analytical chemistry

Article

pubs.acs.org/ac

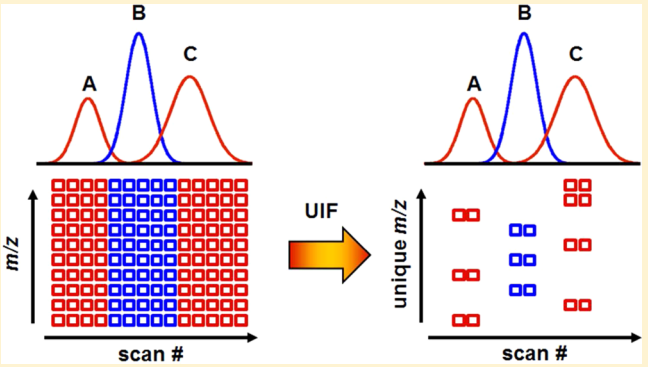
Unique Ion Filter: A Data Reduction Tool for GC/MS Data Preprocessing Prior to Chemometric Analysis

L. A. Adutwum and J. J. Harynuk*

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

Supporting Information

ABSTRACT: Using raw GC/MS data as the X-block for chemometric modeling has the potential to provide better classification models for complex samples when compared to using the total ion current (TIC), extracted ion chromatograms/profiles (EIC/EIP), or integrated peak tables. However, the abundance of raw GC/MS data necessitates some form of data reduction/feature selection to remove the variables containing primarily noise from the data set. Several algorithms for feature selection exist; however, due to the extreme number of variables (10^6–10^8 variables per chromatogram), the feature selection time can be prolonged and computationally expensive. Herein, we present a new prefilter for automated data reduction of GC/MS data prior to feature selection. This tool, termed unique ion filter (UIF), is a module that can be added after chromatographic alignment and prior to any subsequent feature selection algorithm. The UIF objectively reduces the number of irrelevant or redundant variables in raw GC/MS data, while preserving potentially relevant analytical information. In the m/z dimension, data are reduced from a full spectrum to a handful of unique ions for each chromatographic peak. In the time dimension, data are reduced to only a handful of scans around each peak apex. UIF was applied to a data set of GC/MS data for a variety of gasoline samples to be classified using partial least-squares discriminant analysis (PLS-DA) according to octane rating. It was also applied to a series of chromatograms from casework fire debris analysis to be classified on the basis of whether or not signatures of gasoline were detected. By reducing the overall population of candidate variables subjected to subsequent variable selection, the UIF reduced the total feature selection time for which a perfect classification of all validation data was achieved from 373 to 9 min (98% reduction in computing time). Additionally, the significant reduction in included variables resulted in a concomitant reduction in noise, improving overall model quality. A minimum of two um/z and scan window of three about the peak apex could provide enough information about each peak for the successful PLS-DA modeling of the data as 100% model prediction accuracy was achieved. It is also shown that the application of UIF does not alter the underlying chemical information in the data.



Gas chromatography/mass spectrometry (GC/MS) is a versatile tool that has been applied in various fields of chemical analysis including environmental, pharmaceutical, petrochemical, and forensics, among others. This is due to the remarkable separation power of the GC and the rich multivariate data generated by the MS detector. Mass spectrometers such as time-of-flight MS (TOF-MS) or even modern high-speed quadrupole MS (qMS) systems are capable of rapidly acquiring spectra and generating data containing several thousands of spectra per sample. This renders data interpretation daunting, especially when dealing with complex samples. The underlying chemical information can be obscured by the enormity of the data. Chemometric techniques involve the use of statistical and computational methods to extract useful information from complex chemical data and have become very useful.[1,2] Reviews by Levine and Workman have highlighted the application of chemometrics in various fields of analytical chemistry.[3,4] Supervised pattern recognition techniques, for example, partial least-squares discriminant analysis (PLS-DA), and unsupervised exploratory techniques such as principal component analysis (PCA) and cluster analysis have been applied to the interpretation of various types of GC/MS data. Chemometric techniques have been used in the identification of jet fuels,[5] classification and chemical fingerprinting of gasoline,[6−11] tracking and weathering of oil spills,[12,13] classification of casework arson samples,[14] classification of vinegars and wines,[15−17] biomarker identification,[18,19] drug discovery and verification of herbal medicines,[20,21] and compound identification[22,23] as well as metabolomics and breath analysis.[24−28]

Raw GC/MS data presents as a two-dimensional matrix with rows representing mass-to-charge ratio (m/z) and columns representing time (scan #). High data rate mass analyzers are

Received: May 6, 2014
Accepted: July 6, 2014
Published: July 6, 2014

   dx.doi.org/10.1021/ac501660a | Anal. Chem. 2014, 86, 7726−7733

desirable since they allow for rapid separations and provide sufficient data density along the time axis to ensure accurate peak description, especially for very narrow peaks.[29] However, these detectors deliver a huge amount of data (>10^6 data points per chromatogram) which complicates the data analysis. Prior to chemometric analysis, the data are subjected to various preprocessing techniques such as retention alignment, baseline correction, smoothing (noise removal), scaling, and data simplification or reduction.[30,31]

Data reduction is of particular importance for GC/MS data due to the sheer number of variables. Common approaches to data reduction for GC/MS include the use of integrated peak areas based on total ion currents (TICs) or mass spectrally deconvoluted data.[6,13,25,32,33] This approach is very simple and computationally inexpensive but may oversimplify the data, losing the m/z dimension, which could otherwise provide useful information. Selection of signals from one or a few m/z channels, known as extracted ion chromatograms (EICs), is also a common approach. EICs are useful for well-characterized samples in well-understood systems, but there is a risk of accidentally removing informative ions if the system is not well-understood. Additionally, this approach includes many variables containing only noise (baseline variables). Combined, this makes the EIC approach somewhat subjective and of little use when modeling a poorly understood data set (e.g., biomarker discovery). The advantage of using the entire GC/MS chromatogram has been demonstrated and applied to very complex samples.[5,10,11,14,34] In these works, the entire GC/MS chromatogram is unfolded along one axis into a single vector, which makes each m/z at each scan an independent variable. This results in several thousands or millions of variables for each sample and produces a huge data set, which is computationally expensive to manipulate.

The use of such a high number of variables for building chemometric models is prohibitive due to the sheer size of the data; moreover, the majority of the variables will not provide useful information for the chemometric model that is being built and their inclusion will be detrimental to the model.[35] To overcome this challenge, relevant variables are obtained using feature ranking and feature selection protocols.[5,36−39] Synovec et al. employed a threshold-based feature selection based on the Fisher ratio from analysis of variance (ANOVA) and selected a number of top-ranked variables.[5] The use of selectivity ratio as a feature ranking technique has also be reported.[11,36] The ranking metric provides a starting point for identifying the variables with a high potential to provide useful information, though a highly ranked variable may not necessarily be the most useful variable in the chemometric model, and similarly, a lower-ranked variable may prove crucial. Thus, a strategy to test and identify a subset of the most informative variables becomes necessary. While there are multiple feature selection algorithms that could be used, we have previously demonstrated the use of a cluster resolution (CR)-guided, hybrid backward elimination/forward selection (BE/FS) algorithm for feature selection.[10,11,14,34]

Briefly, the algorithm creates an initial model using a fraction of top-ranked variables (e.g., by Fisher scores or selectivity ratio). The quality of the model is evaluated using CR. During the BE step, the effect of discarding a single variable is evaluated. If discarding the lowest-ranked variable improves the model, the variable is discarded; otherwise, it is returned to the model, and then the next-lowest-ranked variable is tested. In the FS step, the variables that were not included in the initial BE step are tested sequentially to see if their inclusion improves the model based on the variables that survived the BE step. CR is based on the calculation of the size of the confidence ellipse or ellipsoid that can be described around each cluster of points without overlap in either PCA or PLS-DA scores space.

In theory, an exhaustive test on all variables should be performed; however, this is impractical and unnecessary in the case of GC/MS data where high data rate detectors are used, as the vast majority of the variables are uninformative. When studying the results of earlier research, it was found that several hundreds or even thousands of variables were selected for a single chromatographic peak.[10,34] This number of variables selected for each peak points to the potential for excessive redundancy in the selected features. In principle, redundancy in the data is helpful as the presence of multiple variables providing identical chemical information adds stability to a model as they reinforce each other. However, there is likely a point where the benefits of redundancy are outweighed by the additional noise and computing requirements needed to handle the extra data. This excessive redundancy in the data could lead to over fitting the training set data and/or confusion of the learning algorithm, in this case, the feature selection process.[39−42] Hence, a reduction in the number of candidate variables and variable redundancy should lead to faster, more effective and efficient variable selection and ultimately contribute to the construction of a more parsimonious chemometric model.

In this paper, we present a preprocessing technique termed unique ion filter (UIF) for automated GC/MS data reduction prior to chemometric analysis (Figure S1, Supporting Information). Data reduction is achieved by reducing the number of ions retained for each peak to a few of the most abundant, unique ions (um/z) within a specified scan window around each peak apex. Essentially, the UIF objectively filters each raw GC/MS chromatogram independently to remove variables that are likely unimportant or redundant in a chromatographic sense. Using this approach, there is the potential for a drastic reduction in the number of variables passed to the feature selection step without losing the multivariate nature of the data. There are two expected outcomes of the variable reduction. Obviously, by reducing the total number of variables under consideration, there should be a significant reduction in computational time for feature selection. The second outcome is more important, though less obvious. The number of included variables in the final model should be decreased, with a concomitant reduction in included noise and artifacts, resulting in more parsimonious models.

EXPERIMENTAL SECTION

A data set used for a previously published work[10] was used in this proof-of-principle work. Briefly, the data comprise a series of GC/MS chromatograms from a set of gasoline samples to be classified according to their octane ratings (87, 89, and 91 octane). For each class of gasoline, 24 chromatograms were obtained.

