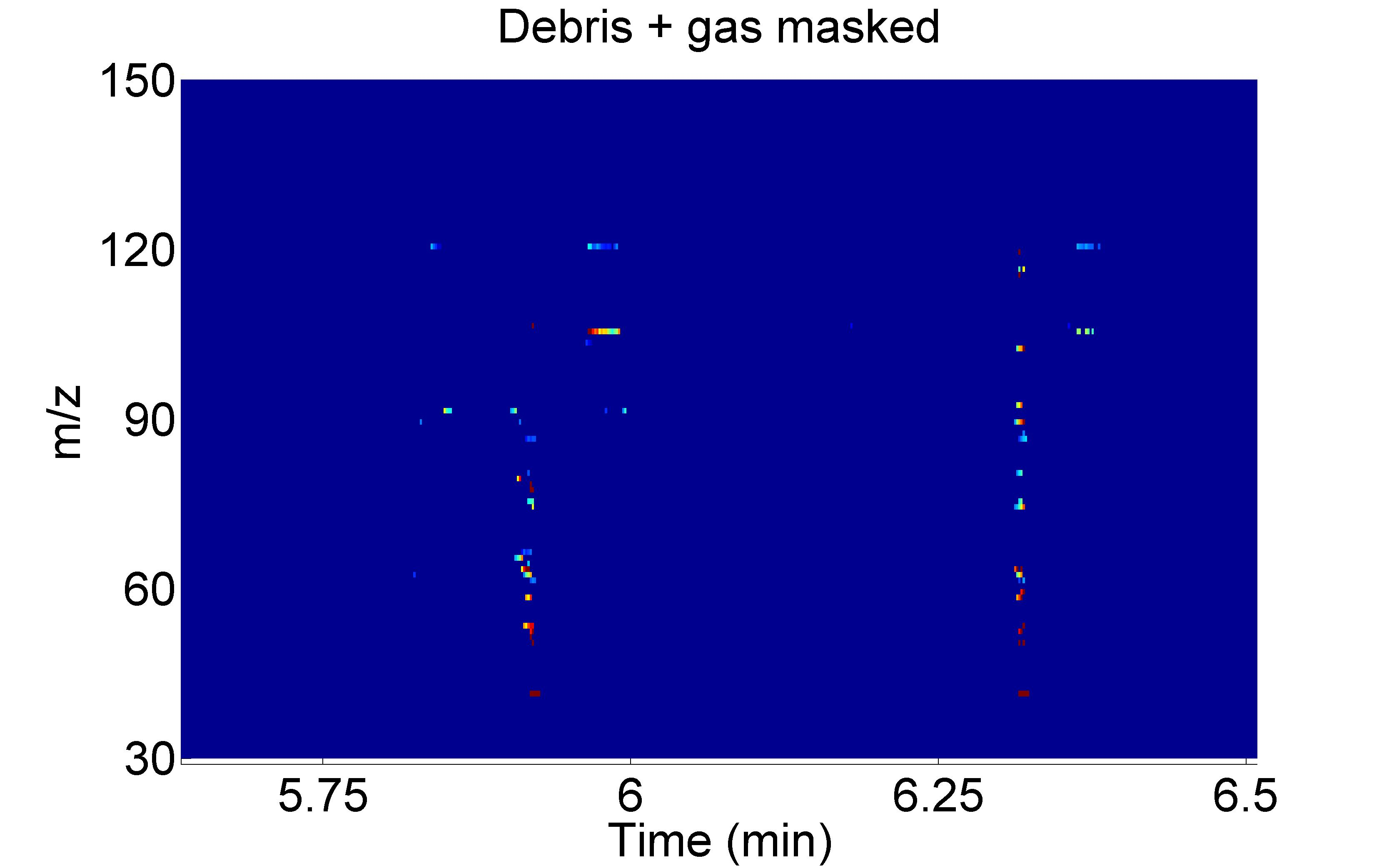
D:



Figure 5:

A:



B:

1. [] P. Doble, P.M.L. Sandercock, E. Du Pasquier, P. Petocz, C. Roux, M. Dawson, Forensic Sci. Int. 132 (2003) 26-39 [↑](#endnote-ref-1)
2. [] R.B. Gaines, G.J. Hall, G.S. Frysinger, W.R. Gronlund, K.L. Juaire, Enviro. Forensics 7 (2006) 77-87 [↑](#endnote-ref-2)
3. [] S. García-Martín, C. Herrero, R.M. Peña, J. Barciela, Food Chem. 118 (2010) 456-461 [↑](#endnote-ref-3)
4. [] M. Barker, W. Rayens, J. Chemometrics 17 (2003) 166-173 [↑](#endnote-ref-4)
5. [] R. t’Kindt, K. Morreel, D. Deforce, W. Boerjan, J. van Bocxlaer, J. Chromatogr. B 877 (2009) 3572-3580 [↑](#endnote-ref-5)
6. [] A. Kende, D. Portwood, A. Senior, M. Earll, E. Bolygo, M. Seymour, J. Chromatogr. A 1217 (2010) 6718–6723 [↑](#endnote-ref-6)
7. [] S. Lin, N. Liu, Z. Yang, W. Song, P. Wang, H. Chen, M. Lucio, P. Schmitt-Kopplin, G. Chen, Z. Cai, Talanta 83 (2010) 262–268 [↑](#endnote-ref-7)
8. [] J. Vial, H. Noçairi, P. Sassiat, S. Mallipatu, G. Cognon, D. Thiébaut, B. Teillet, D.N. Rutledge, J.Chromatogr. A 1216 (2009) 2866-2872 [↑](#endnote-ref-8)
9. [] K. J. Johnson, B. W. Wright, K. H. Jarman, R. E. Synovec, J Chromatogr. A 996 (2003) 141-155 [↑](#endnote-ref-9)
10. [] K. J. Johnson, R. E. Synovec, Chemom. Intell. Lab. Syst. 60 (2002) 225-237 [↑](#endnote-ref-10)
11. [] N-P.V. Nielsen, J. M. Carstensen, J. Smedsgaard J Chromatogr. A 805 (1998) 17-35 [↑](#endnote-ref-11)
12. [] K.M. Åberg, E. Alm, R.J.O. Torgrip, Anal. Bioanal. Chem. 394 (2009) 151-162 [↑](#endnote-ref-12)
13. [] G. Tomasi, F. van den Berg, C. Andersson, J. Chemometrics 18 (2004) 231-241 [↑](#endnote-ref-13)
14. [] C.A. Smith, E.J. Want, G. O’Maille, R. Abagyan, G. Siuzdak, Anal. Chem. 78 (2006) 779–787. [↑](#endnote-ref-14)
15. [] M.D. Robinson, D.P. De Souza, W.W. Keen, E.C. Saunders, M.J. McConville, T.P. Speed, V.A. Likic, BMC Bioinform. 8 (2007) 419. [↑](#endnote-ref-15)
16. [] R.G. Brereton, Applied Chemometrics for Scientists, John Wiley & Sons, Inc. Toronto, 2007. [↑](#endnote-ref-16)
17. [] N. A. Sinkov, J. J. Harynuk, Talanta 83 (2011) 1079–1087 [↑](#endnote-ref-17)
18. [] L.J. Marshall, J.W. McIlroy, V.L. McGuffin, R. Waddell Smith, Anal. Bioanal. Chem. 394 (2009) 2049-2059 [↑](#endnote-ref-18)
19. [] M.D. Krebs, R.D. Tingley, J.E. Zeskind, M.E. Holomboe, J. Kang, C.E. Davis, Chemom. Intell. Lab. Syst. 81 (2006) 74-81. [↑](#endnote-ref-19)
20. [] C.R. Borges, Anal. Chem. 79 (2007) 4805-4813. [↑](#endnote-ref-20)
21. [] N.E. Watson, M.M. VanWingerden, K.M. Pierce, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1129 (2006) 111-118. [↑](#endnote-ref-21)
22. [] K.M. Pierce, J.K. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1096 (2005) 101-110. [↑](#endnote-ref-22)
23. [] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.M. Myhr, O.M. Kvalheim, Anal. Chem. 81 (2009) 2581-2590. [↑](#endnote-ref-23)