The entire chromatogram for each sample was imported into Matlab 2013a (The Mathworks, Natick, MA) as a 7500 × 271 (scan number × m/z) matrix. All data were handled with Matlab algorithms written in-house. Chemometric models were constructed using PLS Toolbox 7.3 (Eigenvector Research Inc., Wenatchee, WA). All chemometric analyses were performed on

dx.doi.org/10.1021/ac501660a | Anal. Chem. 2014, 86, 7726−7733

a MacBook Pro running on a core 2.9 GHz i7 Intel processor and 16 GB RAM.

THEORY

Algorithm for UIF. UIF is an additional preprocessing technique that is applied to individual sample chromatograms after alignment and prior to feature selection (Figure S1, Supporting Information). There are two main inputs, which are the maximum number of unique ions (um/z) to be retained for each peak and the number of scans surrounding the peak apex to be included. In further discussion, the notation of $\mathrm{UIF}_{(p,w)}$ is used where p is the number of unique ions to retain for each peak and w is the width of the window around the peak apex (an odd number). For example, w = 5 would indicate that a window of five scans (the peak apex plus two scans to either side of the apex) would be retained. Accurate peak detection is necessary for effective application of UIF, including retention times and peak widths. In principle, any peak detection algorithm that is capable of detecting peak apexes, starts, and stops can be used.

Determination of Peak Parameters. The main parameters critical to UIF are peak apex locations and the determination of any peak overlap with neighboring peaks. Any robust peak finding algorithm can be used for the determination of these peak parameters. In this proof-of-concept work, a laboratory written peak detection algorithm based on the aligned total ion current (TIC) signal was used.

The TIC was generated by summing the chromatogram in the scan dimension (eq 1), where $\mathbf{X}$ is the raw chromatogram, $\mathbf{z}$ is the TIC vector, i is the scan number, j is the m/z, and J is the total number of ions.

$$\mathbf{z}_i = \sum_{j=1}^{J} \mathbf{X}_{(i,j)} \tag{1}$$

A second-derivative Savitsky-Golay smoothing vector ($\mathbf{s}$) is generated and applied to the TIC vector ($\mathbf{z}$) to generate smoothed second-derivative **sdz**, according to eq 2, where **sdz** is the second derivative vector, $\mathbf{s}$ is second-derivative Savitsky-Golay smoothing vector, $\mathbf{z}$ is the TIC, f is the smoothing window, and n is the length of $\mathbf{z}$.

$$\mathbf{sdz}_i = \mathbf{s}^T \times \mathbf{z}_{\left(i-\frac{f-1}{2}:i+\frac{f-1}{2}\right)};\ \left(\frac{f+1}{2}\right) \le i \le \left(n - \frac{f-1}{2}\right) \tag{2}$$

Subsequently, peak apex and peak inflection points are identified. Peak apexes are determined as the lowest valley point with a negative value on **sdz**. Peak inflection points are obtained from two positive maxima neighboring a negative minimum of an apex location on the **sdz** vector. For this work, peaks were assumed to be Gaussian, and the peak widths (4σ) were estimated from the inflection points of each peak ($\pm\sigma$).

Three different types of peak groups can be identified from peak start and peak stop locations (Figure S2, Supporting Information). Group A are resolved peaks, where peak start and peak stop locations do not overlap with any adjacent peaks. Groups B1 and B2 are peaks with either front or tail overlap only, and Group C are sandwiched peaks; i.e., both start and stop locations overlap with neighboring peaks. The peak resolution information in addition to the user specified number of um/z and scans around peak apexes to be used are then passed to the UIF algorithm (Figure S3, Supporting Information).

It is important to note that with this particular peak detection algorithm there must be sufficient chromatographic resolution between a pair of peaks such that a valley appears between their apexes in order for the peaks to be identified as two separate peaks (Resolution ~0.7 for peaks with equal heights). Severely coeluting peaks (i.e., those with no valley between their apexes) appear to this algorithm as a single peak and are treated together. Thus, it is possible that a minor peak coeluting with a major peak could be lost if throughout every chromatogram in the series its intensity is not high enough to have one of its ions selected as a um/z for the sum of the coeluting spectra. However, this limitation is a reflection on the peak detection algorithm used herein and not on the UIF itself. Improved peak detection algorithms that can deconvolute severely coeluting peaks could also be used in conjunction with the UIF and would be expected to yield improved results in the cases of severely coeluting compounds.

Identification of Unique Ions. The signals at all peak apexes for a chromatogram are extracted into a matrix ($\mathbf{Y}$) with dimension number peaks (n) × m/z. The extracted signals in $\mathbf{Y}$ are converted into a mass spectrum matrix, $\mathbf{Y}_{\mathrm{MS}}$, according to eq 3 where $\mathbf{Y}_{\mathrm{MS}}$ is the mass spectrum at the apexes, n is the peak number, and j is m/z.

$$\mathbf{Y}_{\mathrm{MS}(n,j)} = \frac{\mathbf{Y}_{(n,j)}}{\sum_{j=1}^{J} \mathbf{Y}_{(n,j)}} \tag{3}$$

The group (A, B, C; above) into which a peak falls controls how um/z are identified for that peak. Unique ions are stored in $\mathbf{U}$ (initially, a matrix of zeros having the same dimensions as $\mathbf{Y}_{\mathrm{MS}}$). Thus, for n = 1, 2, 3, ... N, where N is the total number of peaks in the chromatogram, if peak n belongs to Group A, then all m/z in $\mathbf{Y}_{\mathrm{MS}}$ (n, j = 1, 2, 3, ... J) are um/z to peak n and all ions above a minimum threshold are retained in $\mathbf{U}$ by setting their coordinates in $\mathbf{U}$ = 1.

If peak n is a member of B1 or B2, the relative abundance vector $\mathbf{v}$ is generated according to eq 4 or 5, respectively, where j = 1, 2, 3, ... J.

$$\mathbf{v} = \frac{\mathbf{Y}_{\mathrm{NORM}(n,j)}}{\mathbf{Y}_{\mathrm{NORM}(n-1,j)}} \tag{4}$$

$$\mathbf{v} = \frac{\mathbf{Y}_{\mathrm{NORM}(n,j)}}{\mathbf{Y}_{\mathrm{NORM}(n+1,j)}} \tag{5}$$

Since $\mathbf{v}$ is a vector of the relative abundances of m/z, elements of $\mathbf{v}$ greater than 1 have higher abundances in peak n relative to (n − 1) in (4) or (n + 1) in (5). Truly unique ions in $\mathbf{v}$ will have a value of ∞, while pseudounique ions will have a large value. Elements of $\mathbf{v}$ above a certain uniqueness threshold are deemed to be um/z of peak n, and their coordinates in $\mathbf{U}$ are set to a value of 1.

Finally, if peak n is in Group C (i.e., a peak with a coelutant on both sides), two abundance vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are calculated using eqs 4 and 5, respectively, and ions in $\mathbf{v}_1$ and $\mathbf{v}_2$ that exceed the uniqueness threshold are set to a value of 1. A third vector $\mathbf{v}_3$ is then generated from the diagonal of the outer product of $\mathbf{v}_1{}^T$ and $\mathbf{v}_2$. This vector $\mathbf{v}_3$ is composed of zeros, with ones located at positions indicating ions that are unique (or pseudounique) to peak n in the cluster of three peaks. The coordinates of these um/z are set to a value of 1 in $\mathbf{U}$.

dx.doi.org/10.1021/ac501660a | Anal. Chem. 2014, 86, 7726−7733

The resulting matrix **U** is a sparse matrix of zeros and ones with the ones indicating the positions of *um/z* for each peak. A Hadamard product of **U** and $\mathbf{Y}_{MS}$ yields **V** ($\mathbf{V} = \mathbf{U} \circ \mathbf{Y}_{MS}$), a matrix of the raw abundance of each *um/z*. On the basis of the user-input number of unique ions to be chosen, *p*, the *m/z* positions of the *p* most abundant unique ion(s) for each peak can be obtained.

Generation of New Chromatogram. In the final step of the UIF, a mask of zeros, **M**, of same size as the original data is generated and modified such that ones are placed at the coordinates where the *p* most-abundant unique ions in each detected peak for a width of *w* scans in the scan direction, centered on the peak apex. A Hadamard product of **M** and the original data matrix **X** results in the unique ion filtered data, $\mathrm{UIF}_{(p,w)} = \mathbf{M} \circ \mathbf{X}$.

Chemometric Analysis. Chromatograms were imported from .csv files and aligned using an algorithm written in-house[10] which is based on a piecewise alignment algorithm.[43] The total of 72 samples were split into a training set (8 samples per class), optimization set (8 samples per class), and validation set (8 samples per class). In the benchmark work, all the chromatograms were unfolded in the scan dimension yielding a vector of 2 032 500 variables for each chromatogram. A data set matrix, 72 samples × 2 032 500 variables resulted. Variable positions where all samples had no signal intensity above a minimum threshold (in this work, 150 counts) were removed from consideration. Feature ranking was then performed with the training set data using an ANOVA-based ranking technique reported earlier.[10,24] The training and optimization data sets were used for the cluster resolution variable selection procedure as done previously. Variables that passed the feature selection process were used for chemometric analysis. Using the combined training and optimization sets, the selected features in each chromatogram were autoscaled and normalized to a value of 1 and then used to construct PLS-DA models. Model quality was assessed on the basis of the ability of the model to correctly predict the validation set data. The UIF evaluation pathway followed the same process, except that after alignment the UIF was applied to all samples prior to unfolding and subsequent feature ranking and selection steps.

Specificity, sensitivity, and accuracy of each optimized model were calculated on the basis of validation data and used as an objective parameter in comparing model quality for both routes.[44] Sensitivity measures the model's ability to correctly classify positive results, i.e., true positive rate (sensitivity = true positives/number of positives). Specificity is the measure of the model's ability to correctly classify or predict negative results, i.e., true negative rate (specificity = true negatives/number of negatives). Accuracy is the measure of true results (accuracy = (sensitivity + specificity)/2). These parameters present values on a scale of 0 to 1, with 0 being the worst model and 1 being the best model.

RESULTS AND DISCUSSION

The UIF offers a convenient approach for automated, objective binning of GC/MS data that preserves the multivariate information contained in the *m/z* dimension. Two principal inputs, the number of *um/z* (*p*) and the scan window (*w*), are required. Since the user does not decide which ions are unique to each peak, the subjectivity and the risk of losing otherwise relevant data are largely reduced. UIF reduces the number of variables per peak by focusing on ions unique to each peak at the peak apex.

For the data set used in this study, unfolding the 72 chromatograms without UIF application resulted in a matrix of 72 samples × 2 032 500 variables. After removing null variables, i.e., columns having no signal above a minimal threshold (150 counts) for all chromatograms, the number of variables was reduced to 1 668 403 (i.e., 72 samples × 1 668 403 variables). When the UIF was applied and all the *um/z* across the entire width of each peak were retained, the maximum number of variables was reduced to 225 830 (i.e., 72 samples × 225 830 variables) representing an 86% reduction in the number of variables from the original data set (after removal of null variables). Selecting only a few *um/z* for only a few central scans on each peak will further reduce the size of the matrix to be considered by subsequent feature ranking and selection routines.

For comparative purposes, we benchmark this work without the UIF at the minimum number of top-ranked variables that must be tested to achieve an excellent model prediction quality (sensitivity, specificity, and accuracy of 1) for all classes using ANOVA ranking and our hybrid BE/FS approach. We chose this approach because it was readily available and has demonstrated success in handling entire raw GC/MS chromatograms.[10,11,14,34] Fundamentally, the feature selection method used on the GC/MS data is of little-to-no importance to the efficacy or applicability of the UIF. Regardless of the feature selection (and possible variable ranking) methods used, the UIF will improve the situation as it will reduce the number of candidate variables that must be considered, typically by 1−3 orders of magnitude (as will be shown below). In Figure 1, an

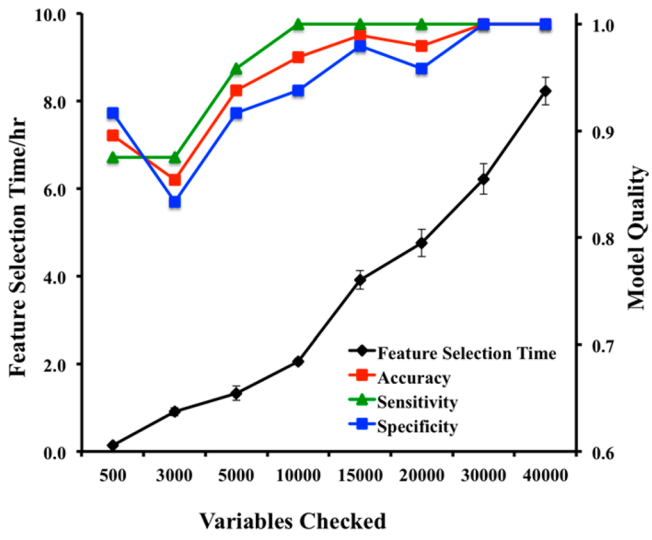


Figure 1. Feature selection time and model quality plot for benchmark pathway.

increase in the model sensitivity, specificity and overall accuracy are observed, commensurate with an increase in the number of top-ranked variables checked during the feature selection process. A model that achieved a sensitivity, specificity, and accuracy of 1.0 was achieved when 30 000 top-ranked variables were tested. It must also be noted that increasing the maximum number of features tested also increases the computation time for the feature selection process. PLS-DA Y-predicted plots for the three octane ratings using features selected by the benchmark algorithm without the UIF are presented in Figure 2. Here, a total of 3001 of the 30 000 tested variables passed the

dx.doi.org/10.1021/ac501660a | Anal. Chem. 2014, 86, 7726−7733

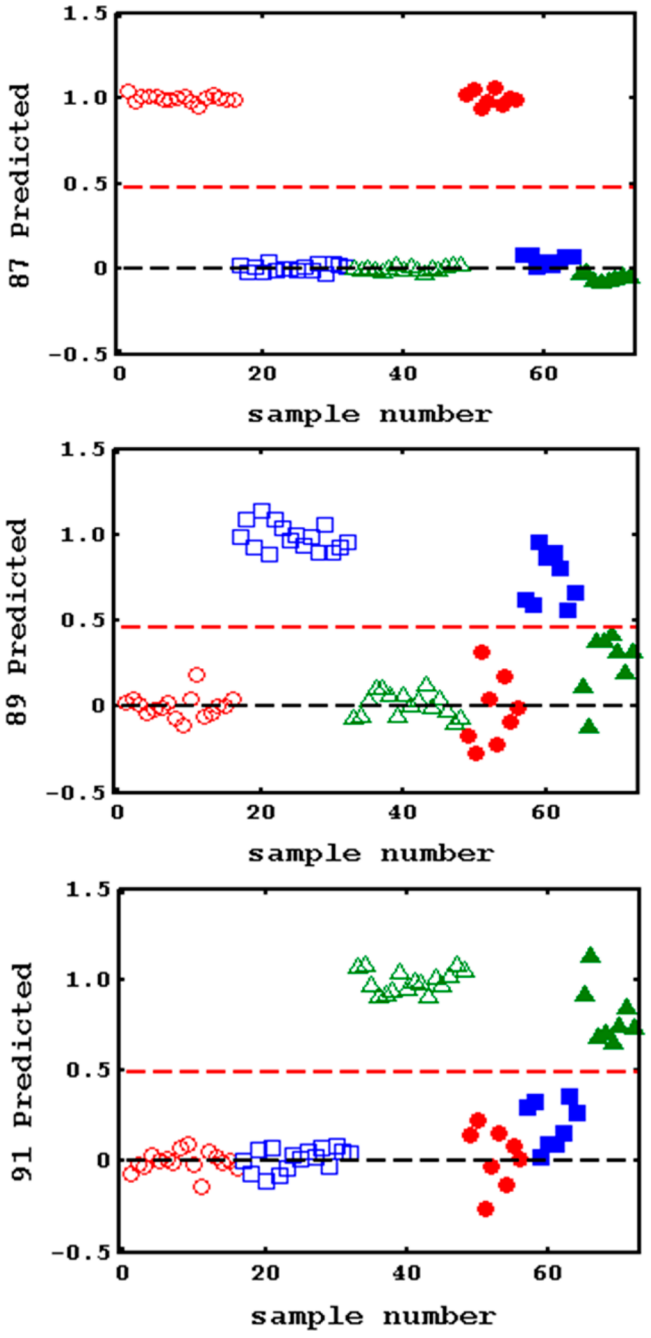


Figure 2. Y-predicted plot for PLS-DA classification of gasoline samples after feature selection but without application of UIF. Red circles, blue squares, and green triangles indicate 87, 89, and 91 octane ratings of gasoline, respectively. Hollow markers indicate training and optimization while solid markers indicate validation set.

feature selection process. The predicted plots for gasoline with 87 (solid red circles), 89 (solid blue squares), and 91 (solid green triangles) octane ratings show that all validation set samples were predicted with a 100% prediction sensitivity, specificity, and accuracy.

To compare the effect of UIF on the feature selection process and ultimately the quality of the chemometric model to that of the benchmark, multiple combinations of p (number of um/z) and w (window about apex) were investigated. um/z ranging from 1 to 10 and scan windows of 1 to 17 (odd numbers only) were investigated. The number of variables to be passed to the feature selection algorithm after the application of the UIF ranged from 3717 for $UIF_{(1,1)}$ to 107 982 for $UIF_{(10,17)}$. Due to this reduction in the total number of variables, the number of top-ranked variables submitted to the variable selection process was limited to 500. These experiments show that, at a scan number of 1 (i.e., only ions at the peak apex are retained), an increase in the number um/z considered does not improve the model (Figure 3). However,

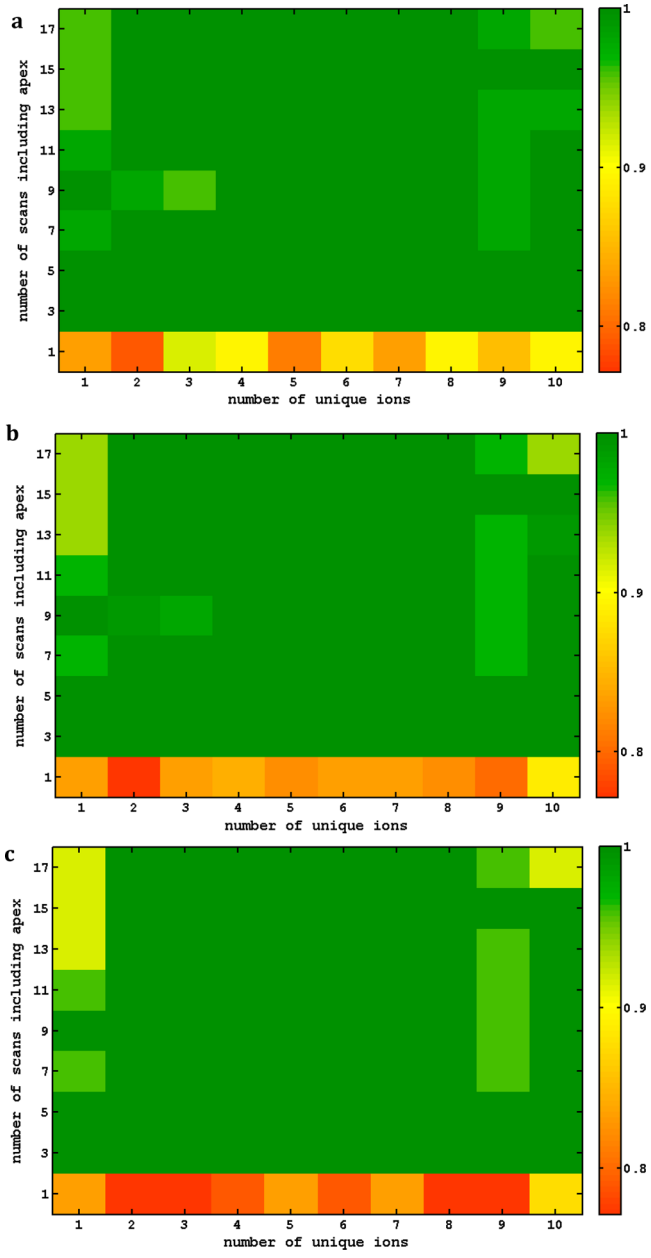


Figure 3. Sensitivity (a), specificity (b), and accuracy (c) of UIF experiments. See Tables S3−S5 in the Supporting Information for numerical results.

increasing w to 3, even when considering a single um/z per peak, significantly improves model quality. This is likely due to lessening the effects of minor shifts in peak position and allowing some additional reinforcing variables containing nearly identical information to be considered. The increase in w may also allow some information about the peak's profile to be retained. For this particular data set, a minimum of two um/z

dx.doi.org/10.1021/ac501660a | Anal. Chem. 2014, 86, 7726−7733

and three scans is necessary to achieve 100% model prediction sensitivity, specificity, and accuracy (Figure S4, Supporting Information).

The PLS-DA Y-predicted plots for the three classes of samples when $UIF_{(2,5)}$ was applied prior to feature selection are shown in Figure 4. This result is comparable to the benchmark when 30 000 top ranked variables were checked during feature selection. However, the model presented in Figure 4 is likely a more robust model since the validation data for 87 , 89, and 91 octane project further away from the class discrimination boundary (red line in plots). Additionally, the Y-predicted positive and negative values for the samples are much closer to the ideal values of 1 and 0, respectively, and have clustered closer together relative to the benchmark case. This indicates a significant reduction in within-class variance, likely due to the exclusion of redundant variables and excess noise.

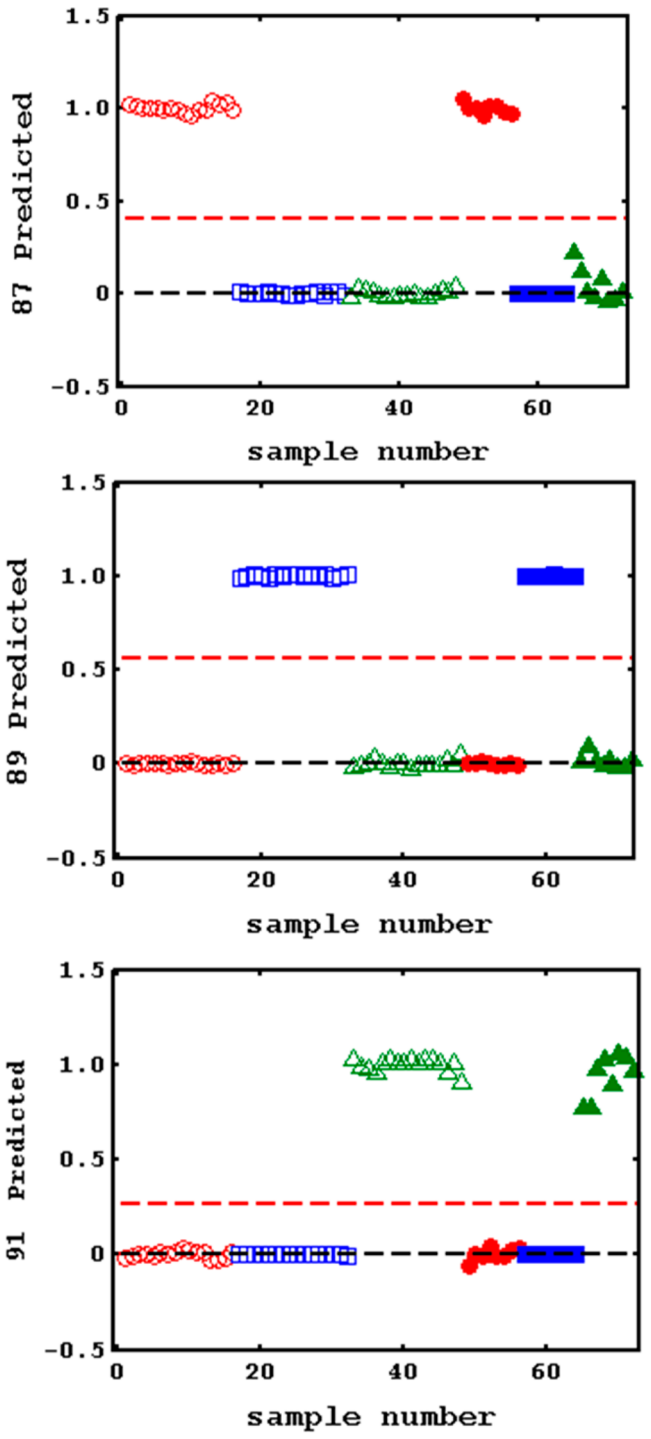


Figure 4. Y-predicted plot for PLS-DA classification of gasoline samples with the application of $UIF_{(2,5)}$ prior to feature selection. Red circles, blue squares, and green triangles indicate 87, 89, and 91 octane ratings of gasoline, respectively. Hollow markers indicate training and optimization while solid markers indicate validation set.

The overall effect of applying $UIF_{(2,5)}$ to a sample region of a chromatogram is shown in Figure 5. The overall reduction in the number of candidate variables is obvious. It is worth noting that the m/z dimension in Figure 5b,c is restricted to that showing the majority of ions. Thus, in some cases where only one um/z is apparent for a given peak in Figure 5, the other um/z is at a m/z value >140.

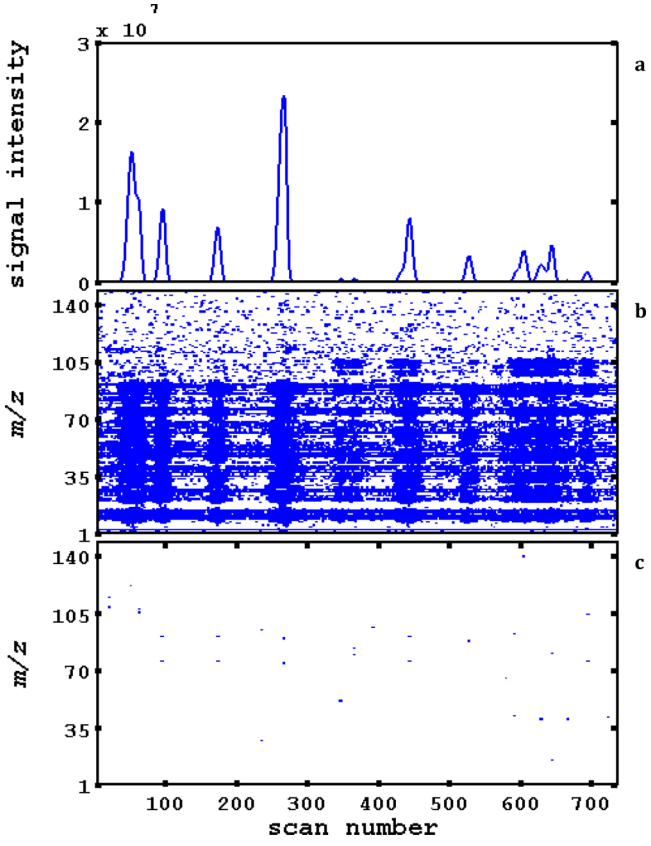


Figure 5. Effect of $UIF_{(2,5)}$ on an example segment of a chromatogram. (a) TIC trace, (b) unfiltered GC/MS data matrix, (c) data matrix after being filtered by $UIF_{(2,5)}$. Blue dots indicate locations of signals greater than 150 count threshold.

Comparing the features selected with and without the application of the UIF, it is apparent that the features correspond largely to the same compounds (Figure S6, Supporting Information). These features have been tentatively identified as 4-methyl heptane, toluene, and an unknown compound. This observation indicates that the use of the UIF does not alter the underlying chemical information in the data.

To demonstrate the need for feature selection, PLS-DA models were generated on the raw chromatograms with no feature selection or filtering. The overall model quality was poor (Figure S7, Supporting Information). UIF was also tested on a more challenging data set. The optimum UIF setting for

dx.doi.org/10.1021/ac501660a | Anal. Chem. 2014, 86, 7726−7733

Table 1. Result of Feature Selection and Model Quality for Selected Conditions

data unfolding		feature selection				model quality
condition	time/sample (s)	total	checked	passed	time/min	accuracy
$UIF_{(2,5)}$	0.56 (0.016)[a]	13 838	500	53	9 (1)[a]	1.00
NO UIF	0.027 (0.002)[a]	1 668 403	500	116	8.4 (0.9)[a]	0.83
NO UIF	0.029 (0.004)[a]	1 668 403	30 000	3001	370 (18)[a]	1.00

[a]Mean and standard deviation at $n = 5$.

this work (i.e., $UIF_{(2,5)}$) was applied to a data set comprising GC/MS chromatograms of casework fire debris samples from a previous study.[14] In this case, features were being selected to permit the identification of gasoline in casework arson data using PLS-DA. A model with similar performance to that found previously was achieved, and the resultant Y-predicted plot is presented in Figure S8, Supporting Information.

Table 1 presents a comparison of the optimum benchmark and UIF conditions. Even though excellent model quality was achieved without the UIF, this required the testing of 30 000 top-ranked variables and prolonged the feature selection process to over 6 h. As expected, data unfolding time when the UIF is applied is slightly longer than for the benchmark algorithm due to the additional computations applied by the UIF. However, the total number of candidate variables was reduced by 2 orders of magnitude over the non-UIF case, and excellent model quality was achieved after testing only 500 variables. This is attributed to the reduction in irrelevant and/or redundant features in the data by the UIF, making it easier for the learning algorithm to focus on the relevant data. Due to this reduction in the variables tested, excellent model prediction accuracies were achieved from the resulting variables when a fewer number of top-ranked variables were tested. This reduced the overall feature selection time to 9 min including application of the UIF. Results in Table 1 also show that, without the use of the UIF, testing only the 500 top-ranked variables led to poorer overall model quality.

CONCLUSIONS

UIF is a novel feature reduction approach for preprocessing of multivariate data. The filter does not require the *a priori* knowledge of the samples being analyzed. Using two major inputs of the number of um/z and the scan window, the algorithm selects unique features that contain the relevant chemical information for each peak, while reducing redundancy in the number of variables considered per peak by at least an order of magnitude. This leads to the reduction in the number of candidate variables for subsequent feature selection and chemometric analysis. Consequently, feature selection time is greatly reduced as is the amount of noise for which the model must account. The reduction in noise results in an overall increase in model quality.

Application of the UIF does not alter the fundamental chemical information in analytical data upon which models are ultimately based. It was also realized that the use of a single m/z or only the peak apex scan does not provide enough information for the classification of the samples we studied. This indicates the need for *some* redundancy in variables of a data set.

With the increase in the use of high data rate mass analyzers, UIF provides an avenue for researchers to reduce the initial number of variables without losing the multivariate nature of the data. It must however be emphasized that UIF also relies on the user having a robust peak detection algorithm.

While UIF was applied to GC/MS data in this study, it can be adapted to other chromatographic data with a multivariate detector (LC/MS, GC/IR, CE/MS, etc.). Readers may contact the corresponding author for more information about the UIF algorithm or a copy of the algorithm.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: +1.780.492.8303. Fax: +1.780.492.9231. E-mail: james.harynuk@ualberta.ca.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The Natural Sciences and Engineering Research Council of Canada (NSERC), Genome Canada, and Genome Alberta are gratefully acknowledged for financial support of this research.

REFERENCES

(1) Wold, S. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 109−115.
(2) Otto, M. *Chemometrics, Statistics and Computer Application in Analytical Chemistry*; Wiley VCH: Weinheim, 1998.
(3) Lavine, B. K.; Workman, J. *Anal. Chem.* **2008**, *80*, 4519−4531.
(4) Lavine, B. K.; Workman, J. *Anal. Chem.* **2013**, *85*, 705−714.
(5) Johnson, K. J.; Synovec, R. E. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 225−237.
(6) Doble, P.; Sandercock, M. L.; Du Pasquier, E.; Petocz, P.; Roux, C.; Dawson, M. *Forensic Sci. Int.* **2003**, *132*, 26−39.
(7) Sandercock, P. M. L.; Du Pasquier, E. *Forensic Sci. Int.* **2003**, *134*, 1−10.
(8) Sandercock, P. M. L.; Du Pasquier, E. *Forensic Sci. Int.* **2004**, *140*, 43−59.
(9) Sandercock, P. M. L.; Du Pasquier, E. *Forensic Sci. Int.* **2004**, *140*, 71−77.
(10) Sinkov, N. A.; Harynuk, J. J. *Talanta* **2011**, *83*, 1079−1087.
(11) Sinkov, N. A.; Harynuk, J. J. *Talanta* **2013**, *103*, 252−259.
(12) Nelson, R. K.; Kile, B. M.; Plata, D. L.; Sylva, S. P.; Xu, L.; Reddy, C. M.; Gaines, R. B.; Frysinger, G. S.; Reichenbach, S. E. *Environ. Forensics* **2006**, *7*, 33−44.
(13) Christensen, J. H.; Tomasi, G. *J. Chromatogr., A* **2007**, *1169*, 1−22.
(14) Sinkov, N. A.; Sandercock, P. M. L.; Harynuk, J. J. *Forensic Sci. Int.* **2014**, *235*, 24−31.
(15) Pizarro, C.; Esteban-Díez, I.; Sáenz-González, C.; González-Sáiz, J. M. *Anal. Chim. Acta* **2008**, *608*, 38−47.
(16) Weldegergis, B. T.; Crouch, A. M. *J. Agric. Food Chem.* **2008**, *56*, 10225−10236.
(17) Ballabio, D.; Skov, T.; Leardi, R.; Bro, R. *J. Chemom.* **2008**, *22*, 457−463.

(18) Li, X.; Xu, Z.; Lu, X.; Yang, X.; Yin, P.; Kong, H.; Yu, Y.; Xu, G. *Anal. Chim. Acta* **2009**, *633*, 257−262.
(19) Beckstrom, A. C.; Humston, E. M.; Snyder, L. R.; Synovec, R. E.; Juul, S. E. *J. Chromatogr., A* **2011**, *1218*, 1899−1906.
(20) Pietracci, E.; Bermejo, A. M.; Álvarez, I.; Cabarcos, P.; Balduini, W.; Tabernero, M.-J. *Forensic Toxicol.* **2012**, *31*, 124−132.
(21) Gad, H. A.; El-Ahmady, S. H.; Abou-Shoer, M. I.; Al-Azizi, M. M. *Phytochem. Anal.* **2013**, *24*, 1−24.
(22) Maree, J.; Kamatou, G.; Gibbons, S.; Viljoen, A.; Van Vuuren, S. *Chemom. Intell. Lab. Syst.* **2014**, *130*, 172−181.
(23) Figueira, J.; Câmara, H.; Pereira, J.; Câmara, J. S. *Food Chem.* **2014**, *145*, 653−663.
(24) Kind, T.; Tolstikov, V.; Fiehn, O.; Weiss, R. H. *Anal. Biochem.* **2007**, *363*, 185−195.
(25) Yang, S.; Nadeau, J. S.; Humston-Fulmer, E. M.; Hoggard, J. C.; Lidstrom, M. E.; Synovec, R. E. *J. Chromatogr., A* **2012**, *1240*, 156−164.
(26) Xu, Y.; Fowler, S. J.; Bayat, A.; Goodacre, R. *Metabolomics* **2013**, *10*, 375−385.
(27) Das, M. K.; Bishwal, S. C.; Das, A.; Dabral, D.; Varshney, A.; Badireddy, V. K.; Nanda, R. *Anal. Chem.* **2014**, *86*, 1229−1237.
(28) Xiong, Y.-H.; Xu, Y.; Yang, L.; Wang, Z.-T. *J. Appl. Toxicol.* **2014**, *34*, 149−157.
(29) Lim, H. K.; Stellingweif, S.; Sisenwine, S.; Chan, K. W. *J. Chromatogr., A* **1999**, *831*, 227−241.
(30) Bro, R.; Smilde, A. K. *J. Chemom.* **2003**, *17*, 16−33.
(31) Kjeldahl, K.; Bro, R. *J. Chemom.* **2010**, *24*, 558−564.
(32) Lavine, B. K. *Anal. Chem.* **1995**, *67*, 3846−3852.
(33) Zekavat, B.; Solouki, T. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 1873−1884.
(34) Sinkov, N. A.; Johnston, B. M.; Sandercock, P. M. L.; Harynuk, J. J. *Anal. Chim. Acta* **2011**, *697*, 8−15.
(35) Guyon, I. *J. Mach. Learn. Res.* **2003**, *3*, 1157−1182.
(36) Rajalahti, T.; Arneberg, R.; Kroksveen, A. C.; Berle, M.; Myhr, K.-M.; Kvalheim, O. M. *Anal. Chem.* **2009**, *81*, 2581−2590.
(37) Khanmohammadi, M.; Bagheri Garmarudi, A.; de la Guardia, M. *Talanta* **2013**, *104*, 128−134.
(38) Wongravee, K.; Heinrich, N.; Holmboe, M.; Schaefer, M. L.; Reed, R. R.; Trevejo, J.; Brereton, R. G. *Anal. Chem.* **2009**, *81*, 5204−5217.
(39) Vieira, S. M.; Sousa, J. M. C.; Kaymak, U. *Fuzzy Sets Syst.* **2012**, *189*, 1−18.
(40) Cadenas, J. M.; Garrido, M. C.; Martínez, R. *Expert Syst. Appl.* **2013**, *40*, 6241−6252.
(41) Duval, B.; Hao, J.-K. *Briefings Bioinf.* **2010**, *11*, 127−141.
(42) Ferreira, A. J.; Figueiredo, M. A. T. *Pattern Recognit.* **2012**, *45*, 3048−3060.
(43) Pierce, K. M.; Hope, J. L.; Johnson, K. J.; Wright, B. W.; Synovec, R. E. *J. Chromatogr., A* **2005**, *1096*, 101−110.
(44) Loong, T. *Br. Med. J.* **2003**, *327*, 716−719.

Supporting Information

Unique Ion Filter – A data reduction tool for GC-MS data preprocessing prior to chemometric analysis

Authors: L. A. Adutwum[1] J. J. Harynuk[1*]

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

*Corresponding author: Department of Chemistry, University of Alberta, Edmonton, Alberta, T6G 2G2 Canada

Phone: +1.780.492.8303. Fax: +1.780.492.9231. Email:james.harynuk@ualberta.ca

Figures:

Tables:

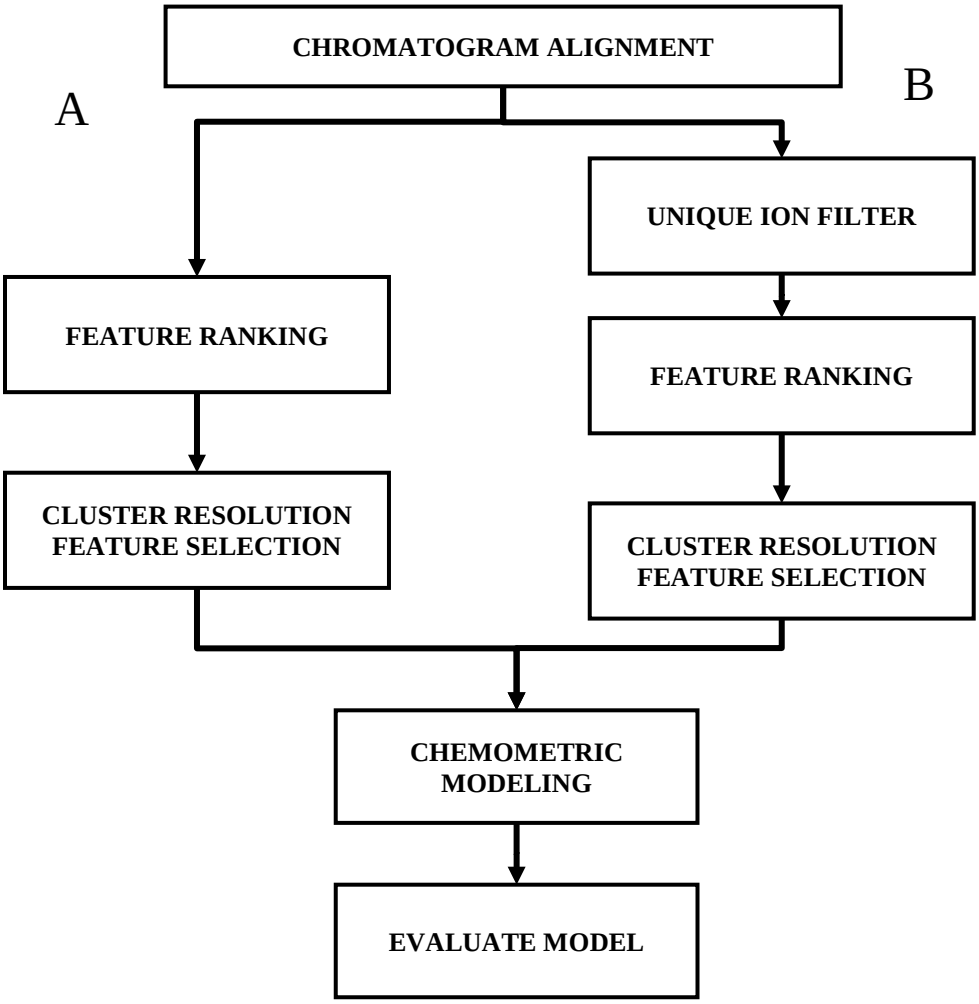


Figure S1. Flow chart showing benchmark pathway (A) and UIF pathway (B)

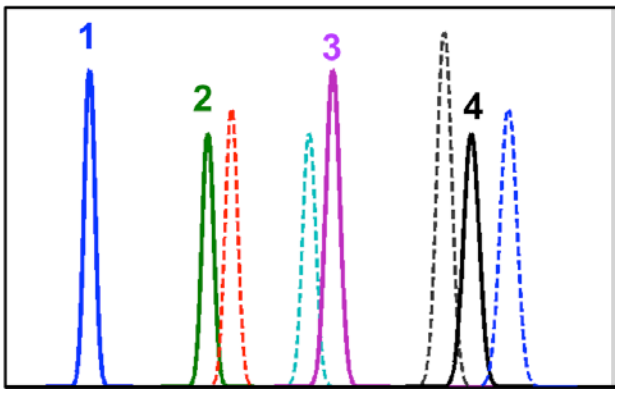


Figure S2. Peak groups for unique ion identification (1 – A, 2 –B1, 3 – B2 and 4 – C)

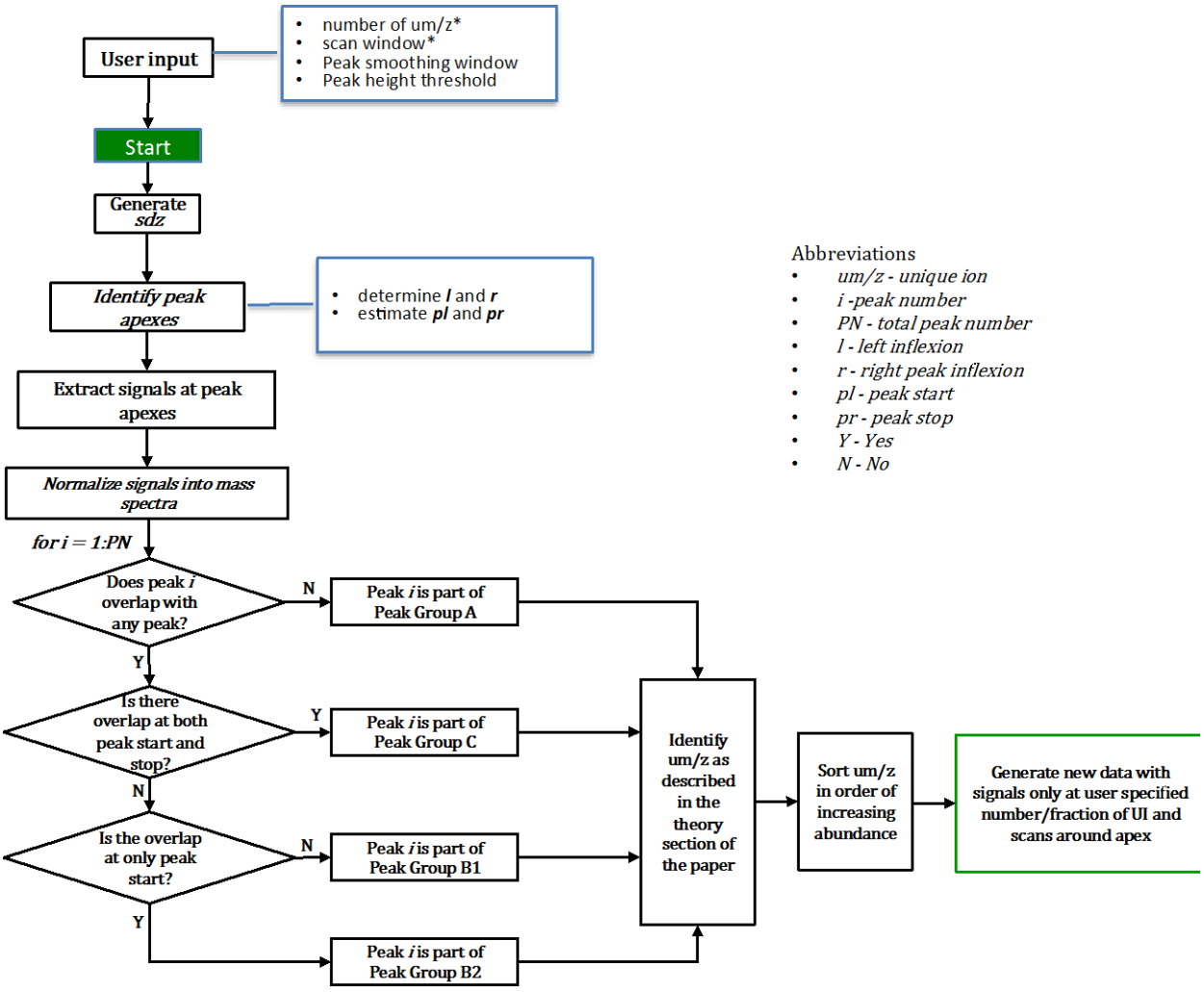


Figure S3. Flow chart for peak detection and UIF application

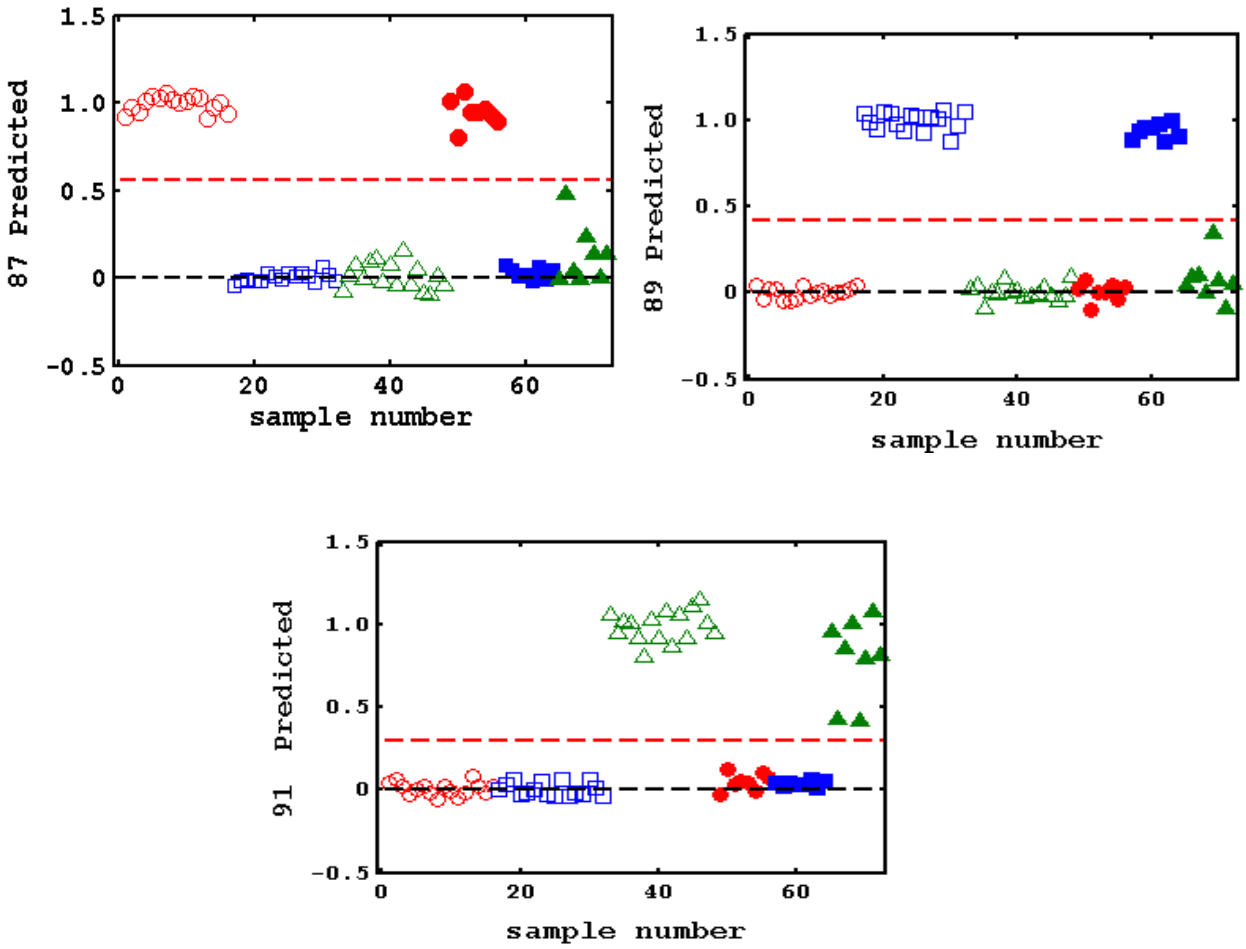


Figure S4. Y-predicted plot for PLS-DA classification of gasoline samples at the minimum UIF conditions that gave excellent model prediction accuracy: UIF $_{(2, 3)}$. Red circles, blue squares and green triangles indicate 87, 89 and 91 octane ratings gasoline, respectively. Hollow markers indicate training and optimization while solid markers indicate validation set.

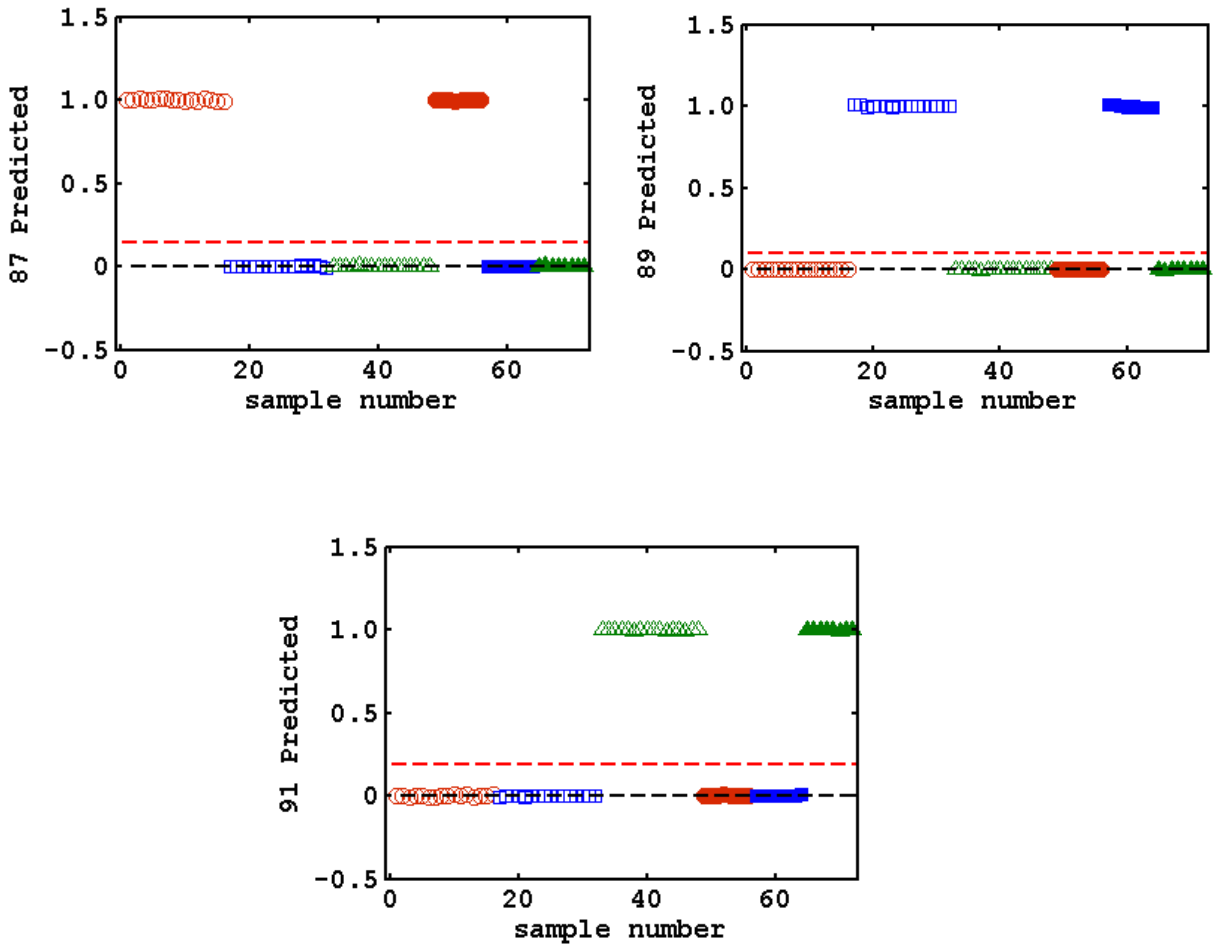


Figure S5. Y-predicted plot for PLS-DA classification of gasoline samples with UIF providing the least number of variables passed when UIF was used: UIF $_{(2, 11)}$. Red circles, blue squares and green triangles indicate 87, 89 and 91 octane ratings gasoline, respectively. Hollow markers indicate training and optimization while solid markers indicate validation set.

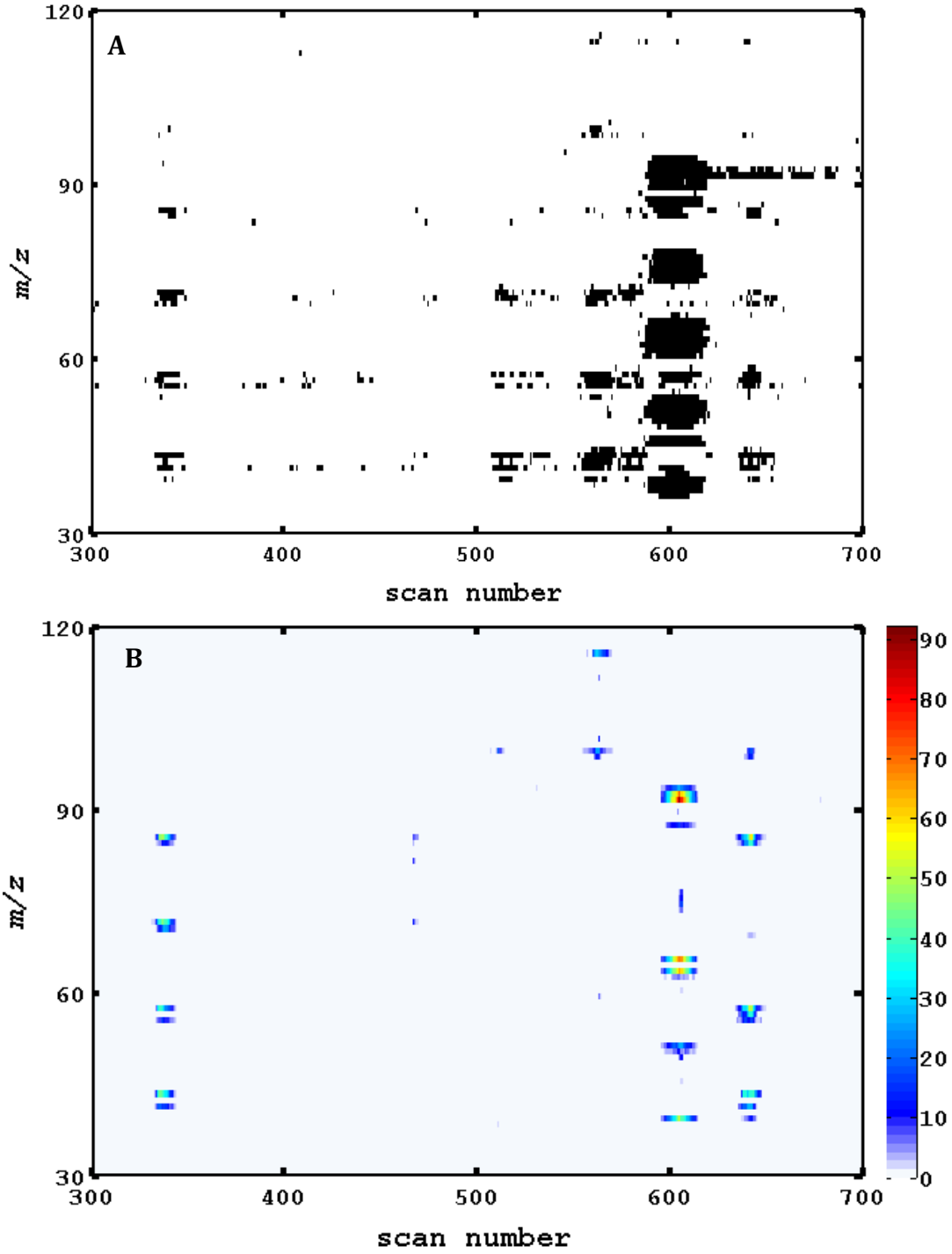


Figure S6. Features selected by the feature selection algorithm without (A) and with (B) the application of UIF. Dark regions in (A) show non-zero variables, color map in (B) shows the number of times each feature was selected for the separate UIF conditions

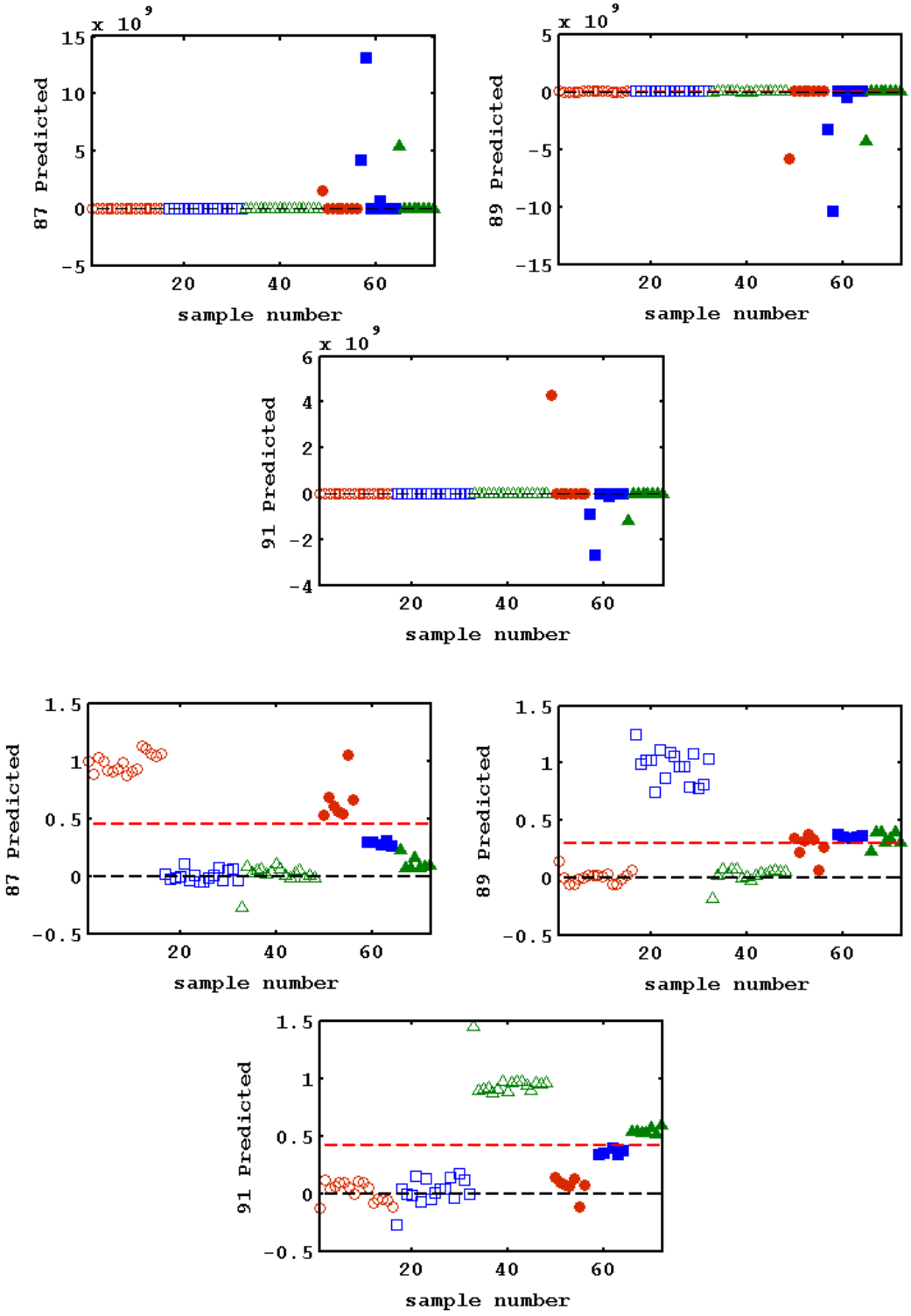


Figure S7. Y-predicted plots for predicting gasoline class using PLS-DA directly on the raw GC-MS data. Top: expanded scale to show all points; Bottom: close-up of -0.5 – 1.5. Red circles, blue squares and green triangles indicate 87, 89 and 91 octane ratings gasoline, respectively. Hollow markers indicate training and optimization while solid markers indicate validation set.

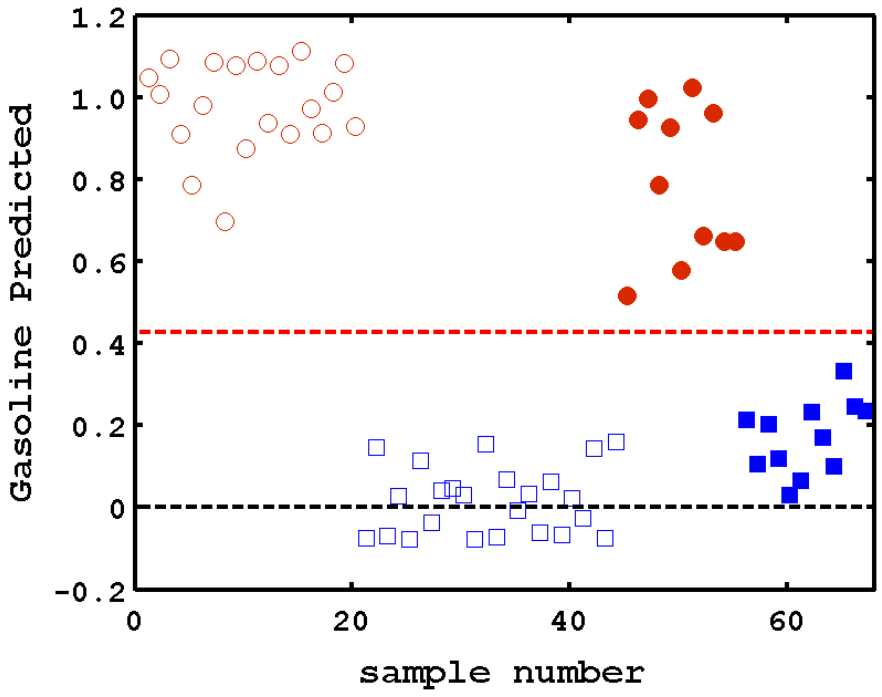


Figure S8. Y-predicted plots for predicting the presence or absence of gasoline in casework fire debris sample from a previous study. Open symbols are training/optimization set; filled symbols are the validation set. Red circles indicate samples containing gasoline; blue squares indicate samples that do not contain gasoline. Experimental details for this data set can be found in Sinkov, N. A.; Sandercock, P. M. L.; Harynuk, J. J. *Forensic Sci. Int.* **2014**, *235*, 24–31.

Table S1. Number of top-ranked variables checked, the number passed and the PLS-DA model quality for the benchmark pathway

Feature Selection Variables		PLS-DA Model Prediction			
Checked	Passed	Sensitivity	Specificity	Accuracy	LV
500	116	0.88	0.92	0.83	6
3000	342	0.88	0.83	0.85	3
5000	565	0.96	0.92	0.94	3
10000	929	1.00	0.94	0.97	3
15000	1425	1.00	0.98	0.99	3
20000	1094	1.00	0.96	0.98	4
30000	3001	1.00	1.00	1.00	5
40000	4101	1.00	1.00	1.00	5

Table S2. Number of variables selected in final model after UIF and feature selection for different combinations of *p* and *w* in UIF tested

um/z / scans	1	2	3	4	5	6	7	8	9	10
1	139	74	85	58	73	68	76	106	104	96
3	88	53	46	61	66	63	34	113	159	125
5	37	51	23	96	100	100	100	61	45	141
7	82	48	22	100	100	100	100	200	200	155
9	50	25	11	100	100	100	100	200	200	135
11	46	5	100	100	100	100	101	200	199	140
13	36	98	100	100	100	101	100	200	200	152
15	36	100	100	100	100	78	104	200	60	148
17	30	99	100	99	92	98	104	180	128	128

Table S3. PLS-DA model sensitivity for various UIF filter combinations

um/z / scans	1	2	3	4	5	6	7	8	9	10
1	0.83	0.75	0.75	0.79	0.83	0.79	0.83	0.75	0.75	0.88
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00
9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00
11	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00
13	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00
15	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
17	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.92

Table S4. PLS-DA model specificity for various UIF filter combinations

um/z / scans	1	2	3	4	5	6	7	8	9	10
1	0.83	0.79	0.92	0.90	0.81	0.88	0.83	0.90	0.85	0.90
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00
9	1.00	0.98	0.96	1.00	1.00	1.00	1.00	1.00	0.98	1.00
11	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00
13	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98
15	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
17	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.96

Table S5. PLS-DA model accuracy for various UIF filter combinations

um/z / scans	1	2	3	4	5	6	7	8	9	10
1	0.83	0.77	0.83	0.84	0.82	0.83	0.83	0.82	0.80	0.89
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00
9	1.00	0.99	0.98	1.00	1.00	1.00	1.00	1.00	0.97	1.00
11	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00
13	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.99
15	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
17	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.94